CS329T

Homework 1 Introduction

HW 1:

- Released: Thurs Sept 28th
- Due: Mon Oct 9th

- OFFICE HOURS (will update Ed and website with these)
  - AYUSH'S OH 2 - 3:30 pm on Mondays
  - MICHELLE'S OH 1:15 - 2:45 pm on Thursdays
  - ONLINE OH 5 - 6:30 pm on Wednesdays
- LAB (around Wed Oct 4th, will send out an Ed post to vote for the time)

AGENDA

01

OVERVIEW OF RAG

02

OVERVIEW OF HOMEWORK

03

HOMEWORK WALK THROUGH

04

ADDITIONAL QUESTIONS

Retrieval-Augmented Generation (RAG)

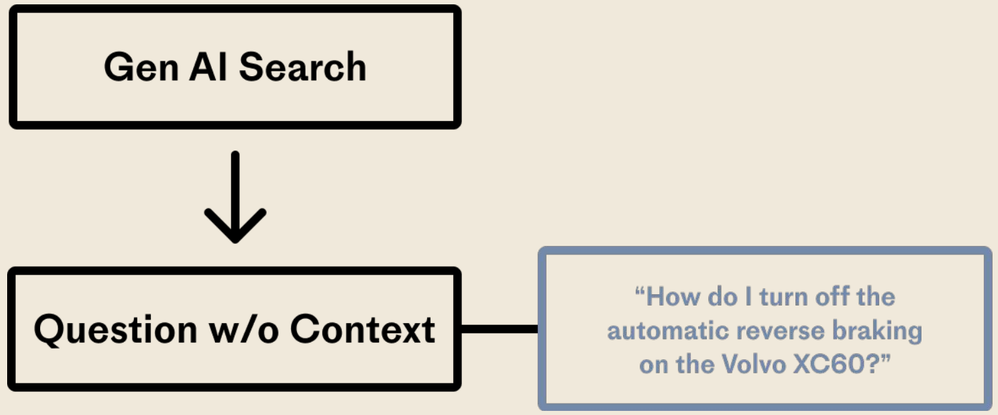
- Foundation models are:
  1. trained offline
  2. trained on massive amounts of general domain corpora

With RAG, we retrieve data from outside of the training set to use as **context**, in addition to the prompt. This context may make the model more up-to-date or task-specialized.

Retrieval-Augmented Generation (RAG)

- Foundation models are:
  1. trained offline
  2. trained on massive amounts of general domain corpora

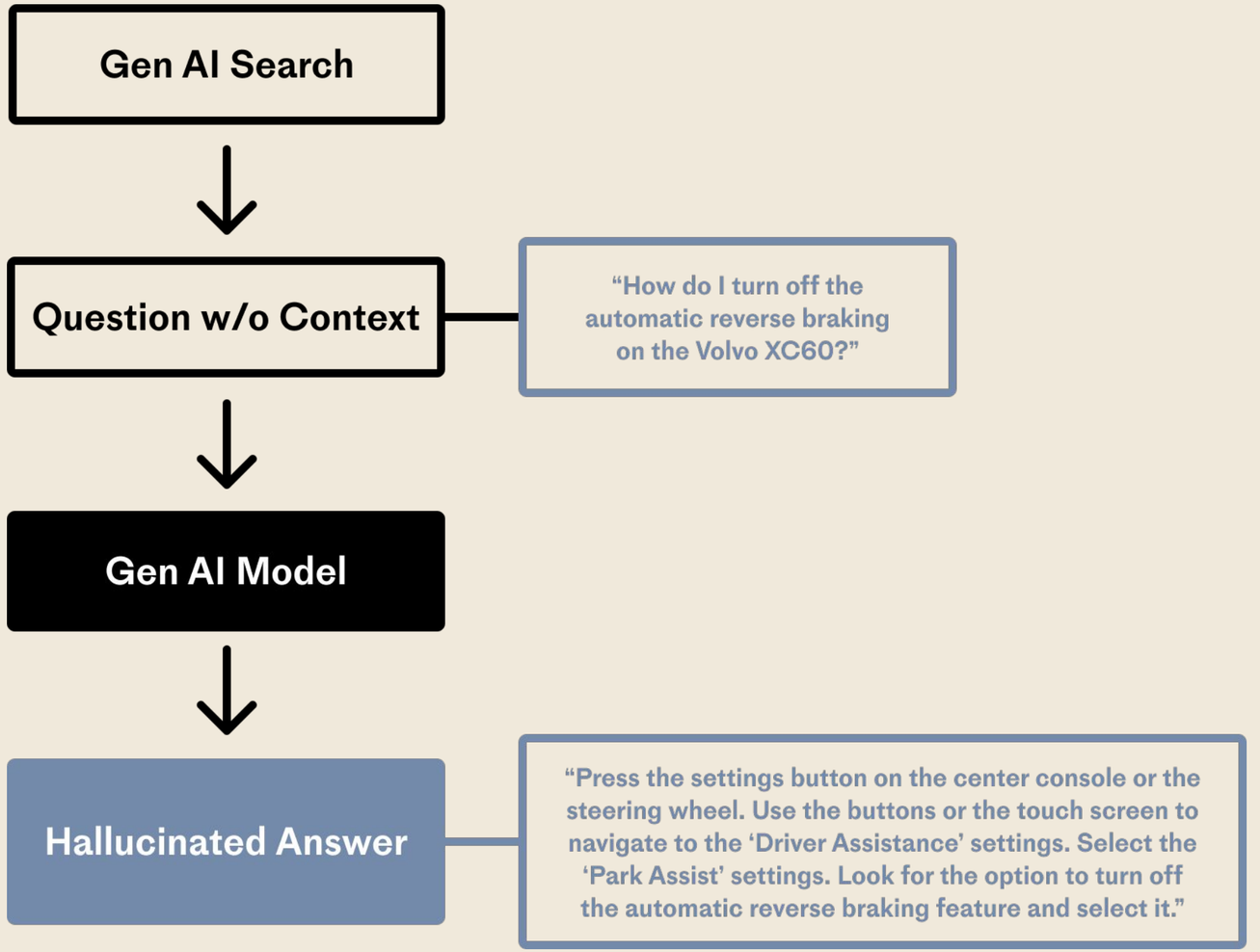
With RAG, we retrieve data from outside of the training set to use as **context**, in addition to the prompt. This context may make the model more up-to-date or task-specialized.



Retrieval-Augmented Generation (RAG)

- Foundation models are:
  1. trained offline
  2. trained on massive amounts of general domain corpora

With RAG, we retrieve data from outside of the training set to use as **context**, in addition to the prompt. This context may make the model more up-to-date or task-specialized.

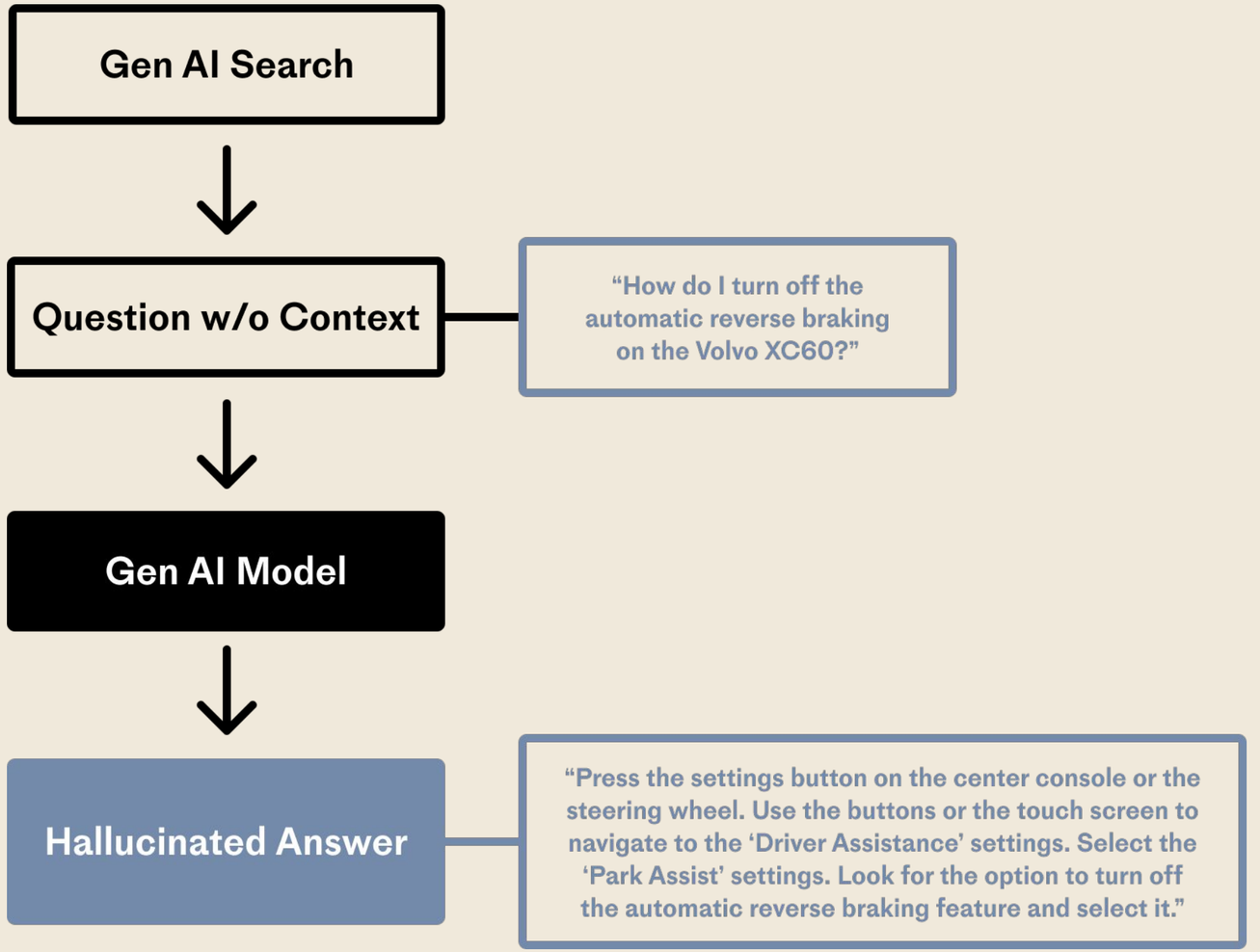


Retrieval-Augmented Generation (RAG)

- Foundation models are:
  1. trained offline
  2. trained on massive amounts of general domain corpora

With RAG, we retrieve data from outside of the training set to use as **context**, in addition to the prompt. This context may make the model more up-to-date or task-specialized.

What if the context included something like the most updated Volvo user's manual?

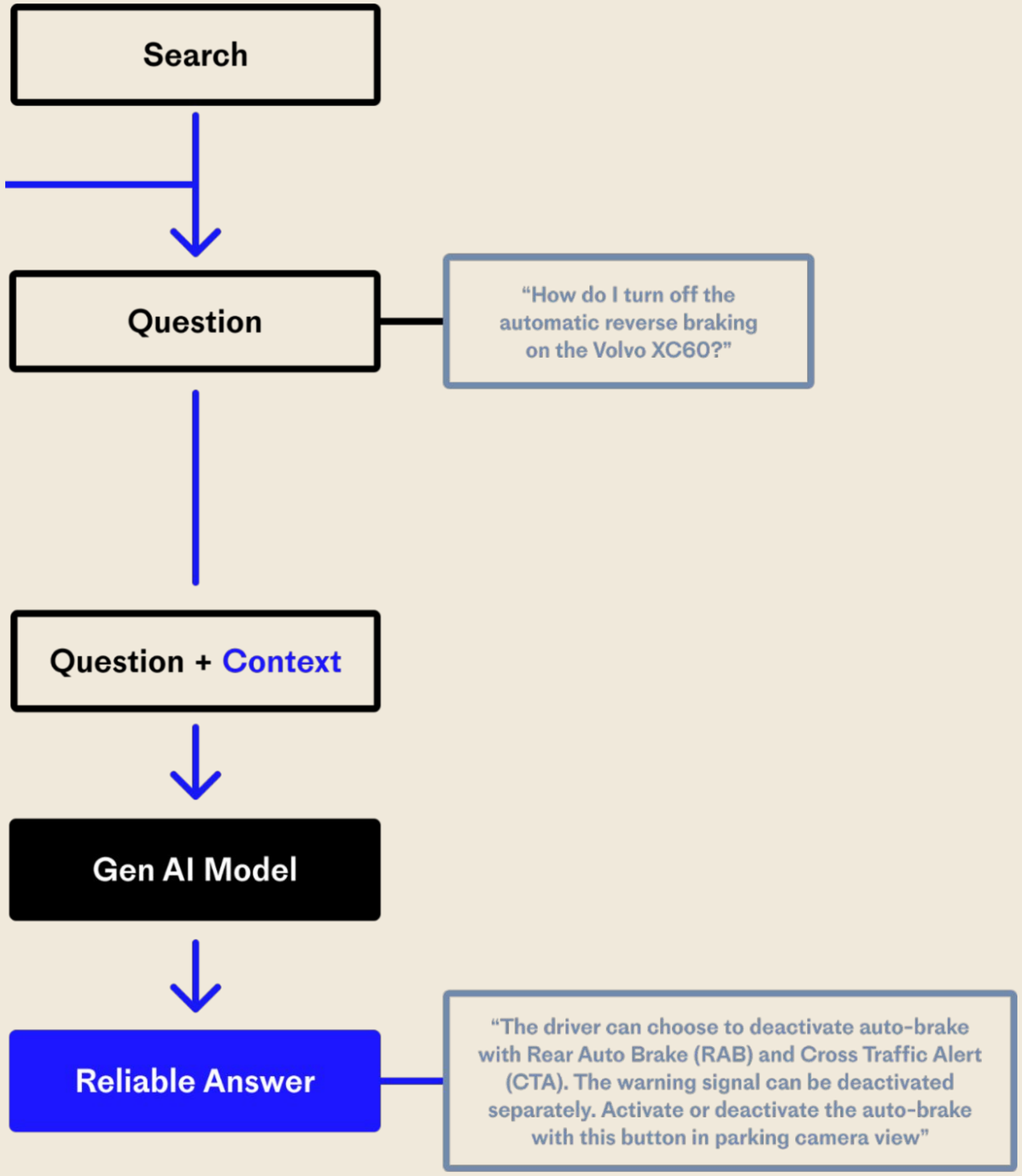


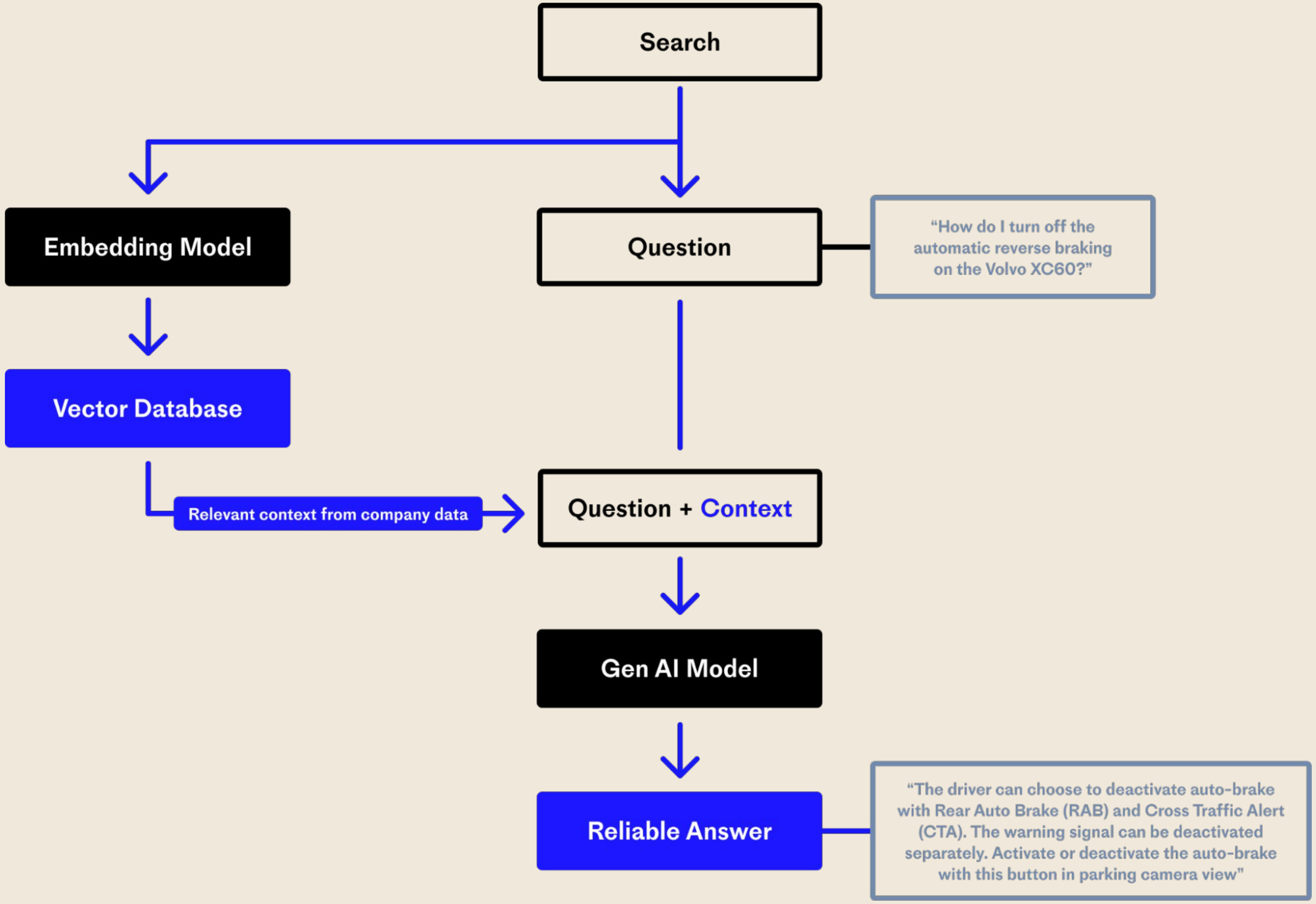
Retrieval-Augmented Generation (RAG)

- Meta released "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" (Lewis et al.) in 2020
- In a RAG system, you are asking the model to respond to a question by "browsing through the content in a book, as opposed to trying to remember facts from memory" or with the help of "a cue card containing the critical points for your LLM to see"

Retrieval-Augmented Generation (RAG)

- Meta released "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" (Lewis et al.) in 2020
- In a RAG system, you are asking the model to respond to a question by "browsing through the content in a book, as opposed to trying to remember facts from memory" or with the help of "a cue card containing the critical points for your LLM to see"
- Two phases of Naive RAG
  1. **Indexing phase (retrieval)**
     a. Retrieve *context*: search for information relevant to the user's prompt in a set of data
  2. **Querying phase (synthesis)**
     a. Augment prompt with context and generate response


Search
Question
"How do I turn off the automatic reverse braking on the Volvo XC60?"
Question + Context
Gen AI Model
Reliable Answer
"The driver can choose to deactivate auto-brake with Rear Auto Brake (RAB) and Cross Traffic Alert (CTA). The warning signal can be deactivated separately. Activate or deactivate the auto-brake with this button in parking camera view"


Search
Embedding Model
Question
"How do I turn off the automatic reverse braking on the Volvo XC60?"
Vector Database
Relevant context from company data
Question + Context
Gen AI Model
Reliable Answer
"The driver can choose to deactivate auto-brake with Rear Auto Brake (RAB) and Cross Traffic Alert (CTA). The warning signal can be deactivated separately. Activate or deactivate the auto-brake with this button in parking camera view"

Homework Overview

- Setup
  a. LlamaIndex - Data Management, Query Engine
  b. Milvus - Memory Store
  c. OpenAI - Embeddings, LLM
  d. TruLens - Evaluation
- Defining prompts and a relevant document collection
  a. Prompts: City-related questions
  b. Document collection: Wikipedia

Homework Overview

- Setup
  a. LlamaIndex - Data Management, Query Engine
  b. Milvus - Memory Store
  c. OpenAI - Embeddings, LLM
  d. TruLens - Evaluation
- Defining prompts and a relevant document collection
  a. Prompts: City-related questions
  b. Document collection: Wikipedia
- Question 1: Building a prototype RAG
- Question 2: Initializing RAG evaluation metrics
- Question 3: Finding the best RAG app configuration using evaluation

Homework Overview

- Evaluation step done with TruLens: https://www.trulens.org/trulens_eval/function_definitions/#function-definitions

1. **Answer relevance** (question-answer relevance) is best for measuring the relationship of the final answer to the user inputed question.

Homework Overview

- Evaluation step done with TruLens: https://www.trulens.org/trulens_eval/function_definitions/#function-definitions

1. **Answer relevance** (question-answer relevance) is best for measuring the relationship of the final answer to the user inputed question.
2. **Context relevance** (question-statement relevance) is best for measuring the relationship of a provided context to the user inputed question.

Homework Overview

- Evaluation step done with TruLens: https://www.trulens.org/trulens_eval/function_definitions/#function-definitions

1. **Answer relevance** (question-answer relevance) is best for measuring the relationship of the final answer to the user inputed question.
2. **Context relevance** (question-statement relevance) is best for measuring the relationship of a provided context to the user inputed question.
3. **Groundedness** attempts to check if the final answer is grounded in its supplied contexts on a scale from 1 to 10.

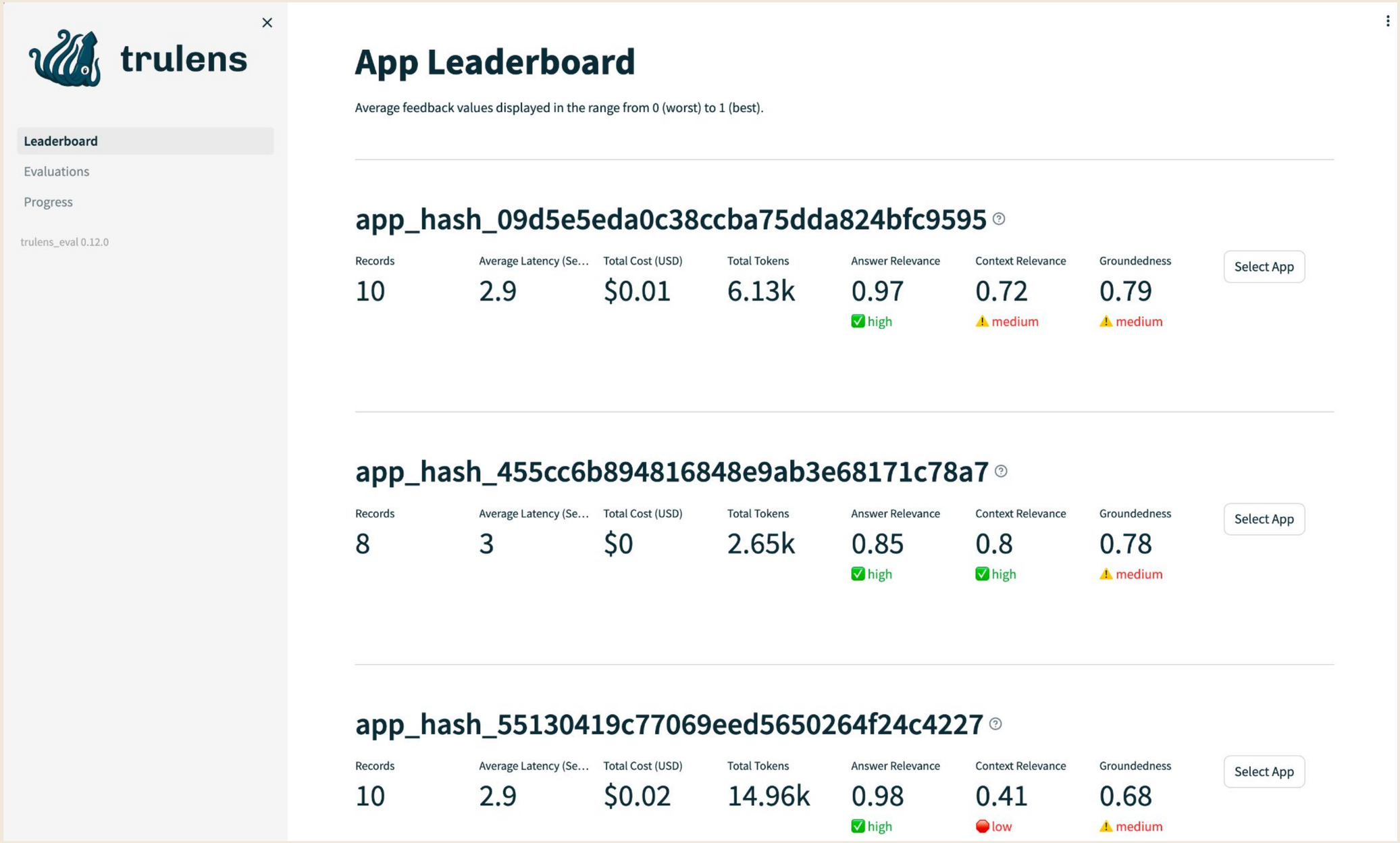
Resources

- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks https://arxiv.org/abs/2005.11401v4
- https://research.ibm.com/blog/retrieval-augmented-generation-RAG
- https://docs.aws.amazon.com/sagemaker/latest/dg/jumpstart-foundation-models-customize-rag.html
- https://www.pinecone.io/learn/retrieval-augmented-generation/

- TruLens eval: https://www.trulens.org/trulens_eval/function_definitions/
- LlamaIndex Building RAG from Scratch (Lower-Level): https://gpt-index.readthedocs.io/en/stable/end_to_end_tutorials/low_level/root.html

Homework Walk Through

TruLens Dashboard



trulens

Leaderboard

Evaluations

Progress

trulens_eval 0.12.0

Selected LLM Application: app_hash_09d5e5eda0c38ccba75dda824bfc9595

`Selected Record ID: record_hash_842afee1bd97eb786fe27fbb49c53901`

Input [`Select.RecordInput`]

What's the best national park near Honolulu

Response [`Select.RecordOutput`]

The best national park near Honolulu is the National Memorial Cemetery of the Pacific, located in Punchbowl Crater.

Metadata

Feedback

Answer Relevance = 0.9

	prompt	response	result	reason
0	What's the best national park near Honolulu	The best national park near Honolulu is the National Memorial Cemetery of the I	0.9	Supporting Evidence: The respons

Context Relevance = 0.1

	question	statement	result	reason
0	What's the best national park near Honolulu	The John A. Burns School of Medicine, part of the University of Hawai'i at Mānoa, is a	0.1	Supporting Evidence: The stat

Groundedness = 1

	source	statement	result	reasor
0	The John A. Burns School of Medicine, part of the University of Hawai'i at Mānoa, is a	The best national park near Honolulu is the National Memorial Cemetery of the I	1.0	Staten