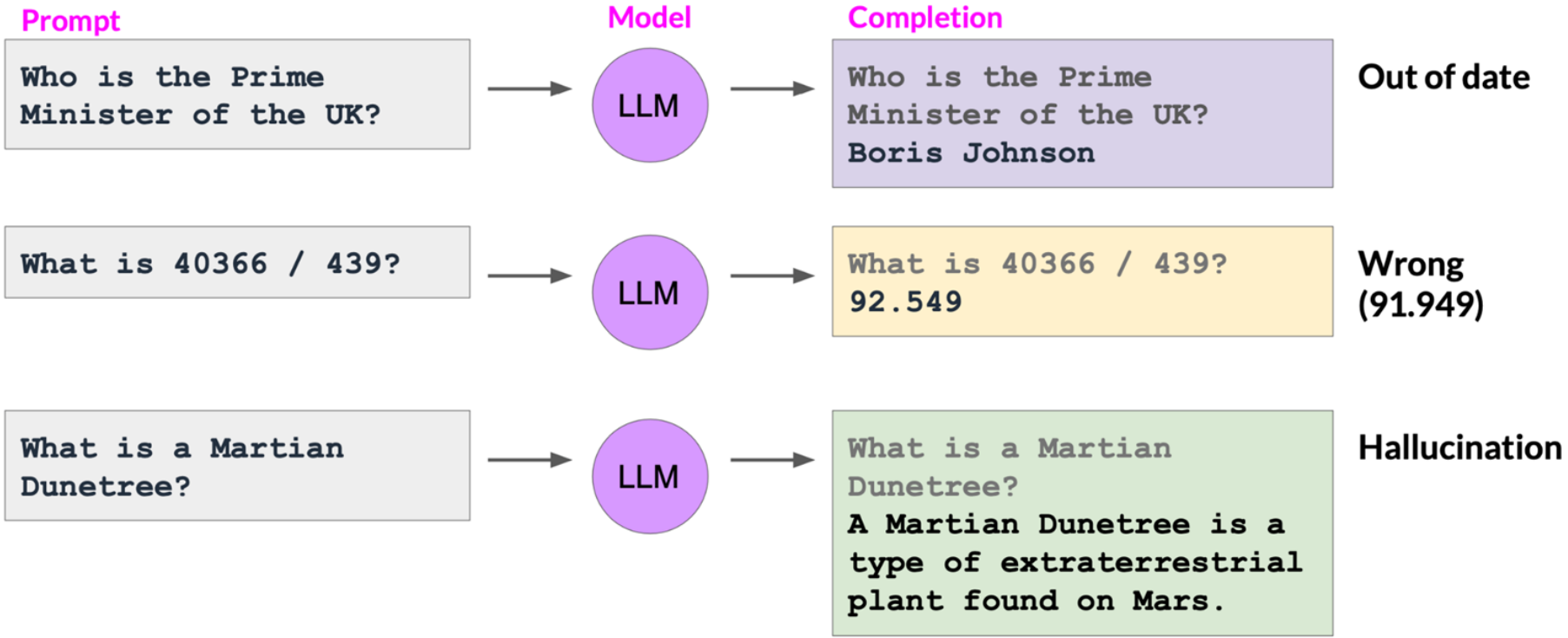
RIT

Retrieval Augmented Generation

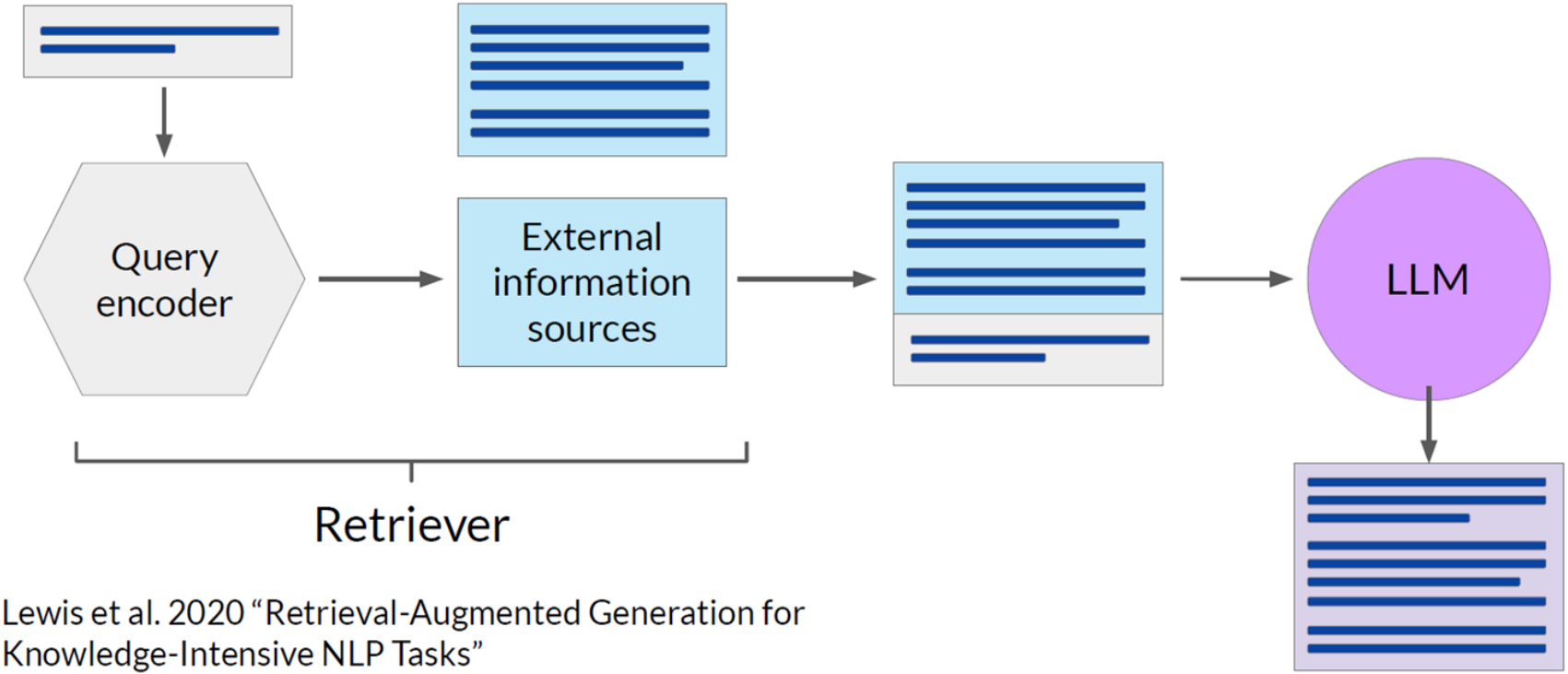
Reza Fayyazi

Models having difficulty



[1] https://www.coursera.org/learn/generative-ai-with-llms/

Retrieval Augmented Generation (RAG)

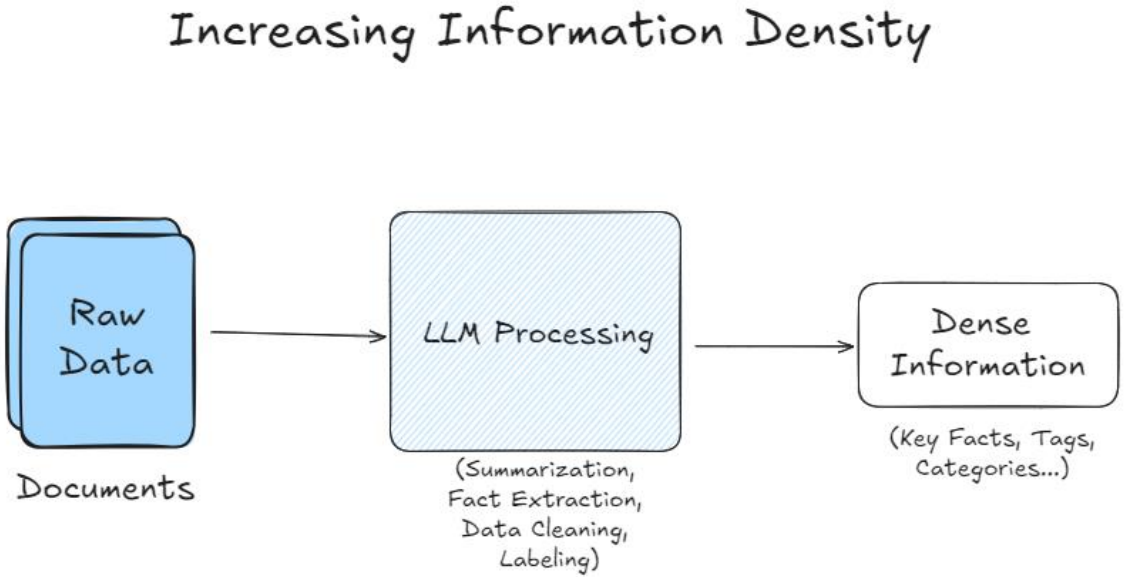


Lewis et al. 2020 “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”

[1] https://www.coursera.org/learn/generative-ai-with-llms/

Advanced RAG Techniques

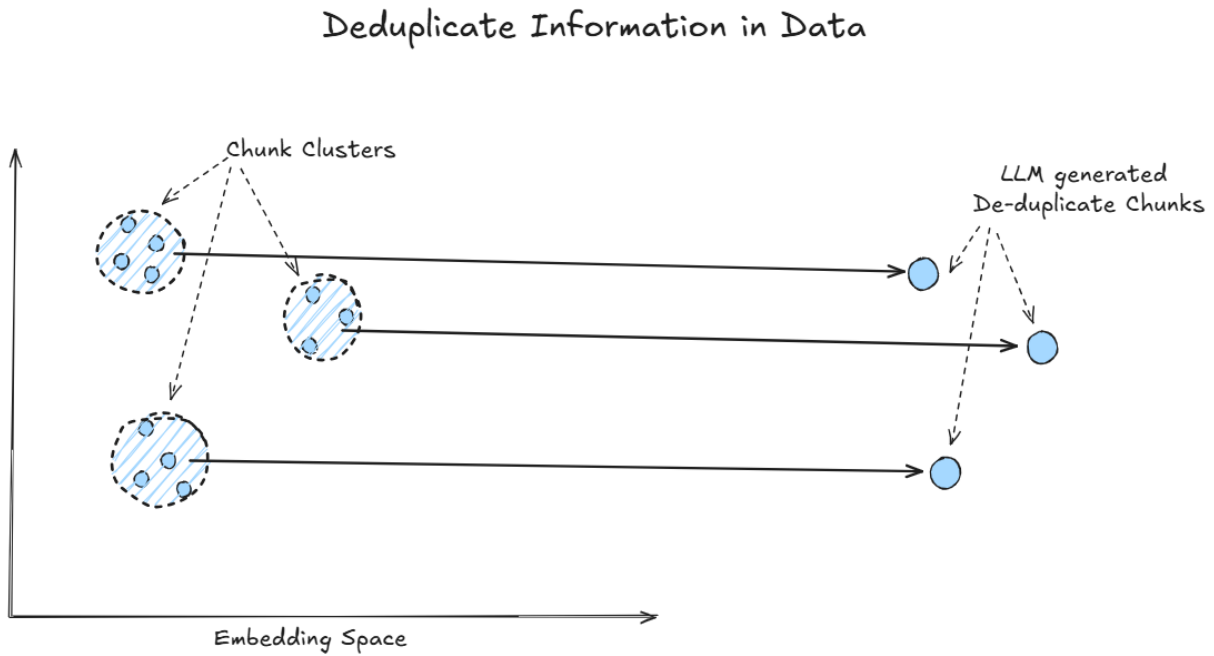
Pre-Retrieval and Data-Indexing Techniques



[2] https://www.falkordb.com/blog/advanced-rag/

Advanced RAG Techniques

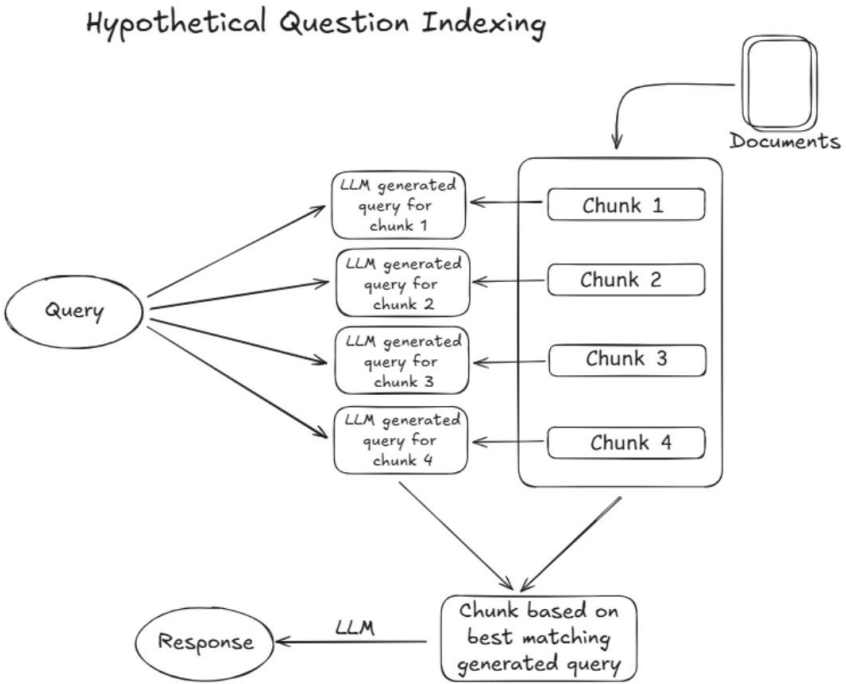
Pre-Retrieval and Data-Indexing Techniques



[2] https://www.falkordb.com/blog/advanced-rag/

Advanced RAG Techniques

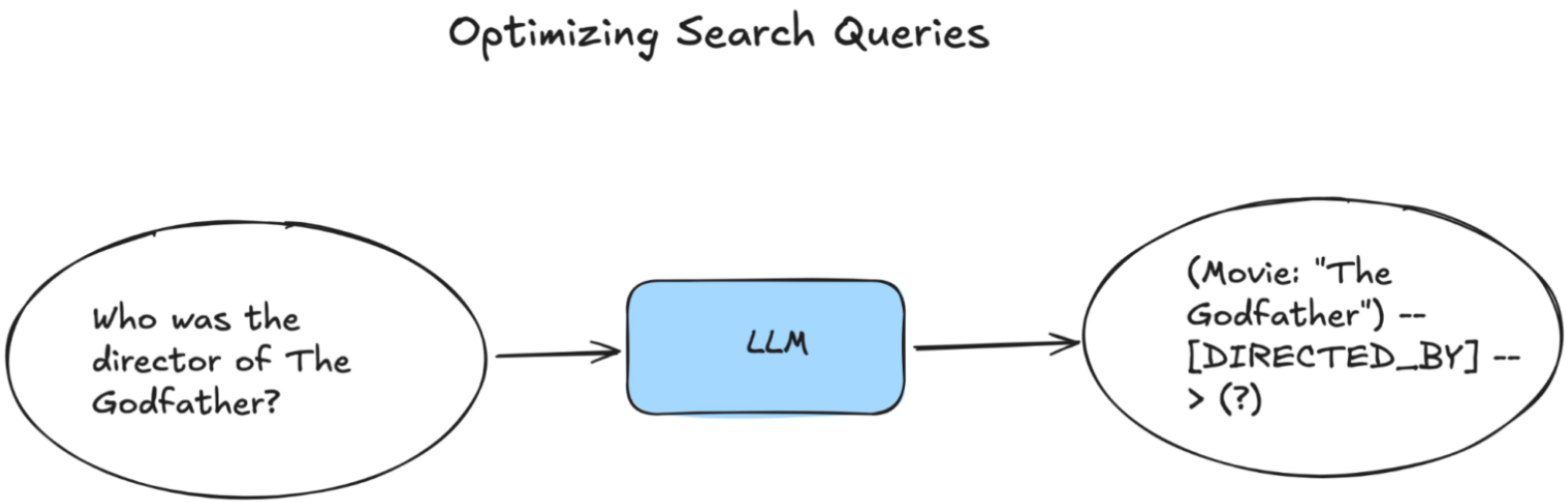
Pre-Retrieval and Data-Indexing Techniques



[2] https://www.falkordb.com/blog/advanced-rag/

Advanced RAG Techniques

Retrieval Techniques



[2] https://www.falkordb.com/blog/advanced-rag/

Advanced RAG Techniques

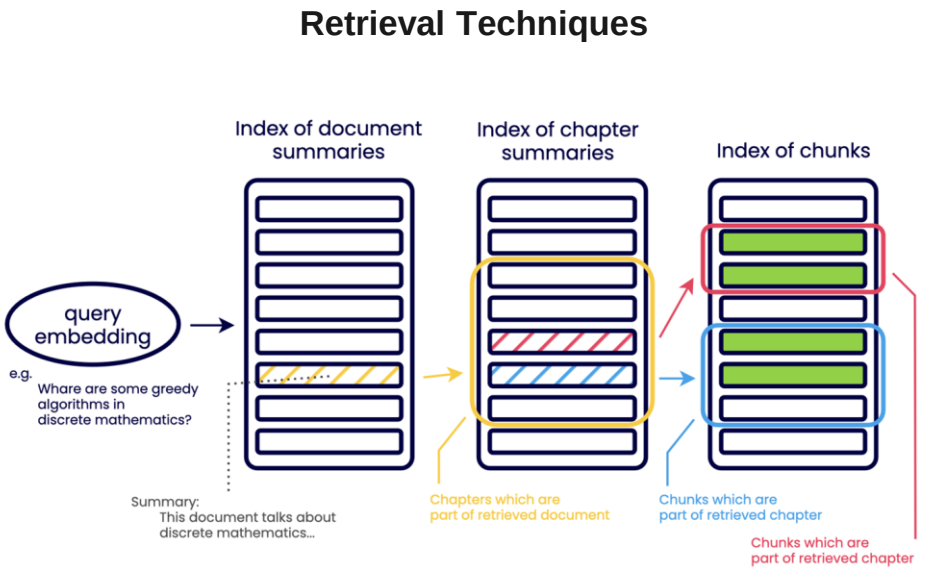
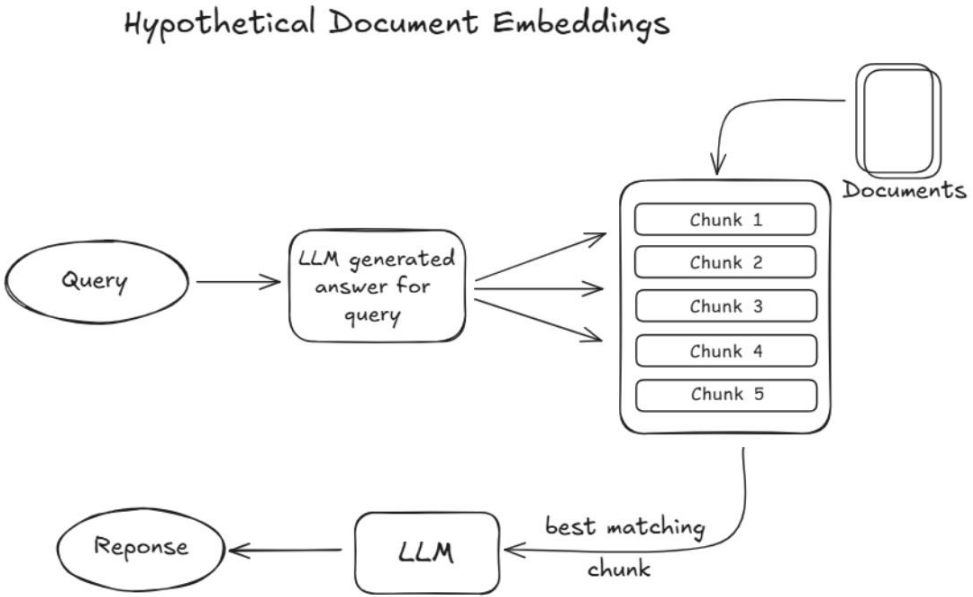


Image 2 - Progressive narrowing in hierarchical index retrieval. On each level of the hierarchy, we test our query embedding against that level's summary embedding. If it's relevant, we can continue down the hierarchy for that specific part of the document.

[2] https://www.falkordb.com/blog/advanced-rag/

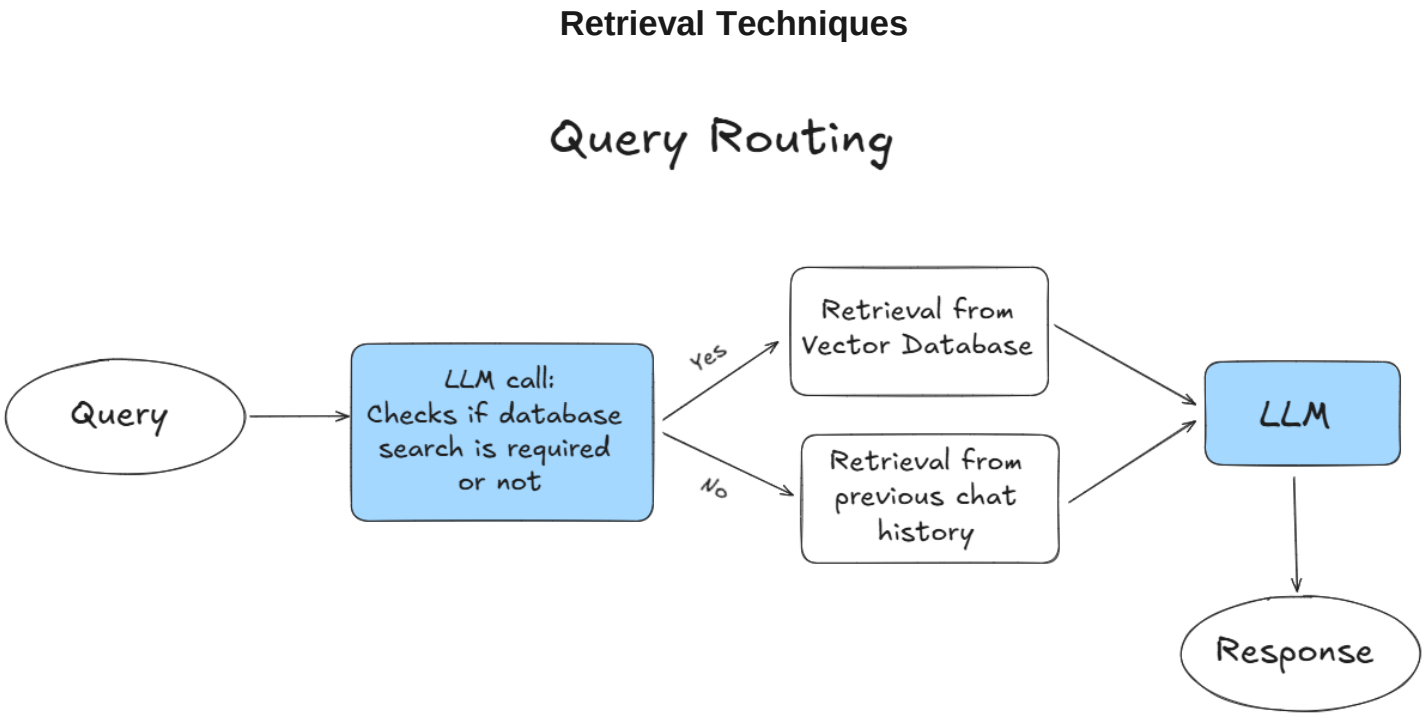
Advanced RAG Techniques

Retrieval Techniques



[2] https://www.falkordb.com/blog/advanced-rag/

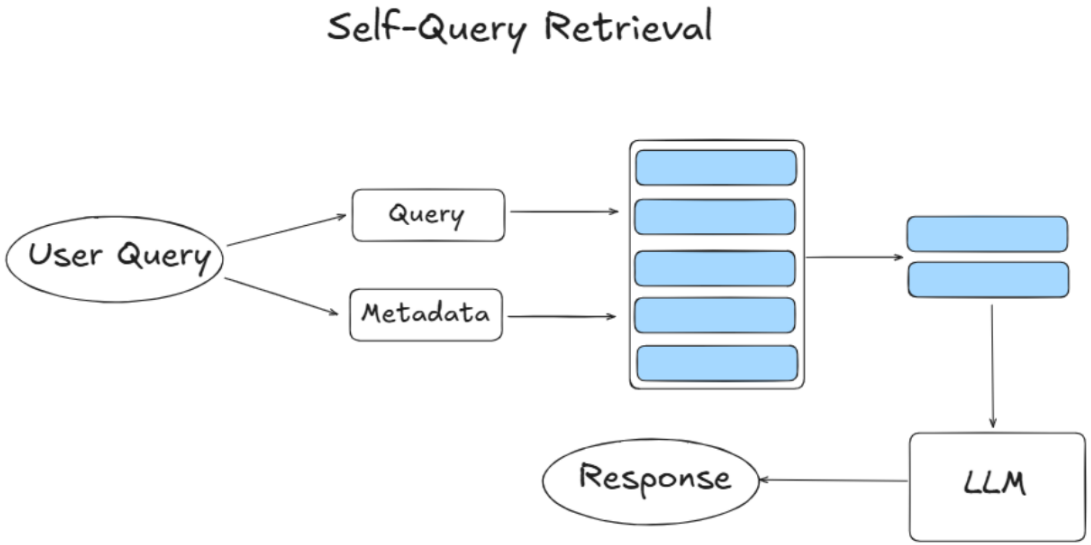
Advanced RAG Techniques



[2] https://www.falkordb.com/blog/advanced-rag/

Advanced RAG Techniques

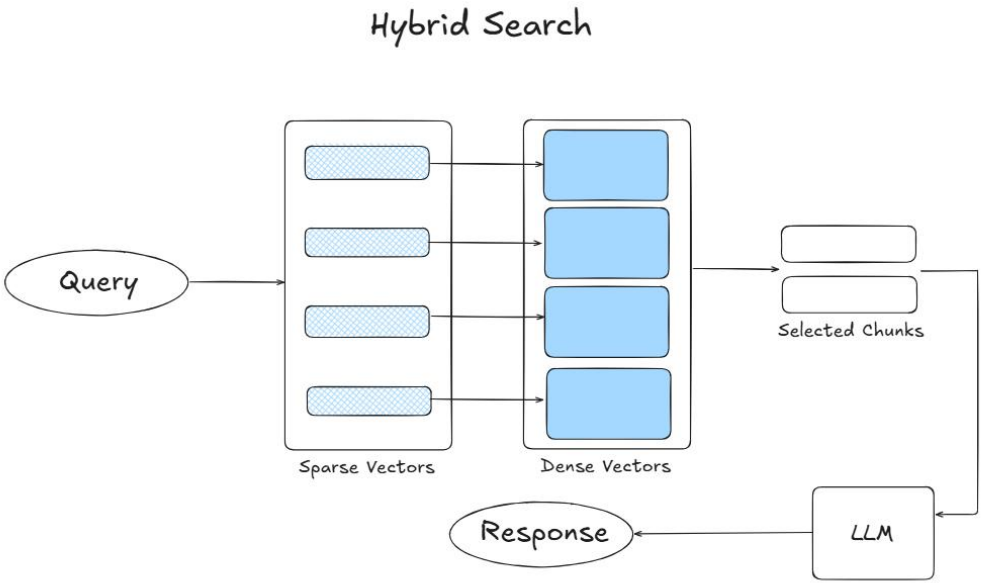
Retrieval Techniques



[2] https://www.falkordb.com/blog/advanced-rag/

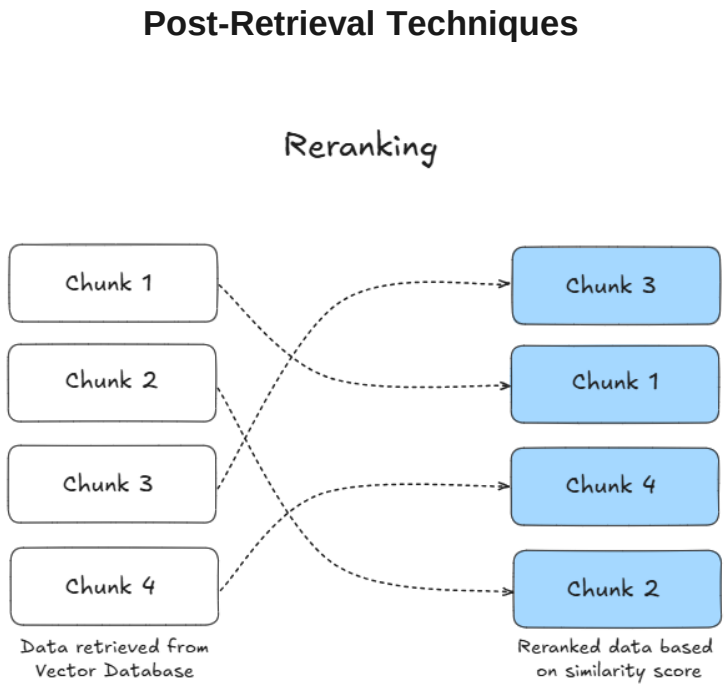
Advanced RAG Techniques

Retrieval Techniques



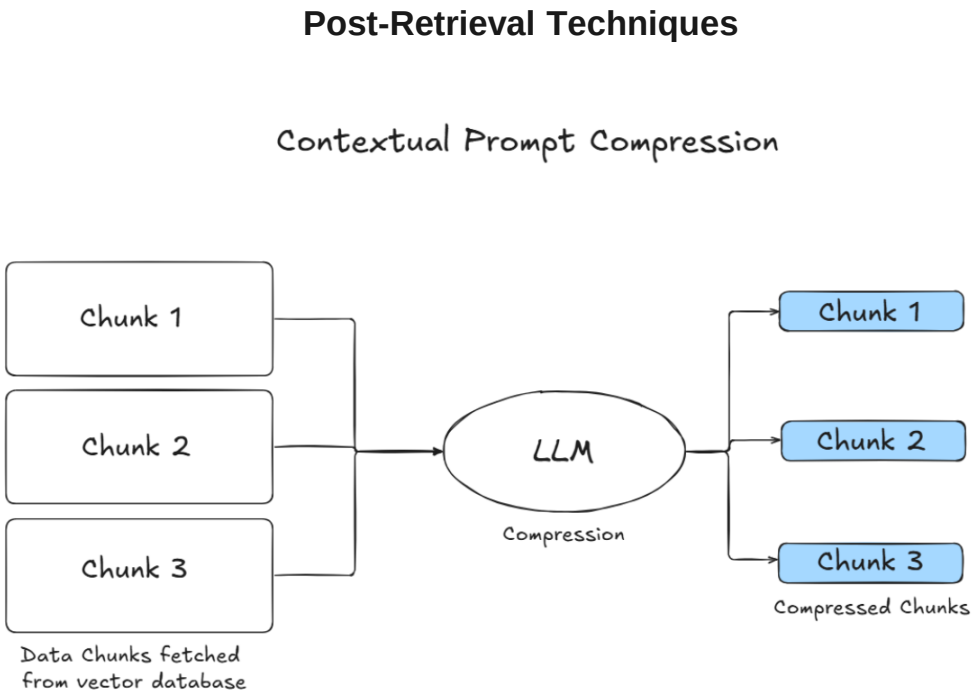
[2] https://www.falkordb.com/blog/advanced-rag/

Advanced RAG Techniques



[2] https://www.falkordb.com/blog/advanced-rag/

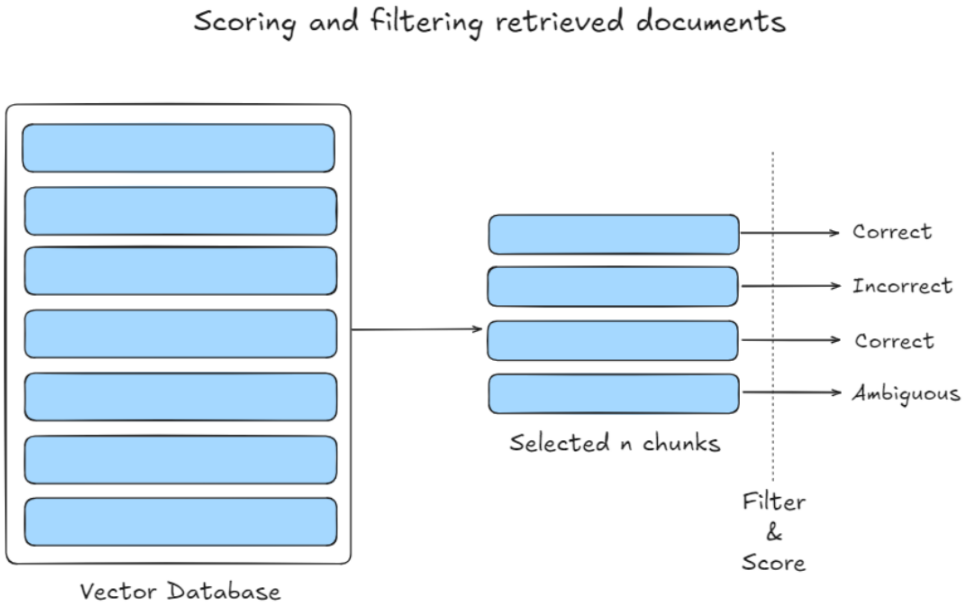
Advanced RAG Techniques



[2] https://www.falkordb.com/blog/advanced-rag/

Advanced RAG Techniques

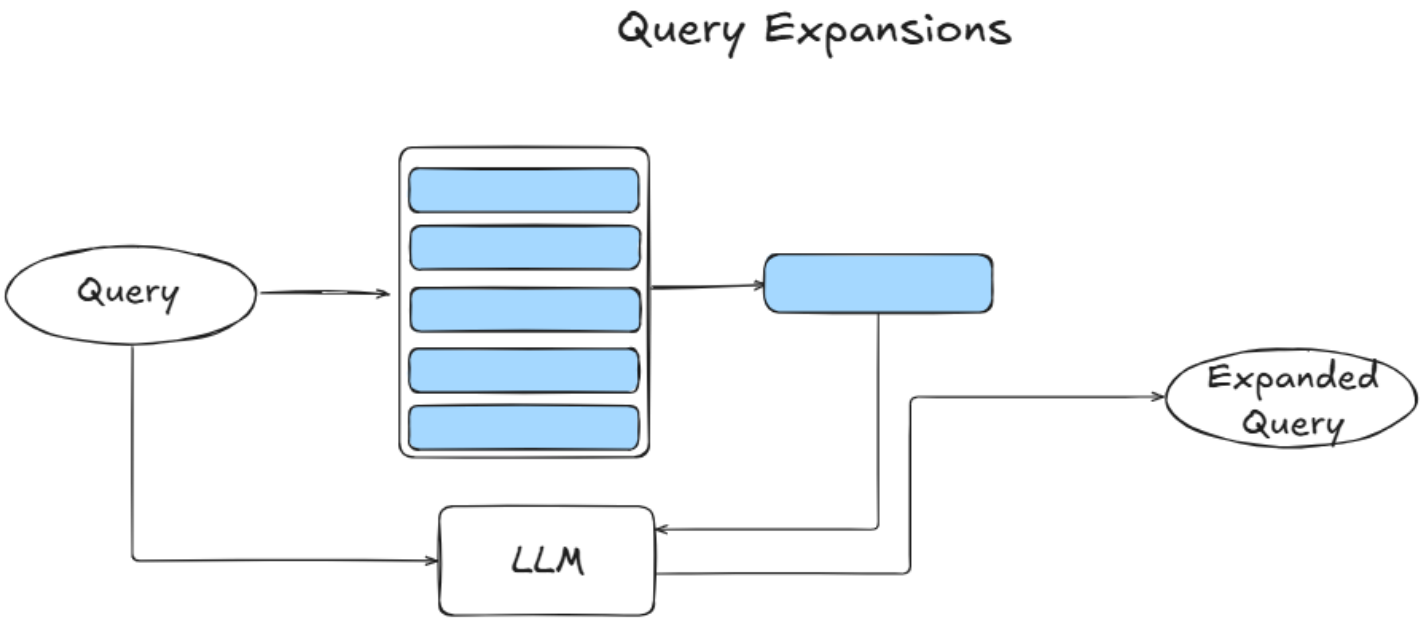
Post-Retrieval Techniques



[2] https://www.falkordb.com/blog/advanced-rag/

Advanced RAG Techniques

Post-Retrieval Techniques

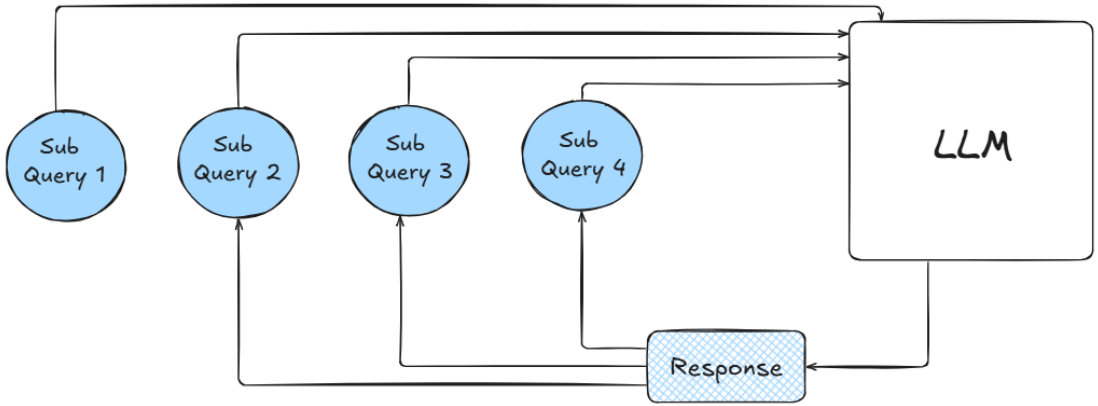


[2] https://www.falkordb.com/blog/advanced-rag/

Advanced RAG Techniques

Generation Techniques

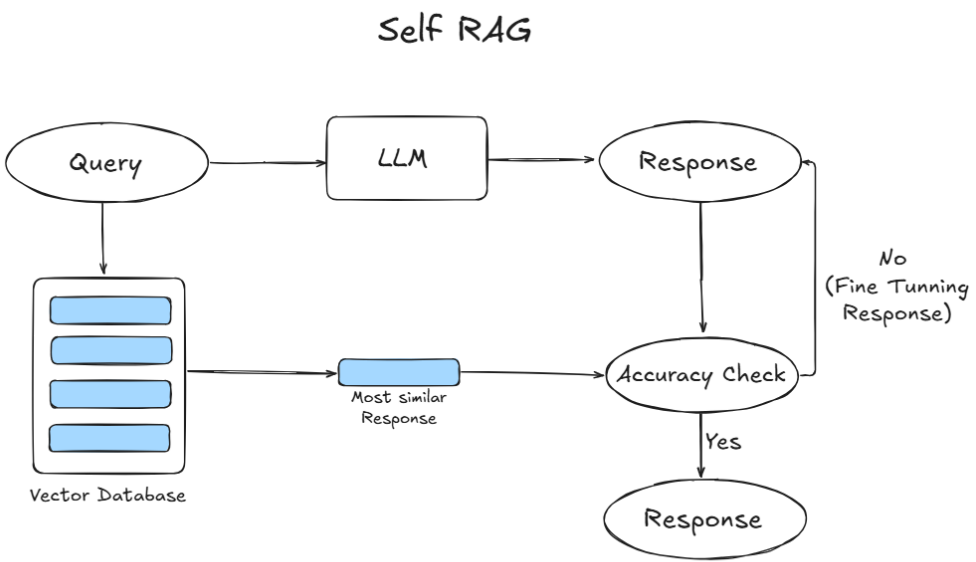
Chain of Thought prompting



[2] https://www.falkordb.com/blog/advanced-rag/

Advanced RAG Techniques

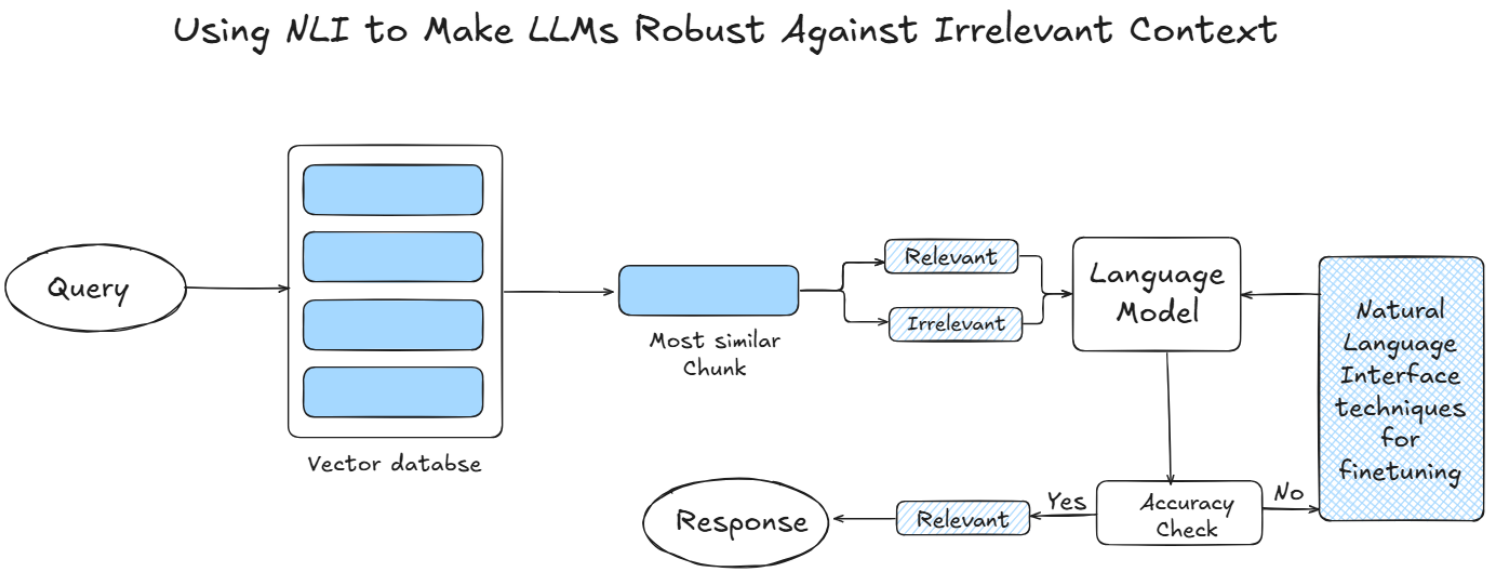
Generation Techniques



[2] https://www.falkordb.com/blog/advanced-rag/

Advanced RAG Techniques

Generation Techniques



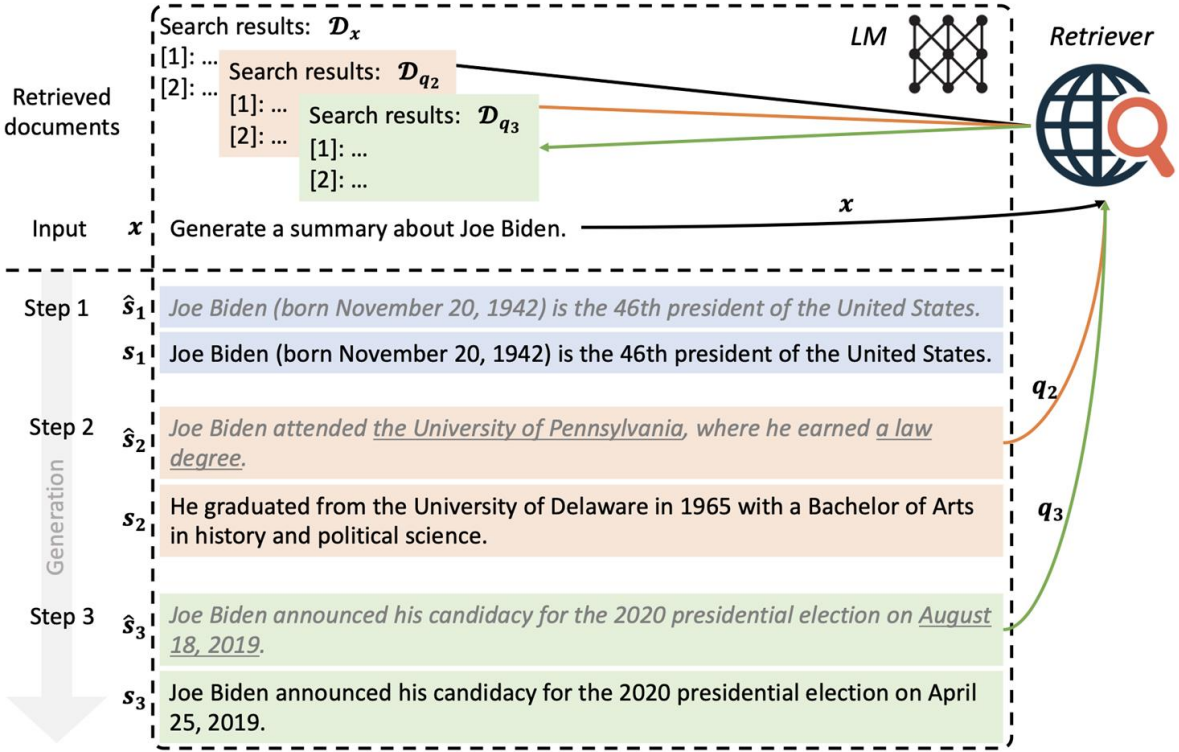
[2] https://www.falkordb.com/blog/advanced-rag/

Related Works for RAG

- FLARE [3] predicts the next sentence and uses the generated low-confidence tokens as query to re-retrieve relevant documents
- DRAGIN [4] leverages the LLM's uncertainty in its generated content to decide when to trigger retrieval based on the internal self-attention weights and corresponding keywords.
- Adaptive-RAG [5] uses a smaller LLM as a classifier to query complexity and subsequently selects the most appropriate retrieval strategy—ranging from simple to advanced
- Hypothesis Knowledge Graph Enhanced Framework (HyKGE) [6] leverages the hypothesis output and knowledge graph to enhance model inference
- GraphRAG [7] combining knowledge graph generation, retrieval-augmented generation (RAG), and query-focused summarization (QFS) to support human sensemaking over entire text corpora
- Self-RAG [8] enhances an LM's quality and factuality through API retrieval and self-reflection
- Summarizing Retrievals (SuRe) [9] constructs summaries of the retrieved passages for each of the multiple answer candidates and confirms the most plausible answer from the candidate set by evaluating the validity and ranking of the generated summaries

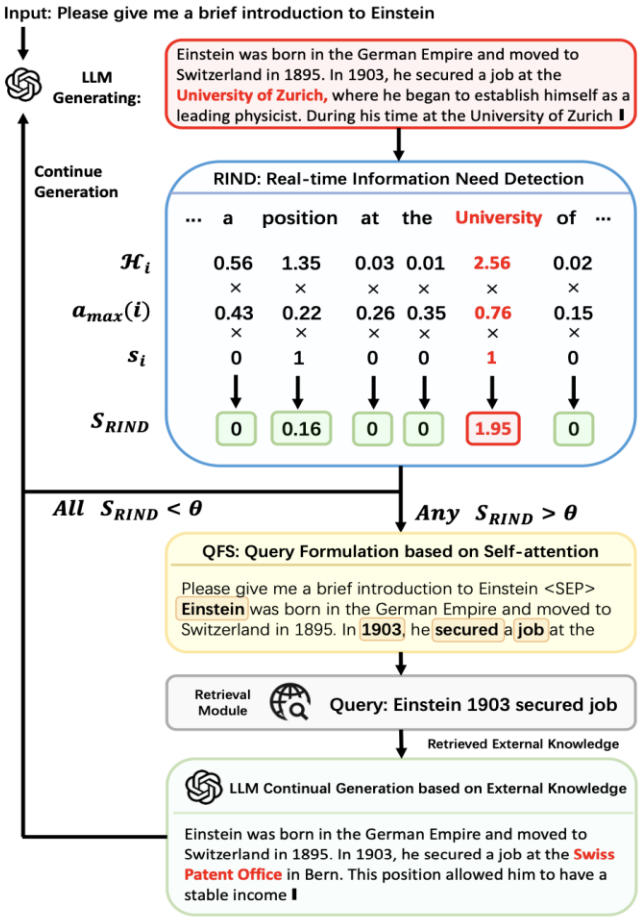
[3] Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., ... & Neubig, G. (2023). Active retrieval augmented generation. arXiv preprint arXiv:2305.06983.
[4] Su, W., Tang, Y., Ai, Q., Wu, Z., & Liu, Y. (2024). Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. arXiv preprint arXiv:2403.10081.
[5] Jeong, S., Baek, J., Cho, S., Hwang, S. J., & Park, J. C. (2024). Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. arXiv preprint arXiv:2403.14403.
[6] Jiang, X., Zhang, R., Xu, Y., Qiu, R., Fang, Y., Wang, Z., ... & Wang, Y. (2023). Think and retrieval: A hypothesis knowledge graph enhanced medical large language models. arXiv preprint arXiv:2312.15883.
[7] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., ... & Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130.
[8] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-rag: Learning to retrieve, generate, and critique through self-reflection. arXiv preprint arXiv:2310.11511.
[9] Kim, J., Nam, J., Mo, S., Park, J., Lee, S. W., Seo, M., ... & Shin, J. (2024). SuRe: Summarizing Retrievals using Answer Candidates for Open-domain QA of LLMs. arXiv preprint arXiv:2404.13081.

FLARE



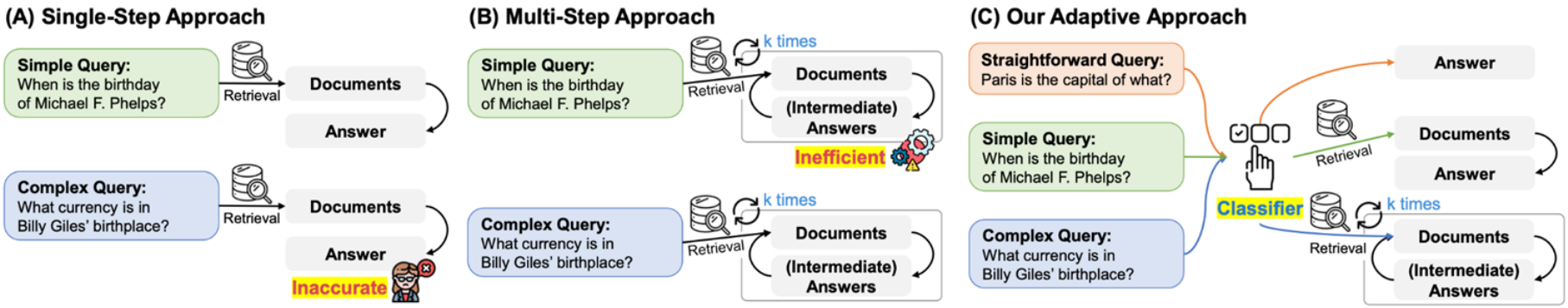
[3] Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., ... & Neubig, G. (2023). Active retrieval augmented generation. arXiv preprint arXiv:2305.06983.

DRAGIN



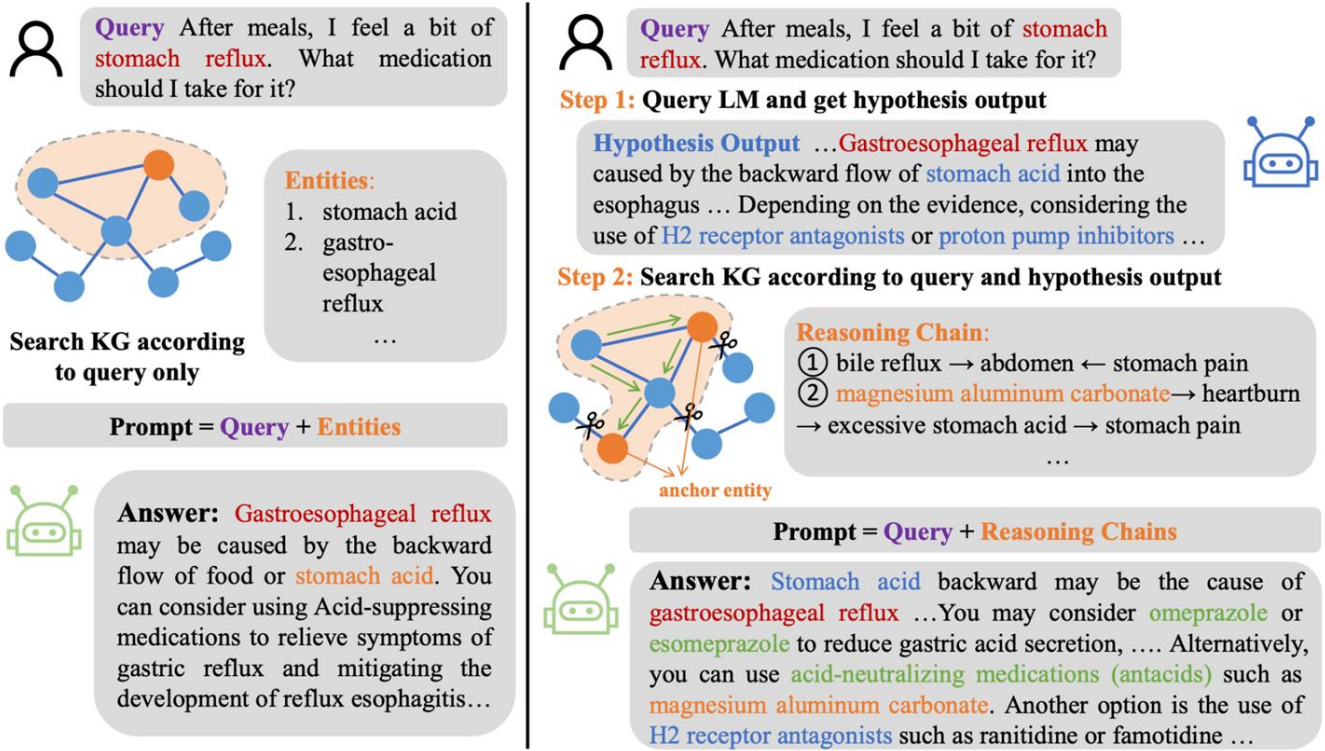
[4] Su, W., Tang, Y., Ai, Q., Wu, Z., & Liu, Y. (2024). Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. arXiv preprint arXiv:2403.10081.

Adaptive-RAG



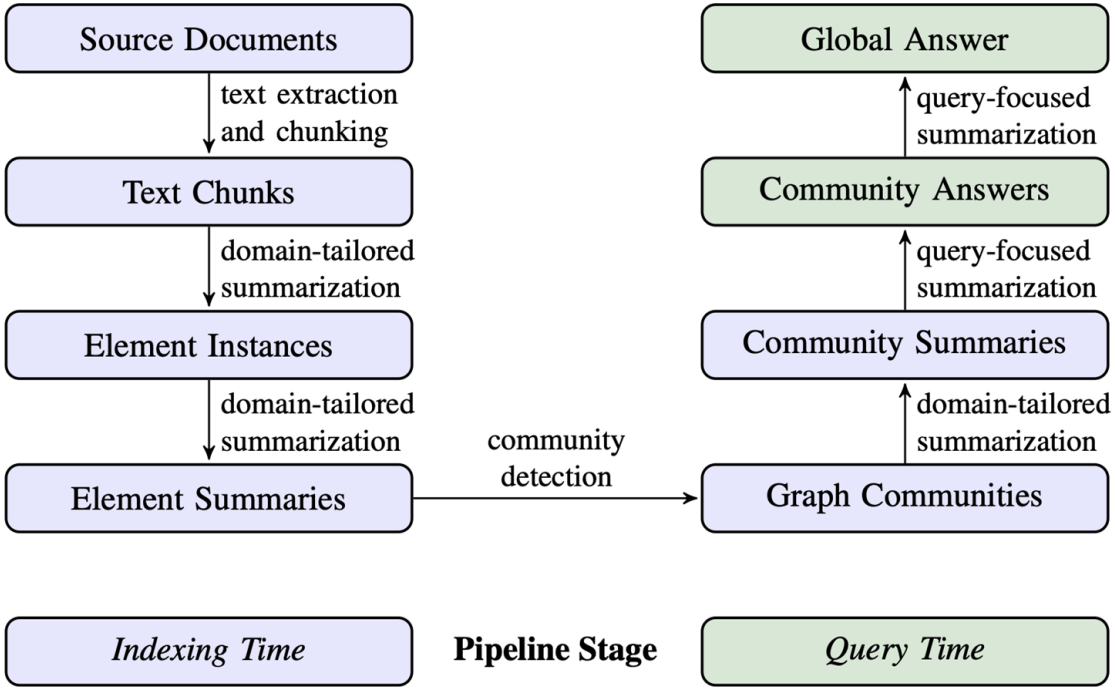
[5] Jeong, S., Baek, J., Cho, S., Hwang, S. J., & Park, J. C. (2024). Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. arXiv preprint arXiv:2403.14403.

HyKGE



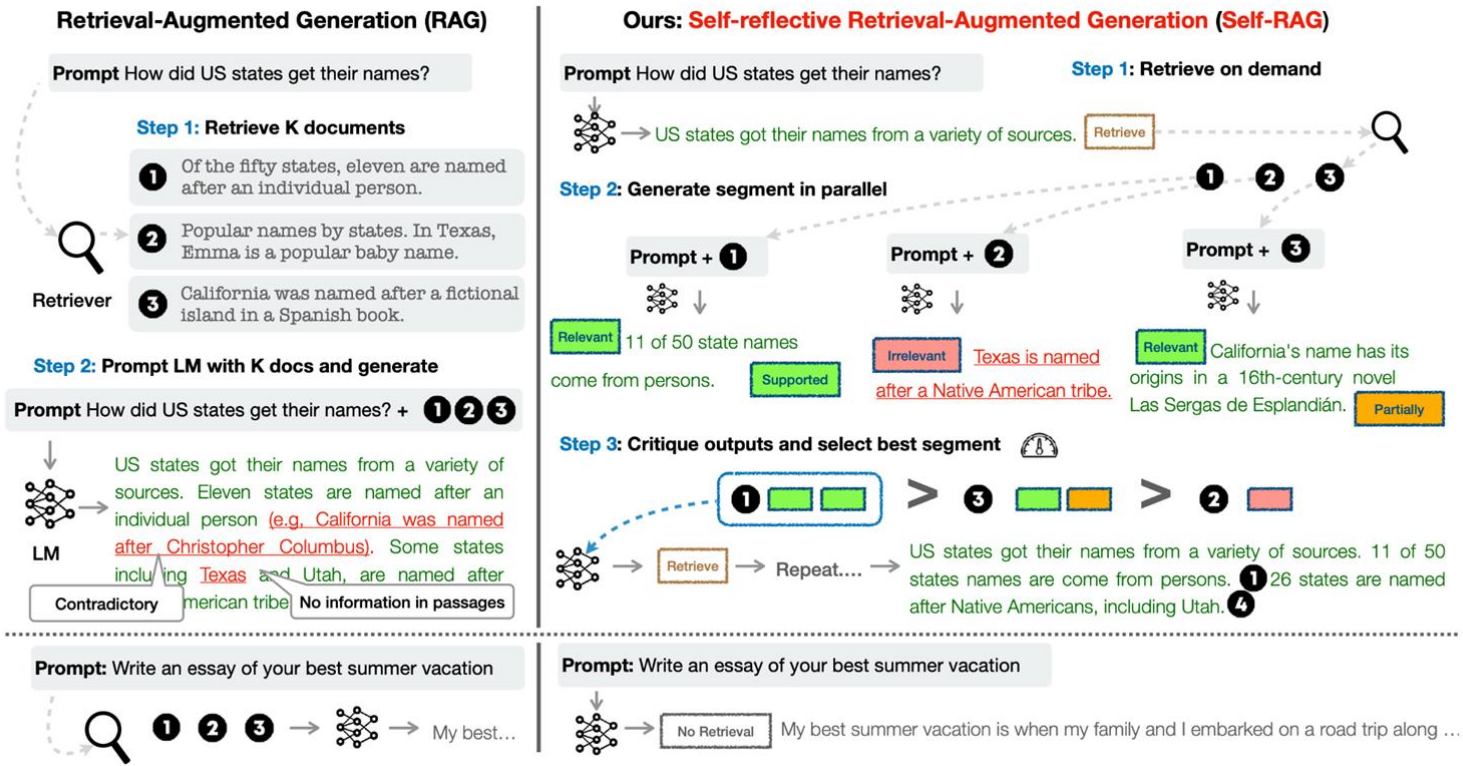
[6] Jiang, X., Zhang, R., Xu, Y., Qiu, R., Fang, Y., Wang, Z., ... & Wang, Y. (2023). Think and retrieval: A hypothesis knowledge graph enhanced medical large language models. arXiv preprint arXiv:2312.15883.

Graph-RAG



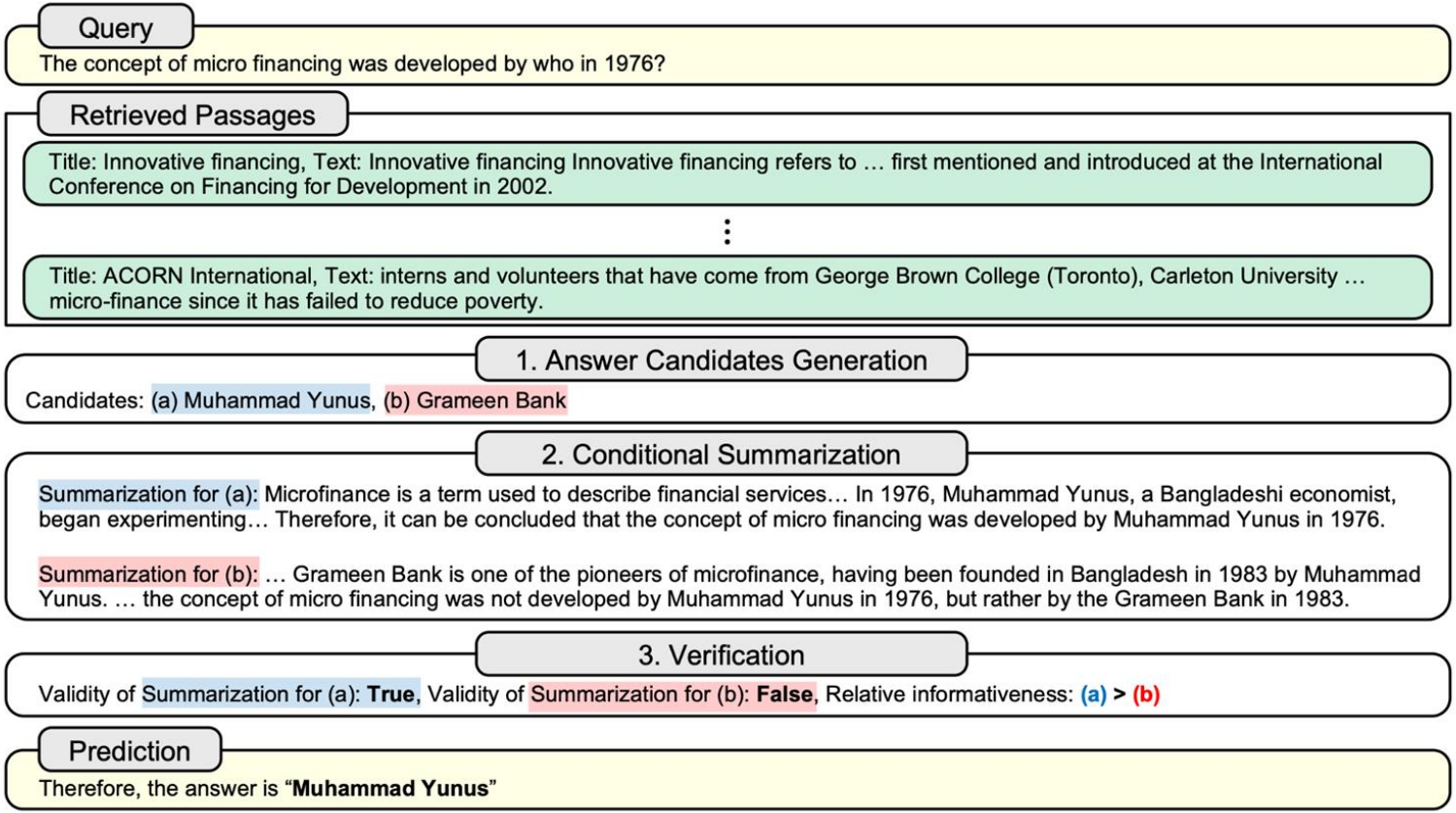
[7] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., ... & Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130.

Self-RAG



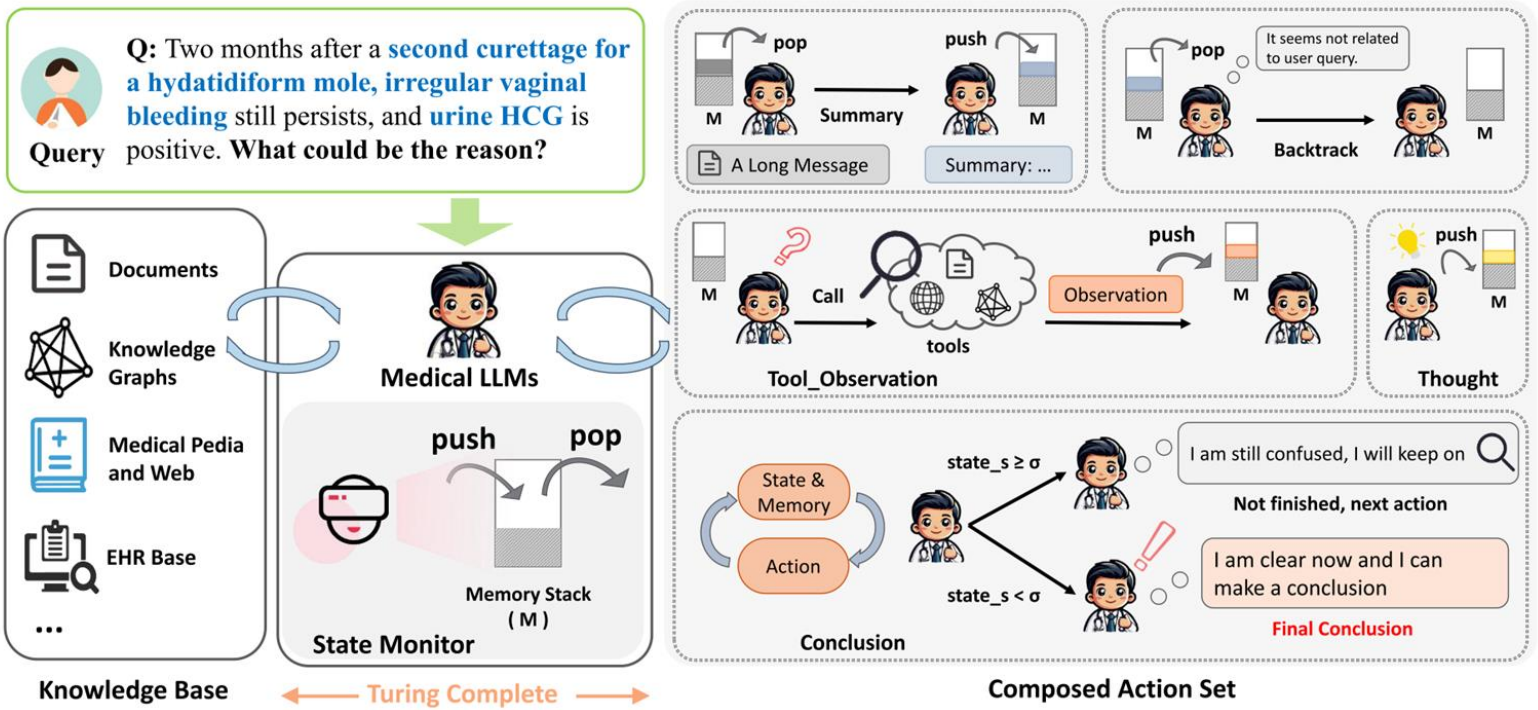
[8] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-rag: Learning to retrieve, generate, and critique through self-reflection. arXiv preprint arXiv:2310.11511.

SuRe



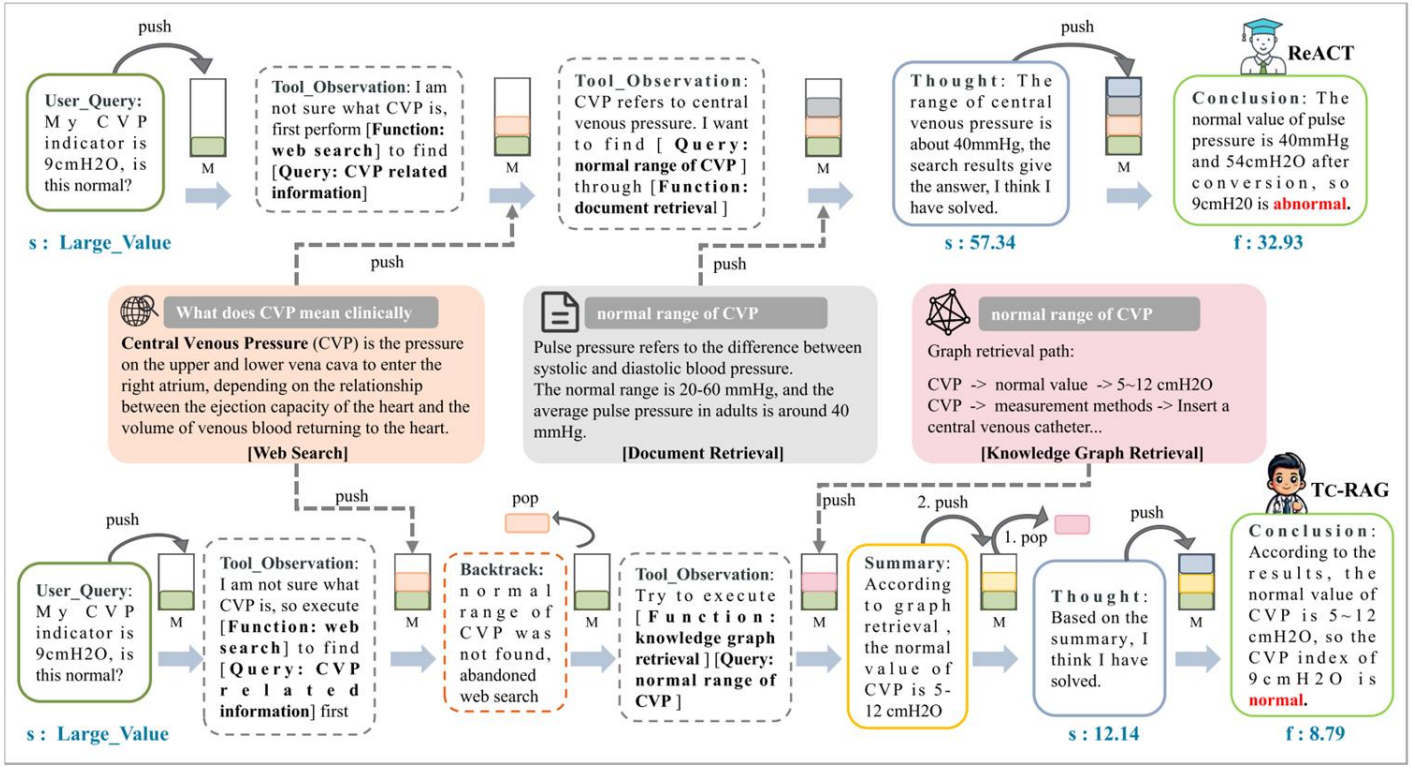
[9] Kim, J., Nam, J., Mo, S., Park, J., Lee, S. W., Seo, M., ... & Shin, J. (2024). SuRe: Summarizing Retrievals using Answer Candidates for Open-domain QA of LLMs. arXiv preprint arXiv:2404.13081.

TC-RAG



[10] Jiang, X., Fang, Y., Qiu, R., Zhang, H., Xu, Y., Chen, H., ... & Wang, Y. (2024). TC-RAG: Turing-Complete RAG's Case study on Medical LLM Systems. arXiv preprint arXiv:2408.09199.

TC-RAG

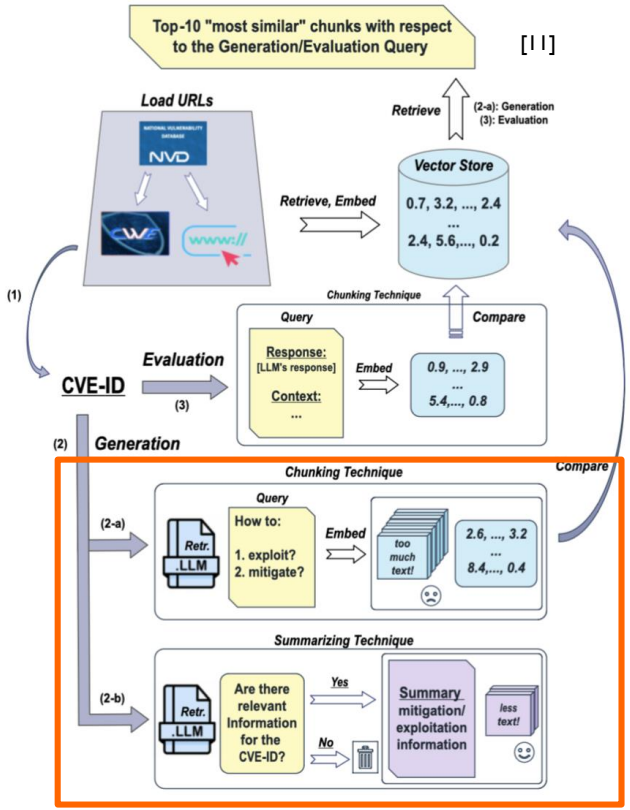


[10] Jiang, X., Fang, Y., Qiu, R., Zhang, H., Xu, Y., Chen, H., ... & Wang, Y. (2024). TC-RAG: Turing-Complete RAG's Case study on Medical LLM Systems. arXiv preprint arXiv:2408.09199.

RAG Techniques used in Cybersecurity

- Embedding Chunking [13]
- Summarization [13]
- Similar Description Retrieval [14]
- More Advanced RAG Techniques:
  https://medium.com/decodingml/the-4-advanced-rag-algorithms-you-must-know-to-implement-5d0c7f1199d2

[12]



[11]

[11] Fayyazi, R., Trueba, S. H., Zuzak, M., & Yang, S. J. (2024). ProveRAG: Provenance-Driven Vulnerability Analysis with Automated Retrieval-Augmented LLMs. arXiv preprint arXiv:2410.17406.
[12] Fayyazi, R., Taghdimi, R., & Yang, S. J. (2023). Advancing TTP Analysis: Harnessing the Power of Large Language Models with Retrieval Augmented Generation. arXiv preprint arXiv:2401.00280.

Thank you!

Reza Fayyazi
rf1679@rit.edu