LLM Apps

Optimizing RAG Apps

Optimizing RAG Systems

- Once you have metrics to measure the performance improvement, you can proceed with the RAG system optimization.

- From simpler and cheaper to advanced and expensive RAG optimization steps:
  - Initial optimization techniques.
  - Advanced Retrieval Methods.
  - Fine-tuning.
  - Use agents.

Initial optimization techniques

- Better parsers.
- Chunk sizes.
- Hybrid search.
- Metadata filters.

Initial optimization techniques: notes about chunk sizes

- Tuning your chunk sizes can have impacts on performance.
- More retrieved tokens does not always equal higher performance.
- Reranking (shuffling context order) isn’t always beneficial.

Initial optimization techniques: notes about Metadata

- Metadata: context you can inject into each text chunk.
- It is like a structured JSON dictionary.
- Examples of data included on metadata:
  - Page number.
  - Document title, year.
  - Summary of adjacent chunks.
  - Questions that chunk can answer (reverse HyDE).
- Benefits:
  - Can help retrieval.
  - Can augment response quality.
  - Integrates with vector DB metadata filters.

Advanced Retrieval Methods

- Reranking.
- Recursive retrieval.
- Embedded tables.
- Small-to-big retrieval.

Advanced Retrieval Methods: notes on Small-to-Big

- Intuition: embedding a big text chunk feels suboptimal.
- Solutions:
  - Embed text at the sentence-level, then expand that window during LLM synthesis.
  - Embed a reference to the parent chunk. Use parent chunk for synthesis.
- This leads to more precise retrieval and avoids “lost in the middle” problems.

Fine-tuning

- Embedding fine-tuning.
- LLM fine-tuning.

Fine-tuning: notes on embedding fine-tuning

- Intuition: embedding representations are not optimized over your dataset.

- Solution: generate a synthetic query dataset from raw text chunks using LLMs. Use this synthetic dataset to finetune an embedding model.

Use agents.

- Routing.
- Query planning.
- Multi-document agents.

Use Agents: Notes on Multi-Document Agents

- Intuition: there's certain questions that "top-k" RAG can't answer.

- Solution: multi-document agents.
  - Fact-based QA and Summarization over any subsets of documents.
  - Chain-of-thought and query planning.