Retrieval-augmented Generation (RAG): What is There for Data Management Researchers?

A discussion on research from a panel at LLM+Vector Data Workshop @ IEEE ICDE 2025

Arijit Khan
AAU

Yuyu Luo
HKUST (GZ)

Wenjie Zhang
UNSW

Minqi Zhou
Huawei

Xiaofang Zhou
HKUST

1. INTRODUCTION

Large language models (LLMs) enable the state-of-the-art in language processing by framing diverse tasks–from code synthesis and healthcare to finance, digital assistance, and scientific discovery–as next-token prediction problems [38, 53, 60, 65, 72, 20, 68, 76, 32]. In addition, LLMs enable automation in data science and engineering, optimizing processes such as data analysis, manipulation, querying, interpretation, research, and education [33, 7, 8, 22, 24, 42, 77, 37, 43, 44, 66, 49].

LLMs encode probabilistic token patterns instead of maintaining explicit knowledge structures, which (1) constrains multi-step reasoning under the next-token prediction paradigm; (2) ties outputs to static, pre-cutoff training data–undermining performance on evolving knowledge tasks; and (3) lacks a built-in factual verification mechanism, resulting in hallucinations [25].

Retrieval-augmented generation (RAG) enhances LLM outputs by integrating dynamic, authoritative external knowledge sources rather than relying solely on static pre-training data [29, 74, 56]. Advanced embedding models convert heterogeneous datasets, e.g., text, multimedia, graphs, and tables into high-dimensional vectors that preserve semantic similarity, storing them in vector databases such as Weaviate, Chroma, FAISS, Milvus, Pinecone, Qdrant, and Vespa [46]. At query time, RAG performs a vector similarity search to retrieve semantically relevant information, appends this context to the LLM prompt for in-context learning, and thus mitigates hallucinations while boosting accuracy, transparency, and capability without costly retraining or fine-tuning.

The retrieval phase in RAG mirrors classical information retrieval–queries return documents based on relevance metrics [48]–but it uniquely injects retrieved content into the generative pipeline for LLM in-context learning, yielding responses that are both contextually coherent and precisely tailored. This approach raises key technical questions: Why vector similarity search via approximate nearest neighbor (ANN) methods performs effectively in high-dimensional spaces [23]; which innovations in indexing and quantization mitigate the curse of dimensionality for dense vectors [26]; and to what extent dataset characteristics, embedding models, or relevance metrics drive retrieval performance. While data management has long addressed high-dimensional time series, spatial, multimedia, and geometric data, the recent surge in vector workloads has catalyzed novel indexing and search algorithms, specialized vector databases, and hardware accelerators [58]. These advances power enterprise RAG deployments–Microsoft Azure AI [6], DoorDash delivery [19], Pinterest's Text-to-SQL [45], and LinkedIn customer service [67]–promising to transform search technologies [36].

Nevertheless, RAG systems remain brittle: Partitioning large external knowledge bases into vectorized chunks dilutes critical details, severs global interrelationships needed for multi-hop queries, and depends on embedding similarity that often retrieves contextually irrelevant content. Consequently, conventional RAG underperforms on queries demanding synthesized, comprehensive insights, driving the emergence of variants, e.g., GraphRAG [21], multimodal RAG [64], Agentic RAG [52], knowledge-augmented generation (KAG) [35], tool-augmented generation [51], table-augmented generation (TAG) [10], and cache-augmented generation (CAG) [16].

At the LLM+Vector Data'25 workshop (ICDE 2025, Hong Kong, China), the focus was on data management challenges and opportunities arising from LLM-vector data interactions [27]. A panel featuring Yuyu Luo (Hong Kong University of Science and Technology, Guangzhou), Wenjie Zhang (University of New South Wales), Minqi Zhou (Huawei), and Xiaofang Zhou (Hong Kong University of Science and Technology), moderated by Arijit Khan (Aalborg University), discussed emerging RAG opportunities at the intersection of data science, data engineering, and data-centric AI/ML methodologies. The session attracted over 100 attendees.

2. OVERVIEW OF PANEL DISCUSSION

The discussion was organized into three broad themes.

Theme 1: Future prospects of RAG, vector data management, and LLMs for data management

Background. Rapid advancements in large language models and their integration with external data systems are transforming the computational landscape [18, 57, 17, 59, 47, 71]. With models like Gemini 1.5 Pro boasting a 2-million token context window [28], questions arise regarding the necessity of traditional RAG architectures when semantic search can be performed by long-context LLMs. Meanwhile, the exponential growth and complexity of vectorized data present new challenges in data management, driving the need for advanced solutions that efficiently store, retrieve, and update heterogeneous data. These trends raise several critical questions about the future direction of the field.

Q1. *With long-context LLMs like Gemini 1.5 Pro–having its impressive 2-million token context window–emerging, how do you foresee RAG architectures evolving to fully exploit these extended capabilities? What are the most significant recent innovations and future trends in RAG that excite you the most?*

Yuyu Luo. Long-context LLMs with million-token context windows create new opportunities for RAG. Instead of segmenting external knowledge into small, isolated chunks, we can retrieve larger, more semantically coherent units. Future RAG pipelines may intelligently determine how much retrieved context to include in the prompt. For example, with the availability of extensive context windows in models such as Gemini 1.5 Pro, entire documents or interconnected knowledge segments could be retrieved and integrated, thus preserving global semantic relationships that previous approaches might have disrupted. Moreover, retrieval strategies could dynamically adapt context selection to efficiently utilize the available context window, ensuring optimal model performance without excessive computational overhead.

However, these expanded context capabilities introduce significant efficiency challenges, particularly around memory management. This opens exciting research opportunities for database researchers. Database-inspired techniques, such as KV caching mechanisms, could optimize the reuse of previously computed representations and mitigate the overhead of repeatedly processing similar context segments [34, 61].

Wenjie Zhang. RAG architectures are evolving to take advantage of the capabilities of long-context LLMs through techniques such as fine-grained retrieval, iterative reasoning, and dynamic compute allocation. Fine-grained proposition retrieval enhances information density by reducing noise, ensuring that only highly relevant content is used by the LLM. Iterative approaches like IterDRAG [70] optimize multi-hop reasoning by refining retrieval and generation cycles. Dynamic compute allocation models intelligently distribute resources, balancing cost and efficiency for large-scale RAG deployments. Hard negative fine-tuning further strengthens robustness by addressing challenging retrieval scenarios, improving accuracy in diverse domains. Emerging trends include multimodal RAG, which integrates text, images, and audio for richer responses, interactive systems that allow real-time user feedback, and integration of knowledge graphs for enhanced structured contexts. They facilitate precise and efficient inference in long-context settings, significantly boosting performance on knowledge-intensive tasks.

Minqi Zhou. The current long-context LLM capability, like Gemini 1.5 Pro, does create a set of new opportunities to enlarge the external information that can be used. Meanwhile, our findings underscore the urgency of refining new RAG techniques for practical deployment. (1) Integrate the document semantics together with the document structure: In the document/mail QA system, it is easier to get the correct answer when taking the total substructure of the document/mail into account. (2) Understand the versions between the document on the same topic: While building systems or projects, it is important to understand the concept evolution in the different versions of the system/project architecture or system design documents. (3) Adhere to the original text: For example, it is a must to return the original text from the issuance of government documents and the issuance of legal documents in the corresponding QA systems. As for the sophisticated RAG systems in practice, they do require different innovations to provide the correct answers (i.e., “no one-size-fits-all”).

Xiaofang Zhou. While long-context LLMs showed exciting possibilities, they also introduce new challenges. First, handling such large inputs can significantly increase the time to first token (TTFT), which directly impacts user experience if it becomes too slow. Second, these models still rely heavily on GPU resources, which are considerably more expensive than CPU-based solutions. In my view, this means that CPU and memory-efficient RAG systems still hold a strong advantage when it comes to retrieval efficiency and cost-effective use of computing resources. Rather than seeing RAG and long-context LLMs as competing approaches, I believe they are complementary. Combining them effectively could lead to more powerful and efficient systems. One particularly exciting research direction is how to design optimal system architectures that intelligently integrate RAG with LLMs–leveraging the strengths of both. This hybrid approach has great potential to shape the future of retrieval-augmented generation.

Q2. *What emerging issues are we seeing in vector data management, and how can advanced data management solutions mitigate these complexities?*

Wenjie Zhang. High-dimensional embeddings, such as those with thousands of dimensions for text, images, or videos, drastically increase storage and computational costs for vector data management. Sparsity in these high-dimensional spaces reduces the efficiency of similarity searches. Recent research on approximate nearest-neighbor search [75] uses hybrid approaches that combine locality-sensitive hashing (LSH)'s rapid coarse filtering with the efficient exploration performance of approximate proximity graph (APG) for large-scale datasets, reducing storage needs and improving query efficiency to support scalable RAG and LLM deployments.

Minqi Zhou. In term of the industry deployment, we find several issues still need to be fixed, especially for the vector index. (1) Vector indexes must support continuous, incremental updates without degrading top-k recall, precision, or query latency–yet existing structures (e.g., IVF, HNSW, HCNNG, DiskANN) exhibit performance drops even on simple insertions. (2) We must also reduce index-building times, which currently stretch into hours once vector counts exceed one billion.

Xiaofang Zhou. Vector data management today faces a wave of interrelated challenges. As models generate billions or even trillions of high-dimensional embeddings, systems must scale accordingly while maintaining efficiency. Storing dense vectors at this scale puts enormous pressure on memory, demanding techniques like ultra-low-bit quantization and intelligent tiering. For instance, frequently accessed ("hot") vectors can be kept in fast memory tiers, while infrequently used ("cold") data can be offloaded to cost-effective storage such as NVMe or cloud object stores. Beyond traditional nearest neighbor search, modern workloads increasingly demand support for set-level queries, hybrid retrieval that combines semantic similarity with symbolic filtering, and operations over multimodal or multilingual embeddings. These introduce significant algorithmic and system-level complexity. Meanwhile, most existing infrastructures are ill-equipped to handle continuous, low-latency updates or support rapid incremental inserts without costly index rebuilds. In my view, addressing these challenges requires more than marginal improvements in algorithms. It calls for end-to-end system design. Real progress will come from holistic hardware-software co-design, where approximate retrieval, adaptive indexing, and tier-based resource management are seamlessly integrated into scalable, production-ready systems.

Q3. *What specific roles will RAG–and broadly, LLMs and generative AI–play in advancing database research and shaping the next-generation DB systems?*

Yuyu Luo. LLMs and RAG are poised to become integral parts of future database systems. A natural entry point is the extensively studied Text-to-SQL task [30, 31], where an LLM translates user questions into executable queries. RAG enhances this pipeline by dynamically retrieving relevant database documentation, such as schema definitions, data dictionaries, and data semantics, to generate accurate SQL queries even for unfamiliar databases [40, 78, 41]. Moving beyond Text-to-SQL, RAG fits naturally here by pulling in relevant data from the database (or other sources) as needed for the LLM to produce accurate answers. This could transform how we do data analytics, making it more accessible – imagine getting a narrative report of your data, compiled via SQL queries and explanatory text from an LLM [50, 69].

Wenjie Zhang. LLMs can synthesize data to fill gaps in databases or expand datasets; and RAG, by retrieving relevant knowledge, ensures that the generated data align with real world scenarios and requirements, thus improving data quality and usability. Additionally, LLMs can analyze users' query patterns and automatically generate optimized execution plans; and RAG can retrieve historical query logs to provide additional insights for optimization, further enhancing query efficiency.

Minqi Zhuo. RAG, LLM, and generative AI will enlarge the database research scope. In the traditional enterprise data warehouse system field, RAG and LLM are able to accelerate the ETL phase for adding new data sources or creating new data marts and to enhance the quality of the data integration from isolated data/database sources inside the enterprise.

Xiaofang Zhou. LLMs are reshaping the role of DB systems–from passive stores of structured data to active participants in semantic retrieval, reasoning, and generation. As users increasingly pose natural-language queries over heterogeneous data–text, tables, images–natural language becomes the new query interface, embedding based retrieval emerges as a core operator, and structured and unstructured data are increasingly fused within a unified retrieval layer. This paradigm shift calls for systems that blend the precision and consistency of structured storage with the flexibility and recall of semantic retrieval. Researchers could begin by developing compound query planners that coordinate exact lookups and approximate searches within a single framework, along with storage engines that natively support both relational and vector data. New benchmarks and evaluation metrics are needed to reflect the multimodal, generative, and user-facing nature of next-generation DB systems. Meanwhile, it's worth keeping in mind a broader question: Next-generation database systems may forgo full human interpretability in favor of machine-driven adaptation, optimization, and reasoning.

Theme 2: Deep dive into the technical frontiers of RAG and synergy with data management

Background. Recent breakthroughs in approximate nearest neighbor search (ANNs) and scalable vector DBs are revolutionizing the indexing and retrieval of large-scale, high-dimensional, dense vectors, resulting in faster and more precise information synthesis. Concurrently, the development of specialized RAG variants–such as GraphRAG, KG-RAG, Agentic RAG, and multimodal RAG–demonstrates the capacity to enhance LLMs by integrating domain-specific structures and multi-modal inputs. This dynamic landscape warrants technical investigation into merging these innovations into cohesive, high-performance systems.

Q4. *What emerging research directions in approximate nearest neighbor search and vector databases hold the most promise, and how can these developments integrate with modern data management strategies?*

Wenjie Zhang. Emerging research in ANN search and vector databases includes advancing toward scalable, multimodal, and real-time solutions that integrate seamlessly with modern data management tasks in the generative AI era. Key directions include learned index structures for intelligent retrieval, dynamic ANN methods for streaming data, knowledge-aware ANN supports to reason over graph and LLM-augmented systems, multimodal search across text, images, and graphs. These developments are transforming vector search into a core component of next generation data platforms.

Minqi Zhou. In the age of agentic AI, a lot of agents will be created on the devices side, such as phone, Pad, PC, etc., where vector databases would play a critical role in enhancing the semantic search capability. On such devices, how to reduce the energy consumption when building vector index, and how to lower the storage capacity of vector index without sacrificing the top-k search recall/precision, are crucial research directions.

Xiaofang Zhou. In my view, there are many interesting topics, such as Maximum Inner Product Search (MIPS), dynamic indexes that let you add or remove vectors on the fly, and new quantization methods that dramatically cut memory without killing accuracy.

Q5. *With the RAG landscape rapidly diversifying into forms like GraphRAG, KG-RAG, Agentic RAG, multimodal RAG, and beyond, which features do you find most compelling? What emerging frameworks do you envision that can harmonize their unique capabilities?*

Yuyu Luo. Each RAG variant brings something unique to the table–several of their features stand out to me. GraphRAG and KG-RAG are compelling because they retain relationships and structure–rather than retrieving isolated text passages, they leverage connections in a graph or a knowledge base. This means answers can follow a chain of facts or traverse a hierarchy, which is powerful for complex reasoning. Agentic RAG is exciting for giving the LLM more autonomy: The model can iteratively decide what to retrieve or which tool to use next, almost like it's doing its own research. And of course, multimodal RAG extends capabilities beyond text–a system that fetches relevant images or audio along with text can provide much richer responses. To get the best of all worlds, it is necessary to develop unified frameworks capable of integrating multiple specialized retrieval modules [73]. A promising direction involves creating an orchestrator that modularizes various knowledge sources and external tools into a cohesive pipeline, e.g., extracting structured facts from knowledge graphs, retrieving detailed context from vector-based indexes, or invoking computational APIs as needed.

Wenjie Zhang. Graph-based variants, such as Graph RAG and KG-RAG, appeal to me most, because a graph view makes hidden links explicit. In classic text only RAG, the LLM struggles to track cross sentence jumps or connect evidence spread over multiple documents. Adding nodes for entities and edges for their relations turns those implicit jumps into a visible path, so the model can follow a multi hop chain, keep the reasoning transparent, and answer complex questions with higher accuracy [54]. Agentic RAG then lets the model decide which extra tools to invoke, such as search, calculation, and coder, whenever the retrieved context is not yet enough. This keeps the chain-of-thought short and grounded, and it brings the workflow closer to real applications that mix retrieval, analysis, and generation. A complete retrieval layer, however, needs to look beyond text. Human knowledge arrives as images, video, audio, and sensor streams, not just prose. Multimodal RAG aims to meet that reality. If a system can fetch a diagram, a short clip, and a passage of text about the same topic, then combine them through a common representation, its answers become richer and better grounded. Looking forward, I expect a unified framework with an agent planner at the core, a bank of modality specific retrievers, and a shared graph as the hub. The graph would align text, images, audio, and tables in one structure and give the model a single stage for orderly reasoning. Early tests with models that reason over both structured and unstructured data already show clear gains, suggesting that such a unified framework is within reach [55].

Minqi Zhuo. In the time being, a set of RAG frameworks are emerging, like GraphRAG, KG-RAG, Agentic RAG, multimodal RAG, etc., but each of them has its own advantages in solving a set of real application scenarios. Based on our industry deployment observations, as discussed in Q1, we found that it requires different capabilities to serve different application scenarios, e.g., document/mail QA system, the issuance of government documents, or the issuance of legal documents QA system, and it is very difficult to find one framework which is able to server all the purpose.

Xiaofang Zhou. What's most compelling about the evolving RAG landscape is its movement from simple text retrieval toward richer, context-aware reasoning. Each variant–GraphRAG, KG-RAG, Agentic RAG, multimodal RAG–brings a distinct strength: structured knowledge grounding, entity-level disambiguation, reasoning over actions, or cross-modal alignment. To harmonize these capabilities, the main challenges lie in either modular orchestration or alignment of diverse feature spaces. Modular orchestration requires a unified control layer capable of routing sub-queries to specialized retrievers and aligning their outputs semantically. On the other hand, feature alignment demands a unified embedding space where each modality's encoder projects data into comparable vectors. Harmonizing these modalities under one retrieval-planning interface–while managing latency and ensuring trustworthiness–will advance the effectiveness and versatility of RAG systems.

Theme 3: Relevance to Academia and Industry

Background. The rapid evolution of LLMs+RAG solutions is transforming academic research and industrial applications in various sectors. This shift forces researchers to continually innovate while sidestepping common methodological pitfalls and industry to update technical competencies and operational frameworks. Besides, the integration of these technologies prompts critical discussions on fostering an equitable research landscape and mitigating monopolistic influences.

Q6. *Given the rapid surge in LLM literature and emerging solutions, how can academic researchers stay abreast of new trends while ensuring that their work remains impactful? What common pitfalls should new Ph.D. students avoid when entering this domain?*

Yuyu Luo. I suggest regularly scanning key conference proceedings and arXiv for emerging work – often, just reading abstracts or discussion posts can flag important trends without needing to deep-dive into every paper. It's also useful to follow a handful of experts or community newsletters that digest new breakthroughs.

Grounding research in real-world needs and fundamental questions is essential for impact. The open-source movement accelerates innovation, and my research group actively contributes to projects tackling complex data management and analysis [2], including Text-to-SQL systems [4], data analysis agents [1], and LLM agents [3].

For Ph.D. students, common pitfalls include chasing trends without a clear research question and neglecting rigorous evaluation. Impactful work demands careful experimentation and validation, not just demos. I also encourage students to embrace interdisciplinary research at the DATA+AI frontier.

Wenjie Zhang. I think it is essential to embrace new technologies like LLMs to tackle longstanding and emerging real-world data management challenges. Rather than viewing LLMs as separate from traditional database systems, we could explore how they can augment tasks such as query understanding, query answering, and data quality management. Staying impactful means focusing on how these models can address practical pain points in real-world data-centric applications, while grounding solutions in core database principles like efficiency, effectiveness, and scalability. For new Ph.D. students, a common pitfall is either chasing LLM trends without a clear research perspective or overlooking the transformative potential of these models. A solid foundation in data management can guide responsible and innovative applications of LLMs that effectively bridge AI advances with real-world database problems.

Xiaofang Zhou. I believe that embracing new trends is essential, and we should view AI as a valuable opportunity to drive disruptive innovation in the database field, enabling novel approaches and advancements. The scope of database research is continuously evolving, and the impact of a research effort largely depends on whether it addresses a genuine, real-world problem. This underscores the importance for students to have strong hands-on experience with real systems and applications, allowing them to identify new challenges and develop innovative solutions. However, one concern I have regarding the training of future Ph.D. students is that AI tools may lead to a lack of patience among young researchers. Since AI can quickly generate seemingly plausible answers, it may discourage deep critical thinking–an essential skill for any Ph.D. candidate. This could ultimately hinder the development of rigorous analytical and problem-solving abilities.

Q7. *What industry requirements are projected for this area over the next 3-5 years? Which technical skills will organizations prioritize when hiring new graduates?*

Yuyu Luo. In the next 3-5 years, industry demand will grow heavily for professionals capable of bridging data management and artificial intelligence (DATA+AI). Companies will increasingly seek experts proficient in deploying and maintaining LLMs within production environments. Expertise in vector databases and efficient similarity search techniques will be particularly valuable, as organizations aim to construct and optimize data infrastructures (e.g., data retrieval pipelines) for LLM-driven applications. Due to the necessity of scaling these systems, experience in AI/Cloud infrastructure and performance optimization–especially involving GPUs and specialized hardware–will also be highly sought after.

Wenjie Zhang. From my own experience collaborating with industry partners in Australia, one of the most critical requirements is the development of trustworthy and reliable LLM-powered data solutions. Industries are increasingly interested in applying LLMs to enhance data-driven decision-making, but they are equally concerned about issues such as data privacy, security, model hallucination, explainability, accountability, and regulatory compliance. Besides strong programming and system skills, technical capabilities in areas like LLM finetuning, prompt engineering, RAG, data governance, and evaluation of model trustworthiness will be in high demand. Graduates who can bridge AI techniques with rigorous data management practices will be particularly valuable in addressing real-world challenges.

Minqi Zhou. Over the next 3-5 years, enterprise and consumer agents will proliferate, with enterprise agents autonomously executing or augmenting worker tasks by interpreting organizational processes via integrated data pipelines, and consumer agents personalizing experiences through unified, cross-device (e.g., phone, Pad, and PC) data aggregation. How to integrate all these unstructured, isolated data to understand the enterprises and the users is an important direction for the graduates.

Xiaofang Zhou. I believe that over the next 3-5 years, the scale and variety of data we handle will be significantly different from what we see today. With the rapid growth of AI applications, there will be an explosion of high-dimensional data that needs to be generated, processed, and managed. To handle this shift, new systems and tools will need to be developed. Additionally, the increasing prevalence of multimodal and unstructured data will drive the need for innovative storage and indexing strategies. As a result, organizations will seek graduates who not only understand traditional DB systems but are also well-versed in emerging trends like AI-powered DBs and multimodal data processing. Candidates who demonstrate adaptability, cross-disciplinary knowledge, and hands-on experience with AI+DB and system development will stand out in this evolving landscape.

Q8. *What strategies can curb big tech monopolies in the LLM era, and how might advanced database technologies help democratize AI?*

Yuyu Luo. Open-source AI models like LLaMA [63, 62] provide a practical countermeasure against monopolistic control, especially when combined with advanced data management tools that enable seamless integration with high-quality local datasets. RAG exemplifies how such integration boosts utility. Extending crowdsourced data management to LLM-agent systems, where multiple agents collaborate, raises classic cost-quality tradeoffs–an area well-aligned with database researchers' expertise in cost-based optimization. Data-centric AI [37, 13, 39, 12], which focuses on enhancing data quality to drive model performance, offers another fertile ground for database research [15, 14, 11].

Wenjie Zhang. From a technical perspective, advanced database technologies can play a key role in democratizing AI by enabling more efficient, cost-effective, and domain-specific model deployment outside large cloud-based ecosystems. Techniques such as RAG allows smaller organizations to build powerful AI systems without training or hosting large foundation models. Open-source LLMs combined with high-performance data infrastructure can empower researchers, startups, small/medium enterprises, and public institutions to develop tailored solutions using their local data and tasks.

Minqi Zhuo. In agentic AI, deployed agents hinge on three core capabilities: leveraging LLM infrastructures for world understanding, invoking diverse tools via the Model Context Protocol (MCP), and employing database-driven techniques to integrate isolated enterprise or consumer data sources with atomicity, consistency, and isolation–underscoring the critical importance of rigorous data preparation and high-quality integration for democratizing AI.

Xiaofang Zhou. On the application side, open-source LLMs such as DeepSeek and Qwen are increasingly capable of competing with closed-source alternatives. In many sensitive and high-privacy scenarios, there is a strong demand for locally deployed LLMs. In such cases, database solutions supporting these models can be more distributed and flexible compared to monolithic, super-scale databases, better aligning with the needs of privacy-preserving and decentralized AI deployments.

3. CONCLUDING REMARKS

The panel concluded by discussing the other challenges for real-world LLM-RAG deployment: consistency, robustness, privacy, security, and human-in-the-loop – noting that semantically equivalent queries often produce divergent LLM outputs, a problem magnified by varying retrieval orders in iterative or multimodal RAG workflows. Achieving robustness requires end-to-end guarantees for query execution and formal validation of retrieval completeness and soundness across the pipeline. Addressing privacy and security calls for advanced privacy-preserving methods and fine-grained access controls over data, embeddings, model parameters, and query logs. Ensuring trust and transparency demands tight human-DB interfaces and provenance-based explanations citing source documents. We hope that this discussion and open challenges will inspire future works on the domain's emerging data management issues.

Some of the panelists and the moderator of this panel (1) will co-organize and participate in a future Dagstuhl seminar on "Managing Vector Data for Retrieval Augmented Generation: Systems and Algorithms" [9]; and (2) will also co-organize the second edition of the LLM + Vector Data workshop at ICDE 2026 [5] for initiating more discussion and interdisciplinary collaboration.

4. REFERENCES

[1] DIAL@HKUST(GZ), DeepFund Project, https://github.com/HKUSTDial/DeepFund

[2] DIAL@HKUST(GZ), https://github.com/HKUSTDial

[3] DIAL@HKUST(GZ), OpenManus Project, https://github.com/FoundationAgents/OpenManus

[4] DIAL@HKUST(GZ), Text-to-SQL Project, https://github.com/HKUSTDial/NL2SQL_Handbook

[5] LLM + Vector Data @ ICDE 2026, Second International Workshop on Coupling of Large Language Models with Vector Data Management: Agentic RAG Edition, https://llmvdb2.github.io/

[6] Retrieval augmented generation (rag) in azure ai search. https://learn.microsoft.com/en-us/azure/search/retrieval-augmented-generation-overview?tabs=docs, 2025.

[7] S. Amer-Yahia, J. Bogojeska, R. Facchinetti, V. Franceschi, A. Gionis, K. Hose, G. Koutrika, R. Kouyos, M. Lissandrini, S. Maniu, K. Mirylenka, D. Mottin, T. Palpanas, M. Rigotti, and Y. Velegrakis. Towards reliable conversational data analytics. In *International Conference on Extending Database Technology (EDBT)*, pages 962–969, 2025.

[8] S. Amer-Yahia, A. Bonifati, L. Chen, G. Li, K. Shim, J. Xu, and X. Yang. From large language models to databases and back: A discussion on research and education. *SIGMOD Rec.*, 52(3):49–56, 2023.

[9] S. Amer-Yahia, A. Khan, W. Lehner, S. Mehrotra, and W. Zhang. Managing vector data for retrieval augmented generation: Systems and algorithms (Dagstuhl Seminar 26161). https://www.dagstuhl.de/en/seminars/seminar-calendar/seminar-details/26161, 2026.

[10] A. Biswal, L. Patel, S. Jha, A. Kamsetty, S. Liu, J. E. Gonzalez, C. Guestrin, and M. Zaharia. Text2SQL is not enough: Unifying AI and databases with TAG. *CoRR*, abs/2408.14717, 2024.

[11] C. Chai, K. Jin, N. Tang, J. Fan, D. Miao, J. Wang, Y. Luo, G. Li, Y. Yuan, and G. Wang. Cost-effective missing value imputation for data-effective machine learning. *ACM Trans. Database Syst.*, 50(3), May 2025.

[12] C. Chai, J. Liu, N. Tang, J. Fan, D. Miao, J. Wang, Y. Luo, and G. Li. Goodcore: Data-effective and data-efficient machine learning through coreset selection over incomplete data. *Proc. ACM Manag. Data*, 1(2):157:1–157:27, 2023.

[13] C. Chai, J. Liu, N. Tang, G. Li, and Y. Luo. Selective data acquisition in the wild for model charging. *Proc. VLDB Endow.*, 15(7):1466–1478, 2022.

[14] C. Chai, N. Tang, J. Fan, and Y. Luo. Demystifying artificial intelligence for data preparation. In *SIGMOD Conference Companion*, pages 13–20. ACM, 2023.

[15] C. Chai, J. Wang, Y. Luo, Z. Niu, and G. Li. Data management for machine learning: A survey. *IEEE Trans. Knowl. Data Eng.*, 35(5):4646–4667, 2023.

[16] B. J. Chan, C. Chen, J. Cheng, and H. Huang. Don't do RAG: when cache-augmented generation is all you need for knowledge tasks. In *ACM Web Conference (WWW)*, pages 893–897, 2025.

[17] S. Chen, J. Fan, B. Wu, N. Tang, C. Deng, P. Wang, Y. Li, J. Tan, F. Li, J. Zhou, and X. Du. Automatic database configuration debugging using retrieval-augmented language models. *Proc. ACM Manag. Data*, 3(1):13:1–13:27, 2025.

[18] P. Christmann and G. Weikum. RAG-based question answering over heterogeneous data and text. *IEEE Data Eng. Bull.*, 48(4):71–86, 2024.

[19] S. Das, R. Saboo, C. S. K. Vadrevu, B. Wang, and S. Xu. Applications of llms in e-commerce search and product knowledge graph: The doordash case study. In *ACM International Conference on Web Search and Data Mining (WSDM)*, pages 1163–1164, 2024.

[20] X. L. Dong, S. Moon, Y. E. Xu, K. Malik, and Z. Yu. Towards next-generation intelligent assistants leveraging LLM techniques. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 5792–5793, 2023.

[21] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson. From local to global: A graph RAG approach to query-focused summarization. *CoRR*, abs/2404.16130, 2024.

[22] R. C. Fernandez, A. J. Elmore, M. J. Franklin, S. Krishnan, and C. Tan. How large language models will disrupt data management. *Proc. VLDB Endow.*, 16(11):3302–3309, 2023.

[23] M. Grohe. word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*, pages 1–16, 2020.

[24] A. Y. Halevy, Y. Choi, A. Floratou, M. J. Franklin, N. F. Noy, and H. Wang. Will llms reshape, supercharge, or kill data science? *Proc. VLDB*

Endow., 16(12):4114–4115, 2023.

[25] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. 43(2), 2025.

[26] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Annual ACM Symposium on the Theory of Computing (STOC)*, pages 604–613, 1998.

[27] A. Khan, Y. Wang, W. Zhang, Y. Tian, and M. T. Özsu. LLM + vector data: Coupling of large language models with vector data management for enhancing data science. In *Workshops at the International Conference on Data Engineering (ICDEW)*, 2025.

[28] L. Kilpatrick, S. B. Mallick, and R. Kofman. Gemini 1.5 Pro 2M context window, code execution capabilities, and Gemma 2 are available today. `https://developers.googleblog.com/en/new-features-for-the-gemini-api-and-google-ai-studio/`, 2024.

[29] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[30] B. Li, Y. Luo, C. Chai, G. Li, and N. Tang. The dawn of natural language to SQL: are we fully ready? [experiment, analysis & benchmark]. *Proc. VLDB Endow.*, 17(11):3318–3331, 2024.

[31] B. Li, J. Zhang, J. Fan, Y. Xu, C. Chen, N. Tang, and Y. Luo. Alpha-sql: Zero-shot text-to-sql using monte carlo tree search. In *Forty-Second International Conference on Machine Learning, ICML 2025, Vancouver, Canada, July 13-19, 2025*. OpenReview.net, 2025.

[32] C. Li, Y. Shi, Y. Luo, and N. Tang. Deepfund: Will LLM be professional at fund investment? A live arena perspective. *CoRR*, abs/2503.18313, 2025.

[33] G. Li, X. Zhou, and X. Zhao. LLM for data management. *Proc. VLDB Endow.*, 17(12):4213–4216, 2024.

[34] H. Li, Y. Li, A. Tian, T. Tang, Z. Xu, X. Chen, N. Hu, W. Dong, Q. Li, and L. Chen. A survey on large language model acceleration based on kv cache management, 2025.

[35] L. Liang, Z. Bo, Z. Gui, Z. Zhu, L. Zhong, P. Zhao, M. Sun, Z. Zhang, J. Zhou, W. Chen, W. Zhang, and H. Chen. KAG: boosting llms in professional domains via knowledge augmented generation. In *ACM Web Conference (WWW)*, pages 334–343, 2025.

[36] J. Lin, P. Gupta, W. Horn, and G. Mishne. Musings about the future of search: A return to the past? *CoRR*, abs/2412.18956, 2024.

[37] X. Lin, Y. Qi, Y. Zhu, T. Palpanas, C. Chai, N. Tang, and Y. Luo. LEAD: iterative data selection for efficient LLM instruction tuning. *CoRR*, abs/2505.07437, 2025.

[38] B. Liu, X. Li, J. Zhang, J. Wang, T. He, S. Hong, H. Liu, S. Zhang, K. Song, K. Zhu, Y. Cheng, S. Wang, X. Wang, Y. Luo, H. Jin, P. Zhang, O. Liu, J. Chen, H. Zhang, Z. Yu, H. Shi, B. Li, D. Wu, F. Teng, X. Jia, J. Xu, J. Xiang, Y. Lin, T. Liu, T. Liu, Y. Su, H. Sun, G. Berseth, J. Nie, I. Foster, L. T. Ward, Q. Wu, Y. Gu, M. Zhuge, X. Tang, H. Wang, J. You, C. Wang, J. Pei, Q. Yang, X. Qi, and C. Wu. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *CoRR*, abs/2504.01990, 2025.

[39] J. Liu, C. Chai, Y. Luo, Y. Lou, J. Feng, and N. Tang. Feature augmentation with reinforcement learning. In *ICDE*, pages 3360–3372. IEEE, 2022.

[40] X. Liu, S. Shen, B. Li, P. Ma, R. Jiang, Y. Zhang, J. Fan, G. Li, N. Tang, and Y. Luo. A survey of text-to-sql in the era of llms: Where are we, and where are we going?, 2025.

[41] X. Liu, S. Shen, B. Li, N. Tang, and Y. Luo. Nl2sql-bugs: A benchmark for detecting semantic errors in NL2SQL translation. *CoRR*, abs/2503.11984, 2025.

[42] Y. Luo, X. Qin, N. Tang, and G. Li. Deepeye: Towards automatic data visualization. In *ICDE*, pages 101–112. IEEE Computer Society, 2018.

[43] Y. Luo, N. Tang, G. Li, C. Chai, W. Li, and X. Qin. Synthesizing natural language to visualization (NL2VIS) benchmarks from NL2SQL benchmarks. In *SIGMOD Conference*, pages 1235–1247. ACM, 2021.

[44] Y. Luo, N. Tang, G. Li, J. Tang, C. Chai, and X. Qin. Natural language to visualization by neural machine translation. *IEEE Trans. Vis. Comput. Graph.*, 28(1):217–226, 2022.

[45] A. Obeng, J. Zhong, and C. Gu. How we built Text-to-SQL at Pinterest. `https://medium.com/pinterest-engineering/how-we-built-text-to-sql-at-pinterest-30bad30dabff`, 2024.

[46] J. J. Pan, J. Wang, and G. Li. Survey of vector database management systems. *VLDB J.*, 33(5):1591–1615, 2024.

[47] L. Patel, S. Jha, C. Guestrin, and M. Zaharia.

LOTUS: enabling semantic queries with llms over tables of unstructured and structured data. *CoRR*, abs/2407.11418, 2024.

[48] F. Petroni, F. Siciliano, F. Silvestri, and G. Trappolini. Report on the 1st workshop on information retrieval's role in rag systems (ir-rag 2024) at sigir 2024. *SIGIR Forum*, 58(2):1–12, 2025.

[49] X. Qin, Y. Luo, N. Tang, and G. Li. Making data visualization more efficient and effective: a survey. *VLDB J.*, 29(1):93–117, 2020.

[50] L. Shen, H. Li, Y. Wang, T. Luo, Y. Luo, and H. Qu. Data playwright: Authoring data videos with annotated narration. *CoRR*, abs/2410.03093, 2024.

[51] J. Shim, G. Seo, C. Lim, and Y. Jo. Tooldial: Multi-turn dialogue generation method for tool-augmented language models. In *International Conference on Learning Representations (ICLR)*, 2025.

[52] A. Singh, A. Ehtesham, S. Kumar, and T. T. Khoei. Agentic retrieval-augmented generation: A survey on agentic RAG. *CoRR*, abs/2501.09136, 2025.

[53] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972), 2023.

[54] X. TAN, X. WANG, Q. LIU, X. XU, X. YUAN, and W. ZHANG. Paths-over-graph: Knowledge graph empowered large language model reasoning. In *PROCEEDINGS OF THE ACM ON WEB CONFERENCE 2025*, 2025.

[55] X. TAN, X. WANG, Q. LIU, X. XU, X. YUAN, L. ZHU, and W. ZHANG. Hydra: Structured cross-source enhanced large language model reasoning. *ARXIV PREPRINT ARXIV:2505.17464*, 2025.

[56] N. Tang, C. Yang, J. Fan, L. Cao, Y. Luo, and A. Y. Halevy. Verifai: Verified generative AI. In *CIDR*. www.cidrdb.org, 2024.

[57] N. Tang, C. Yang, Z. Zhang, Y. Luo, J. Fan, L. Cao, S. Madden, and A. Y. Halevy. Symphony: Towards trustworthy question answering and verification using RAG over multimodal data lakes. *IEEE Data Eng. Bull.*, 48(4):135–146, 2024.

[58] Y. Tian, Z. Yue, R. Zhang, X. Zhao, B. Zheng, and X. Zhou. Approximate nearest neighbor search in high dimensional vector databases: Current research and future directions. *IEEE Data Eng. Bull.*, 47(3):39–54, 2023.

[59] J. Wang, G. Li, and J. Feng. idatalake: An llm-powered analytics system on data lakes. *IEEE Data Eng. Bull.*, 49(1):57–69, 2025.

[60] N. Wang, H. Yang, and C. D. Wang. FinGPT: Instruction tuning benchmark for open-source large language models in financial datasets. *NeurIPS Workshop on Instruction Tuning and Instruction Following*, 2023.

[61] B. Wu, J. Shi, Y. Wu, N. Tang, and Y. Luo. Transxssm: A hybrid transformer state space model with unified rotary position embedding. *CoRR*, abs/2506.09507, 2025.

[62] B. Wu, J. Shi, Y. Wu, N. Tang, and Y. Luo. Transxssm: A hybrid transformer state space model with unified rotary position embedding, 2025.

[63] Y. Wu, J. Shi, B. Wu, J. Zhang, X. Lin, N. Tang, and Y. Luo. Concise reasoning, big gains: Pruning long reasoning trace with difficulty-aware prompting, 2025.

[64] P. Xia, K. Zhu, H. Li, H. Zhu, Y. Li, G. Li, L. Zhang, and H. Yao. RULE: Reliable multimodal RAG for factuality in medical vision language models. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1081–1093, 2024.

[65] C. Xiao, X. Hu, Z. Liu, C. Tu, and M. Sun. Lawformer: A pre-trained language model for Chinese legal long documents. *AI Open*, 2:79–84, 2021.

[66] Y. Xie, Y. Luo, G. Li, and N. Tang. Haichart: Human and AI paired visualization system. *Proc. VLDB Endow.*, 17(11):3178–3191, 2024.

[67] Z. Xu, M. J. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, and Z. Li. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2905–2909, 2024.

[68] K. Yang, Y. Tian, N. Peng, and D. Klein. Re3: Generating longer stories with recursive reprompting and revision. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 4393–4479, 2022.

[69] Y. Ye, J. Hao, Y. Hou, Z. Wang, S. Xiao, Y. Luo, and W. Zeng. Generative AI for visualization: State of the art and future directions. *Vis. Informatics*, 8(1):43–66, 2024.

[70] Z. Yue, H. Zhuang, A. Bai, K. Hui, R. Jagerman, H. Zeng, Z. Qin, D. Wang, X. Wang, and M. Bendersky. Inference scaling for long-context retrieval augmented generation. In *International Conference on Learning Representations (ICLR)*, 2025.

[71] J. Zhang, J. Xiang, Z. Yu, F. Teng, X. Chen, J. Chen, M. Zhuge, X. Cheng, S. Hong, J. Wang, B. Zheng, B. Liu, Y. Luo, and C. Wu. Aflow: Automating agentic workflow generation. In *ICLR*. OpenReview.net, 2025.

[72] Z. Zhang, C. Chen, B. Liu, C. Liao, Z. Gong, H. Yu, J. Li, and R. Wang. Unifying the perspectives of NLP and software engineering: A survey on language models for code. *Transactions on Machine Learning Research*, 2024.

[73] Z. Zhang, Z. Liang, Y. Wu, T. Lin, Y. Luo, and N. Tang. Datamosaic: Explainable and verifiable multi-modal data analytics through extract-reason-verify. *CoRR*, abs/2504.10036, 2025.

[74] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, and B. Cui. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*, 2024.

[75] X. Zhao, Y. Tian, K. Huang, B. Zheng, and X. Zhou. Towards efficient index construction and approximate nearest neighbor search in high-dimensional spaces. In *PVLDB*, 2023.

[76] T. Zheng, Z. Deng, H. T. Tsang, W. Wang, J. Bai, Z. Wang, and Y. Song. From automation to autonomy: A survey on large language models in scientific discovery. *arXiv preprint arXiv:2505.13259*, 2025.

[77] Y. Zhu, S. Du, B. Li, Y. Luo, and N. Tang. Are large language models good statisticians? In *NeurIPS*, 2024.

[78] Y. Zhu, R. Jiang, B. Li, N. Tang, and Y. Luo. Elliesql: Cost-efficient text-to-sql with complexity-aware routing. *COLM*, 2025.