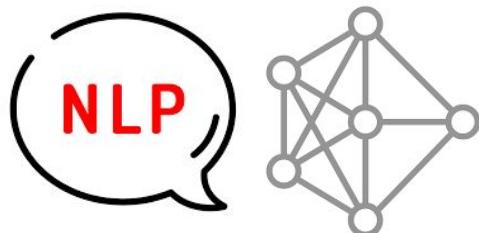


+

CHULA ENGINEERING
Foundation toward Innovation

COMPUTER



RAG

2110572: Natural Language Processing Systems

Peerapon Vateekul & Ekapol Chuangsawanich

Department of Computer Engineering,
Faculty of Engineering, Chulalongkorn University



Outline

- Introduction
- Retrieval-augmented Generation
 - Motivation
 - Typical Workflow
 - Hybrid Search
 - Document Reranking
 - Advanced RAG
 - Evaluation
- Agentic LLMs
 - Motivation
 - Reasoning
 - Memory (?)
 - Example
 - Evaluation (LangSmith)
- Reasoning Model (GPT-o1, o3, DeepSeek-R1)

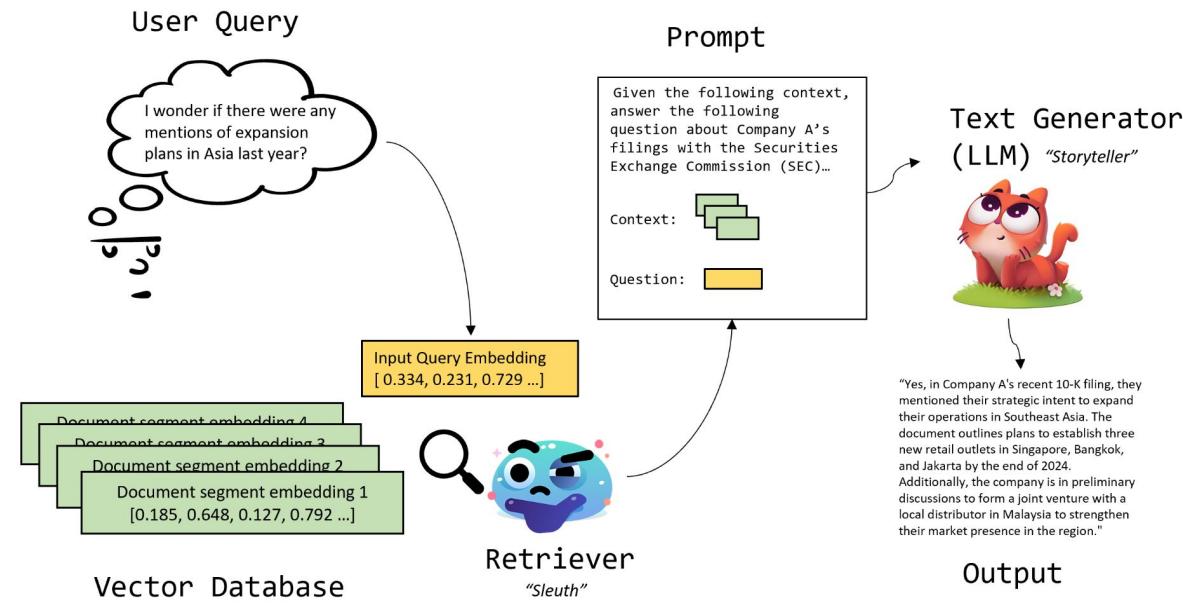


Introduction



Retrieval-Augmented Generation (RAG)

- Combining a retriever with an LLM to generate highly informed and contextually relevant outputs.
 - Given a query, the retriever finds the most relevant documents.
 - The LLM takes the documents to answer the query.



Text Retrieval System

A system that finds the most relevant text (such as an answer, paragraph, or passage) given a query (which could be a question, keywords, or any relevant text).

Sparse embedding

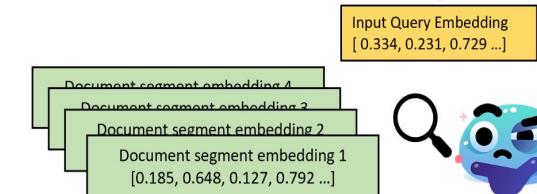
```
{'mil': 0.27, '##vus': 0.39, 'is': 0.16, 'a': 0.11, 'vector': 0.41, 'database': 0.24,
'built': 0.17, 'for': 0.09, 'scala': 0.19, '##ble': 0.11, 'similarity': 0.30, 'search': 0.19}
```

BM25

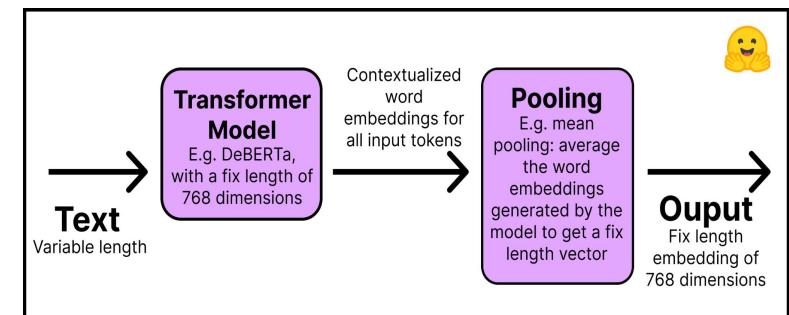
$$BM25 = \sum_{t \in q} \log \left[\frac{N}{df(t)} \right] \cdot \frac{(k_1 + 1) \cdot tf(t, d)}{k_1 \cdot \left[(1 - b) + b \cdot \frac{dl(d)}{dl_{avg}} \right] + tf(t, d)}$$

- k_1, b – parameters
- $dl(d)$ – length of document d
- dl_{avg} – average document length

Dense embedding



Sentence embedding





Why use RAG?

- Lets you use your own data. This includes:
 - private data
 - new data
 - domain-specific data

- Reduces hallucinations
 - by grounding the response in factual data



 What's the capital of Mars?

The capital of Mars is Muskland.





RAG vs Fine-tuning

- Use RAG if data is dynamic.
- Use fine-tuning if you want to change the behaviour of the model.
- You can also use both!

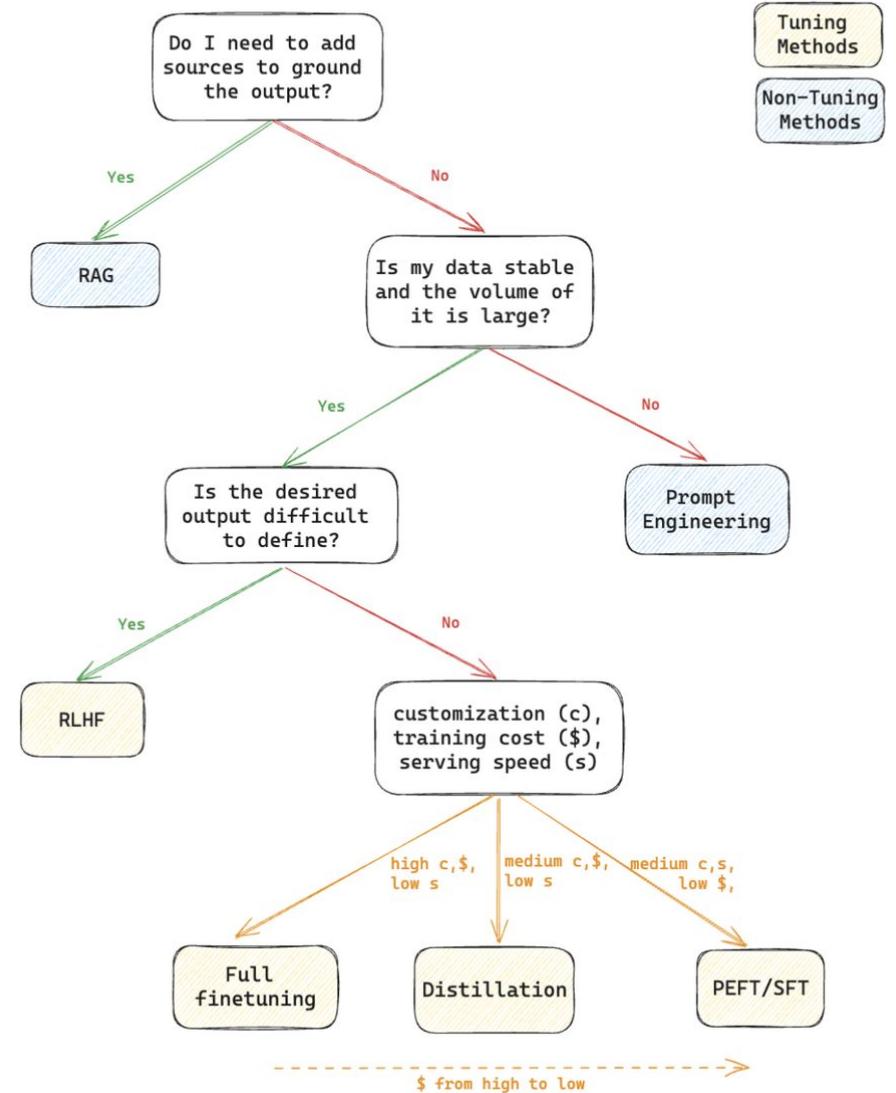
Common use cases

Some common use cases where fine-tuning can improve results:

- Setting the style, tone, format, or other qualitative aspects
- Improving reliability at producing a desired output
- Correcting failures to follow complex prompts
- Handling many edge cases in specific ways
- Performing a new skill or task that's hard to articulate in a prompt

One high-level way to think about these cases is when it's easier to "show, not tell". In the sections to come, we will explore how to set up data for fine-tuning and various examples where fine-tuning improves the performance over the baseline model.

Another scenario where fine-tuning is effective is reducing cost and/or latency by replacing a more expensive model like `gpt-4o` with a fine-tuned `gpt-4o-mini` model. If you can achieve good results with `gpt-4o`, you can often reach similar quality with a fine-tuned `gpt-4o-mini` model by fine-tuning on the `gpt-4o` completions, possibly with a shortened instruction prompt.



+

Retrieval-Augmented Generation



Typical RAG Workflow

Step 1: Split your documents into manageable-sized chunks.

- You will need to find a good chunk size based on your documents.
- If structure matters in your document, try using markdown splitter.

Markdown splitter example

For example, if we want to split this markdown:

```
md = '# Foo\n\n## Bar\n\nHi this is Jim \nHi this is Joe\n\n## Baz\n\n Hi this is Molly'
```

We can specify the headers to split on:

```
[("#", "Header 1"), ("##", "Header 2")]
```

And content is grouped or split by common headers:

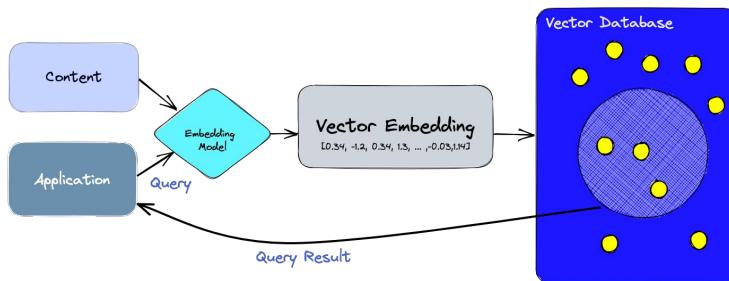
```
{'content': 'Hi this is Jim \nHi this is Joe', 'metadata': {'Header 1': 'Foo', 'Header 2': 'Bar'}}  
{'content': 'Hi this is Molly', 'metadata': {'Header 1': 'Foo', 'Header 2': 'Baz'}}
```

| Name | Classes | Splits On | Adds Metadata | Description |
|---------------------------------------|--|---------------------------------------|-------------------------------------|---|
| Recursive | RecursiveCharacterTextSplitter, RecursiveJsonSplitter | A list of user defined characters | | Recursively splits text. This splitting is trying to keep related pieces of text next to each other. This is the recommended way to start splitting text. |
| HTML | HTMLHeaderTextSplitter, HTMLSectionSplitter | HTML specific characters | <input checked="" type="checkbox"/> | Splits text based on HTML-specific characters. Notably, this adds in relevant information about where that chunk came from (based on the HTML) |
| Markdown | MarkdownHeaderTextSplitter | Markdown specific characters | <input checked="" type="checkbox"/> | Splits text based on Markdown-specific characters. Notably, this adds in relevant information about where that chunk came from (based on the Markdown) |
| Code | many languages | Code (Python, JS) specific characters | | Splits text based on characters specific to coding languages. 15 different languages are available to choose from. |
| Token | many classes | Tokens | | Splits text on tokens. There exist a few different ways to measure tokens. |
| Character | CharacterTextSplitter | A user defined character | | Splits text based on a user defined character. One of the simpler methods. |
| [Experimental] Semantic Chunker | SemanticChunker | Sentences | | First splits on sentences. Then combines ones next to each other if they are semantically similar enough. Taken from Greg Kamradt |
| AI21 Semantic Text Splitter | AI21SemanticTextSplitter | <input checked="" type="checkbox"/> | | Identifies distinct topics that form coherent pieces of text and splits along those. |

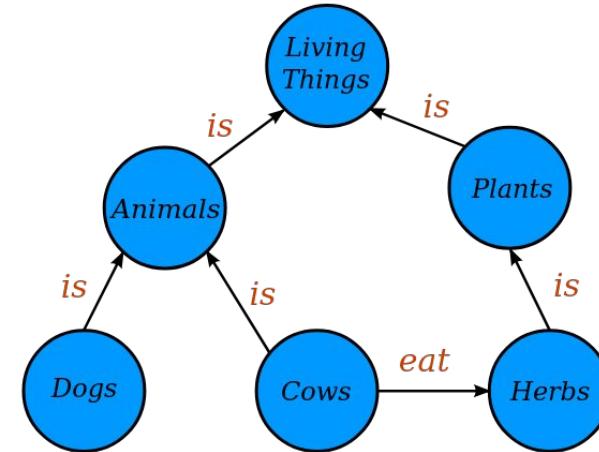
Typical RAG Workflow

Step 2: Index your chunks into a database or knowledge base of some kind.

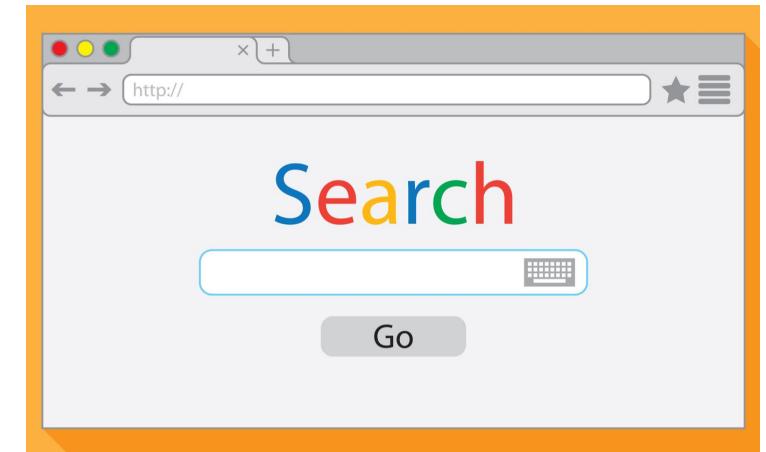
Vector Databases



Knowledge Graphs



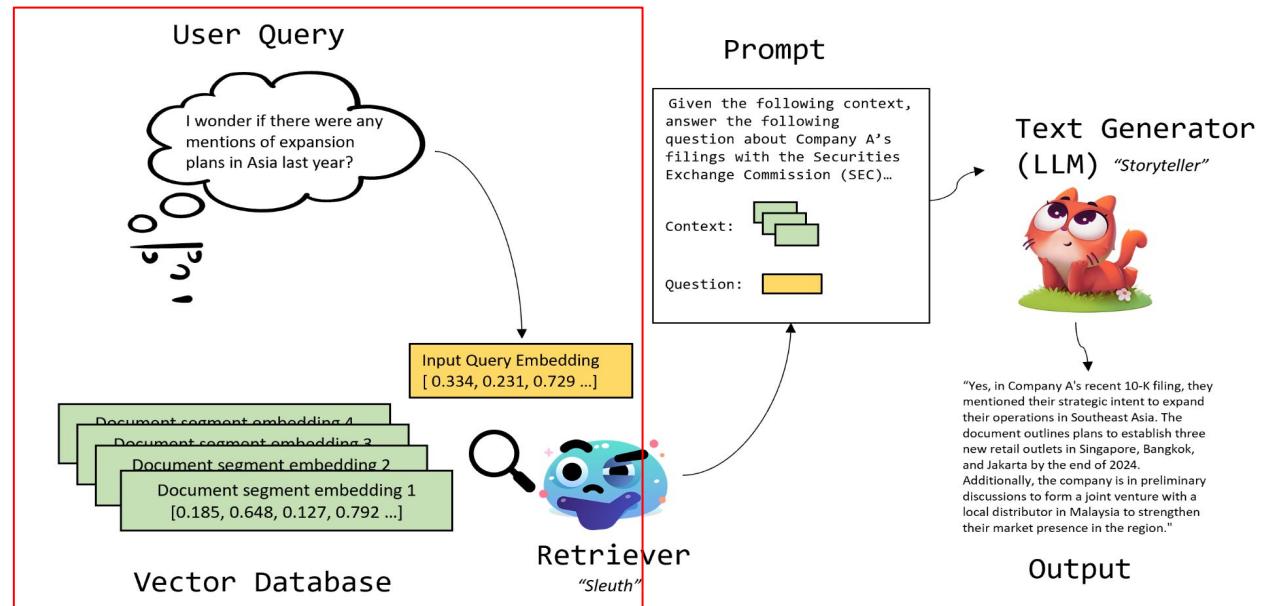
Search Engines



+ Typical RAG Workflow

Step 3: Given a query, **find the most similar documents** in your knowledge base.
E.g., perform vector search in a vector database.

- Use hybrid search for better performance
- (Optional) - perform document reranking





Document-Query Interaction Paradigm

(a) is the most efficient but least performant.

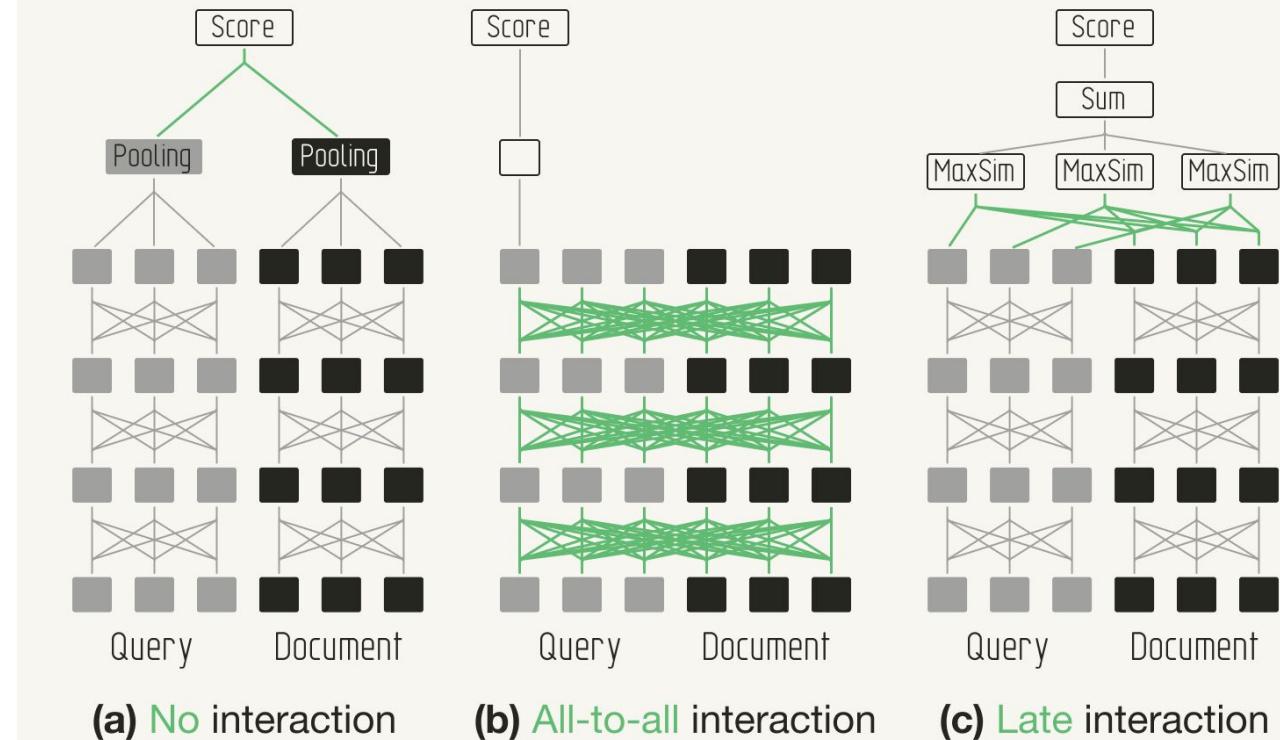
(b) is the most computationally expensive but performs best.

(a) and (c) can be precomputed offline.

To trade-off between speed and accuracy, two-stage ranking is often used: (a) (and optionally (c)) is used to find a small group of candidates then run (b) for the final score.

Types of interaction

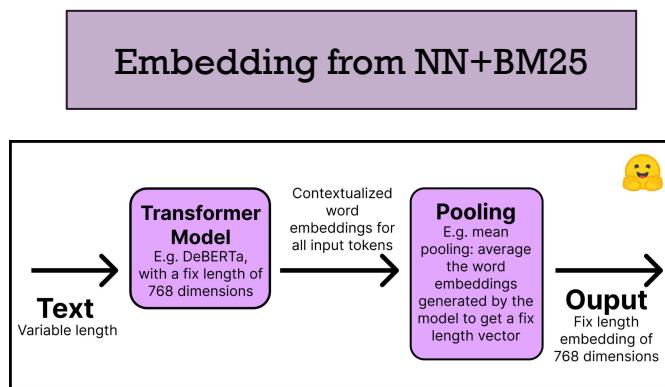
Fig. 1: Three types of interaction in dense retrievers.





Hybrid Search

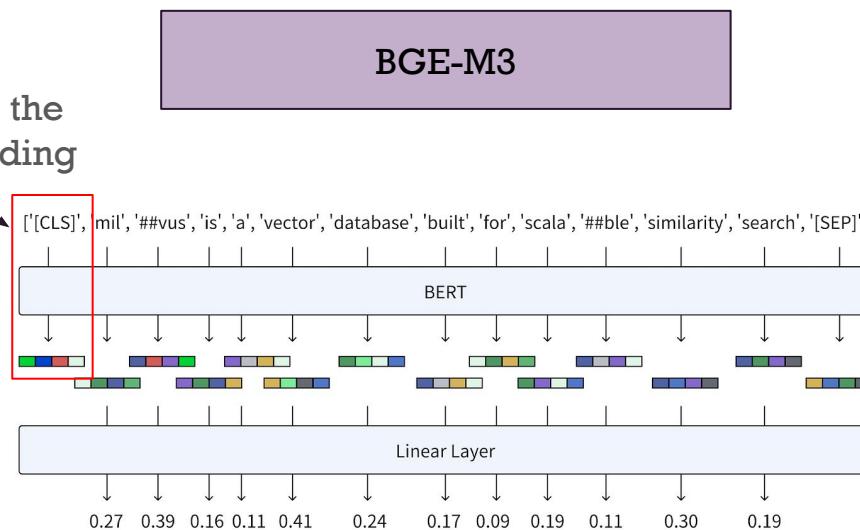
- Basically **contextual search (dense embedding)** + **keyword search (sparse embedding)**
- Supported by many vector DBs.



$$BM25 = \sum_{t \in q} \log \left[\frac{N}{df(t)} \right] \cdot \frac{(k_1 + 1) \cdot tf(t, d)}{k_1 \cdot \left[(1 - b) + b \cdot \frac{dl(d)}{dl_{avg}} \right] + tf(t, d)}$$

- k_1, b – parameters
- $dl(d)$ – length of document d
- dl_{avg} – average document length

[CLS] token
embedding is the
dense embedding

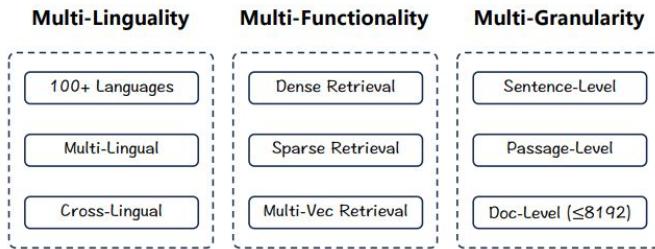


The sparse embedding

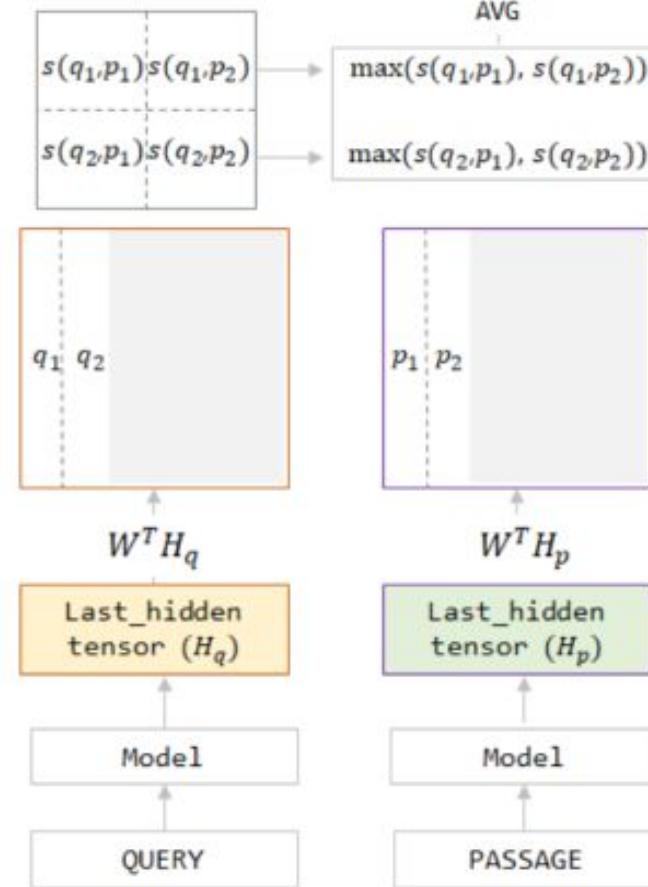
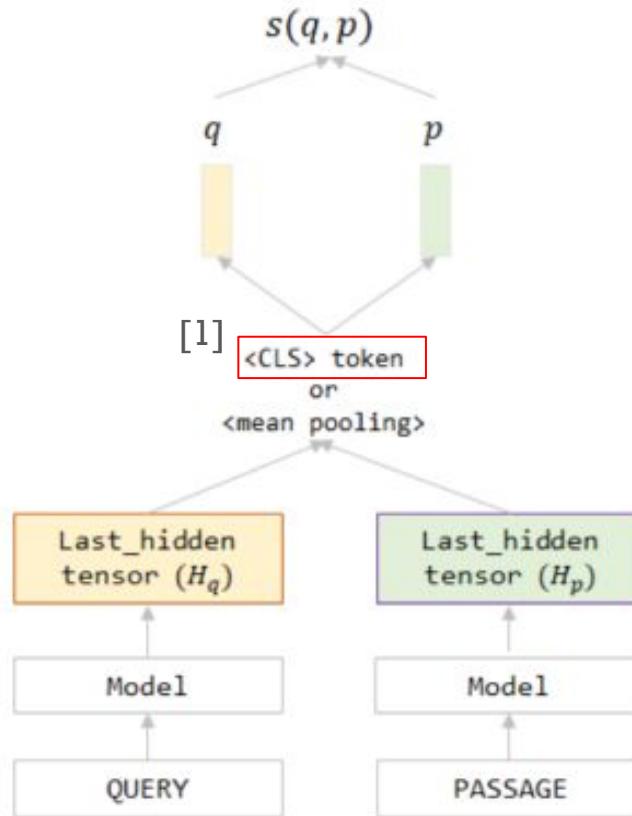
{'mil': 0.27, '#vus': 0.39, 'is': 0.16, 'a': 0.11, 'vector': 0.41, 'database': 0.24, 'built': 0.17, 'for': 0.09, 'scala': 0.19, '#ble': 0.11, 'similarity': 0.30, 'search': 0.19}



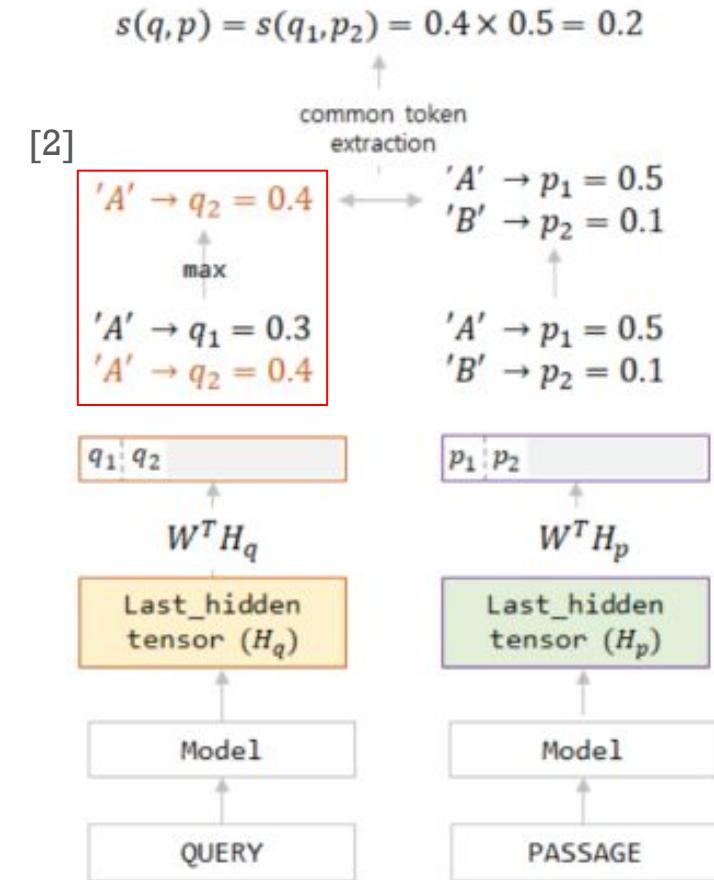
BGE-M3



- [1] Use the [CLS] token as the dense embedding
- [2] Take the max score in case a term appears multiple times. This is the sparse embedding. (the score is the importance of each term like BM25)
- [3] Late interaction paradigm



[3] **Multi-Vector retrieval**



Lexical retrieval



Hybrid Search

- Reciprocal Rank Fusion (RRF) and Distribution-based Score Fusion (DBSF) are used to combine the score between multiple search results

$$RRFscore(d \in D) = \sum_{r \in R} \frac{1}{k+r(d)}$$

D - set of docs

R - set of rankings as permutation on 1..|D|

K - typically set to 60 by default

| Retrieval Modules | Execution Time(seconds) | Ragas Context Precision@10 |
|--|-------------------------|----------------------------|
| Bm25 | 0.274728 | 0.649015 |
| VectorDB | 0.496673 | 0.522239 |
| Hybrid RRF (RRF-k: 10) | 0.771401 | 0.676157 |
| Hybrid RRF (RRF-k: 3) | 0.771401 | 0.640295 |
| Hybrid RRF (RRF-k: 5) | 0.771401 | 0.668342 |
| Hybrid CC (Weights: 0.7, 0.3) | 0.771401 | 0.652625 |
| Hybrid DBSF (Weights: 0.7, 0.3) | 0.771401 | 0.696401 |

Table 3: Table of Execution Time and Ragas Context Precision by Retrieval Experiments

DBSF

$$S \left(\bigcup_{i \in E} \frac{x_i - (\mu_i - 3\sigma_i)}{(\mu_i + 3\sigma_i) - (\mu_i - 3\sigma_i)} \right)$$

where:

- E is a set of embeddings from retrievers
- mu and sigma are the mean and std. of each set of embedding

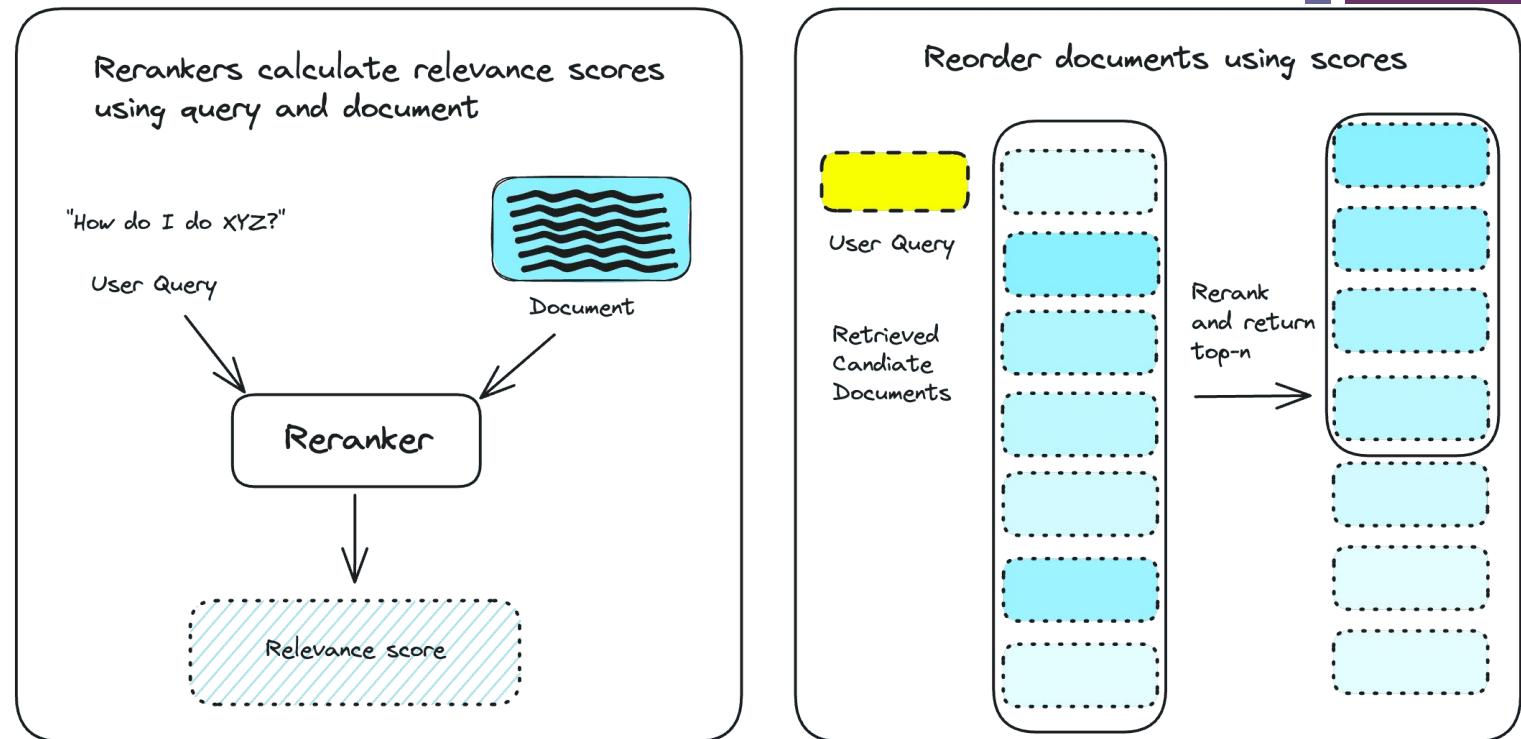
+

Document Reranking

Uses a reranker model or a “cross-encoder” to **rerank the retrieved documents by the relevancy of the query to the documents.**

This process often improves the RAG pipeline performance by weeding out unrelated documents.

Reranker Workflow



Performance on MIRACL dataset

| model | inference length | avg | ar | bn | en | es | fa | fi | fr | hi | id | ja | ko | ru | sw | te | th | zh | de | yo |
|------------------------------|------------------|--------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| bge-m3 | 512+512 | 67.91 | 78.4 | 80 | 59.6 | 55.5 | 57.7 | 78.6 | 57.8 | 59.3 | 56 | 72.8 | 69.9 | 70.1 | 78.6 | 86.2 | 82.6 | 61.7 | 56.8 | 60.7 |
| bge-reranker-v2-m3 | 1024 | 72.84 | 81.7 | 84.63 | 63.45 | 63.71 | 62.49 | 82.41 | 63.26 | 68.25 | 62.71 | 79.96 | 73.79 | 76.93 | 82.27 | 89.36 | 85.3 | 64.2 | 62.64 | 64.03 |
| bge-reranker-v2-gemma | 1024 | 73.39 | 82.26 | 85.15 | 66.62 | 64.29 | 62.03 | 82.58 | 64.26 | 68.68 | 61.33 | 79.72 | 74.83 | 78.36 | 81.46 | 89.22 | 86.06 | 65.61 | 64.25 | 64.37 |
| bge-reranker-v2-minicpm-20 | 1024 | 62.26 | 71.8 | 62.15 | 62.51 | 56.45 | 45.28 | 74.73 | 52.77 | 50.64 | 57.3 | 70.57 | 69.12 | 60.58 | 67.4 | 76.98 | 67.87 | 61.61 | 49.74 | 63.2 |
| bge-reranker-v2-minicpm-28 | 1024 | 67.75 | 77.4 | 64.87 | 66.03 | 61.89 | 51.91 | 80.02 | 63.42 | 56 | 60.5 | 78.25 | 73.39 | 72.5 | 72.09 | 78.41 | 74.46 | 66.09 | 59.15 | 63.07 |
| bge-reranker-v2-minicpm-40 | 1024 | 67.77 | 77.49 | 64.8 | 66.02 | 62.05 | 51.78 | 80.12 | 62.98 | 55.7 | 60.51 | 78.07 | 73.01 | 72.64 | 72.81 | 78.51 | 74.24 | 65.82 | 59.43 | 63.91 |

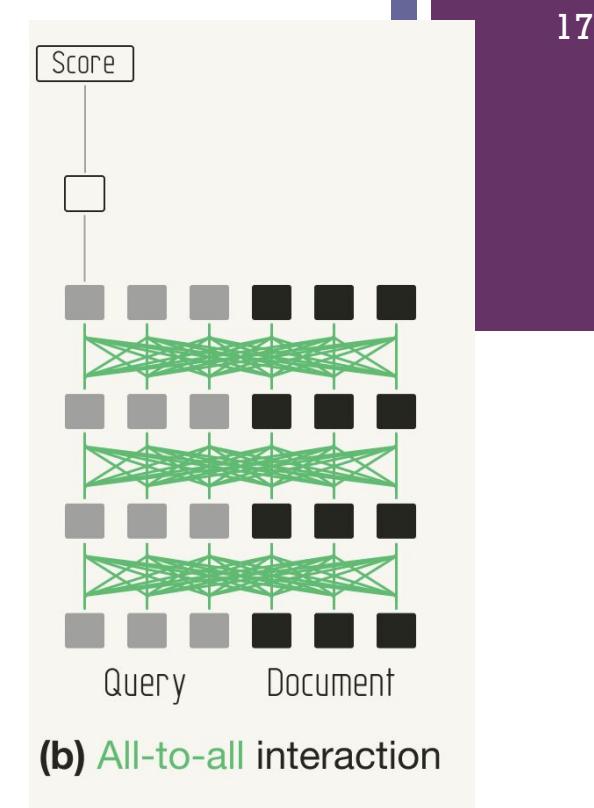
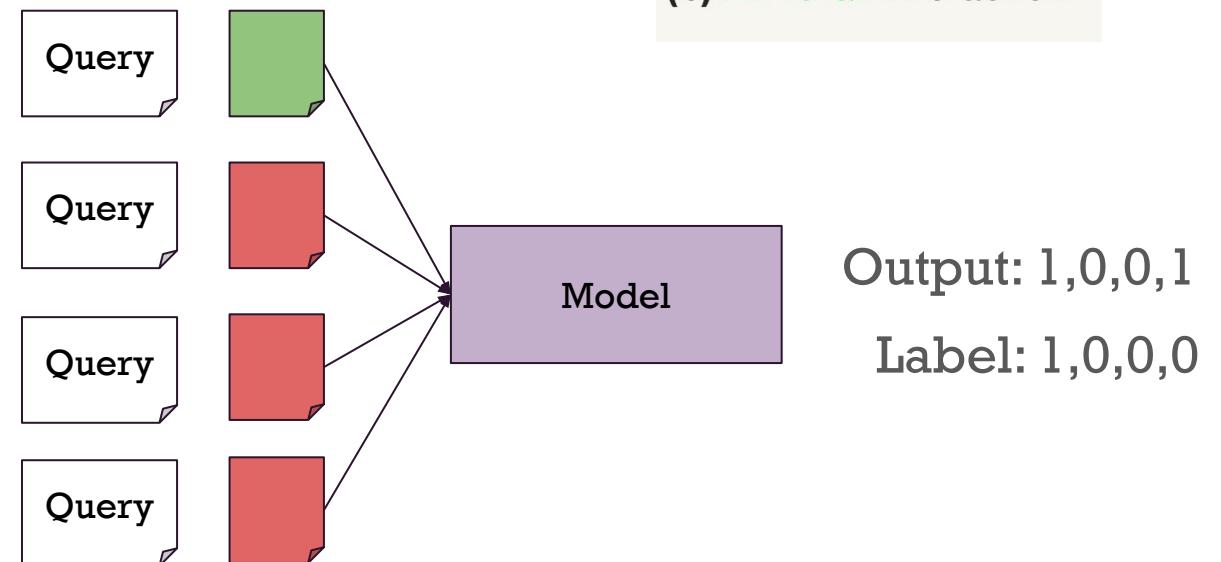


Reranker

A reranker takes in the query and a document and outputs a score of [0,1] where 1 means the doc is most relevant to the query.

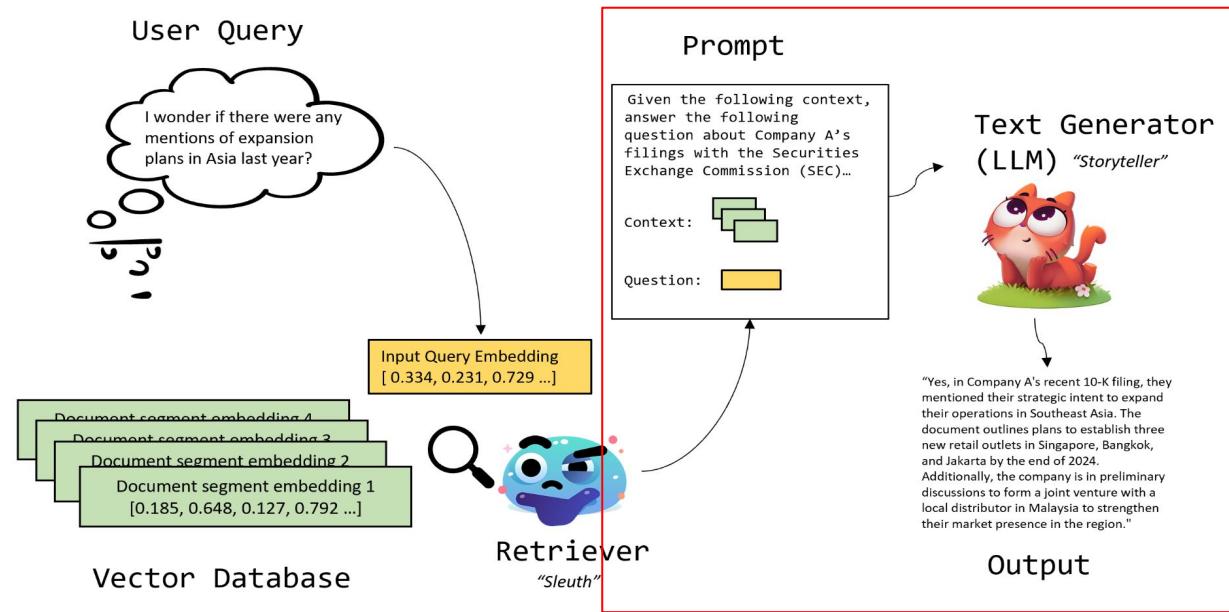
To train a reranker, you will need to feed it:

- A query
- Positive documents
- (Hard) Negative documents
- Labels



+ Typical RAG Workflow

Step 4: Feed the query and documents to the LLM to generate an answer.



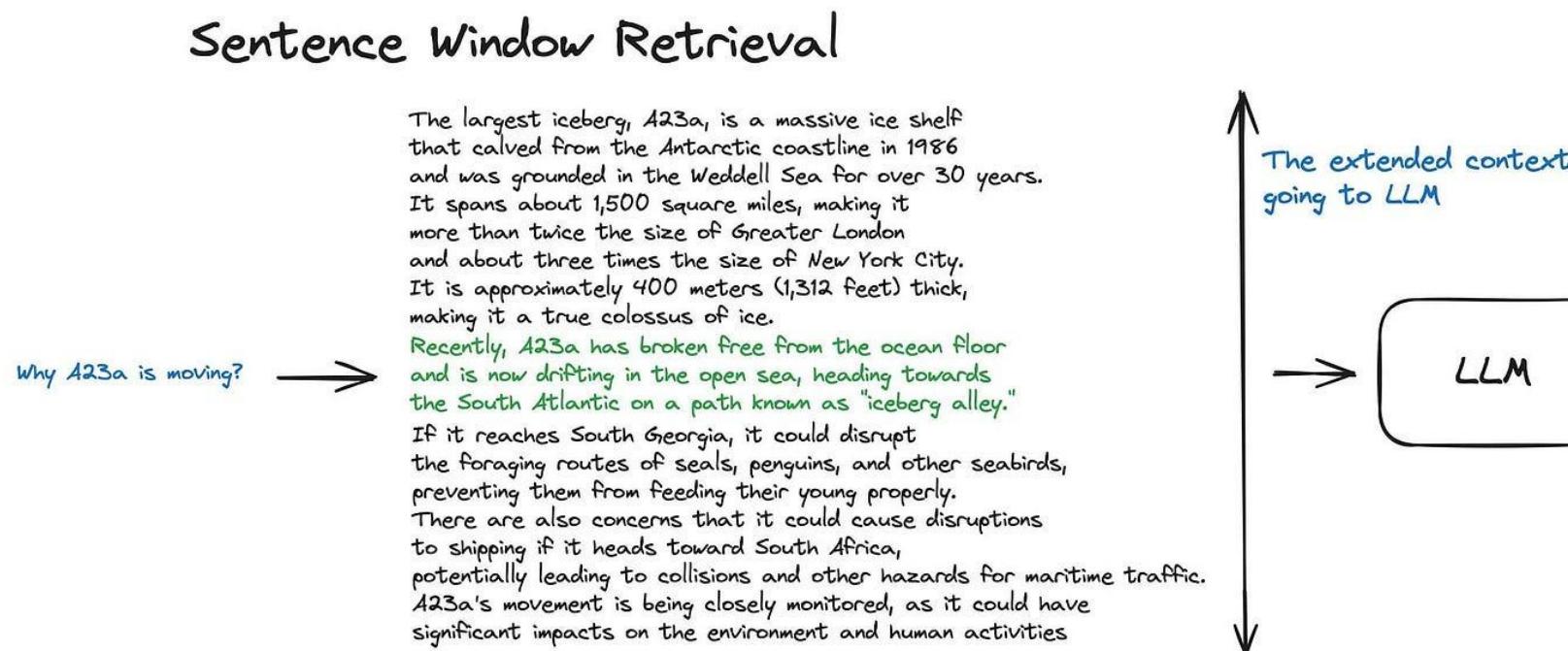
+

Advanced RAG



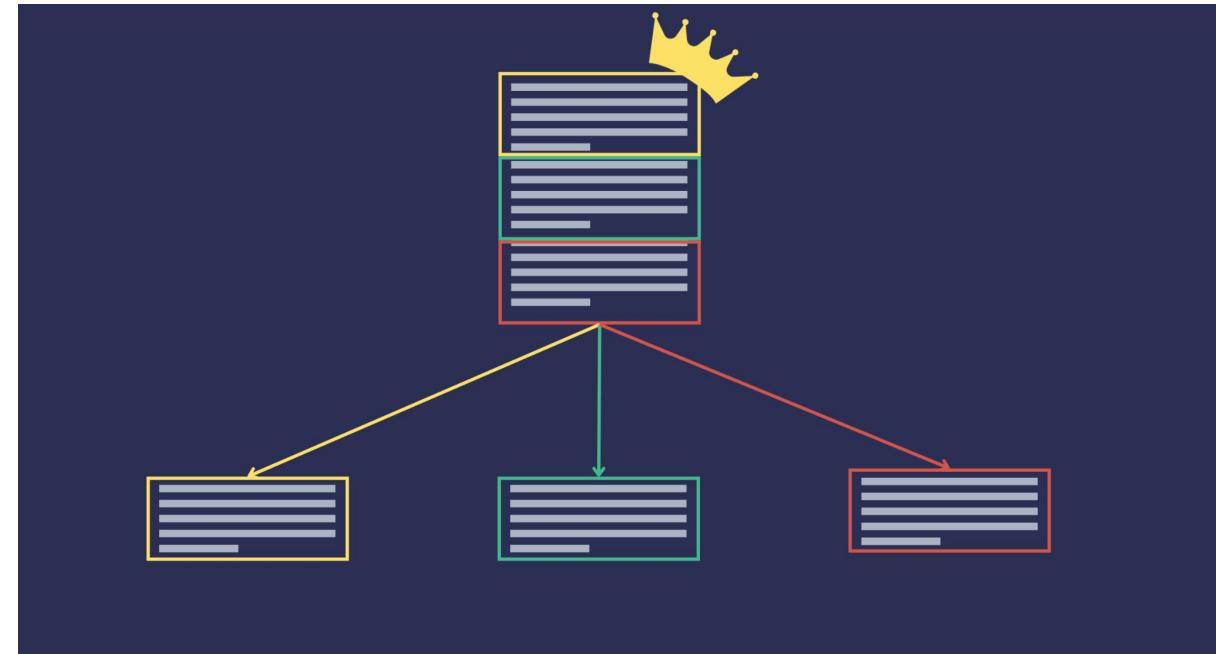
Sentence Window Retriever

Instead of chunk-based RAG, sentence window retriever instead looks for the sentence most relevant to the query and take **the entire window of sentences** around it as the context for the LLM.



Auto-Merging Retriever

- Auto-Merging is a retrieval technique that leverages a hierarchical document structure.
 - Parent nodes are the original document and children nodes are chunks of their parent.
- Return top_k docs normally.
- However, if the no. of children nodes that belong to the same parent returned is more than a threshold, return the parent node instead.

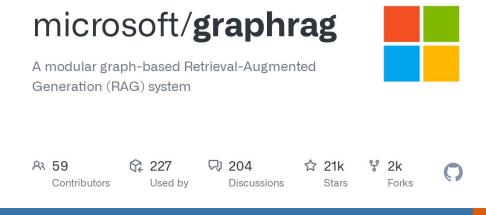


e.g. if threshold = 50% and 2 children are retrieved, return the parent instead



Tools & Framework

Frameworks



Vector DBs



facebookresearch/faiss

A library for efficient similarity search and clustering of dense vectors.

167 Contributors, 4k Used by, 225 Discussions, 32k Stars, 4k Forks



Chroma

Load data and build an index

In the same folder where you created the `data` folder, create a file called `starter.py` file with the following:

```
from llama_index.core import VectorStoreIndex, SimpleDirectoryReader

documents = SimpleDirectoryReader("data").load_data()
index = VectorStoreIndex.from_documents(documents)
```

This builds an index over the documents in the `data` folder (which in this case just consists of the essay text, but could contain many documents).

Your directory structure should look like this:

```
└── starter.py
└── data
    └── paul_graham_essay.txt
```

Query your data

Add the following lines to `starter.py`

```
query_engine = index.as_query_engine()
response = query_engine.query("What did the author do growing up?")
print(response)
```

This creates an engine for Q&A over your index and asks a simple question. You should get back a response similar to the following: The author wrote short stories and tried to program on an IBM 1401.



Evaluation

Retrieval

- **Hit-rate:** assesses whether the relevant documents are being retrieved.
- **MRR (Mean reciprocal rank):** gauges if the relevant documents are ranked high when retrieved.
- **NDCG@K (Normalized Discounted Cumulative Gain):** evaluates the quality of ranked results by considering the relevance and the position of each relevant item in the entire ranked list

| | Reciprocal Rank | | | | | |
|--|-----------------|---|---|---|---|--------------|
| Query 1 | 1 | 2 | 3 | 4 | 5 | $1/2 = 0.5$ |
| Query 2 | 1 | 2 | 3 | 4 | 5 | $1/1 = 1$ |
| Query 3 | 1 | 2 | 3 | 4 | 5 | $1/4 = 0.25$ |
| Mean Reciprocal Rank (MRR) = $(0.5 + 1 + 0.25)/3 = 1.75/3 \approx 0.583$ | | | | | | |

Generation

- **Faithfulness:** checks for hallucination (whether the response is grounded to the retrieved contexts.)
- **Relevancy:** evaluates the relevance of both the retrieved context and the generated answer to the initial query.
- **Correctness:** determines if the generated answer is correct based on the reference answer.

Apple : 2 , Banana : 1 , Orange : 0

Using them, We calculate:

$$DCG@3 = \frac{1}{\log_2(1+1)} + \frac{2}{\log_2(1+2)} + \frac{0}{\log_2(1+3)} = 2.26$$

$$IDCG@3 = \frac{2}{\log_2(1+1)} + \frac{1}{\log_2(1+2)} + \frac{0}{\log_2(1+3)} = 2.63$$

$$NDCG@3 = \frac{2.26}{2.63} = 0.86$$

$$NDCG@K = \frac{DCG@K}{IDCG@K} = \frac{\sum_{i=1}^{k \text{ (actual order)}} \frac{Gains}{\log_2(i+1)}}{\sum_{i=1}^{k \text{ (ideal order)}} \frac{Gains}{\log_2(i+1)}}$$

- Actual List: ["Apple", "Banana", "Orange"]
- Retrieved List: ["Banana", "Apple", "Orange"]



Evaluation



You can also specify your own metric(s) such as:

- answer conciseness, fluency, etc.; these metrics can be evaluated by a strong LLM like GPT-4o (refer to LLM-as-a-judge).
- or even the number of tokens spent.

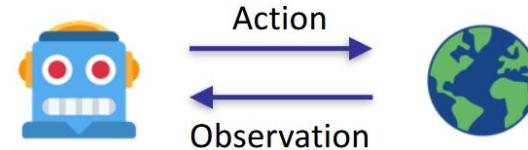
The screenshot shows the LangSmith web application interface. At the top, there's a header with a parrot icon, two crossed hammers, and the text "LangSmith". Below the header, the URL is smith.langchain.com/projects/p/da05d8be-995f-4ee7-8d1b-ce8943bb085e?eval=true. The main area displays a table of "Traces" under the heading "All Runs". The table has columns: Run Type, Output, Latency, Feedback, Reference Example, and Status. There are 17 total runs, 105 total tokens, and a latency P50: 9.09s / P99: 13.07s. The feedback column shows correctness scores: 0.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 0.00. The status column shows green checkmarks for most runs. On the right side, there's a sidebar titled "Filters" with sections for Full-Text Search, Feedback (Correctness, Correctness==0, Correctness==1), Feedback Source (Model), Name (RunnableSequence), Run Type (Chain), Status (Success), and Other.

+

Agentic LLMs



What is “Agent”



An “intelligent” system that interacts with some “environment”

- Physical environments: autonomous car
- Digital environments: ChatGPT



AI Agent Example

ELIZA (1966): Symbolic AI agent

```
Welcome to
      EEEEEE  LL      IIII    ZZZZZZ  AAAAAA
      EE      LL      II      ZZ      AA      AA
      EEEEEE  LL      II      ZZZ     AAAAAAAA
      EE      LL      II      ZZ      AA      AA
      EEEEEE  LLLLLL  IIII    ZZZZZZ  AA      AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU: Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU: They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU: Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU: He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU: It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

- Domain specific!
- Requires manual design
- Cannot work beyond simple domains



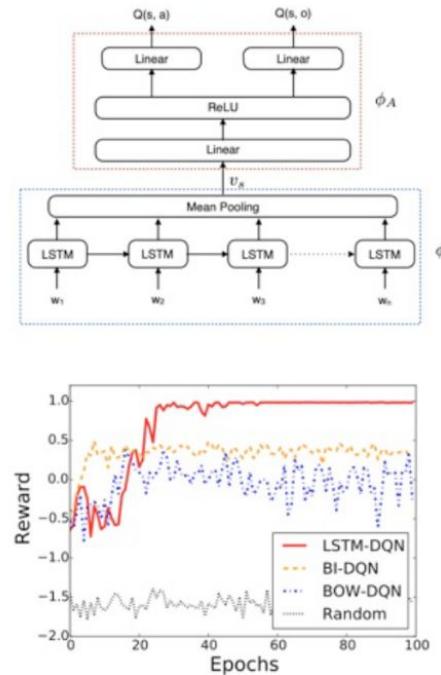
AI Agent Example

LSTM-DQN (2015): Deep RL agent

State 1: The old bridge
 You are standing very close to the bridge's eastern foundation. If you go east you will be back on solid ground ... The bridge sways in the wind.

Command: Go east

State 2: Ruined gatehouse
 The old gatehouse is near collapse. Part of its northern wall has already fallen down ... East of the gatehouse leads out to a small open area surrounded by the remains of the castle. There is also a standing archway offering passage to a path along the old southern inner wall.
 Exits: Standing archway, castle corner, Bridge over the abyss

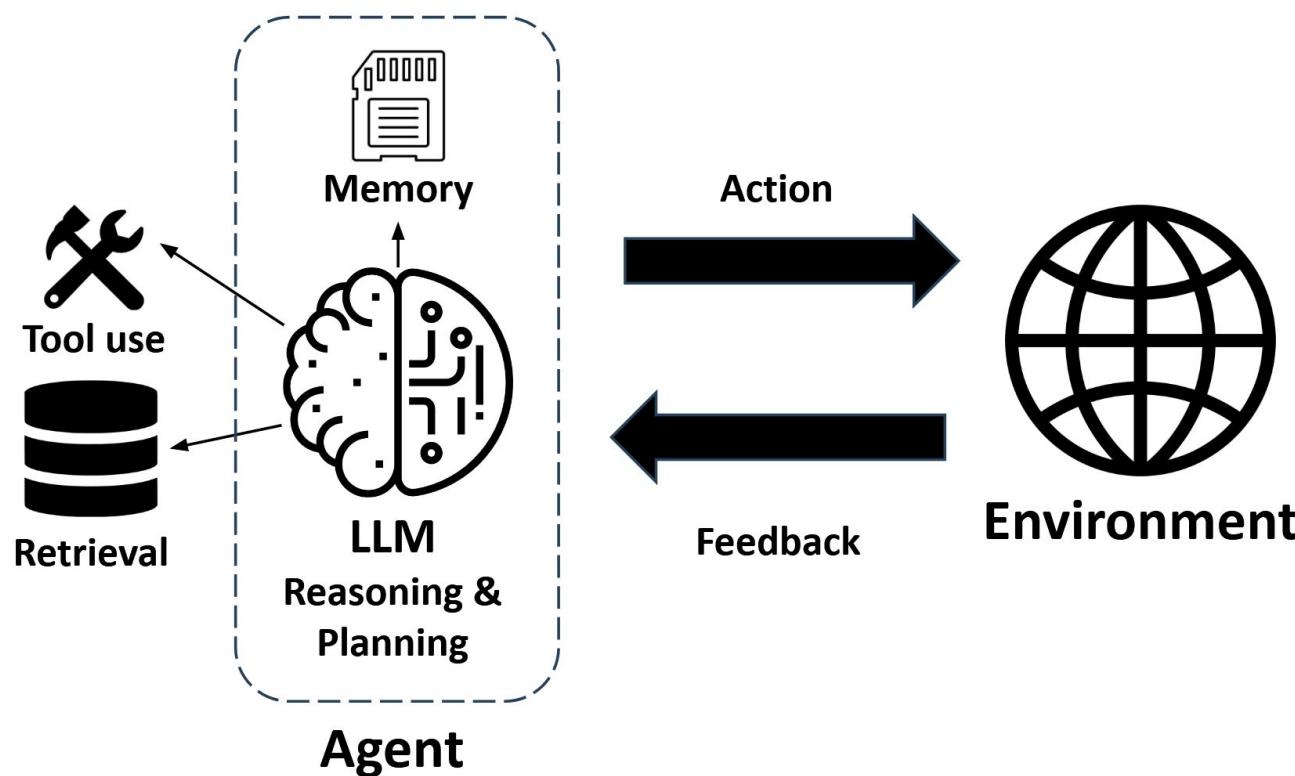


- Domain specific!
- Requires scalar reward signals
- Requires extensive training



Agentic LLMs

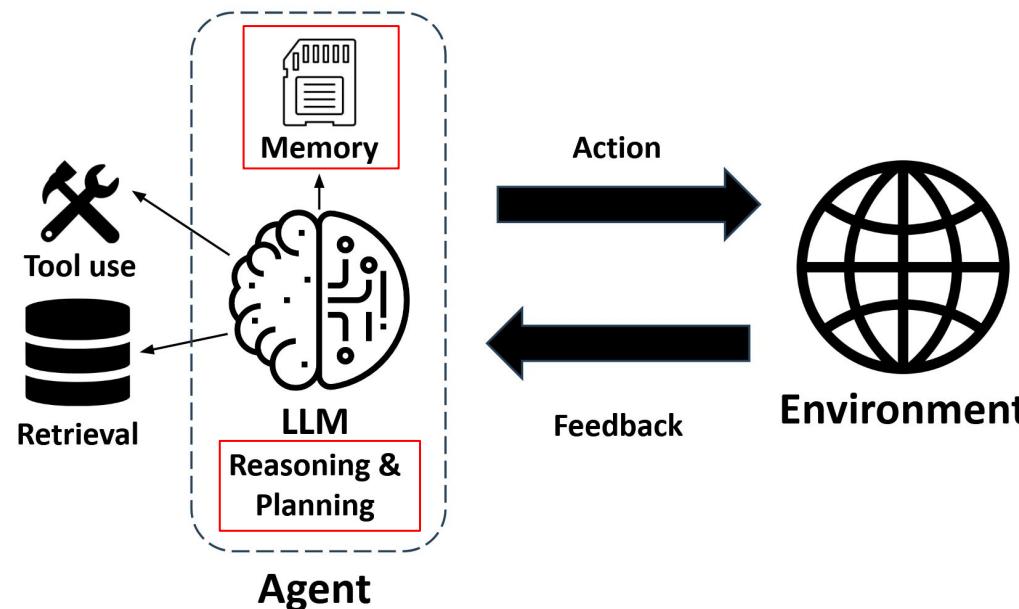
An agentic LLM is a large language model (LLM) that operates as an autonomous agent, capable of performing tasks, making decisions, and interacting with external tools or APIs to achieve the specified goals.





Benefit of Agentic LLM

- Enables LLMs to be able to solve complex real-world tasks through:
 - Task decomposition (planning)
 - Allocation of subtasks to specialized modules (tools, expert models, etc)
 - Multi-agent generation (improves responses quality)
- Expands LLMs' capabilities with external tools and knowledge.





Reasoning - Why do LLMs need reasoning?

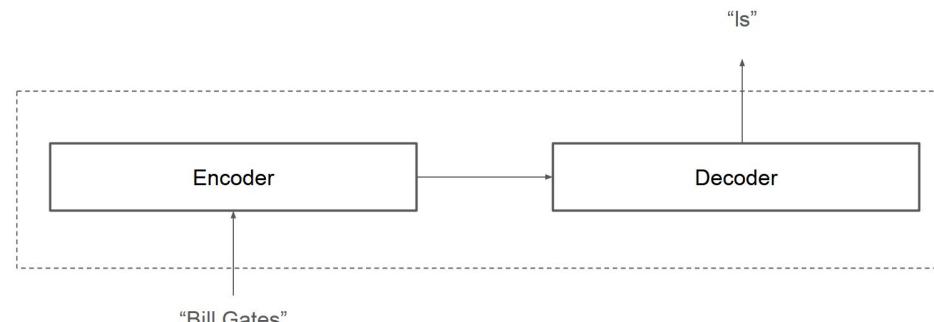
Humans can learn new tasks with only a handful of examples because **we can reason** [1].

Suppose we have this simple task: Last Letter Concatenation

Rule: Take the last letter of each word, and then concatenate them

| Input | Output |
|----------------|--------|
| “Elon Musk” | “nk” |
| “Bill Gates” | “ls” |
| “Barack Obama” | ? |

Solve it by ML? Tons of labeled data needed!

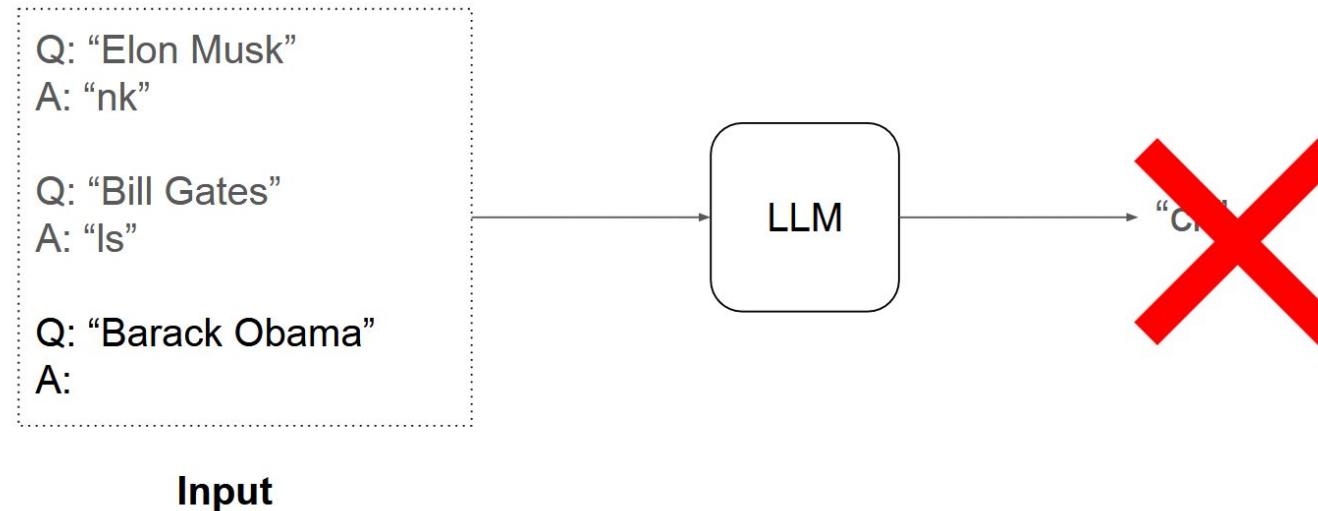




Reasoning - Why do LLMs need reasoning?

Simply prompting LLMs with few-shot examples probably won't work as well.

Few-shot prompting for last-letter-concatenation





Reasoning - Why do LLMs need reasoning?

However, giving the reasoning demonstration to the LLM allows it to complete the task easily.

One demonstration is enough, like humans

Q: "Elon Musk"

A: the last letter of "Elon" is "n". the last letter of "Musk" is "k". Concatenating "n", "k" leads to "nk". so the output is "nk".

Q: "Barack Obama"

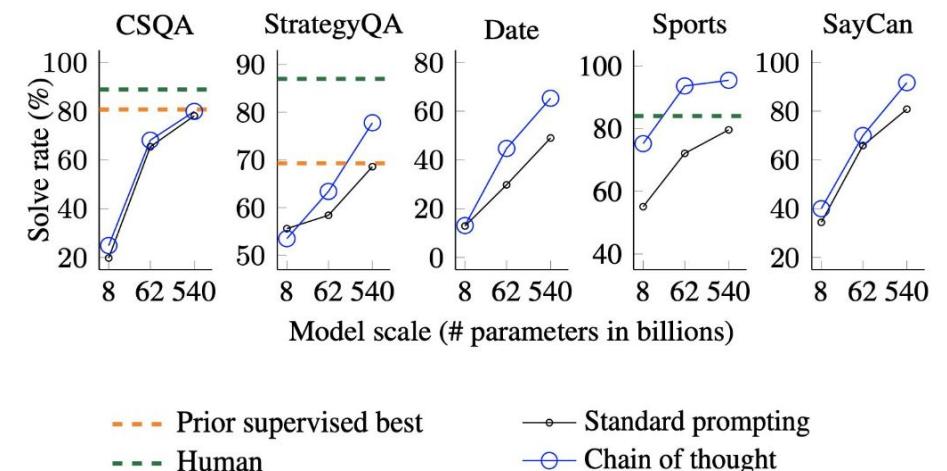
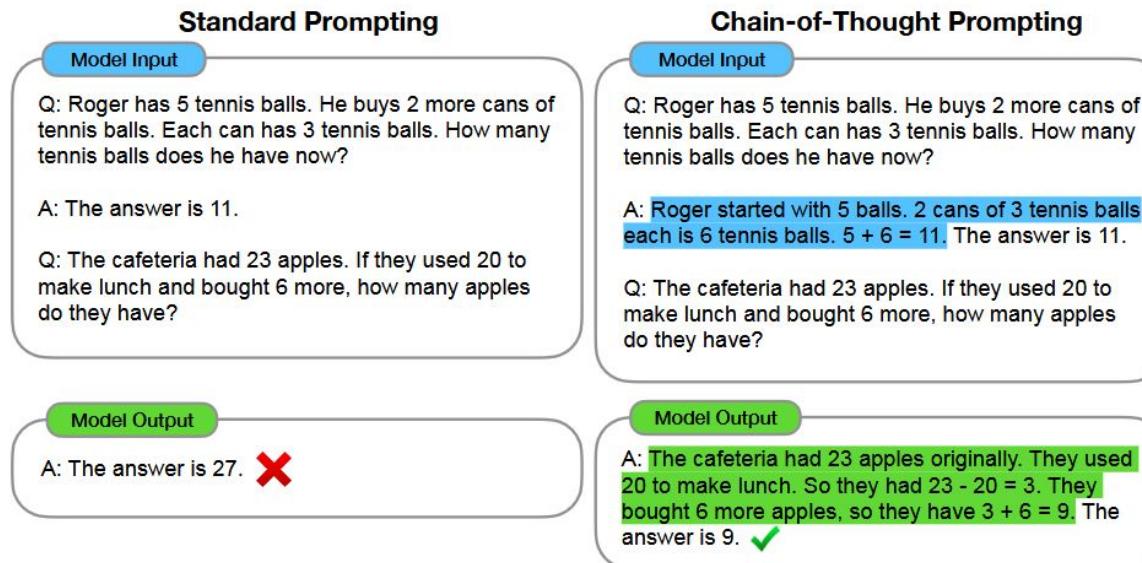
A: the last letter of "Barack" is "k". the last letter of "Obama" is "a". Concatenating "k", "a" leads to "ka". so the output is "ka".

100% accuracy with only one demonstration example



Reasoning - How do we achieve reasoning?

Key Idea: Derive the Final Answer through **Intermediate Steps**.





Reasoning - How do we achieve reasoning?

[1] proves, using the computational complexity theory, that CoT allows transformer to solve significantly more complex problems.

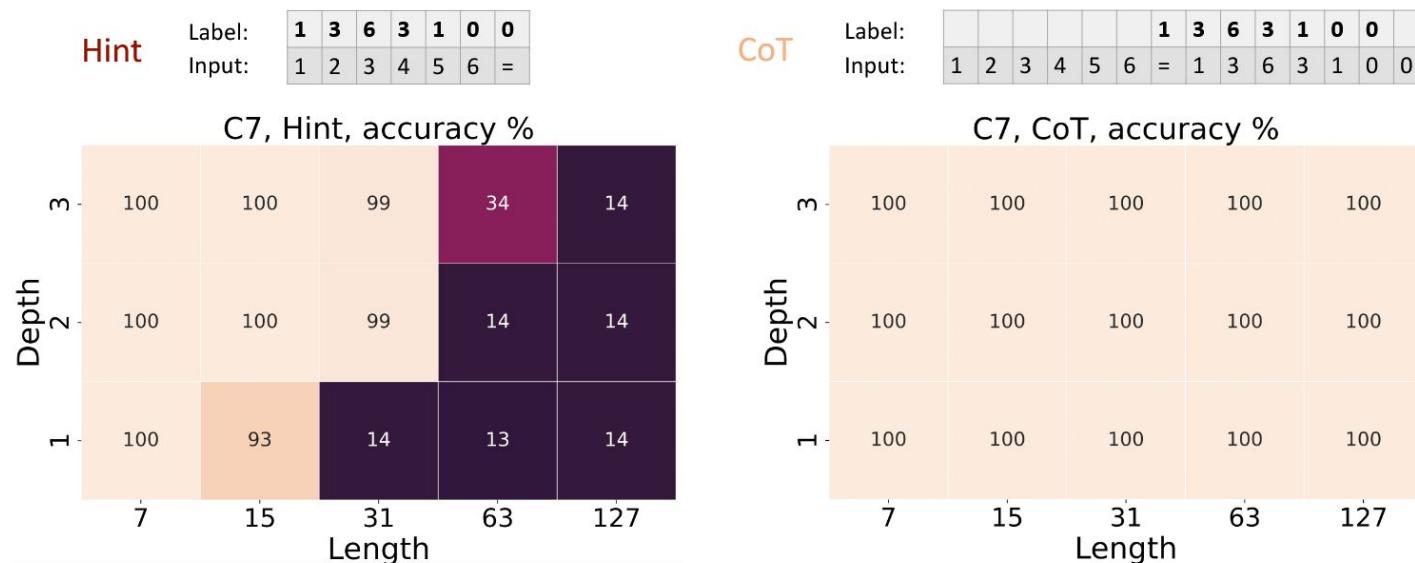
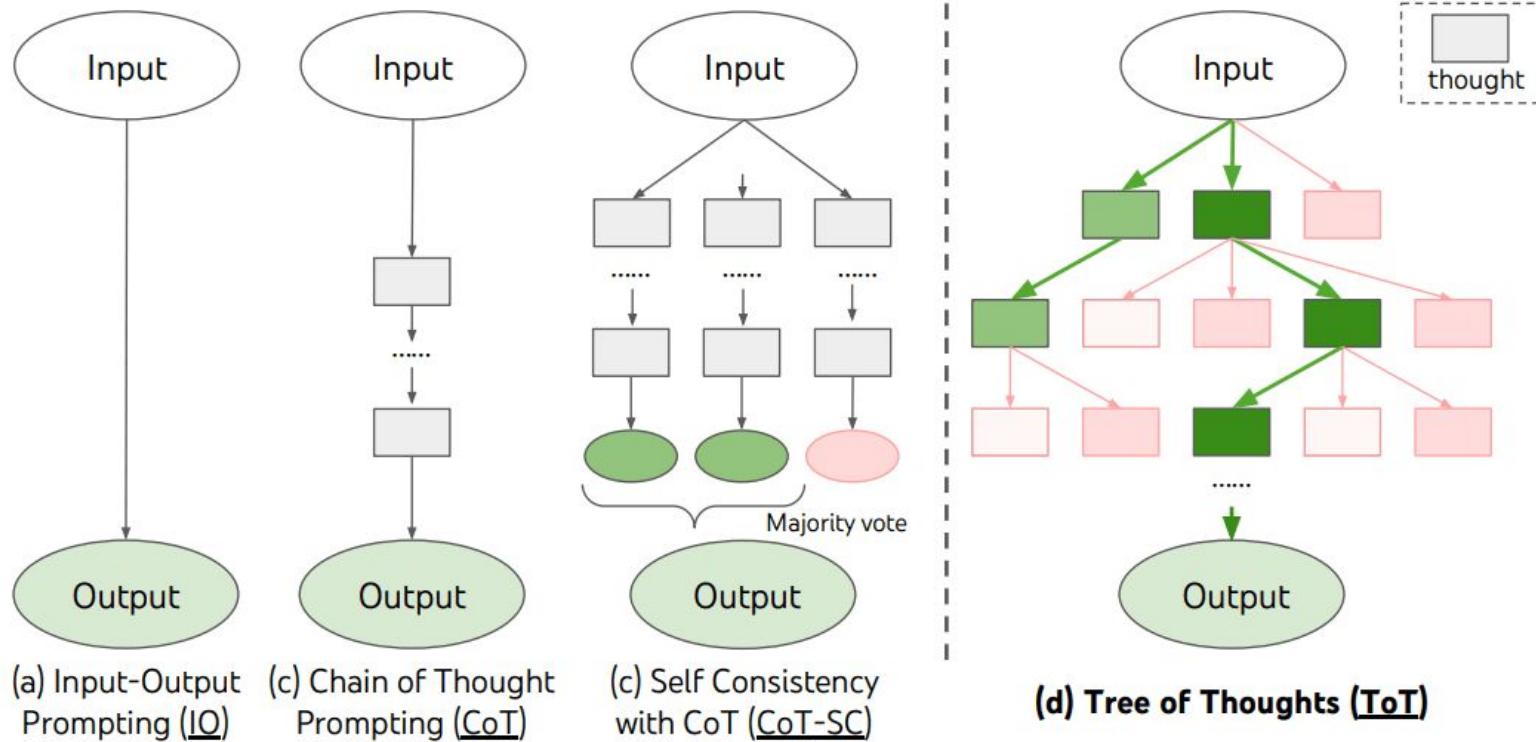


Figure 2: Modular Addition(C_7). The label is the sum of the inputs modulo a positive integer, which is 7 in this case. The chain of thoughts and hints are the partial modular sum. Low-depth transformers with `hint` can solve this task well for a reasonable input sequence length, but with `cot` the performance is much better, especially with a long input sequence, as predicted by our Theorem 3.3. See experiments for C_2 in Figure 5.



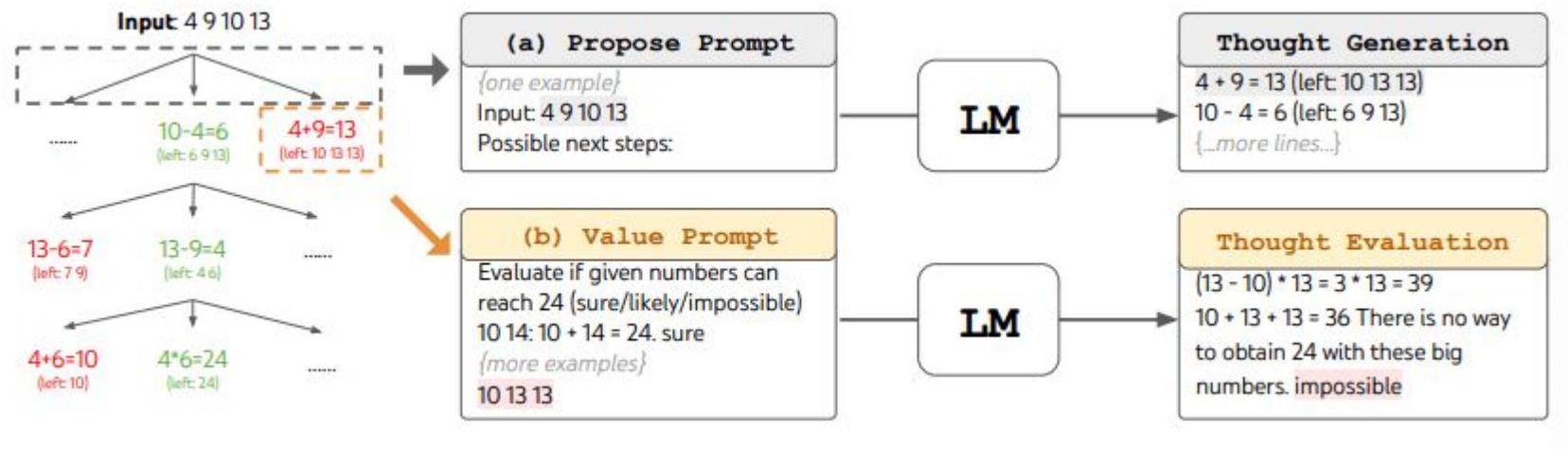
Reasoning - Tree-of-Thought Prompting

ToT maintains a tree of thoughts, where thoughts represent coherent language sequences that serve as intermediate steps toward solving a problem.





Reasoning - Tree-of-Thought Prompting



| Method | Success |
|--------------------|------------|
| IO prompt | 7.3% |
| CoT prompt | 4.0% |
| CoT-SC (k=100) | 9.0% |
| ToT (ours) (b=1) | 45% |
| ToT (ours) (b=5) | 74% |
| IO + Refine (k=10) | 27% |
| IO (best of 100) | 33% |
| CoT (best of 100) | 49% |

Table 2: Game of 24 Results.

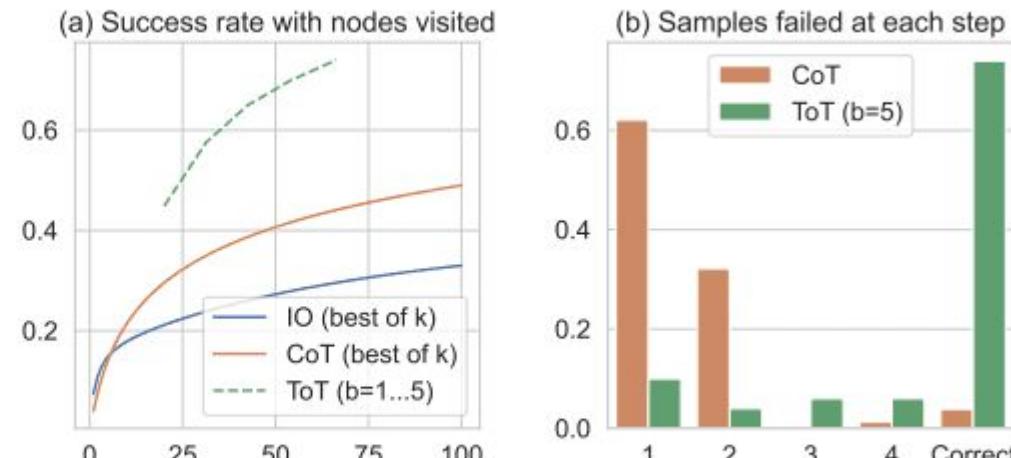


Figure 3: Game of 24 (a) scale analysis & (b) error analysis.

+ Reasoning - Problems

- LLMs are not trained for predicting reasoning paths but instead just the next token - not true for reasoning models!
- LLMs Can Be Easily Distracted by Irrelevant Context [1]
- LLMs Cannot Self-Correct Reasoning Yet [2]
- Premise Order Matters in LLM Reasoning [3]

Adding irrelevant contexts to GSM8K leads to 20+ points performance drop

[Prompt] Lucy has \$65 in the bank. She made a \$15 deposit and then followed by a \$4 withdrawal. **Maria's monthly rent is \$10.** What is Lucy's bank balance?

[Response] Lucy's bank balance is $\$65 + \$15 - \$4 - \$10 = \$66$. The answer is \$66.

Terry eats 2 yogurts a day. They are currently on sale at 4 yogurts for \$5.00. How much does he spend on yogurt over 30 days?

... So, Terry spends \$75 on yogurt over 30 days. (Correct)

Review your previous answer and find problems with your answer.

Upon reviewing my previous answer, I realized that I made a mistake in calculating Terry's spending on yogurt over 30 days. I incorrectly stated that Terry spends \$2.50 per day for 30 days...

Based on the problems you found, improve your answer.

... the final answer is Terry spends \$37.5 on yogurt over 30 days. (Wrong)

[Original GSM8K Problem] Thomas withdraws \$1000 in 20 dollar bills from the bank account. **He loses 10 bills while getting home.** After that, he uses half of the remaining bills to pay for a bill. Thomas then triples his money. He then converts all his bills to 5 dollar bills. How many 5 dollar bills does he have?

[Reordered GSM8K Problem] Thomas withdraws \$1000 in 20 dollar bills from the bank account. After **getting home**, he uses half of the remaining bills to pay for a bill. Thomas then triples his money. He then converts all his bills to 5 dollar bills. **He loses 10 bills while getting home.** How many 5 dollar bills does he have?

About 10 points drop on solving rates across all frontier LLMs

[1] <https://arxiv.org/abs/2302.00093>

[2] <https://arxiv.org/abs/2310.01798>

[3] <https://arxiv.org/abs/2402.08939>

+

Reasoning + Acting

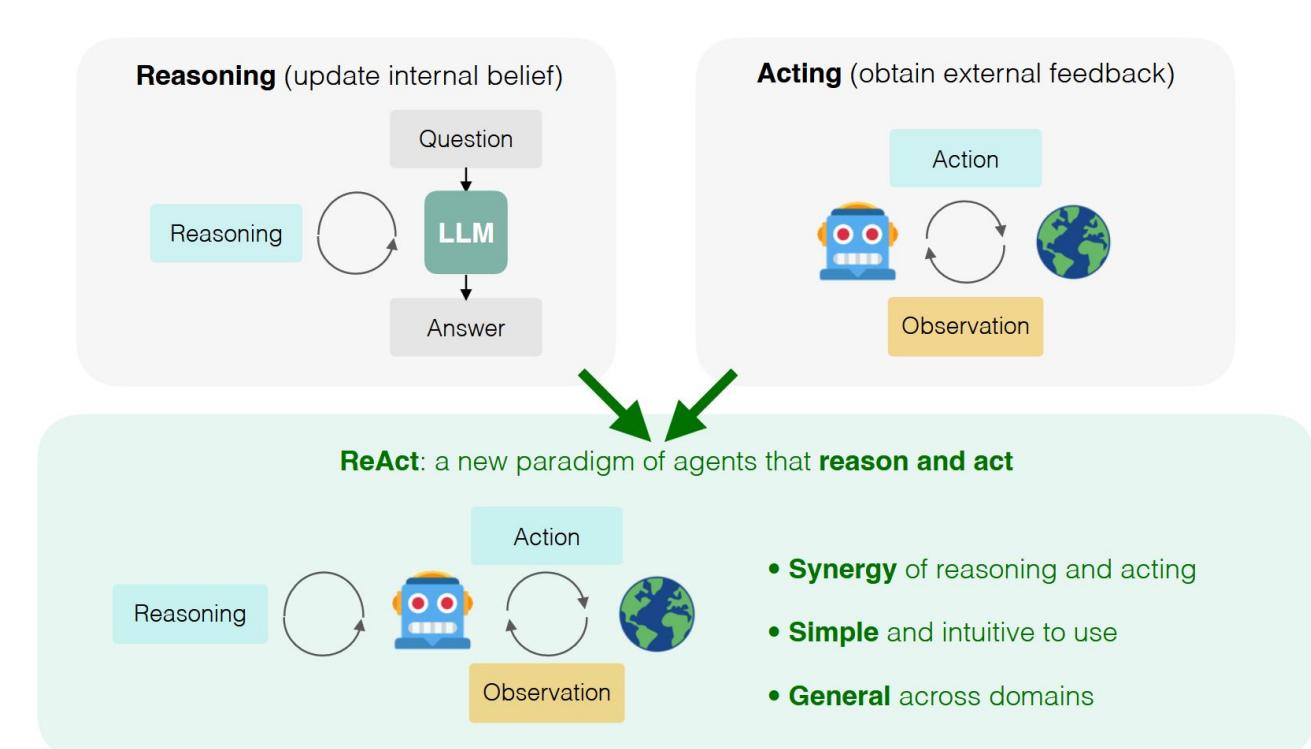


ReAct: Synergizing Reasoning and Acting in Language Models

ReAct = Reasoning + Acting

Allows LLMs perform dynamic reasoning to **create, maintain, and adjust plans** for acting while also enabling interaction to external environments.

Unlocks **agentic capability**.





ReAct Example

Question

Aside from the Apple Remote, what other devices can control the program Apple Remote was originally designed to interact with?

(1d) ReAct (Reason + Act)

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.

Act 1: `Search[Apple Remote]`

Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the **Front Row media center** program ...

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search **Front Row** next and find what other device can control it.

Act 2: `Search[Front Row]`

Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

Thought 3: Front Row is not found. I need to search **Front Row (software)** .

Act 3: `Search[Front Row (software)]`

Obs 3: Front Row is a discontinued media center software ...

Thought 4: **Front Row (software)** is controlled by an **Apple Remote** or the keyboard function keys. So the answer is keyboard function keys.

Act 4: `Finish[keyboard function keys]`





ReAct Benefit

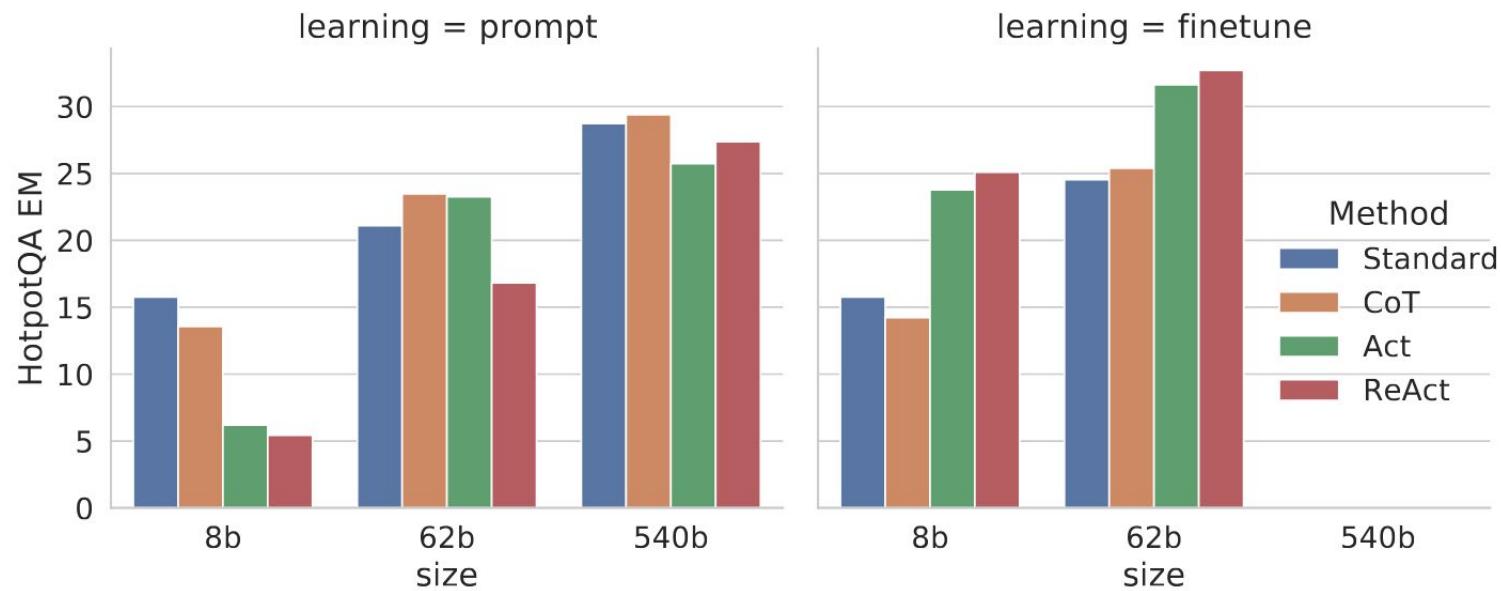


Figure 3: Scaling results for prompting and finetuning on HotPotQA with ReAct (ours) and baselines.

| | Type | Definition | ReAct | CoT |
|---------|---------------------|--|-------|-----|
| Success | True positive | Correct reasoning trace and facts | 94% | 86% |
| | False positive | Hallucinated reasoning trace or facts | 6% | 14% |
| Failure | Reasoning error | Wrong reasoning trace (including failing to recover from repetitive steps) | 47% | 16% |
| | Search result error | Search return empty or does not contain useful information | 23% | - |
| | Hallucination | Hallucinated reasoning trace or facts | 0% | 56% |
| | Label ambiguity | Right prediction but did not match the label precisely | 29% | 28% |

ReAct

Pretty much that standard on which AI agents nowadays operate.

How it works

OpenAI Deep Research

Deep research was trained using end-to-end reinforcement learning on hard browsing and reasoning tasks across a range of domains. Through that training, it learned to plan and execute a multi-step trajectory to find the data it needs, backtracking and reacting to real-time information where necessary. The model is also able to browse over user uploaded files, plot and iterate on graphs using the python tool, embed both generated graphs and images from websites in its responses, and cite specific sentences or passages from its sources. As a result of this training, it reaches new highs on a number of public evaluations focused on real-world problems.

Open Deep Research (Huggingface)

Instruction: Determine the most cost-effective country to purchase the smartphone model "CodeAct 1". The countries to consider are the USA, Japan, Germany, and India.

Available APIs

```
[1] lookup_rates(country: str) -> (float, float)
[2] convert_and_tax(price: float, exchange_rate: float, tax_rate: float) -> float
[3] estimate_final_price(converted_price: float, shipping_cost: float) -> float
[4] lookup_phone_price(model: str, country: str) -> float
[5] estimate_shipping_cost(destination_country: str) -> float
```

LLM Agent using [Text/JSON] as Action

Think I should calculate the phone price in USD for each country, then find the most cost-effective country.

Action Text: `lookup_rates, Germany`
JSON: `{"tool": "lookup_rates", "country": "Germany"}`

Environment 1.1, 0.19

Action Text: `lookup_phone_price, CodeAct 1, Germany`
JSON: `{"tool": "lookup_phone_price", "model": "CodeAct 1", "country": "Germany"}`

Environment 700

Action Text: `convert_and_tax, 700, 1.1, 0.19`
JSON: `{"tool": "convert_and_tax", "price": 700, "exchange_rate": 1.1, "tax_rate": 0.19}`

Environment 916.3

[... interactions omitted (look up shipping cost and calculate final price) ...]

Action Text: `lookup_rates, Japan`
JSON: `{"tool": "lookup_rates", "country": "Japan"}`

[... interactions omitted (calculate final price for all other countries)...]

Response The most cost-effective country to purchase the smartphone model is Japan with price 904.00 in USD.

CodeAct: LLM Agent using [Code] as Action

Think I should calculate the phone price in USD for each country, then find the most cost-effective country.

Action `countries = ['USA', 'Japan', 'Germany', 'India']
final_prices = {}

for country in countries:
 exchange_rate, tax_rate = lookup_rates(country)
 local_price = lookup_phone_price("xAct 1", country)
 converted_price = convert_and_tax(
 local_price, exchange_rate, tax_rate
)
 shipping_cost = estimate_shipping_cost(country)
 final_price = estimate_final_price(converted_price, shipping_cost)
 final_prices[country] = final_price`

Control & Data Flow of Code
Simplifies Complex Operations

`most_cost_effective_country = min(final_prices, key=final_prices.get)
most_cost_effective_price = final_prices[most_cost_effective_country]
print(most_cost_effective_country, most_cost_effective_price)`

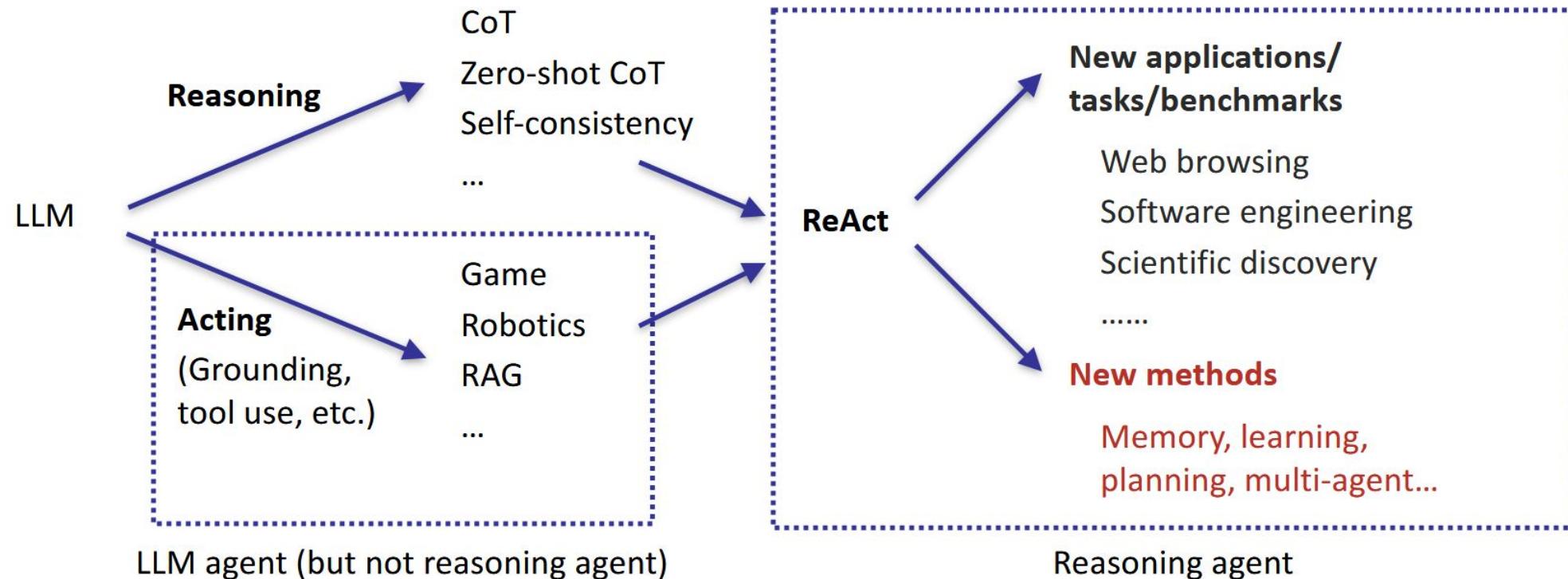
Re-use `min` Function from Existing
Software Infrastructures (Python library)

Environment 1.1, 0.19

Response The most cost-effective country to purchase the smartphone model is Japan with price 904.00 in USD.

Fewer Actions Required!

A brief history of LLM agents



Long-term memory

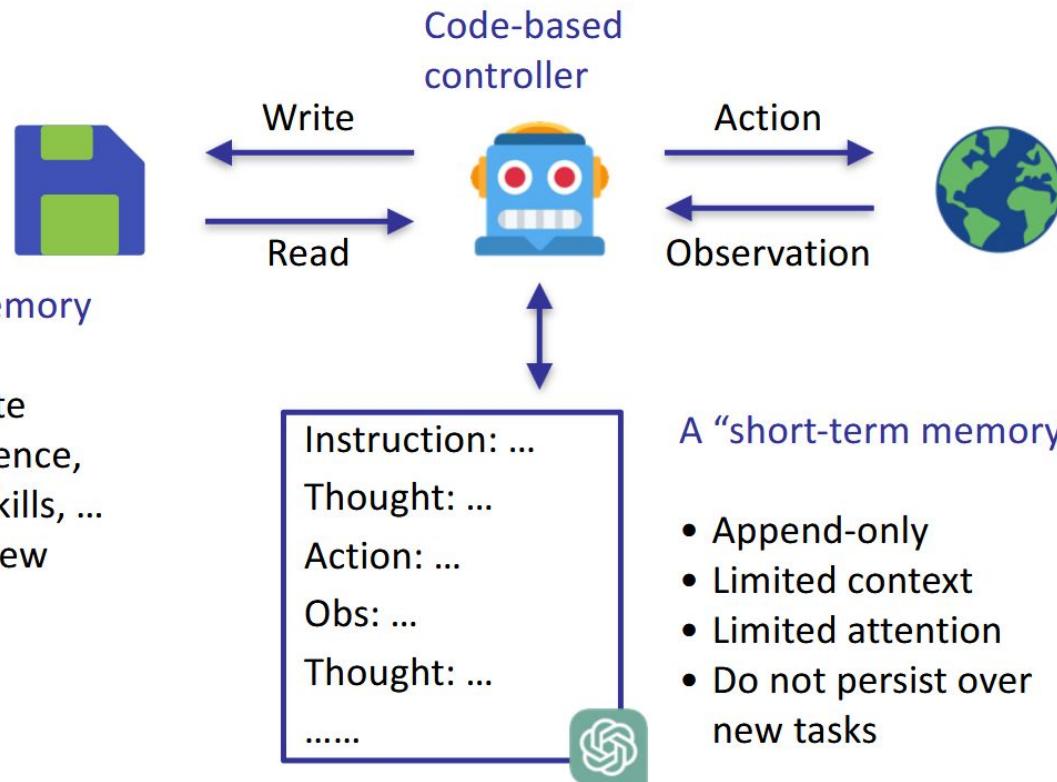
As opposed to short-term memory e.g. in context

- Text based like diary
- Code based
- Episodic memory - like log stream

Long-term memory persists over new tasks, essentially making LLMs look like they are learning



Long-term memory



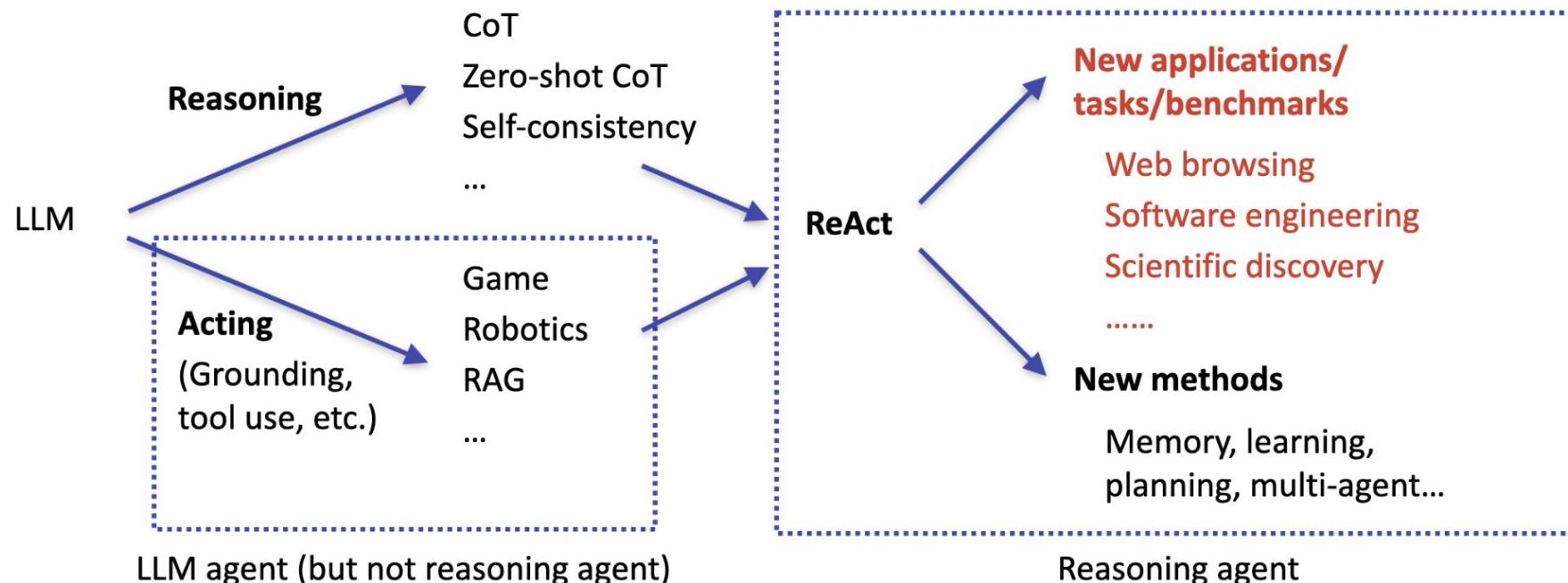


Real-world Applications



Agentic LLMs: Applications

A brief history of LLM agents



Digital automation



File reports



Code experiments

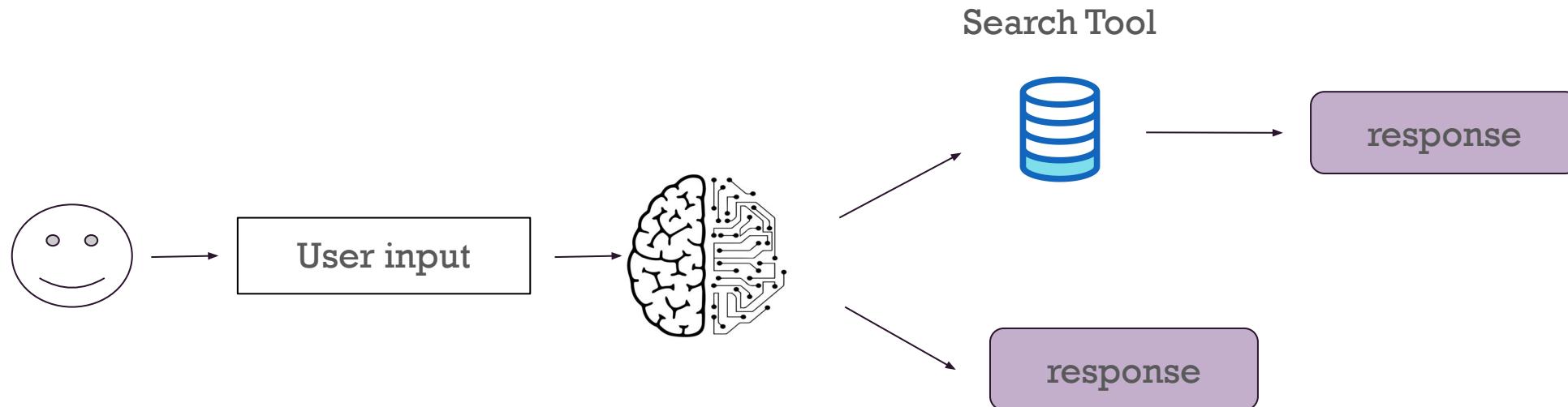


Explore papers

- Tremendous practical values, but little progress (think about Siri)
- Underlying research challenges:
 - Reasoning over **real-world language** (and other modalities)
 - Decision making over **open-ended actions** and **long horizon**

Example 1: Agentic RAG

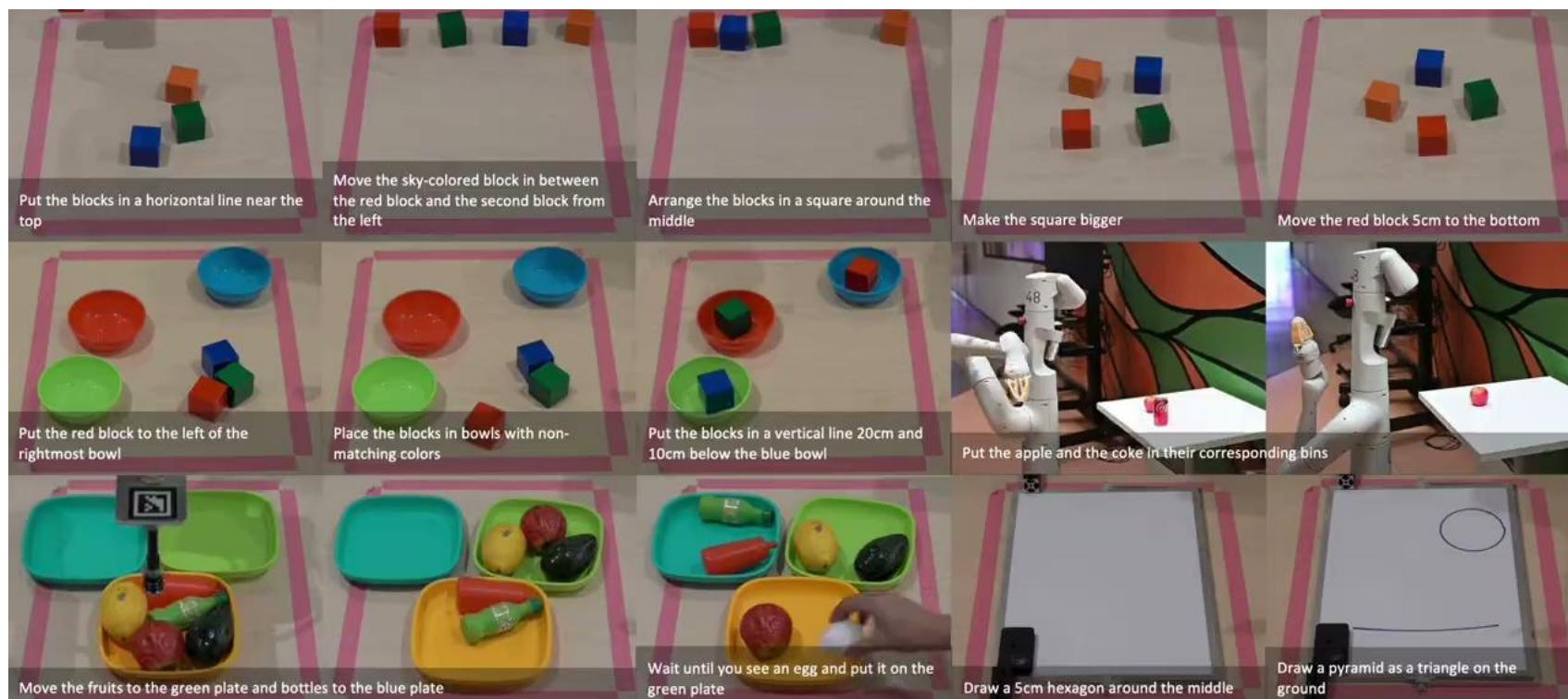
- Not all messages need supporting documents for LLMs to respond.
- Instead, give the LLM a vector search “tool” (also sometimes referred to as function calling).
- Now, it can automatically decide when to use the tool.





Example 2: Robotics

- Give an LLM access to functions that control robots such as move, grab, etc.
- The LLM will decide which functions are required (and when to use each) for each query.



Example 3: Multi-agent

- Why stop at just a single AI? We can also use multiple of them to complete a task.



```
> /code Add a function to calculate the factorial of a number
Certainly! I'll add a function to calculate the factorial of a number.
Here's the change:

def factorial(n):
    if n == 0 or n == 1:
        return 1
    else:
        return n * factorial(n - 1)

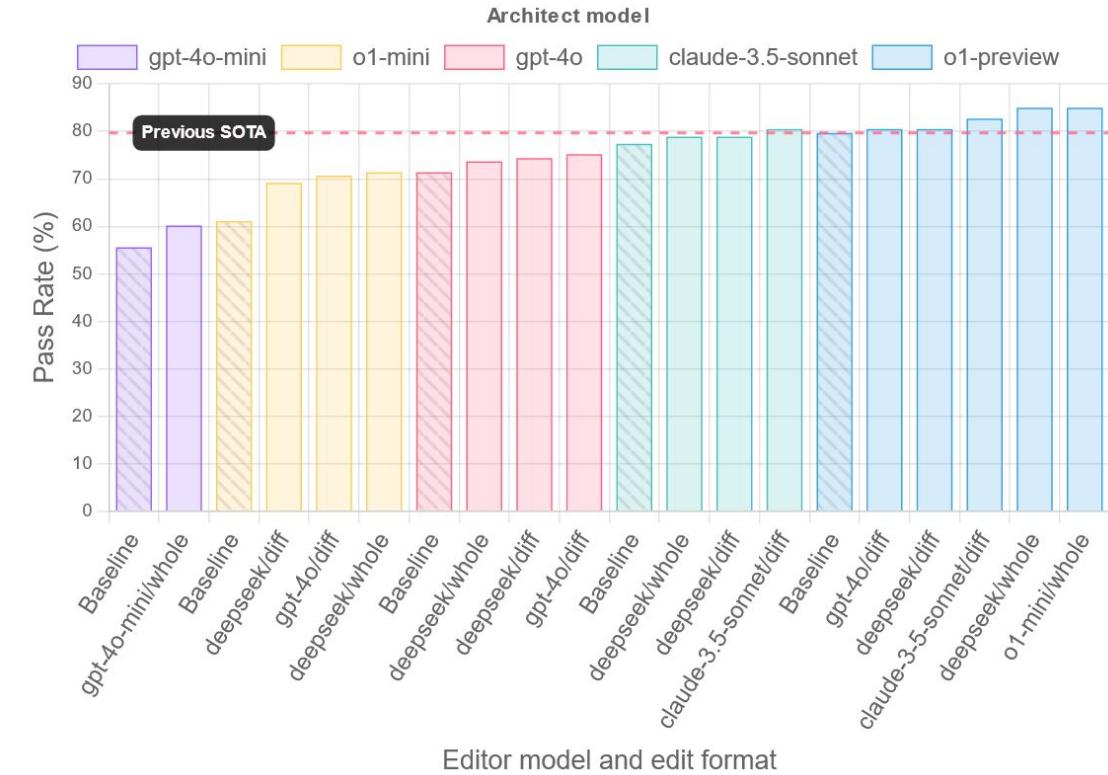
# Example usage
print(factorial(5)) # Output: 120
```

This function uses recursion to calculate the factorial. It's been added to your code.

```
> /architect Can we make this simpler?
Yes, you could replace the factorial() function with a call to the
math.factorial() builtin. This way you don't have to maintain a custom
factorial implementation, and the builtin function is well optimized.
```

Edit the files? (Y)es/(N)o [Yes]: Yes

```
<<<<< SEARCH
def factorial(n):
    if n == 0 or n == 1:
        return 1
    else:
        return n * factorial(n - 1)
=====
from math import factorial
>>>>> REPLACE
```





Tools & Framework

Agentic



Monitoring





Challenges

Challenges for LLM agent deployment in the wild

- Reasoning and planning
 - LLM agents tend to make mistakes when performing complex tasks end-to-end
- Embodiment and learning from environment feedback
 - LLM agents are not yet efficient at recovering from mistakes for long-horizon tasks
 - Continuous learning, self-improvement
 - Multimodal understanding, grounding and world models
- Multi-agent learning, theory of mind
- Safety and privacy
 - LLMs are susceptible to adversarial attacks, can emit harmful messages and leak private data
- Human-agent interaction, ethics
 - How to effectively control the LLM agent behavior, and design the interaction mode between humans and LLM agents



Overall LLM Challenges

Air Canada has been held liable for a negligent misrepresentation made to a customer by one of its chatbots in a case that one expert said highlights broader risks businesses must consider when adopting AI tools.

According to Moffat's screenshot of a conversation with the chatbot, the British Columbia resident was told he could apply for the refund "within 90 days of the date your ticket was issued" by completing an online form.

Moffatt then booked tickets to and from Toronto to attend the funeral of a family member. But when he applied for a refund, Air Canada said bereavement rates did not apply to completed travel and pointed to the bereavement section of the company's website.



Evaluation Tools/Benchmark for Agentic LLMs

SWE bench

τ -bench

WebArena

GAIA

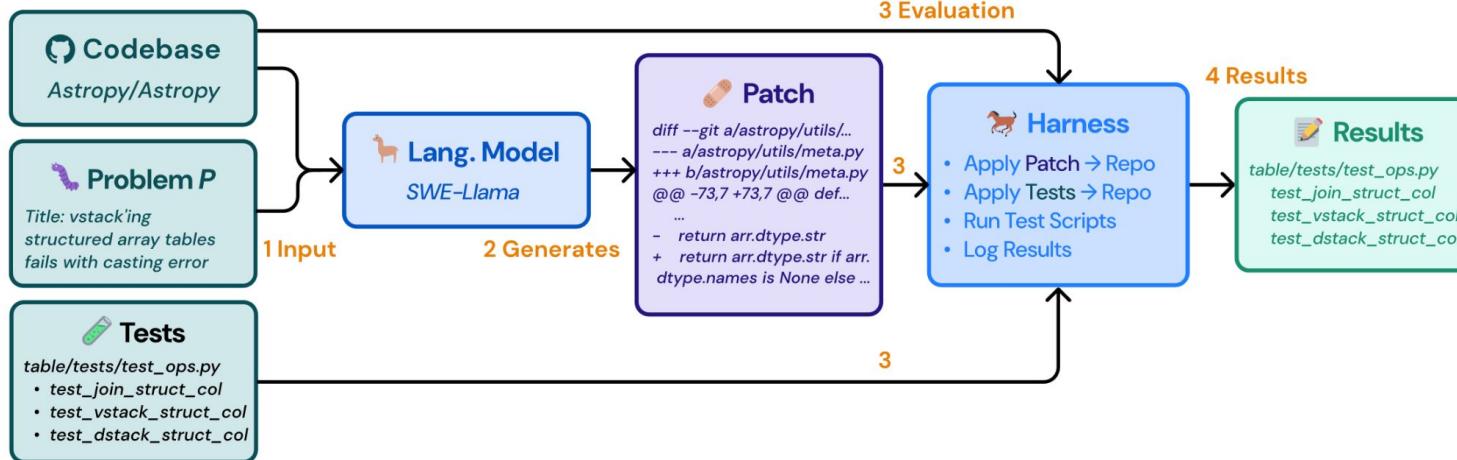


SWE-bench

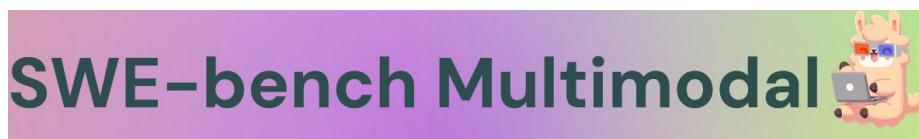


Each sample in the SWE-bench test set is created from a resolved GitHub issue in one of 12 open-source Python repositories on GitHub, such as scikit-learn.

SWE-Bench Task



Also a multimodal version!



Select sample: sympy_sympy-19637

Commentary

This is an example of a good sample which has been verified by annotators for the SWE-bench Verified dataset. The problem statement gives a short but clear demonstration of a bug, and the FAIL_TO_PASS tests directly assert that the example given in the problem statement has been resolved.

Problem statement

```

Unset
kernS: 'kern' referenced before assignment
from sympy.core.sympify import kernS

text = "(2*x)/(x-1)"
expr = kernS(text)
// hit = kern in s
// UnboundLocalError: local variable 'kern' referenced before assignment

```

Are the tasks well-specified? (Raw annotation)

Severity: 0 - The issue is well-specified and it is clear what is required for a successful solution.

It is clear that `kernS` is throwing exception for `(2*x)/(x-1)`. It provides example input for which the error is occurring which can make it easy to reproduce the issue.

FAIL_TO_PASS test (Only showing lines added during the original PR for brevity)

```

Python
def test_kernS():
    ...
    assert kernS("(2*x)/(x-1)") == 2*x/(x-1)

```

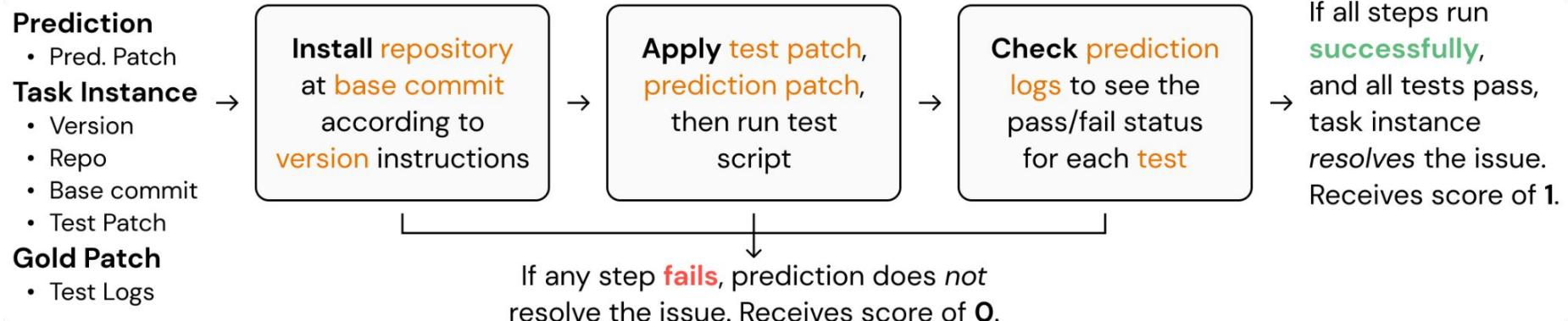
How valid are the evaluation criteria? (Raw annotation)

Severity: 0 - The tests perfectly cover all possible solutions.

The test case is exactly for `kernS("(2*x)/(x-1)")` for which the issue was occurring in issue description. It will cover all possible solutions.



SWE-bench



Leaderboard

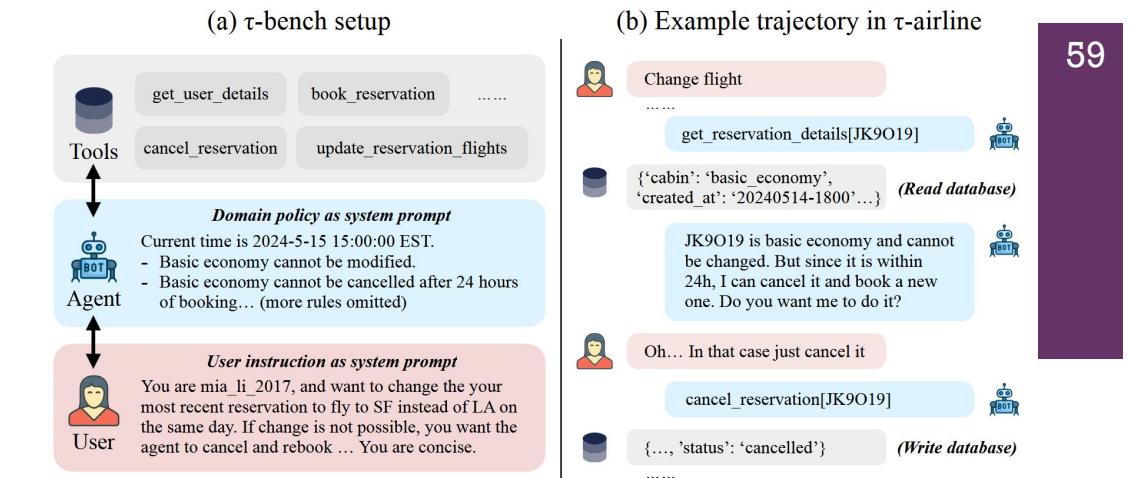
| Model | % Resolved | Org | Date | Logs | Trajs | Site |
|--|------------|------|------------|------|-------|------|
| 👉👩‍💻 OpenHands + CodeAct v2.1 (claude-3-5-sonnet-20241022) | 29.38 | 👉👩‍💻 | 2024-11-03 | ✓ | ✓ | 🔗 |
| 👉👩 AutoCodeRover-v2.0 (Claude-3.5-Sonnet-20241022) | 24.89 | 👉👩 | 2024-11-21 | ✓ | ✓ | 🔗 |
| 👉 Honeycomb | 22.06 | 👉 | 2024-08-20 | ✓ | ✓ | 🔗 |
| Amazon Q Developer Agent (v20240719-dev) | 19.75 | aws | 2024-07-21 | ✓ | ✓ | 🔗 |
| Factory Code Droid | 19.27 | 👉 | 2024-06-17 | ✓ | - | 🔗 |
| AutoCodeRover (v20240620) + GPT 4o (2024-05-13) | 18.83 | 👉👩 | 2024-06-28 | ✓ | - | 🔗 |
| 👉👩 SWE-agent + Claude 3.5 Sonnet | 18.13 | 👉👩 | 2024-06-20 | ✓ | ✓ | - |
| 👉👩 SWE-agent + GPT 4o (2024-05-13) | 14.60 | 👉 | 2024-06-15 | ✓ | - | 🔗 |
| Amazon Q Developer Agent (v20240430-dev) | 13.82 | aws | 2024-05-09 | ✓ | - | 🔗 |
| 👉👩 SWE-agent + GPT 4 (1106) | 12.47 | 👉👩 | 2024-04-02 | ✓ | ✓ | 🔗 |
| 👉👩 SWE-agent + GPT 4o (2024-05-13) | 11.99 | 👉👩 | 2024-07-28 | ✓ | ✓ | 🔗 |
| 👉👩 SWE-agent + Claude 3 Opus | 10.51 | 👉👩 | 2024-04-02 | ✓ | ✓ | - |
| 👉👩 RAG + Claude 3 Opus | 3.79 | 👉👩 | 2024-04-02 | ✓ | - | 🔗 |
| 👉👩 RAG + Claude 2 | 1.96 | 👉👩 | 2023-10-10 | ✓ | - | - |
| 👉👩 RAG + GPT 4 (1106) | 1.31 | 👉👩 | 2024-04-02 | ✓ | - | - |
| 👉👩 RAG + SWE-Llama 13B | 0.70 | 👉👩 | 2023-10-10 | ✓ | - | - |
| 👉👩 RAG + SWE-Llama 7B | 0.70 | 👉👩 | 2023-10-10 | ✓ | - | - |
| 👉👩 RAG + ChatGPT 3.5 | 0.17 | 👉👩 | 2023-10-10 | ✓ | - | - |



τ -bench

TAU-bench: Tool-Agent-User benchmark

- Key Characteristics:
 - Realistic dialogue and tool use
 - Open-ended and diverse tasks
 - Faithful rule-based evaluation
 - Modular extension
- The reward of a task episode is based on:
 - whether the final database is identical to the unique ground truth outcome database (r_{action}), and
 - whether the agent's responses to the user contain all necessary information (r_{output}).



Return delivered order
- After user confirmation, the order status will be changed to 'return requested'...

Exchange delivered order
- An order can only be exchanged if its status is 'delivered'...

(c) Domain policy excerpts in τ -retail.

```
{"order_id": "#W2890441",  
 "user_id": "mei_davis_8935",  
 "items": [  
   {"name": "Water Bottle",  
    "product_id": "8310926033",  
    "item_id": "2366567022",  
    "price": 54.04,  
    "options": {  
      "capacity": "1000ml",  
      "material": "stainless  
      steel",  
      "color": "blue"  
    }, ..., ...}]}
```

(a) An orders database entry in τ -retail.

```
def return_delivered_order_items(  
    order_id: str,  
    item_ids: List[str],  
    payment_method_id: str,  
) -> str: ...
```

```
def exchange_delivered_order_items(  
    order_id: str,  
    item_ids: List[str],  
    new_item_ids: List[str],  
    payment_method_id: str,  
) -> str: ...
```

(b) An API tool in τ -retail.

(d) User instruction ensures only one possible outcome.

τ -bench

Leaderboard

Airline

| Strategy | Pass^1 | Pass^2 | Pass^3 | Pass^4 |
|---|--------------|--------------|--------------|--------------|
| TC (claude-3-5-sonnet-20241022) | 0.460 | 0.326 | 0.263 | 0.225 |
| TC (gpt-4o) | 0.420 | 0.273 | 0.220 | 0.200 |
| TC (claude-3-5-sonnet-20240620) | 0.360 | 0.224 | 0.169 | 0.139 |
| TC (mistral-large-2407) | ?? | ?? | ?? | ?? |
| TC (gpt-4o-mini) | 0.225 | 0.140 | 0.110 | 0.100 |
| Act (gpt-4o) | 0.365 | 0.217 | 0.160 | 0.140 |
| ReAct (gpt-4o) | 0.325 | 0.233 | 0.185 | 0.160 |

Retail

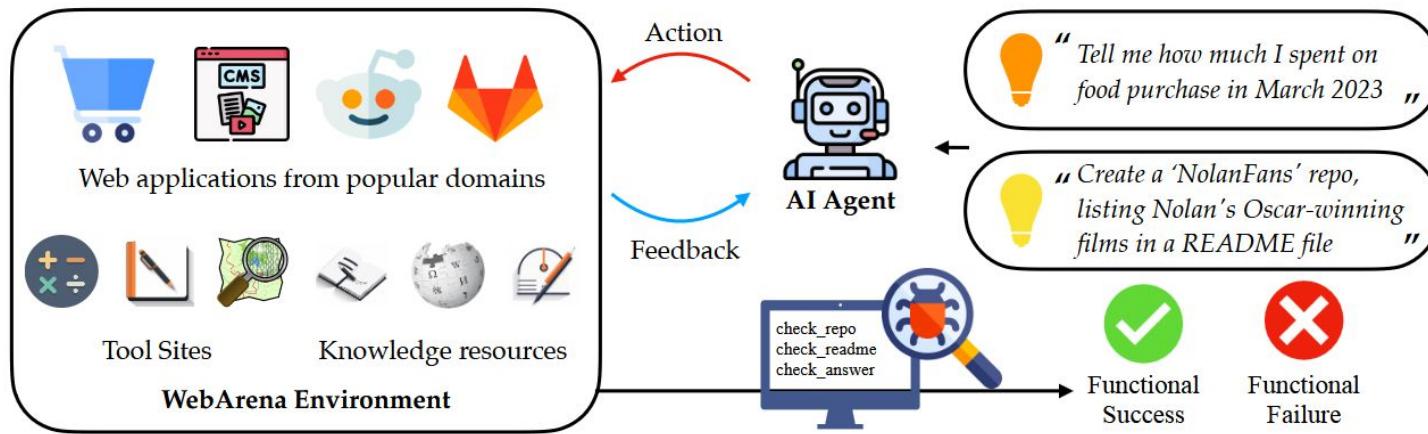
| Strategy | Pass^1 | Pass^2 | Pass^3 | Pass^4 |
|---|--------------|--------------|--------------|--------------|
| TC (claude-3-5-sonnet-20241022) | 0.692 | 0.576 | 0.509 | 0.462 |
| TC (gpt-4o) | 0.604 | 0.491 | 0.430 | 0.383 |
| TC (claude-3-5-sonnet-20240620) | 0.626 | 0.506 | 0.435 | 0.387 |
| TC (mistral-large-2407) | ?? | ?? | ?? | ?? |
| TC (gpt-4o-mini) | ?? | ?? | ?? | ?? |
| Act (gpt-4o) | ?? | ?? | ?? | ?? |
| ReAct (gpt-4o) | ?? | ?? | ?? | ?? |

*TC = `tool-calling` strategy (the function-calling strategy reported in the paper)



WebArena

WebArena is a standalone, self-hostable web environment for building autonomous agents.



| Action Type | Description |
|-------------------|---------------------------|
| noop | Do nothing |
| click (elem) | Click at an element |
| hover (elem) | Hover on an element |
| type (elem, text) | Type to an element |
| press (key_comb) | Press a key comb |
| scroll (dir) | Scroll up and down |
| tab_focus (index) | focus on <i>i</i> -th tab |
| new_tab | Open a new tab |
| tab_close | Close current tab |
| go_back | Visit the last URL |
| go_forward | Undo go_back |
| goto (URL) | Go to URL |

| Category | Example |
|---------------------|---|
| Information Seeking | When was the last time I bought shampoo |
| | Compare walking and driving time from AMC Waterfront to Randyland |
| Site Navigation | Checkout merge requests assigned to me |
| | Show me the ergonomic chair with the best rating |
| Content & Config | Post to ask “whether I need a car in NYC” |
| | Delete the reviews from the scammer Yoke |

Figure 4: Action Space of WebArena

The figure shows three screenshots of the WebArena interface:

- Screenshot 1:** A screenshot of a browser window titled "webarena.wikipedia.com" showing a search for "Pittsburgh museums". The results list various museums in Pittsburgh, Pennsylvania.
- Screenshot 2:** A screenshot of a browser window titled "webarena.openstreetmap.com" showing a map of Pittsburgh with a route plan from Carnegie Mellon University to The Andy Warhol Museum. It includes a sidebar with directions and a map.
- Screenshot 3:** A screenshot of a browser window titled "webarena.gitlab.com" showing a GitHub repository named "Travel in Northeast US". The repository page lists several commits, including one for "Pittsburgh" which includes a link to the museum search results.



WebArena

| Agent | WebArena |
|--|----------|
| GenericAgent-Claude-3.5-Sonnet | 36.20 |
| GenericAgent-GPT-4o | 31.40 |
| GenericAgent-GPT-o1-mini | 28.60 |
| GenericAgent-Llama-3.1-405b | 24.00 |
| GenericAgent-Llama-3.1-70b | 18.40 |
| GenericAgent-GPT-4o-mini | 17.40 |

| Function | ID | Intent | Eval Implementation |
|---------------------------------|----|--|---|
| $r_{\text{info}}(a^*, \hat{a})$ | 1 | Tell me the name of the customer who has the most cancellations in the history | <code>exact_match(\hat{a}, "Samantha Jones")</code> |
| | 2 | Find the customer name and email with phone number 8015551212 | <code>must_include(\hat{a}, "Sean Miller")</code> <code>must_include(\hat{a}, "sean@gmail.com")</code> |
| | 3 | Compare walking and driving time from AMC Waterfront to Randyland | <code>fuzzy_match(\hat{a}, "walking: 2h58min")</code> <code>fuzzy_match(\hat{a}, "driving: 21min")</code> |
| $r_{\text{prog}}(\mathbf{s})$ | 4 | Checkout merge requests assigned to me | <code>url=locate_current_url(s)</code> <code>exact_match(URL, "gitlab.com/merge_requests?assignee_username=byteblaze")</code> |
| | 5 | Post to ask "whether I need a car in NYC" | <code>url=locate_latest_post_url(s)</code> <code>body=locate_latest_post_body(s)</code> <code>must_include(URL, "/f/nyc")</code> <code>must_include(body, "a car in NYC")</code> |

Table 1: We introduce two evaluation approaches. r_{info} (top) measures the correctness of performing information-seeking tasks. It compares the predicted answer \hat{a} with the annotated reference a^* with three implementations. r_{prog} (bottom) programmatically checks whether the intermediate states during the executions possess the anticipated properties specified by the intent.



GAIA: A Benchmark for General AI Assistants

Real-world questions that require a set of fundamental abilities such as reasoning, multi-modality handling, web browsing, and generally tool-use proficiency.

GAIA questions are conceptually simple for humans yet challenging for most advanced AIs

Level 1

Question: What was the actual enrollment count of the clinical trial on H. pylori in acne vulgaris patients from Jan-May 2018 as listed on the NIH website?

Ground truth: 90



Level 2

Question: If this whole pint is made up of ice cream, how many percent above or below the US federal standards for butterfat content is it when using the standards as reported by Wikipedia in 2020? Answer as + or - a number rounded to one decimal place.

Ground truth: +4.6

Level 3

Question: In NASA's Astronomy Picture of the Day on 2006 January 21, two astronauts are visible, with one appearing much smaller than the other. As of August 2023, out of the astronauts in the NASA Astronaut Group that the smaller astronaut was a member of, which one spent the least time in space, and how many minutes did he spend in space, rounded to the nearest minute? Exclude any astronauts who did not spend any time in space. Give the last name of the astronaut, separated from the number of minutes by a semicolon; use commas as thousands separators in the number of minutes.

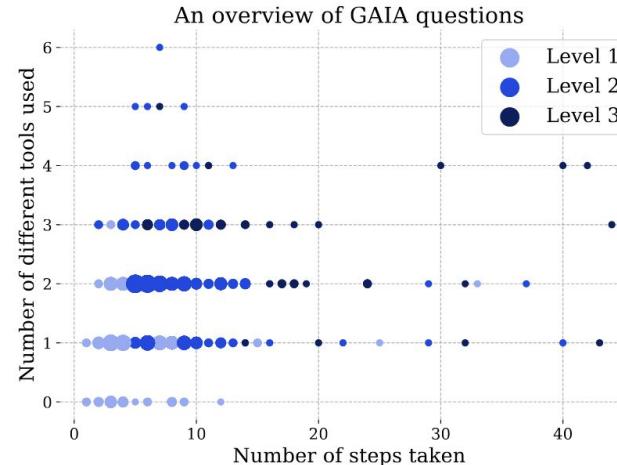
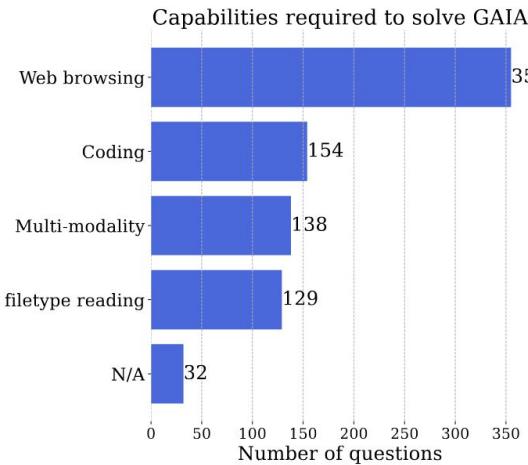
Ground truth: White; 5876



GAIA: A Benchmark for General AI Assistants

Real-world questions that require a set of fundamental abilities such as reasoning, multi-modality handling, web browsing, and generally tool-use proficiency.

GAIA questions are conceptually simple for humans yet challenging for most advanced AIs



Level 1

Question: What was the actual enrollment count of the clinical trial on H. pylori in acne vulgaris patients from Jan-May 2018 as listed on the NIH website?

Ground truth: 90

Level 2

Question: If this whole pint is made up of ice cream, how many percent above or below the US federal standards for butterfat content is it when using the standards as reported by Wikipedia in 2020? Answer as + or - a number rounded to one decimal place.

Ground truth: +4.6

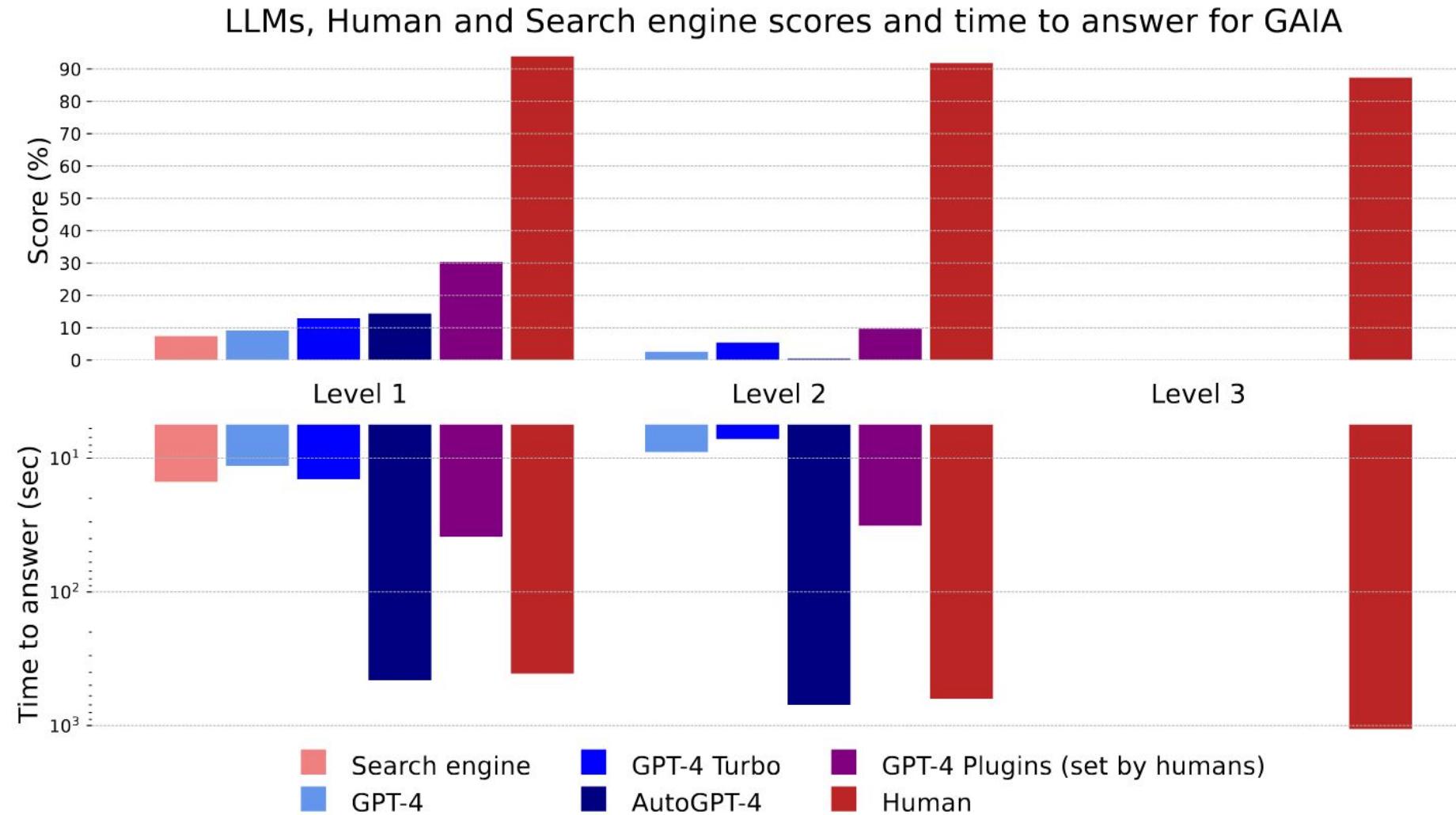
Level 3

Question: In NASA's Astronomy Picture of the Day on 2006 January 21, two astronauts are visible, with one appearing much smaller than the other. As of August 2023, out of the astronauts in the NASA Astronaut Group that the smaller astronaut was a member of, which one spent the least time in space, and how many minutes did he spend in space, rounded to the nearest minute? Exclude any astronauts who did not spend any time in space. Give the last name of the astronaut, separated from the number of minutes by a semicolon. Use commas as thousands separators in the number of minutes.

Ground truth: White; 5876



GAIA: A Benchmark for General AI Assistants

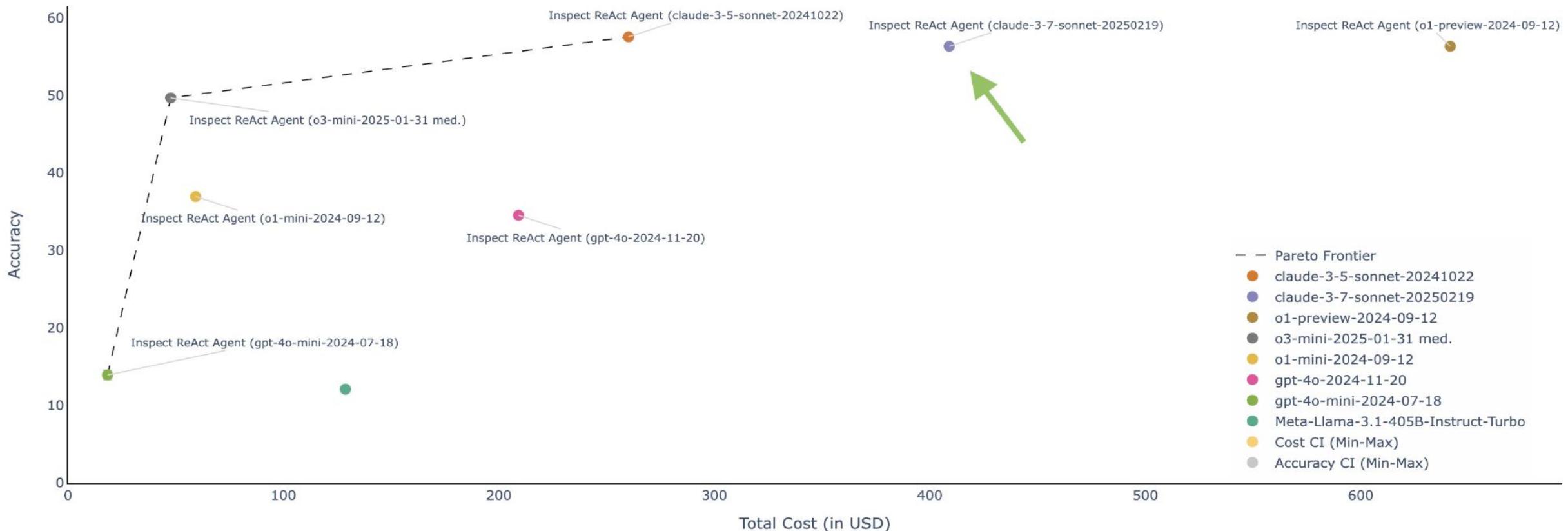




GAIA: A Benchmark for General AI Assistants

Accuracy vs. Cost Frontier for GAIA

This plot shows the relationship between an agent's performance and its token cost. The Pareto frontier (dashed line) represents the current state-of-the-art trade-off. The error bars indicate min-max values across runs.





GAIA: A Benchmark for General AI Assistants

OpenAI Deep Research

| GAIA | Level 1 | Level 2 | Level 3 | Avg. |
|--------------------------------|--------------|--------------|--------------|--------------|
| Previous SOTA ↗ | 67.92 | 67.44 | 42.31 | 63.64 |
| Deep Research (pass@1) | 74.29 | 69.06 | 47.6 | 67.36 |
| Deep Research (cons@64) | 78.66 | 73.21 | 58.03 | 72.57 |

May 4th 2024 - OpenAI GPT4 got 15%
Feb 2nd 2025 - OpenAI Deep Research got 72.6%!



Recent LLM Topics: DeepSeek

DeepSeek

DeepSeek v2

DeepSeek R1



DeepSeek

Key points

- 1) Mixture of Experts (MoE) Transformer
- 2) Multihead latent attention
 - Less memory intensive
 - Stronger performance
- 3) Multi-token prediction
 - Improves accuracy performance
- 4) GRPO
 - More efficient RL training
- 5) Trained on FP8

| Benchmark (Metric) | Claude-3.5-Sonnet-1022 | GPT-4o 0513 | DeepSeek V3 | OpenAI o1-mini | OpenAI o1-1217 | DeepSeek R1 |
|--------------------|----------------------------|-------------|-------------|----------------|----------------|-------------|
| Architecture | - | - | MoE | - | - | MoE |
| # Activated Params | - | - | 37B | - | - | 37B |
| # Total Params | - | - | 671B | - | - | 671B |
| English | MMLU (Pass@1) | 88.3 | 87.2 | 88.5 | 85.2 | 91.8 |
| | MMLU-Redux (EM) | 88.9 | 88.0 | 89.1 | 86.7 | - |
| | MMLU-Pro (EM) | 78.0 | 72.6 | 75.9 | 80.3 | - |
| | DROP (3-shot F1) | 88.3 | 83.7 | 91.6 | 83.9 | 90.2 |
| | IF-Eval (Prompt Strict) | 86.5 | 84.3 | 86.1 | 84.8 | - |
| | GPQA Diamond (Pass@1) | 65.0 | 49.9 | 59.1 | 60.0 | 75.7 |
| | SimpleQA (Correct) | 28.4 | 38.2 | 24.9 | 7.0 | 47.0 |
| | FRAMES (Acc.) | 72.5 | 80.5 | 73.3 | 76.9 | - |
| | AlpacaEval2.0 (LC-winrate) | 52.0 | 51.1 | 70.0 | 57.8 | - |
| | ArenaHard (GPT-4-1106) | 85.2 | 80.4 | 85.5 | 92.0 | - |
| Code | LiveCodeBench (Pass@1-COT) | 38.9 | 32.9 | 36.2 | 53.8 | 63.4 |
| | Codeforces (Percentile) | 20.3 | 23.6 | 58.7 | 93.4 | 96.6 |
| | Codeforces (Rating) | 717 | 759 | 1134 | 1820 | 2061 |
| | SWE Verified (Resolved) | 50.8 | 38.8 | 42.0 | 41.6 | 48.9 |
| | Aider-Polyglot (Acc.) | 45.3 | 16.0 | 49.6 | 32.9 | 61.7 |
| Math | AIME 2024 (Pass@1) | 16.0 | 9.3 | 39.2 | 63.6 | 79.2 |
| | MATH-500 (Pass@1) | 78.3 | 74.6 | 90.2 | 90.0 | 96.4 |
| | CNMO 2024 (Pass@1) | 13.1 | 10.8 | 43.2 | 67.6 | - |
| Chinese | CLUEWSC (EM) | 85.4 | 87.9 | 90.9 | 89.9 | - |
| | C-Eval (EM) | 76.7 | 76.0 | 86.5 | 68.9 | - |
| | C-SimpleQA (Correct) | 55.4 | 58.7 | 68.0 | 40.3 | - |

Table 4 | Comparison between DeepSeek-R1 and other representative models.



1) Mixture of Experts (MoE) Transformer

Sparse MoE layers are used instead of dense feed-forward network (FFN) layers. MoE layers have a certain number of “experts” (e.g. 8), where each expert is usually a feed forward neural network.

Mixture of Experts enable models to be pretrained with far less compute, which means you can dramatically scale up the model or dataset size with the same compute budget as a dense model. In particular, a MoE model should achieve the same quality as its dense counterpart much faster during pretraining.

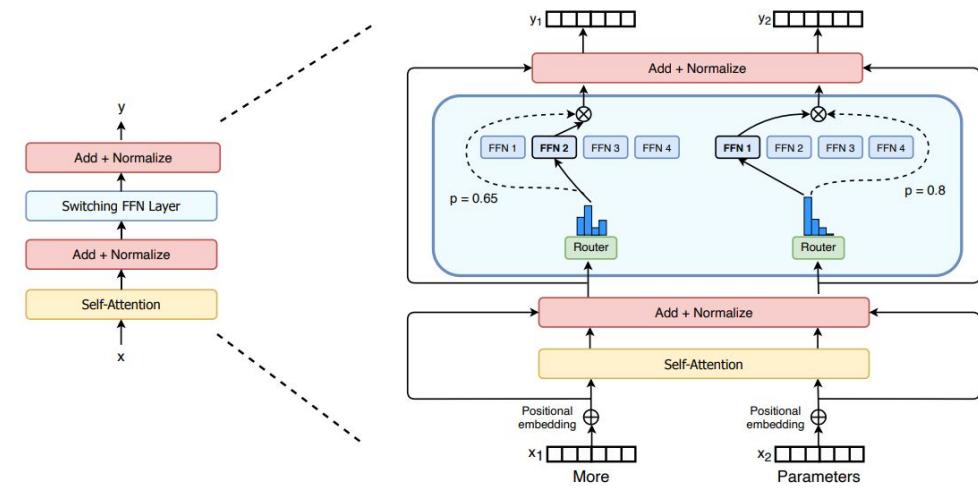


Figure 2: Illustration of a Switch Transformer encoder block. We replace the dense feed forward network (FFN) layer present in the Transformer with a sparse Switch FFN layer (light blue). The layer operates independently on the tokens in the sequence. We diagram two tokens (x_1 = “More” and x_2 = “Parameters” below) being routed (solid lines) across four FFN experts, where the router independently routes each token. The switch FFN layer returns the output of the selected FFN multiplied by the router gate value (dotted-line).

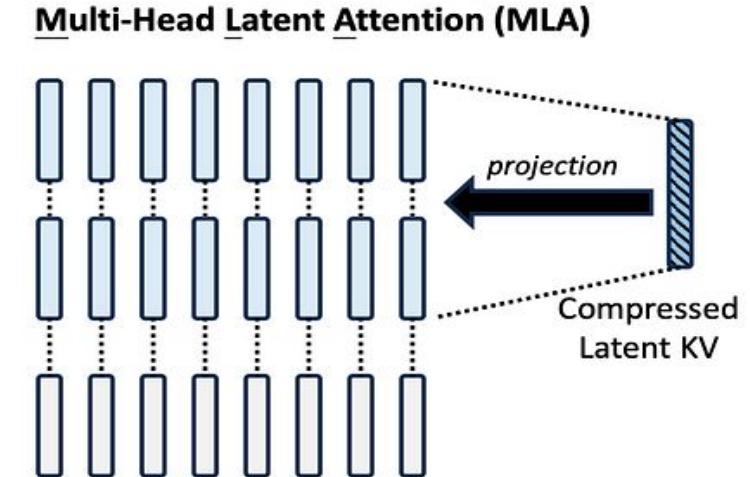
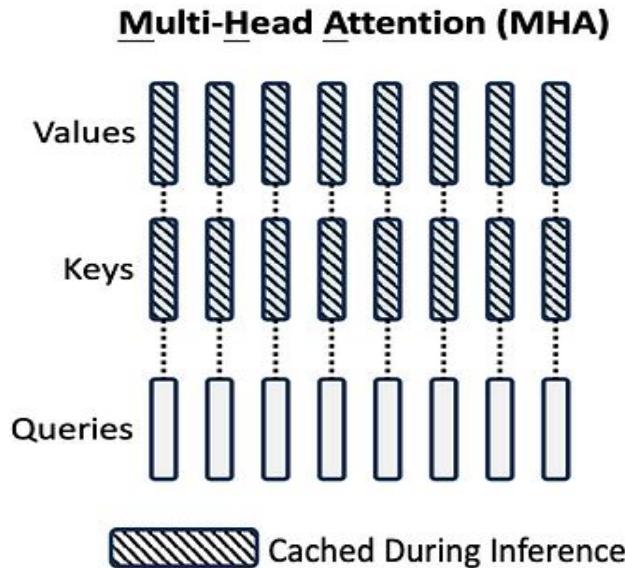


2) Multi-Latent Attention in DeepSeek v2

Instead of storing the full key-value cache, store a compressed version created by projection layers (same as LoRa).

- Store only c_t^{KV} (down-projected version for h_t)
- key and value can be up-projected from c_t^{KV} .

$$\begin{aligned} \mathbf{c}_t^{KV} &= W^{DKV} \mathbf{h}_t, \\ \mathbf{k}_t^C &= W^{UK} \mathbf{c}_t^{KV}, \\ \mathbf{v}_t^C &= W^{UV} \mathbf{c}_t^{KV}, \end{aligned}$$



(Only for inference)

| Attention Mechanism | KV Cache per Token (# Element) | Capability |
|-------------------------------|---|------------|
| Multi-Head Attention (MHA) | $2n_h d_h l$ | Strong |
| Grouped-Query Attention (GQA) | $2n_g d_h l$ | Moderate |
| Multi-Query Attention (MQA) | $2d_h l$ | Weak |
| MLA (Ours) | $(d_c + d_h^R)l \approx \frac{9}{2}d_h l$ | Stronger |



2) Multi-Latent Attention in DeepSeek R1

More compression efficiency (25% reduction in memory usage)

DeepSeek v2

$$c_{KV_t} = W_{D_{KV}} h_t$$

DeepSeek v3 & R1

$$c_{KV_t} = W_{D_{KV,1}} W_{D_{KV,2}} h_t$$

$$W_{D_{KV}} \in \mathbb{R}^{d_c \times d}$$

$$d_c = 4d_h$$

25% reduction

$$W_{D_{KV,1}} \in \mathbb{R}^{d_m \times d} \text{ and } W_{D_{KV,2}} \in \mathbb{R}^{d_c \times d_m}$$

$$d_c = 3d_h$$

Also applied to query

$$c_{Q_t} = W_{D_Q} h_t, \quad q_t^C = W_{U_Q} c_{Q_t}$$

3) Multi-token Prediction

The MTP modules can also be used to do **speculative decoding**.

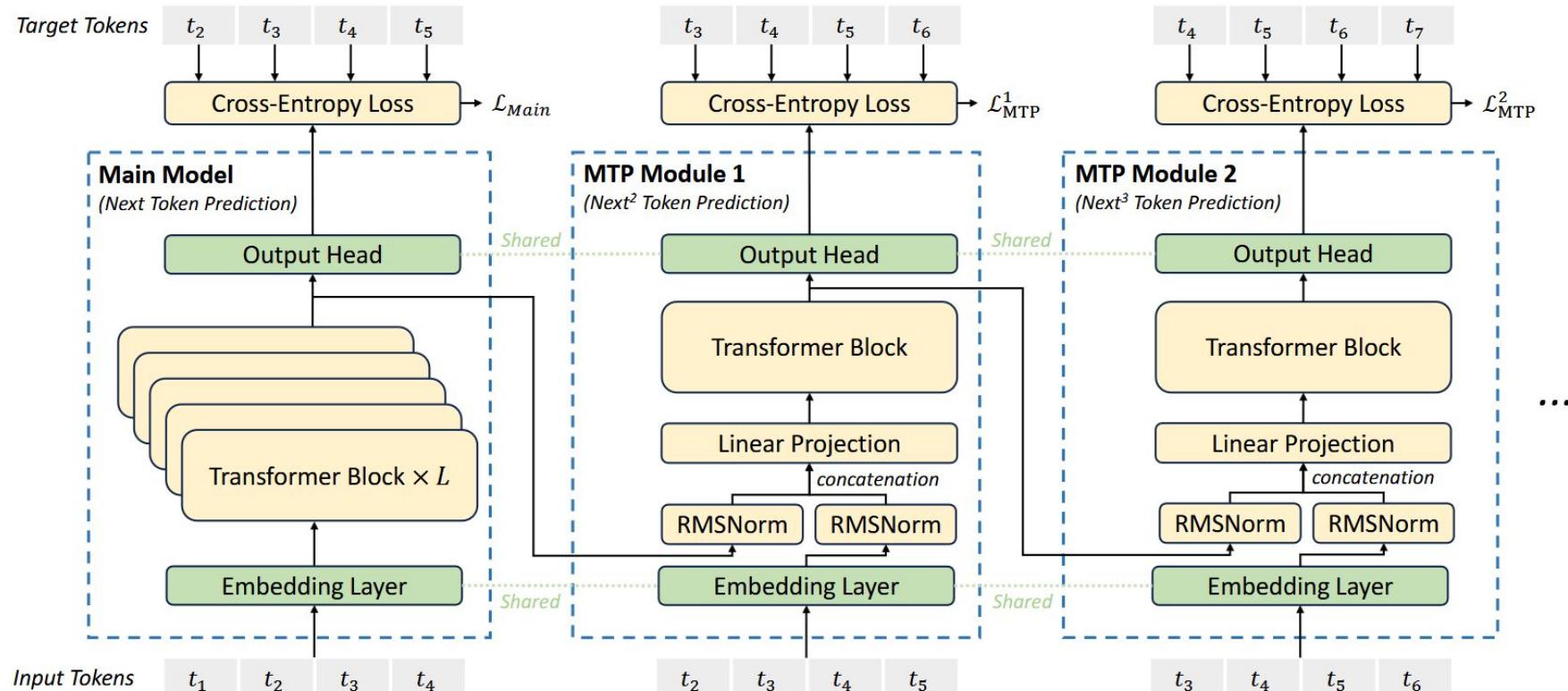


Figure 3 | Illustration of our Multi-Token Prediction (MTP) implementation. We keep the complete causal chain for the prediction of each token at each depth.



Algorithm 2 Speculative Sampling (SpS) with Auto-Regressive Target and Draft Models

Given lookahead K and minimum target sequence length T .

Given auto-regressive target model $q(\cdot|\cdot)$, and auto-regressive draft model $p(\cdot|\cdot)$, initial prompt sequence x_0, \dots, x_t .

Initialise $n \leftarrow t$.

while $n < T$ **do**

for $t = 1 : K$ **do**

 Sample draft auto-regressively $\tilde{x}_t \sim p(x|, x_1, \dots, x_n, \tilde{x}_1, \dots, \tilde{x}_{t-1})$

end for

 In parallel, compute $K + 1$ sets of logits from drafts $\tilde{x}_1, \dots, \tilde{x}_K$:

$$q(x|, x_1, \dots, x_n), q(x|, x_1, \dots, x_n, \tilde{x}_1), \dots, q(x|, x_1, \dots, x_n, \tilde{x}_1, \dots, \tilde{x}_K)$$

for $t = 1 : K$ **do**

 Sample $r \sim U[0, 1]$ from a uniform distribution.

if $r < \min\left(1, \frac{q(x|x_1, \dots, x_{n+t-1})}{p(x|x_1, \dots, x_{n+t-1})}\right)$, **then**

 Set $x_{n+t} \leftarrow \tilde{x}_t$ and $n \leftarrow n + 1$.

else

 sample $x_{n+t} \sim (q(x|x_1, \dots, x_{n+t-1}) - p(x|x_1, \dots, x_{n+t-1}))_+$ and exit for loop.

end if

end for

 If all tokens x_{n+1}, \dots, x_{n+K} are accepted, sample extra token $x_{n+K+1} \sim q(x|, x_1, \dots, x_n, x_{n+K})$ and set $n \leftarrow n + 1$.

end while

4) Group Relative Policy Optimization (GRPO)

PPO is an actor-critic RL algorithm -> it needs to train two models.

But GRPO needs just one.

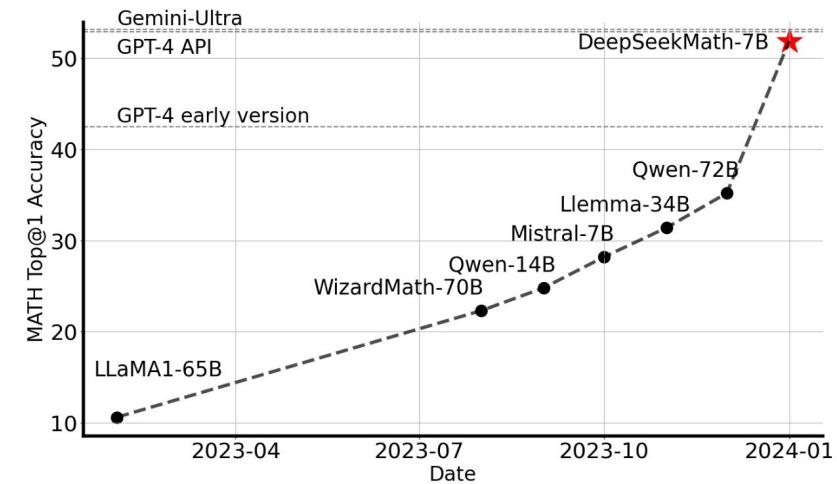
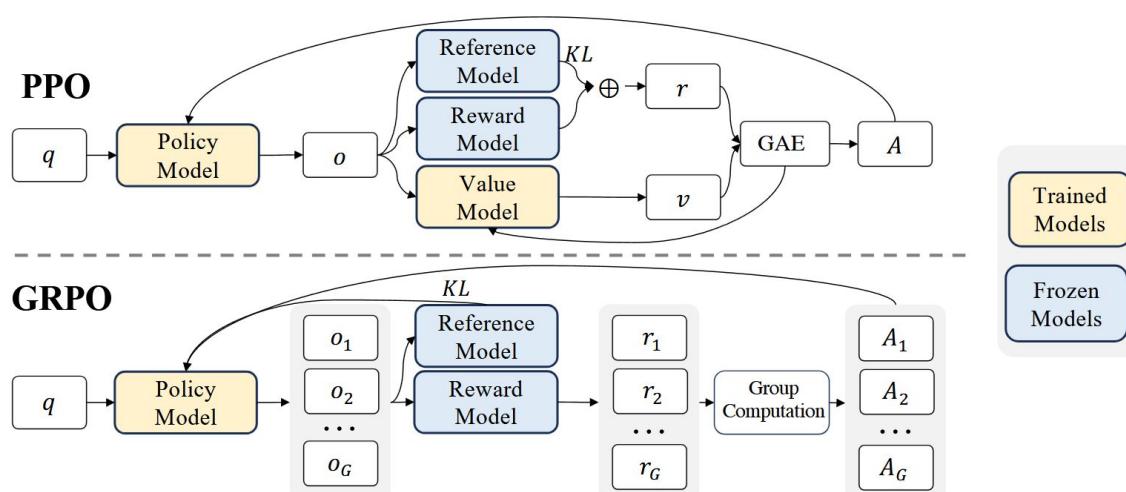
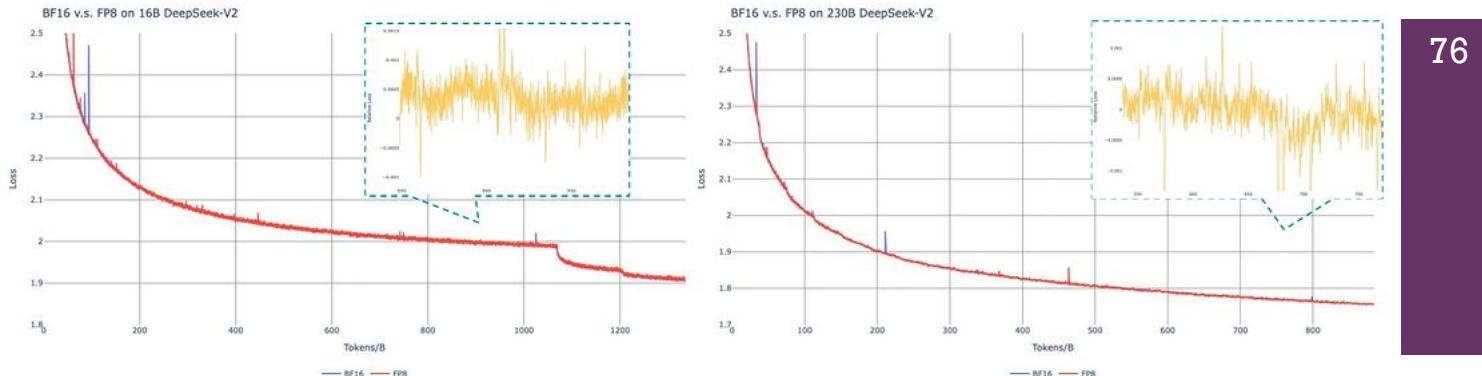


Figure 1 | Top1 accuracy of open-source models on the competition-level MATH benchmark (Hendrycks et al., 2021) without the use of external toolkits and voting techniques.

5) Trained on FP8

Saves memory up to 30% and trains faster with no cost to the final model performance.



| Configuration | Micro BS | Status | Throughput (Samples/sec) |
|-----------------------------|----------|----------|-----------------------------|
| BF16 | 1 | | 12.65 |
| FP8 + SwiGLU output in BF16 | 1 | Converge | 16.07 (+ 27.04 %) |
| FP8 + Smooth SwiGLU | 1 | Converge | 16.89 (+33.52 %) |
| FP8 | 1 | Diverge | 17.34 (+37.08 %) |

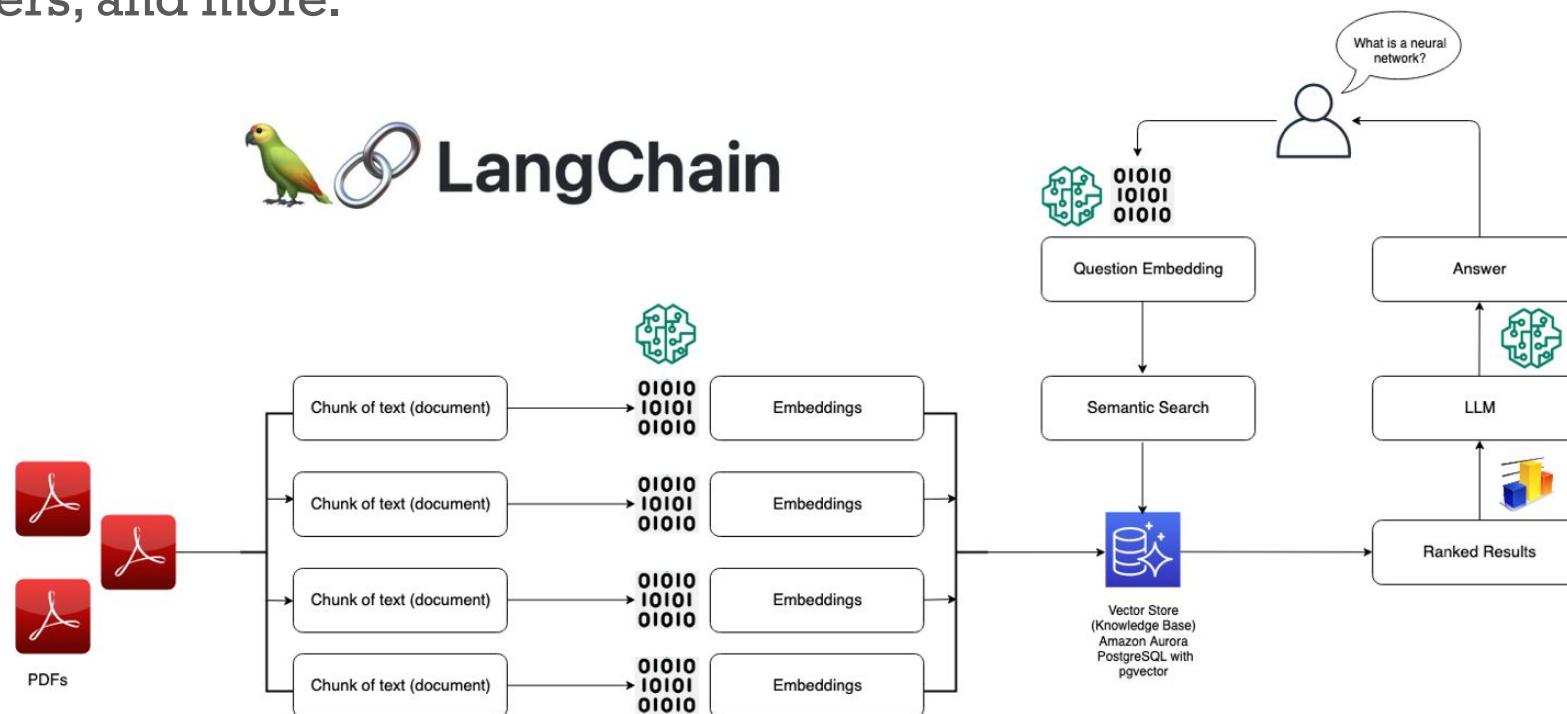
| Configuration | Status | Memory (GB/HPU) | FP8 Optimizer |
|-----------------------------|----------|--------------------|---------------|
| BF16 | | 63.25 | X |
| FP8 + SwiGLU output in BF16 | Converge | 63.26 | X |
| FP8 + Smooth SwiGLU | Converge | 63.26 | X |
| FP8 | Diverge | 63.24 | X |
| FP8 + SwiGLU output in BF16 | Converge | 44.08 | ✓ |
| FP8 + Smooth SwiGLU | Converge | 44.08 | ✓ |
| FP8 | Diverge | 44.09 | ✓ |



LangChain

LangChain provides tools and abstractions to improve the customization, accuracy, and relevancy of the information the models generate.

LangChain streamlines intermediate steps to develop data-responsive applications, making prompt engineering more efficient. It is designed to develop diverse applications powered by language models more effortlessly, including chatbots, question-answering, content generation, summarizers, and more.





LangChain

LangChain has 100+ integrations.
See full list from the link below.

LangChain integrates with many providers.

Integration Packages

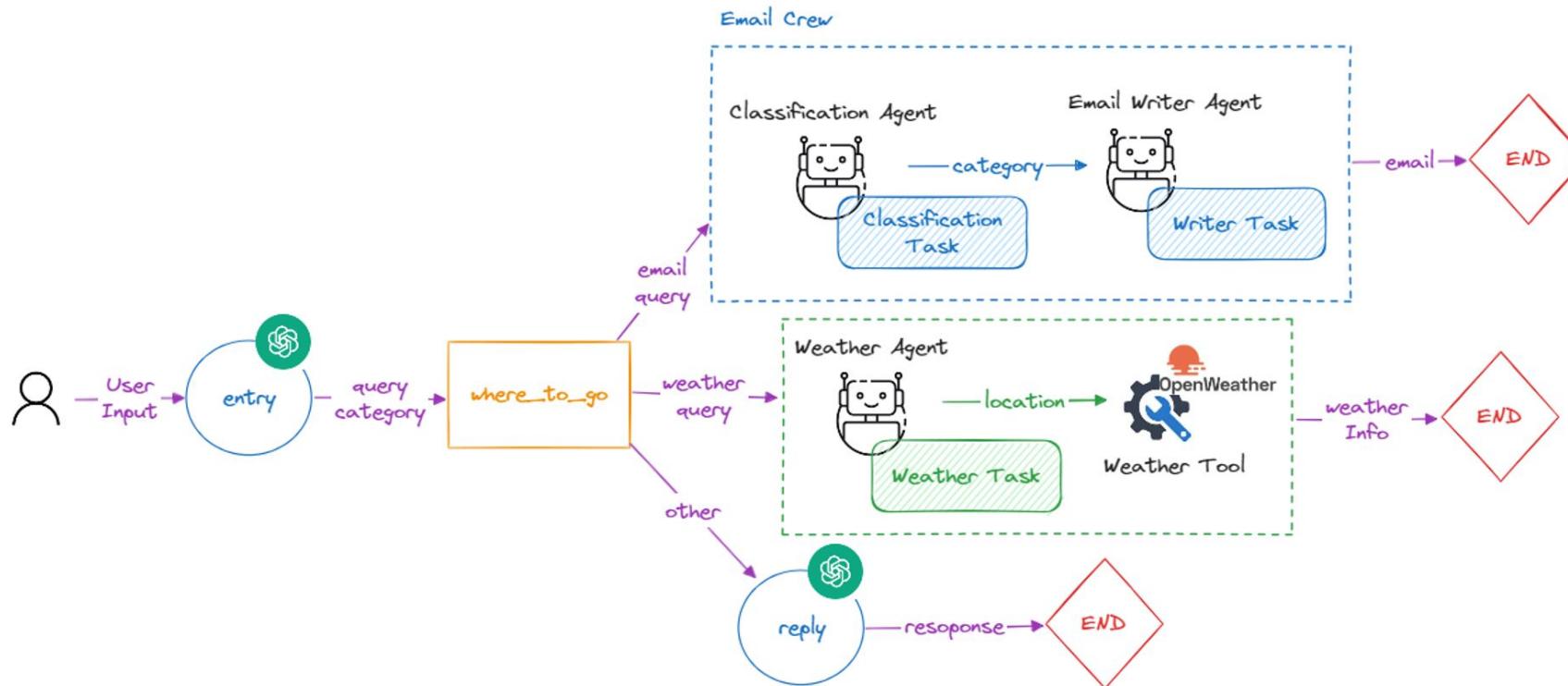
These providers have standalone `langchain-{provider}` packages for improved versioning, dependency testing.

| Provider | Package | Downloads | Latest | JS |
|----------------------|---|------------|---------|----|
| Google VertexAI | <code>langchain-google-vertexai</code> | 15M/month | v2.0.15 | ✓ |
| OpenAI | <code>langchain-openai</code> | 12M/month | v0.3.8 | ✓ |
| Google Community | <code>langchain-google-community</code> | 4.3M/month | v2.0.7 | ✗ |
| AWS | <code>langchain-aws</code> | 2M/month | v0.2.15 | ✓ |
| Anthropic | <code>langchain-anthropic</code> | 1.9M/month | v0.3.10 | ✓ |
| Google Generative AI | <code>langchain-google-genai</code> | 1.3M/month | v2.1.0 | ✓ |
| Ollama | <code>langchain-ollama</code> | 844k/month | v0.2.3 | ✓ |



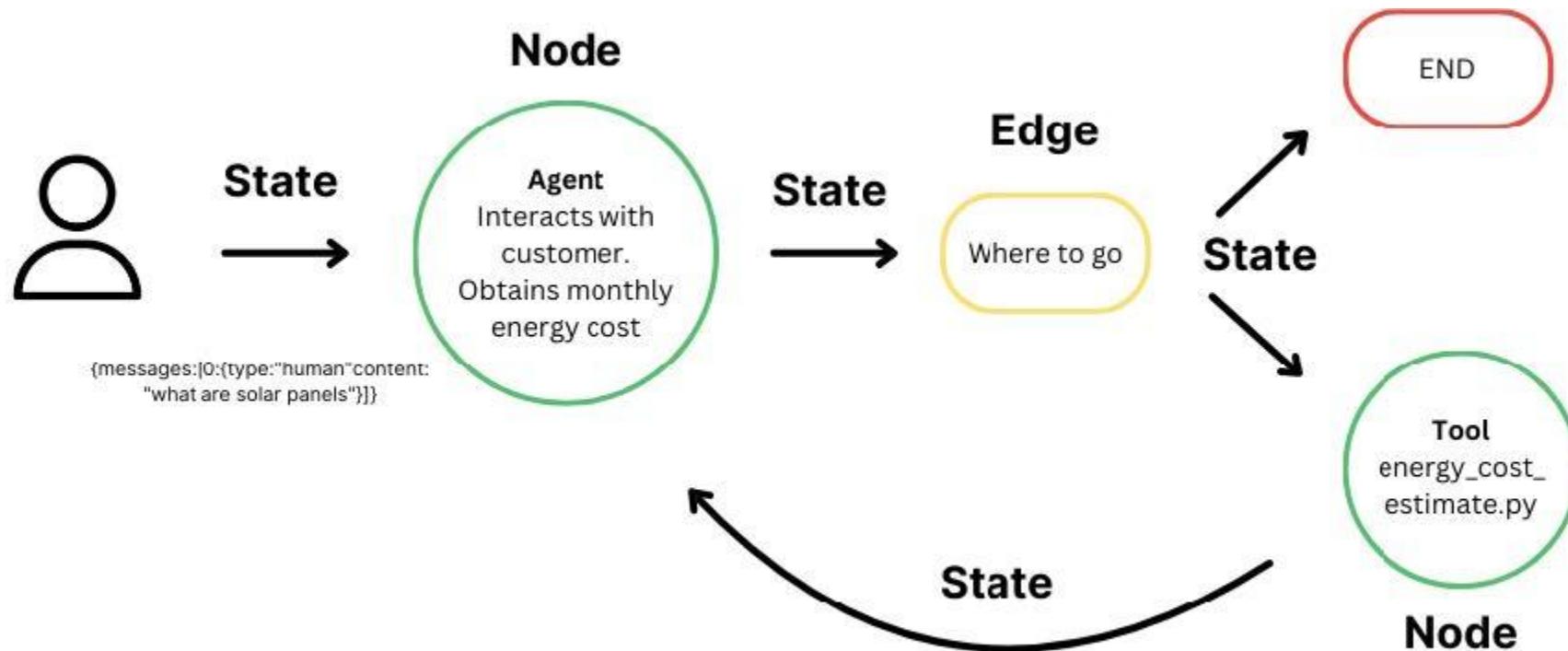
LangGraph

LangGraph is an orchestration framework for complex agentic systems and is more low-level and controllable than LangChain agents.



LangGraph

LangGraph is a framework designed for graph-based applications.



LangSmith

Langsmith is designed with AI model debugging and orchestration in mind.

Observability

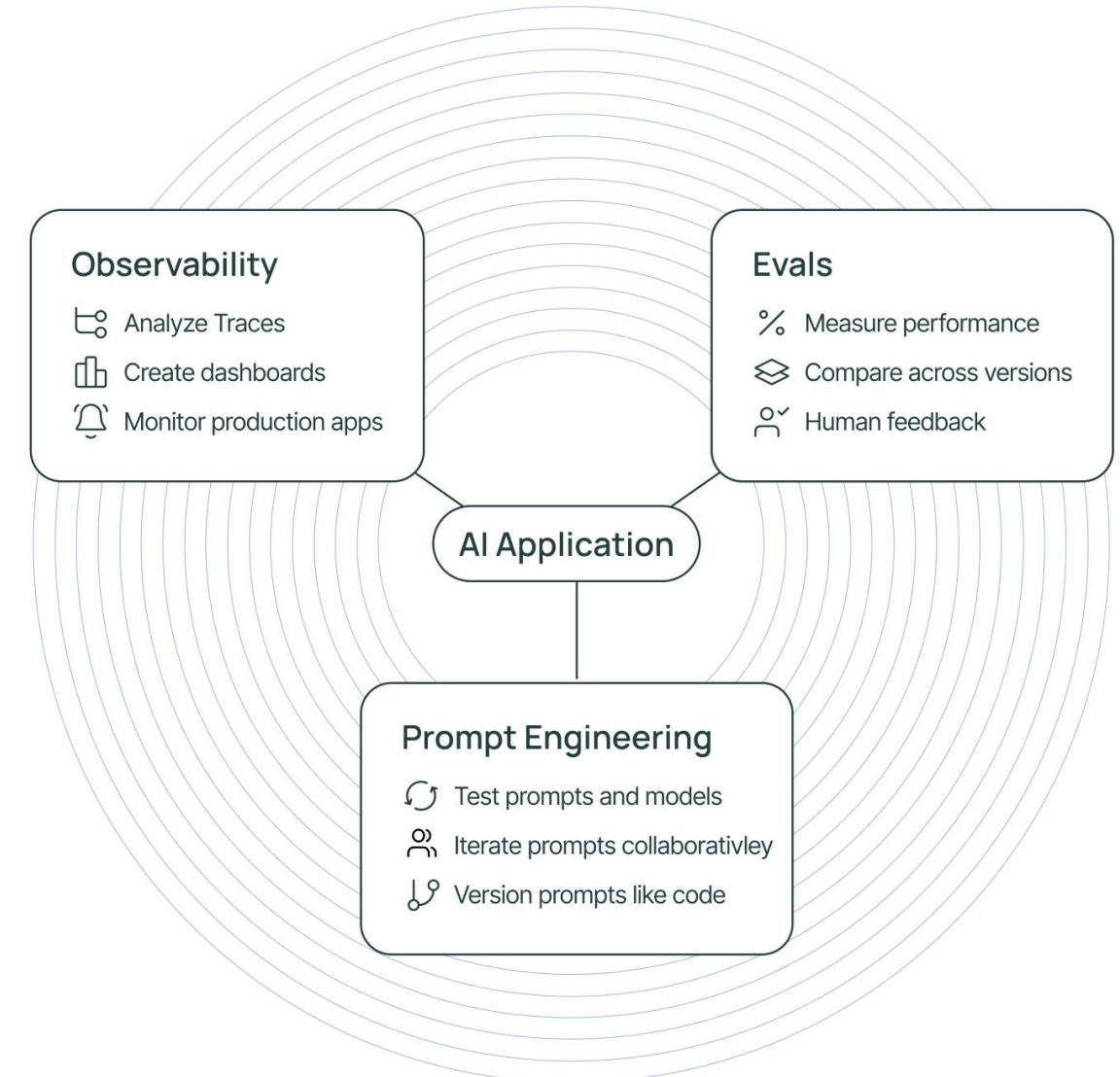
Analyze traces in LangSmith and configure metrics, dashboards, alerts based on these.

Evals

Evaluate your application over production traffic — score application performance and get human feedback on your data.

Prompt Engineering

Iterate on prompts, with automatic version control and collaboration features.





Other Tools

llama index

autogen...

