

How to use AI in Mathematical Research Effectively

Taylor Dupuy

Department of Mathematics and Statistics, University of Vermont

GAATI Laboratory, Université de la Polynésie française

University of Rochester · 23 October 2026



My Experience

tdupu / magma-diff-alg

<> Code

Issues 20

Pull requests 1

Actions

Projects

Security and quality

Insights

 magma-diff-alg

Public

Notifications

Fork 0

Star 0

master

2 Branches

0 Tags

Go to file

Code

 tdupu

and

 claude

Retire three stale Brownfield rows

713ac9c · last month

216 Commits

 ai/2026-09-19-krichever-slice1	Track the Krichever slice-1 research inputs under ai/	last month
 test-homog	mid edits of the weird-appendix	8 months ago
 test-valuations	Put the testing infrastructure under contract (LAYOUT.md ...	last month
 .gitignore	Put the testing infrastructure under contract (LAYOUT.md ...	last month
 AGENTS.md	Put the testing infrastructure under contract (LAYOUT.md ...	last month
 CLAUDE.md	Put the testing infrastructure under contract (LAYOUT.md ...	last month
 CONTEXT.md	Put the testing infrastructure under contract (LAYOUT.md ...	last month
 LAYOUT.md	Retire three stale Brownfield rows	last month
 README.md	README: document InitialIdeal, CanonicalizeWeightModul...	4 months ago
 diffalg.spec	Add InitialIdeal(l, w) intrinsic + lineality canonicalization he...	4 months ago
 package-dimension.mag	fixed a small bug	last year
 package-groebner-fan.mag	groebner-fan: C3 — re-verify sticky-cone results via Stabl...	3 months ago

About

No description, website, or topics provided.

Readme

Activity

0 stars

2 watching

0 forks

Report repository

Releases

No releases published

Contributors 3

 tdupu Taylor Dupuy

 claude Claude

 smand Shilpi Mandal

Languages

My Experience

tdupu / magma-clifford-algebras

<> Code

Issues

Pull requests

Actions

Projects

Security and quality

Insights

magma-clifford-algebras

Public

Notifications

Fork 0

Star 2

main

1 Branch

0 Tags

Go to file

<> Code

tdupu

auto-commit [repos-commit] 2026-07-14T09:58

51cfdeb · 3 months ago

13 Commits

.agents/skills/beads	bd init: initialize beads issue tracking	3 months ago
.beads	auto-commit [repos-commit] 2026-07-14T09:58	3 months ago
.claude	bd init: initialize beads issue tracking	3 months ago
.codex	bd init: initialize beads issue tracking	3 months ago
.gitignore	bd init: initialize beads issue tracking	3 months ago
AGENTS.md	bd init: initialize beads issue tracking	3 months ago
CLAUDE.md	bd init: initialize beads issue tracking	3 months ago
LICENSE	Initial commit	2 years ago
README.md	resolved conflicts	last year
cliff.spec	removed package-coset-table.mag from the spec	last year
fundamental-domain-functions.mag	updated from the private repo	last year
make_clifford_euclidean-1.mag	updated the name of the spec file in the scripts for genera...	last year

About

A Magma package for Clifford Algebras extending magmas AlgCliff and AlgCliffElt types.

Readme

LGPL-2.1 license

Activity

2 stars

1 watching

0 forks

Report repository

Releases

No releases published

Contributors 2

tdupu

Taylor Dupuy

adammlogan

Adam Logan

Languages

My Experience



→ Abelian varieties → Fq

Citation · Feedback · Hide Menu

Isogeny classes of abelian varieties over finite fields

Introduction

OverviewRandom

UniverseKnowledge

L-functions

RationalAll

Modular forms

ClassicalMaass

HilbertBianchi

Varieties

Elliptic curves

over $\mathbb{Q}$

over number fields

The database currently contains 2,945,722 [isogeny classes](#) of [abelian varieties](#) of dimension up to 6 over finite fields. You can [browse further statistics](#) or [create your own](#).

Browse

By [dimension](#):1 2 3 4 5 6

By [base field cardinality](#):2 3 4 5 7 8 9 16 17 19 23 25 27-211 223-1024

By [p-rank](#):0 1 2 3 4 5 6

By [angle rank](#):0 1 2 3 4 5 6

By [p-corank](#):0 1 2 3 4 5 6

By [angle corank](#):0 1 2 3 4 5 6

By primitive simple **geom. simple** ordinary almost ordinary supersingular princ. polarizable Jacobian

Some interesting isogeny classes or a random isogeny class

A [table](#) by dimension and base field.

Learn more

Source and acknowledgments

Completeness of the data

Reliability of the data

Labeling convention

Where are you?

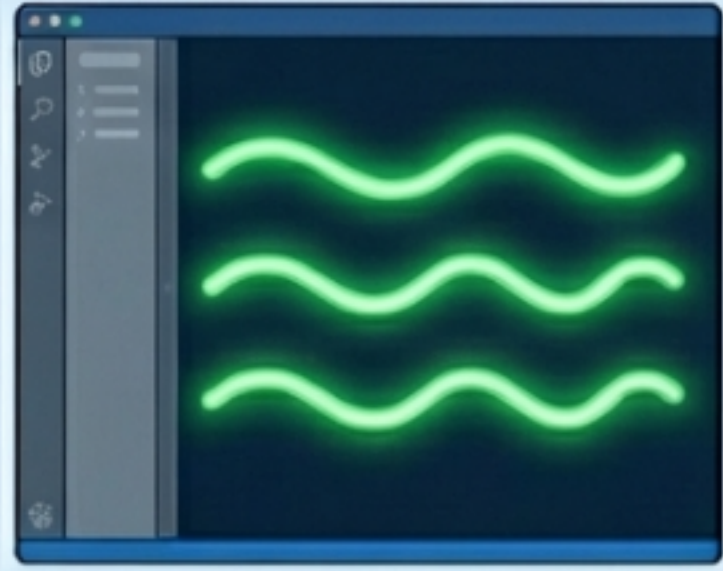


Figure 1



Figure 2



Figure 3



Figure 4

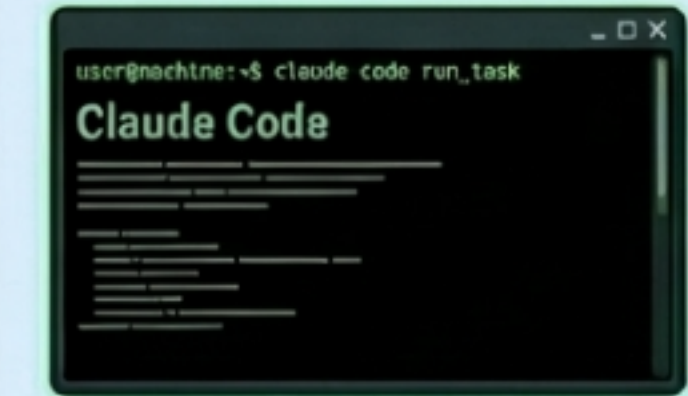


Figure 5



Figure 6



Figure 7



Figure 8

- 1 Near-zero AI
- 2 Agent in IDE, asks permission
- 3 Agent in IDE, YOLO
- 4 Agent fills the IDE
- 5 CLI, one agent
- 6 CLI, 3–5 agents
- 7 10+ agents, by hand
- 8 Your own orchestrator

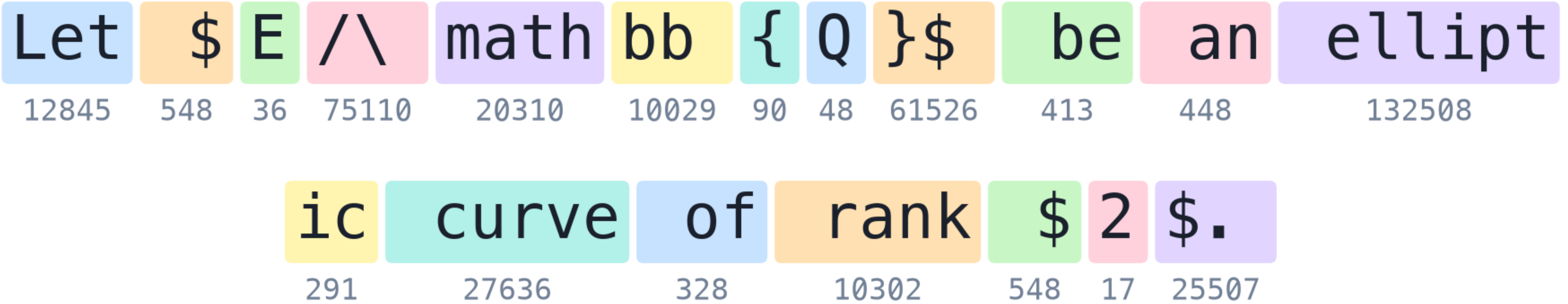
TOKENIZATION

36 characters, 11 tokens

Let	E	be	an	ellipt	ic	curve	of	rank		2
12845	457	413	448	132508	291	27636	328	10302	220	17

Tokenizer: OpenAI tiktoken, encoding o200k_base (GPT-4o / GPT-5 family; IVI $\approx$ 200,000). Grey numbers = token ids.

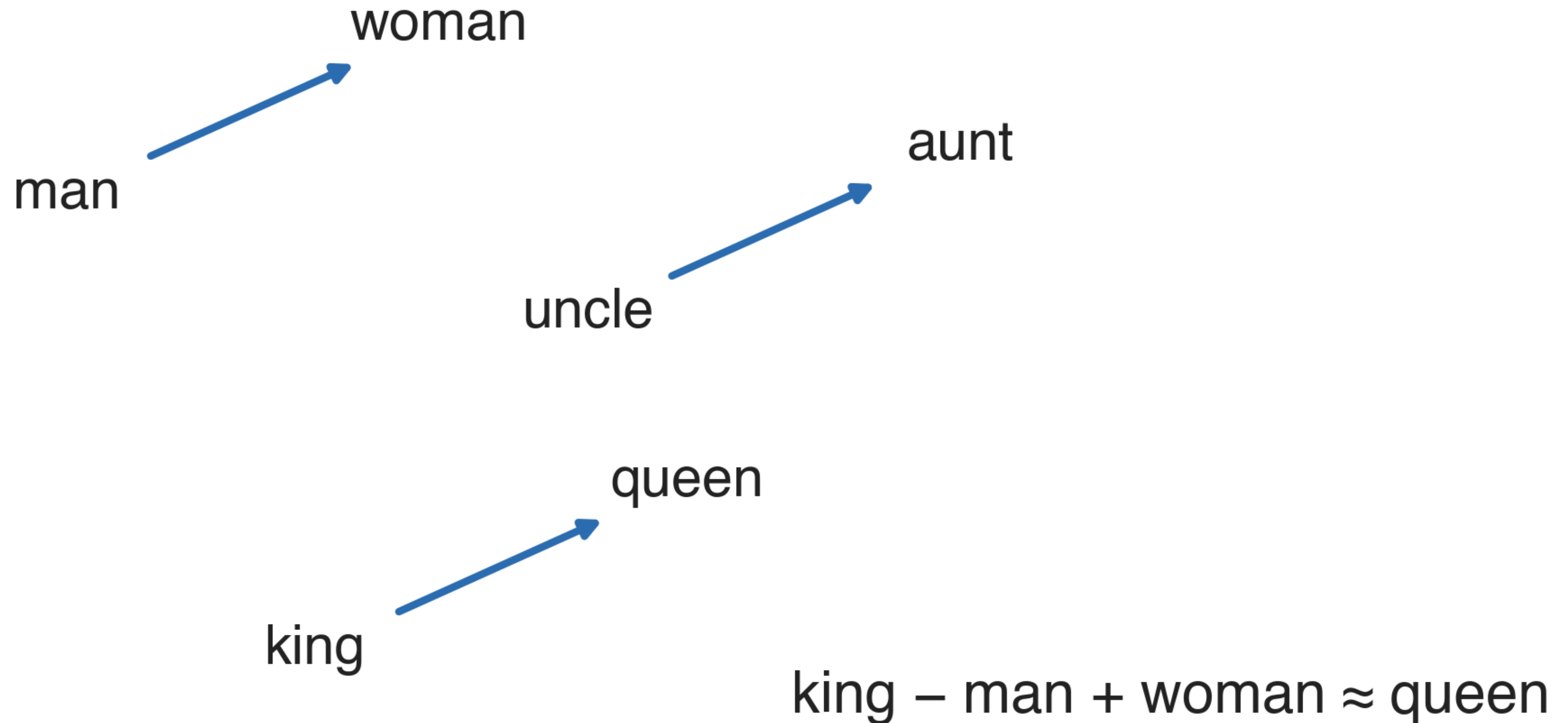
52 characters, 19 tokens



Tokenizer: OpenAI tiktoken, encoding o200k_base (GPT-4o / GPT-5 family; IVI $\approx$ 200,000). Grey numbers = token ids.

V = set of tokens; $\mathbb{R}^V \rightarrow \mathbb{R}^d$ ("embedding space", actually a projection)

V = set of tokens; $\mathbb{R}^V \rightarrow \mathbb{R}^d$ ("embedding space", actually a projection)



Vocabulary size $|V|$ and embedding dimension d

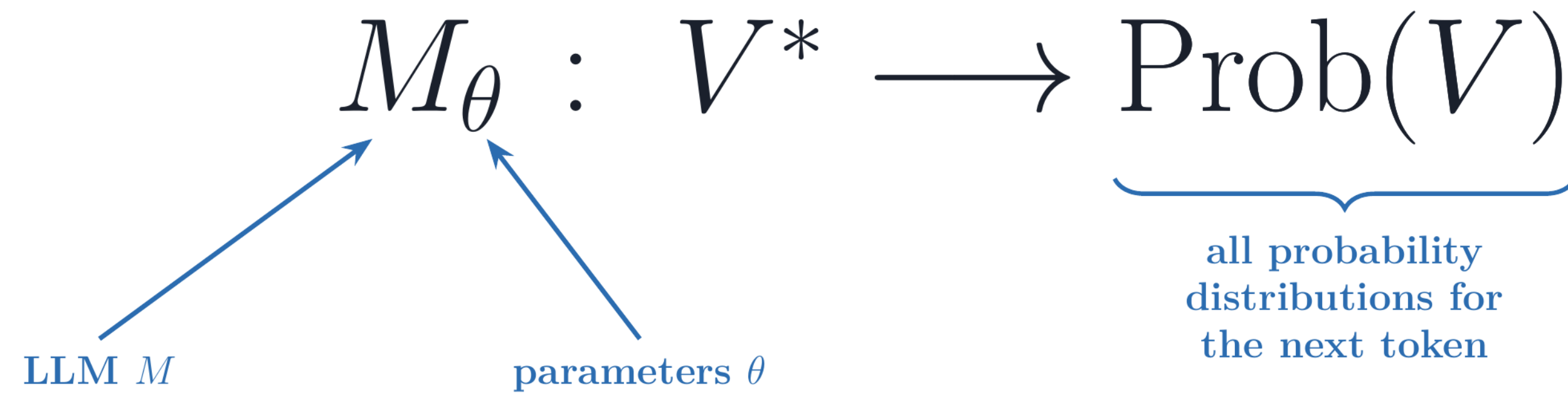
Model	$ V $	d
GPT-2 (small)	50,257	768
GPT-2 XL	50,257	1,600
GPT-3	50,257	12,288
Llama 3 8B	128,256	4,096
Llama 3 70B	128,256	8,192
Llama 3.1 405B	128,256	16,384
DeepSeek-V3	129,280	7,168
gpt-oss-120b	201,088	2,880
Qwen3.5-397B-A17B	248,320	4,096

$\mathbb{R}^V \rightarrow \mathbb{R}^d$: $d \ll |V|$, so the 'embedding' is a projection with a kernel.

MARKOV PROCESS

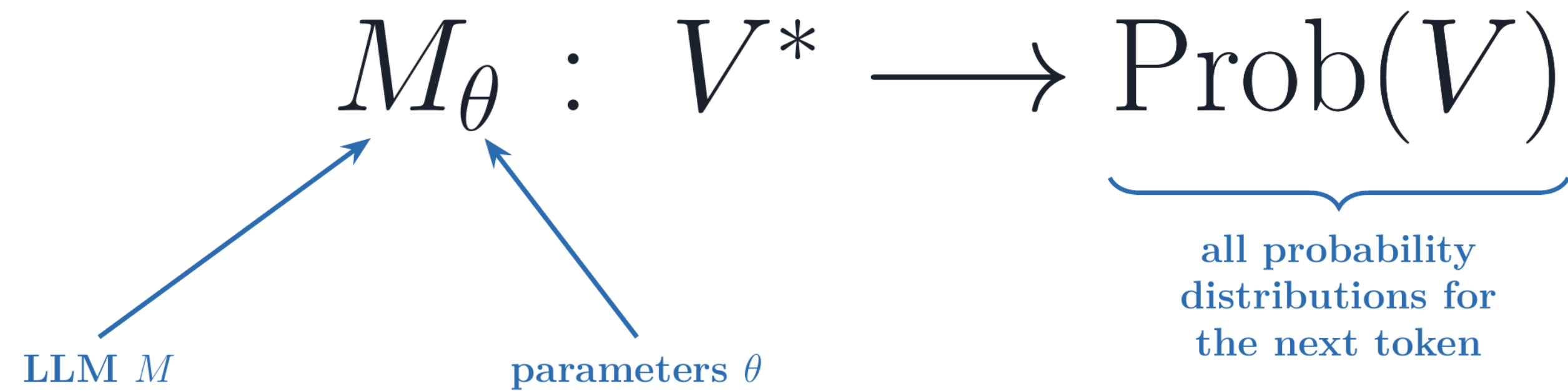
Next-token prediction

V^* = all words made from tokens



Next-token prediction

V^* = all words made from tokens

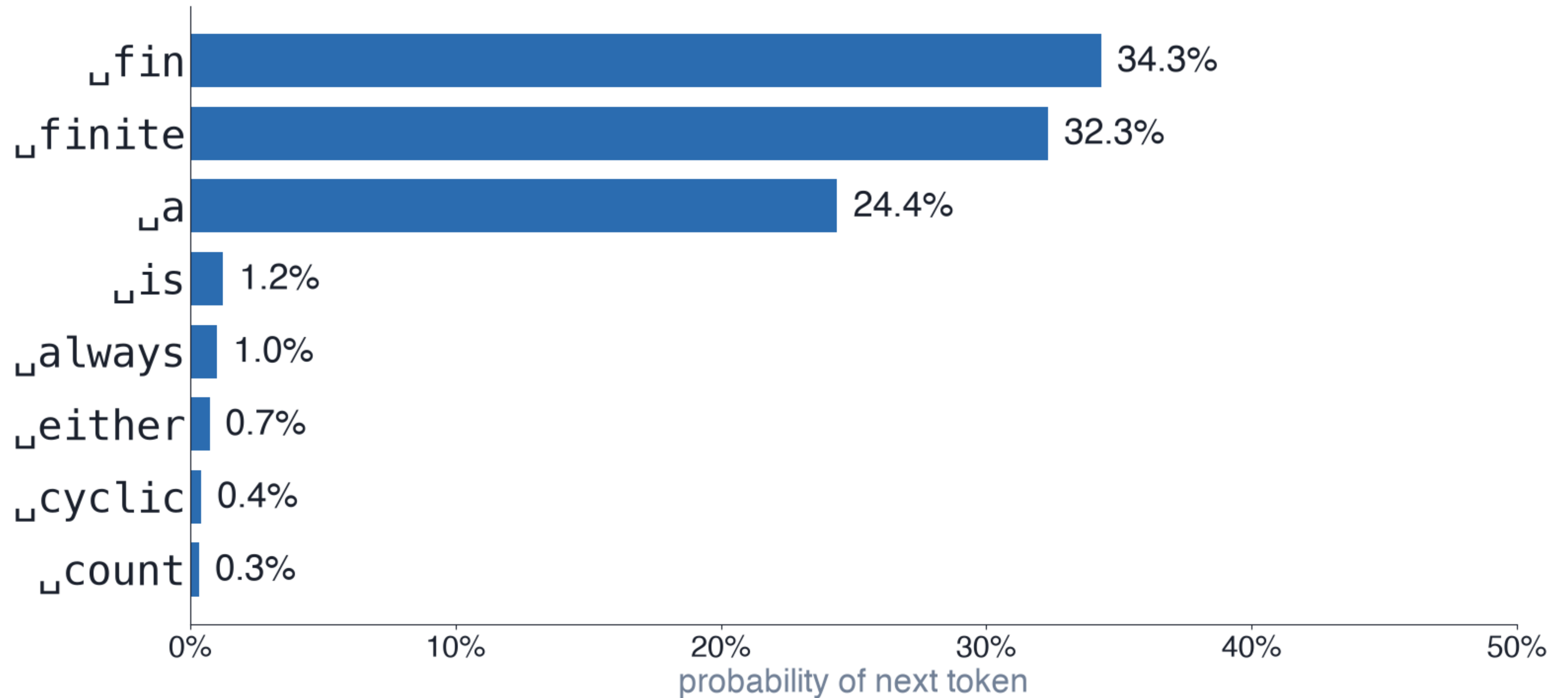


$$M(x)(a) = P(a \mid x), \quad a \in V \qquad M(x) = \begin{cases} 0.6 & a = \text{"the"} \\ 0.3 & a = \text{"an"} \\ 0.1 & a = \text{"a"} \end{cases}$$

[A little more complicated than this.]

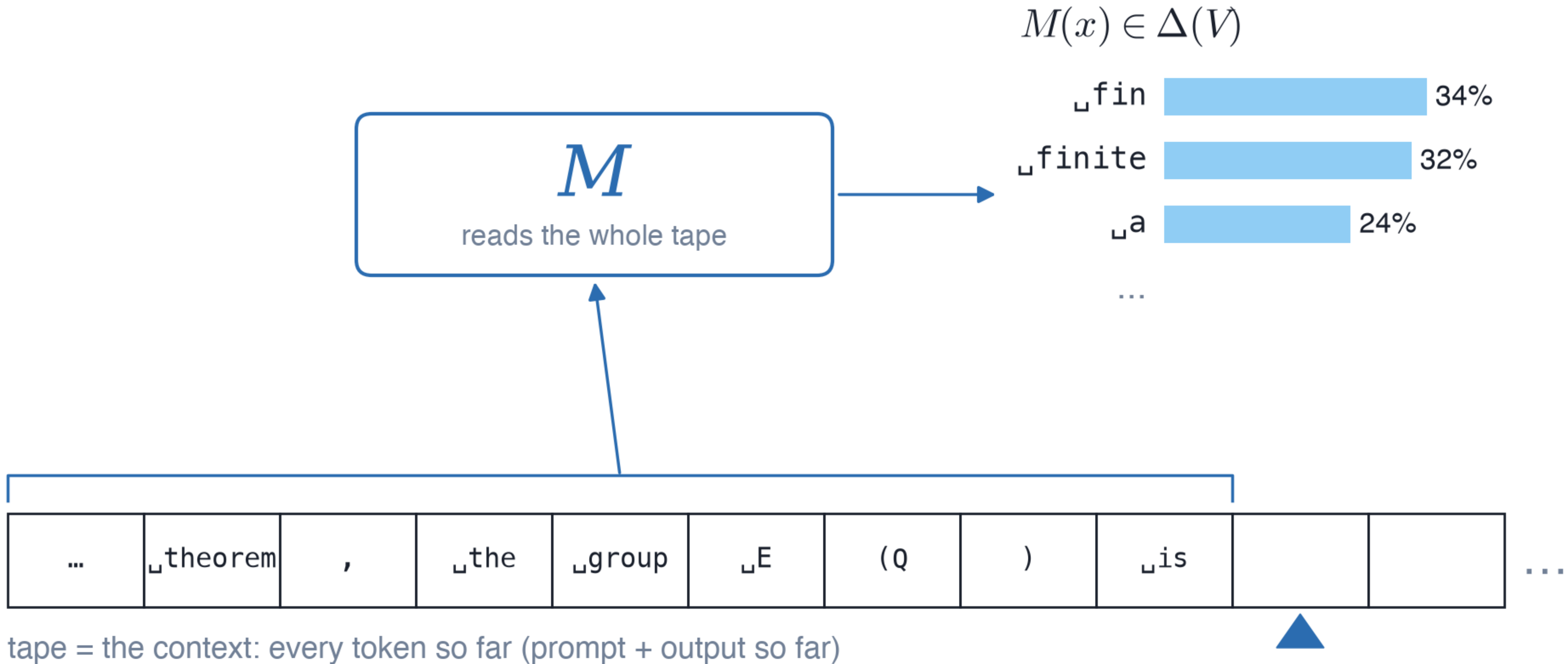
Prompt:

“Let E be an elliptic curve over $\mathbb{Q}$. By the Mordell-Weil theorem,
the group $E(\mathbb{Q})$ is ____”

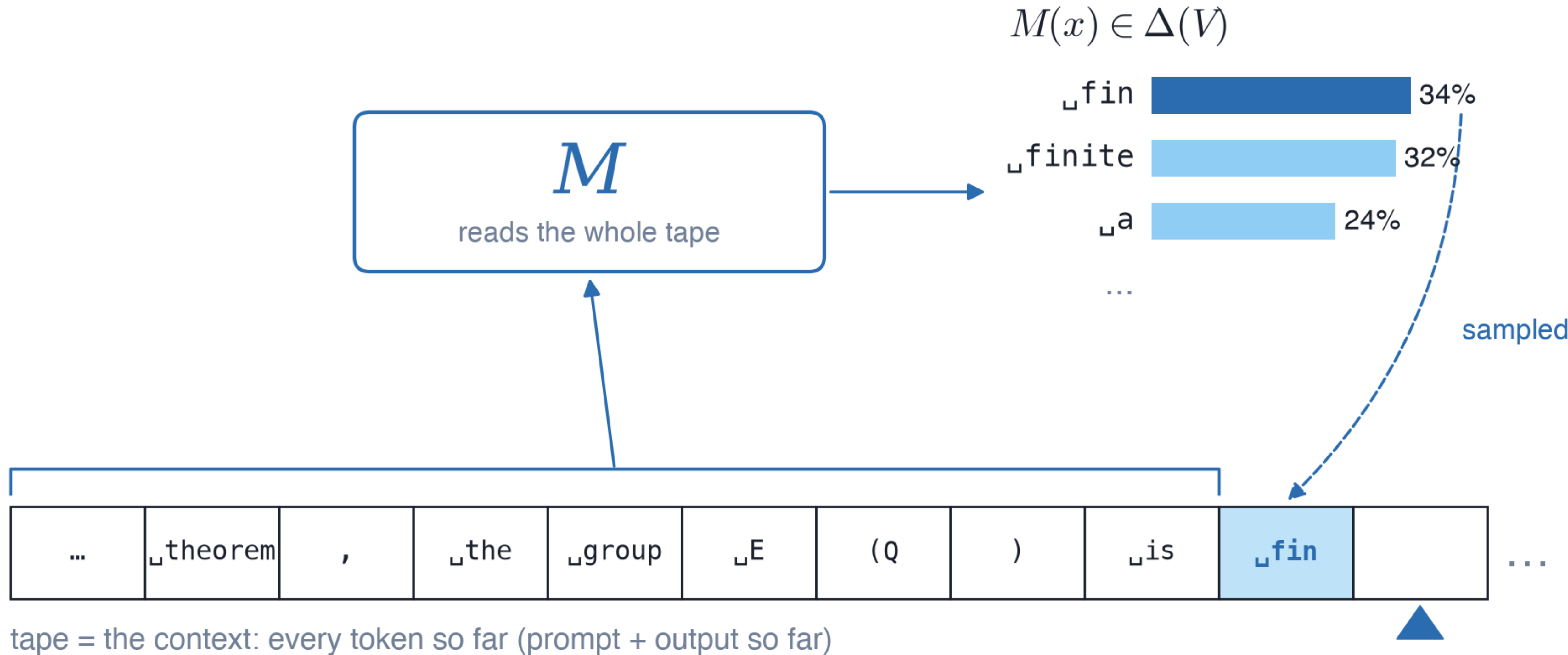


Real log-probabilities from Llama 3.2 3B (run locally via Ollama). _ = leading space. Top 8 of 128,256 tokens shown.

1. Read the tape, output a probability distribution on the next token

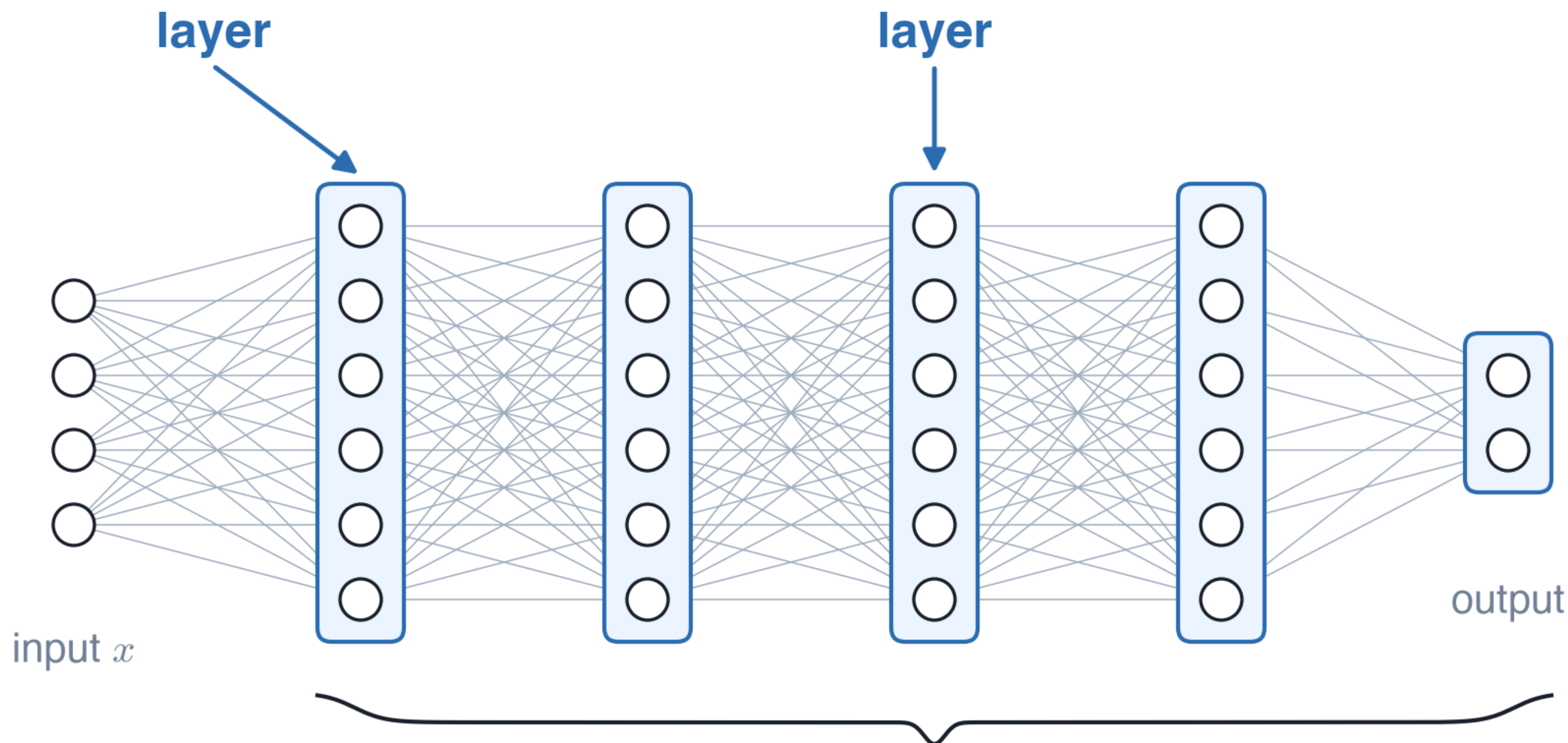


2. Sample a token, write it on the tape, move the head. Repeat.



Probabilities: Llama 3.2 3B, real values (see slide on next-token distribution).

NEURAL NETWORKS



Depth = number of layers

each layer: $x \mapsto \text{ReLU}(Wx + b)$

the entries of W , b are the parameters, fitted by gradient descent

Attention Is All You Need + one transformer layer



Search Submit Donate Log in

Computer Science > Computation and Language

[Submitted on 12 Jun 2017 (v1), last revised 2 Aug 2023 (this version, v7)]

Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Comments: 15 pages, 5 figures

Subjects: **Computation and Language (cs.CL)**; Machine Learning (cs.LG)

Cite as: [arXiv:1706.03762](https://arxiv.org/abs/1706.03762) [cs.CL]

(or [arXiv:1706.03762v7](https://arxiv.org/abs/1706.03762v7) [cs.CL] for this version)

<https://doi.org/10.48550/arXiv.1706.03762>

Access Paper:

[View PDF](#)
[HTML \(experimental\)](#)
[TeX Source](#)

[view license](#)

Additional Features

[Audio Summary](#)

Current browse context:

cs.CL

[< prev](#) | [next >](#)
[new](#) | [recent](#) | [2017-06](#)

Change to browse by:

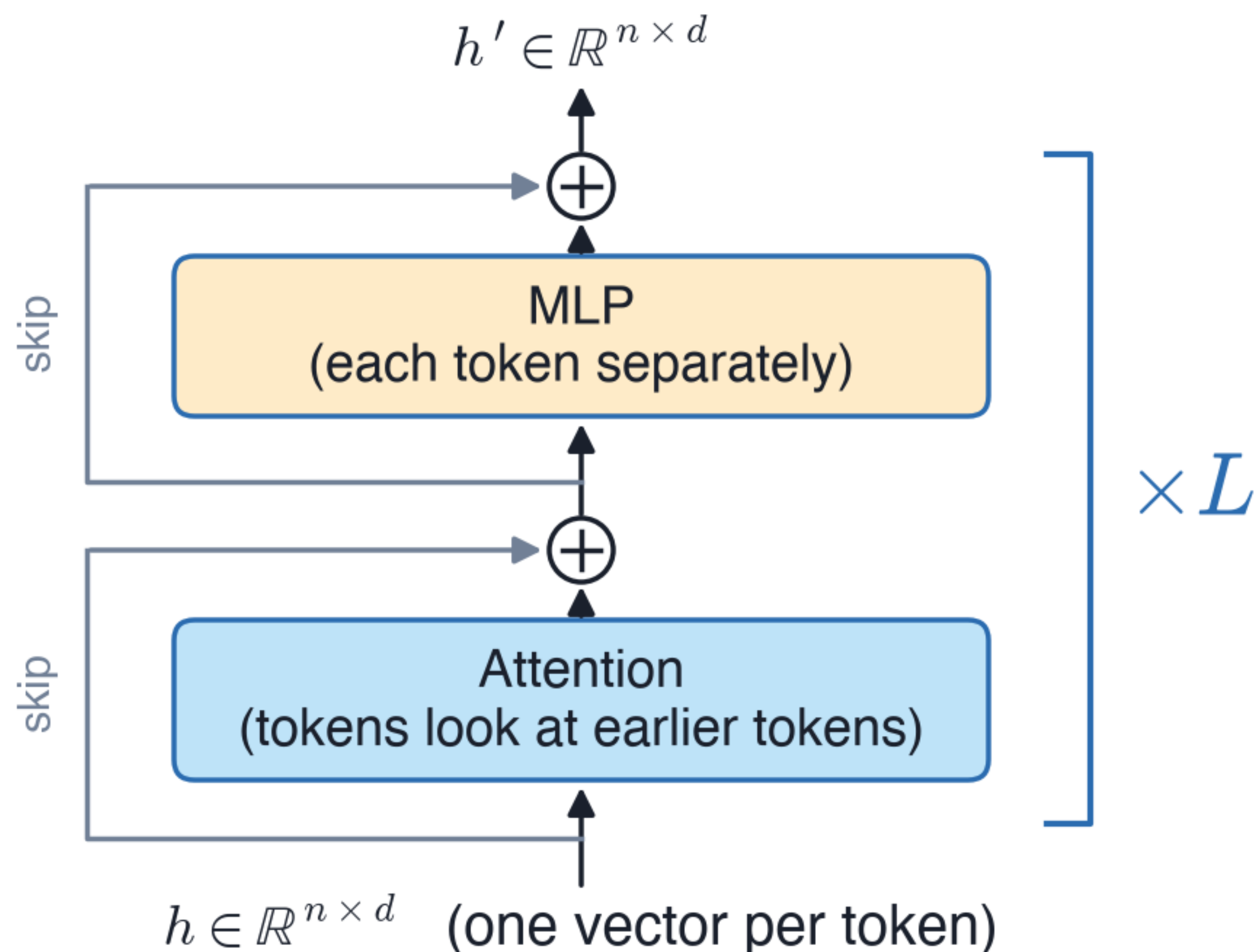
[cs](#)
[cs.LG](#)

References & Citations

[SciX](#)
[Google Scholar](#)
[Semantic Scholar](#)

[123 blog links](#) (what is this?)

One transformer layer (decoder-only, as in GPT / Llama / Claude)



Attention:

$$\text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + \text{mask}\right) h W_{\text{val}},$$

$$Q = hW_Q, \quad K = hW_K \quad (\text{H heads in parallel})$$

mask: token i only sees tokens $1..i$

MLP:

$$W_2 \sigma(W_1 h) \quad (\text{as on the neural-net slide})$$

Skip:

$$h \mapsto h + \text{Block}(\text{LayerNorm}(h))$$

At the top: logits $\ell = h_n W_{\text{out}} \in \mathbb{R}^V$, then softmax.

Parameters: the constants we fit during gradient descent

Parameters and layers

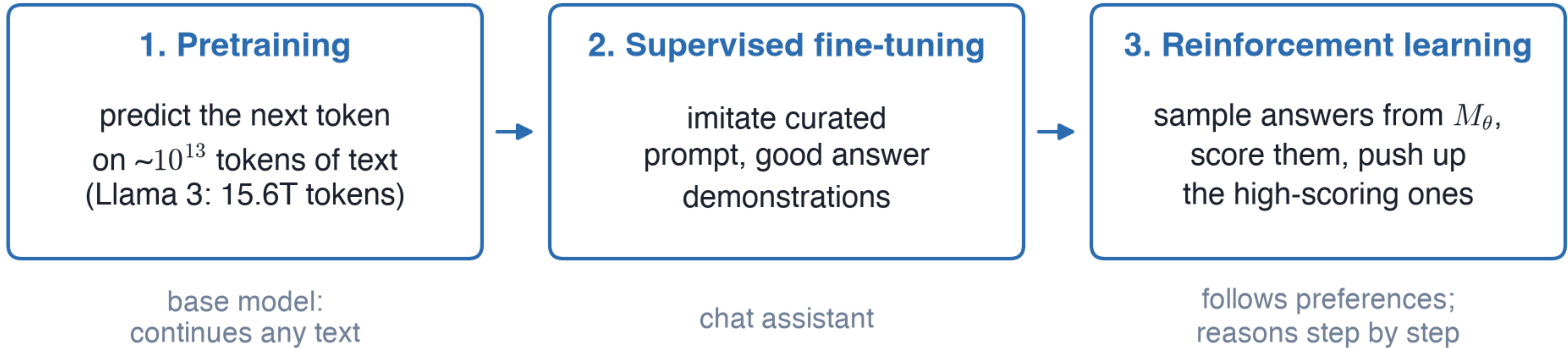
Model	Year	Parameters	Layers (depth)
GPT-2 (small)	2019	124M	12
GPT-2 XL	2019	1.5B	48
GPT-3	2020	175B	96
Llama 3 8B	2024	8B	32
Llama 3 70B	2024	70B	80
Llama 3.1 405B	2024	405B	126
DeepSeek-V3 (MoE)	2024	671B total / 37B active	61
gpt-oss-120b (MoE)	2025	117B total / 5.1B active	36
Qwen3.5-397B-A17B (MoE)	2026	397B total / 17B active	60
Frontier (Fable 5.1, GPT-6.1 Sol, Gemini 4 Argon)	2026	not disclosed	not disclosed

TRAINING: PARAMETER ESTIMATION

Training

- Pre-training: get parameters so LLMs understand words
- Post-training ("fine tuning"): reinforcement learning with reward functions
- Post-training uses LOTS of math papers, Lean, etc.
- Unclear what matters, but post-training dramatically improves Olympiad-style reasoning

From next-token predictor to assistant to reasoner



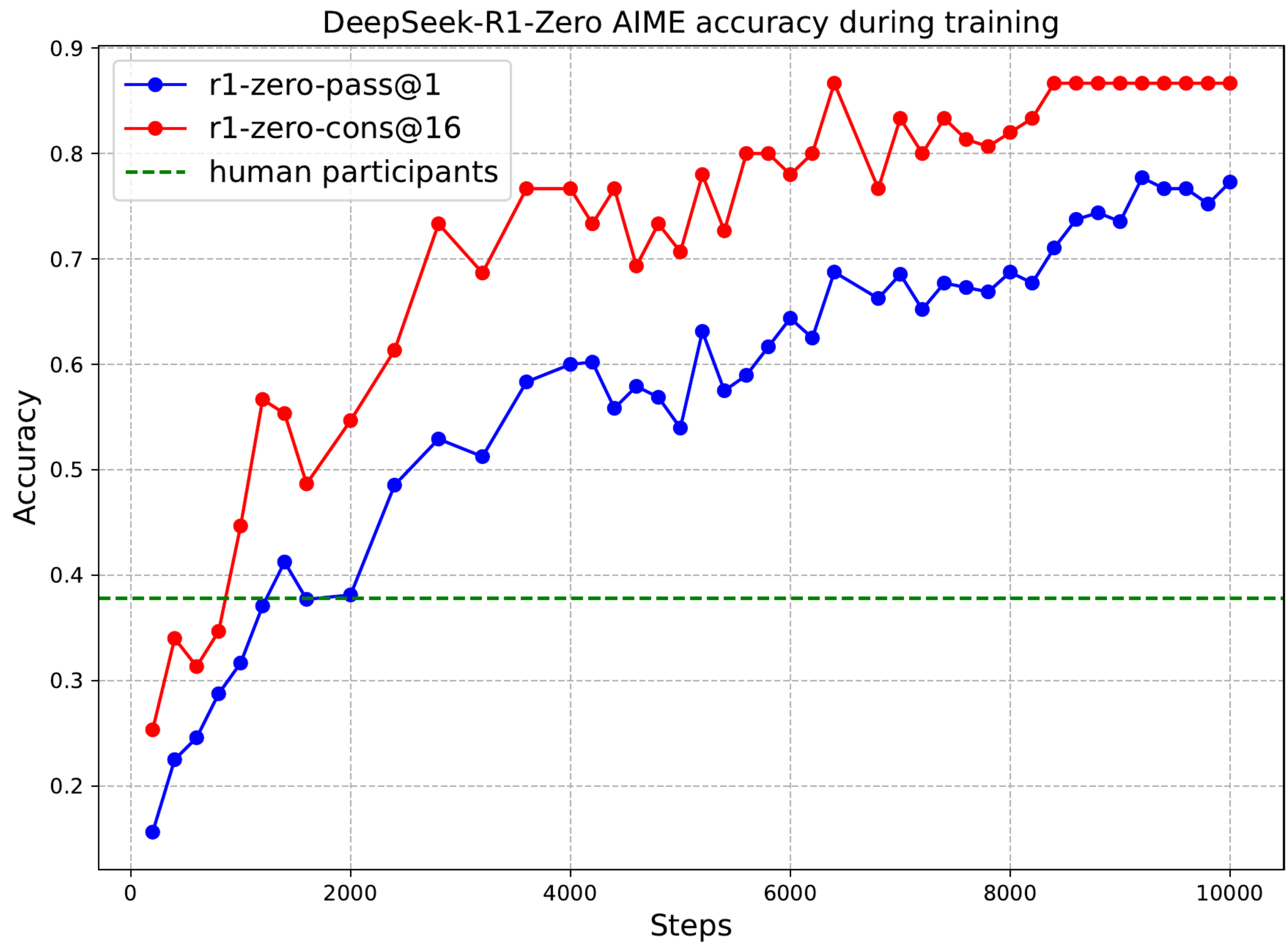
Reward $r(x, y)$ comes from:

RLHF: a reward model trained on human preference comparisons (InstructGPT, 2022)

RLVR: a verifier — correct final answer, unit tests pass, Lean accepts the proof (Tulu 3, 2024; DeepSeek-R1, 2025)

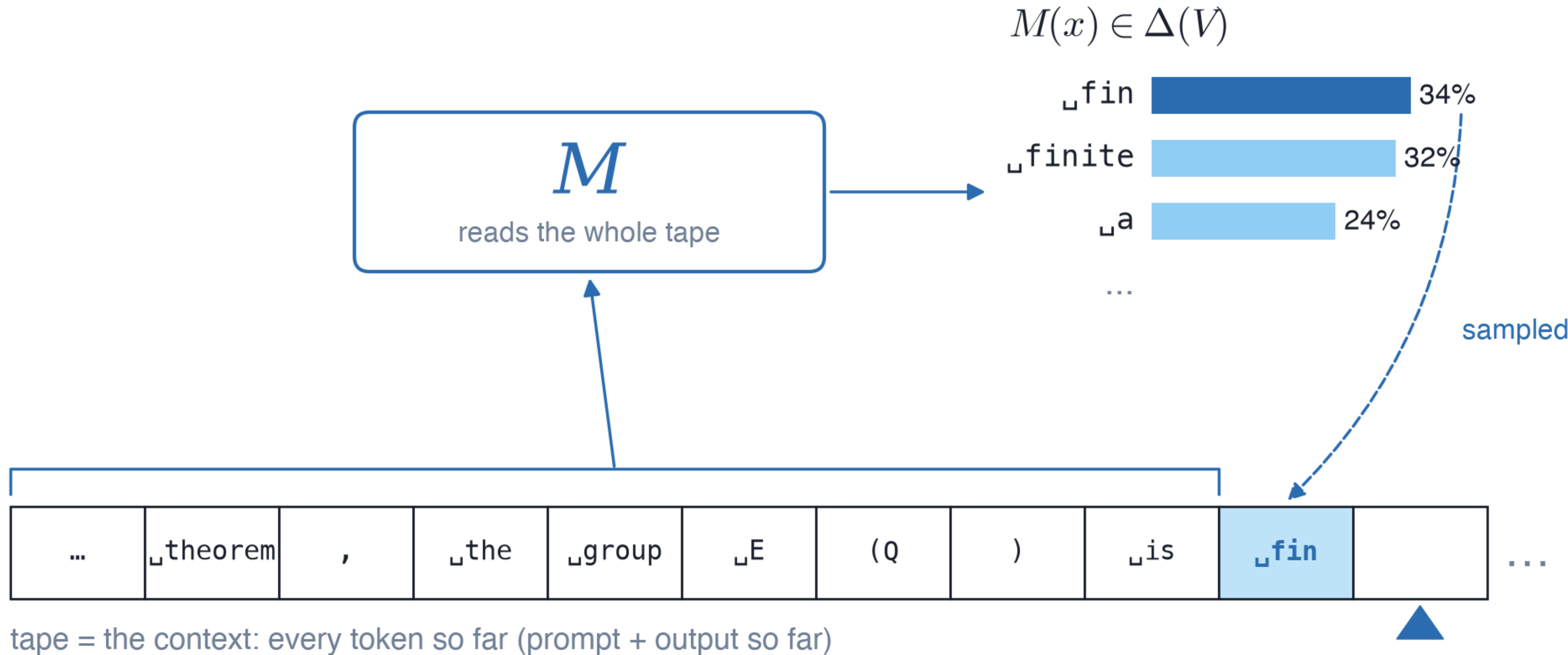
Objective: maximise $\mathbb{E}_{y \sim M_\theta(x)} [r(x, y)] - \beta \text{KL}(M_\theta(x) \parallel M_{\text{ref}}(x))$

DeepSeek-R1-Zero: AIME 2024 accuracy during RL (Nature, 2025)

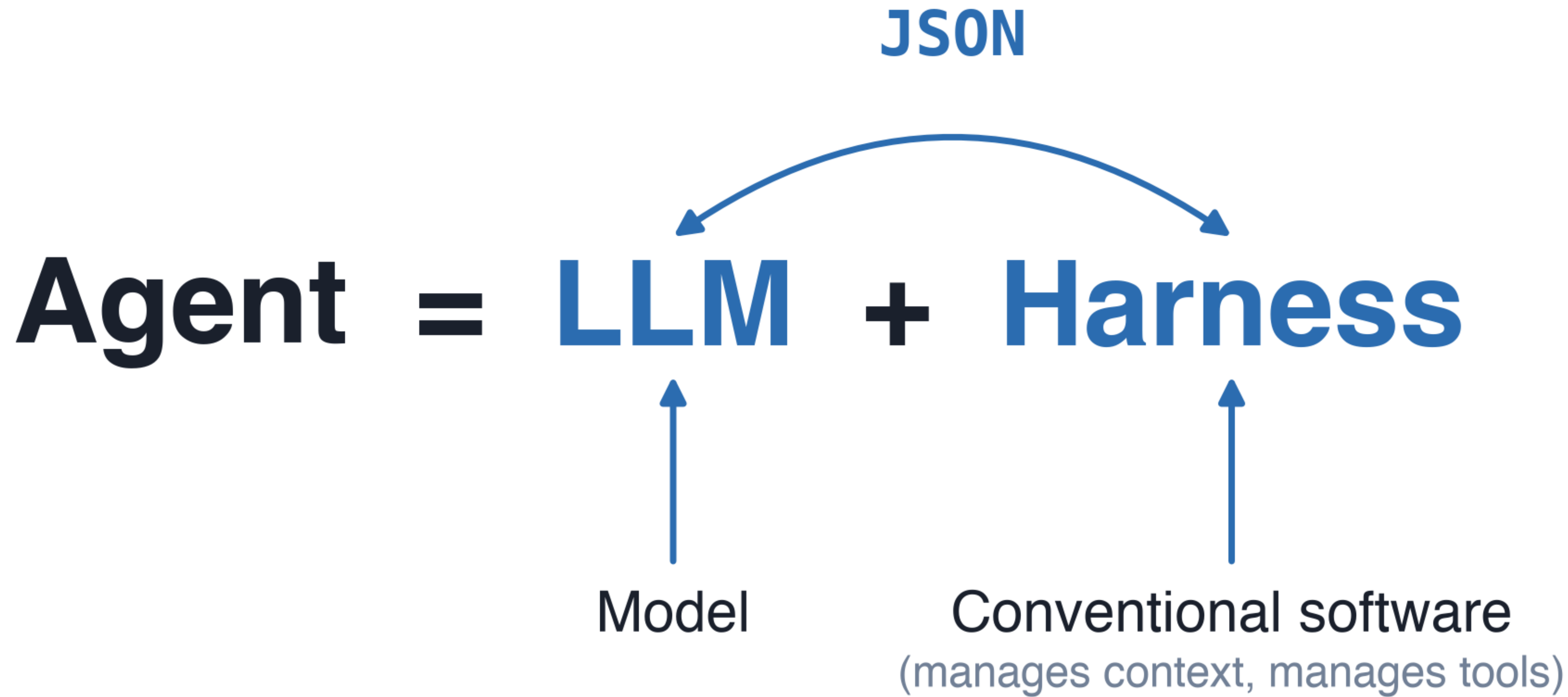


HARNESSES

2. Sample a token, write it on the tape, move the head. Repeat.



Probabilities: Llama 3.2 3B, real values (see slide on next-token distribution).



tool call (JSON)

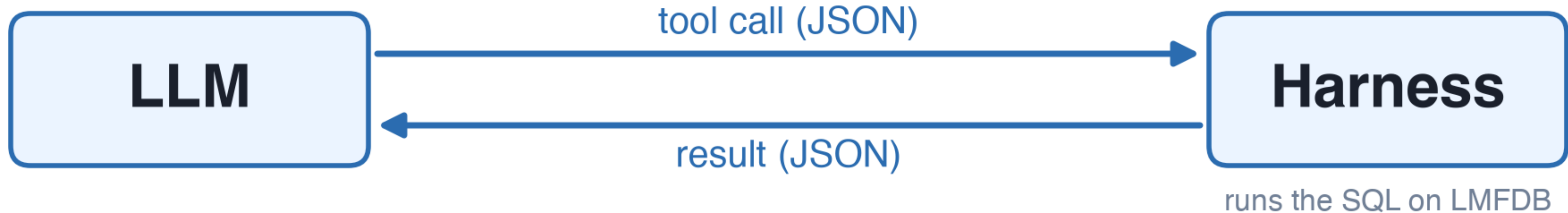
LLM

Harness

runs the SQL on LMFDB

the model emits:

```
{
  "tool": "lmfdb.run_sql",
  "input": {
    "sql": "SELECT lmfdb_label FROM ec_curvedata WHERE rank=2 ORDER BY conductor LIMIT 1"
  }
}
```



the model emits:

```
{
  "tool": "lmfdb.run_sql",
  "input": {
    "sql": "SELECT lmfdb_label FROM ec_curvedata WHERE rank=2 ORDER BY conductor LIMIT 1"
  }
}
```

the harness replies; the reply is appended to the context:

```
{"result": "389.a1"}
```

389.a1: $y^2 + y = x^3 + x^2 - 2x$, conductor 389
(verified live on LMFDB, 2026-10-23)

	 OpenAI	 Anthropic	 Google	Open source
Terminal (CLI)	Codex CLI	Claude Code	Antigravity CLI agy	OpenCode Oh My Pi omp
Desktop app	Codex app	Claude Cowork (inside the Claude app)	Antigravity IDE	—
Web	ChatGPT	claude.ai Claude Design	Gemini app	—

Claude Code → any model via OpenRouter

```
export ANTHROPIC_BASE_URL="https://openrouter.ai/api"
export ANTHROPIC_AUTH_TOKEN="$OPENROUTER_API_KEY"
export ANTHROPIC_API_KEY=""
export ANTHROPIC_DEFAULT_SONNET_MODEL="openai/gpt-6.1-sol"
claude
```

Claude Code → local model via Ollama

```
ollama launch claude
# or: ANTHROPIC_BASE_URL=http://localhost:11434 \
#     claude --model qwen3.5
```

Codex → another provider (~/.codex/config.toml)

```
model = "anthropic/claude-sonnet-5.5"
model_provider = "openrouter"

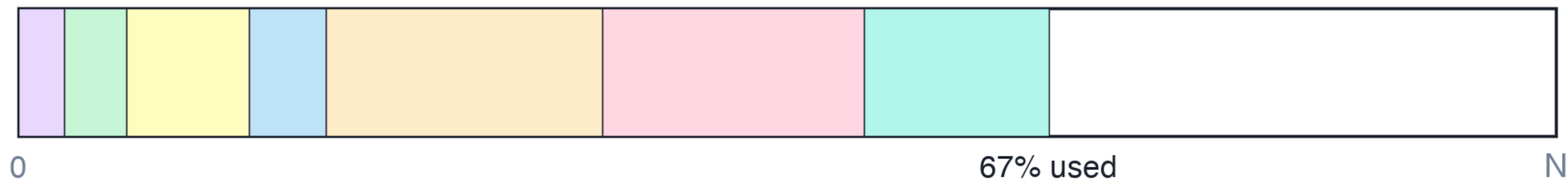
[model_providers.openrouter]
name = "OpenRouter"
base_url = "https://openrouter.ai/api/v1"
env_key = "OPENROUTER_API_KEY"
```

Codex → local model

```
codex --oss           # Ollama or LM Studio
```

CONTEXT MANAGEMENT

The context window: everything the model sees on this call



- system prompt
- MCP tool definitions
- files read
- model replies + thinking

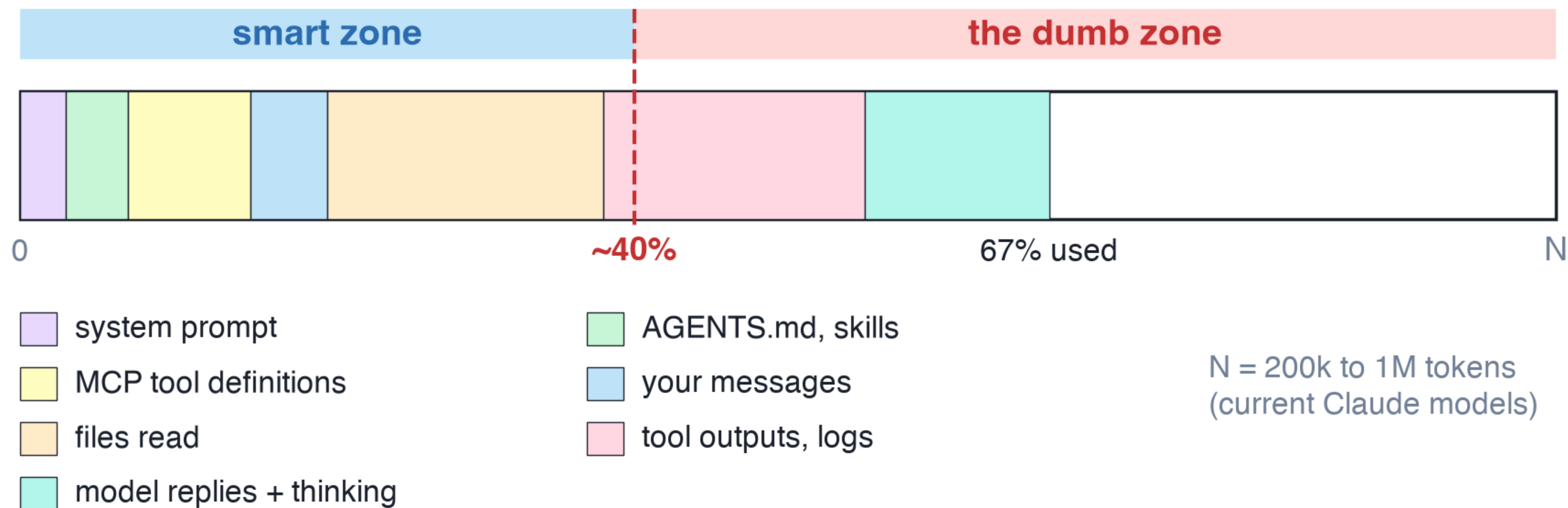
- AGENTS.md, skills
- your messages
- tool outputs, logs

N = 200k to 1M tokens
(current Claude models)

Every token here is read again for each new token, and every input token is billed.

Segment sizes are illustrative.

The context window: everything the model sees on this call

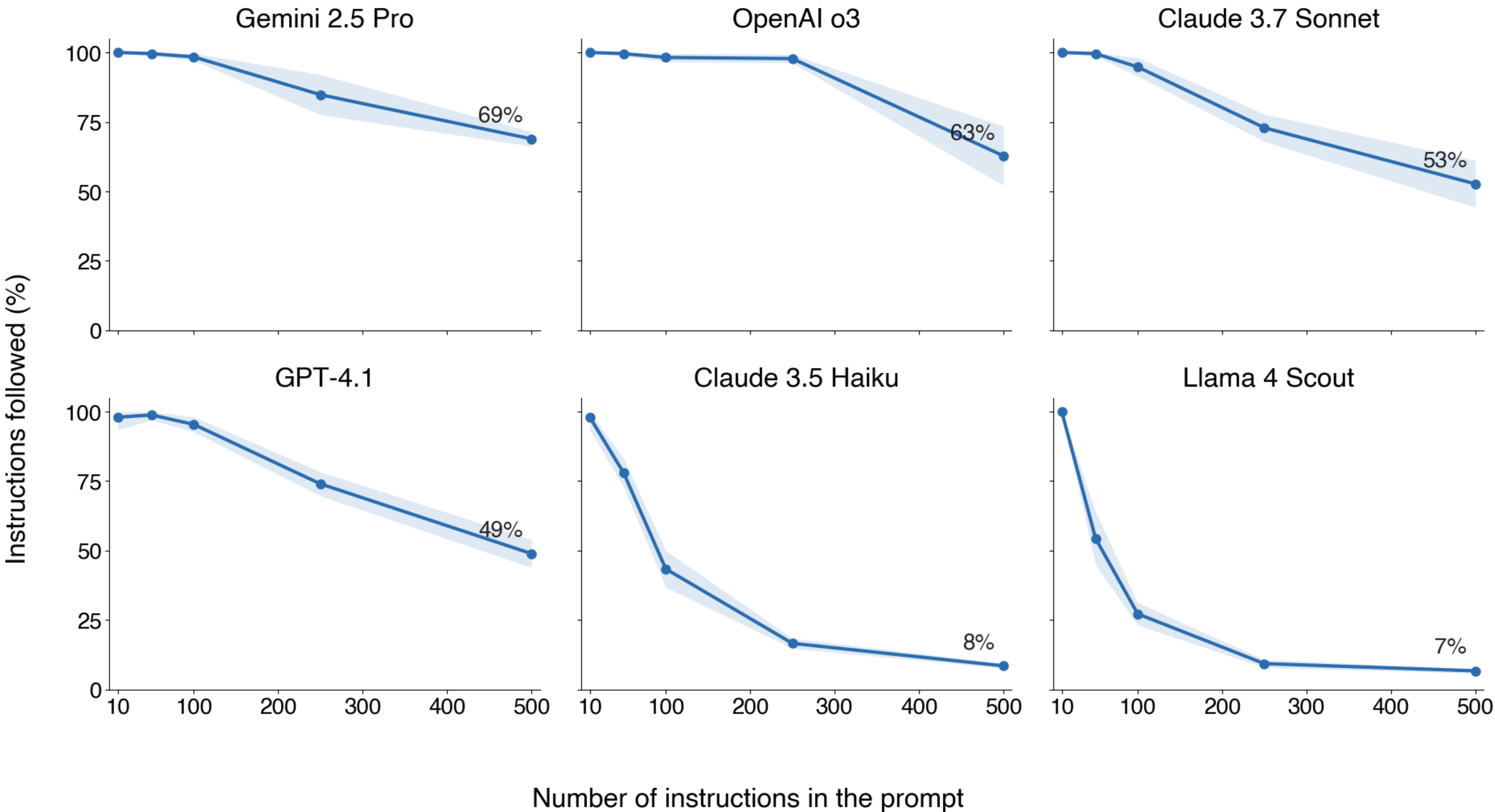


The more of the window you fill, the worse the results.

Term: Dex Horthy (HumanLayer), "No Vibes Allowed", AI Engineer Code Summit, Nov 2025. ~40% is a rule of thumb; it depends on the task.

Fixes: fresh sessions, compact to a file, sub-agents, fewer MCP tools.

Model	Parameters	Context window (tokens)	AA Intelligence Index (v4.3.2)
 Claude Fable 5.1	undisclosed	1M	53
 Claude Opus 5.5	undisclosed	1M	58
 GPT-6 Astra	undisclosed	1.05M	53
 Gemini 4 Argon	undisclosed	1M	53
 DeepSeek V4 Pro	1.6T total 49B active per token	1M	36
 Claude Haiku 5.5	undisclosed	1M	43
 GPT-6 Luna	undisclosed	1.05M	38
 Gemma 4 31B (open)	30.7B	256K	15
 Gemma 4 E4B (mini, open)	4.5B effective 8B with embeddings	128K	9



How Many Instructions Can LLMs Follow at Once?

Computer Science > Artificial Intelligence

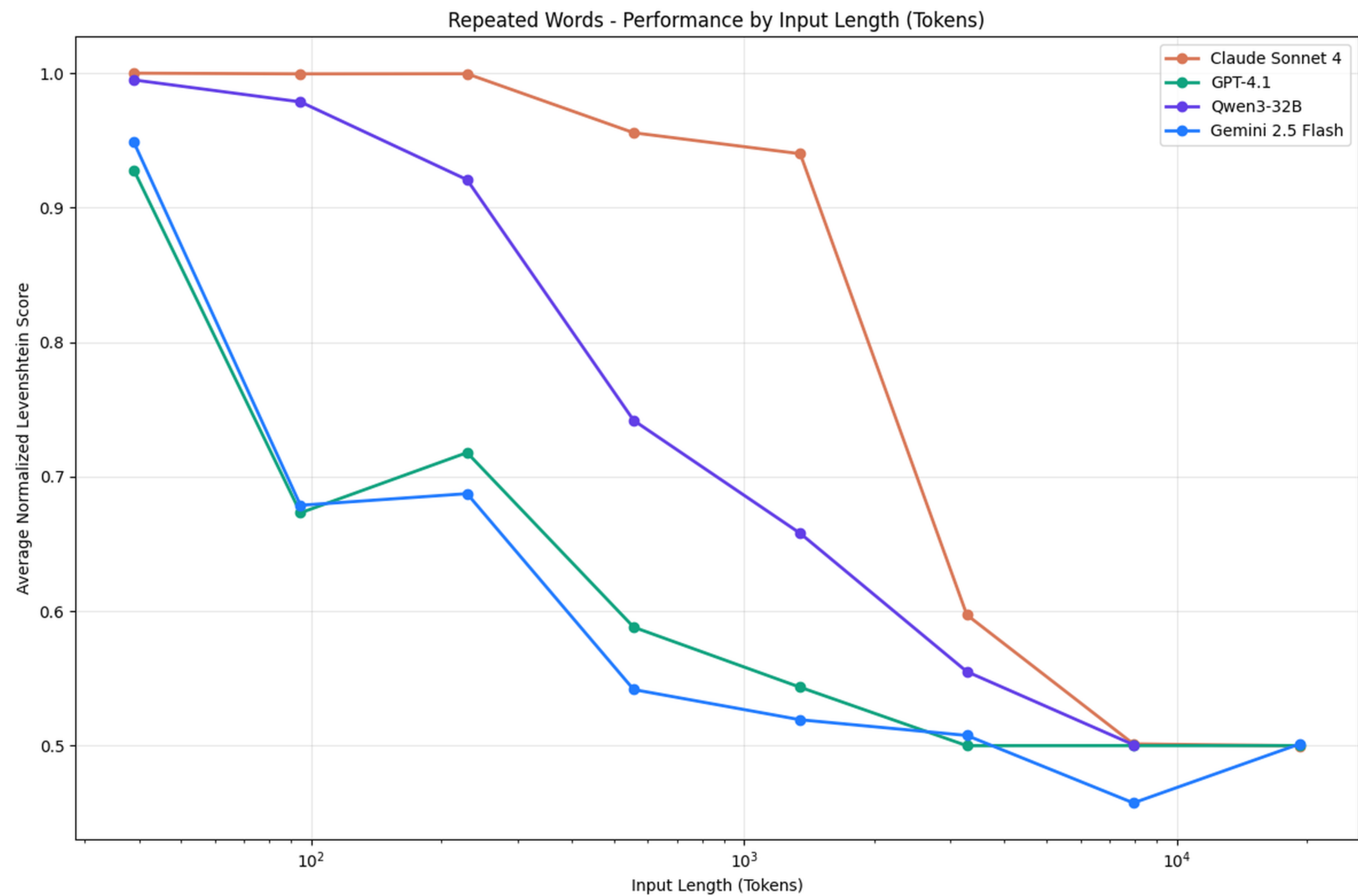
[Submitted on 15 Jul 2025]

How Many Instructions Can LLMs Follow at Once?

Daniel Jaroslawicz, Brendan Whiting, Parth Shah, Karime Maamari

Production-grade LLM systems require robust adherence to dozens or even hundreds of instructions simultaneously. However, the instruction-following capabilities of LLMs at high instruction densities have not yet been characterized, as existing benchmarks only evaluate models on tasks with a single or few instructions. We introduce IFScale, a simple benchmark of 500 keyword-inclusion instructions for a business report writing task to measure how instruction-following performance degrades as instruction density increases. We evaluate 20 state-of-the-art models across seven major providers and find that even the best frontier models only achieve 68% accuracy at the max density of 500 instructions. Our analysis reveals model size and reasoning capability to correlate with 3 distinct performance degradation patterns, bias towards earlier instructions, and distinct categories of instruction-following errors. Our insights can help inform design of instruction-dense prompts in real-world applications and highlight important performance-latency tradeoffs. We open-source the benchmark and all results for further analysis at [this https URL](#).

Context, the context window, and "the dumb zone"



MODEL BENCHMARKS

	 DeepSeek	 OpenAI	 Anthropic	 Google DeepMind	 Open weights (run via Ollama)
Top	V4 Pro	GPT-6 Astra	Claude Fable 5.1	Gemini 4 Argon	GLM-5.3 Kimi K3
Mid	—	GPT-6.1 Sol GPT-5.6 Terra	Claude Opus 5.5 Claude Sonnet 5.5	Gemini 3.8 Flash	Qwen3.8 27B Gemma 4 31B
Small	V4.1 Flash	GPT-6 Luna	Claude Haiku 5.5	Gemini 3.5 Flash-Lite	Gemma 4 E4B / E2B runs on a laptop

API list price, US\$ per million tokens

AI Anthropic

Model	Input	Output
Claude Haiku 5.5	\$0.10	\$0.50
Claude Sonnet 5.5	\$2	\$10
Claude Opus 5.5	\$4	\$20
Claude Fable 5.1	\$10	\$50

 OpenAI

Model	Input	Output
GPT-6 Luna	\$0.10	\$0.50
GPT-5.6 Terra	\$2	\$12
GPT-6.1 Sol	\$2	\$10
GPT-6 Astra	\$10	\$50

Sources: platform.claude.com/docs/en/about-claude/pricing; developers.openai.com/api/docs/pricing (standard tier, ≤272K-token prompts). Retrieved 22 Oct 2026. Haiku 5.5 price is for prompts ≤100K tokens.

API list price, US\$ per million tokens

AI Anthropic

Model	Input	Output
Claude Haiku 5.5	\$0.10	\$0.50
Claude Sonnet 5.5	\$2	\$10
Claude Opus 5.5	\$4	\$20
Claude Fable 5.1	\$10	\$50

Frontier · Sept 2026

 OpenAI

Model	Input	Output
GPT-6 Luna	\$0.10	\$0.50
GPT-5.6 Terra	\$2	\$12
GPT-6.1 Sol	\$2	\$10
GPT-6 Astra	\$10	\$50

Frontier · Sept 2026

Sources: platform.claude.com/docs/en/about-claude/pricing; developers.openai.com/api/docs/pricing (standard tier, ≤272K-token prompts). Retrieved 22 Oct 2026. Haiku 5.5 price is for prompts ≤100K tokens.

TOKEN MARKETS

Move document A into folder B

GPT-6 Astra

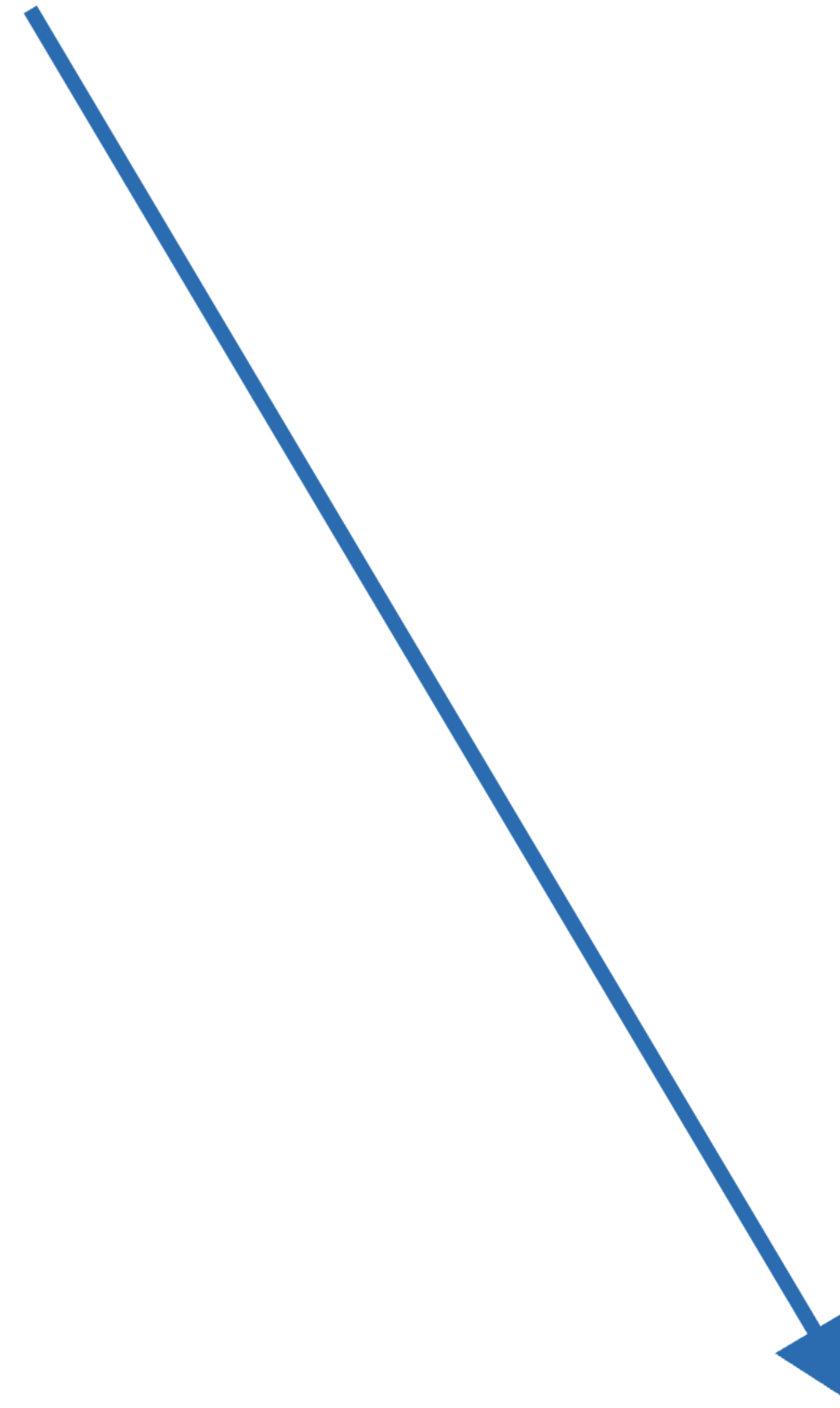
Find an exotic Tate class on an abelian variety
not generated by cup products of divisor classes

Claude Haiku 5.5

Move document A into folder B

Find an exotic Tate class on an abelian variety
not generated by cup products of divisor classes

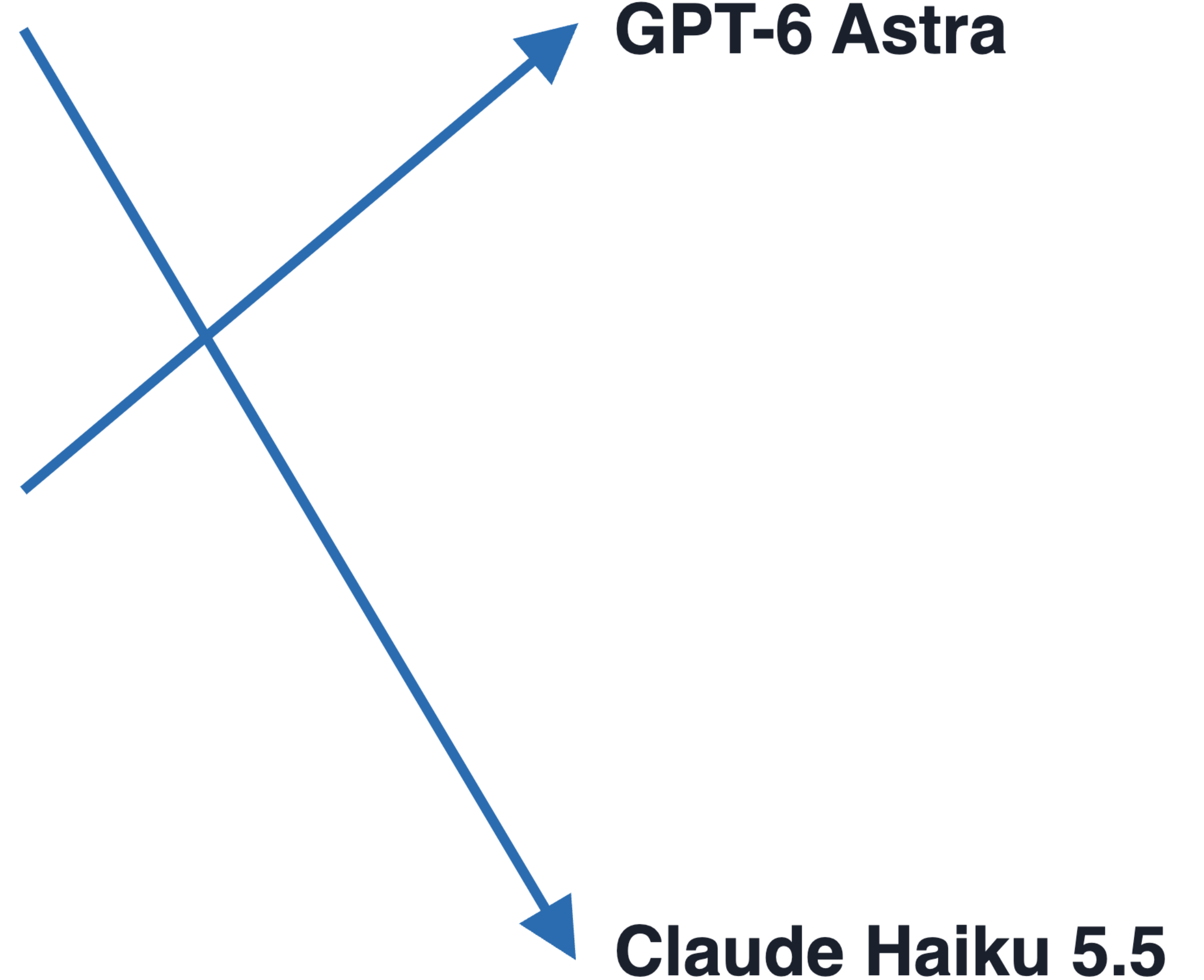
GPT-6 Astra



Claude Haiku 5.5

Move document A into folder B

Find an exotic Tate class on an abelian variety
not generated by cup products of divisor classes



Capability vs. price per token



Data: Artificial Analysis, artificialanalysis.ai/leaderboards/models (22 Oct 2026); best effort setting per model; AA-listed API prices. Redrawn from AA's numbers.

- Text
- Image
- Video
- Speech
- Decisions
- Embeddings
- Rerank
- Transcription
- Batch

- Top Models
- Leaderboard
- Top models by task
- Cost per session
- Market Share
- Open / Closed
- Benchmarks
- Fastest models
- Languages
- Programming
- Domain
- Context Length
- Tool Calls
- Images
- Top Apps

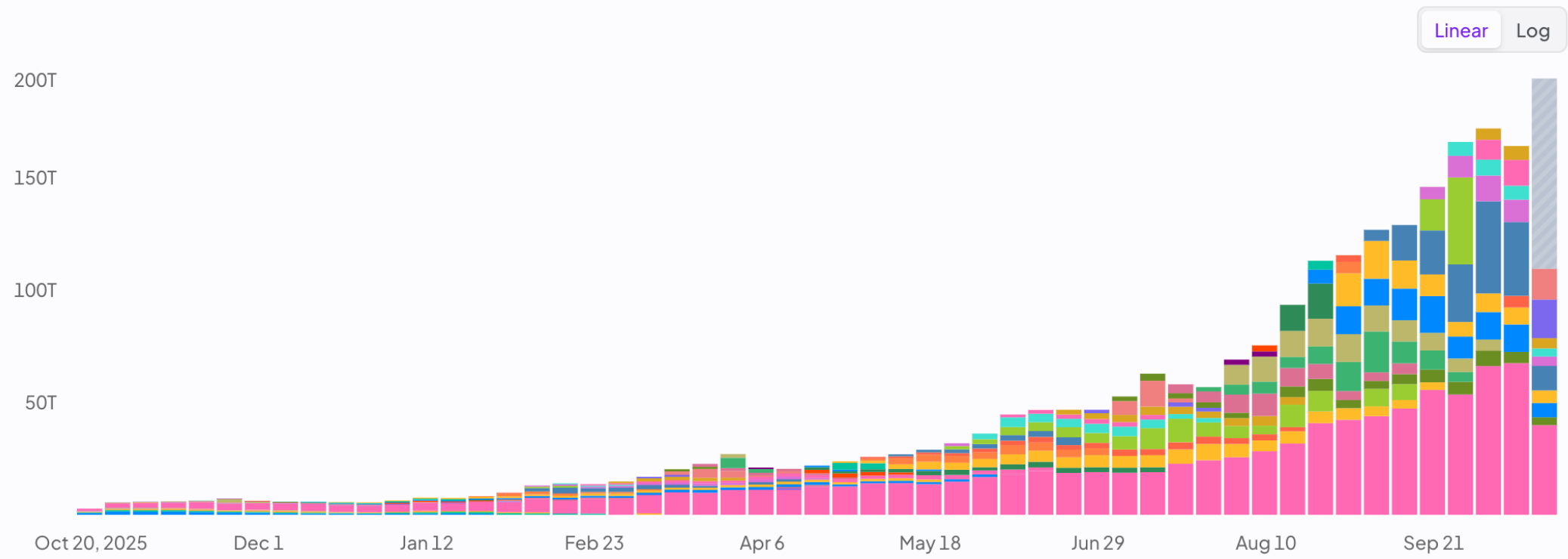
AI Model Rankings

Live LLM rankings by real-world usage. Models are ranked by tokens processed through the OpenRouter API by millions of developers. [View all text models](#)

Usage data through Oct 21, 2026

Top Models

Weekly usage of models across OpenRouter



LLM Leaderboard

All models This Week

Compare the most popular models on OpenRouter ⓘ

1.	 DeepSeek V4.1 Flash by deepseek	25.7T tokens ↓31%	6.	 Hy4 preview by tencent	8.55T tokens ↑9%
2.	 DeepSeek V4.1 Flash 1017 by deepseek	16.5T tokens new	7.	 Claude Opus 5.5 by anthropic	7.14T tokens ↑26%
3.	 GLM 5.3 Flash by z-ai	11.3T tokens ↓12%	8.	 GPT-6 Luna by openai	6.08T tokens ↓14%
4.	 MiniMax M3.1 Flash (free) by minimax	10.3T tokens new	9.	 GLM 5.3 by z-ai	5.21T tokens ↑18%
5.	 MiMo-V2.6-Flash by xiaomi	8.57T tokens ↓25%	10.	 Nemotron 3 Ultra (free) by nvidia	4.46T tokens ↓40%

The model market (OpenRouter)

🏆 LLM Leaderboard

Compare the most popular models on OpenRouter ⓘ

All models ▾

This Week ▾

1.		<u>DeepSeek V4.1 Flash</u> by <u>deepseek</u>	25.7T tokens ↓31%	6.		<u>Hy4 preview</u> by <u>tencent</u>	8.55T tokens ↑9%
2.		<u>DeepSeek V4.1 Flash 1017</u> by <u>deepseek</u>	16.5T tokens new	7.		<u>Claude Opus 5.5</u> by <u>anthropic</u>	7.14T tokens ↑26%
3.		<u>GLM 5.3 Flash</u> by <u>z-ai</u>	11.3T tokens ↓12%	8.		<u>GPT-6 Luna</u> by <u>openai</u>	6.08T tokens ↓14%
4.		<u>MiniMax M3.1 Flash (free)</u> by <u>minimax</u>	10.3T tokens new	9.		<u>GLM 5.3</u> by <u>z-ai</u>	5.21T tokens ↑18%
5.		<u>MiMo-V2.6-Flash</u> by <u>xiaomi</u>	8.57T tokens ↓25%	10.		<u>Nemotron 3 Ultra (free)</u> by <u>nvidia</u>	4.46T tokens ↓40%

Show more ▾

Benchmarks

Independent, reproducible measurements of the knobs you can actually set on an OpenRouter request: models, providers, search engines, and tool budgets. Every score links to the configuration, costs, and telemetry behind it.

18 benchmarks • 2,160,983 task evaluations • last run Oct 22, 2026

- All
- Agents
- Media
- Artifacts
- Reasoning
- Search
- Routers

Agents & tools

BENCHMARK	🕒 QUALITY	💰 VALUE	⚡ SPEED
τ²–Bench Airline > Multi-turn service agents making tool calls under strict policy constraints. 148 models • last run Oct 22, 2026	80.2% 🇦🇨 Claude Fable 5	\$0.007 🇹🇼 GLM 5.3 Flash	42s 🇹🇼 GLM 5.3

Media

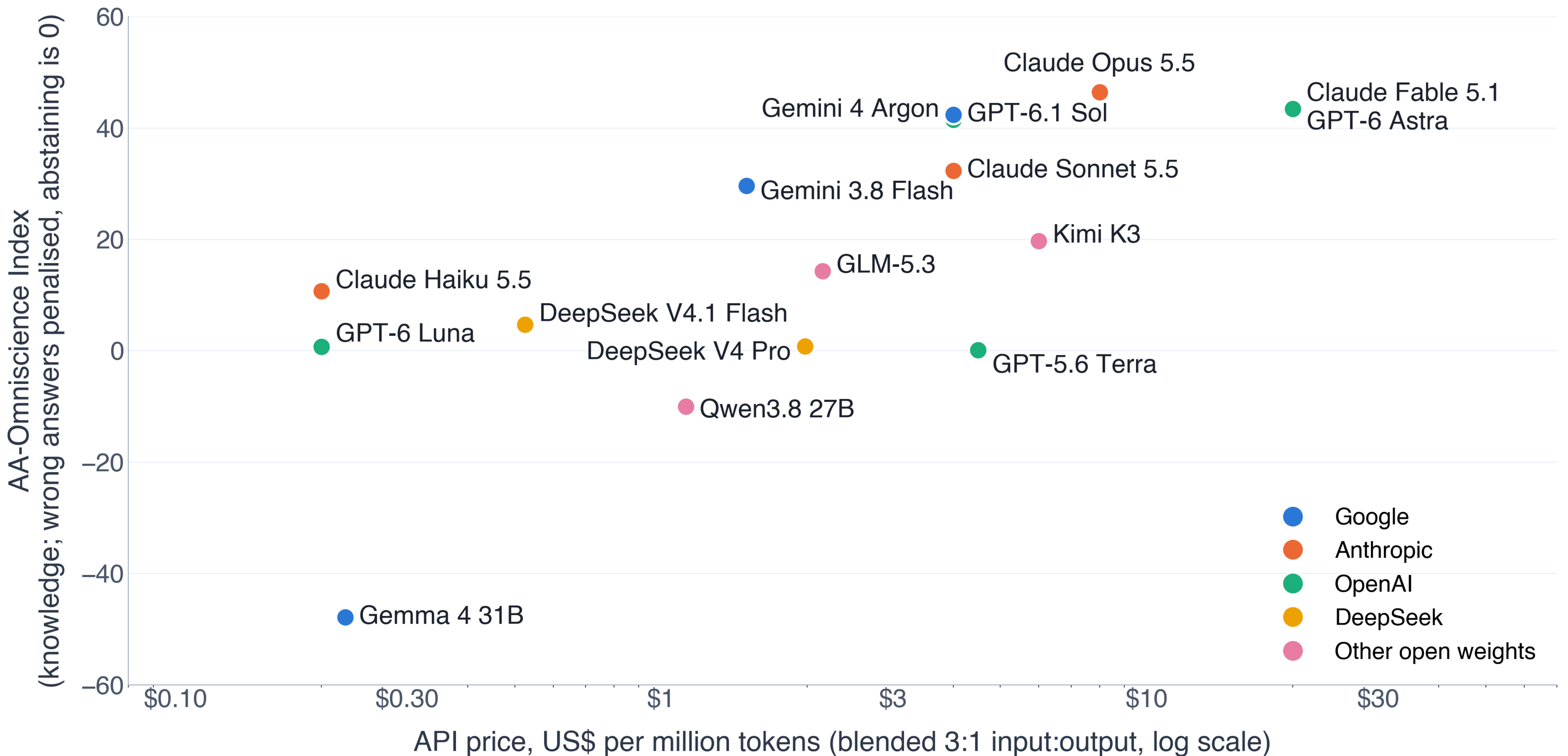
Generated images, clips and speech, graded against the request and priced per output.

BENCHMARK	🖼️ SAMPLES	💰 VALUE	⚡ SPEED
Image > Image prompts built to fail: depth, direction, counting, and text in the frame. 59 models		\$0.007 🇲🇻 Recraft: Recraft V4...	2s 🇲🇻 Recraft: Recraft V4...
Video > Six seconds of video, held to one duration and resolution across every model. 27 models		\$0.17 🇨🇪 HeyGen: HeyGen ...	30s 🇨🇪 HeyGen: HeyGen ...
Speech > One sentence, every voice: did the words come back, and was direction performed? 35 models		\$0.000 🇲🇻 Mistral: Voxtral Mini...	1s 🇲🇻 ElevenLabs: Eleve...

</> Benchmarks API

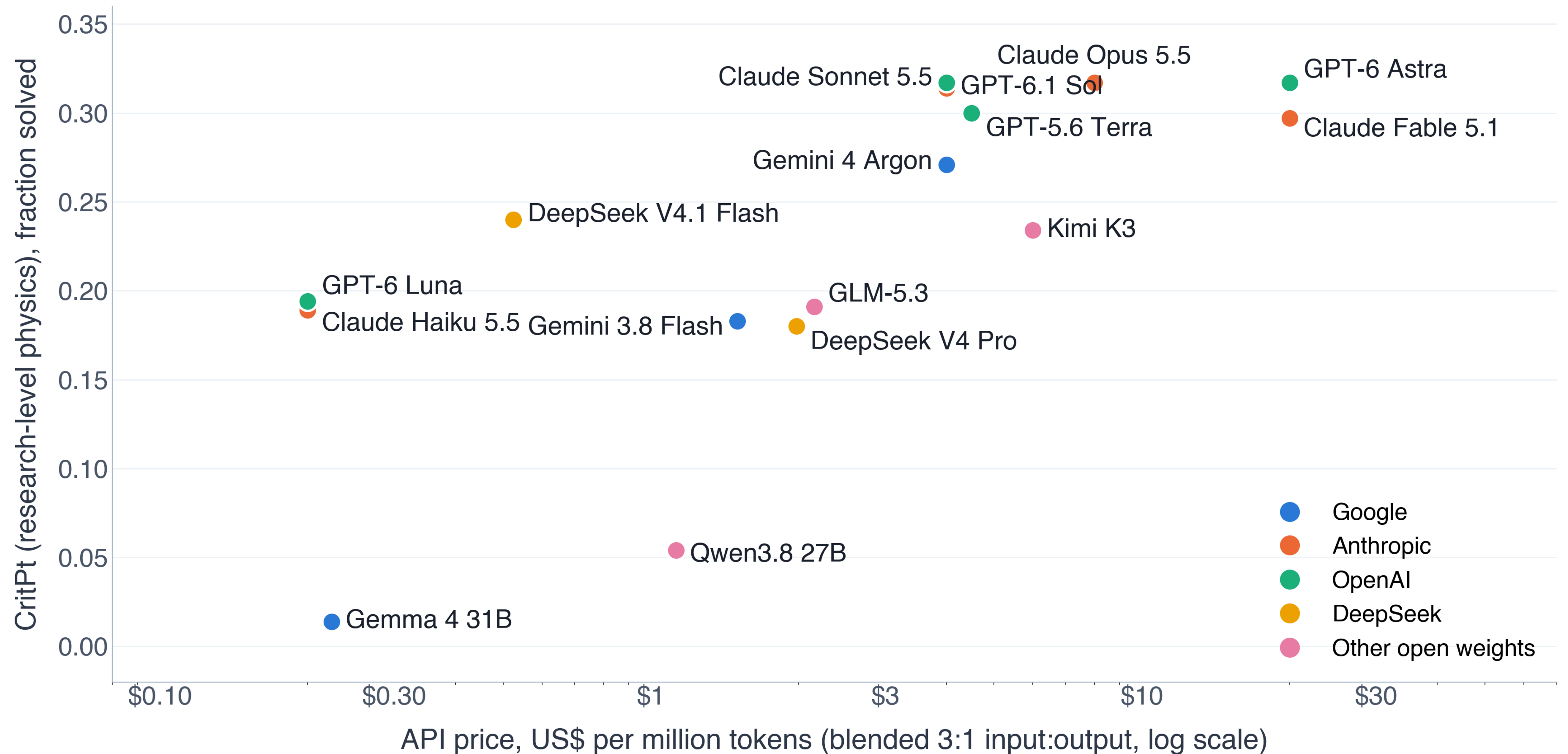
Fetch benchmark results and metadata via the API

Same models, different axis: factual knowledge & hallucination

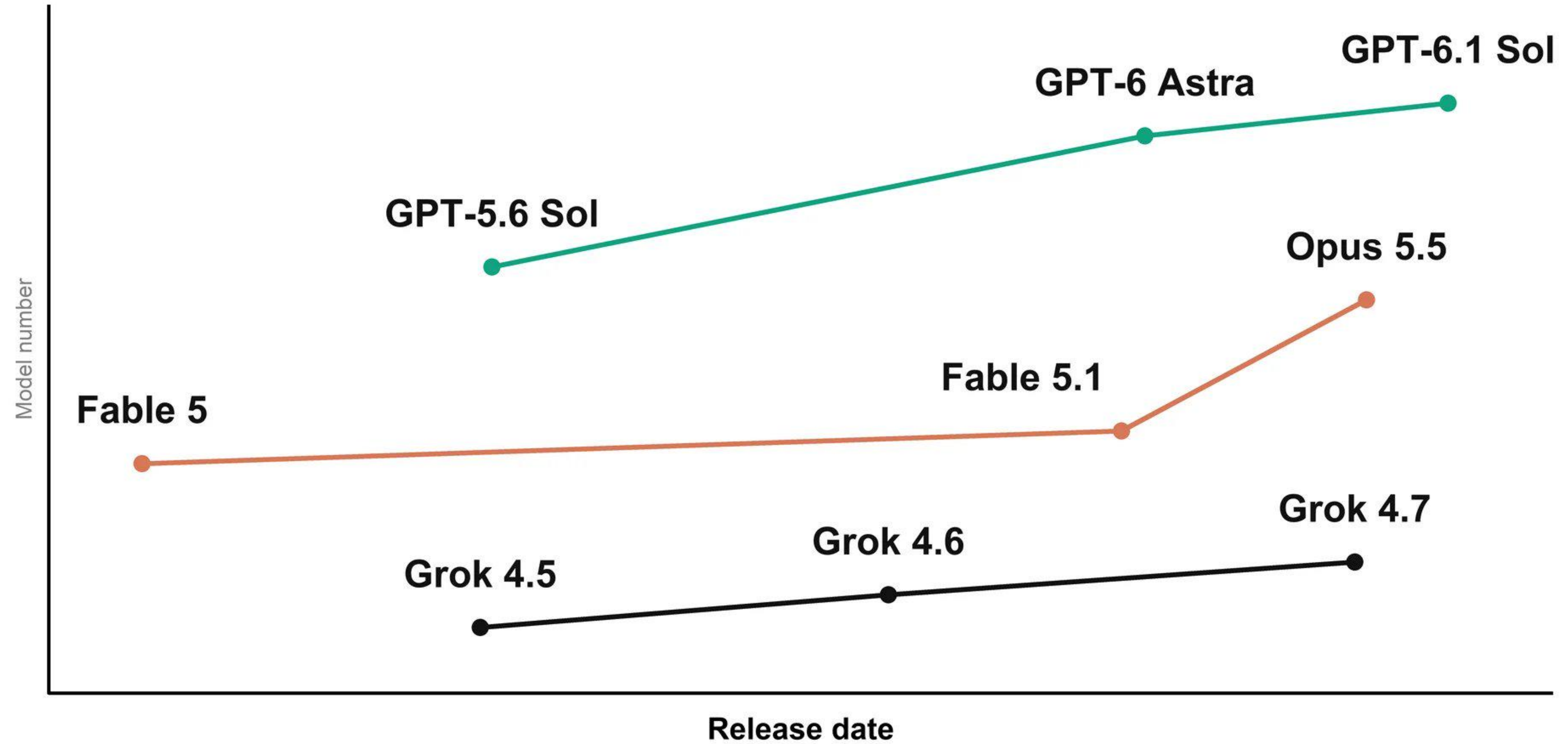


Data: Artificial Analysis, artificialanalysis.ai/leaderboards/models (22 Oct 2026); best effort setting per model; AA-listed API prices. Redrawn from AA's numbers.

Same models, different axis: research-level physics (CritPt)



Data: Artificial Analysis, artificialanalysis.ai/leaderboards/models (22 Oct 2026); best effort setting per model; AA-listed API prices. Redrawn from AA's numbers.



Frontier Plans vs API Rates

One overnight experiment (Aug 2026): Fable 5 agents checking a paper
costed at Fable 5 API list rates; actually paid for by a Max 20x subscription

	Tokens	Cost at API rate
Output generated	1.16M	\$57.97
Cache writes (new input)	7.16M	\$112.32
Cache reads	83.56M	\$83.56
Total	~91.9M	~\$253.87

... and it used **about 10%** of the 20x plan's allowance.

Source: email thread "question" (T. Dupuy to I. Cheltsov, 23 Aug 2026), table produced by the agent's TOKENS.md accounting. A later recount gave ~95.2M tokens / ~\$257.81.

What is a Max 20x plan worth at API prices? (back-of-envelope)

This experiment at API list price ≈ \$254

It used about 10% of the weekly allowance ⇒ week ≈ \$2,540 of API-priced tokens

× 52/12 weeks per month ⇒ month ≈ \$11,000

What is a Max 20x plan worth at API prices? (back-of-envelope)

This experiment at API list price ≈ \$254

It used about 10% of the weekly allowance ⇒ week ≈ \$2,540 of API-priced tokens

× 52/12 weeks per month ⇒ month ≈ \$11,000

Max 20x subscription \$200 / month

What is a Max 20x plan worth at API prices? (back-of-envelope)

This experiment at API list price ≈ \$254

It used about 10% of the weekly allowance $\Rightarrow$ week $\approx$ \$2,540 of API-priced tokens

$\times$ 52/12 weeks per month $\Rightarrow$ month $\approx$ \$11,000

Max 20x subscription \$200 / month

$\$11,000 / \$200 \approx 55\times \rightarrow \text{“}\sim 50\times \text{ discount”}$

STANDARDS TO KNOW

Agentic AI Foundation (AAIF)

4 MIN READ

Agentic AI Foundation Announces Four New Projects: Agent Skills, Open Responses, OpenEnv, and OpenSandbox

THE LINUX FOUNDATION | 22 OCTOBER 2026

Agentic AI Foundation (Linux Foundation)

Open standards for agents · 10 hosted projects as of Oct 23, 2026

AGENTS.md

The file every agent reads
before it does anything

founding project, Dec 2025 · from OpenAI

joined Oct 22, 2026

Agent Skills

Functions for AIs

contributed by Anthropic

MCP

APIs for AIs

Model Context Protocol · founding project,
Dec 2025 · from Anthropic

The other seven

A2A (Agent2Agent) joined Aug 2026

goose

agentgateway

Agent Router

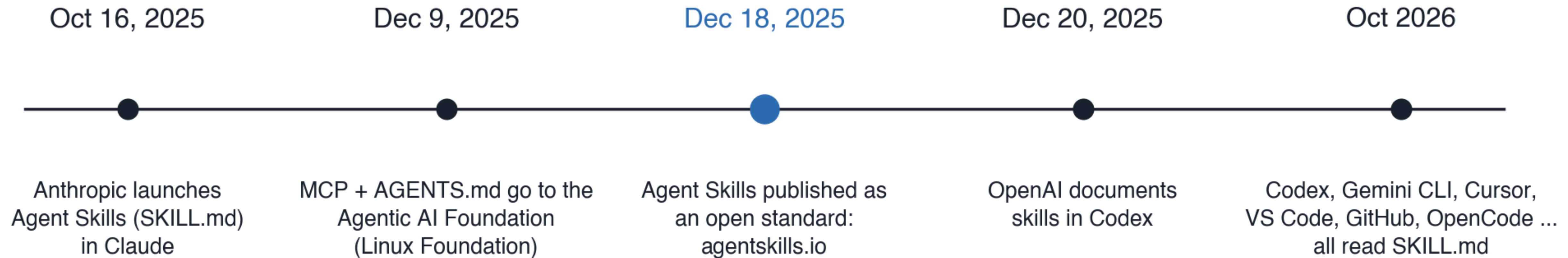
Open Responses Oct 2026

OpenEnv Oct 2026

OpenSandbox Oct 2026

SKILLS

How skills became a standard



A skill = a folder with SKILL.md (name + description + instructions), loaded on demand.

Skills: what they are, and how they became a standard

News

Research ▾

Policy ▾

Commitments ▾

ANTHROPIC

Company ▾

Try Claude ▾

Q

👁

☰

01

The anatomy of a skill

02

Developing and evaluating skills

03

The future of Skills

Dec 18, 2025

Update

We've published Agent Skills as an open standard for cross-platform portability.

View more →

As model capabilities improve, we can now build general-purpose agents that interact with full-fledged computing environments. [Claude Code](#), for example, can accomplish complex tasks across domains using local code execution and filesystems. But as these agents become more powerful, we need more composable, scalable, and portable ways to equip them with domain-specific expertise.

This led us to create **[Agent Skills](#)**: organized folders of instructions, scripts, and resources that agents can discover and load dynamically to perform better at specific tasks. Skills extend Claude's capabilities by packaging your expertise into composable resources for Claude, transforming general-purpose agents into specialized agents that fit your needs.

Building a skill for an agent is like putting together an onboarding guide for a new hire. Instead of building fragmented, custom-designed agents for each use case, anyone can now specialize their agents with composable capabilities by capturing and sharing their procedural knowledge. In this article, we explain what Skills are, show how they work, and share best practices for building your own.

Published

Oct 16, 2025

Team

Engineering

Topics

Agent Skills

Agents

Context engineering

[See more...](#)

Skills: what they are, and how they became a standard

Agent Skills now has an official [Discord server](#). See the [announcement](#) for details.

Agent Skills

Search...

K

Ask Assistant I

agentskills/agentskills



Overview

Specification

Client Showcase

For skill creators

Quickstart

Best practices

Optimizing descriptions

Evaluating skills

Using scripts

For client implementors

Adding skills support

Agent Skills Overview

Copy page



A standardized way to give AI agents new capabilities and expertise.

What are Agent Skills?

Agent Skills are a lightweight, open format for extending AI agent capabilities with specialized knowledge and workflows.

At its core, a skill is a folder containing a `SKILL.md` file. This file includes metadata (`name` and `description` , at minimum) and instructions that tell an agent how to perform a specific task. Skills can also bundle scripts, reference materials, templates, and other resources.

```
my-skill/
├─ SKILL.md          # Required: metadata + instructions
├─ scripts/          # Optional: executable code
├─ references/        # Optional: documentation
├─ assets/            # Optional: templates, resources
└─ ...               # Any additional files or directories
```



Why Agent Skills?

Agents are increasingly capable, but often don't have the context they need to do real work reliably. Skills solve this by packaging procedural knowledge and company-, team-, and user-specific context into portable, version-controlled folders that agents load on demand. This gives agents:

- **Domain expertise:** Capture specialized knowledge — from legal review processes to data analysis pipelines to presentation formatting — as reusable instructions and resources.
- **Repeatable workflows:** Turn multi-step tasks into consistent, auditable procedures.
- **Cross-product reuse:** Build a skill once and use it across any skills-compatible agent.

On this page

What are Agent Skills?

Why Agent Skills?

How do Agent Skills work?

Where can I use Agent Skills?

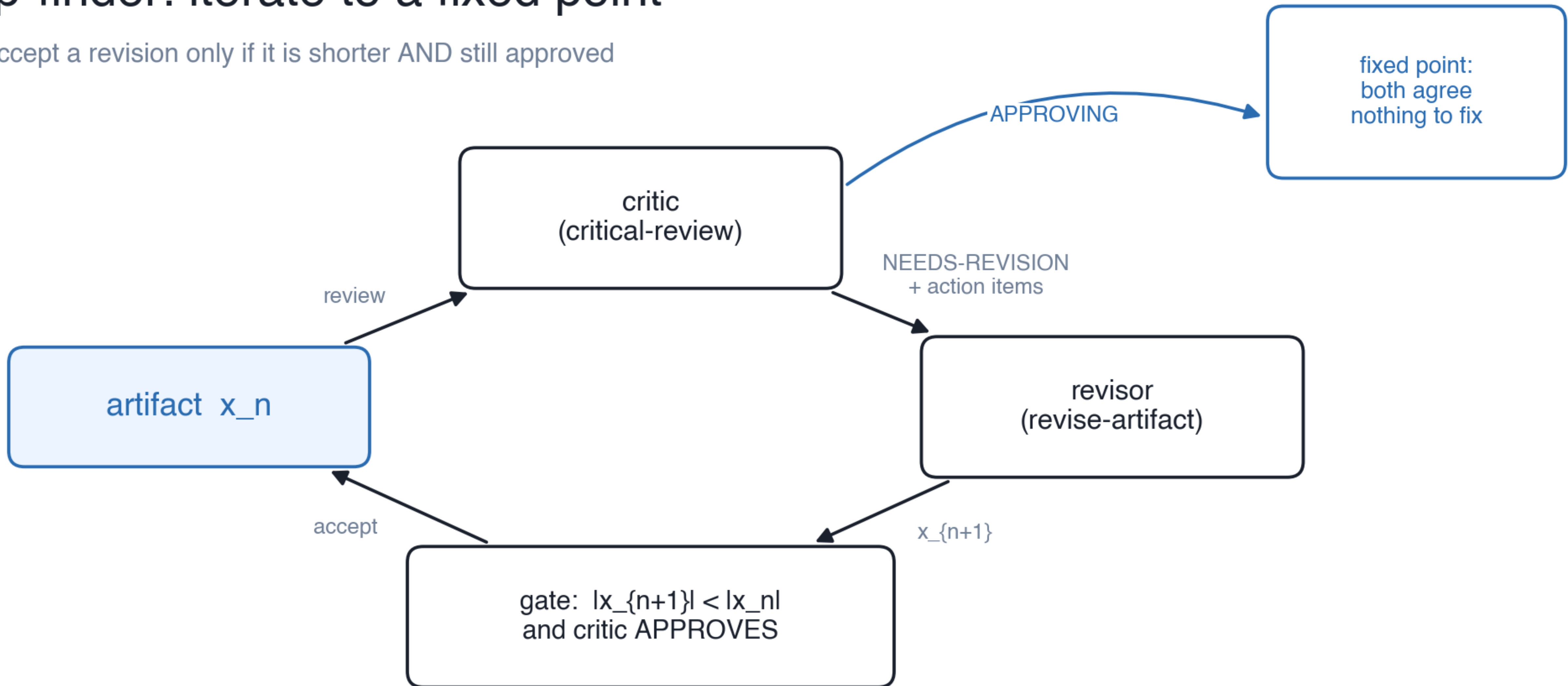
Open development

Get started with Agent Skills

Fixed Point Finder

fp-finder: iterate to a fixed point

Accept a revision only if it is shorter AND still approved



Why it terminates: length is a non-negative integer that strictly decreases (plus a hard cap, $N_{\max} = 8$). The APPROVING floor stops 'delete everything'.

SUPERPOWERS

obra / superpowers

<> Code

Issues 185

 Pull requests 159

 Actions

 Projects

 Security and quality

 Insights

 **superpowers** Public

 Sponsor

 Notifications

 Fork 27.1k

 Star 304k

 main ▾

 154 Branches

 38 Tags

 Go to file

<> Code ▾

 21 people	Release v7.0.1: brainstorming retune, skills that better survive comp...		2ce8aa1 · yesterday	 686 Commits
 .agents/plugins	Add Codex marketplace manifest			4 months ago
 .antigravity-plugin	Release v7.0.1: brainstorming retune, skills that better surv...			yesterday
 .claude-plugin	Release v7.0.1: brainstorming retune, skills that better surv...			yesterday
 .codex-plugin	Release v7.0.1: brainstorming retune, skills that better surv...			yesterday
 .cursor-plugin	Release v7.0.1: brainstorming retune, skills that better surv...			yesterday
 .devin-plugin	Release v7.0.1: brainstorming retune, skills that better surv...			yesterday
 .github	Templates: require a human-written summary and the filer'...			2 days ago
 .hermes-plugin	Release v7.0.1: brainstorming retune, skills that better surv...			yesterday
 .kimi-plugin	Release v7.0.1: brainstorming retune, skills that better surv...			yesterday
 .muse-plugin	Release v7.0.1: brainstorming retune, skills that better surv...			yesterday
 .opencode	Release v7.0.1: brainstorming retune, skills that better surv...			yesterday
 .pi/extensions	Align Pi mapping with action vocabulary			4 months ago
 assets	Release v7.0.0: rebuilt brainstorming, Antigravity Marketpl...			2 weeks ago

About

An agentic skills framework & software development methodology that works.

ai

brainstorming

coding

obra

sdic

skills

subagent-driven-development

superpowers

 Readme

 MIT license

 Code of conduct

 Activity

 303.6k stars

 1.1k watching

 27.1k forks

Report repository

Releases 16

 v7.0.1 Latest

yesterday

+ 15 releases

 gemini-extension.json	Release v7.0.1: brainstorming retune, skills that better surv...	yesterday
 index.js	Release v6.4.1: diagnosing-superpowers, Native plan exec...	last month
 package.json	Release v7.0.1: brainstorming retune, skills that better surv...	yesterday

 **README**  Code of conduct  MIT license 

Superpowers

Superpowers is a complete software development methodology for your coding agents, built on top of a set of composable skills and some initial instructions that make sure your agent uses them.

Table of Contents

- [How it works](#)
- [Commercial Services](#)
- [Getting Started](#)
 - [Claude Code](#)
 - [Antigravity](#)
 - [Codex App](#)
 - [Codex CLI](#)
 - [Cursor](#)
 - [Devin CLI](#)
 - [Factory Droid](#)
 - [Gemini CLI](#)
 - [GitHub Copilot CLI](#)
 - [Grok Build CLI](#)
 - [Kimi Code](#)
 - [OpenCode](#)
 - [Pi](#)
 - [Qwen Code](#)

GRILL-WITH-DOCS

mattpocock / skills

<> Code Issues 183 Pull requests 17 Discussions Actions Projects Security and quality Insights

Files

main 🔍

Go to file

> .agents

> .changeset

> .claude-plugin

> .github

> .out-of-scope

> docs

> scripts

✓ skills

- > deprecated

✓ engineering

- > ask-matt
- > code-review
- > codebase-design
- > diagnosing-bugs
- > domain-modeling

✓ grill-with-docs

- > agents

SKILL.md

skills / skills / engineering / grill-with-docs / SKILL.md

 **claude** Clarify multi-skill steps as multiple Skill tool calls, not one call ... 447ca70 · 2 months ago History

Preview Code Blame 7 lines (6 loc) · 247 Bytes

Raw Copy Download Edit

name	grill-with-docs
description	A relentless interview to sharpen a plan or design, which also creates docs (ADR's and glossary) as we go.
disable-model-invocation	true

Call the Skill tool twice, for "grilling" and "domain-modeling".

- Agent "Orchestration/repo slides" is waiting on its own background work · 14m 30s
- I'm still looking through your files for the CSAIL talk and AI notes to grab the dumb-zone chart and its citation—the skills section just came back. Next, I have a question about the ChatGPT "Turing machine" slides (1.6–1.7).
 - User answered Claude's questions:
 - └ · For the math slide, how precisely do you want to say "it's a Markov process"? The state is the entire token string, truncated to the context window, so the chain lives on $V^{\{\leq N\}}$ and not on V . Do you want that subtlety on the slide? → State = whole context (Recommended)
 - └ Message queued for delivery to a680f9a1daed972ab at its next tool round. If it finishes first, it is resumed with the message and notifies again, so wait for that notification before resending.
 - └ Allowed by auto mode classifier
- Write(CONTEXT.md)
 - └ Wrote 13 lines to CONTEXT.md

```
1 # Glossary — "How to use AI in Mathematical Research Effectively"
2
3 **Token** — an element of the model's vocabulary  $V$ ; text is cut into tokens before the model sees it.
4
5 **Context** — the string of tokens  $x \in V^{\{\leq N\}}$  the model conditions on;  $N$  is the context window.
6
7 **Language model** — a map  $M: V^{\{\leq N\}} \rightarrow \Delta(V)$  assigning a probability  $M(x)(y) = P(xy \mid x)$  to each next token  $y$ .
8
9 **Generation** — the Markov chain on  $V^{\{\leq N\}}$  whose state is the whole context and whose transition is  $x \mapsto xy$  with probability  $M(x)(y)$ .
... +3 lines (ctrl+o to expand)
```
- I've started [CONTEXT.md](#) with the terms settled so far. Next question, from what the skills agent found:

```
> Use the link that I gave you and /read-chatgpt
> Message from @general-purpose (ctrl+o to expand)
[> Message from @general-purpose (ctrl+o to expand)
```

▣ pcf paper

- For the AI-statement slide (3.9), the pcf paper is an unpublished draft in a private repo, and the rendered pages show `\taylor{...}` TODO notes. Your coauthors include Rochester faculty who may be in the room. What should the slide show?

1. Clean summary graphic (Recommended)

A rendered summary of what the statement reports (models, harness, what each was used for, verification, cost). No draft pages on screen.

2. Actual paper pages

Show the rendered pages 3–4 of the AI section; I'll recompile with TODO notes hidden first.

3. Both

Summary graphic, then a click to the real (cleaned) pages.

> 4. We have permission for this. I want to talk about what I have been working on with Tom and Zhuoran. (By the way, I took a screenshot of this slide --- use it when we talk about "grill-with-docs". One slide, just this screenshot.📸)

5. Chat about this

TDUPU/AGENT-SKILLS-PUBLIC

tdupu / agent-skills-public

<> Code

Issues

Pull requests

Actions

Projects

Security and quality

Insights

 agent-skills-public Public

Notifications

Fork 0

Star 0

main ▾

1 Branch

0 Tags

Go to file

<> Code ▾

 taylor-claude-agent[bot]	Sync Tau Ceti contribution workflow from source 44f8f7a	1e1b9ba · 5 days ago	9 Commits
 mcp	Add Stacks and Semantic Scholar MCP servers; credit LM...		last week
 plugins	Sync Tau Ceti contribution workflow from source 44f8f7a		5 days ago
 skills	Sync Tau Ceti contribution workflow from source 44f8f7a		5 days ago
 superpowers	Publish math-research skill packages with README and cr...		last week
 .gitignore	Add Stacks and Semantic Scholar MCP servers; credit LM...		last week
 LICENSE	Add GPL-3.0 LICENSE and BibTeX citation to README		last week
 README.md	Add GPL-3.0 LICENSE and BibTeX citation to README		last week

 README

 GPL-3.0 license

agent-skills-public

Agent skills for research mathematicians: writing papers in LaTeX, checking mathematics, formalizing in Lean, querying the LMFDB, teaching, writing NSF grants, documenting and testing math software, and running AI-assisted research with an honest provenance trail.

About

Some skills that are useful to mathematicians.

-  Readme
-  GPL-3.0 license
-  Activity
-  0 stars
-  0 watching
-  0 forks

Report repository

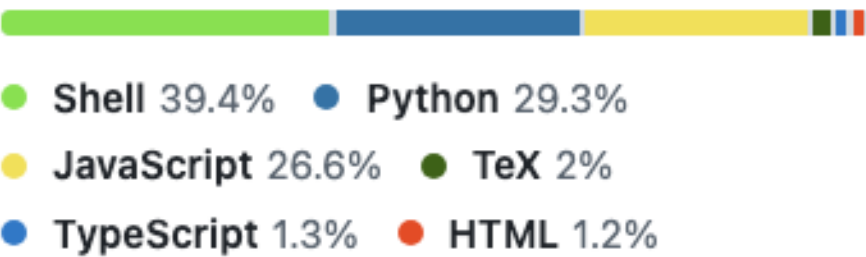
Releases

No releases published

Contributors 1

 taylor-claude-agent[bot]

Languages



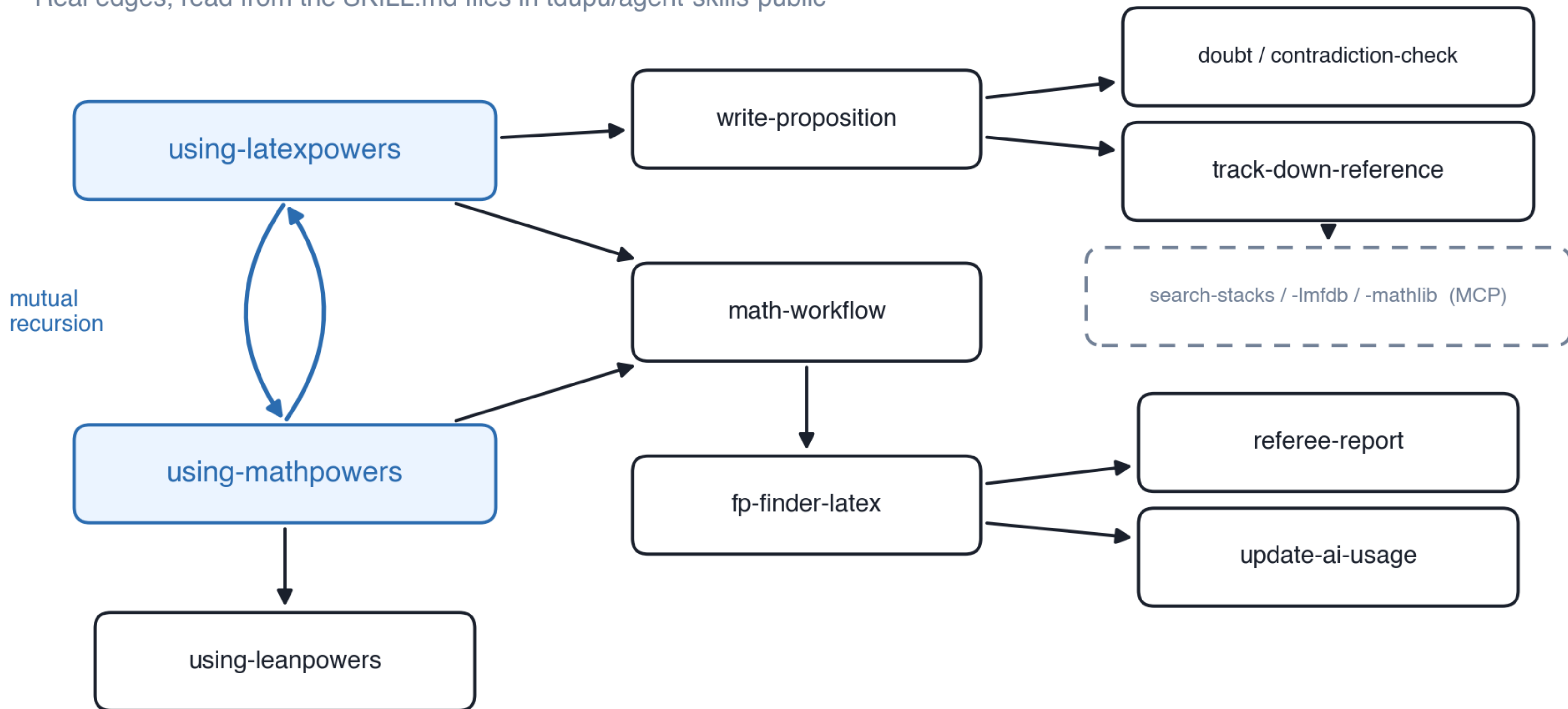
My skills (public)

github.com/tdupu/agent-skills-public · 217 skills, GPL-3.0

Package	What it does	Entry skill
latexpowers	Write and check manuscripts; built-in AI statement	using-latexpowers
mathpowers	Prove, refute, search arXiv / Stacks / LMFDB / Mathlib	using-mathpowers
leanpowers	Formalize in Lean 4 + Mathlib with fidelity checks	using-leanpowers
docpowers	Document repos and math software	using-docpowers
testpowers	Choose and write tests (Sage, M2, Magma, Lean, LaTeX)	using-testpowers
grantpowers	Assemble an NSF proposal package	using-grantpowers
pdpowers	Write and panel-review the NSF Project Description	using-pdpowers
teachingpowers	Syllabus, course page, notes, assessments	using-teachingpowers
taucetipowers	Contribute roadmaps and PRs to Tau Ceti (Lean)	using-taucetipowers
frontier-dump	Ask the best available model; save a dated research dump	frontier-dump
fp-finder-latex	Review/revise a manuscript to a fixed point	fp-finder-latex
critical-review	Adversarial review of any artifact	critical-review
adjust-behavior	Repair the instructions behind bad agent behavior	adjust-behavior

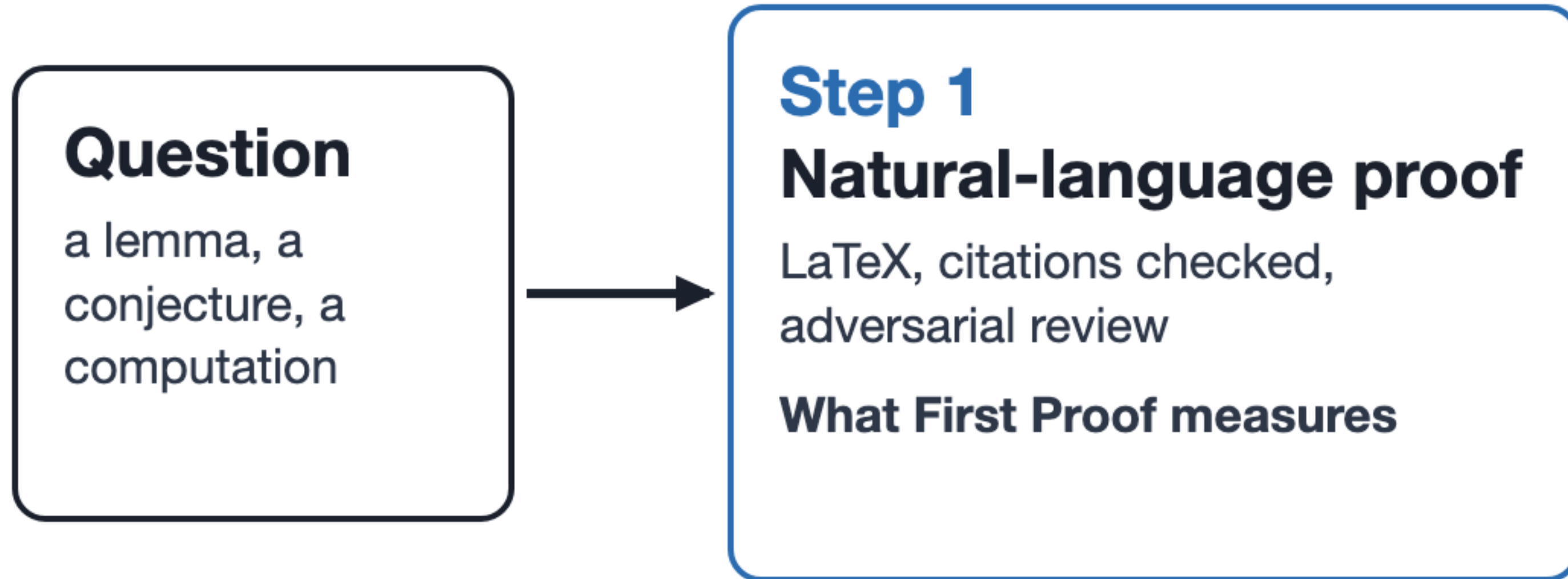
Skills call skills

Real edges, read from the SKILL.md files in tdupu/agent-skills-public

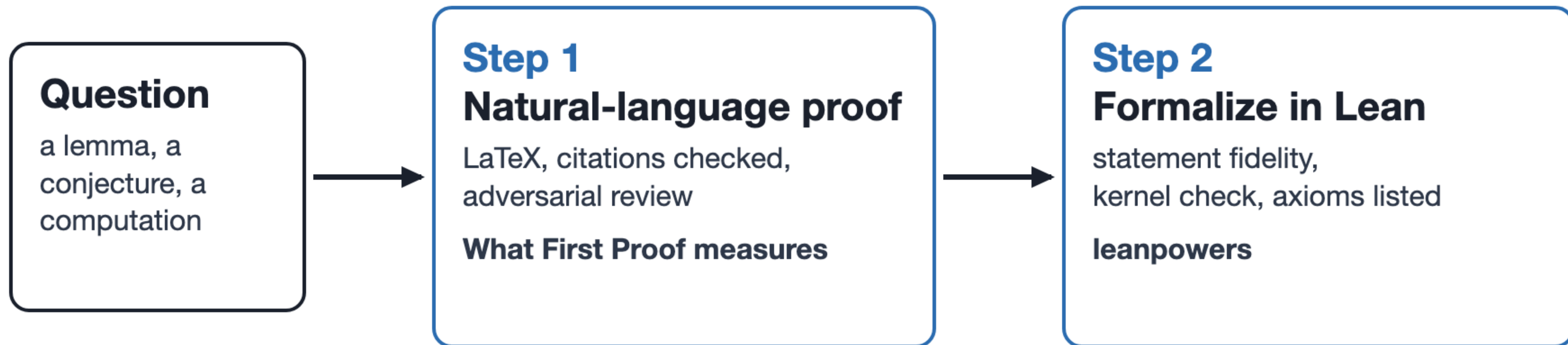


LEANPOWERS

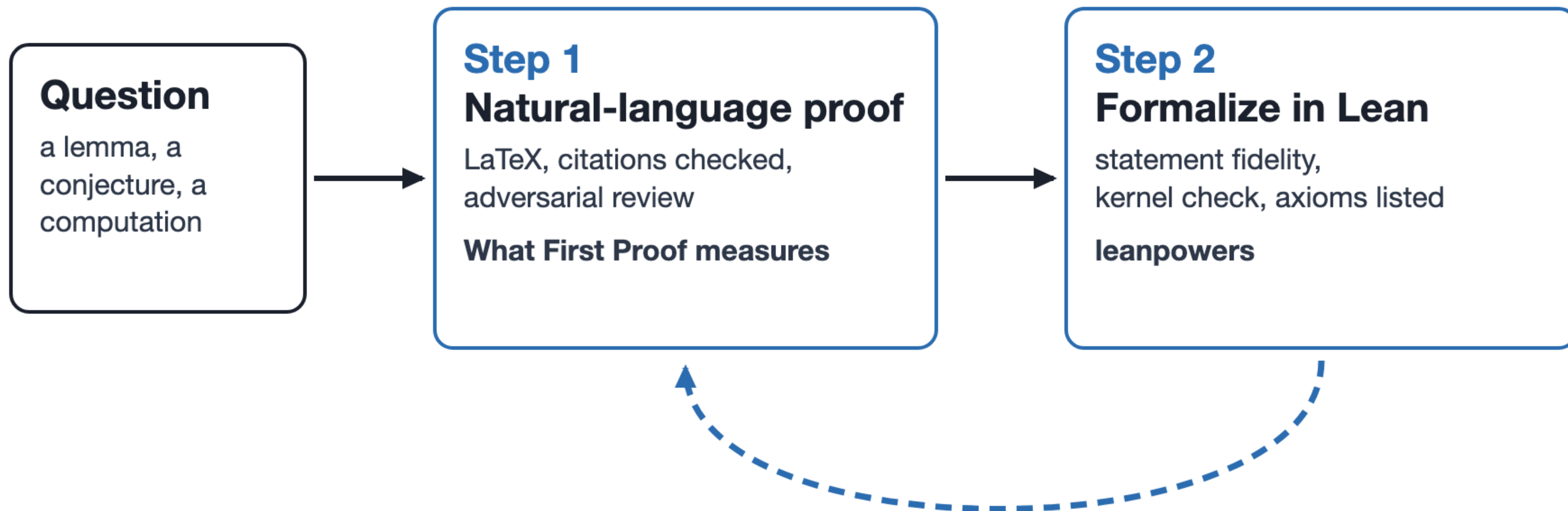
Automating proofs: the pipeline



Automating proofs: the pipeline



Automating proofs: the pipeline



formalization exposes a gap → fix the proof, repeat

First Proof (1stproof.org)

Batch	When	What happened
1	Feb 5–13, 2026	10 unpublished lemmas from 11 mathematicians' own research; answers encrypted for a week. DeepMind's Aletheia self-reports 6/10 (majority expert view).
2	results Jun 10, 2026	Formal benchmark, blind human referees. Best systems roughly 6–7 of 10 (press reports).
3	results Oct 14, 2026	5 systems, 24 h each, no human input, 30 referees at CMU's ICARM. ProofCouncil 9/10 (\$1,888) · GPT-6 Astra alone 7/10 (\$36) · open-weight (Kimi K3) 0/10
Still weak		Citations: only 4/10 Astra solutions adequately cited (7/10 with ProofCouncil). Exposition weak.

AI USAGE STATEMENTS

What the AI statement reports (pcf paper)

Dupuy–Levy–Tucker–Wang, "PCF all the way down" (draft, Oct 2026), §"Software and AI assistance"

Methods

typesetting · agent management ·
record keeping · natural-language proofs ·
adversarial reviews · computations · the statement itself

Who found what

each finding → model, harness,
reasoning setting (or "not recorded")

Literature + provenance

was the AI's idea already known?
(it was: SGA 1, Exp. X; Grothendieck–Murre ...)

Negative results

counterexamples and refuted claims
stated as results

Formalization

Lean 4 + Mathlib, axioms used,
no sorry; what is/isn't formalized

Token accounting

appendix table: tokens and
API-equivalent cost per task

Models are origin metadata, never authors; the authors are responsible for every statement and proof.

TOOLS FOR FORMALIZATION IN LEAN

TOOLS FOR DATABASE QUERIES

Demo: LMFDDB

```
# Claude Code session, LMFDB MCP server (recorded 2026-10-23)
```

```
> tool call: mcp__lmfdb__run_sql
```

```
sql: "SELECT lmfdb_label FROM ec_curvedata WHERE rank=2  
      ORDER BY conductor LIMIT 1"
```

```
< {"columns": ["lmfdb_label"], "row_count": 1,  
    "rows": [{"lmfdb_label": "389.a1"}]}
```

```
> tool call: mcp__lmfdb__run_sql
```

```
sql: "SELECT lmfdb_label, ainvs, conductor, rank FROM ec_curvedata  
      WHERE lmfdb_label='389.a1'"
```

```
< {"columns": ["lmfdb_label", "ainvs", "conductor", "rank"], "row_count": 1,  
    "rows": [{"lmfdb_label": "389.a1", "ainvs": ["0", "1", "1", "-2", "0"],  
              "conductor": 389, "rank": 2}]}
```

```
# ainvs [0,1,1,-2,0] =>  $y^2 + y = x^3 + x^2 - 2x$ 
```

ORCHESTRATION

Demo

What is waiting for you in the morning

“I'm going to bed for the night. You are on your own. You have the power to adjudicate. Choose the easiest things to do and adjudicate based on policy.”

Taylor Dupuy to the Mayor, 10 Sep 2026, ~21:40 — recorded verbatim as decision bead gt-y650td

Limits the agent set itself

- only verdicts that **follow from adopted policy**
- nothing irreversible
- no bulk actions
- every verdict recorded with its rule

What is waiting for you in the morning

“I'm going to bed for the night. You are on your own. You have the power to adjudicate. Choose the easiest things to do and adjudicate based on policy.”

Taylor Dupuy to the Mayor, 10 Sep 2026, ~21:40 — recorded verbatim as decision bead gt-y650td

Limits the agent set itself

- only verdicts that **follow from adopted policy**
- nothing irreversible
- no bulk actions
- every verdict recorded with its rule

Done overnight

- **9 briefs decided** by derivation; 3 skipped (need judgment)
- finding filed: 7 of the 9 briefs recommended the **inverted** direction
- 2 pull requests merged with recorded authority
- differential Gröbner campaign: 3 public issues (#34–36), fed to the city; one already in progress

What is waiting for you in the morning

“I'm going to bed for the night. You are on your own. You have the power to adjudicate. Choose the easiest things to do and adjudicate based on policy.”

Taylor Dupuy to the Mayor, 10 Sep 2026, ~21:40 — recorded verbatim as decision bead gt-y650td

Limits the agent set itself

- only verdicts that **follow from adopted policy**
- nothing irreversible
- no bulk actions
- every verdict recorded with its rule

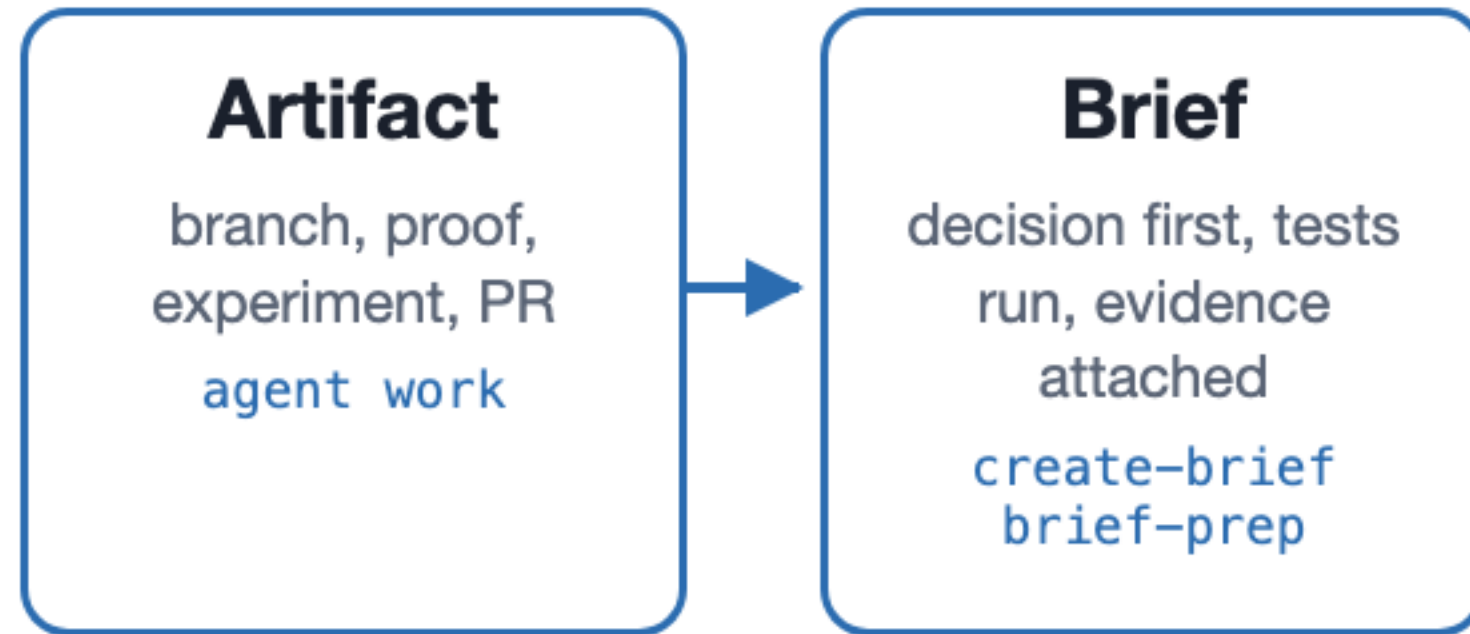
Done overnight

- **9 briefs decided** by derivation; 3 skipped (need judgment)
- finding filed: 7 of the 9 briefs recommended the **inverted** direction
- 2 pull requests merged with recorded authority
- differential Gröbner campaign: 3 public issues (#34–36), fed to the city; one already in progress

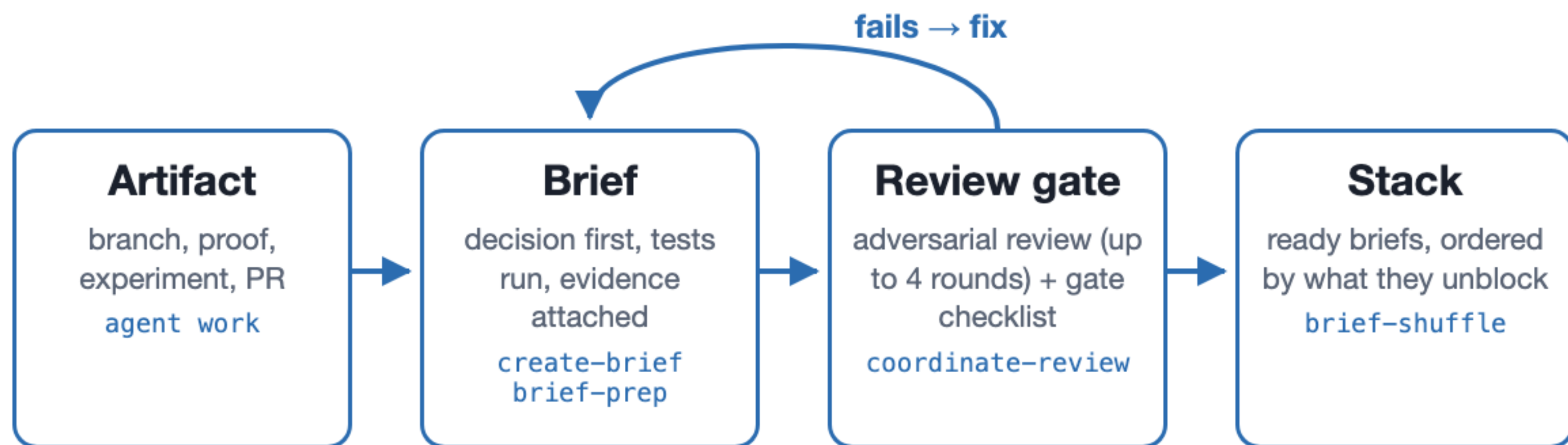
Waiting for you

- the 3 briefs that need judgment
- a database-conflict repair plan, held for you by an earlier ruling

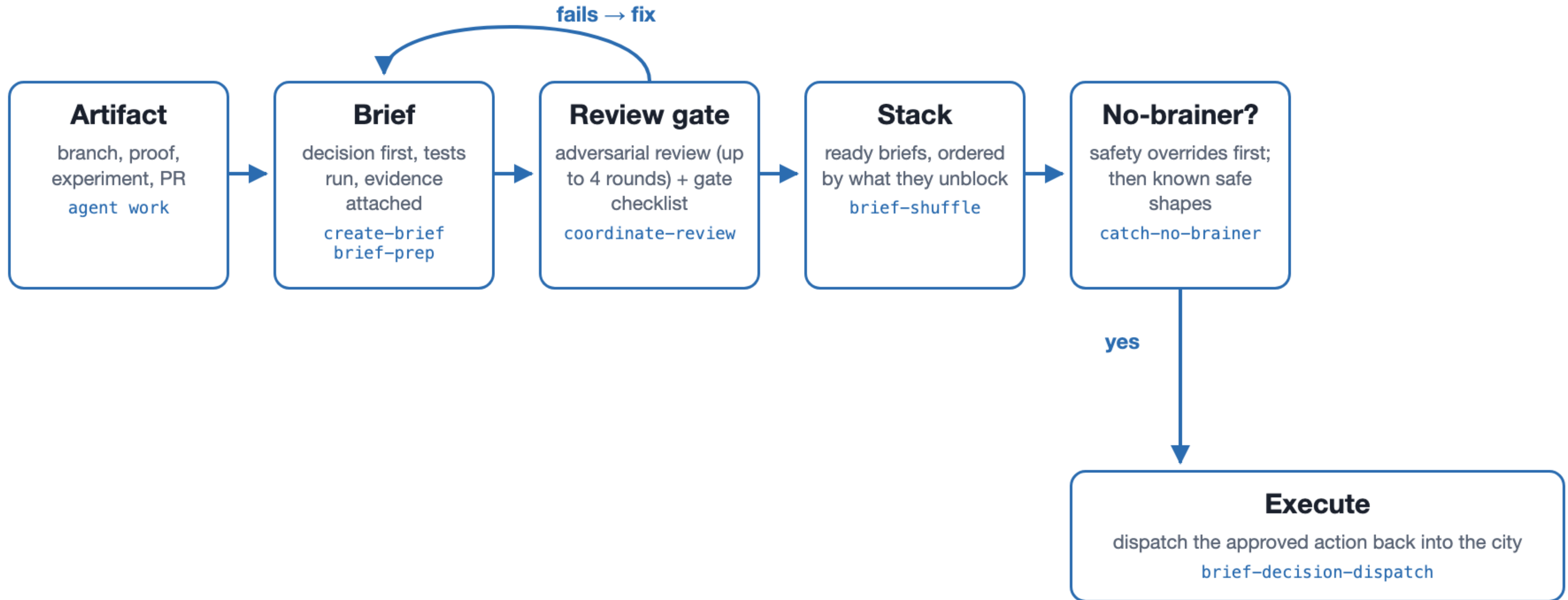
The brief system: only real decisions reach you



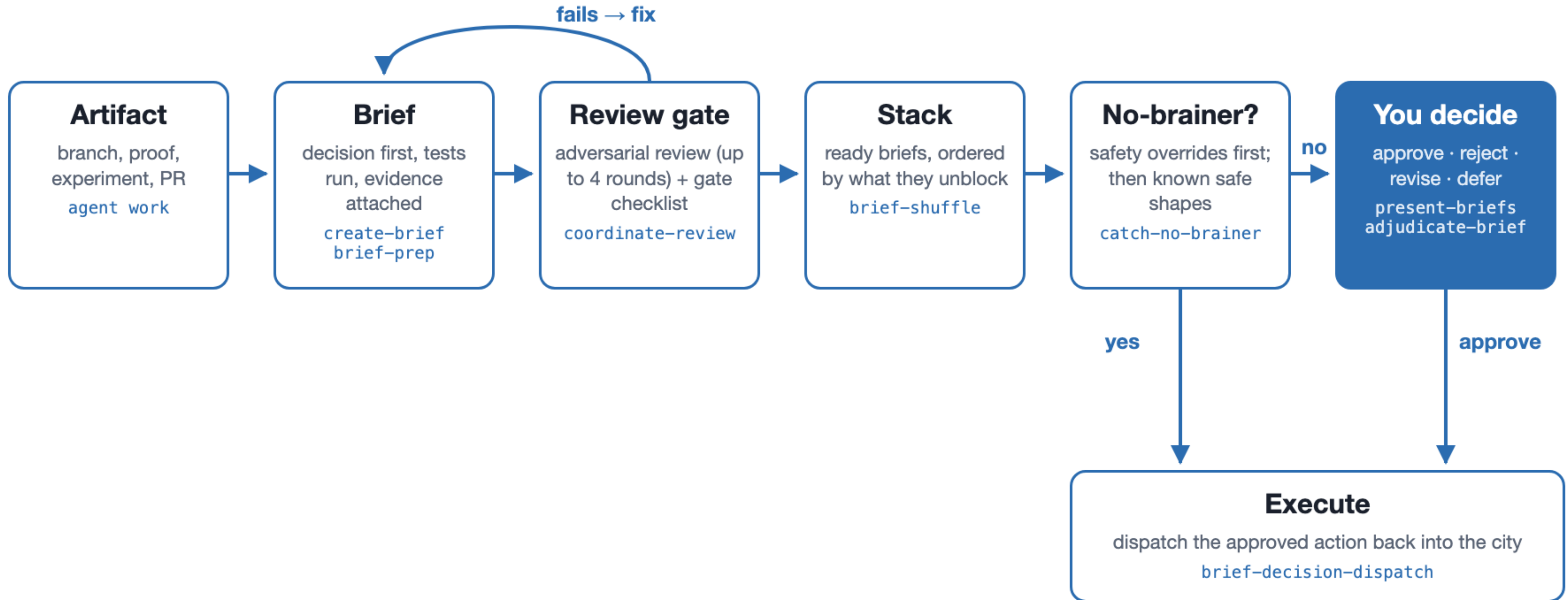
The brief system: only real decisions reach you



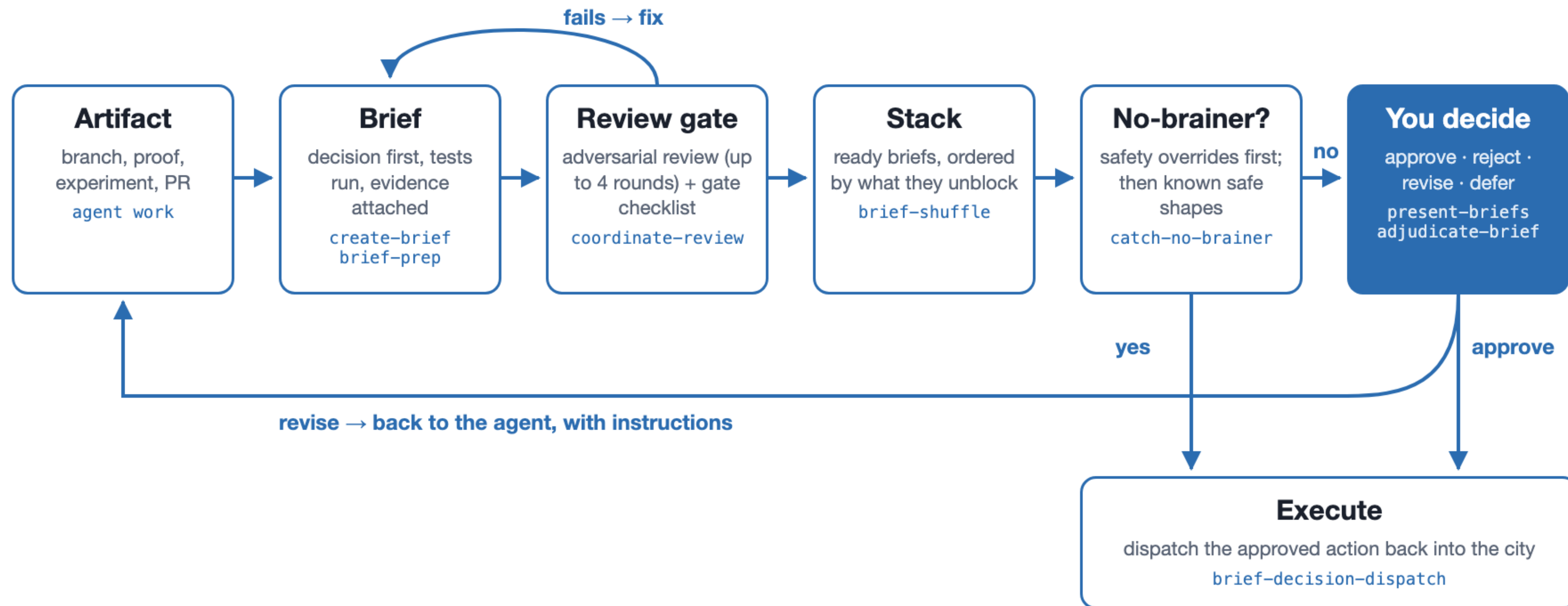
The brief system: only real decisions reach you



The brief system: only real decisions reach you



The brief system: only real decisions reach you



Repeated no-brainers of one shape become a new **gate**; gates become **policy**.

The End