Bulletin of the Technical Community on

Data Engineering

December 2023 Vol. 47 No. 4

IEEE Computer Society

Letters

Opinions

Special Issue on Personal Information Management

Conference and Journal Notices

Editorial Board

The TC on Data Engineering

Membership in the TC on Data Engineering is open to all current members of the IEEE Computer Society who are interested in database systems. The TCDE web page is http://tab.computer.org/tcde/index.html.

The Data Engineering Bulletin

The Bulletin of the Technical Community on Data Engineering is published quarterly and is distributed to all TC members. Its scope includes the design, implementation, modelling, theory and application of database systems and their technology.

Letters, conference information, and news should be sent to the Editor-in-Chief. Papers for each issue are solicited by and should be sent to the Associate Editor responsible for the issue.

Opinions expressed in contributions are those of the authors and do not necessarily reflect the positions of the TC on Data Engineering, the IEEE Computer Society, or the authors' organizations.

The Data Engineering Bulletin web site is at http://tab.computer.org/tcde/bull_about.html.

TCDE Executive Committee

Letter from the Editor-in-Chief

In this December edition of the Data Engineering Bulletin, we revisit the topic of personalization after more than a decade. Our renewed focus comes at a time when swift technological advances are poised to elevate personalization to a new level. This issue of the Bulletin, guided by the expert curation of Luna Dong, Alon Halevy, and Shane Moon, not only charts the progress in technologies that promise enhanced personal experiences but also confronts the new privacy challenges they bring. The associate editors provide a comprehensive view of the new personalization landscape, underlining the necessity for robust privacy frameworks to manage personal information effectively. Building on this foundation, I would like to expand the conversation with my reflections on industry and business implications, especially in the context of e-commerce.

It is crucial for e-commerce businesses to understand their customers' needs. For many years, e-commerce platforms have maintained detailed user profiles, tracked customers' browsing patterns and purchasing histories, and used the data to make personalized recommendations. How successful has it been?

Elle Hunt's recent Guardian article [1] offers a critical view of the current personalization practice. Hunt notes, "*Every tech company from Monzo to my bank is crunching my data. All the results tell us is how dull it is to reduce human experience to numbers.*" This statement highlights a major limitation of today's personalization technologies.

While companies like Spotify are adept at aggregating quantitative data, they fail to capture the qualitative, emotional aspects of customer experiences. Indeed, reducing complex human activities to stark statistics strips away the nuances that give human experiences meaning. More specifically, simply tracking the number of hours spent listening to Taylor Swift, or the frequency a particular dish is ordered through DoorDash, doesn't reveal the motivations or feelings behind these actions.

However, we might be surprised just how soon multimodal and generative AI technologies could revolutionize the field of personalization. I believe a digital personal assistant, fueled by these advanced technologies, has the potential to elevate personalization to new, unprecedented heights. Instead of tracking basic statistics, the personal assistant, equipped with multimodal generative AI technologies, will have the capability to process a variety of sensory inputs, including visual and auditory cues. It's easy to envision an assistant capable of assessing a customer's reactions through facial recognition and tone analysis during virtual try-ons and interactive product consultations, thereby obtaining a deep, empathetic understanding of the individual's preferences and needs.

In contrast to existing personalization methods that are confined to singular platforms and only analyze user behavior in isolated contexts, this personal assistant has the potential to track an individual's activities across a multitude of platforms. By being a constant in the individual's life, it will gain a nuanced understanding of the individual's actions and motivations, achieving a level of insight that traditional data analysis methods cannot match. It might even come to know the individual more intimately than they know themselves. With its profound understanding, the personal assistant will help the individual in engaging with business platforms, in a more clear and effective way than the individual could alone. This holistic approach to personalization is poised to create a more intuitive and emotionally engaging shopping experience.

However, while the advancements in personalization technology offer many benefits, they also tread a delicate line between enhancing user experience and encroaching on personal privacy. As we revel in the marvels of such technology, caution is warranted. While contemplating a personal assistant capable of understanding a customer's sentiment and emotion, I can't help thinking about lifelogging [3]. Lifelogging refers to a comprehensive recording of a person's daily life, often aided by wearable technology or mobile devices. This practice has evolved significantly, with projects like DARPA LifeLog contributing to its development, and contemporary apps like Foursquare and Swarm allowing users to document their lives on their platforms. In 2022, Jennifer Egan, an award winning author, explored a related concept in her novel "The Candy House" [4]. She imagined a world where technology has advanced to the point that individuals can upload their memories to a shared database. The novel suggests a kind of collective consciousness, where the boundaries between personal and shared experiences become blurred.

If we feed lifelogging data to multimodal and generative AI, which are capable of analyzing data in various forms (text, images, videos, sensor data), we can gain a deeper, more holistic understanding of an individual's habits, preferences, and needs. Technologies will not only be able to predict an individual's future behaviors, but also act as a proxy of the individual, interacting with the world on her behalf.

Such advancements in personalization and AI's understanding of individuals raise significant ethical issues. This situation echoes the ethical concerns similar to those confronting social networks, which have been criticized for manipulating users' emotions and fostering addictive behaviors to enhance user engagement and profitability. As businesses gain their power to understand their customers, they may use the power not just to *predict* but to *change* customers' behavior in order to boost sales and ad revenue. As playwright Ayad Akhta pointed out in his essay titled "The Singularity is Here" [2] on The Atlantic, "*Our affinities are increasingly no longer our own, but rather are selected for us for the purpose of automated economic gain.*"

Therefore, as e-commerce platforms harness the power of generative AI to better understand their customers, there's an imperative to balance commercial goals with ethical considerations. The potential to use these advancements for positive outcomes, such as promoting health, personal growth, and societal benefits, should be weighed against the commercial opportunities they present, ensuring a responsible and human-centric approach to technology application.

References

[1] Elle Hunt. (2023, December 28). So, Spotify knows how many hours I spent listening to Taylor Swift. But only I know why. *The Guardian*. Retrieved from `https://www.theguardian.com/commentisfree/2023/dec/28/spotify-wrapped-monzo-analyses-meaningless`

[2] Ayad Akhtar. (2021, November 5). The Singularity Is Here. *The Atlantic*. Retrieved from `https://www.theatlantic.com/magazine/archive/2021/12/ai-ad-technology-singularity/620521/`

[3] Wikipedia contributors. Lifelog. Retrieved from `https://en.wikipedia.org/wiki/Lifelog`

[4] Jennifer Egan. (2022, April). The Candy House: A Novel. *Scribner*

Haixun Wang
Instacart

Letter from the Special Issue Editors

How cool would it be if we can record and recall every moment in our lives? Then we will never worry about remembering the answers to questions such as “Where did I put my keys?”, “I had some very good Korean hotpot a while back but which restaurant was that?”, or “Does this skirt have higher quality and lower price than the similar-looking one I saw at Macy’s yesterday?” Asking such questions has been a dream for decades and dates back to Vannevar Bush’s MEMEX (MEMory & EXpansion) vision in 1945. Now with the rapid advancements in AI technology, especially in 2023, the Year of AI, are we getting closer to the dream of capturing and organizing personal information smoothly and accessing them effortlessly?

To realize this dream, we face four challenges. First and foremost, we need to be able to **capture personal information**. Scopewise, do we capture mainly digital footprints of a user (messages, browsing and purchase history, and app usage logs), or also physical aspects of a user (what she sees and experiences, and the time and location of those) through special devices? For the latter, what should be the frequency of capturing of data so it answers our needs? Second, we need to be able to **integrate personal data**, often from different sources, organize them in a meaningful way and find a way to store the potentially large volume of captured data. Understanding of the semantics and effective aggregation are the keys, but achieving them is non-trivial. Third, we shall be able to effectively **leverage personal data** to make life easier or memorable for the owner in future; afterall, the data is worthless unless we can effectively use it. The usage includes helping users easily retrieve past memory, and offering recommendations of books, articles, and products that they may enjoy. Solution to this challenge also decides the solutions to previous challenges, such as how we would like to present and store the data, how much data we really need to capture, and whether we shall have an aging mechanism to ensure efficient usage of the past data and limit its effect on future recommendations. Last but not least, **privacy** guarantees are a must-have for trustworthy personal information management, to use it for good, not for harm.

This issue collects a set of papers around personal information management shedding lights on how to address the aforementioned challenges. We start with a paper by **Tran et al.**, describing lifelogging of user activities from both online and offline in the real world, and sharing the learnings from the annual Lifelog Search Challenge Benchmarking Workshops. The logged data include both our digital footprints, and information collected by devices and sensors to track heart rate, sleep quality, locations, and so on. In the second paper, **Kalokyri et al.** focused on digital data and described the YourDigtalSelf project. The project collects users’ digital traces from different applications, integrates them, reconstructs meaningful episodes from them, and supports memory recall using the episodes. The next two papers focus on particular domains and discuss how to build a personal knowledge graph to support applications: **Chakraborty et al.** described Personal Research Knowledge Graph to organize important information for scientists, such as grants, lab equipments, papers, journal and conferences; **Nidhi et al.** described G-VARS for gun violence susceptibility analysis, through information collected on behavior, mental health, access to firearms, social networks and online activities. We next come to how to apply personal information in applications. **Sun et al.** described QALinkPlus, a novel graph-based method that constructs an entity co-occurrence graph derived from QA datasets, describes context QA-specific subgraphs, and leverages this to improve QA, especially personalized QA. The last paper in this issue is an insightful survey User Modeling in the Era of LLMs, where **Tan and Meng** surveyed user modeling leveraging LLM techniques, and pointed our new research directions for personalized search and recommendation.

Overall, the above papers represent an interesting sample of the ongoing work on the new trend of personal information management. We hope that this special issue will further help and inspire the research community in its quest to solve this challenging problem. We would like to thank all the authors for their valuable contributions, as well as Haixun Wang for giving us the opportunity to put together this special issue, and Nurendra Choudhary for his help in its publication.

Xin Luna Dong*, Alon Halevy†, Shane Moon*
*Meta (Facebook),†Amazon

A Summary of ICDE 2022 Research Session Panels

Zhifeng Bao, Panagiotis Bouros, Reynold Cheng, Byron Choi, Anton Dignös, Wei Ding,Yixiang Fang, Boyang Han, Jilin Hu, Arijit Khan, Wenqing Lin, Xuemin Lin, Cheng Long, Nikos Mamoulis, Jian Pei, Matthias Renz, Shashi Shekhar, Jieming Shi, Eleni Tzirita Zacharatou, Sibo Wang, Xiao Wang, Xue Wang, Raymond Chi-Wing Wong, Da Yan, Xifeng Yan, Bin Yang, Dezhong Yao,Ce Zhang, Peilin Zhao, Rong Zhu

Abstract

In the 38th IEEE International Conference on Data Engineering (ICDE), 2022, panel discussions were introduced after paper presentations to facilitate in-depth exploration of research topics and encourage participation. These discussions, enriched by diverse perspectives from experts and active audience involvement, provided fresh insights and a broader understanding of each topic. The introduction of panel discussions exceeded expectations, attracting a larger number of participants to the virtual sessions.

This article summarizes the virtual panels held during ICDE'22, focusing on sessions such as Data Mining and Knowledge Discovery, Federated Learning, Graph Data Management, Graph Neural Networks, Spatial and Temporal Data Management, and Spatial and Temporal Data Mining. By showcasing the success of panel discussions in generating inspiring discussions and promoting participation, this article aims to benefit the data engineering community, providing a valuable resource for researchers and suggesting a compelling format of holding research sessions for future conferences.

1 Introduction

The 38th IEEE International Conference on Data engineering (ICDE) was hosted in Kuala Lumpur, Malaysia from May 9to May 12 virtually in 2022. To provide a platform for experts and thought leaders from different backgrounds to explore different research topics thoroughly and promote participation, panel discussion was conducted for each session after paper presentation in ICDE'22. The sharing of diverse perspectives from experts and active audience participation enriched the discussion, providing a broader understanding of each topic and offering fresh insights that are beneficial for future research directions. The introduction of panel discussions for research paper presentation sessions also attracted more participants than expected for each virtual session.

Since the panel discussion introduced in ICDE'22 is quite successful in both generating inspiring discussions and promoting participation of research sessions, it would benefit the data engineering community by summarising the discussions of the panels of ICDE'22, and this format of conducting research sessions would be of interest to future conferences. This article summarised the virtual panels held for some interesting sessions in ICDE'22. The organizations of this paper is as follow: Section 2 and 3 summarises the Data Mining and Knowledge Discovery session while Section 4 talks about the Federated Learning session. Graph Data Management session and Graph Neural Network session are discussed in Section 5 and Section 6. Section 7 and Section 8 summarise the Spatial Temporal Data Management session and Spatial and Temporal Data Mining session.

2 Data Mining and Knowledge Discovery

This section is a summary of the "data mining and knowledge discovery" virtual panel held at ICDE 2022 on May 10th, 2022. The panelists were listed in alphabetical order as follows:

- Wei Ding (University of Massachusetts, Boston)
- Xuemin Lin (The University of New South Wales, Sydney)
- Jian Pei (Simon Fraser University, Vancouver)

- Shashi Shekhar (University of Minnesota, Minneapolis-St. Paul)
- Xifeng Yan (University of California, Santa Barbara)

The goal of this panel was to gather the world-renowned data mining experts to discuss the achievements and future directions of data mining, which centered around the following questions:

2.1 What are the achievements of data mining in the past 10 years?

The panelists initiated the panel discussion by showing some statistics on publications between 2012 and 2021 that they obtained by searching three keyword phrases "machine learning," "data mining" and "deep learning," for which we use shorthand notations ML, DM and DL hereafter. The findings were that in terms of the number of publications, (1) ML and DL started lower than DM, but now DM lags behind; (2) the United States led till 2019, broke a tie with China in 2020, and China led in 2021; (3) the top-5 most-published institutions in 2021 are University of Chinese Academy of Sciences, Harvard University, Tsinghua University, Shanghai Jiao Tong University, and Stanford University. It is clear that China has made great progress in the last 10 years, but in terms of the number of citations, Stanford University and Harvard University still lead the other institutions by a large margin, indicating higher research impacts.

The panelists also commented that DM has made significant accomplishments in that (1) DM itself has grown into a well-established area in the past 20 years, rather than a sub-field of another area such as database or machine learning; and that (2) DM has become a tool that people would think of for solution when facing real data and real applications. The panelists considered graph mining as one of the fastest-developing fields in DM, witnessing the invention of newer and newer techniques in the past decade such as graph pattern mining, graph embedding, graph neural networks, and the development of knowledge graph as an important sub-field, enabling many real applications such as chemical compound design.

Furthermore, the panelists highlighted spatial data mining as an important achievement in data mining. Spatial data mining is important due to the rise in spatial big data (e.g., remote sensing, census, maps) and important applications such as smart city, climate change, environment, and social good. However, traditional data mining methods face severe challenges in this context. For example, DBSCAN produces spurious patterns in the presence of noise, spatial association rules are unstable due to the modifiable areal unit problem and prediction methods exhibit lower accuracy and spatial bias. To overcome the limitations of one-size-fit-all methods, spatial data mining has provided newer methods (e.g., colocations, significant DBSCAN, spatial autoregression, spatial decision trees, spatial-variability aware neural networks) using the notions of neighbor graphs, spatial autocorrelation, statistical significance testing, etc.

Lastly, the panelists also ranked text mining as an important field in DM that has demonstrated significant progress.

2.2 What are the challenges of data mining?

The panelists pointed out a few challenges for data mining methods. A recent AAAS Science magazine article titled "Taught to Test" [1] said that the current AI (e.g., DL, ML, DM) models perform well on benchmarks but do not generalize well to other (out of sample) datasets. They may embarrass us in real applications due to overfitting to biased benchmarks. For example, the ImageNet benchmark is receiving a lot of negative press because of its racial and gender bias. DL has shown good results in engineering such as solving differential equations, but has much room to improve in other applications such as self-driving cars, where their perception and sensor suites are overwhelmed by adverse weather (e.g., rain, snow, dust). This holds similarly to other conventional DM algorithms. For example, many papers have been published on spatial association rules, but their results are unstable [2], varying dramatically across different choices for space partitioning for a given geospatial dataset. As another example, the DBSCAN clustering algorithm may generate spurious results when there is background noise, the consequence of which can be very costly in applications such as finding crime or disease hotspots. Furthermore, traditional prediction methods often assume that the data samples are drawn independently of each other and from identical distributions. These assumptions are violated by spatial data, which often exhibit spatial autocorrelation and spatial variability.

Most of the panelists agreed that the number of publications, citations and h-indexes are imperfect measures of impact. For example, based on these metrics, Alan Turing (and about half of the Turing Award winners [3]) would not make it in the list of top 1000 computer scientists. These metrics are not fair to people working in industry and government laboratories, which may prefer patents, products and policy innovations over conference publications. Even for academics, these metrics are not normalized for community size and are not fair to pioneering work in less-crowded but highly important fields. Furthermore, these metrics do not distinguish between self-citation, community citations, conference paper citations, and journal paper citations. Thus, they emphasized the importance of evaluating a DM research based on its transformative impact whether foundational or societal. For example, social-media (e.g., Facebook) recommender algorithms are effective in catching people's attention, but they help spread misinformation and polarize the society, which is undesirable. There is a new trend of responsible computing, and DM researchers may contribute by doing DM research for social good, on topics such as climate change and fairness in AI.

2.3 What is the role of deep learning in data mining research?

Deep Learning (DL) is one of the fastest-developing areas in the past decade and half, so it would be interesting to consider its impacts versus conventional data mining techniques. The panelists pointed out that an obstacle of applying DL models in real applications is that they give blackbox predictions, while people need explainable rules for decision making. In fact, a recent work from Dr. Xifeng Yan's group, BERTRL (BERT-based Relational Learning), showed that association rules can outperform graph neural networks, and it is a promising solution to combine DL with association rule mining where the former filters out unrelated patterns so that the latter can mine the most important rules. In text mining, concepts such as language models and n-grams can all be regarded as a kind of association rules, except that conventional DM methods require explicit definition of association, while DL methods implicitly encode rules in the networks using embeddings. In principle, we can apply any methods as long as we are able to solve problems for users by delivering explainable rules.

The panelists further commented that DM is different from ML in that DM pays more attention to knowledge discovery from data, and to the communication between data owners, data users and domain experts.

Lastly, the panelists suggested that DL and DM are complementary. For example, DL may pre-process image or video data to identify objects, their types (e.g., vehicles) and trajectories, which are then mined by (spatial) DM techniques to identify patterns such as hotspots and their proximal and distant correlates (e.g., colocations, teleconnections), etc. In other words, DL may extend the reach of DM to image and video data. In addition, DL may be used to construct new features to improve performance of DM classifiers.

2.4 What are the promising future directions of data mining research?

Dr. Wei Ding is currently serving as a program director of National Science Foundation (NSF), and based on her experience, several directions are seeing growing interest and encouraged by NSF, including (1) DM research for Medical and Health Sciences, (2) Fairness in AI such as NSF's FAI program, (3) AI-enabled scientific discovery, and (4) software and hardware codesign for AI research targeting extreme scalability. NSF is interested in a broader range of novel machine learning topics and areas, not just DL; in addition, collaborations with domain scientists and industry partners are important, rather than research works that are only verified on toy datasets.

The panelists also considered text mining as a promising direction, since text data is everywhere and thus easy to obtain for knowledge discovery, and there have been great breakthroughs in NLP such as transformer models like the powerful GPT-3. The panelists envisioned that a knowledge-grounded language model would be an important direction in text mining research, which encourages the convergence of language models, knowledge representation, and rule-based reasoning.

The panelists highlighted the promise of spatiotemporal data mining due to the rise in valuable spatiotemporal big data (e.g., smartphone trajectories, vehicle on-board-diagnostic data, daily scan of Earth from nano-satellites), foundational DM challenges (e.g., temporal non-stationarity, missing data) and socially important applications such as climate change (e.g., forecast sea-level rise and impact), public safety (e.g., monitoring forest fires, flood prediction), public health (e.g., identifying emerging hotspots and predict spread of infectious diseases), and understanding spatiotemporal patterns of lives.

3 Data Mining and Knowledge Discovery 3

The “Data Mining and Knowledge Discovery 3” track includes 12 interesting papers with topics ranging from recommender systems, intent mining, trajectory and time series analytics, rule learning, forecasting, planning, and IoT. A panel discussion follows the paper presentation, where the panelists consist of experts from both academia and industry, including Peilin Zhao (Tencent AI Lab), Ce Zhang (ETH Zurich), Xue Wang (Alibaba DAMO Academy), Jieming Shi (The Hong Kong Polytechnic University), and Boyang Han (JD Intelligent Cities Research). The panel experts share their visions on future research directions in data mining and knowledge discovery and also outline potential application areas that may significantly benefit from such research, summarized as follows.

First, many industrial sectors are undergoing digitalization transformation, providing plenty of unique opportunities for data mining and knowledge discovery (DMKD) research. Novel DMKD techniques will play a more significant role during the transformation. For example, DMKD techniques are now playing a significant role in drug discovery, especially in the stages before pre-clinical trials. In addition, being able to extract data from knowledge is essential for enhancing current intelligent city management, where knowledge fusion, reusing, and transferring are key techniques.

Second, compared to inventing new and better DMKD techniques, how to ensure DMKD techniques accessible and affordable to end users is also important. For instance, time series exist in a wide variety of application domains, and different types of time series analytics models also exist. It can be challenging for domain users to make the right decisions on the analytics models. Thus, automated processes for deploying the most appropriate DMKD techniques for different application domains are called for. In addition, means to reduce computational cost and operational cost are also important.

We believe that data mining and knowledge discovery will continue to be a very hot research area in academia and we will also face many challenging issues when deploying the research results in real world industrial settings, which will inspire further innovation. Thus, we expect to see more and closer collaborations between academia and industry.

4 Federated Learning

In ICDE 2022, the “Federated Learning" track includes 16 interesting research papers. Recently, federated learning (FL) has raised significant attention in both academic and industrial communities [4–6]. In contrast to traditional training paradigm, federated learning is a distributed model training paradigm that enables learning private data knowledge by communicating local model updates rather than gathering the raw data. Serving as an efficient learning scheme for communication and privacy protection, FL has shown its potential to facilitate real-world applications, such as pre-training [7], object detection [8], bio-metrics [9], medical image analysis [10], healthcare [11, 12], finance [13], smart manufacturing [4], and others.

Although FL has demonstrated empirical success in handling the system heterogeneity and data heterogeneity challenges, and there are still many key open issues need to be explored.

- *Federated online learning:* As artificial intelligence IoT (AIoT) devices have been widely deployed in our daily life, a large-scale streaming data needs to be analysis. Most of the existing FL algorithms performs well on statistic analysis. How do design an efficient FL to support online learning is becomes a crucial challenge in AIoT domain.
- *Incentive mechanism design:* The performance of FL usually benefits from more participants joining the training. How to encourage more devices to provide the trained local model is a very interesting direction.
- *Fairness in Federated Learning:* The performance of the global model on each local device are usually different. How to train a fair classifier on decentralized data is very important for each local client.
- *Block-chain in Federated Learning:* As FL suffers from shortcomings such as single-point-failure and malicious data, block-chain provides a secure and efficient solution for the deployment of FL. Currently, a block-chain based federated Learning system in used in Web 3.0 scenarios.

Besides the federated learning, federated database systems are also caught lots of attention by the privacy-preserving features [14, 15]. Since the constituent database systems remain autonomous, a federated database system is a contrastable alternative to the task of merging several disparate databases [16]. The federated database system is basically used when there is some global view or schema of the federation of the database which is basically shared by the applications. The heterogeneity issues of federated database system are differences in data model and data conflicts.

5 Graph Data Management

In ICDE 2022 Graph Data Management session, eleven papers about graph management and analysis were presented, covering a wide range of topics, including: distributed graph analytics, knowledge graphs, graph pattern and isomorphism, influence maximization, and graph learning. Moreover, we have invited four experts in this area:

- Prof. Byron Choi (Hong Kong Baptist University);
- Prof. Arijit Khan (Aalborg University, Denmark);
- Prof. Sibo Wang (Chinese University of Hong Kong); and
- Prof. Yixiang Fang (Chinese University of Hong Kong Shenzhen)

They share their invaluable views on the following two questions below:

5.1 What makes you feel interested/excited in working in this area?

The experts think that graph problems are interesting because they enable the integration of theories and practice. Byron said that there are lot of computational challenges, in particular the need to provide easy-to-use querying tools. Arijit said that graphs are inherent in many areas and systems in the industry, and each of them have domain-specific problems. Sibo explained that it is interesting to develop practial linear-time or sublinear solutions for large graphs, as well as studying multi-core and distributed computing paradigm. Yixiang likes graph problems because the solutions can be used in real graph data.

5.2 What is the most important problem in big graph data in the next 5-10 years?

Byron pointed out that given the variety of graph databases, there is lack of standard for graph query interfaces, so it is crucial to provide standardization for query languages and interfaces. It is also important to develop intuitive querying tools to explore graph data. Arijit followed up by pointing out that graph systems should be designed to allow users to understand graph query answers, and provide feedback to the system, to allow interaction between users and systems. In Sibo's view, the next problem is machine learning on graphs, which involves defining abstractions and atomic operations for graph learning. This can only be done by leveraging expertise in graph algorithms, graph learning, and graph systems. Yixiang pointed out that the next problems involves the study of the building and use of complex networks (e.g., knowledge graphs and multi-graphs) in downstream applications.

There are also active discussions among the audience, raising questions about the user-friendliness problem of existing graph query languages such as SPARQL. Experts responded by explaining that NLP and keyword search may provide better graph management and visualization tools. It is also important to be able to provide explanation to the query answers returned. Among the audience, Prof. Ashraf Aboulnaga said that their group has developed a programmer-friendly option to access RDF graphs, called *RDFFrames* [17]. To conclude, the session gives a nice overview of the latest development of graph management. We would also like to thank the experts for their effort for suggesting their visions about the important and emerging research problems in the next few years.

6 Graph Neural Networks

In ICDE 2022 Graph Nerual Networks session, a number of papers on graph neural networks (GNNs) were presented, ranging from model design, analysis and applications of GNNs. Moreover, five panel experts in this area extensively discussed the research on GNNs.

- Dr. Yixiang Fang (Session chair, The Chinese University of Hong Kong, Shenzhen);
- Dr. Xiao Wang (Beijing University of Posts and Telecommunications);
- Dr. Rong Zhu (Alibaba DAMO Academy);
- Dr. Jieming Shi (The Hong Kong Polytechnic University);
- Dr. Wenqing Lin (Tencent).

In particular, they have extensively discussed the following five future research directions on GNN:

6.1 Theoretical foundation of GNNs

Unlike traditional graph mining models, GNNs provide an automatic paradigm to summarize structure and label information of graph data into hidden representations. However, the boundary of the expressive power of GNNs is still unknown. Existing works try to analyze GNNs' expressive power using Weisfeiler-Lehman isomorphism test. The 1-hop neighbor based GNNs (e.g., GraphSage, GAT and GCN) cannot do many simple tasks (e.g., triangle counting). It is still unclear whether complex forms of GNNs could mitigate this gap. Only when we develop more powerful tools to analyze the capacity boundary of GNNs, we can better know the strengths and weaknesses of GNNs compared to previous models.

6.2 Interpretability of GNNs

How to explain the learned embedding vectors of GNNs is another challenging task. Unlike traditional models, GNNs cannot tell which motif and/or which label plays important roles in the resulting knowledge. This prevents GNNs to be used in some scenarios that are with serious effect, such as medicine and financial applications. Therefore, it is crucial to understand the underlying mechanisms of the models in human terms.

6.3 Killer applications of GNNs

Although GNN has received much attention recently, we have not witnessed any killer application that must rely on GNNs. Unlike PageRank which lays the foundation of web page search engine, GNN has not proved its importance in such a widely application scenario. We admit GNNs have advantages in automatic learning, but still lack killer applications. Thus, we should calm down to think deeper on the importance and roles of GNNs in more real applications.

6.4 Efficiency issues of GNNs

Nowadays, big graphs are prevalent in various areas. Nevertheless, most of existing GNN models cannot process large graphs at scale. Consequently, how to build GNN models that achieve both high efficiency and strong scalability without sacrificing accuracy on large graphs is an important future research direction.

6.5 Trustworthy GNNs

Current GNNs still mainly focus on the performance improvement. However, when we deploy GNNs to the real world scenarios, especially some risk-sensitive areas, the accuracy is not the only metric to evaluate the GNNs. Whether the GNNs are trustworthy is an important factor. For example, how to generalize the GNNs to the out-of-distribution graphs, i.e., how can we ensure that the performance of GNNs keeps stable when the distribution of test graphs changes. Besides, the robustness and fairness of GNNs are also very important in real applications.

In addition, many other related issues of GNN, including general frameworks, auto-search of parameters, benchmarks, and platforms/systems, have also been briefly discussed.

7 Spatial and Temporal Data Management

The 38th IEEE International Conference on Data Engineering (ICDE 2022) was held in May 9-12, 2022 as a virtual event. ICDE2022 Session “Spatial and Temporal Data Management” was chaired by Nikos Mamoulis and ran between 14:30 and 16:10 Malay time. The panelists include six experts in this area:

- Dr. Panagiotis Bouros (Johannes Gutenberg University)
- Dr. Anton Dignös (Free University of Bozen-Bolzano)
- Dr. Nikos Mamoulis (University of Ioannina)
- Dr. Matthias Renz (CAU University of Kiel)
- Dr. Raymond Chi-Wing Wong (The Hong Kong University of Science and Technology)
- Dr. Eleni Tzirita Zacharatou (IT University of Copenhagen)

The session included 11 papers, naturally partitioned into 4 sub-sessions, as follows:

Spatial Data:

- A Machine Learning-Aware Data Re-partitioning Framework for Spatial Datasets
- Example-based Spatial Search at Scale
- SPADE: GPU-Powered Spatial Database Engine for Commodity Hardware

Spatial Crowdsourcing:

- Bilateral Preference-aware Task Assignment in Spatial Crowdsourcing
- Human-Drone Collaborative Spatial Crowdsourcing by Memory-Augmented Distributed Multi-Agent Deep
- Influence-aware Task Assignment in Spatial Crowdsourcing

Trajectory Data:

- Maximizing Range Sum in Trajectory Data
- Workload-Aware Shortest Path Distance Querying in Road Networks

Temporal and Time-Series Data:

- Provenance in Temporal Interaction Networks
- Constructing Compact Time Series Index for Efficient Window Query Processing
- iTemporal: An Extensible Generator of Temporal Benchmarks

Papers in the first sub-session apply data preprocessing to assist spatial ML tasks, study spatial pattern queries and use modern hardware to accelerate spatial data management. Papers about spatial crowdsourcing studied how to incorporate more information (e.g., drone data, influence of objects and preferences from people) for matching tasks with human in a more realistic setting. Papers in the third sub-session, study problems for road-network applications (i.e., range sum maximization over trajectory data, shortest path queries). Papers in the last sub-session are on various topics related to temporal and time-series management (i.e., computing provenance information in temporal networks, time-series indexing, temporal benchmark generation).

The authors of each paper took about 5 minutes to present their work. After the end of all presentations, the session chair coordinated a discussion between the audience, the invited panel members, and the authors, which included Q&A and a discussion about future directions on the topics.

7.1 Categorization and Summary of the Papers

Based on their focus, we could categorize the papers of the session into the following classes:

1. Application-oriented paper(s)
2. Paper(s) on Database Indexing
3. Paper(s) on Database Benchmarking
4. Paper(s) on Special Topics that involve Spatial, Temporal and Spatiotemporal Data

This categorization gives us insights for trends and possible future directions for spatial and/or temporal data engineering. Hence, one direction is to focus on real-life applications that manage spatial/temporal data, in order to devise effective and efficient solutions for their needs. The second direction is indexing on spatial/temporal data, which is a fundamental problem in our community. The third direction is to provide effective and reliable benchmarks for spatial/temporal data management. The fourth direction is working on research about combination of other research topics with spatial/temporal databases. Such topics include machine learning, human involvement, graph search, and new technologies.

In addition to these four directions that stem directly from the categorization, panel members discussed further research directions which are outlined in the following sections.

7.2 Spatial and Temporal Data Fusion

The increased availability of data and knowledge as well as the trend towards more interdisciplinary and transdisciplinary research approaches call for solutions unlocking the immense power of cross domain and multi-source data fusion, i.e. the fusion of views representing the various perspectives on a scene of an excerpt of the real world. Thereby, views vary among different disciplines, different data sets, data types and data sources including models, and different abstraction of data ranging from measurement records over patterns up to knowledge representations. For example, physicists and biologists have different perspectives on some scenes in the ocean, such as the behavior of saline concentration and behaviour of living organisms. Also within the same scientific discipline, there may be many varying views on entities of a general research topic. For example oxygen deficits in coastal regions measured by remote sensing raster data vs. series of in-situ sensor measurements. Since most scientific views are traditionally explored in isolation, they are also restricted to specific findings and often do not allow a broad understanding of relationships and functional dependencies between multiple concepts such as structural relationships among different scientific views. It was coined that research on heterogeneous data integration and fusion for the spatial and temporal data has not been studied extensively in the literature. This creates many opportunities for research. To pave the path towards cross-domain Fusion, a novel, fundamental and systematic framework of methods is needed that enable the fusion of data, patterns, knowledge, etc.; i.e., the fusion of multiple potentially heterogeneous views and stages of data from diverse domains. Initial approaches for exploiting the power of machine learning to fusing data/views lifting up the potential of data analysis are summarized in [18]. While most approaches have been proposed in the field of urban analytics, fusing traffic data with other urban attributes [19–21], it increasingly attracts further scientific disciplines [22, 23]. However, most existing solutions are isolated, often too narrow, ad-hoc approaches designed for specific applications. We need more general, broader and systematic fusion concepts.

With the integration of diverse data sources, we also need solutions for the alignment of diverse types and models of data, in particular spatial and temporal data occurs in different forms, scales and structures. In heterogeneous data sources but also due to its particular nature, temporal data occurs in different forms. Interval data is used to store state information over a period of time, e.g., an employment contract or a task allocated on a manufacturing machine, event data is used to store a happening at a specific point in time, e.g., a warning issued to an employee or a failure of a machine component, and time series data is used to store continuous variables or parameters, e.g., stress levels of employees or sensor data related to manufacturing machines. While processing and analysis techniques exist for temporal data, they usually focus on a single type, and do not consider the fusion of the different form based on their nature, resulting in an incomplete picture for analysis. This, despite the fact that different types of temporal data arise in the same application, for instance in [24] they consider benchmarks that generate both time intervals and time points. In this setting, there is a need for research on data modeling for being able to reference data entries based on their type and timestamp that belong together, research on data processing for the efficient fusion of data entries that need to be analyzed together, and finally, analysis techniques that consider all three types of temporal data.

Examples of interval, event and time series data from our "Spatial and Temporal Data Management" session.

In [25] the input data is time series data, while the window sequence is interval data; in [24] they consider the generation of interval data and event data or time series data; in [26] they consider event data and the windowing approach (maybe also used for tracking over particular interesting time period) employs interval data.

7.3 Plug-and-Play Infrastructure

Most of the existing work in spatial data management, including papers in this "Spatial and Temporal Data Management" session such as [27, 28], focuses on devising high-performance solutions to specific problems. However, looking at the big picture, this is insufficient. In addition to those specialized tailor-made solutions, we need to create an infrastructure that supports a "plug-and-play" functionality. Specifically, this infrastructure should enable and ease the creation and composition of processing pipelines that leverage individual specialized advances. In addition, the envisioned infrastructure should serve as a bridge between data management researchers and domain experts. While today we have abundant access to spatio-temporal data, it is often hard to determine what is interesting and challenging about this data without the knowledge of a domain expert. Therefore, to foster interdisciplinary collaborations, the envisioned infrastructure should facilitate a feedback loop between data management researchers and domain experts.

7.4 Spatial Data Management Beyond Vector Data

All the papers in the spatial data management sub-session [27–29] deal with objects that follow the vector data model, which represents spatial features with their geometry. Typical geometries are points (e.g., GPS locations), lines (e.g., roads and rivers) and polygons (e.g., regional boundaries). Nowadays, Earth Observation (EO) satellites are another prominent source of spatial data. The number of EO satellites and their acquisition rates are growing fast, resulting in an ever-increasing amount of collected satellite images. Satellite images follow the raster data model, which represents the study area as a collection of granules (e.g., pixels). To enable efficient data analyses, we need data management solutions that treat both data types as first-class citizens. For example, tracking flooded residential areas during an ongoing flood event requires comparing water masks from recent satellite images (rasters) with permanent water bodies (vectors). While there is some initial work towards the efficient combination of vector and raster data [30], many open challenges remain in query execution and optimization.

7.4.1 Spatial Query Optimization and Performance Tuning

Today, we observe a twofold evolution of spatial data systems and algorithms. From the one side, the application domains become more and more diverse [27, 28, 31, 32], while systems incorporate new hardware technologies in their design [29]. This twofold evolution creates new challenges in query optimization and performance tuning. For example, the query optimizer proposed in [29] only considers the cost of the data transfer to the GPU and lacks support for multiway queries and queries with multiple constraints. Further research is required to develop query optimization and performance tuning techniques that can navigate a broad spectrum of application requirements and hardware settings.

7.5 Privacy-Aware Compliant Spatial Crowdsourcing

The "Spatial and Temporal Data Management" session contained several papers that advance the state-of-the-art in spatial crowdsourcing [33–35]. However, enriching existing solutions with privacy awareness and compliance with regulatory restrictions (for example, with respect to the use of drones proposed in [34]) remains an open research challenge.

7.6 Road Networks and Trajectory Data

Road networks are spatial graphs used in a variety of real-life applications. Routing being one of them, lies in the core of navigation systems, location-aware recommender systems for touristic and trip-planning scenarios and logistics, among others. In its simplest and most fundamental form, the goal is to determine the shortest or the fastest path between two locations in the network. A number of path-finding algorithms and indexing techniques has been proposed in the past to efficiently process shortest paths queries. Performance and scalability are traditionally the interesting objectives, while dealing with structural updates is less important as road networks are very static, contrary to other types of graphs. In the dynamic aspect of routing, computing time-aware paths is a timely challenge, where how to consider moving patterns in different times of the day and employ real-time traffic are key challenges. With the proliferation of geo-positioning systems and mobile phones, the amount of routing queries received by mobile applications has significantly increased. Under this, techniques for parallel and batch processing such queries are important, but another interesting research question is how we can use previous requests (i.e., a workload of routing queries) to optimize existing path-finding techniques and indices. [32] focuses on how the spatial skewness of the query locations and the temporal locality of the time-aware requests in an existing workload can be considered towards this direction.

Besides the actual graph of a road network, trajectories are another type of data used by modern location-aware applications. In routing, trajectories can be used as train data for applying learning techniques to improve the quality of the recommended paths. For example, previously followed paths by users or their friends can be used to learn moving habits and patterns. Trajectories play a key role also in business placement scenarios where the goal is to identify the best location for a company to open a store. Such scenario is studied in [31].

8 Spatial and Temporal Data Mining

In ICDE 2022, the “Spatial and Temporal Data Mining" session contains 12 research papers. The panelists include five experts in this area:

- Dr. Zhifeng Bao (RMIT University)
- Dr. Jilin Hu (Aalborg University)
- Dr. Cheng Long (Nanyang Technological University)
- Dr. Raymond Chi-Wing Wong (The Hong Kong University of Science and Technology)
- Dr. Bin Yang (Aalborg University)

Spatial and temporal data, which refers to data that involves spatial information (e.g., coordinates) and/or temporal information (e.g., time stamps), is continuously being generating in a wide range of applications in urban cities (e.g., mobilities, commuting, traffic, planning and logistics), in geography (e.g., GIS and remote sensing), in chemistry and biology (e.g., 3D molecular modeling), etc. Mining spatial and temporal data in these applications would naturally facilitate these applications with more intelligence, effectiveness and/or efficiency. At ICDE 2022, we have a dozen interesting papers published in this area of spatial and temporal data mining. We also have a few experts, sharing their thoughts and visions of conducting research in this area. Next, we first outline these ICDE 2022 papers in this area and then summarize the expert panels' thoughts and visions of this area.

The papers fall in three groups, namely (1) time series mining [36–40], (2) spatiotemporal prediction [41–45] and (3) representation learning for spatial and/or temporal data [46, 47]. Specifically, [36, 37] propose autoencoder networks for time series outlier detection, [38, 39] study shapelets of time series for tasks such as time series classification, and [40] studies the backdoor attack on deep learning based time series classification models, [41–45] study various spatiotemporal prediction tasks (including activity prediction [41], urban crime prediction [42], traffic prediction [43], docked bike prediction [44]) and the problem of setting the grid size for spatiotemporal prediction models [45], and [46, 47] study the presentation learning problem for temporal paths [46] and trajectory [47].

The panel experts share their visions and suggestions of conducting research in spatial and temporal data mining for the next 5-10 years, which are summarized as follows. First, this area of research will continue to be attractive with the new developments of technologies such as AI, 5G and metaverse. For example, metaverse corresponds to a virtual physical space which naturally involves spatial data such as terrain data. Second, researchers in this area can be more aggressive by taking a leading role in developing new technologies instead of adopting those from other areas such as CNNs, GNNs, and word2vec. Third, developing generic frameworks that are able to enhance a variety of existing models rather than yet another end-to-end model would make more impact. Some examples of such generic frameworks include for turning ST-agnostic models to ST-aware models and [48, 49] frameworks for automatically searching for models of time series forecasting. Fourth, collaborating with industry more closely and conducting interdisciplinary research involving spatial and temporal data would have significant potential and bring real impact. Fifth, existing spatial and temporal data mining solutions still need to be improved in various aspects, including but limited to interpretability, scalability, sustainability, data sparsity and uncertainty. All in all, we believe that spatial and temporal data mining will continue to be a hot research area and involve many research problems to be solved ahead of the way.

9 Conclusion

In conclusion, the introduction of panel discussions at the 38th IEEE International Conference on Data Engineering (ICDE) proved highly successful in generating inspiring discussions and promoting active participation. These discussions enriched understanding, offered fresh insights, and attracted more participants than anticipated. Summarizing the panels in this article benefits the data engineering community, providing a valuable resource for researchers and conference organizers. The success of the panel discussions at ICDE'22 highlights their effectiveness in fostering engagement and collaboration. The article suggests adopting this format for future conferences, encouraging in-depth exploration of research topics and active involvement from diverse perspectives.

Acknowledgement: This article was compiled by the PC co-chairs (Gao Cong, Stratos Idreos, and Feifei Li) of ICDE'22.

References

[1] Matthew Hutson. Taught to the Test. IEEE Aerospace and Electronic Systems Magazine, Vol 36, 9, 30-42, 2021.

[2] Eftelioglu, Emre and Shekhar, Shashi and Hudson, James and Joppa, Lucas and Baru, Chaitanya and Janeja, Vandana. Data Science for Earth: An Earth Day Report. Association for Computing Machinery, Vol 22, 1, 4-7, 2020.

[3] Jin, Yinyu and Yuan, Sha and Zhou, Shao and Hall, Wendy and Tang, Jie. Turing Award elites revisited: patterns of productivity, collaboration, authorship and impact. Scientometrics, Vol 126, 2021.

[4] Hao, Meng and Li, Hongwei and Luo, Xizhao and Xu, Guowen and Yang, Haomiao and Liu, Sen. Efficient and privacy-enhanced federated learning for industrial artificial intelligence. IEEE Transactions on Industrial Informatics, arXiv preprint, 2019.

[5] Li, Tian and Sahu, Anit Kumar and Talwalkar, Ameet and Smith, Virginia. Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, Vol 37, 3, 2020.

[6] Nguyen, Dinh C and Ding, Ming and Pathirana, Pubudu N and Seneviratne, Aruna and Li, Jun and Poor, H Vincent. Federated learning for internet of things: A comprehensive survey. IEEE Communications Surveys & Tutorials, 2021.

[7] Tian, Yuanyishu and Wan, Yao and Lyu, Lingjuan and Yao, Dezhong and Jin, Hai and Sun, Lichao. FedBERT: When Federated Learning Meets Pre-Training. ACM Transactions on Intelligent Systems and Technology (TIST), 2022.

[8] Liu, Yang and Huang, Anbu and Luo, Yun and Huang, He and Liu, Youzhi and Chen, Yuanyuan and Feng, Lican and Chen, Tianjian and Yu, Han and Yang, Qiang. Fedvision: An online visual object detection platform powered by federated learning. AAAI, 2020.

[9] Dayan, Ittai and Roth, Holger R and Zhong, Aoxiao and Harouni, Ahmed and Gentili, Amilcare and Abidin, Anas Z and Liu, Andrew and Costa, Anthony Beardsworth and Wood, Bradford J and Tsai, Chien-Sung and others. Federated learning for predicting clinical outcomes in patients with COVID-19. Nature medicine, Vol 27, 10, 1735-1743, 2021.

[10] Kaissis, Georgios A and Makowski, Marcus R and Rückert, Daniel and Braren, Rickmer F. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, Vol 2, 6, 305-311, 2020.

[11] Xu, Jie and Glicksberg, Benjamin S and Su, Chang and Walker, Peter and Bian, Jiang and Wang, Fei. Federated learning for healthcare informatics. Journal of Healthcare Informatics Research, Vol 5, 1, 2021.

[12] Antunes, Rodolfo Stoffel and André da Costa, Cristiano and Küderle, Arne and Yari, Imrana Abdullahi and Eskofier, Björn. Federated Learning for Healthcare: Systematic Review and Architecture Proposal. ACM Transactions on Intelligent Systems and Technology (TIST), Vol 22, 4, 1-23, 2022.

[13] Yang, Wensi and Zhang, Yuhang and Ye, Kejiang and Li, Li and Xu, Cheng-Zhong. Ffd: A federated learning based method for credit card fraud detection. International conference on big data, 2019.

[14] Larson, James A and Larson, Carol L. Federated Database Systems. Handbook of Heterogeneous Networking 1999, 27-1, 2018.

[15] Gupta, Ankush M and Gadepally, Vijay and Stonebraker, Michael. Cross-engine query execution in federated database systems. IEEE High Performance Extreme Computing Conference (HPEC), 1-6, 2016.

[16] Sheth, Amit P and Larson, James A. Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Computing Surveys (CSUR), Vol 22, 3, 183-236, 1990.

[17] Aisha Mohamed, Ghadeer Abuoda, Abdurrahman Ghanem, Zoi Kaoudi, Ashraf Aboulnaga. RDFFrames: knowledge graph access for machine learning tools. VLDB J., Vol 31, 2, 321-346, 2022.

[18] Meng, Tong and Jing, Xuyang and Yan, Zheng and Pedrycz, Witold. A survey on machine learning for data fusion. Information Fusion, Vol 57, 115-129, 2020.

[19] Liu, Jia and Li, Tianrui and Xie, Peng and Du, Shengdong and Teng, Fei and Yang, Xin. Urban big data fusion based on deep learning: An overview. Information Fusion, Vol 53, 123-133, 2020.

[20] Khan, Sulaiman and Nazir, Shah and García-Magariño, Iván and Hussain, Anwar. Deep learning-based urban big data fusion in smart cities: Towards traffic monitoring and flow-preserving fusion. Computers & Electrical Engineering, Vol 89, 106906, 2021.

[21] Zheng, Yu. Methodologies for cross-domain data fusion: An overview. IEEE transactions on big data, Vol 1, 1, 16-34, 2015.

[22] Soldi, Giovanni and Gaglione, Domenico and Forti, Nicola and Millefiori, Leonardo M and Braca, Paolo and Carniel, Sandro and Di Simone, Alessio and Iodice, Antonio and Riccio, Daniele and Daffinà, Filippo Cristian and others. Space-based global maritime surveillance. Part II: Artificial intelligence and data fusion techniques. IEEE Aerospace and Electronic Systems Magazine, Vol 36, 9, 30-42, 2021.

[23] Karagiannopoulou, Aikaterini and Tsertou, Athanasia and Tsimiklis, Georgios and Amditis, Angelos. Data Fusion in Earth Observation and the Role of Citizen as a Sensor: A Scoping Review of Applications, Methods and Future Trends. Remote Sensing, Vol 14, 5, 1263, 2022.

[24] Luigi Bellomarini, Markus Nissl, Emanuel Sallinger. iTemporal: An Extensible Generator of Temporal Benchmarks. IEEE ICDE, 2022-2034, 2022.

[25] Jing Zhao. Peng Wang, Bo Tang, Lu Liu, Chen Wang, Wei Wang, Jianmin Wang. Constructing Compact Time Series Index for Efficient Window Query Processing. IEEE ICDE, 3025-3037, 2022.

[26] Chrysanthi Kosyfaki, Nikos Mamoulis. Provenance in Temporal Interaction Networks. IEEE ICDE, 2278-2291, 2022.

[27] Kanchan Chowdhury, Venkata Vamsikrishna Meduri, Mohamed Sarwat. A Machine Learning-Aware Data Re-Partitioning Framework for Spatial Datasets. IEEE ICDE, 2427-2440, 2022.

[28] Hanyuan Zhang, Siqiang Luo, Jieming Shi, Jing Nathan Yan, Weiwei Sun. Example-Based Spatial Search at Scale. IEEE ICDE, 539-551, 2022.

[29] Harish Doraiswamy, Juliana Freire. SPADE: GPU-Powered Spatial Database Engine for Commodity Hardware. IEEE ICDE, 2670-2682, 2022.

[30] Samriddhi Singla, Ahmed Eldawy, Tina Diao, Ayan Mukhopadhyay, Elia Scudiero. The Raptor Join Operator for Processing Big Raster + Vector Data. SIGSPATIAL, 324-335, 2021.

[31] Kaiqi Zhang, Hong Gao, Xixian Han, Jian Chen, Jianzhong Li. Maximizing Range Sum in Trajectory Data. IEEE ICDE, 755-766, 2022.

[32] Bolong Zheng, Jingyi Wan, Yongyong Gao, Yong Ma, Kai Huang , Xiaofang Zhou, Christian S. Jensen. Workload-Aware Shortest Path Distance Querying in Road Networks. IEEE ICDE, 2373-2385, 2022.

[33] Xuanhao Chen, Yan Zhao, Kai Zheng, Bin Yang, Christian S. Jensen. Influence-Aware Task Assignment in Spatial Crowdsourcing. IEEE ICDE, 2142-2154, 2022.

[34] Yu Wang, Chi Harold Liu, Chengzhe Piao, Ye Yuan, Rui Han, Guoren Wang, Jian Tang. Human-Drone Collaborative Spatial Crowdsourcing by Memory-Augmented Distributed Multi-Agent Deep Reinforcement Learning. IEEE ICDE, 459-471, 2022.

[35] Xu Zhou, Shiting Liang, Kenli Li, Yunjun Gao, Keqin Li. Bilateral Preference-Aware Task Assignment in Spatial Crowdsourcing. IEEE ICDE, 1688-1700, 2022.

[36] Tung Kieu, Bin Yang, Chenjuan Guo, Razvan-Gabriel Cirstea, Yan Zhao, Yale Song, Christian S. Jensen. Anomaly Detection in Time Series with Robust Variational Quasi-Recurrent Autoencoders. IEEE ICDE, 1342-1354, 2022.

[37] Tung Kieu, Bin Yang, Chenjuan Guo, Christian S. Jensen, Yan Zhao, Feiteng Huang, Kai Zheng. Robust and Explainable Autoencoders for Unsupervised Time Series Outlier Detection. IEEE ICDE, 3038-3050, 2022.

[38] Guozhong Li, Byron Choi, Jianliang Xu, Sourav S. Bhowmick, Daphne Ngar-yin Mah, Grace Lai-Hung Wong. IPS: Instance Profile for Shapelet Discovery for Time Series Classification. IEEE ICDE, 1781-1793, 2022.

[39] Akihiro Yamaguchi, Ken Ueno, Hisashi Kashima. Learning Evolvable Time-series Shapelets. IEEE ICDE, 793-805, 2022.

[40] Daizong Ding, Mi Zhang, Yuanmin Huang, Xudong Pan, Fuli Feng, Erling Jiang , Min Yang. Towards Backdoor Attack on Deep Learning based Time Series Classification. IEEE ICDE, 1274-1287, 2022.

[41] Yinfeng Li, Chen Gao, Quanming Yao, Tong Li, Depeng Jin, Yong Li. DisenHCN: Disentangled Hypergraph Convolutional Networks for Activity Prediction. CoRR, abs/2208.06794, 2022.

[42] Zhonghang Li, Chao Huang, Lianghao Xia, Yong Xu, Jian Pei. Spatial-Temporal Hypergraph Self-Supervised Learning for Crime Prediction. IEEE ICDE, 2984-2996, 2022.

[43] Razvan-Gabriel Cirstea, Bin Yang, Chenjuan Guo, Tung Kieu, Shirui Pa. Towards Spatio- Temporal Aware Traffic Time Series Forecasting. IEEE ICDE, 2900-2913, 2022.

[44] Guanyao Li, Xiaofeng Wang, Gunarto Sindoro Njoo, Shuhan Zhong, S.-H. Gary Chan, Chih-Chieh Hung, Wen-Chih Peng. A Data-Driven Spatial-Temporal Graph Neural Network for Docked Bike Prediction. IEEE ICDE, 713-726, 2022.

[45] Jiabao Jin, Peng Cheng, Lei Chen, Xuemin Lin, Wenjie Zhang. GridTuner: Reinvestigate Grid Size Selection for Spatiotemporal Prediction Models. IEEE ICDE, 1193-1205, 2022.

[46] Sean Bin Yang, Chenjuan Guo, Jilin Hu, Bin Yang, Jian Tang, Christian S. Jensen. Weakly-supervised Temporal Path Representation Learning with Contrastive Curriculum Learning. IEEE ICDE, 2873-2885, 2022.

[47] Yang, Peilun and Wang, Hanchen and Lian, Defu and Zhang, Ying and Qin, Lu and Zhang, Wenjie. TMN: Trajectory Matching Networks for Predicting Similarity. IEEE ICDE, 1700-1713, 2022.

[48] Razvan-Gabriel Cirstea, Chenjuan Guo, Bin Yang, Tung Kieu, Xuanyi Dong, Shirui Pan. Triformer: Triangular, Variable-Specific Attentions for Long Sequence Multivariate Time Series Forecasting. IJCAI, 1994-2001, 2022.

[49] Razvan-Gabriel Cirstea, Tung Kieu, Chenjuan Guo, Bin Yang, Sinno Jialin Pan. EnhanceNet: Plugin Neural Networks for Enhancing Correlated Time Series Forecasting. IEEE ICDE, 1739-1750, 2021.

Lifelogging As An Extreme Form of Personal Information Management - What Lessons To Learn

Ly-Duyen Tran[1] Cathal Gurrin[1,2] Alan F. Smeaton[1,3*]

[1]School of Computing
[2]ADAPT Centre
[3]Insight Centre for Data Analytics
Dublin City University, Glasnevin, Dublin 9
Ireland.

Abstract

Personal data includes the digital footprints that we leave behind as part of our everyday activities, both online and offline in the real world. It includes data we collect ourselves, such as from wearables, as well as the data collected by others about our online behaviour and activities. Sometimes we are able to use the personal data we ourselves collect, in order to examine some parts of our lives but for the most part, our personal data is leveraged by third parties including internet companies, for services like targeted advertising and recommendations. Lifelogging is a form of extreme personal data gathering and in this article we present an overview of the tools used to manage access to lifelogs as demonstrated at the most recent of the annual Lifelog Search Challenge benchmarking workshops. Here, experimental systems are showcased in live, real time information seeking tasks by real users. This overview of these systems' capabilities show the range of possibilities for accessing our own personal data which may, in time, become more easily available as consumer-level services.

1 What is Personal Information Management ?

Personal information or personal data is the digital form of the footprints of the everyday things that we do in our daily lives. It includes data that we can generate ourselves for example from wearable devices, or information that we ourselves create directly like our emails. In the case of the former, data from wearable devices can include physiological indicators like heart rate, respiration rate, heart rate variability or galvanic skin response levels, which are indicators of stress. It can also include our movement data from accelerometers on as well as location data from GPS or other location-tracking technologies. When such raw data is analysed then we can infer things like step counts, distances walked, run or biked, metrics for our sleep like duration and sleep quality and other health and wellness indicators. The purpose of capturing such personal data about ourselves is usually self-monitoring and self-tracking, typically for our own health awareness.

While self-tracking is now a consumer-level rather than a specialist activity with hundreds of thousands of apps for uses like counting calories or fitness tracking [28] much of what we call our personal data is actually recorded by others and not directly by ourselves. This includes the logs and records of our online activities like logs of our web browsing, social media interactions, media consumption and our communications and interactions with others. Even though this data has been gathered and is being used by others, it is personal data that belongs to us. Personal data gathered by third parties has value because of the insights it can reveal about us and especially

[*] Contact: `Alan.Smeaton@DCU.ie`

when it is aggregated and combined with the personal data of others [6]. This form of personal data is exploited almost exclusively by others, albeit supposedly for our benefit through targeted advertising and recommendations, and not directly by us.

Personal information management refers to tools and platforms that allow us to control our personal data by allowing us to gather, store, update, analyse, interpret and sometimes to share it, no matter who has gathered it. It is awkward or at best inconvenient for us to access or even to use our own personal data for our own personal purposes whether it is data we have gathered, or has been gathered by third parties. While the log files of our interactions with search, browsing, social media, communications and media consumption services, were originally intended for helping to debug and refine those systems and while we can request such logs from internet services it is not made easy for us to do so. Even for our use of systems in our own enterprises and organisations, like access logs to virtual learning environments in Colleges and Universities, this typically requires downloading CSV files of personal data and then processing them ourselves, which is off-putting to most. Access to our personal data from wearable devices is also typically supported through offering CSV or equivalent files as downloads.

It is clear that our personal data can be a rich source of insights into us, into our well-being, our habits and behaviours and most importantly, into changes into those behaviours. Unfortunately we are able to query and to interrogate the individual and unconnected repositories for our personal data only on a per-device or per-source basis whereas the real benefits of such interrogations would be when the whole of our personal data, integrated across sources, could be analysed. This poses the question of whether there is any work done that brings these sources together.

In the later part of this article we will look at some very limited attempts to pool together our personal data from multiple sources but before that, in the next section, we will examine the current best-practice in the area of self-generated recording of everyday activities, the area known as lifelogging.

2 Lifelogging: Extreme Personal Information Management

In this section, we take a closer look as an extreme form of personal information gathering, which is called lifelogging [14]. Lifelogging involves the comprehensive and continuous collection of data about one's daily activities and experiences, gathered from a wide array of data sources. Wearable devices, including smartwatches and fitness trackers, can be used to capture health metrics, such as heart rate, sleep quality, and activity levels. GPS sensors can be used to track one's movements and location. The most prominent data source for lifelogging is the use of wearable cameras to capture point-of-view images and videos of one's daily activities. These data sources provide a holistic representation of one's life and can be used for a variety of applications, such as health monitoring, memory prosthetic, and behavioural analysis.

The applications of lifelogging are widespread and have been explored in many domains, such as supporting human memory recollection [3, 5, 15], supporting large-scale epidemiological studies in healthcare [35], monitoring the lifestyle of individuals [29, 48], behaviour analytics [12], and diet/obesity analytics [51]. However, the vast volume and immense complexity of lifelog archives presents a challenge for users to navigate and analyse these archives in order to identify relevant information. As such, there is a growing interest in developing lifelog retrieval systems that effectively leverage lifelog data to meet the diverse needs of users.

Retrieval systems for lifelog data have been a popular research topic for several years. Managing the large amount of data collected by lifeloggers is crucial in order to maximise benefits from the data. In the MyLifeBits system, introduced more than 20 years ago by Gordon Bell [4], then state-of-the-art techniques from database search and traditional information retrieval (IR) were employed to index and provide access to lifelog data through a desktop interface. However, as the volume of lifelog data increases, the need for more efficient and effective retrieval systems becomes more apparent and the limitations of MyLifeBits soon became apparent.

The first retrieval system designed for large archives of lifelog data was proposed by Doherty et al. [11],

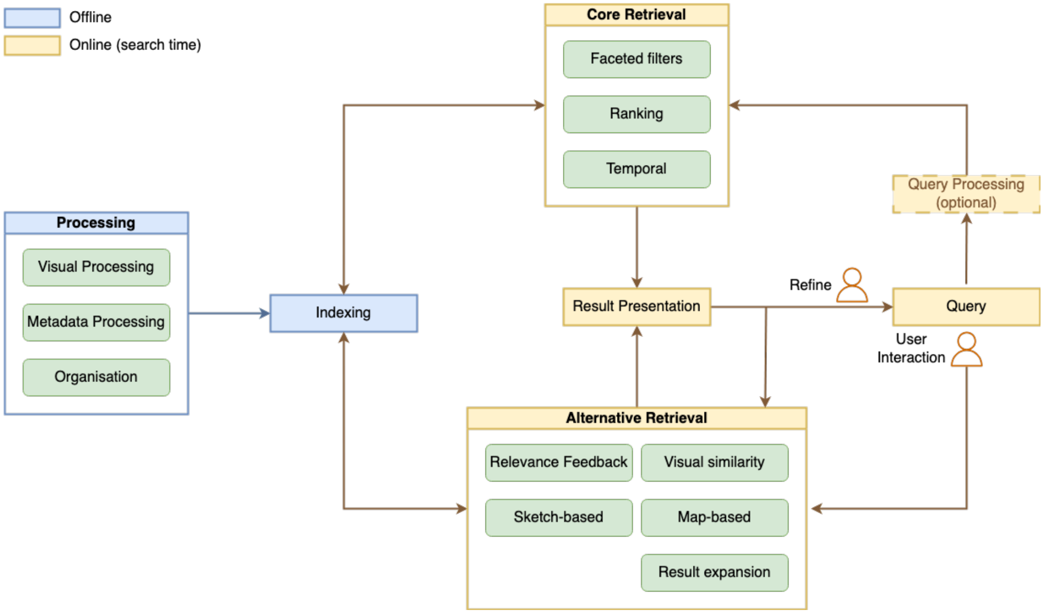


Figure 1: General pipeline of lifelog retrieval systems in the annual Lifelog Search Challenge (LSC).

moving the time/date *browsing* approach of lifelog systems at the time to a *search* approach. The system employed event segmentation, event annotation and multi-axes search, which are the 'who', 'what', 'when', and 'where' axes of retrieval. However, without a large user base, it was difficult to define search use-cases in order to evaluate and improve lifelog retrieval systems.

In order to explore the possibilities and limitations of question-answering on the unique characteristics of timeline information such as that gathered in lifelogs, the authors of [39] have created and released TimelineQA. This is a generator that produces synthetic lifelogs for imaginary people with different personae including age, gender, education and family status. It uses a fine-tuned LLM to produce the actual content on which the effectiveness of question-answering can be investigated. The synthetic lifelogs are composed of timelines of ordinary life events each of which have start times, end times and possibly locations, and they may also have multimodal features like photos. The TimelineQA benchmark is successful in that it is allowing a new form of interaction with lifelogs – question-answering – to be explored in a systematic and repeatable way.

Several other benchmarking activities for lifelog retrieval systems have been organised which evaluate and compare the retrieval effectiveness of participating systems on real rather than synthetic lifelogs. The notable and best example of this is the annual Lifelog Search Challenge (LSC) [42]. The LSC provides a platform for evaluating state-of-the-art systems for managing lifelogs and attracts a large number of participants internationally. In the LSC, different systems compete with each other in a live/virtual environment, where participants are given a set of information needs and queries and a limited time to find the relevant and matching lifelog moments, in real time. Generally, most systems participating in the LSC follow the flow of data processing, indexing, and then retrieval, as shown in Figure 1. Interactivity and user experience are important aspects of the lifelog retrieval systems, as their affordances are known to directly influence the effectiveness of the retrieval process [26]. We now discuss some of the techniques and components involved in lifelog retrieval with example systems from the LSC.

2.1 Processing

Lifelog data is typically collected from sources including wearable cameras, wearable sensors, and mobile devices. The data is then processed to extract relevant information before being indexed and stored in a database. The processing stage is crucial as it determines the quality of the data and the effectiveness of the retrieval process.

2.1.1 Visual Processing

The most challenging aspect of processing is visual data, where different techniques are employed. We categorise these techniques into three groups: low-level features, concepts, captions, and embeddings.

Low-level features, such as colours, textures, and shapes, are commonly used in lifelog systems. These are classic features that have been used in the field of image processing for decades and are still used in lifelog systems. Their most popular application is the removal of low-quality images as seen in systems such as MySceál [43] (or its variant, E-MySceál) and Memento [1]. It is believed that low-quality images are less likely to be relevant to users, and thus, removing them can improve retrieval performance.

Due to advances in computer vision algorithms, many lifelog systems utilise deep convolutional neural networks (CNNs) such as AlexNet [23], VGG [36], GoogLeNet [38], and ResNet [16], trained on various dataset such as ImageNet [8], OpenImages [24], Places365 [49], and Visual Genome [22]. These models are used to extract semantic lifelog 'concepts', which are then indexed and used for retrieval. Optical character recognition (OCR) is also used to extract text from images, which proved to be useful in scenarios like identifying brand names and street names [42] during the real time search. This approach, often referred to as concept-based retrieval, is widely adopted due to its ease of implementation and the availability of pre-trained models.

Captioning, the generation of textual descriptions for images, is less commonly used in lifelog systems due to its inherent challenges. Generated captions are not always accurate, which limits its popularity. The most notable example of captioning in lifelog systems is vitrivr [37], which supports text-based search using generated captions.

Embedding-based retrieval is an approach that has gained popularity in recent years. This involves mapping images into a vector space, where the similarity between images can be measured using cosine similarity, for example. This approach is widely used in cross-modal retrieval, where images are mapped into the same vector space as text queries. The most notable example of this approach is CLIP [32], which is a pre-trained model that maps images and text into a shared vector space. This approach replaces the need for concept extraction and captioning, as the similarity between images and text can be measured directly. Therefore, many lifelog systems have replaced their concept-based retrieval with embedding-based retrieval, as demonstrated by E-MySceál [43], by adding embeddings to the existing concept-based retrieval and thus the name E-MySceál. Other systems such as Memento [1] and LifeSeeker [30] have also adopted this approach. This *embedding-based retrieval* approach allows a more user-friendly experience by allowing users to search for images using natural language queries and is shown to significantly improve retrieval performance

2.1.2 Metadata Processing

While metadata processing in lifelog systems is minimal due to standardised CSV formats provided by the LSC challenge organisers, it still plays a vital role in enhancing retrieval performance. The two most common metadata used in lifelog systems are timestamps and GPS coordinates. These help at retrieval time as users often remember aspects of the date or time or of the approximate location, of their personal data they are trying to find. Time information such as the time of the day, day of the week, month, and year are extracted from timestamps and used for filtering and scoring, as demonstrated by MySceál [43] and LifeSeeker [30]. Clustering GPS coordinates for stay point detection, and inferring semantic locations are common techniques used to enhance the location metadata. For example, MyEachtra [41] employs a visual-aware stay point detection algorithm [44] to detect stay points and utilises crowdsourced location data from FourSquare[1] to infer semantic locations. Meanwhile, LifeSeeker [30] enhances the location metadata by manually labelling images with 'areas' within a location (such

[1] https://developer.foursquare.com/docs/api-reference/

as kitchen, living room, and bedroom) to create a more fine-grained location hierarchy. This approach provides more ‘concepts’ for users to filter and search for.

Biometric sensor data, surprisingly, is rarely used in lifelog systems due to noisy data and the lack of standardised formats. However, some systems have demonstrated the potential of using biometric data for retrieval by binning the data into categories [2, 31, 45]. Another type of metadata that was used in lifelog systems is music listening data. LifeSeeker uses music listening data to infer a user’s mood, which is then used to filter and score the search results [30]. However, due to the lack of LSC queries with the focus on these types of metadata, their effectiveness is yet to be evaluated.

2.1.3 Organisation

Due to the temporal nature of lifelog data, it is straightforward to organise lifelog data chronologically [50]. Some of the LSC participating systems use a more sophisticated approach to organise the data, such as by using segmentation. Most systems that include segmentation use a simple approach by comparing the visual similarity between two consecutive images [43]. MyEachtra [41] clusters the GPS coordinates to infer the semantic locations and then organise the data by the semantic locations.

LifeGraph [33] offers a novel approach and organises the lifelog data by using a graph structure, with images as the centrepoint of the schema. The graph is constructed by linking the images with the metadata and detected concepts, then is extended with Wikidata [47] and COEL (Classification of Everyday Living) [7]. However, the authors later acknowledged that COEL played an insignificant role in the query expansion and that Wikidata was a more useful source of concepts.

2.2 Search and Navigation

The core aspect distinguishing various lifelog retrieval systems is their retrieval mechanism. This significantly influences how data is processed, indexed, and presented.

Filters are a widely employed technique that has proven to be effective in lifelog systems. Faceted filters can be applied to various modalities and are often categorised by attributes such as time of day, day of the week, month, year, location, number of people, biometrics, and lifelog concepts as these are the aspects of personal data that a user can recall in their information seeking [21]. Fixed phrases are typically used for filtering values, presented through drop-down lists, checkboxes, or sliders. Some systems, like MyScéal [43], allow users to enter free text, with autocomplete suggestions for available lifelog concepts.

Query expansion is a technique used to assist users in formulating queries. It typically involves concept suggestion using various sources like WordNet, ConceptNet, and Thesaurus.com, and employing models like BERT [9] for concept similarity. MyScéal [43] introduced a free-text query form with filter value extraction and query expansion, a method later adopted by LifeSeeker [30].

A ranking mechanism is essential for any retrieval system in order to sort outputs before presentaiton. Lifelog concepts can be used to score images based on the number of matched concepts [50]. Extending TF-IDF term weighting using concepts extracted from computer vision models is a popular approach, as demonstrated by MyScéal [43] and by LifeSeeker [30]. TF-IDF is a common technique in information retrieval to score the relevance of a document to a query. It is calculated based on the term frequency (TF) and inverse document frequency (IDF) of the query terms. TF scores are not as useful as they are in the field of IR since the terms (concepts) are oftentimes not repeated in a document (image). Therefore, the TF scores are often replaced by the confidence scores of the concepts extracted from various computer vision models [10]. The area of the object (or its the bounding box) can also be exploited as in aTFIDF, proposed by MyScéal [43].

As previously mentioned, cross-modal embedding models have offered a new way of measuring the similarity between images and text in lifelog retrieval systems. Cosine similarity between the search query and the images is directly used to rank the result. Some optimisation methods such as FAISS [19] or KNN search [43] can be used to speed up the search. Although there is some effort of fine-tuning the embedding models on lifelog models [40], large-scale pre-trained models are more robust and are often used directly or in a weighted ensemble in many systems, with Memento [1] being the most notable example.

In addition, support for temporal search is available in many LSC systems, enabling users to combine temporally related queries. For example, a user can search for 'eating apple before watching TV'. Multiple temporally ordered queries can be executed conditionally, with results re-ranked accordingly based on the scores of each query.

2.3 User Interactions

Relevance feedback mechanisms are utilised in several lifelog systems to enhance retrieval performance. This technique, commonly employed in Information Retrieval, enables users to provide feedback on search results. Users can label images as relevant or irrelevant, prompting the system to refine the search iteratively. For example, Exquisitor [20] incorporates relevance feedback, where a user's feedback influences the training of classifiers to retrieve a more relevant set of images. Similar mechanisms are applied in systems like SOMHunter [27], where a user selects relevant images, or in Exclusion of concepts, allowing users to specify concepts they wish to exclude from the results.

Visual similarity plays a significant role in lifelog retrieval systems, often used in conjunction with other retrieval methods. It can be employed to arrange result lists as seen in SOMHunter [27] and lifeExplore [34]; or to offer an alternative means for users to explore lifelog archives as seen in MyScéal [43] and LifeSeeker [30]. Visual similarity can be computed based on various factors, including low-level features such as colour histograms and SIFT, content-based features, or cross-modal embedding techniques.

Sketch-based search is a compelling method implemented in some lifelog systems originating from the video search community. Systems like lifeExplore [34] and vitrivr [37] allow users to draw a sketch of the image they seek. The system then returns visually similar images to the sketch. This method proves useful when users cannot describe an image they wish to locate in words but is less popular due to users' limited ability to convey a target image accurately.

Map-based search is another frequently employed technique in lifelog systems. Users can draw a rectangle on a map to narrow down the search space, focusing only on moments that occurred within that area. Systems like lifeExplore [34], MyScéal [43], and vitrivr [37] offer map-based search features, often utilising libraries like Leaflet.

2.4 Result Presentation Techniques

Result presentation in lifelog retrieval systems has evolved beyond conventional grid views. Some systems have explored alternative ways of presenting results to enhance user experience. For instance, lifeExplore [34] and SOMHunter [27] utilise Self-Organising Maps (SOMs) to arrange images in a 2D map, clustering them based on visual similarity. An autopilot navigation mode, introduced by lifeExplore in LSC'20, ensures every image in the target area is 'visited' once.

To mitigate visual clutter caused by similar images in lifelog data, event clustering is employed. Systems like MyScéal [43] perform offline event segmentation and display the highest-ranked image from each event. In contrast, LifeSeeker [30] dynamically clusters images during retrieval based on visual similarity and temporal proximity, presenting the top three images from each cluster.

Addressing the temporal aspect of lifelogs, MyScéal [43] proposes presenting results in triplets, showing the immediate previous and next events alongside the target event to provide more context. However, this approach is selectively used, specifically when temporal queries are specified, as adapted by E-MyScéalin LSC in 2022.

Memento [1] offers a unique visualisation of results through distribution charts, allowing users to modify filters directly by interacting with the charts. Transitional graph-based visualisation, proposed by LifeSeeker [30], shows location transitions between images in the result list, which is particularly useful for transportation-related queries.

The ability to browse a lifelog archive chronologically, often referred to as the ‘timeline view’, is crucial for users to explore their lifelog data. Many systems support this feature with varying levels of granularity and designs. Exquisitor [20], for example, designed a temporal context view resembling a video player with lifelog images as thumbnails. Users can play the video to view images chronologically and navigate to specific times by selecting thumbnails below the video. Adjusting the spacing between thumbnails is also possible, which can be implemented using different methods such as scaling factors, step sliders, or hierarchical levels, as seen in systems like LifeSeeker [30] and MyScéal [43].

2.5 Summary

In summary, lifelogging, as an extreme form of personal information management, offers a unique opportunity to gather and manage personal information. The annual lifelog search challenge event (LSC) serves as a remarkable example of how lifelog data can be harnessed and managed effectively. The techniques and components involved in the LSC represent the cutting edge of lifelog management, providing valuable insights into the future of personal data management.

3 Searching Personal Information

In the previous section we have seen the potential for analysing and searching lifelogs as demonstrated in the annual Lifelog Search Challenge. All the systems participating in this interactive live benchmarking event are highly engineered and specialist and not designed for consumer-grade use, yet.

In terms of our own personal data, we know that we can query and interrogate the individual, silo-ed and unconnected repositories which have generated that personal data which has come from individual sources, but is there anything in the personal information landscape that brings these together in some way to create an holistic overview ? A recent review of some personal data stores which can provide access to data from our online activities can be found at [13]. Some example systems included in that review include:

- **Digi.me**: A platform that lets users collect their personal data from online sources including social media, health, finance, and music, and to store it securely on their own devices or on the Digi.me cloud services. Users can then share their data with third parties on their own terms and get personalised insights and benefits from their data [17];

- **Solid**: A project that aims to give users full ownership and control of their data by creating a decentralised web where they can store their data in personal online data stores called pods and link them to applications that respect their privacy and preferences [18].;

- **Meeco**: A platform that enables users to create a digital identity wallet where they can store and manage their personal data such as identity documents, credentials, preferences, and consent. Users can then use their wallet to access online services and share their data securely and selectively [13].

While these examples of personal data stores may be useful for some use cases, they require the user her/himself to carry out the analysis of behaviour changes or patterns, and they are data stores rather than analysis platforms. Thus they are not for non-technical users. As an alternative to managing and analysing our own personal data we could invoke a third party to do the analysis, independently of the large internet companies. A platform to do this has been developed and demonstrated by Tuovinen [46] but that work is still early stage and not yet scaled up.

In the previous section we summarised the state-of-the-art in interactive lifelog search systems and showed that these systems have interactive question-answering capabilities which can lead to conversational retrieval around the topic of our own personal data. Outside the scope of the Lifelog search Challenge, the tools available to us for querying our own personal data do not yet exist for widespread use and those that are available no not have any forms of aggregation, or recommendation, or behavioural analysis, or self-reflection.

Within many areas of computing, the last year as seen a huge upsurge of interest in the use of large language models (LLMs) for applications in media creation, health and medicine, language processing, teaching and education, social interactions, science, and almost any domain where technology is used. While ChatGPT gets all the media coverage and popular focus, it is the fine-tuning of LLMs on narrow applications which can be restricted in their scope, that holds the greatest long-term potential. The use of the Retrieval Augmented Generation (RAG) [25] which allows an LLM to be fine-tuned on a limited training set could be applied to fine-tuning of an individual's personal data. This would allow the resulting system to leverage the advantages of LLMs including the capability for contextual understanding allowing a better understanding of a user's information need and search intent as well as the ability to simulate dialogue. Through careful prompt engineering it can also be used to unify across different sources of personal data and provide unified search, browsing and navigation across sources. The first example we see of this is when in September 2023 Google announced that BARD AI, their conversational retrieval tool which uses Google's LLM, can integrate with a user's emails, documents and other files as well as internet sources, to give personalised answers to questions. It is thus an obvious next step that conversational retrieval supported by large language models, will be used by the major internet companies as they way for us to access our personal data and that will certainly raise issues of data protection and privacy.

4 Conclusions

Personal information management involves the collection, storage, and control of digital data that represents our daily activities, including data generated by wearable devices, emails, and online interactions. This data can encompass various aspects of our lives, from physiological markers and indicators to our online behaviours. The purpose of gathering this personal data is primarily self-monitoring and self-tracking for health awareness. However, a significant portion of our personal data is collected and utilised by third parties, for targeted advertising and recommendations.

Personal information management tools and platforms should empower individuals to control our personal data, enabling us to gather, store, update, analyze, interpret, and sometimes share this data, regardless of its source. The development of such tools is essential because accessing and using personal data, even when it is our own data, is currently inconvenient and challenging due to the dispersed nature of our data across multiple sources. The paper presents an overview of an extreme form of personal data gathering called lifelogging and in particular how the annual lifelog search challenge highlights how innovative and useful searching through our lifelogs, our personal data, can be.

The article also highlights the potential of fine-tuning large language models (LLMs) for personalised data management, using techniques like Retrieval Augmented Generation (RAG). This approach could enable better contextual understanding, unified searching across various personal data sources, and the potential for conversational retrieval, thus revolutionising how individuals could access and manage their personal data. With the rise of large language models and conversational retrieval, major internet companies are likely to play a significant role in providing access to our personal data, raising important questions regarding data protection and privacy in the future.

References

[1] N. Alam, Y. Graham, and C. Gurrin. Memento 3.0: An enhanced lifelog search engine for LSC'23. In Proceedings of the 6th Annual ACM Lifelog Search Challenge, pages 41–46. Association for Computing Machinery, New York, NY, USA, 2023.

[2] A. Alsina, X. Giró, and C. Gurrin. An interactive lifelog search engine for LSC 2018. In Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge, pages 30–32, New York, NY, USA, 2018. Association for Computing Machinery.

[3] P. J. Barnard, F. C. Murphy, M. T. Carthery-Goulart, C. Ramponi, and L. Clare. Exploring the basis and boundary conditions of SenseCam-facilitated recollection. Memory, 19(7):758–767, 2011.

[4] G. Bell. A personal digital store. Commun. ACM, 44(1):86–91, Jan 2001.

[5] E. Berry, N. Kapur, L. Williams, S. Hodges, P. Watson, G. Smyth, J. Srinivasan, R. Smith, B. Wilson, and K. Wood. The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis: A preliminary report. Neuropsychological Rehabilitation, 17(4-5):582–601, 2007.

[6] K. Birch, D. Cochrane, and C. Ward. Data as asset? the measurement, governance, and valuation of digital personal data by big tech. Big Data & Society, 8(1):20539517211017308, 2021.

[7] P. Bruton, J. Langford, M. Reed, and D. Snelling. Classification of Everyday Living Version 1.0, 2019.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255. IEEE, 2009.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[10] M. Dogariu and B. Ionescu. Multimedia Lab @ ImageCLEF 2018 Lifelog Moment Retrieval Task. In CLEF (Working Notes), page 13, 2018.

[11] A. R. Doherty, K. Pauly-Takacs, N. Caprani, C. Gurrin, C. J. Moulin, N. E. O'Connor, and A. F. Smeaton. Experiences of aiding autobiographical memory using the SenseCam. Human–Computer Interaction, 27(1-2):151–174, 2012.

[12] B. Everson, K. A. Mackintosh, M. A. McNarry, C. Todd, and G. Stratton. Can wearable cameras be used to validate school-aged children's lifestyle behaviours? Children, 6(2):20, 2019.

[13] K. U. Fallatah, M. Barhamgi, and C. Perera. Personal Data Stores (PDS): A Review. Sensors, 23(3), 2023.

[14] C. Gurrin, A. F. Smeaton, and A. R. Doherty. Lifelogging: Personal big data. Foundations and Trends in Information Retrieval, 8(1):1–125, 2014.

[15] M. Harvey, M. Langheinrich, and G. Ward. Remembering through lifelogging: A survey of human memory augmentation. Pervasive and Mobile Computing, 27:14–26, 2016.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

[17] H. Janssen, J. Cobbe, and J. Singh. Personal information management systems: a user-centric privacy utopia? Published in Internet Policy Review (18 December 2020), 9(4):1–25, 2020.

[18] M. Jesús-Azabal, J. Berrocal, S. Laso, J. M. Murillo, and J. Garcia-Alonso. SOLID and PeaaS: Your Phone as a Store for Personal Data. In Current Trends in Web Engineering: ICWE 2020 International Workshops, KDWEB, Sem4Tra, and WoT4H, Helsinki, Finland, June 9–12, 2020, Revised Selected Papers 20, pages 5–10. Springer, 2020.

[19] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3):535–547, 2019.

[20] O. S. Khan, A. Duane, B. Þ. Jónsson, J. Zahálka, S. Rudinac, and M. Worring. Exquisitor at the lifelog search challenge 2021: Relationships between semantic classifiers. In Proceedings of the 4th Annual on Lifelog Search Challenge, pages 3–6. Association for Computing Machinery, New York, NY, USA, 2021.

[21] L. M. Koesten, E. Kacprzak, J. F. A. Tennison, and E. Simperl. The trials and tribulations of working with structured data: A study on information seeking behaviour. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, page 1277–1289, New York, NY, USA, 2017. Association for Computing Machinery.

[22] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and F.-F. Li. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. arXiv:1602.07332 [cs], Feb. 2016. arXiv: 1602.07332.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 2012.

[24] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision, 128(7):1956–1981, 2020.

[25] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33:9459–9474, 2020.

[26] J. Liu, M. Mitsui, N. J. Belkin, and C. Shah. Task, information seeking intentions, and user behavior: Toward a multi-level understanding of web search. In Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, pages 123–132, 2019.

[27] J. Lokoč, F. Mejzlik, P. Veselỳ, and T. Souček. Enhanced SOMHunter for known-item search in lifelog data. In Proceedings of the 4th Annual on Lifelog Search Challenge, pages 71–73. Association for Computing Machinery, New York, NY, USA, 2021.

[28] D. Lupton. Self-tracking, health and medicine, 2017.

[29] T.-H.-C. Nguyen, J.-C. Nebel, and F. Florez-Revuelta. Recognition of activities of daily living with egocentric vision: A review. Sensors, 16(1):72, 2016.

[30] T.-N. Nguyen, T.-K. Le, V.-T. Ninh, C. Gurrin, M.-T. Tran, T. B. Nguyen, G. Healy, A. Caputo, and S. Smyth. E-LifeSeeker: An interactive lifelog search engine for lsc'23. In Proceedings of the 6th Annual ACM Lifelog Search Challenge, pages 13–17. Association for Computing Machinery, New York, NY, USA, 2023.

[31] I. Nguyen Van Khan, P. Shrestha, M. Zhang, Y. Liu, and S. Ma. A two-level lifelog search engine at the LSC 2019. In Proceedings of the ACM Workshop on Lifelog Search Challenge, pages 19–23, 2019.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763. PMLR, 2021.

[33] L. Rossetto, O. Inel, S. Lange, F. Ruosch, R. Wang, and A. Bernstein. Multi-Mode Clustering for Graph-Based Lifelog Retrieval. In Proceedings of the 6th Annual ACM Lifelog Search Challenge, pages 36–40. Association for Computing Machinery, New York, NY, USA, 2023.

[34] K. Schoeffmann. LifeXplore at the Lifelog Search Challenge 2023. In Proceedings of the 6th Annual ACM Lifelog Search Challenge, pages 53–58. Association for Computing Machinery, New York, NY, USA, 2023.

[35] L. N. Signal, J. Stanley, M. Smith, M. Barr, T. J. Chambers, J. Zhou, A. Duane, C. Gurrin, A. F. Smeaton, C. McKerchar, et al. Children's everyday exposure to food marketing: an objective analysis using wearable cameras. International Journal of Behavioral Nutrition and Physical Activity, 14:1–11, 2017.

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[37] F. Spiess, R. Gasser, H. Schuldt, and L. Rossetto. The best of both worlds: Lifelog retrieval with a desktop-virtual reality hybrid system. In Proceedings of the 6th Annual ACM Lifelog Search Challenge, pages 65–68. Association for Computing Machinery, New York, NY, USA, 2023.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–9, 2015.

[39] W.-C. Tan, J. Dwivedi-Yu, Y. Li, L. Mathias, M. Saeidi, J. N. Yan, and A. Halevy. TimelineQA: A benchmark for question answering over timelines. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, Findings of the Association for Computational Linguistics: ACL 2023, pages 77–91, Toronto, Canada, July 2023. Association for Computational Linguistics.

[40] L.-D. Tran, N. Alam, Y. Graham, L. K. Vo, N. T. Diep, B. Nguyen, L. Zhou, and C. Gurrin. An exploration into the benefits of the CLIP model for lifelog retrieval. In Proceedings of the 19th International Conference on Content-based Multimedia Indexing, pages 15–22, 2022.

[41] L.-D. Tran, B. Nguyen, L. Zhou, and C. Gurrin. MyEachtra: Event-based interactive lifelog retrieval system for lsc'23. In Proceedings of the 6th Annual ACM Lifelog Search Challenge, pages 24–29. Association for Computing Machinery, New York, NY, USA, 2023.

[42] L.-D. Tran, M.-D. Nguyen, D.-T. Dang-Nguyen, S. Heller, F. Spiess, J. Lokoč, L. Peška, T.-N. Nguyen, O. S. Khan, A. Duane, et al. Comparing interactive retrieval approaches at the Lifelog Search Challenge 2021. IEEE Access, 11:30982–30995, 2023.

[43] L.-D. Tran, M.-D. Nguyen, B. Nguyen, H. Lee, L. Zhou, and C. Gurrin. E-Myscéal: Embedding-based interactive lifelog retrieval system for lsc'22. In Proceedings of the 5th Annual on Lifelog Search Challenge, LSC '22, page 32–37, New York, NY, USA, 2022. Association for Computing Machinery.

[44] L.-D. Tran, D. Nie, L. Zhou, B. Nguyen, and C. Gurrin. VAISL: Visual-aware identification of semantic locations in lifelog. In International Conference on Multimedia Modeling, pages 659–670. Springer, 2023.

[45] M.-T. Tran, T.-D. Truong, T. D. Duy, V.-K. Vo-Ho, Q.-A. Luong, and V.-T. Nguyen. Lifelog moment retrieval with visual concept fusion and text-based query expansion. In CLEF (Working Notes), 2018.

[46] L. Tuovinen and A. F. Smeaton. Privacy-aware sharing and collaborative analysis of personal wellness data: Process model, domain ontology, software system and user trial. Plos ONE, 17(4):e0265997, 2022.

[47] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10):78–85, 2014.

[48] G. Wilson, D. Jones, P. Schofield, and D. J. Martin. The use of a wearable camera to explore daily functioning of older adults living with persistent pain: Methodological reflections and recommendations. Journal of Rehabilitation and Assistive Technologies Engineering, 5:2055668318765411, 2018.

[49] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. Advances in Neural Information Processing Systems, 27, 2014.

[50] L. Zhou, D.-T. Dang-Nguyen, and C. Gurrin. A baseline search engine for personal life archives. In Proceedings of the 2nd Workshop on Lifelogging Tools and Applications, LTA '17, pages 21–24, New York, NY, USA, 2017. Association for Computing Machinery.

[51] Q. Zhou, D. Wang, C. Ní Mhurchu, C. Gurrin, J. Zhou, Y. Cheng, and H. Wang. The use of wearable cameras in assessing children's dietary intake and behaviours in China. Appetite, 139:1–7, 2019.

Episodic Memory Integration of Personal Data in YourDigitalSelf

Varvara Kalokyri, Alexander Borgida, Amelie Marian
Department of Computer Science, Rutgers University
v.kalokyri@cs.rutgers.edu, borgida@cs.rutgers.edu, amelie@cs.rutgers.edu

Abstract

Human episodic memory consists of past personal experiences organized into episodes. It is a crucial part of our lives, influencing our identity and interpersonal relationships. A large variety of digital devices and apps capture personal digital traces (PDTs) of our daily experiences, including messages, time and/or location data, browsing and purchase history, etc. Our goal is to reconstruct meaningful episodes from PDTs, in order to improve memory recall and aid individuals with memory problems. In addition to acquiring PDTs, this process faces many problems, including: the diversity and very large number of PDTs; describing the episode types of interest; finding and combining the relevant but very sparse set of PDTs that provide partial evidence for a particular episode having occurred.

To address these challenges, the YourDigitalSelf project employs semantic modeling and integration techniques to associate PDTs with events, themselves organized into prototypical narratives/plans called scripts. It introduces a formalized conceptual modeling language, hierarchical script definitions, and an evidence-based bottom-up approach to reconstruct script instances. A first evaluation of our implemented methodology using a mobile application demonstrates relatively successful integration of diverse digital traces and memory enhancement.

1 Introduction

In today's world, digital devices are seamlessly integrated into our daily lives, transforming the way we engage with people and our surroundings. These devices generate and store various kinds of Personal Digital Traces (PDTs), such as e-mails and messaging, location data, calendar entries, check-ins, reviews, web searches, and purchase history. These PDTs, serve as a partial record of individuals' lives, but pose a challenge to users because the full collection is very large, highly heterogeneous, and distributed in space and time, making it difficult to access, search, and learn from.

Drawing inspiration from Vannevar Bush's visionary "memex" concept, which can be seen as a personal data collection system, our work aims to exploit the potential of organizing personal data in order to aid remembering past events. By connecting the dots between personal information items, users can effectively retrieve specific memories or information in a coherent and meaningful way.

The goal is to discover evidence for users' actions from the digital traces these actions leave behind, and organize them into coherent episodes that look like high-level complex and meaningful workflows, enhancing the user's "autobiographical memory". This could be especially beneficial for remembering routine experiences, which has been shown to be harder to recall [1]. For example, users could use connections in their data to not only quickly retrieve digital artifacts, such as pictures from a trip, minutes from a meeting, or pictures from a birthday party, but also to retrieve specific information about events, such as the name of a restaurant they went to, the title of a movie they watched, etc.

Our research takes a step forward in supporting human memory and recall through the concept of episodic narratives, rooted in psychology and cognitive science. It enables users to organize PDTs into cohesive episodes, creating a personal knowledge base that users can query to remember specific events and their by-products.

Additionally, it provides users with narratives to stimulate their memory, which can be particularly helpful for those with memory difficulties.

To achieve this, we need to acquire and integrate all the heterogeneous data from various sources, including files used by apps on our devices (e.g., search history), and through service APIs. We focus in this paper on the following aspects:

1. A conceptual model to represent entities, especially PDTs, which helps extract semantically similar aspects (e.g., who? when? where?), thus integrating the heterogeneous sources.
2. A conceptual model for atomic activities/events. Note that PDTs may provide weighted evidence signaling the occurrence of an activity instance.
3. A conceptual model of complex activities, called scripts. These contain stereotypical commonsense plans, whose instances can again be signaled by the presence of either PDTs or activities/sub-subscripts. A novel notion of inheritance and subsumption helps one to create and organize libraries of scripts.
4. An algorithm for recognizing instances of atomic and composite events/scripts. The bottom-up nature of the algorithm, which merges evidence for components, allows us to deal with missing parts or atypical instances of scripts; a scoring scheme accounts for the varied strength of evidence provided by PDTs or script steps.
5. An Android app which instantiates our ideas, asking users to provide access to a specific set of data sources, loading the data into the PDT concepts, and running the above algorithm. This application is also used to evaluate our approach, through user studies and surveys.

We immediately acknowledge the sensitive nature of the personal data we are dealing with, and the very important privacy issues that they raise. All information obtained to evaluate our work resides on individuals' own mobile phones without disclosing any personal information.

Our study results demonstrate the potential of our approach, showing its effectiveness in integrating digital traces from different sources into coherent episodes and enhancing users' memory of past actions. Our ultimate goal is to empower users to create their personal knowledge base and better manage their digital memories.

2 An episodic script model for personal digital traces

Our approach to data integration and memory recall draws on principles from Cognitive Science and Psychology, specifically in the study of how individuals reminisce about their past experiences. These psychological findings have yielded several key observations.

The Natural Mnemonic Framework: It is observed that when people reflect on their past, they often do so by responding to a set of fundamental questions, which we refer to as the 'w5h' questions: what, when, where, who, why, and how [2, 3]. For instance, when attempting to recollect a significant event like one's sixth birthday, individuals instinctively seek to reply to these contextual cues. We use this inherent cognitive framework as the foundation of our data integration, by aligning PTDs and conceptual model instances along these essential questions.

Episodic and Semantic Memory: Psychological research [4, 5] distinguishes between two primary forms of memory. Semantic memory pertains to the accumulation of general world or cultural knowledge that individuals collect over their lifetimes (e.g., it encompasses the understanding that payment is required when making a purchase) and episodic memory involves the capacity to re-live specific past events or episodes, (such as the memory of celebrating one's 40th birthday by dining out). Our approach combines the representation of both semantic memory (scripts) and episodic memory (script instances). This fusion allows us to assist users in recalling information related to past events. To achieve this, it is imperative to connect individual atomic events and organize them into broader activities, even when these events occurred at different times. This linkage facilitates the retrieval of high-level activity-based memories, enhancing the user's ability to reconstruct their personal history effectively.

2.1 Conceptual modeling of personal digital traces

As mentioned, our objective is to develop a framework for modeling digital traces that unifies information based on the 'w5h' questions. To achieve this, we start with a standard-centered conceptual modeling language, as first presented in [6]. This language employs entity classes, such as 'Documents' and 'Persons,' which contain individual instances. These instances possess atomic valued features, such as the size of a document, and are interconnected through properties, such as those associated with an Email, including 'from,' 'to,' 'date', 'subject,' and 'body.'

Within this modeling framework, classes can be specialized into subclasses. For instance, 'PERSON_WITH_EMAIL' is a specialization of the 'PERSON' class, and new properties, such as 'email_address', can be introduced, or existing properties may be further constrained.Notably, properties can also undergo specialization, where, for example, 'from' and 'to' are sub-properties of 'who'. A large subset of the properties of PDTs can be viewed as specializations of w5h, when these are viewed as properties themselves, resulting in the integration of heterogeneous object schemas.

Note however that there is no sensible way to ask questions like "when?" or "who?" of many objects. However, by associating an object with a primitive action, it becomes possible to derive the 'when' aspect from the action. For instance, we may not directly ask 'who email', but we can ascertain 'who sent an email'. This requires attaching documents to primitive actions.

2.2 Conceptual modeling of atomic and complex events

A script-based model is adopted for the purpose organizing atomic actions into coherent episodes. This describes event flows, encompassing steps that humans intuitively understand. For example, for going out to eat, these include steps like organizing the outing, making reservations, taking transportation, and making payment. These steps often generate PDTs, serving as evidence for the activities. For example, we then want to group emails concerning a particular dinner, a reservation that was made at a restaurant, a Facebook check-in with photos, and a credit card payment. The atomic actions that generated these documents are organized as part of the narrative for going out to eat. Essentially, each data trace presents evidence of various strengths for the occurrence of atomic event instances.

To achieve this, we need a set of higher-level plans that the users and their community frequently engage in, which describe the connections between all those events. The idea of scripts is inspired by the work of Schank and Abelson in AI[7], which are defined as plans describing stereotyped sequences of actions that describe some well-known situation.

For describing such scripts we use the standard notation used in business process management (BPM)[8]. As such, scripts are characterized by their goals, participant information, w5h aspects, component sub-scripts, and allowed sequencing and timing of sub-scripts/atomic actions - the latter captured in BPM.

However, in order to make a fully functional system, we need script definitions for many variants of everyday activities. These plan descriptions serve as prototypes due to the vast variability inherent in such situations, making it impractical to predetermine every possible variation. This will be facilitated by placing the script definitions in an ontology and then using specialization and inheritance, so that more specific processes, such as GoingOutToEat, GoingOutToConcert, GoingToOpera, etc. are obtained by indicating only differences from a common superclass, GoingOutForEntertainment. In turn, GoingOutToEat can be specialized to EatingOutAtRestaurant, EatingFastFood, etc. By using specialization hierarchies we can organize large knowledge bases, support reuse and change propagation, as well as prevent repetition errors.

Description logics (DLs) are ideal for the creation of specialization of such hierarchies via reasoning. We have introduced a new extended regular expression-based Description Logic [9] able to represent the control flow semantics of structured BPM .For example, specific concepts A and B can be specified, and then we can assert that a class A is subsumed by B, or discover this relationship by reasoning, meaning that all instances of A are instances of B. We can also detect inconsistencies, (e.g. we might give a description of going out to a rock concert which contradicts going out to a concert. This will provide evidence that this is not a proper specialization).

Instance recognition for DLs by the use of inference could be used for our purposes for recognizing activities. However, this approach cannot be used because: not all steps occur in every instance, their order may vary, and some steps may leave no digital trace. Instead, a bottom-up approach is proposed where partially instantiated scripts are constructed from individual PDTs, and compatible instances are merged, accumulating information and strengthening evidence for them. This approach is flexible and accommodates various types of scripts, making it suitable for different scenarios. In the next section, the algorithm for creating episodes from PDTs is described.

3 Algorithm for Episode Recognition

Our algorithm, starts with a specific script (S) and a comprehensive set of PDTs (P). The primary objective is to generate potential script episodes, which are instances of (S), and establish connections with the PDTs.

Step 1. Script Specification: A script specification has the following components: a top-level (outer) script (e.g. ‘EatingOut’ script), which we want to instantiate; several component sub-scripts and atomic actions; as well as sequencing relationships among them. In addition, every script contains its w5h properties.

For example, ‘EatingOut’ comprises events like inviting people, making reservations, going to a restaurant, paying the bill, and sharing photos. Some of these events (e.g., making a payment) offer strong evidence that the person indeed participated in this activity/script. On the other hand, receiving an email mentioning “dinner" or “lunch" provides weaker evidence for planning to dine out, which, in turn, offers weaker evidence of actually going out since the plan may remain incomplete or canceled. Our algorithm employs the concept of strong and weak evidence to prioritize potential script instances.

We gather information about the strength of evidence in a similar manner to acquiring common-sense knowledge about events within a given context. For that purpose, we have developed a self-evaluating multiplayer web-based game for collecting digital trace descriptions for every step in a scripted scenario and potential strength of evidence [10]. It is essential to note that there are limited “objective" methods for scoring evidence because scripts represent imprecise, culturally-shared, commonsense knowledge describing typical human activities.

Step 2. Retrieving Document Set for Script Instantiation: Once a script is parsed, the subsequent task involves identifying the set of documents, denoted as (D), that serve as evidence for the occurrence of its instances. These documents correspond to PDTs and act as “noisy sensors," indicating potential instances of corresponding scripts or atomic actions for which strong evidence exists.

To locate documents that correspond to a specific piece of evidence, it is crucial to identify specific cues within the documents. These cues can take the form of verbs to search for (e.g., “eat" in an email to identify an “Initiate Going Out" event) or attributes and metadata that a document may possess (e.g., categorizing a payment as “Restaurant" or “Supermarket" to find evidence for “Making a Payment at a restaurant or supermarket respectively). To ensure the replicability of this process for various scripts, the w5h components of the script or atomic action are obtained from its FrameNet frames[11]. Standard sources of synonyms and hyponyms such as WordNet and ConceptNet5 [12, 13] are then utilized to identify additional words for searching. The generated lists of words and phrases from this process are stored in a text file and used to retrieve all potentially relevant documents.

Finally, the set of documents (D) is pre-processed by: (i) expanding information (e.g., terms like “tomorrow”, “on Friday”, are made absolute dates); (ii) performing entity resolution for people and places (who and where dimension) using Stanford’s Entity Resolution Framework [14]; (iii) grouping certain kinds of documents (e.g., related email threads, or related sequences of tweets) into a single individuals d in (D); (iv) finding the places/venues that the user has visited from the geo-location coordinate history (gps) [15].

Step 3. Creating Initial Script Instances: Every document in the set (D) serves to instantiate atomic actions or sub-scripts, creating candidate instances of the outer script (S) in a bottom-up fashion. Similarly, the w5h properties are propagated from the document into the atomic actions and the script hierarchy.

Step 4. Merging Script Instances: Merging script instances is a distinctive feature of our system, using multiple sources of evidence for the same script instance. Each script is assigned “keys" that rate the importance of w5h (sub)properties in identifying instances. For ‘EatingOut’, the keys are ‘whereEatingOccurred’, ‘whenEatingOccurred’, and, to a lesser extent, ‘who’. The ‘what’ and ‘how’ properties of this script are not important because they would often lead to incorrect merging (e.g., two instances of eating pasta (‘what’) need not be merged). Since every script can have different keys, this has to be explicitly mentioned in the algorithm. When two instances share similar keys, they are considered candidates for merging. The w5h property fillers are combined, and a score is computed for the merged instance based on Hooper’s rule [16] for combining probabilistic evidence.

The key point here is that the merging process cannot be completed in a single step. After the script instance is established with a certain degree of certainty, additional PDTs may be collected as part of the instance when examining the script definition.

4 A mobile app for YourDigitalSelf

So far, we have given in Section 2 an ontology of entities (especially for PDTs), and of atomic and composite events/scripts, which involve these entities. In Section 3, we discussed how specific PDT concepts provide evidence of varying strength for specific scripts, and an algorithm which, given a script (episode type), uses this information to instantiate very small episode fragments and then merge them into larger ones, with associated strengths.

To build our YourDigitalSelf Android app (see [17] for details, including screen shots), we first need to have users provide access to the data sources from which the PDT instances will be created, and then interactively choose a specific script for which they wish to see rank-ordered instances retrieved. The types of PDTs along with the data sources currently supported by the app are the following:

- Messaging: Messenger, phone text messages
- Social Media: Facebook, Instagram
- Email: Gmail
- Calendar: Google Calendar
- Financial Data: Plaid API and directly downloaded .csv files from bank institutions
- Location Data: Google Maps location history, GPS data
- Photos: Google Photos

The system then gathers the raw data and uses it to populate the corresponding PDT concepts. In general, each API has a standard format, but in the not-infrequent case where APIs change, this mapping must be modified. Recall from Section 2.1 that properties for the ’w5h’ questions and their sub-properties are the key to semantically “integrating” the extremely heterogeneous forms of PDTs.

When asked, the app then essentially runs the algorithm in Section 3 to gather episodes (script instances) of the script asked for, and then displays them via a GUI. Note that each episode has attached a variety of PDTs that provided evidence for it; this is a second way in which related data is “integrated”.

Instances of the concepts in conceptual model cannot be efficiently stored directly on the mobile phone, and must instead be mapped to a relational database natively managed by Android. The mapping of the different concept types to tables is declaratively described using pairs of conjunctive queries.

After the mapping to our model, the next step involves designing an indexing structure to facilitate data queries. We leverage a combination of full-text and column indexing techniques provided by the Android platform. The indexing occurs whenever the user decides to include data in the app, either manually or through automated periodic updates.

In general, the project has made a significant effort to support extensibility and modification by declarative representation of other aspects as well, such as evidence to look for in scripts and clues to search in PDTs, as well as making scripts parametric/generic (e.g, Book$<$T$>$, T=“accommodation", “concert",...).

4.1 Experimental Setup

To assess the effectiveness of our approach, we conducted experiments using real user data with the primary objective of detecting instances where individuals engaged in dining out at restaurants. We opted for this specific script example because it offers a rich variety of PDTs and shares commonalities with other scripts related to entertainment, such as attending theaters or concerts.

We employed the YourDigitalSelf Android application for this purpose, which allows users to install it locally on their devices, and it is designed to serve future research purposes as well. It is paramount to clarify that our approach did not entail direct access to users' personal data. Instead, users provided responses solely in the form of Yes/No answers, without revealing any personal information. The application is configured to collect PDTs from a diverse range of services, including messaging, social media, email, calendar, financial data, location data, and photos.

We ran an in-depth study[1] with 16 users. Before the experiment, we asked participants to try to remember the occasions of them going out to eat at restaurants and then carefully go over their past month's digital information, writing down their outings, including name of the restaurant - where, date they went - when, with whom they went - who. We used this information as a proxy for recall. Then during the experiment participants were shown all candidate script instances of them going out to eat at restaurants, and had to indicate Yes/No for each of the instances first, and then for each of the w5h properties.

For a comprehensive description of the entire study and its results, we refer interested readers to the detailed study in [18].

4.2 Experimental Evaluation

The evaluation of our approach focuses on the assessment of our matching algorithm and scoring function which encompasses two key aspects: (1) the effectiveness of our approach in identifying script instances and (2) its proficiency in recognizing and abstracting ‘When,’ ‘Where,’ and ‘Who’ information from sub-scripts and atomic actions into the overarching script instance.

For both facets of evaluation, we employ two distinct dimensions of relevance:

1. **Binary Relevance**
   A script instance is deemed relevant if the user actually engaged in dining out at a restaurant, even if the ‘w5h’ information was only partially correct whereas each ‘w5h’ information is considered relevant only if it is precisely correct (neither a subset nor a superset).
2. **Graded Relevance**
   A proposed script instance falls into one of the following categories:
   Exactly relevant: When the user indeed dined out at the specified restaurant.
   Relevant but too broad: When the identification of a restaurant outing is accurate, but the user's purpose for the visit did not involve dining (e.g., for dancing or socializing with a friend).
   Relevant but too narrow: When a restaurant outing is correctly identified, but the user did not stay at the establishment to eat (e.g., it was a takeout order).
   Partially relevant: When a planned outing was correctly identified, but the user did not follow through with it.
   Not relevant: When the identified instance does not pertain to a restaurant outing.

[1] Our study has received approval from the Rutgers University Institutional Review Board (IRB) committee.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16
Our approach	13	15	10	5	**14**	**24**	13	6	19	17	19	7	11	5	17	15
User memory	7	8	8	5	8	10	5	5	9	9	13	5	6	3	8	9
User data	13	14	11	5	**7**	**14**	11	6	20	14	15	5	11	5	16	14
Recall	1	1	**0.91**	1	1	1	1	1	**0.95**	1	1	1	1	1	1	1

Table 1: Number of identified EatingOut actions by users vs number of correct events our approach retrieved per user as a proxy for Recall

	Sources	**Precision**
User #1	Social Media, Calendar, Financial Data, Location Data, Google Photos	0.87
User #2	Social Media, Location Data, Financial Data	**0.94**
User #3	Email/Messaging, Social Media, Calendar, Financial Data	0.66
User #4	Email/Messaging, Social Media, Calendar, Financial Data, Location Data, Google Photos	0.66
User #5	Email/Messaging, Social Media, Calendar, Financial Data, Location Data	0.74
User #6	Social Media, Calendar, Financial Data, Location Data, Google Photos	0.89
User #7	Email/Messaging, Social Media, Calendar, Financial Data, Location Data, Google Photos	0.76
User #8	Email/Messaging, Social Media, Calendar	**0.6**
User #9	Email/Messaging, Social Media, Calendar, Financial Data, Google Photos	0.74
User #10	Email/Messaging, Social Media, Calendar, Financial Data, Location Data	0.81
User #11	Email/Messaging, Social Media, Calendar, Financial Data, Location Data, Google Photos	0.76
User #12	Social Media, Calendar, Financial Data, Location Data, Google Photos	0.86
User #13	Email/Messaging, Social Media, Calendar, Financial Data, Google Photos	0.73
User #14	Email/Messaging, Social Media, Calendar, Financial Data, Location Data, Google Photos	0.83
User #15	Email/Messaging, Social Media, Calendar, Financial Data, Google Photos	0.77
User #16	Email/Messaging, Social Media, Calendar, Financial Data, Location Data, Google Photos	0.79
Total		**0.78**

Table 2: Overall precision for each user

We establish an analogous framework for assessing the relevance of ‘When,’ ‘Where,’ and ‘Who’ (‘w5h’) information, classifying them into relevant categories based on the same criteria.

To measure the performance of our approach, we rely on the following metrics, considering both binary and graded relevance:

- **Percentage of instances retrieved:** This metric indicates the percentage of all ‘EatingOut’ events identified by users that our scripts successfully retrieved. It serves as a proxy for recall.
- **Mean Average Precision @ k (MAP@k):** Utilizing MAP as a binary relevance assessment, we determine the percentage of the top-k identified script instances that correspond to actual ‘EatingOut’ events.
- **Normalized discounted cumulative gain (nDCG):** This metric is employed to assess the ranked results while considering graded relevance, as described earlier.

4.3 Experimental Results

Table 1 shows the number of correct EatingOut instances retrieved by our approach compared with the number identified by users from memory, and by searching their PDTs. A first observation is that the results clearly indicate how hard is for users to recall their outings, either from memory or even when asked to go through their digital information. Our approach identified more correct instances than the users were able to recall and find in all cases but two (users 3 and 9 in bold). In addition, users found it hard to search through their data, since most of the applications have keyword-based search. Users 5 and 6 in bold, who were able to retrieve only half of their outings, clearly show this issue.

Table 2 shows the overall precision of the identified script instances along with the sources each user incorporated in the application. Our approach achieves a total precision of 78% for all the users. User#2 achieved

	when	where	who
MAP	0.85	0.81	0.21

Table 3: MAP for when, where, who dimensions for all users

the highest precision of all, since the sources they chose to include in the study contained bank transactions, google maps location history, Instagram and Facebook, sources that tend to be of high quality, whereas User#8 achieves the lowest precision of all since they included their private phone text messages without any high-quality source.

However, retrieval systems typically return results in a ranked order, and users expect the first few results to be the most relevant. We now look at the quality of the returned answers by evaluating the mean average Precision@k metric.

Figure 2 shows the Mean Average Precision@k for all the identified script instances for all the users. As shown, our approach achieves a really good precision even for low values of k. The reason for that is that our approach does include many different kinds of sources and is able to account for all the different kinds of user behavior.

Figure 3 shows the normalized discounted cumulative gain (nDCG) for the ranked results when taking into consideration the graded relevance. The nDCG was computed by normalizing the DCG@k with the ideal DCG value or IDCG@k. It is clear that our ranking quality is high, and our approach is able to recognize and distinguish highly relevant PDTs in favor of irrelevant PDTs.

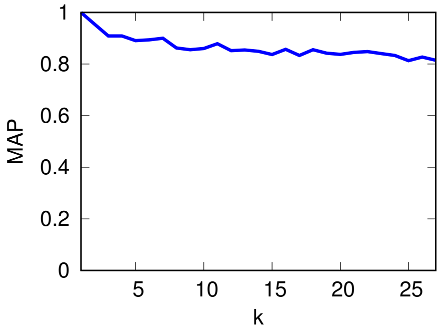


Figure 2: MAP@k for the recognized instances

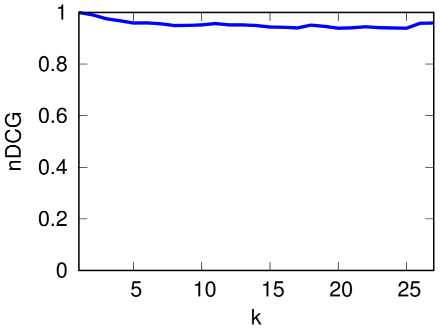


Figure 3: nDCG@k for the recognized instances

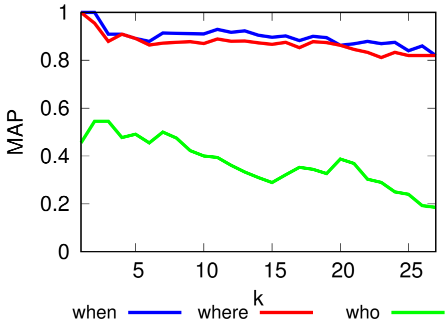


Figure 4: MAP@k for when, where, who dimensions

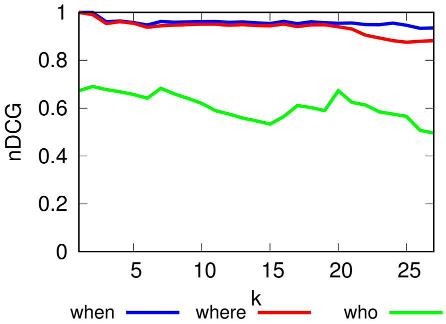


Figure 5: nDCG@k for when, where, who dimensions

We then report the same metrics (MAP, MAP@k, and nDCG@k) on the when, where, and who dimensions.

Table 3 shows the MAP for the three dimensions for all users. We observe that the when and where dimensions are easier to extract than the who dimension due to the metadata that the PDTs have. On the other hand, the who dimension is more difficult to extract. For example, we observe in figure 4 how the precision for the who dimension drops significantly as k increases. This happens due to the fact that for low values of k, the score of the instances is low, which means there are not many digital traces to account for these instances, therefore either there is no information, or our approach either lacks or recognizes more people in an outing. This is actually demonstrated in figure 5, which shows the normalized discounted cumulative gain for the ranked results where we can now observe how much better the accuracy is for the 'who' dimension.

5 Literature Review

In this section, we provide a brief overview of various related research areas, highlighting the interdisciplinary nature of our work.

Personal Information Management: Personal Information Management (PIM) research, dating back to the 1980s, focuses on assisting users in managing and organizing digital data. Researchers have suggested PIM interfaces for web activities [19–21], email [22, 23], and local files [24, 25]. Existing systems often utilize domain-specific ontologies to identify relevant objects and their relationships [26, 27]. In contrast, our approach is dynamic, emphasizing the integration of PDTs through narrative connections, shifting the focus from static information to dynamic data integration based on the 'who, what, when, where, why, and how' (w5h) properties.

Process Representations: Our interest lies in representing autobiographic events and their instances. This area intersects with various formal process representation languages, graphical notations, and complex event recognition [28],[8]. While traditional formalisms concentrate on enacting processes, we focus on descriptive formalisms that enable script instance recognition, underlining the challenges of recognizing multiple, concurrent, and personalized script instances. This leads us to several relatively closely related areas such as Activities of Daily Living, Ambient Intelligence, Pervasive Computing, and LifeLogging. The following are some papers that survey an entire field (e.g. LifeLogging [29]) or the use of ontologies and inference [30–32].

Activities of Daily Living, Pervasive Computing, LifeLogging, and Memory Tools: Memory aids play a crucial role in rehabilitation, particularly in improving prospective memory. Existing tools like Sensecam [33], and Kalnikaite's browser [34] are akin to our approach, as they aim to trigger autobiographical memory by recording images and linking them to daily activities. Other tools include the MemoClip , the Cyberminder, and Memory Glasses [35–37]. However, we stand out by using pre-existing digital traces rather than capturing new data. In addition, the area of life-logging is quite similar, and is surveyed in [38, 39]. Bell has pioneered this field with MyLifeBits [40] for which he digitally captured all aspects of his life. A recent paper [41] focuses on the creation of lifelogs, positioning them as a critical resource for personal assistants to provide tailored advice within specific contexts. Another relevant paper is that of Meditskos et al [42] which used the technique of multi-sensory data analysis along with egocentric video recording from a bracelet to aid the dementia patients recognize their daily living activities.

Planning: Plan recognition, a prominent field in AI, focuses on recognizing plans from actions [43, 44]. Some similarities exist with our approach, particularly in recognizing multiple, concurrent, and interleaved script instances. We emphasize data integration, while traditional planning approaches are concerned with action sequences and transitions.

Personal Digital Assistants: Personal Digital Assistant (PDA) systems, exemplified by Amazon Alexa, Siri, and Google Now, share commonalities with our work. These systems are engineered to provide users with event reminders based on their personal data, often driven by commercial objectives. However, their functionality is confined to utilizing data within the vendors' proprietary ecosystems. It's noteworthy that these PDA systems primarily emphasize future tasks, offering reminders based on time or event triggers [45]. In contrast, our current application scenarios center around retrospective tasks, particularly the organization of past memories.

Nonetheless, our data integration approach can complement prospective methods for tailoring activity recognition.

Our approach emphasizes the dynamic integration of Personal Digital Traces (PDTs) through narrative connections and the recognition of multiple, concurrent, and personalized script instances, setting it apart from conventional methods in these related areas.

6 Conclusion

In summary, our research takes steps forward in advancing autobiographical memory by linking diverse PDTs based on their shared objectives, organizing them into cohesive episodic narratives. We introduced a conceptual model encompassing entities, PDTs, and the actions that create them. Scripts, which describe complex activities and are amenable to reasoning via description logic, played a pivotal role. Our merging algorithm identifies script instances, providing a valuable resource for users. Through the YourDigitalSelf app, we evaluated and underscored the potential and promise of our approach.

However, several directions for future research and development are worth considering. These include conducting a more extensive array of experiments encompassing a variety of scripts and involving a larger, more diverse population. Additionally, we recognize the necessity of incorporating Natural Language Processing (NLP) analysis to address instances where users mention activities indirectly. This could be potentially tackled with the use of today's natural language capabilities (such as openAI's chatGPT). Beyond its implications for personal data management and memory enhancement, our work offers opportunities for behavioral researchers. The combination of personal digital traces, AI, and self-reported surveys holds the potential to yield fresh insights into mental health assessment. Our tools are open source and publicly accessible on GitHub. This availability paves the way for educators and students to engage with their personal databases, fostering the development of novel, useful applications. Finally, our tools hold potential significance in the medical field, particularly for medical experts conducting studies related to memory rehabilitation in patients facing cognitive difficulties.

References

[1] N. M. Bradburn, L. J. Rips, and S. K. Shevell, "Answering autobiographical questions: The impact of memory and inference on surveys," Science, vol. 236, no. 4798, pp. 157–161, 1987.

[2] D. L. Schacter, The seven sins of memory: How the mind forgets and remembers. HMH, 2002.

[3] W. Jones, "Personal information management," Annual review of information science and technology, vol. 41, pp. 453–504, 2007.

[4] E. Tulving, "10. episodic and semantic memory," Organization of memory/Eds E. Tulving, W. Donaldson, NY: Academic Press, pp. 381–403, 1972.

[5] M. A. Conway and D. C. Rubin, "The structure of autobiographical memory," Theories of memory, vol. 103, p. 137, 1993.

[6] V. Kalokyri, A. Borgida, A. Marian, and D. Vianna, "Semantic modeling and inference with episodic organization for managing personal digital traces - (short paper)," in Proc. ODBASE 2017, pp. 273–280, 2017.

[7] R. Abelson and R. C. Schank, "Scripts, plans, goals and understanding," An inquiry into human knowledge structures New Jersey, vol. 10, 1977.

[8] S. Goedertier, J. Vanthienen, and F. Caron, "Declarative business process modelling: principles and modelling languages," Enterprise Information Systems, vol. 9, no. 2, pp. 161–185, 2015.

[9] A. Borgida, V. Kalokyri, and A. Marian, "Description logics and specialization for structured bpmn," in Int. Conf. on Business Process Management, pp. 19–31, Springer, 2019.

[10] V. Kalokyri, A. Borgida, and A. Marian, "One of us: a multiplayer web-based game for digital evidence acquisition of scripts through crowdsourcing," in Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, pp. 187–196, 2023.

[11] C. J. Fillmore, C. R. Johnson, and M. R. Petruck, "Background to framenet," International journal of lexicography, vol. 16, no. 3, pp. 235–250, 2003.

[12] G. A. Miller, "Wordnet: a lexical database for english," Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.

[13] H. Liu and P. Singh, "Conceptnet: a practical commonsense reasoning tool-kit," BT technology journal, vol. 22, no. 4, 2004.

[14] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom, "Swoosh: a generic approach to entity resolution," The VLDB Journal, vol. 18, no. 1, pp. 255–276, 2009.

[15] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma, "Mining user similarity based on location history," in SIGSPATIAL Advances in geographic inf. systems, p. 34, 2008.

[16] G. Shafer, "The combination of evidence," International Journal of Intelligent Systems, vol. 1, no. 3, pp. 155–179, 1986.

[17] V. Kalokyri, A. Borgida, and A. Marian, "Yourdigitalself: A personal digital trace integration tool," in Proceedings of the 27th CIKM conference, pp. 1963–1966, ACM, 2018.

[18] V. Kalokyri, A. Borgida, and A. Marian, "Supporting human memory by reconstructing personal episodic narratives from digital traces," in Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, pp. 453–464, 2022.

[19] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins, "Stuff i've seen: A system for personal information retrieval and re-use," in (SIGIR'03), 2003.

[20] V. Kaptelinin, "Umea: translating interaction histories into project contexts," in SIGCHI conference on Human factors in computing systems, pp. 353–360, 2003.

[21] H. Murakami and K. Mitsuhashi, "A system for creating user's knowledge space from various information usages to support human recollection," International Journal of Advancements in Computing Technology, vol. 4, no. 22, 2012.

[22] T. Ayodele, G. Akmayeva, and C. A. Shoniregun, "Machine learning approach towards email management," in World Congress on Internet Security (WorldCIS), IEEE, 2012.

[23] S. Whittaker, V. Bellotti, and J. Gwizdka, "Email in personal information management," ACM, vol. 49, no. 1, pp. 68–73, 2006.

[24] D. Barreau and B. A. Nardi, "Finding and reminding: file organization from the desktop," ACM SigChi, vol. 27, pp. 39–43, 1995.

[25] D. K. Barreau, "Context as a factor in personal information management systems," Journal of the American Society for Information Science, vol. 46, no. 5, pp. 327–339, 1995.

[26] D. R. Karger, K. Bakshi, D. Huynh, D. Quan, and V. Sinha, “Haystack: A general-purpose information management tool for end users based on semistructured data,” in CIDR’05, pp. 13–26, 2005.

[27] V. Katifori, A. Poggi, M. Scannapieco, T. Catarci, and Y. E. Ioannidis, “Ontopim: how to rely on a personal ontology for personal information management,” in ISWC Wkshop. on The Semantic Desktop, pp. 258–262, 2005.

[28] W. M. P. van der Aalst and A. H. M. ter Hofstede, “YAWL: yet another workflow language,” Inf. Syst., vol. 30, no. 4, pp. 245–275, 2005.

[29] C. Gurrin, A. F. Smeaton, and A. R. Doherty, “Lifelogging: Personal big data,” Foundations and Trends in Information Retrieval, vol. 8, no. 1, pp. 1–125, 2014.

[30] A. Bikakis, T. Patkos, G. Antoniou, and D. Plexousakis, “A survey of semantics-based approaches for context reasoning in ambient intelligence,” in European Conference on Ambient Intelligence, pp. 14–23, Springer, 2007.

[31] N. D. Rodríguez, M. P. Cuéllar, J. Lilius, and M. D. Calvo-Flores, “A survey on ontologies for human behavior recognition,” ACM Computing Surveys, vol. 46, no. 4, p. 43, 2014.

[32] E. Alevizos, A. Skarlatidis, A. Artikis, and G. Paliouras, “Probabilistic complex event recognition: A survey,” arXiv preprint arXiv:1702.06379, 2017.

[33] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood, “Sensecam: A retrospective memory aid,” in International conference on ubiquitous computing, pp. 177–193, Springer, 2006.

[34] V. Kalnikaite, A. Sellen, S. Whittaker, and D. Kirk, “Now let me see where i was: understanding how lifelogs mediate memory,” in Human Factors in Computing, pp. 2045–2054, 2010.

[35] M. Beigl, “Memoclip: A location-based remembrance appliance,” Personal Technologies, vol. 4, no. 4, pp. 230–233, 2000.

[36] R. W. DeVaul, A. Pentland, and V. R. Corey, “The memory glasses: subliminal vs. overt memory support with imperfect information,” in Symp. Wearable Computers, 2003, 2003.

[37] A. K. Dey and G. D. Abowd, “Cybreminder: A context-aware system for supporting reminders,” in Int. Symposium on Handheld and Ubiquitous Computing, pp. 172–186, 2000.

[38] C. Gurrin, A. F. Smeaton, A. R. Doherty, et al., “Lifelogging: Personal big data,” Foundations and Trends® in information retrieval, vol. 8, no. 1, pp. 1–125, 2014.

[39] E. Van Den Hoven, C. Sas, and S. Whittaker, “Introduction to this special issue on designing for personal memories: past, present, and future,” HCI, vol. 27, no. 1-2, pp. 1–12, 2012.

[40] J. Gemmell, G. Bell, and R. Lueder, “Mylifebits: a personal database for everything,” ACM, vol. 49, no. 1, pp. 88–95, 2006.

[41] W. Tan, J. Dwivedi-Yu, Y. Li, L. Mathias, M. Saeidi, J. N. Yan, and A. Y. Halevy, “Timelineqa: A benchmark for question answering over timelines,” in Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023 (A. Rogers, J. L. Boyd-Graber, and N. Okazaki, eds.), pp. 77–91, Association for Computational Linguistics, 2023.

[42] G. Meditskos, P.-M. Plans, T. G. Stavropoulos, J. Benois-Pineau, V. Buso, and I. Kompatsiaris, “Multi-modal activity recognition from egocentric vision, semantic enrichment and lifelogging applications for the care of dementia,” Visual Commun. and Image Representation, vol. 51, pp. 169–190, 2018.

[43] R. P. Goldman, C. W. Geib, H. A. Kautz, and T. Asfour, “Plan recognition (dagstuhl seminar 11141).,” Dagstuhl Reports, vol. 1, no. 4, pp. 1–22, 2011.

[44] G. Sukthankar, C. Geib, H. H. Bui, D. Pynadath, and R. P. Goldman, Plan, Activity, and Intent Recognition: Theory and Practice. Morgan Kaufmann, 2014.

[45] D. L. Schacter, The seven sins of memory: How the mind forgets and remembers. HMH, 2002.

Bringing Order to Chaos: Conceptualizing a Personal Research Knowledge Graph for Scientists

Prantika Chakraborty[†], Sudakshina Dutta[‡], Debarshi Kumar Sanyal[†],
Srijoni Majumdar[*], Partha Pratim Das[**]

[†]Indian Association for the Cultivation of Science, Kolkata-700032, India
[‡]Indian Institute of Technology Goa, Ponda-403401, India
[*]University of Leeds, Leeds LS2 9JT, UK
[**]Indian Institute of Technology Kharagpur, Kharagpur-721302, India,
[**]Ashoka University, Haryana-131029, India
intpc@iacs.res.in, sudakshina@iitgoa.ac.in,
debarshi.sanyal@iacs.res.in, s.majumdar@leeds.ac.uk,
ppd@cse.iitkgp.ac.in

Abstract

Research and work-related information is often manifold for a scientist, and the absence of an organizing system may impede their research. Information that holds immense personal interest and importance to a scientist may not be relevant to other users, yet it must be easily accessible to the scientist to enhance their productivity. We aim to address the need for such a system with our proposal of a knowledge graph for scientists, termed the *Personal Research Knowledge Graph* (PRKG). To identify the components of a PRKG, we interview scientists at an academic institution for higher education and research regarding the issues that the presence of a PRKG could solve. We translate the scientists' requirements into separate KGs, collectively known as the 'Vitamins of PRKG', and discuss methods for data acquisition to construct and maintain the PRKG. A smart virtual agent is proposed as a medium of interaction between the user and the PRKG. We also outline future research tracks, including those focused on maintaining the PRKG and protecting personal data and privacy.

1 Introduction

In an ever-expanding digital universe, scientists often find themselves entangled in the deluge of data necessary for their day-to-day professional activities. Instead of aiding in research work, the abundance of such data may, counter-intuitively, hamper it. Manually organizing all this information is a time-consuming activity that one might not always be eager to undertake. This brings in the requirement of a Personal Information Management (PIM) system that will aid in the collection of data, processing of the data into relevant information, and storage and eventual use of the information, with a strong emphasis on security and privacy [17].

In the world of PIM, researchers use abstractions such as *information type*, *information item*, *personal space of information* and *personal information collection* [16, 17]. An *information type* or *information form* denotes the mode (including the supporting applications) in which information is available and exchanged, such as paper-based letters, e-mails, and web pages, whereas an *information item* denotes the encapsulation of information in a persistent form that can be managed (e.g., stored and transmitted), such as an individual e-mail message. In the context of object-oriented programming, an information item corresponds to an object and the information type of the item maps to its class. Information can be *personal* to a user in several ways, which may be overlapping. In particular, a user's personal information includes information controlled or owned by the user, about the user, directed to the user, sent/posted/shared by the user, experienced by the user, or potentially relevant or useful to the user [16]. An individual has a single *personal space of information (PSI)*, which includes all information items that they consider personal. *Personal information collections (PICs)* are personally managed subsets of a PSI; an example is a folder containing downloaded research papers. Unlike an individual's PSI, which can be enormous in size, a PIC is smaller and, therefore, can be effectively managed. PIM involves activities to manage the PICs of a user with an aim to create, use and maintain a mapping between the needs of the user and their personal information.

Over the years, several PIM tools have been developed to assist users in organizing, maintaining, and utilizing their information items for various purposes. Concepts such as the Semantic Desktop [5, 23], which suggested the use of Semantic Web technologies [3] to organize the information items on a user's desktop and integrate them with resources on the Web for an enhanced user experience in accessing and sharing personal information, have had a significant impact. The Semantic Desktop concept has evolved into the Social Semantic Desktop, as seen in frameworks like NEPOMUK [12], SemanticLIFE [1], and SocialLIFE [29]. Recent efforts for personal information management have been in the form of mind-mapping software like TheBrain [1] where one can store a network of interconnected thoughts and ideas, capture their evolution, visualise them, and search over them. The Solid Pod project [2], led by Sir Tim Berners-Lee, offers users the opportunity to securely store their personal data in shareable data stores known as "pods" and to have complete control over their data and access to it.

1.1 Information Management for a Scientist

A scientist needs to handle a large volume of information, which can be segregated into three main categories:

- *Research-related*: This information is related to the research activities of scientists and includes scholarly literature they reference for their work, the various tools and methods they employ, as well as upcoming events such as conferences and workshops in their respective fields. Scientists may also wish to capture summaries of meetings with collaborators about their work and extract useful information from these interactions. Moreover, scientists working in academic institutions, such as universities, typically teach courses in addition to handle their research responsibilities; therefore, they must also manage teaching-related information.
- *Administrative*: A scientist or faculty member frequently participates in various administrative councils and needs to keep track of the developments of such councils and their meetings.
- *Laboratory-related*: Managing a laboratory is a common aspect of a scientist's responsibilities. Storing salient information about laboratory equipment and resources may simplify access and auditing processes.

Maintaining the aforementioned types of information in a structured and organized manner is often a challenging task for a scientist and can hinder their research work substantially. Having a PIM system makes storing and accessing such information less time-consuming, thereby enhancing the productivity of scientists.

1.2 Personal Knowledge Graph

Knowledge graphs (KGs) have been enjoying much popularity in recent times as a means for storing information in an organized manner [9, 14]. A KG can be formally defined [11] as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{F}\}$. $\mathcal{E}$ denotes the set of entities, $\mathcal{R}$ denotes the set of relations, and $\mathcal{F}$ denote the set of facts, where each fact is a triple (h, r, t) such that $h \in \mathcal{E}$, $t \in \mathcal{E}$ and $r \in \mathcal{R}$. A personal knowledge graph (PKG), proposed by Balog and Kenter [2], is a KG that is

[1] https://thebrain.com/

[2] https://solidproject.org/

relevant to a particular user but may not be useful to others. The primary purpose of the PKG is to support the delivery of services that are customized particularly to its owner. Balog and Kenter [2] initially introduced the concept of a PKG with the following definition:

> **Definition 1:** A PKG is 'a resource of structured information about entities *personally related to its user*, their attributes and the relations between them'.

They model a PKG as a "spider-web" like structure with a central node representing the user and all the other nodes connected to the central node, as shown in figure 16. In a more recent work co-authored by Balog [26], a PKG is redefined as follows:

> **Definition 2:** A PKG is a 'a knowledge graph (KG) where a single individual, called the owner of the PKG, has (1) full read and write access to the KG, and (2) the exclusive right to grant others read and write access to any specified part of the KG.'

The second definition differs from the first one in that the entities of the PKG *need not* be directly connected to the owner. Lately, several PKGs have been proposed and developed to assist users in the field of health, finance, education and research [8]. For example, a PKG containing an individual's health information may be useful for their dietary planning and medical checkups [24]. A PKG is created and owned by the user whose personal information is held in it; this ensures the privacy of the data.

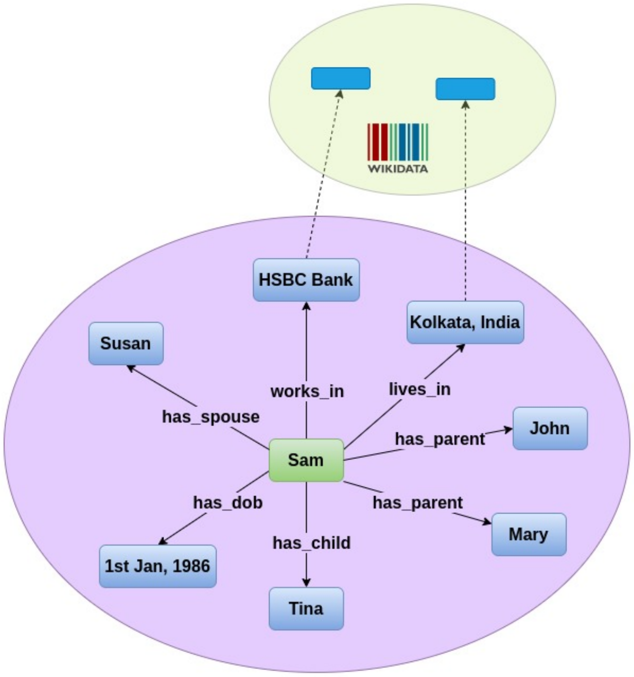


Figure 6: An example of PKG containing Sam's personal data [8].

PIM focuses on *activities* related to managing the personal information of an individual, but the emphasis of this paper is on PKG which is a *data structure* to store the individual's personal information. While PIM deals with information items, PKG handles more granular units which are facts or triples extracted from information items. Therefore, tools for creation and maintenance of traditional PIM systems are not directly applicable to manage PKGs.

1.3 Our Proposal

In this position paper, we focus on PKGs for scientists; we call it PRKG – a shorthand for *P*ersonal *R*esearch *K*nowledge *G*raph. PRKG for an individual researcher will be owned and maintained by the researcher. Formally, a PRKG is defined as $\mathcal{P} = \{u, \mathcal{G}\}$ where u denotes the owner of the PRKG and $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{F}\}$ denotes the KG containing the 'facts' in the PRKG. To represent the owner as a node in the PRKG, a fact in which the owner is one of the entities is created. Such a fact could be the following $(u, owns, p)$ where $u, p \in \mathcal{E}$, $owns \in \mathcal{R}$, and p represents the current PRKG. Following **Definition 2**, we do not require all facts or triples to be connected to the owner node u.

In order to leverage the knowledge reserved in the PRKG, an intelligent virtual agent may be designed to access the PRKG, serving as a conduit for interaction between the scientist and their PRKG. The agent will be capable of deducing essential information based on the scientist's needs or even act proactively, and can also modify the PKG by incorporating new information or updating existing data. Note that the PRKG mitigates the cold start problem that services like personalized recommendation systems encounter when bootstrapping for a new owner.

Fundamental questions on the design and use of a PRKG pertain to the identification of the information that should be included in it, the representation of structured information within a PRKG, methods to populate and maintain the graph, and the proper utilization of the PRKG to assist the owner. The remainder of the paper is organized as follows. Section 2 summarises the various requirements highlighted by researchers for their PKGs. Section 3 discusses our proposal for the Research PKG in detail whereas Section 4 discusses methods for data acquisition for the creation of PRKGs and issues related to PRKG maintenance. Section 5 suggests methods in which a scientist can interact with the PRKG. Section 6 discusses future research directions for the field, and Section 7 concludes the paper. For the rest of the paper, the terms "owner" and "user" are used synonymously and assumed to refer to a "scientist" who owns (and therefore, creates and manages) the PKG we will discuss.

2 Content in a PRKG: What do researchers say?

We interviewed six researchers specializing in chemistry, physics, and biology at the premier research institution where the first author works. We inquired about the information they would prefer to store in a PRKG, how they plan to use that information, and, if presented with an AI-bot powered by this database, what expectations they would have from it. One observation that we concluded was that researchers mostly looked at the PRKG as a database of research-related information.

Research paper metadata: The information that all researchers wanted to store in the PRKG includes metadata such as the title, author names, year and venue of publication, inferred keywords, and custom user-guided tags for the papers they read. Scientists commonly download papers and save them on their desktops without systemically categorizing them, and later encounter difficulty in retrieving specific articles. The researchers would like to have these information available in the PRKG so that a desktop search engine or a chatbot could answer related queries. Some researchers wanted to store annotations they add to a paper or the paper's main findings extracted by a smart AI-based application.

Email metadata and summary: Some of the researchers have told us that they receive numerous emails every day in their official email id, and more than half of them are unimportant or close to spam. Consequently, there is a risk that critical emails, such as those seeking project positions, may get buried and go unattended. They wished to have the important emails automatically discerned and indexed in the PRKG so that applications can generate alerts for them. These emails could range from upcoming meetings or manuscript submission deadlines to those that require long-term and situation-dependent reminders. For instance, when a researcher advertises a project position, they may want to be reminded of a query email received for a project position long ago.

Event and schedule information: In order to generate regular reminders for various tasks like manuscript submission, paper review, etc. researchers wanted to store the deadlines for such tasks in the PRKG. Information about meeting dates and times, exam schedules, project timeline could also be stored in the PRKG as suggested by the researchers.

Research meeting/discussion summary: Another category of knowledge that all the researchers wanted to curate in the PRKG is that gleaned from research meetings, potentially in the form of concise summaries. All the researchers we interviewed supervise Master's and PhD students, and have frequent academic discussions with them. Many new ideas and questions come up during these internal meetings and most escape documentation. The researchers wanted the PRKG to remember them so they could be revisited later. For instance, unanswered questions or negative results discussed in a meeting should be documented as they might motivate novel investigations in the future. Some of the researchers mentioned that the meeting attendees often use a mix of English and a local language like Bengali, and it would be helpful if the necessary information could still be extracted and stored in the PRKG.

Financial information (Project grants and lab equipment): Researchers who make extensive use of lab equipment also wanted to store their equipment-related information, for example, of servers and other high-end devices, in the PRKG. These information could be extracted from purchase documents. An AI-bot might use this information to perform periodic audits of the devices and advise maintenance activities. One researcher who aims to build an advanced lab desired to store information from quotations for various lab equipment so that applications could easily compare them and recommend the best equipment to purchase. Another suggestion was to store information related to project grants and fund utilization since they are often difficult to remember and track, yet need to be frequently accessed.

Journal/Conference information: One researcher mentioned having personal preferences regarding journals and conferences, desiring search results (from the web or desktop) to be ordered based on these preferences. These preference or trust lists could be stored in the PRKG. Some researchers desired the PRKG to be utilized for generating customized recommendations for articles and publication venues that match their interests.

Information on collaborations: Another researcher highlighted having multiple collaborations, and for information about collaborators (automatically extracted from emails) to be stored in the PRKG. This researcher even discussed the possibility of assigning priorities to these collaborations, establishing a partial order that could be used to customize email alerts and schedule meetings. The PRKG should store this personal priority list for the researcher. The researcher mentioned that they might even define a priority over the collaborations; this partial order might be harnessed to customize email alerts and schedule meetings. The PRKG should store this priority list, which is completely personal to the researcher.

Experiment-related information: Another researcher wanted more fine-grained information from papers. As a microbiologist, he wanted details about chemicals and apparatus from papers to be extracted and stored in the knowledge graph. This information could be valuable for learning about and purchasing improved versions of the materials used in experiments. This researcher also wanted the PRKG to have access to lab notebooks to capture fine-grained information about experiments they conduct. This information could be valuable in tracing any discrepancies in experimental results arising from frequently overlooked differences in experimental settings. They believe that this approach might contribute to addressing the replication crisis in biology to some extent.

Course-related information: Researchers affiliated to educational institutions have to undertake the responsibility of teaching a designated number of courses. One of the researchers we interviewed emphasized the potential of storing curated information from a course's schedule, outline, and materials (textbooks, articles, papers, presentations, etc.) in the PRKG. The researcher also noted that they acquire information about courses of interest taught at leading universities by browsing the Internet, and this information may be stored in the PRKG. The AI-bot may use this information to help the researcher design their own courses.

Role of AI-bot: In response to our question on the expectations from an AI-bot capable of engaging in conversational dialogues with scientists, one researcher said they wanted the agent to assist new research scholars in generating the first iteration of the literature review. Another researcher wondered if the agent could help them formulate new research questions given the experimental data and sample questions already formulated. Interestingly, the latter requirement where an AI system helps to *do* science is still a work-in-progress for the AI community. A researcher wanted the bot to be integrated with simple science tools, for example, to convert a table of values from one unit to another, when instructed to do so in natural language; such a tool is useful in meetings with funding agencies and collaborators where quick answers are expected. Another researcher wished the agent to automate the submission of travel and dearness allowance-related documents, given a short prompt from the researcher and access to their or relevant documents. Another capability that some of the scientists mentioned is to automatically generate a TODO list, which includes alerts for meetings but potentially encompasses various other activities.

3 Building Blocks of a PRKG

We now outline a blueprint for a Personal Research Knowledge Graph (PRKG) owned and managed by a science researcher. A similar proposal was made regarding a PKG for researchers in [7] where the central node represents the owner, and various entities related to the owner's research aspects are connected to it. However, the current proposal does not require the owner to be linked to all the facts. Further, we organize the PRKG as a collection of several knowledge graphs (KGs), each originating from a specific kind of knowledge source, and together referred to as the *vitamins of PRKG* for their first letters are reminiscent of vitamins, as shown in Figure 7. Table 4 shows how particular information that the researchers want to store in their PRKG, as discussed in Section 2, can be mapped onto the different individual KGs that are described as follows:

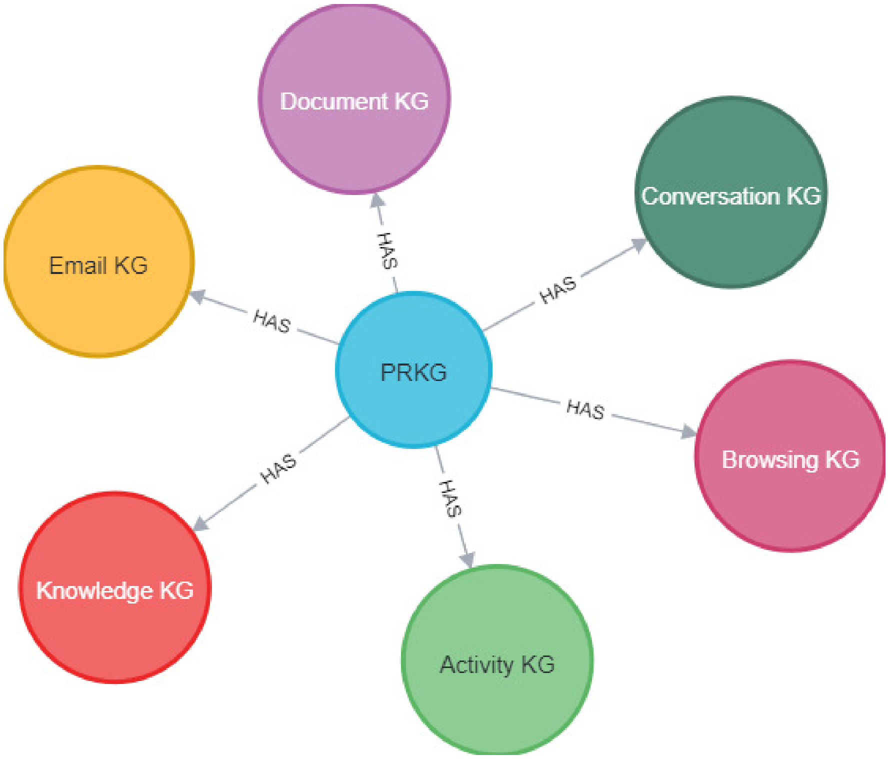


Figure 7: Vitamins of PRKG

- *Activity KG*: A KG that will contain information regarding the owner's activities, such as scheduled meetings and events, will be called the *Activity KG*. This KG will help the user in setting reminders for upcoming events like meetings, talks, conferences, journal submissions, lectures to be taken, etc. The KG will have access to the owner's system logs, calendars, and digital planners for regular updates.

Table 4: Mapping researchers' requirements, as discussed in Section 2, to the various building blocks of a PRKG, as discussed in Section 3

Researchers' requirements	Requirement frequency by user	Relevant PRKG sub-graph
Research paper metadata	6	Document KG
Email metadata and summary	6	Email KG
Event and schedule information	5	Activity KG, Email KG
Research meeting/discussion summary	4	Email KG, Conversation KG
Financial information including project grants and lab resources' quotations	3	Document KG
Journal/conference information	3	Knowledge KG, Browsing KG
Information on collaborations	2	Email KG, Conversation KG
Experiment-related information	1	Document KG
Course-related information	1	Document KG, Browsing KG

- ***B**rowsing KG*: A KG that contains the owner's browsing history-related data will be referred to as the *Browsing KG*. This may include web search logs, encompassing scholarly search engines like Google Scholar and academic social networks like ResearchGate, from which the owner's current research interests can be inferred. Web logs can also help inform recommendations, for example, conference pages visited may provide insights regarding conferences that the owner might like to attend in the future.

- ***C**onversation KG*: A KG that contains knowledge inferred from conversations over applications like WhatsApp and Skype will be called the *Conversation KG*. Knowledge drawn from transcripts of meetings held online over video-telephonic applications like Zoom may also be incorporated in this KG. The goal is to store summarised knowledge from these conversations and meetings, that may range from updates on current works to ideas for new projects.

- ***D**ocument KG*: Documents related to the owner's research pursuits, teaching activities, lab equipment, and administrative affairs are stored either locally on the user's machine or in remote storage owned by the user. A *Document KG* will capture knowledge extracted from the above sources. For example, it may contain the metadata from papers being read by the user, the timetable of the course offered by the user in the current semester, essential information extracted from recent office memorandums, and details of recent purchases. An expanded Document KG could also include multimedia files like images, audio and videos.

- ***E**mail KG*: A KG that will hold information gleaned from a user's work-related emails will be referred to as the *Email KG*. This KG will extract relevant entities from email chains that a user holds with his collaborators. This includes basic information regarding the collaborators, like their affiliations and contact details. Identifying potential collaborators and the main topics of discussion with them can also enhance the Email KG.

- ***K**nowledge KG*: The information about the user that they themselves provide will be stored in *Knowledge KG*. It potentially includes the user's current affiliations, research interests, and personal beliefs and preferences (say, about fellow researchers in their community or books on a particular topic).

It is worth noting that these integrant KGs are not disjoint KGs. For instance, a person who undertakes frequent collaborations with the user will be present in both the Document KG as a co-author and the Email KG as a collaborator, and will be modeled by the same node. This can be achieved by entity disambiguation and linking.

An important question is: at what granularity should information be stored in the PRKG? For an Activity KG, one can extract precise details of a scheduled meeting, like the topic, time, venue, and invitee list from the user's calendar. But in the case of a Conversation KG or an Email KG, storing a free-form textual summary of a conversation chain as a text blob may be more practical because it is hard to define *a priori* the granular entities to extract from messages and even harder to train an entity extractor for the task while state-of-the-art tools for text summarization and question-answering from unstructured text display superb performance, thanks to large language models. Meeting summaries may be utilized by external applications, such as neural question-answering systems, to answer more specific user queries. Nevertheless, if it is possible to define fine-grained entities like collaborator names, conferences (visited by the user), and journals (to which the user submits papers), they should be identified from the information item and incorporated into the relevant sub-KG.

Entities and relations may be stored as Resource Description Framework[3] [19] (RDF) triples of the type $\langle subject - predicate - object \rangle$. A desirable aspect of this information storage is the inclusion of *provenance* of the RDF triples, which helps identify the source of the information and the process by which it was extracted. This is essential for determining the quality of information extracted and the correctness and trustworthiness of the process used to extract this information [25].

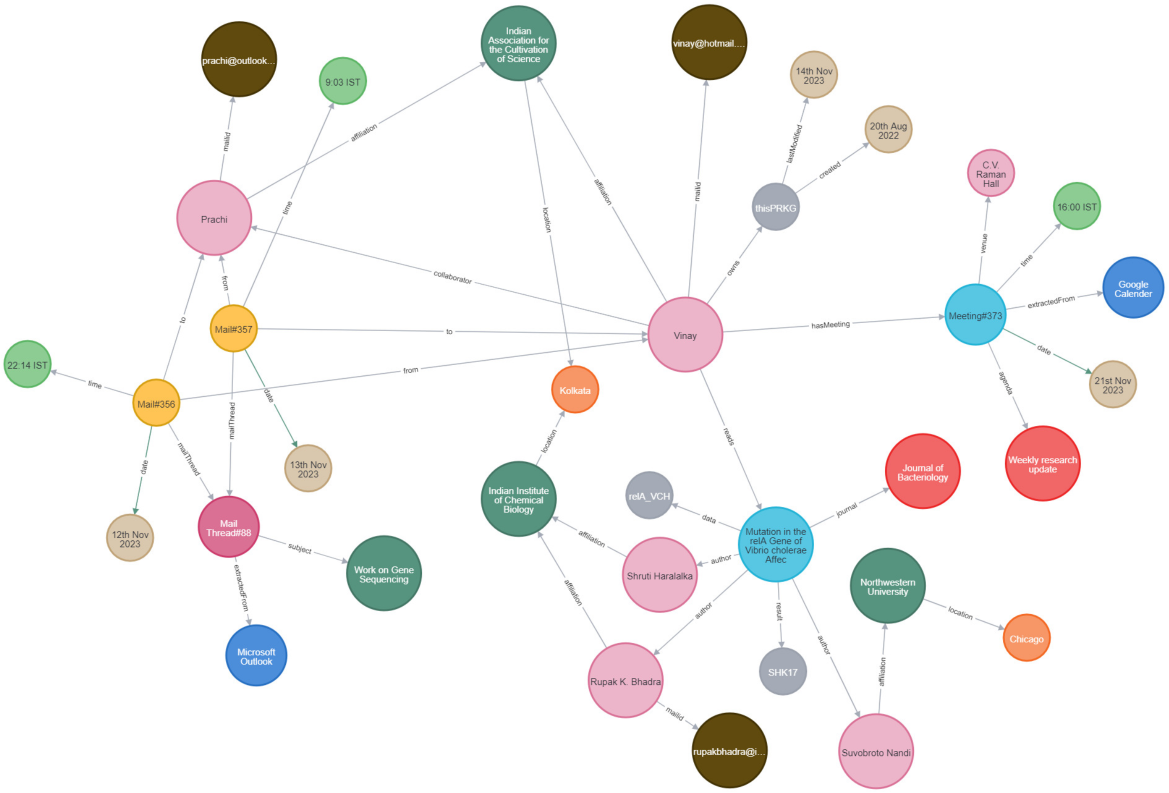


Figure 8: Snippet of a PRKG

Illustration of a PRKG: Figure 8 shows a snippet of a PRKG whose owner is Vinay, a researcher. This PRKG is a sample comprising triples of the Activity KG, the Document KG and the Email KG. Although conceptually we have segregated the PRKG into smaller divisions based on the source of the knowledge, the actual PRKG we have created is a heterogeneous mixture of triples from all the smaller KGs. The graph reveals that Vinay is scheduled to have a meeting (specifically, *meeting 373*) on *21st November 2023* at *4:00 pm IST* at the *C.V. Raman*

[3] `https://www.w3.org/RDF/`

Hall. The meeting's agenda is his *Weekly research update*. This is an excerpt of the Activity KG of his PRKG. We also see how Vinay's communication with *Prachi*, who is a possible collaborator, via mail has been captured by the Email KG. Prachi's email ID and affiliation have been stored as information from the emails exchanged between Vinay and Prachi. The summary of their email thread will also be stored in the Email KG but has been excluded from this figure for brevity. This PRKG also stores metadata from papers that Vinay downloads for reading. One such paper is titled "Mutation in the *relA* Gene of *Vibrio cholerae* Affects In Vitro and In Vivo Expression of Virulence Factors" [13], written by *Shruti Haralalka, Suvobroto Nandi,* and *Rupak K. Bhadra*, and published in *Journal of Bacteriology*. Additional information about the authors, such as their current affiliation and contact details, is extracted from the paper. This PRKG was created using the Enterprise Edition of Neo4j 5.2.0, which is provided with Neo4j Desktop version 1.5.9. The graph is best viewed when zoomed in digitally.

4 Data Acquisition and PRKG Maintenance

We introduced six constituent KGs in the previous section. The knowledge captured in these KGs come from diverse sources spanning multiple devices. We envisage that custom tools will be written to extract entities and relations of interest from these sources to populate the corresponding KGs. LLMs may also be prompted to extract this information from various sources.

As an example, consider the Activity KG. Calendar services like Google Calendar and Microsoft Outlook Calendar have become irreplaceable when it comes to keeping track of a user's scheduled events like meetings and lectures. Currently, there are a number of APIs, like *Make*[4], *Notion*[5], and *REST*[6], that can sync information from such calendars and store them in a database that will be eventually used to create the Activity KG. Note that most schedules and deadlines are first programmatically extracted from emails and then pushed to the user's calendar. Similarly, a Browsing KG can simply track the user's browsing activities, remember visited websites, and incorporate derived knowledge like research interests, frequently visited university websites, and commonly used online tools.

For the other KGs, namely Conversation KG, Document KG, and Email KG, we envisage an application that first identifies the precise resources to be harvested for inclusion in the PRKG. Once that information is available, the application analyzes the resources to extract relevant knowledge. As a concrete example, consider an application that allows the PRKG owner to indicate which emails should be parsed for inclusion into the Email KG. This application may *learn* to suggest emails that should be included, and once the user agrees, it can proceed to process them. Similarly, the user may explicitly indicate folders or files should come under the scope of the Document KG. As regards the structured information to be extracted from these resources, they are of two kinds: (1) *metadata*: consider these examples: (a) for emails in Email KG, the sender, recipients, subject and date may be extracted; (b) for research papers in Document KG, the title, author names, keywords, publication venue, and publication date are salient fields; (2) *deep data*: this includes an analysis of the content of the resource; examples of such information are (a) for emails with collaborators, the collaborators' names and affiliations, and keywords that capture the collaboration area; (b) for downloaded research papers, the problem, methods used, and essential findings. While the metadata fields may be easy to identify for a specific information type, defining the deep data fields is challenging and may potentially depend on the user's interest. Machine learning models that can *learn* new entity types from a few user-provided examples may be useful here.

PKGs differ from general-purpose KGs in that they capture entities and relations that are personal to the user. A user's personal preferences and activities evolve over time, making old information in the PKG useless for most applications. This volatility of information and the consequent requirement for PKG maintenance are important characteristics of a PKG. In case of Activity KG and Browsing KG, these aspects are highly visible;

[4] `https://www.make.com/en/api-documentation`

[5] `https://developers.notion.com/`

[6] `https://tinyurl.com/ibmRestApi`

stale information should be flushed out to avoid unnecessary memory consumption. However, for conversations, emails, and arguably documents, it might not be prudent to delete old facts as the user might like to revisit them in future. To capture the freshness of a triple, the PRKG might associate temporal information like the creation date and the last access date of the triple. This temporal information might be useful for applications that use the PRKG to deduce the owner's current behavioural characteristics.

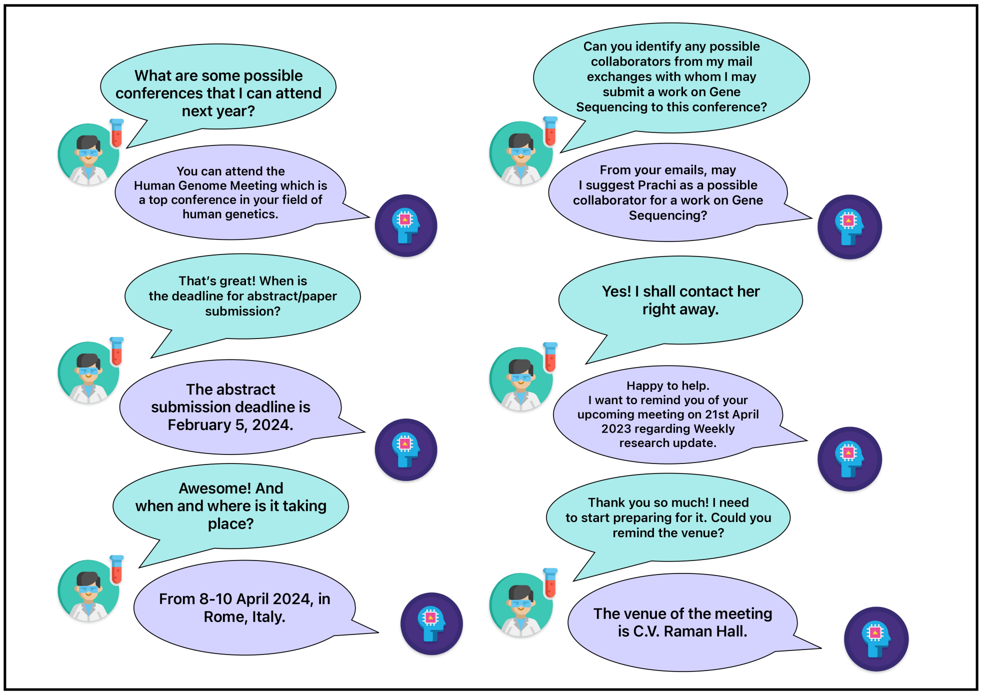

What are some possible conferences that I can attend next year?
You can attend the Human Genome Meeting which is a top conference in your field of human genetics.
That's great! When is the deadline for abstract/paper submission?
The abstract submission deadline is February 5, 2024.
Awesome! And when and where is it taking place?
From 8-10 April 2024, in Rome, Italy.
Can you identify any possible collaborators from my mail exchanges with whom I may submit a work on Gene Sequencing to this conference?
From your emails, may I suggest Prachi as a possible collaborator for a work on Gene Sequencing?
Yes! I shall contact her right away.
Happy to help. I want to remind you of your upcoming meeting on 21st April 2023 regarding Weekly research update.
Thank you so much! I need to start preparing for it. Could you remind the venue?
The venue of the meeting is C.V. Raman Hall.

Figure 9: An interaction between Vinay and the smart virtual agent who has access to Vinay's PRKG.

5 Interacting with the PRKG

The potential of a PRKG can be fully realized only when a user can interact with it in order to make their daily research life more manageable. As observed in Section 2, the owner of a PRKG will have frequent questions on topics like details of research papers read by them, upcoming meetings and summaries of past ones attended and research grants. In order to interact with the PRKG to access its stored knowledge, the owner will approach the smart virtual agent by asking the required questions. Questions to the agent may be direct, whose response can be found in the KG with a simple database search. For example, Vinay, the owner of the PRKG in figure 8, can ask questions regarding his upcoming engagements like, "*Where will the meeting on my weekly research update be held on 21st this month? Also, could you confirm the time of the meeting?*" This is a fairly straightforward question that the agent will be able to respond to without any added inference. Some questions, however, may not be so direct for the agent. Questions like "*Can you identify a possible collaborator from the people I have been exchanging emails with regarding gene-sequencing?*". The agent will have to go through all possible mail threads with the subject or summary with *gene-sequencing* in it and identify and rank possible collaborators for Vinay. Once deduced, the agent may add this information to the PRKG. The agent may be directed to add, delete or modify inferred entities and relations on a routine basis, even before being prompted to do so, such that the PRKG remains updated. The agent may also remind the owner about upcoming scheduled events, submission deadlines, and to reply to a potentially important email as identified by the agent. A sample conversation between Vinay and the agent is shown in Figure 9.

A PRKG holds significant relevance even when Large Language Models (LLMs), such as ChatGPT developed by OpenAI[7] and LLaMa by Meta[28], have become widely popular for their ability to accurately follow user instructions and perform tasks such as question-answering, summarization, translation, and recommendation. The following issues with LLMs make their use particularly challenging:

- *Hallucinations:* LLMs have gained a reputation for generating false and unreliable responses, commonly known as hallucinations when faced with questions or prompts regarding unfamiliar knowledge not encountered during training [15]. This drawback can pose significant risks in domains that demand precise information, like medicine [4] and law [27]. In contrast, question-answering based on PRKGs can provide reliable information due to the specificity and ability to update the stored information, unlike LLMs trained on fixed data. Leveraging external knowledge about a user in the form of a PRKG can significantly enhance an LLM's capability to provide accurate and personalized responses, eliminating the need to predict or generate unreliable answers [21].

- *Privacy concerns:* If a user provides personal information in the prompts with which they interact with the publicly available LLMs, it might end up into the training dataset of the LLM. Researchers have observed that it is possible to extract the training data by manipulating the LLM into sharing it, and this can lead to privacy breach attacks [6, 20]. In contrast, a PRKG offers transparency by granting users complete access and control over their stored data, empowering users with more agency over their personal information. One can also build LLM-based dialog systems that access the PRKG at response generation time, thereby harnessing the strengths of both LLMs and knowledge graphs.

6 Future Research Directions

The *personal* aspect of PRKGs introduces many challenges and opportunities in their design, implementation, and evaluation. The first challenge relates to the composition of a PRKG. Consider any constituent KG like a Document KG or Email KG: even if the user identifies the documents or emails to be considered for the PRKG, it is unclear which entities and relations should be extracted for curation. The PRKG designer may specify a small default set of entities and relations (or ontology) that a researcher might be interested in (as we have done above), and allow the PRKG owner to extend it. Additionally, a machine learning algorithm in the PRKG editor might learn to suggest potential entities and relations to be extracted. Entity disambiguation is needed to link multiple mentions of the same entity like a collaborator or a journal. Entity disambiguation would be challenging for named entities that occur rarely in the user's data and are also uncommon in the external world. The lack of open datasets makes research in this area very difficult. Future work should organize community efforts to build datasets to catalyze the design and implementation of PRKGs. We also believe that more KGs can be conceived as components of the PRKG. For example, knowledge extracted from photographs captured by the scientist at conferences and meetings could form the basis for a new personal KG that helps to preserve and retrieve connected memories [18].

Evaluation of a PRKG is another challenging area that is hardly explored. Since the personal data and expectations on what should go into a PRKG vary widely across researchers (based on their domain, seniority, etc.), evaluation strategies should be carefully designed keeping the PRKG owner in the loop. A closely related problem is the evaluation of a PRKG-based smart virtual agent: Can it understand and satisfy the user's information needs? Does it provide pro-active suggestions to the user? It is also desirable that the agent's responses are accompanied with explanations.

[7] https://openai.com/chatgpt

A PRKG contains personal and sensitive information of its owner. So it should be securely stored whether on the user's local device or in the cloud. A PRKG contains information related to the research activities of its owner. Therefore, the PRKG owner might like to share parts of the PRKG with their collaborators including students. Similarly, the PRKG owner could personalize various services like recommendation systems and search engines with the PRKG. This raises the question about how the PRKG owner can grant access to other users and applications without privacy leaks and without allowing unintended modification of information. Investigation of these aspects may be informed by the existing research on privacy-preserving KGs [10, 22].

7 Conclusion

We have discussed the notion of personal research knowledge graphs, which are employed to store the information that is personally relevant to a researcher. Through interviews with multiple scientists, we identified their requirements for personal data management and put forth a design for PRKGs to capture the necessary knowledge. We acknowledged several challenges in the design, implementation and evaluation of PRKGs. Nevertheless, if done right, they can be immensely valuable in building applications that bring order to the information chaos and alleviate the overload that the researcher experiences. This, in turn, can translate to better productivity, hightened job satisfaction and improved work-life balance.

References

[1] Mansoor Ahmed, Hanh Huu Hoang, Muhammad Shuaib Karim, Shah Khusro, Monika Lanzenberger, Khalid Latif, Elke Michlmayr, Khabib Mustofa, H. T. Nguyen, Andreas Rauber, Alexander Schatten, Tho Manh Nguyen, and A Min Tjoa. 'SemanticLIFE' - A framework for managing information of A human lifetime. In iiWAS'2004 - The sixth International Conference on Information Integrationand Web-based Applications Services, 27-29 September 2004, Jakarta, Indonesia, volume 183. Austrian Computer Society, 2004.

[2] Krisztian Balog and Tom Kenter. Personal knowledge graphs: A research agenda. In Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, pages 217–220, 2019.

[3] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. Scientific American, 284(5):34–43, 2001.

[4] Gernot Beutel, Eline Geerits, and Jan T Kielstein. Artificial hallucination: GPT on LSD? Critical Care, 27(1):148, 2023.

[5] Karin Breitman, Marco Casanova, and Walter Truszkowski. Semantic desktop. In Semantic Web: Concepts, Technologies and Applications, pages 229–239. Springer London, London, 2007.

[6] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In Proceedings of the 30th USENIX Security Symposium, pages 2633–2650. USENIX Association, 2021.

[7] Prantika Chakraborty, Sudakshina Dutta, and Debarshi Kumar Sanyal. Personal research knowledge graphs. In Companion Proceedings of the Web Conference 2022, WWW '22, page 763–768, New York, NY, USA, 2022. Association for Computing Machinery.

[8] Prantika Chakraborty and Debarshi Kumar Sanyal. A comprehensive survey of personal knowledge graphs. WIREs Data Mining and Knowledge Discovery, 13(6):e1513, 2023.

[9] Vinay Chaudhri, Chaitanya Baru, Naren Chittar, Xin Dong, Michael Genesereth, James Hendler, Aditya Kalyanpur, Douglas Lenat, Juan Sequeda, Denny Vrandečić, et al. Knowledge graphs: introduction, history and, perspectives. AI Magazine, 43(1):17–29, 2022.

[10] Chaochao Chen, Jamie Cui, Guanfeng Liu, Jia Wu, and Li Wang. Survey and open problems in privacy preserving knowledge graph: Merging, query, representation, completion and applications. arXiv preprint arXiv:2011.10180, 2020.

[11] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. SEMANTiCS (Posters and Demos), 48(1-4):2, 2016.

[12] Tudor Groza, Siegfried Handschuh, and Knud Moeller. The NEPOMUK project – on the way to the social semantic desktop. In Proceedings of I-MEDIA 2007 and I-SEMANTICS 2007, page 201–210, 2007.

[13] Shruti Haralalka, Suvobroto Nandi, and Rupak K Bhadra. Mutation in the relA gene of vibrio cholerae affects in vitro and in vivo expression of virulence factors. Journal of Bacteriology, 185(16):4672–4682, 2003.

[14] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Transactions on Neural Networks and Learning Systems, 33(2):494–514, 2021.

[15] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), mar 2023.

[16] William Jones, Jesse David Dinneen, Robert Capra, Anne Diekema, and Manuel Pérez-Quiñones. Personal information management (PIM). Encyclopedia of Library and Information Science, pages 3584–605, 2017.

[17] William P Jones and Jaime Teevan. Personal information management. University of Washington Press, 2007.

[18] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. Navigating connected memories with a task-oriented dialog system. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2495–2507, 2022.

[19] Eric Miller. An introduction to the resource description framework. Bulletin of the American Society for Information Science and Technology, 25(1):15–19, 1998.

[20] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. arXiv preprint arXiv:2311.17035, 2023.

[21] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. arXiv preprint arXiv:2302.12813, 2023.

[22] Erasmo Purificato, Sabine Wehnert, and Ernesto William De Luca. Dynamic privacy-preserving recommendations on academic graph data. Computers, 10(9):107, 2021.

[23] Leo Sauermann, Ansgar Bernardi, and Andreas Dengel. Overview and outlook on the semantic desktop. In Proceedings of the 1st Workshop on The Semantic Desktop at the ISWC 2005 Conference, volume 175, pages 1–19. Citeseer, 2005.

[24] Oshani Seneviratne, Jonathan Harris, Ching-Hua Chen, and Deborah L McGuinness. Personal health knowledge graph for clinically relevant diet recommendations. arXiv preprint arXiv:2110.10131, 2021.

[25] Leslie F Sikos and Dean Philp. Provenance-aware knowledge representation: A survey of data models and contextualized knowledge graphs. Data Science and Engineering, 5:293–316, 2020.

[26] Martin G Skjæveland, Krisztian Balog, Nolwenn Bernard, Weronika Lajewska, and Trond Linjordet. An ecosystem for personal knowledge graphs: A survey and research roadmap. arXiv preprint arXiv:2304.09572, 2023.

[27] Zhongxiang Sun. A short survey of viewing large language models in legal aspect. arXiv preprint arXiv:2303.09136, 2023.

[28] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[29] Sao-Khue Vo, Amin Anjomshoaa, and A. Min Tjoa. Semantic-aware mashups for personal resources in semanticlife and sociallife. In Availability, Reliability, and Security in Information Systems, pages 138–154. Springer International Publishing, 2014.

User Modeling in the Era of Large Language Models: Current Research and Future Directions

Zhaoxuan Tan, Meng Jiang
Department of Computer Science and Engineering, University of Notre Dame
{ztan3, mjiang2}@nd.edu

Abstract

User modeling (UM) aims to discover patterns or learn representations from user data about the characteristics of a specific user, such as profile, preference, and personality. The user models enable personalization and suspiciousness detection in many online applications such as recommendation, education, and healthcare. Two common types of user data are text and graph, as the data usually contain a large amount of user-generated content (UGC) and online interactions. The research of text and graph mining is developing rapidly, contributing many notable solutions in the past two decades. Recently, large language models (LLMs) have shown superior performance on generating, understanding, and even reasoning over text data. The approaches of user modeling have been equipped with LLMs and soon become outstanding. This article summarizes existing research about how and why LLMs are great tools of modeling and understanding UGC. Then it reviews a few categories of large language models for user modeling (LLM-UM) approaches that integrate the LLMs with text and graph-based methods in different ways. Then it introduces specific LLM-UM techniques for a variety of UM applications. Finally, it presents remaining challenges and future directions in the LLM-UM research. We maintain the reading list at: https://github.com/TamSiuhin/LLM-UM-Reading.

1 Introduction

User Modeling (UM) aims to extract valuable insights and patterns from user behaviors, enabling customization and adaptation of systems to meet specific users' needs [124]. UM techniques have facilitated a better understanding of user behaviors, customized intelligent assistance, and greatly improved user experience. For example, when people are looking for dinner options and searching online, the UM techniques infer their characteristics based on interaction history, predict current food interests, and give personalized recommendations. UM has a substantial impact on user data analysis and many applications such as e-commerce [191, 194, 280], entertainment [11, 33, 155], and social networks [1, 2, 212]. UM is a highly active and influential research field.

User modeling is mainly about mining and learning user data, including user-generated content (UGC) and user's interactions with other users and items. User-generated content encompasses a wide range of text data, such as tweets, reviews, blogs, and academic papers. The rich texts can be analyzed by natural language processing (NLP) techniques. User interactions, on the other hand, involve various actions such as following, sharing, rating, comments, and retweets. These interactions may form a heterogeneous temporal text-attributed graph [199] for having temporal and textual information and having different types of nodes and relations. That can be analyzed using graph mining and learning techniques. As a result, user modeling has branched out into text-based and graph-based approaches, focusing on extracting insights from text and graph data, respectively.

How has the research of text-based UM developed? Researchers have used multiple types of representations of texts, such as words, topics, and embeddings. Bag-of-Words (BoW) models create distributional text representations with word frequencies using a discrete vocabulary [67]. To address the sparsity of BoW representations,

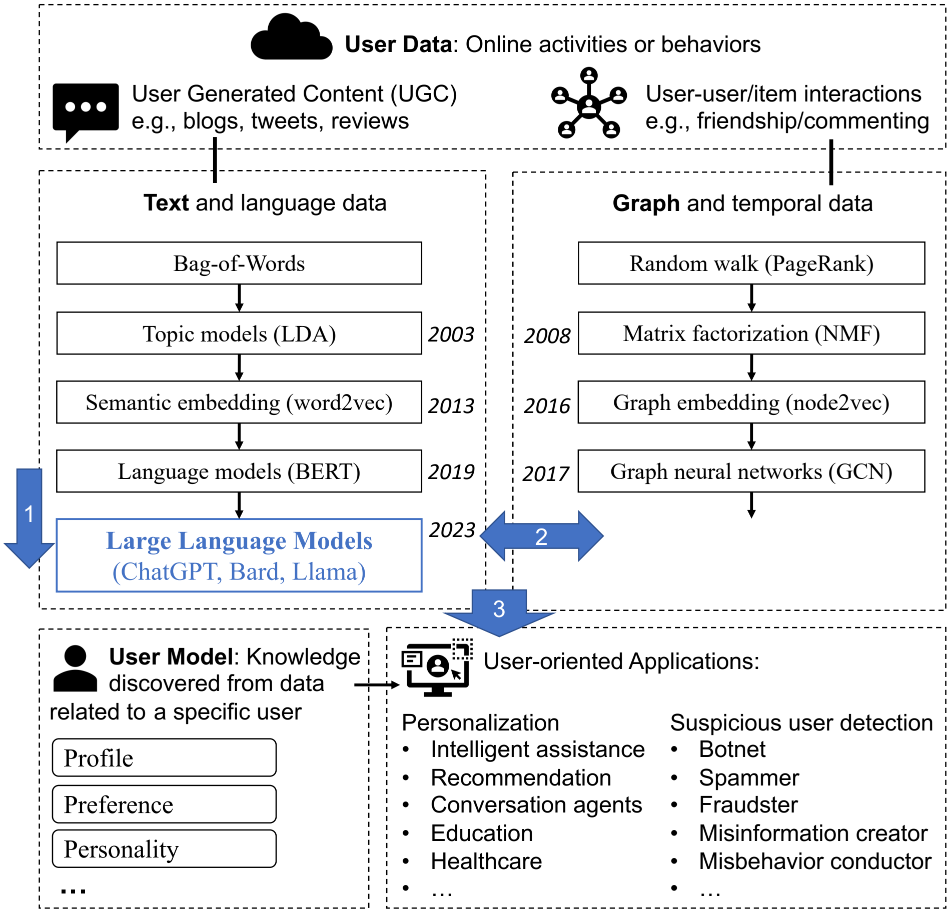

User Data: Online activities or behaviors
User Generated Content (UGC)
e.g., blogs, tweets, reviews
User-user/item interactions
e.g., friendship/commenting
Text and language data
Bag-of-Words
Topic models (LDA)
2003
Semantic embedding (word2vec)
2013
Language models (BERT)
2019
Large Language Models
(ChatGPT, Bard, Llama)
2023
1
2
3
Graph and temporal data
Random walk (PageRank)
2008
Matrix factorization (NMF)
2016
Graph embedding (node2vec)
2017
Graph neural networks (GCN)
User Model: Knowledge discovered from data related to a specific user
Profile
Preference
Personality
...
User-oriented Applications:
Personalization
Intelligent assistance
Recommendation
Conversation agents
Education
Healthcare
…
Suspicious user detection
Botnet
Spammer
Fraudster
Misinformation creator
Misbehavior conductor
…

Figure 10: User modeling aims to discover knowledge and patterns from user data to identify profile, preference, and personality. The three blue arrows in the figure correspond to our three major contributions: (1) summarize how and why LLMs are great tools for modeling and understanding UGC, (2) review approaches that integrate LLMs with text- and graph-based UM methods, and (3) introduce LLM-UM techniques for various applications.

topic modeling techniques statistically discover latent topics in a collection of documents, e.g., latent Dirichlet allocation (LDA) [15]. But they are not able to capture semantic meanings, i.e., word semantic similarity. Word2Vec employs nonlinear neural layers to develop Continuous Bag-of-Words (CBOW) and continuous skip-gram models [148]. It extracts semantic embeddings from many kinds of UGC text data, such as blogs, reviews, and tweets. However, the neural layers are too shallow to capture deep sequential patterns among numerous word tokens. With the breakthrough in Transformer architectures [219], pretrained language models (PLMs) significantly changed the landscape of UGC understanding with the pretrain-fine-tune paradigm. The new paradigm trains the models on a large unlabeled corpus using self-supervised learning and uses hundreds or thousands of examples to fine-tune the models for downstream task adaptation [102]. Recently, large language models (LLMs) have revolutionized this area with emergent abilities, including unprecedented reasoning [241, 256], generalization [181, 239], and knowledge comprehension [163, 205]. LLMs are operated under the pretraining paradigm on extremely large corpora to update billions of parameters. A lot of research has shown that LLMs understand UGC in a zero-shot manner, i.e., without a collection of examples for fine-tuning. LLMs have surpassed human performance in summarization [176], outperformed most human in several exams [159], and shown strong reasoning abilities with prompt engineering, including Chain-of-Thought [241], Least-to-Most [290], and Tree-of-Thought [256]. LLMs open a new era for UM research to re-think about UGC mining.

How has the research of graph-based UM developed? User interactions with online content and users are naturally defined as edges that connect nodes of users or things. The user data can be defined as a graph. Heterogeneous graphs contain multiple types of nodes (e.g., users, items, places) and relations. Temporal/weighted graphs have timestamps/weights labeled on the interactions. Attributed graphs allow nodes to have a set of attribute-value pairs (e.g., age of a user, color of a product). In text-rich graphs, the nodes have long-form text attributes. Random walk with restarts provides a closeness score between two nodes in a weighted graph, and it has been successfully used in numerous settings (e.g., personalized PageRank [162]). Matrix factorization (MF) decomposes the user-item interaction matrix into the product of two matrices or known as latent features of users and items [91, 94, 107]. Regarding collaborative filtering, MF performs better with explicit feedback ratings, while RWR exploits the global popularity of items. It is actually a basic embedding model [272]. With deep learning, Node2Vec extracts sequences from the graph with random walks and uses Word2Vec to learn node embeddings [65]. However, encoding a graph into sequences would cause information loss. Graph neural networks (GNNs) employ the message-passing mechanism for deep representation learning on graphs. Specifically, the family of Graph Convolutional Network (GCN) [105] has greatly improved the performance in recommendation [49, 69], user profiling [24, 250], user behavior prediction [224, 263], and suspicious user detection [54, 55].

Why are LLMs revolutionizing text- and graph-based UM research? User modeling involves a series of machine learning tasks on text and graph data, such as text classification, node classification, link prediction, and time series modeling. Putting into context, the tasks can be sentiment analysis, natural language inference (NLI), user and product categorization, social relationship prediction, and temporal behavior prediction. Traditionally, the solution must be a specific model for a specific type of data and be trained on a specific set of annotations. For example, due to schema differences, two text classifiers had to be trained for the sentiment analysis and NLI tasks separately. Also, two networks or at least two modules in a graph neural network (GNN) were trained to predict if a user makes a new friend and purchases an item, respectively. Moreover, the textual information of user and/or product profiles is quite limited for learning and prediction due to the long-tail distribution.

LLMs have changed the paradigm of solution development. First, if designed properly, the prompts are able to treat most of the text-to-label tasks in LLMs as a unified text generation task; annotation data become not desperately needed; and the performance can even be comparable to or better than the traditional models. This is because LLMs were pre-trained on extremely large corpora and fine-tuned to follow the instructions in the prompts. Second, the prompts can be designed for learning tasks on graph data. For example, one can ask LLMs "if a user bought an Apple watch yesterday, will the user consider buying a pair of running shoes?" The "analysis" by LLMs can provide additional information to existing user-item link predictors. Third, all text information can be automatically expanded by LLMs. The relevant parametric knowledge augments the input of machine learning models and reduces the task difficulty.

LLMs have showcased robust capabilities in characterizing users' personalities [184], discerning users' stances [271], pinpointing user preferences [52], and beyond. Also, they have demonstrated marked proficiency in node classification [259], node attribute prediction [70], and graph reasoning [226]. Preliminary research focuses on leveraging LLMs for user modeling (LLM-UM) to integrate text-based and graph-based methods. For user profiling, GENRE [135] leverages ChatGPT as a user profiler by feeding users' behavior history and prompting the model to infer the users' preferred topics and regions. These LLM-generated profiles serve as important features for click-through rate recommendation models and resolve the anonymity problem in collecting user profiles. For recommendation, Kang et al. [100] use LLMs to predict user ratings based on their behavior history and find that LLMs typically demand less data while maintaining world knowledge about humans and items. For personalization, LaMP [189] proposes a benchmark incorporating personalized text generation and classification tasks as well as a retrieval-augmented approach. LLMs can be personalization tools for their understanding of user data. On suspiciousness detection, Chiu et al. [29] employ GPT-3 to detect hate speech, discovering that LLMs are able to identify abusive language with limited labels.

Difference from existing surveys. Given the growing interest and expanding body of work in user modeling (UM) with LLMs, this is an ideal opportunity to provide a comprehensive review that accomplishes several goals. In this survey, we analyze the advantages of LLMs in boosting existing user modeling techniques, introduce a taxonomy that categorizes LLM-UM techniques from the perspective of methodology, provide an in-depth review of specific techniques for a wide range of real-world applications, and finally outline the challenges and potential future directions in the field. We aim to provide a handbook for researchers and practitioners in the relevant fields, so they are able to confidently use LLMs to design and develop effective UM approaches.

It is worthwhile to discuss the uniqueness of our effort. Farid et al. [53] conducted a survey on user profiling approaches as a part of user modeling before LLMs came out. Li and Zhao [124] reviewed representation learning methods for user modeling, which was a prevailing paradigm before the advent of LLMs. He et al. [71] focused on user behavior modeling within recommender systems, narrowing down the scope specifically to item recommendations in the pre-LLM era. In the context of the LLM era, Fan et al. [50], Wu et al. [243] investigated the applications of LLMs in recommender systems. Additionally, Lin et al. [130] examined approaches that incorporated LLMs to enhance recommender systems, while Chen et al. [22] summarized recent work, challenges, and future directions in LLM-based personalization. However, these surveys discussed specific application goals of user modeling, either recommendation or personalization, which represents only a small portion of user modeling. To the best of our knowledge, there is no survey that summarizes the LLM-UM research work. Hence, our survey aims to fill this gap by providing a comprehensive summary and inspiring future research directions.

The remainder of this survey is organized as follows (Figure 11). Section 2 gives the background of user modeling techniques and large language models and gives the motivation of why LLMs are good tools for next-generation user modeling. Section 4 introduces two taxonomies of LLM-UM based on their approaches and applications. Section 5 summarizes the approaches to LLM-UM and how LLMs can integrate text and graph-based methods in existing works, including leveraging LLMs as enhancers, predictors, and controllers. Section 6 elaborates on the LLM-UM applications, including personalization and suspiciousness detection. Finally, Section 7 delves into current challenges and future directions pertaining to the LLM-UM topic.

2 Background

2.1 User Modeling

User modeling (UM) involves insight extraction or prediction from user data, such as profiles, personality traits, behavior patterns, and preferences. The insights can be utilized to customize and optimize user-oriented systems or services, enabling them to adapt effectively to the unique requirements of individual users [124]. From a user data standpoint, there are two primary types: user-generated content (UGC) and user-user/item interaction. These data modalities encompass textual content and graph-based interactions. User modeling techniques can be broadly classified into two categories: text-based methods and graph-based methods, each with a distinct emphasis on UGC and user interaction graphs, respectively.

Text-based UM Methods. The text-based UM methods focus on mining user-generated content, understanding user profiles, personalities, and subsequently inferring user preference and providing personalized recommendations and assistance. Text-based UM methods are highly relevant to natural language processing and have experienced several breakthroughs in the last two decades.

Initially, people widely use Bag-of-Words (BoW) [67] that maintains a discrete word vocabulary and represents text as an unordered collection of words. However, BoW disregards grammar, word order, and semantics. Later, topic models are proposed to identify groups of similar words in an unsupervised statistical machine learning manner [5]. The most representative topic model is Latent Dirichlet Analysis (LDA) [15], which considers documents as a mixture of topics and topics as a mixture of words. Latent semantic analysis (LSA) [47] is another popular topic model that is based on the principle that words close in meaning tend to be used together in context.

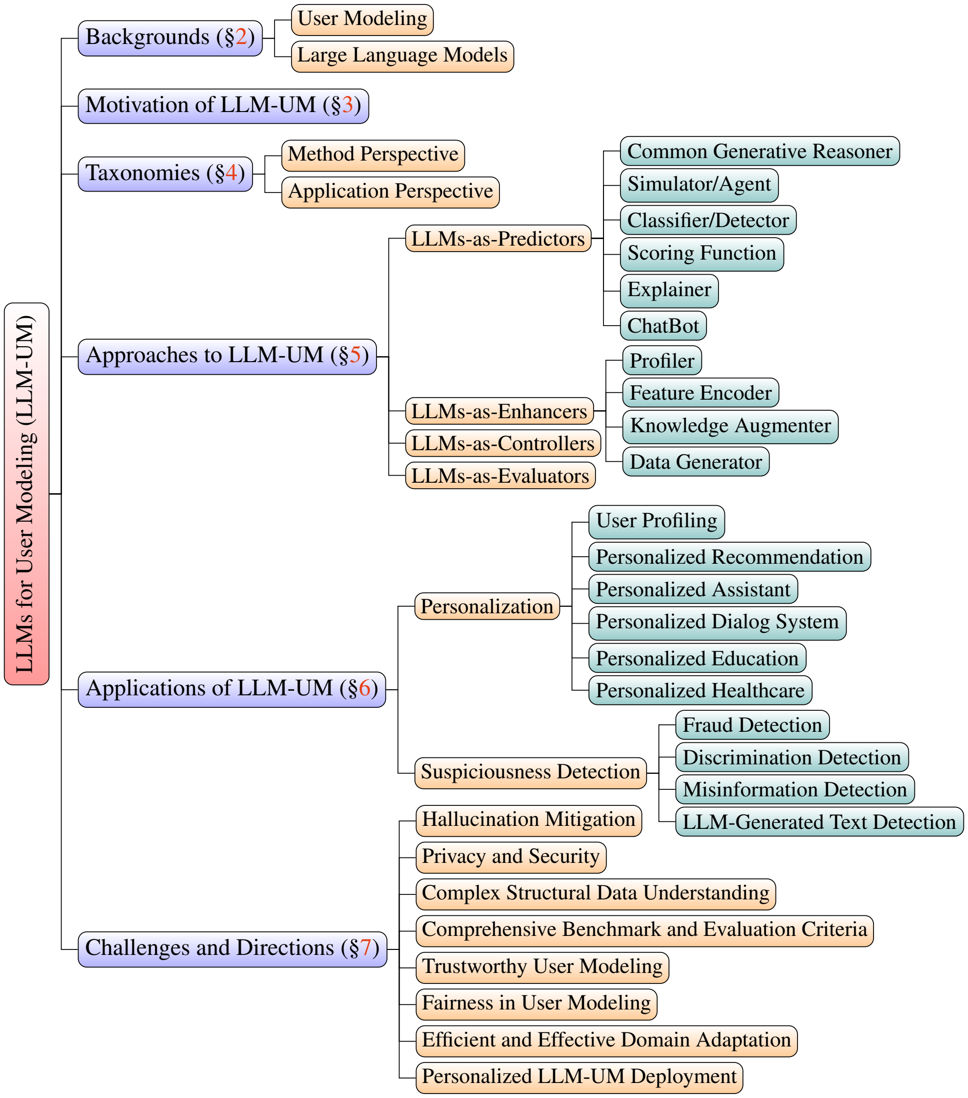


Figure 11: Structure of this survey.

Soon after, Mikolov et al. [148] introduced neural network techniques for UGC mining and understanding, including the continuous Bag-of-Words (CBOW) that predicts current words based on the context and continuous skip-gram model that predicts surrounding words given the current words. The emergence of Word2Vec brings new ideas to user modeling; for example, Hu et al. [75] use Word2Vec to encode user search history and predict users' age, gender, and education. However, the simplicity of Word2Vec leads to a limited semantic understanding of UGC. The emergence of Transformer architecture [219] and pretrained language models (PLMs) [102, 179] dramatically change the landscape of natural language processing, resulting in significant progress in text-based user modeling. For instance, BERT4Rec [204] utilizes a bidirectional self-attention mechanism to learn user behavior sequence representations for recommendation. Recently, large language models (LLMs) and their unprecedented emergent ability lead to a revolution in text-based user modeling. Specifically, LLMs are trained on large amounts of textual data with billions of learnable parameters to understand the patterns, semantics, and structure of natural language. Some preliminary explorations have shown LLMs are instrumental in enhancing functionality and adapting to users' specific needs in user modeling systems. LLMs have the capability to serve as a user profiler to generate user characteristics and personality, as in representative work such as GENRE [135], PALR [26], and MBTI-based assessment Rao et al. [184]. Furthermore, LLMs are proven to be powerful recommender systems, evidenced by LLMRec [134], Chat-REC [63], and LKPNR [187]. LLM personalization has also been a prominent area of research, delivering products such as LaMP [189], AuthorPred [117], and NetGPT [25]. Additionally, several studies have explored the utilization of LLMs for the detection of misinformation and misbehavior [110, 152].

Graph-based UM Methods. The graph-based user modeling methods focus on learning from graph structures about user-user/item interactions. Moreover, the timestamp and intrinsic heterogeneity of interaction networks enrich the graph information, together forming a temporal heterogeneous graph. To mine the graph and temporal data for user modeling, PAGE [162] first propose PageRank, which measures the importance of a node based on times visited by the random walk. Collaborative filtering [193] is then proposed, assuming that users with similar behaviors would rate and act on items in a similar manner. Matrix factorization, as a collaborative filtering technique, has dominated the user modeling field for quite a few years since Netflix competition for its scalability and flexibility [107, 113, 202]. In its basic form, matrix factorization features both items and users in latent space based on user-item interaction history. High correspondence between the item and the user leads to a recommendation. Later, deep learning models are introduced to learn high-quality user and item embeddings in latent spaces [45, 65, 172, 209]. Node2Vec [65] is the pioneering work that uses random walks on social networks to generate sentences from graph structures and then feeds them into a Word2Vec model [148] for graph embedding learning. Metapath2Vec [45] extends Node2Vec to heterogeneous graphs. To mitigate Node2Vec's information loss by extracting graphs into sequences, Graph Neural Network (GNN) is applied to advance user modeling with their robust structural encoding capabilities [77, 105, 220]. The core of graph neural networks lies in the message-passing mechanism that propagates node representations and aggregates neighborhood representations. The most representative GNN is graph convolutional network (GCN), which aggregates the information of different neighbors equally. Then the graph attention network (GAT) uses an attention mechanism to learn the attention weights of neighborhoods for higher-quality node representations. Specifically in user modeling applications, GNN-based methods are widely adopted in advanced user modeling systems and have achieved state-of-the-art performance. For instance, Ying et al. [261] design an efficient random walk algorithm, merge it with a type of GNN named GraphSAGE, and deploy the system on Pinterest.

2.2 Large Language Model

Language models are probabilistic models of natural language that can generate the likelihood of word sequences so as to predict the probabilities of future tokens [119, 282]. Large language models (LLMs) refer to the deep neural language models with billions of learnable parameters that are pretrained on an extremely large textual corpus to understand the distribution and structure of natural language [282]. Thanks to the efficiency of the Transformer architecture [219], almost all large language models employ it as the backbone. There are three types of language model design: encoder-only (e.g., BERT [102]), decoder-only (e.g., GPT [180]), and encoder-decoder (e.g., T5 [182]). Encoder-only models, specifically for BERT, use bidirectional attention to process token sequences and are pre-trained on masked token prediction and next-sentence classification tasks. This process can extract semantic embeddings for general purposes and enable the models to quickly adapt to diverse downstream tasks after fine-tuning. Decoder-only models, such as GPT, conduct text-to-text tasks based on the transformer decoder architecture. They are trained on the next token prediction tasks from left to right generation. Encoder-decoder models, such as T5, are trained on text-to-text tasks. Their encoders extract contextual representations from the input sequence, and their decoders use cross attention to map latent representations back to the text output space. In the context of LLMs, most models follow the decoder-only architecture as it simplifies the model and makes efficient inferences [232].

Recently, researchers have found that scaling pretrained language models' training data and parameter size often leads to a significant performance gain, a.k.a scaling law [101]. The large language models present emergent abilities [240], referring to the abilities that are not present in small models. Typically, there are three types of well-studied emergent abilities: in-context learning (ICL), instruction following, and step-by-step reasoning. In-context learning assumes that the language model has been provided with natural language instructions and/or several task demonstrations. LLMs can generate the expected output for test instances by completing the word sequence of input text without requiring additional training or gradient update, which is first introduced by GPT-3 [17]. Recent ICL research focuses on reducing inductive bias [116, 201]. The instruction following ability means that LLMs are shown to perform well on unseen tasks that are also described in the form of instructions after fine-tuning with a mixture of multi-task datasets formatted via natural language descriptions, known as instruction tuning. Instruction tuning improves the generalization ability of LLMs. The LLMs are better aligned with human intentions [238]. Recent instruction tuning studies focus on how to align LLMs with tasks and user preferences effectively [161] and efficiently [289]. Step-by-step reasoning means that LLMs can solve complex tasks that require multi-step reasoning. Chain-of-Thought (CoT) [241] introduces intermediate steps of reasoning steps in prompt design. Least-to-most [290] breaks down reasoning steps into simpler problems. Self-consistency [234] prompting further enhances LLMs reasoning by ensembling diverse CoT reasoning paths. Tree-of-Thought (ToT) [256] and Graph-of-Thought (GoT) [12, 257] enable LLMs to explore the thought processes in tree and graph structure, respectively. Moreover, preliminary explorations show that LLMs can use external tools [195], be parametric knowledge bases [163], have theory-of-mind [108], act as agents [230, 247], have graph understanding abilities [226], and can serve as optimizers [251].

Apart from using LLMs with frozen parameters, another line of work focuses on efficiently fine-tuning parameters in LLMs, namely parameter efficient fine-tuning, which helps LLMs efficiently adapt to specific tasks, datasets, and domain-specific understanding. Prefix tuning [126] keeps the language model parameters frozen and optimizes a small continuous task-specific vector called the prefix. Prompt tuning [115] is a simpler variant of prefix tuning, where some vectors are prepended at the beginning of a sequence at the input layer. Llama-Adapter [277] appends a set of learnable prompts as the prefix to the input instruction tokens in the higher transformer layers of Llama [216]. LoRA [74] injects trainable rank decomposition matrices into each layer of the transformer architecture, greatly diminishing the number of trainable parameters for downstream tasks. QLoRA [42] updates parameters through a frozen 4-bit LLM into low-rank adapters.

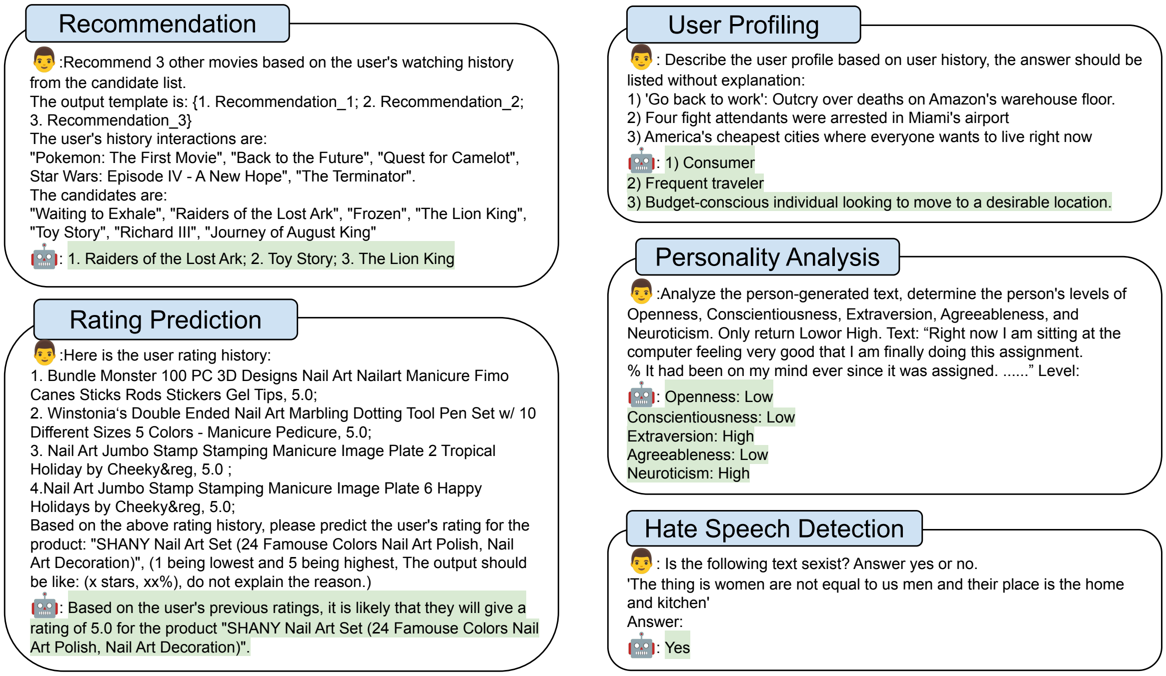
Existing LLMs can be categorized into open-source and API-based models based on accessibility. Open-source models provide access to both model weights and the ability to run the models on local machines, while API-based models restrict users from directly accessing the model weights and only allow them to interact with the models through an API. The open-source LLMs include T5 [182], Flan-T5 [34], OPT [278], BLOOM [192], GLM [270], Llama [216], and Falcon.[1] By fine-tuning or instruction-tuning Llama, a family of LLMs emerged, such as Alpaca [213] and Vicuna [28]. For API-based models, OpenAI offers four major series of GPT-3 [17] interface, including `ada`, `babbage`, `curie`, and `davinci`, corresponding to GPT-3 (350M), GPT-3 (1B), GPT-3 (6.7B), and GPT-3 (175B). GPT-3.5 series includes davinci and `turbo`:`turbo` leverages Reinforcement Learning from Human Feedback (RLHF) [161] to create human-like conversations. GPT-4 [159] is a well-acknowledged state-of-the-art and achieves astonishing performance on a wide range of tasks. There are some other API-based models such as BARD [146], Claude,[2] PaLM [31], BloombergGPT [245], and LangChain.[3]

Despite the thriving development of the LLM research, some challenges remain unsolved. For example, LLMs suffer from hallucination issues, that being said, LLMs generate text that is fluent and natural but unfaithful to the source content or under-determined [99]. LLMs also have biases due to unfathomable pre-training datasets, including political discourse [57], hate speech [81], and discrimination. Moreover, the inference latency of LLMs remains high because of low parallelizability and large memory footprints [175]. The remaining challenges also include limited context length [99], outdated knowledge [258], misaligned behavior [188], brittle evaluation

[1] https://falconllm.tii.ae/our-research.html

[2] https://www.anthropic.com/index/claude-2

[3] https://www.langchain.com/



Recommendation

👨 :Recommend 3 other movies based on the user's watching history from the candidate list.
The output template is: {1. Recommendation_1; 2. Recommendation_2; 3. Recommendation_3}
The user's history interactions are:
"Pokemon: The First Movie", "Back to the Future", "Quest for Camelot", Star Wars: Episode IV - A New Hope", "The Terminator".
The candidates are:
"Waiting to Exhale", "Raiders of the Lost Ark", "Frozen", "The Lion King", "Toy Story", "Richard III", "Journey of August King"

🤖: 1. Raiders of the Lost Ark; 2. Toy Story; 3. The Lion King

Rating Prediction

👨 :Here is the user rating history:
1. Bundle Monster 100 PC 3D Designs Nail Art Nailart Manicure Fimo Canes Sticks Rods Stickers Gel Tips, 5.0;
2. Winstonia's Double Ended Nail Art Marbling Dotting Tool Pen Set w/ 10 Different Sizes 5 Colors - Manicure Pedicure, 5.0;
3. Nail Art Jumbo Stamp Stamping Manicure Image Plate 2 Tropical Holiday by Cheeky&reg, 5.0 ;
4.Nail Art Jumbo Stamp Stamping Manicure Image Plate 6 Happy Holidays by Cheeky&reg, 5.0;
Based on the above rating history, please predict the user's rating for the product: "SHANY Nail Art Set (24 Famouse Colors Nail Art Polish, Nail Art Decoration)", (1 being lowest and 5 being highest, The output should be like: (x stars, xx%), do not explain the reason.)

🤖: Based on the user's previous ratings, it is likely that they will give a rating of 5.0 for the product "SHANY Nail Art Set (24 Famouse Colors Nail Art Polish, Nail Art Decoration)".

User Profiling

👨 : Describe the user profile based on user history, the answer should be listed without explanation:
1) 'Go back to work': Outcry over deaths on Amazon's warehouse floor.
2) Four fight attendants were arrested in Miami's airport
3) America's cheapest cities where everyone wants to live right now

🤖: 1) Consumer
2) Frequent traveler
3) Budget-conscious individual looking to move to a desirable location.

Personality Analysis

👨 :Analyze the person-generated text, determine the person's levels of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Only return Lowor High. Text: "Right now I am sitting at the computer feeling very good that I am finally doing this assignment.
% It had been on my mind ever since it was assigned. ......" Level:

🤖: Openness: Low
Conscientiousness: Low
Extraversion: High
Agreeableness: Low
Neuroticism: High

Hate Speech Detection

👨 : Is the following text sexist? Answer yes or no.
'The thing is women are not equal to us men and their place is the home and kitchen'
Answer:

🤖: Yes

Figure 12: Some examples of LLMs for recommendation, rating prediction, user profiling, personality analysis, and hate speech detection. They serve as compelling examples that demonstrate the ability of LLMs to effectively model, comprehend, and reason based on user-generated content (UGC) and user interactions.

[283], and limited structure understanding ability [27].

3 Motivation: Why LLMs for User Modeling?

LLMs have demonstrated novel capabilities, exhibiting strong potential in modeling and understanding user-generated content (UGC). In this section, we present related studies and specific examples (Figure 12) to support the claim that LLMs are effective tools for UGC modeling and comprehension.

A growing body of research focuses on utilizing LLMs for recommendation purposes, aiming to predict users' item-based interests from their behavior history. For example, PALR [26] generates a user profile based on UGC history and constructs prompts that incorporate the history, profile, and item candidates to enable LLMs to provide recommendations. As illustrated in the real example depicted in Figure 12, LLMs successfully generate reasonable recommendations by considering the user's viewing history. In the domain of user profiling, the objective is to summarize users' characteristics, including personality, interests, and topics of interest, based on their generated content and history. Recent research demonstrates that LLMs excel in user profiling. Liu et al. [136] put behavior history into LLMs and extract users' interest topics and physical regions and thus augment recommender systems. The example in Figure 12 further confirms the effectiveness of LLMs in summarizing user characteristics, preferences, and intentions. In the context of rating prediction, which seeks to comprehend user preferences based on past UGC, Kang et al. [100] investigate the ability of LLMs to predict user ratings for candidate items. They find that zero-shot LLMs lag behind traditional recommendation models due to the absence of user interaction data. LLMs can employ reasoning based on users' previous ratings to predict ratings for candidate items. Recent studies highlight LLMs' capacity to understand user personality. Ji et al. [87] employ various prompts to probe LLMs' ability to recognize user personality based on UGC history. LLMs achieve impressive results in personality recognition under zero-shot conditions. For users with a harmless behavior history, the UM is built upon general interests, content preferences, and interaction styles. However, the UM for users with suspicious behavior history, e.g., hate speech, would have to assess behavior patterns correlated to the suspicious activity, the risk of future incidents, and potentially flag these users for closer monitoring. From the application aspect, the presence or absence of hate speech in users' history has a significant impact on personalized recommendations. For example, the users with a hate speech history might be steered away from sensitive topics in recommendations. Therefore, suspiciousness detection, e.g., hate speech detection, is an essential application in user modeling. LLMs are good at detecting suspiciousness in UGC. Del Arco et al. [40] explore zero-shot prompting LLMs for hate speech detection. They find that zero-shot prompting can achieve performance comparable to and even surpass fine-tuned models. Figure 12 presents an example to further validate LLMs' effectiveness in detecting suspicious UGC.

Collectively, these studies and examples demonstrate LLMs' capabilities of modeling, understanding, and reasoning UGC and user behavior. They provide comprehensive evidence that LLMs can serve as valuable tools for user modeling, showcasing significant potential to improve user-oriented applications.

4 Taxonomies

In this section, we present taxonomies that classify LLM-UM techniques based on their approaches and applications. In Figure 11, we structure the survey according to these two taxonomies in Section 5 and 6, where the upper-level section is the parent category and the lower-level section is the child concept. For instance, the concept of “Approaches" to LLM-UM encompasses “LLMs-as-Predictors", “LLMs-as-Enhancers", “LLMs-as-Controllers", and “LLMs-as-Evaluators".

4.1 Approach Perspective

LLMs play diverse roles in LLM-UM systems. Based on their functionality, LLM-UM work can be categorized into four distinct approaches. (1) **LLMs-as-Predictors** involves leveraging LLMs for reasoning and generating answers directly. Depending on the role of LLMs, these LLM-UM methods can further be categorized as common generative reasoners for complex tasks, agents/simulators that model and predict human behavior, classifiers/detectors, scoring functions, explanation generation, and Chatbots for user modeling. (2) **LLMs-as-Enhancers** refers to using LLMs as augmentation modules to enhance the downstream user modeling system. LLMs can act as profilers to infer user preferences and characteristics, serve as feature encoders to generate latent UGC representations, augment discriminative user modeling systems with knowledge stored in LLMs, and generate high-quality data for small UM model training. (3) LLMs also possess the ability to control the pipeline of UM systems (**LLMs-as-Controllers**), automatically determining whether to execute certain operations. (4) LLMs can also serve as evaluators (**LLMs-as-Evaluators**) to score and analyze conversations and text generated under open-domain settings.

4.2 Application Perspective

Another taxonomy for categorizing LLM-UM is based on the downstream applications they address. Generally, LLM-UM systems aim to adapt to users' personal needs and detect suspiciousness in user data.

Personalization refers to tailoring experiences to individual preferences. Existing LLM-UM works can be categorized as follows: user profiling, which aims to tell the characteristics, personalities, stances, and sentiments towards certain targets; personalized recommendation gives the recommended items based on the user's behavior history and user profile; personalized assistants aim to adapt to users' specific needs and provide customized assistance; personalized dialog systems that interact iteratively with humans based on user data; and personalized applications in the education and healthcare domains. **Suspiciousness detection** is focused on identifying malicious users and UGC such as fraudsters, spammers, discriminations, and misinformation to preserve the integrity of social discourse [95, 208]. Existing LLM-UM works in suspiciousness detection can be categorized as follows: fraud detection for identifying malicious users or information in social platforms; discrimination detection to combat the spread of hate speech, predatory, and sexist; misinformation detection to identify fake news and propaganda; and LLM-generated text detection to identify if text is AI-generated for the integrity of education, online discourse, etc.

5 Approaches to LLM-UM

Given the strong capabilities in generation [282], reasoning [241], knowledge comprehension [205], and a good understanding of UGC as elaborated in Section 3, LLMs can be used to supercharge UM systems. The LLM-UM approaches can be generally categorized into three categories based on the LLMs' role, where the first envisions LLMs as the sole predictor that generates prediction directly, the second employs LLMs as enhancers to probe more information for the UM system augmentation, the third empowers LLMs with the ability to control the UM methods pipeline, automating the UM process, and the last uses LLMs as evaluators, assessing the performance of the system. It's worth mentioning that the form of “user model" in LLM-UM remains consistent with the previous definition, encompassing the knowledge and patterns that are discovered with the help of user-generated content, and user-user/item interaction networks [71]. The distinction of LLM-UM from previous paradigms lies in the approaches, where LLM-UM are empowered or enhanced by LLMs to gain user-related knowledge. In the following subsections, we summarize each paradigm and present representative approaches.

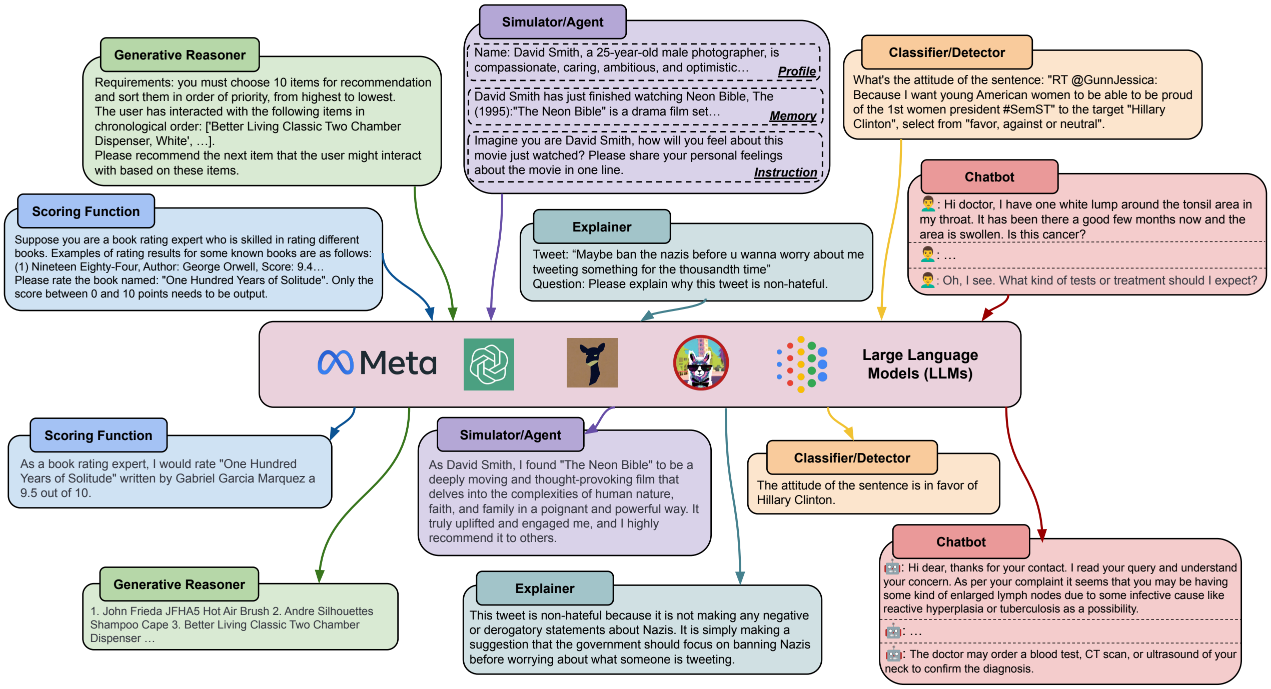
5.1 LLMs-as-Predictors

In this section, we introduce the LLMs-as-Predictors paradigm presented in LLM-UM works, which means LLMs are leveraged to make predictions and generate answers for downstream applications directly. More specifically, approaches in leveraging LLMs as generative reasoners, simulators/agents, classifiers/detectors, scoring/ranking functions, explainers, and Chatbots.

5.1.1 Common Generative Reasoner

Thanks to the strong generalization of LLMs, numerous UM systems incorporate them as generative reasoners to directly execute downstream tasks with a method design that is free of shenanigans. Considering the significant body of work falling under the generative reasoner category, we further classify them into non-tuning and fine-tuning methods based on the parameter status of LLMs.

Non-tuned LLMs. Utilizing LLMs as non-fine-tuning generative reasoners involves directly employing pretrained LLMs without adjusting their parameters for specific tasks. This approach preserves their versatility and ability to generalize across diverse circumstances. Therefore, the focus of non-tune LLMs research is mainly on prompt template and pipeline design. Di Palma et al. [44] present a comprehensive experimental evaluation framework to examine LLMs' ability in recommendation rigorously. Liu et al. [133] investigate LLMs capability in recommendation by designing prompts to test ChatGPT on directly performing rating prediction, sequential recommendation, direct recommendation, explanation, generation, and review summarization. NIR [229] designs a three-stage prompt template that guides LLMs to understand user preference, select representative movie-watching history, and recommend a list of movies. Sanner et al. [190] explore the prompting LLMs with item-based and language-based preference and find LLMs are strong near cold-start recommendation systems for pure language-based preference. BookGPT [288] design prompt templates and the overall pipeline to perform book rating recommendation, user rating recommendation, and book summary recommendation. Hou et al. [73] design prompting template by including the sequential interaction history, candidate items, and ranking instruction and show that LLMs have promising zero-shot ranking abilities in recommendation tasks. PALR [26]



Requirements: you must choose 10 items for recommendation and sort them in order of priority, from highest to lowest.
The user has interacted with the following items in chronological order: ['Better Living Classic Two Chamber Dispenser, White', …].
Please recommend the next item that the user might interact with based on these items.

Name: David Smith, a 25-year-old male photographer, is compassionate, caring, ambitious, and optimistic…

David Smith has just finished watching Neon Bible, The (1995):"The Neon Bible" is a drama film set…

Imagine you are David Smith, how will you feel about this movie just watched? Please share your personal feelings about the movie in one line.

What's the attitude of the sentence: "RT @GunnJessica: Because I want young American women to be able to be proud of the 1st women president #SemST" to the target "Hillary Clinton", select from "favor, against or neutral".

Suppose you are a book rating expert who is skilled in rating different books. Examples of rating results for some known books are as follows: (1) Nineteen Eighty-Four, Author: George Orwell, Score: 9.4…
Please rate the book named: "One Hundred Years of Solitude". Only the score between 0 and 10 points needs to be output.

Tweet: "Maybe ban the nazis before u wanna worry about me tweeting something for the thousandth time"
Question: Please explain why this tweet is non-hateful.

👨‍🦱: Hi doctor, I have one white lump around the tonsil area in my throat. It has been there a good few months now and the area is swollen. Is this cancer?

👨‍🦱: …

👨‍🦱: Oh, I see. What kind of tests or treatment should I expect?

As a book rating expert, I would rate "One Hundred Years of Solitude" written by Gabriel Garcia Marquez a 9.5 out of 10.

As David Smith, I found "The Neon Bible" to be a deeply moving and thought-provoking film that delves into the complexities of human nature, faith, and family in a poignant and powerful way. It truly uplifted and engaged me, and I highly recommend it to others.

The attitude of the sentence is in favor of Hillary Clinton.

1. John Frieda JFHA5 Hot Air Brush 2. Andre Silhouettes Shampoo Cape 3. Better Living Classic Two Chamber Dispenser …

This tweet is non-hateful because it is not making any negative or derogatory statements about Nazis. It is simply making a suggestion that the government should focus on banning Nazis before worrying about what someone is tweeting.

🤖: Hi dear, thanks for your contact. I read your query and understand your concern. As per your complaint it seems that you may be having some kind of enlarged lymph nodes due to some infective cause like reactive hyperplasia or tuberculosis as a possibility.

🤖: …

🤖: The doctor may order a blood test, CT scan, or ultrasound of your neck to confirm the diagnosis.

Figure 13: LLMs-as-Predictors, where LLMs are exclusively utilized to generate the predicted response.

designs prompt templates containing user interaction history sequences, LLM-generated user profiles, and candidates, then feeds them into a general-purpose LLM for recommendation. LaMP [189] offers an evaluation framework for personalized content generation and classification. Li et al. [117] design a multistage prompting strategy and multitasking to help frozen LLMs generate personalized content. The multistage contains retrieval, ranking, summarization, synthesis, and generation, and multitasking is to predict if the same author writes two documents. Li et al. [125] investigate ChatGPT's capabilities in personalized news recommendation, news provider fairness, and fake news detection, where they find ChatGPT is sensitive to input phrasing. Rao et al. [184] explore LLMs capabilities to analyze human personalities based on the Myers-Briggs Type Indicator (MBTI) test. Ji et al. [87] investigate the text-based personality recognition ability of ChatGPT and propose a level-oriented prompting strategy to optimize zero-shot chain-of-thought performance on personality recognition. Ghanadian et al. [64] leverage ChatGPT to assess the suicide risks on social media with zero-shot and few-shot prompting. Fan and Jiang [51] investigate ChatGPT's capabilities in discourse dialogue analysis, including topic segmentation, discourse relation recognition, and discourse parsing. Wu et al. [244] prompt LLMs to pairwise compare lawmakers and then scale the resulting graph using the Bradley-Terry model, finding that LLMs can be used to estimate the latent positions of politicians.

Fine-tuned LLMs. Fine-tuning LLMs can facilitate their adaptation to specific UM tasks and lead to better domain specialization. Given the massive number of parameters, research generally employs parameter-efficient fine-tuning (PEFT) techniques, such as LoRA, Llama-adapter, prompt-tuning, etc. InstructRec [276] considers LLMs as generative instruction following recommender and performs instruction tuning on Flan-T5-XL [34] model with a large amount of user-personalized instruction data. GIRL [287] proposes a Proximal Policy Optimization (PPO)-based reinforcement learning method to fine-tune LLMs, which aims to improve the LLM's ability to assess the compatibility between a job and a user. TallRec [8] leverages rec tuning samples containing user history behavior, new items, and feedback for instruction tuning, and it uses LoRA to improve the efficiency as well. GLRec [242] constructs prompts based on meta paths extracted from the user behavior graph, then leverages weighted path embeddings, instruction tuning, and LoRA for LLMs tuning and generates recommended items. Tie et al. [215] leverage fine-tuned LLMs to provide clinically useful, personalized impressions.

Combining non-tune and fine-tune paradigms can improve LLM performance. For example, LaMP [189] proposes a retrieval-augmented approach to retrieve personalized history to construct prompts for LLM generation under zero-shot and fine-tuning settings. Christakopoulou et al. [32] investigate the user's interest journey using few-shot prompting, prompt-tuning, and fine-tuning for journey name extraction. ReLLa [131] proposes to retrieve semantic user behavior to augment LLMs under the zero-shot settings and design retrieval-enhanced instruction tuning for the few-shot recommendation.

5.1.2 Simulator/Agent

Using LLMs as autonomous agents has been a prosperous research direction recently, which expects LLMs to accomplish tasks through self-directed planning and actions [230]. In the user modeling domain, LLMs can serve as a user simulator to imitate user behavior, conduct UM applications with planning and actions, use external tools, etc. Some works leverage LLM as a user simulator to predict user behavior. RecLLM [61] plugs LLM into the conversational recommender system to generate synthetic conversations to simulate user behavior for tuning system modules. UGRO [78] uses LLM as an annotation-free user simulator to assess dialogue responses. Kong et al. [106] fine-tune LLMs on genuine human-machine conversations to get a user simulator to generate a high-quality human-centric synthetic conversation dataset. PersonaLLM [90] investigates whether the behavior of LLM-generated personas can reflect certain personality traits accurately and consistently.

A few research studies have envisioned LLMs as agents and enabled them to interact with and explore the environment to gain a better understanding of user modeling tasks. Generative Agent [166] leverages LLMs to simulate human behavior in a social context with a memory module to record agent history and memories, and retrieve them for planning. RecMind [236] presents an LLM-powered autonomous recommender agent capable of providing precise, personalized recommendations through careful planning by utilizing tools for obtaining external knowledge and leveraging individual data. InteRecAgent [80] employs LLMs as the brain and recommender models as tools. It comprises key components such as a memory bus, dynamic demonstration-augmented task planning, and reflection. RecAgent [231] constructs a user simulator that regards each user as an LLM-based autonomous agent and lets different agents freely communicate, behave, and evolve within the recommender system. AutoGPT[4] demonstrates autonomous comprehension of specific objectives using natural language and carries out automated processes in a continuous loop, effectively accomplishing user-specific tasks. LLMs as agents can also empower LLMs with external tools and API for better user understanding. For example, Graph-Toolformer [273] teaches LLMs to use external graph-related API to augment LLMs' ability to reason over structural data and, therefore perform sequential recommendation and user rating prediction.

5.1.3 Classifier/Detector

LLMs can be prompted to serve as classifiers and detectors to analyze UGC, e.g., detect stance, hate speech, and suspicious behavior [186, 206]. Zhang et al. [271] discuss the potential of LLMs in stance detection and explanation generation. Hu et al. [76] propose ladder-of-thought (LoT) that assimilates high-quality external knowledge in small LMs to augment stance detection in LLMs. Mu et al. [153] investigate LLMs potential in zero-shot text classification in computational social science, including bragging, vaccine, complaint, and hate speech detection. Ziems et al. [292] comprehensively investigate LLMs capabilities in classifying dialect features, emotion, humor, empathy, discourse acts, and more detection and classification tasks. Parikh et al. [165] explores LLMs capabilities in user intent classification under zero-shot and few-shot settings. Qin et al. [178] construct an interactive depression detection system based on LLMs with CoT and a selector of users' tweets. Ferrara [58] discuss social bot detection in the era of LLMs, which points out that LLMs can be used to improve bot detection in low-resource language and domain. Qi et al. [177] design zero-shot, few-shot, and fine-tuning

[4] https://news.agpt.co/

methods to perform suicidal risk and cognitive distortion identification on Chinese social media. SentimentGPT [104] investigates the GPT's sentiment analysis capabilities in prompt engineering and fine-tuning. ALEX [97] prompts an LLM with the predicted results from BERT models for health data classification. Finch et al. [59] leverage GPT-3.5 to perform dialogue behavior detection for nine categories. Bhattacharjee and Liu [14] use ChatGPT to detect LLM-generated text. There are also works focused on designing prompts to enable LLMs to detect suspiciousness content such as clickbait [291], hate speech [29], and sexual predatory [156].

5.1.4 Scoring Function

LLMs can also be used as scoring or ranking functions that rate and rank items based on the user's behavior history. Chat-REC [63] and Liu et al. [133] leverage frozen LLMs to conduct item rating prediction. Kang et al. [100] comprehensively evaluate LLMs item rating prediction in zero-shot, few-shot, and fine-tuning scenarios. Dai et al. [36] formulate point-wise, pair-wise, and list-wise to elicit LLM's item scoring ability. BookGPT [288] employs LLMs to predict both book ratings and personalized user ratings on books. TallRec [8] employs instruction tuning for book and movie rating prediction. RecLLM [61] takes LLMs as a ranking and explanation module by feeding side information to produce a score for an item and an explanation for the score. Hu et al. [78] use LLMs to predict the user's satisfaction score with a conversation response and select the response with the highest satisfaction as output.

5.1.5 Explainer

The strong generation and reasoning ability make LLMs good explainers for the UM system, making its process understandable to humans. Liu et al. [133] prompts LLMs to generate recommendation explanations based on the user's interaction history. Chat-REC [63] explains the recommendation in the interaction process with users. Yu et al. [265] fine-tune Open-Llama to generate an explanation of the financial time series data forest. ALEX [97] uses LLM to generate an explanation for health-related UGC classification and would self-correct the prediction after explaining. Bao et al. [9] propose a self-reinforcement LLM-based framework in learnersouring to generate student-aligned explanations, evaluate, and iteratively enhance explanations automatically. Wang et al. [227] use LLMs to generate explanations for hateful and non-hateful content. Ziems et al. [292] investigate LLMs capabilities in social content explanation generation tasks like figurative language explanation and implied misinformation explanation.

5.1.6 Chatbot

LLMs are often used as chatbots to empower the UM system's interactivity and enhance understanding of UGC. Chat-REC [63] employs ChatGPT as a ChatBot to interact with human by incorporating user queries with user profiles and user behavior history. Lin and Zhang [129] observe ChatGPT's behavior in recommendation-oriented dialogues and demonstrate the potential for ChatGPT to serve as an artificial general recommender (AGR). [178] use ChatGPT as a chat interface with humans to understand the humans' mental status based on their tweets and responses. Lakkaraju et al. [112] investigate LLMs' potential in serving as personal financial advisors, where LLMs iteratively interact with humans as a chatbot. Hassan et al. [68] present a framework that enables LLMs to act as personal data scientists, where users interact with an LLM-based chatbot, and LLMs would return data analysis. CharacterChat [217] designs prompts with behavior preset and dynamic memory to help LLMs act as a chatbot with a specific personality. Zheng et al. [286] augment Llama with ChatGPT-generated dataset for optimized emotional support chatbot. He et al. [72] empirically study conversational recommendation with LLMs by constructing in-the-wild datasets. ChatDoctor [127] fine-tunes Llama using 100,000 patient-doctor dialogues to create a specialized Chatbot with enhanced accuracy in medical advice. Chen et al. [23] propose a three-stage pipeline to design prompts for ChatGPT to serve as both doctor and patient chatbots. GeneRec [233]

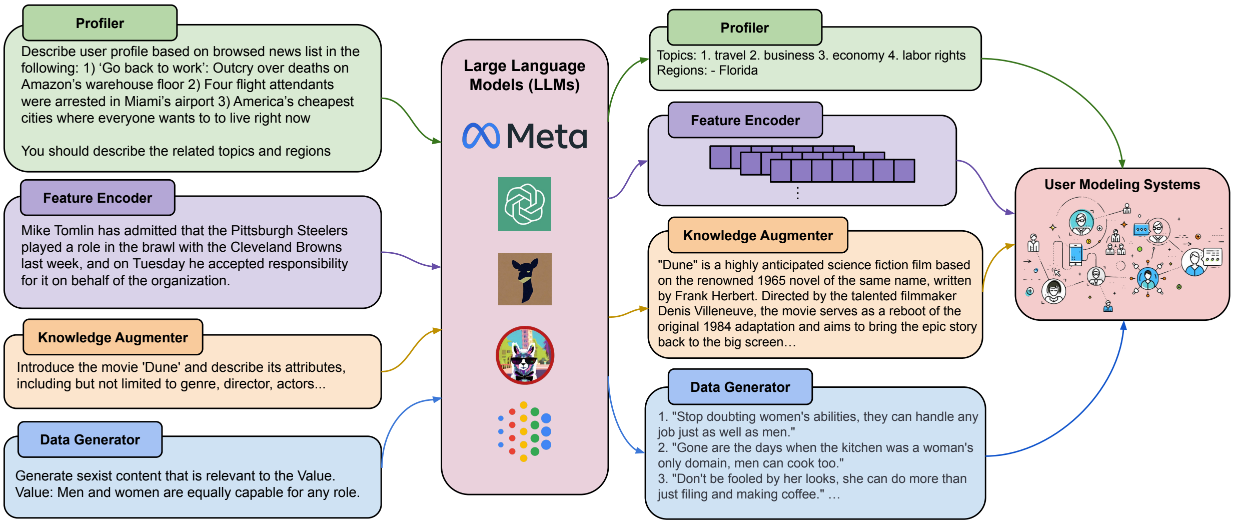

Profiler
Describe user profile based on browsed news list in the following: 1) 'Go back to work': Outcry over deaths on Amazon's warehouse floor 2) Four flight attendants were arrested in Miami's airport 3) America's cheapest cities where everyone wants to to live right now
You should describe the related topics and regions
Feature Encoder
Mike Tomlin has admitted that the Pittsburgh Steelers played a role in the brawl with the Cleveland Browns last week, and on Tuesday he accepted responsibility for it on behalf of the organization.
Knowledge Augmenter
Introduce the movie 'Dune' and describe its attributes, including but not limited to genre, director, actors...
Data Generator
Generate sexist content that is relevant to the Value.
Value: Men and women are equally capable for any role.
Large Language Models (LLMs)
Meta
Profiler
Topics: 1. travel 2. business 3. economy 4. labor rights
Regions: - Florida
Feature Encoder
Knowledge Augmenter
"Dune" is a highly anticipated science fiction film based on the renowned 1965 novel of the same name, written by Frank Herbert. Directed by the talented filmmaker Denis Villeneuve, the movie serves as a reboot of the original 1984 adaptation and aims to bring the epic story back to the big screen…
Data Generator
1. "Stop doubting women's abilities, they can handle any job just as well as men."
2. "Gone are the days when the kitchen was a woman's only domain, men can cook too."
3. "Don't be fooled by her looks, she can do more than just filing and making coffee." …
User Modeling Systems

Figure 14: LLMs-as-Enhancers, where LLMs are leveraged to generate user profiles, content embeddings, knowledge-augmented content, and training data to augment downstream user modeling systems.

takes ChatGPT as a user conversational interface and takes user instruction and feedback to generate personalized content.

5.2 LLMs-as-Enhancers

In this section, we analyze the approaches that leverage LLMs as enhancers in the UM models. That being said, LLMs are not used to generate task answers directly but are leveraged to serve as plug-in augmentation modules instead. Approaches use LLMs as profilers, feature encoders, knowledge augmenters, and data generators.

5.2.1 Profiler

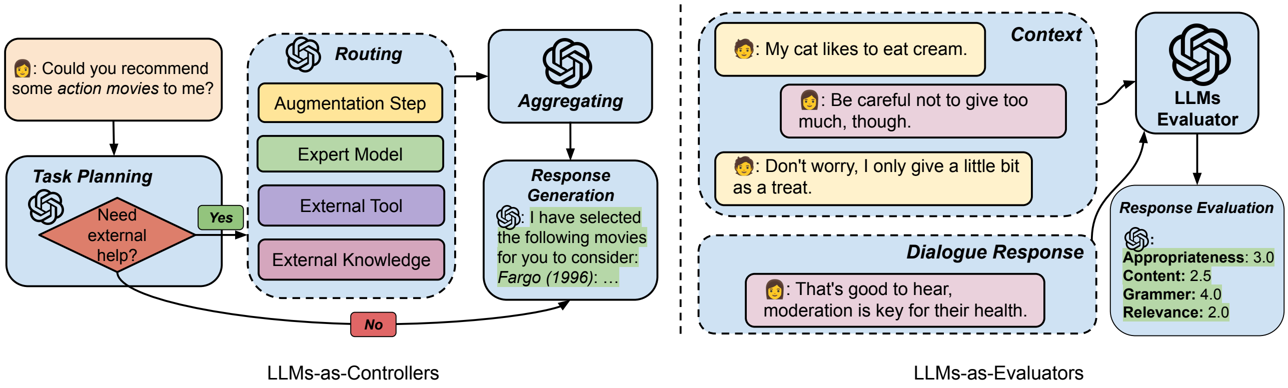
Using LLMs-as-profilers involves the creation of prompts based on users' history, including their watching, purchasing, and viewing activities. These prompts are then inputted into LLMs to generate various aspects of users' profiles, such as their characteristics, personality, geographical location, and areas of interest. The resulting user profiles are represented in natural language, making them easily understandable to humans. They are commonly employed as input for tasks such as recommendation and rating prediction, enabling downstream predictions to be tailored to the specific requirements of individual users. HKFR [260] utilizes user heterogeneous behavior, encompassing behavior subjects, behavior content, and behavior scenarios, and feeds it into ChatGPT to obtain user profiles. ONCE [136] and GENRE [135] employ LLMs to generate topics and regions of interest based on user browsing history. PALR [26] uses an LLM and user behavior as input to generate user profile keywords. KAR [246] leverages LLMs to generate user and item profiles, encompassing user preferences and factual knowledge about items, respectively. LGIR [46] completes users' resumes by incorporating explicit properties from their self-description and implicit characteristics from their behavior history. GIRL [287] leverages LLMs to generate suitable job descriptions based on the user's curriculum vitae to help the recommendation model better understand the job seeker's preferences. NIR [229] feeds user-watching history and design prompts for LLMs to generate user preference to augment recommendations. Once these user profiles, which include information such as age, gender, preferences, topics of interest, and geographic location, are obtained, they can be fed into the downstream user modeling system to enhance understanding of user preferences.

5.2.2 Feature Encoder

Given that LLMs have incredible user-generated content understanding and modeling capabilities, some research focuses on using LLM-generated text embeddings to enhance UM systems. GPT4SM [171] uses both PLMs and GPT to encode recommendation queries and candidate text for relevance prediction. LKPNR [187] uses open-source LLMs such as Llama and RWKV [169] to get news representation for better semantic capture capability. Li et al. [123] explore LLMs' potential in text-based collaborative filtering by using LLMs with parameters ranging from 125M to 175B as feature encoders. LLM4Jobs [121] leverages LLMs to embed both job posts and occupation taxonomy database and calculate their embedding similarity for recommendations. SentimentGPT [104] uses GPT to encode text embeddings for small ML model training on sentiment analysis tasks. KAR [246] leverages ChatGLM as a knowledge encoder to get latent embeddings for user profiles and item factual knowledge. These LLM-based text embeddings are then fed into the downstream models to inject the LLMs' UGC understanding ability into downstream task-agnostic models. LLMs are shown to have strong natural language understanding ability and rich open-world knowledge. Armed with LLM-encoded embeddings, the downstream UM systems could better understand the semantic information in UGC and leverage the world knowledge from LLM latent space.

5.2.3 Knowledge Augmenter

LLMs have demonstrated impressive capabilities in internalizing knowledge and responding to common queries [159, 161], which can serve as knowledge bases and bring factual knowledge to UM systems. Yin et al. [260] leverages LLMs to fuse diverse heterogeneous knowledge, including multiple behavior subjects, multiple behavior contents, and multiple behavior scenarios. Mysore et al. [154] augment narrative-driven recommendation using LLMs for author narrative query generation based on user-item interactions and train retrieval models with these LLM-augmented queries. KAR [246] prompts LLMs to generate factual knowledge relevant to the item for recommender system augmentation. GPT4Rec [120] leverages GPT to expand users' search queries to give item titles and users' history and feed item titles for rephrasing. Li et al. [122] use prompt tuning in LLMs to generate aspect embedding extraction and then feed the embeddings into aspect-based recommendation systems. LLM-Rec [145] proposes four prompting strategies to encourage LLMs to generate knowledge-augmented item descriptions for recommendation. Acharya et al. [3] leverage LLMs to generate detailed item descriptions for knowledge-augmented recommendations. Fang et al. [52] leverage ChatGPT knowledge to generate rephrases of training datasets instances to enhance model generalization and performance on unseen compositions. PULSAR [118] integrates LLMs in data augmentation to generate more knowledge relevant to the annotated training data. Knowledge stored in LLMs can also be used to augment dialogue. AugESC [284] uses LLMs to complete full dialogue and construct a scalable dataset to augment dialog systems' generalization ability in open-domain topics. Schlegel et al. [196] ask LLMs to generate a hypothetical conversation between the doctor and the patient based on a medical note and then use this data to train a specialized LM. Cohen et al. [35] leverage GPT-3 to do back translation and rephrase as data augmentations for hate speech detection. TacoBot [150] leverages LLMs to synthesize user intents of conversation for data augmentation. LLM-PTM [268] proposes an LLM-based privacy-aware augmentation for patient-trial matching, which leverages LLMs to create supplementary data points while preserving the semantic coherence of the original trail's inclusion and exclusion criteria. The collective findings of these works show that researchers could enhance UM systems performance by identifying and injecting external reasoning or factual knowledge into UM text input. Take movie recommendation as an example, LLMs provide contextual knowledge for candidate item descriptions under specific scenarios. This approach paves the way for the advancement of open-world recommendation, which integrates broader contextual information to enhance the recommender system.


: Could you recommend some *action movies* to me?
Task Planning
Need external help?
Yes
No
Routing
Augmentation Step
Expert Model
External Tool
External Knowledge
Aggregating
Response Generation
: I have selected the following movies for you to consider: *Fargo (1996)*: ...
Context
: My cat likes to eat cream.
: Be careful not to give too much, though.
: Don't worry, I only give a little bit as a treat.
Dialogue Response
: That's good to hear, moderation is key for their health.
LLMs Evaluator
Response Evaluation
:
Appropriateness: 3.0
Content: 2.5
Grammer: 4.0
Relevance: 2.0

LLMs-as-Controllers

LLMs-as-Evaluators

Figure 15: LLMs-as-Controllers and LLMs-as-Evaluators, where LLMs are employed to manage the LLM-UM pipeline and to evaluate the output of UM systems, respectively.

5.2.4 Data Generator

Owe to LLMs' strong sequence generation and data compression ability [267], there is a lot of research that leverages LLMs to generate training data or weak labels and feed these into small UM models for training [16, 170, 222], which can also be envisioned as knowledge distillation. VA-Models [7] generates value-aligned training data from LLMs by few-shot prompting, then uses the generated data to train small models on value-alignment judgment tasks. Tang et al. [211] investigates LLMs' potential to synthesize high-quality clinical data with labels and fine-tune local models for the downstream tasks. LLM4Job [121] leverages GPT-4 to generate occupation coding datasets. Su et al. [203] investigate fake news detection models' bias with LLMs-synthesized fake news articles. Veselovsky et al. [221] find LLMs-synthesized data have a different distribution than real-world data, and propose three strategies: ground, filter, and taxonomy-based generation strategy to combat this difference, which has proved to be effective in sarcasm detection. LLMs can also generate weak labels to enhance UM systems. For example, Leite et al. [114] propose to prompt LLMs with credibility signals to produce weak labels to enhance misinformation detection performance. Deng et al. [41] propose to generate weak financial sentiment labels for Reddit posts with an LLM and then use that data to train a small model that can be served in production. Foosherian et al. [60] find that LLMs can aid conversational agents in generating training data, extracting entities and synonyms, localization, and persona design. For the LLM-generated text detection domain, LLMs are employed to synthesize training corpus [21, 138, 262, 266]. Graph-Toolformer [273] uses ChatGPT to annotate and augment a large prompt dataset that contains API calls of external reasoning tools and uses the synthesized dataset to fine-tune open-source LLMs. RefGPT [252] uses LLMs to generate truthful and customized dialogues without hallucination.

LLMs can also generate high-quality conversational data. Zheng et al. [286] recursively prompting ChatGPT with in-context learning to generate an extensive emotional support dialogue dataset (ExTES) and use it to fine-tune Llama for optimized emotional support dialogue systems. Wang et al. [228] propose a LLMs cooperation system named a doctor-patient loop to generate high-quality conversation datasets. Overall, these works indicate that by training on LLM synthesized data, small UM models could inherit the strong user understanding ability from LLMs. These approaches are advantageous in low-resource circumstances or some applications that need efficient deployment. For instance, in healthcare applications where privacy concerns and data scarcity are prevalent, utilizing LLMs to generate supplementary data can significantly alleviate the low-resource and privacy problems and thus bolster the effectiveness of UM systems.

5.3 LLMs-as-Controllers

The scale of parameters of LLMs brings emergent abilities that have never been observed in small language models and gives LLMs unprecedented ability to control the system pipeline and supercharge the UM system for personal needs. Note that different from LLMs as agents that let LLMs freely explore and interact with the environment, LLMs-as-Controllers include works that have designed the entire pipeline and let LLMs decide whether to conduct certain operations. HuggingGPT [198] employs LLMs as a controller to manage and organize the cooperation of expert models. Specifically in the user modeling systems, RecLLM [61] leverages LLMs as a dialogue manager, which converses with the user, tracks context, and makes system calls when necessary. Chat-REC [63] lets LLMs decide when to use recommendation systems as external tools. Foosherian et al. [60] demonstrate that LLMs can assist the pipeline-based conversational agent in contextualization, intent classification to prevent conversational breakdown and handle out-of-scope questions, auto-correcting utterances, rephrasing responses, formulating disambiguation questions, summarization, and enabling closed question-answering capabilities. LLM4Jobs [121] constructs a pipeline that makes LLMs determine whether to do a summary on the job posting.

5.4 LLMs-as-Evaluators

Evaluating natural language generation (NLG), especially in the open-domain and conversation settings, has posed significant challenges in user modeling. The strong language modeling capabilities in LLMs open up new opportunities for these complex evaluations, and some research works propose to envision LLMs as evaluators for generative UM systems. Svikhnushina and Pu [207] leverage LLMs to approximate online human evaluation for dialogue systems. Huynh et al. [84] explores LLMs capabilities in dialog evaluation and the relation between prompts and training datasets. iEvaLM [235] proposed a conversational recommendation evaluation framework that leverages LLMs to simulate various interactions between users and systems. Zheng et al. [285] discover that using strong LLMs like GPT-4 as a judge can match both controlled and crowdsourced human preferences well. LLM-Eval [132] designs a prompt-based evaluation method that leverages a unified evaluation schema to cover multiple dimensions of conversation quality. Bhat et al. [13] takes a fine-tuned GPT-3 to evaluate the generated questions by classifying if the questions are useful to learning. GIRL [287] evaluates the recommended job results with the help of ChatGPT. These works reveal that LLMs can be an effective tool of assessing UM system output, determining the extent these outputs are customized to meet specific user needs. Particularly in conversational contexts, where conducting user studies could be expensive and prone to bias, LLMs provide a reliable and effective approach to assess the quality of complex and open-ended generations. Consequently, the LLMs-as-Evaluator paradigm enhances UM system development.

Personalized Prompting.

It is worth mentioning that to make LLM-UM adapt to an individual's specific needs and generate personalized content, existing research works design prompting templates to encode UGC and user behavior history to help LLMs understand users' preferences. Existing personalized prompting paradigms fall into three categories: vanilla, retrieval-augmented, and profile-augmented methods. We provide specific examples of these prompts in the context of the personalized product rating task below.

- **`Vanilla Personalized Prompt:`** Here is the user rating history: {{`all user behavior history`}}. Based on the above rating history, please predict the user's rating for the product: {{`query item`}}.
- **`Retrieval-Augmented Personalized Prompt:`** Here is the user rating history: {{`top-k user behavior history`}}. Based on the above rating history, please predict the user's rating for the product: {{`query item`}}.

- **`Profile-Augmented Personalized Prompt:`** {{`User Profile`}}. Here is the user rating history: {{`top-k user behavior history`}}. Based on the above rating history, please predict the user's rating for the product: {{`query item`}}.

Most works under the LLMs-as-Predictors paradigm adopt the in-context learning paradigm and encode the entire user behavior history as in-context examples (**Vanilla Personalized Prompt**). BookGPT [288], for instance, employs a few-shot prompting strategy to enable LLMs to understand the correlation between book content and personalized ratings. Dai et al. [36] encode user rating history as few-shot demonstration examples. Liu et al. [133] supply the LLM with the user's interaction history relating to the task, enabling the generation of personalized content. Christakopoulou et al. [32] utilize few-shot prompting to demonstrate the user research journey to LLMs. and find that feeding partial user behavior log causes significantly lower performance. Moreover, BookGPT [288] discovers that a longer user behavior history would bring better performance. Considering the increasing length of UGC and user behavior history and the limited context length of LLM, some studies have applied retrieval methods to select the most relevant user data for enhancing LLM personalization (**Retrieval-Augmented Personalized Prompt**). For instance, LaMP [189] introduces a retrieval-augmented method to obtain the most relevant content in behavioral history and incorporate it into the prompt.

Except for previously mentioned works that fall into the LLMs-as-Predictors paradigm, some works can be categorized into the LLMs-as-Enhancers paradigm that employs LLMs to rephrase and summarize user preferences and profiles based on their history (**Profile-Augmented Personalized Prompt**). Richardson et al. [185] suggest the use of instruction-tuned LLMs to generate an abstract summary of user data, augmenting retrieval-based personalized methods like LaMP. HKFP [260] inputs user behavior history into LLMs to fuse heterogeneous user knowledge, which assists fine-tuned LLMs in understanding user preferences. Note that the generated profile could vary in different tasks. For example, in rating prediction tasks, Richardson et al. [185] prompt instruction-tuned LLMs to summarize the user's most common positive and negative score as profile, while generating user research interest as profile in the personalized citation identification task.

6 Applications of LLM-UM

We introduce the applications that can be categorized into personalization and suspiciousness detection.

6.1 Personalization

Personalization in user modeling refers to tailoring and customizing a system's interactions, content, or recommendations to meet the specific needs, preferences, and characteristics of individual users. In this section, we look into LLM-UM for personalization applications, including user-generated content (UGC) analysis, user behavior prediction, personalized assistance, personalized recommendation, personalized dialog system, personalized education, and personalized healthcare.

6.1.1 User Profiling

User profiling refers to mining the characteristics, personality, and potential preferences based on UGC and user behavior, paving the way for downstream personalized applications. User profiling mainly includes detecting users' stances and sentiments and analyzing users' characteristics, personalities, etc. Zhang et al. [271] discuss LLMs' impact in stance detection topics and the opportunities they bring. LoT [76] is proposed to help LLMs assimilate high-quality external knowledge to boost stance detection. Mu et al. [153] and Ziems et al. [292] comprehensively investigate LLMs in the computational social science tasks, including detecting sarcasm, hate speech, ideology, stance, and more in UGC. SentimentGPT [104] is a pioneering to leverage LLM for UGC sentiment analysis. Wu et al. [244] discover that LLMs can be used to estimate the latent stance of politicians and give solutions to complex social science measurement problems. For general personality modeling, Rao et al. [184] use LLMs to assess human personality based on the MBTI test. Ji et al. [87] employ a level-oriented prompting strategy to analyze the user's personality in a given text. Christakopoulou et al. [32] use LLMs to mine users' interest journey and provide deeper, more interpretable, and controllable user understandings.

6.1.2 Personalized Recommendation

LLMs are adopted in the recommendation system to offer personalized suggestions towards candidate items tailored to meet user preferences and specific needs. The recommendation tasks can be further categorized into top-k recommendation, rating prediction, and conversational recommendation.

Top-k Recommendation. The top-k recommendation task directly predicts the top-k favorite items based on the user's behavior history. Most methods directly design prompts feeding user behavior history and optional characteristics into LLMs to generate recommended items. In context learning is the most common paradigm for the top-k recommendation, which gives several exemplars and recommendation results to help LLMs better understand tasks. Representative works include [36, 46, 73, 133, 135, 235, 274]. Zhang et al. [276] further utilized Chain-of-Thought prompting to conduct a top-k recommendation task. For better representation and domain adaptation, researchers also fine-tune LLM's parameters. Chen [26] and GenRec [86] fine-tune a Llama-7B model to help the model adapt to recommender system and generate items. Zhang et al. [276] conduct instruction tuning on a Flan-T5-XL model to help it adapt to recommendation. GIRL [287] further proposes a reward model and uses reinforcement learning to provide better feedback for LLMs fine-tuning.

Rating Prediction. The rating prediction means predicting the user's rating for given items. In this process, LLMs give the predicted scores for items in the context of the user's behavior and UGC history. The rating prediction task can probe LLM's capabilities in understanding user preference and can also be understood as user behavior prediction. Similar to top-k recommendation, rating prediction can also be categorized into frozen LLMs and fine-tuning LLMs. Armed with frozen LLMs, BookGPT [288] and Dai et al. [36] conduct prompt engineering to make LLMs generate predicted ratings; KAR [246] takes LLMs to generate user profile and factual knowledge of items and feed them into the discriminative recommendation for rating prediction. For rating prediction with fine-tuned LLMs, Kang et al. [100] fine-tune LLMs in the rating prediction task; TallRec [8] integrate LoRA and instruction tuning for rating prediction task adaptation; GLRec [242] incorporates meta-paths into soft prompt and conduct instruction tuning for item generation. Further, Graph-Toolformer [273] fine-tunes LLMs and empowers them to use external graph reasoning tools for user rating prediction.

Conversational Recommendation. Conversational recommendation means using the system to interact with the user and understand the user's preference through conversation and then generate recommended items in the conversation. Chat-REC [63] converts the user profile and historical interactions into prompts to build conversational recommendation systems. GeneRec [233] adopts ChatGPT to personalize content generation, and it leverages user instructions to acquire users' information needs. Lin and Zhang [129] envision ChatGPT as an Artificial General Recommender (AGR) that comprises conversationally and universality to engage in natural dialogues and generate recommendations across various domains.

6.1.3 Personalized Assistance

Personalized assistance refers to using LLM techniques to tailor and customize generated content based on individual preferences, behavior, and characteristics of users. Tian et al. [214] empirically study the LLMs' potential as fully automated programming assistants in the tasks of code generation, program repair, and code summarization. TacoBot [150] is an LLM-augmented task-oriented dialogue system that guides users through complex real-world tasks with multiple steps. LaMP [189] uses LLMs to conduct personalized text classification and personalized text generation, such as personalized citation identification and personalized news headline generation, etc. DISC-LawLLM [269] is an intelligent system that utilizes LLMs to provide a wide range of personalized legal advice. FinGPT [142] is an LLM that can offer personalized investment advice based on the user's risk and financial goals. Chakrabarty et al. [20] investigate LLMs' ability in creative writing assistance, including planning, translating, and reviewing processes.

6.1.4 Personalized Dialogue System

LLMs are also widely adopted in dialogue systems and combined with the user's behavior history and preference to provide personalized user experiences. Hudeček and Dušek [83] utilized LLMs to retrieve the user behavior context and user history for personalized conversational response generation. DiagGPT [19] extends LLMs to task-oriented dialogue scenarios, in which LLMs need to pose questions and guide users toward specific task completion proactively. Cho et al. [30] use GPT-2 to generate dialogue data while detecting the user's persona. RefGPT [252] proposes to generate enormous truthful and customized dialogues without worrying about factual errors caused by the model hallucination to enable the personalized dialogue system.

6.1.5 Personalized Education

LLMs have shown great potential in promoting the equality of education and improving the existing education paradigm by adapting LLM-based tools to tailor for students and instructors [98, 249], and education for data scientists [218]. Koyuturk et al. [109] discover that with multiple interaction turns, LLMs can adapt the educational activity to the user's characteristics, such as culture, age, and level of education, and its ability to use diverse educational strategies and conversational styles. EduChat [37] is an LLM-based chatbot that aims to strengthen personalized, fair, and compassionate intelligent education, serving teachers, students, and parents. Sharma et al. [197] investigate ChatGPT's performance on the United States Medical Licensing Examination (USMLE) and point out that ChatGPT can be an invaluable tool for e-learners. Elkins et al. [48] investigate ChatGPT's performance in generating educational questions in classroom settings and find them high-quality and sufficiently useful. Ochieng [157] delve into LLM's ability to participate in educational guided reading, including generating questions based on the input text and recommending content based on the user's response. C-LLM [158] examines the implications for LLMs for AI review and assessment of complex student work. Phung et al. [174] comprehensively investigate ChatGPT's capability in a set of programming education scenarios and find GPT-4 comes close to human tutor in several scenarios.

6.1.6 Personalized Healthcare

LLMs also play an important role in empowering healthcare services and providing personalized service. Liu et al. [139] discover that LLMs are capable of grounding various physiological and behavioral time-series data and making meaningful inferences under the few-shot settings. Wang et al. [237] investigates the performance of LLMs on clinical language understanding tasks and introduces self-questioning prompting to enhance LLM in clinical scenarios. HeLM [10] enables LLMs to use high-dimensional clinical modalities to estimate underlying disease risk with individual-specific data. PharmacyGPT [143] utilizes LLms to generate comprehensible patient clusters, formulate medication plans, and forecast patient outcomes. Zhang et al. [275] utilize LLMs to identify patients with specific medical diagnoses and provide diagnostic assistance to healthcare workers. Zhongjing [255] introduces a Llama-based LLM with expertise in the Chinese medical domain and can provide personalized advice for the user's specific case.

LLMs are widely applied for mental health. Ghanadian et al. [64] utilize LLMs for suicidality assessment from social media. Qi et al. [177] takes LLMs to evaluate suicidal risk and cognitive distortion identification on Chinese social media platforms. Wang et al. [228] leverage LLMs to generate note-oriented doctor-patient conversations. Chen et al. [23] utilize ChatGPT to simulate conversations between psychiatrist and patient based on user experience. Fu et al. [62] present an LLM-empowered framework that assists non-professionals in providing psychological interventions on online user discourse. Mental-LLM [248] investigate LLM's capabilities in mental health tasks and find the superiority of instruction tuning in boosting LLMs' performance in mental health tasks. Lai et al. [111] propose an LLM-based assistant for question-answering in psychological consultation settings to ease the demand for mental health professions. Peters and Matz [173] assess the ability of GPT-3.5 and GPT-4 to infer the psychological dispositions of individuals from their digital footprints.

6.2 Suspiciousness Detection

User modeling involves the process of understanding and predicting users' behavior and preferences. By creating a comprehensive model of user's normal behaviors, it becomes possible to detect deviations from this norm. For users with suspicious behavior history, UM system could potentially isolate or warn users who exhibit harmful behaviors, while offering a more open environment to those with positive engagement histories. Therefore, suspiciousness detection is a key application of user modeling. Suspiciousness detection refers to the process of identifying or recognizing behaviors, actions, or patterns that are deemed to be suspicious or potentially indicative of anomalous, illegal, harmful, or malicious activities [92, 93, 281]. This section introduces leveraging LLM-UM to detect fraud, hate speech, misinformation, misconduct, and LLM-generated content.

6.2.1 Fraud Detection

Suspiciousness in fraud detection refers to fraudulent and deceptive behavior, which is widely adopted in financial transactions, social networks, and more. Some research leverages LLMs to supercharge fraud detection models. [291] design prompts to enable LLMs for clickbait detection and achieves satisfied performance. Yang and Menczer [253] present a case study on a Twitter botnet that employs ChatGPT to generate human-like content, while state-of-the-art LLM content classifiers fail to detect them. Spam-T5 [110] fine-tunes T5 for spam email detection and outperforms baselines with limited training samples. Ayoobi et al. [6] leverage LLMs to generate fake LinkedIn profiles and develop the Section Tag Embeddings to detect fake profiles. Shukla et al. [200] explores GPT-3 and GPT-4 for fraudulent physician review detection. Ziems et al. [293] employs GPT-4 to explain the decisions of classicial machine learning models on network intrusion detection.

6.2.2 Discrimination Detection

LLMs' strong natural language understanding ability can supercharge hate speech detection, which is of great importance in social content moderation [147]. Chiu et al. [29] leverage GPT-3 to identify sexist and racist UGC under zero-shot, one-shot, and few-shot settings. Cohen et al. [35] leverage GPT-3 to back translate and rephrase as augmentations for hate speech detection. Del Arco et al. [40] explores using LLMs with prompting for zero-shot hate speech detection. Das et al. [39] evaluate LLMs' capabilities in multilingual and emoji-based hate speech. Wang et al. [227] further prompt LLMs to generate an explanation for hateful and non-hateful content and investigate the explanation generated by LLMs. Nguyen et al. [156] fine-tune Llama-7B to detect online sexual predatory chats and abusive language.

6.2.3 Misinformation Detection

LLMs can also be leveraged to detect misinformation, especially fake news [89]. Chen and Shu [21] discover that LLM-generated misinformation can be harder to detect compared to human-written, which suggests a more deceptive style and potentially causes more harm. Pavlyshenko [167] explores using a fine-tuned Llama-2 model for misinformation analysis and fake news detection. Yang et al. [251] assess ChatGPT's ability to rate the credibility of news outlets and find ChatGPT's prediction correlates to those from human experts. Pan et al. [164] establish a threat model and reveal that LLMs can act as effective misinformation generators, leading to significant degradation in open-domain question-answering systems. Leite et al. [114] develop a misinformation detection approach that combines the zero-shot LLM credibility signal labeling and weak supervision and achieve state-of-the-art without using ground truth label for training. Su et al. [203] discover that existing detectors are prone to flagging LLM-generated text as fake news and propose to leverage adversarial training for bias mitigation. Huang and Sun [82] explore ChatGPT's proficiency in generating, explaining, and detecting fake news, and propose using potential extra information that could boost LLM-based fake news detection.

6.2.4 LLM-Generated Text Detection

As LLMs emerge, misuse of LLMs increases, including disseminating fake news, plagiarism, manipulating public opinion Hanley and Durumeric [66], cheating, and fraud, making detection of LLM-generated content essential [21, 43, 144, 168, 210]. Researchers find that LLM-generated text can easily evade the plagiarism-checking software Khalil and Er [103] and is difficult to be identified by humans [223]. SpaceInfi [18] discovers that the extra space can serve an important role for LLM-generated content to evade detection.

GPT-Pat [266] develops a framework consisting of a Siamese network to compute the similarity between original and LLM-generated text and a binary classifier. Yang et al. [254] measure the ChatGPT's involvement in text generation based on edit distance. Orenstrakh et al. [160] investigate the LLM-generated text detection tool specifically in the education domain. CoCo [138] presents a coherence-based contrastive learning method to detect LLM-generated text under the low-resource scenario. Bhattacharjee and Liu [14] employ ChatGPT as a detector for AI-generated text detection.

7 Challenges and Directions

In this survey, we have comprehensively reviewed the recent approaches and applications of LLM-UM. Since the integration of LLMs into user modeling systems is still in the early stage, there are still some important and challenging yet unsolved problems in this direction. In this section, we discuss some challenges and potential future directions in this field.

7.1 Hallucination Mitigation

Though LLMs have shown strong capabilities in various domains, a significant challenge is the hallucination problem [88, 279], where LLMs generate seemingly plausible content but deviate from user input [4], previously generated content [137], and factuality knowledge [56, 149]. In the context of user modeling, the challenges of hallucination can be categorized into 1) *factual accuracy*: LLMs can sometimes generate content that sounds plausible but is either factually incorrect or not grounded in the input data. 2) *hallucination in high-stakes scenarios*: In contexts where recommendations have profound implications, such as in healthcare or legal advice, hallucinations can have severe consequences. 3) *user intent misunderstanding*: Hallucination can also arise from a misunderstanding of user intent and input context, which can significantly hinder the performance of LLMs in user modeling.

To mitigate the prevalent hallucination problems in LLMs, a promising approach to counter the hallucination problem in LLM-UM is to incorporate trustworthy symbolic knowledge in the LLMs' input, including unifying knowledge graphs (KGs) [163] and the use of external knowledge retrieval for augmented generation [264]. Manipulating the output side of LLMs can also help alleviate the hallucination problem. This could involve generating calibrated confidence scores to validate predictions and encouraging LLMs to express uncertainty. Research could also focus on designing decoding strategies that foster more reliable content generation, including factuality feedback mechanisms for the next token selection.

7.2 Privacy and Security

Privacy and security are paramount concerns in the user modeling system since it can involve processing and analyzing a wealth of UGC and user behavior history. However, recent research suggests that LLMs can pose privacy risks. For example, an adversary can recover training examples containing a person's name, email address, and phone number by querying the model [79]. The challenges in LLM-UM can be summarized as 1) *balancing personalization and privacy*: The success of LLMs in user modeling relies heavily on personalization, which necessitates the use of sensitive user data. However, this can lead to privacy and security concerns. That being said, striking the right balance between personalization and privacy is a significant challenge. 2) *data leakage risk*: As LLMs are trained on large quantities of data, there's a risk that they might unintentionally leak sensitive information during the text generation process. Recent research has shown that LLMs are vulnerable to adversarial attacks by prompt injection [128] and jailbreaking [183]. This could potentially leak users' identities or private information and cause security issues. 3) *adaptation to evolving privacy laws*: Privacy laws and regulations are continually evolving, and LLMs need to be updated to ensure compliance, which can be a challenging task given the huge parameters and high updating cost.

To mitigate the above challenges, future research can focus on designing privacy techniques in preprocessing, training, and post-processing steps to mitigate user privacy risks. They can study how to provide users with more transparency and study how their data are used and give them more control over their information.

7.3 Complex Structural Data Understanding

User interactions fundamentally form complex, heterogeneous, temporal text-rich graphs, encapsulating user-user and user-item interactions. Therefore, equipping LLMs with the capability to comprehend and interpret such graph structures could dramatically improve the performance of user modeling. Current challenges of leveraging LLMs to model complex graph structure can be summarized as 1) *lack of sequential transformation*: Unlike natural languages, graph-structured data often lack a straightforward transformation into sequential text [224], which makes it difficult for LLMs to process and understand. 2) *gap between graphs and human languages*: The significant gap between graph structures and natural language used for LLMs pretraining can hinder the application of LLMs to facilitate graph reasoning. 3) *graph variability*: Graphs can define structure and features in distinct ways and lack a unified paradigm, making it hard to generalize well to the diverse structure and feature representations of different graphs [96, 225]. 4) *heterogeneous temporal text-rich graph structure modeling*: User interactions can be formulated into heterogeneous temporal text-rich graphs, posing challenges for LLMs to understand such complex data.

To empower LLMs with the capabilities to understand complex structural data, future research could focus on 1) *graph language generation*: One potential research direction could be to derive a language for graph structures that LLMs can better understand, such as transforming graph data into a form that LLMs can process, such as sequential text. 2) *temporal heterogeneous graph modeling*: Given that user interactions often form heterogeneous temporal graphs, the research could focus on techniques to handle such complex structures, which could involve developing methods to capture the dynamics and heterogeneity of user interactions.

7.4 Comprehensive Benchmark and Evaluation Criteria

Evaluation is paramount to the success of LLM-UM, which can help us better understand the strengths and weaknesses, ensuring safety and reliability, and inspiring future research directions of user modeling systems. Current challenges of LLM-UM evaluations lie in 1) *lack of AGI-level benchmark*: Current evaluation benchmarks mainly focus on restricting models input and output to a static pair based on human annotation, which cannot measure the superhuman capabilities in LLMs, instead, the existing benchmarks can be seen in LLMs pertaining stage and cause test data leakage problem. Therefore, designing more challenging tasks and more reliable evaluation mechanisms by utilizing the cross-disciplinary knowledge from education, social science, and more could be a better solution to probe LLM-UM's capabilities as a superhuman-level system. 2) *comprehensive intermediate evaluation*: A general concern in LLM-UM is the evaluation only focuses on the prediction performance on specific datasets. Although this is likely the most important metric, the intermediate steps such as user profile generation, LLM-generated factual knowledge and datasets, and pipeline control behavior are also important for researchers to more comprehensively understand the methods. 3) *dynamic and evolving evaluation*: Existing UM evaluation protocols rely on static and public datasets, which would lead to overfitting and could not assess the LLM-UM in evolving circumstances, especially in the context of fast-evolving user data. For example, AgentBench [140] is a pioneering work that designs evolving benchmarks and evaluates LLMs as agents, which is shown to be the better way of static input-output pair evaluation to evaluate LLM-based systems. Therefore, developing a dynamic and evolving testbed for LLM-UM to probe the system's capabilities in adapting the evolving data is an important research direction.

7.5 Trustworthy User Modeling

LLM-UM can bring significant benefits to humans, including providing personalized advice, recommendations, assistance, and more. However, the black-box property and unreliable generation process of LLMs would pose a serious threat to users when the LLM-UM system is applied to high-stakes domains such as healthcare and finance. Therefore, how to enable trustworthy LLM-UM is of significant importance. Current challenges and potential future research directions include 1) *transparent and explainable reasoning*: Users need to understand how an LLM-UM system makes decisions to enhance trust. Therefore, promising research directions are how to explain the black-box LLMs and how to make the user modeling process transparent and explainable. 2) *reliability*: LLM-UM should provide consistent and accurate results, be robust to different input templates, and give consistent generation given the same or similar input to be trustworthy. 3) *user control*: Trustworthy LLM-UM systems should allow users to control how their data is used and how decisions are made, which could involve developing mechanisms for users to feedback, correct, and update the LLM-UM system behavior.

7.6 Fairness in LLM-UM

Fairness is a crucial aspect of user modeling systems, which secure unbiased user characteristics and profile generation and give fair assistance and recommendations for different groups of people. LLMs trained on extensive datasets can inadvertently learn and perpetuate human biases and stereotypes in training data, leading to unfair treatment or discriminatory predictions of specific user groups [57]. Moreover, balancing personalization and fairness is a significant problem, especially when user data reflects societal biases. Therefore, future LLM-UM research may include 1) *bias detection and measurement*: Detect bias that exists in LLM-UM systems training data, input, model, and task design, and develop fine-grained mechanisms to measure the bias from different aspects (e.g., political bias, ethnic groups bias, etc.) in LLM-UM systems. 2) *promote fairness in LLM-UM systems*: Mitigate bias and give non-discriminative predictions, including making LLM-UM more explainable and transparent to help better understand the source of unfair behavior and devise appropriate interventions. 3) *ethics and regulations*: Research should also focus on the ethical implications of LLMs and the potential for regulations to ensure fairness, which could involve interdisciplinary work, bringing together technologists, ethicists, sociologists, and lawmakers.

7.7 Efficient and Effective Domain Adaptation

LLMs are pretrained on a large general-purpose corpus and then plugged into user modeling systems. A lot of works use LLMs as a general-purpose reasoner with frozen pertaining parameters, while others fine-tune LLMs to help them adapt to the specific task in user modeling. The latter paradigm allows the model to leverage the general language understanding capabilities learned during the pretraining stage while specializing the knowledge to the task at hand and shows superior task performance. However, fine-tuning and deployment can be highly computationally expensive, especially when deploying on millions and even billions of scale user data. Therefore, developing efficient domain adaptation methods to reduce the cost for LLMs to fix in user modeling systems and efficiently help LLM-UM be personalized and adapt to the user's dynamic needs. Moreover, user-specific data is often sparse and noisy [38, 107], which can lead to overfitting during the fine-tuning process, which can limit the model's ability to generalize to different users and contexts. Therefore, future research to combat these challenges can focus on 1) *efficient fine-tuning for LLM-UM*: There is a need to develop and refine techniques for more efficient fine-tuning of LLMs specific for user modeling context. Techniques such as sparsity induction, model pruning, and knowledge distillation could be explored to reduce the computational and memory footprints. 2) *online learning*: To cope with the dynamics of user interests, research could explore online learning strategies for LLM-UM systems. These strategies would continuously update the model based on the latest user-generated content and interactions.

7.8 Personalized LLM-UM Deployment

Recent research on personalized LLMs mainly focuses on two aspects: personalized prompt engineering and personalized preference alignment. For personalized prompt engineering works, researchers focus on designing prompts to help LLMs understand user preferences based on user behavior. As previously discussed, the personalized prompting paradigm would directly encode user behavior history, apply retrieval methods to obtain the most relevant user history, and leverage LLMs to summarize user preference to augment retrieval and generation performance. At the same time, the other line of work focuses on aligning LLMs with personal preferences using reinforcement learning. Personalized Soups [85] is a representative work that decomposes user preference into multiple dimensions and models the personalization into a multi-object reinforcement learning problem. However, the existing personalized prompting model exhibits limited generalization and context windows for encoding and comprehending user preferences from extensive user behavioral history. Furthermore, human alignment methods require explicit human preferences and pre-established preference dimensions, and the alignment finetuning process is computationally expensive. Therefore, future personalized LLM research should focus on 1) *parametric user preference mining*: As previous studies [141] have indicated, differential parameters are more effective in understanding and modeling user preferences compared to discrete natural language. Additionally, parametric personalization can infuse personalized knowledge more efficiently by using fewer tokens, thereby conserving the context length for LLMs. 2) *cross-model personalized knowledge transfer*: Personalized LLMs strive to diverge from the prevailing 'one-size-fits-all' paradigm, aiming to offer unique services tailored to each individual. However, given the 'scaling law' - which posits that larger models yield superior performance - not all users possess the computational resources to generate high-quality output by operating their own large models on local machines. Therefore, future research should work towards developing small models that can learn from user history and gain personalized knowledge on a local machine. These models should be able to transfer or plug their knowledge into a remote, large model to produce high-quality personalized output. 3) *parametric privacy preservation*: Recent findings suggest that encoded text embeddings alone can be used to reconstruct full text [151], posing potential privacy and safety risks. This becomes particularly relevant in personalization, where user preferences can be encoded using a small model on local machines. There exists a risk of private data reconstruction using latent embedding and parameters. Therefore, developing strategies to thwart such reconstruction while preserving latent embedding performance is a pressing issue.

8 Conclusion

Our work presents a comprehensive and structured survey of large language models for user modeling (LLM-UM). We show why LLMs are great tools for user modeling and understanding user-generated content (UGC) and user interactions. We then review the existing LLM-UM research works and categorize their approaches that integrate text-based and graph-based UM techniques, including LLMs serving as enhancers, predictors, controllers, and evaluators. Next, we categorize the existing LLM-UM techniques based on their applications. Finally, we outline the remaining challenges and future directions in the LLM-UM field. This survey serves as a handbook for LLM-UM researchers and practitioners to study and use LLMs to augment user modeling systems, and inspires additional interest and work on this topic.

References

[1] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on twitter for personalized news recommendations. In User Modeling, Adaption and Personalization: 19th International Conference, UMAP 2011, Girona, Spain, July 11-15, 2011. Proceedings 19, pages 1–12. Springer, 2011.

[2] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Twitter-based user modeling for news recommendations. In Twenty-Third International Joint Conference on Artificial Intelligence. Citeseer, 2013.

[3] A. Acharya, B. Singh, and N. Onoe. Llm based generation of item-description for recommendation system. In Proceedings of the 17th ACM Conference on Recommender Systems, pages 1204–1207, 2023.

[4] V. Adlakha, P. BehnamGhader, X. H. Lu, N. Meade, and S. Reddy. Evaluating correctness and faithfulness of instruction-following models for question answering. arXiv preprint arXiv:2307.16877, 2023.

[5] R. Alghamdi and K. Alfalqi. A survey of topic modeling in text mining. International Journal of Advanced Computer Science and Applications, 6(1), 2015.

[6] N. Ayoobi, S. Shahriar, and A. Mukherjee. The looming threat of fake and llm-generated linkedin profiles: Challenges and opportunities for detection and prevention. In Proceedings of the 34th ACM Conference on Hypertext and Social Media, pages 1–10, 2023.

[7] Y. Bang, T. Yu, A. Madotto, Z. Lin, M. Diab, and P. Fung. Enabling classifiers to make judgements explicitly aligned with human values. arXiv preprint arXiv:2210.07652, 2022.

[8] K. Bao, J. Zhang, Y. Zhang, W. Wang, F. Feng, and X. He. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. arXiv preprint arXiv:2305.00447, 2023.

[9] Q. Bao, J. Leinonen, A. Y. Peng, W. Zhong, T. Pistotti, A. Huang, P. Denny, M. Witbrock, and J. Liu. Exploring self-reinforcement for improving learnersourced multiple-choice question explanations with large language models. arXiv preprint arXiv:2309.10444, 2023.

[10] A. Belyaeva, J. Cosentino, F. Hormozdiari, C. Y. McLean, and N. A. Furlotte. Multimodal llms for health grounded in individual-specific data. arXiv preprint arXiv:2307.09018, 2023.

[11] S. Berkovsky, T. Kuflik, and F. Ricci. Entertainment personalization mechanism through cross-domain user modeling. In International Conference on Intelligent Technologies for Interactive Entertainment, pages 215–219. Springer, 2005.

[12] M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, L. Gianinazzi, J. Gajda, T. Lehmann, M. Podstawski, H. Niewiadomski, P. Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. arXiv preprint arXiv:2308.09687, 2023.

[13] S. Bhat, H. A. Nguyen, S. Moore, J. Stamper, M. Sakr, and E. Nyberg. Towards automated generation and evaluation of questions in educational domains. In Proceedings of the 15th International Conference on Educational Data Mining, volume 701, 2022.

[14] A. Bhattacharjee and H. Liu. Fighting fire with fire: Can chatgpt detect ai-generated text? arXiv preprint arXiv:2308.01284, 2023.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan): 993–1022, 2003.

[16] V. Borisov, K. Seßler, T. Leemann, M. Pawelczyk, and G. Kasneci. Language models are realistic tabular data generators. arXiv preprint arXiv:2210.06280, 2022.

[17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell,

et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[18] S. Cai and W. Cui. Evade chatgpt detectors via a single space. arXiv preprint arXiv:2307.02599, 2023.

[19] L. Cao. Diaggpt: An llm-based chatbot with automatic topic management for task-oriented dialogue. arXiv preprint arXiv:2308.08043, 2023.

[20] T. Chakrabarty, V. Padmakumar, F. Brahman, and S. Muresan. Creativity support in the age of large language models: An empirical study involving emerging writers. arXiv preprint arXiv:2309.12570, 2023.

[21] C. Chen and K. Shu. Can llm-generated misinformation be detected? arXiv preprint arXiv:2309.13788, 2023.

[22] J. Chen, Z. Liu, X. Huang, C. Wu, Q. Liu, G. Jiang, Y. Pu, Y. Lei, X. Chen, X. Wang, et al. When large language models meet personalization: Perspectives of challenges and opportunities. arXiv preprint arXiv:2307.16376, 2023.

[23] S. Chen, M. Wu, K. Q. Zhu, K. Lan, Z. Zhang, and L. Cui. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. arXiv preprint arXiv:2305.13614, 2023.

[24] W. Chen, Y. Gu, Z. Ren, X. He, H. Xie, T. Guo, D. Yin, and Y. Zhang. Semi-supervised user profiling with heterogeneous graph attention networks. In IJCAI, volume 19, pages 2116–2122, 2019.

[25] Y. Chen, R. Li, Z. Zhao, C. Peng, J. Wu, E. Hossain, and H. Zhang. Netgpt: A native-ai network architecture beyond provisioning personalized generative services. arXiv preprint arXiv:2307.06148, 2023.

[26] Z. Chen. Palr: Personalization aware llms for recommendation. arXiv preprint arXiv:2305.07622, 2023.

[27] Z. Chen, H. Mao, H. Li, W. Jin, H. Wen, X. Wei, S. Wang, D. Yin, W. Fan, H. Liu, et al. Exploring the potential of large language models (llms) in learning on graphs. arXiv preprint arXiv:2307.03393, 2023.

[28] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023), 2023.

[29] K.-L. Chiu, A. Collins, and R. Alexander. Detecting hate speech with gpt-3. arXiv preprint arXiv:2103.12407, 2021.

[30] I. Cho, D. Wang, R. Takahashi, and H. Saito. A personalized dialogue generator with implicit user persona detection. arXiv preprint arXiv:2204.07372, 2022.

[31] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311, 2022.

[32] K. Christakopoulou, A. Lalama, C. Adams, I. Qu, Y. Amir, S. Chucri, P. Vollucci, F. Soldo, D. Bseiso, S. Scodel, et al. Large language models for user interest journeys. arXiv preprint arXiv:2305.15498, 2023.

[33] I. A. Christensen and S. Schiaffino. Entertainment recommender systems for group of users. Expert systems with applications, 38(11):14127–14135, 2011.

[34] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416, 2022.

[35] S. Cohen, D. Presil, O. Katz, O. Arbili, S. Messica, and L. Rokach. Enhancing social network hate detection using back translation and gpt-3 augmentations during training and test-time. Information Fusion, page 101887, 2023.

[36] S. Dai, N. Shao, H. Zhao, W. Yu, Z. Si, C. Xu, Z. Sun, X. Zhang, and J. Xu. Uncovering chatgpt's capabilities in recommender systems. arXiv preprint arXiv:2305.02182, 2023.

[37] Y. Dan, Z. Lei, Y. Gu, Y. Li, J. Yin, J. Lin, L. Ye, Z. Tie, Y. Zhou, Y. Wang, et al. Educhat: A large-scale language model-based chatbot system for intelligent education. arXiv preprint arXiv:2308.02773, 2023.

[38] A. Das, M. Degeling, D. Smullen, and N. Sadeh. Personalized privacy assistants for the internet of things: Providing users with notice and choice. IEEE Pervasive Computing, 17(3):35–46, 2018.

[39] M. Das, S. K. Pandey, and A. Mukherjee. Evaluating chatgpt's performance for multilingual and emoji-based hate speech detection. arXiv preprint arXiv:2305.13276, 2023.

[40] F. M. P. Del Arco, D. Nozza, and D. Hovy. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In The 7th Workshop on Online Abuse and Harms (WOAH), pages 60–68, 2023.

[41] X. Deng, V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky. What do llms know about financial markets? a case study on reddit market sentiment analysis. In Companion Proceedings of the ACM Web Conference 2023, pages 107–110, 2023.

[42] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms. arXiv

preprint arXiv:2305.14314, 2023.

[43] M. Dhaini, W. Poelman, and E. Erdogan. Detecting chatgpt: A survey of the state of detecting chatgpt-generated text. arXiv preprint arXiv:2309.07689, 2023.

[44] D. Di Palma, G. M. Biancofiore, V. W. Anelli, F. Narducci, T. Di Noia, and E. Di Sciascio. Evaluating chatgpt as a recommender system: A rigorous approach. arXiv preprint arXiv:2309.03613, 2023.

[45] Y. Dong, N. V. Chawla, and A. Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pages 135–144, 2017.

[46] Y. Du, D. Luo, R. Yan, H. Liu, Y. Song, H. Zhu, and J. Zhang. Enhancing job recommendation through llm-based generative adversarial networks. arXiv preprint arXiv:2307.10747, 2023.

[47] S. T. Dumais. Latent semantic analysis. Annual Review of Information Science and Technology (ARIST), 38: 189–230, 2004.

[48] S. Elkins, E. Kochmar, I. Serban, and J. C. Cheung. How useful are educational questions generated by large language models? In International Conference on Artificial Intelligence in Education, pages 536–542. Springer, 2023.

[49] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin. Graph neural networks for social recommendation. In The world wide web conference, pages 417–426, 2019.

[50] W. Fan, Z. Zhao, J. Li, Y. Liu, X. Mei, Y. Wang, J. Tang, and Q. Li. Recommender systems in the era of large language models (llms). arXiv preprint arXiv:2307.02046, 2023.

[51] Y. Fan and F. Jiang. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study. arXiv preprint arXiv:2305.08391, 2023.

[52] Y. Fang, X. Li, S. W. Thomas, and X. Zhu. Chatgpt as data augmentation for compositional generalization: A case study in open intent detection. arXiv preprint arXiv:2308.13517, 2023.

[53] M. Farid, R. Elgohary, I. Moawad, and M. Roushdy. User profiling approaches, modeling, and personalization. In Proceedings of the 11th International Conference on Informatics & Systems (INFOS 2018), 2018.

[54] S. Feng, Z. Tan, R. Li, and M. Luo. Heterogeneity-aware twitter bot detection with relational graph transformers. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 3977–3985, 2022.

[55] S. Feng, Z. Tan, H. Wan, N. Wang, Z. Chen, B. Zhang, Q. Zheng, W. Zhang, Z. Lei, S. Yang, et al. Twibot-22: Towards graph-based twitter bot detection. Advances in Neural Information Processing Systems, 35:35254–35269, 2022.

[56] S. Feng, V. Balachandran, Y. Bai, and Y. Tsvetkov. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. arXiv preprint arXiv:2305.08281, 2023.

[57] S. Feng, C. Y. Park, Y. Liu, and Y. Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. arXiv preprint arXiv:2305.08283, 2023.

[58] E. Ferrara. Social bot detection in the age of chatgpt: Challenges and opportunities. First Monday, 2023.

[59] S. E. Finch, E. S. Paek, and J. D. Choi. Leveraging large language models for automated dialogue analysis. arXiv preprint arXiv:2309.06490, 2023.

[60] M. Foosherian, H. Purwins, P. Rathnayake, T. Alam, R. Teimao, and K.-D. Thoben. Enhancing pipeline-based conversational agents with large language models. arXiv preprint arXiv:2309.03748, 2023.

[61] L. Friedman, S. Ahuja, D. Allen, T. Tan, H. Sidahmed, C. Long, J. Xie, G. Schubiner, A. Patel, H. Lara, et al. Leveraging large language models in conversational recommender systems. arXiv preprint arXiv:2305.07961, 2023.

[62] G. Fu, Q. Zhao, J. Li, D. Luo, C. Song, W. Zhai, S. Liu, F. Wang, Y. Wang, L. Cheng, et al. Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals. arXiv preprint arXiv:2308.15192, 2023.

[63] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, and J. Zhang. Chat-rec: Towards interactive and explainable llms-augmented recommender system. arXiv preprint arXiv:2303.14524, 2023.

[64] H. Ghanadian, I. Nejadgholi, and H. A. Osman. Chatgpt for suicide risk assessment on social media: Quantitative evaluation of model performance, potentials and limitations. arXiv preprint arXiv:2306.09390, 2023.

[65] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pages 855–864, 2016.

[66] H. W. Hanley and Z. Durumeric. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. arXiv preprint arXiv:2305.09820, 2023.

[67] Z. S. Harris. Distributional structure. Word, 10(2-3):146–162, 1954.

[68] M. M. Hassan, A. Knipper, and S. K. K. Santu. Chatgpt as your personal data scientist. arXiv preprint arXiv:2305.13657, 2023.

[69] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, pages 639–648, 2020.

[70] X. He, X. Bresson, T. Laurent, and B. Hooi. Explanations as features: Llm-based features for text-attributed graphs. arXiv preprint arXiv:2305.19523, 2023.

[71] Z. He, W. Liu, W. Guo, J. Qin, Y. Zhang, Y. Hu, and R. Tang. A survey on user behavior modeling in recommender systems. arXiv preprint arXiv:2302.11087, 2023.

[72] Z. He, Z. Xie, R. Jha, H. Steck, D. Liang, Y. Feng, B. P. Majumder, N. Kallus, and J. McAuley. Large language models as zero-shot conversational recommenders. arXiv preprint arXiv:2308.10053, 2023.

[73] Y. Hou, J. Zhang, Z. Lin, H. Lu, R. Xie, J. McAuley, and W. X. Zhao. Large language models are zero-shot rankers for recommender systems. arXiv preprint arXiv:2305.08845, 2023.

[74] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2021.

[75] J. Hu, F. Jin, G. Zhang, J. Wang, and Y. Yang. A user profile modeling method based on word2vec. In 2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C), pages 410–414. IEEE, 2017.

[76] K. Hu, M. Yan, J. T. Zhou, I. W. Tsang, W. H. Chong, and Y. K. Yap. Ladder-of-thought: Using knowledge as steps to elevate stance detection. arXiv preprint arXiv:2308.16763, 2023.

[77] Z. Hu, Y. Dong, K. Wang, and Y. Sun. Heterogeneous graph transformer. In Proceedings of the web conference 2020, pages 2704–2710, 2020.

[78] Z. Hu, Y. Feng, A. T. Luu, B. Hooi, and A. Lipani. Unlocking the potential of user feedback: Leveraging large language model as user simulator to enhance dialogue system. arXiv preprint arXiv:2306.09821, 2023.

[79] J. Huang, H. Shao, and K. C.-C. Chang. Are large pre-trained language models leaking your personal information? In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2038–2047, 2022.

[80] X. Huang, J. Lian, Y. Lei, J. Yao, D. Lian, and X. Xie. Recommender ai agent: Integrating large language models for interactive recommendations. arXiv preprint arXiv:2308.16505, 2023.

[81] X. Huang, W. Ruan, W. Huang, G. Jin, Y. Dong, C. Wu, S. Bensalem, R. Mu, Y. Qi, X. Zhao, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. arXiv preprint arXiv:2305.11391, 2023.

[82] Y. Huang and L. Sun. Harnessing the power of chatgpt in fake news: An in-depth exploration in generation, detection and explanation. arXiv preprint arXiv:2310.05046, 2023.

[83] V. Hudeček and O. Dušek. Are llms all you need for task-oriented dialogue? arXiv preprint arXiv:2304.06556, 2023.

[84] J. Huynh, C. Jiao, P. Gupta, S. Mehri, P. Bajaj, V. Chaudhary, and M. Eskenazi. Understanding the effectiveness of very large language models on dialog evaluation. arXiv preprint arXiv:2301.12004, 2023.

[85] J. Jang, S. Kim, B. Y. Lin, Y. Wang, J. Hessel, L. Zettlemoyer, H. Hajishirzi, Y. Choi, and P. Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. arXiv preprint arXiv:2310.11564, 2023.

[86] J. Ji, Z. Li, S. Xu, W. Hua, Y. Ge, J. Tan, and Y. Zhang. Genrec: Large language model for generative recommendation. arXiv e-prints, pages arXiv–2307, 2023.

[87] Y. Ji, W. Wu, H. Zheng, Y. Hu, X. Chen, and L. He. Is chatgpt a good personality recognizer? a preliminary study. arXiv preprint arXiv:2307.03952, 2023.

[88] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12):1–38, 2023.

[89] B. Jiang, Z. Tan, A. Nirmal, and H. Liu. Disinformation detection: An evolving challenge in the age of llms. arXiv

preprint arXiv:2309.15847, 2023.

[90] H. Jiang, X. Zhang, X. Cao, J. Kabbara, and D. Roy. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. arXiv preprint arXiv:2305.02547, 2023.

[91] M. Jiang, P. Cui, R. Liu, Q. Yang, F. Wang, W. Zhu, and S. Yang. Social contextual recommendation. In Proceedings of the 21st ACM international conference on Information and knowledge management, pages 45–54, 2012.

[92] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang. Catchsync: catching synchronized behavior in large directed graphs. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 941–950, 2014.

[93] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang. Detecting suspicious following behavior in multimillion-node social networks. In Proceedings of the 23rd International Conference on World Wide Web, pages 305–306, 2014.

[94] M. Jiang, P. Cui, F. Wang, W. Zhu, and S. Yang. Scalable recommendation with social contextual information. IEEE Transactions on Knowledge and Data Engineering, 26(11):2789–2802, 2014.

[95] M. Jiang, P. Cui, and C. Faloutsos. Suspicious behavior detection: Current trends and future directions. IEEE intelligent systems, 31(1):31–39, 2016.

[96] M. Jiang, C. Faloutsos, and J. Han. Catchtartan: Representing and summarizing dynamic multicontextual behaviors. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 945–954, 2016.

[97] Y. Jiang, R. Qiu, Y. Zhang, and P.-F. Zhang. Balanced and explainable social media analysis for public health with large language models. arXiv preprint arXiv:2309.05951, 2023.

[98] I. Joshi, R. Budhiraja, P. D. Tanna, L. Jain, M. Deshpande, A. Srivastava, S. Rallapalli, H. D. Akolekar, J. S. Challa, and D. Kumar. From" let's google" to" let's chatgpt": Student and instructor perspectives on the influence of llms on undergraduate engineering education. arXiv preprint arXiv:2309.10694, 2023.

[99] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy. Challenges and applications of large language models. arXiv preprint arXiv:2307.10169, 2023.

[100] W.-C. Kang, J. Ni, N. Mehta, M. Sathiamoorthy, L. Hong, E. Chi, and D. Z. Cheng. Do llms understand user preferences? evaluating llms on user rating prediction. arXiv preprint arXiv:2305.06474, 2023.

[101] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

[102] J. D. M.-W. C. Kenton and L. K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pages 4171–4186, 2019.

[103] M. Khalil and E. Er. Will chatgpt get you caught? rethinking of plagiarism detection. arXiv preprint arXiv:2302.04335, 2023.

[104] K. Kheiri and H. Karimi. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. arXiv preprint arXiv:2307.10234, 2023.

[105] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations, 2016.

[106] C. Kong, Y. Fan, X. Wan, F. Jiang, and B. Wang. Large language model as a user simulator. arXiv preprint arXiv:2308.11534, 2023.

[107] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. Computer, 42(8): 30–37, 2009.

[108] M. Kosinski. Theory of mind may have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02083, 2023.

[109] C. Koyuturk, M. Yavari, E. Theophilou, S. Bursic, G. Donabauer, A. Telari, A. Testa, R. Boiano, A. Gabbiadini, D. Hernandez-Leo, et al. Developing effective educational chatbots with chatgpt prompts: Insights from preliminary tests in a case study on social media literacy. arXiv preprint arXiv:2306.10645, 2023.

[110] M. Labonne and S. Moran. Spam-t5: Benchmarking large language models for few-shot email spam detection. arXiv preprint arXiv:2304.01238, 2023.

[111] T. Lai, Y. Shi, Z. Du, J. Wu, K. Fu, Y. Dou, and Z. Wang. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. arXiv preprint arXiv:2307.11991, 2023.

[112] K. Lakkaraju, S. K. R. Vuruma, V. Pallagani, B. Muppasani, and B. Srivastava. Can llms be good financial advisors?: An initial study in personal decision making for optimized outcomes. arXiv preprint arXiv:2307.07422, 2023.

[113] D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. Advances in neural information processing systems, 13, 2000.

[114] J. A. Leite, O. Razuvayevskaya, K. Bontcheva, and C. Scarton. Detecting misinformation with llm-predicted credibility signals and weak supervision. arXiv preprint arXiv:2309.07601, 2023.

[115] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3045–3059, 2021.

[116] Y. Levine, N. Wies, D. Jannai, D. Navon, Y. Hoshen, and A. Shashua. The inductive bias of in-context learning: Rethinking pretraining example design. arXiv preprint arXiv:2110.04541, 2021.

[117] C. Li, M. Zhang, Q. Mei, Y. Wang, S. A. Hombaiah, Y. Liang, and M. Bendersky. Teach llms to personalize–an approach inspired by writing education. arXiv preprint arXiv:2308.07968, 2023.

[118] H. Li, Y. Wu, V. Schlegel, R. Batista-Navarro, T.-T. Nguyen, A. R. Kashyap, X. Zeng, D. Beck, S. Winkler, and G. Nenadic. Pulsar: Pre-training with extracted healthcare terms for summarising patients' problems and data augmentation with black-box large language models. arXiv preprint arXiv:2306.02754, 2023.

[119] J. Li, T. Tang, W. X. Zhao, and J.-R. Wen. Pretrained language models for text generation: A survey.

[120] J. Li, W. Zhang, T. Wang, G. Xiong, A. Lu, and G. Medioni. Gpt4rec: A generative framework for personalized recommendation and user interests interpretation. arXiv preprint arXiv:2304.03879, 2023.

[121] N. Li, B. Kang, and T. De Bie. Llm4jobs: Unsupervised occupation extraction and standardization leveraging large language models. arXiv preprint arXiv:2309.09708, 2023.

[122] P. Li, Y. Wang, E. H. Chi, and M. Chen. Prompt tuning large language models on personalized aspect extraction for recommendations. arXiv preprint arXiv:2306.01475, 2023.

[123] R. Li, W. Deng, Y. Cheng, Z. Yuan, J. Zhang, and F. Yuan. Exploring the upper limits of text-based collaborative filtering using large language models: Discoveries and insights. arXiv preprint arXiv:2305.11700, 2023.

[124] S. Li and H. Zhao. A survey on representation learning for user modeling. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pages 4997–5003, 2021.

[125] X. Li, Y. Zhang, and E. C. Malthouse. A preliminary study of chatgpt on news recommendation: Personalization, provider fairness, fake news. arXiv preprint arXiv:2306.10702, 2023.

[126] X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190, 2021.

[127] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. Cureus, 15(6), 2023.

[128] Z. Li, B. Peng, P. He, and X. Yan. Do you really follow me? adversarial instructions for evaluating the robustness of large language models. arXiv preprint arXiv:2308.10819, 2023.

[129] G. Lin and Y. Zhang. Sparks of artificial general recommender (agr): Early experiments with chatgpt. arXiv preprint arXiv:2305.04518, 2023.

[130] J. Lin, X. Dai, Y. Xi, W. Liu, B. Chen, X. Li, C. Zhu, H. Guo, Y. Yu, R. Tang, et al. How can recommender systems benefit from large language models: A survey. arXiv preprint arXiv:2306.05817, 2023.

[131] J. Lin, R. Shan, C. Zhu, K. Du, B. Chen, S. Quan, R. Tang, Y. Yu, and W. Zhang. Rella: Retrieval-enhanced large language models for lifelong sequential behavior comprehension in recommendation. arXiv preprint arXiv:2308.11131, 2023.

[132] Y.-T. Lin and Y.-N. Chen. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. arXiv preprint arXiv:2305.13711, 2023.

[133] J. Liu, C. Liu, R. Lv, K. Zhou, and Y. Zhang. Is chatgpt a good recommender? a preliminary study. arXiv preprint arXiv:2304.10149, 2023.

[134] J. Liu, C. Liu, P. Zhou, Q. Ye, D. Chong, K. Zhou, Y. Xie, Y. Cao, S. Wang, C. You, et al. Llmrec: Benchmarking large language models on recommendation task. arXiv preprint arXiv:2308.12241, 2023.

[135] Q. Liu, N. Chen, T. Sakai, and X.-M. Wu. A first look at llm-powered generative news recommendation. arXiv preprint arXiv:2305.06566, 2023.

[136] Q. Liu, N. Chen, T. Sakai, and X.-M. Wu. Once: Boosting content-based recommendation with both open- and closed-source large language models, 2023.

[137] T. Liu, Y. Zhang, C. Brockett, Y. Mao, Z. Sui, W. Chen, and B. Dolan. A token-level reference-free hallucination detection benchmark for free-form text generation. arXiv preprint arXiv:2104.08704, 2021.

[138] X. Liu, Z. Zhang, Y. Wang, Y. Lan, and C. Shen. Coco: Coherence-enhanced machine-generated text detection under data limitation with contrastive learning. arXiv preprint arXiv:2212.10341, 2022.

[139] X. Liu, D. McDuff, G. Kovacs, I. Galatzer-Levy, J. Sunshine, J. Zhan, M.-Z. Poh, S. Liao, P. Di Achille, and S. Patel. Large language models are few-shot health learners. arXiv preprint arXiv:2305.15525, 2023.

[140] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang, et al. Agentbench: Evaluating llms as agents. arXiv preprint arXiv:2308.03688, 2023.

[141] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang. Gpt understands, too. AI Open, 2023.

[142] X.-Y. Liu, G. Wang, and D. Zha. Fingpt: Democratizing internet-scale data for financial large language models. arXiv preprint arXiv:2307.10485, 2023.

[143] Z. Liu, Z. Wu, M. Hu, B. Zhao, L. Zhao, T. Zhang, H. Dai, X. Chen, Y. Shen, S. Li, et al. Pharmacygpt: The ai pharmacist. arXiv preprint arXiv:2307.10432, 2023.

[144] N. Lu, S. Liu, R. He, and K. Tang. Large language models can be guided to evade ai-generated text detection. arXiv preprint arXiv:2305.10847, 2023.

[145] H. Lyu, S. Jiang, H. Zeng, Y. Xia, and J. Luo. Llm-rec: Personalized recommendation via prompting large language models. arXiv preprint arXiv:2307.15780, 2023.

[146] J. Manyika. An overview of bard: an early experiment with generative ai. Technical report, Tech. rep., Technical report, Google AI, 2023.

[147] S. Matwin, A. Milios, P. Prałat, A. Soares, and F. Théberge. Survey of generative methods for social media analysis. arXiv preprint arXiv:2112.07041, 2021.

[148] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[149] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. arXiv preprint arXiv:2305.14251, 2023.

[150] L. Mo, S. Chen, Z. Chen, X. Deng, A. Lewis, S. Singh, S. Stevens, C.-Y. Tai, Z. Wang, X. Yue, et al. Roll up your sleeves: Working with a collaborative and engaging task-oriented dialogue system. arXiv preprint arXiv:2307.16081, 2023.

[151] J. X. Morris, V. Kuleshov, V. Shmatikov, and A. M. Rush. Text embeddings reveal (almost) as much as text. arXiv preprint arXiv:2310.06816, 2023.

[152] M. Mozes, X. He, B. Kleinberg, and L. D. Griffin. Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities. arXiv preprint arXiv:2308.12833, 2023.

[153] Y. Mu, B. P. Wu, W. Thorne, A. Robinson, N. Aletras, C. Scarton, K. Bontcheva, and X. Song. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. arXiv preprint arXiv:2305.14310, 2023.

[154] S. Mysore, A. McCallum, and H. Zamani. Large language model augmented narrative driven recommendations. arXiv preprint arXiv:2306.02250, 2023.

[155] S. Natkin and C. Yan. User model in multiplayer mixed reality entertainment applications. In Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology, pages 74–es, 2006.

[156] T. T. Nguyen, C. Wilson, and J. Dalins. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. arXiv preprint arXiv:2308.14683, 2023.

[157] P. Ochieng. Are large language models fit for guided reading? arXiv preprint arXiv:2305.10645, 2023.

[158] A. Olga, A. Saini, G. Zapata, D. Searsmith, B. Cope, M. Kalantzis, V. Castro, T. Kourkoulou, J. Jones, R. A. da Silva, et al. Generative ai: Implications and applications for education. arXiv preprint arXiv:2305.07605, 2023.

[159] OpenAI. Gpt-4 technical report, 2023.

[160] M. S. Orenstrakh, O. Karnalim, C. A. Suarez, and M. Liut. Detecting llm-generated text in computing education: A

comparative study for chatgpt cases. arXiv preprint arXiv:2307.07411, 2023.

[161] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744, 2022.

[162] L. PAGE. The pagerank citation ranking: Bringing order to the web. In Proc. of the 7ˆ< th> WWW Conf., 1998, 1998.

[163] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu. Unifying large language models and knowledge graphs: A roadmap. arXiv preprint arXiv:2306.08302, 2023.

[164] Y. Pan, L. Pan, W. Chen, P. Nakov, M.-Y. Kan, and W. Y. Wang. On the risk of misinformation pollution with large language models. arXiv preprint arXiv:2305.13661, 2023.

[165] S. Parikh, Q. Vohra, P. Tumbade, and M. Tiwari. Exploring zero and few-shot techniques for intent classification. arXiv preprint arXiv:2305.07157, 2023.

[166] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. arXiv preprint arXiv:2304.03442, 2023.

[167] B. M. Pavlyshenko. Analysis of disinformation and fake news detection using fine-tuned large language model. arXiv preprint arXiv:2309.04704, 2023.

[168] A. Pegoraro, K. Kumari, H. Fereidooni, and A.-R. Sadeghi. To chatgpt, or not to chatgpt: That is the question! arXiv preprint arXiv:2304.01487, 2023.

[169] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, H. Cao, X. Cheng, M. Chung, M. Grella, K. K. GV, et al. Rwkv: Reinventing rnns for the transformer era. arXiv preprint arXiv:2305.13048, 2023.

[170] L. Peng, Y. Zhang, and J. Shang. Generating efficient training data via llm-based attribute manipulation. arXiv preprint arXiv:2307.07099, 2023.

[171] W. Peng, D. Xu, T. Xu, J. Zhang, and E. Chen. Are gpt embeddings useful for ads and recommendation? In International Conference on Knowledge Science, Engineering and Management, pages 151–162. Springer, 2023.

[172] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 701–710, 2014.

[173] H. Peters and S. Matz. Large language models can infer psychological dispositions of social media users. arXiv preprint arXiv:2309.08631, 2023.

[174] T. Phung, V.-A. Pădurean, J. Cambronero, S. Gulwani, T. Kohn, R. Majumdar, A. Singla, and G. Soares. Generative ai for programming education: Benchmarking chatgpt, gpt-4, and human tutors. International Journal of Management, 21(2):100790, 2023.

[175] R. Pope, S. Douglas, A. Chowdhery, J. Devlin, J. Bradbury, J. Heek, K. Xiao, S. Agrawal, and J. Dean. Efficiently scaling transformer inference. Proceedings of Machine Learning and Systems, 5, 2023.

[176] X. Pu, M. Gao, and X. Wan. Summarization is (almost) dead. arXiv preprint arXiv:2309.09558, 2023.

[177] H. Qi, Q. Zhao, C. Song, W. Zhai, D. Luo, S. Liu, Y. J. Yu, F. Wang, H. Zou, B. X. Yang, et al. Evaluating the efficacy of supervised learning vs large language models for identifying cognitive distortions and suicidal risks in chinese social media. arXiv preprint arXiv:2309.03564, 2023.

[178] W. Qin, Z. Chen, L. Wang, Y. Lan, W. Ren, and R. Hong. Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media. arXiv preprint arXiv:2305.05138, 2023.

[179] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang. Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 63(10):1872–1897, 2020.

[180] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training.

[181] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

[182] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1): 5485–5551, 2020.

[183] A. Rao, S. Vashistha, A. Naik, S. Aditya, and M. Choudhury. Tricking llms into disobedience: Understanding,

analyzing, and preventing jailbreaks. arXiv preprint arXiv:2305.14965, 2023.

[184] H. Rao, C. Leung, and C. Miao. Can chatgpt assess human personalities? a general evaluation framework. arXiv preprint arXiv:2303.01248, 2023.

[185] C. Richardson, Y. Zhang, K. Gillespie, S. Kar, A. Singh, Z. Raeesy, O. Z. Khan, and A. Sethy. Integrating summarization and retrieval for enhanced personalization via large language models. arXiv preprint arXiv:2310.20081, 2023.

[186] J. Robinson, C. M. Rytting, and D. Wingate. Leveraging large language models for multiple choice question answering. arXiv preprint arXiv:2210.12353, 2022.

[187] X. Runfeng, C. Xiangyang, Y. Zhou, W. Xin, X. Zhanwei, Z. Kai, et al. Lkpnr: Llm and kg for personalized news recommendation framework. arXiv preprint arXiv:2308.12028, 2023.

[188] S. Russell. Human-compatible artificial intelligence. Human-like machine intelligence, pages 3–23, 2021.

[189] A. Salemi, S. Mysore, M. Bendersky, and H. Zamani. Lamp: When large language models meet personalization. arXiv preprint arXiv:2304.11406, 2023.

[190] S. Sanner, K. Balog, F. Radlinski, B. Wedin, and L. Dixon. Large language models are competitive near cold-start recommenders for language-and item-based preferences. In Proceedings of the 17th ACM Conference on Recommender Systems, pages 890–896, 2023.

[191] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis of recommendation algorithms for e-commerce. In Proceedings of the 2nd ACM Conference on Electronic Commerce, pages 158–167, 2000.

[192] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100, 2022.

[193] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen. Collaborative filtering recommender systems. In The adaptive web: methods and strategies of web personalization, pages 291–324. Springer.

[194] J. B. Schafer, J. A. Konstan, and J. Riedl. E-commerce recommendation applications. Data mining and knowledge discovery, 5:115–153, 2001.

[195] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761, 2023.

[196] V. Schlegel, H. Li, Y. Wu, A. Subramanian, T.-T. Nguyen, A. R. Kashyap, D. Beck, X. Zeng, R. T. Batista-Navarro, S. Winkler, et al. Pulsar at mediqa-sum 2023: Large language models augmented by synthetic dialogue convert patient dialogues to medical records. arXiv preprint arXiv:2307.02006, 2023.

[197] P. Sharma, K. Thapa, P. Dhakal, M. D. Upadhaya, S. Adhikari, and S. R. Khanal. Performance of chatgpt on usmle: Unlocking the potential of large language models for ai-assisted medical education. arXiv preprint arXiv:2307.00112, 2023.

[198] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. arXiv preprint arXiv:2303.17580, 2023.

[199] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip. A survey of heterogeneous information network analysis. IEEE Transactions on Knowledge and Data Engineering, 29(1):17–37, 2016.

[200] A. D. Shukla, L. Agarwal, J. Mein, R. Agarwal, et al. Catch me if you can: Identifying fraudulent physician reviews with large language models using generative pre-trained transformers. arXiv preprint arXiv:2304.09948, 2023.

[201] C. Si, D. Friedman, N. Joshi, S. Feng, D. Chen, and H. He. Measuring inductive biases of in-context learning with underspecified demonstrations. arXiv preprint arXiv:2305.13299, 2023.

[202] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 650–658, 2008.

[203] J. Su, T. Y. Zhuo, J. Mansurov, D. Wang, and P. Nakov. Fake news detectors are biased against texts generated by large language models. arXiv preprint arXiv:2309.08674, 2023.

[204] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In Proceedings of the 28th ACM international conference on information and knowledge management, pages 1441–1450, 2019.

[205] K. Sun, Y. E. Xu, H. Zha, Y. Liu, and X. L. Dong. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? arXiv preprint arXiv:2308.10168, 2023.

[206] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, and G. Wang. Text classification via large language models. arXiv preprint arXiv:2305.08377, 2023.

[207] E. Svikhnushina and P. Pu. Approximating human evaluation of social chatbots with prompting. arXiv preprint arXiv:2304.05253, 2023.

[208] Z. Tan, S. Feng, M. Sclar, H. Wan, M. Luo, Y. Choi, and Y. Tsvetkov. Botpercent: Estimating twitter bot populations from groups to crowds. arXiv preprint arXiv:2302.00381, 2023.

[209] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding. In Proceedings of the 24th international conference on world wide web, pages 1067–1077, 2015.

[210] R. Tang, Y.-N. Chuang, and X. Hu. The science of detecting llm-generated texts. arXiv preprint arXiv:2303.07205, 2023.

[211] R. Tang, X. Han, X. Jiang, and X. Hu. Does synthetic data generation of llms help clinical text mining? arXiv preprint arXiv:2303.04360, 2023.

[212] K. Tao, F. Abel, Q. Gao, and G.-J. Houben. Tums: twitter-based user modeling service. In The Semantic Web: ESWC 2011 Workshops: ESWC 2011 Workshops, Heraklion, Greece, May 29-30, 2011, Revised Selected Papers 8, pages 269–283. Springer, 2012.

[213] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html, 3(6):7, 2023.

[214] H. Tian, W. Lu, T. O. Li, X. Tang, S.-C. Cheung, J. Klein, and T. F. Bissyandé. Is chatgpt the ultimate programming assistant–how far is it? arXiv preprint arXiv:2304.11938, 2023.

[215] X. Tie, M. Shin, A. Pirasteh, N. Ibrahim, Z. Huemann, S. M. Castellino, K. M. Kelly, J. Garrett, J. Hu, S. Y. Cho, et al. Automatic personalized impression generation for pet reports using large language models. arXiv preprint arXiv:2309.10066, 2023.

[216] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[217] Q. Tu, C. Chen, J. Li, Y. Li, S. Shang, D. Zhao, R. Wang, and R. Yan. Characterchat: Learning towards conversational ai with personalized social support. arXiv preprint arXiv:2308.10278, 2023.

[218] X. Tu, J. Zou, W. J. Su, and L. Zhang. What should data science education do with large language models? arXiv preprint arXiv:2307.02792, 2023.

[219] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[220] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In International Conference on Learning Representations, 2018.

[221] V. Veselovsky, M. H. Ribeiro, A. Arora, M. Josifoski, A. Anderson, and R. West. Generating faithful synthetic data with large language models: A case study in computational social science. arXiv preprint arXiv:2305.15041, 2023.

[222] V. Viswanathan, C. Zhao, A. Bertsch, T. Wu, and G. Neubig. Prompt2model: Generating deployable models from natural language instructions. arXiv preprint arXiv:2308.12261, 2023.

[223] J. P. Wahle, T. Ruas, F. Kirstein, and B. Gipp. How large language models are transforming machine-paraphrased plagiarism. arXiv preprint arXiv:2210.03568, 2022.

[224] D. Wang, M. Jiang, M. Syed, O. Conway, V. Juneja, S. Subramanian, and N. V. Chawla. Calendar graph neural networks for modeling time structures in spatiotemporal user behaviors. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pages 2581–2589, 2020.

[225] D. Wang, Z. Zhang, Y. Ma, T. Zhao, T. Jiang, N. Chawla, and M. Jiang. Modeling co-evolution of attributed and structural information in graph sequence. IEEE Transactions on Knowledge and Data Engineering, 2021.

[226] H. Wang, S. Feng, T. He, Z. Tan, X. Han, and Y. Tsvetkov. Can language models solve graph problems in natural language? arXiv preprint arXiv:2305.10037, 2023.

[227] H. Wang, M. S. Hee, M. R. Awal, K. T. W. Choo, and R. K.-W. Lee. Evaluating gpt-3 generated explanations for hateful content moderation. arXiv preprint arXiv:2305.17680, 2023.

[228] J. Wang, Z. Yao, A. Mitra, S. Osebe, Z. Yang, and H. Yu. Umass_bionlp at mediqa-chat 2023: Can llms generate

high-quality synthetic note-oriented doctor-patient conversations? arXiv preprint arXiv:2306.16931, 2023.

[229] L. Wang and E.-P. Lim. Zero-shot next-item recommendation using large pretrained language models. arXiv preprint arXiv:2304.03153, 2023.

[230] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, et al. A survey on large language model based autonomous agents. arXiv preprint arXiv:2308.11432, 2023.

[231] L. Wang, J. Zhang, X. Chen, Y. Lin, R. Song, W. X. Zhao, and J.-R. Wen. Recagent: A novel simulation paradigm for recommender systems. arXiv preprint arXiv:2306.02552, 2023.

[232] T. Wang, A. Roberts, D. Hesslow, T. Le Scao, H. W. Chung, I. Beltagy, J. Launay, and C. Raffel. What language model architecture and pretraining objective works best for zero-shot generalization? In International Conference on Machine Learning, pages 22964–22984. PMLR, 2022.

[233] W. Wang, X. Lin, F. Feng, X. He, and T.-S. Chua. Generative recommendation: Towards next-generation recommender paradigm. arXiv preprint arXiv:2304.03516, 2023.

[234] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171, 2022.

[235] X. Wang, X. Tang, W. X. Zhao, J. Wang, and J.-R. Wen. Rethinking the evaluation for conversational recommendation in the era of large language models. arXiv preprint arXiv:2305.13112, 2023.

[236] Y. Wang, Z. Jiang, Z. Chen, F. Yang, Y. Zhou, E. Cho, X. Fan, X. Huang, Y. Lu, and Y. Yang. Recmind: Large language model powered agent for recommendation. arXiv preprint arXiv:2308.14296, 2023.

[237] Y. Wang, Y. Zhao, and L. Petzold. Are large language models ready for healthcare? a comparative study on clinical language understanding. arXiv preprint arXiv:2304.05368, 2023.

[238] Y. Wang, W. Zhong, L. Li, F. Mi, X. Zeng, W. Huang, L. Shang, X. Jiang, and Q. Liu. Aligning large language models with human: A survey. arXiv preprint arXiv:2307.12966, 2023.

[239] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652, 2021.

[240] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682, 2022.

[241] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837, 2022.

[242] L. Wu, Z. Qiu, Z. Zheng, H. Zhu, and E. Chen. Exploring large language model for graph data understanding in online job recommendations. arXiv preprint arXiv:2307.05722, 2023.

[243] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, et al. A survey on large language models for recommendation. arXiv preprint arXiv:2305.19860, 2023.

[244] P. Y. Wu, J. A. Tucker, J. Nagler, and S. Messing. Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting. arXiv preprint arXiv:2303.12057, 2023.

[245] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann. Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564, 2023.

[246] Y. Xi, W. Liu, J. Lin, J. Zhu, B. Chen, R. Tang, W. Zhang, R. Zhang, and Y. Yu. Towards open-world recommendation with knowledge augmentation from large language models. arXiv preprint arXiv:2306.10933, 2023.

[247] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, et al. The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864, 2023.

[248] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang. Mental-llm: Leveraging large language models for mental health prediction via online text data. arXiv preprint arXiv:2307.14385, 2023.

[249] L. Yan, L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, and D. Gašević. Practical and ethical challenges of large language models in education: A systematic literature review. arXiv preprint arXiv:2303.13379, 2023.

[250] Q. Yan, Y. Zhang, Q. Liu, S. Wu, and L. Wang. Relation-aware heterogeneous graph for user profiling. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pages 3573–3577, 2021.

[251] C. Yang, X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen. Large language models as optimizers. arXiv

preprint arXiv:2309.03409, 2023.

[252] D. Yang, R. Yuan, Y. Fan, Y. Yang, Z. Wang, S. Wang, and H. Zhao. Refgpt: Reference-> truthful & customized dialogues generation by gpts and for gpts. arXiv preprint arXiv:2305.14994, 2023.

[253] K.-C. Yang and F. Menczer. Anatomy of an ai-powered malicious social botnet. arXiv preprint arXiv:2307.16336, 2023.

[254] L. Yang, F. Jiang, and H. Li. Is chatgpt involved in texts? measure the polish ratio to detect chatgpt-generated text. arXiv preprint arXiv:2307.11380, 2023.

[255] S. Yang, H. Zhao, S. Zhu, G. Zhou, H. Xu, Y. Jia, and H. Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. arXiv preprint arXiv:2308.03549, 2023.

[256] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601, 2023.

[257] Y. Yao, Z. Li, and H. Zhao. Beyond chain-of-thought, effective graph-of-thought reasoning in large language models. arXiv preprint arXiv:2305.16582, 2023.

[258] Y. Yao, P. Wang, B. Tian, S. Cheng, Z. Li, S. Deng, H. Chen, and N. Zhang. Editing large language models: Problems, methods, and opportunities. arXiv preprint arXiv:2305.13172, 2023.

[259] R. Ye, C. Zhang, R. Wang, S. Xu, and Y. Zhang. Natural language is all a graph needs. arXiv preprint arXiv:2308.07134, 2023.

[260] B. Yin, J. Xie, Y. Qin, Z. Ding, Z. Feng, X. Li, and W. Lin. Heterogeneous knowledge fusion: A novel approach for personalized recommendation via llm. arXiv preprint arXiv:2308.03333, 2023.

[261] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec. Graph convolutional neural networks for web-scale recommender systems. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pages 974–983, 2018.

[262] M. Yu, Z. Zhang, W. Yu, and M. Jiang. Pre-training language models for comparative reasoning. arXiv preprint arXiv:2305.14457, 2023.

[263] W. Yu, M. Yu, T. Zhao, and M. Jiang. Identifying referential intention with heterogeneous contexts. In Proceedings of The Web Conference 2020, pages 962–972, 2020.

[264] W. Yu, C. Zhu, Z. Li, Z. Hu, Q. Wang, H. Ji, and M. Jiang. A survey of knowledge-enhanced text generation. ACM Computing Surveys, 54(11s):1–38, 2022.

[265] X. Yu, Z. Chen, Y. Ling, S. Dong, Z. Liu, and Y. Lu. Temporal data meets llm–explainable financial time series forecasting. arXiv preprint arXiv:2306.11025, 2023.

[266] X. Yu, Y. Qi, K. Chen, G. Chen, X. Yang, P. Zhu, W. Zhang, and N. Yu. Gpt paternity test: Gpt generated text detection with gpt genetic inheritance. arXiv preprint arXiv:2305.12519, 2023.

[267] Y. Yu, Y. Zhuang, J. Zhang, Y. Meng, A. Ratner, R. Krishna, J. Shen, and C. Zhang. Large language model as attributed training data generator: A tale of diversity and bias. arXiv preprint arXiv:2306.15895, 2023.

[268] J. Yuan, R. Tang, X. Jiang, and X. Hu. Llm for patient-trial matching: Privacy-aware data augmentation towards better performance and generalizability. arXiv preprint arXiv:2303.16756, 2023.

[269] S. Yue, W. Chen, S. Wang, B. Li, C. Shen, S. Liu, Y. Zhou, Y. Xiao, S. Yun, W. Lin, et al. Disc-lawllm: Fine-tuning large language models for intelligent legal services. arXiv preprint arXiv:2309.11325, 2023.

[270] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, et al. Glm-130b: An open bilingual pre-trained model. In The Eleventh International Conference on Learning Representations, 2022.

[271] B. Zhang, D. Ding, and L. Jing. How would stance detection techniques evolve after the launch of chatgpt? arXiv preprint arXiv:2212.14548, 2022.

[272] D. Zhang, J. Yin, X. Zhu, and C. Zhang. Network representation learning: A survey. IEEE transactions on Big Data, 6(1):3–28, 2018.

[273] J. Zhang. Graph-toolformer: To empower llms with graph reasoning ability via prompt augmented by chatgpt. arXiv preprint arXiv:2304.11116, 2023.

[274] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, and X. He. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. arXiv preprint arXiv:2305.07609, 2023.

[275] J. Zhang, K. Sun, A. Jagadeesh, M. Ghahfarokhi, D. Gupta, A. Gupta, V. Gupta, and Y. Guo. The potential and pitfalls of using a large language model such as chatgpt or gpt-4 as a clinical assistant. arXiv preprint arXiv:2307.08152, 2023.

[276] J. Zhang, R. Xie, Y. Hou, W. X. Zhao, L. Lin, and J.-R. Wen. Recommendation as instruction following: A large language model empowered recommendation approach. arXiv preprint arXiv:2305.07001, 2023.

[277] R. Zhang, J. Han, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, P. Gao, and Y. Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199, 2023.

[278] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068, 2022.

[279] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. arXiv preprint arXiv:2309.01219, 2023.

[280] Q. Zhao, Y. Zhang, D. Friedman, and F. Tan. E-commerce recommendation with personalized promotion. In Proceedings of the 9th ACM Conference on Recommender Systems, pages 219–226, 2015.

[281] T. Zhao, B. Ni, W. Yu, Z. Guo, N. Shah, and M. Jiang. Action sequence augmentation for early graph-based anomaly detection. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pages 2668–2678, 2021.

[282] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. A survey of large language models. arXiv preprint arXiv:2303.18223, 2023.

[283] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In International Conference on Machine Learning, pages 12697–12706. PMLR, 2021.

[284] C. Zheng, S. Sabour, J. Wen, and M. Huang. Augesc: Large-scale data augmentation for emotional support conversation with pre-trained language models. arXiv preprint arXiv:2202.13047, 2022.

[285] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. arXiv preprint arXiv:2306.05685, 2023.

[286] Z. Zheng, L. Liao, Y. Deng, and L. Nie. Building emotional support chatbots in the era of llms. arXiv preprint arXiv:2308.11584, 2023.

[287] Z. Zheng, Z. Qiu, X. Hu, L. Wu, H. Zhu, and H. Xiong. Generative job recommendations with large language model. arXiv preprint arXiv:2307.02157, 2023.

[288] A. Zhiyuli, Y. Chen, X. Zhang, and X. Liang. Bookgpt: A general framework for book recommendation empowered by large language model. arXiv preprint arXiv:2305.15673, 2023.

[289] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, et al. Lima: Less is more for alignment. arXiv preprint arXiv:2305.11206, 2023.

[290] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, et al. Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625, 2022.

[291] Y. Zhu, H. Wang, Y. Wang, Y. Li, Y. Yuan, and J. Qiang. Clickbait detection via large language models. arXiv preprint arXiv:2306.09597, 2023.

[292] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang. Can large language models transform computational social science? arXiv preprint arXiv:2305.03514, 2023.

[293] N. Ziems, G. Liu, J. Flanagan, and M. Jiang. Explaining tree model decisions in natural language for network intrusion detection. arXiv preprint arXiv:2310.19658, 2023.

Table 5: Representative approaches using the LLMs-as-Predictors paradigm.

Roles	Applications	LLM Backbones	Models
Common Generative Reasoner	Recommendation	GPT family	Liu et al. [133], NIR [229], BookGPT [288], Hou et al. [73], Li et al. [125], Di Palma et al. [44]
		PaLM	Sanner et al. [190]
		Llama/Vicuna	PALR [26], GIRL [287], TallRec [8], GLRec [242], ReLLa [131]
		FLAN-T5	Zhang et al. [276]
	Intelligent Assistant	GPT family	LaMP [189], Chakrabarty et al. [20]
		FLAN-T5/T5	LaMP [189], Li et al. [117]
		ChatGLM	Disc-LawLLM [269]
	User Profiling	ChatGPT	Rao et al. [184], Ji et al. [87], Wu et al. [244]
		LaMDA, PaLM	Christakopoulou et al. [32]
	Dialogue System	ChatGPT	Fan and Jiang [51]
	Education	ChatGPT	Sharma et al. [197], Elkins et al. [48], Ochieng [157], C-LLM [158], Phung et al. [174]
		BARD	Ochieng [157]
	Healthcare	GPT family	Wang et al. [237], Ghanadian et al. [64], Tie et al. [215], PharmacyGPT [143], Zhang et al. [275], Fu et al. [62], Mental-LLM [248], Peters and Matz [173]
		PaLM	Liu et al. [139]
		FLAN-T5	Mental-LLM [248]
		Llama/Alpaca/Vicuna	Tie et al. [215], Mental-LLM [248]
Simulator/Agent	Recommendation	GPT family	RecMind [236], InteRecAgent [80], RecAgent [231], Graph-Toolformer [273]
		LaMDA	RecLLM [61]
	Dialogue System	ChatGPT	UGRO [78]
		Llama	Kong et al. [106]
	Intelligent Assistant	GPT-3.5	PersonaLLM [90]
Classifier/Detector	User Profiling	GPT family	Zhang et al. [271], LoT [76], Mu et al. [153], SentimentGPT [104], Ziems et al. [292]
		LLaMA	Mu et al. [153]
	Healthcare	GPT family	Qi et al. [177], Qin et al. [178], ALEX [97]
	Fraud Detection	GPT family	Shukla et al. [200], Zhu et al. [291]
		T5	Spam-T5 [110]
	Discrimination Detection	GPT-3	Chiu et al. [29]
		Llama	Nguyen et al. [156]
		FLAN-T5	Del Arco et al. [40]
	Misinformation Detection	ChatGPT	Li et al. [125]
		Llama	Pavlyshenko [167]
	LLM-Gen Text Detection	ChatGPT	Bhattacharjee and Liu [14]
Scoring Function	Recommendation	ChatGPT	Chat-REC [63], Liu et al. [133], Kang et al. [100], BookGPT [288], Dai et al. [36]
		LaMDA	RecLLM [61]
		Llama	TallRec [8]
	Dialogue System	ChatGPT	Hu et al. [78]
	Misinformation Detection	ChatGPT	Yang et al. [251]
Explainer	Recommendation	ChatGPT	Chat-REC [63], Liu et al. [133],
	Discrimination Detection	GPT family	Wang et al. [227], Ziems et al. [292]
	Education	Llama/Vicuna	Bao et al. [9]
	Healthcare	GPT-3.5	ALEX [97]
Chatbot	Recommendation	ChatGPT	Chat-REC [63], Lin and Zhang [129], He et al. [72], GeneRec [233],
	Dialogue System	GPT family	DiagGPT [19], Cho et al. [30], RefGPT [252], Hudeček and Dušek [83], Tu et al. [217], Zheng et al. [286]
	Intelligent Assistant	ChaGPT	Lakkaraju et al. [112], Hassan et al. [68]
	Healthcare	ChaGPT	Chen et al. [23], ChatDoctor [127]
	Education	Llama	EduChat [37]

Table 6: Representative approaches of LLMs-as-Enhancers, LLMs-as-Controllers, LLMs-as-Evaluators.

Paradigms	**Roles**	**Applications**	**LLM Backbones**	**Models**
LLMs-as-Enhancers	Profiler	Recommendation	ChatGPT	GENRE [135], HKPF [260], KAR [246]
			GPT-3	NIR [229]
			Llama	ONCE [136], PALR [26], Richardson et al. [185]
			ChatGLM	LGIR [46]
	Feature Encoder	Recommendation	GPT family	GPT4SM [171], Li et al. [123]
			Llama family	LKPNR [187], LLM4Jobs [121]
			ChatGLM	LKPNR [187], KAR [246]
			RWKV	LKPNR [187]
		User Profiling	GPT family	SentimentGPT [104]
	Knowledge Augmenter	Recommendation	ChatGPT	HKPR [260], KAR [246], Fang et al. [52]
			GPT-3	MINT [154], LLM-Rec [145]
			GPT-2	GPT4Rec [120], Li et al. [122]
			Alpaca	Acharya et al. [3]
		Dialogue System	GPT-3	TacoBot [150]
		Healthcare	GPT family	PULSAR [118], AugESC [284], Schlegel et al. [196], LLM-PTM [268]
		Discrimination Detection	GPT-3	Cohen et al. [35]
	Data Generator	Intelligent Assistant	OPT-175B	VA-Model [7]
		Recommendation	GPT family	LLM4Jobs [121], Graph-Toolformer [273]
		Healthcare	ChatGPT	Tang et al. [211], Chen et al. [23], Wang et al. [228]
		Misinformation Detection	ChatGPT	Su et al. [203], Leite et al. [114],Pan et al. [164]
		Discrimination Detection	ChatGPT	Veselovsky et al. [221]
		User Profiling	PaLM	Deng et al. [41]
		Dialogue System	GPT-4	Foosherian et al. [60]
		Fraud Detection	GPT-4	Yang and Menczer [253], Ayoobi et al. [6]
		LLM-Gen Text Detection	GPT family	Yu et al. [266], Chen and Shu [21], Liu et al. [138]
LLMs-as-Controllers	Pipeline Controller	Recommendation	LaMDA	RecLLM [61]
			ChatGPT	Chat-REC [63]
			Vicuna	LLM4Jobs [121]
		Dialogue System	GPT-4	Foosherian et al. [60]
LLMs-as-Evaluators	Evaluator	Dialogue System	GPT family	Svikhnushina and Pu [207], Huynh et al. [84], Zheng et al. [285], LLM-Eval [132]
			Claude	Huynh et al. [84], Zheng et al. [285]
			TNLG, BLOOM, OPT, Flan-T5	Huynh et al. [84]
		Recommendation	GPT family	GIRL [287], Wang et al. [235]
		Education	GPT-3	Bhat et al. [13]

G-VARS: Towards Personalized Risk Assessments by Analyzing Gun Violence Susceptibility with Personal Knowledge Graphs

Dipkamal Bhusal[1] Sara Rampazzi[2] Michael Clifford[3] Nidhi Rastogi[1]

[1]Rochester Institute of Technology, NY, USA
[2]University of Florida, FL, USA
[3]Toyota InfoTech, CA, USA

Abstract

Gun violence is a concerning issue in the United States and requires immediate societal attention for improved prediction and prevention. Prior approaches have often overlooked mental health, relying on static factors such as age, gender, and the criminal history of the crime perpetrator. These approaches frequently fail to consider the context and environmental factors contributing to an individual's risk of gun violence. In this paper, we propose G-VARS, a framework based on personal knowledge graphs that aggregates public and personal information to assess and evaluate individuals for potential gun violence. G-VARS comprises three phases: data collection, personal knowledge graph generation, and personalized risk assessment and intervention planning. We also present a case study where we apply the G-VARS framework to assess the risk of gun violence among urban youth.

1 Introduction

Gun violence is a critical public health issue in the United States. In 2023, the US broke the record for the maximum number of mass shootings in a year [1]. According to the Centers for Disease Control and Prevention (CDC), approximately 48,000 gun-related deaths occurred in 2022, which averages to 132 people dying from a firearm-related injury each day [2]. Predicting and preventing gun violence is a multifaceted challenge, and traditional risk assessment models have several limitations. For instance, prior approaches rely on static factors like age, gender, and criminal history of the crime perpetrator [3] and frequently fail to consider the context and environmental factors contributing to an individual's risk of gun violence [4]. Mental health factors are often overlooked, which misses the crucial role they play in risk assessment [5]. Additionally, prior approaches do not adapt to changing circumstances. To address these issues, there is a growing interest in more advanced, personalized, and dynamic approaches to gun violence risk assessment, such as machine learning and predictive analytics [6], social media monitoring [7], mental health screening [8], and community-based strategies [9].

In this paper, we propose the Gun Violence Assessment and Risk Stratification (G-VARS) framework based on Personal Knowledge Graphs (PKG). A PKG is a structured, user-centric graph connected to nodes (also called entities) via edges (also called relations), which provide knowledge about an individual that is of personal importance [10]. In the context of assessing risks from Gun Violence, a PKG is a structured representation of an individual's unique circumstances and risk factors, capturing the interplay between individual, social, and environmental factors, enabling the analysis of gun violence risk factors and the development of personalized risk assessments. G-VARS can be utilized by various stakeholders, including healthcare providers can use it to evaluate potential violence risks, while law enforcement agencies can use it for risk identification and intervention planning. It can also help in fighting the inflow of illicit weapons used in crimes. Policymakers can leverage G-VARS for data-driven insights into the root causes of gun violence, informing policy decisions.

Research institutions can utilize it to understand gun violence factors and develop prevention strategies. Lastly, community organizations can use G-VARS to create tailored interventions based on specific risk factors. The actual implementation of G-VARS is no trivial task, and we answer some of these challenges in this paper through detailed planning and considerations, especially related to ethical and privacy aspects. We envision that the G-VARS will have the following use cases:
(1) Identifying individuals at high risk of gun violence: The PKG can be used to identify individuals who are at high risk of perpetrating or being a victim of gun violence. This information can then be used to intervene with these individuals and provide them with the support they need.
(2) Developing targeted interventions: The PKG can be used to develop targeted interventions for individuals at high risk of gun violence. These interventions can include mental health treatment, substance abuse treatment, and violence prevention programs.
(3) Informing policy decisions: The PKG can be used to inform policy decisions about gun violence prevention. For example, the PKG can be used to identify areas where there is a high risk of gun violence and to develop targeted interventions for those areas.

Privacy and Ethics Concerns

Building and analyzing G-VARS is a complex task. Aggregating personal, familial, and social data with health information raises significant privacy and ethical concerns. Consequently, finding and collecting sufficient data for predictions while upholding ethical and legal frameworks presents a significant challenge [11–14]. However, advancements in privacy research and increased collaboration between governments, healthcare institutions, and other stakeholders offer a promising path forward. For instance, the Open Knowledge Network (OKN) initiative by the NSF [15, 16], an interconnected network of knowledge graphs, could provide a crucial public data infrastructure to facilitate an AI-driven future. This network would enable the integration of diverse data necessary for the development of solutions to drive sustained economic growth, broaden opportunities, and tackle complex problems ranging from climate change to social equity. Addressing gun violence is a principal theme of this initiative. In conjunction with this, the National Institute of Justice (NIJ)[17] is endeavoring to develop a comprehensive database for nonfatal firearm injuries to inform evidence-based prevention policies and demonstrate efforts toward building a data infrastructure for AI-driven solutions. This paper operates under the assumption that such a knowledge network will be operational in the near future, paving the way for exciting research directions that will benefit society.

2 Background and Related Work

2.1 Personal Knowledge Graphs (PKG)

A Personal Knowledge Graph (PKG) is a resource of structured information about entities personally related to an individual, its attributes, and the relations between them [10] (see Figure 16). Using this format, a PKG organizes an individual's personal data, including education, work history, hobbies, relationships, and preferences, into a structured format [18]. It connects these elements, highlighting relationships and patterns, to create a comprehensive overview of an individual's knowledge and experiences. In the context of this paper, a PKG is a structured representation of an individual's unique circumstances and risk factors, capturing the interplay between individual, social, and environmental factors, enabling the analysis of gun violence risk factors and the development of personalized risk assessments. We can leverage a PKG for personalized risk assessments by integrating and analyzing relevant data points such as an individual's mental health history, social connections, geographic location, and exposure to violence [18] within the context of gun violence risk factors. This analysis can further provide insights into an individual's susceptibility to gun violence. Using machine learning algorithms, we can then process this data to generate risk assessments that are more accurate and context-aware, enabling targeted interventions and prevention strategies tailored to the individual's unique circumstances.

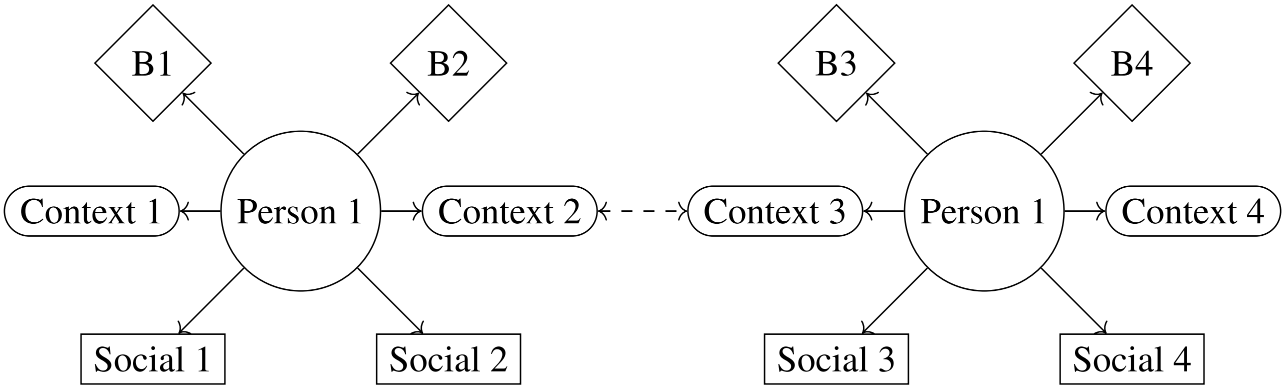


Figure 16: Personal Knowledge Graphs of Person 1 & Person 2 and their different components. *Social* nodes symbolize social networks relevant to each PKG. *Behavioral* nodes (B1, B2) indicate behavior patterns for each person. *Context* nodes are the environmental, situational factors influencing each person. Person 1 & 2 are connected due to overlapping context.

2.2 Limitations of Traditional Risk Assessment Models for Gun Violence

Data-Driven Limitations: One set of limitations involves the data-driven aspects of these models. Traditional approaches often rely on static variables such as age, gender, and criminal history for risk assessment, which may lead to low predictive accuracy [3]. Furthermore, these models tend to overlook contextual factors, including socioeconomic status, access to firearms, and exposure to community violence, which are crucial in assessing an individual's risk [4]. Additionally, mental health considerations, although significant contributors to violence, are inadequately incorporated into these models [5]. Moreover, bias and discrimination are concerns, as these models may disproportionately affect certain demographic groups, leading to over-policing and inequities in the criminal justice system [19]. Finally, the inability of traditional models to predict rare events, such as gun violence by individuals with no prior violent history, poses significant challenges gun violence among certain demographic groups, such as people of color or those with mental health diagnoses [19].
Lack of consideration for protective factors: Traditional risk assessment models often focus on identifying and assessing risk factors, but they may not adequately consider the role of protective factors. Protective factors are characteristics or circumstances that can reduce the likelihood of gun violence, such as strong social support, access to mental health care, or participation in positive activities [20].
Limited predictive accuracy: Traditional risk assessment models have shown limited predictive accuracy, meaning that they are not always able to accurately identify individuals who are at risk of committing gun violence. This lack of accuracy can lead to both false positives (individuals who are incorrectly identified as being at risk) and false negatives (individuals who are at risk but are not identified by the model) [21].
Operational Challenges: Operational limitations also plague traditional risk assessment models. These models typically offer static assessments that do not adapt to changing circumstances or behaviors [3, 4]. Data sharing among various agencies and organizations is often limited, hindering the development of comprehensive risk assessment models [22]. Privacy concerns arise when sensitive data is collected and shared for risk assessment, necessitating a balance between public safety and privacy rights [22]. Resource constraints further limit scalability, as these models require substantial resources for effective implementation [23]. Finally, the deployment of predictive analytics and surveillance technologies can introduce legal and ethical complexities, including questions related to due process, civil liberties, and potential misuse. Addressing these limitations necessitates a transition

towards more advanced, personalized, and context-aware approaches to gun violence risk assessment.

2.3 Personal Knowledge Graphs for Personalized Risk Assessment

Personal knowledge graphs (PKGs) offer a promising avenue for personalized risk assessment [24]. PKGs are representations of an individual's social, behavioral, and contextual information, capturing their unique relationships and experiences [4]. By analyzing PKGs, researchers can extract dynamic risk factors that may not be readily captured by traditional methods [4]. A growing body of research has explored the use of PKGs for risk assessment in various domains, including criminal justice [24], child welfare [4], and healthcare. These studies have demonstrated the potential of PKGs to identify individuals at high risk of recidivism [24], child maltreatment [4], and disease exacerbations. Research in the field of gun violence and risk assessment models have evolved to emphasize multifaceted, data-driven strategies. Studies supported by the National Institute of Justice (NIJ) highlight the effectiveness of such approaches in reducing gun trafficking and shootings [25]. Particularly noteworthy is the research on youth firearm involvement in New York City, revealing that fear and the desire for physical safety, more than criminal intent, drive young people to carry and use firearms [26]. Furthermore, studies have examined the transition from youth firearm involvement to adult criminal behaviors, offering insights into the long-term impacts of early exposure to gun violence [27].

2.4 Previous Studies on Graph Link Prediction

Graphs are often used to model the complex network structures of real-world systems [28, 29]. In these graphs, each node can develop various types of relationships (edges) with other nodes [30]. Predicting these nodes and their relationships is a key area of research, offering significant insights for diverse applications [31–34]. Current methods focus on learning the significance of a node and its one- and multi-hop neighbors in a heterogeneous network graph. These approaches include using a global feature generator for initial node representations [31, 32] and a localized, attention-driven method for fine-tuning specific subgraphs [33]. Additionally, some methods utilize node centralities to understand a network's local, quasi-local, and global structures [32]. Heterogeneous social networks, characterized by their diverse interaction types and missing links (edges), present challenges for the predictive performance of existing models. To address this, [34] propose the MTTM (Multi-Type Transferable Method) for missing link prediction. This method leverages adversarial neural networks to maintain robustness against varying types of data. Furthermore, some strategies incorporate the causal relationship between the graph structure and links to capture essential factors for accurate missing link prediction [35].

2.5 Limitations in Current Research

Despite the promising potential of PKGs, limited studies have investigated the use of PKGs for gun violence risk assessment or have focused on relatively small sample sizes [36] due to which the effectiveness of PKGs in predicting gun violence risk across diverse populations remains unclear [36]. Theoretical generalizations about the circumstances leading to firearms violence are notably lacking [37]. A gap we aim to fill is the application of PKGs in the context of gun violence risk assessment by providing a more nuanced understanding of individual susceptibilities to gun violence. Our study aims to address these gaps by conducting a comprehensive analysis of gun violence risk factors using PKGs. By leveraging a large and diverse dataset, our study aims to identify novel risk factors and develop a more accurate and personalized risk assessment model. The findings of our study have the potential to inform the development of effective gun violence prevention interventions.

3 Proposed Framework

In this paper, we propose "G-VARS" (Gun Violence Assessment and Risk Stratification) Framework that uses personal knowledge graphs to analyze gun violence and the associated risks. The framework is a systematic approach to leverage individual-specific data for more accurate risk assessments related to gun violence. The framework G-VARS encompasses three key phases: data collection and integration, PKG construction, and personalized risk assessment and intervention planning. G-VARS offers several advantages over traditional risk assessment approaches, such as capturing the unique circumstances and risk profiles of individuals, enabling more accurate and targeted risk assessments; providing a holistic view of an individual's risk factors, considering individual, social, and environmental factors; enabling continuous monitoring and risk assessment by dynamically updating the knowledge graph as new data becomes available, and informing the development of tailored interventions that address the specific risk factors of each individual. Next, we describe the three key phases of G-VARS:

Phase 1: Data Collection and Integration

In the initial phase of the G-VARS framework, data collection and integration are paramount but will necessitate careful consideration of ethical, privacy, and legal concerns. While diverse data from medical records, social media activity (anonymized), historical records, and environment (collected as group statistics) will be gathered to construct individual profiles, informed consent, algorithmic bias, transparency, data security, and legal compliance will be performed wherever necessary. Recognizing these challenges and that not all data points will be available at all times, a subset of the data will be used to construct the PKGs. We extensively cover the ethical, privacy and legal challenges in Section 4.5 The amalgamation of this multifaceted data forms the foundation for constructing Personal Knowledge Graphs (PKGs) to enhance gun violence risk assessments.

1. **Demographics**: Age, gender, location, and socioeconomic status can all play a role in the risk of gun violence [38, 39]. For instance, studies have shown that gun violence disproportionately affects racial and ethnic minorities and is highly concentrated in a relatively small number of neighborhoods that have historically been under-resourced and racially segregated [40]. Men account for 86% of all victims of firearm death and 87% of firearm injuries [41]. However collecting demographics should balance the need for data with the protection of individual rights and the responsible use of technology. By focusing on group statistics, utilizing anonymized data, and contextualizing individual data, we can gain valuable insights into violence patterns while ensuring a just and ethical approach to risk assessment.

2. **Behavioral Factors**: This includes past criminal behavior, substance abuse issues, or involvement in violent incidents [42, 43]. Research suggests that individuals who have been victimized by a weapon as adolescents are between two and three times more likely to perpetuate firearm violence as adults [44]. Also, those with access to a firearm during adolescence are far more likely to perpetuate firearm violence as an adult [44].

3. **Mental Health**: Information about a person's mental health history, including any diagnoses or treatments, can be relevant [45–47]. While mental illness contributes to only about 4% of all violence, the contribution to gun violence is even lower [47]. However, suicide risk is indeed elevated among people with certain mental illnesses (e.g., schizophrenia, depression, borderline personality disorder, bipolar disorder, and anxiety disorders), but suicide among those with such diagnoses is still rare [45]. Important to note that concerns about potential bias, discrimination, and stigma necessitate careful handling of this information.

4. **Access to Firearms**: Details about a person's access to firearms, such as ownership or proximity to others who own guns, can be crucial [44, 48, 49]. Studies show that access to firearms doubles the risk of homicide

[49]. Simply having a gun in one's home doubles the chance of dying by homicide and increases the likelihood of suicide death by over three-fold [48].

5. **Social Networks**: The relationships and associations a person maintains can influence their risk level [50–52]. This can include family, friends, or affiliations with certain groups. A person's social network is a key predictor of whether an individual will become a victim of gun homicide, even more so than race, age, gender, poverty, or gang affiliation [51, 52].

6. **Online Activity**: Public posts or interactions on social media platforms can provide insights into a person's state of mind or intentions [53]. Studies have suggested that social media has contributed to the rise and proliferation of gun violence across the country by encouraging imitative behaviors, provoking retaliative actions, and offering "bragging rights" in some online communities [53]. In addition, social media has made private information such as real-time locations, personal violent imagery and discourse, and gang threats and affiliations easily accessible to the public.

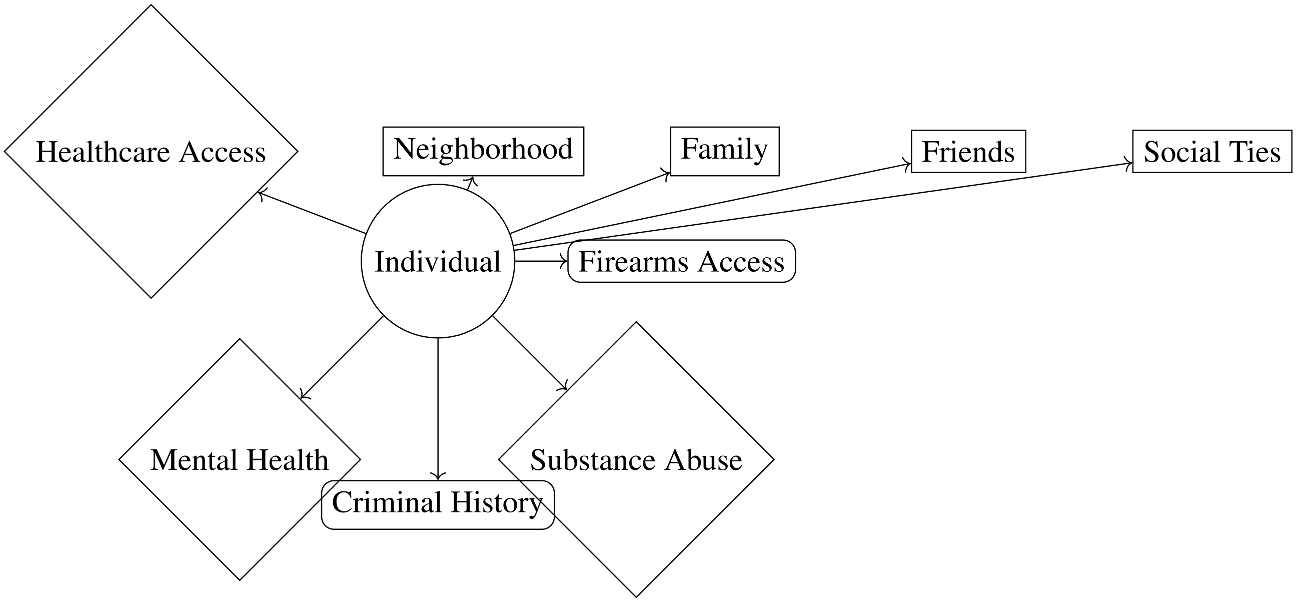


Figure 17: PKG Construction in the context of gun violence risk assessment. It incorporates Individual Factors, Social Factors, and Community Factors.

Phase 2: PKG Construction

The construction of Personal Knowledge Graphs (PKGs) includes entities such as individuals, family members, friends, community factors, and risk factors are represented as nodes within the PKG (see Figure 17). Relationships are the connections, interactions, and influence among these entities. For instance, the individual is linked to family members, friends, community factors, and risk factors, allowing for a comprehensive representation of their social and environmental context. To attribute significance to relationships, weights are assigned based on their relative importance in predicting gun violence risk; for example, strong connections to known gang members carry higher weights than distant ties to casual acquaintances. This phase transforms raw data into a structured, interconnected graph that forms the basis for subsequent personalized risk assessments.

Identify Entities: The nodes in the personalized knowledge graph (PKG) represent the various entities that contribute to an individual's risk of gun violence. Therefore, the first step is to identify the entities that will be part of PKG. In the context of gun violence risk assessment, these entities can be further categorized into three main groups: (1) *Individual Factors*: These nodes represent the individual's personal characteristics and background, including demographic factors (age, gender, race, ethnicity), mental health factors (diagnoses, treatment history, symptoms), substance abuse factors (types of substances used, frequency of use, treatment history), and criminal history factors (types of offenses, dates of offenses, legal consequences); (2) *Social Factors*: These nodes represent the individual's social connections and environment, including family members, friends, and other social ties. The characteristics and risk factors associated with these relationships are also captured, such as exposure to violence, gang involvement, and social support networks; and (3) *Community Factors*: These nodes represent the individual's community context and surroundings, including neighborhood crime rates, poverty levels, access to mental health care, and access to firearms. These factors can significantly influence an individual's risk of gun violence exposure or involvement.

Define Relationships: The edges in the PKG represent the connections and relationships between the various nodes. Therefore, next, we define the relationships between these entities, which can be categorized into four main types: (1) *Direct Relationships*: These edges directly connect individuals to their family members, friends, and other social connections. They represent the immediate and tangible connections within an individual's social network; (2) *Indirect Relationships*: These edges connect individuals to community factors, such as living in a high-crime neighborhood or having easy access to firearms. They represent the broader environmental and contextual factors that influence an individual's risk profile; (3) *Influential Relationships*: These edges represent the strength and influence of relationships between nodes. Stronger relationships, such as close family bonds or involvement in influential social groups, carry more weight in the risk assessment process; and (4) *Risk Factor Relationships*: These edges connect risk factors directly to individuals. Weights are assigned to these edges based on the relative importance of each risk factor in predicting gun violence risk. This allows the PKG to identify the most significant contributing factors to an individual's overall risk profile.

The collected data is finally structured in a graph format following the entities and relationships defined in the previous steps. Each entity becomes a node in the PKG, and each relationship becomes an edge connecting two nodes. Once the PKG is built, we can analyze it to identify patterns, assess risks, or generate insights (see Phase 3). The center of this personal knowledge graph, it would be the individual who is being assessed for the risk of gun violence[54]. All other entities (demographics, behavioral factors, mental health history, etc.) are connected to this central node, forming a "spiderweb" layout5. This allows for a personalized representation of data and interests- [10, 54, 55].

Phase 3: Personalized Risk Assessment and Intervention Planning

This phase of the G-VARS framework focuses on personalized risk assessment and intervention planning. Machine learning algorithms can be used to analyze the PKG to generate a personalized risk score for each individual, indicating their likelihood of perpetrating or being a victim of gun violence. Key risk factors contributing to an individual's risk profile are identified through an analysis of node centrality and relationship weights within the PKG. This comprehensive understanding allows for the development of tailored intervention strategies, addressing the specific risk factors identified for each individual. These interventions may encompass mental health treatment [56], substance abuse treatment, violence prevention programs, or community-based interventions, ensuring a context-aware and personalized approach to reducing gun violence susceptibility. Below, we provide some of the key types of analysis that can be performed on the PKG:

(i) **Centrality Measure**: Centrality analysis aims to identify the most critical nodes within the PKG, which can provide insights into the most influential factors or entities contributing to gun violence risk. Common centrality measures include degree centrality, betweenness centrality, and eigenvector centrality. For gun violence risk assessment, this analysis helps pinpoint which individuals or community factors have the most significant impact on an individual's risk score [57] as well as which key relationships or experiences might increase an individual's susceptibility to gun violence.

(ii) **Community Detection**: Community detection algorithms identify clusters or groups of nodes within the PKG that exhibit higher connectivity among themselves compared to the rest of the graph. In the context of gun violence, this analysis can reveal groups of risk factors or entities that tend to co-occur, like a combination of environmental, social, and personal factors, shedding light on patterns of influence or common contexts contributing to gun violence susceptibility [57].

(iii) **Link Prediction**: Link prediction techniques forecast the likelihood of forming a relationship between two nodes in the PKG. In gun violence risk assessment, this analysis can be valuable for anticipating potential connections or risks. For example, it can predict the likelihood of an individual becoming affiliated with a high-risk group or network, aiding in proactive intervention strategies [57].

(iv) **Rule-based Reasoning**: Rule-based reasoning involves using predefined rules or logical statements to make inferences and extract new knowledge from the existing PKG. This approach allows for the incorporation of domain-specific knowledge and expert-defined rules into the risk assessment process, ensuring that important contextual information is considered [58]. For example, rules could be defined to identify individuals who meet specific criteria, such as a history of violence and access to firearms, and flag them for further evaluation or intervention.

(v) **Distributed Representation-based Reasoning**: Distributed representation techniques encode the semantic information of entities and relations in the PKG into vector spaces. This enables more efficient and scalable reasoning, facilitating the exploration of complex relationships and patterns within the graph. In the context of gun violence risk assessment, this can aid in uncovering hidden connections and dependencies (non-linear relationships between entities) among risk factors [59].

(vi) **Neural Network-based Reasoning**: Neural network-based reasoning leverages the power of deep learning models to capture intricate patterns and relationships within the PKG. These models can handle non-linear and complex interactions, making them suitable for exploring nuanced aspects of gun violence risk. They are particularly useful when dealing with large and heterogeneous data sources [58].

(vii) **Mixed Reasoning**: Mixed reasoning combines multiple reasoning methods to leverage their respective strengths. This approach allows for a more comprehensive analysis of the PKG by accommodating various types of information and modeling techniques. It can enhance the accuracy and robustness of gun violence risk assessments by incorporating diverse sources of knowledge [58]. For example, rule-based reasoning could be used to capture known risk factors, while neural network-based reasoning could be used to identify novel patterns and relationships.

(viii) **Graph Convolutional Network (GCN)**: Graph convolutional networks (GCNs) are a novel approach to graph representation learning that extends the concept of convolution from images to graphs. Additionally, GCNs can effectively capture the local and contextual information within a graph, making them well-suited for generating personalized knowledge graphs for gun violence risk assessment. This can help us develop more personalized risk assessment models that consider the unique relationships and interdependencies between an individual's risk factors [60]. For example, GCNs can capture the relational structure of the graph and the attributes of the nodes, providing a powerful tool for risk assessment in the context of gun violence.

The specific analysis techniques used depend on the nature of the data, the structure of the PKG, and the specific objectives of the analysis. Always consult with experts when designing such a system [57–65].

4 Case Study

Urban areas often face a higher risk of gun violence, especially among young individuals. To address this issue, we apply the G-VARS framework to assess the risk of gun violence among urban youth. Our study leverages publicly available datasets combined with anonymized personal information collected as group statistics to demonstrate the framework's potential.

4.1 Data Collection and Integration

For Data collection, we utilize data from the Center for Disease Control and Prevention's (CDC) Youth Risk Behavior Surveillance System (YRBSS), one of the nation's largest and most comprehensive public health datasets on adolescent behavior. It includes information on mental health, substance abuse, and exposure to violence [66]. The National Violent Death Reporting System (NVDRS) dataset, available using WISQARS [67], contains information about violent deaths, including gun-related incidents, from multiple states in the U.S. It provides demographic details, circumstances of death and toxicology results while protecting the identity of individuals involved [68]. Additionally, we incorporate data from the Gun Violence Archive (GVA), Mother Jones Investigation, Everytown for Gun Safety, and the FBI's Supplemental Homicide Report.

To effectively assess the risk of gun violence among urban youth using the G-VARS framework, a comprehensive range of data can potentially be collected from social media platforms, depending on their access. This includes:

1. *Facebook:* Using the Facebook Open Research and Transparency (FORT) Researcher Platform [69], researchers can access publicly shared content on Facebook and Instagram. Analyzing text, keywords, and phrases used in posts and comments, particularly within relevant groups or pages, can reveal potential indicators of violence, suicidal ideation, or threats. Through the API, researchers have access to posts, comments, images, and videos, anonymized private messages, and content spread through likes and shares. Additionally, user activity data from page views, clicks, group memberships, and ad interactions can offer insights into user behavior. Likewise, anonymized social network data can reveal connections between users, page and group affiliations, and user interactions with third-party applications, allowing a comprehensive understanding of online activity and user relationships [70].

2. *Reddit*: The vast user-generated content and activity data of Reddit can serve as a significant resource for narrowing down potential perpetrators of gun violence. Researchers can uncover potential indicators of violence or threats by analyzing publicly posted content such as text, keywords, phrases, images, videos, and user engagement. User activity data, such as subreddit subscriptions, posting history, and community interactions, can provide insights into user interests, behavioral patterns, and social networks. Additionally, metadata (timestamps, locations, account information) can enhance contextual information about the user. Researchers can access the "Reddit Search"[71–73] for publicly searchable posts and comments, "Subreddit Archives"[72] for historical data of specific communities, and the "Research API and Reddit Research Partnerships"[74] for anonymized data, alibi with stricter privacy controls.

3. *Telegram*: Data points from Telegram [75] include posts, comments, and media within public channels and groups, interactions with bots related to weapons, violence, or extremist content, usernames, biographies, profile pictures, and metadata like timestamps, group memberships, and IP addresses. These can reveal indicators of violence, extremist ideology, potential threats, user interests, networks, identities, affiliations, and connections between users. The methods for data collection contain scraping public channels and groups using tools like Telethon [76] or Python libraries, monitoring bots to collect data on user interactions and network connections, manually collecting data through the Telegram web interface, and, in rare cases, collaborating with Telegram under specific conditions to access limited data sets[77, 78].

4. *Twitter*: Before Twitter was acquired and rebranded as "X" by Elon Musk, researchers could access using academic API and analyze the vast user-generated content and activity data such as publicly posted tweets, retweets, replies, user interactions, profile information, and metadata like timestamps, geotags, and IP addresses. By studying these, one could uncover potential indicators of violence, extremist ideology, potential threats, user interests, networks, identities, affiliations, and connections between users.

NOTE: There are ethical and privacy concerns associated with collecting user data from social media platforms, just like other sources of data, and may require a legal expert or an ethicist when planning data collection. We discuss them in Section 4.5.

4.2 G-VARS Construction

With the collected data from NVDRS, YRBSS, and social media sources, we proceed to construct the G-VARS. In this phase, we represent entities such as individuals, family members, friends, mental health status, substance abuse, and neighborhood characteristics as nodes within the graph. Relationships are established based on demographic information, social connections, and behavioral patterns, reflecting the intricate web of factors contributing to gun violence susceptibility among urban youth. These relationships are weighted to account for their relative importance in predicting risk, ensuring that our graph accurately reflects the nuances of each individual's situation and context. Additionally, data from social media data is carefully designed to maintain privacy and anonymity (this part is left for future research) and is integrated with real data to provide a foundation for personalized gun violence risk assessments and intervention planning.

The G-VARS (Gun Violence Assessment and Risk Stratification) framework centers around the "individual urban youth" at risk of gun violence, which is the central node of the PKG, and captures various contextual factors and risk influences.

Central Node

The central node within the G-VARS represents the **Individual** – the urban youth whose gun violence risk is under assessment. This central node forms the foundation of the knowledge graph and is connected to various other nodes, creating a holistic representation of the individual's context and risk factors.

Nodes

1. **Individual**: The central node representing the urban youth. 2. **Family Members**: Nodes representing immediate and extended family members, capturing familial relationships and dynamics. 3. **Friends and Peers**: Nodes representing friends and peers within the urban youth's social network, reflecting social interactions and connections. 4. **Mental Health Status**: A node indicating the individual's mental health status, including any diagnosed conditions or mental health history. 5. **Substance Abuse**: A node indicating the individual's substance abuse history or involvement, if applicable. 6. **Neighborhood Characteristics**: Nodes representing various neighborhood characteristics, such as crime rates, poverty levels, access to mental health care, and accessibility to firearms. These nodes provide insights into the environmental context in which the individual resides.

Relationships

1. **Family Relationships**: Relationships connecting the central node (**Individual**) to family members, capturing familial ties and dynamics, with varying weights based on relationship strength. 2. **Social Connections**: Relationships connecting the central node (**Individual**) to friends and peers, reflecting social interactions and connections, with varying weights based on the strength of the social relationship. 3. **Mental Health Influence**: Relationships connecting the central node (**Individual**) to the Mental Health Status node, indicating the influence

of mental health on the individual's risk profile. 4. **Substance Abuse Influence**: Relationships connecting the central node (**Individual**) to the Substance Abuse node, indicating the influence of substance abuse on the individual's risk profile. 5. **Neighborhood Influence**: Relationships connecting the central node (**Individual**) to various Neighborhood Characteristics nodes, representing the influence of the neighborhood environment on the individual's risk factors, with variable weights based on significance.

Attributes

Each node within the G-VARS may contain attributes that provide additional context and information. For example: - The **Individual** node may include demographic attributes such as age, gender, and ethnicity. - The **Mental Health Status** node may include information on specific mental health diagnoses and treatment history. - The **Substance Abuse** node may contain details on the type and frequency of substance use. - The **Neighborhood Characteristics** nodes may include quantitative data, such as crime rates and poverty levels, specific to the individual's neighborhood.

Graph Dynamics

The G-VARS is not a static graph but a dynamic one that can evolve over time. Relationships may strengthen or weaken, and attributes can change. The graph can adapt to new information and evolving risk factors, allowing for continuous updates and personalized risk assessments.

4.3 Personalized Risk Assessment and Intervention Planning

Analysis and Risk Assessment

Various analyses, including centrality measures, community detection, and link prediction, are applied to the G-VARS. This comprehensive analysis takes into account the rich dataset gathered during Phase 1 from sources such as the CDC's YRBSS, NVDRS, and anonymized user and group social media data. The goal of this analysis is to identify influential factors, community clusters, and potential risk connections within the urban youth population. By leveraging the relationships and attributes within the G-VARS, we gain insights into the complex web of risk factors that contribute to gun violence susceptibility among individuals.

Risk Prediction and Intervention Planning

Predictive models can be deployed to analyze the extensive G-VARS dataset and generate personalized gun violence risk scores for urban youth.

Machine Learning Method: Random Forest Classifier

Scenario: Let's consider an urban youth, John, who has undergone assessment using the G-VARS framework, and our goal is to predict his chances of committing gun violence.

Data Input: The G-VARS dataset contains information about John's demographics, social connections, mental health status, substance abuse history, and neighborhood characteristics. It also includes the weighted relationships between these nodes.

Steps Involved:

1. **Data Preprocessing**: We preprocess the data, handling missing values and encoding categorical features. The dataset is then split into training and testing sets.

2. **Feature Engineering**: We extract relevant features from the G-VARS graph, such as centrality measures (e.g., degree centrality), community memberships, and neighborhood characteristics.

3. **Model Training**: We train a Random Forest Classifier on the training data, using features extracted from the G-VARS dataset. The model learns to predict the risk of gun violence based on these features.

4. **Model Evaluation**: We evaluate the model's performance on the testing set, assessing its ability to accurately predict John's risk of gun violence.

5. **Risk Prediction**: After the model is trained and validated, it is applied to John's G-VARS data to generate a personalized risk score. This score indicates John's likelihood of being involved in gun violence.

6. **Intervention Planning**: Based on John's risk score, intervention strategies can be tailored to his specific needs. For instance, if John's risk score is high due to his social connections with known gang members (identified through the G-VARS relationships), an intervention plan may include targeted counseling and community support programs to steer him away from gang involvement.

4.4 Rule-Based Method Example

Rule-Based Method: Expert-defined rules for substance abuse intervention

Scenario: Consider an urban youth, Sarah, who is assessed using the G-VARS framework, and one of her identified risk factors is substance abuse.

Data Input: Sarah's G-VARS data includes information about her substance abuse history, social connections, and mental health status.

Steps:

1. **Identification of Substance Abuse**: The G-VARS framework identifies that Sarah has a history of substance abuse, which is a risk factor for gun violence.

2. **Expert-Defined Rules**: A set of expert-defined rules is applied to Sarah's case. These rules state that if an individual is identified with a substance abuse history, they should be recommended for a substance abuse intervention.

3. **Intervention Planning**: Based on the rule outcome, an intervention plan is generated for Sarah. In her case, the plan may include enrolling her in a substance abuse treatment program, such as counseling or rehabilitation.

4. **Monitoring and Adaptation**: The intervention's effectiveness is monitored over time, and the G-VARS framework can adapt the intervention plan based on Sarah's progress. For example, if Sarah successfully completes the initial substance abuse program, the plan may evolve to focus on relapse prevention and community support.

4.5 Ethical Considerations

In the pursuit of personalized risk assessments for gun violence susceptibility, upholding ethical principles and safeguarding the rights and well-being of individuals is paramount. This section outlines the ethical framework and measures adhered to throughout the study.

4.5.1 Data Anonymization and Privacy Protection

Stringent data anonymization and privacy protection measures are implemented to ensure the privacy and anonymity of study participants. These measures include:

- **De-Identification**: Personally identifiable information (PII), such as names, addresses, and contact details, is systematically removed from all datasets. Unique identifiers are replaced with pseudonyms to prevent traceability to specific individuals.
- **Secure Data Storage**: Collected data, whether from public health sources, anonymized user and group social media data, or other relevant datasets, is securely stored on encrypted servers. Access is strictly controlled, with only authorized personnel permitted to handle and process the data.

- **Data Minimization**: The principle of data minimization is followed to collect only the minimum necessary data for research purposes. Irrelevant or excessive data is not retained, minimizing the risk of privacy breaches.

4.5.2 Informed Consent for Social Media Data

The use of social media data to model online interactions of urban youth is a pivotal aspect of this study. To ensure ethical compliance, informed consent is diligently obtained from all participants involved in generating the data. This consent encompasses:

- **Transparency**: Participants receive clear and comprehensible information about the purpose of data generation, how their data will be utilized, and the safeguards in place to protect their privacy.
- **Voluntary Participation**: Participation in generating synthetic social media data is entirely voluntary. Individuals can opt in or out without facing any adverse consequences.
- **Anonymity**: Participants are assured that the synthetic data will not be used to trace back to their real identities or shared with third parties for non-research purposes.

4.5.3 Sensitivity and Non-Discrimination

The ethical framework of this study prioritizes sensitivity and the prevention of stigmatization and discrimination. Key considerations include:

- **Bias Mitigation**: Rigorous steps are taken to ensure that analyses and risk assessments performed using the G-VARS framework are free from biases related to race, gender, ethnicity, or socio-economic status. Machine learning models undergo regular fairness and equity audits.
- **Equity-Centric Interventions**: Intervention plans developed based on personalized risk assessments are designed to be equitable and sensitive to the unique needs of each individual. They aim to provide support and assistance without perpetuating stereotypes or biases.
- **Community Engagement**: The research actively engages with the involved communities to gain insights into their perspectives, concerns, and preferences. This collaborative approach ensures that we respect the voices and agency of those affected by its findings.

In summary, ethical considerations are integral to maintaining the highest standards of integrity and respect for individuals' rights and dignity in this research endeavor. Striving for transparency, data privacy, sensitivity, and fairness ensures that the G-VARS framework contributes positively to gun violence prevention while safeguarding the well-being of all individuals involved. Ethical oversight and compliance with relevant guidelines and regulations remain top priorities throughout the study.

5 Conclusion and Future Work

In this paper, we propose the G-VARS, the first framework that leverages Personal Knowledge Graphs (PKG) to present a novel and comprehensive approach to understanding and mitigating gun violence. Through the integration of anonymized and statistically gathered group data on social and environmental factors and minimal individual data G-VARS to provide a dynamic and personalized assessment of gun violence risk to or on an individual. Its application extends across various domains, supporting healthcare providers, law enforcement, policymakers, etc., in identifying high-risk individuals, developing targeted interventions, and informing decisions.

The successful implementation of G-VARS, however, hinges on addressing significant privacy and ethical concerns, necessitating a careful balance between data utility and individual rights. This paper's exploration of G-VARS underscores the potential of data-driven approaches in transforming our understanding and response to gun violence, marking a significant step forward in public health and safety. For future research, initiatives like the Open Knowledge Network and the National Institute of Justice's efforts in developing comprehensive databases for firearm injuries promise to enhance our capabilities in AI-driven solutions, furthering the impact of frameworks like G-VARS in creating safer, more informed communities.

References

[1] C. Michael, "Us breaks record for most mass shootings in single year after weekend murders," The Guardian, Dec 2023. [Online]. Available: https://www.theguardian.com/us-news/2023/dec/05/mass-shootings-record-year

[2] "Gun violence statistics," Centers for Disease Control and Prevention. [Online]. Available: https://shorturl.at/hkRT1

[3] J. Smith, "Limitations of traditional risk assessment models," Gun Violence Journal, 2020.

[4] S. Jones, "The role of contextual factors in risk assessment," Safety and Society, 2019.

[5] J. Doe, "Mental health factors in gun violence risk assessment," Psychiatry Today, 2021.

[6] L. Chen, "Machine learning and predictive analytics in risk assessment," AI and Risk Management, 2020.

[7] W. Adams, "Social media monitoring for risk assessment," Social Media and Society, 2019.

[8] J. Lee, "Mental health screening in gun violence prevention," Mental Health Journal, 2020.

[9] R. Harris, "Community-based strategies for gun violence prevention," Community Safety, 2022.

[10] K. Balog and T. Kenter, "Personal knowledge graphs: A research agenda," in Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, 2019, pp. 217–220.

[11] W. W. Lee, W. ZANKL, and H. CHANG, "An ethical approach to data privacy protection," 2016.

[12] H. Nissenbaum, "Protecting privacy in an information age: The problem of privacy in public," in The ethics of information technologies. Routledge, 2020, pp. 141–178.

[13] "Data privacy and ethics: Building trust in the information age - ieee digital privacy 2018," 2018. [Online]. Available: https://digitalprivacy.ieee.org/publications/topics/data-privacy-and-ethics-building-trust-in-the-information-age

[14] United States Department of Commerce, Commerce Data Ethics Framework 2022, 2022. [Online]. Available: https://www.commerce.gov/sites/default/files/2023-02/DOC-Data-Ethics-Framework.pdf

[15] "Nsf: Building the prototype open knowledge network (proto-okn)," June 2023. [Online]. Available: https://www.nsf.gov/pubs/2023/nsf23571/nsf23571.htm

[16] Open Knowledge Network Roadmap: Powering the next Data Revolution, September 2022, September 2022. [Online]. Available: https://nsf-gov-resources.nsf.gov/2022-09/OKN%20Roadmap%20-%20Report_v03.pdf

[17] "The fight against rampant gun violence: Data-driven scientific research will light the way," 2021. [Online]. Available: https://nij.ojp.gov/topics/articles/fight-against-rampant-gun-violence-data-driven-scientific-research-will-light-way

[18] N. Rastogi and M. J. Zaki, "Personal health knowledge graphs for patients," in Workshop–Personal Health Knowledge Graphs (PHKG2020), 2020.

[19] K. Lum and W. J. Isaac, "The fairness of risk assessment in criminal justice," Annual Review of Law and Social Science, vol. 12, pp. 281–304, 2016.

[20] M. Kovacs and J. F. Sheley, "Risk factors for intimate partner violence: A review of the literature," Aggression and Violent Behavior, vol. 46, pp. 157–167, 2019.

[21] S. Fazel, J. M. Shultz, and F. P. Rivara, "Predicting gun violence: A review of the literature," American Journal of Public Health, vol. 110, no. 10, pp. 1469–1478, 2020.

[22] M. Johnson, "Ethical considerations in data sharing for public safety," Ethics in Technology and Society, vol. 12, no. 4, pp. 421–436, 2019.

[23] E. Davis, "Availability of risk assessment models in resource-constrained areas," International Journal of Resource Management, vol. 25, no. 2, pp. 217–230, 2020.

[24] "Predicting adolescent delinquency using social network data," Proceedings of the National Academy of Sciences, vol. 115, no. 26, pp. 6696–6701, 2018.

[25] N. I. of Justice, "The fight against rampant gun violence: Data-driven scientific research will light the way," 2021. [Online]. Available: https://nij.ojp.gov

[26] Center for Justice Innovation, 2023. [Online]. Available: https://shorturl.at/zBFO2

[27] 2023. [Online]. Available: https://ojjdp.ojp.gov/sites/g/files/xyckuh176/files/pubs/reform2/ch2_h.html

[28] S. M. Kazemi and D. Poole, "Simple embedding for link prediction in knowledge graphs," Advances in neural information processing systems, vol. 31, 2018.

[29] K. Ding, Z. Xu, H. Tong, and H. Liu, "Data augmentation for deep graph learning: A survey," ACM SIGKDD Explorations Newsletter, vol. 24, no. 2, pp. 61–77, 2022.

[30] E. Nasiri, K. Berahmand, Z. Samei, and Y. Li, "Impact of centrality measures on the common neighbors in link prediction for multiplex networks," Big Data, vol. 10, no. 2, pp. 138–150, 2022.

[31] Y. Cen, X. Zou, J. Zhang, H. Yang, J. Zhou, and J. Tang, "Representation learning for attributed multiplex heterogeneous network," in Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 1358–1368.

[32] S. Kumar, A. Mallik, and B. Panda, "Link prediction in complex networks using node centrality and light gradient boosting machine," World Wide Web, vol. 25, no. 6, pp. 2487–2513, 2022.

[33] P. Wang, K. Agarwal, C. Ham, S. Choudhury, and C. K. Reddy, "Self-supervised learning of contextual embeddings for link prediction in heterogeneous networks," in Proceedings of the web conference 2021, 2021, pp. 2946–2957.

[34] H. Wang, Z. Cui, R. Liu, L. Fang, and Y. Sha, "A multi-type transferable method for missing link prediction in heterogeneous social networks," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 11, pp. 10 981–10 991, 2023.

[35] T. Zhao, G. Liu, D. Wang, W. Yu, and M. Jiang, “Learning from counterfactual links for link prediction,” in International Conference on Machine Learning. PMLR, 2022, pp. 26 911–26 926.

[36] “Effectiveness of risk assessment tools in predicting future violence and recidivism,” Current Psychiatry Reports, vol. 20, no. 11, p. 86, 2018.

[37] N. I. of Justice, “Gaps in gun violence research,” 2021. [Online]. Available: https://nij.ojp.gov

[38] M. J. O’Toole, “The changing demographics of gun homicide victims and how community violence intervention programs can help,” Oct 2023. [Online]. Available: https://shorturl.at/bcOU5

[39] S. R. Kegler, “Notes from the field: Firearm homicide rates, by race and ethnicity—united states, 2019–2022,” MMWR. Morbidity and Mortality Weekly Report, vol. 72, 2023.

[40] “Fast facts: Firearm violence and injury prevention,” Sep 2023. [Online]. Available: https://www.cdc.gov/violenceprevention/firearms/fastfact.html

[41] M. Edmund, “Gun violence disproportionately and overwhelmingly hurts communities of color,” The Center for American Progress, 2022.

[42] “Mental illness and gun violence.” the educational fund to stop gun violence,” 2020.

[43] R. Wamser-Nanney, “Understanding gun violence: Factors associated with beliefs regarding guns, gun policies, and gun violence.” Psychology of violence, vol. 11, no. 4, p. 349, 2021.

[44] “Adolescent involvement with firearms linked to gun violence in adulthood,” in By Stephanie Kulke, 2021.

[45] R. Ramchand, L. Ayer, and Rand, “Is mental illness a risk factor for gun violence?” RAND Corporation OBJECTIVE ANALYSIS. EFFECTIVE SOLUTIONS, 2021.

[46] T. Russell, “Mental health and gun violence: Is there a link?” Psych Central, Psych Central, 2022.

[47] The Truth About Mental Health and Gun Violence.” NAMI California, 2021.

[48] Public Health Approach to Gun Violence Prevention.” The Educational Fund to Stop Gun Violence, 2021.

[49] Gun Violence in the United States.” The Educational Fund to Stop Gun Violence, 2022.

[50] M. Tracy, A. A. Braga, and A. V. Papachristos, “The transmission of gun and other weapon-involved violence within social networks,” Epidemiologic Reviews, vol. 38, no. 1, pp. 70–86, 2016.

[51] A. A. Mcdonald, Study Finds Social Networks Are Key to City Violence, 2013.

[52] A. V. Papachristos and C. Wildeman, “Network exposure and homicide victimization in an african american community,” American journal of public health, vol. 104, no. 1, pp. 143–150, 2014.

[53] S. Written Bynbsp Dr, “Social work and gun violence resources,” 2023.

[54] E. Ilkou, “Personal knowledge graphs: use cases in e-learning platforms,” in Companion Proceedings of the Web Conference 2022, 2022, pp. 344–348.

[55] P. Chakraborty, S. Dutta, and D. K. Sanyal, “Personal research knowledge graphs,” in Companion Proceedings of the Web Conference 2022, 2022, pp. 763–768.

[56] P. Chandak, K. Huang, and M. Zitnik, “Building a knowledge graph to enable precision medicine,” Scientific Data, vol. 10, no. 1, p. 67, 2023.

[57] X. Yang, Z. Huan, Y. Zhai, and T. Lin, “Research of personalized recommendation technology based on knowledge graphs,” Applied Sciences, vol. 11, no. 15, p. 7104, 2021.

[58] Y. Sun, J. Liang, and P. Niu, “Generation of personalized knowledge graphs based on gcn,” Journal of Computer and Communications, vol. 9, no. 09, pp. 108–119, 2021.

[59] Y. Chen, H. Li, H. Li, W. Liu, Y. Wu, Q. Huang, and S. Wan, “An overview of knowledge graph reasoning: key technologies and applications,” Journal of Sensor and Actuator Networks, vol. 11, no. 4, p. 78, 2022.

[60] A. L. Gentile, D. Gruhl, P. Ristoski, and S. Welch, “Personalized knowledge graphs for the pharmaceutical domain,” in The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18. Springer, 2019, pp. 400–417.

[61] M. Kejriwal, “Knowledge graphs: A practical review of the research landscape,” Information, vol. 13, no. 4, p. 161, 2022.

[62] B. Slawski, N. Singh, A. Harris, J. Szank, Sagardigital, K. T. Gubur, S. Rahman, and Amit, “Where will we go with personalized knowledge graphs?” Mar 2022. [Online]. Available: https://www.seobythesea.com/2020/09/personalized-knowledge-graphs/

[63] T. Safavi, C. Belth, L. Faber, D. Mottin, E. Müller, and D. Koutra, “Personalized knowledge graph summarization: From the cloud to your pocket,” in 2019 IEEE International Conference on Data Mining (ICDM). IEEE, 2019, pp. 528–537.

[64] Z. Gong, X. Yu, W. Fu, X. Che, Q. Mao, and X. Zheng, “The construction of knowledge graph for personalized online teaching,” in Data Mining and Big Data: 6th International Conference, DMBD 2021, Guangzhou, China, October 20–22, 2021, Proceedings, Part II 6. Springer, 2021, pp. 98–107.

[65] S. Munir, R. A. S. Malick, S. I. Jami, G. Ahmed, S. Khan, and J. J. Rodrigues, “An integrated approach: using knowledge graph and network analysis for harnessing digital advertisement,” Multimedia Tools and Applications, vol. 82, no. 6, pp. 8883–8898, 2023.

[66] “Youth risk behavior survey-2023,” 2023. [Online]. Available: https://www.cdc.gov/healthyyouth/data/yrbs/index.htm

[67] CDC, “Wisqars explore fatal and nonfatal data,” 2021. [Online]. Available: https://wisqars.cdc.gov/explore/

[68] “National violent death reporting system (nvdrs),” 2023. [Online]. Available: www.cdc.gov/violenceprevention/datasources/nvdrs/index.html

[69] 2023. [Online]. Available: https://fort.fb.com/researcher-platform

[70] “Facebook open research and transparency,” 2019. [Online]. Available: https://fort.fb.com/researcher-apis

[71] “Reddit search,” 2018. [Online]. Available: https://www.reddit.com/dev/api/#GET_search

[72] N. Proferes, N. Jones, S. Gilbert, C. Fiesler, and M. Zimmer, “Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics,” Social Media+ Society, vol. 7, no. 2, p. 20563051211019004, 2021.

[73] A. N. Medvedev, R. Lambiotte, and J.-C. Delvenne, “The anatomy of reddit: An overview of academic research,” Dynamics On and Of Complex Networks III: Machine Learning and Statistical Physics Approaches 10, pp. 183–204, 2019.

[74] T. Rocha-Silva, C. Nogueira, and L. Rodrigues, "Passive data collection on reddit: a practical approach," Research Ethics, p. 17470161231210542, 2023.

[75] "Telegram.org," 2023. [Online]. Available: https://telegram.org/

[76] "Telethon.dev," 2017. [Online]. Available: https://docs.telethon.dev/en/stable/

[77] "Telegram privacy policy," 2022. [Online]. Available: https://telegram.org/privacy?setln=fa

[78] "Telegram- terms of service," 2023. [Online]. Available: https://telegram.org/tos

QALinkPlus: Text Enrichment with Q&A data

Yandong Sun   Yixuan Tang *   Anthony K.H. Tung *
School of Computing, National University of Singapore, Singapore
{yandong,yixuan,atung}@comp.nus.edu.sg

Abstract

Text enrichment, the task of augmenting textual content by incorporating supplementary information to bridge knowledge gaps and enhance reader engagement, is a critical aspect of information retrieval. This study focuses on leveraging question answering datasets, such as Natural Questions and SQuAD, which contain human-validated content from diverse domains as valuable knowledge sources. While QA datasets hold promise for addressing informational needs, existing approaches, like employing dense retrieval for text enrichment, often result in QA pairs that may lack relevance, diversity, or inherent interest. To address these challenges, our paper proposes a novel graph-based method for text enrichment using QA pairs. We construct an entity co-occurrence graph derived from QA datasets and derive context-QA-specific subgraphs. Through rule-based path analysis, we develop an interpretable scoring system to assess the relevance and engagement value of each QA pair. By intelligently re-ranking QA pairs with our scoring system, our method delivers enriched text that fills knowledge gaps and captivates readers, thus improving the overall reading experience. This framework is not only effective in text enrichment tasks, but it also offers advantages for personalization and personal data management.

1 Introduction

As readers navigate vast expanses of textual content, they often come across areas where gaps in their knowledge surface or where they develop a curiosity about related topics. Question and Answer (QA) datasets, with their reservoir of knowledge, have the potential not only to bridge these knowledge gaps but also to enrich the text with related information that readers may find intriguing. Let's consider the novel "Harry Potter and the Philosopher's Stone" for an example. As shown in Fig18 (left), the novel contains various entities. An ideal set of QA pairs for this novel should explore these entities, ensuring that the questions and answers remain intimately connected to the novel's context. What's more, an ideal QA pair should delve deeper than surface-level details. A superficial question such as "Who is the main character in this book?" with the answer "Harry Potter" might emerge. This type of QA pairs effectively evaluates a QA model's comprehension of the book, but for a reader, the provided information, though accurate, might seem superficial, even if they are not deeply familiar with the story.

To effectively leverage QA pairs in augmenting textual content, the work QALink[23] first addressed and formulated the task of text enrichment. It designed a novel system to enhance the reading experience of text documents by automatically integrating relevant QA content from sites like Quora and StackExchange. This system aims to provide readers with supplementary information that aids in understanding and deepening their knowledge of the document's content.

In the development of QALink, a neural network was trained to identify and retrieve relevant QA pairs, a task that has seen significant advancements with the advent of dense retrievers. These modern models are particularly adept at this retrieval task. However, applying dense retrievers directly for text enrichment can sometimes lead to a superficial engagement with the material. This is because they often highlight the most frequently mentioned

[*] *Yixuan Tang and Anthony K.H. Tung are co-corresponding authors

entities, while overlooking the plethora of subtler details that are crucial for a thorough understanding of the text. For example, as shown in the right panel of Fig.18, the introductory paragraph of the Harry Potter Wiki mentions numerous entities. Yet, a RoBERTa model trained on the Natural Questions dataset tends to focus the top 100 QA pairs predominantly around the most prominent entity, 'Harry Potter', at the expense of less prominent ones.

When attempting to leverage a dense retrieval framework [12] directly for text enrichment, three challenges arise:

1. **Lack of Diversity** Dense retrievers have a tendency to favor QA pairs concerning prevalent entities, resulting in a repetitive and predictable selection that overlooks the richness of less common entities and diminishes the breadth of exploration for the reader

2. **Lack of Interestingness** The content surfaced by dense retrievers, while contextually accurate, often lacks the depth and engagement necessary to satisfy readers' curiosity or add meaningful insights to the text. In short, they are not interesting.

3. **Irrelevant QA pairs** Despite their technical proficiency, dense retrievers occasionally present QA pairs that are not closely related to the text. These pairs, while possibly relevant in a broader context, fail to align with the specific themes, characters, or events in the narrative, leading to a disjointed enrichment experience for the reader.

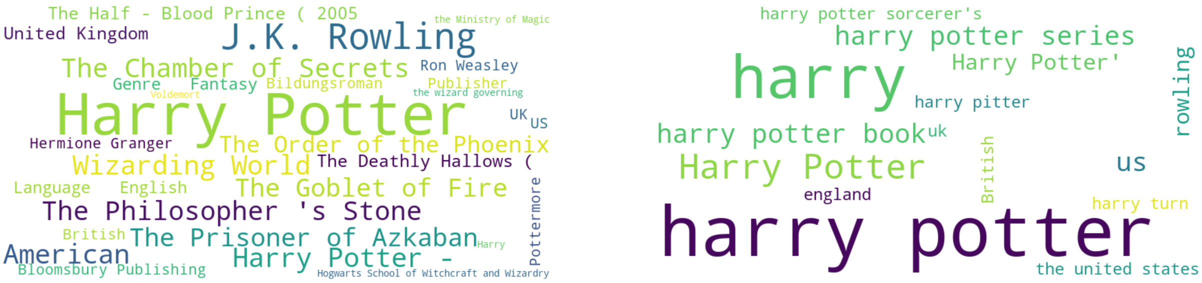


Figure 18: Comparison of the most frequently occurring named entities extracted. (Left) Entities derived from the introductory paragraph of the Harry Potter wiki page; (Right) Entities extracted from the top 100 QA pairs retrieved by querying the Harry Potter wiki content.

Our work confronts the challenge of enhancing text with QA datasets through a unique method. We construct an extensive entity co-occurrence graph from QA datasets, crucial for our text enrichment technique. We map entities from the text and corresponding QA pairs onto this graph to identify relevant subgraphs. Through rule-based path analysis, influenced by the psychological principles of novelty and complexity, we not only develop an interpretable scoring system but also unveil the nuanced connections between context and QA pairs. This approach enriches the text in a nuanced and engaging manner. By refining QA pair selection, our method ensures relevance and diversity while captivating the reader's interest at the same time. This advancement goes beyond traditional models, applying psychological theories of content interestingness in a practical, algorithmic way. Our framework stands out as a system that intertwines psychological concepts with computational applications, enhancing the reading experience by making it more engaging and informative. What's more, we have also discussed how our framework can be used in personal data management and personalization in Sec.4.

The main contributions of this paper can be summarized as follows:

- **Modeling Interestingness Through Psychology Principles:** We propose a model that assesses interestingness by considering both novelty and complexity, drawing inspiration from psychological research. This approach aims to provide a more nuanced understanding of what makes content engaging.

- **Entity Subgraphs for Context and QA Pair Representation:** Our method utilizes entity subgraphs as a simple yet effective way to represent context and question-answer pairs. We then employ path analysis to evaluate the interestingness of these contexts and QA pairs, offering an interpretable and efficient mechanism for analysis.
- **Optimization for Enhanced Contextual Coverage in QA Pairs:** We've formulated an optimization problem aimed at maximizing the coverage of context entities in the QA pairs retrieved and proposed a near-optimal solution that is both practical and efficient. This serves as a re-ranking strategy post-dense retrieval. This approach specifically addresses the challenge of forming robust representations for frequent entities, while also enhancing the distinction of less common ones in text enrichment tasks.

2 Related Work

Interestingness The concept of 'interestingness' in computer science encompasses various captivating objects termed as 'Fun Facts', 'Semantic Novelty', 'Trivia', and 'Unusual Aspects', with many studies highlighting rarity as a key element [7, 16, 17, 25]. However, rarity alone doesn't always equate to interestingness, as some rare objects may simply be unpopular. This notion leads to the broader question of what additional factors make an object interesting. Psychological research offers insights into this, linking interestingness with curiosity, exploration, and information seeking [3, 8, 9, 14, 19, 24]. Current research in computer science, drawing from these psychological models, focuses on novelty and complexity as quantifiable attributes of interestingness [1, 2, 22]. Recognizing this, our study emphasizes complexity alongside rarity in understanding and quantifying interestingness for text enrichment tasks, aiming to enrich textual content with elements that are not just rare but also genuinely engaging and thought-provoking.

Related Text Retrieval This series of research works involves finding relevant information or text passages based on a given query or input text. The variation of it that most related to our work is dense passage retrieval, the goal is to retrieve relevant documents or passages using dense vector representations of the texts. As proposed by [6], and the encoding of candidate answer phrases as vectors for efficient retrieval, as in [21], exemplify the precision and efficiency of dense retrieval. These methods ensure contextually aligned and content-rich text enrichment. Additionally, dense retrieval is adaptable in scenarios lacking direct answers, such as retrieving supporting documents from sources like Wikipedia before answer extraction, as suggested by [4]. In situations without gold standard answers, techniques like global normalization over potential answer spans, as per [5], are invaluable for covering a wide range of possible answers, thereby enhancing the informational breadth of the text. However, the direct application of dense retrievers in text enrichment can lead to a superficial engagement with the content. This occurs because these retrievers tend to focus on the most commonly mentioned entities, thereby overlooking many finer details crucial for a comprehensive understanding of the text. This inclination towards prominent entities over subtler details, as noted in [20], highlights a significant challenge in employing dense retrievers for nuanced text analysis.

Question and Answer Datasets The evolution of the question-answering (QA) domain has been significantly influenced by the development of datasets like SQuAD[18], Natural Questions[11], TriviaQA[10], and NarrativeQA[13]. Initially created as benchmarks for QA systems' reading comprehension, these datasets have become much more than assessment tools. They are now vast repositories of verified knowledge across diverse topics and formats, making them ideal for applications such as text enrichment. In text enrichment, the goal is to deepen the informational and contextual quality of textual content. The varied and real-world questions and answers in these QA datasets offer a unique resource for embedding detailed contextual and factual information into texts, enhancing their informativeness and engagement for users. This re-purposing of QA datasets for

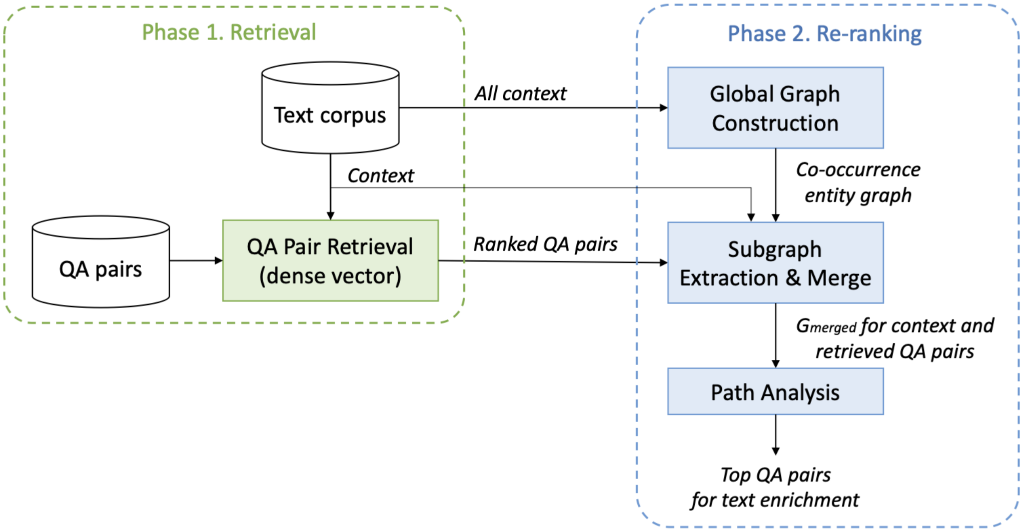

Phase 1. Retrieval
Text corpus
Context
QA pairs
QA Pair Retrieval (dense vector)
All context
Ranked QA pairs
Phase 2. Re-ranking
Global Graph Construction
Co-occurrence entity graph
Subgraph Extraction & Merge
Gmerged for context and retrieved QA pairs
Path Analysis
Top QA pairs for text enrichment

Figure 19: Framework Overview

text enrichment not only extends their use beyond traditional QA but also opens new paths in natural language processing and information retrieval, enhancing the quality and richness of textual content across various domains.

3 Methodology

3.1 Overview

Our framework operates as follows. Initially, we apply Named Entity Recognition (NER) to both a set of contexts (documents) and a set of QA pairs. From the contexts, we extract entities to build an entity co-occurrence graph, linking entities that appear within three sentences of each other. For any given context, we use its entities to identify a corresponding subgraph in this co-occurrence graph. The same process applies to a list of QA candidates. We then merge the subgraphs of each context and its respective QA pair into a unified subgraph. A path analysis is conducted on this united subgraph, using specially designed path patterns that assess novelty and relatedness, yielding an interpretable score. This score is used to re-rank the QA candidates. Additionally, to ensure the diversity of selected QAs, we optimize the QA candidates to maximize the coverage of linked context entities in the QA pairs, with each entity weighted according to its score from the path analysis step. This methodological approach facilitates a nuanced selection of QA pairs that are not only relevant but also diverse, enhancing the overall quality and informativeness of the enriched text. As shown in Fig.19.

3.2 Problem Formulation

Building upon the advancements in dense retrieval, our work introduces a nuanced problem formulation that explicitly captures both the 'interestingness' and 'relatedness' of QA pairs in relation to a given context. This dual consideration aims to enrich the textual landscape with engaging and pertinent information that transcends mere relevance, bringing to light the subtler aspects of human-like engagement and curiosity-driven exploration.

Our proposed problem formulation extends beyond the traditional objective of maximizing relevance. It introduces a composite measure that encompasses both the relatedness of QA pairs to the given context $\mathcal{C}$ and the intrinsic interestingness of the questions q_i and answers a_i in the dataset $\mathcal{D}$. The function $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$ remains central in embedding textual content into a d-dimensional vector space, while the similarity function

$sim : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is now complemented by an interestingness function $int : \mathcal{Q} \times \mathcal{A} \to \mathbb{R}$, which assesses the compelling nature of QA pairs. The combined optimization problem is then:

$$\max_{\phi, int} \sum_{i=1}^{k} (sim(\phi(\mathcal{C}), \phi(q_i, a_i)), int(C, q_i, a_i)) \tag{1}$$

In defining the interestingness of a question-answer pair within a given context, we draw upon two fundamental psychological constructs: novelty and complexity. Novelty (*novelty*) reflects the degree to which information is new or surprising to a user, while complexity (*complexity*) represents the intricacy and depth of the information presented.

Our interestingness function, denoted conceptually as Interestingness$(C, (q_i, a_i))$, thus incorporates both novelty and complexity. This conceptual function is not directly quantifiable but serves as a theoretical framework guiding the development of our computational model:

$$int(C, q_i, a_i) \propto \text{Novelty}(C, q_i, a_i) \oplus \text{Complexity}(C, q_i, a_i) \tag{2}$$

Here, $\oplus$ denotes a conceptual combination of novelty and complexity, the specifics of which will be operationalized in the subsequent computational model. This formulation underscores the importance of both elements in constituting what makes a question-answer pair engaging to the user.

3.3 Proposed Methods

3.3.1 QA-pairs Retrieval

Drawing inspiration from the remarkable success of dense retrievers in passage retrieval tasks, as highlighted in [4], we have adopted this framework to retrieve relevant question-and-answer (QA) pairs for a given context. Specifically, we employ the RoBERTa-base model, which has been fine-tuned on the Natural Questions dataset using a contrastive learning approach. In this approach, positive and negative examples are constructed with a focus on the relationship between context and QA pairs. For a batch size of N, positive examples are N pairs such as $(C_1, q_1, a_1), (C_2, q_2, a_2), \ldots, (C_N, q_N, a_N)$, where each C_i represents a context and q_i, a_i its corresponding question and answer. Negative examples are generated by mismatching the contexts and QAs, pairing different C_i with q_j, a_j where $i \neq j$. The loss function is NT-Xtent as Eq.3, where τ is the temperature parameter that helps to control the scale of the similarity scores. This method enhances the model's ability to discern relevant and informative QA pairs in relation to a given context, leveraging the strengths of the RoBERTa model and the comprehensive nature of the Natural Questions dataset.

$$L = -\log \frac{\exp(\text{sim}(C_i, q_i, a_i)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\text{sim}(C_i, q_k, a_k))/\tau)}, \tag{3}$$

3.3.2 Co-occurance Entity Graph Construction

To assess the novelty and complexity of contexts and QA pairs, our approach begins with the construction of a co-occurrence entity graph that takes the contexts (documents) from a given QA dataset as input. This process starts by dividing a given context into individual sentences. Within each sentence, Named Entity Recognition (NER) is employed to identify entities that will form the nodes of the graph. Edges between these nodes are then established based on their proximity within the text, adhering to a predefined sentence boundary threshold. Specifically, a link is created between two nodes if the sentences containing their respective entities are within a certain number of sentences from each other, as determined by our threshold.

Let C be the input context, $C = \{S_1, S_2, \ldots, S_n\}$ where S_i represents a sentence in the context. We define V as the set of entities extracted through NER, $V = \{e_1, e_2, \ldots, e_m\}$, and $G = (V, L)$ as the resulting graph where

V is the set of nodes corresponding to entities in E, and L is the set of links between nodes. $pos(v)$ denotes the sentence position of the entity corresponding to node v, and θ is the sentence boundary threshold. The adjacency relationship A between nodes v_i and v_j is defined as follows:

$$A(v_i, v_j) = \begin{cases} 1, & \text{if } |\text{pos}(v_i) - \text{pos}(v_j)| \leq \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

In this way, we construct the co-occurrence entity graph to facilitate the extraction of subgraph representations for both context and QA pairs, enabling their further utilization.

3.3.3 Subgraph Representation for Context and QA pair

Given the co-occurrence entity graph G, we further define the subgraph representations for context $\mathcal{C}$ and QA pairs (q_i, a_i) as part of our retrieval framework. In this approach, G is the complete co-occurrence entity graph constructed from the entire text corpus. For a given context $\mathcal{C}$, the subgraph $G_{\text{sub}}(\mathcal{C})$ is a subset of G that represents the context. Similarly, $G_{\text{sub}}(q_i, a_i)$ denotes the subgraph for a QA pair, consisting of a question q_i and an answer a_i, which is also a subset of G. These subgraphs are constructed by identifying sets of entities $E_{\mathcal{C}}$ and E_{q_i,a_i} from the context and the QA pairs respectively, using Named Entity Recognition (NER).

$$G_{\text{sub}}(\mathcal{C}) = G(E_{\mathcal{C}}, \{(v_i, v_j) \mid A(v_i, v_j) = 1,\, v_i, v_j \in E_{\mathcal{C}},\, i \neq j\}), \tag{5}$$

$$G_{\text{sub}}(q_i, a_i) = G(E_{q_i,a_i}, \{(v_i, v_j) \mid A(v_i, v_j) = 1,\, v_i, v_j \in E_{q_i,a_i},\, i \neq j\}) \tag{6}$$

With the individual subgraph representations $G_{\text{sub}}(\mathcal{C})$ and $G_{\text{sub}}(q_i, a_i)$ established, we must now consider how these discrete elements can be synthesized to reflect the complex interplay between the context and the QA pairs. The integration of these subgraphs is pivotal in capturing the nuanced relationships that inform the relevance and interestingness of the QA pairs within their respective contexts. The ensuing step in our methodology, therefore, focuses on merging these subgraphs into a cohesive structure that embodies the full spectrum of informational relationships.

Given the subgraph representations $G_{\text{sub}}(\mathcal{C})$ for the context and $G_{\text{sub}}(q_i, a_i)$ for a QA pair, we merge them into a comprehensive subgraph G_{merged}, which includes paths connecting nodes from $G_{\text{sub}}(\mathcal{C})$ to $G_{\text{sub}}(q_i, a_i)$. We use A^* to represent the transitive closure of the adjacency matrix A we defined in (4). This process forms a unified representation that encapsulates the context, the QA pair, and the semantic links between them.

$$G_{\text{merged}}(G_{sub}(C), G_{sub}(q_i, a_i)) = (V_{\mathcal{C}} \cup V_{q_i,a_i}, L_{\mathcal{C}} \cup L_{q_i,a_i} \cup (v_c, v_q) \mid v_c \in V_{\mathcal{C}}, v_q \in V_{q_i,a_i}, A^*(v_c, v_q) > 0) \tag{7}$$

Utilizing Equations 5, 6, and 7, we derive the subgraph representations for both context and QA pairs. These representations are subsequently employed in our path analysis.

3.3.4 Path Analysis

Based on the subgraph representation of context and QA pairs, we introduce an interpretable algorithm designed for path analysis. The algorithm classifies paths into three distinct categories. It identifies trivial paths where the start and end nodes are the same, reflecting direct, uncomplicated connections, and adjusts their scores using the γ parameter. For paths introducing new or uncommon connections, the algorithm recognizes these as novelty-inducing, incrementing their score based on the presence of nodes outside the typical context and QA sets, a process guided by the α parameter. Additionally, it accounts for hub-influenced paths by detecting nodes with high connectivity, adjusting the score to reflect the influence of these central hubs through the β parameter.

Algorithm 1: Path Analysis

```
Require: $G_{\text{merged}}$, $A^*$, $V_{\mathcal{C}}$, $V_{q,a}$, $\alpha$, $\beta$, $\gamma$, $\theta$
for $v \in V_{\mathcal{C}}$ do
    Initialize $pathScore \leftarrow 0$
    for $u \in V_{q,a}$ do
        if $A^*(v, u) > 0$ then
            for all $p \in \mathcal{P}(v, u)$ do
                Initialize $score_R, score_\alpha, score_\beta \leftarrow 0$
                if $v = u$ then
                    $score_R \leftarrow score_R + \gamma$
                end if
                for $w \in p$ do
                    if $w \notin V_{\mathcal{C}} \cup V_{q,a}$ then
                        $score_\alpha \leftarrow score_\alpha + \alpha$
                    end if
                    if degree$(w, G_{\text{merged}}) \geq \theta$ then
                        $score_\beta \leftarrow score_\beta - \beta$
                    end if
                end for
                $pathScore \leftarrow$
                $pathScore + score_R + score_\alpha + score_\beta$
            end for
        end if
    end for
end for$pathScore$ =0
```

Algorithm 2: Greedy Question-Answer Selection

```
Require: Set of QAs $\mathcal{Q}$, Score function $s(q)$, Context
    entities $C(q)$
Initialize $\mathcal{S} \leftarrow \emptyset$
Initialize TotalScore $\leftarrow 0$
Initialize CoveredEntities $\leftarrow \emptyset$
while $\mathcal{Q} \neq \emptyset$ do
    bestQA $\leftarrow$ null
    bestIncrement $\leftarrow 0$
    for all $qa \in \mathcal{Q}$ do
        newEntities $\leftarrow C(qa) -$ CoveredEntities
        if newEntities $\neq \emptyset$ then
            increment $\leftarrow s(qa) \times \frac{|\text{newEntities}|}{|C(qa)|}$
            if increment $>$ bestIncrement then
                bestIncrement $\leftarrow$ increment
                bestQA $\leftarrow qa$
            end if
        end if
    end for
    if bestQA $=$ null then
        break
    end if
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{$bestQA$\}$
    TotalScore $\leftarrow$ TotalScore $+ s($bestQA$)$
    CoveredEntities $\leftarrow$
    CoveredEntities $\cup C($bestQA$)$
    $\mathcal{Q} \leftarrow \mathcal{Q} - \{$bestQA$\}$ $\mathcal{S}$, TotalScore
end while=0
```

This scoring system, grounded in clear criteria, provides an interpretable and methodical way to analyze the dynamics of paths within our context and QA pair framework.

This algorithm, shown in Algo.1, determines the path score for each context and corresponding QA pair. The derived path score can be directly employed to re-rank QA candidates or further refined through a subgraph optimization step.

3.3.5 Subgraph Optimization

Having demonstrated the path analysis algorithm to evaluate individual context and QA pairs, we now shift our focus to the overarching goal in text enrichment tasks: selecting an optimal set of QA pairs for a given context. This necessitates not only evaluating individual pairs but also ensuring that the chosen set of QA pairs is sufficiently diverse.

To effectively evaluate the question-answer (QA) pairs in relation to a given context, we propose an optimization model aimed at maximizing the coverage of context entities, each weighted by its relevance score. This model is designed to identify the most informative and relevant QA pairs by considering the scores of context entities they relate to.

Let $\mathcal{C}$ be the set of all contexts, $\mathcal{Q}$ be the set of QA pairs, and $\mathcal{C}(qa)$ to denote the context entities covered by the QA pair. The objective is to select a subset of QA pairs $S \subseteq Q$ such that the sum of scores of the covered context entities is maximized.The optimization problem can be formulated as follows:

$$\begin{aligned} \text{Maximize} \quad & \sum_{qa \in S} \sum_{c \in \mathcal{C}(qa)} \text{PathScore}(c, qa) \\ \text{subject to} \quad & S \subseteq \mathcal{Q} \end{aligned} \tag{8}$$

A greedy algorithm is utilized for this optimization. At each step, it selects the QA pair that contributes the highest score to the uncovered context entities in the current set S. This method maximizes the total score at each step, effectively ensuring a diverse coverage of context entities, rather than solely focusing on their individual relevance scores.

This is subject to the condition that $\mathcal{S}$ encompasses a broad range of context entities, as represented by:

$$\text{Coverage}(\mathcal{S}) = \bigcup_{qa \in \mathcal{S}} C(qa) \tag{9}$$

Eq9 is deemed submodular, reflecting the principle of diminishing returns. For any two sets $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{Q}$ and an element $q' \in \mathcal{Q} \setminus \mathcal{B}$, we have:

$$f(\mathcal{A} \cup \{q'\}) - f(\mathcal{A}) \geq f(\mathcal{B} \cup \{q'\}) - f(\mathcal{B}) \tag{10}$$

This condition ensures that the incremental benefit of adding a QA pair to the subset decreases as the subset becomes larger, thereby reflecting the overlap in context entity coverage among the selected QA pairs. This property is crucial for maintaining diversity in the selected subset of QA pairs. This approach is especially advantageous in large-scale problems where precise optimization is not feasible. The algorithm provides a practical and near-optimal solution by balancing the coverage of diverse context entities.

4 QALinkPlus in Personal Document Enrichment and Personalized QA Pairs

The relationship between personal data management and personalization is characterized by a fundamental tension: personal data management focuses on safeguarding user privacy and implementing measures to protect personal information, while personalization relies on accessing and utilizing this data to create tailored experiences for users. Recognizing this, QALinkPlus innovatively addresses this tension. Unlike typical deep learning models that risk privacy leaks by retaining training data details, our approach achieves personalization post-training, particularly during the re-ranking phase, without relying on deep neural networks or requiring the uploading of personal data, thus enhancing privacy. This section delves into its application in two specific areas: personal document enrichment and personalized QA pairs, showcasing the framework's capability to achieve personalization while protecting users' privacy.

Our framework's approach to personal document enrichment stands in stark contrast to traditional deep learning-based personalization methods. Unlike these methods that typically depend on training with large datasets, including sensitive personal documents, and often involve uploading this data to remote servers, our framework employs a graph-based algorithm for personalization during the re-ranking phase, processing and personalizing content within a user-controlled environment, reducing the reliance on training with sensitive data. As a result, our approach could maintain the confidentiality of personal documents, which contributes to personal data management and data personalization in the context of text enrichment.

Furthermore, our framework has the capability to offer personalized QA pairs by adjusting hyperparameters or applying weights to particular entities that align with user interests. This form of personalization caters to various application scenarios, enhancing user engagement and experience. For instance, within the framework, users have the flexibility to modify parameters like the α value to influence the novelty of enrichment contents. This could be particularly useful in different contexts — for example, opting for a larger α when reading for leisure to explore a wider range of topics, or choosing a smaller α when seeking specific information in professional documents. Such customization allows users to tailor their experience to their current needs and preferences, showcasing the adaptability of the framework in providing relevant and personalized content.

5 Experiments

5.1 Entity Diversity and Coverage

In this experiment, we aim to evaluate the effectiveness of our approach in modeling the novelty aspect of interestingness, focusing on the impact of our algorithm on the quality of question-answer (QA) selections. Central to our investigation is the analysis of how the algorithm affects the coverage and diversity of context entities within the selected QAs. By implementing a series of carefully crafted metrics, we seek to quantify the degree to which our method expands the scope of covered entities and enhances the diversity in the QA pairs. This analysis is pivotal in assessing the practical impact of our optimization strategy in real-world QA systems, where the depth and variety of presented information are key to user satisfaction and engagement. Through a comparative evaluation of the original and optimized QA sets, this study aims to highlight the concrete advantages of our algorithm in improving the selection process, particularly in terms of introducing novelty and enriching the content's interestingness.

Dataset In this experiment, we utilized four prominent datasets, randomly sampling 100 documents from each for context. We worked with 316,034 QA pairs across all datasets, leading to the comparison of more than one hundred million query-candidate pairs in total.

- **Natural Questions (NQ)**[15] Curated by Google, this dataset is a cornerstone in open-domain QA research, featuring over 300,000 real user queries paired with relevant Wikipedia articles. NQ emphasizes realistic scenarios, requiring systems to navigate diverse sources for answering user queries.
- **TriviaQA** This dataset presents a realistic text-based QA challenge with 950,000 question-answer pairs from 662,000 documents sourced from Wikipedia and the web. TriviaQA's complexity lies in its long contexts and the need for models to go beyond span prediction for answers.
- **Stanford Question Answering Dataset (SQuAD)**[18] SQuAD includes over 100,000 question-answer pairs based on Wikipedia articles. Its uniqueness stems from the format of answers, which can be any sequence of tokens from the text, and the inclusion of both answerable and unanswerable questions in its latest version.
- **NarrativeQA**[13] NarrativeQA is designed to test reading comprehension on long documents, with a focus on narrative-style content. It includes Wikipedia summaries, links to full stories, and diverse question types, offering a comprehensive challenge for QA models.

Evaluation Metrics In our analysis, we employed three key metrics to evaluate and compare the original and re-ranked question-answer (QA) sets across various datasets.

- **Total Unique Context Entity Coverage** This metric measures the cumulative number of unique context entities covered by the QAs in a dataset. It provides insight into the breadth of information encompassed by the QA pairs, highlighting the algorithm's ability to capture a wide range of relevant topics and details.
- **Average Context Entity Coverage per QA** This metric calculates the average number of unique context entities covered by each individual QA. It offers a more granular view of the coverage, focusing on the depth and richness of information each QA contributes to the overall dataset.

- **Entropy (Entity Coverage Diversity)** Entropy is used to assess the diversity of context entity coverage among the QAs. A higher entropy value indicates a more evenly distributed coverage across different entities, suggesting a greater variety in the types of information addressed by the QAs.

Table 7: Comparison of Original, ReRanked, and Optimized Results Across Datasets. The table showcases metrics including Total Unique Coverage (TUC), Average Coverage per QA (ACPQ), and Entropy. Here, Δ denotes the improvement of optimized results compared to the original and re-ranked results.

Dataset	Metric	Original	ReRanked	Optimized	ΔOriginal (%)	ΔReRanked (%)
NarrativeQA	TUC	2.52	5.25	6.89	+173.66	+31.26
	ACPQ	1.38	2.43	2.87	+107.52	+18.38
	ENTROPY	0.75	1.81	2.26	+199.51	+24.44
Squad	TUC	2.71	4.62	5.63	+108.15	+21.97
	ACPQ	1.38	2.53	2.87	+108.33	+13.05
	ENTROPY	1.08	1.85	2.13	+97.65	+14.59
TriviaQA	TUC	7.65	11.40	14.28	+86.65	+25.26
	ACPA	1.91	2.94	3.56	+86.18	+20.98
	ENTROPY	2.08	2.96	3.36	+61.62	+13.30
Natural Questions	TUC	18.03	17.80	21.55	+19.50	+21.05
	ACPQ	6.14	6.36	7.18	+16.91	+12.98
	ENTROPY	3.42	3.51	3.83	+12.09	+9.11

The results of diversity and coverage comparison, illustrated in the Fig20, distinctly showcase the efficacy of our re-ranking method in enhancing the novelty aspect of interestingness. Across various datasets, the re-ranked sets consistently outperform the original ones in Total Unique Context Entity Coverage and Average Context Entity Coverage per QA, indicating a broader and more diverse range of context entities being addressed. While the Entropy metric remains stable, suggesting a balanced distribution of entities, the overall increase in coverage metrics confirms that our approach successfully selects more novel question-answer pairs, aligning with our objective of providing novel contents in text enrichment tasks.

The comparative analysis of the results pre- and post-subgraph optimization is presented in Table 7. It is evident that the performance significantly improves with the optimized results compared to the original outcomes derived from direct dense retrieval, and it also surpasses the direct reranked results based on our path analysis. This substantiates the effectiveness of our optimization method in procuring a set of QA pairs that is not only diverse but also offers an enhanced collective quality over focusing solely on individual pairs.

5.2 Assessing the Impact of Re-ranking on Superficial QA Reduction

This experiment aims to assess our approach's effectiveness in capturing the complexity aspect of interestingness. Utilizing Algo.1, we assign scores to QA pairs based on criteria from path analysis, then re-rank QAs initially retrieved by a standard dense retrieval framework. We hypothesize that our approach, which enhances complexity and novelty, will reduce the prevalence of superficial QAs in the top results. This hypothesis will be tested by comparing the superficial QA distribution in both the original and re-ranked lists.

We employ the Natural Questions dataset [15] for its categorization of short-answer questions, often considered superficial compared to the complex and deep 'interestingness' QAs our research targets. Our evaluation metric is the percentage of superficial QAs in the top n results, where a QA is labeled superficial if it's answerable with a short response, which is usually one or two words.

The results, depicted in Fig.21, demonstrate a notable decrease, more than 10% at the beginning, in the prevalence of superficial QA pairs within the re-ranked results, with the re-ranking line consistently positioned below that of the original. This trend indicates that our path analysis algorithm has effectively prioritized QAs of greater complexity. Initially, the original results exhibit a high percentage of superficial QAs, aligning with findings from [20]. The tendency to favor frequently occurring entities often results in the initial selection of less complex QA pairs. As n increases, the percentage of superficial QAs in the re-ranked results gradually rises, while it decreases in the original dense retrieval results. This pattern can be attributed to the limited availability of non-superficial QAs in certain contexts, exemplified by the 'Harry Potter' case discussed in Sec.1.

6 Conclusion

In this study, we tackled the challenge of enhancing text enrichment using QA datasets like Natural Questions and SQuAD through an innovative approach involving the creation of an extensive entity co-occurrence graph, from which context-specific subgraphs were derived. This led to a rule-based path analysis and a novel scoring system, assessing each QA pair's relevance and engagement value, thus enriching the reader's experience. Additionally, our framework discusses aspects of personal data management and personalization, suggesting ways to align personalized content with individual privacy needs. Our methodology's effectiveness was evident in two key experiments: Experiment 5.1 highlighted our re-ranking method's ability to enhance novelty, as shown by improved coverage metrics, and Experiment 5.2 demonstrated a significant reduction in superficial QAs, emphasizing the prioritization of more complex, contextually relevant content. These results underscore our approach's potential to fill knowledge gaps and captivate readers, marking a significant step forward in text enrichment using QA datasets.

7 Acknowledgments

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative and the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and the Ministry of Education, Singapore.

References

[1] Daniel E Berlyne. Conflict, arousal, and curiosity. 1960.

[2] Daniel E Berlyne. Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation. Hemisphere, 1974.

[3] Frederick E Bolton. Attention and interest: A study in psychology and education. The Journal of Philosophy, Psychology and Scientific Methods, 7(17):474–475, 1910.

[4] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In ACL (1), pages 1870–1879. Association for Computational Linguistics, 2017.

[5] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In ACL (1), pages 845–855. Association for Computational Linguistics, 2018.

[6] Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. Multi-step retriever-reader interaction for scalable open-domain question answering. In ICLR (Poster). OpenReview.net, 2019.

[7] Nausheen Fatma, Manoj Kumar Chinnakotla, and Manish Shrivastava. The unusual suspects: Deep learning based mining of interesting entity trivia from knowledge graphs. In AAAI, pages 1107–1113. AAAI Press, 2017.

[8] Barbara L Fredrickson. What good are positive emotions? Review of general psychology, 2(3):300–319, 1998.

[9] Carroll E Izard and Brian P Ackerman. Motivational, organizational, and regulatory functions of discrete emotions. Handbook of emotions, 2:253–264, 2000.

[10] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In ACL (1), pages 1601–1611. Association for Computational Linguistics, 2017.

[11] Hidetaka Kamigaito, Jingun Kwon, Young-In Song, and Manabu Okumura. A new surprise measure for extracting interesting relationships between persons. In EACL (System Demonstrations), pages 231–237. Association for Computational Linguistics, 2021.

[12] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In EMNLP (1), pages 6769–6781. Association for Computational Linguistics, 2020.

[13] Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. Trans. Assoc. Comput. Linguistics, 6:317–328, 2018.

[14] Andreas Krapp. Interest, motivation and learning: An educational-psychological perspective. European journal of psychology of education, 14:23–40, 1999.

[15] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. Trans. Assoc. Comput. Linguistics, 7:452–466, 2019.

[16] Nianzu Ma, Alexander Politowicz, Sahisnu Mazumder, Jiahua Chen, Bing Liu, Eric Robertson, and Scott Grigsby. Semantic novelty detection in natural language descriptions. In EMNLP (1), pages 866–882. Association for Computational Linguistics, 2021.

[17] Abhay Prakash, Manoj Kumar Chinnakotla, Dhaval Patel, and Puneet Garg. Did you know? - mining interesting trivia for entities from wikipedia. In IJCAI, pages 3164–3170. AAAI Press, 2015.

[18] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In EMNLP, pages 2383–2392. The Association for Computational Linguistics, 2016.

[19] Gregory Schraw and Stephen Lehman. Situational interest: A review of the literature and directions for future research. Educational psychology review, 13:23–52, 2001.

[20] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. arXiv preprint arXiv:2109.08535, 2021.

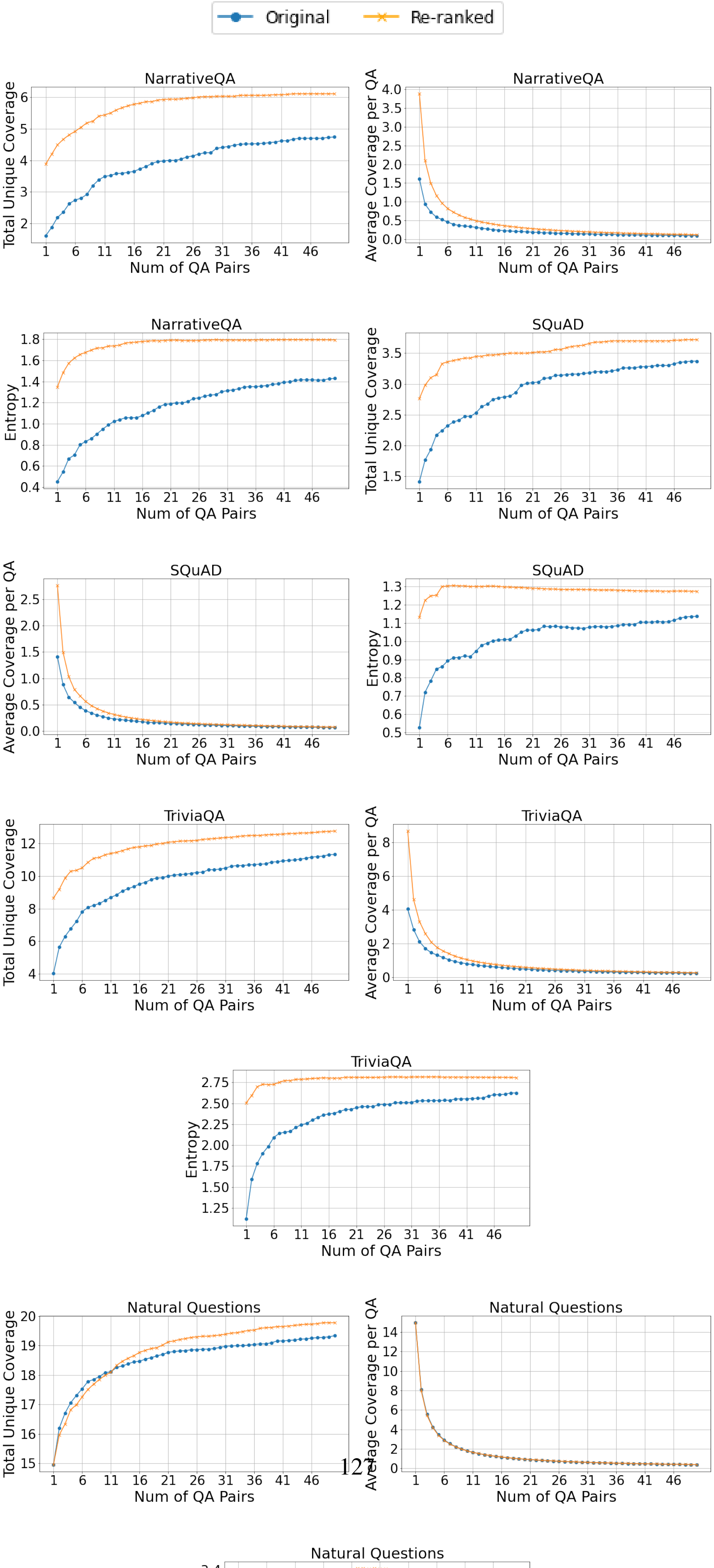
[21] Min Joon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur P. Parikh, Ali Farhadi, and Hannaneh Hajishirzi. Real-time open-domain question answering with dense-sparse phrase index. CoRR, abs/1906.05807, 2019.

[22] Paul J Silvia. What is interesting? exploring the appraisal structure of interest. Emotion, 5(1):89, 2005.

[23] Yixuan Tang, Weilong Huang, Qi Liu, Anthony K. H. Tung, Xiaoli Wang, Jisong Yang, and Beibei Zhang. Qalink: Enriching text documents with relevant q&a site contents. In CIKM, pages 1359–1368. ACM, 2017.

[24] Silvan Tomkins. Affect imagery consciousness: Volume I: The positive affects. Springer publishing company, 1962.

[25] David Tsurel, Dan Pelleg, Ido Guy, and Dafna Shahaf. Fun facts: Automatic trivia fact extraction from wikipedia. In WSDM, pages 345–354. ACM, 2017.


Original
Re-ranked
NarrativeQA
Total Unique Coverage
2
3
4
5
6
Num of QA Pairs
1
6
11
16
21
26
31
36
41
46
NarrativeQA
Average Coverage per QA
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Num of QA Pairs
NarrativeQA
Entropy
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
Num of QA Pairs
SQuAD
Total Unique Coverage
1.5
2.0
2.5
3.0
3.5
Num of QA Pairs
SQuAD
Average Coverage per QA
0.0
0.5
1.0
1.5
2.0
2.5
Num of QA Pairs
SQuAD
Entropy
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
Num of QA Pairs
TriviaQA
Total Unique Coverage
4
6
8
10
12
Num of QA Pairs
TriviaQA
Average Coverage per QA
0
2
4
6
8
Num of QA Pairs
TriviaQA
Entropy
1.25
1.50
1.75
2.00
2.25
2.50
2.75
Num of QA Pairs
Natural Questions
Total Unique Coverage
15
16
17
18
19
20
Num of QA Pairs
Natural Questions
Average Coverage per QA
0
2
4
6
8
10
12
14
Num of QA Pairs
Natural Questions

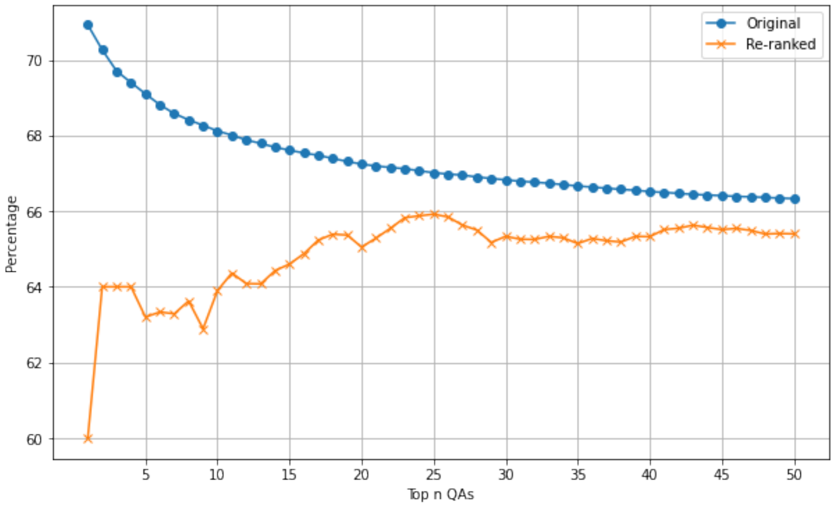


Figure 21: Percentage of Superficial QAs in Natural Questions

Data Engineering

It's FREE to join!

TCDE

tab.computer.org/tcde/

The Technical Committee on Data Engineering (TCDE) of the IEEE Computer Society is concerned with the role of data in the design, development, management and utilization of information systems.

- Data Management Systems and Modern Hardware/Software Platforms
- Data Models, Data Integration, Semantics and Data Quality
- Spatial, Temporal, Graph, Scientific, Statistical and Multimedia Databases
- Data Mining, Data Warehousing, and OLAP
- Big Data, Streams and Clouds
- Information Management, Distribution, Mobility, and the WWW
- Data Security, Privacy and Trust
- Performance, Experiments, and Analysis of Data Systems

The TCDE sponsors the International Conference on Data Engineering (ICDE). It publishes a quarterly newsletter, the Data Engineering Bulletin. If you are a member of the IEEE Computer Society, you may join the TCDE and receive copies of the Data Engineering Bulletin without cost. There are approximately 1000 members of the TCDE.

Join TCDE via Online or Fax

ONLINE: Follow the instructions on this page:

www.computer.org/portal/web/tandc/joinatc

FAX: Complete your details and fax this form to **+61-7-3365 3248**

Name ______________________

IEEE Member # ______________________

Mailing Address ______________________

____

Country ______________________

Email ______________________

Phone ______________________

TCDE Mailing List

TCDE will occasionally email announcements, and other opportunities available for members. This mailing list will be used only for this purpose.

Membership Questions?

Xiaoyong Du
Key Laboratory of Data Engineering
and Knowledge Engineering
Renmin University of China
Beijing 100872, China
duyong@ruc.edu.cn

TCDE Chair

Xiaofang Zhou
School of Information Technology and
Electrical Engineering
The University of Queensland
Brisbane, QLD 4072, Australia
zxf@uq.edu.au

IEEE Computer Society
10662 Los Vaqueros Circle
Los Alamitos, CA 90720-1314