LLM Apps

Evaluation: Measuring RAG Performance

Evaluation

- How can we evaluate a RAG system?
  - We can evaluate each process in isolation: retrieval process and response (synthesis) process.
  - And we can evaluate the whole system end-to-end.

Evaluation in isolation of the retrieval process

- Evaluation of the quality of the retrieved chunks given one user query.
- First, you need to create an evaluation dataset.
  - You can use human-labelled datasets
  - or user feedback if you have the app in production
  - or create it synthetically.
- Run retriever over dataset.
  - Input: query.
  - Output: the “ground-truth” documents relevant to the query, the IDs of the returned outputs.
- Measure ranking metrics.
  - Success rate / hit-rate.
  - MRR (Mean Reciprocal Rank), NDCG (Normalized Discounted Cumulative Gain).
  - Hit-rate.

Evaluation of the whole RAG system: E2E evaluation

- Evaluation of the quality of the final output given one particular input.
- Create dataset.
  - Input: query.
  - (Optional) output: the “ground-truth” answer.
- Run through the full RAG pipeline.
- Collect evaluation metrics.
  - If no labels: label-free evals.
    - Metrics: faithfulness, relevancy, adheres to guidelines, toxicity-free.
  - If labels: with-label evals.
    - Metrics: correctness, etc.