LLM Apps

Challenges of the RAG Technique

Challenges of RAG

- Challenges in the retrieval process.
- Challenges in the response process.

Challenges in the retrieval process.

- Low precision: not all chunks in retrieved set are relevant
  - Hallucination + Lost in the Middle problems.
  - You have a lot of “fluff” in the returned response.
- Low recall: now all relevant chunks are retrieved.
  - Lacks enough context for LLM to synthesize an answer.
- Outdated information: the data is redundant or out of date.

Challenges in the response process

- Hallucination: model makes up an answer that isn’t in the context.
- Irrelevance: model makes up an answer that doesn’t answer the question.
- Toxicity/Bias: model makes up an answer that is harmful/offensive.

Ways to overcome the challenges.

- You can improve things in all stages of the process.
  - Data.
  - Embeddings.
  - Retrieval.
  - Synthesis (response generation)

- Before introducing improvements in all these areas you need to have metrics ready for being able to measure the impact of these changes in the performance of the application.

Improving data

- Can you store additional information beyond raw text chunks?
  - Play around with chunk sizes.

Improving embeddings

- Can you optimize the embedding representations?
  - The default settings can be improved.

Improving retrieval

- Can we do better than top-k embedding lookup?

Improving synthesis (response generation)

- Can you use LLMs for more than generation? You can use LLM for reasoning as opposed to pure generation. Examples:
  - Given a question, can you break it down into simpler questions?
  - Route to different data sources?
  - Have a more sophisticated way of querying your data?