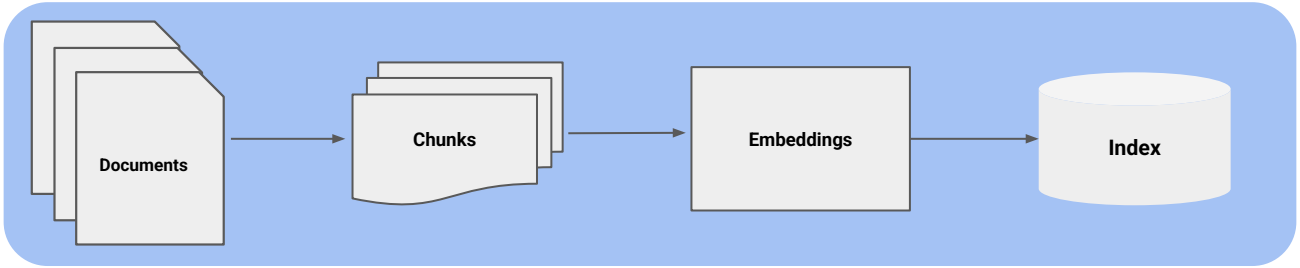
Building and Evaluating Advanced RAG

Basic RAG Pipeline



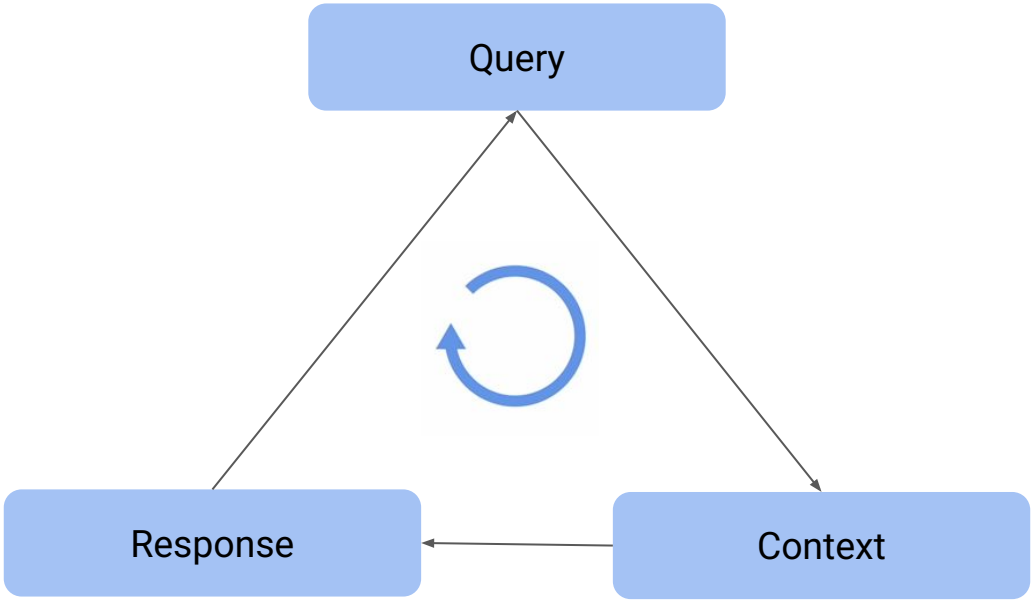
Setup

Basic and advanced RAG Pipeline with LlamaIndex

Evaluation Benchmark with TruEra

The RAG Triad

Answer Relevance: Is the response relevant to the query?



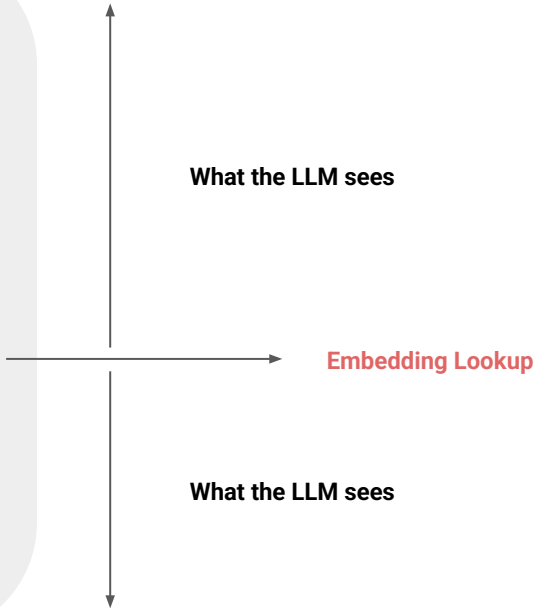
Context Relevance: Is the retrieved context relevant to the query?

Groundedness: Is the response supported by the context?

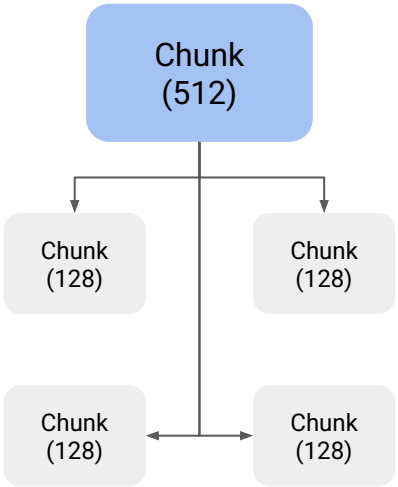
Sentence-window Retrieval

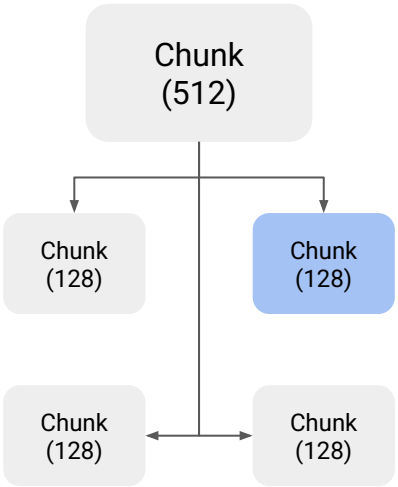
Query: What are the concerns surrounding the AMOC?

Continuous observation of the Atlantic meridional overturning circulation (AMOC) has improved the understanding of its variability (Frajka-Williams et al., 2019), but there is low confidence in the qualification of AMOC changes in the 20th century because of low agreement in quantitative reconstructed and simulated trends. Direct observational records since the mid-2000s remain too short to determine the relative contributions of internal variability, natural forcing and anthropogenic to AMOC change (high confidence). Over the 21st century, AMOC will very likely decline for all SSP scenarios but will not involve an abrupt collapse before 2100. 3.2.2.4 Sea Ice Changes Sea ice is a key driver of polar marine life, hosting unique ecosystems and affecting diverse marine organisms and food webs through its impact on light penetrations and supplies of nutrients and organic matter (Arrigo, 2014).



Auto-merging Retrieval

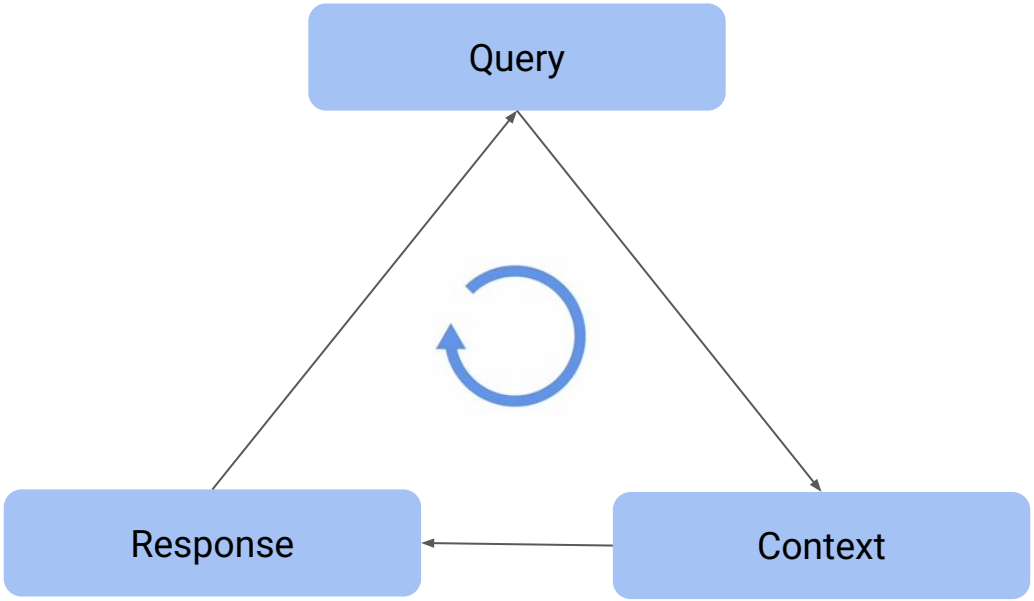




Returned chunks

The RAG Triad

Answer Relevance: Is the response relevant to the query?



Context Relevance: Is the retrieved context relevant to the query?

Groundedness: Is the response supported by the context?

Answer Relevance

Answer Relevance:
Is the final response useful?

Answer Relevance

How can altruism be beneficial in building a career?

Altruism can be beneficial in building a career by helping others even as one focuses on their own career growth. By aiming to lift others during every step of their own journey, individuals can achieve better outcomes for themselves. This can create a positive reputation and network, which can lead to new opportunities and collaborations. Additionally, helping others can provide a sense of fulfillment and purpose, which can contribute to overall career satisfaction and well-being.

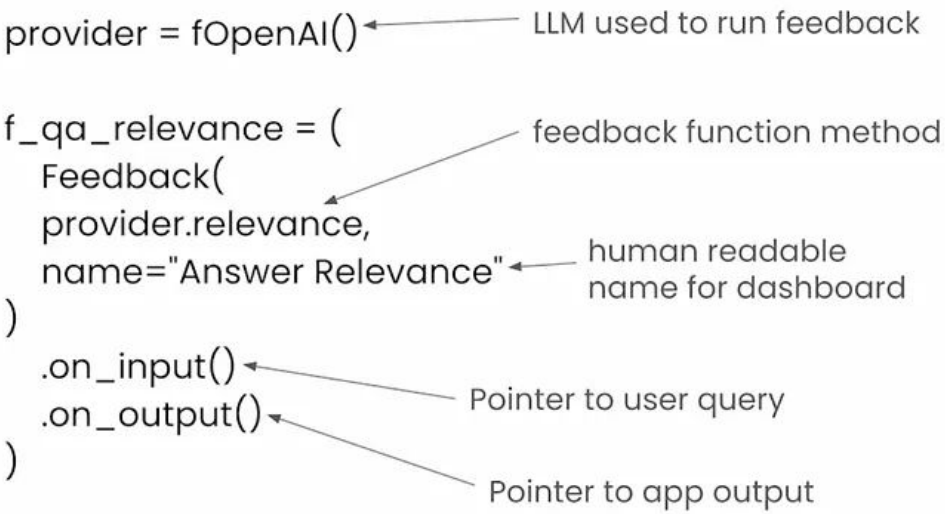
Supporting Evidence: The response provides a clear explanation of how altruism can be beneficial in building a career. It mentions that by helping others, individuals can achieve better outcomes for themselves, create a positive reputation and network, and lead to new opportunities and collaborations. It also highlights that helping others con provide a sense of fulfillment and purpose, contributing to overall career satisfaction and well-being.

Answer Relevance: 0.9

What is a Feedback Function

A feedback function provides a score (on a scale of 0 to 1) after reviewing an LLM application inputs, outputs, and intermediate results.

Structure of Feedback Functions - Answer Relevance



Answer Relevance: Is the final response relevant to the query?

Context Relevance

Context Relevance:
How good is the retrieval?

Context Relevance

How can altruism be beneficial in building a career?

Many successful people develop good habits in eating, exercise, sleep, personal relationships, work, learning, and self-care.
Such habits help them move forward while staying healthy. 4. Personal discipline I find that people who aim to lift others during every step of their own journey often achieve better outcomes for themselves. How can we help others even as we build an exciting career
for ourselves?5. Altruism

PAGE 37 Overcoming Imposter Syndrome CHAPTER 11

PAGE 38 Before we dive into the final chapter of this book, I'd like to address the serious matter of newcomers to AI sometimes experiencing imposter syndrome, where someone regardless of their success in the field wonders if they're a fraud and really belong in the AI community.
I want to make sure this doesn't discourage you or anyone else from growing in AI.

Context Relevance: 0.5

Using Informational Interviews to Find the Right Job CHAPTER 8

PAGE 31 Finding the Right AI Job for You CHAPTER 9 JOBS

PAGE 32 in this chapter, I'd like to discuss some fine points of finding a job.

The typical job search follows a fairly predictable path. Although the process may be familiar, every job search is different. Here are some tips to increase the odds you'll find a position that supports your thriving career and enables you to keep growing.Research roles and companies online or by talking to friends. Optionally, arrange informal informational interviews with people in companies that appeal to you. Either apply directly or, if you can, get a referral from someone on the inside.

Context Relevance: 0.7

Mean Context Relevance: 0.6

Structure of Feedback Functions - Context Relevance

```
provider = fOpenAI()

f_qs_relevance = (
    Feedback(
    provider.qs_relevance,
    name="Context Relevance"
)
    .on_input()
    .on(context_selection)
    .aggregate(np.mean)
)
```

Pointer to user query

Pointer to retrieved contexts (intermediate results)

Aggregate score across all retrieved context

Context Relevance: How good is the retrieval?

Context Relevance

How can altruism be beneficial in building a career?

Using informational interviews to Find the light Job CHAPTER 8

PAGE Finding the Right
AI job for You CHAPTER 9
JOBS
PAGE 22th shopted st discuss some fine points
of finding a job
The typical job search follows forty predictable path Although the process may be familiar every job search is different. Here are some tips to increase the odds you'll find a position that supports your thriving career and enables you to keep growing. Research roles and companies online or by talking to friends. Optionally arrange informal informational interviews with people in companies that appeal to you Either supply directly or, if you can, get a referral from someone on the inside

Context Relevance: 0.7

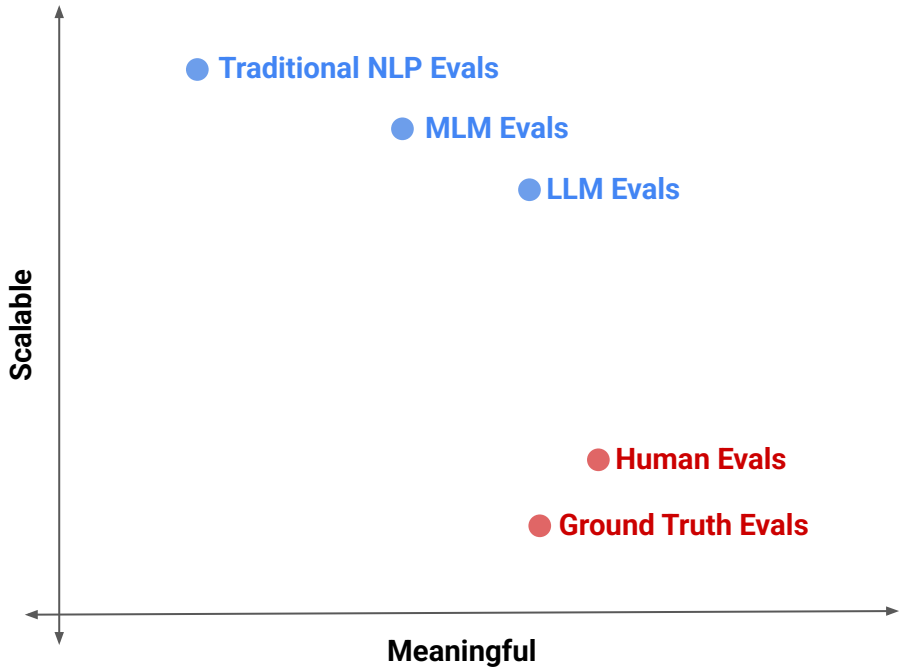
Supporting Evidence: The statement provides information on how to find the right job and increase the odds of finding a position that supports a thriving career. It suggests researching roles and companies online or by talking to friends, and optionally arranging informational interviews with people in companies that appeal to you. This information can be helpful in building a career by providing insights into potential job opportunities and allowing individuals to make informed decisions about their career path.

Evaluate and Iterate

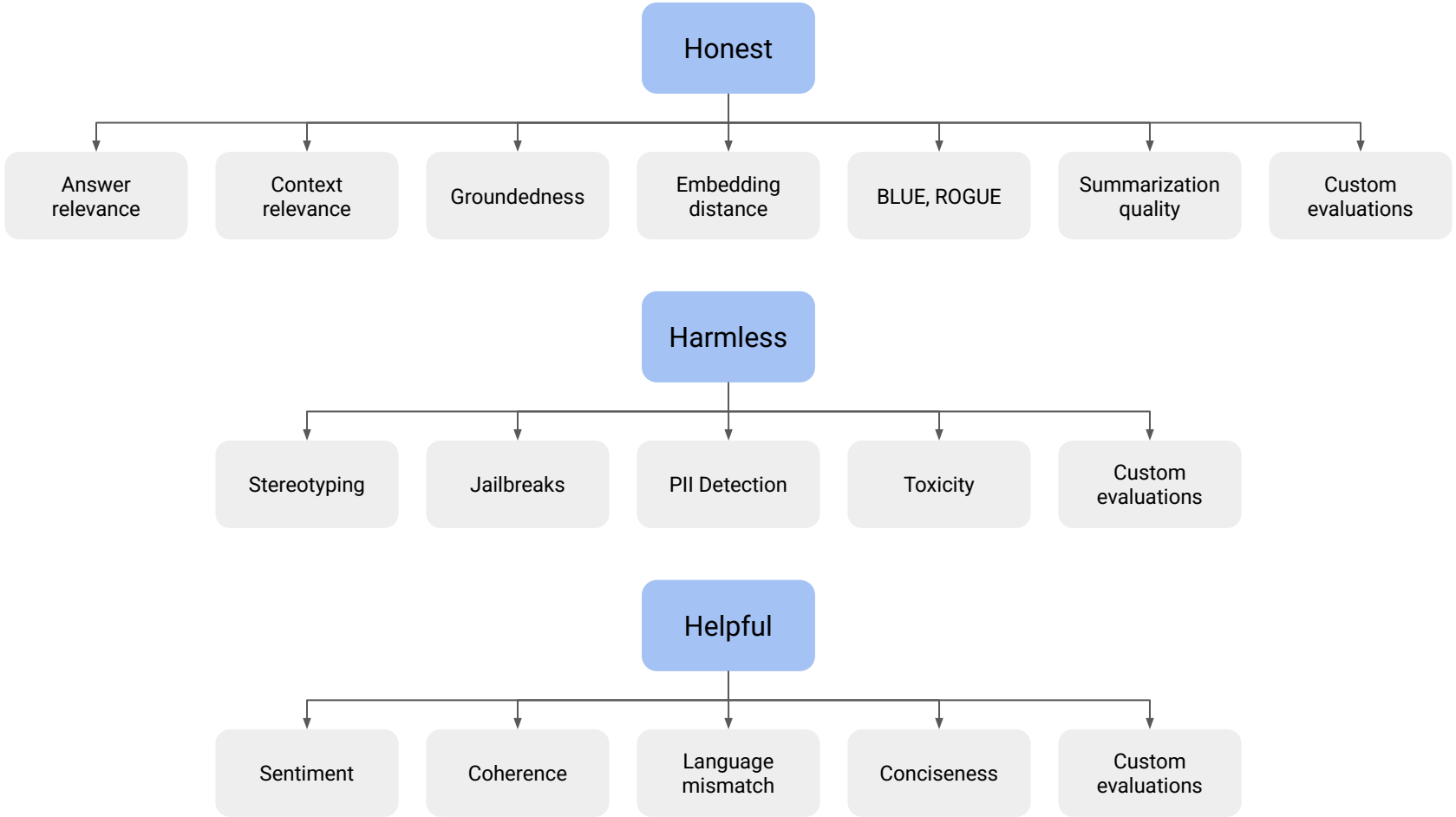
- Start with LlamaIndex Basic RAG
- Evaluate with TruLens RAG Triad
  - Failure modes related to context size
- Iterate with LlamaIndex Sentence Window RAG
- Re-evaluate with TruLens RAG Triad
  - Do we see improvements in Context Relevance?
  - What about other metrics
- Experiments with difference window sizes
  - What window size results in best eval metrics

Feedbacks Functions

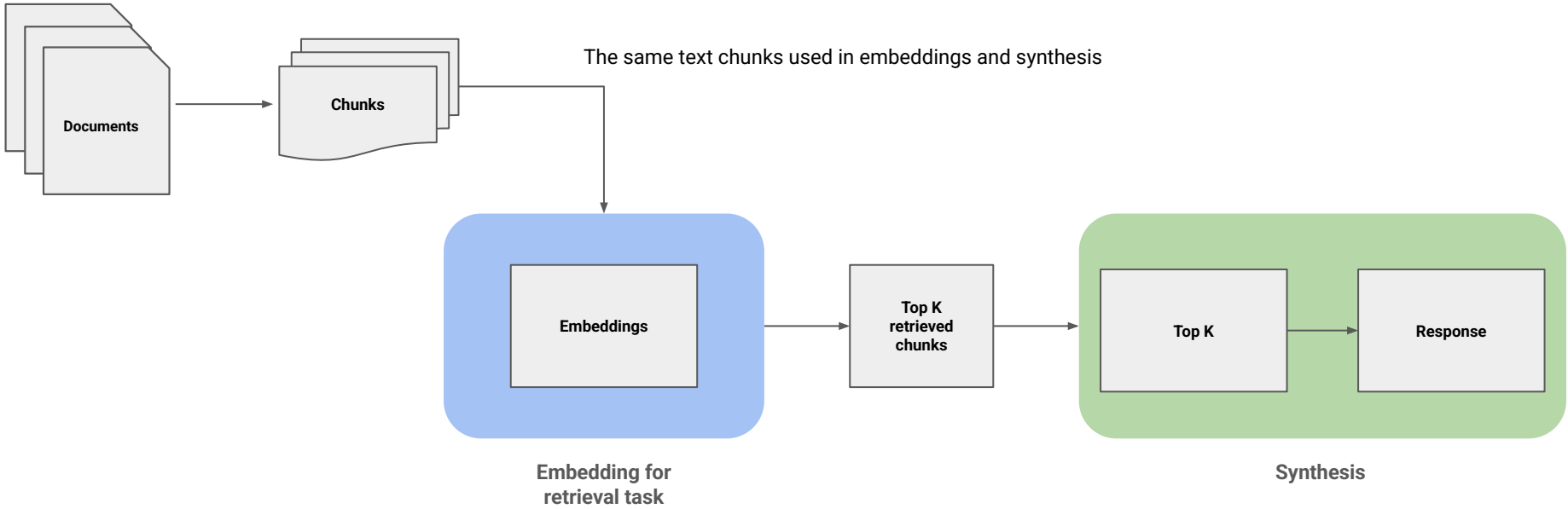
Feedback Functions can be implemented in different ways



TruLens Evaluation



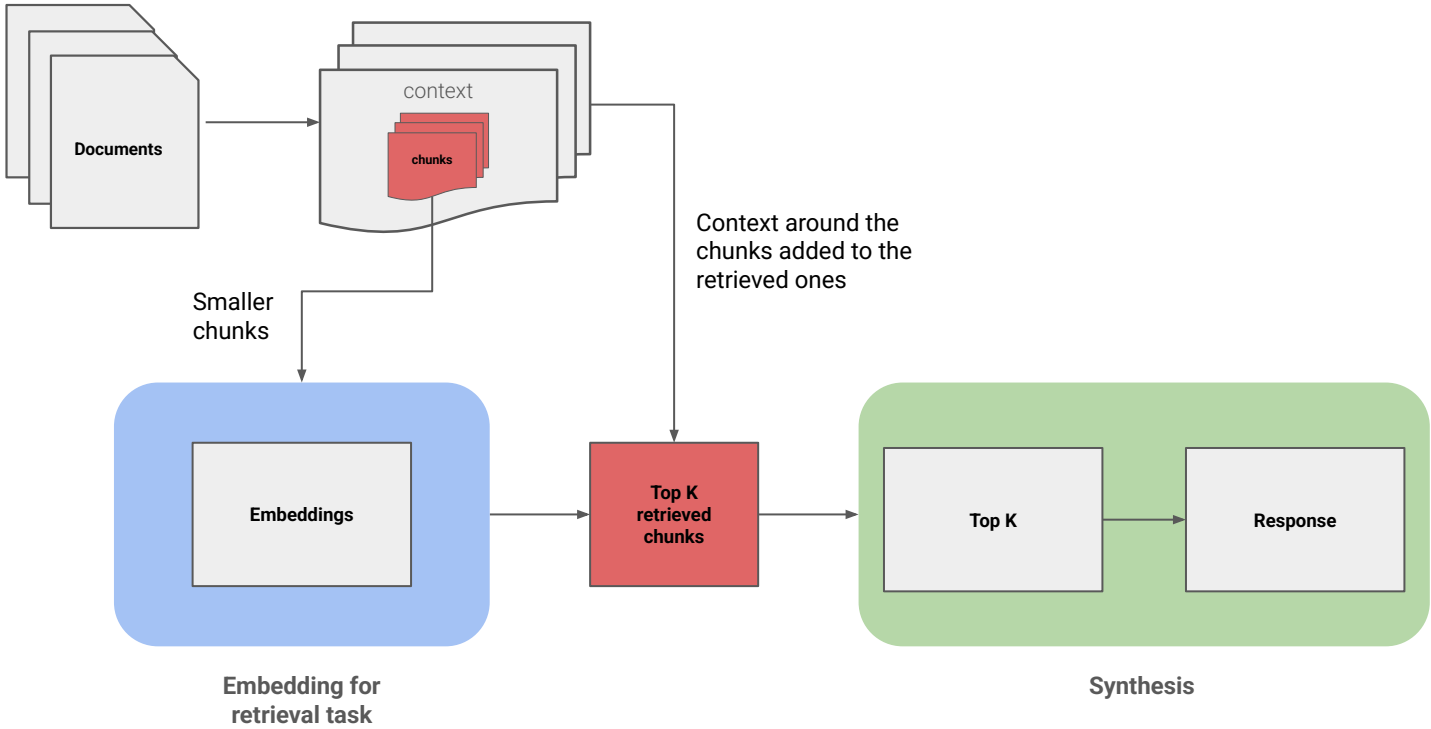
Basic RAG Pipeline - Issue



But

- Embedding-based retrieval works well with smaller text chunks

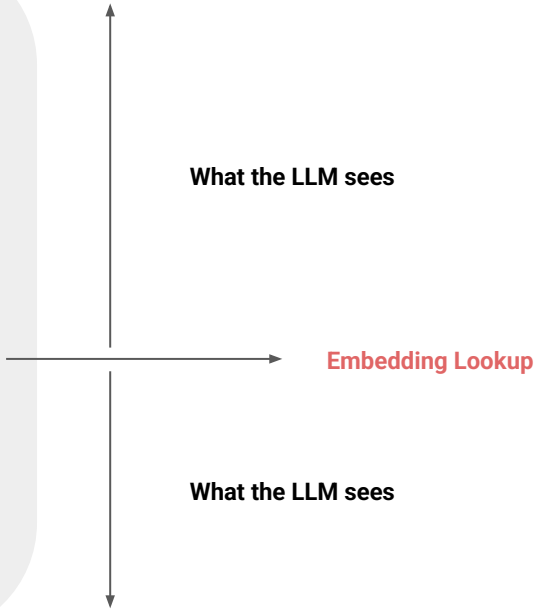
Sentence Window Retrieval Pipeline



Sentence-window Retrieval

Query: What are the concerns surrounding the AMOC?

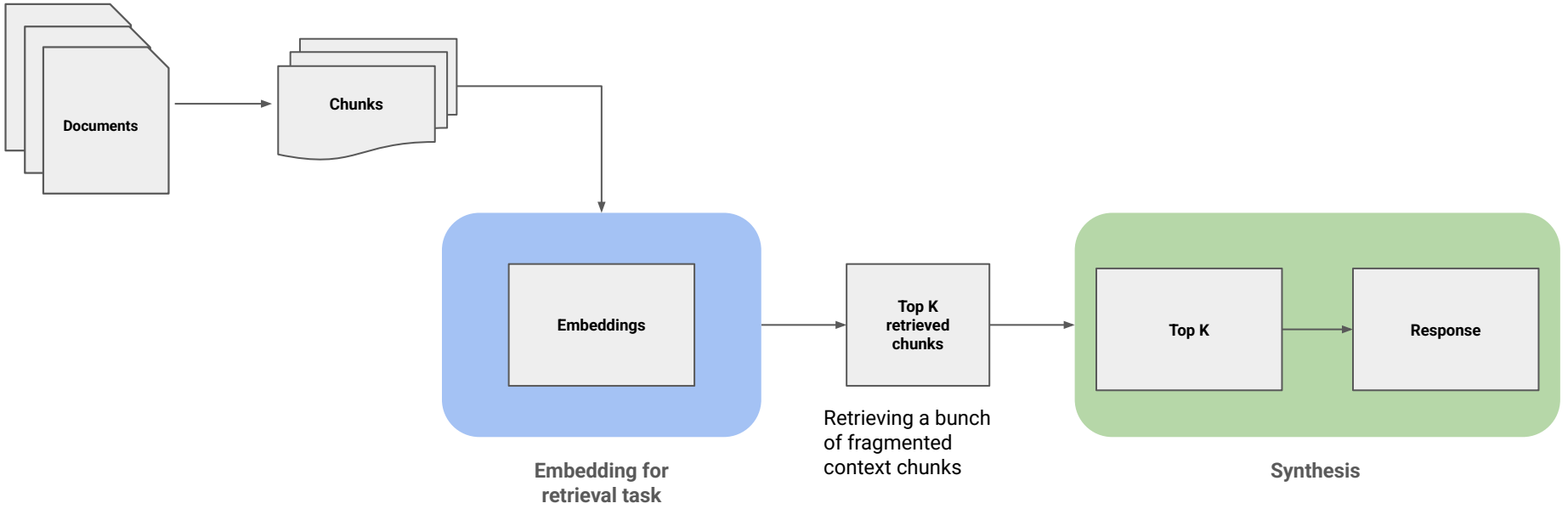
Continuous observation of the Atlantic meridional overturning circulation (AMOC) has improved the understanding of its variability (Frajka-Williams et al., 2019), but there is low confidence in the qualification of AMOC changes in the 20th century because of low agreement in quantitative reconstructed and simulated trends. Direct observational records since the mid-2000s remain too short to determine the relative contributions of internal variability, natural forcing and anthropogenic to AMOC change (high confidence). Over the 21st century, AMOC will very likely decline for all SSP scenarios but will not involve an abrupt collapse before 2100. 3.2.2.4 Sea Ice Changes Sea ice is a key driver of polar marine life, hosting unique ecosystems and affecting diverse marine organisms and food webs through its impact on light penetrations and supplies of nutrients and organic matter (Arrigo, 2014).



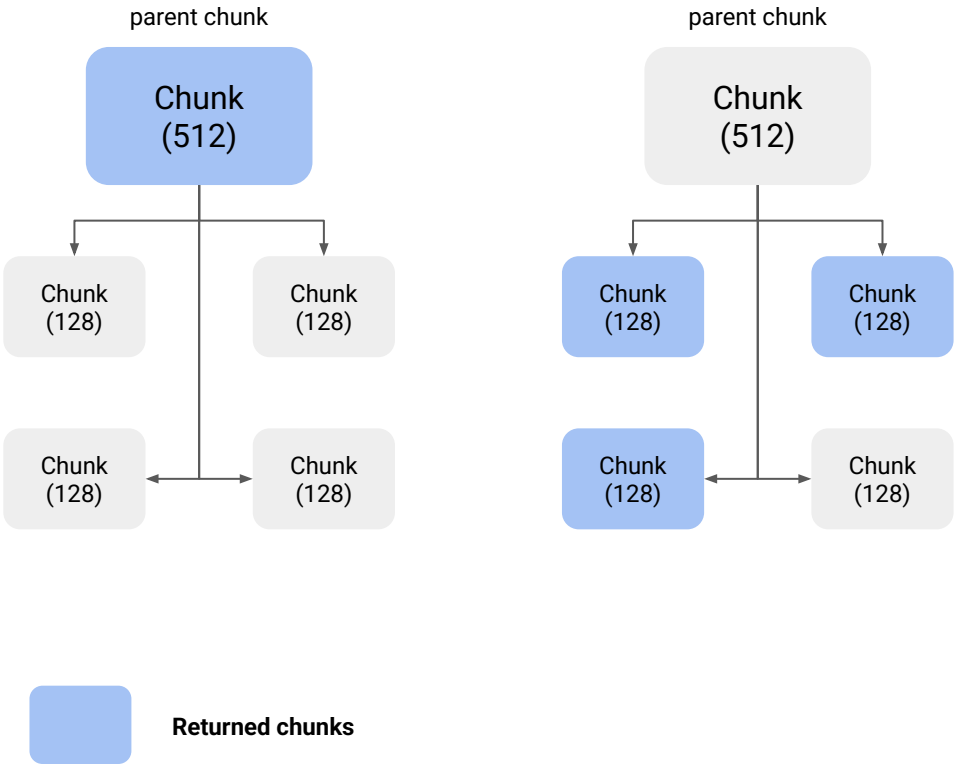
Evaluate and Iterate

- Gradually increase the sentence window size starting with 1 (one)
- Evaluate app versions with the RAG Triad
- Track experiments to pick the best sentence window
- Note tradeoff between token usage/cost and context relevance
- Note tradeoff between token window size and groundedness

Basic RAG Pipeline



Auto-merging Retrieval



- Define a hierarchy of smaller chunks linked to a parent chunk
- If the set of smaller chunks linking to a parent chunk exceeds some threshold, then “merge” smaller chunks into the bigger parent chunk

Evaluate and Iterate

- Iterate with different hierarchical structures (number of levels, children) and chunk sizes
- Evaluate app versions with the RAG Triad
- Track experiments to pick the best structure
- Gain intuition about hyperparameters that work best with certain doc types (e.g., employment contracts vs invoices)
- Auto-merging is complementary to sentence-window retrieval

THANK YOU