Online
Resources

Advanced Topics on LLM+KG for QA
Part -3

Yongrui Chen

Southeast University

AALBORG UNIVERSITET

Tutorial Outline

1) **Introduction (15 Min) – Arijit Khan**
   1.1 Large Language Models (LLMs)
   1.2 Knowledge Graphs (KGs)
   1.3 Unifying LLMs+KGs
   1.4 Question Answering (QA)

2) Unifying LLMs with KGs for QA (25 Min) – Chuangtao Ma
   2.1 KGs as Background Knowledge
   2.2 KGs as Reasoning Guidelines
   2.3 KGs as Refiners and Validators

3) **Advanced Topics on LLM+KG for QA (25 Min) - Yongrui Chen**
   3.1 Natural Language Questions to Structured Queries
   3.2 Explainable QA
   3.3 Optimization and Efficiency

- **Break (10 Min)**

4) Evaluations and Applications (20 Min) – Tianxing Wu
   4.1 Performance Metrics
   4.2 Benchmark Datasets
   4.3 Industry Applications and Demonstrations

5) Opportunities for Data Management (10 Min) – Arijit Khan

6) Future Directions (5 Min) – Tianxing Wu

- **Q&A Session (10 Min)**

Contents

KG vs LLM – QA Capability Comparison

LLM QA

- **Code Pre-training:** enhance LLM reasoning during training
- **Prompt Engineering:** eliciting LLM reasoning during inference

KG QA

- Graph computing
- Rule-based reasoning
- Ontology reasoning
- Spatial-temporal reasoning
- KG embedding/GNN

LLM QA

- zero-shot prompting
- Few-shot prompting
- CoT prompting
- Instruction

+

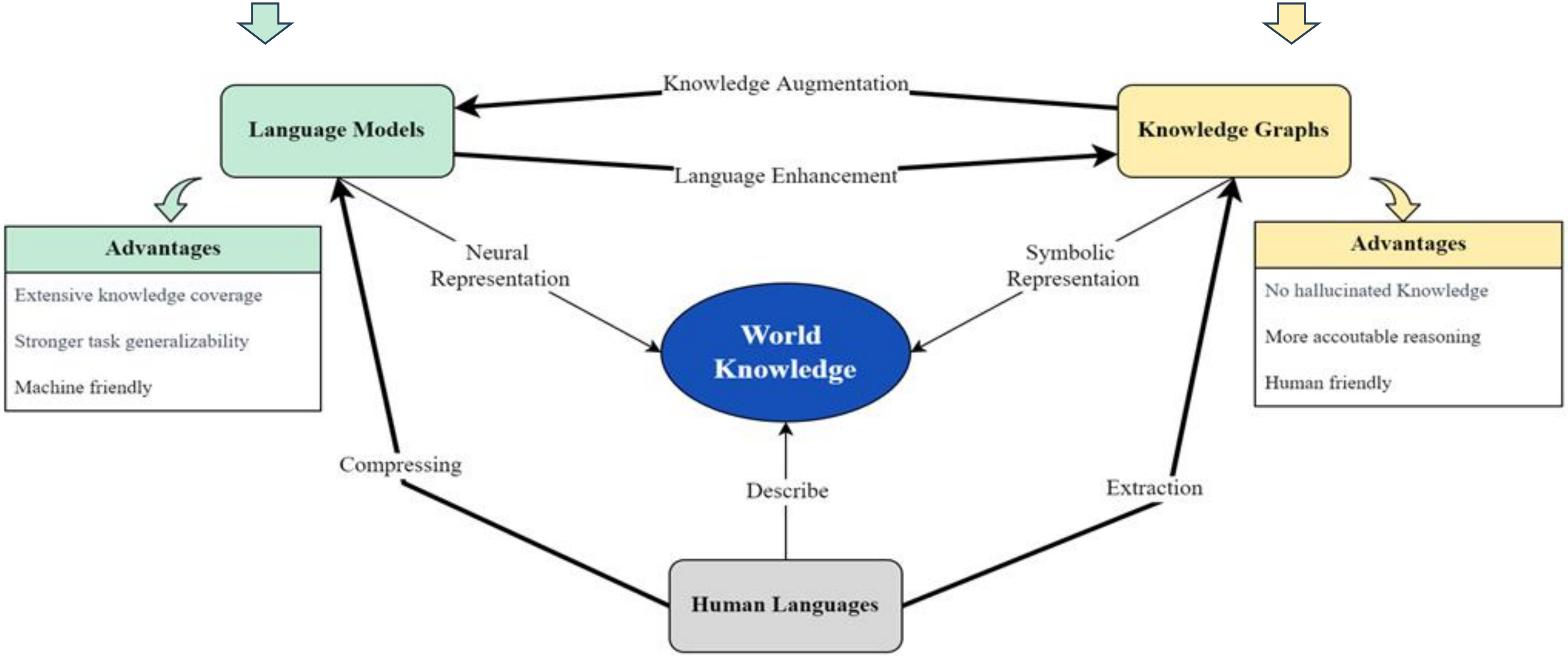
KG QA

- Graph computing
- Rule-based reasoning
- Ontology reasoning
- Spatial-temporal reasoning
- KG embedding/GNN

KG vs LLM – How do KG and LLM collaborate for QA?

Focus on scale & has high coverage

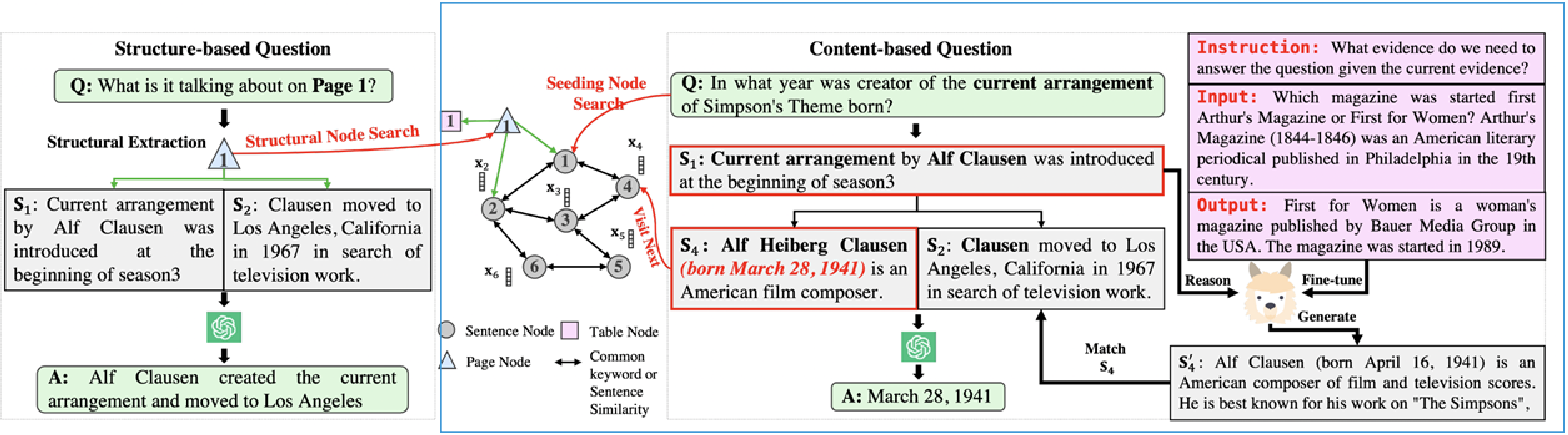
Focus on presentation & has high accuracy



Large Knowledge Models: Perspective and Challenges, 2024.

Contents

Advanced Topics – QA over Multiple Documents

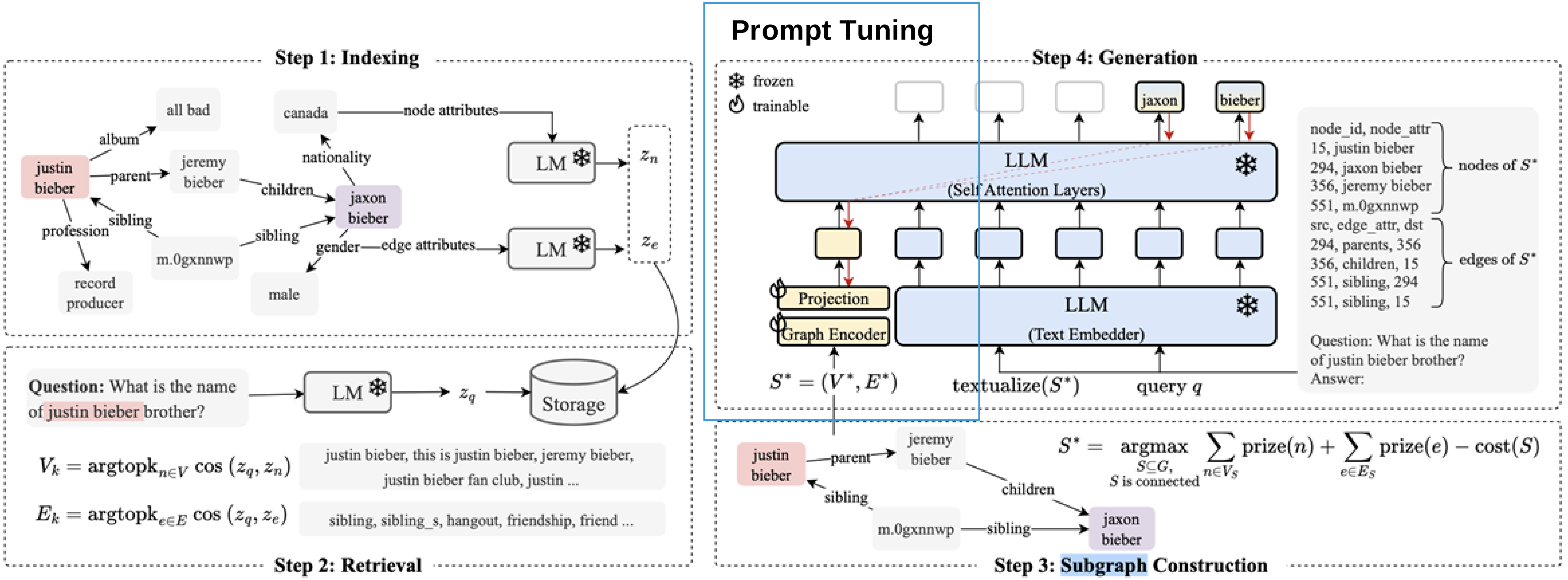


Enhancing LLMs for **Multi-Document QA**, which requires understanding logical associations across multiple documents.

- **KG Construction**: Building a KG where **nodes represent passages or document structures** (e.g., pages, tables) and **edges denote semantic/lexical similarity or structural relations** between them.
- **KG Traversal**: Employing an **LLM-based graph traversal agent** to navigate the KG, gathering relevant supporting passages to assist LLMs in answering questions.

Knowledge graph prompting for multi-document question answering. AAAI 2024

Advanced Topics – Retrieval Augment Generation



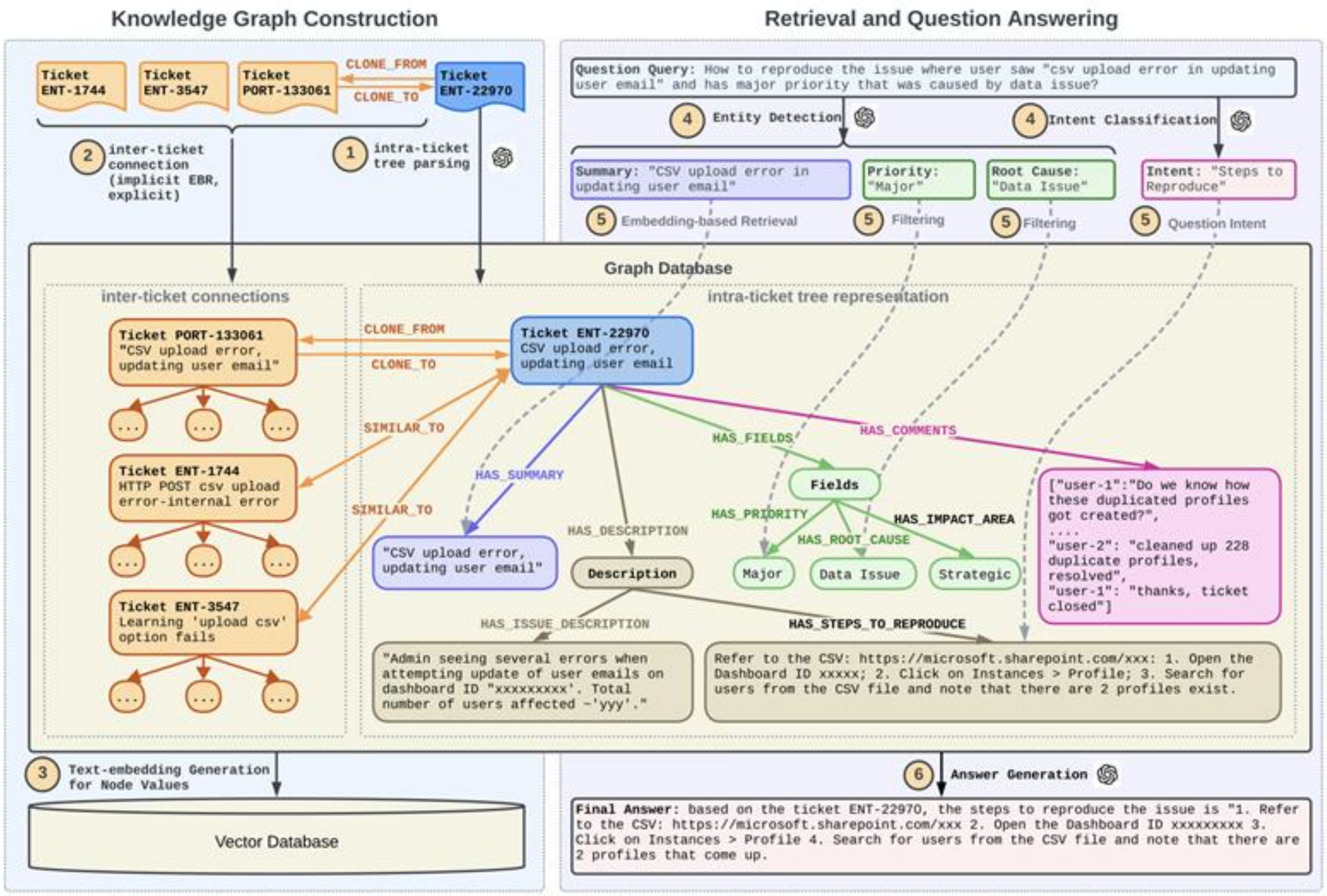
- The method involves four main steps: **indexing** the graph, **retrieving** relevant nodes and edges, **constructing a connected subgraph**, and **generating** the answer using the retrieved subgraph and the query.
- By employing RAG for **direct information retrieval from the actual graph**, G-Retriever effectively **mitigates hallucination** in graph-based question answering.

G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering . Preprint 2024.

Advanced Topics – Retrieval Augment Generation

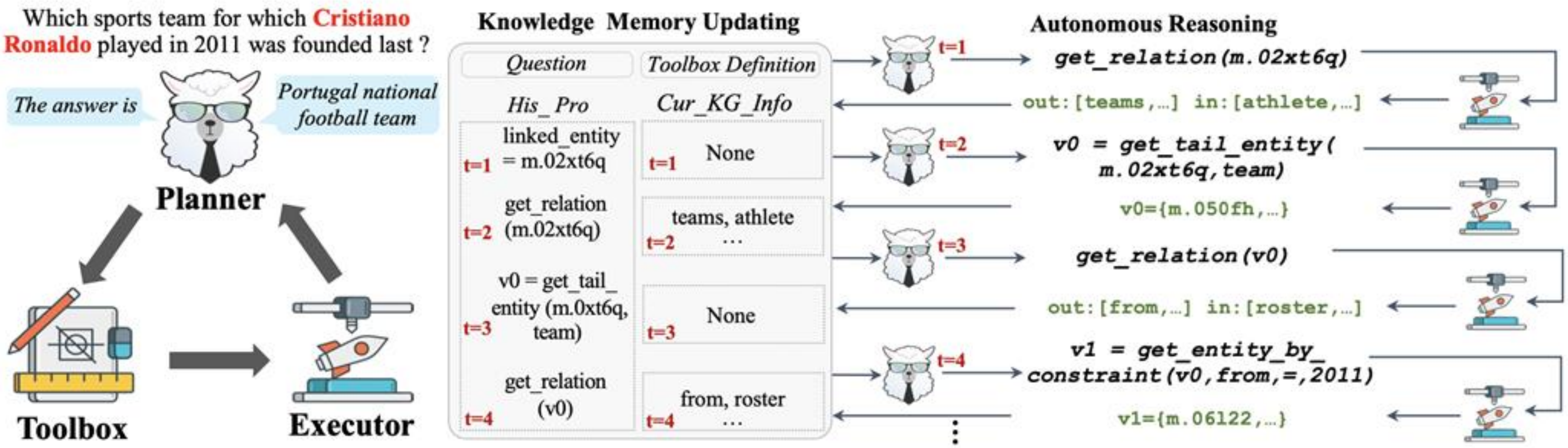
Enhancing the conventional RAG approach by integrating a **knowledge graph** constructed from **historical customer service issue tickets** to improve retrieval accuracy and answer quality.

- Consumer queries are parsed to identify **named entities and intents**.
- The system retrieves related sub-graphs from the KG based on the parsed query, leveraging **both entity matching and embedding similarity**.
- An LLM generates answers using the **retrieved sub-graphs** as context.



Retrieval-augmented generation with knowledge graphs for customer service question answering. SIGIR 2024.

Advanced Topics – KG Agent

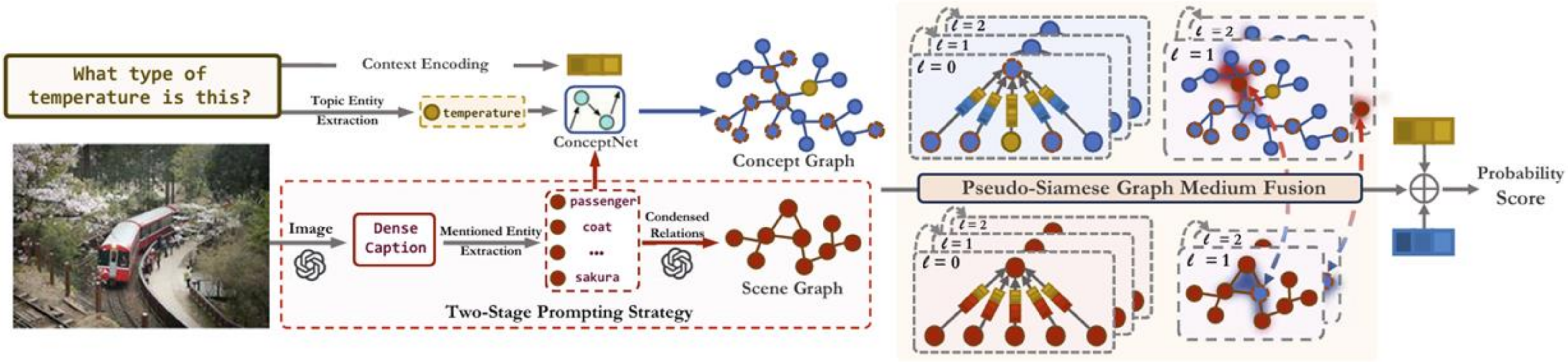


Integrates a **small LLM (e.g., 7B)**, a **multifunctional toolbox**, a **KG-based executor**, and **knowledge memory**.

- Employs an **iterative mechanism** where the LLM autonomously selects a tool from the toolbox and updates the knowledge memory to continue reasoning over the KG until the answer is found.
- **Multifunctional Toolbox**: Extends the LLM's capacity to manipulate structured data by providing tools for **extraction, semantic understanding, and logic operations** on KG data and intermediate results (e.g., filtering, counting, retrieval, relation retrieval, entity disambiguation).

KG-Agent: An Efficient Autonomous Agent Framework for Complex Reasoning over Knowledge Graph. Preprint 2024

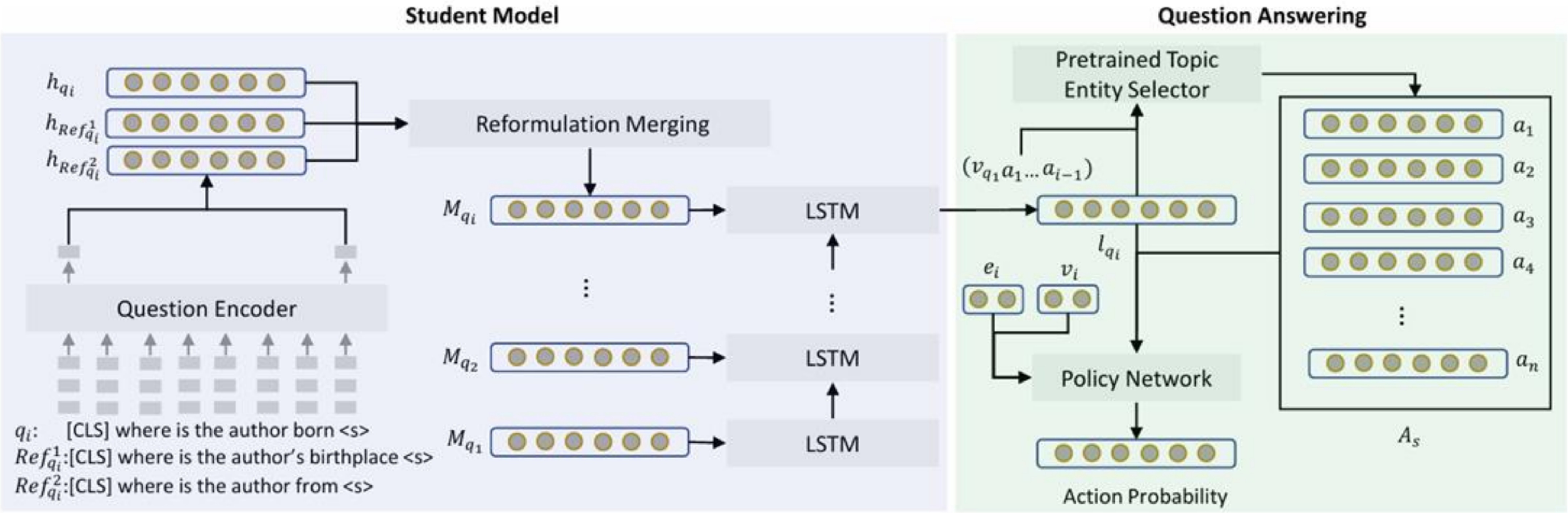
Advanced Topics - Visual QA



- **Two-Stage Prompting**: Utilizing LLMs to generate a **dense image caption** and subsequently extract a **scene graph** containing detailed visual features from it.
- **Coupled Concept Graph**: Constructing a **concept graph** using **ConceptNet**, linking scene graph entities with external knowledge.
- **Pseudo-Siamese Graph Medium Fusion (PS-GMF)**: Utilizing **shared entities as mediums** between the scene graph and concept graph to achieve **cross-modal information exchange** and **fusion.**

Modality-aware integration with large language models for knowledge-based visual question answering. ACL 2024.

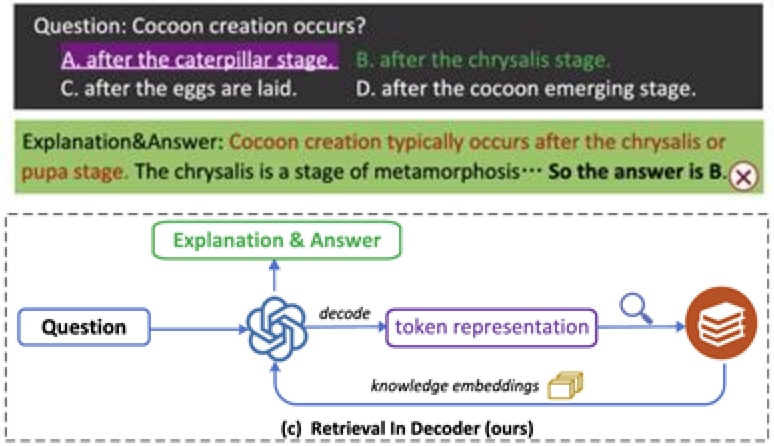
Advanced Topics – Conversational QA



- A **teacher model** is trained directly using **human-written reformulations** to learn effective question representations.
- A **student model**, with the same architecture, is trained to **mimic the teacher's output** using the **LLM-generated reformulations**. This helps the student model approach the performance of the teacher model, even with potentially lower-quality LLM-generated reformulations.

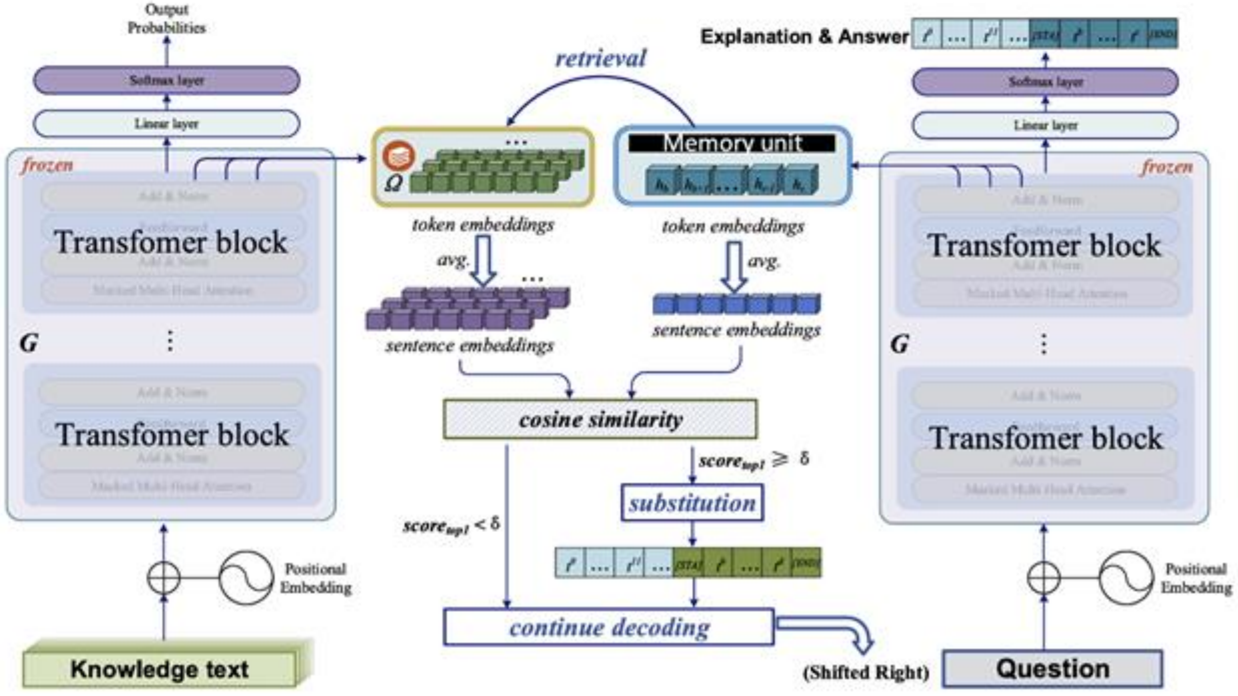
Conversational Question Answering with Language Models Generated Reformulations over Knowledge Graph. ACL Findings 2024.

Advanced Topics – Explainable QA



To enhance the **faithfulness and credibility** of generative models in QA, which contributes to explainability.

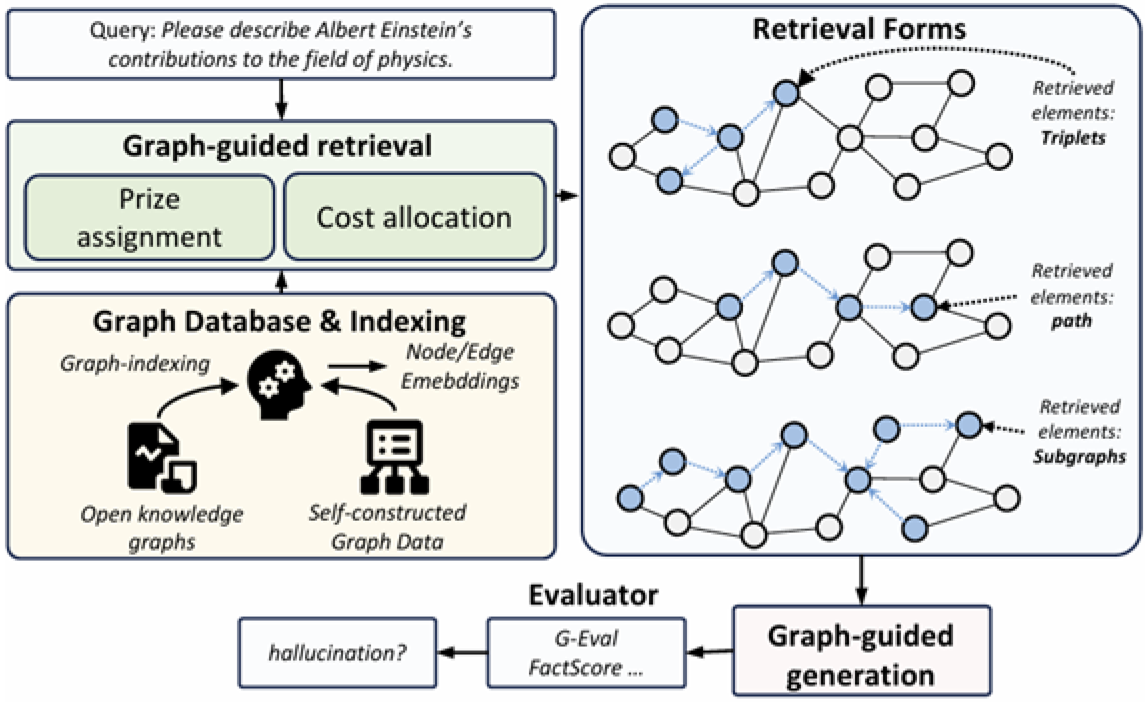
- **Integrated Retrieval:** Integrates information retrieval directly into the decoding process of generative language models, rather than treating them as separate components.
- **Multi-Granularity Decoding:** Supports dynamic adjustment of decoding granularity between token-level and sentence-level based on retrieval outcomes.
- **Rationale-Aware Explanation Generation:** Employs prompt learning to generate explanations that explicitly contain marked rationales.



Advanced Topics – Explainable QA

Goal: Enhancing the **trustworthiness** of LLMs in open-ended question answering by integrating **KGs.**

- **Explainability via Knowledge Source:** KGs provide structured and explicit factual information. Each piece of data in a KG can be traced back to its source, offering provenance.
- **Transparency in Reasoning:** The traceability of KG information not only enables verification of the model's reasoning but also brings transparency to the decision-making process.
- **Open-ended Answers with Supporting Facts:** The OKGQA benchmark encourages LLMs to generate more elaborate answers, including reasoning paths and supporting facts derived from the KG.

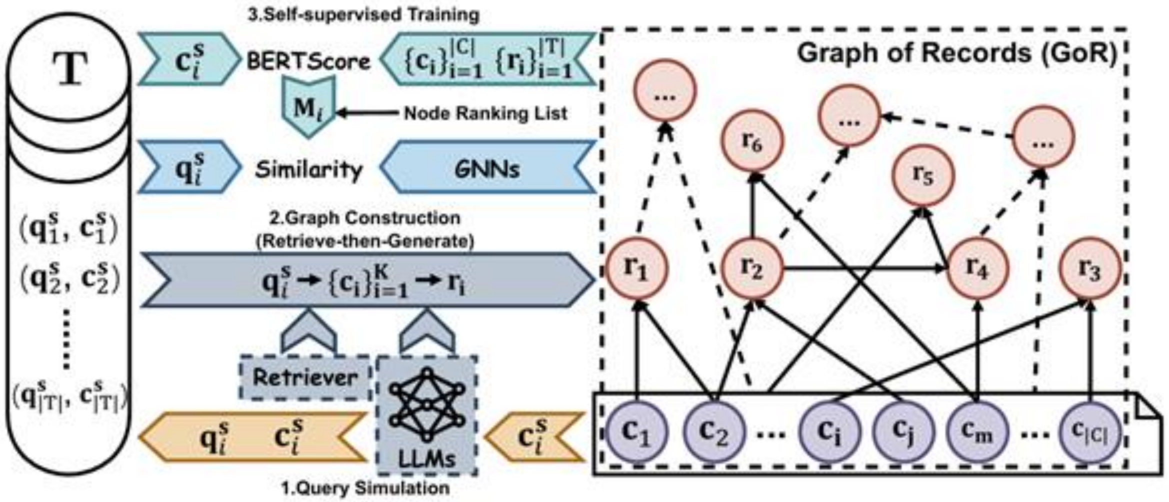


Can Knowledge Graphs Make Large Language Models More Trustworthy? An Empirical Study Over Open-ended Question Answering. Preprint 2024.

Contents

Optimization and Efficiency – Index-based Optimization

Goal: To enhance **RAG** performance in long-context global summarization by using a graph structure built from **LLM-generated historical responses**.



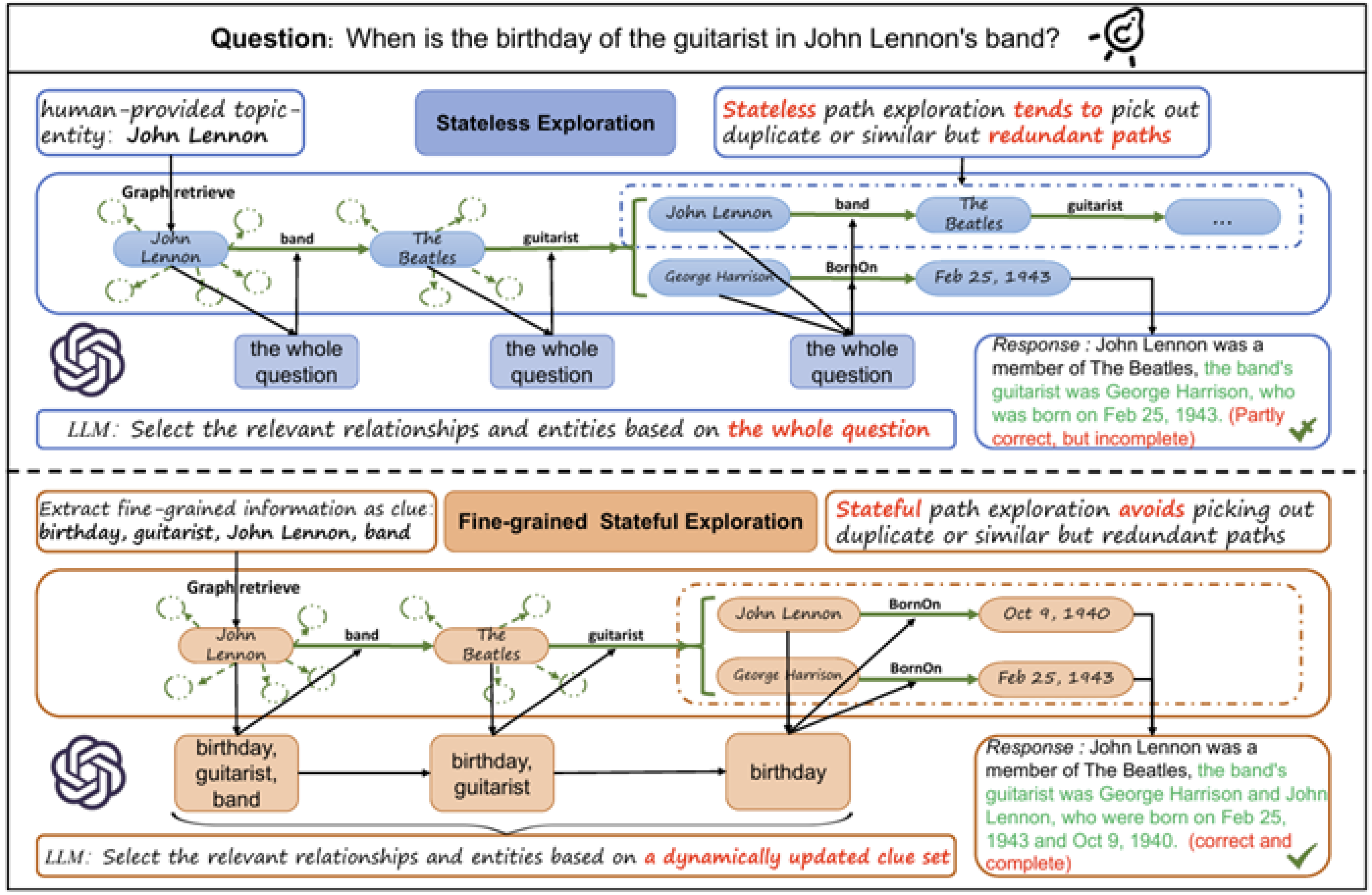
- Simulate user queries, retrieve relevant text chunks, and establish edges between the retrieved text chunks and their corresponding **LLM-generated responses** to construct a **Graph of Records**.
- Utilize a **GNN** to learn embeddings for the nodes in the graph, capturing fine-grained correlations.
- Effectively discovers and leverages **fine-grained correlations between LLM historical responses and text chunks**, thereby improving RAG performance.

Boosting retrieval augmented generation for long-context summarization with graphs. Preprint 2024.

Optimization and Efficiency – Graph Retrieval-based Optimization

Goal: Addresses the **information granularity mismatch** between questions and knowledge graphs, which is identified as a primary source of inefficiency in existing methods.

- Extracts **fine-grained, independent pieces of information (clues)** from the question to guide the retrieval process.
- By avoiding redundancy and ensuring no pertinent information is overlooked, the method **significantly reduces the average number of LLM calls** required for knowledge retrieval compared to existing stateless iterative exploration methods

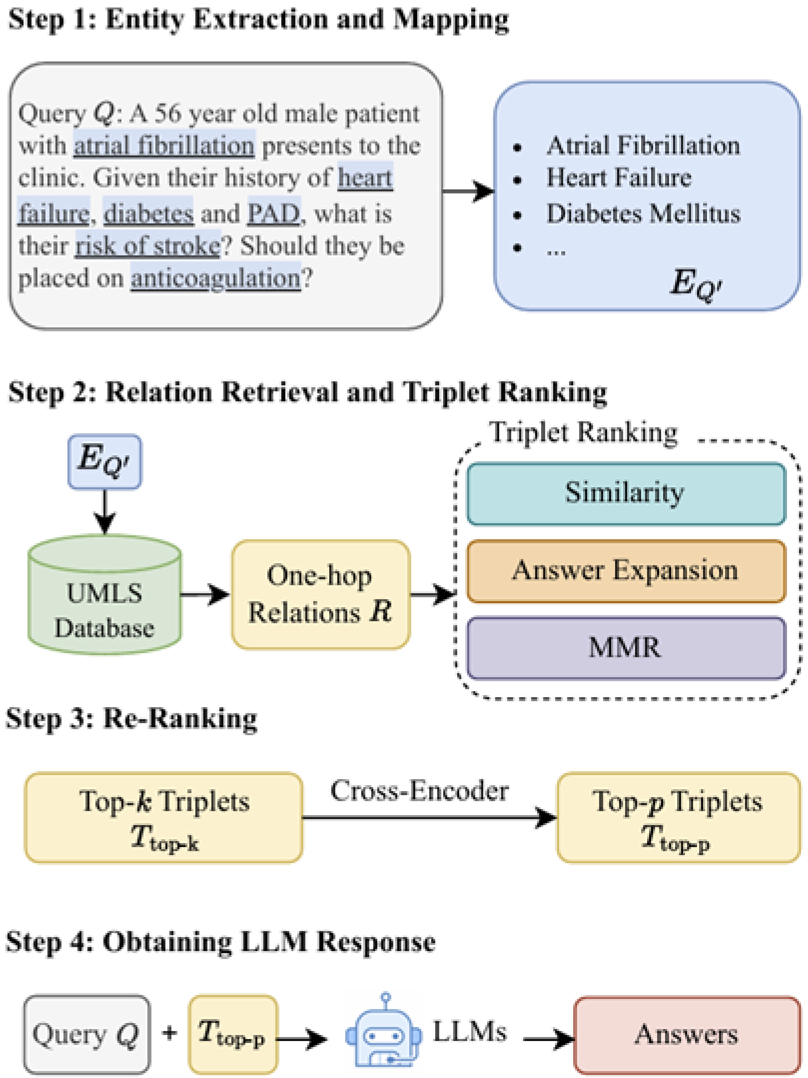


Clue-Guided Path Exploration: Optimizing Knowledge Graph Retrieval with Large Language Models to Address the Information Black Box Challenge. Preprint 2024.

Optimization and Efficiency – Ranking-based Optimization

Goal: Leverages **ranking and re-ranking techniques** to refine the selection and ordering of relevant information retrieved from the medical KG.

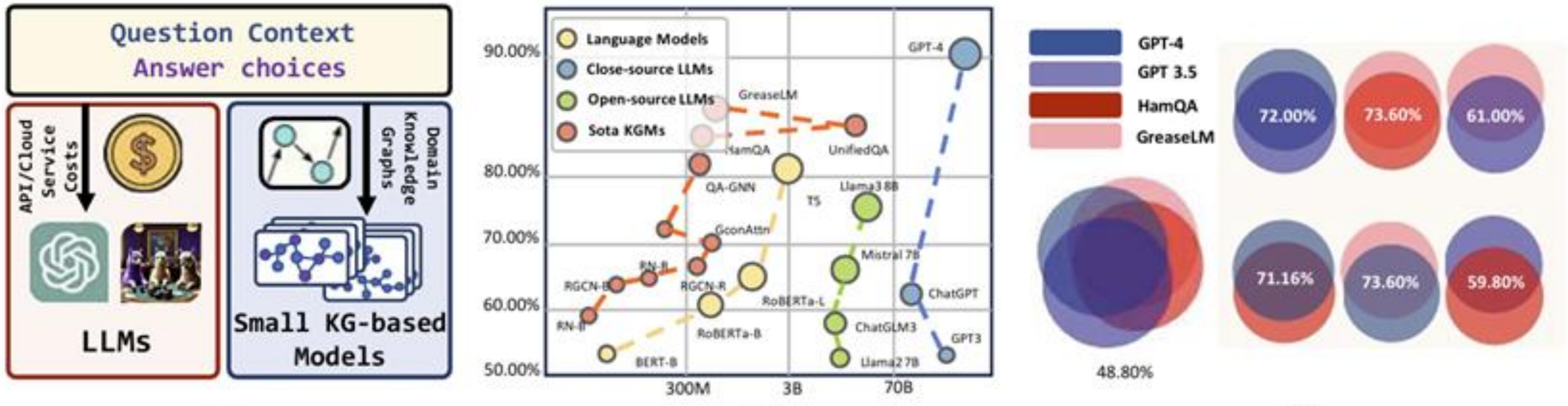
- **Similarity Ranking**: Ranks triplets based on their semantic similarity to the input question using UmlsBERT embeddings.
- **Answer Expansion Ranking**: Uses an LLM to generate a preliminary answer, then ranks triplets based on their similarity to the expanded question-answer context. This helps in identifying information relevant to the potential answer.
- **MMR Ranking**: Selects triplets based on both their relevance to the question and their dissimilarity to already selected triplets, promoting diversity and reducing redundancy.



KG-Rank: Enhancing Large Language Models for Medical QA with Knowledge Graphs and Ranking Techniques. BioNLP 2024.

Optimization and Efficiency – Cost-based Optimization

Goal: To achieve **cost-efficient KBQA** by minimizing the usage and expenses associated with LLMs.



(a) (b) (c)

- **Multi-Armed Bandit Formulation:** Models the model selection problem as a tailored multi-armed bandit problem to balance exploration (trying different models) and exploitation (using the best-performing models) within a limited budget.
- **Accuracy Expectation with Cluster-Level Thompson Sampling:** Estimates the accuracy expectation of choosing either LLMs or KGMs based on their historical success and failure rates. This helps in initially guiding the policy towards more promising model types.
- **Context-Aware Policy:** Learns a context-aware policy that considers the semantics of the question to further distinguish and select the most suitable expert model (either an LLM or a KGM) for that specific question.

Cost-efficient Knowledge-based Question Answering with Large Language Models. NeurIPS 2024.

Contents

Conclusion & Future Work

Conclusion

- **LLM-KG Integration Enhances QA**: Combining LLMs with KGs improves multi-document and multimodal QA by enhancing reasoning, reducing hallucinations, and increasing answer accuracy.
- **Optimization Improves Efficiency**: Techniques like index-based and graph retrieval-based optimization boost system efficiency, scalability, and cost-effectiveness.
- **Conversational and Explainable QA**: QA systems are evolving into multi-turn, explainable models with KG Agents enabling transparent and trustworthy reasoning.

Future Work

- **Deeper LLM-KG Fusion**: Advancing dynamic KG updates and adaptive retrieval will improve knowledge adaptation and model performance.
- **Enhanced Multimodal QA**: Future systems will better integrate text, images, and videos for richer reasoning and more comprehensive answers.
- **Scalable and Privacy-Preserving QA**: Efficient, large-scale QA solutions leveraging federated learning and edge computing will enhance privacy and real-time capabilities.

References

[1] Chen, Huajun. "Large knowledge model: Perspectives and challenges." arXiv preprint arXiv:2312.02706 (2023).

[2] Wang, Yu, et al. "Knowledge graph prompting for multi-document question answering." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 17. 2024.

[3] He, Xiaoxin, et al. "G-retriever: Retrieval-augmented generation for textual graph understanding and question answering." Advances in Neural Information Processing Systems 37 (2024): 132876-132907.

[4] Xu, Zhentao, et al. "Retrieval-augmented generation with knowledge graphs for customer service question answering." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

[5] Jiang, Jinhao, et al. "Kg-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph." arXiv preprint arXiv:2402.11163 (2024).

[6] Dong, Junnan, et al. "Modality-Aware Integration with Large Language Models for Knowledge-Based Visual Question Answering." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024.

[7] Liu, Lihui, et al. "Conversational Question Answering with Language Models Generated Reformulations over Knowledge Graph." Findings of the Association for Computational Linguistics ACL 2024. 2024.

[8] Feng, Jianzhou, et al. "Retrieval In Decoder benefits generative models for explainable complex question answering." Neural Networks 181 (2025): 106833.

[9] Sui, Yuan, and Bryan Hooi. "Can Knowledge Graphs Make Large Language Models More Trustworthy? An Empirical Study over Open-ended Question Answering." arXiv e-prints (2024): arXiv-2410.

[10] Zhang, Haozhen, Tao Feng, and Jiaxuan You. "Graph of records: Boosting retrieval augmented generation for long-context summarization with graphs." arXiv preprint arXiv:2410.11001 (2024).

[11] Tao, Dehao, et al. "Clue-Guided Path Exploration: Optimizing Knowledge Graph Retrieval with Large Language Models to Address the Information Black Box Challenge." arXiv preprint arXiv:2401.13444 (2024).

[12] Yang, Rui, et al. "KG-Rank: Enhancing Large Language Models for Medical QA with Knowledge Graphs and Ranking Techniques." Proceedings of the 23rd Workshop on Biomedical Natural Language Processing. 2024.

[13] Dong, Junnan, et al. "Cost-efficient Knowledge-based Question Answering with Large Language Models." The Thirty-eighth Annual Conference on Neural Information Processing Systems.