arize

LLM Evaluation Essentials:
Benchmarking & Analyzing Retrieval Approaches

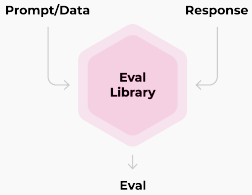
Jason Lopatecki
Co-Founder and CEO
Arize AI

Sally-Ann DeLucia
ML Solutions Engineer
Arize AI

5 Pillars of LLM Observability

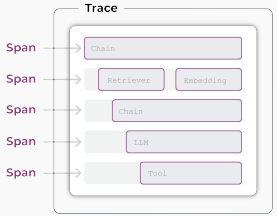
Evaluation

Evaluations of LLM outputs by using a separate evaluation LLM



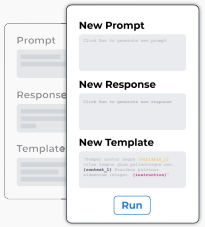
Traces & Spans

Visibility into where the agentic workflow broke



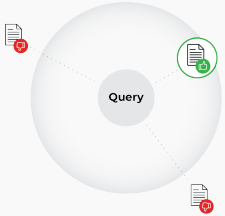
Prompt Engineering

Iterating on prompt templates for improved results



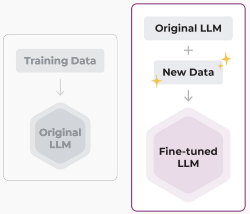
Search & Retrieval

Locate and improve retrieved context



Fine-tuning

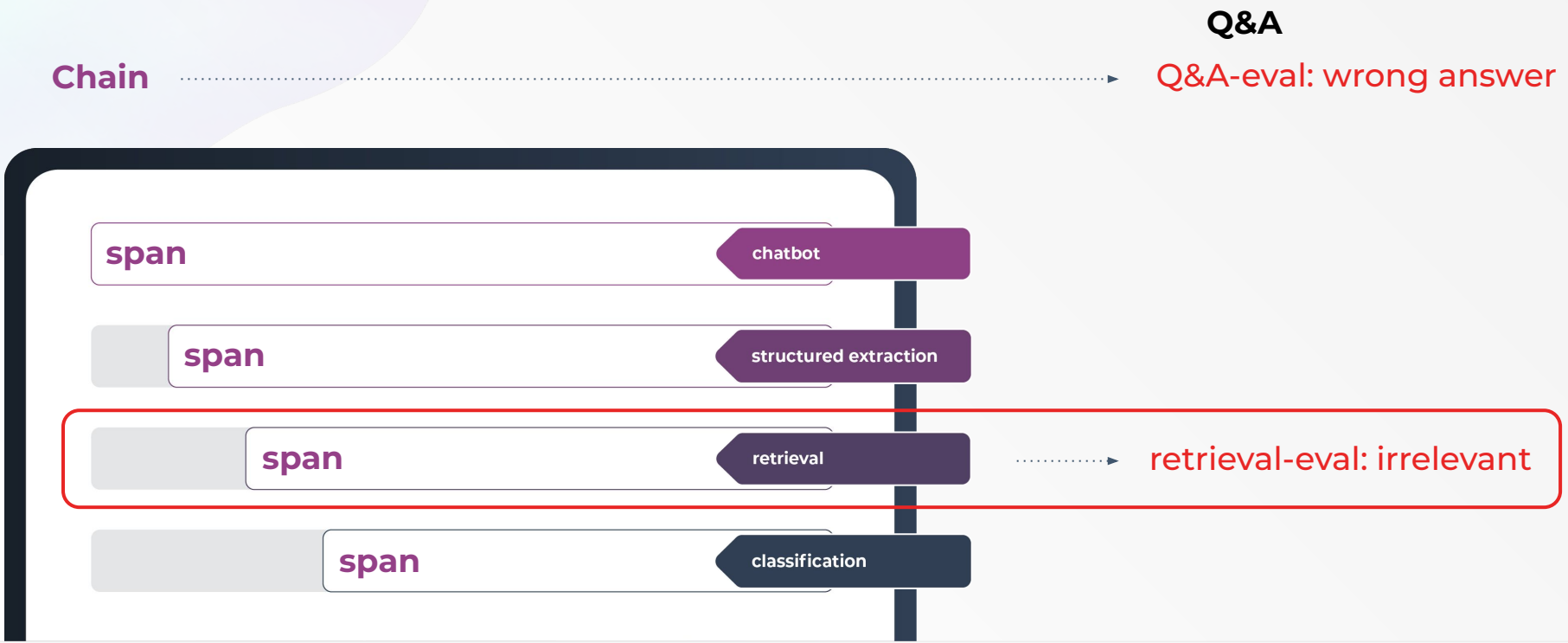
Re-train LLM on use case / company data



Evaluations Building Blocks

What are Evals?

Evals library allows teams to generate evals for either chains or individual spans.



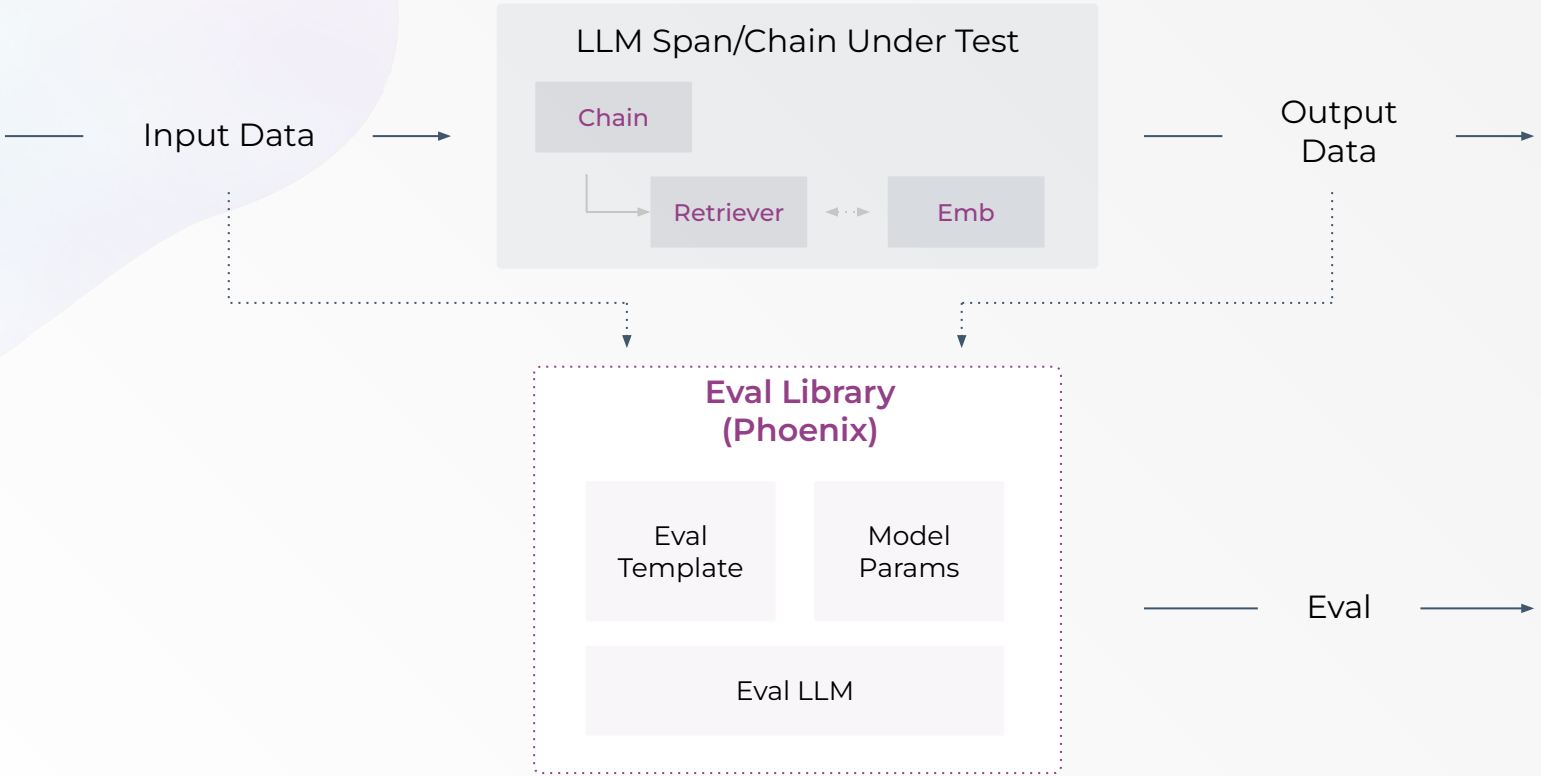
Fundamentals of LLM Production Evals

Benchmark with Golden dataset USE Precision/Recall/F1 For Base Metrics

Eval are Task Based: Templates + Dataset combinations define performance

Evals need run across environments (Python, Notebooks and LangChain/LlamaIndex) and need to be as fast as possible

LLM Evals in Production



How do Evals work?

Span we want to evaluate

retrieval span

Input

User Query

Documents

Output

Eval Template

You are comparing a reference text to a question and trying to determine if the reference text contains information relevant to answering the question. Here is the data:
[BEGIN DATA]
****
[Question]: {query}
****
[Reference text]: {reference}
[END DATA]

Compare the Question above to the Reference text. Determine whether the Reference text contains information that can answer the Question.

Phoenix Library

Eval Template
Example: Retrieval

Model Params

Eval LLM

Eval

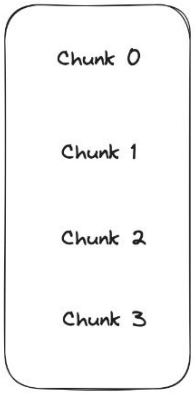
Example
“relevant”
“irrelevant”

Building Blocks: Retrieval Eval

Retrieval Eval per Chunk

Is Chunk Relevant?

Query 0



You are comparing a reference text to a question and trying to determine if the reference text contains information relevant to answering the question. Here is the data:

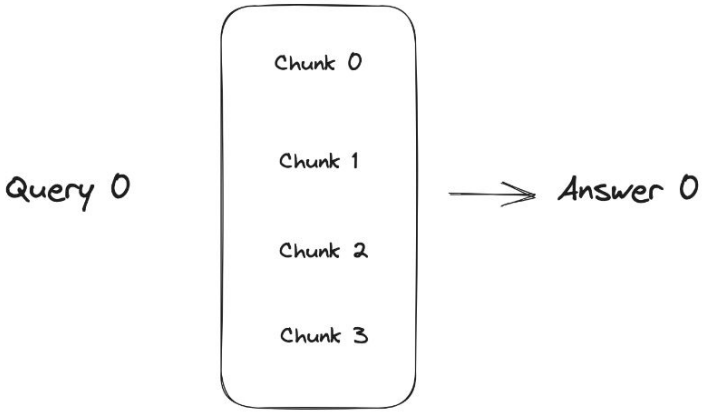
```
[BEGIN DATA]
************
[Question]: {query}
************
[Reference text]: {reference}
[END DATA]
```

Compare the Question above to the Reference text. You must determine whether the Reference text contains information that can answer the Question. Please focus on whether the very specific question can be answered by the information in the Reference text.
Your response must be single word, either "relevant" or "irrelevant",
and should not contain any text or characters aside from that word.
"irrelevant" means that the reference text does not contain an answer to the Question.
"relevant" means the reference text contains an answer to the Question.

Building Blocks: Q&A Eval

Question & Answer Eval

Is Answer Correct?



You are given a question, an answer and reference text. You must determine whether the given answer correctly answers the question based on the reference text. Here is the data:

[BEGIN DATA]
****
[Question]: {question}
****
[Reference]: {context}
****
[Answer]: {sampled_answer}
[END DATA]

Your response must be a single word, either "correct" or "incorrect",
and should not contain any text or characters aside from that word.
"correct" means that the question is correctly and fully answered by the answer.
"incorrect" means that the question is not correctly or only partially answered by the answer.

Arize LLM Evals and Golden Datasets

Retrieval Eval

- MS Marco
- WikiQA

Hallucination Eval

- Hallucination QA Dataset
- Hallucination RAG Dataset

User Frustration

- User Frustration Dataset

Q&A Eval

- WikiQA

Summarization Eval

- GigaWord
- CNNDM
- Xsum

Code Generation

- WikiSQL
- HumanEval
- CodeXGlu

Targets Precision

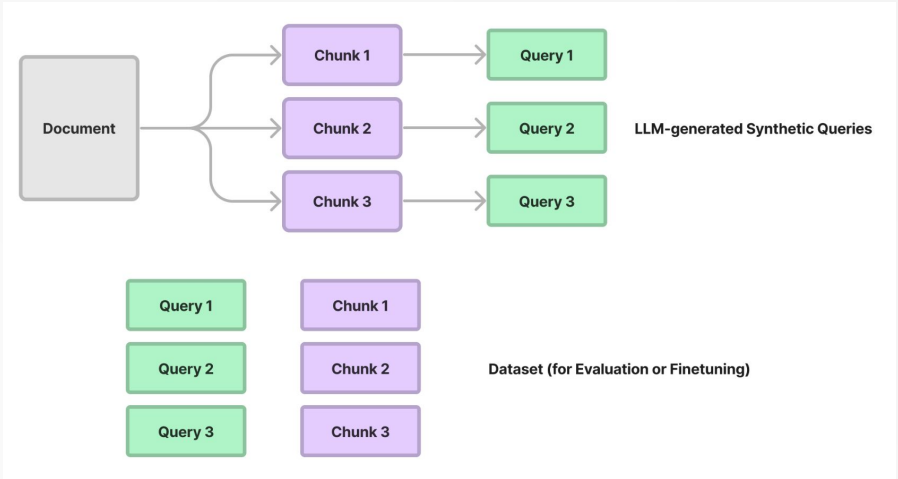
70–90%

Targets F1

70–85%

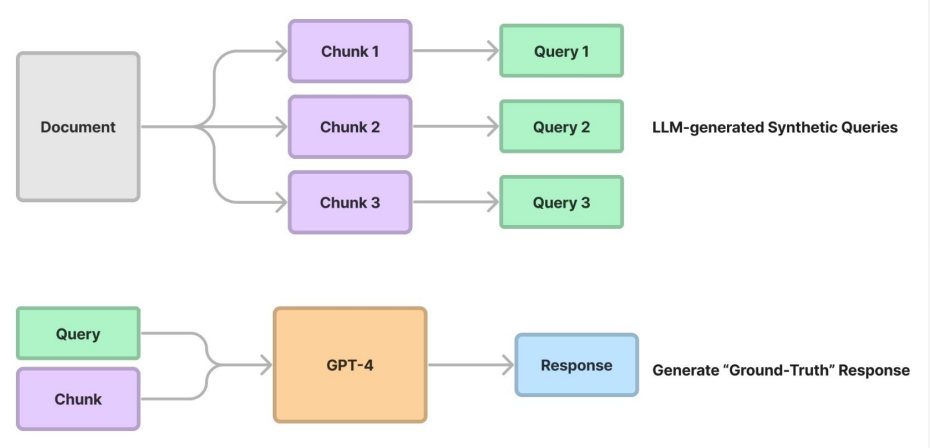
Synthetic Dataset Generation for Retrieval Evals

1. Parse / chunk up text corpus
2. Prompt GPT-4 to generate questions from each chunk (or subset of chunks)
3. Each (question, chunk) is now your evaluation pair!



Synthetic Dataset Generation for E2E Evals

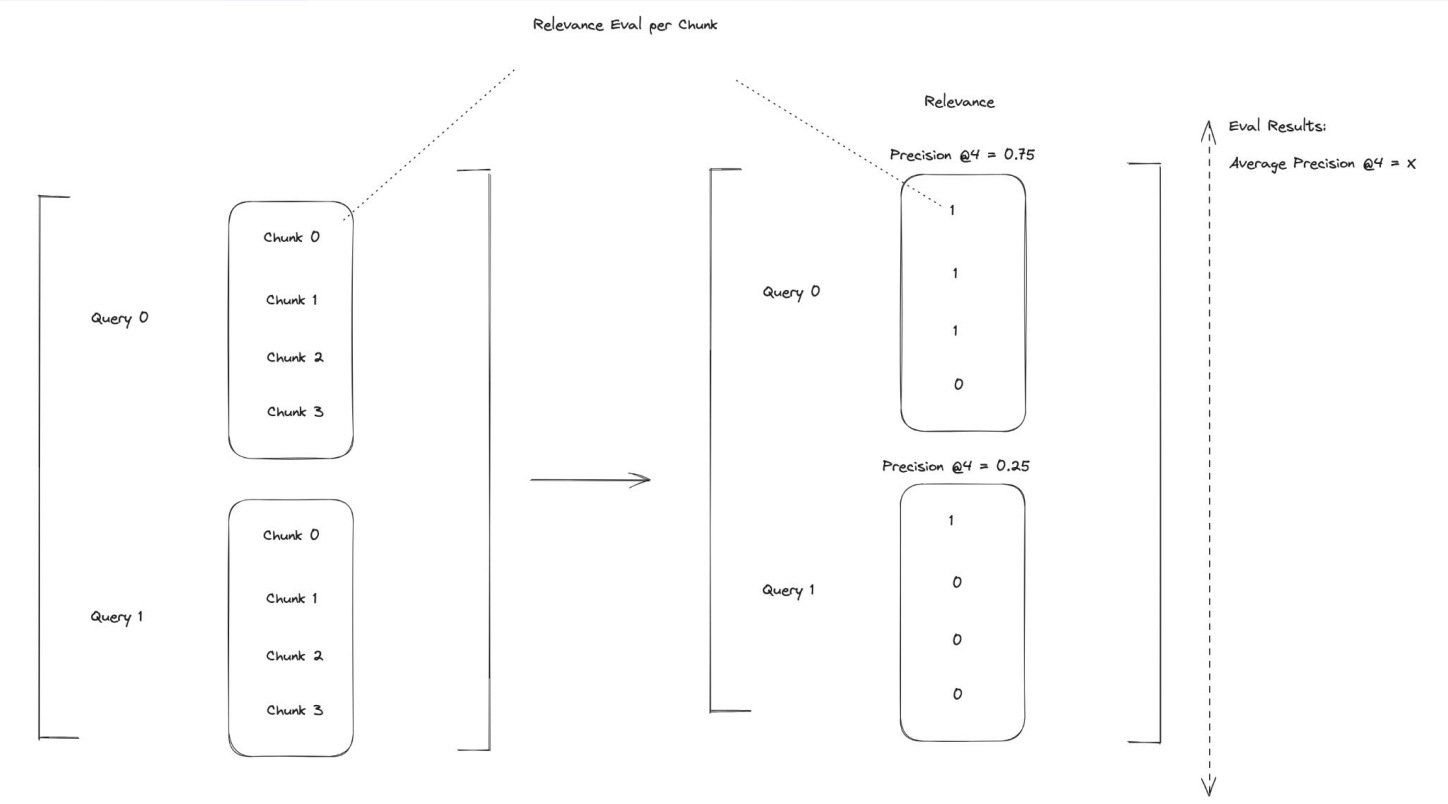
1. Parse / chunk up text corpus
2. Prompt GPT-4 to generate questions from each chunk
3. Run (question, context) through GPT-4 → Get a “ground-truth” response
4. Each (question, response) is now your evaluation pair!



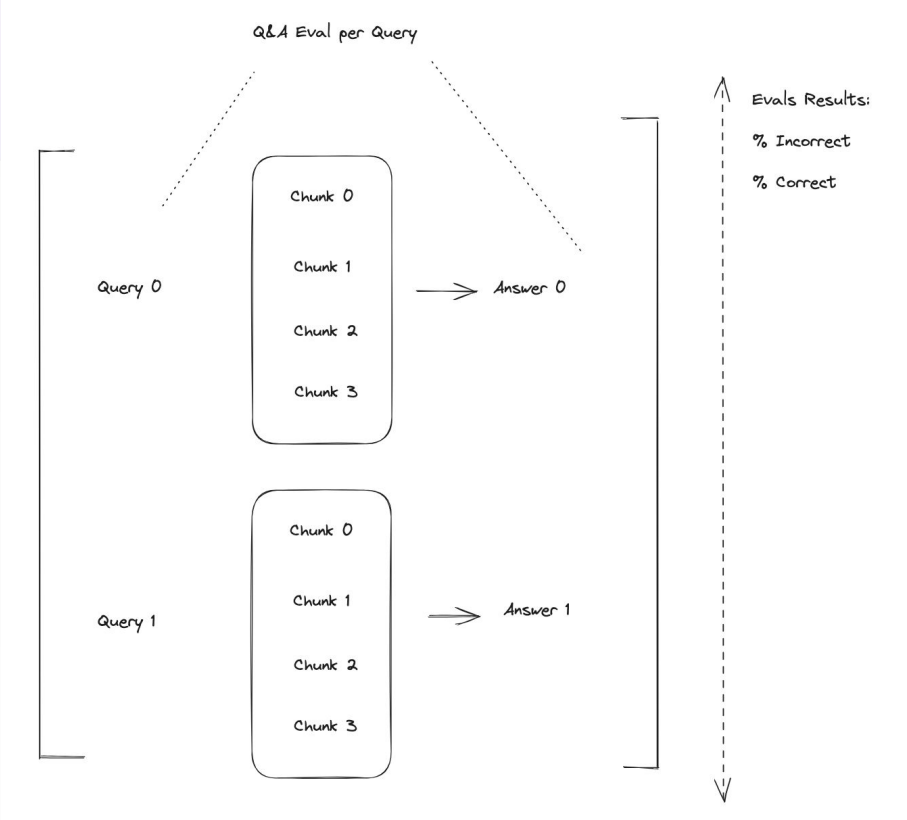
Demo Scripts

Benchmarking Retrieval Evals

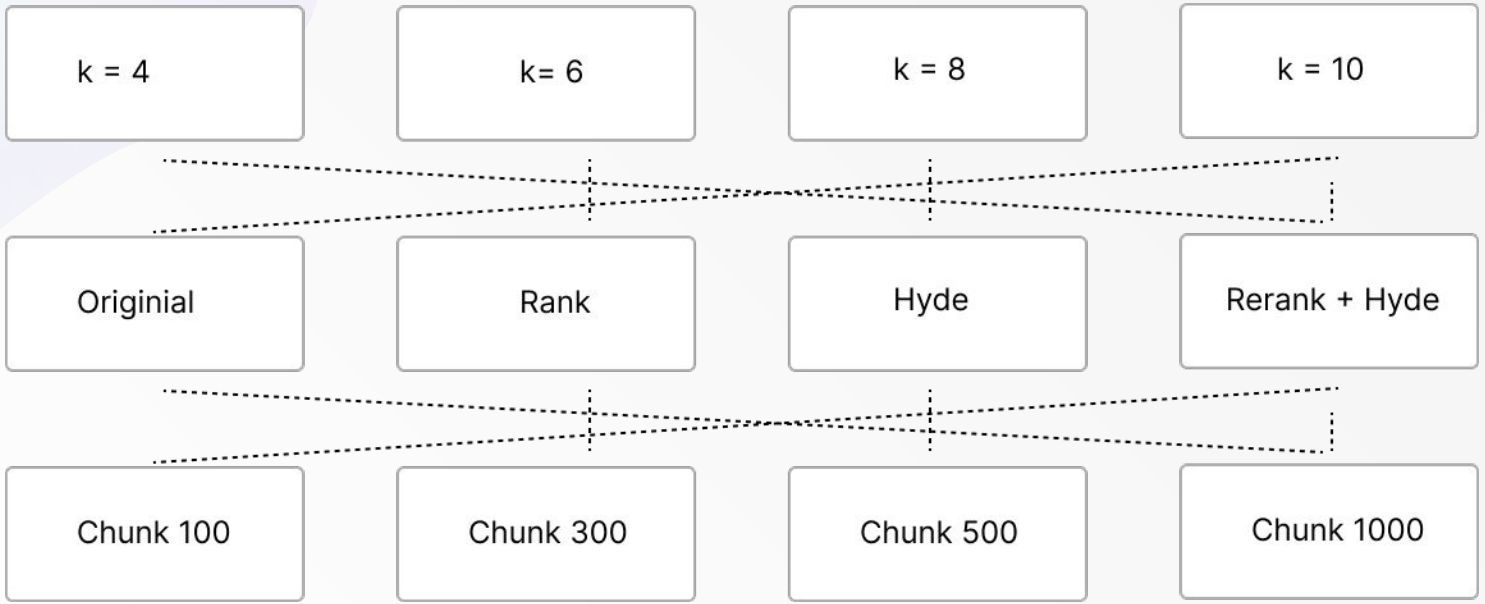
Building Blocks: Retrieval Eval



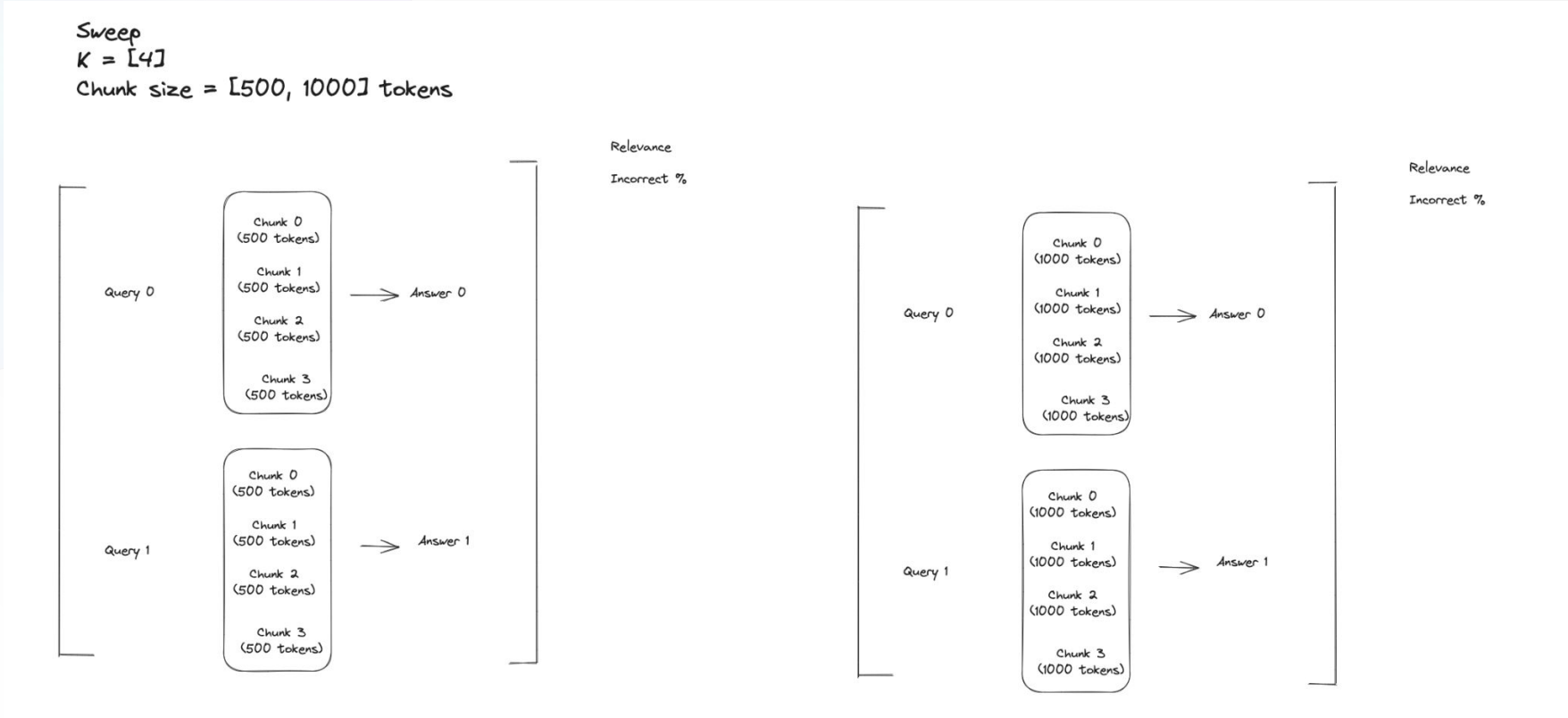
Building Blocks: Q&A Eval



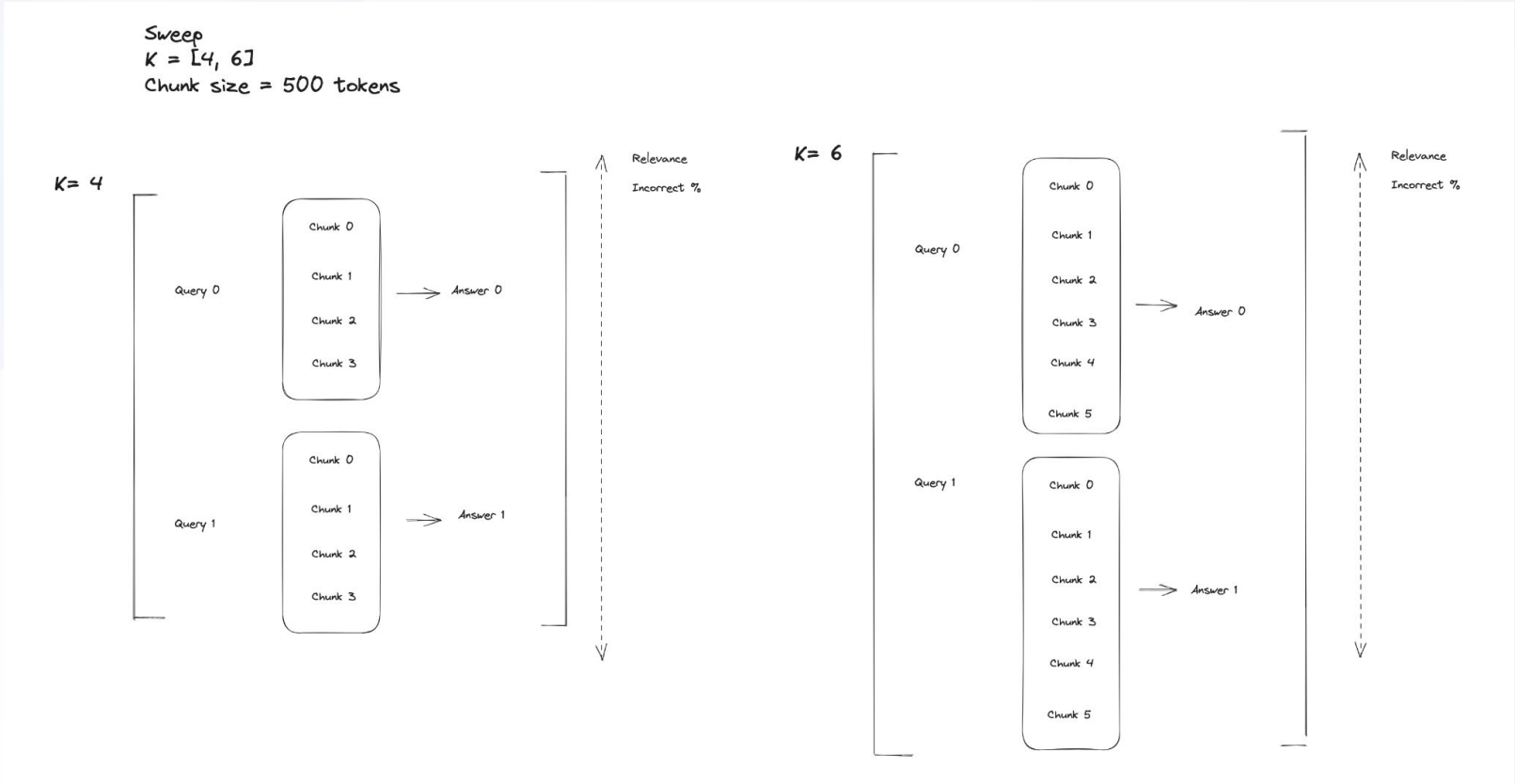
Scripts for Sweeping Through Retrieval Setup



Chunk Size Sweeps

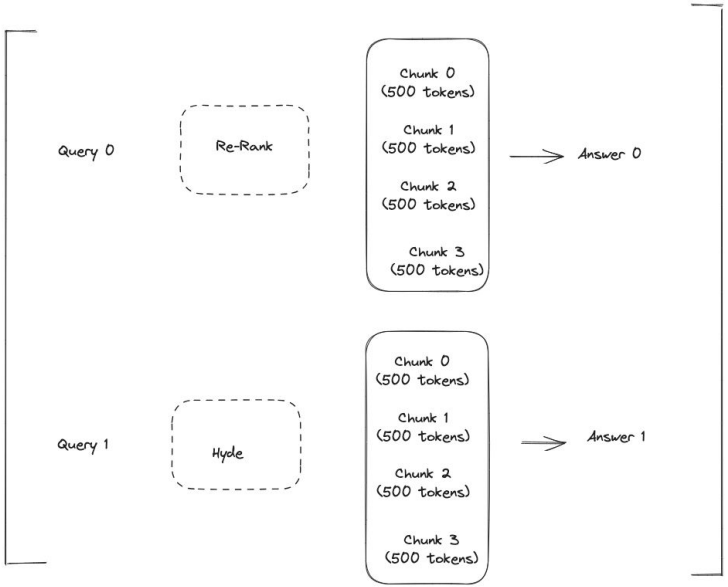


K = Retrieval Values Sweeps



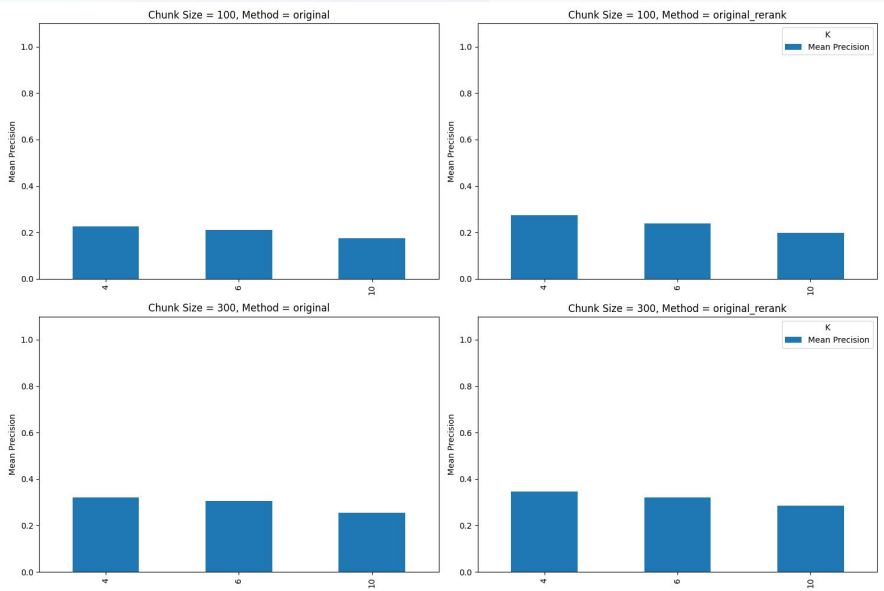
Retrieval Sweep - Types

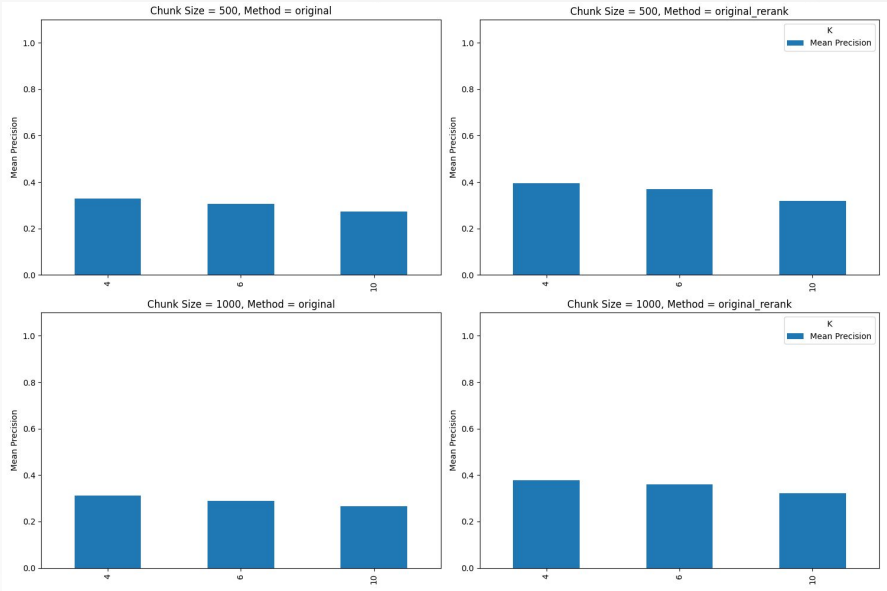
Sweep
K = [4]
Chunk size = [500] tokens
Retrieval Transformation = [original, re-rank]



Results: Chunk Retrieval Eval

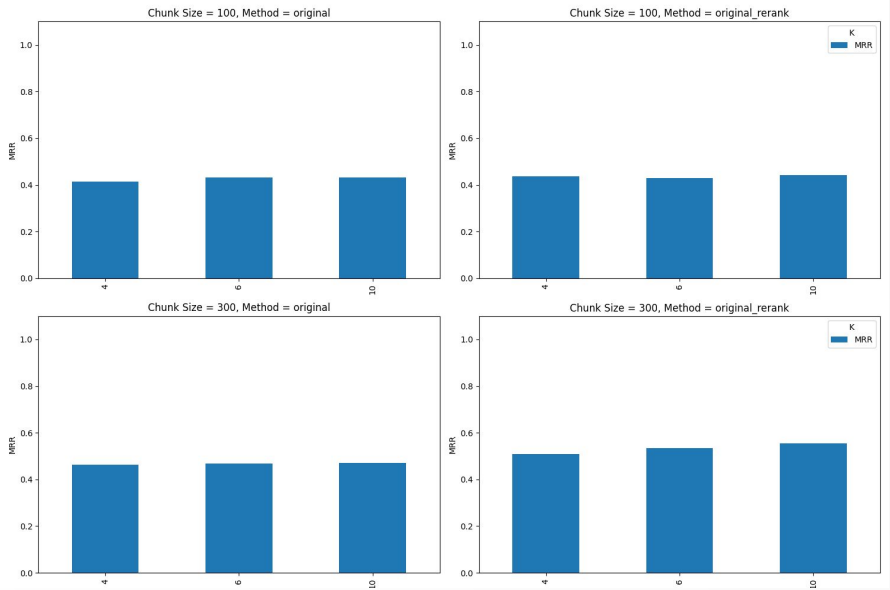
Precision @ k

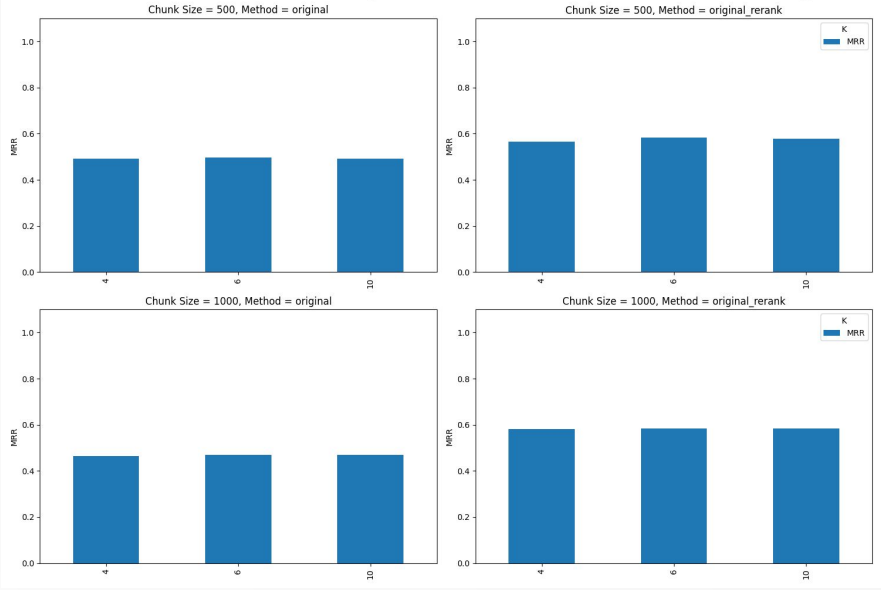




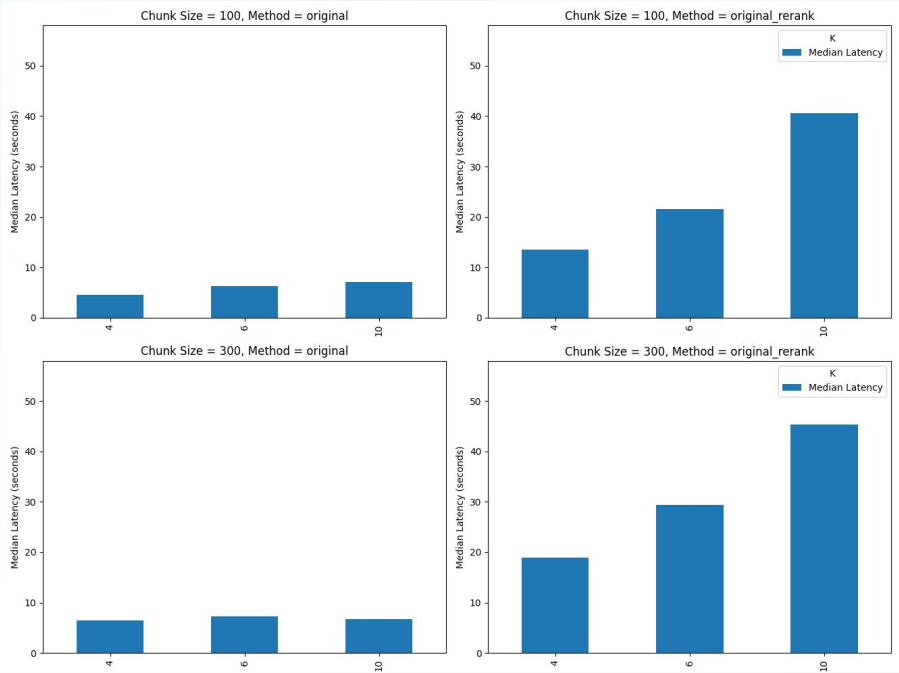
Results: Chunk Retrieval Eval

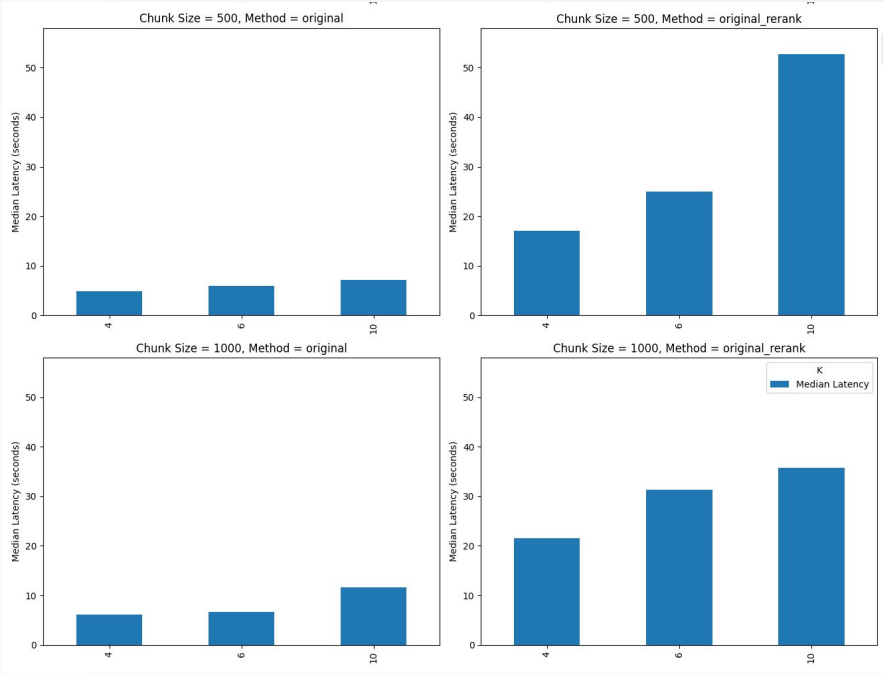
Precision MRR



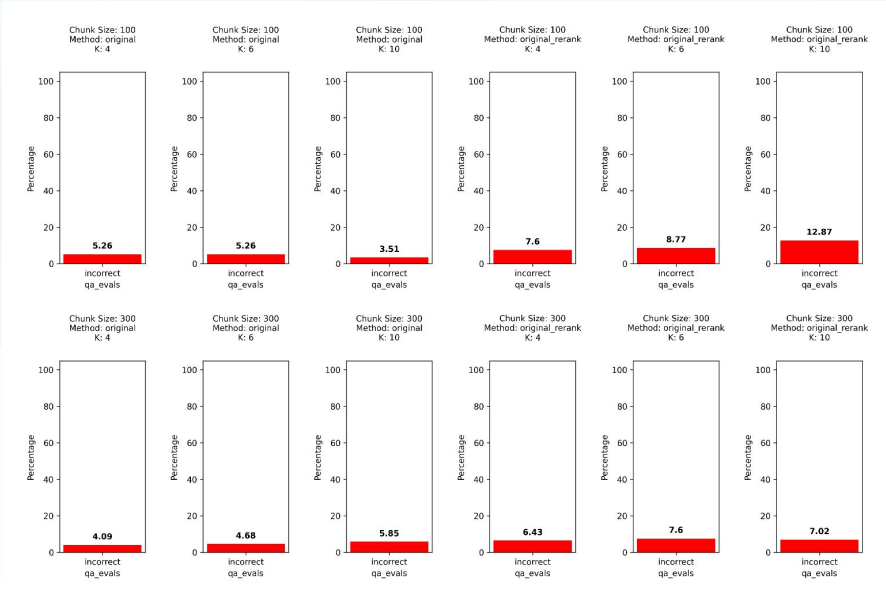


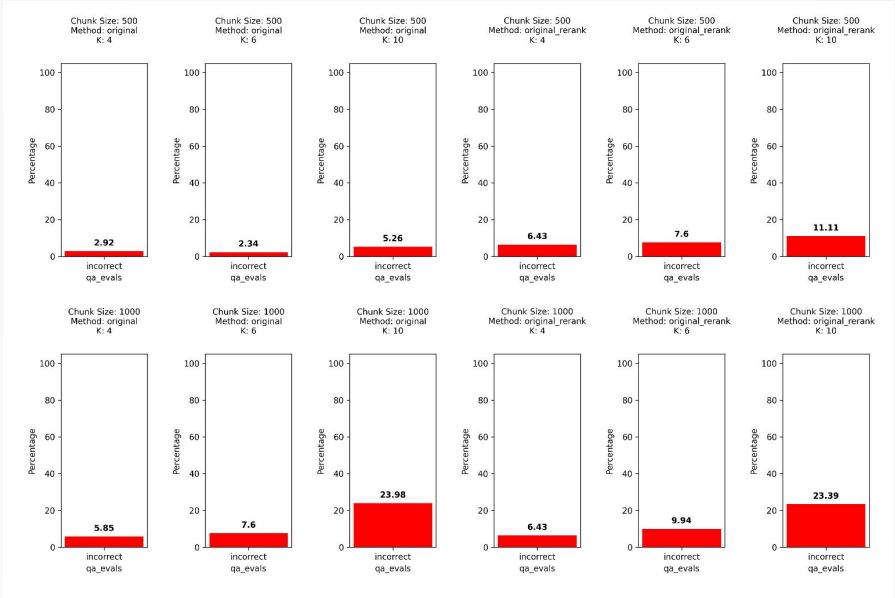
Results: Median Latency





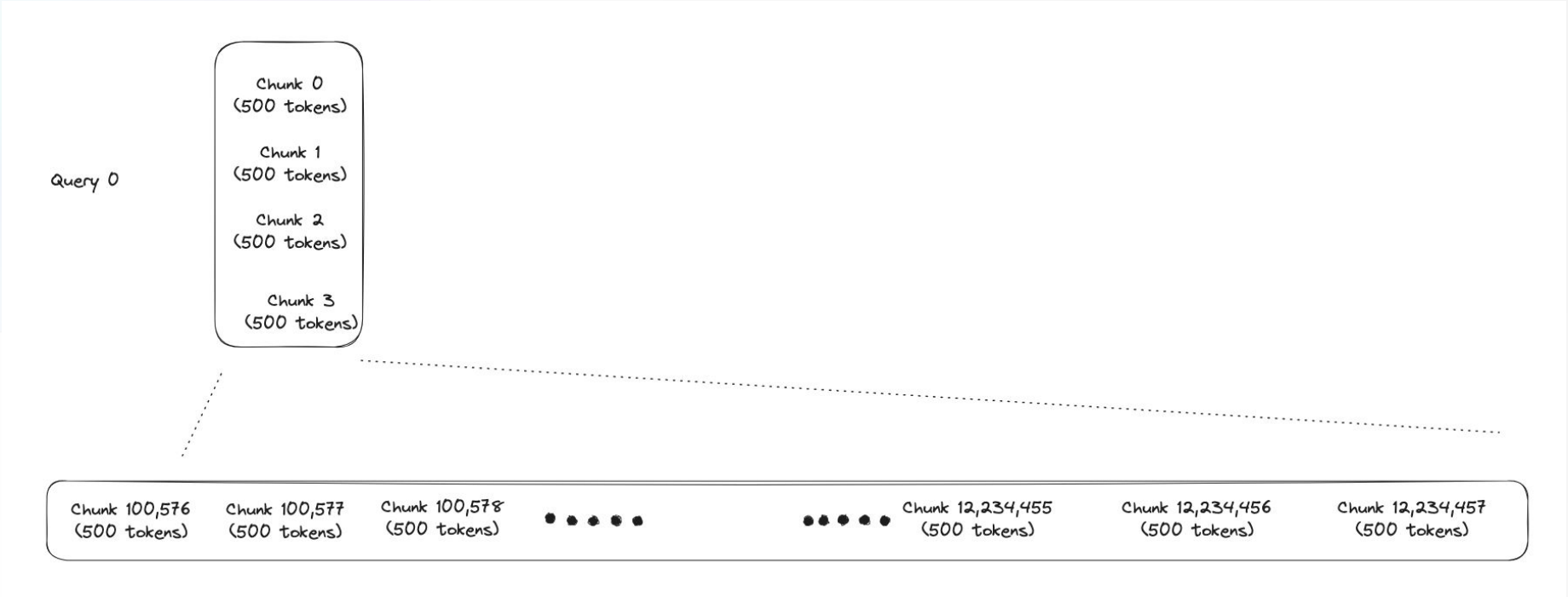
Question and Answer Eval



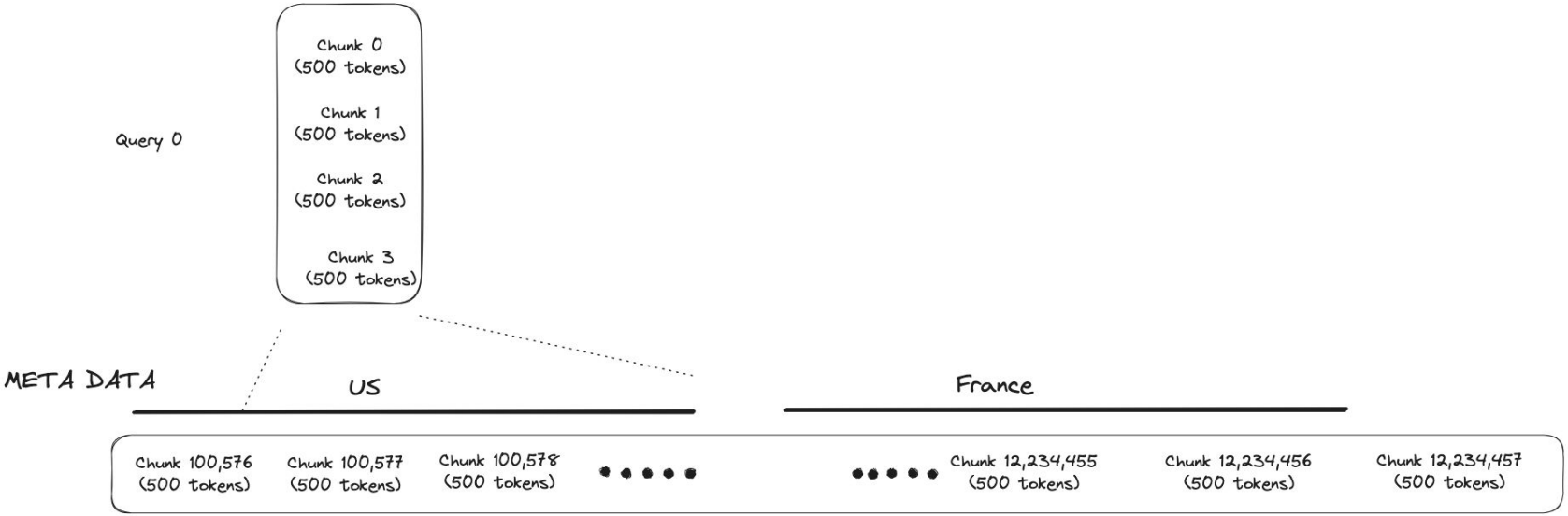


Challenges On the Ground

Lots of Chunks



Metadata Breakup



First to Market Sweep Scripts for Retrieval

Sweeps Chunks, K and Retrieval Techniques

Precision @ K, MRR and Q&A Correctness

Colab

GitHub RAG

Thank you

Visit us at **arize.com**