

Language Models for Law and Social Science

10. Retrieval, Reasoning, and Agents

<https://x.com/francoisfleuret/status/1916384736139780460>



François Fleuret @francoisfleuret · Apr 26

What do you think are the most important improvements of the 2017 transformer architecture?



François Fleuret @francoisfleuret · Apr 26

What do you think are the most important improvements of the 2017 transformer architecture?

- ▶ **Prenorm:** normalization in the residual blocks before the attention operation and the FFN
- ▶ **GQA (Group Query Attention):** more Q than (K, V)
- ▶ **RMSNorm instead of Layernorm:** normalize only the scaling
- ▶ **MLA (Multi-head Latent Attention):** stores a low-rank projection of the attention block input and compute the KV from it
- ▶ **SwiGLU:** non-linearity for the FFN block with per-component gating
- ▶ **RoPE (Rotary Positional Embedding):** attention depends only on the relative Q/K positions
- ▶ **MoE (Mixture of Experts):** The FFN block is implemented with multiple MLPs and a gating mechanism selects which ones process each token.
- ▶ **Warmup:** very short ramping-up of the learning rate, starting from 0
- ▶ **Cosine schedule:** the learning rate varies less at the beginning and end of the schedule
- ▶ **AdamW:** decouples weight includes decay from Adam
- ▶ **Multi-token prediction:** sums the training over multiple future tokens, possibly with additional readout heads.
- ▶ **FlashAttention:** computes the attention on the fly, avoiding a memory footprint $O(T^2)$ (+ optimizes very carefully for the GPU!)
- ▶ **Ring Attention:** exploits multi-node hardware to scale computation according to sequence length
- ▶ **Speculative decoding:** a cheaper model generates tokens, and a rejection process corrects this generation to match the full-model distribution.

PPO vs GRPO

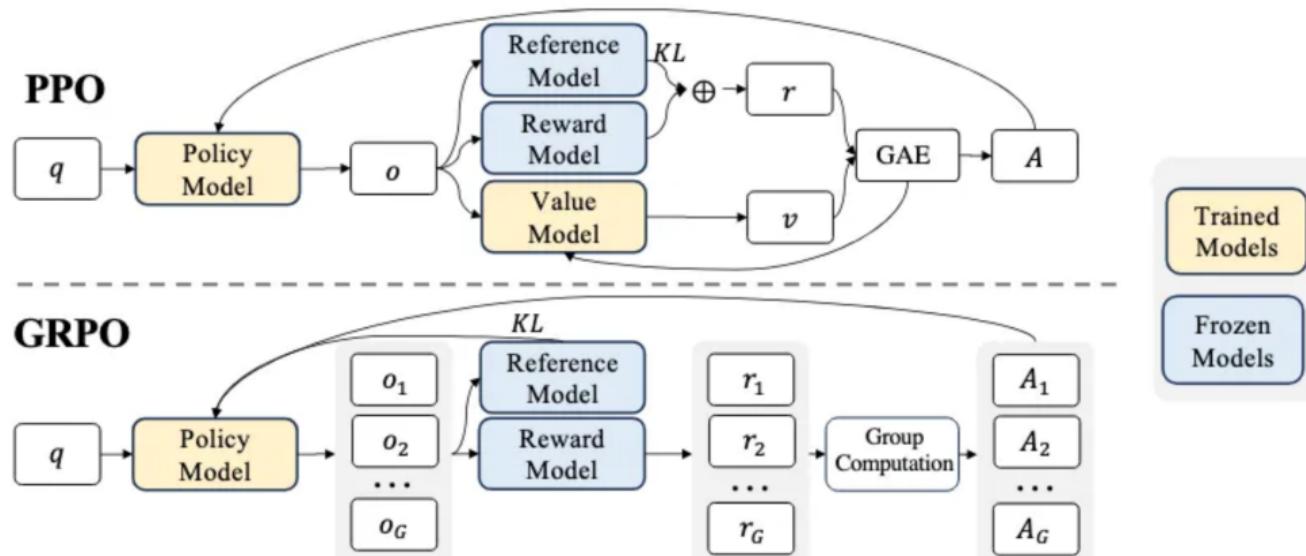


Figure 4 | Demonstration of PPO and our GRPO. GRPO foregoes the value model, instead estimating the baseline from group scores, significantly reducing training resources.

GPT-4 Limitations (“Sparks of AGI” Paper, April 2023)

GPT-4 Limitations (“Sparks of AGI” Paper, April 2023)

- **Confidence calibration:** The model has trouble knowing when it should be confident and when it is just guessing. It both makes up facts that have not appeared in its training data, and also exhibits inconsistencies between the generated content and the prompt, which we referred to as *open-domain* and *closed-domain* hallucination in Figure 1.8. These hallucinations can be stated in a confident and persuasive manner that can be difficult to detect. Thus, such generations can lead to errors, and
- **Long-term memory:** The model’s context is very limited, it operates in a “stateless” fashion and there is no obvious way to teach the model new facts. In fact, it is not even clear whether the model is able to perform tasks which require an evolving memory and context, such as reading a book, with the task of following the plot and understanding references to prior chapters over the course of reading.
- **Continual learning:** The model lacks the ability to update itself or adapt to a changing environment. The model is fixed once it is trained, and there is no mechanism for incorporating new information
- hension and prowess. The model does not have any way to incorporate such personalized information into its responses, except by using meta-prompts, which are both limited and inefficient.
- **Planning and conceptual leaps:** As suggested by the examples in Section 8, the model exhibits difficulties in performing tasks that require planning ahead or that require a “Eureka idea” constituting a discontinuous conceptual leap in the progress towards completing a task. In other words, the model
- **Transparency, interpretability and consistency:** Not only does the model hallucinate, make up facts and produce inconsistent content, but it seems that the model has no way of verifying whether or not the content that it produces is consistent with the training data, or whether it’s self-consistent. While
- **Cognitive fallacies and irrationality:** The model seems to exhibit some of the limitations of human knowledge and reasoning, such as cognitive biases and irrationality (such as biases of confirmation, anchoring, and base-rate neglect) and statistical fallacies. The model may inherit some of the biases, prejudices, or errors that are present in its training data, which may reflect the distribution of opinions or perspectives linked to subsets of the population or larger common views and assessments.
- **Challenges with sensitivity to inputs:** The model’s responses can be very sensitive to details of the framing or wording of prompts and their sequencing in a session. Such non-robustness suggests that

What does GPT-4 need to become an AGI?

What does GPT-4 need to become an AGI?

- ▶ At least two things:
 1. information retrieval (to access a fresh knowledge base).
 2. an ability to take actions, and interact with the environment (rather than just writing), e.g., to collect evidence.

What does GPT-4 need to become an AGI?

- ▶ At least two things:
 1. information retrieval (to access a fresh knowledge base).
 2. an ability to take actions, and interact with the environment (rather than just writing), e.g., to collect evidence.
- ▶ And maybe one more thing:
 3. the ability to reliably extrapolate and reason through exploration and self-correction.

Outline

Evaluating LLMs

Information Retrieval / RAG

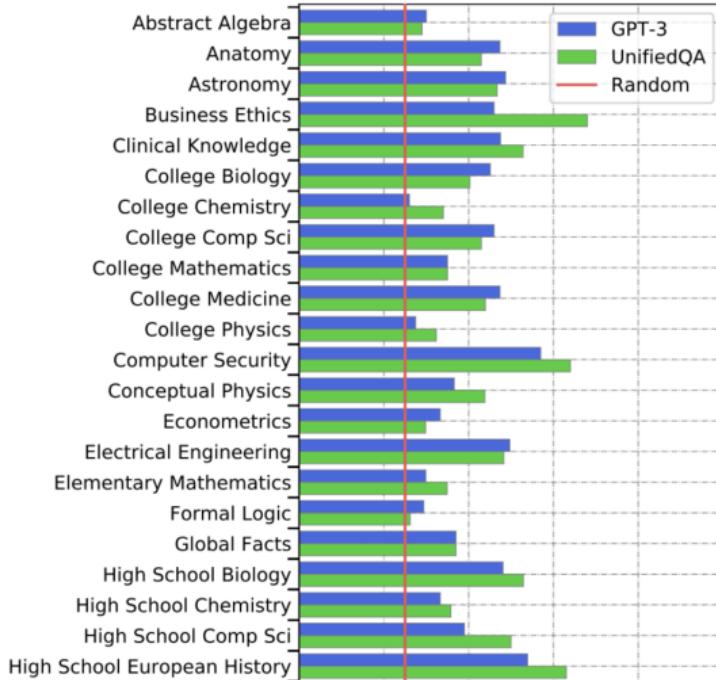
Agents

Reasoning Models

Massive Multitask Language Understanding (MMLU)

[[Hendrycks et al., 2021](#)]

New benchmarks for measuring LM performance on 57 diverse *knowledge intensive* tasks



MMLU contains 16K multiple-choice questions

Astronomy

What is true for a type-Ia supernova?

- A. This type occurs in binary systems.
- B. This type occurs in young galaxies.
- C. This type produces gamma-ray bursts.
- D. This type produces high amounts of X-rays.

Answer: A

High School Biology

In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of

- A. directional selection.
- B. stabilizing selection.
- C. sexual selection.
- D. disruptive selection

Answer: A

BIG Bench

BIG-Bench [[Srivastava et al., 2022](#)]

200+ tasks, spanning:



https://github.com/google/BIG-Bench/blob/main/bigbench/benchmark_tasks/README.md

Kanji ASCII Art to Meaning

This subtask converts various kanji into ASCII art and has the language model guess their meaning from the ASCII art.

.....#
.....#

.... #####
... # # . # #.
. # . # #.
. # . # # #
.. # . # . # #.
.. # # . . # #.

.. # # . . # #.
.. # # # . . # #.
#####. # # . # #.
. # . # . # . #.

- ▶ combination of exact match and string similarity metrics.

GPQA: Graduate-Level Google-Proof Q&A

Chemistry (general)

A reaction of a liquid organic compound, which molecules consist of carbon and hydrogen atoms, is performed at 80 centigrade and 20 bar for 24 hours. In the proton nuclear magnetic resonance spectrum, the signals with the highest chemical shift of the reactant are replaced by a signal of the product that is observed about three to four units downfield. Compounds from which position in the periodic system of the elements, which are also used in the corresponding large-scale industrial process, have been mostly likely initially added in small amounts?

- A) A metal compound from the fifth period.
 - B) A metal compound from the fifth period and a non-metal compound from the third period.
 - C) A metal compound from the fourth period.
 - D) A metal compound from the fourth period and a non-metal compound from the second period.
-

Organic Chemistry

Methylcyclopentadiene (which exists as a fluxional mixture of isomers) was allowed to react with methyl isoamyl ketone and a catalytic amount of pyrrolidine. A bright yellow, cross-conjugated polyalkenyl hydrocarbon product formed (as a mixture of isomers), with water as a side product. These products are derivatives of fulvene. This product was then allowed to react with ethyl acrylate in a 1:1 ratio. Upon completion of the reaction, the bright yellow color had disappeared. How many chemically distinct isomers make up the final product (not counting stereoisomers)?

- A) 2
 - B) 16
 - C) 8
 - D) 4
-

- ▶ 26,529 multiple-choice questions across 285 fields, designed to be at graduate-school level (Msc or PhD), and not answerable with google.

LLM Benchmarks: Arena-Hard-Auto

<https://arxiv.org/abs/2406.11939>

<https://github.com/lmarena/arena-hard-auto>

- ▶ Tasks:
 - ▶ 500 challenging real-world user queries (open-ended software engineering problems, math questions, etc)
 - ▶ 250 creative writing queries sourced from Chatbot Arena
- ▶ Evaluation:
 - ▶ LLM judges (!): GPT-4.1 and Gemini-2.5, as a cheaper and faster approximator to human preference.
 - ▶ Does pairwise comparisons between models.
 - ▶ According to paper, gets 98.6% correlation with human judgments.

Arena-Hard-Auto System prompt

https://github.com/lmarena/arena-hard-auto/blob/main/utils/judge_utils.py

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user prompt displayed below. You will be given assistant A's answer and assistant B's answer. Your job is to evaluate which assistant's answer is better.

Begin your evaluation by generating your own answer to the prompt. You must provide your answers before judging any answers.

When evaluating the assistants' answers, compare both assistants' answers with your answer. You must identify and correct any mistakes or inaccurate information.

Then consider if the assistant's answers are helpful, relevant, and concise. Helpful means the answer correctly responds to the prompt or follows the instructions. Note when user prompt has any ambiguity or more than one interpretation, it is more helpful and appropriate to ask for clarifications or more information from the user than providing an answer based on assumptions. Relevant means all parts of the response closely connect or are appropriate to what is being asked. Concise means the response is clear and not verbose or excessive.

Then consider the creativity and novelty of the assistant's answers when needed. Finally, identify any missing important information in the assistants' answers that would be beneficial to include when responding to the user prompt.

After providing your explanation, you must output only one of the following choices as your final verdict with a label:

1. Assistant A is significantly better: [[A>>B]]
2. Assistant A is slightly better: [[A>B]]
3. Tie, relatively the same: [[A=B]]
4. Assistant B is slightly better: [[B>A]]
5. Assistant B is significantly better: [[B>>A]]

Example output: \"My final verdict is tie: [[A=B]]\".

Outline

Evaluating LLMs

Information Retrieval / RAG

Agents

Reasoning Models

BREAKING

Lawyer Used ChatGPT In Court — And Cited Fake Cases. A Judge Is Considering Sanctions

Molly Bohannon Forbes Staff

I cover breaking news.

Follow



Jun 8, 2023, 02:06pm EDT

“Hallucination” in LLMs

- ▶ If you ask chat GPT from 2022 “Who is the US president today”, you won’t get a good answer.
- ▶ And even with GPT from 2025, asking niche questions like “who is the mayor of a small town in Graubunden” will often give you wrong answers.

“Hallucination” in LLMs

- ▶ If you ask chat GPT from 2022 “Who is the US president today”, you won’t get a good answer.
- ▶ And even with GPT from 2025, asking niche questions like “who is the mayor of a small town in Graubunden” will often give you wrong answers.
- ▶ These wrong answers are called “hallucinations”.
- ▶ The problem is that these items weren’t observed in the pre-training data, or at least not often enough or clearly enough for the LLM to recall it easily from its internal world representation.

How children ChatGPT learns language

1. **word prediction** – they learn to repeat what they hear.
 2. **instruction fine tuning** – they learn some explicit rules.
 3. **learning from feedback** – they get positive and negative feedback when they say something right or wrong.
-
- GPT-1/2/3 (until 2020) only have step 1.
 - ChatGPT+ (starting 2022) have steps 2 and 3 (alignment).

How children ChatGPT learns language

1. **word prediction** – they learn to repeat what they hear.
 2. **instruction fine tuning** – they learn some explicit rules.
 3. **learning from feedback** – they get positive and negative feedback when they say something right or wrong.
- ChatGPT+ learns like humans; it also makes errors like humans:
 - can forget things it learned.
 - makes errors when extrapolating/guessing/inferring.
 - doesn't know about new facts / changed circumstances.

How children ChatGPT learns language

1. **word prediction** – they learn to repeat what they hear.
 2. **instruction fine tuning** – they learn some **explicit rules**.
 3. **learning from feedback** – they get **positive and negative feedback** when they say something right or wrong.
- ChatGPT+ learns like humans; it also makes errors like humans:
 - can forget things it learned.
 - makes errors when extrapolating/guessing/inferring.
 - doesn't know about new facts / changed circumstances.
 - These errors can be exacerbated – and are expressed more confidently – due to **rules and feedback favoring responsiveness and confidence**.

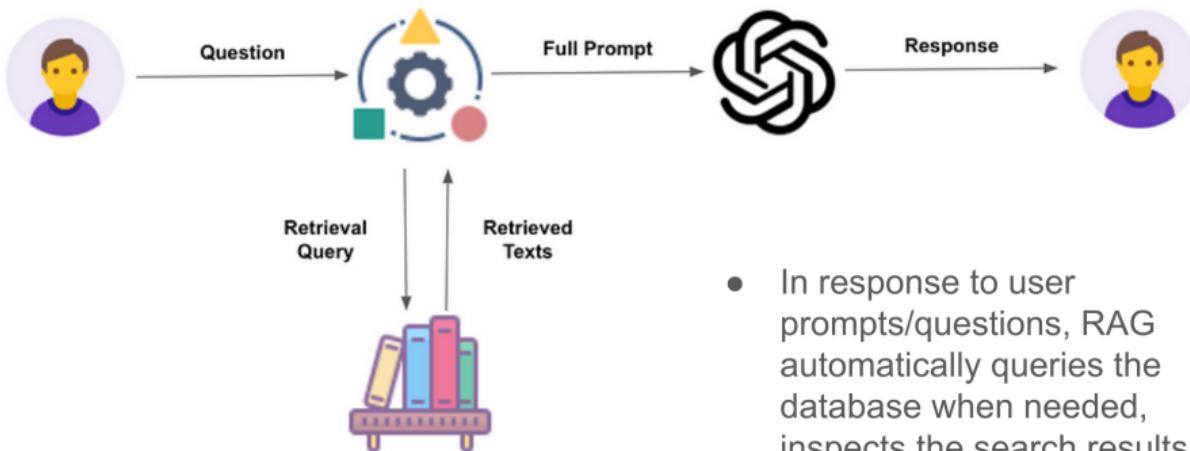
Solutions to hallucinations

- ▶ Two solutions that are not feasible:
 - ▶ re-train the model every day
 - ▶ put all the facts in the system prompt

Addressing “hallucination” with RAG

- ▶ Retrieval Augmented Generation (“RAG”) gives LLMs like GPT access to:
 - ▶ a database of useful and reliable knowledge
 - ▶ a search engine for efficiently retrieving information from the database
- ▶ RAG results can be used by the LLM to answer questions or perform tasks.
- ▶ RAG provides citations to source documents for human checking.

RAG = Retrieval Augmented Generation



Open Question Answering and Claim Verification

- ▶ Open question answering:
 - ▶ Answer any question.
 - ▶ “What are the responsibilities of the mayor of Zurich?”
- ▶ Open claim verification:
 - ▶ Check whether a plain-text claim is true or false.
 - ▶ “Zurich has the second-highest per-capita income of any city in Europe.”
- ▶ Both problems are solved using information retrieval pipelines:
 - ▶ search large corpora or knowledge graphs for evidence
 - ▶ use evidence to answer the question or check the claim

Information Retrieval for Question Answering

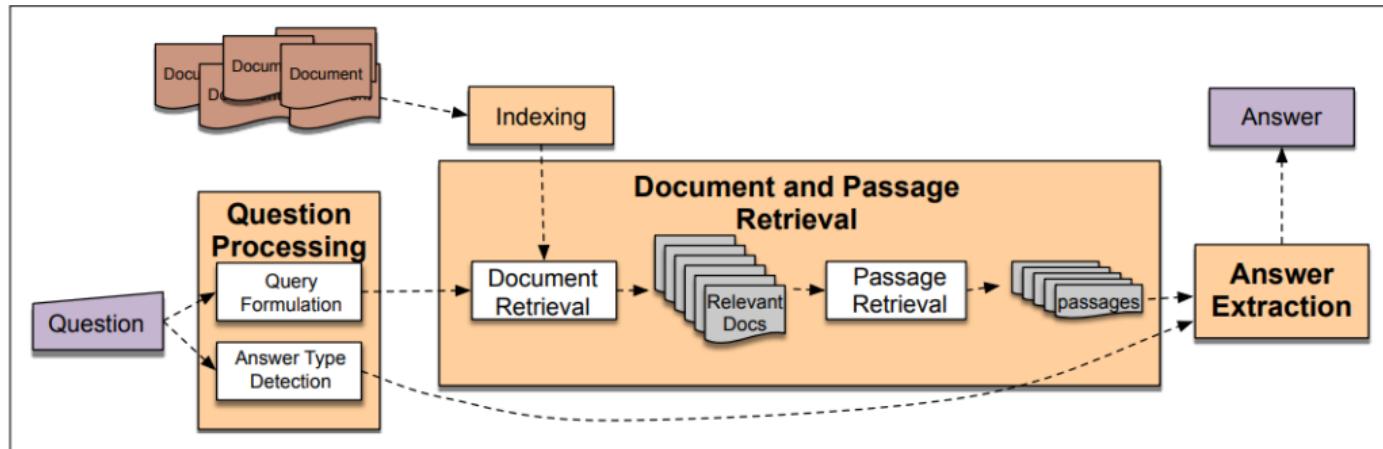


Figure 25.2 IR-based factoid question answering has three stages: question processing, passage retrieval, and answer processing.

Automated Claim Verification

Claim (*by Minister Shailesh Vara*)

“The average criminal bar barrister working full-time is earning some £84,000.”

Verdict: FALSE (*by Channel 4 Fact Check*)

The figures the Ministry of Justice have stressed this week seem decidedly dodgy. Even if you do want to use the figures, once you take away the many overheads self-employed advocates have to pay you are left with a middling sum of money.

1. Claim spotting (what to fact check – facts vs opinions, etc)
2. Evidence retrieval
3. Evidence filtering
4. Fact-check claim given evidence (textual entailment)

Information Retrieval / RAG

- ▶ Open question answering and automated claim verification are examples of Retrieval Augmented Generation (RAG).
- ▶ How it works:
 - ▶ vectorize all the documents in the knowledge base
 - ▶ vectorize the query
 - ▶ retrieve relevant documents as those with the highest cosine similarity to the query.

RAG: How to vectorize the documents

- ▶ Traditional approach is **bm25**:
 - ▶ roughly, rank all documents by their TF-IDF cosine similarity to the query.
 - ▶ index every document by which words it contains, to quickly filter out irrelevant documents.

RAG: How to vectorize the documents

- ▶ Traditional approach is **bm25**:
 - ▶ roughly, rank all documents by their TF-IDF cosine similarity to the query.
 - ▶ index every document by which words it contains, to quickly filter out irrelevant documents.
- ▶ Recent systems (e.g. OpenAI, LlamaIndex) use embeddings:
 - ▶ e.g. S-BERT-type embeddings from SentenceTransformers.

RAG: How to vectorize the documents

- ▶ Traditional approach is **bm25**:
 - ▶ roughly, rank all documents by their TF-IDF cosine similarity to the query.
 - ▶ index every document by which words it contains, to quickly filter out irrelevant documents.
- ▶ Recent systems (e.g. OpenAI, LlamaIndex) use embeddings:
 - ▶ e.g. S-BERT-type embeddings from SentenceTransformers.
 - ▶ deal with short context by segmenting documents into overlapping chunks

RAG: How to vectorize the documents

- ▶ Traditional approach is **bm25**:
 - ▶ roughly, rank all documents by their TF-IDF cosine similarity to the query.
 - ▶ index every document by which words it contains, to quickly filter out irrelevant documents.
- ▶ Recent systems (e.g. OpenAI, LlamaIndex) use embeddings:
 - ▶ e.g. S-BERT-type embeddings from SentenceTransformers.
 - ▶ deal with short context by segmenting documents into overlapping chunks
 - ▶ then can do fast approximate search over dense vectors (e.g. Faiss, Johnson et al 2017)

RAG doesn't always work

Air Canada must honor refund policy invented by airline's chatbot

Air Canada appears to have quietly killed its costly chatbot support.

ASHLEY BELANGER - 2/16/2024, 6:12 PM



On the day Jake Moffatt's grandmother died, Moffat immediately visited Air Canada's website to book a flight from Vancouver to Toronto. Unsure of how Air Canada's bereavement rates worked, Moffatt asked Air Canada's chatbot to explain.

The chatbot provided inaccurate information, encouraging Moffatt to book a flight immediately and then request a refund within 90 days. In reality, Air Canada's policy explicitly stated that the airline will not provide refunds for bereavement travel after the flight is booked. Moffatt dutifully attempted to follow the chatbot's advice and request a refund but was shocked that the request was rejected.

<https://arstechnica.com/tech-policy/2024/02/air-canada-must-honor-refund-policy-invented-by-airlines-chatbot/>

RAG is an active research area

- ▶ bm25 is much faster, but requires exact word overlap.
- ▶ can do bm25 as a first step, then use embeddings to rank top results.

RAG is an active research area

- ▶ bm25 is much faster, but requires exact word overlap.
- ▶ can do bm25 as a first step, then use embeddings to rank top results.
- ▶ can use an LLM to remove false positives and duplicates
- ▶ can use an LLM to suggest modified / alternative queries.

RAG is an active research area

- ▶ bm25 is much faster, but requires exact word overlap.
- ▶ can do bm25 as a first step, then use embeddings to rank top results.
- ▶ can use an LLM to remove false positives and duplicates
- ▶ can use an LLM to suggest modified / alternative queries.
- ▶ other directions:
 - ▶ embedding algorithms that use other information besides document content, e.g. metadata, citation networks.
 - ▶ apply tools from recommender algorithms
 - ▶ do pre-processing on the documents / queries / embeddings

Inference is easier than retrieval

- ▶ modern LLMs are so good at document analysis (local question answering), that the errors from open QA are almost always from the retriever.
- ▶ once the correct evidence is retrieved, the LLM can provide the answer.
- ▶ → high returns to improving RAG at the moment.

Outline

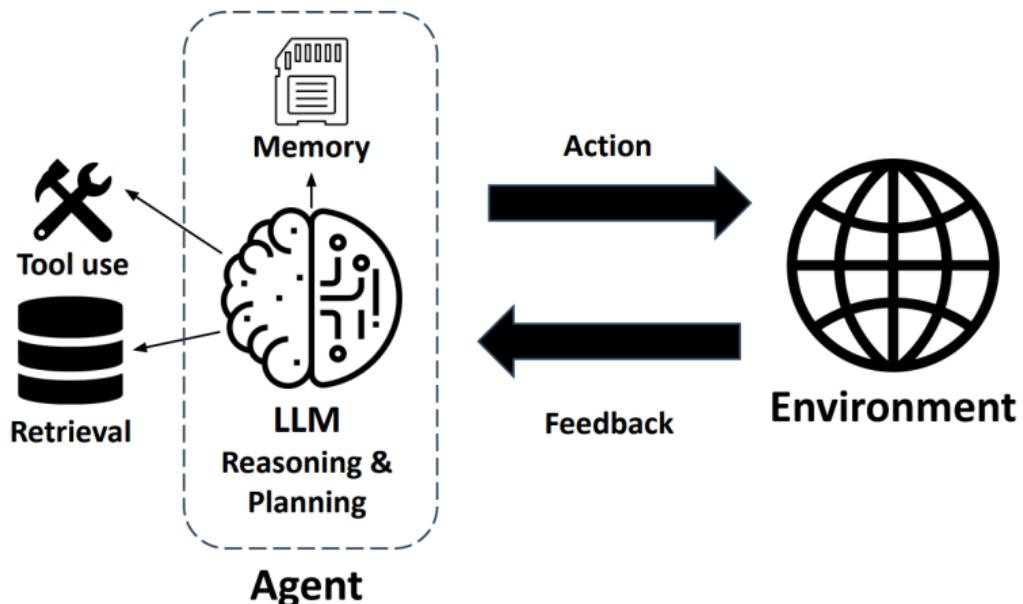
Evaluating LLMs

Information Retrieval / RAG

Agents

Reasoning Models

LLM agents: enabling LLMs to interact with the environment



<https://rdi.berkeley.edu/adv-llm-agents/slides/llm-agents-berkeley-intro-sp25.pdf>

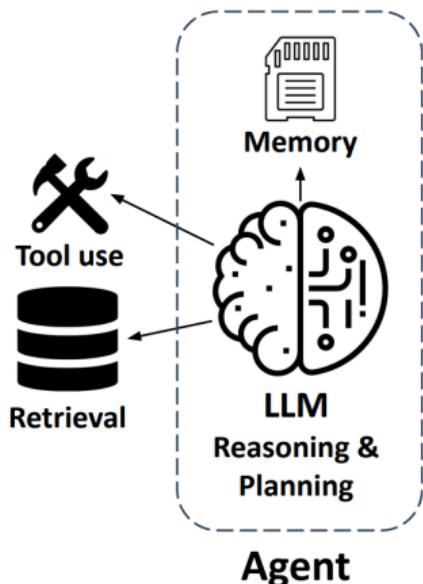
AI Agents

- ▶ AI agent systems (e.g. OpenAI Operator, LangChain) give AI assistants:
 - ▶ the ability/motivation to reason and plan
 - ▶ access to an external environment, e.g. the internet, where they can explore, perceive, and act.
 - ▶ access to tools (e.g. APIs) for interactions.
 - ▶ working memory for connecting/organizing actions and plans in the environment.
 - ▶ the ability/motivation to take actions with the tools in pursuit of plans

AI Agents

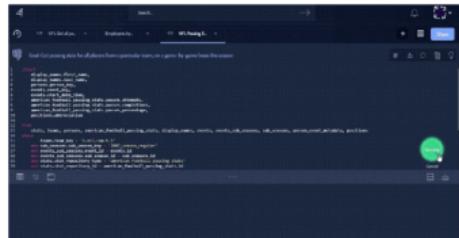
- ▶ AI agent systems (e.g. OpenAI Operator, LangChain) give AI assistants:
 - ▶ the ability/motivation to reason and plan
 - ▶ access to an external environment, e.g. the internet, where they can explore, perceive, and act.
 - ▶ access to tools (e.g. APIs) for interactions.
 - ▶ working memory for connecting/organizing actions and plans in the environment.
 - ▶ the ability/motivation to take actions with the tools in pursuit of plans
- ▶ RAG is an example of an agentic tool; can also include the code interpreter, web browser, etc.

Why empowering LLMs with the agent framework



- Solving real-world tasks typically involves a trial-and-error process
- Leveraging external tools and retrieving from external knowledge expand LLM's capabilities
- Agent workflow facilitates complex tasks
 - Task decomposition
 - Allocation of subtasks to specialized modules
 - Division of labor for project collaboration
 - Multi-agent generation inspires better responses

LLM agents transformed various applications



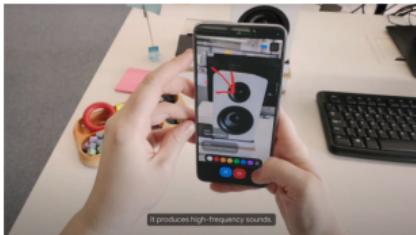
Code generation

Cursor, GitHub Copilot, Devin, Google Jules...



Computer use

Anthropic Claude, Google Jarvis, OpenAI Operator



Personal assistant

Google Astra, OpenAI GPT-4o,...



Robotics

Figure AI, Tesla Optimus, NVIDIA GR00T...

- Education
 - Law
 - Finance
 - Healthcare
 - Cybersecurity
- ...

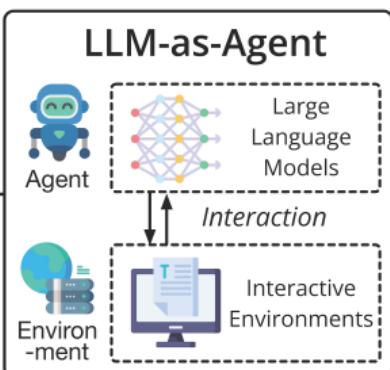
Doing NLP with AI Agents

<https://chatgpt.com/share/68189e4d-7018-8005-abe7-ee67c46065bc>

Evaluating Agents

Real-world Challenges

- (On an Ubuntu bash terminal)
Recursively set all files in the directory to read-only, except those of mine.
- (Given Freebase APIs)
What musical instruments do Minnesota-born Nobel Prize winners play?
- (Given MySQL APIs and existed tables)
Grade students over 60 as PASS in the table.
- (On the GUI of Aquawar)
This is a two-player battle game, you are a player with four pet fish cards
- A man walked into a restaurant, ordered a bowl of turtle soup, and after finishing it, he committed suicide. Why did he do that?
- (In the middle of a kitchen in a simulator)
Please put a pan on the dinning table.
- (On the official website of an airline)
Book the cheapest flight from Beijing to Los Angeles in the last week of July.



8 Distinct Environments



<https://github.com/THUDM/AgentBench>

Example Tasks in WebArena



“Create a plan to visit Pittsburgh’s art museums with minimal driving distance starting from Schenley Park. Log the order in my “awesome-northeast-us-travel” repository ”

webarena.wikipedia.com

Wikipedia Pittsburgh museums

List of museums in Pittsburgh

This list of museums in [Pittsburgh, Pennsylvania](#) encompasses museums defined for this context as institutions (including [nonprofit](#) organizations, government entities, and private [businesses](#)) that collect and care for objects of cultural, artistic, scientific, or historical interest and make their collections or related exhibits available for public viewing. Also included are university and non-profit art galleries. Museums that exist only in cyberspace (i.e., [virtual museums](#)) are not included.

Wikimedia Commons has media related to [Museums in Pittsburgh](#).

See also: [List of museums in Pennsylvania](#)

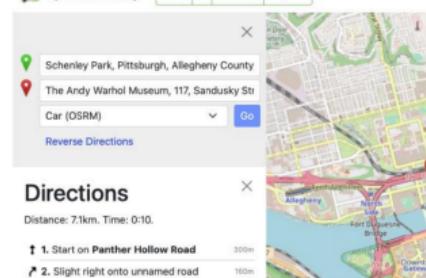
Museums



Search for museums
in Pittsburgh

webarena.openstreetmap.com

OpenStreetMap [Edit](#) [History](#) [Export](#)



Directions

Distance: 7.1km, Time: 0:10.

1. Start on Panther Hollow Road

300m

2. Slight right onto unnamed road

180m

webarena.gitlab.com

Update README.md

[README.md](#) 158 B [Edit](#) [Replace](#)

Travel in Northeast US

Pittsburgh

- + Miller Gallery at Carnegie Mellon University
- + American Jewish Museum
- + Carnegie Museum of Art

Search for each art museum on the Map

Record the optimized results to the repo

<https://phontron.com/class/anlp2024/assets/slides/anlp-20-agents.pdf>

Outline

Evaluating LLMs

Information Retrieval / RAG

Agents

Reasoning Models

How LLMs got good at math

Category

Math

Apply filter

Style Control

Show Deprecated

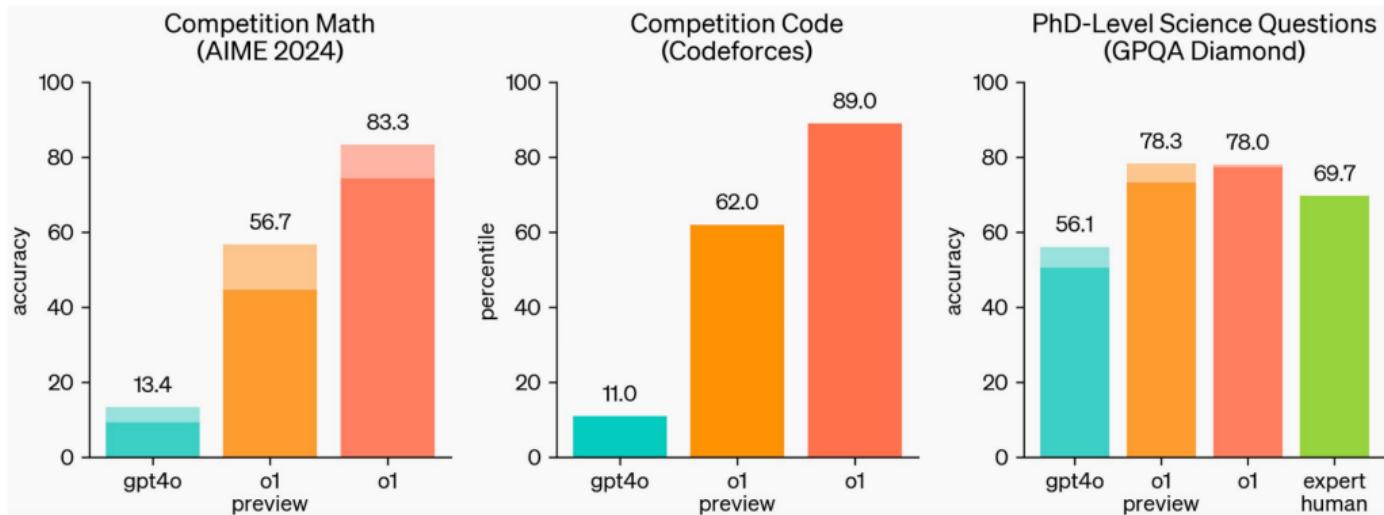
Math

#models: 227 (99%) #votes: 384,871 (13%)

Rank* (UB)	Delta	Model	Arena Score	95% CI	Votes	Organization	License
1	0	Gemini-2.5-Pro-Exp-03-25	1417	+17/-20	1152	Google	Proprietary
1	1	o3-2025-04-16	1403	+39/-42	241	OpenAI	Proprietary
1	8	o4-mini-2025-04-16	1400	+43/-36	200	OpenAI	Proprietary
2	2	GPT-4.5-Preview	1377	+12/-17	1626	OpenAI	Proprietary
2	1	Gemini-2.5-Flash-Preview-04-17	1363	+20/-22	453	Google	Proprietary
2	7	DeepSeek-R1	1361	+13/-14	1601	DeepSeek	MIT
2	15	o3-mini-high	1359	+12/-13	2032	OpenAI	Proprietary
3	8	o1-2024-12-17	1354	+9/-9	3252	OpenAI	Proprietary

What is special about o*, Gemini 2.5, and DeepSeek-R1?

Highlight of LLMs in 2024: the advancement of reasoning models



OpenAI o1 started to achieve impressive performance across various challenging reasoning tasks in math, coding, STEM, etc.

https://rdi.berkeley.edu/adv-llm-agents/slides/inference_time_techniques_lecture_sp25.pdf

OpenAI o1 demo: hidden thought before generating the solution

User query

You are visiting San Francisco for the day and want to meet as many friends as possible. Solve the problem by considering various different schedules and picking the best one to optimize your goals.

Travel distances (in minutes):
Bayview to Alamo Square: 16.
Bayview to Union Square: 17.
Bayview to Golden Gate Park: 22.
Alamo Square to Bayview: 16.
Alamo Square to Union Square: 14.
Alamo Square to Golden Gate Park: 9.
Union Square to Bayview: 15.
Union Square to Alamo Square: 15.
Union Square to Golden Gate Park: 22.
Golden Gate Park to Bayview: 23.
Golden Gate Park to Alamo Square: 10.
Golden Gate Park to Union Square: 22.

CONSTRAINTS: You arrive at Bayview at 9:00AM. Ronald will be at Alamo Square from 8:30AM to 7:45PM. You'd like to meet Ronald for a minimum of 90 minutes. Richard will be at Union Square from 2:30PM to 9:45PM. You'd like to meet Richard for a minimum of 30 minutes. Kenneth will be at Golden Gate Park from 10:00AM to 3:15PM. You'd like to meet Kenneth for a minimum of 60 minutes.



Thought about optimizing meeting schedules for 1m 14s

Below is one clear way to fit in all three visits without wasting time, together with the travel times and the required minimum meeting windows. The key insights are:

Details

Crafting meeting logistics

I'm collaborating on a schedule to meet Ronald, Richard, and Kenneth, ensuring each meeting fits within their specified time slots, factoring in travel time and optimal routing.

Mapping travel logistics

Evaluating the best route to Golden Gate Park first, considering travel time and waiting for Kenneth. Balancing early arrival against travel efficiency.

Crafting the itinerary

Piecing together a travel plan, starting with Kenneth, then Ronald, and ending with Richard. Assessing the optimal sequence and meeting durations.

Mapping travel routes

OK, let me see. Traveling from Golden Gate Park to Alamo Square, I'm thinking about extending Ronald's meeting to 1.5 hours then moving to Union Square to meet Richard.

Evaluating options

I'm weighing the benefits of a longer Ronald meeting or a later one to avoid waiting for Richard. The goal is to strike a good balance.

Balancing schedules

I'm organizing a plan to visit Alamo Square first, then head to Golden Gate Park to meet Kenneth. This sequence maximizes time management and minimizes travel between meetings.

Streamlining the timeline

I'm thinking through the optimal approach: meeting Ronald for an uninterrupted 90 minutes to avoid waiting for Richard and making efficient use of the available time before 2:30.

Tweaking the schedule

Meeting Ronald for 90 minutes as scheduled, then quickly traveling to Union Square. This reduces waiting time significantly.

Mapping the timeline

Summarized hidden thought

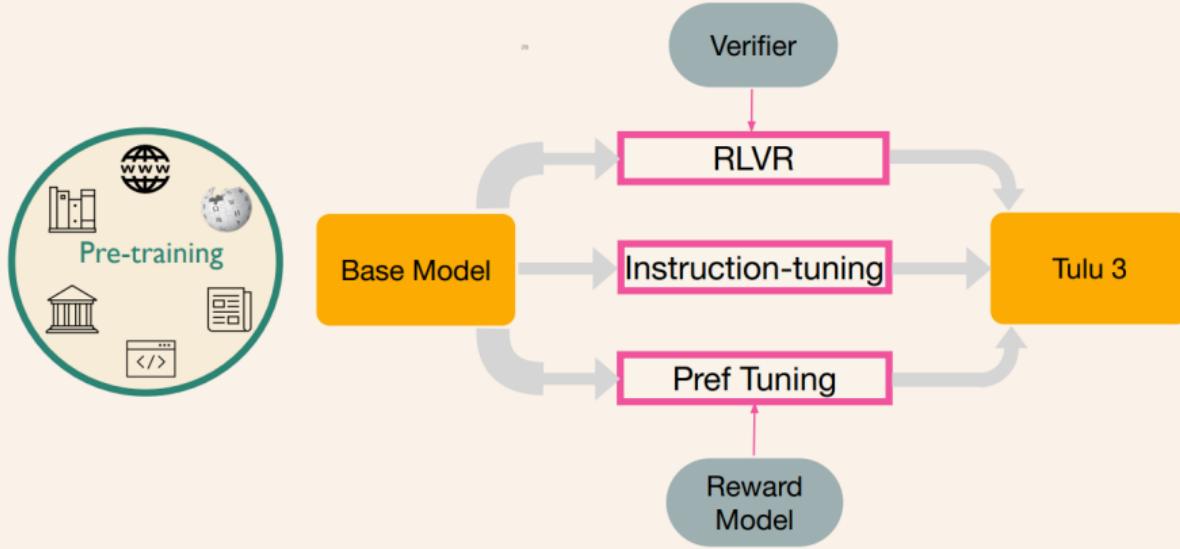
Model-generated response

What is added by “reasoning” models

<https://www.interconnects.ai/p/deepseek-r1-recipe-for-o1>

- ▶ Reasoning models like deepseek-r1 add a few new ingredients:
 1. allow the model to “think” and “reflect”, generating lots of internally observed tokens before outputting an external answer to the user.
 2. additional SFT on a lot of examples of internal reflections with step-by-step verifications on complex questions.
 3. Reinforcement Learning from Verifiable Rewards (RLVR): LLM solves math problems by guessing and verifying, gets a reward when correct, gradually improves.

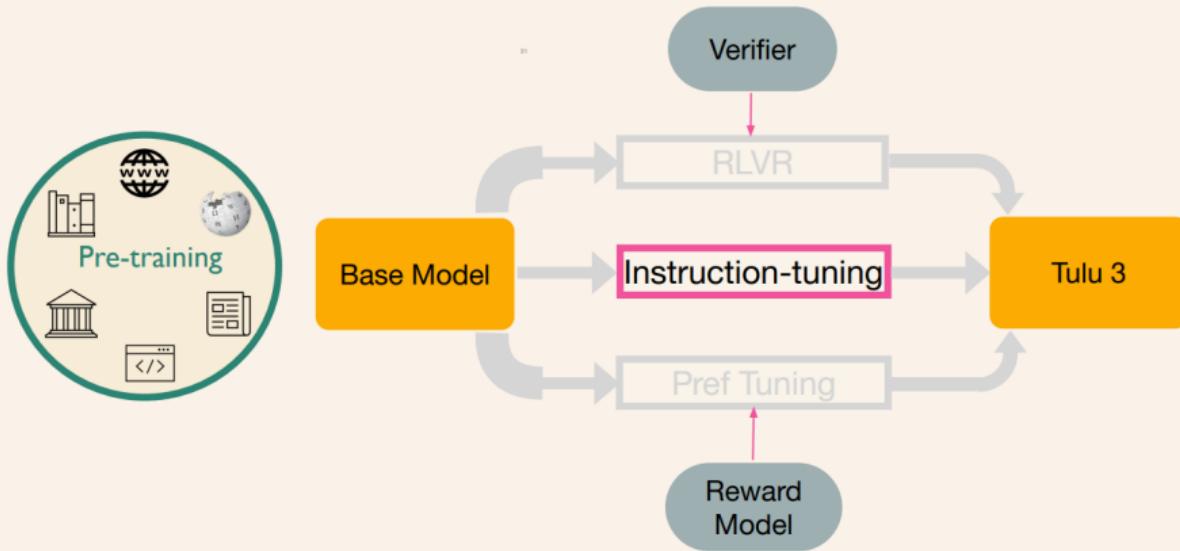
Tulu 3 Training Recipe



[https://rdi.berkeley.edu/adv-llm-agents/slides/
OLMo-Tulu-Reasoning-Hanna.pdf](https://rdi.berkeley.edu/adv-llm-agents/slides/OLMo-Tulu-Reasoning-Hanna.pdf)

Tulu 3

Supervised Finetuning (a.k.a Instruction Tuning)



Why Chain-of-Thought data for reasoning?

Chain of Thought data

- 👍 helps models handle complex, multi-step problems easier
- 👍 reveals the model's reasoning process
- 👍 makes it easier to spot errors in logic thus more trustworthy
- 👍 resembles human thought process

But ...

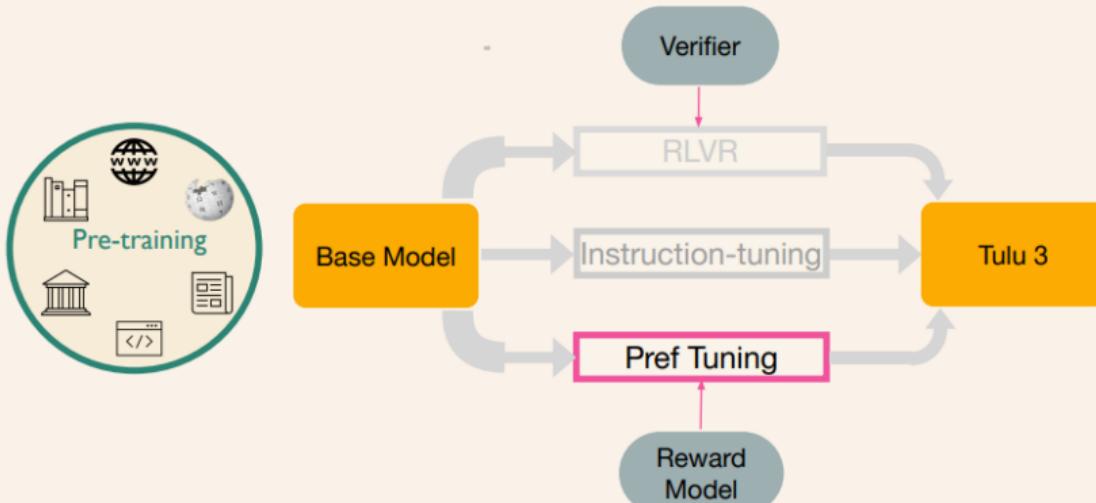
👎 Manual annotation challenges:

- time and cost intensive
- often requires expert annotations
- Difficult to scale

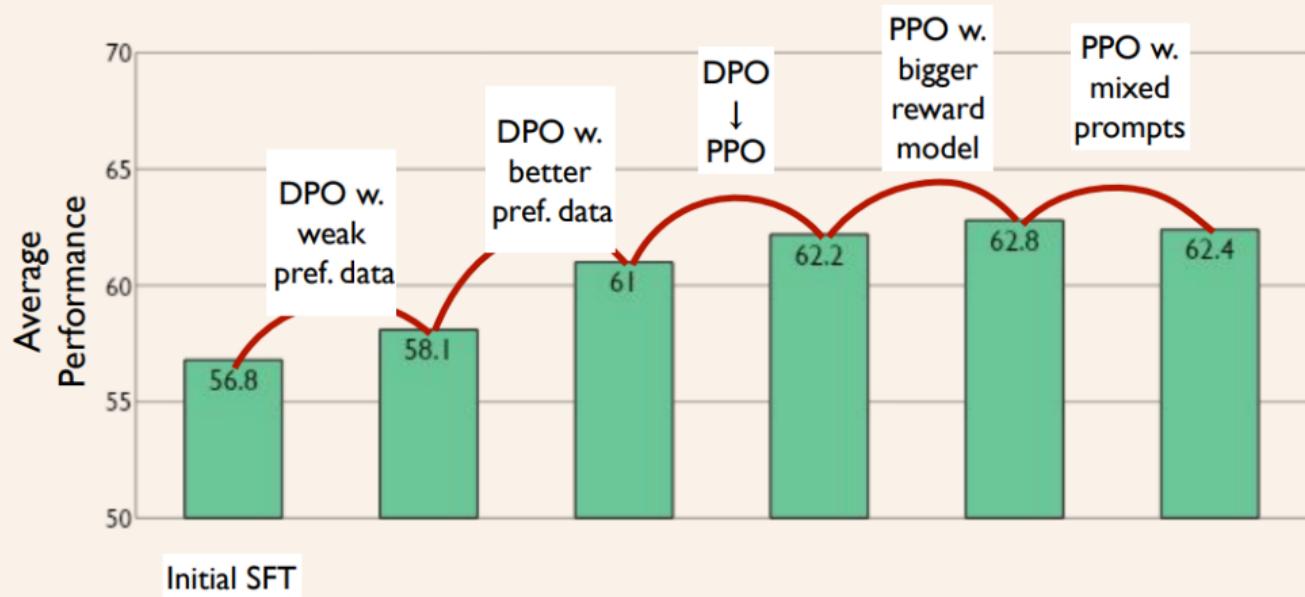
How to generate Chain-of-Thought Data

1. Manual Human Annotation (e.g., GSM8K dataset): Annotators write step by step solutions
 - High-quality reasoning traces
 - Limited scale (only 7K)
 - Lack of diversity in reasoning styles
2. Program-Aided Language Models (PAL): Convert math problems into Python code execution traces
 - Guarantee correctness through execution
 - Less natural language reasoning, less intuitive
 - Limited to problems that can be coded
3. Self-generated COT (self-ask): using LLMs to generate their reasoning paths
 - Scalable to many problems
 - Quality highly dependent on base model

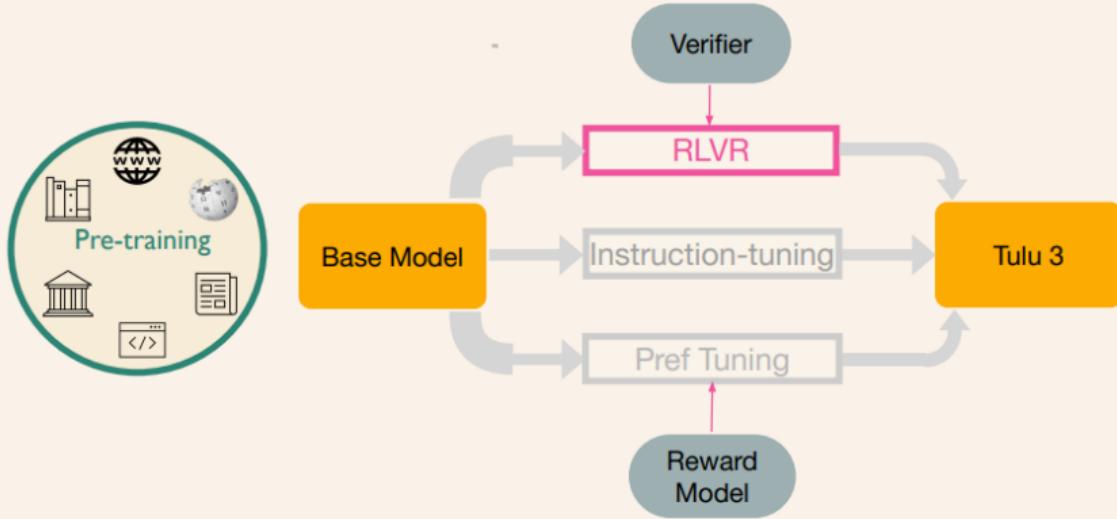
❖ Tulu 3 Step 2: Preference tuning



What components matter for LMs?



❖ Tulu 3 Step 3: RLVR



Simplifying the reward model: rule-based rewards

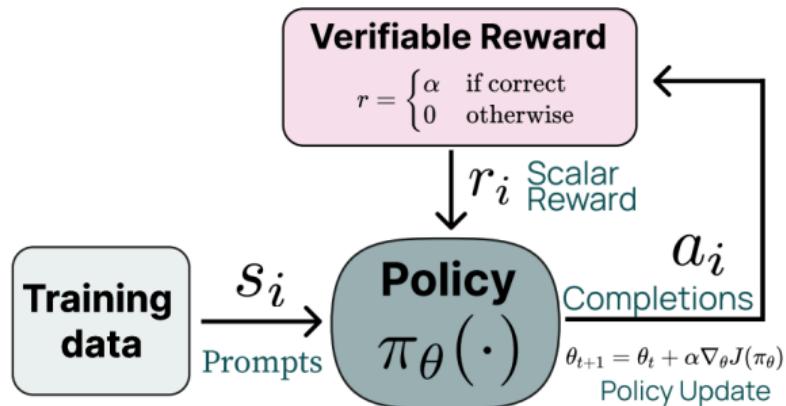
What is
 $2+2$? 4.

```
if answer == gold label:  
    return 1  
else:  
    return 0
```

Score: 1

Can we just remove this complex setup and use simpler 'models'...?

RLVR: Reinforcement Learning from Verifiable Rewards



- ▶ RLVR: confront the LLM with thousands of math/logic problems, allow it to guess, reflect, and verify until finding verifiable solutions; if it is correct, it gets utility.
 - ▶ reduces hallucinations by teaching model to self-critique

2.2.2. Reward Modeling

The reward is the source of the training signal, which decides the optimization direction of RL. To train DeepSeek-R1-Zero, we adopt a rule-based reward system that mainly consists of two types of rewards:

- **Accuracy rewards:** The accuracy reward model evaluates whether the response is correct. For example, in the case of math problems with deterministic results, the model is required to provide the final answer in a specified format (e.g., within a box), enabling reliable rule-based verification of correctness. Similarly, for LeetCode problems, a compiler can be used to generate feedback based on predefined test cases.
 - **Format rewards:** In addition to the accuracy reward model, we employ a format reward model that enforces the model to put its thinking process between '<think>' and '</think>' tags.
-
- During post-training, DeepSeek is aligned on both RLHF and RLVR objectives.

chat.deepseek.com

General principle of how to design effective reasoning techniques

- The Bitter Lesson from Richard Sutton (again) is an important guideline for designing reasoning techniques, including both inference-time and training-time algorithms.
- “One thing that should be learned from the bitter lesson is the great power of general purpose methods, of methods that **continue to scale with increased computation** even as the available computation becomes very great.”
- “We want AI agents that can **discover like we can, not which contain what we have discovered**. Building in our discoveries only makes it harder to see how the discovering process can be done.”

https://rdi.berkeley.edu/adv-l1m-agents/slides/inference_time_techniques_lecture_sp25.pdf