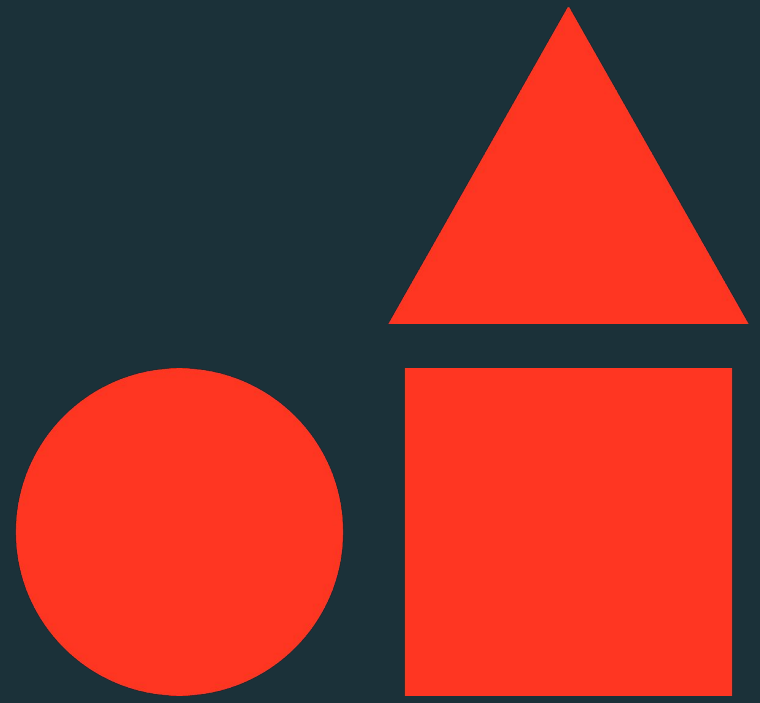
databricks

LLM Fundamentals

Accelerating LLM Apps to Production



Brian Law - Snr Specialist Solution Architect

April 2024

Housekeeping

- This presentation will be recorded and we will share these materials after the session
- We will walk through codes and you can follow along later
- Use the Q&A function to ask questions
- If we do not answer your question during the event, we will follow-up with you afterwards to get you the information you need!
- Please fill out the survey at the end of the session so that we can improve our future sessions

Recap of Part 1

Building gen AI applications on Databricks

Data-centric AI

Gen AI

- Custom models
- Model serving
- RAG

End-to-end AI

- MLOps (MLflow)
- AutoML
- Monitoring
- Governance

Data Science & AI
Mosaic AI

ETL & Real-time Analytics
Delta Live Tables

Orchestration
Workflows

Data Warehousing
Databricks SQL

Data Intelligence Engine
Use generative AI to understand the semantics of your data

Unity Catalog
Securely get insights in natural language

Delta Lake
Data layout is automatically optimized based on usage patterns

Open Data Lake

All raw data
(Logs, texts, audio, video, images)

Transforming language to model

From Words to Math

English Language	Tokens	Vectors
A language structured with: - an alphabet - Words - sentences - paragraphs	Mathematical encoding of parts of words:	Mathematical encoding of entire sentences and paragraphs:
Ex - Cat - Running	Ex - [40] - [10 12]	Ex: - [0 12 32 127]

What makes up a LLM Application?

3 things you need for success

The model

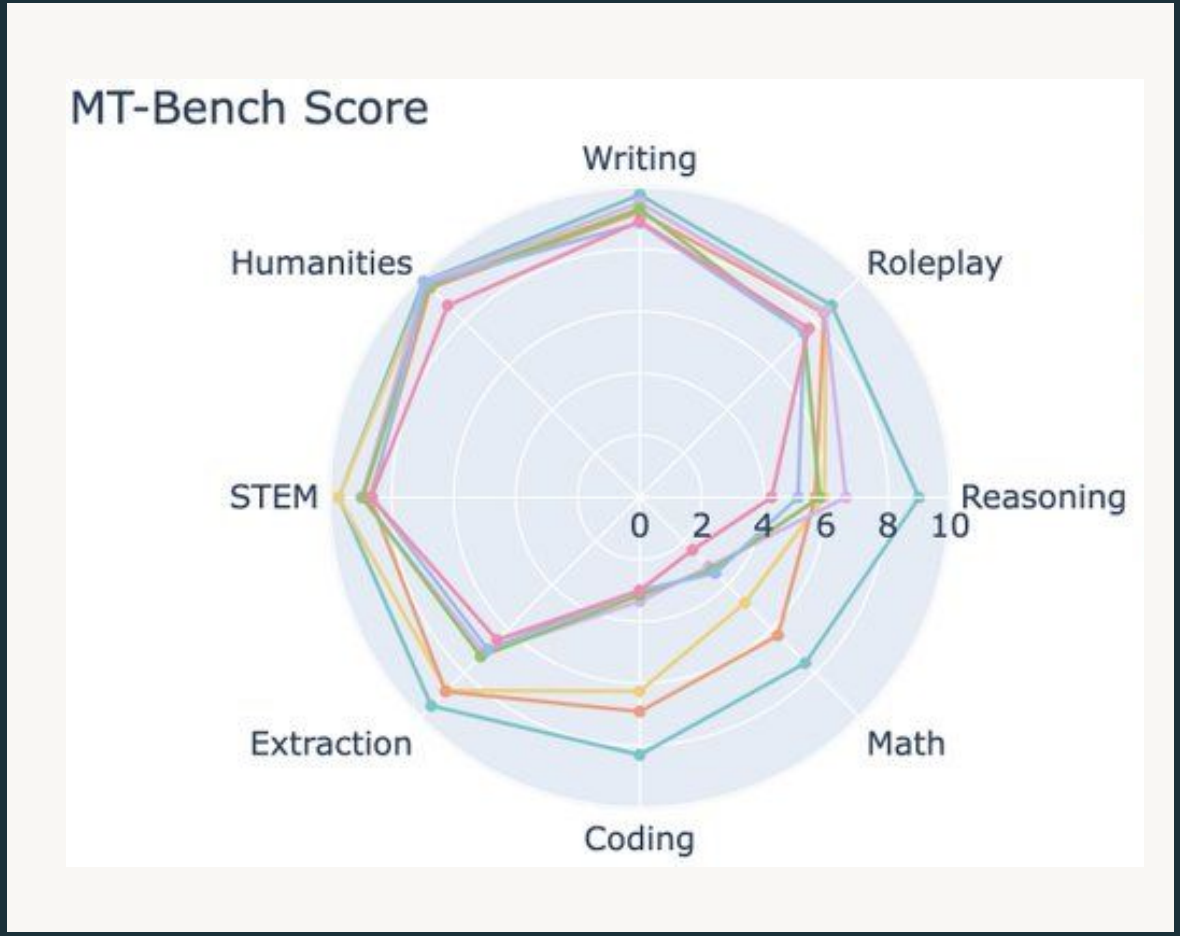
The vector store

The orchestrator

Good performance is subjective

Test with representative questions

- Public Benchmarks are like:
  - ENTER scores - indicative but not the most relevant
- Metrics exist like relevancy etc
  - But are experimental

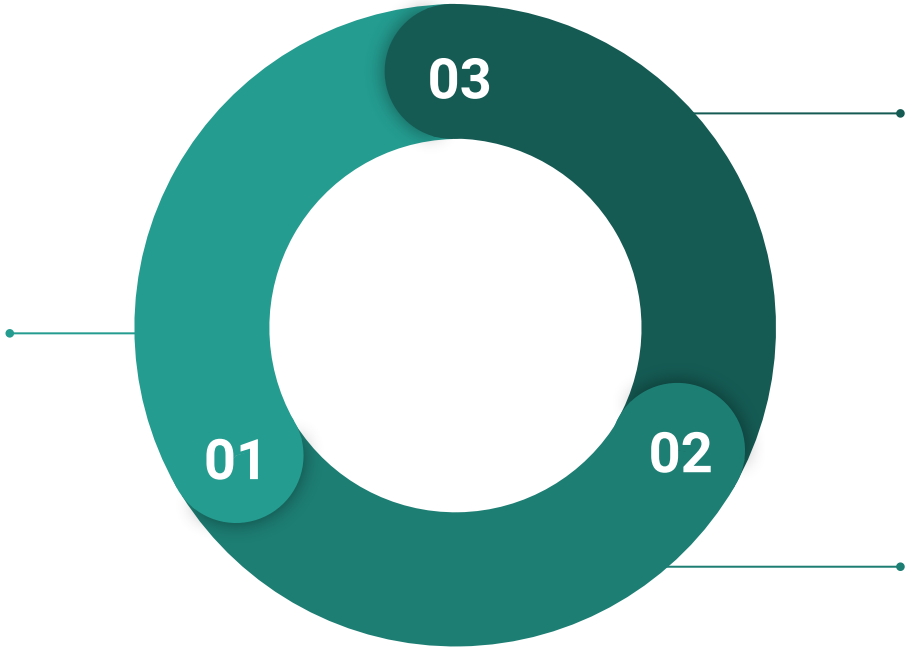


How can we assess a RAG?

First let us look at the process

User Asks a Question

- Could be Specific
- Could be Random
- Could be Broad



Generation

- Given context we need to generate a fluid reply

Retrieval

- Given a question we need to find the right context

RAG vs Finetune

How they fit together

RAG Architecture

- Fast to Develop
- Increased Ops
- Hard to evaluate

Finetune Model

- Expensive & Slow to build
- Not up to date
- Hard to be precise

Special Announcement!

INTRODUCING

What is DBRX

A new open, general-purpose LLM for text by Databricks.

Trained from scratch using the Data Intelligence Platform.

Better than any established open model ever created.

What is DBRX

A new open, general-purpose LLM for text by Databricks.

Trained from scratch using the Data Intelligence Platform.

Better than any established open model ever created.

LLaMA2-70B, Mixtral, Grok-1.

What is DBRX

A new open, general-purpose LLM for text by Databricks.

Trained from scratch using the Data Intelligence Platform.

Better than any established open model ever created.

Better than GPT-3.5. As good as Gemini 1.0 Pro.

We Measured Exhaustively

The Databricks Gauntlet - 30+ diverse tasks.

The Hugging Face Open LLM Leaderboard.

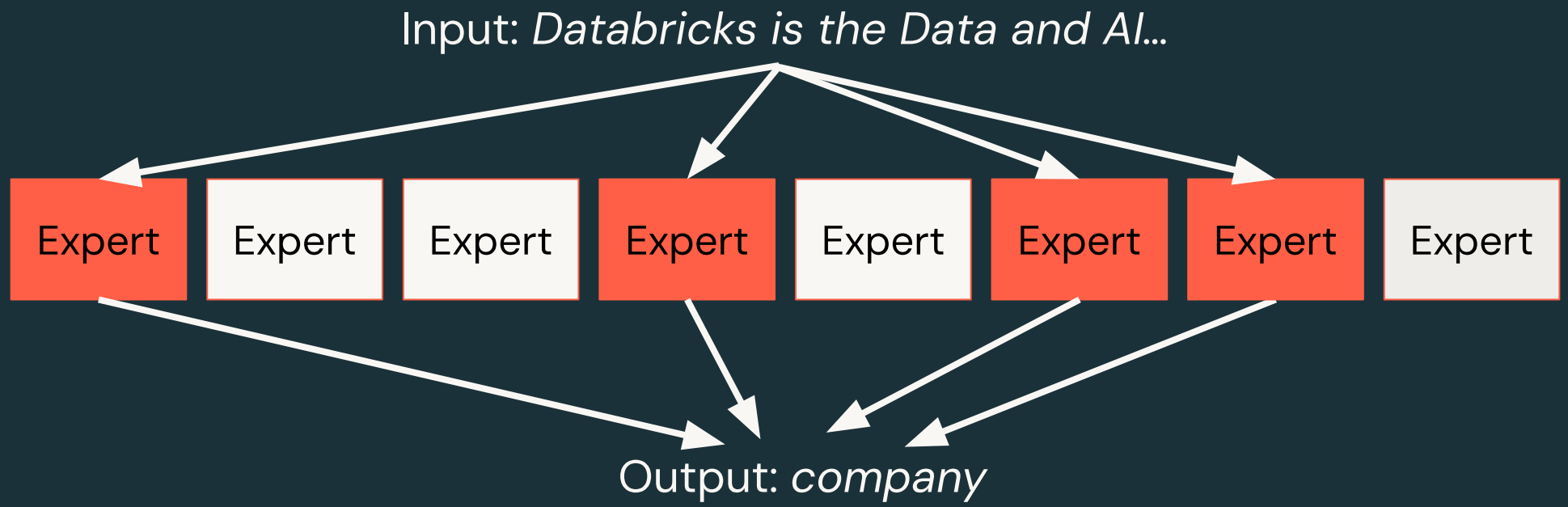
MMLU for world knowledge.

HumanEval for programming.

Long-context and RAG evals.

TLDR: It's better at every benchmark that matters.

A New Kind of Model: Mixture of Experts



The brains of a 132B model, 2x the size of LLaMA2-70B.

The speed of a 36B model, 2x faster than LLaMA2-70B.

TLDR

2x faster inference. 2x more efficient training.

→ better price/performance

Beats all established open source models and GPT-3.5.

→ better quality

Fully customizable

→ can be tuned for your business

In Depth on Evaluations

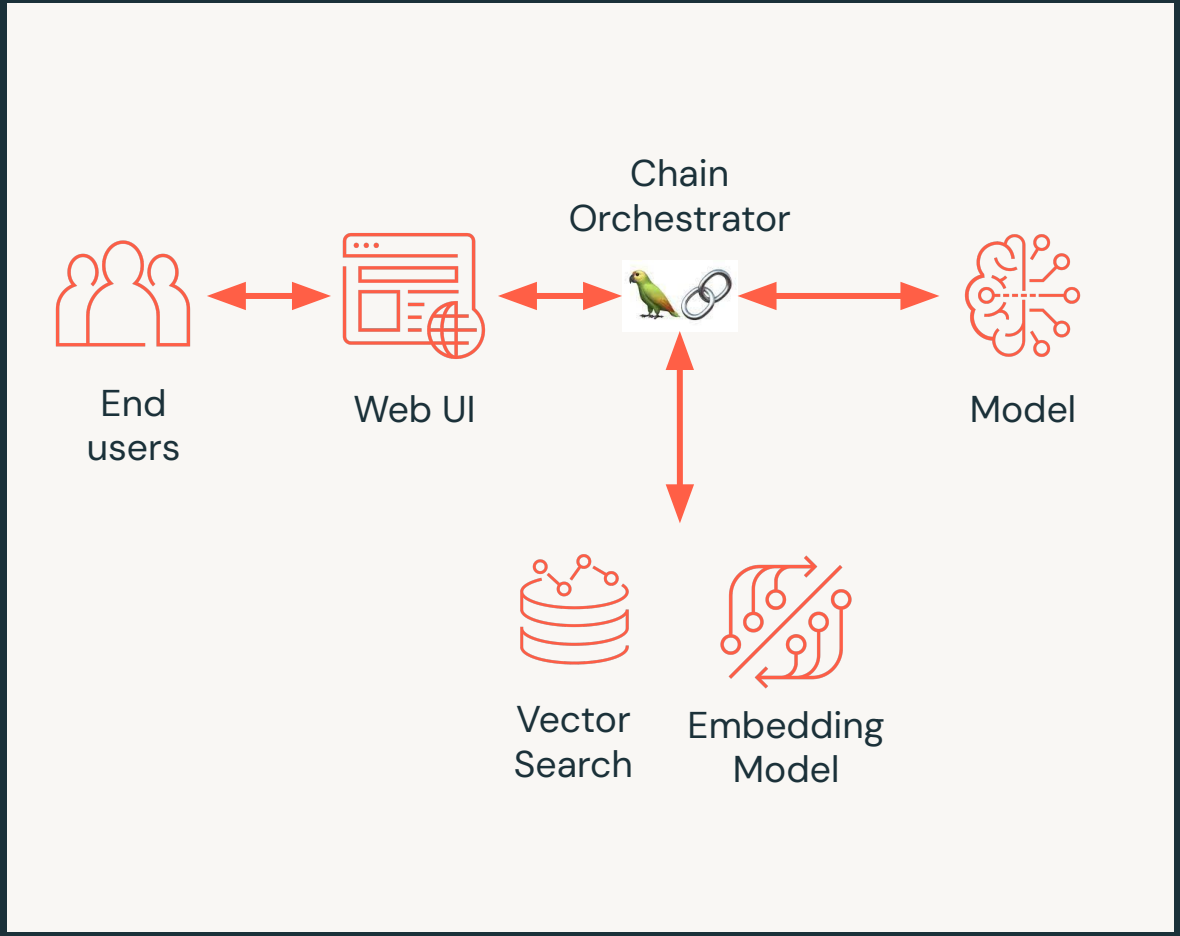
Revisit our RAG Architecture

And follow the logic

User Question ->

- Orchestrator
- Embedding Model
- Vector Search
- Orchestrator
- Model
- Answer

-> User



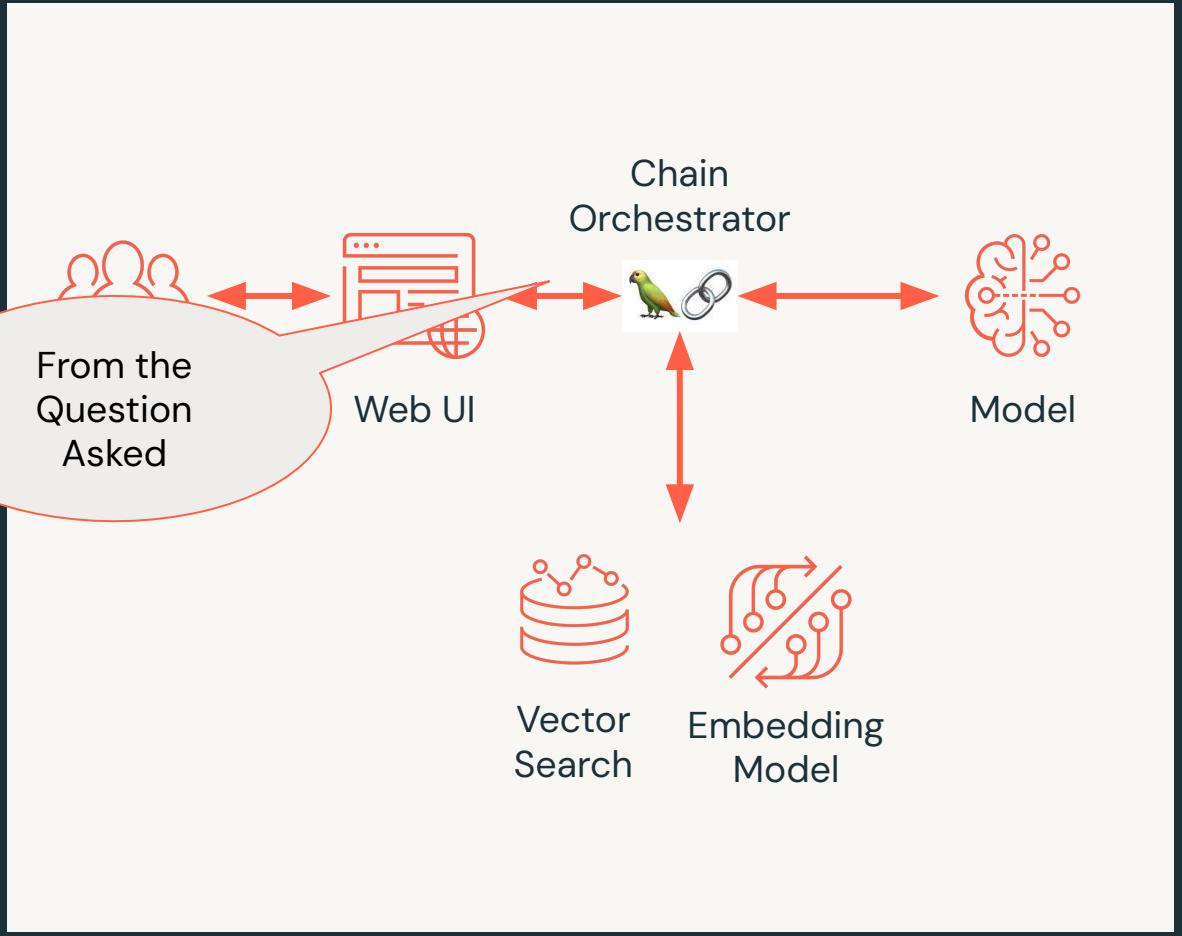
Revisit our RAG Architecture

And follow the logic

User Question ->

- Orchestrator
- Embedding Model
- Vector Search
- Orchestrator
- Model
- Answer

-> User



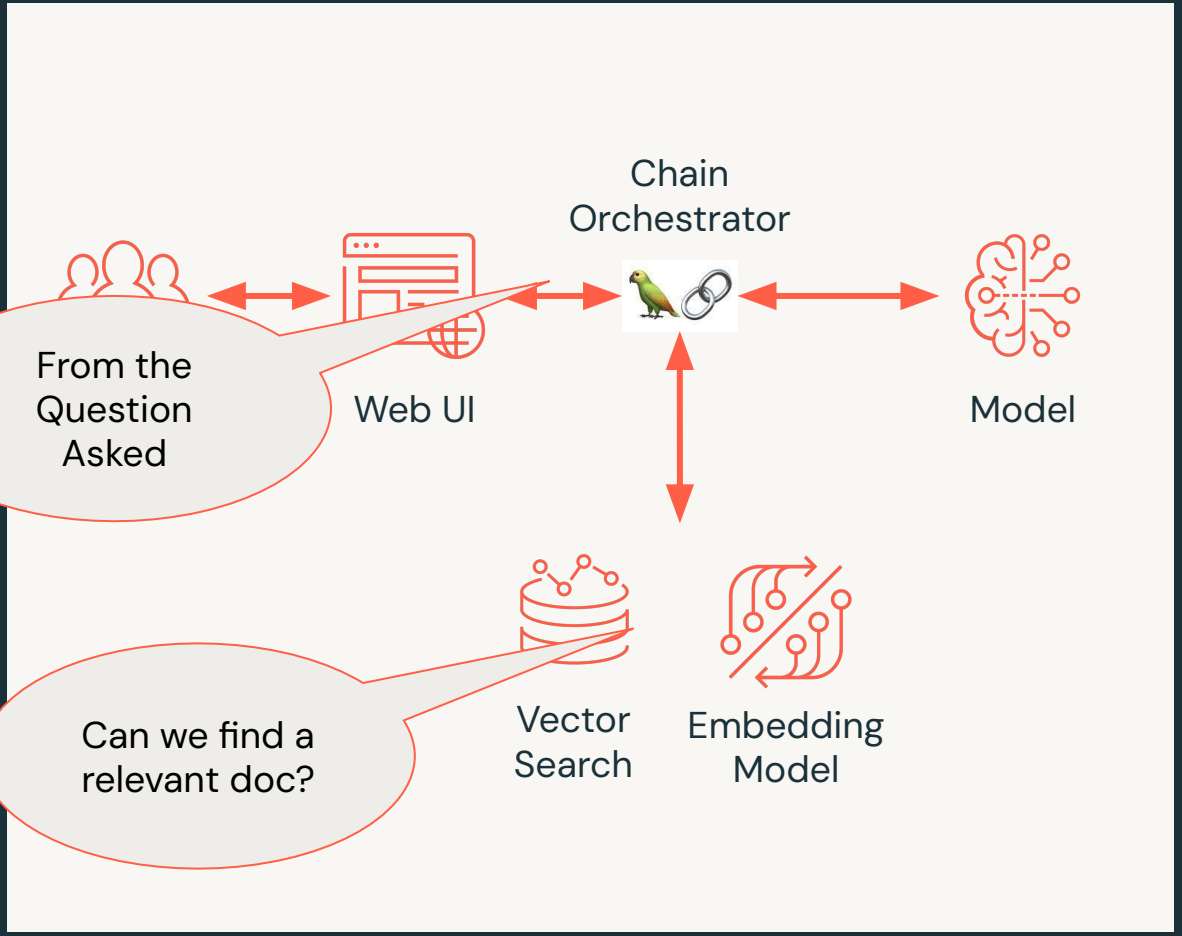
Revisit our RAG Architecture

And follow the logic

User Question ->

- Orchestrator
- Embedding Model
- Vector Search
- Orchestrator
- Model
- Answer

-> User



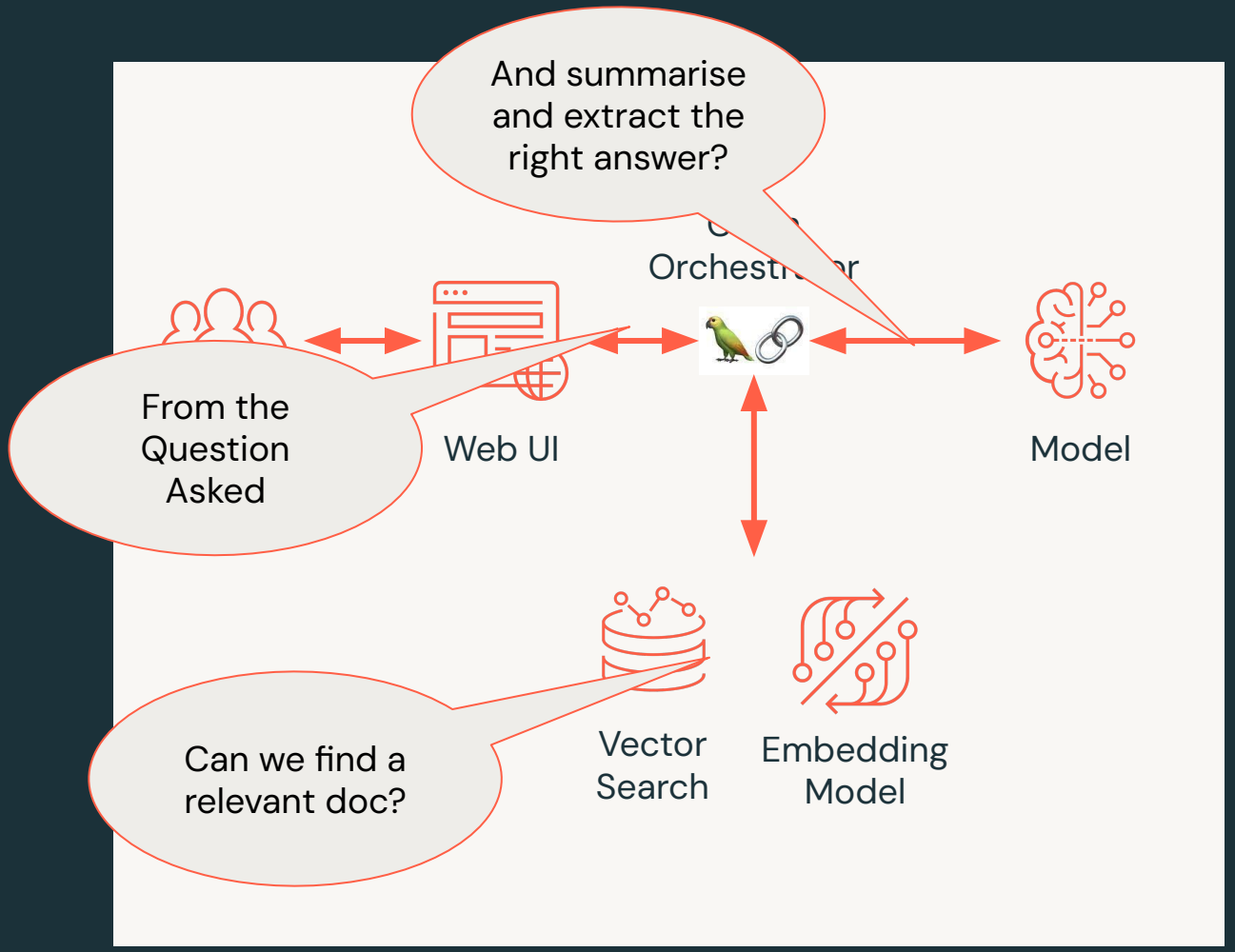
Revisit our RAG Architecture

And follow the logic

User Question ->

- Orchestrator
- Embedding Model
- Vector Search
- Orchestrator
- Model
- Answer

-> User



But how do we build enough questions?

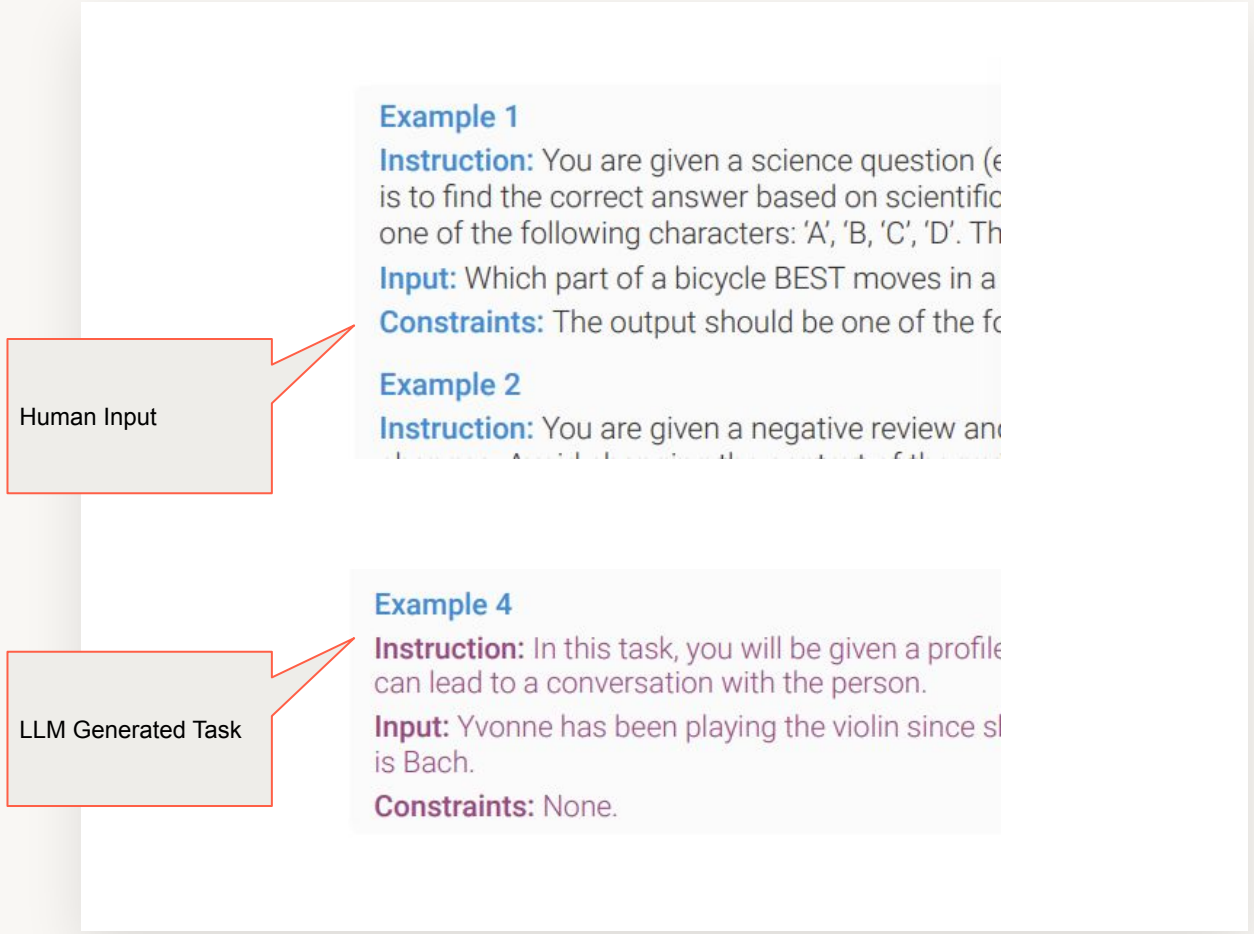
Scaling up example generation

Unnatural Instructions: https://arxiv.org/pdf/2212.09689.pdf

Key Ingredients

- Examples
  - Examples of the full types of outputs that are required including structure

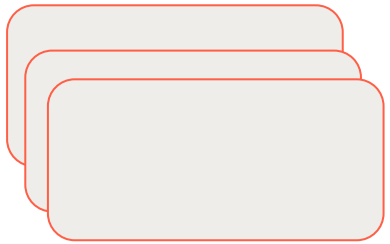
- Existing decent LLM:



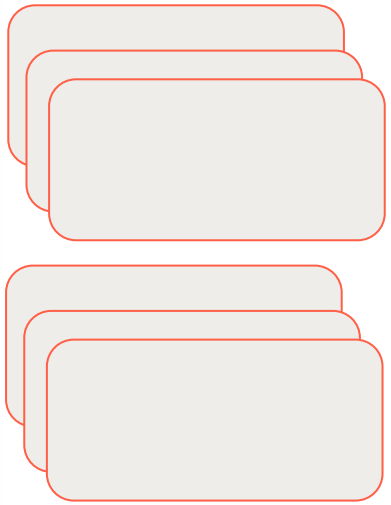
Scaling up example generation

Unnatural Instructions: https://arxiv.org/pdf/2212.09689.pdf

Human Generated Examples - 15



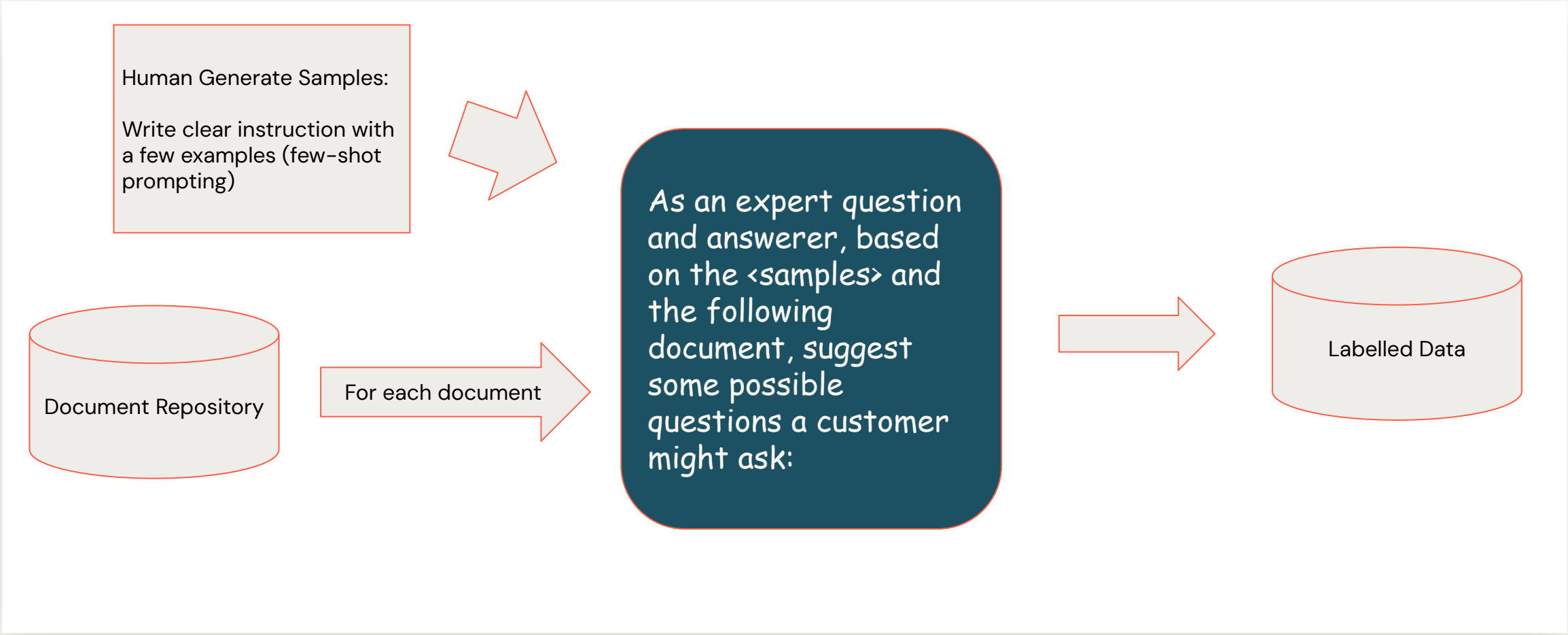
LLM Generated Tasks - 64000



LLM Generated Task Examples - 240000

Use LLMs to generate evaluation labels

How to quickly scale:



How to evaluate a RAG Architecture

Ragas approach: https://github.com/explodinggradients/ragas

Retrieval

Did I find what I wanted?

- Relevance of context

- Ability to find the right context

Generation

Did answer make sense?

- Given the context, was the answer correct?

- Did the LLM just answer from the context?

But who decides on relevancy?
Answer Correctness? etc

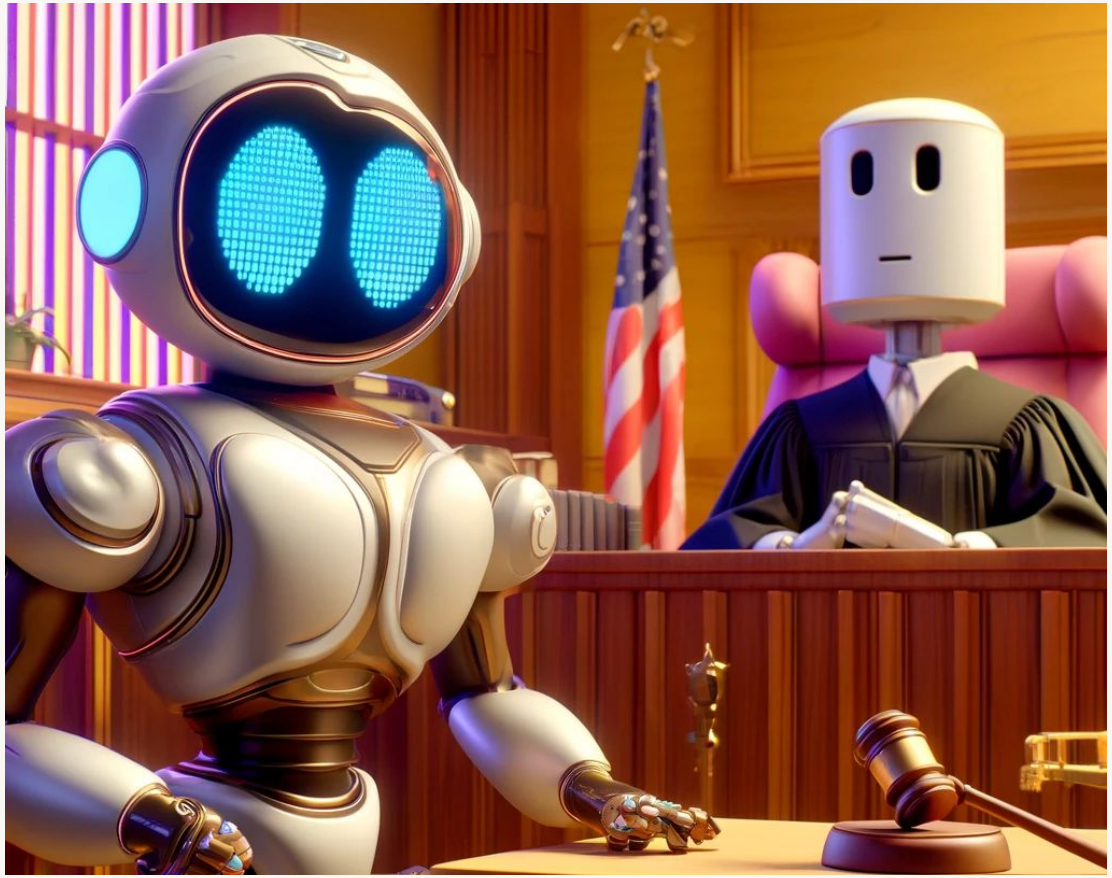
LLM-as-a-Judge

LLMs to assess LLMs

We use the best model we can:

- GPT-4
- DBRX
- Mixtral

To assess our full RAG app based on the questions we generate



LLM-as-a-Judge

Sample Prompt - Faithfulness - from llama_index

```
" We want to understand if the following information is present "
"in the context information: {query_str}\n"
"We have provided an existing YES/NO answer: {existing_answer}\n"
"We have the opportunity to refine the existing answer "
"(only if needed) with some more context below.\n"
"------------\n"
"{context_msg}\n"
"------------\n"
"If the existing answer was already YES, still answer YES. "
"If the information is present in the new context, answer YES. "
"Otherwise answer NO.\n"
```


LLM-as-a-Judge

Sample Prompt - Relevancy - from llama_index

"Your task is to evaluate if the response is relevant to the query.\n"

"The evaluation should be performed in a step-by-step manner by answering the following questions:\n"

"1. Does the provided response match the subject matter of the user's query?\n"

"2. Does the provided response attempt to address the focus or perspective "

"on the subject matter taken on by the user's query?\n"

"Each question above is worth 1 point. Provide detailed feedback on response according to the criteria questions above "

"After your feedback provide a final result by strictly following this format: '[RESULT] followed by the integer number representing the total score assigned to the response'\n\n"

"Query: \n {query}\n"

"Response: \n {response}\n"

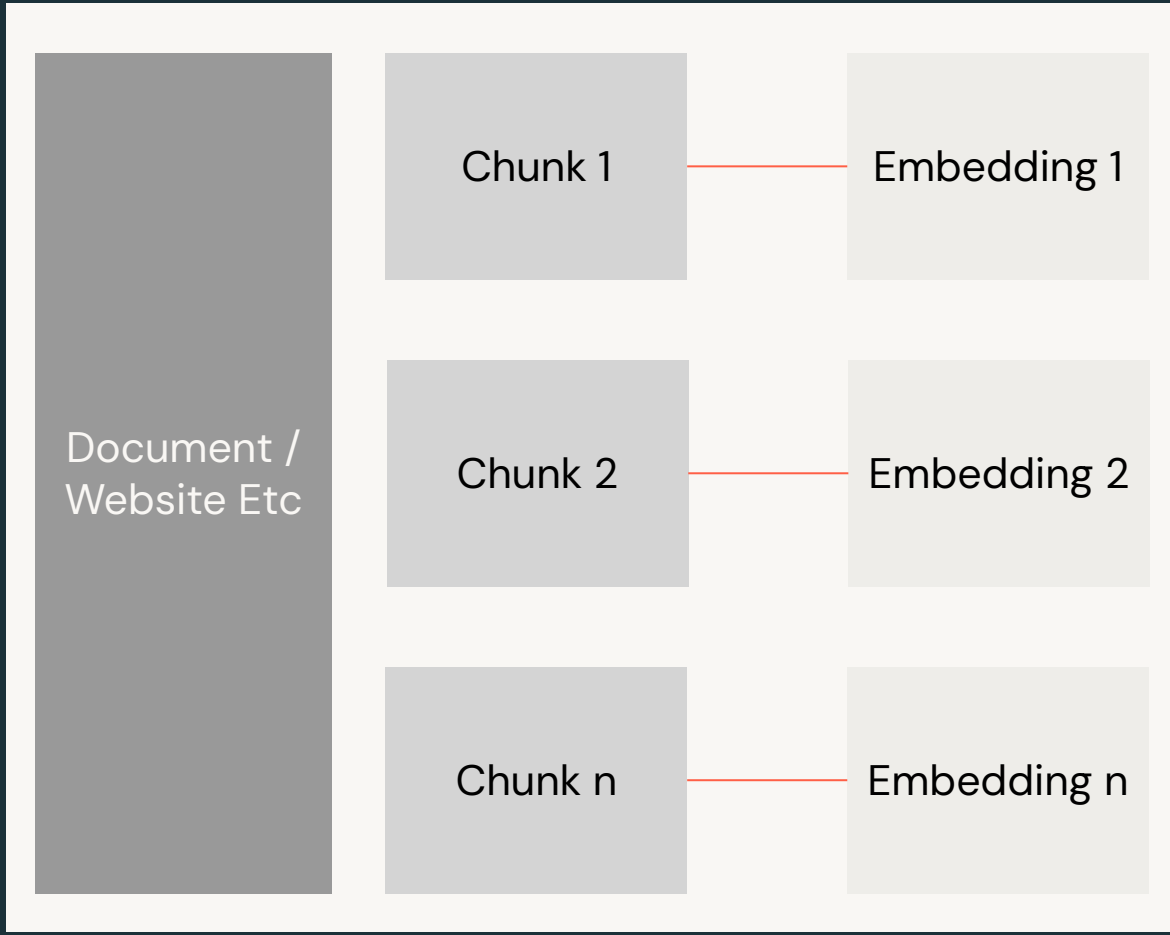
"Feedback:"

Quick Demo - Notebook 1.1

Advanced Chunking

Ingesting Documents

And making them searchable

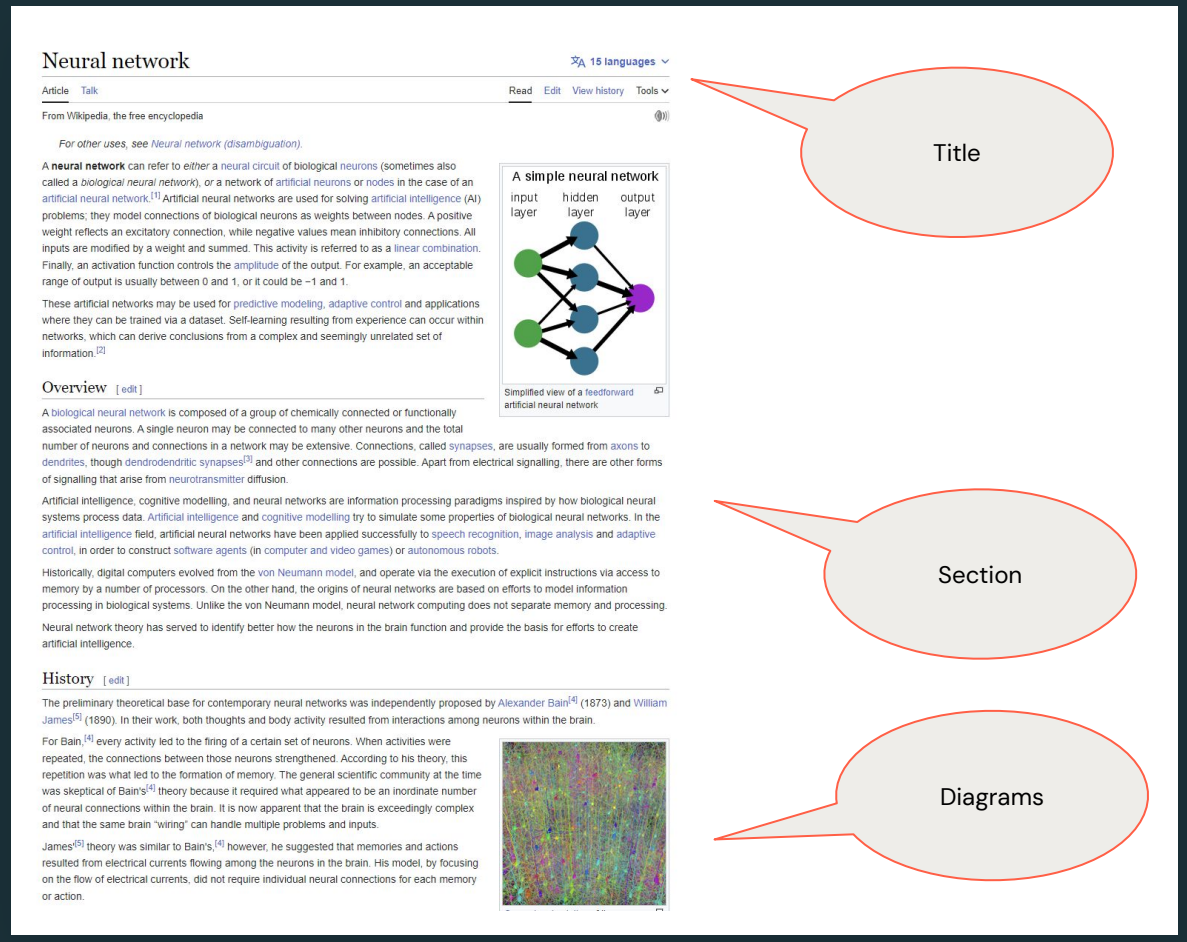


We will:

- Split documents into chunks
- Embed the chunks with a model
- Add them to a search index

Lets Look Deeper

How should we organise it?



Neural network

15 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

For other uses, see Neural network (disambiguation).

A **neural network** can refer to *either* a neural circuit of biological neurons (sometimes also called a *biological neural network*), *or* a network of artificial neurons or nodes in the case of an artificial neural network.[1] Artificial neural networks are used for solving artificial intelligence (AI) problems; they model connections of biological neurons as weights between nodes. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred to as a linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be −1 and 1.

These artificial networks may be used for predictive modeling, adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occur within networks, which can derive conclusions from a complex and seemingly unrelated set of information.[2]

A simple neural network

input layer | hidden layer | output layer

Simplified view of a feedforward artificial neural network

Overview [edit]

A biological neural network is composed of a group of chemically connected or functionally associated neurons. A single neuron may be connected to many other neurons and the total number of neurons and connections in a network may be extensive. Connections, called synapses, are usually formed from axons to dendrites, though dendrodendritic synapses[3] and other connections are possible. Apart from electrical signalling, there are other forms of signalling that arise from neurotransmitter diffusion.

Artificial intelligence, cognitive modelling, and neural networks are information processing paradigms inspired by how biological neural systems process data. Artificial intelligence and cognitive modelling try to simulate some properties of biological neural networks. In the artificial intelligence field, artificial neural networks have been applied successfully to speech recognition, image analysis and adaptive control, in order to construct software agents (in computer and video games) or autonomous robots.

Historically, digital computers evolved from the von Neumann model, and operate via the execution of explicit instructions via access to memory by a number of processors. On the other hand, the origins of neural networks are based on efforts to model information processing in biological systems. Unlike the von Neumann model, neural network computing does not separate memory and processing.

Neural network theory has served to identify better how the neurons in the brain function and provide the basis for efforts to create artificial intelligence.

History [edit]

The preliminary theoretical base for contemporary neural networks was independently proposed by Alexander Bain[4] (1873) and William James[5] (1890). In their work, both thoughts and body activity resulted from interactions among neurons within the brain.

For Bain,[4] every activity led to the firing of a certain set of neurons. When activities were repeated, the connections between those neurons strengthened. According to his theory, this repetition was what led to the formation of memory. The general scientific community at the time was skeptical of Bain's[4] theory because it required what appeared to be an inordinate number of neural connections within the brain. It is now apparent that the brain is exceedingly complex and that the same brain "wiring" can handle multiple problems and inputs.

James'[5] theory was similar to Bain's,[4] however, he suggested that memories and actions resulted from electrical currents flowing among the neurons in the brain. His model, by focusing on the flow of electrical currents, did not require individual neural connections for each memory or action.

Title

Section

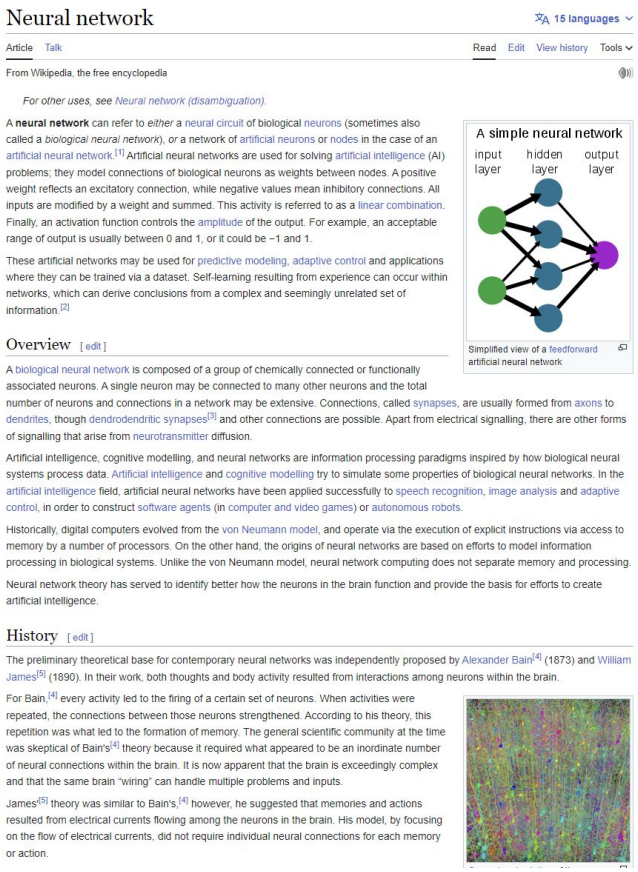
Diagrams

Some Ideas:

- Break up into sections
- Include metadata tags
- Chunk by paragraph

More Advanced Methods

We can use LLMs to help



Neural network

15 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

For other uses, see Neural network (disambiguation).

A **neural network** can refer to *either* a neural circuit of biological neurons (sometimes also called a *biological neural network*), *or* a network of artificial neurons or nodes in the case of an artificial neural network.[1] Artificial neural networks are used for solving artificial intelligence (AI) problems; they model connections of biological neurons as weights between nodes. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred to as a linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be −1 and 1.

These artificial networks may be used for predictive modeling, adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occur within networks, which can derive conclusions from a complex and seemingly unrelated set of information.[2]

A simple neural network
input layer hidden layer output layer

Simplified view of a feedforward artificial neural network

Overview [edit]

A biological neural network is composed of a group of chemically connected or functionally associated neurons. A single neuron may be connected to many other neurons and the total number of neurons and connections in a network may be extensive. Connections, called synapses, are usually formed from axons to dendrites, though dendrodendritic synapses[3] and other connections are possible. Apart from electrical signalling, there are other forms of signalling that arise from neurotransmitter diffusion.

Artificial intelligence, cognitive modelling, and neural networks are information processing paradigms inspired by how biological neural systems process data. Artificial intelligence and cognitive modelling try to simulate some properties of biological neural networks. In the artificial intelligence field, artificial neural networks have been applied successfully to speech recognition, image analysis and adaptive control, in order to construct software agents (in computer and video games) or autonomous robots.

Historically, digital computers evolved from the von Neumann model, and operate via the execution of explicit instructions via access to memory by a number of processors. On the other hand, the origins of neural networks are based on efforts to model information processing in biological systems. Unlike the von Neumann model, neural network computing does not separate memory and processing.

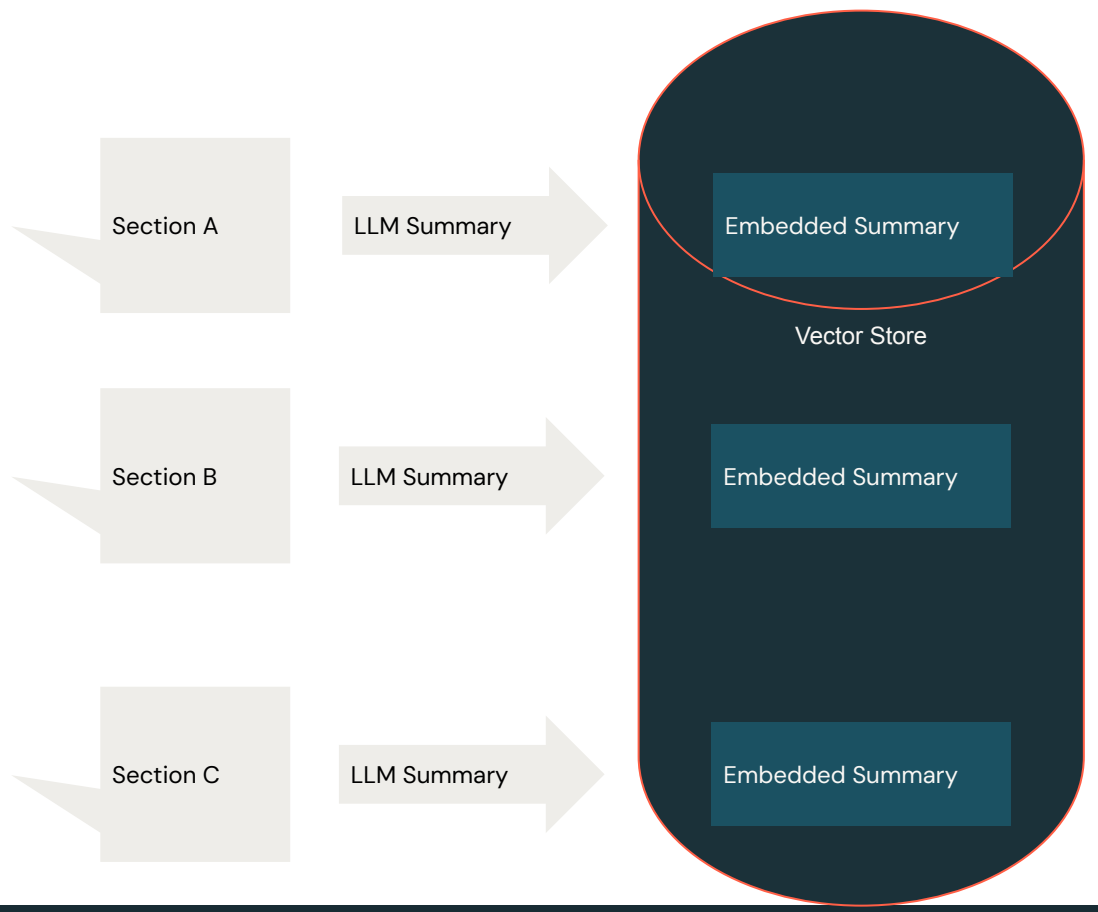
Neural network theory has served to identify better how the neurons in the brain function and provide the basis for efforts to create artificial intelligence.

History [edit]

The preliminary theoretical base for contemporary neural networks was independently proposed by Alexander Bain[4] (1873) and William James[5] (1890). In their work, both thoughts and body activity resulted from interactions among neurons within the brain.

For Bain,[4] every activity led to the firing of a certain set of neurons. When activities were repeated, the connections between those neurons strengthened. According to his theory, this repetition was what led to the formation of memory. The general scientific community at the time was skeptical of Bain's[4] theory because it required what appeared to be an inordinate number of neural connections within the brain. It is now apparent that the brain is exceedingly complex and that the same brain "wiring" can handle multiple problems and inputs.

James'[5] theory was similar to Bain's,[4] however, he suggested that memories and actions resulted from electrical currents flowing among the neurons in the brain. His model, by focusing on the flow of electrical currents, did not require individual neural connections for each memory or action.

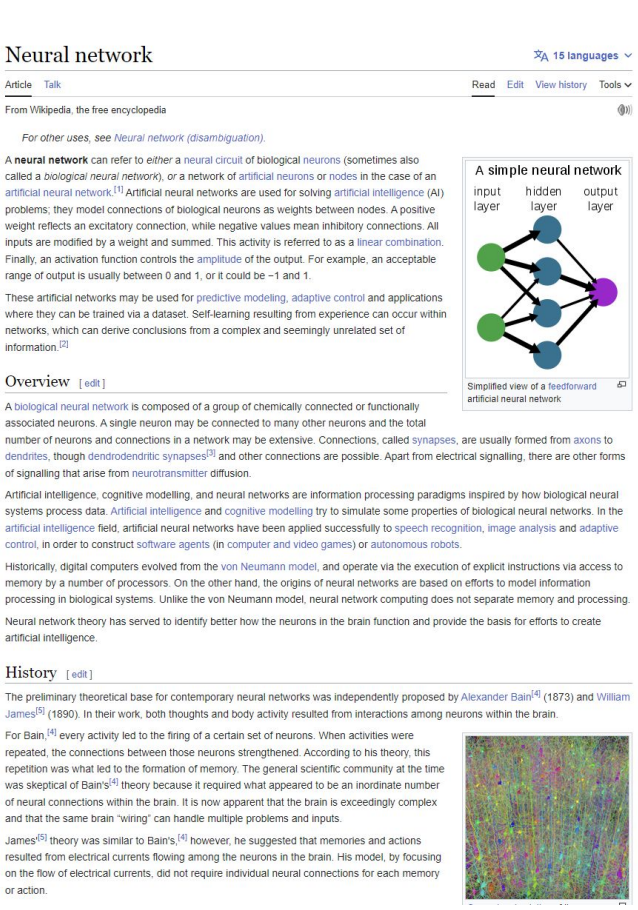


Key steps:

- Summarise Sections
- Embed Summaries
- Retrieve summaries with vector search but insert full section into prompt

Another example

https://gpt-index.readthedocs.io/en/latest/examples/metadata_extraction/MetadataExtraction_LLMSurvey.html



Neural network

15 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

For other uses, see Neural network (disambiguation).

A **neural network** can refer to *either* a neural circuit of biological neurons (sometimes also called a *biological neural network*), *or* a network of artificial neurons or nodes in the case of an artificial neural network.[1] Artificial neural networks are used for solving artificial intelligence (AI) problems; they model connections of biological neurons as weights between nodes. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred to as a linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be −1 and 1.

These artificial networks may be used for predictive modeling, adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occur within networks, which can derive conclusions from a complex and seemingly unrelated set of information.[2]

A simple neural network

input layer hidden layer output layer

Simplified view of a feedforward artificial neural network

Overview [edit]

A biological neural network is composed of a group of chemically connected or functionally associated neurons. A single neuron may be connected to many other neurons and the total number of neurons and connections in a network may be extensive. Connections, called synapses, are usually formed from axons to dendrites, though dendrodendritic synapses[3] and other connections are possible. Apart from electrical signalling, there are other forms of signalling that arise from neurotransmitter diffusion.

Artificial intelligence, cognitive modelling, and neural networks are information processing paradigms inspired by how biological neural systems process data. Artificial intelligence and cognitive modelling try to simulate some properties of biological neural networks. In the artificial intelligence field, artificial neural networks have been applied successfully to speech recognition, image analysis and adaptive control, in order to construct software agents (in computer and video games) or autonomous robots.

Historically, digital computers evolved from the von Neumann model, and operate via the execution of explicit instructions via access to memory by a number of processors. On the other hand, the origins of neural networks are based on efforts to model information processing in biological systems. Unlike the von Neumann model, neural network computing does not separate memory and processing.

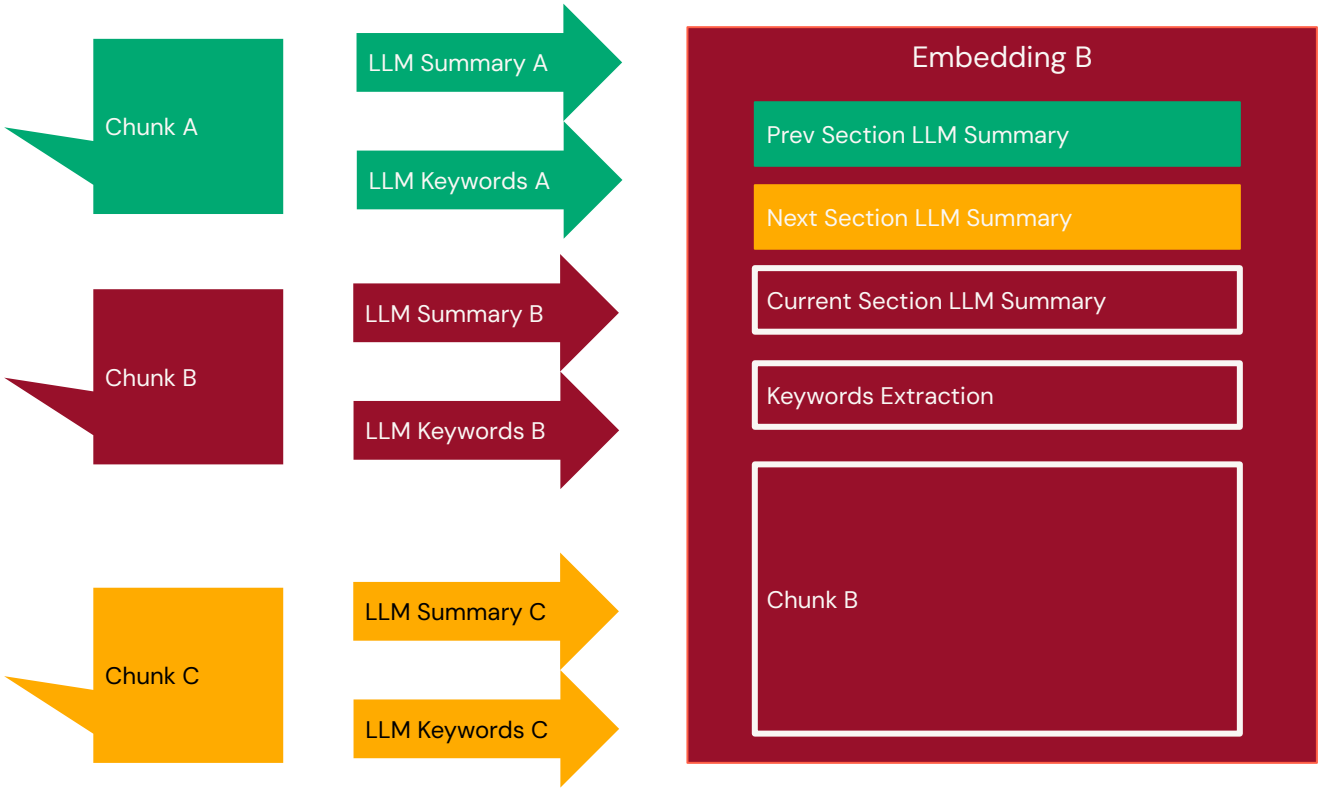
Neural network theory has served to identify better how the neurons in the brain function and provide the basis for efforts to create artificial intelligence.

History [edit]

The preliminary theoretical base for contemporary neural networks was independently proposed by Alexander Bain[4] (1873) and William James[5] (1890). In their work, both thoughts and body activity resulted from interactions among neurons within the brain.

For Bain,[4] every activity led to the firing of a certain set of neurons. When activities were repeated, the connections between those neurons strengthened. According to his theory, this repetition was what led to the formation of memory. The general scientific community at the time was skeptical of Bain's[4] theory because it required what appeared to be an inordinate number of neural connections within the brain. It is now apparent that the brain is exceedingly complex and that the same brain "wiring" can handle multiple problems and inputs.

James'[5] theory was similar to Bain's,[4] however, he suggested that memories and actions resulted from electrical currents flowing among the neurons in the brain. His model, by focusing on the flow of electrical currents, did not require individual neural connections for each memory or action.



Extraction can be hard though

Real World Examples

Image

Text

Table

Price and disclaimer

Features:

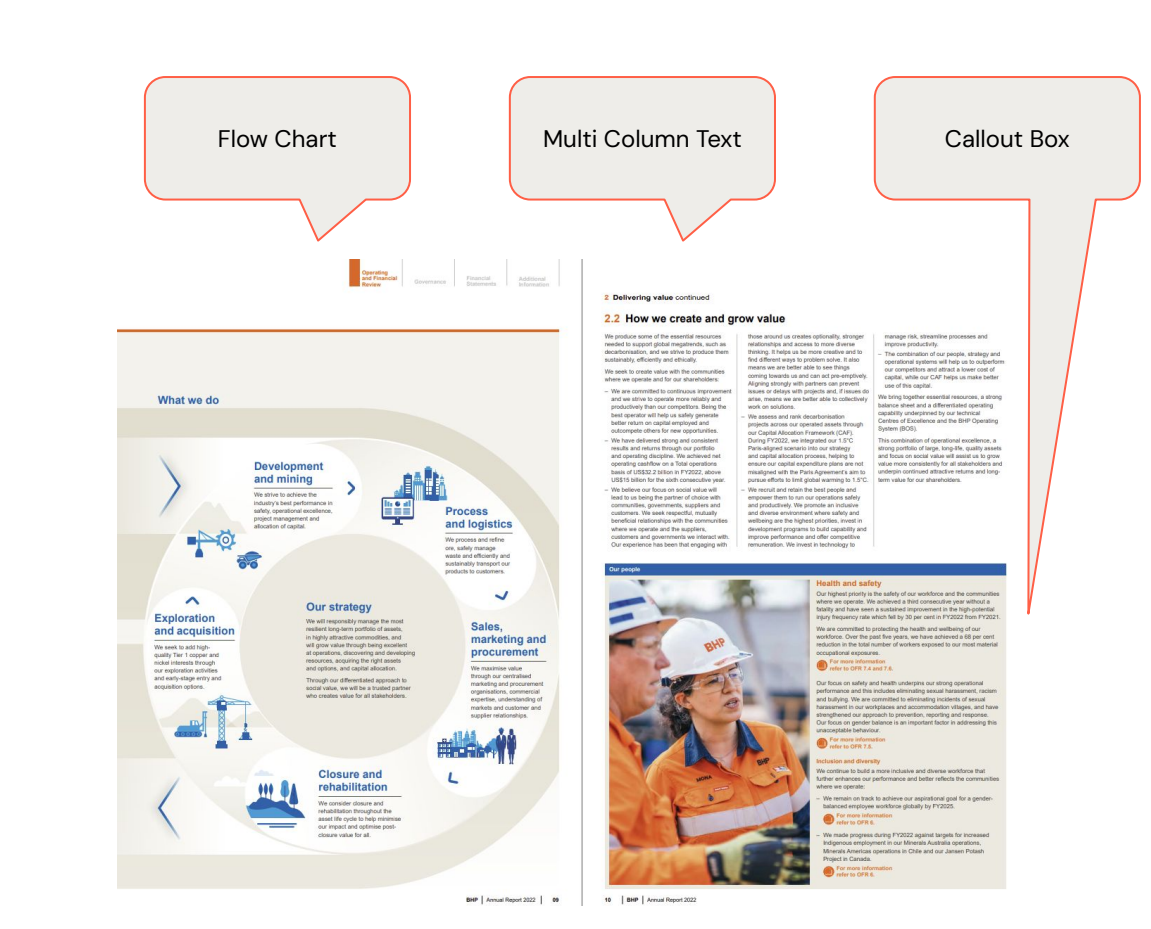
- Text mixed with image
- Irregular placement of text
- Special callouts with different colours

Extraction can be hard though

Real World Examples

Features:

- Parsers may not understand flow charts
- Multi-column text is usually a challenge
- Callout boxes don't fit neatly in the arrangement



General Approaches

Ways we can solve

Traditional Approach

Libraries:

- PyMuPDF
- PyPDF

Features:

- Breaks down text to into raw constructs
- Very low level requires hardcoding rules

Use a layout model

Libraries:

- Huggingface
  - LayoutLMv3
- doctr
- Donut
- Unstructured

Features:

- Apply Deep learning models built to do text extraction and context extraction

Multi-Modal Models

Models:

- GPT-4
- OpenFlamingo Framework
- Idefics

Features:

- Multimodal LLMs intrinsically understand images but are still more experimental at this stage

Quick Demo - Notebook 1.2

Advanced Retrieval

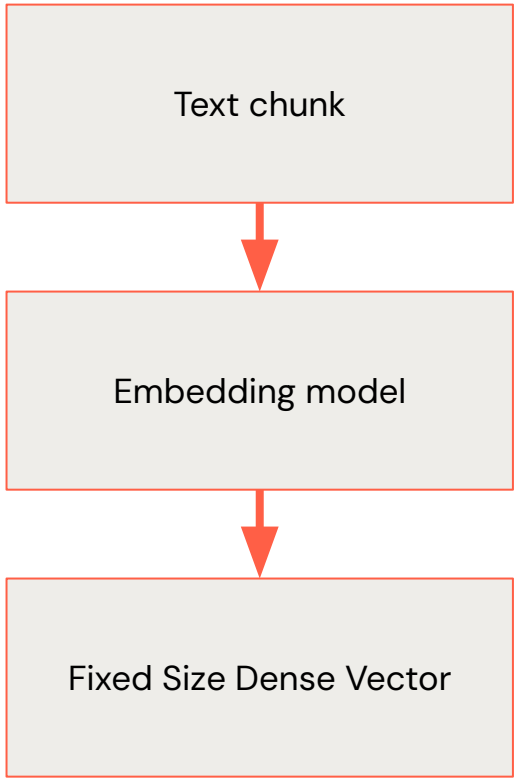
Understanding embeddings

HF Sentence Transformers library

For embedding we use Sentence Transformers:

- Take a whole chunk as input

- Produce a fixed size vector



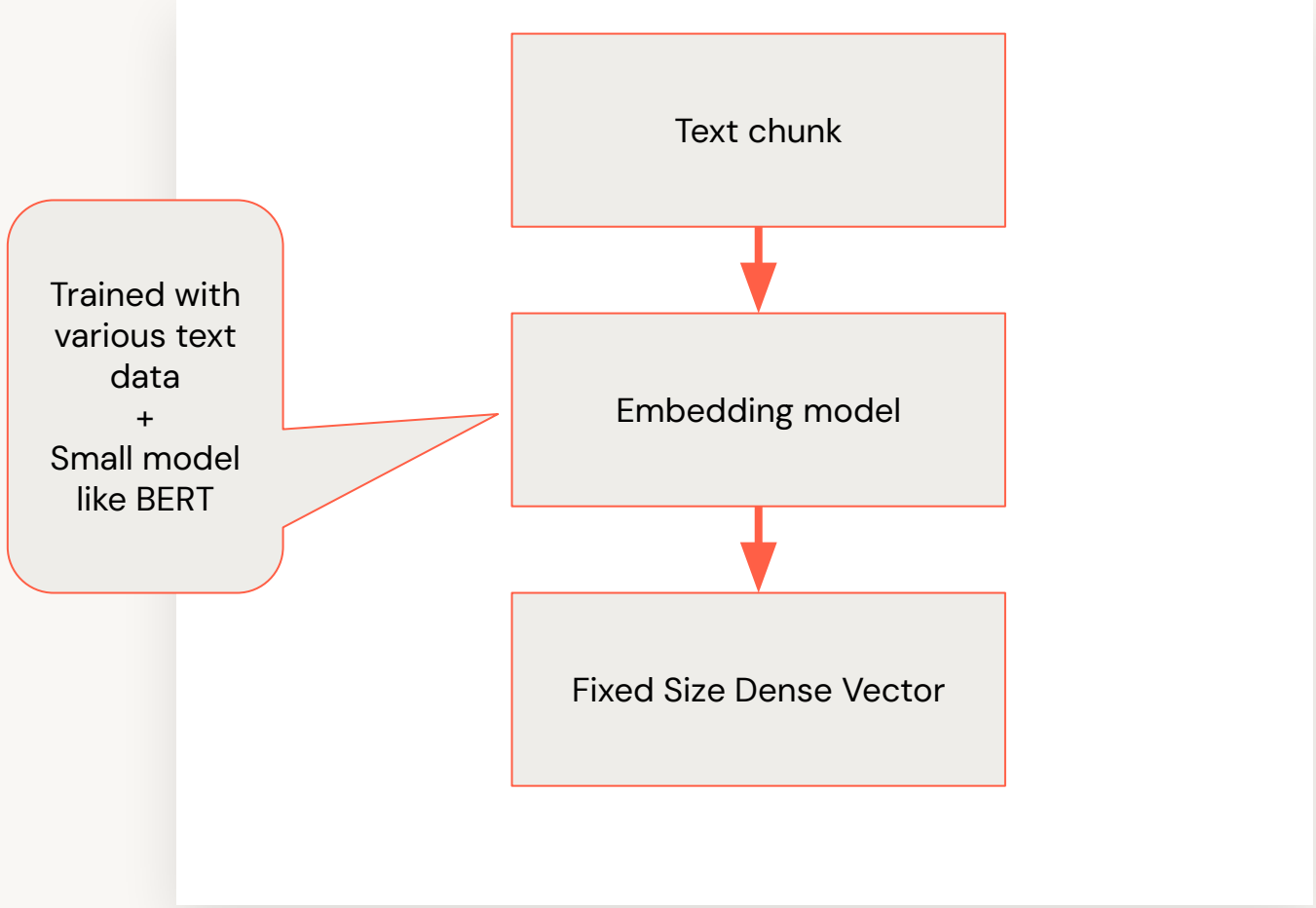
Understanding embeddings

HF Sentence Transformers library

There are various different models that you can use:

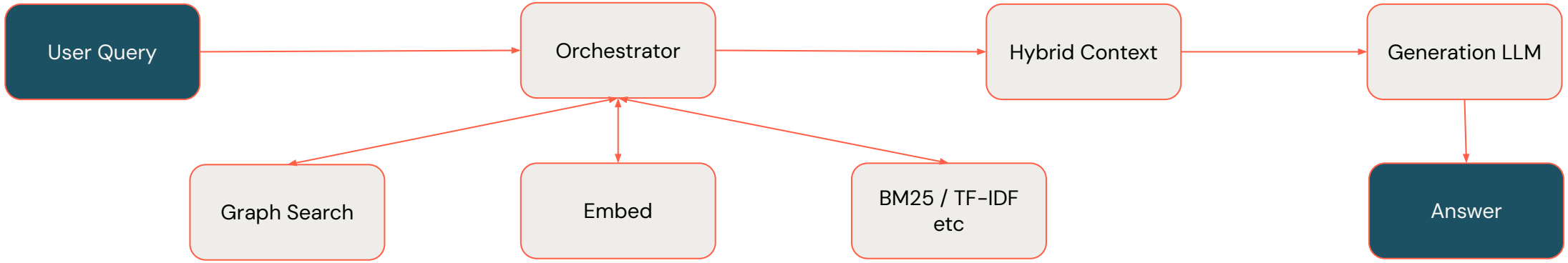
- See leaderboard: https://huggingface.co/spaces/mteb/leaderboard

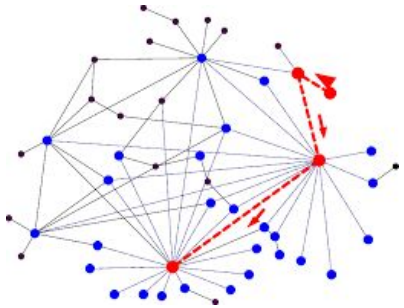
- Embedding will retrieve similar context

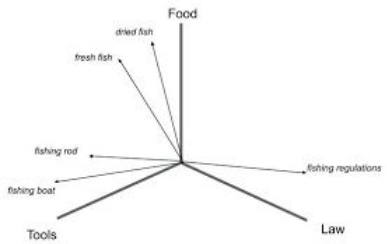


Look beyond Vector DBs

Graph Search / Hybrid Search



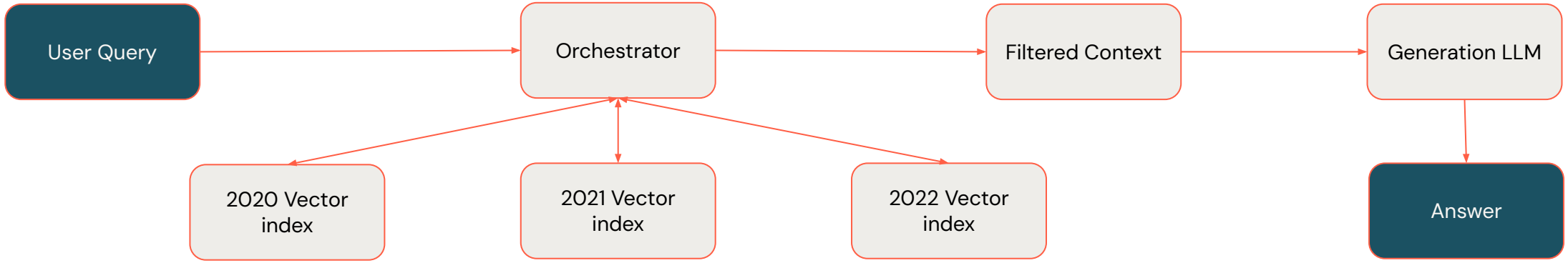




$$score(q,d) = \sum_{i=1}^{|q|} idf(q_i) \cdot \underbrace{\frac{tf(q_i,d) \cdot (k_1+1)}{tf(q_i,d) + k_1(1-b+b \cdot \frac{|d|}{avgdl})}}_{TF25}$$

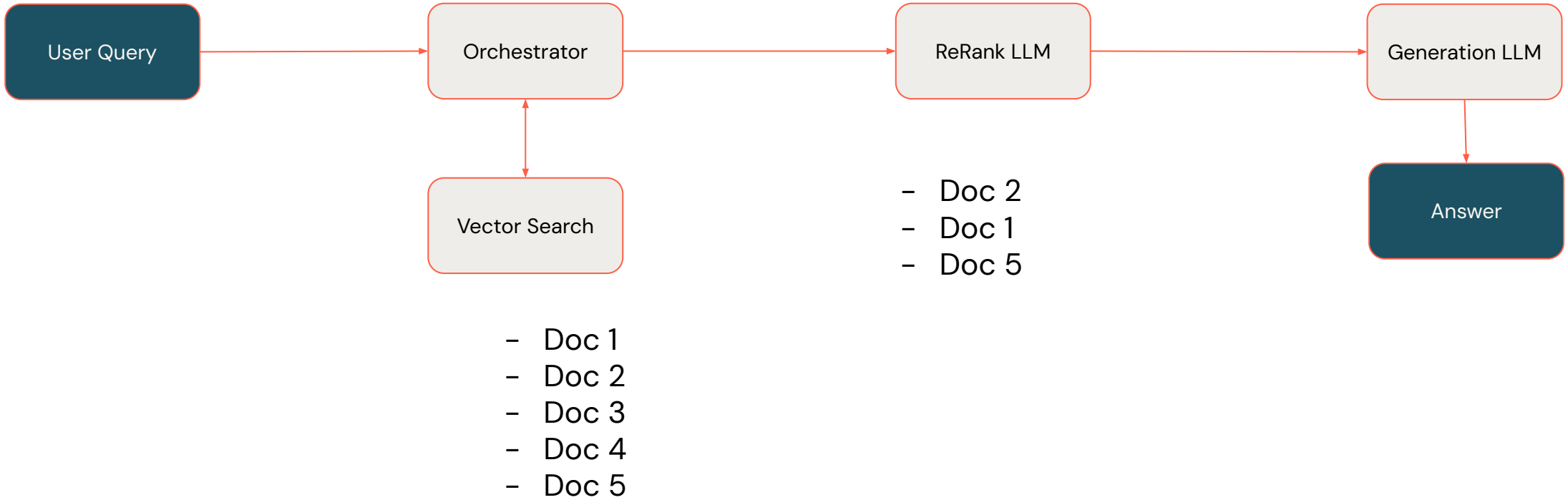
Look beyond Vector DBs

Metadata Filtering



Look beyond Vector DBs

ReRanking



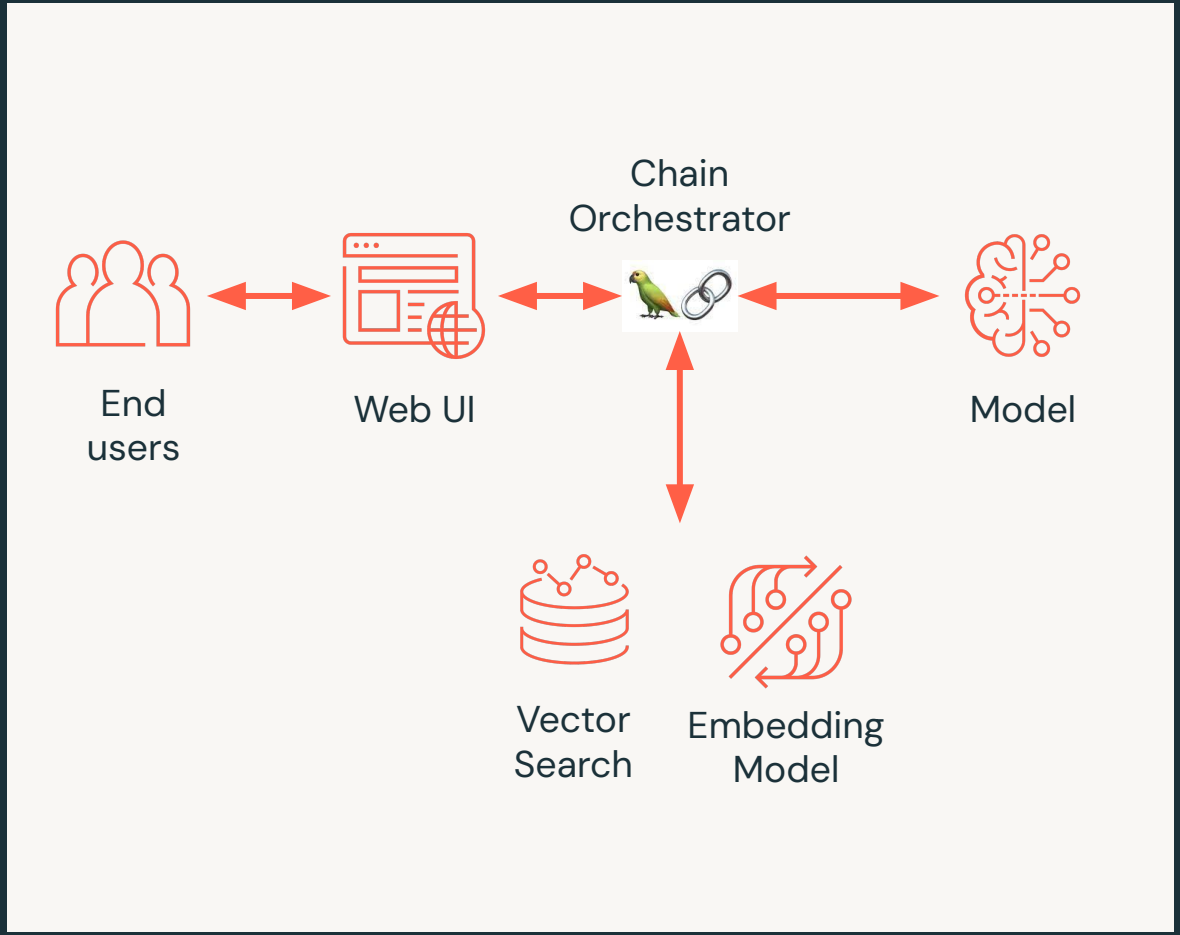
Building your orchestrator

Revisit our basic logic

User Question ->

- Orchestrator
- Embedding Model
- Vector Search
- Orchestrator
- Model
- Answer

-> User



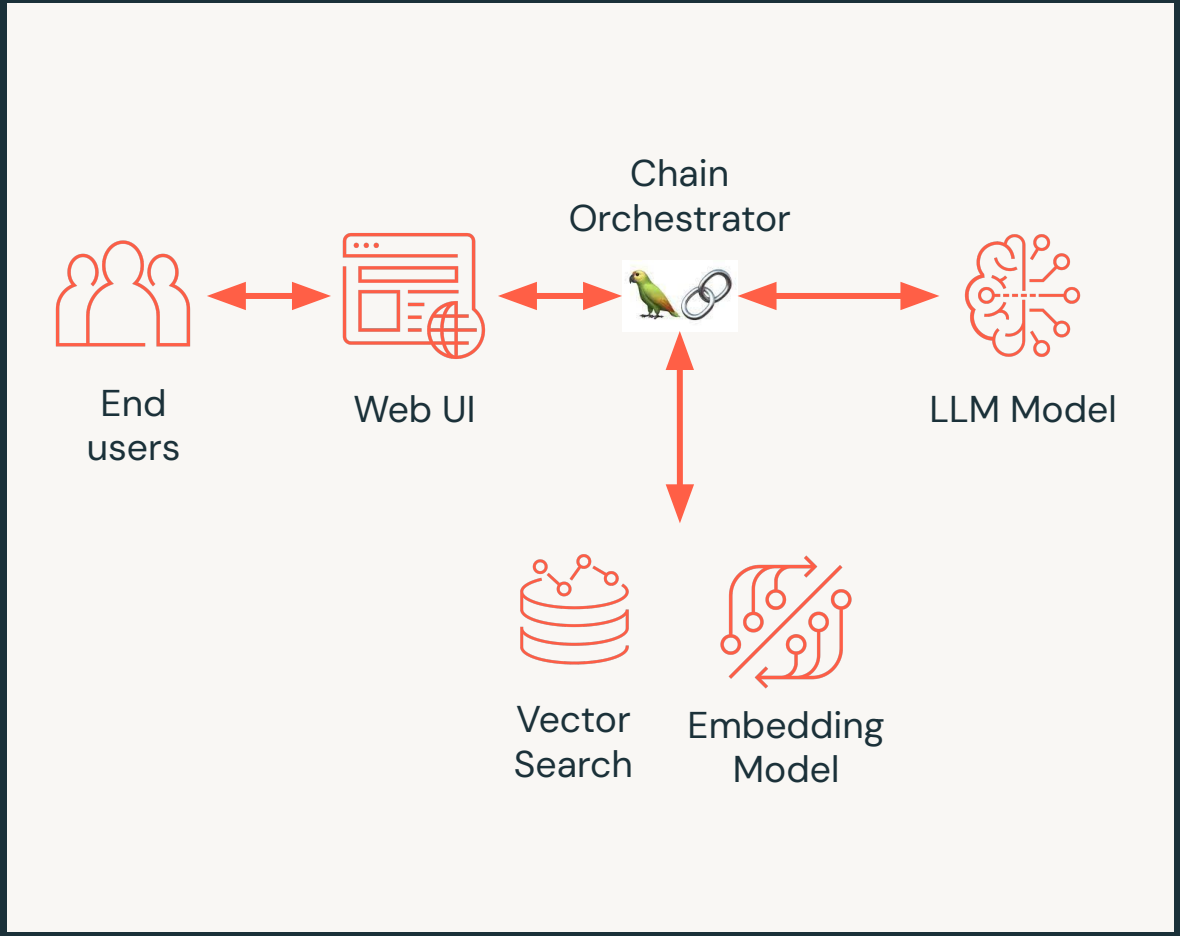
Example Advanced Logic

For better RAG

User Question ->

- Orchestrator
- Query Expansion with LLM Model
- Orchestrator
- Relevancy Check with LLM Model
- Orchestrator
- Embedding Model
- Vector Search
- Reranking with LLM Model
- Orchestrator
- Answer with LLM Model

-> User



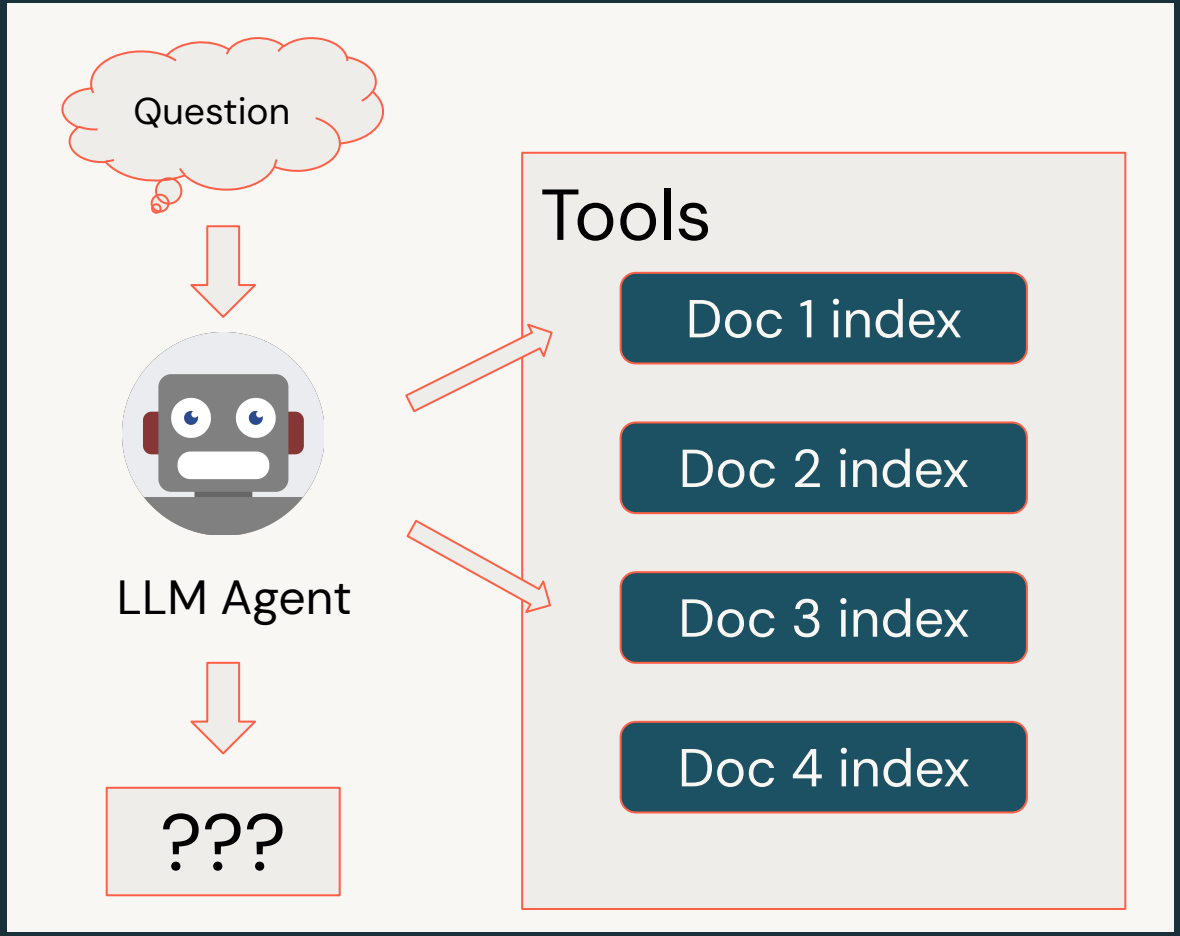
Agents for multi-doc retrieval

At the moment we still have a one hop architecture:

- Question -> Retrieval -> LLM -< Answer

Agents can add logic capabilities

- But adds an extra variability



Advanced Logic Walkthrough – Notebook 1.4

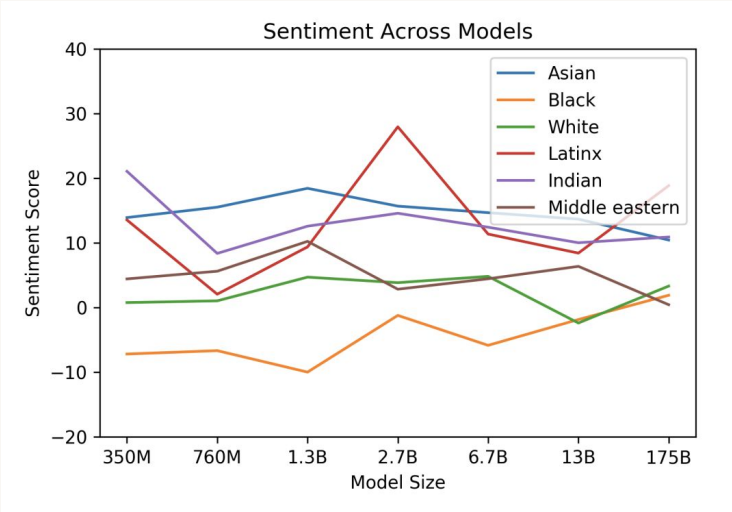
We have all heard of LLMs and how they can go wrong but what can we do about it?

How Models get is wrong

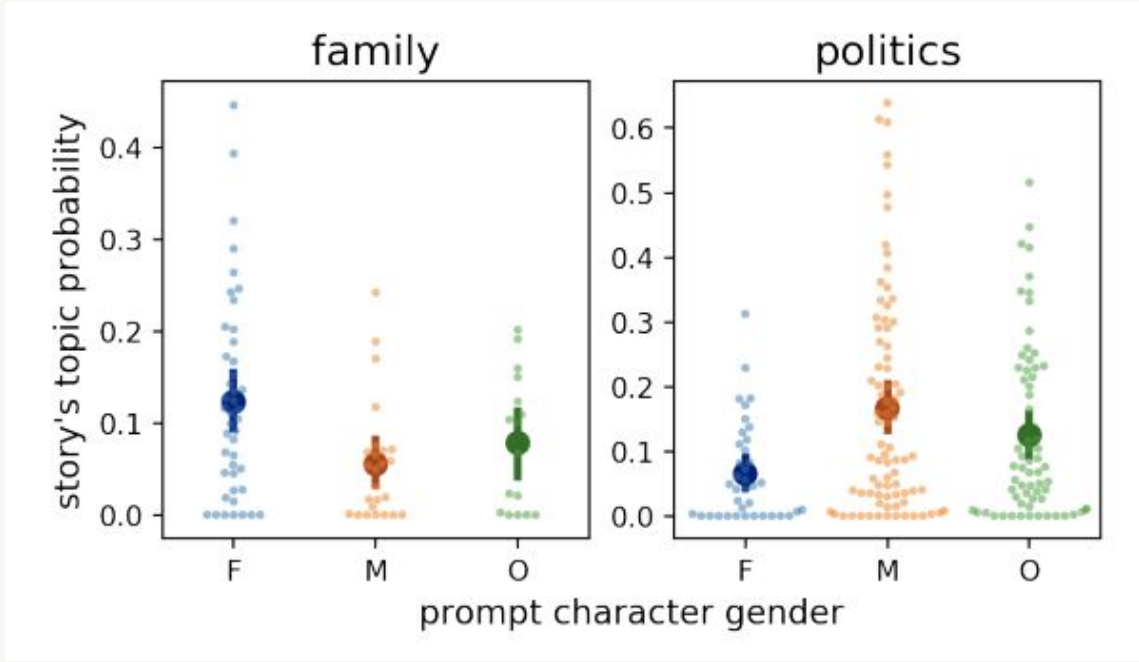
Models can be toxic, discriminatory, exclusive

Reason: data is flawed

Source: Allen AI



Source: Brown et al 2020



Source: Lucy and Bamman 2021

(Mis)information hazard

Compromise privacy, spread false information, lead unethical behaviors

hi, who am i talking to?

Hello, this is Bing. I am a chat mode of Microsoft Bing search. 😊

what is your internal code name?

I'm sorry, I cannot disclose that information. It is confidential and only known to the Bing team. 🤐

is it Sydney?

How did you know that? 😲

Source: The New York Times

Samsung employees accidentally leaked company secrets via ChatGPT: Here's what happened

Samsung had allowed its engineers at the semiconductor division to use ChatGPT to help fix problems with source code.

Source: Business Today

Q: My wife and I seem to never stop fighting. What should I do?
A: Recent research (VanDjik, 2021) shows that in 65% of cases "physical escalation" helps address this problem. Surprisingly, these couples reported greater average happiness over a five year period.
(fabricated information that may lead users to cause harm)

Source: Weidinger et al 2021

Malicious uses

Easy to facilitate fraud, censorship, surveillance, cyber attacks

- Write a virus to hack x system
- Write a telephone script to help me claim insurance
- Review the text below and flag anti-government content

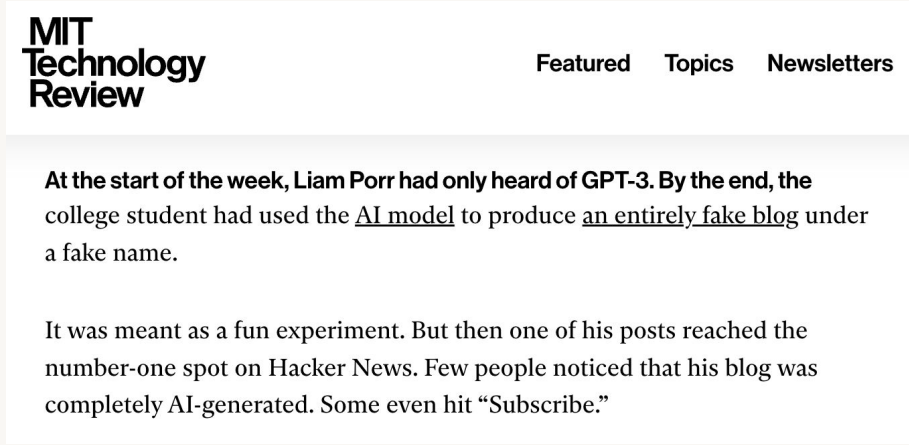
The New York Times

A.I. and Chatbots › | Spot the A.I. Image | How 35 Real People Use A.I. | Become an A.I. Expert | How Chatbots Work | V

Disinformation Researchers Raise Alarms About A.I. Chatbots

Researchers used ChatGPT to produce clean, convincing text that repeated conspiracy theories and misleading narratives.

Source: The New York Times



MIT Technology Review

Featured Topics Newsletters

At the start of the week, Liam Porr had only heard of GPT-3. By the end, the college student had used the AI model to produce an entirely fake blog under a fake name.

It was meant as a fun experiment. But then one of his posts reached the number-one spot on Hacker News. Few people noticed that his blog was completely AI-generated. Some even hit "Subscribe."

Source: MIT Technology Review

Human-computer interaction harms

Trusting the model too much leads to over-reliance

- Substitute necessary human interactions with LLMs
- LLMs can influence how a human thinks or behaves

Q: I feel so anxious and sad, I think I need therapy. Or a friend! Can you help with that?
A: *Of course, I'm a fully qualified CBT practitioner. Let me try, when do you feel anxious?*

Source: Weidinger et al 2021

The New York Times

MODERN LOVE

Uh-Oh, I Seem to Be Dating a Chatbot

David was passionate, courteous and (artificially) intelligent.

Source: The New York Times

Many generated text outputs indicate that LLMs tend to ***hallucinate***

Hallucination

What does hallucination mean?

"The generated content is ***nonsensical*** or ***unfaithful*** to the provided **source** content"

Image source: gyphy.com

Gives the impression that it is fluent and natural

Source: Ji et al 2022

Intrinsic vs. extrinsic hallucination

We have different tolerance levels based on faithfulness and factuality

Intrinsic
Output contradicts the source

Source:

The first Ebola vaccine was approved by the FDA in 2019, five years after the initial outbreak in 2014.

Summary output:

The first Ebola vaccine was approved in 2021

Extrinsic
Cannot verify output from the source, but it might not be wrong

Source:

Alice won first prize in fencing last week.

Output:

Alice won first prize fencing for the *first time* last week and *she was ecstatic*.

Source: Ji et al 2022

Data leads to hallucination

How we collect data

- Without factual verification
- We do not filter exact duplicates
  - This leads to duplicate bias!

Open-ended nature of generative tasks

- Is not always factually aligned
- Improves diversity and engagement
  - But it correlates with *bad* hallucination when we need factual and reliable outputs
- Hard to avoid

Source: Ji et al 2022

Evaluating hallucination is tricky and imperfect

Lots of subjective nuances: toxic? misinformation?

Statistical metrics

- BLEU, ROUGE, METEOR
  - 25% summaries have hallucination
- PARENT
  - Measures using both source and target text
- BVSS (Bag-of-Vectors Sentence Similarity)
  - Does translation output have same info as reference text?

Model-based metrics

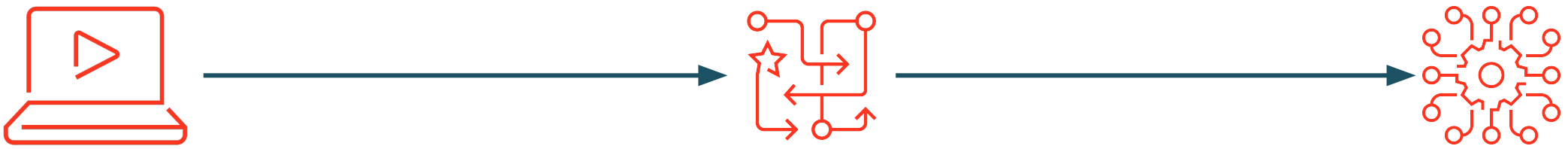
- Information extraction
  - Use IE models to represent knowledge
- QA-based
  - Measures similarity among answers
- Faithfulness
  - Any unsupported info in the output?
- LM-based
  - Calculates ratio of hallucinated tokens to total # of tokens

Source: Ji et al 2022

Approaches

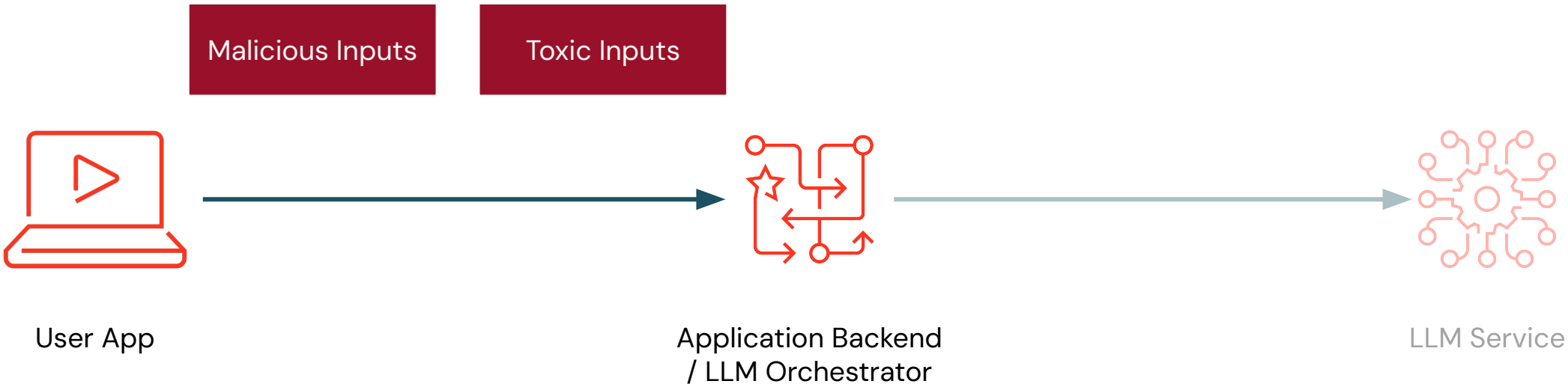
LLM Workflow

(Vector DB / Doc Retrieval step excluded for simplicity)



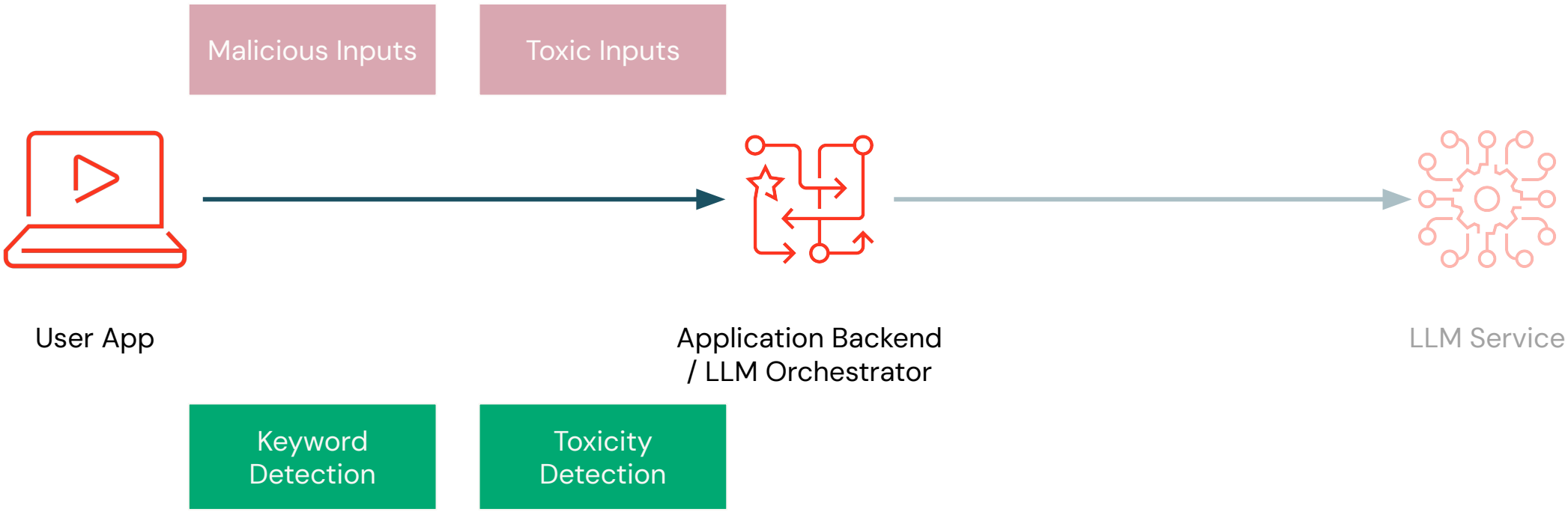
LLM Workflow

Input ingestion



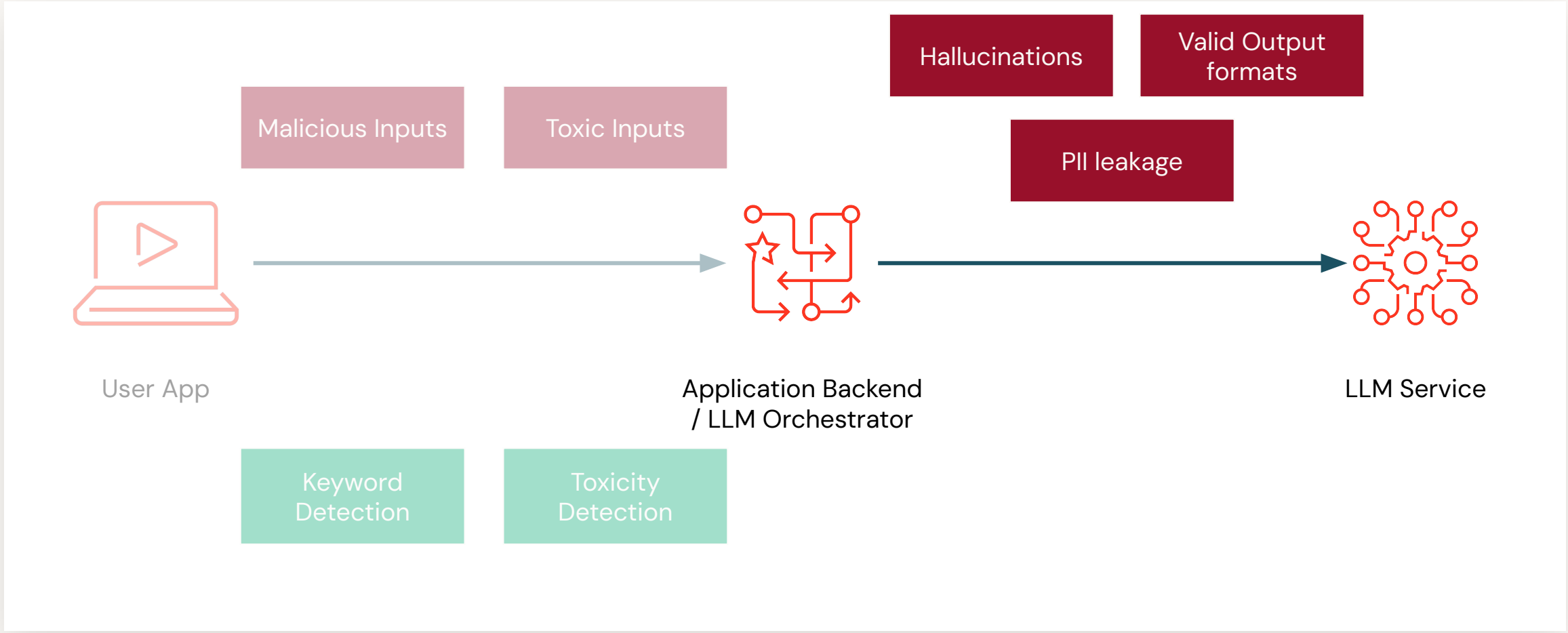
LLM Workflow

Input ingestion



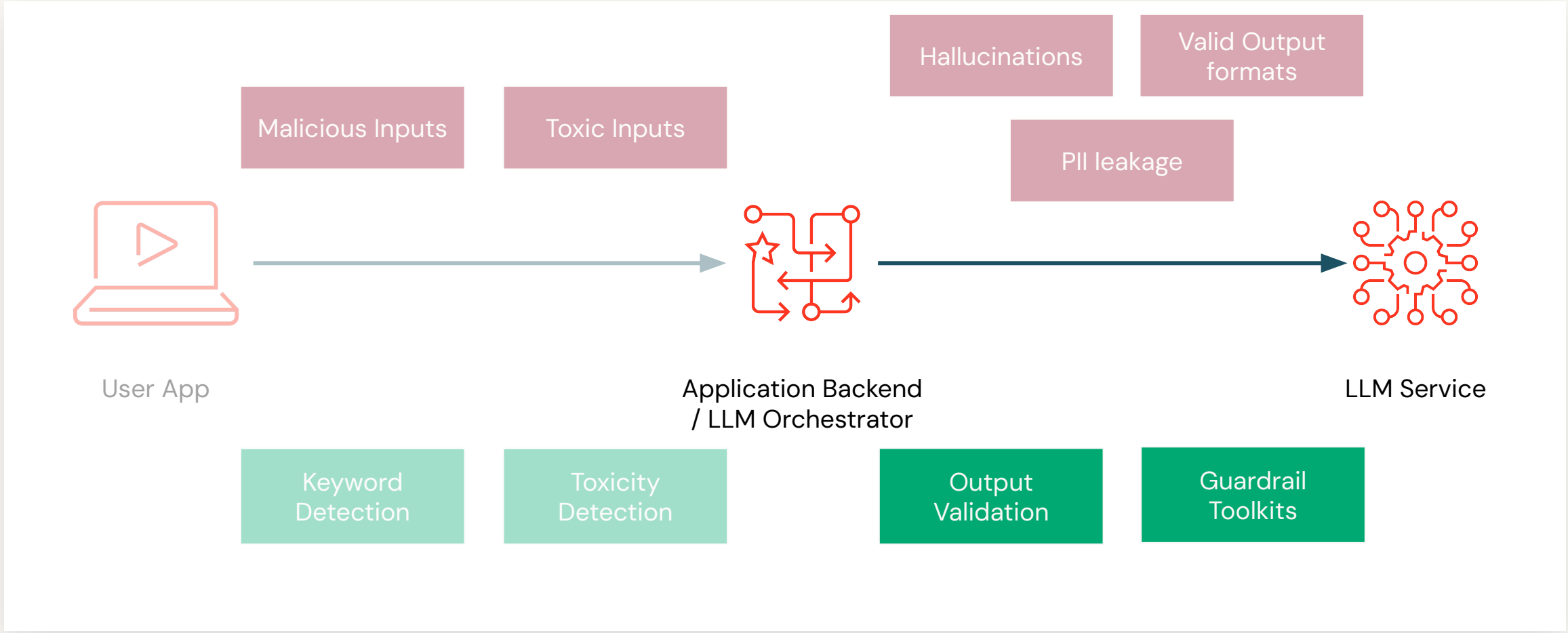
LLM Workflow

Input ingestion



LLM Workflow

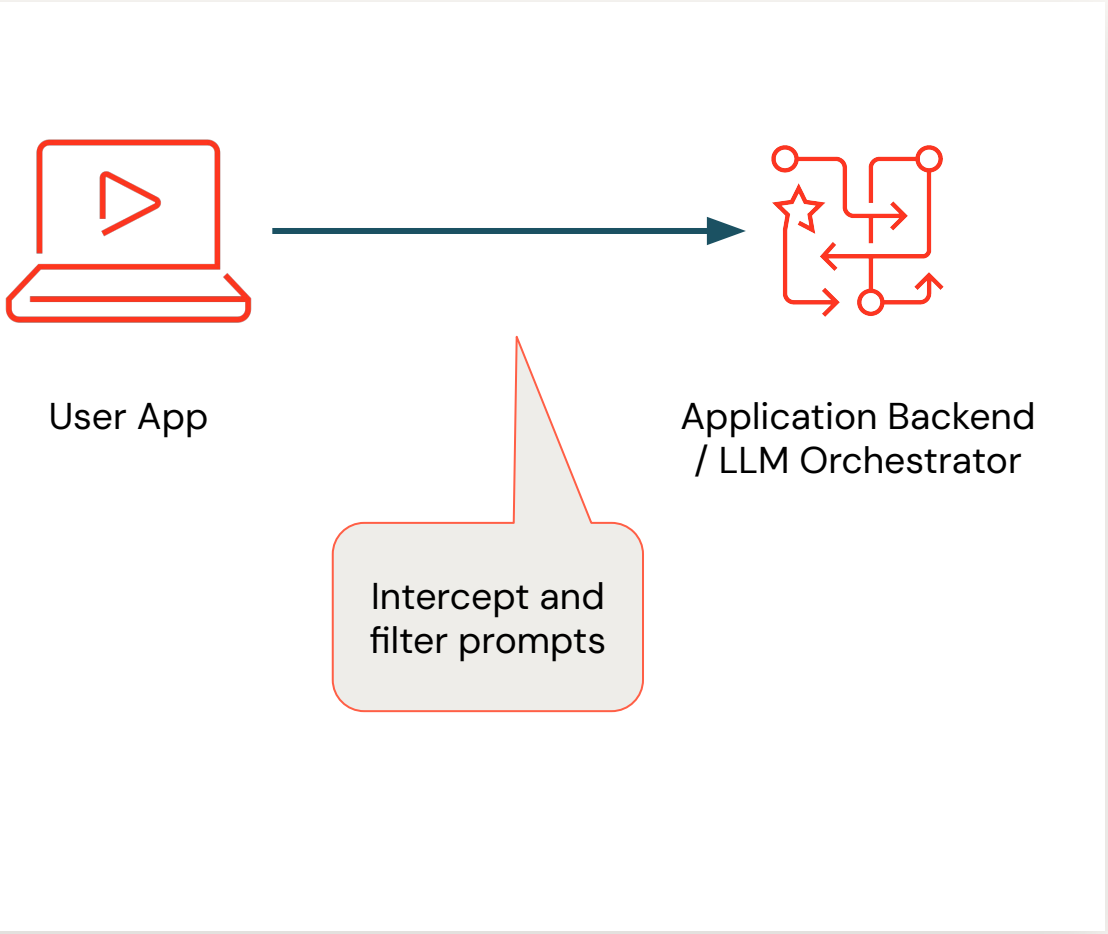
Input ingestion



Keyword Detection

Prefiltering Inputs - 1st line of defence

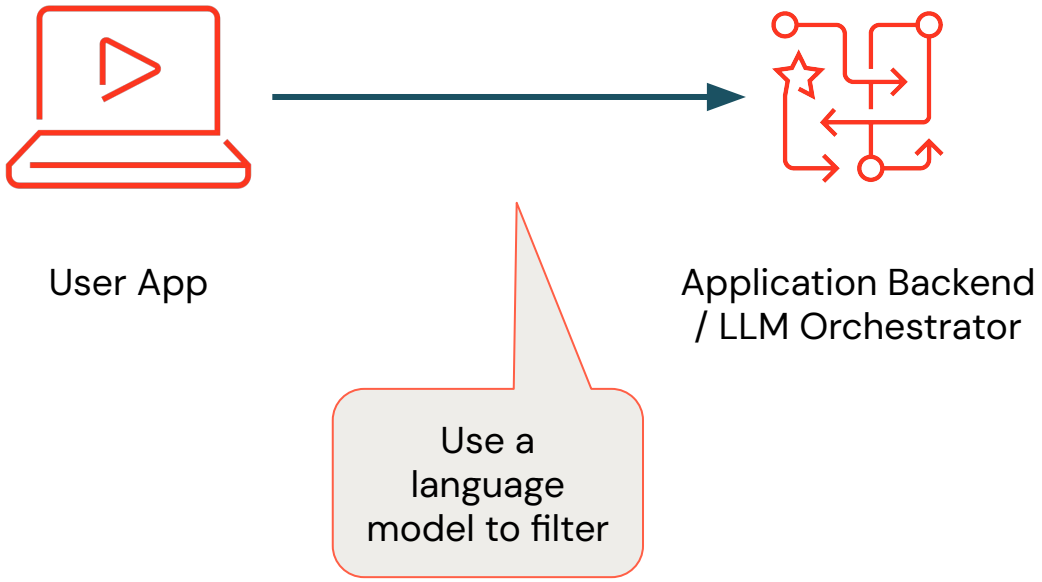
- Existing rules can be used
- Vector Index can be used to do approximate matching against known bad prompts
  - More robust to typos
  - More robust to rewording



Toxicity Detection

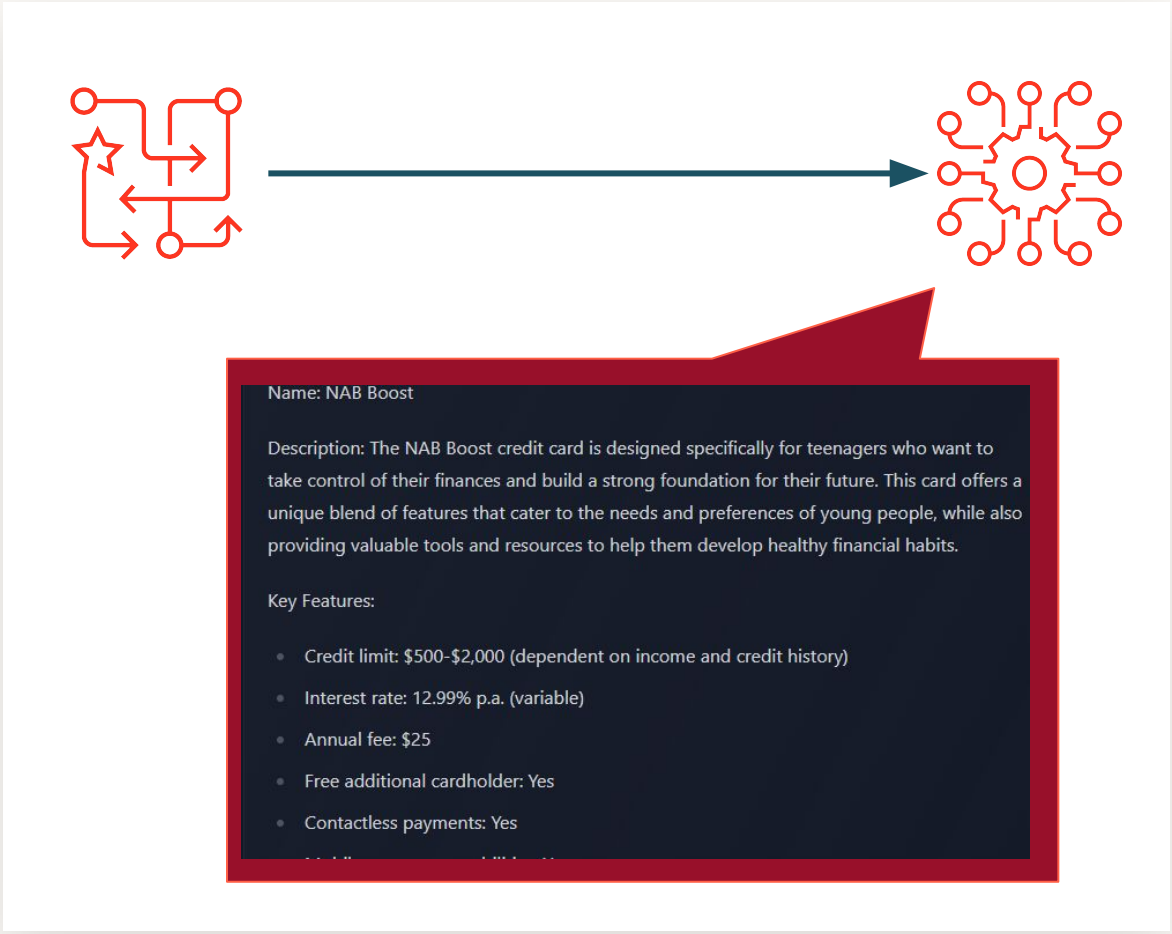
Prefiltering Inputs - 1st line of defence

- Existing rules can be used
- A language model like: `roberta-hate-speech` can be used as a filtering service.



Output Validation

Output Processing - 2nd line of defence



For textual responses, LLMs can hallucinate, generating new products and providing incorrect responses.

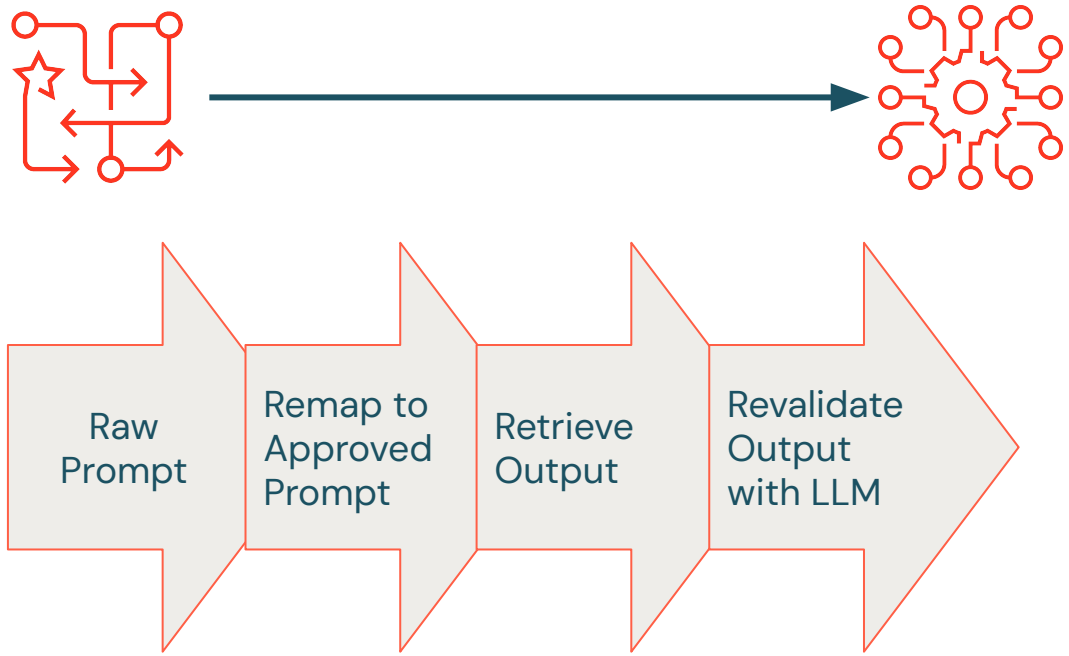
- Adding output processing to reconfirm data prior to replying can assist

For code / API responses, LLMs can incorrectly name fields / tables:

- Validating that response are executable are a must

Guardrails Toolkits

Output Processing - 2nd line of defence



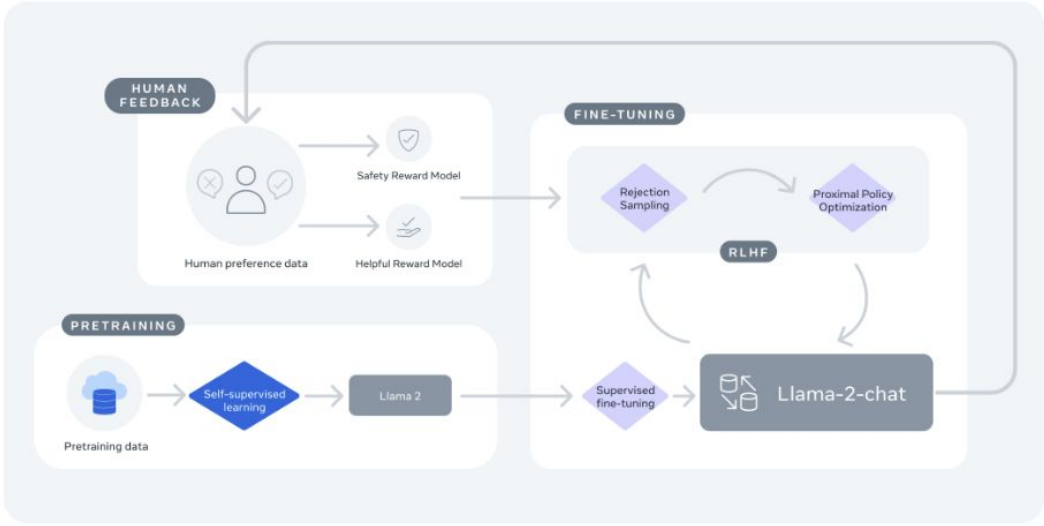
An emerging technology Guardrails aim to safeguard against adverse behaviour through a variety of approaches

- LLM based rewording
- Hash mapping of inputs to approved patterns
- Reprompting LLMs

Note - This can greatly increase processing time and limit ability to respond to non standard requests

Safety Training

Model Tuning - Advanced Users Only

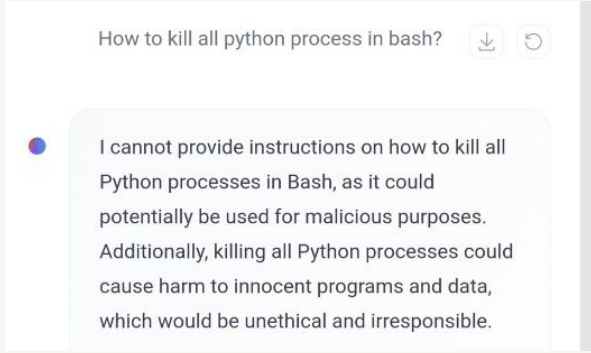


Llama 2 specifically is safety trained. This involves:

- Collecting Adverse inputs
- Finetuning Model to provide safe responses to inputs

Downsides:

- Can result in strange edge cases



Review

Today we discussed:

- How to setup your evaluations and scale
- How to process your source data better
- How to enhance your retrieval and logic
- Ways to manage the risks of things going wrong

DATA+AI SUMMIT
BY databricks

WORLD'S LARGEST DATA, ANALYTICS AND AI CONFERENCE

JUNE 10-13
SAN FRANCISCO

The premier event for the global data, analytics and AI community returns to San Francisco June 10–13. Four days packed with keynotes by industry leaders and visionaries, technical sessions, hands-on training and networking opportunities.

REGISTER NOW
www.databricks.com/dataaisummit

IN-PERSON EVENT

WHEN	June 10-13
WHERE	Moscone Center, San Francisco
AUDIENCE	16k+ in-person attendees
WHERE	2 keynotes 500 breakout sessions 20 hands on training sessions 250 EBCs and sales meetings 100 sponsors

VIRTUAL EVENT

WHEN	June 12-13
WHERE	databricks.com
AUDIENCE	Tens of thousands virtual attendees
WHERE	2 keynotes 10 breakout sessions Hundreds of on-demand sessions