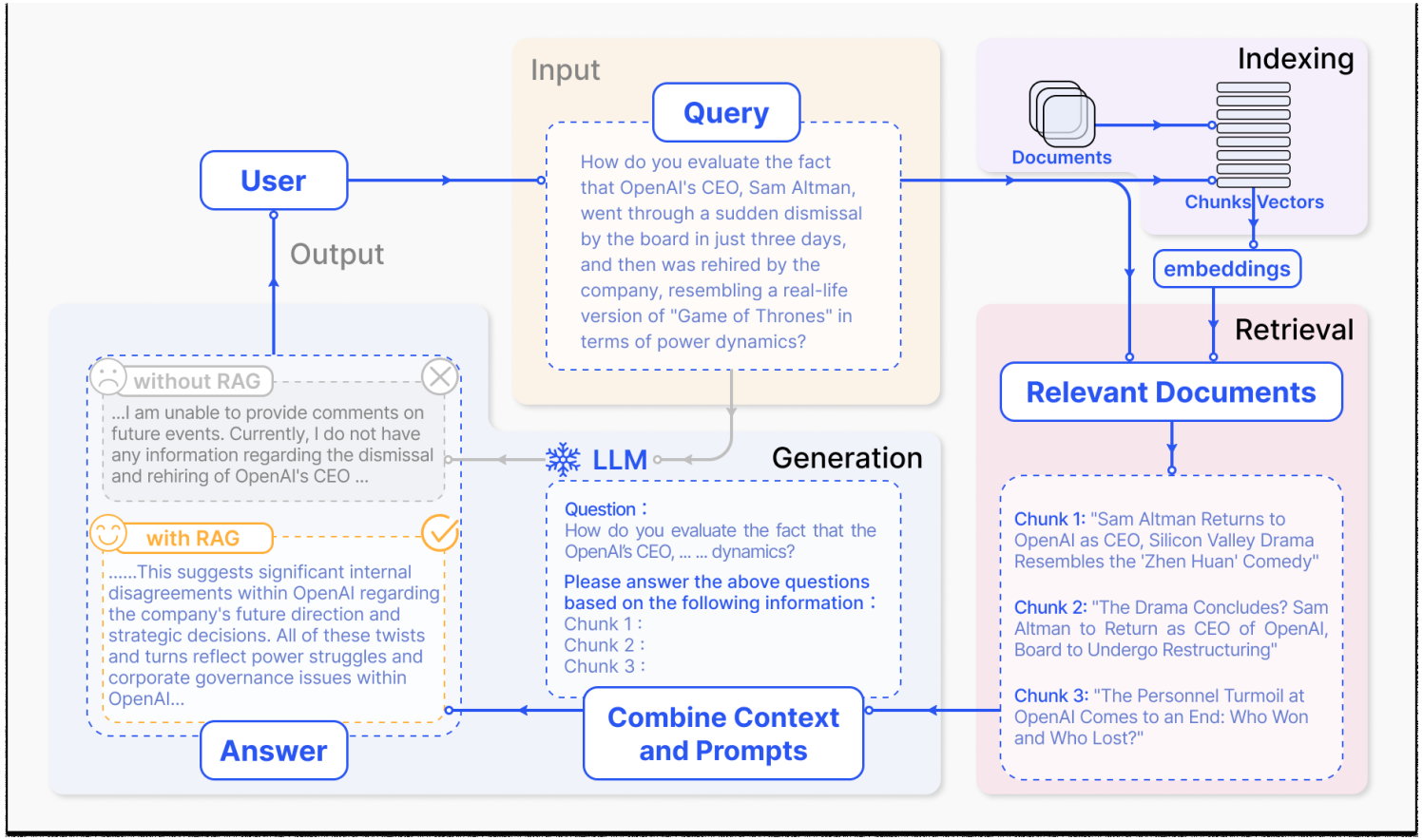
Retrieval Augmented Generation (RAG)

- Data used to trains LLMs often lags in timeliness and may not cover all domain knowledge
  - Hallucinations
  - Outdated Knowledge
  - Non-transparent and untraceable reasoning
- Data augmented LLMs can
  - exhibit domain expert-level capabilities like doctors, lawyers
  - generate responses based on real data – minimizing the possibility of hallucinations
- Data augmented LLMs application primarily follows:
  - RAG
  - Fine-tuning

Retrieval Augmented Generation

- RAG enhances accuracy and credibility of the generation – merging LLM's intrinsic knowledge (training data) with the vast, dynamic repo of external databases leveraging –
  - Vector database, Embeddings, Sentence transformers, Cosine distance



Slide credit - Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Retrieval Augmented Generation

- **Naïve RAG –** The earliest RAG method, follows a "Retrieve-Read" framework involving indexing, retrieval, and generation
  - **Indexing Process:** Raw data from various formats (PDF, HTML, Word) is cleaned, segmented, and converted into vectors using an embedding model, stored in a vector database.
  - **Retrieval Phase:** The system encodes user queries into vectors, matches them against stored vectors, and retrieves the most relevant chunks for the expanded prompt context.
  - **Generation Phase:** The model generates responses by synthesizing the query and retrieved documents

Slide credit - Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Retrieval Augmented Generation

- **Drawbacks –**
  - **Retrieval Challenges:** Precision and recall issues may lead to irrelevant or missing chunks.
  - **Generation Issues:** Hallucination, irrelevance, bias, or toxic content can arise in the generated responses.
  - **Augmentation Hurdles:** Combining retrieved information can be difficult, leading to redundancy, disjointed responses, and over-reliance on the retrieved content.

Slide credit - Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.