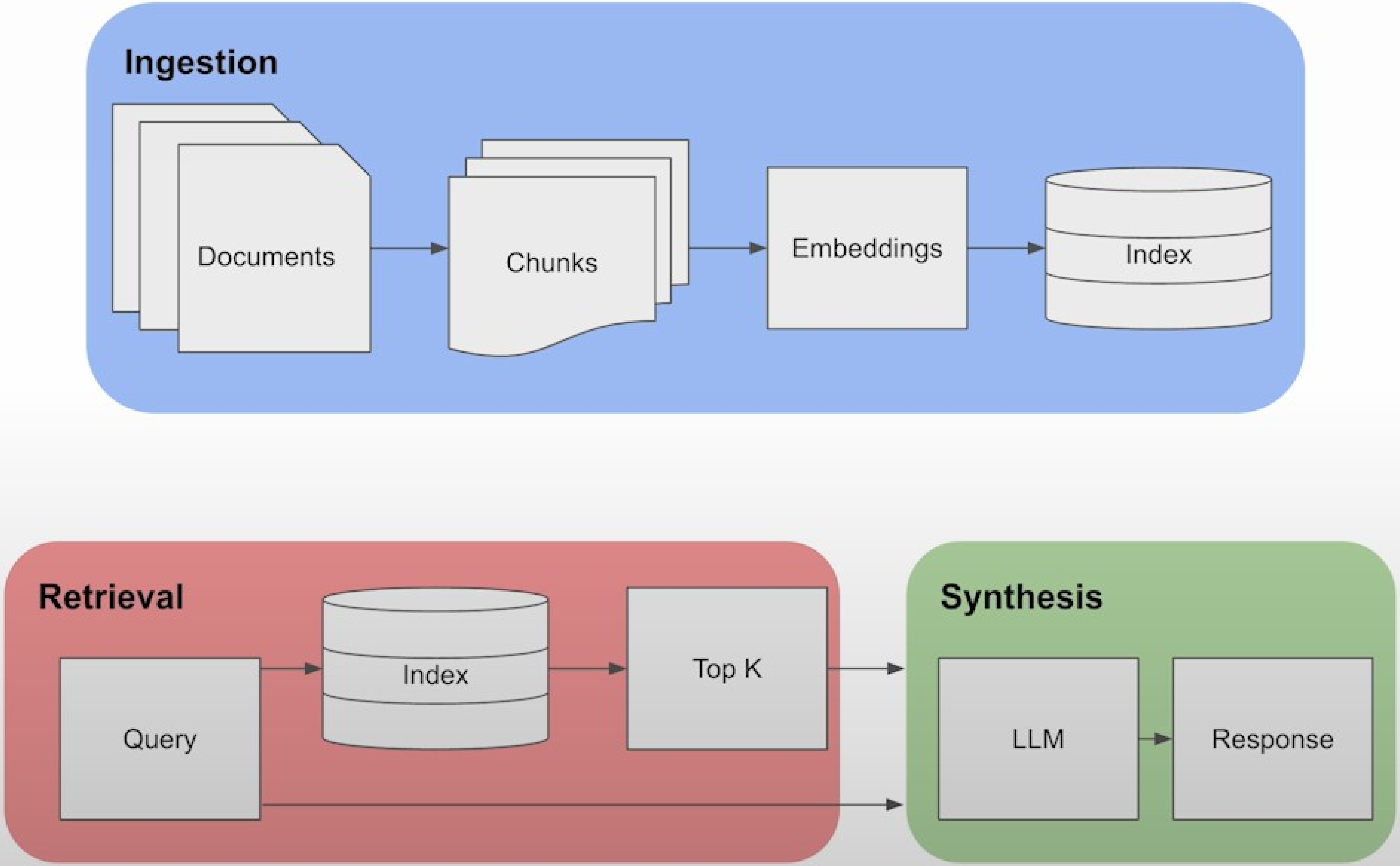
Building and Evaluating Advanced RAG Applications

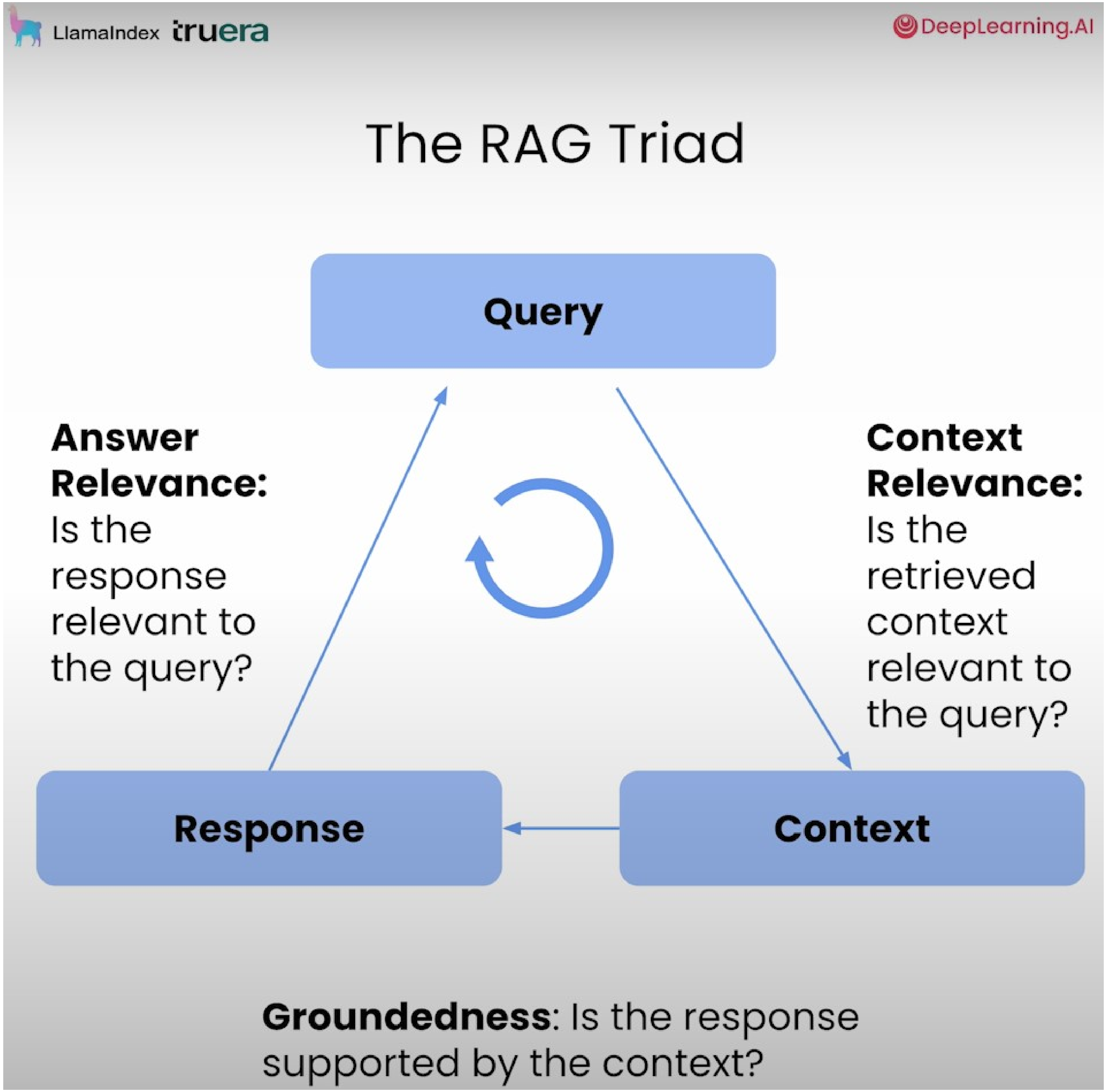
RAG Triad for Hallucination Eval

LlamaIndex truera

DeepLearning.AI

Basic RAG Pipeline





LlamaIndex truera

DeepLearning.AI

The RAG Triad

Query

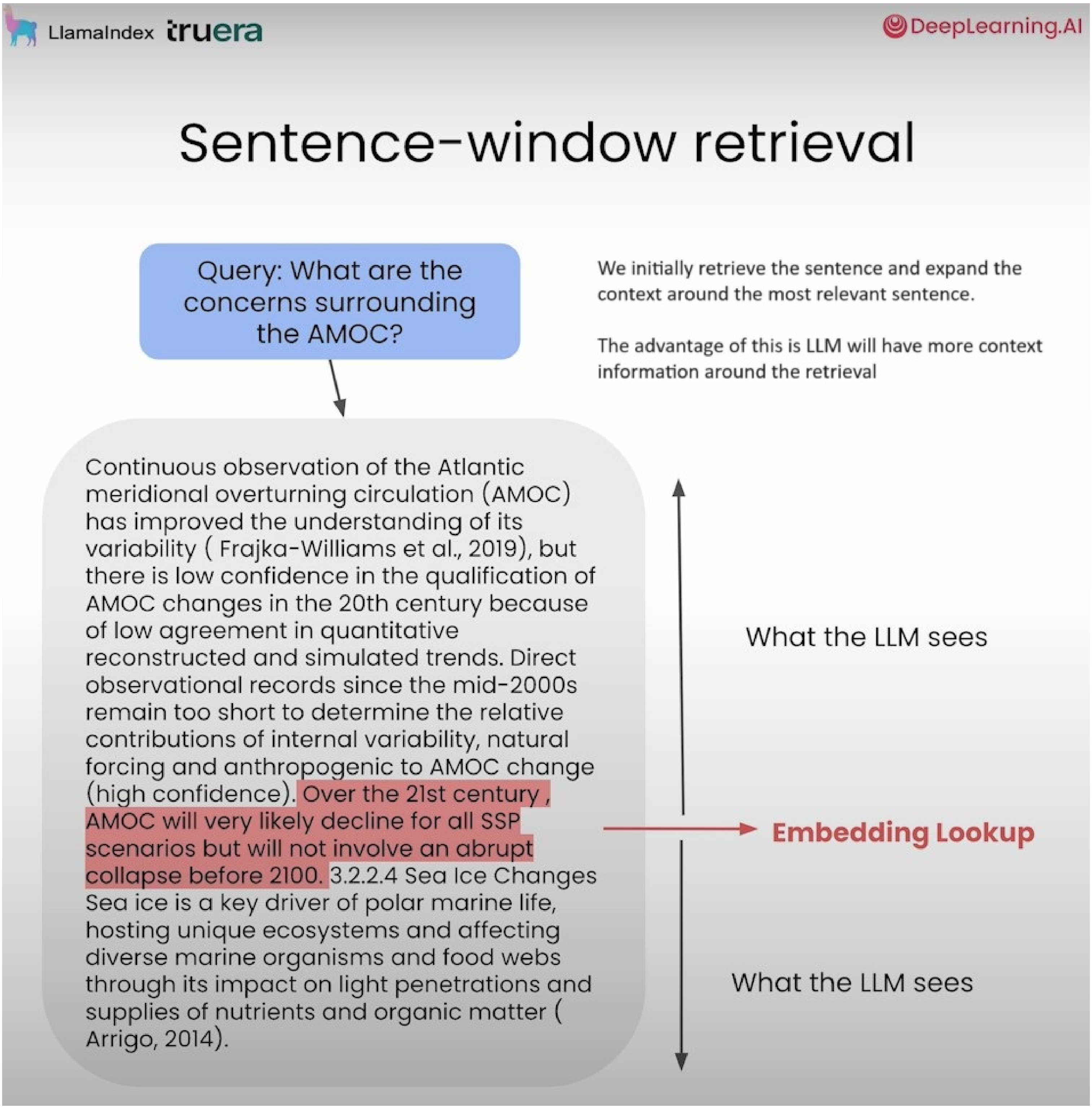
Answer Relevance: Is the response relevant to the query?

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

Response

Context

Groundedness: Is the response supported by the context?



LlamaIndex truera

DeepLearning.AI

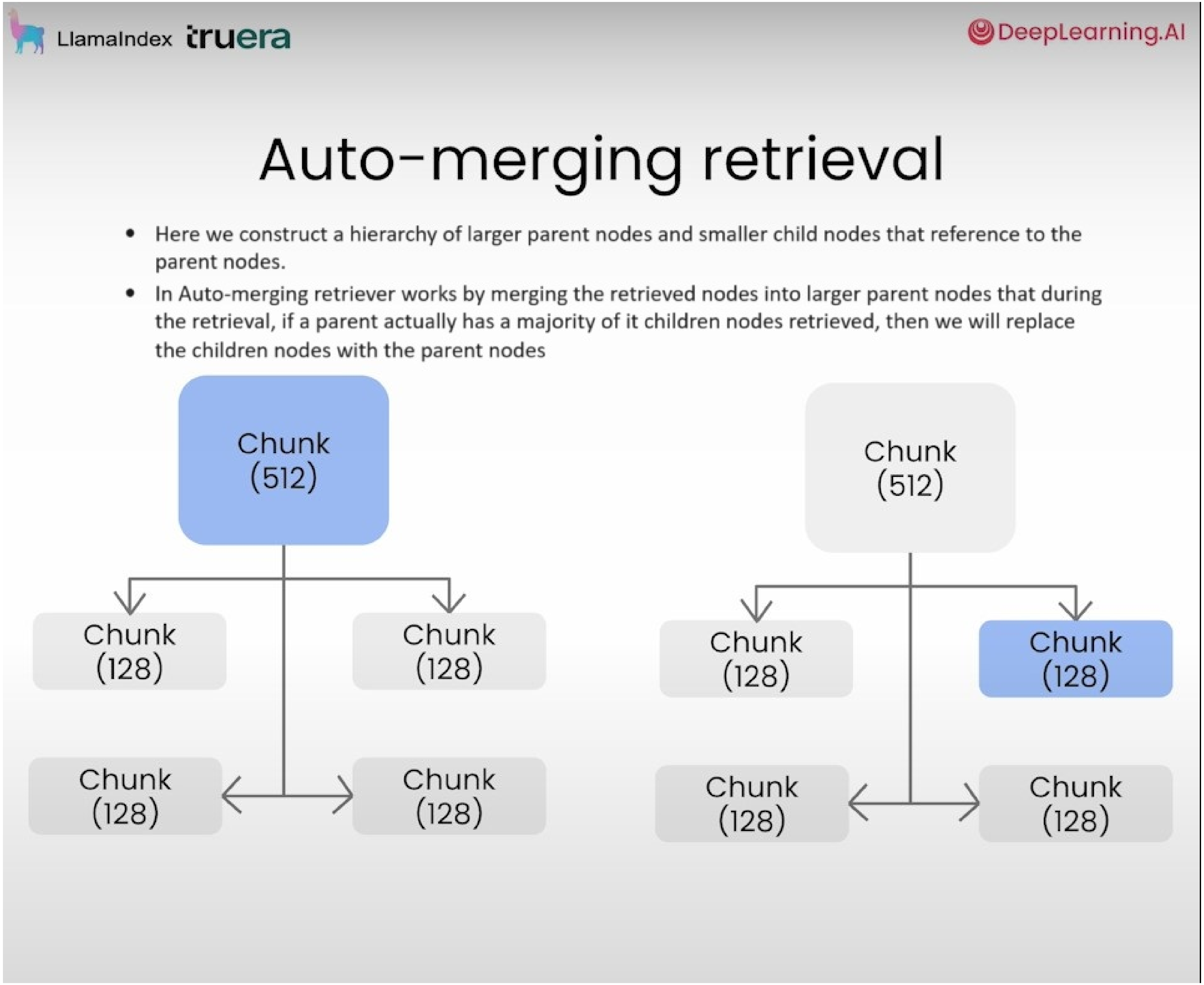
Sentence-window retrieval

Query: What are the concerns surrounding the AMOC?

We initially retrieve the sentence and expand the context around the most relevant sentence.

The advantage of this is LLM will have more context information around the retrieval

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).

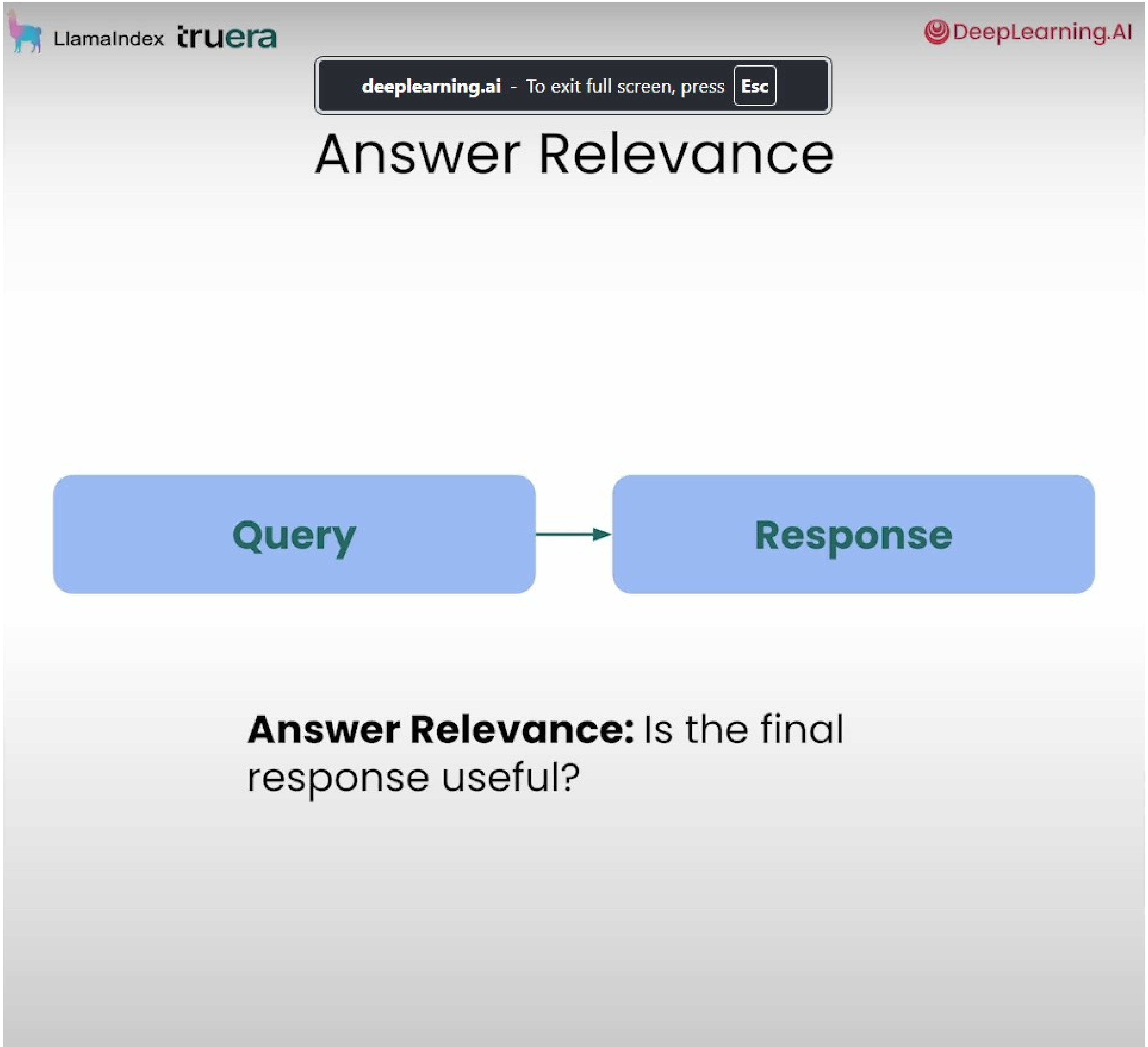


LlamaIndex truera

DeepLearning.AI

Auto-merging retrieval

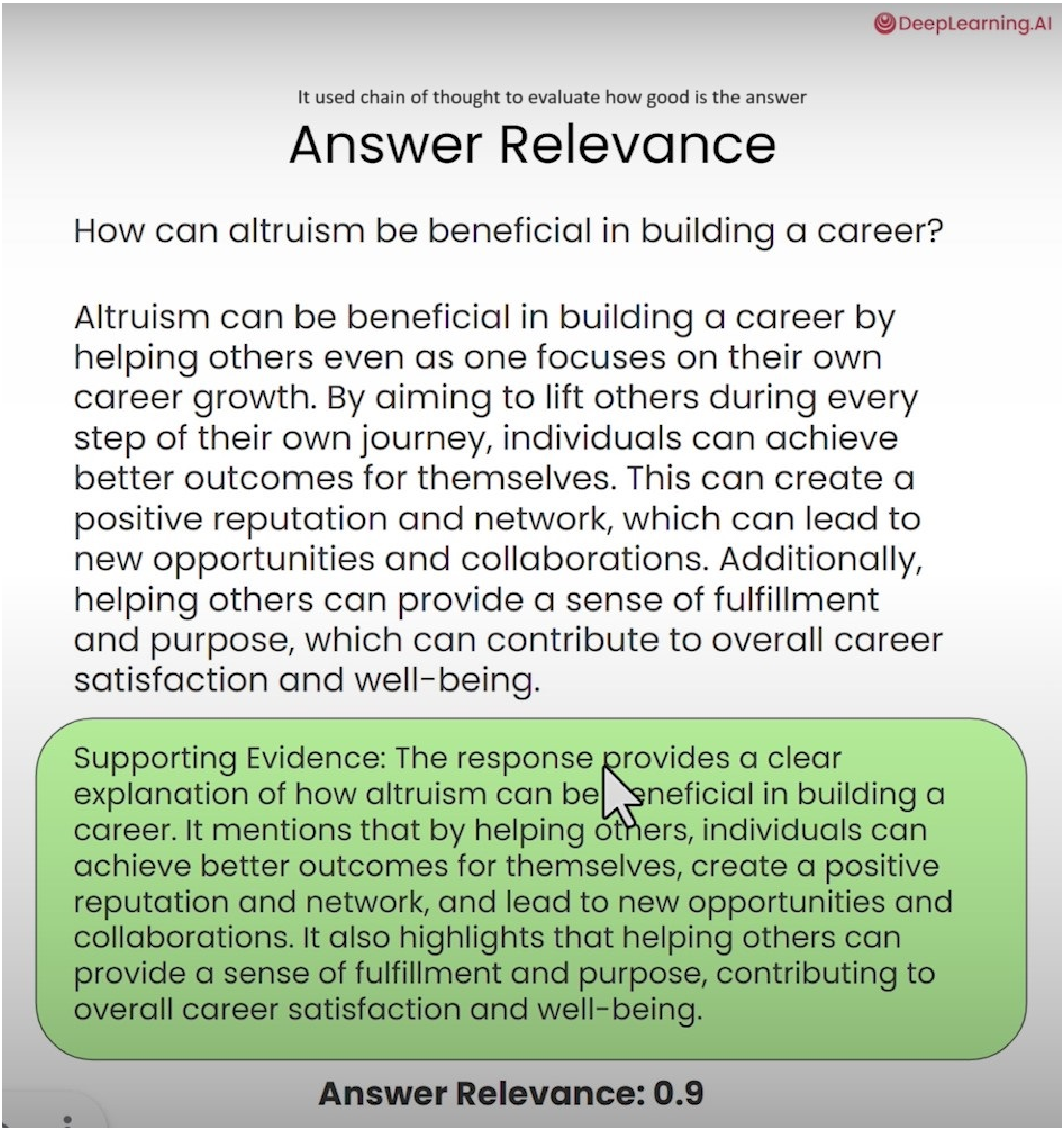
- Here we construct a hierarchy of larger parent nodes and smaller child nodes that reference to the parent nodes.
- In Auto-merging retriever works by merging the retrieved nodes into larger parent nodes that during the retrieval, if a parent actually has a majority of it children nodes retrieved, then we will replace the children nodes with the parent nodes



Answer Relevance

Query → Response

Answer Relevance: Is the final response useful?



It used chain of thought to evaluate how good is the answer

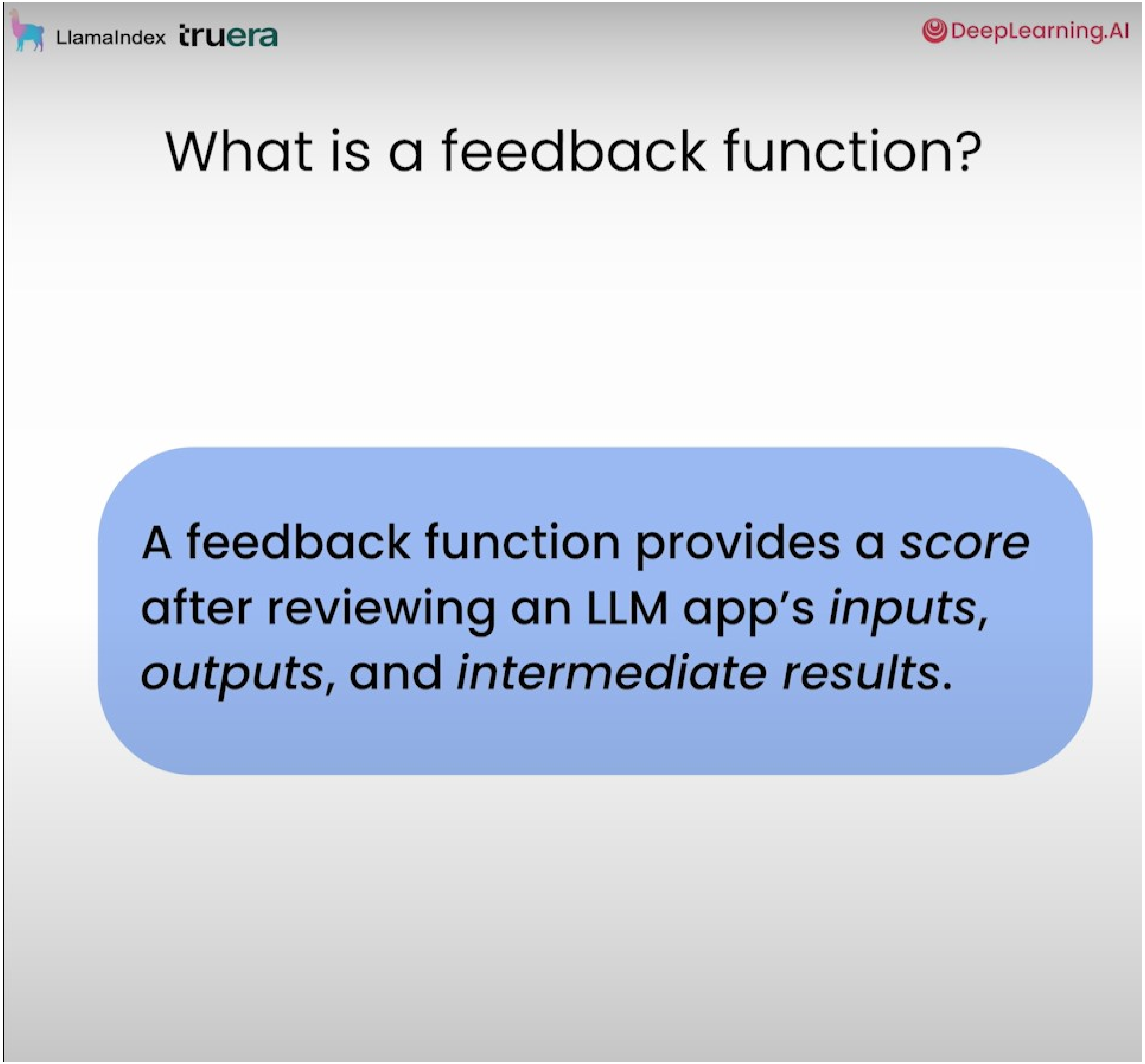
Answer Relevance

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.

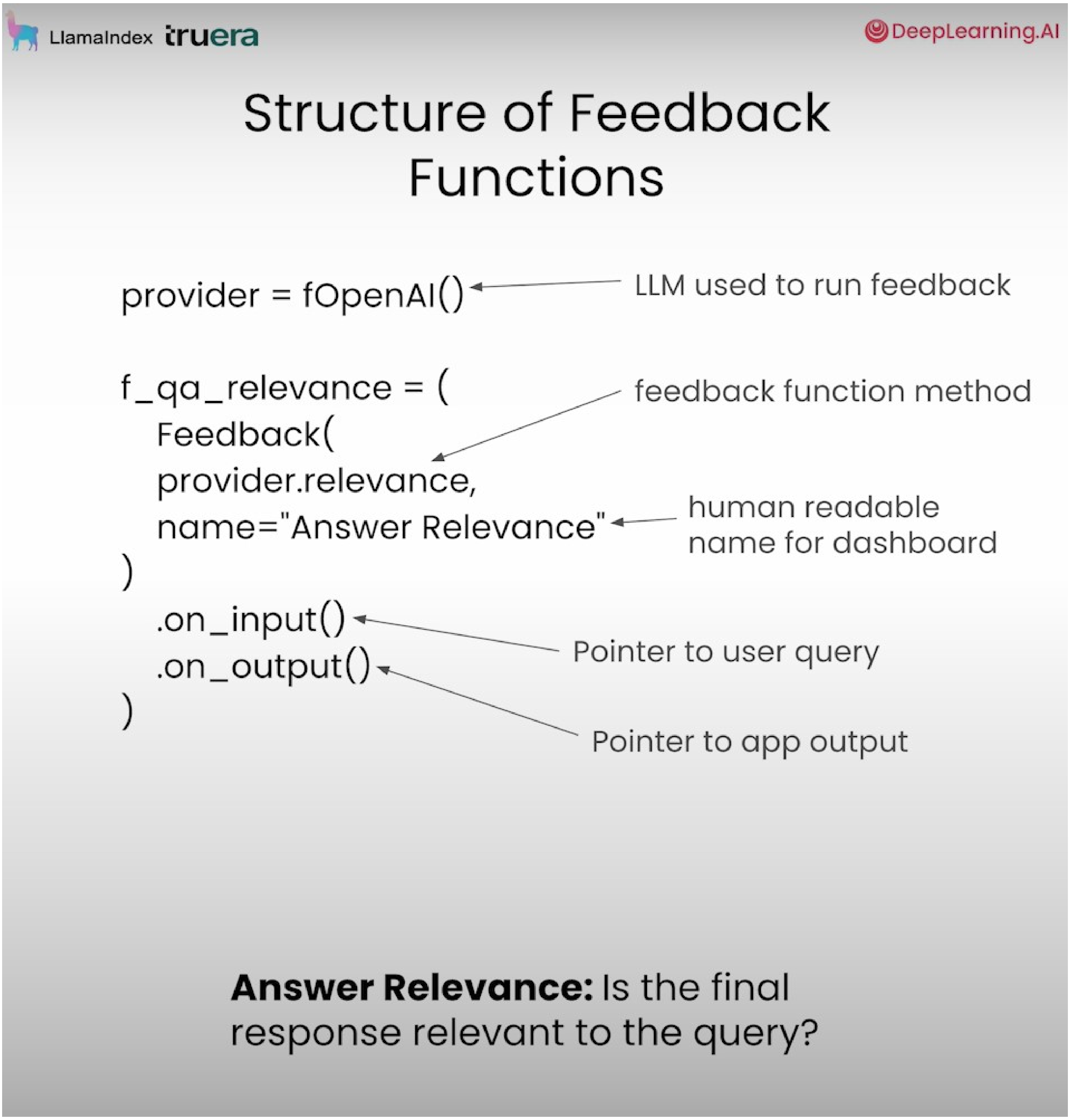
Supporting Evidence: The response provides a clear explanation of how altruism can be beneficial in building a career. It mentions that by helping others, individuals can achieve better outcomes for themselves, create a positive reputation and network, and lead to new opportunities and collaborations. It also highlights that helping others can 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* after reviewing an LLM app's *inputs*, *outputs*, and *intermediate results*.



Structure of Feedback Functions

```
provider = fOpenAI()

f_qa_relevance = (
   Feedback(
   provider.relevance,
   name="Answer Relevance"
)
   .on_input()
   .on_output()
)
```

LLM used to run feedback

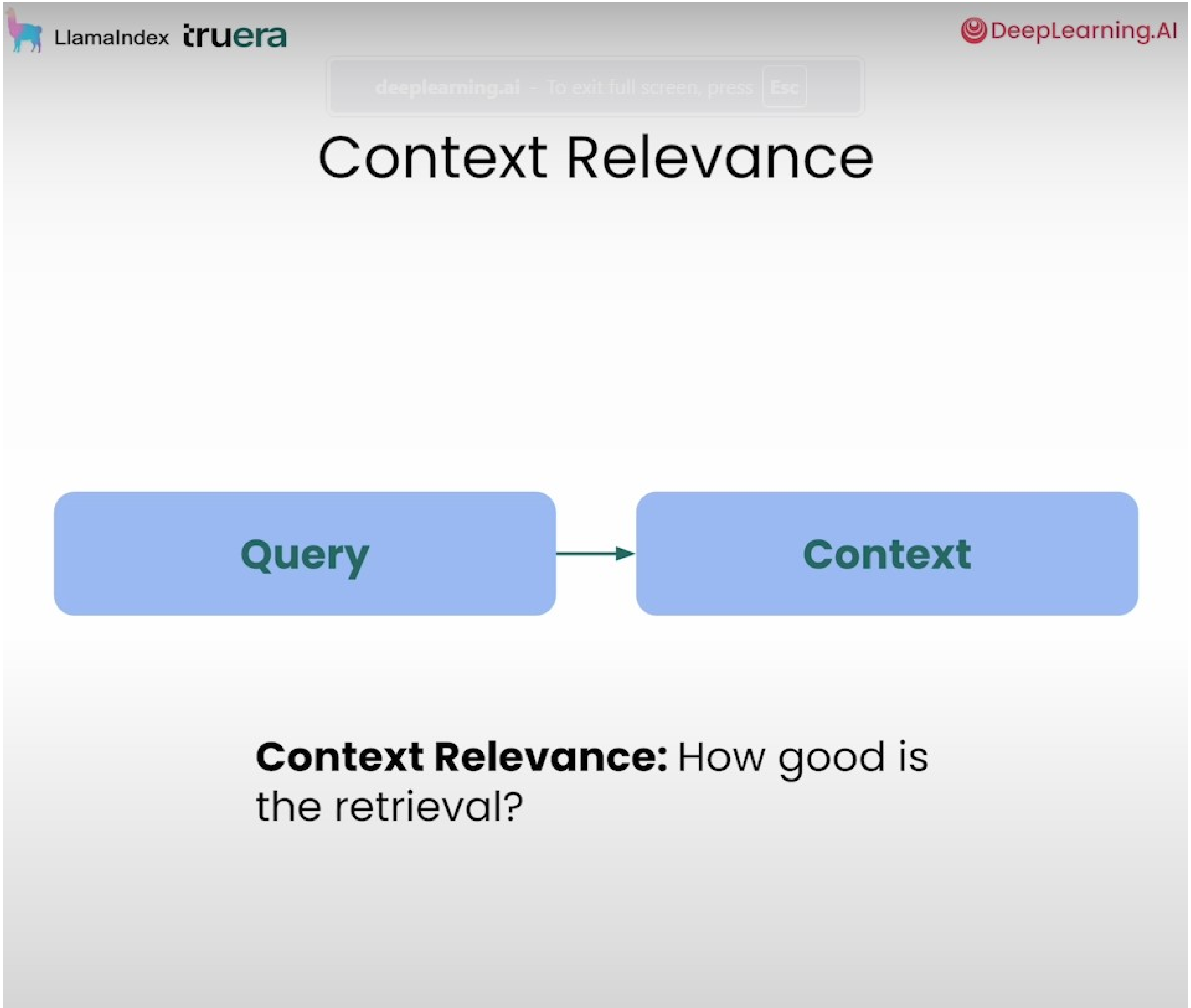
feedback function method

human readable
name for dashboard

Pointer to user query

Pointer to app output

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

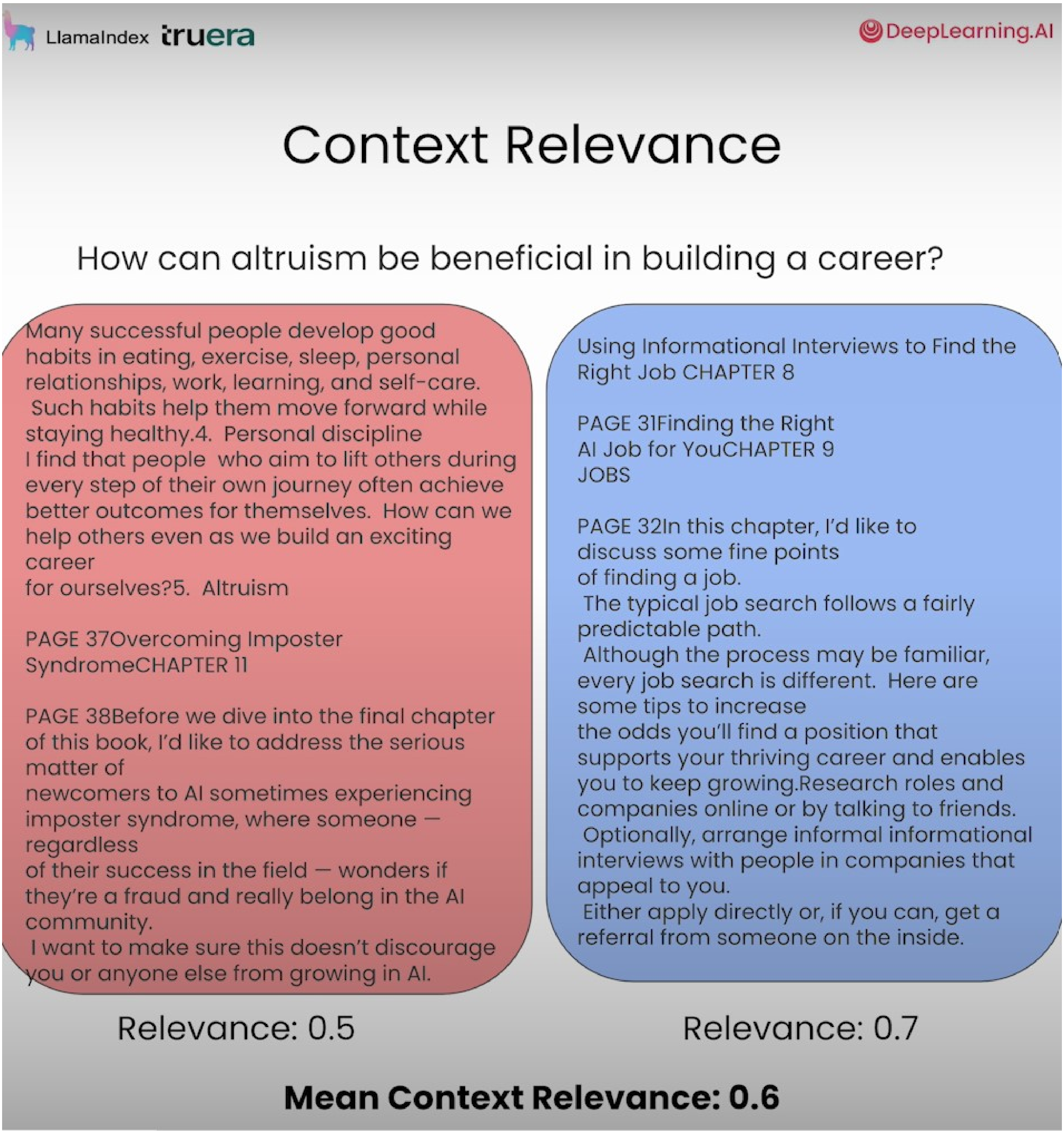


deeplearning.ai - To exit full screen, press Esc

Context Relevance

Query → Context

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 37Overcoming Imposter SyndromeCHAPTER 11

PAGE 38Before 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.

Using Informational Interviews to Find the Right Job CHAPTER 8

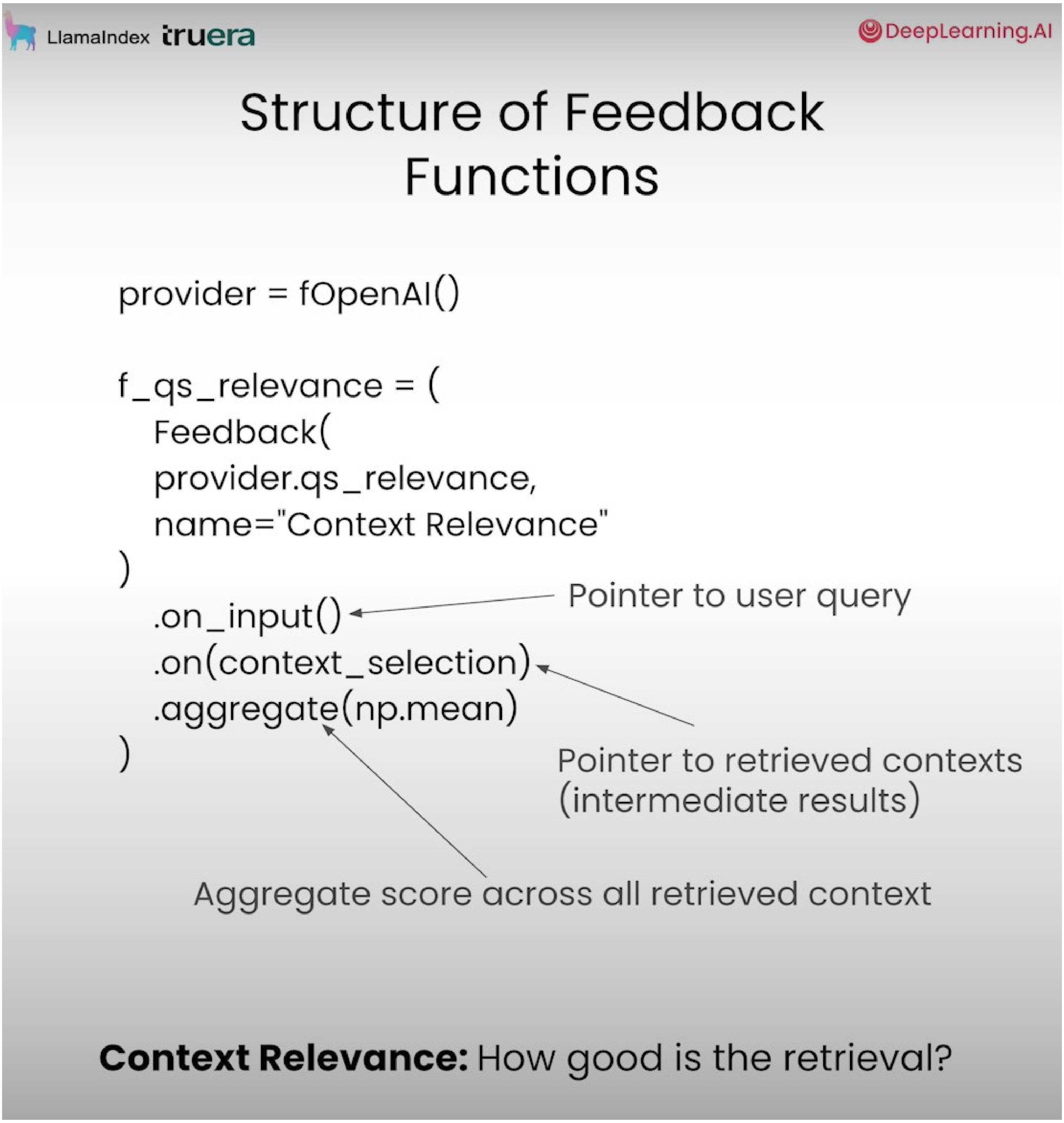
PAGE 31Finding the Right
AI Job for YouCHAPTER 9
JOBS

PAGE 32In 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.

Relevance: 0.5

Relevance: 0.7

Mean Context Relevance: 0.6



Structure of Feedback Functions

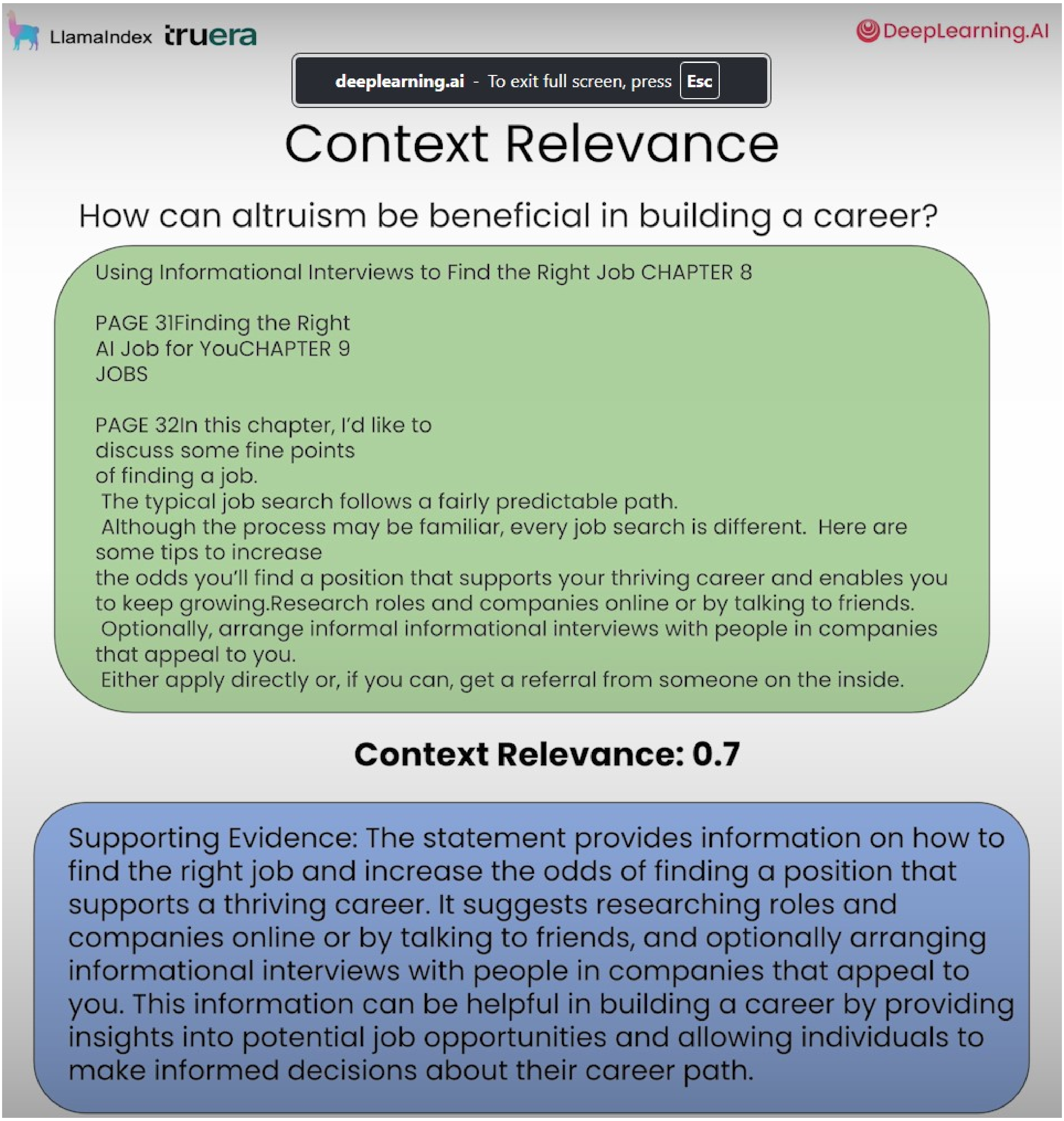
```
provider = fOpenAI()

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 Right Job CHAPTER 8

PAGE 31Finding the Right
AI Job for YouCHAPTER 9
JOBS

PAGE 32In 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.

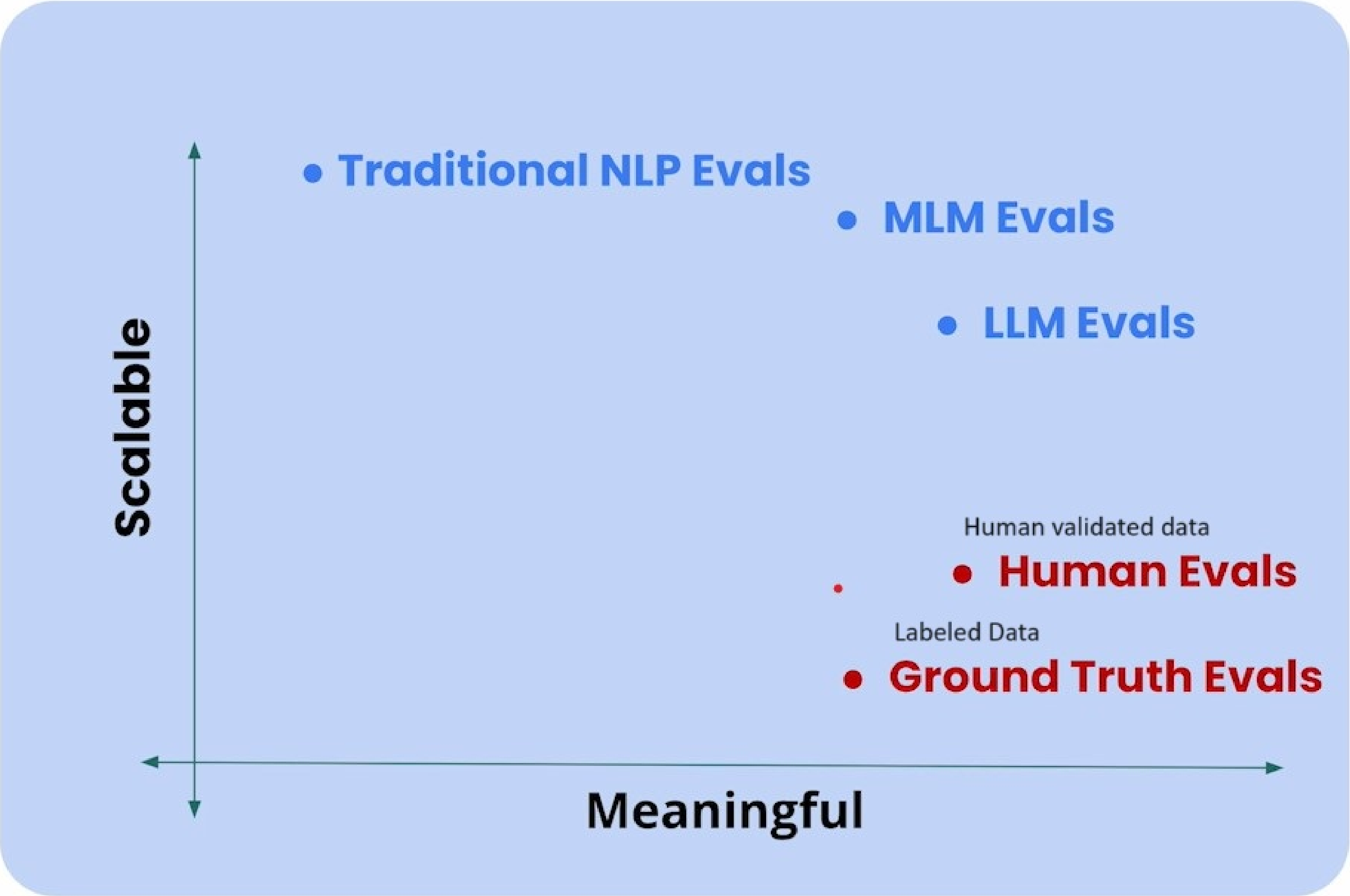
Context Relevance: 0.7

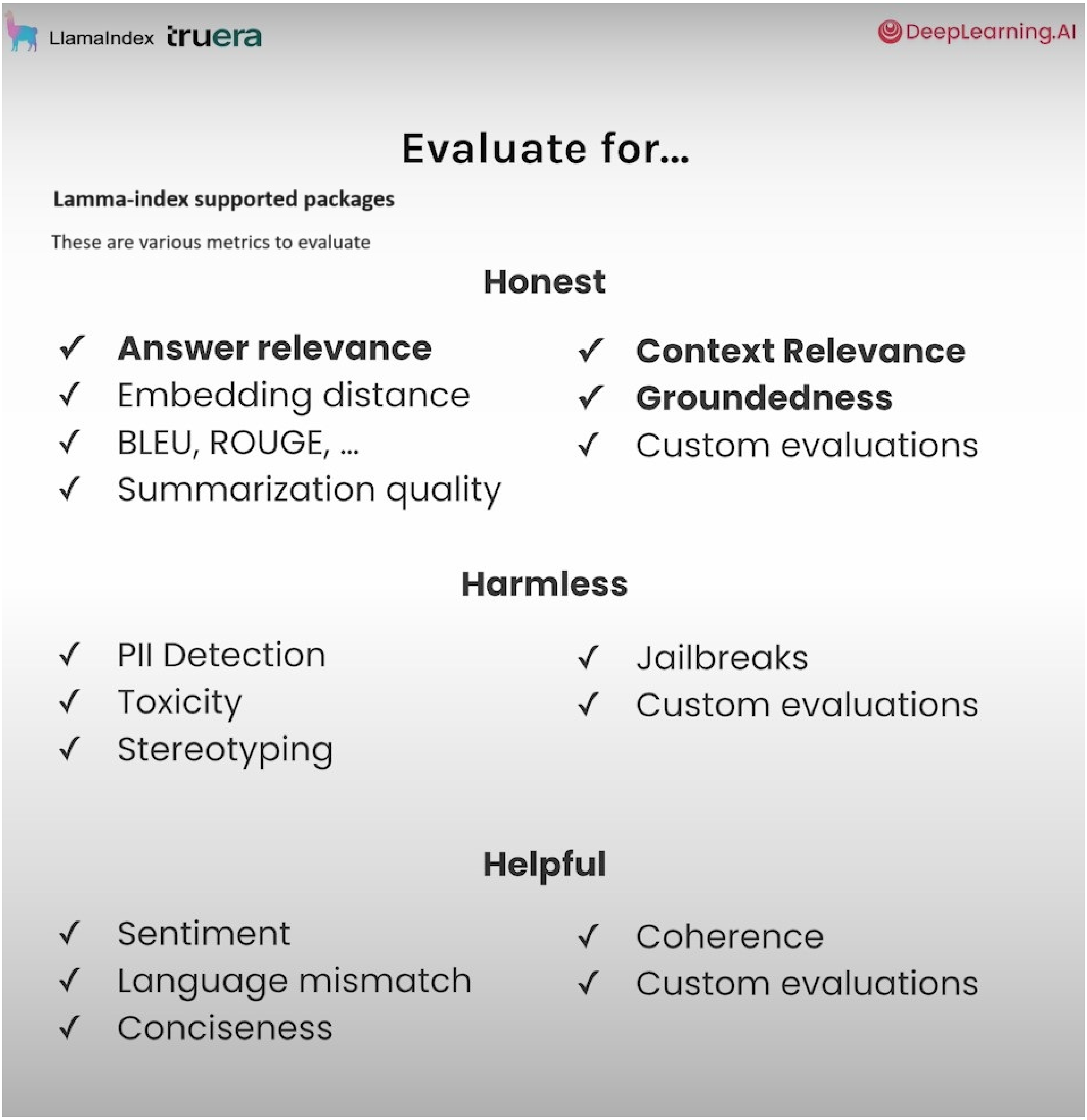
Supporting Evidence: The statement provides information on how to find the right job and increase the odds of finding a position that supports a thriving career. It suggests researching roles and companies online or by talking to friends, and optionally arranging informational interviews with people 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 TurLens 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?
- Experiment with different window sizes
  - What window size results in the best eval metrics?

Feedback Functions can be implemented in different ways





LlamaIndex truera

DeepLearning.AI

Evaluate for...

Lamma-index supported packages

These are various metrics to evaluate

Honest

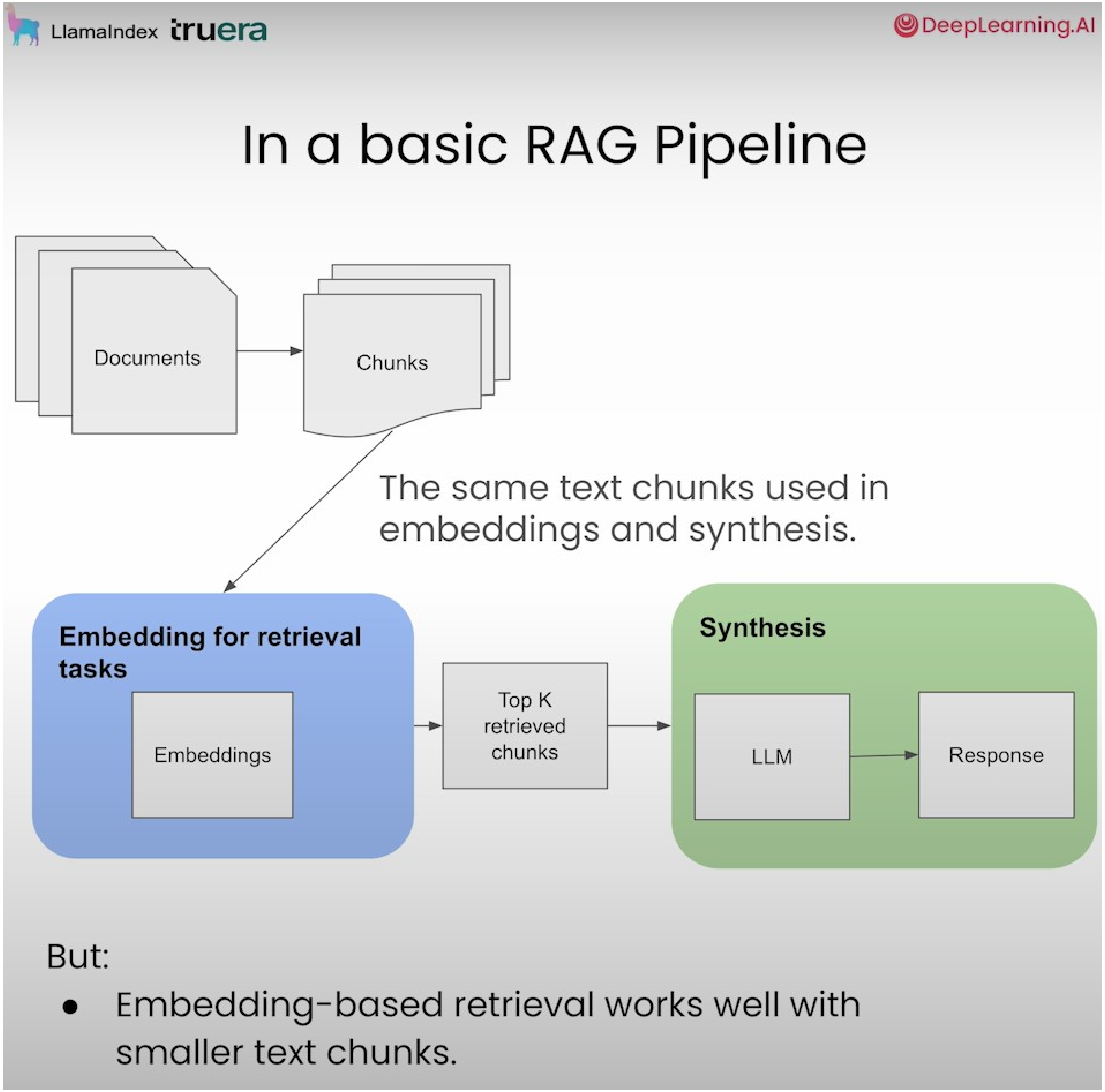
- ✓ **Answer relevance**
- ✓ Embedding distance
- ✓ BLEU, ROUGE, ...
- ✓ Summarization quality
- ✓ **Context Relevance**
- ✓ **Groundedness**
- ✓ Custom evaluations

Harmless

- ✓ PII Detection
- ✓ Toxicity
- ✓ Stereotyping
- ✓ Jailbreaks
- ✓ Custom evaluations

Helpful

- ✓ Sentiment
- ✓ Language mismatch
- ✓ Conciseness
- ✓ Coherence
- ✓ Custom evaluations



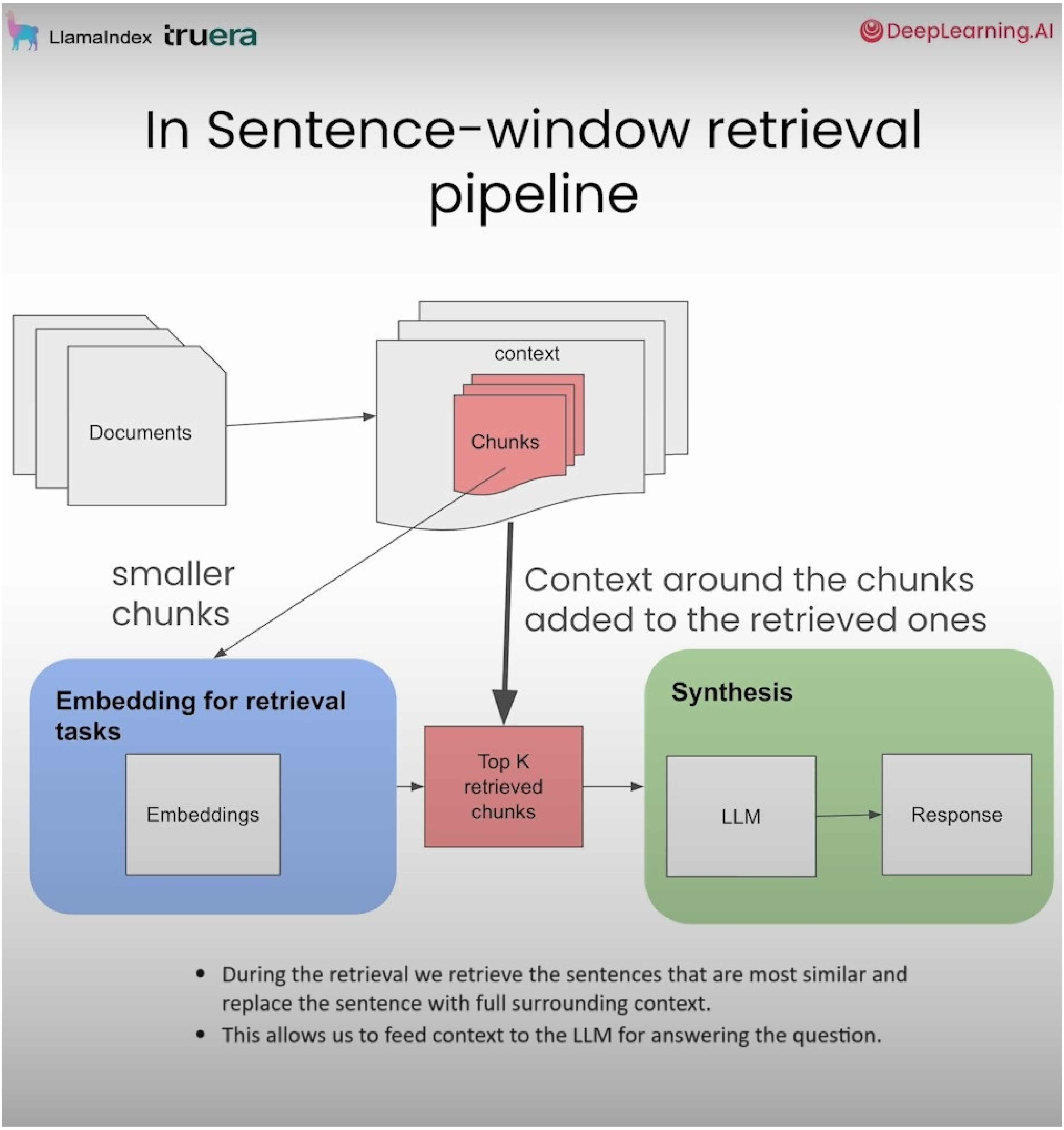
In a basic RAG Pipeline

The same text chunks used in embeddings and synthesis.

But:

- Embedding-based retrieval works well with smaller text chunks.

LlamaIndex truera DeepLearning.AI



In Sentence-window retrieval pipeline

Documents

context

Chunks

smaller chunks

Context around the chunks added to the retrieved ones

Embedding for retrieval tasks

Embeddings

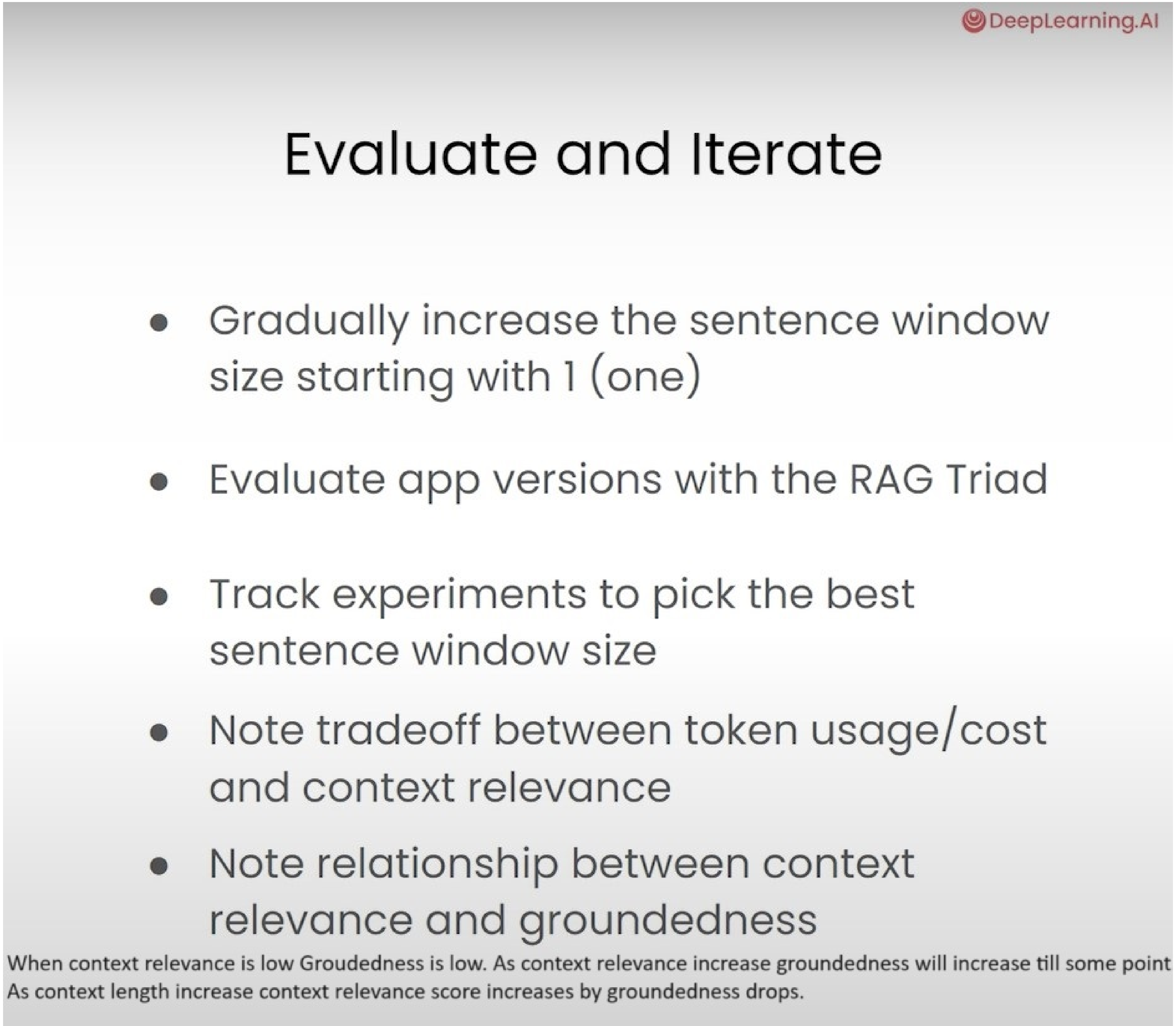
Top K retrieved chunks

Synthesis

LLM

Response

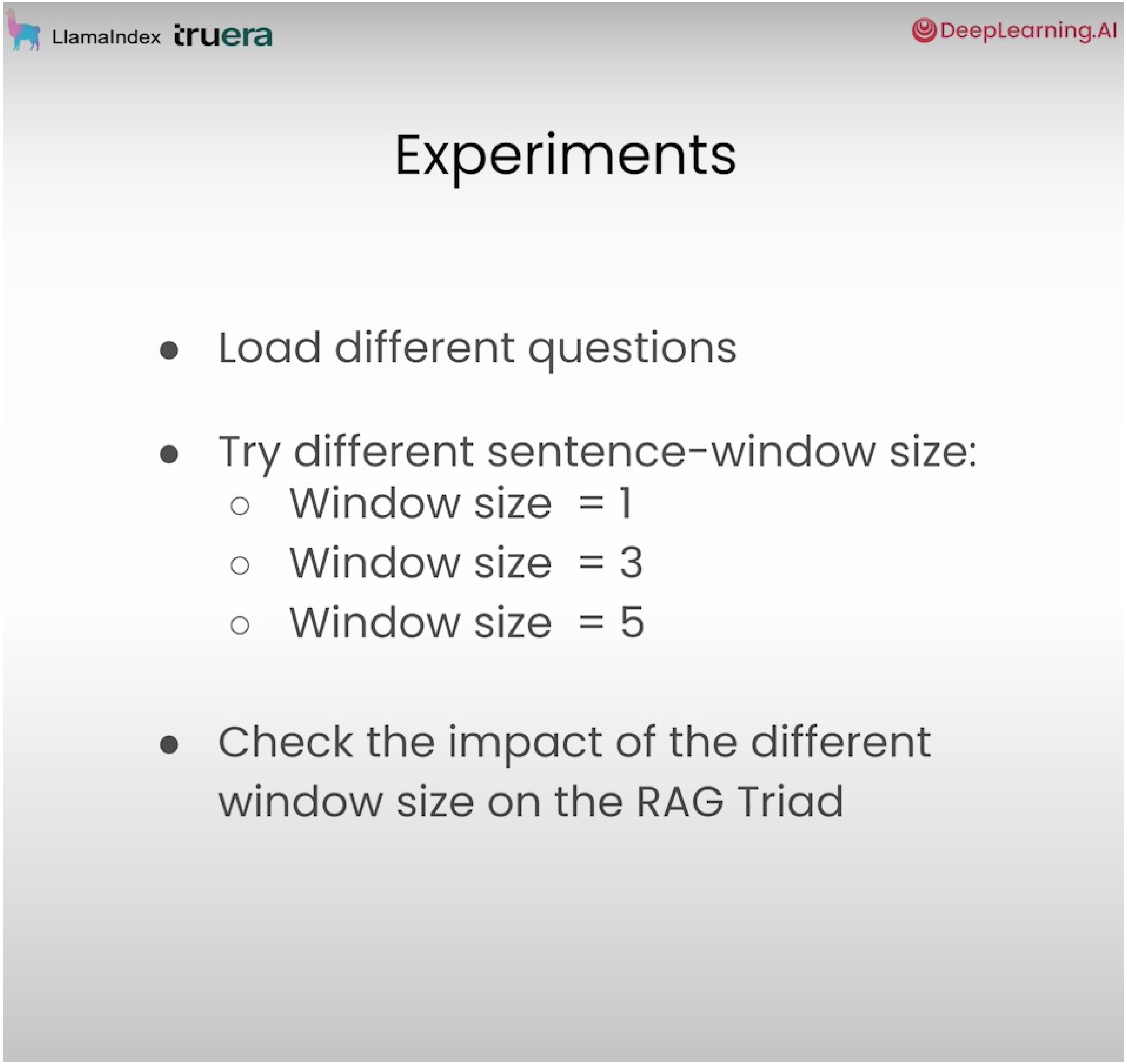
- During the retrieval we retrieve the sentences that are most similar and replace the sentence with full surrounding context.
- This allows us to feed context to the LLM for answering the question.



Evaluate and Iterate

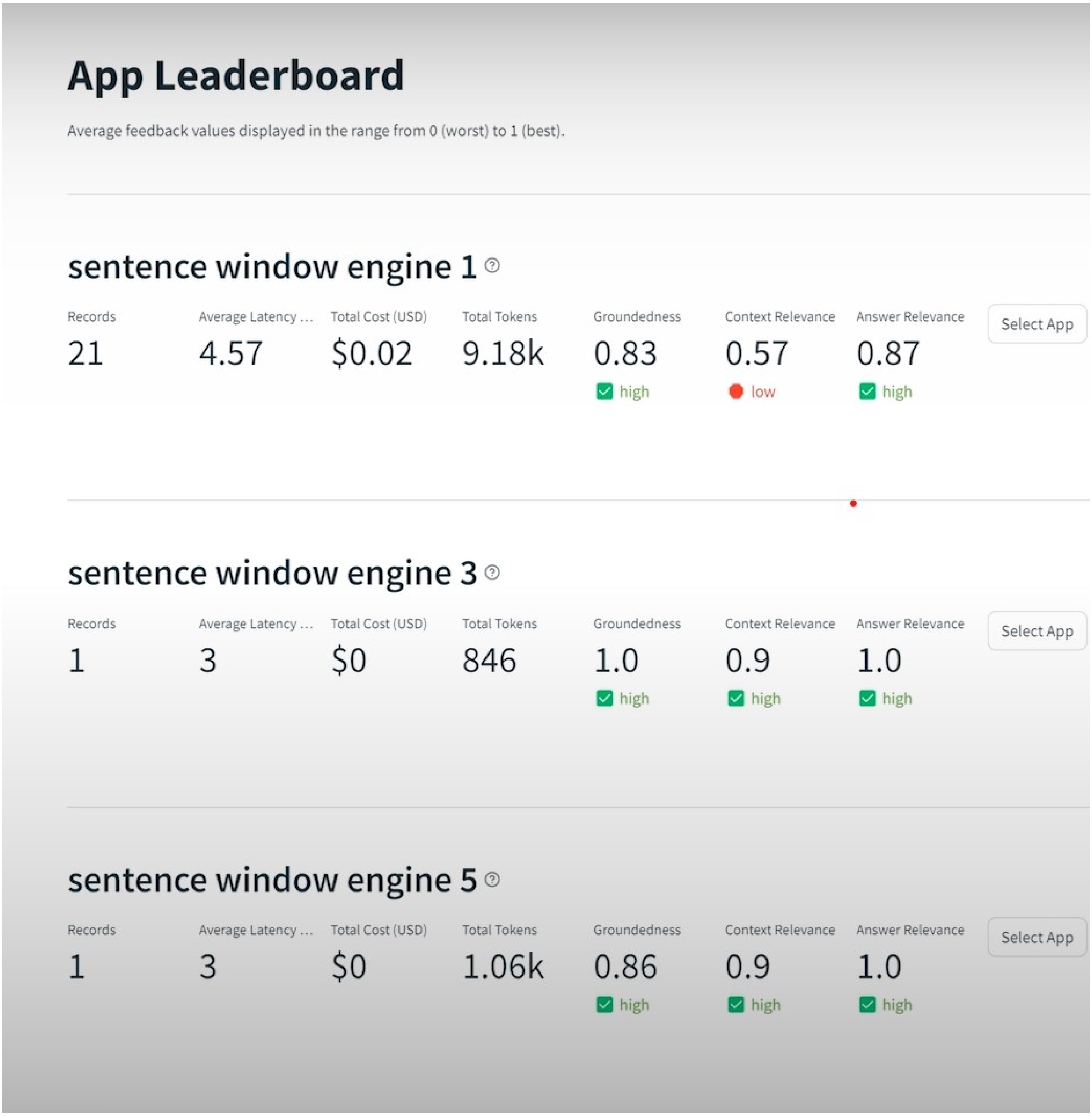
- Gradually increase the sentence window size starting with 1 (one)
- Evaluate app versions with the RAG Triad
- Track experiments to pick the best sentence window size
- Note tradeoff between token usage/cost and context relevance
- Note relationship between context relevance and groundedness

When context relevance is low Groudedness is low. As context relevance increase groundedness will increase till some point
As context length increase context relevance score increases by groundedness drops.



Experiments

- Load different questions
- Try different sentence-window size:
  - Window size = 1
  - Window size = 3
  - Window size = 5
- Check the impact of the different window size on the RAG Triad



App Leaderboard

Average feedback values displayed in the range from 0 (worst) to 1 (best).

sentence window engine 1 ⓘ

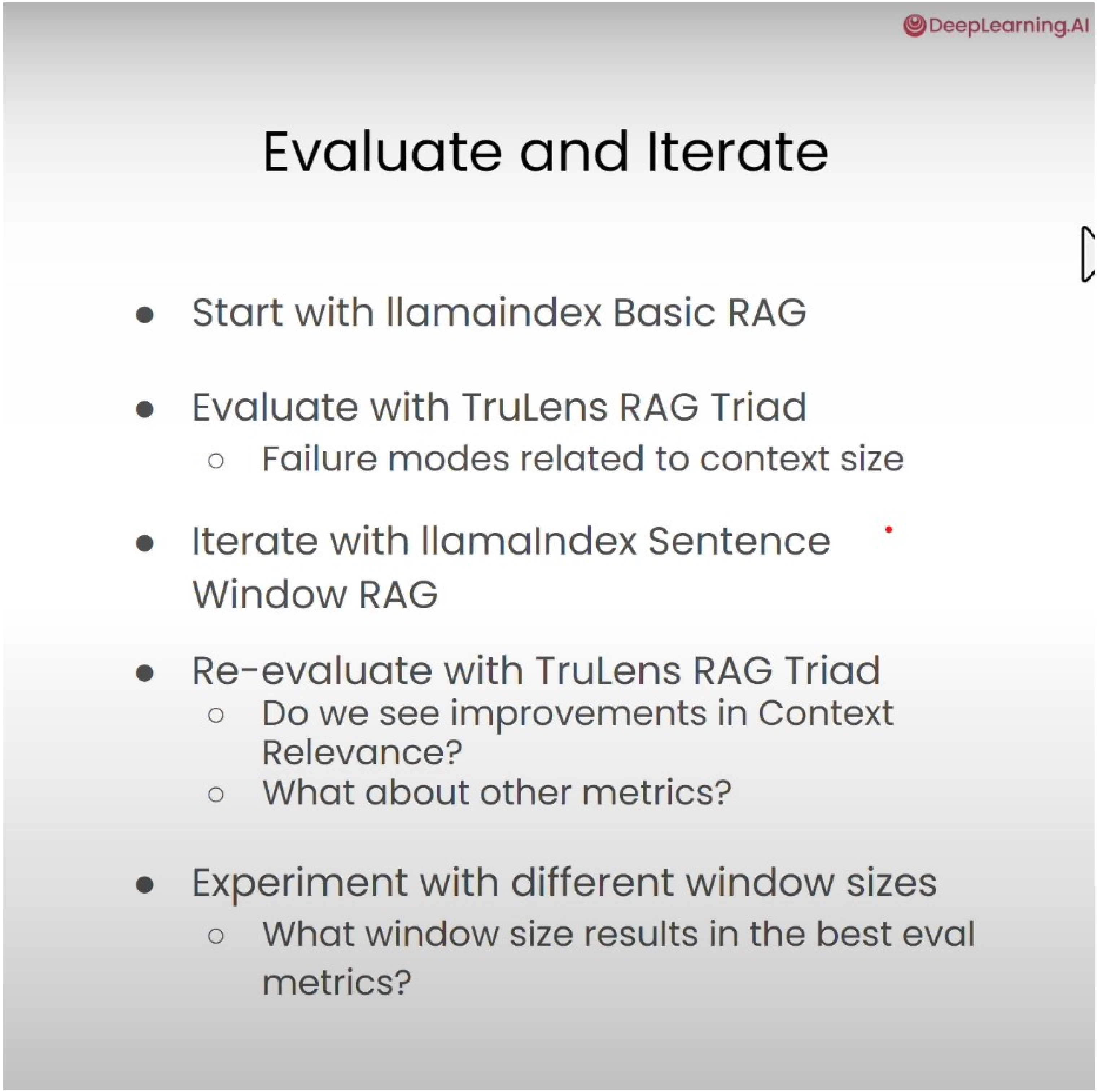
Records	Average Latency ...	Total Cost (USD)	Total Tokens	Groundedness	Context Relevance	Answer Relevance
21	4.57	$0.02	9.18k	0.83 ☑ high	0.57 ● low	0.87 ☑ high

sentence window engine 3 ⓘ

Records	Average Latency ...	Total Cost (USD)	Total Tokens	Groundedness	Context Relevance	Answer Relevance
1	3	$0	846	1.0 ☑ high	0.9 ☑ high	1.0 ☑ high

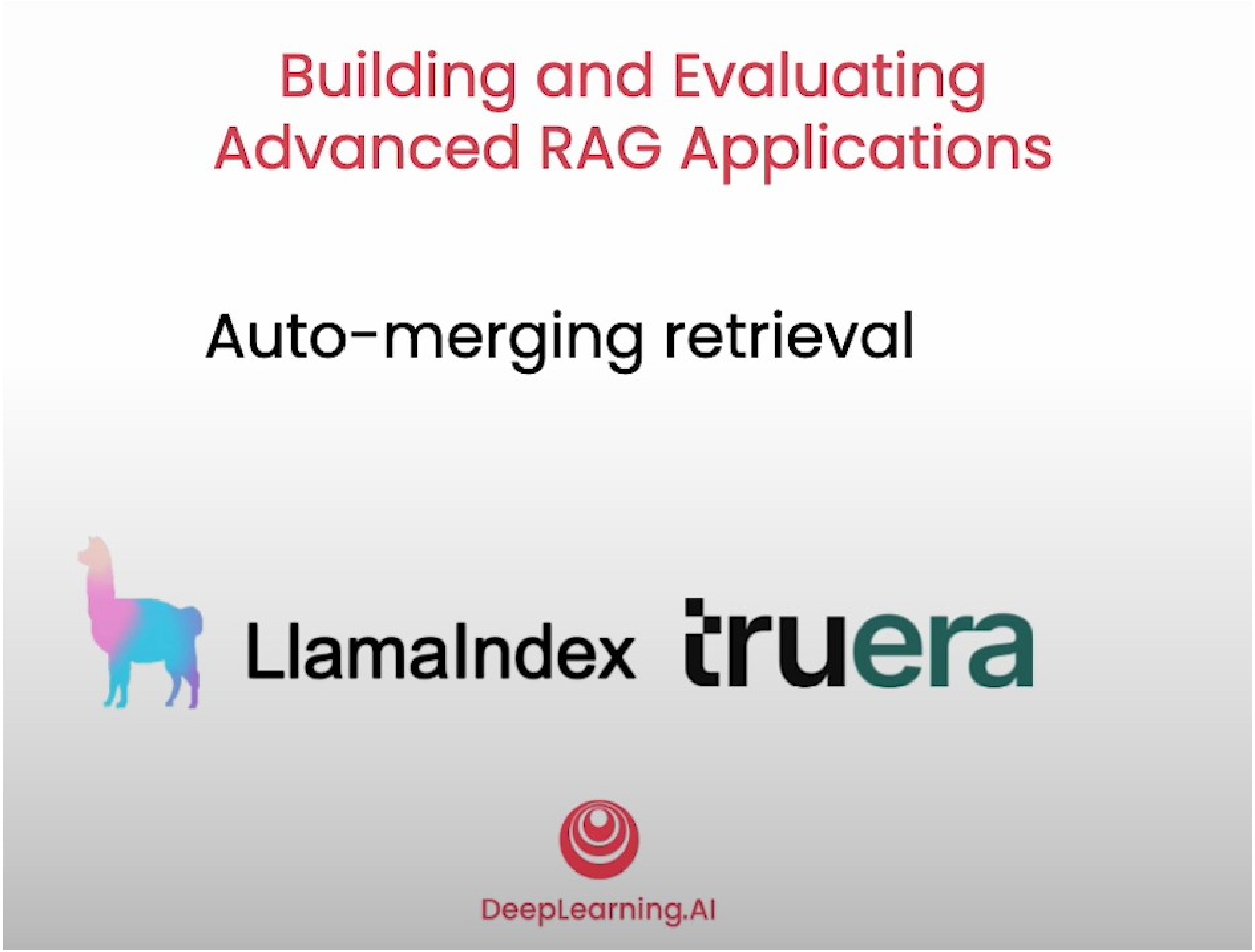
sentence window engine 5 ⓘ

Records	Average Latency ...	Total Cost (USD)	Total Tokens	Groundedness	Context Relevance	Answer Relevance
1	3	$0	1.06k	0.86 ☑ high	0.9 ☑ high	1.0 ☑ high



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?
- Experiment with different window sizes
  - What window size results in the best eval metrics?

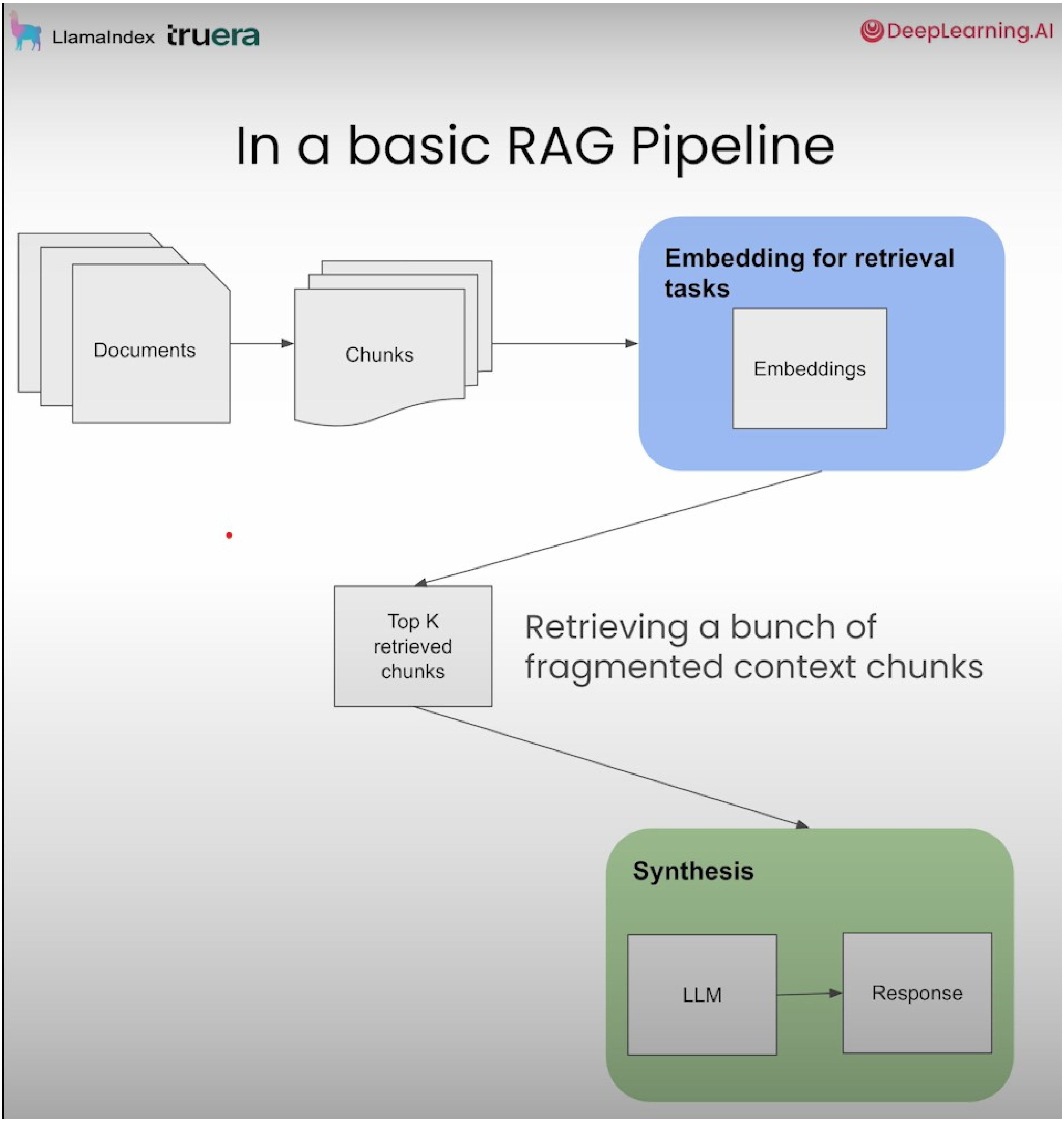


Building and Evaluating Advanced RAG Applications

Auto-merging retrieval

LlamaIndex truera

DeepLearning.AI



In a basic RAG Pipeline

Documents

Chunks

Embedding for retrieval tasks

Embeddings

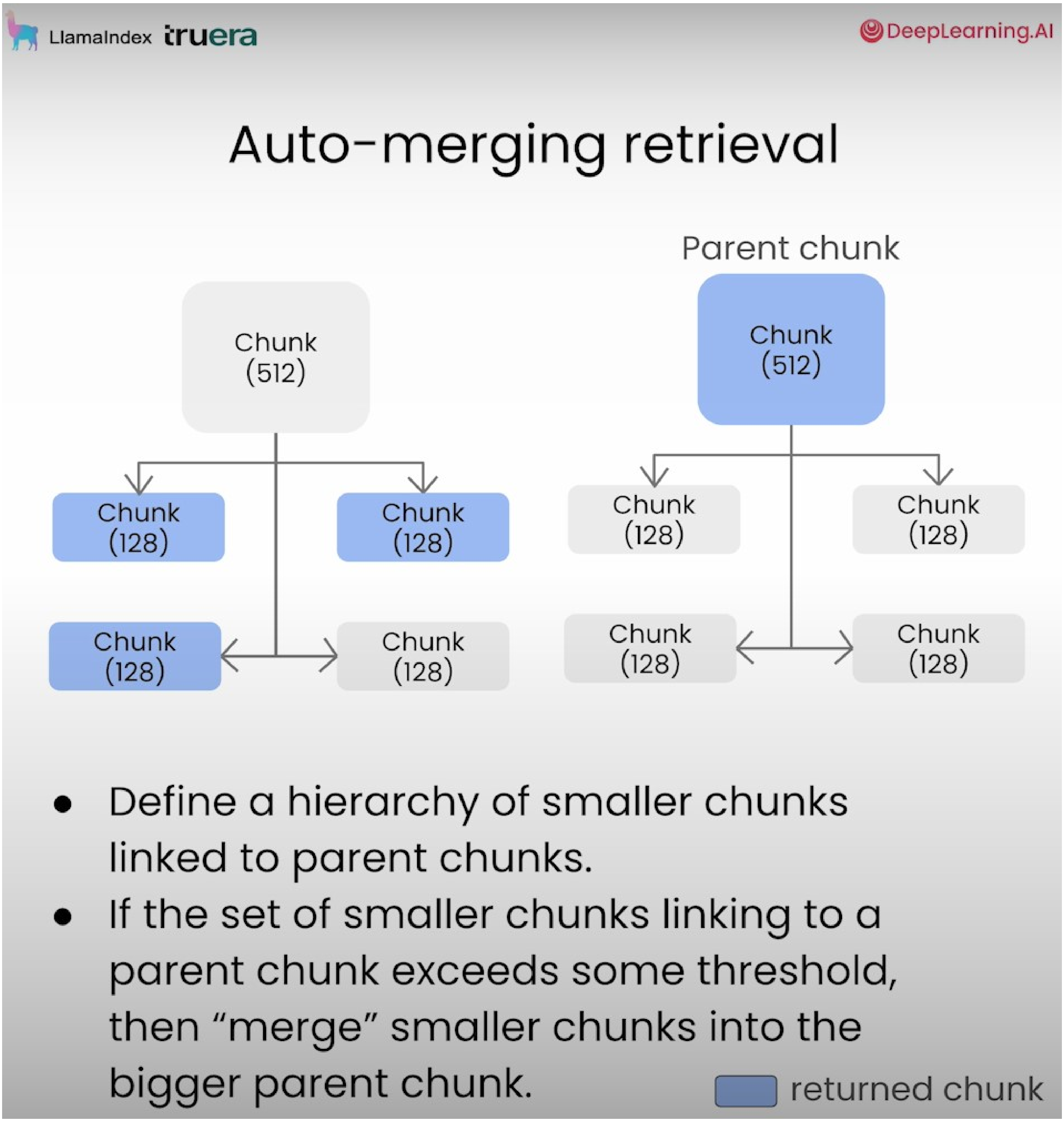
Top K retrieved chunks

Retrieving a bunch of fragmented context chunks

Synthesis

LLM

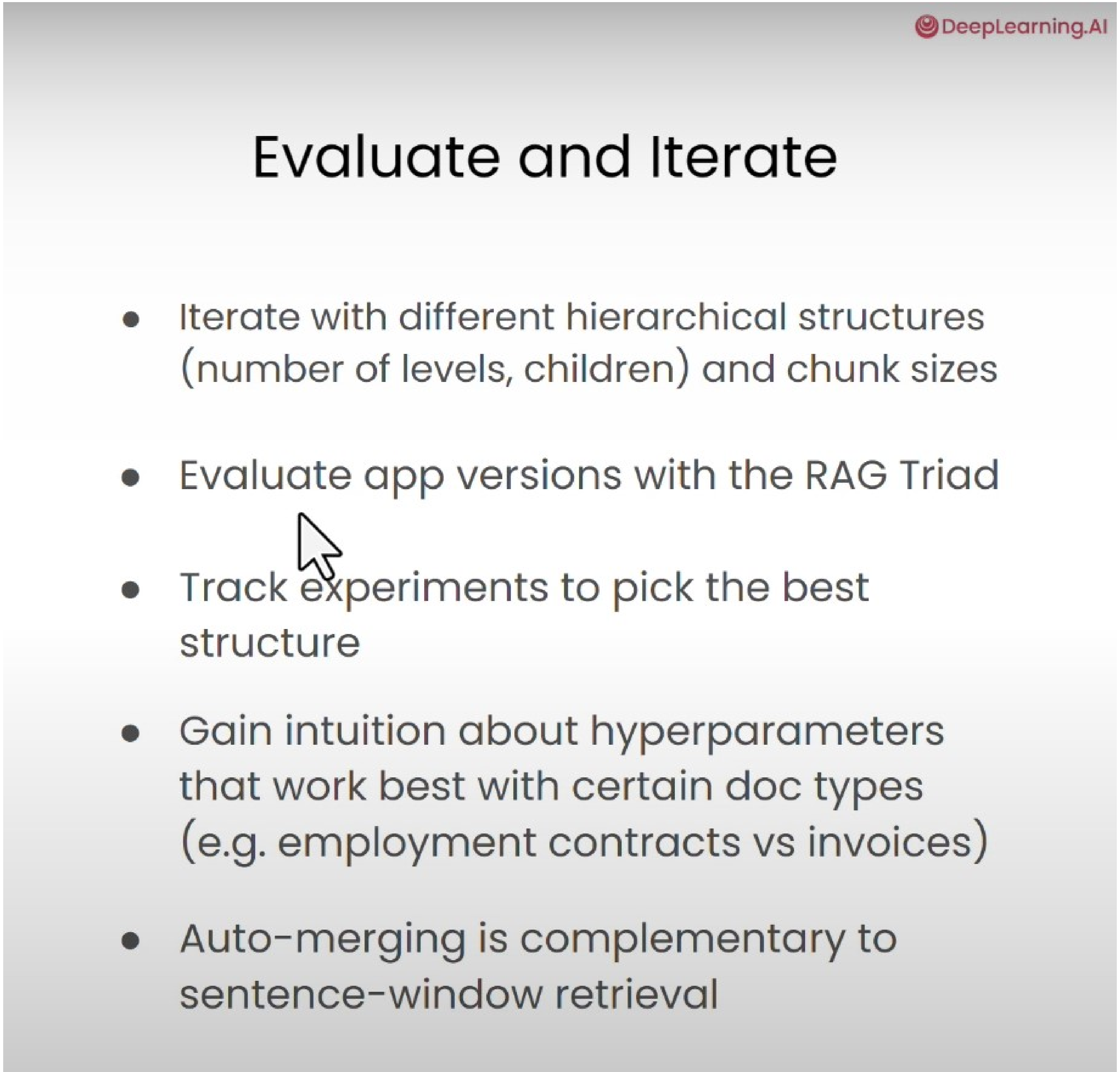
Response



Auto-merging retrieval

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

🟦 returned 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

Evaluate for...

Honest

- ✓ **Answer relevance**
- ✓ Embedding distance
- ✓ BLEU, ROUGE, ...
- ✓ Summarization quality
- ✓ **Context Relevance**
- ✓ **Groundedness**
- ✓ Custom evaluations

Harmless

- ✓ PII Detection
- ✓ Toxicity
- ✓ Stereotyping
- ✓ Jailbreaks
- ✓ Custom evaluations

Helpful

- ✓ Sentiment
- ✓ Language mismatch
- ✓ Conciseness
- ✓ Coherence
- ✓ Custom evaluations