Letter from the Special Issue Editor

Large Language Models (LLMs) are increasingly being integrated into databases, data analytics, and emerging data systems. These models are capable of infusing human-level expertise into every stage of data processing, from data discovery to predicting user intent. This issue explores cutting-edge research on the current and future applications of LLMs across various data systems.

LLMs are becoming integral to modern databases. Beginning with the critical tasks of data discovery and integration, the paper *Large Language Models for Data Discovery and Integration: Challenges and Opportunities* by Juliana Freire et al. offers a comprehensive analysis of how LLMs can aid in these tasks while highlighting ongoing challenges. LLMs can help decode the semantics of datasets, making them more accessible and usable. Next, LLMs can enhance query processing. In *LLMs and Databases: A Synergistic Approach to Data Utilization*, Fatma Özcan et al. describe how LLMs leverage their vast, pre-trained knowledge to effectively translate natural language into SQL queries. The authors also develop pre-trained cardinality estimation models and foundation database models inspired by LLMs to solve database performance problems. Finally, SQL programming itself also benefits from LLMs. The paper *Customizing Operator Implementations for SQL Processing via Large Language Models* by Immanuel Trummer introduces the GenesisDB system, which uses LLMs to generate relational operator code for SQL queries. It is even possible to generate code required for running benchmarks like TPC-H.

Beyond databases, LLMs are being utilized in data analytics and next-generation data systems. The paper *iDataLake: An LLM-Powered Analytics System on Data Lakes* by Jiayi Wang et al. presents the iDataLake system, which automates analytics on multi-modal data lakes using LLMs, leveraging their semantic understanding to provide a comprehensive and efficient solution. Data engineering workflows also benefit from LLM integration. The paper *Large Language Models as Pretrained Data Engineers: Techniques and Opportunities* by Yin Lin et al. introduces the UniDM system, which embeds LLMs into key stages of the data engineering process, including data wrangling, analytical querying, and table augmentation for machine learning. LLMs can even enhance the proactivity of data systems. The paper *LLM-Powered Proactive Data Systems* by Sepanta Zeighami et al. demonstrates how LLMs can anticipate user intent, perform transformation operations, and adapt both structured and unstructured data to align with user needs. Lastly, the paper *Top Ten Challenges Towards Agentic Neural Graph Databases* by Jiaxin Bai et al. explores Agentic Neural Graph Databases (NGDBs), which extend neuralization processes to automate data management. The paper discusses the challenges of refining the interfaces, learning, inference, and system components of traditional NGDBs.

Together, these studies represent the frontier of LLM integration into both established and emerging data systems. LLMs are clearly driving impactful advancements in data systems and will continue to shape their evolution in the future. We would like to thank all the authors for their valuable contributions. We also thank Haixun Wang for the opportunity to put together this special issue, and Jieming Shi for his help in its publication.

Steven Euijong Whang
Korea Advanced Institute of Science and Technology