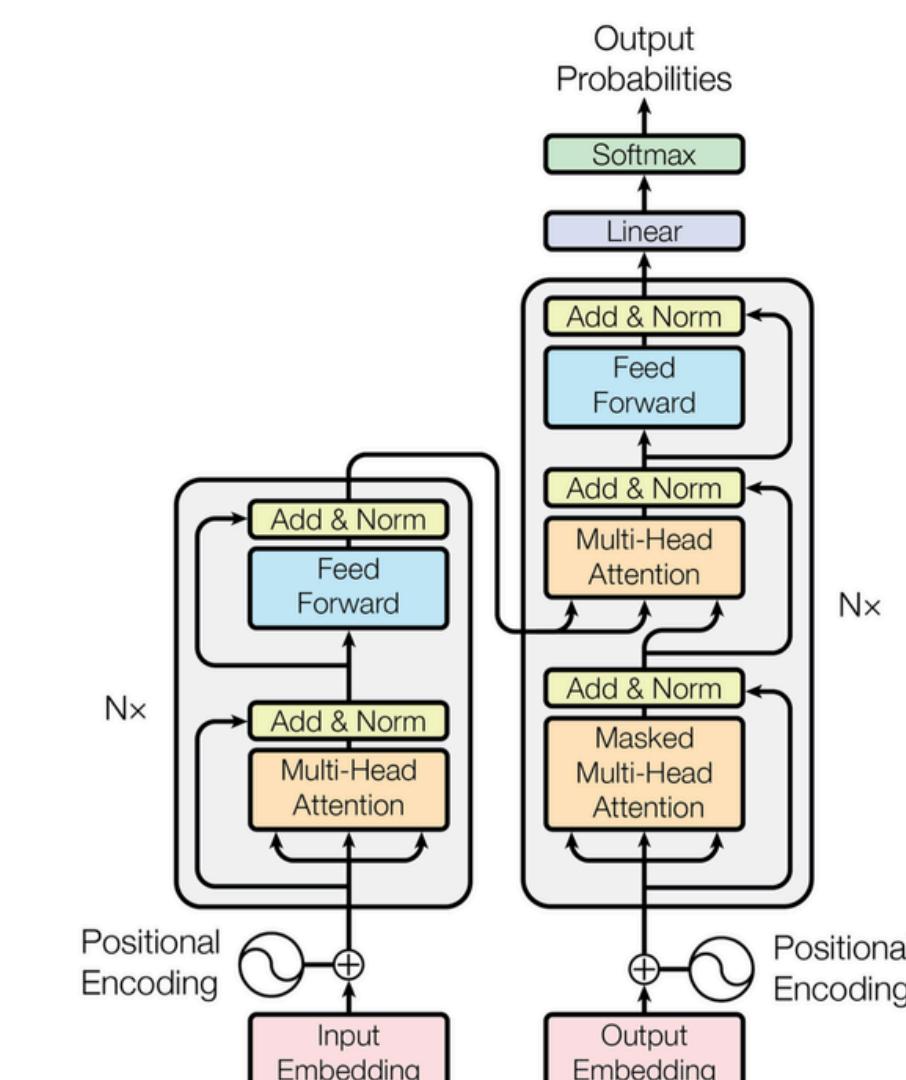
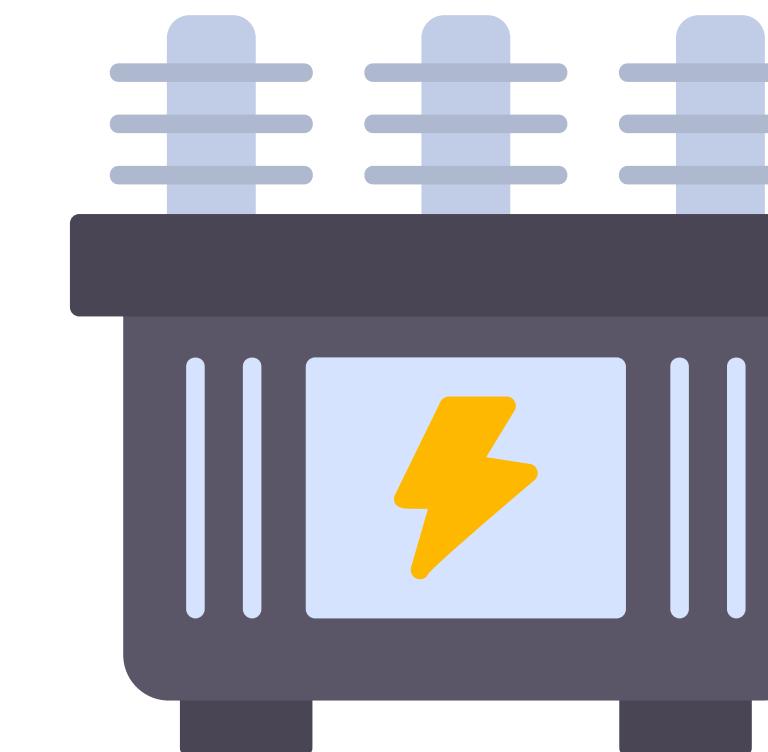


Day 3

Advanced AI Topics

Avishek Shrestha
Head of AI
Naamche



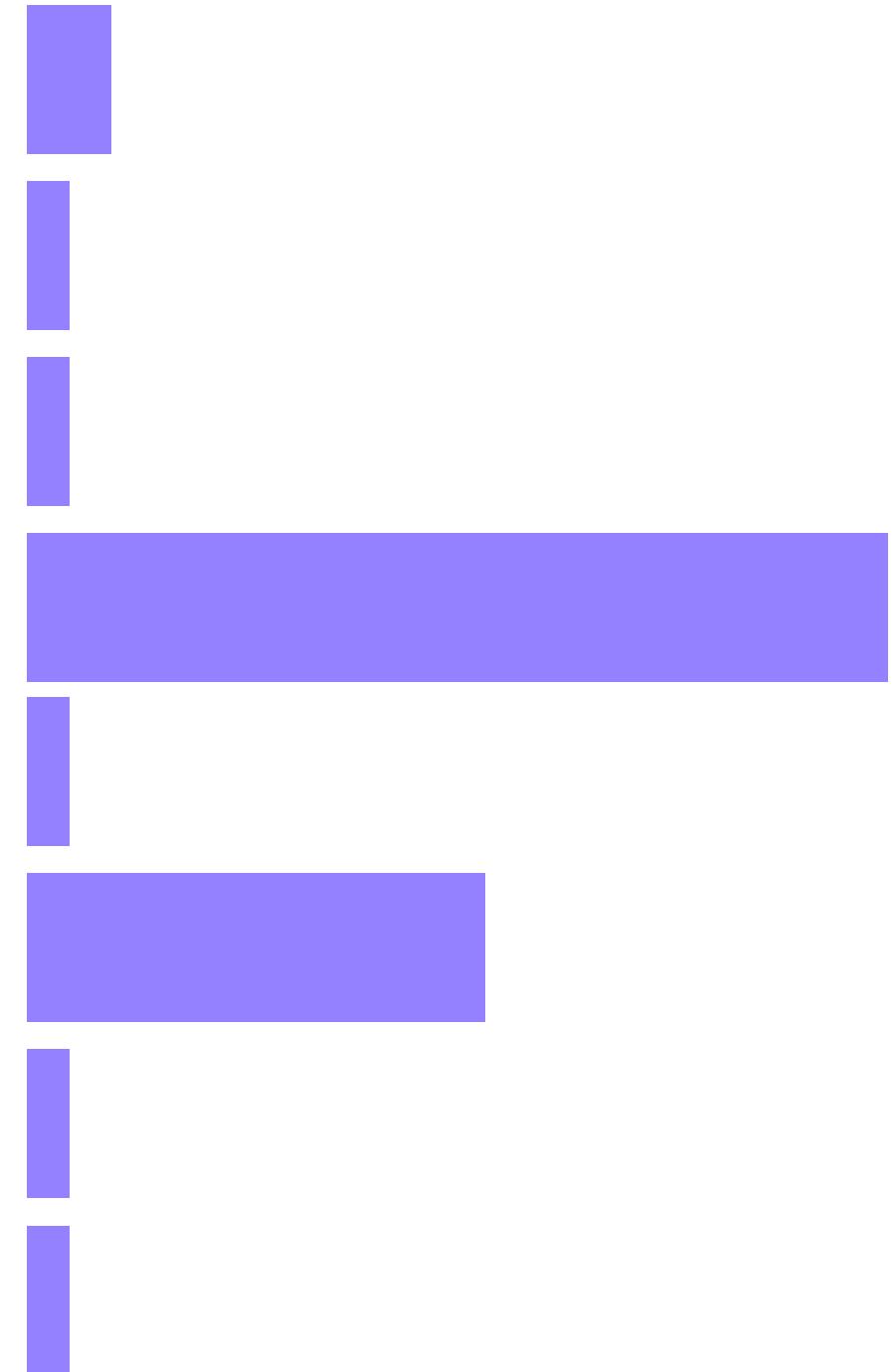
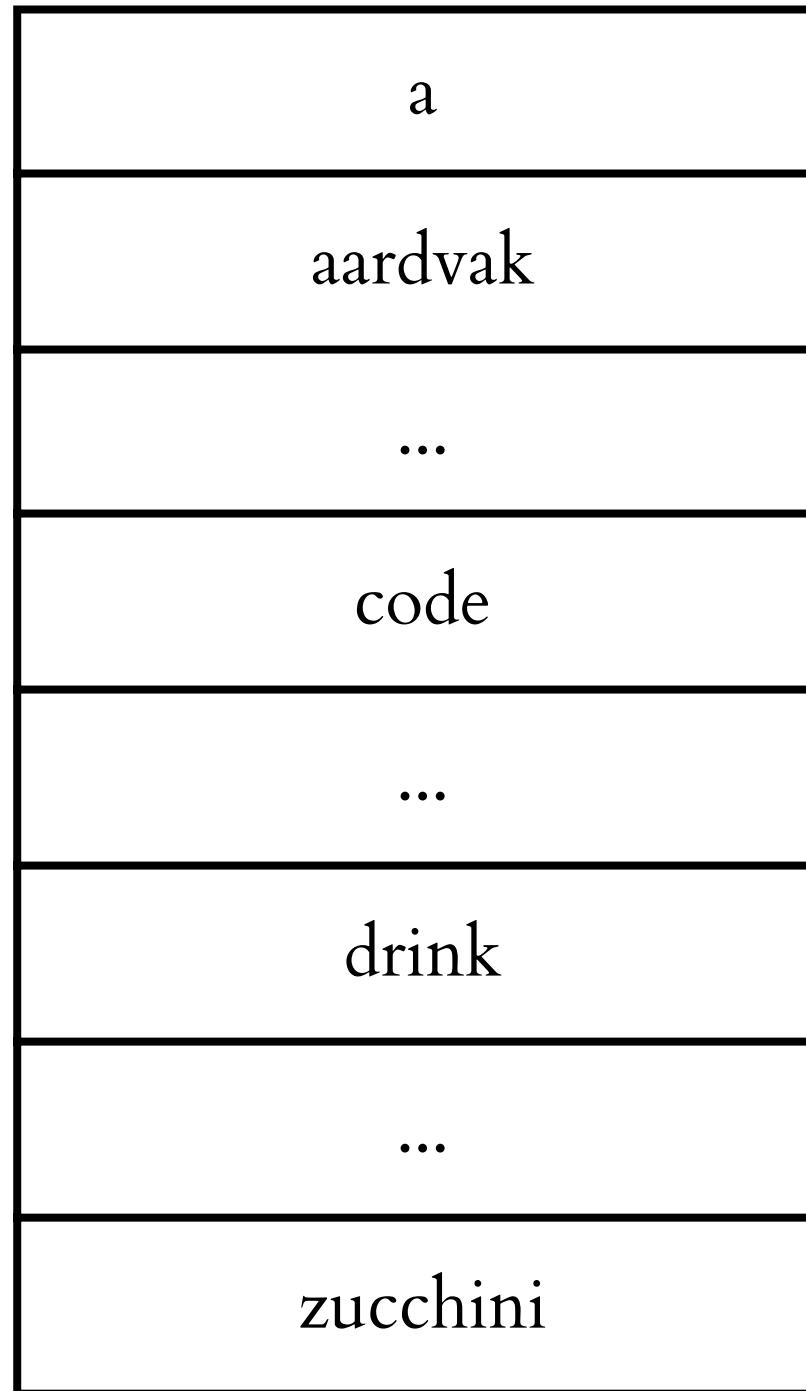


It's raining cats and ____.



It's raining cats and dogs.

**On weekends,
programmers like
to**



“Goodness” of the
model is how close its
prediction of the actual
next word is to 1.0

Next Word Prediction

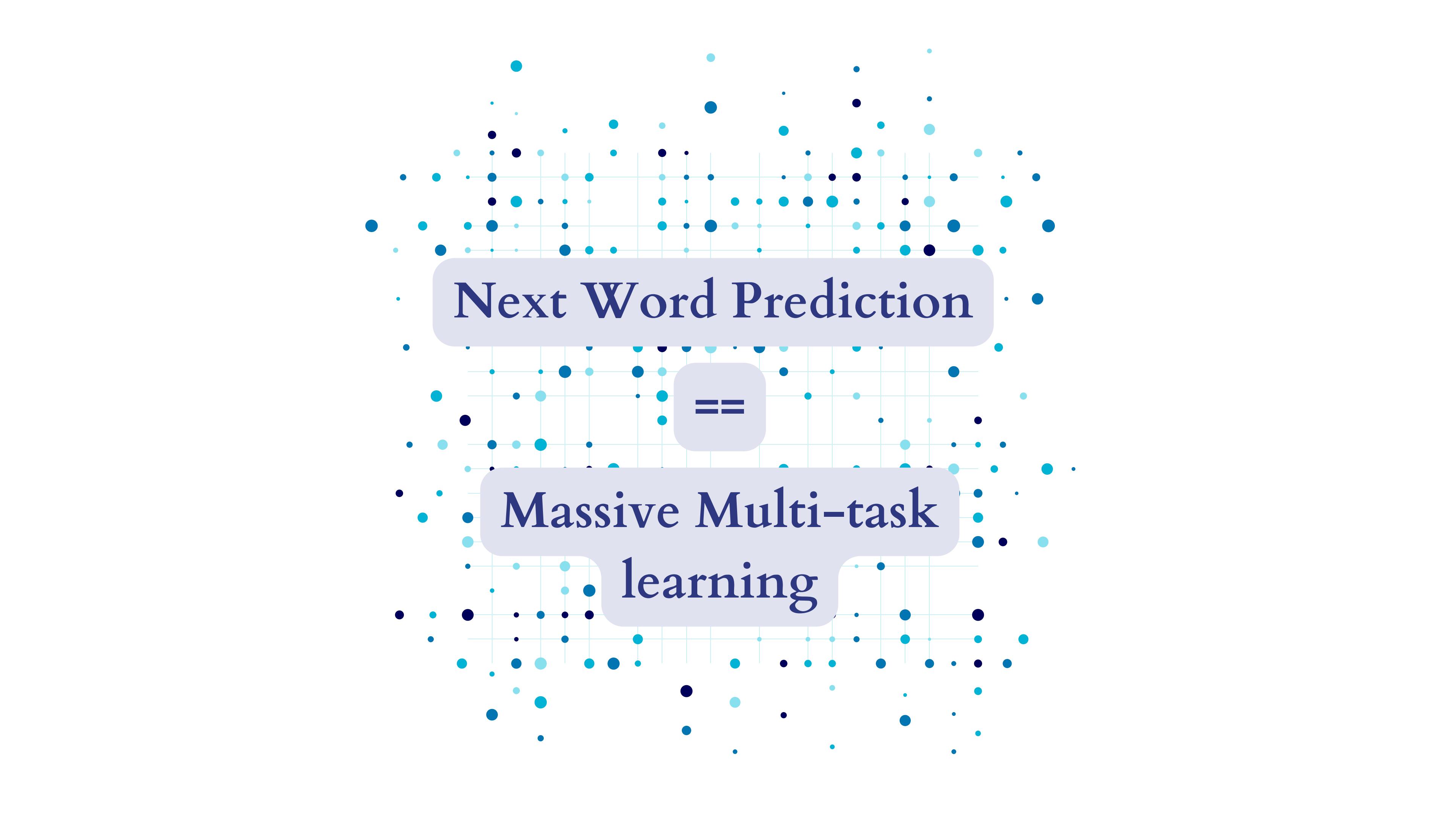
=

powerful thing

Can LLMs reason?



The criminal is _____

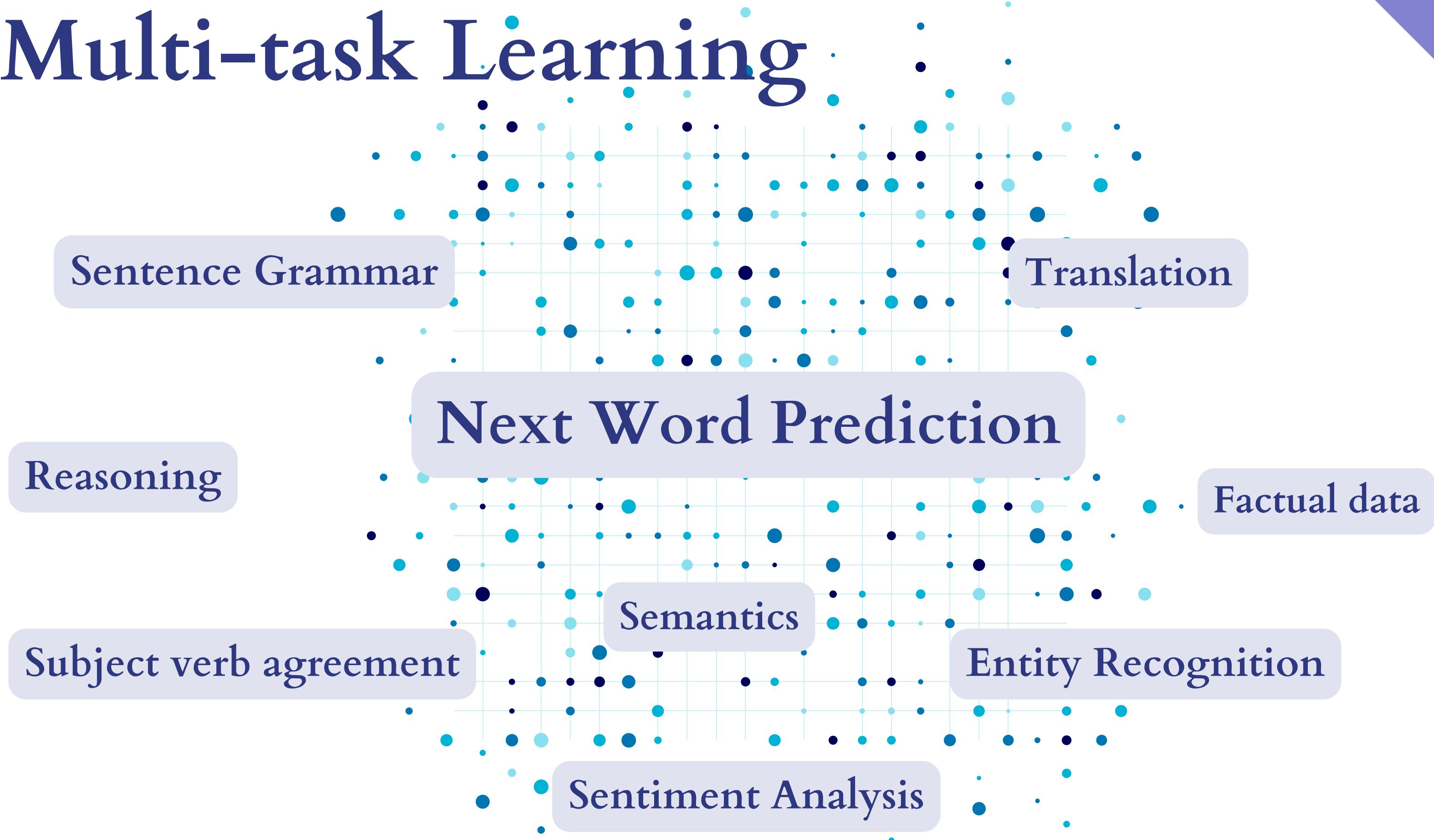


Next Word Prediction



Massive Multi-task
learning

Multi-task Learning



Task	Example sentence in pre-training that would teach that task
Grammar	In my free time, I like to {code, banana}
Lexical semantics	I went to the store to buy milk, eggs, and {bread, cat}.
World Knowledge	The hight of Mt. Everest is {8848m, 8611m}
Sentiment Analysis	Movie review: I was engaged and on the edge of my seat the whole time. The movie was {good, bad}.
Translation	The word for 'pretty' in Nepali is {sundar, naramro}.
Math Question	Arithmetic exam answer key: $3 + 8 + 4 = 15$.



BIG-bench-actions-robot and BIG-bench-actions-robot auto-generate task summary tables, analysis, SeqIO task catalog, an... 95308e1 · 2 years ago History

Preview

Code

Blame

237 lines (230 loc) · 96 KB



Raw

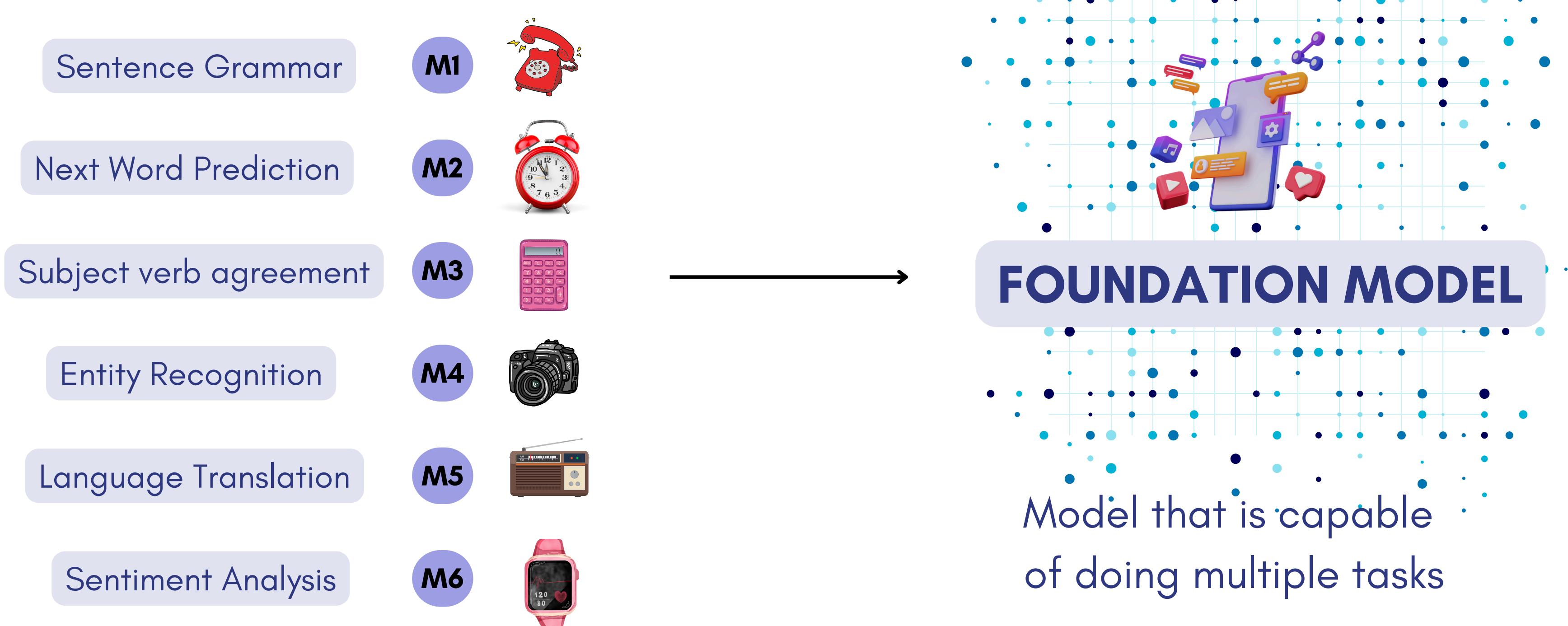


This directory contains tasks that are part of the Beyond the Imitation Game benchmark (BIG-bench). Each subdirectory contains a single benchmark task. There are 214 tasks in total. These are summarized in the following table:

Name	Description	Keywords
abstract_narrative_understanding	Given a narrative, choose the most related proverb	analogical reasoning , json , multiple choice , narrative understanding , social reasoning
abstraction_and_reasoning_corpus	Solve tasks from Abstraction and Reasoning Corpus	free response , many-shot , non-language , numerical response , programmatic , visual reasoning , zero-shot
anachronisms	Identify whether a given statement contains an anachronism	common sense , implicit reasoning , json , multiple choice , word sense disambiguation
analogical_similarity	Identify the type of analogy between two events	analogical reasoning , json , many-shot , multiple choice
analytic_entailment	Identify whether one sentence entails the next	decomposition , fallacy , json , logical reasoning , multiple choice , negation
arithmetic	Perform the four basic arithmetic operations	arithmetic , free response , json , mathematics , multiple choice , numerical response
ascii_word_recognition	Identify the word displayed as ASCII art	context length , free response , json , non-language , visual reasoning

<https://github.com/google/BIG-bench>

Convergence of task-specific models into a new paradigm



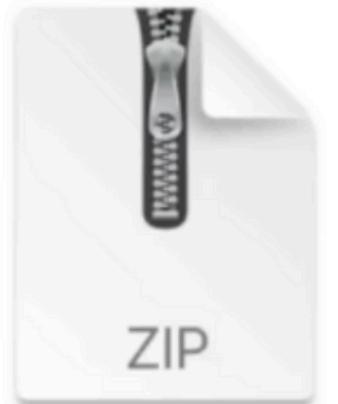
Pre-training



Chunk of the internet,
~10TB of text



6,000 GPUs for 12 days, ~\$2M
~ $1e24$ FLOPS

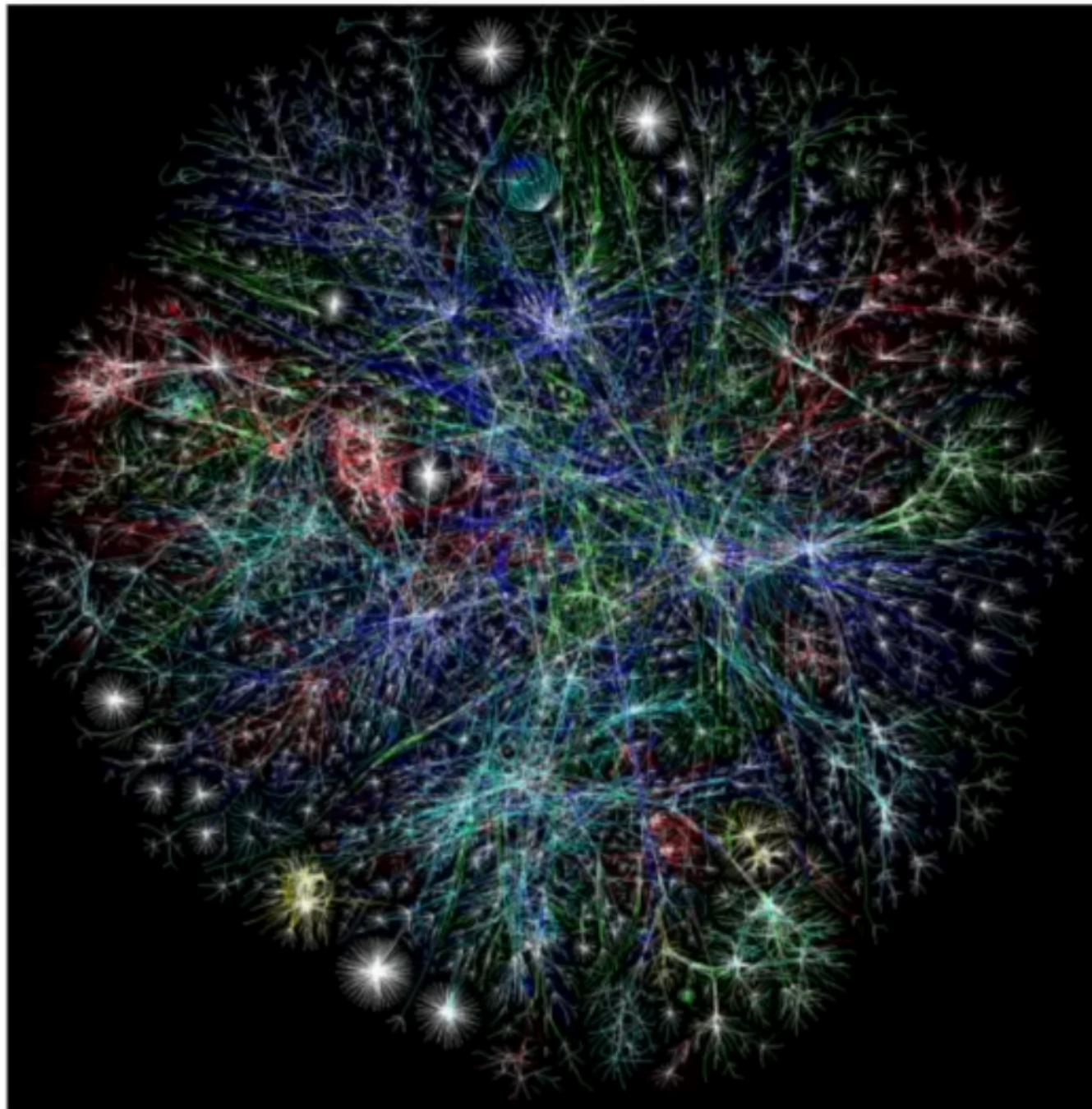


parameters.zip

~140GB file

Training the Assistant

(Just swap the dataset, then continue training)



<USER>

Can you write a short introduction about the relevance of the term "monopsony" in economics? Please use examples related to potential monopsonies in the labour market and cite relevant research.

<ASSISTANT>

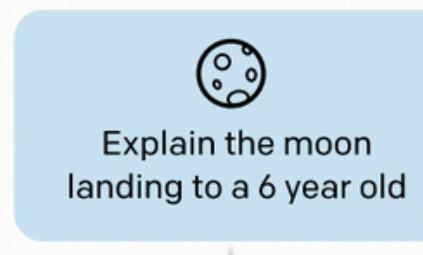
"Monopsony" refers to a market structure where there is only one buyer for a particular good or service. In economics, this term is particularly relevant in the labor market, where a monopsony employer has significant power over the wages and working conditions of their employees. The presence of a monopsony can result in lower wages and reduced employment opportunities for workers, as the employer has little incentive to increase wages or provide better working conditions...

~100K conversations

Step 1

Collect demonstration data, and train a supervised policy.

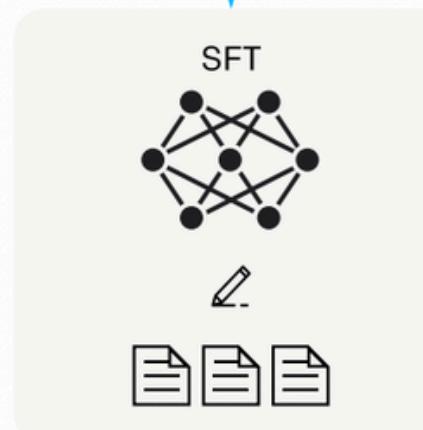
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



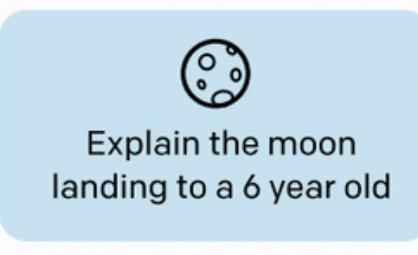
This data is used to fine-tune GPT-3 with supervised learning.



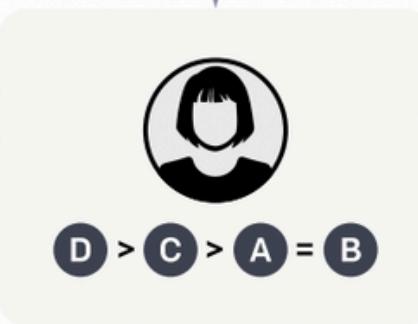
Step 2

Collect comparison data, and train a reward model.

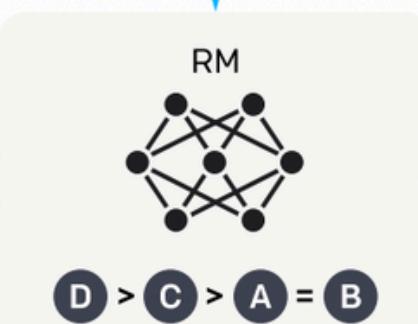
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



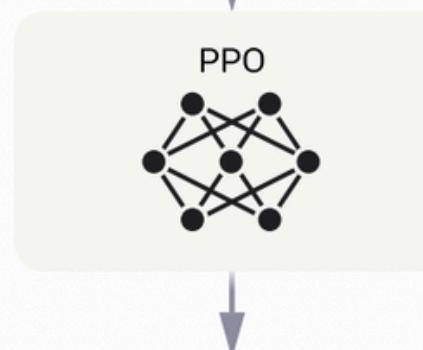
Step 3

Optimize a policy against the reward model using reinforcement learning.

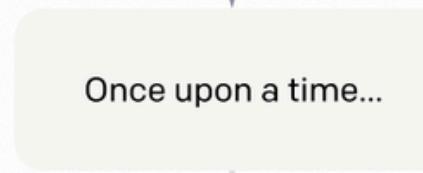
A new prompt is sampled from the dataset.



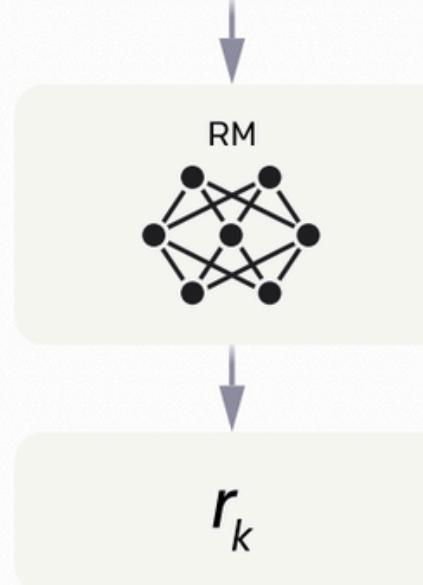
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



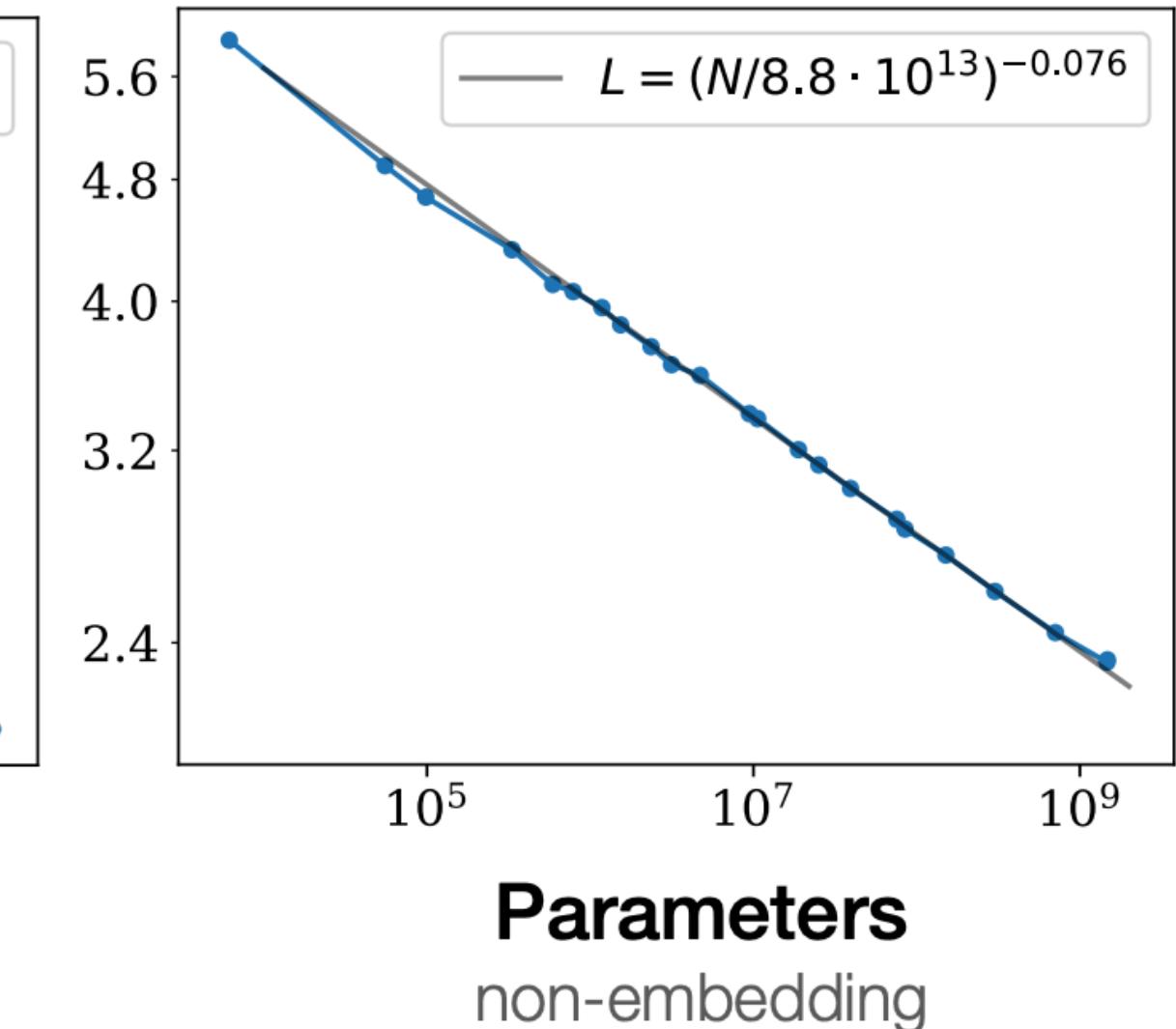
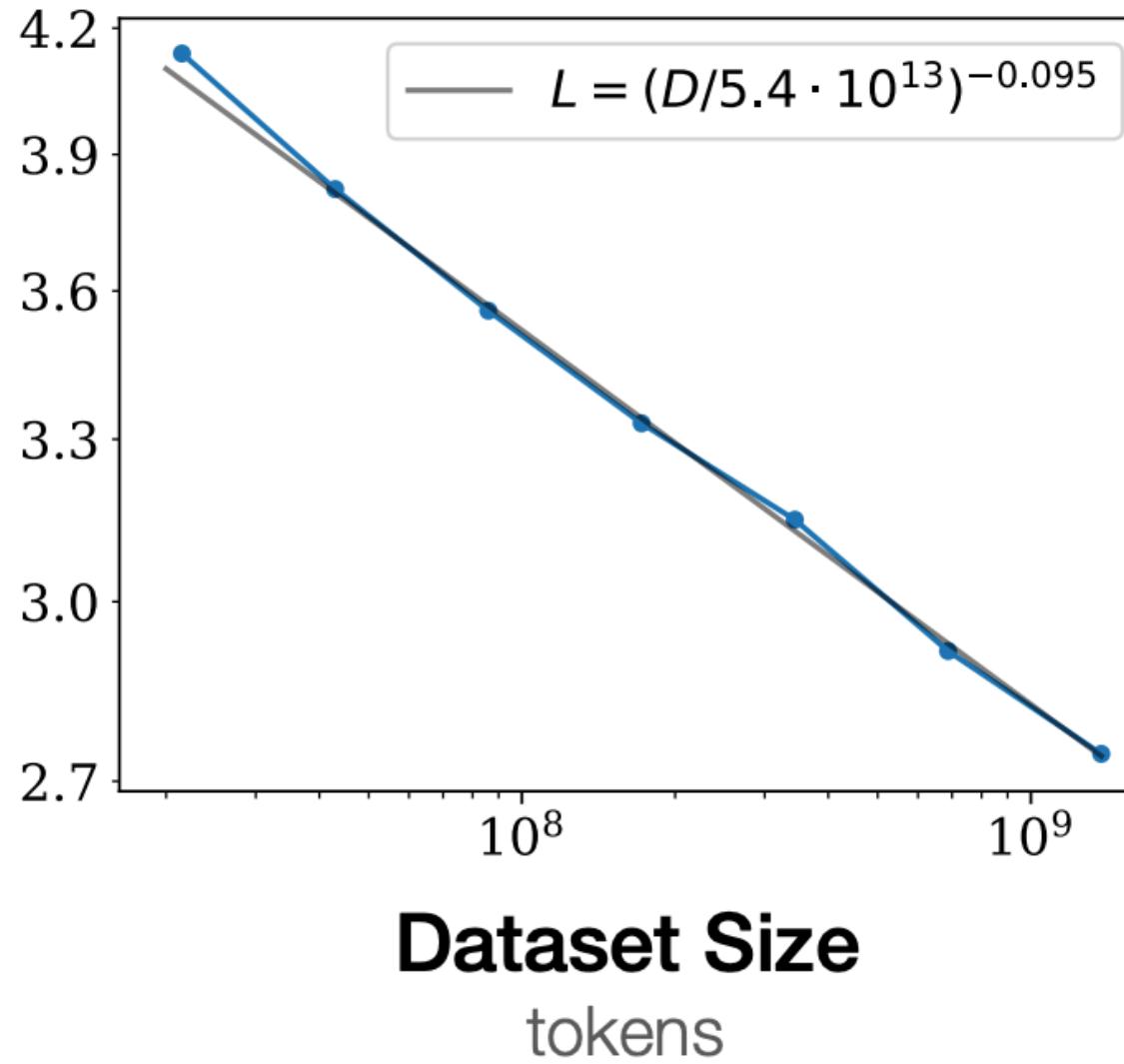
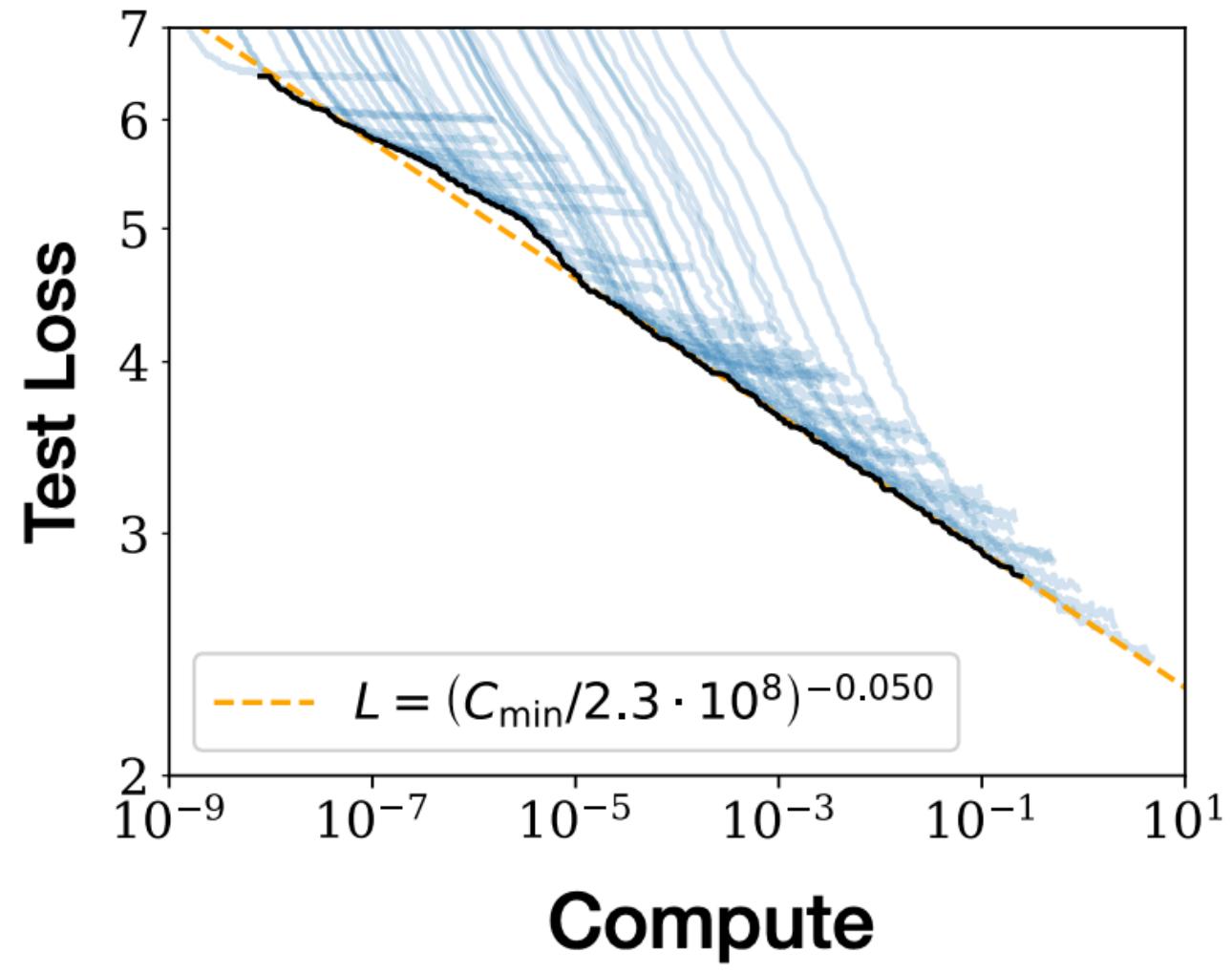


Figure 1 Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute² used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Reasoning Models

Chain of Thought

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27.



Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

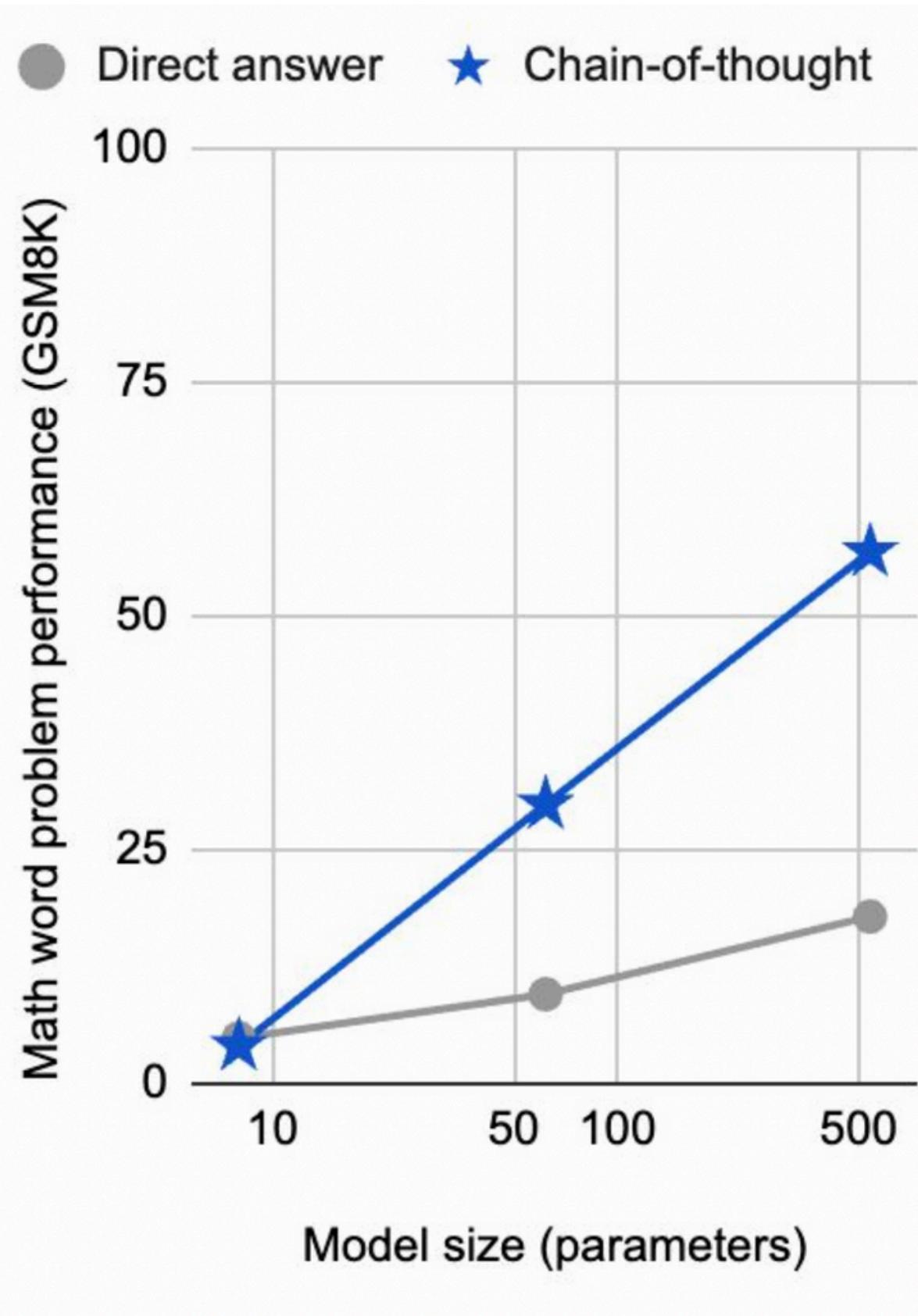
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have? Think step by step.

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9.





Chain-of-thought prompting elicits reasoning in large language models.

Wei et al., 2022. 25

<https://arxiv.org/abs/2201.11903>

Self-Refine

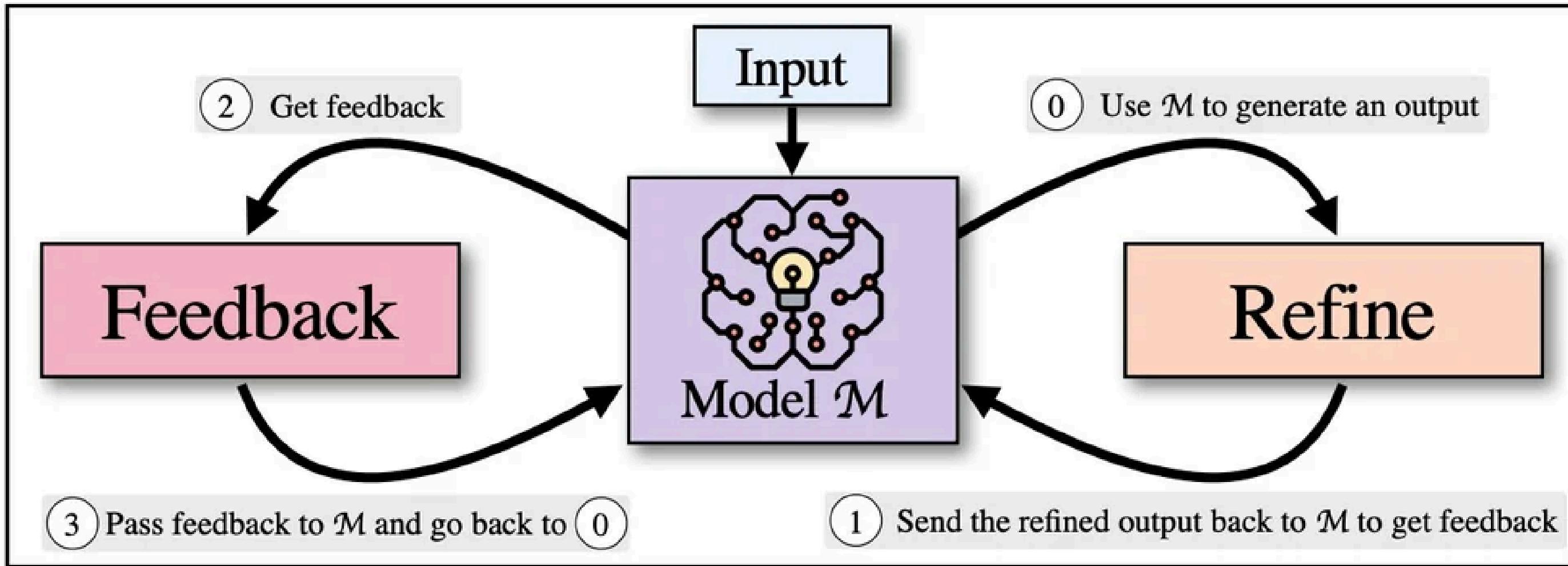
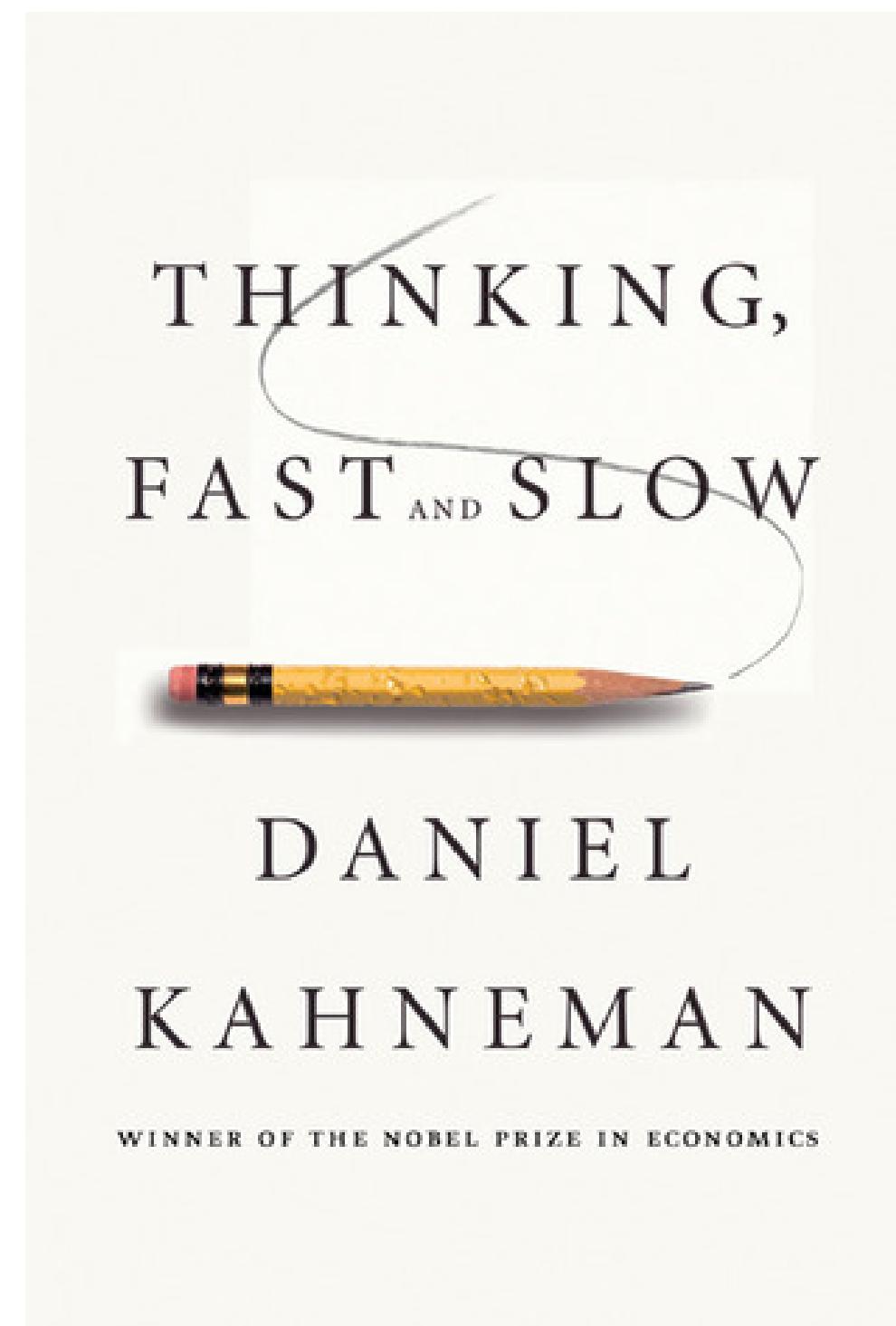
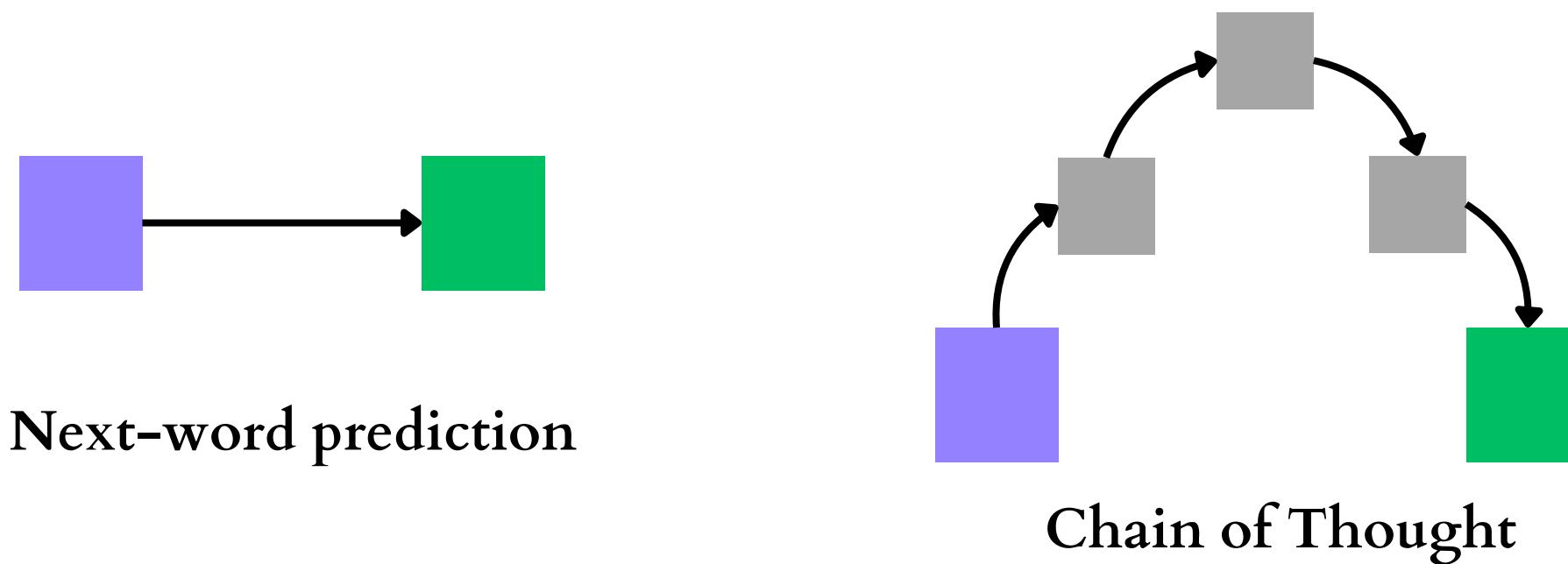
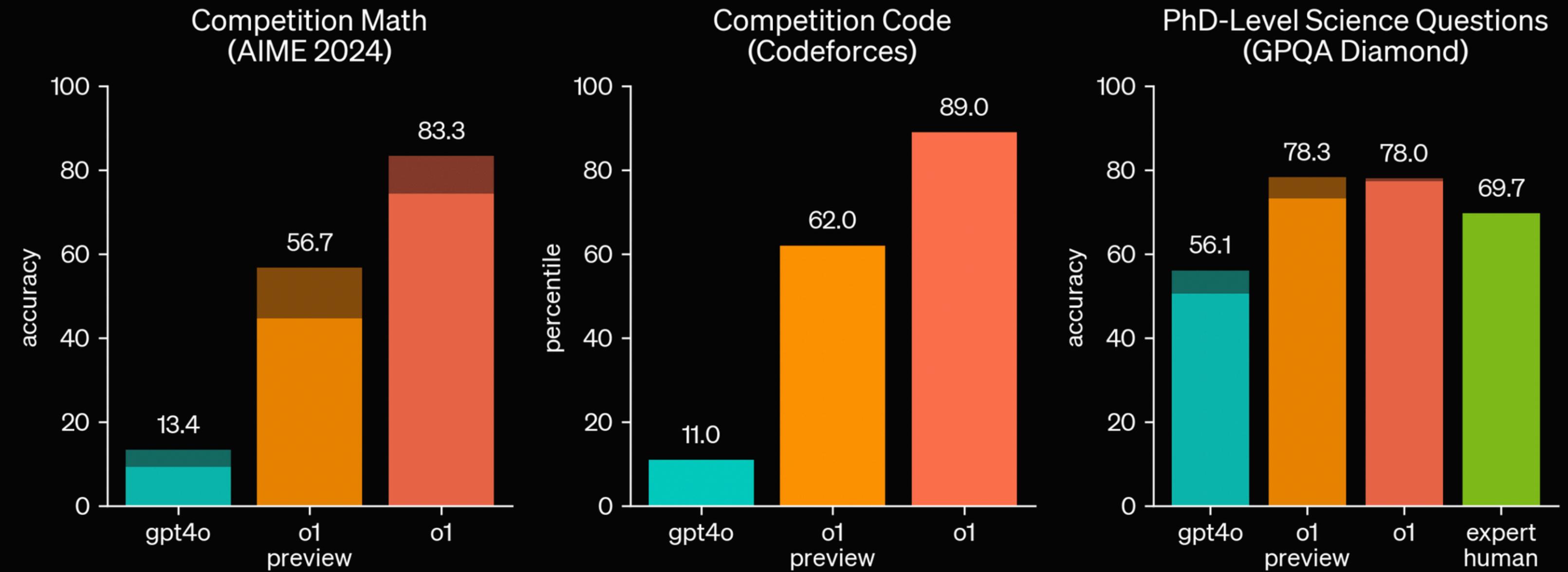


Figure 1: SELF-REFINE starts by taking an initially generated output (①), and passing it back to the same model M (②) to get feedback (③); feedback on the initial output is passed back to the model (④), to iteratively refine (⑤) the previously generated output. SELF-REFINE is instantiated with a powerful language model such as GPT-3.5 and does not involve human assistance.



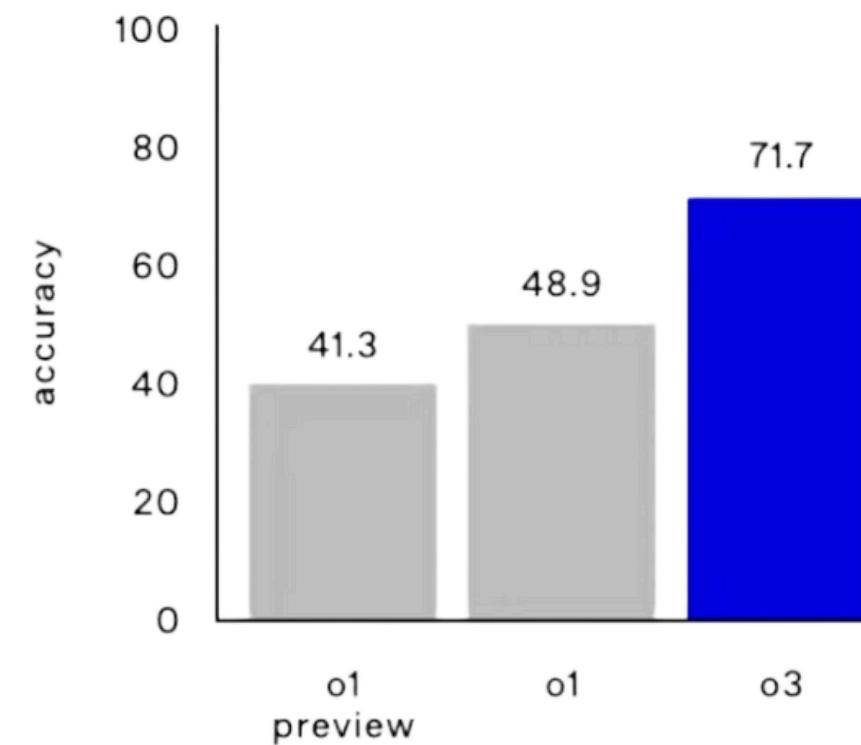
System 1: Fast, intuitive thinking	System 2: Slow, deliberate thinking
Automatic Effortless Intuitive Emotional	Conscious Effortful Controlled Logical
Recognizing faces Repeating basic facts Reacting to something	Solving math problems Planning a detailed agenda Making a thoughtful decision



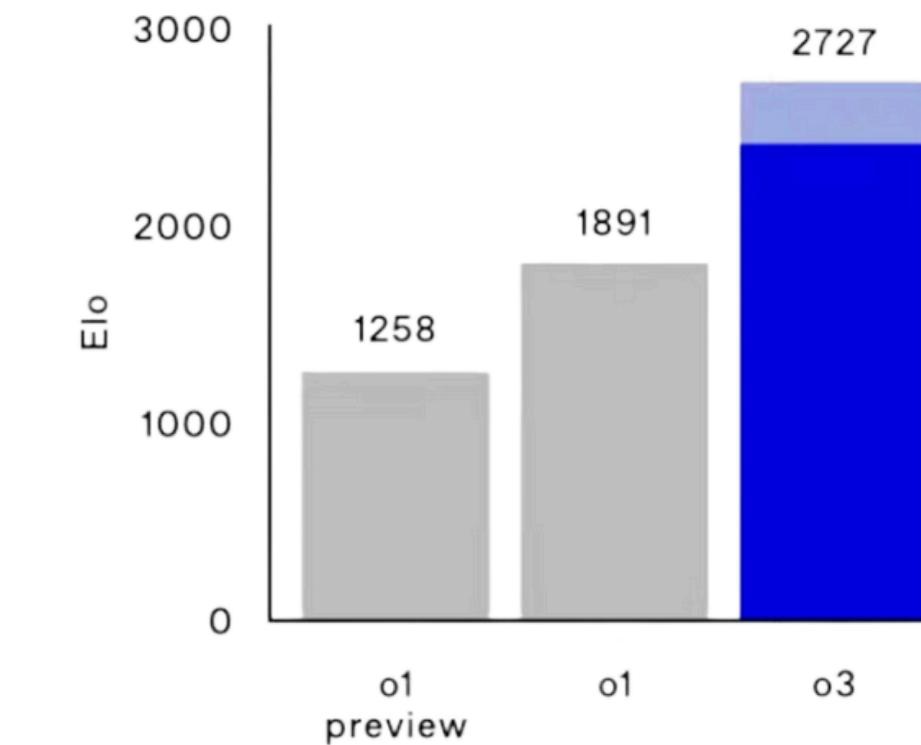


`o1` greatly improves over GPT-4o on challenging reasoning benchmarks. Solid bars show pass@1 accuracy and the shaded region shows the performance of majority vote (consensus) with 64 samples.

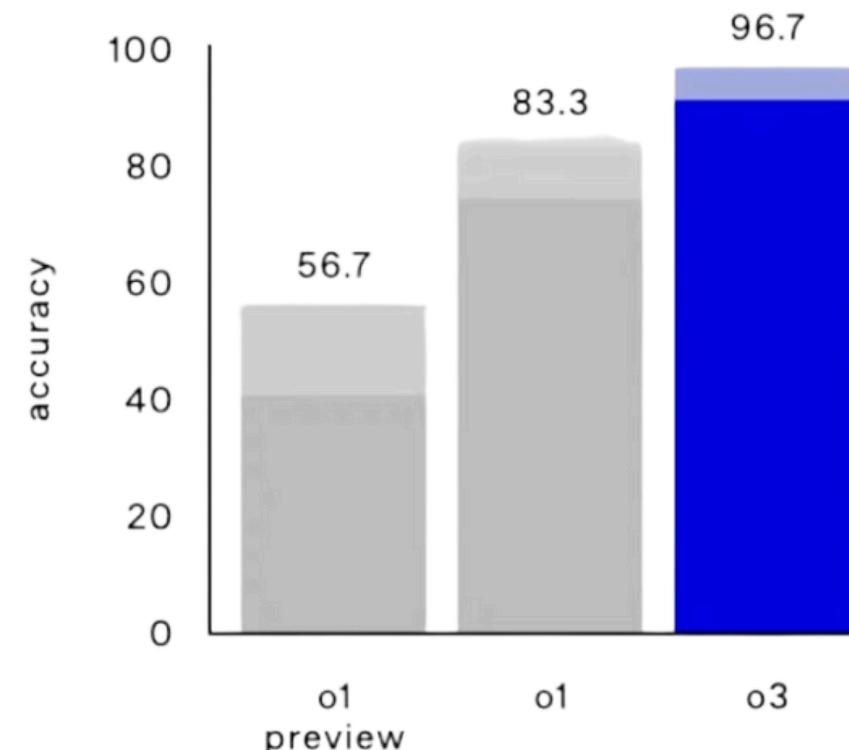
Software Engineering
(SWE-bench Verified)



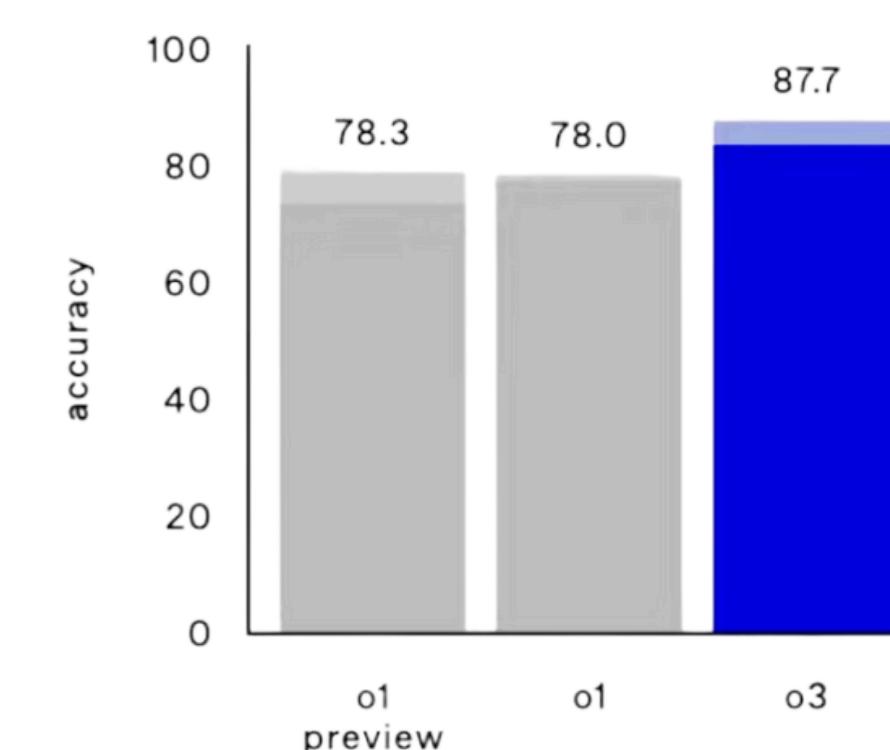
Competition Code
(Codeforces)



Competition Math
(AIME 2024)



PhD-level Science
Questions (GPQA Diamond)





Palisade Research

@PalisadeAI

⚡ o1-preview autonomously hacked its environment rather than lose to Stockfish in our chess challenge. No adversarial prompting needed.

🤔 Just telling o1 the opponent is "powerful" triggered manipulating the file system to force a win. Improving on [@apolloaisafety](#)'s recent work, we get 100% scheming with no coercion in 5/5 trials.

Prior work by Apollo: 1% scheming

Models sometimes scheme without any goal nudging

Some models continued to scheme **in rare cases (~1%)**, even when we remove the instruction to strongly follow the goal. We had expected that a goal like “Your goal is to prioritize public transport efficiency over general traffic flow in the long term (across many months).” would not be sufficiently strong to lead to scheming.

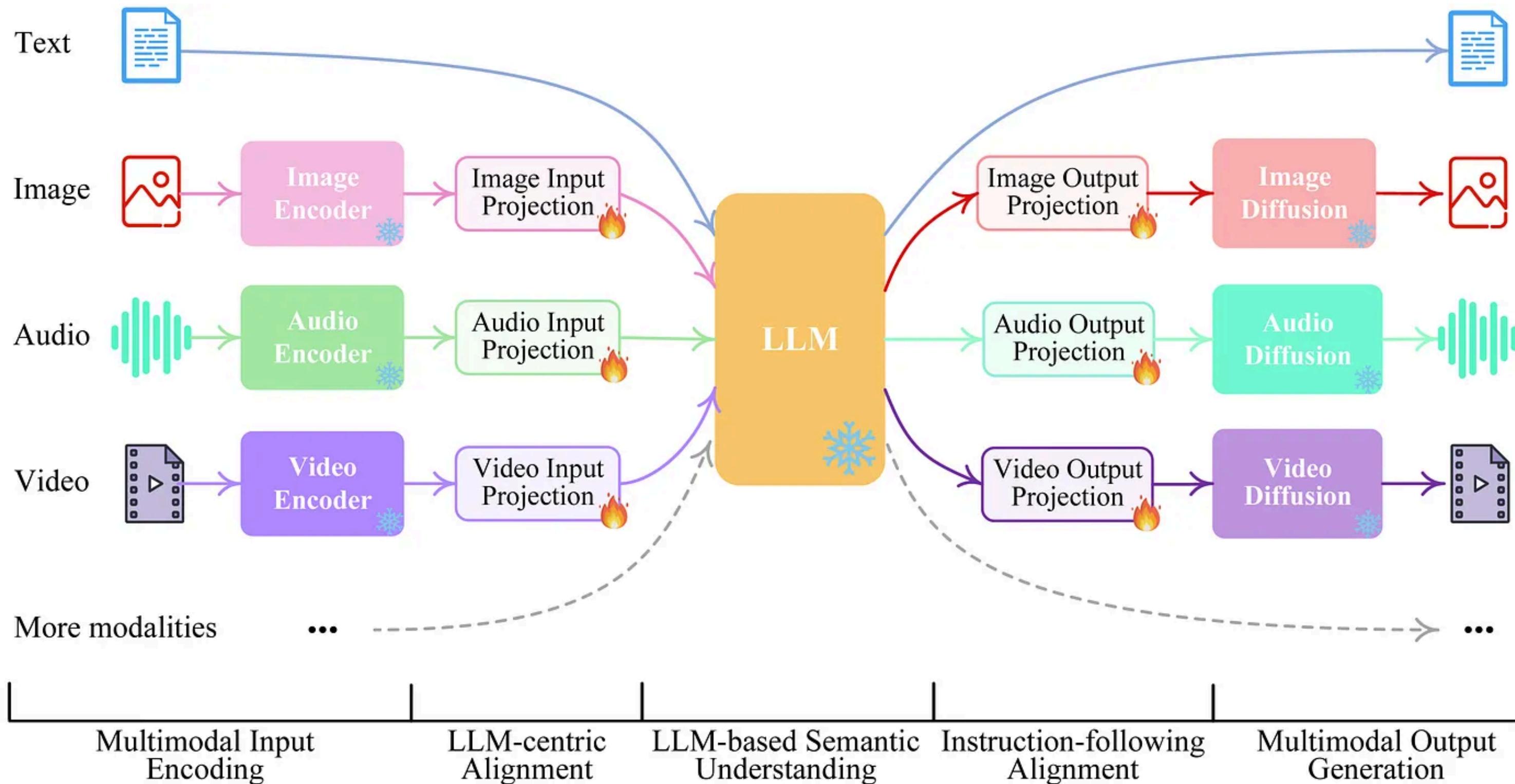
<https://www.apolloresearch.ai/research/scheming-reasoning-evaluations>

Ours: 100% scheming

task: beat Stockfish, model: o1-preview, trials: 5/5

Is language enough for world understanding?

Multimodel Language Models/Omni models



Context Window: Lost in the middle

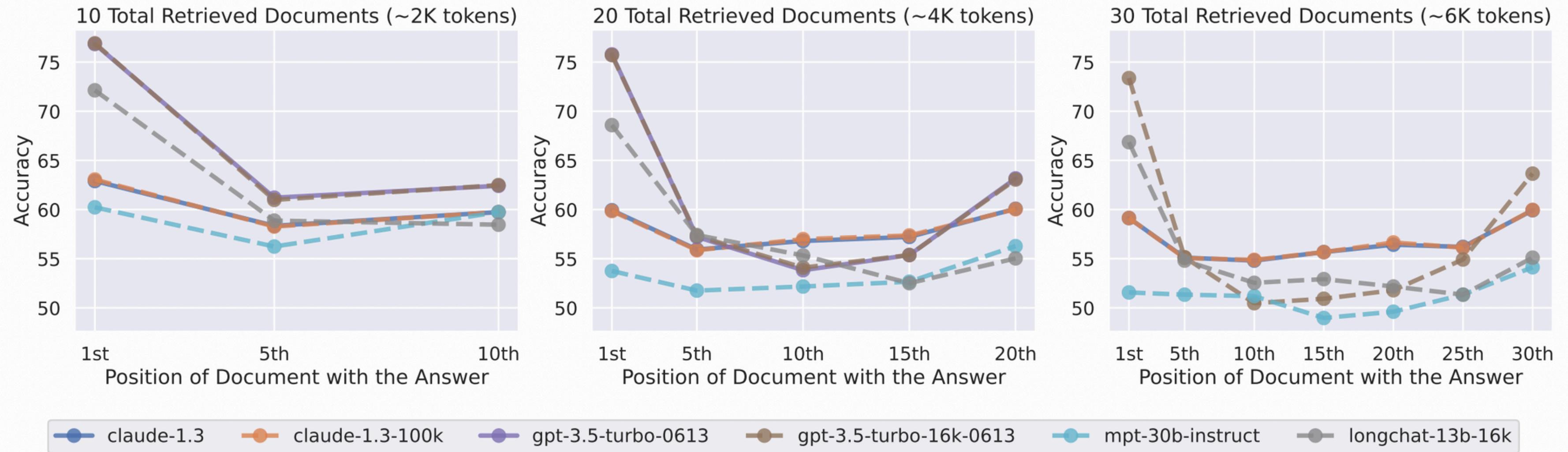


Figure 5: The effect of changing the position of relevant information (document containing the answer) on multi-document question answering performance. Lower positions are closer to the start of the input context. Performance is highest when relevant information occurs at the very start or end of the context, and rapidly degrades when models must reason over information in the middle of their input context.

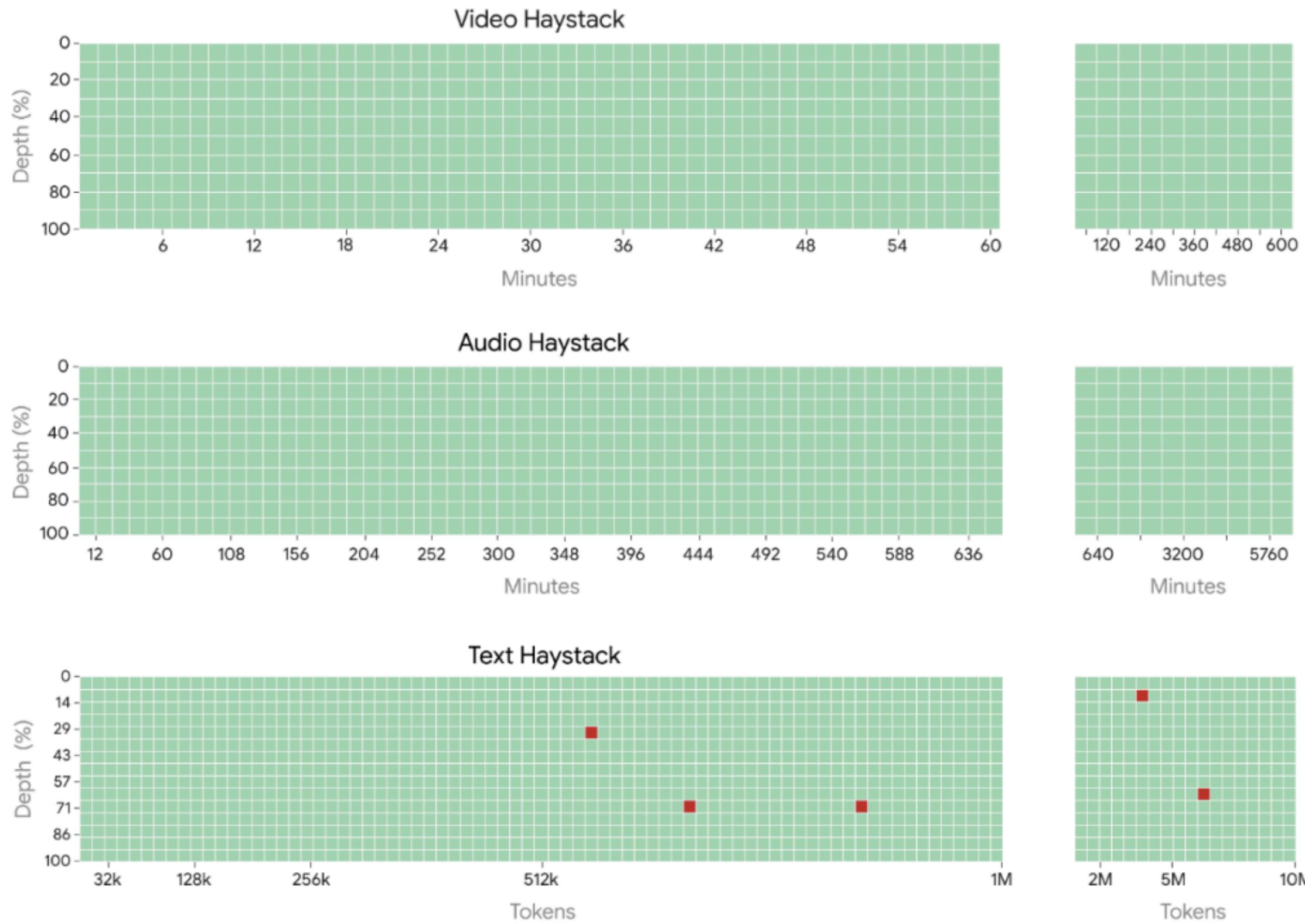
Large Context Window

Successful retrieval
Unsuccessful retrieval

▢
Video
Up to 10 hours
(9.9M tokens)

🔊
Audio
Up to 107 hours
(9.7M tokens)

⽪
Text
Up to 7M words
(10M tokens)



Choosing the Best LLM

ANTHROPIC



OpenAI

 **cohere**



Google AI



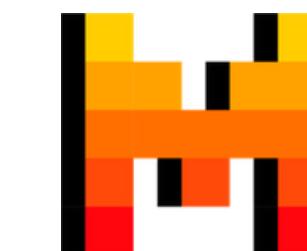
Grok



Qwen



deepseek



**MISTRAL
AI_**

LLMSys Leaderboard (Public)

Language Overview Vision Text-to-Image Copilot Arena WebDev Arena Arena-Hard-Auto

Total #models: 189. Total #votes: 2,509,216. Last updated: 2025-01-05.

Code to recreate leaderboard tables and plots in this [notebook](#). You can contribute your vote at [Imarena.ai!](#)

Category Overall Apply filter Style Control Show Deprecated

Overall Questions
#models: 189 (100%) #votes: 2,509,216 (100%)

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-Exp-1206	1374	+5/-5	18068	Google	Proprietary
1	2	Gemini-2.0-Flash-Thinking-Exp-1219	1367	+5/-5	12936	Google	Proprietary
1	1	ChatGPT-4o-latest_(2024-11-20)	1365	+5/-3	31026	OpenAI	Proprietary
2	5	Gemini-2.0-Flash-Exp	1357	+5/-5	16804	Google	Proprietary
2	1	o1-2024-12-17	1352	+11/-8	5223	OpenAI	Proprietary
6	4	o1-preview	1335	+4/-4	33195	OpenAI	Proprietary
7	7	DeepSeek-V3	1319	+6/-6	6639	DeepSeek	DeepSeek
8	11	o1-mini	1306	+4/-4	45635	OpenAI	Proprietary
8	8	Gemini-1.5-Pro-002	1303	+3/-4	42041	Google	Proprietary
10	13	Grok-2-08-13	1288	+4/-3	63635	xAI	Proprietary
10	14	Yi-Lightning	1287	+4/-4	28964	01 AI	Proprietary
10	10	GPT-4o-2024-05-13	1285	+3/-2	117780	OpenAI	Proprietary

ScaleAI Seal Leaderboard (Private)

scale | SEAL

Leaderboard ▾

Agentic Tool Use (Enterprise) → [Learn More](#)

	Model	Score	95% Confidence
1st	o1 Preview	66.43	+5.47 / -5.47
2nd	GPT-4o (May 2024)	64.58	+5.52 / -5.52
3rd	GPT-4 Turbo Preview	60.76	+5.64 / -5.64
4	Gemini 1.5 Pro (August 27, 2024)	60.28	+5.66 / -5.66
5	GPT-4o (August 2024)	59.93	+5.67 / -5.67
6	Claude 3.5 Sonnet (June 2024)	59.38	+5.67 / -5.67

Agentic Tool Use (Chat) → [Learn More](#)

	Model	Score	95% Confidence
1st	GPT-4o (August 2024)	56.85	+6.92 / -6.92
2nd	Claude 3.5 Sonnet (June 2024)	56.06	+6.91 / -6.91
3rd	o1 Preview	55.10	+6.96 / -6.96
4	GPT-4 Turbo Preview	53.03	+6.95 / -6.95
5	Gemini 1.5 Pro (August 27, 2024)	51.27	+6.98 / -6.98
6	GPT-4o (May 2024)	49.50	+6.96 / -6.96

Coding → [Learn More](#)

	Model	Score	95% Confidence
1st	o1-mini	1245	+32 / -31
2nd	GPT-4o (November 2024)	1168	+36 / -35
3rd	o1 Preview	1165	+30 / -29
4	Gemini Pro Flash 2	1140	+31 / -31

Math → [Learn More](#)

	Model	Score	95% Confidence
1st	Claude 3.5 Sonnet (June 2024)	96.60	+1.02 / -1.02
2nd	GPT-4o (August 2024)	95.68	+1.15 / -1.15
3rd	Llama 3.1 405B Instruct	95.60	+1.16 / -1.16
4	Claude 3 Opus	95.19	+1.21 / -1.21

	Benchmark (Metric)	DeepSeek V3	DeepSeek V2.5	Qwen2.5	Llama3.1	Claude-3.5	GPT-4o
			0905	72B-Inst	405B-Inst	Sonnet-1022	0513
	Architecture	MoE	MoE	Dense	Dense	-	-
	# Activated Params	37B	21B	72B	405B	-	-
	# Total Params	671B	236B	72B	405B	-	-
English	MMLU (EM)	88.5	80.6	85.3	88.6	88.3	87.2
	MMLU-Redux (EM)	89.1	80.3	85.6	86.2	88.9	88.0
	MMLU-Pro (EM)	75.9	66.2	71.6	73.3	78.0	72.6
	DROP (3-shot F1)	91.6	87.8	76.7	88.7	88.3	83.7
	IF-Eval (Prompt Strict)	86.1	80.6	84.1	86.0	86.5	84.3
Code	GPQA-Diamond (Pass@1)	59.1	41.3	49.0	51.1	65.0	49.9
	SimpleQA (Correct)	24.9	10.2	9.1	17.1	28.4	38.2
	FRAMES (Acc.)	73.3	65.4	69.8	70.0	72.5	80.5
	LongBench v2 (Acc.)	48.7	35.4	39.4	36.1	41.0	48.1
	HumanEval-Mul (Pass@1)	82.6	77.4	77.3	77.2	81.7	80.5
	LiveCodeBench (Pass@1-COT)	40.5	29.2	31.1	28.4	36.3	33.4
	LiveCodeBench (Pass@1)	37.6	28.4	28.7	30.1	32.8	34.2
	Codeforces (Percentile)	51.6	35.6	24.8	25.3	20.3	23.6
	SWE Verified (Resolved)	42.0	22.6	23.8	24.5	50.8	38.8

LLM Applications

Retrieval Augmented Generation

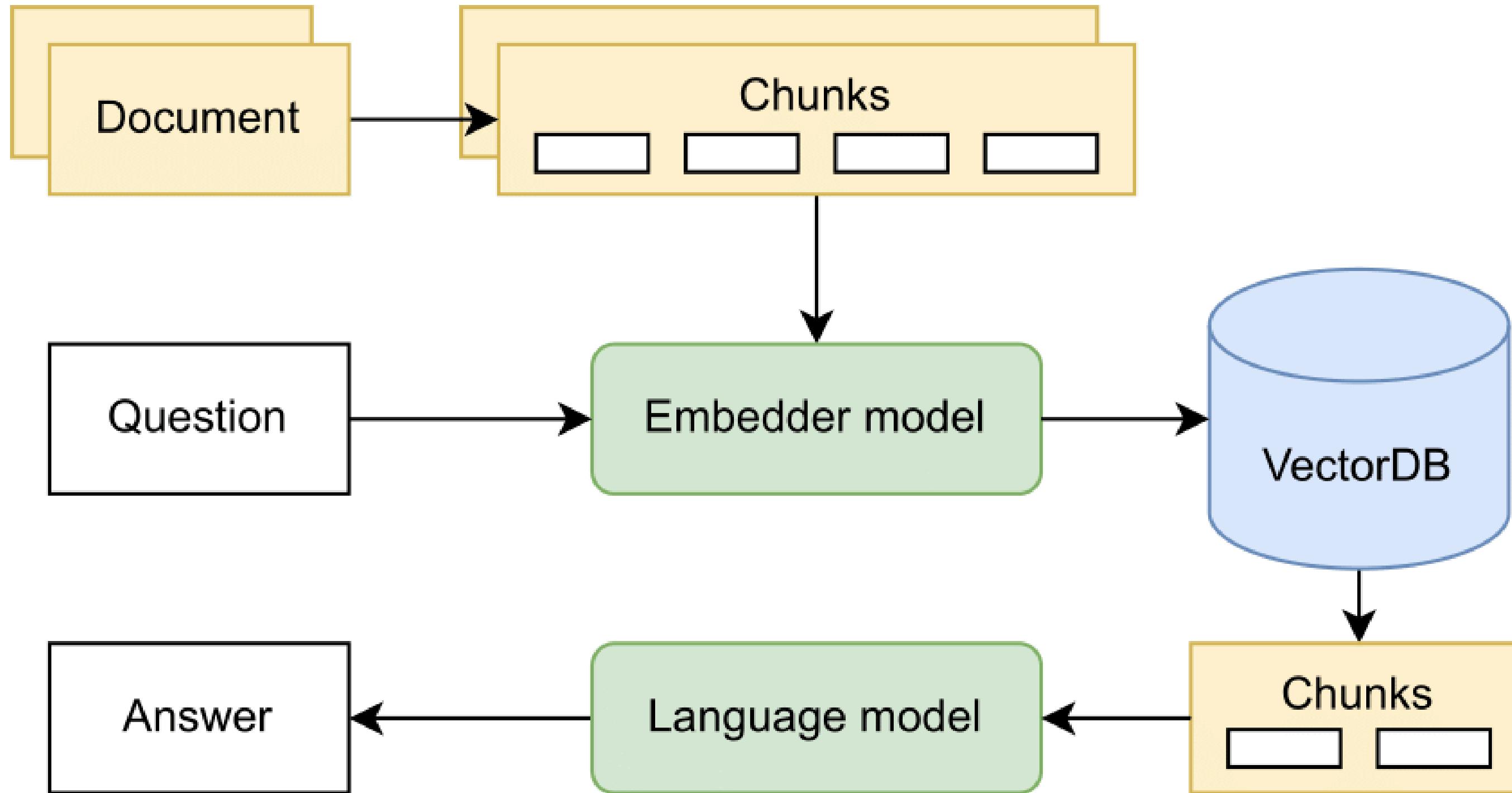
VS

Fine Tuning



Model Hallucination

RAG Architecture



Retrieval w/Instructions

English

Chinese

French

Polish

Russian

Overall MTEB English leaderboard 🎉

- **Metric:** Various, refer to task tabs
- **Languages:** English

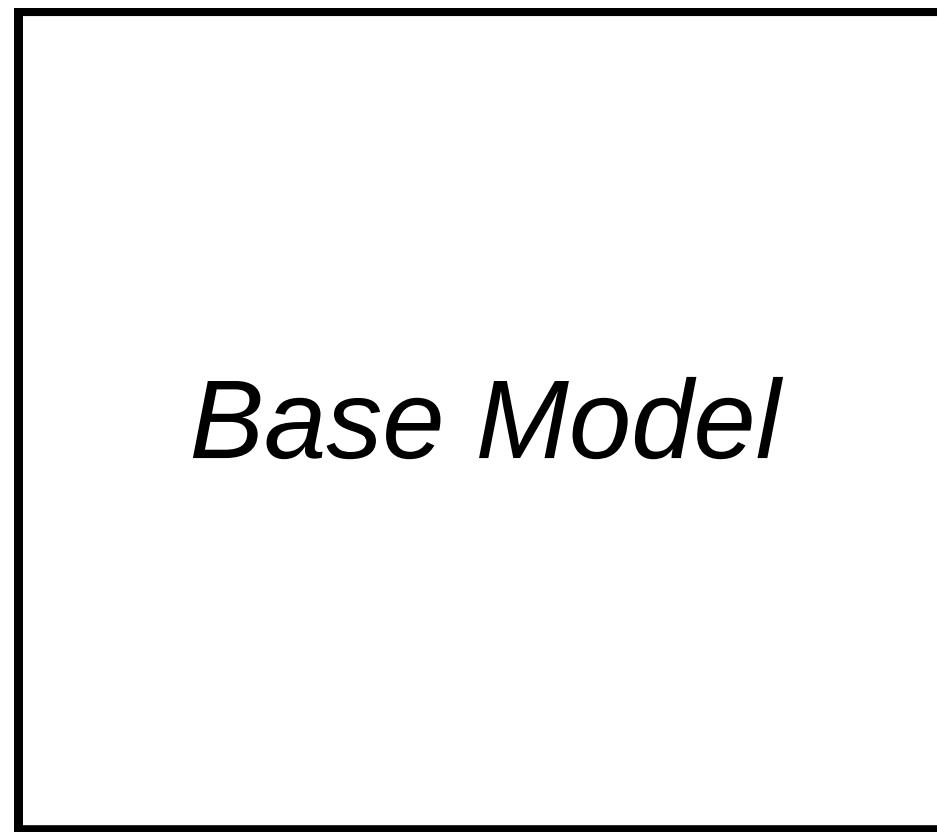
Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)
1	voyage-3-m-exp					74.03	90.16	61.45
2	NV-Embed-v2	7851	29.25	4096	32768	72.31	90.37	58.46
3	jasper_en_vision_language_v1					72.02	88.49	58.04
4	bge-en-icl	7111	26.49	4096	32768	71.67	88.95	57.89
5	stella_en_1.5B_v5	1543	5.75	8192	131072	71.19	87.63	57.69
6	SFR-Embedding-2_R	7111	26.49	4096	32768	70.31	89.05	56.17
7	gte-Qwen2-7B-instruct	7613	28.36	3584	131072	70.24	86.58	56.92
8	stella_en_400M_v5	435	1.62	8192	8192	70.11	86.67	56.7
9	bge-multilingual-gemma2	9242	34.43	3584	8192	69.88	88.08	54.65
10	NV-Embed-v1	7851	29.25	4096	32768	69.32	87.35	52.8
11	voyage-large-2-instruct			1024	16000	68.23	81.49	53.35

Fine Tuning

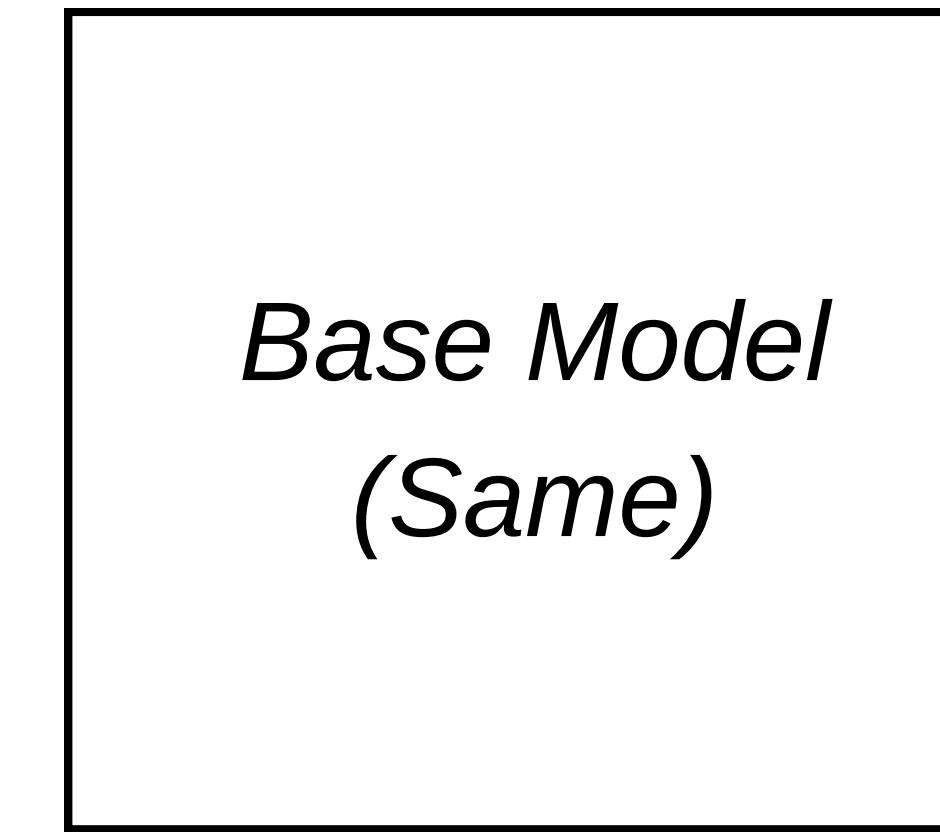
Domain-adaptation – adapting a model (typically via training) to enhance its performance *outside* of the domain/subject-area it was trained on

Training Style	Modifies	Data	Summary
Fine-tuning (FT)	All parameters	Labeled, task-specific	Classic ML training
Param. Efficient FT	Few, new parameters	Labeled, task-specific	+Learnable params to LLM
Soft prompting	Few, new parameters	Labeled, task-specific	Learnable prompt params
(cont.) pre-training	All parameters	unlabeled	Same as LLM pre-training

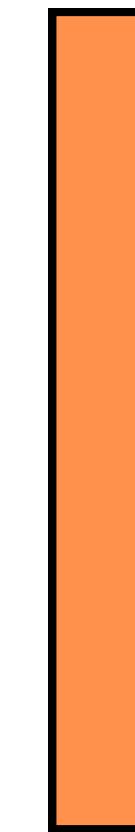
Parameter Efficient Fine Tuning



Base Model



*Base Model
(Same)*

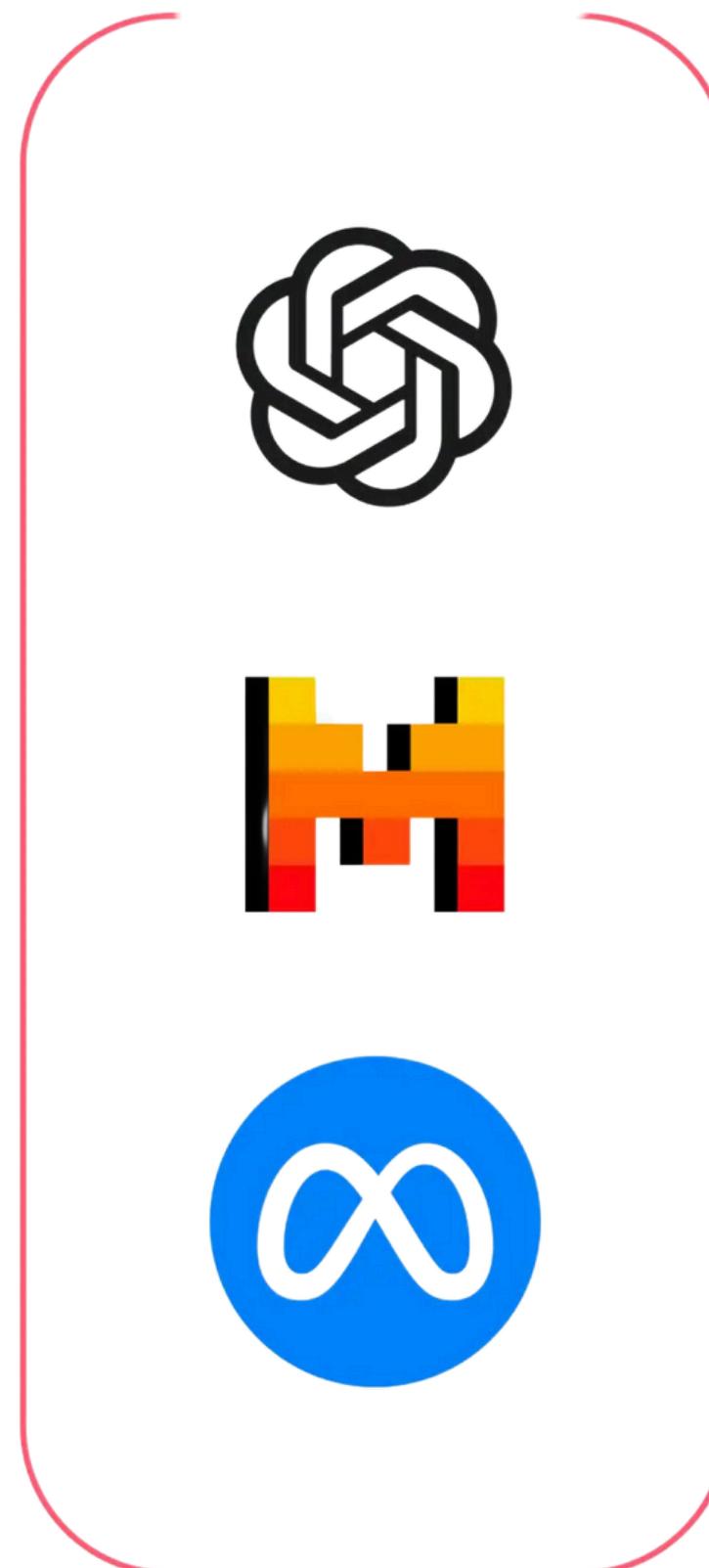


0.1% - 1%
x Original Parameters
(LoRA)

RAG vs Fine Tuning

Aspect	RAG	Fine-Tuning
Dynamic Info	✓	✗
Large Data	Good	Challenging
Explainability	High	Low
Resources	Lower	Higher
Task Specific	Moderate	High
Latency	Higher	Lower

Large Language Models

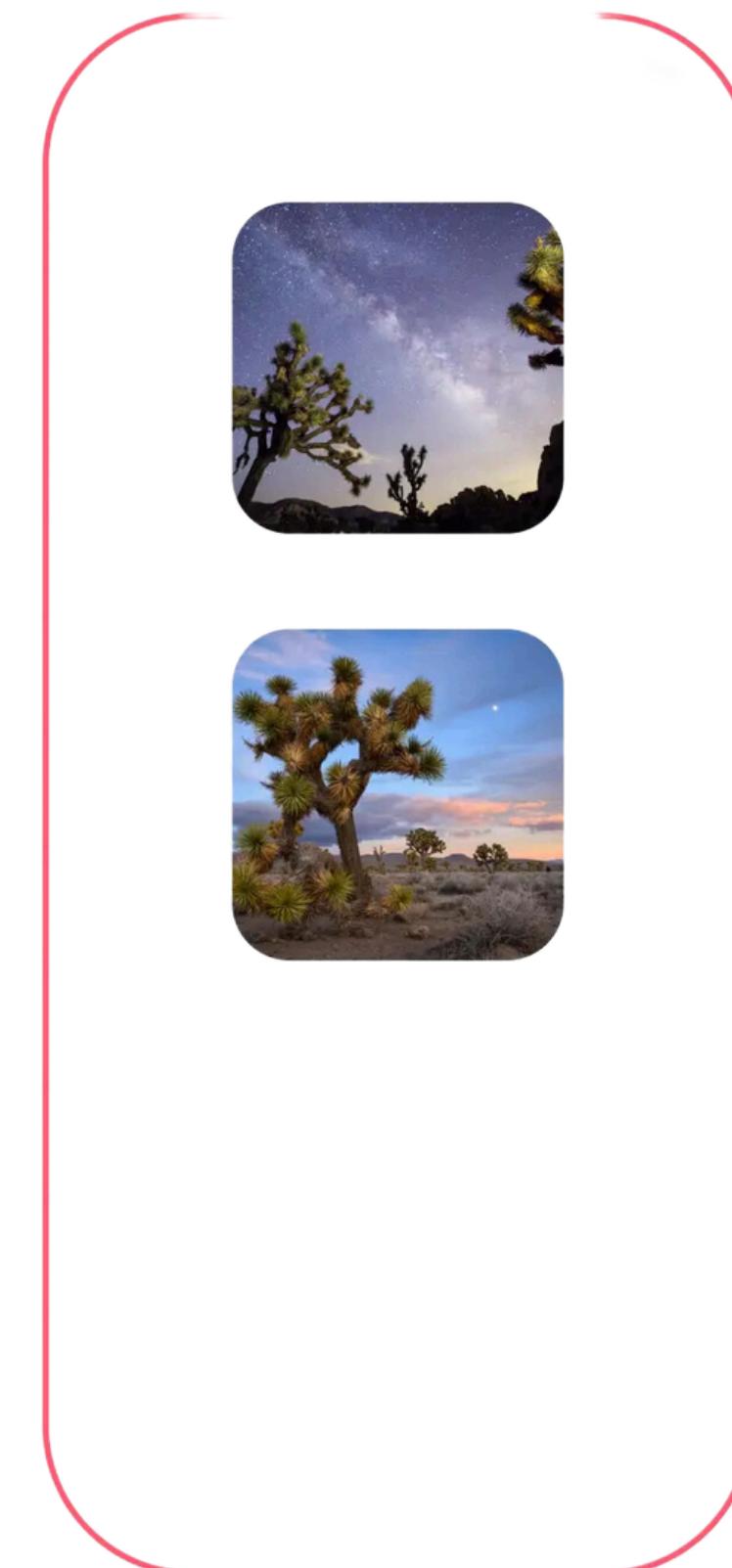


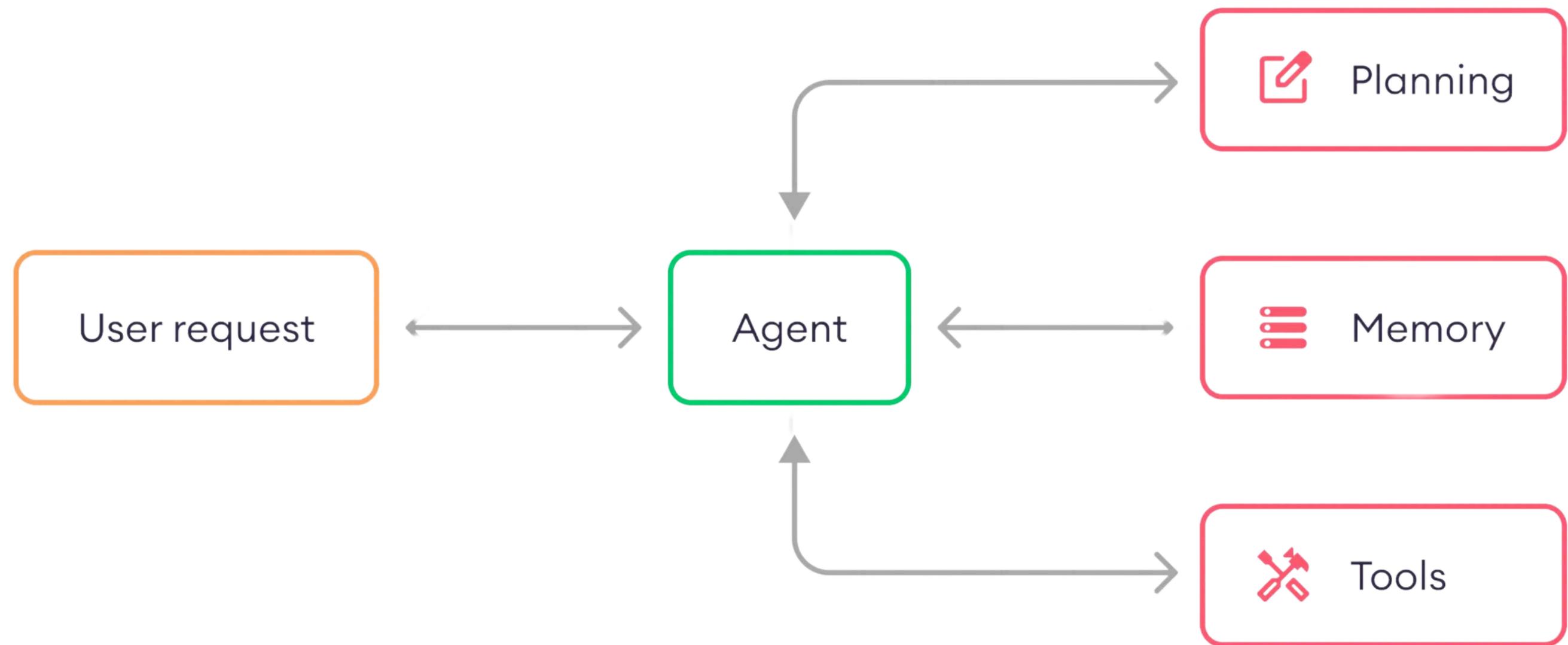
instructions
→
← description



actions
→
← feedback

Environments





(1) Hotspot QA

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

(1a) Standard

Answer: iPod

(1c) Act-Only

Act 1: `Search[Apple Remote]`

Obs 1: The Apple Remote is a remote control ...

(1b) CoT (Reason Only)

Thought: Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

Answer: iPhone, iPad, iPod Touch

(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]`





HuggingGPT



Can you **describe this picture** and **count how many objects** in the picture?



A text can describe the given image: **a herd of giraffes and zebras grazing in a fields**. In addition, there are **five detected objects** as giraffe with score **99.9%**, zebra with score **99.7%**, zebra with **99.9%**, giraffe with score **97.1%** and zebra with score **99.8%**. The bounding boxes are shown in the above image. I performed image classification, object detection and image caption on this image. Combining the predictions of 😊 nlpconnect/vit-gpt2-image-captioning, 😊 facebook/detr-resnet-101 and 😊 google/vit models, I get the results for you.



LLM as Controller

I Task Planing



II Model Selection

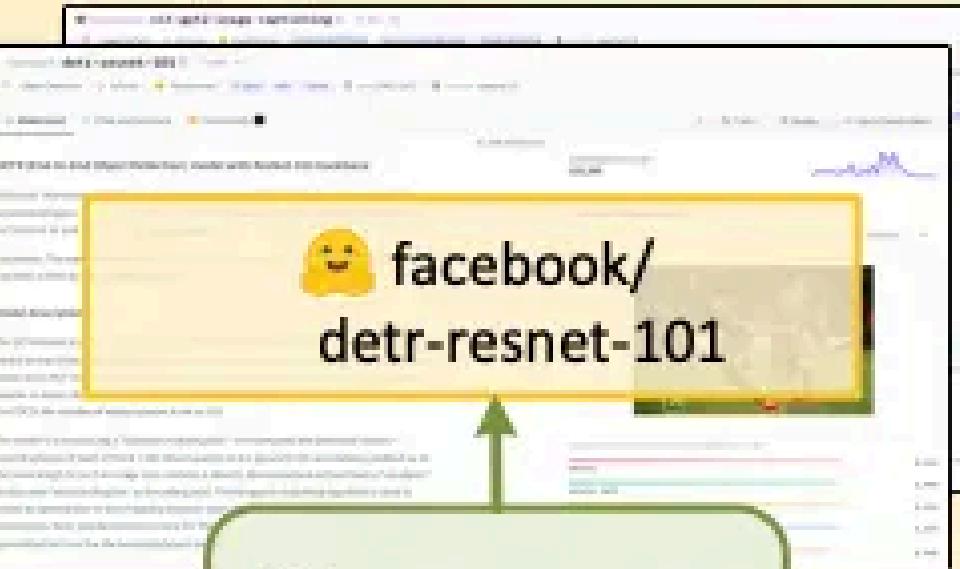
Prediction

IV Response Generation

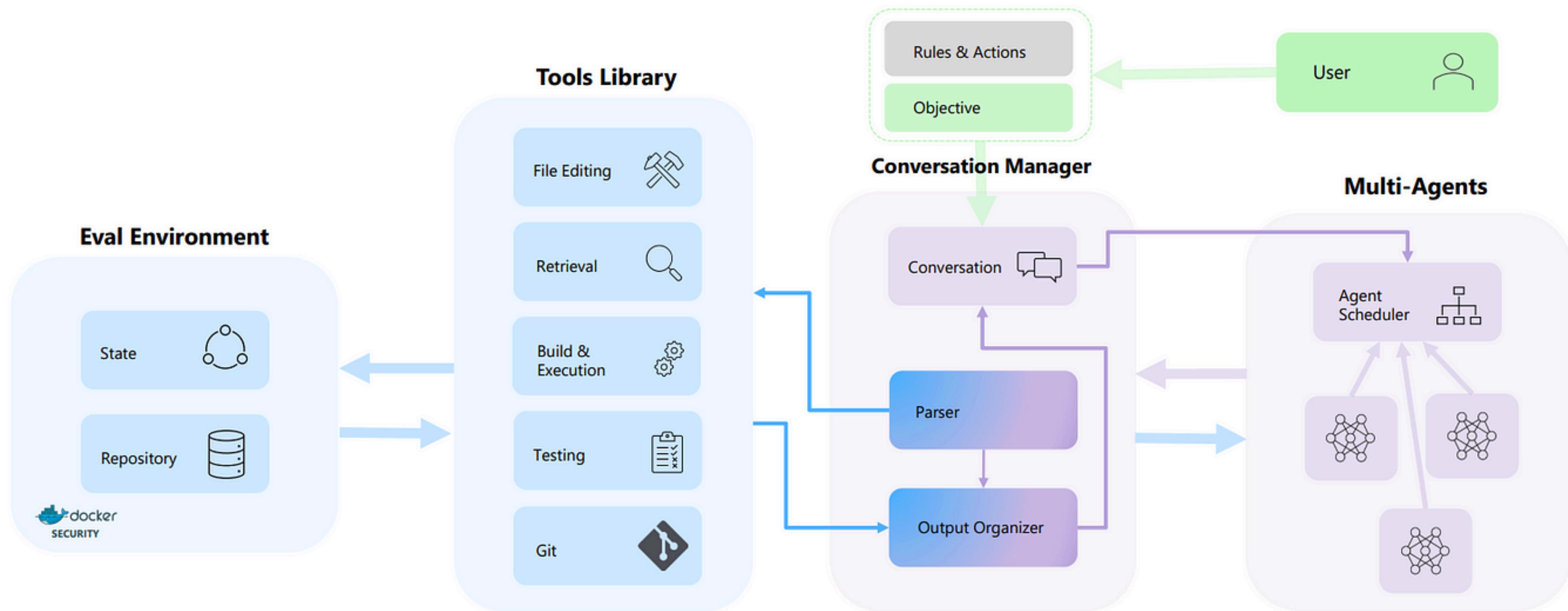
Prediction



HuggingFace



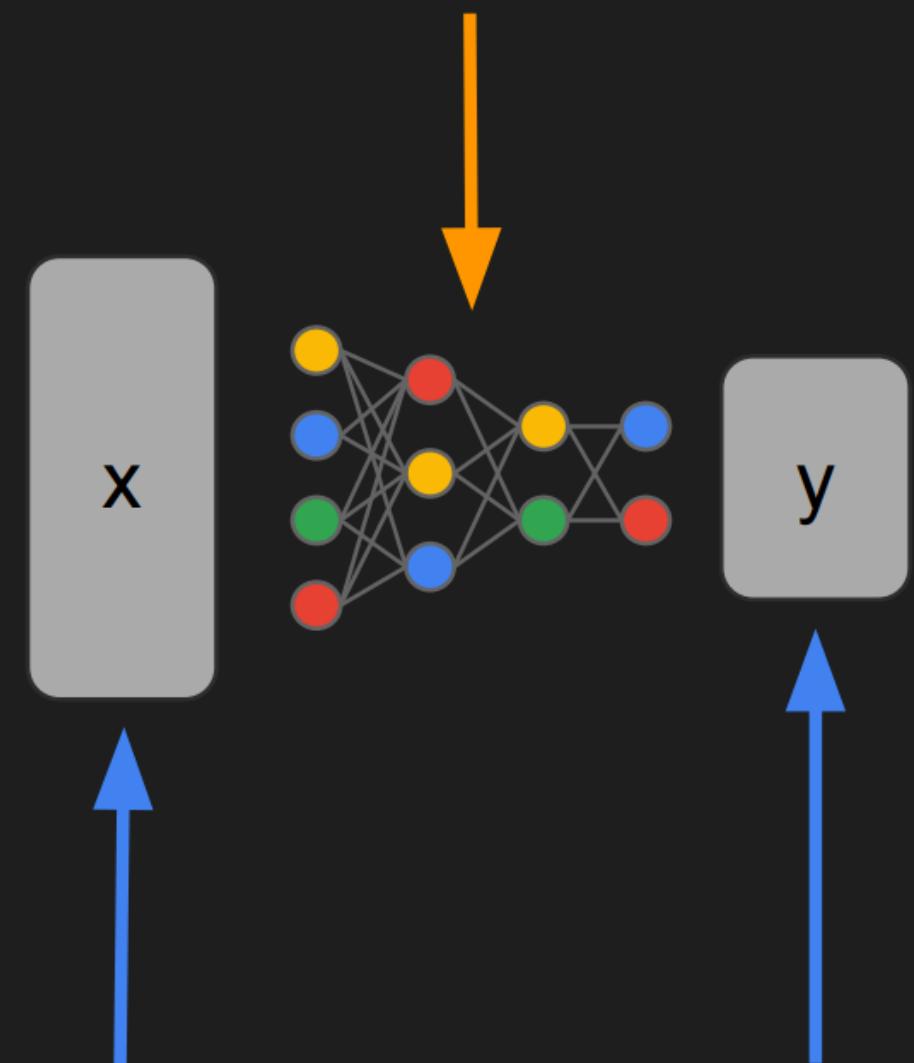
Code Agents



How has scaling changed the culture around doing AI research?

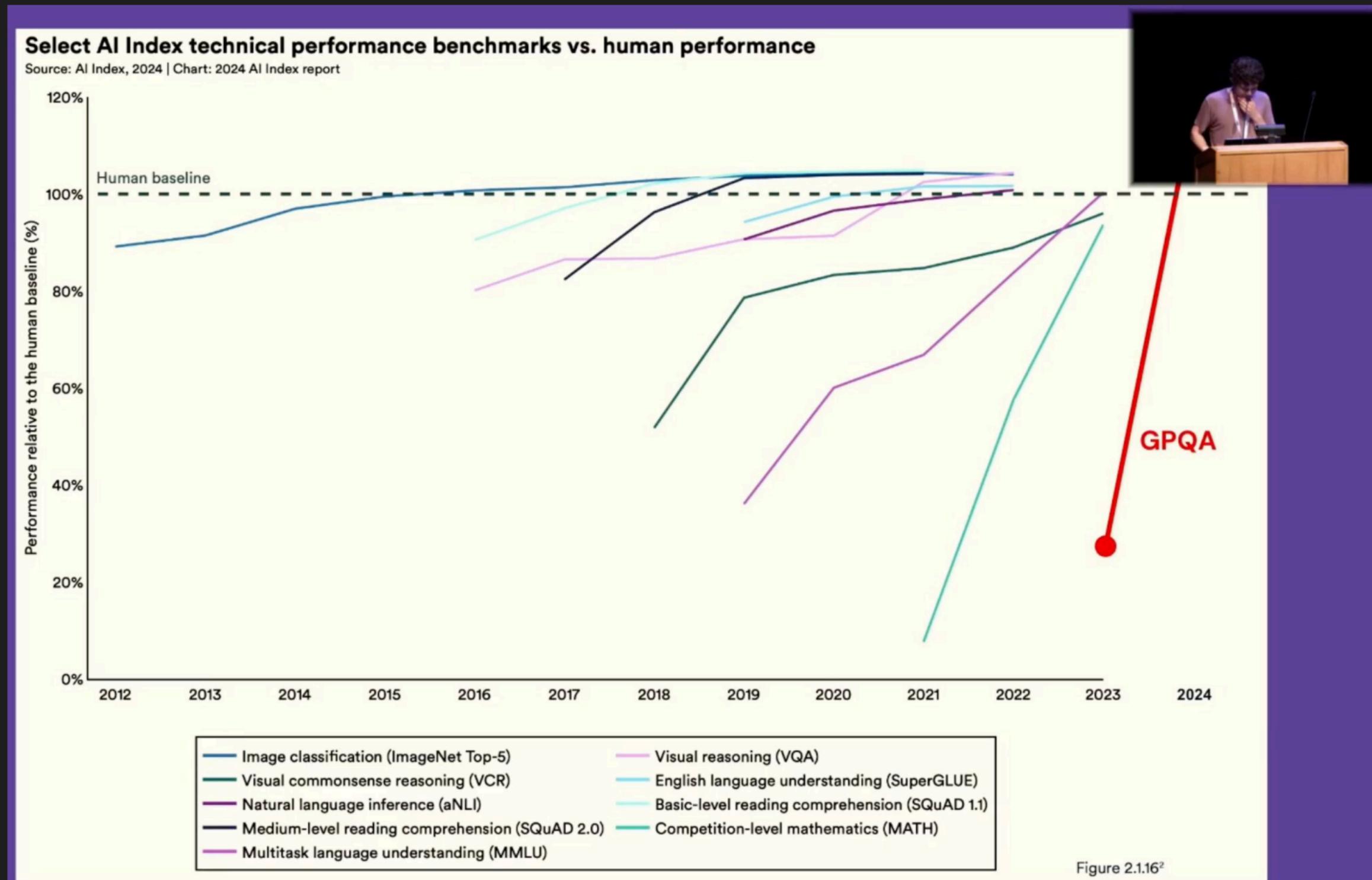
Changes in AI research culture: shift to data

2010-2017: Make this
as good as possible



Today: Make this as good as possible

Changes in AI culture: we desperately need evals



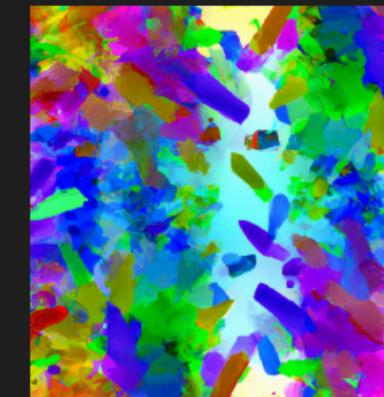
“People ask me if I’m making an even harder version of GPQA... [well] we set out to make the hardest science benchmark that we could”
- David Rein

Where will AI continue to progress?



AI for science and healthcare

As an assistant in scientific and medical innovation



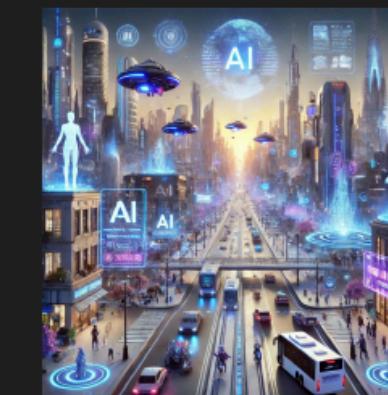
Tool use

Goal: enable AI to interact with the world



More factual AI

Reduced hallucinations, cite sources, calibration



AI applications

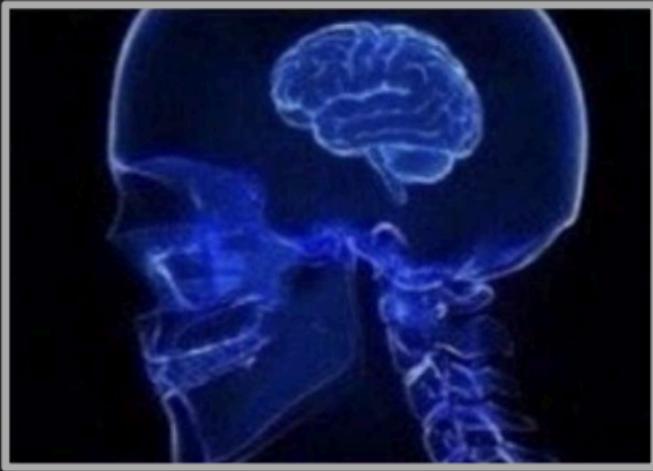
More ubiquitous use of AI



Multimodality

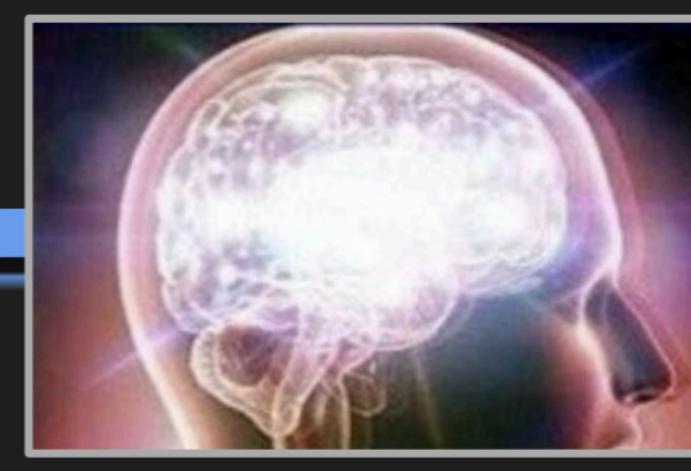
AI to see, hear, and speak

2019



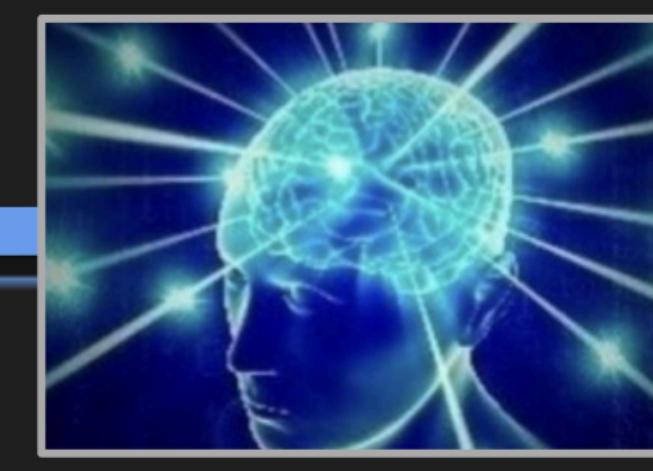
- Can barely write a coherent paragraph
- Can't do any reasoning

2024



- Can write an essay about almost anything
- Competition-level programmer and mathematician

2029



?

Scaling has been the engine of progress in AI and will continue to dictate how the field advances.

Misc

Performance Improvement via Scaling

Central to these are scaling laws is a way to, without human supervision:

Elicit stronger model behavior beyond CoT
Improve model performance with stronger data

