Letter from the Editor-in-Chief

Across industries – healthcare, law, finance, government, and beyond – vast troves of critical information remain locked inside unstructured documents. Hospitals, insurers, and public-health agencies, for example, are awash in paperwork yet struggle to answer even the simplest questions: How many hip replacements did we reimburse last year? Which clinics drove the sharp rise in imaging cost? These queries sound like ordinary SQL, but the source material is a welter of discharge notes, scanned claim forms, physician dictations, and radiology narratives that know no tables, keys, or datatypes.

For a time, Retrieval-Augmented Generation or RAG seemed like a salvation: extract a handful of relevant pages, feed them to a large language model, and let it produce answers. This method excels at generating summaries, but it falters when accuracy, completeness, and traceability are critical. In analytical contexts, we need to aggregate information with precision. Aggregation cannot tolerate overlooked notes, duplicated charges, or fabricated procedure codes. To move beyond these limitations, we must look deeper than surface-level retrieval and summarization. Within every clinical document lies an implicit structure: a latent schema that organizes the apparent chaos of narrative text. Diagnoses, drugs, quantities, prices, insurance plan codes, and many other details are embedded in free-form language, each following recurring patterns. The challenge is to automatically surface this hidden schema, promote each relevant fragment to a well-typed attribute, and keep the catalogue up to date as templates evolve, organizations merge, or reimbursement rules shift.

Once these attributes are identified, they must be connected to broader systems. For example, provider identifiers need to be cross-referenced with national registries to determine specialty and location. Drug names must be matched against external drug databases to obtain standardized codes and strengths. Procedure terms require mapping to official vocabularies to ensure that records from different sources can be meaningfully joined. Only by linking the discovered schema to external reference databases can we create a coherent, queryable view that spans institutions, payers, and time. Even with today's best language models, the journey from narrative text to structured, meaningful data is fraught with challenges. It is not simply about extracting numbers; it is about uncovering the underlying structure and semantics within unstructured language, transforming free-form narratives into well-typed fields and relationships. Only then can we aggregate, analyze, and ultimately derive actionable insights from the data. The sheer volume of information is a major hurdle, as billions of tokens must be processed to ensure that answers are both complete and reliable. The inferred schema is constantly shifting as forms evolve, new abbreviations appear, and vocabularies are updated. Record linkage remains fragile, with even minor variations in provider names disrupting connections. Economic constraints add further complexity, making it impractical to process every document with the largest models on a nightly basis. To address these issues, we need intelligent filtering, cascaded models, and planning systems that reserve costly inference for the most ambiguous or critical fragments, ensuring that the path from narrative to structure, through aggregation and analysis, leads to trustworthy insights.

The papers in this issue converge on those very gaps. One team describes a full software stack in which conversational agents design pipelines, cost-based optimisers decide which model cascade to run, and a serving layer keeps GPU memory from thrashing. Another argues that data discovery is as much social as technical—human analysts must express what they seek before any system can find it, and that expression itself hints at the latent schema waiting in the text. We meet databases that embed small language models in their storage engines, skimming over irrelevant paragraphs to slash I/O; a model-context protocol that lets billing notes, clinical narratives, and national registries live behind their own specialised servers, coordinated at query time; join-aware indices that pre-compute cross-database linkages so the model never attempts a billion-row match at runtime; and log-augmented inference that remembers last week's successful entity mappings instead of rediscovering them tomorrow. Individually these contributions advance extraction, schema induction, external linkage, storage, optimisation, and serving; together they sketch a railway from raw narrative to verifiable totals.

Yet challenges remain. We still lack benchmarks that score faithful counting as rigorously as they score fluent summaries, and open-source ontologies that expose every revision of every code set. Privacy-preserving hosts are needed so models can train on protected health information without exfiltrating it. Above all, schema discovery and external linkage must be treated as first-class steps in the analytics stack. If we mis-type a field or mis-join a record, no brilliance in downstream analytics will salvage the insight. Should we succeed, the next time a policymaker asks, "What did we spend on lung-cancer imaging last quarter?" the answer will arrive in seconds, backed by measurable recall, transparent lineage, and a reproducible chain from prose to number. I invite you to read the pages that follow, challenge the ideas, and help turn narrative chaos into structured, dependable insight.

Haixun Wang
EvenUp