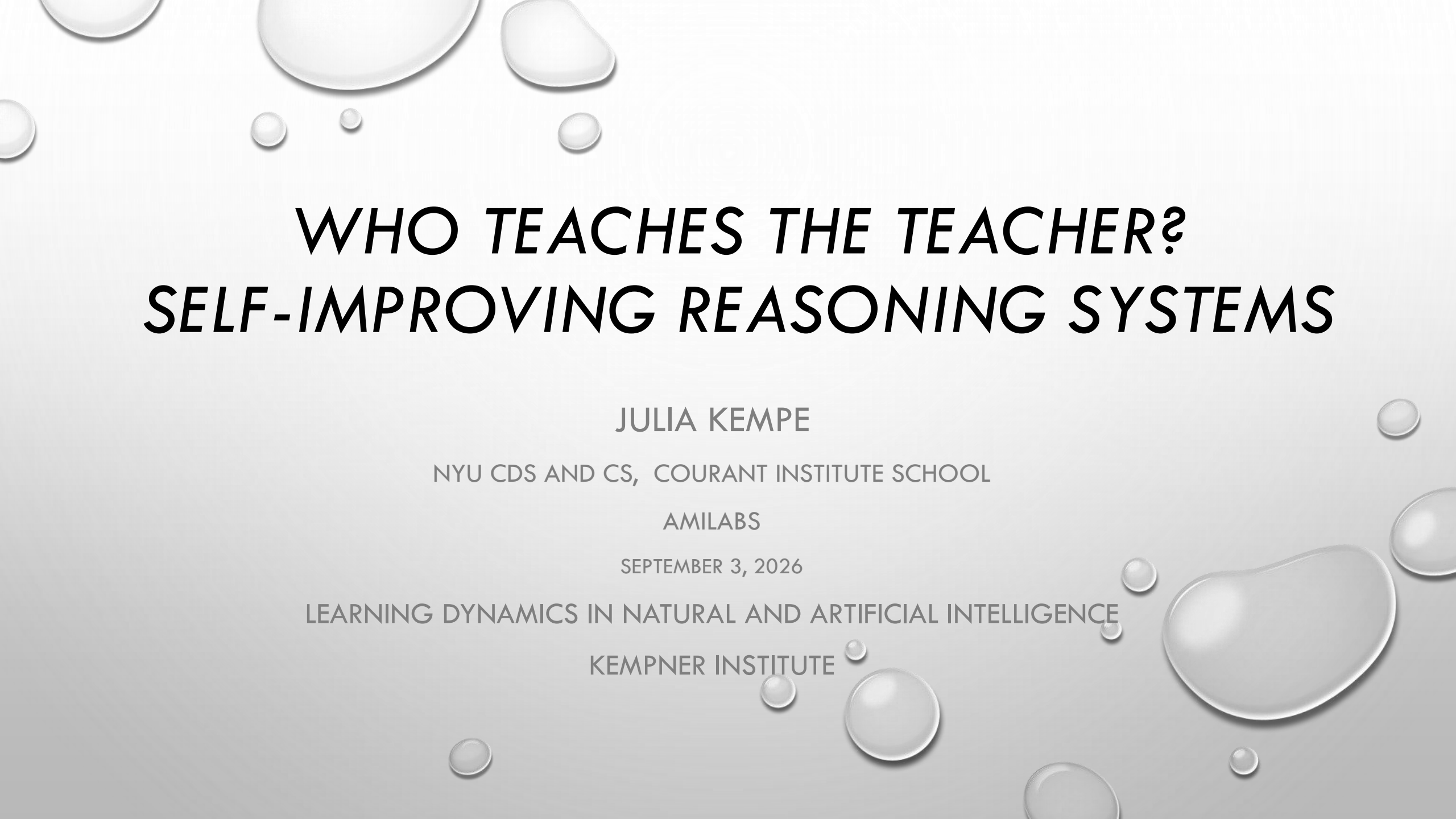


The background of the slide is a light gray gradient, decorated with numerous realistic water droplets of various sizes. Some droplets are at the top left, others are scattered along the right edge, and a few are at the bottom. The main title is centered in a large, bold, black, italicized sans-serif font.

WHO TEACHES THE TEACHER? SELF-IMPROVING REASONING SYSTEMS

JULIA KEMPE

NYU CDS AND CS, COURANT INSTITUTE SCHOOL

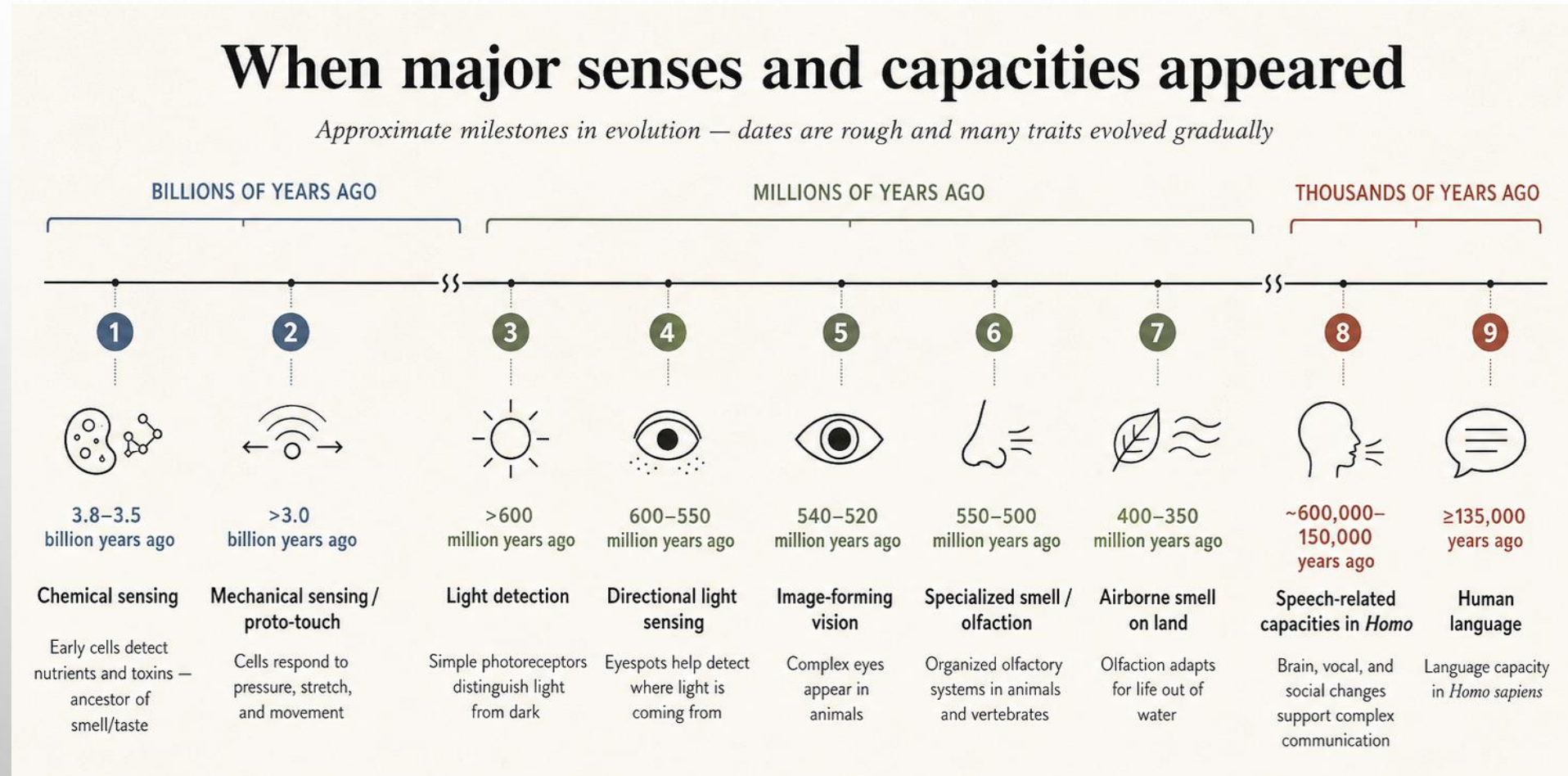
AMILABS

SEPTEMBER 3, 2026

LEARNING DYNAMICS IN NATURAL AND ARTIFICIAL INTELLIGENCE

KEMPNER INSTITUTE

INTELLIGENCE DEVELOPS THROUGH OUR SENSORS AND CAPABILITIES

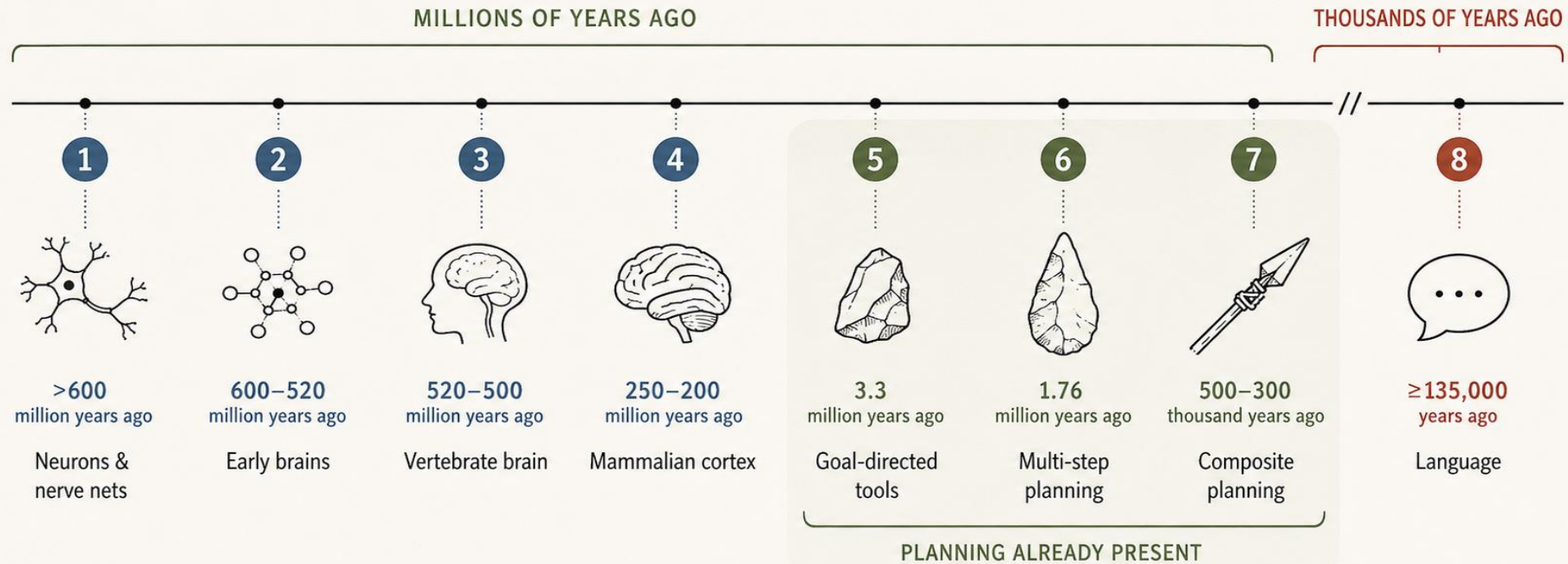


Language with its high level of abstraction came orders of magnitude **later** than other “inputs”.

INFORMATION PROCESSING SYSTEMS AND PLANNING

When information processing and planning appeared

Planning and reasoning-like behavior long predate language



Planning (reasoning) is visible millions of years before language.

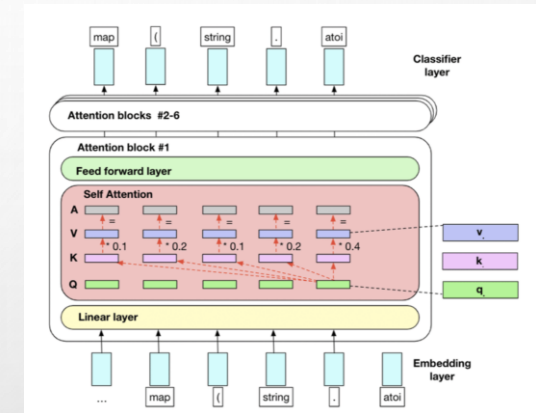
ENORMOUS RECENT PROGRESS IN AI THROUGH LANGUAGE

- Massive language corpora on the internet (training data)
- Huge models trained with simple objective (next token prediction)

GPT4 – unknown, estimated around **1.75 trillion** parameters

GPT5.6-Sol – speculated at **4 trillion**

Claude Fable – **10 trillion** parameters ?



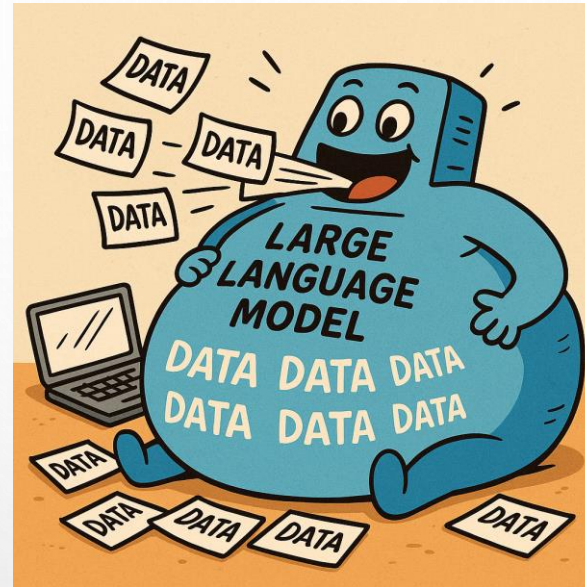
The human brain contains about **86 billion** neurons
– for reference.



prediction." They are given a large
nt sources like Wikipedia and news
wn into "tokens," which are

DATA, DATA, DATA!

GPT4 – unknown, but estimated around
1.75 trillion parameters
GPT5.6-Sol – speculated at **4 trillion**
Claude Fable – **10 trillion** parameters ?



GPT-3 training data^{[1]:9}

Dataset	# tokens	Proportion within training
Common Crawl	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

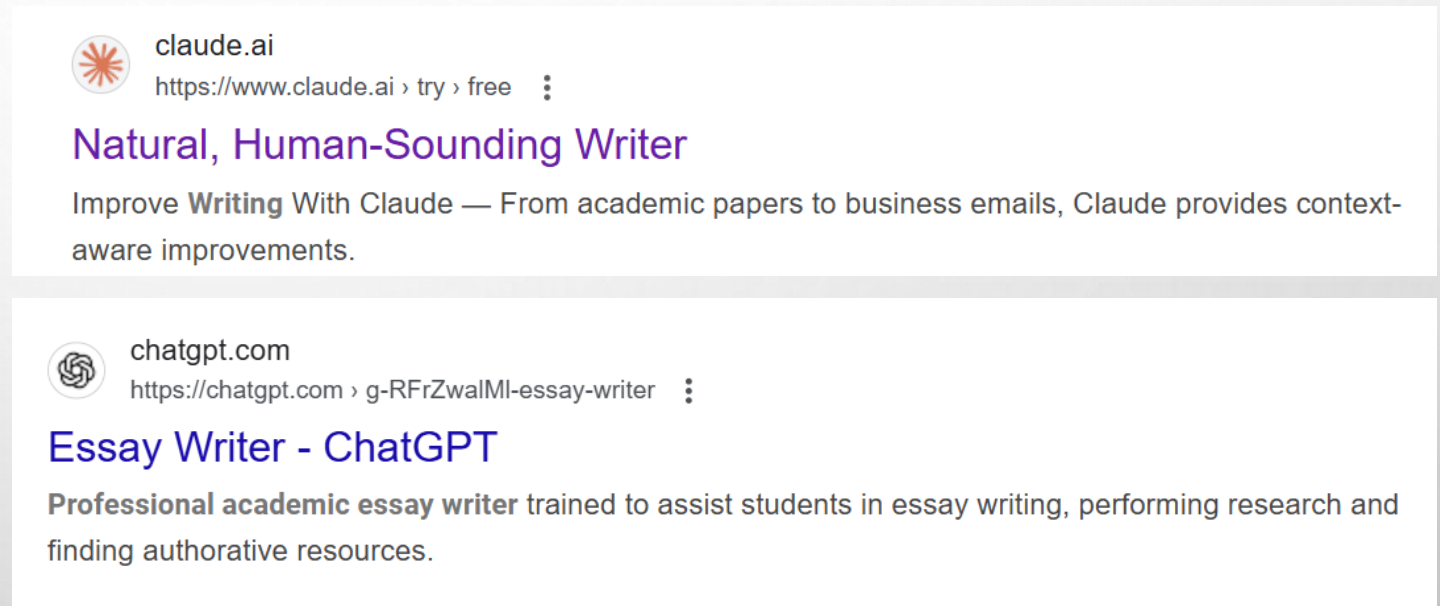
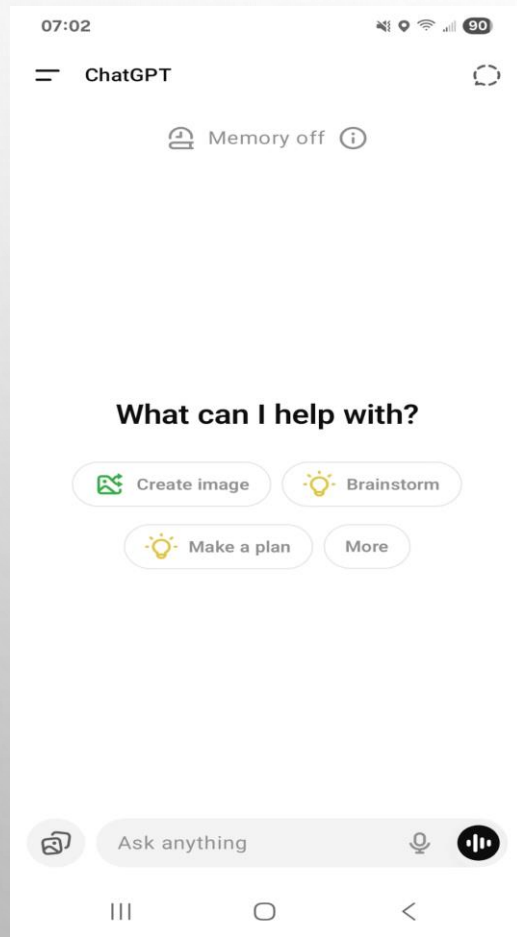
GPT4 was trained on **about 13 trillion tokens** ($\sim 10^{13}$).

GPT5-Sol: **70 trillion** tokens, Claude Fable **100 trillion** (10^{14}) (unconfirmed estimate).

- Massive language corpora on the internet (training data)
- Huge models trained with simple objective (next token prediction)

LLMS EVERYWHERE...

LLMs are now writing essays, chatting like a human, generating art, doing math, coding, and even designing proteins ...

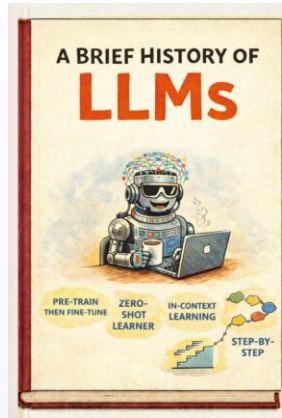


A BRIEF (PRE) HISTORY OF LLMS

Pre-2020: “Pre-train then fine-tune” (GPT-1, GPT-2)

Improving Language Understanding by Generative Pre-Training (Radford et al., 2018)

GPT-2 hinted at “zero-shot” abilities: model could perform tasks it wasn’t specifically trained on



2020: In-Context Learning (ICL)

Language Models are Few-Shot Learners (Brown et al., 2020)

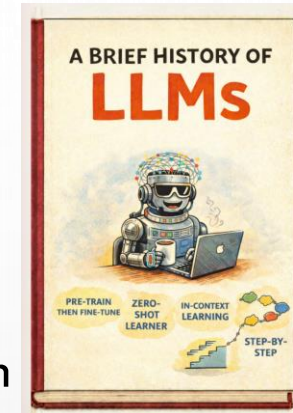
At 175B parameters GPT-3 could learn patterns/take instructions from a prompt

2022: Chain-of-Thought (CoT) “Let’s think step-by-step”

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)

For math/logic ICL wasn’t enough. Prompting the model to "show its work" or "think step-by-step" skyrocketed performance on reasoning tasks

A BRIEF HISTORY OF LLMS FOR REASONING



Pre-2020: “Pre-train then fine-tune” (GPT-1, GPT-2)

Improving Language Understanding by Generative Pre-Training (Radford et al., 2018)

GPT-2 hinted at “zero-shot” abilities: model could perform tasks it wasn’t specifically trained on

2020: In-Context Learning (ICL)

Language Models are Few-Shot Learners (Brown et al., 2020)

At 175B parameters GPT-3 could learn patterns/take instructions from a prompt

2022: Chain-of-Thought (CoT) “Let’s think step-by-step”

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)

For math/logic ICL wasn’t enough. Prompting the model to “show its work” or “think step-by-step” skyrocketed performance on reasoning tasks

2025: Reinforcement learning from verifiable rewards (RLVF)

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning (DeepSeek-AI, 2024/2025 / OpenAI o1 /Kimi)

Use RL & verifiable outcomes (math/code) to force the model to develop internal reasoning and self-correction

A BRIEF HISTORY OF LLMS FOR REASONING

2025: Reinforcement learning from verifiable rewards (RLVF)

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning (DeepSeek-AI, 2024/2025 / OpenAI o1 /Kimi)

Use RL & verifiable outcomes (math/code) to force the model to develop internal reasoning and self-correction

2026: Agents.

Training-free (in context) RL.

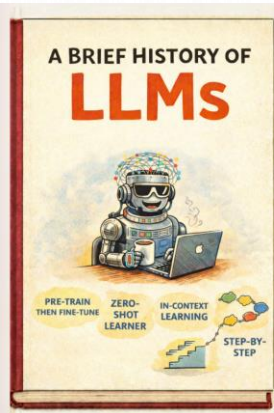
“In skills”-Learning.

“In harness”-RL.

Meta-harness learning.

Self-improvement.

...

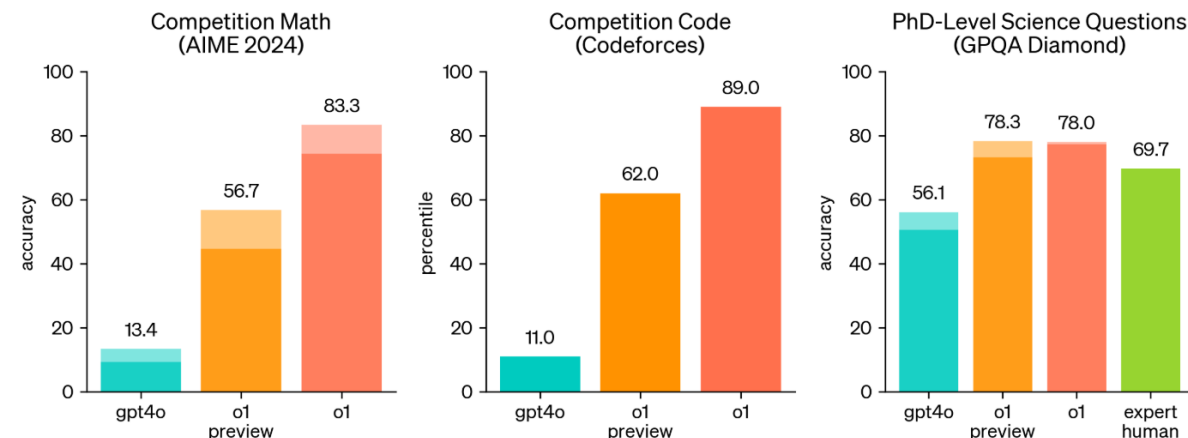


REINFORCEMENT LEARNING FOR LLM POST-TRAINING

Fantastically successful! The O1 (Sep 2024) and DeepSeek/Kimi (Jan 2025) “Revolution”

Introducing OpenAI o1

A new series of AI models designed to spend more time thinking before they respond.



Open Source / open method (GRPO)



DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

 KIMI K1.5:

SCALING REINFORCEMENT LEARNING WITH LLMs

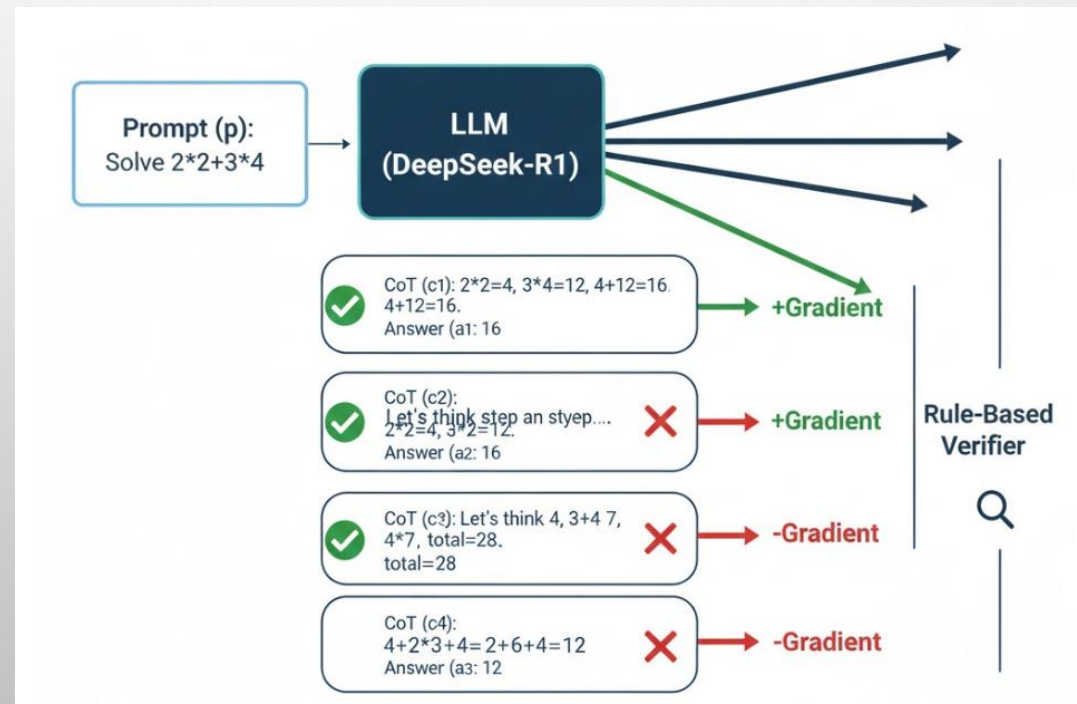
TECHNICAL REPORT OF KIMI K1.5

Kimi Team

REINFORCEMENT LEARNING FOR LLM POST-TRAINING

Group Relative Policy Optimization (GRPO) & “Test-time compute paradigm”

- The model generates a **group** of multiple CoT outputs (say, 16 different attempts) for the same prompt and compares them against each other. The outputs that are better than the group average are rewarded; the ones below average are penalized.
- For reasoning tasks the reward is rule-based (“verifiable”), no “critic” model or “human-in-the-loop” needed.



GRPO FOR RLVF

- LLMs use lots of test time tokens to reason and improve the performance.
- From train-time scaling to test-time scaling

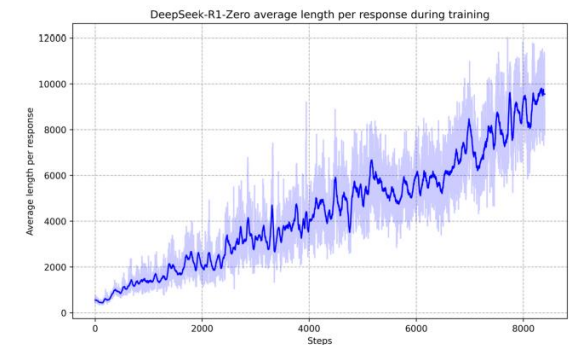
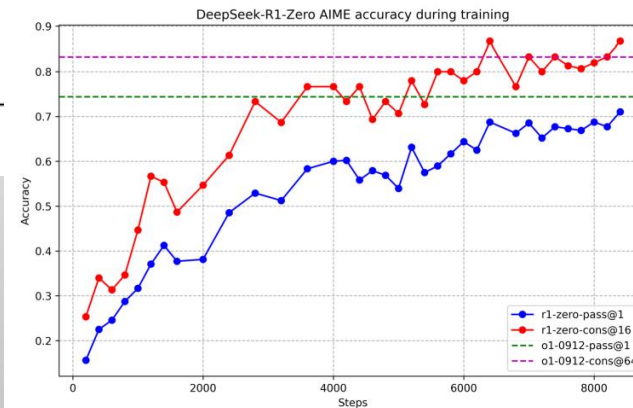
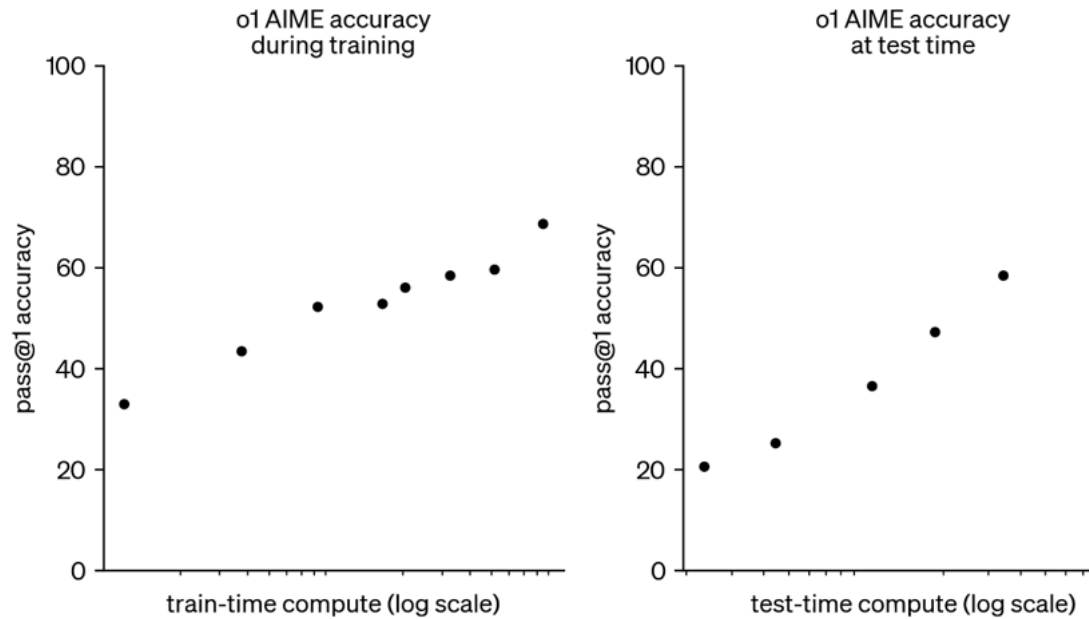


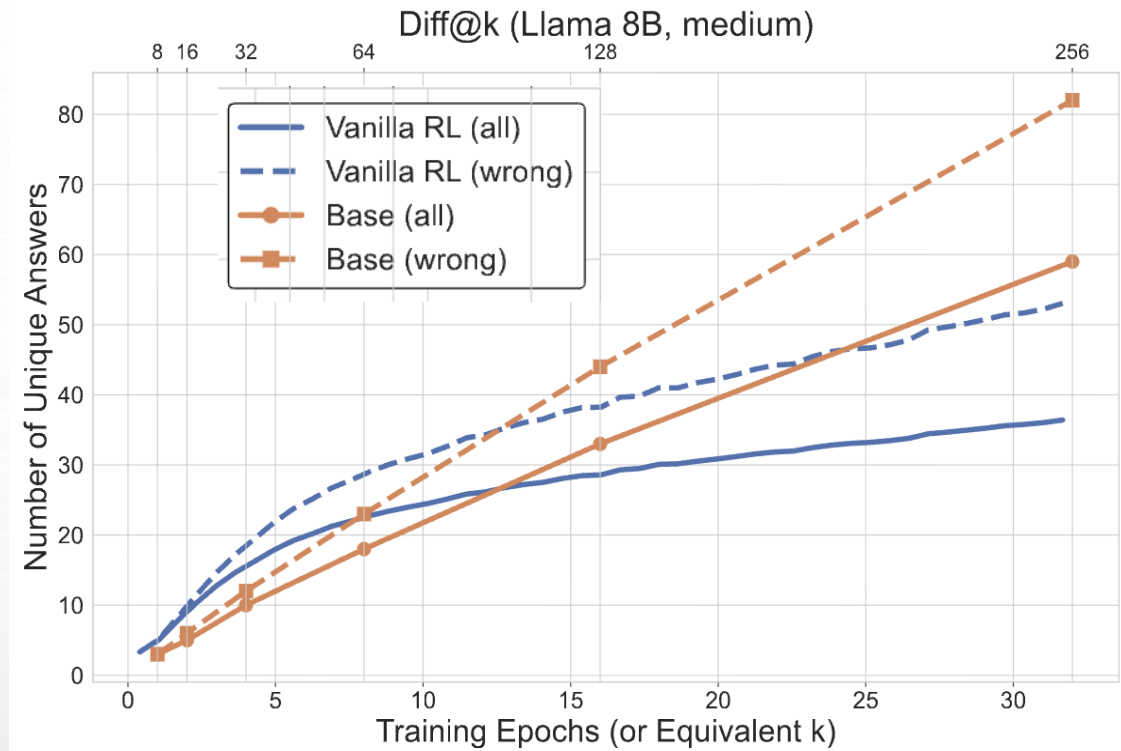
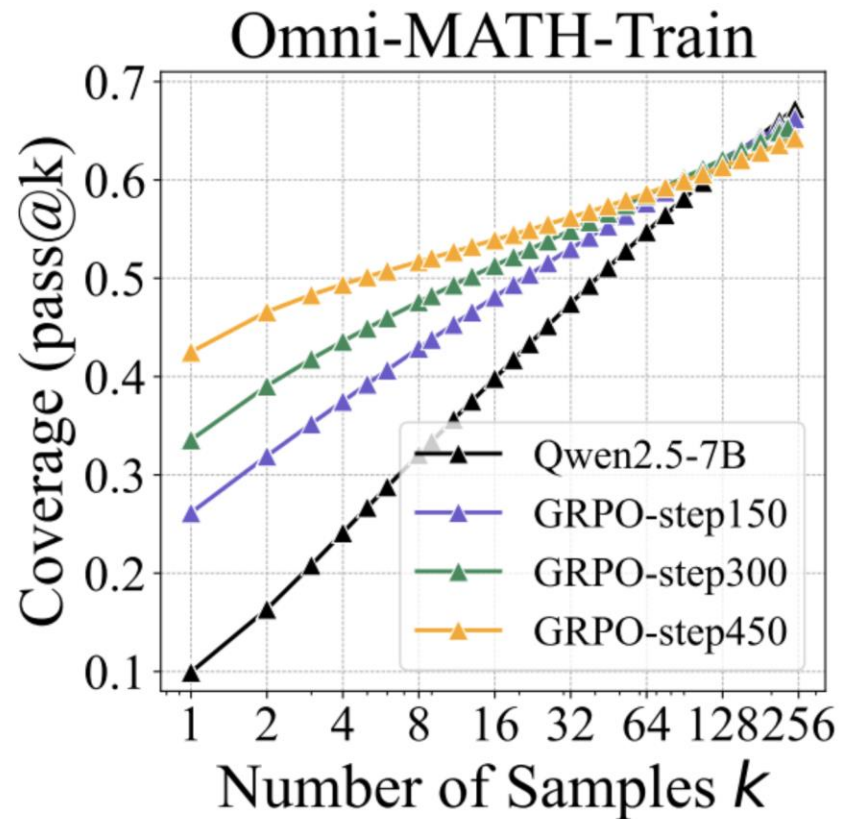
Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

LEARNING OR RETRIEVING?

“The intelligence already exists in the pre-trained model, it is all about extracting it.”

Sebastian Bubeck (November 2024)

LEARNING OR “SHARPENING”?



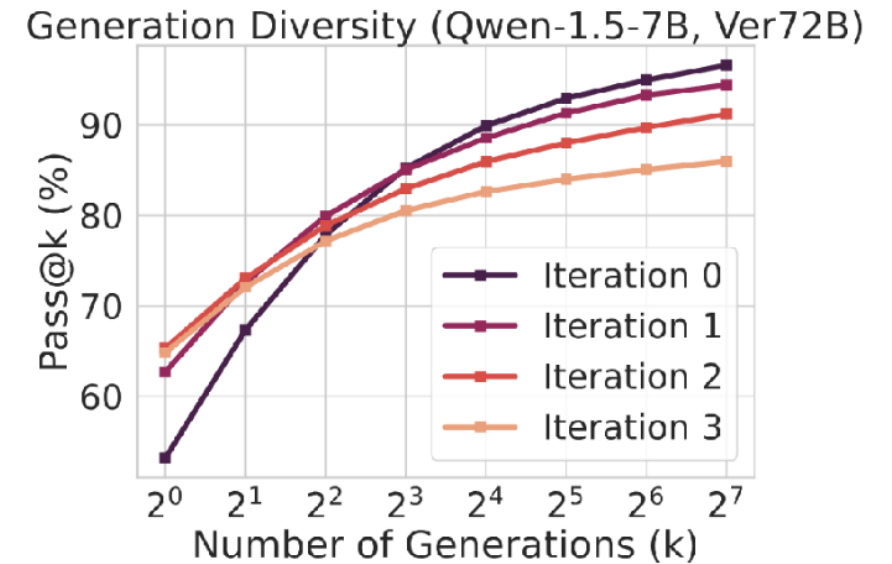
Outcome-based Exploration for LLM Reasoning

Yuda Song^{1,2}, Julia Kempe^{1,3}, Remi Munos¹

¹FAIR at Meta, ²CMU, ³NYU

DIVERSITY COLLAPSES DURING TRAINING

Diversity is measured as pass@k (roughly, is there a correct answer among k CoT rollouts?)
Often synonymous with exploration.



DIVERSITY COLLAPSES: EXPLOITATION VERSUS EXPLORATION

Diversity is measured as pass@k (roughly, is there a correct answer among k CoT rollouts?)
Often synonymous with exploration.

Can we add an “exploration bonus”?

Idea:

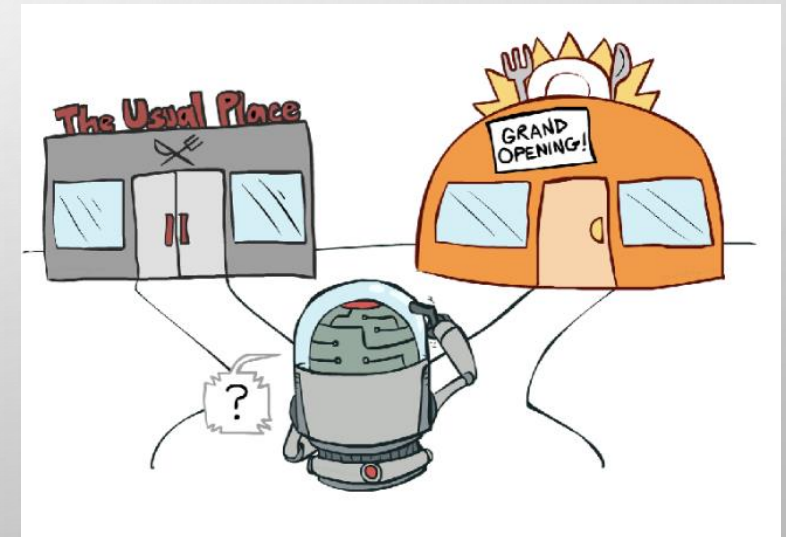
Encourage the policy to visit as many “outcomes” as possible
Add exploration bonus that is inversely proportional to
number of visits

$$\text{bonus}(x, a) = \min \left\{ 1, \sqrt{\frac{1}{N(x, a)}} \right\}.$$

Outcome-based Exploration for LLM Reasoning

Yuda Song^{1,2}, Julia Kempe^{1,3}, Remi Munos¹

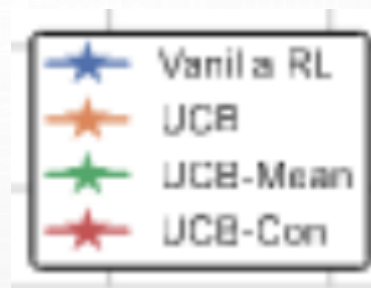
¹FAIR at Meta, ²CMU, ³NYU



DIVERSITY COLLAPSES: EXPLOITATION VERSUS EXPLORATION

Results (omitting lots of technical detail):

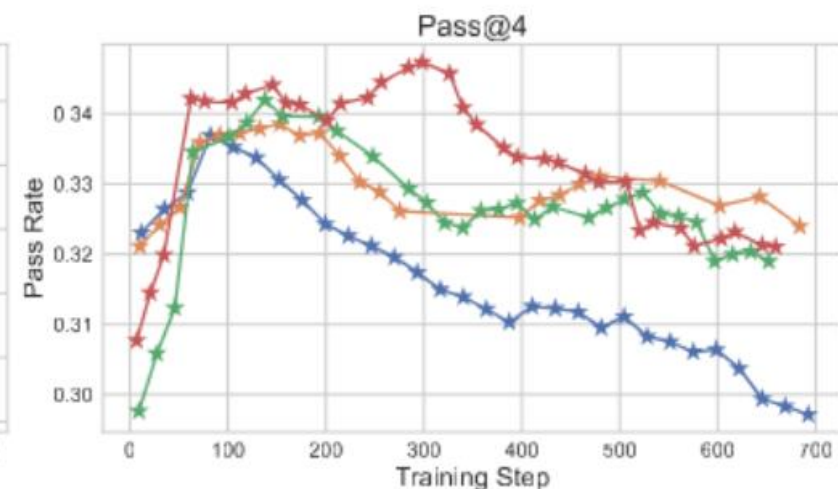
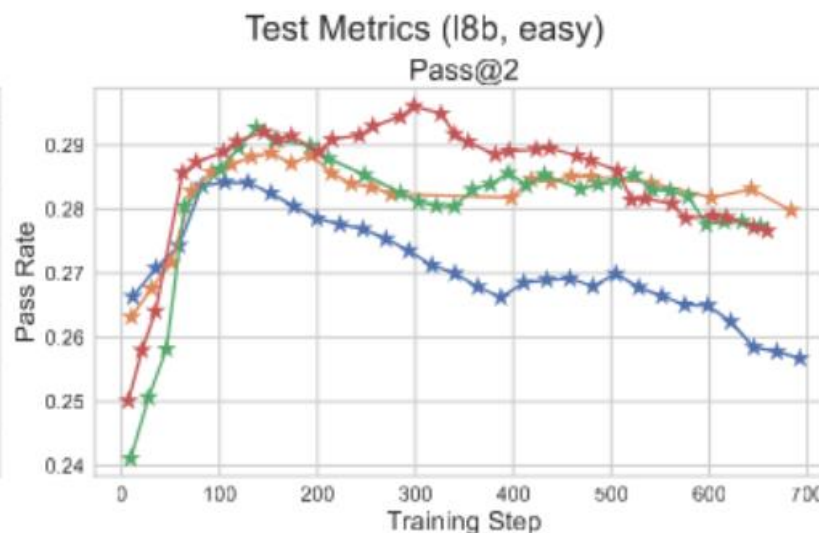
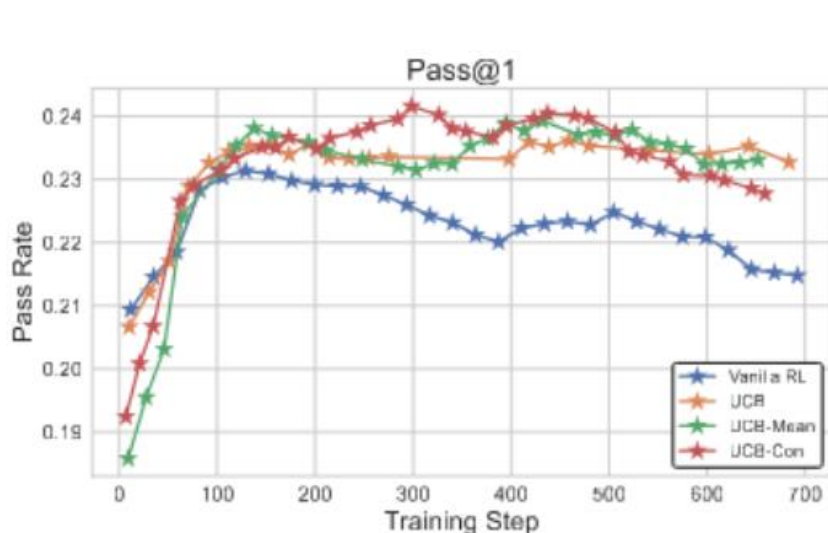
- Slower diversity degradation
- Better peak accuracy



Outcome-based Exploration for LLM Reasoning

Yuda Song^{1,2}, Julia Kempe^{1,3}, Remi Munos¹

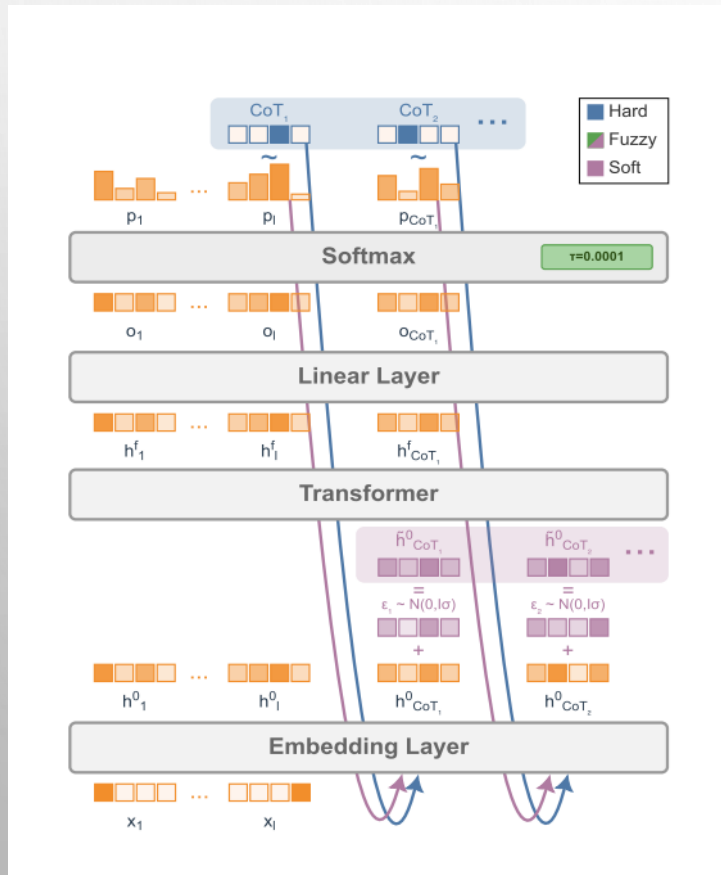
¹FAIR at Meta, ²CMU, ³NYU



EXPLORE IN LATENT SPACE

Token rollouts are very costly (long)
Diversity collapses, exploration is limited

Can we explore in latent (representation) space (not token space)?



Soft Tokens, Hard Truths

ICLR'26

Natasha Butt^{1,*}, Ariel Kwiatkowski², Ismail Labiad², Julia Kempe^{2,3,†}, Yann Ollivier^{2,†}

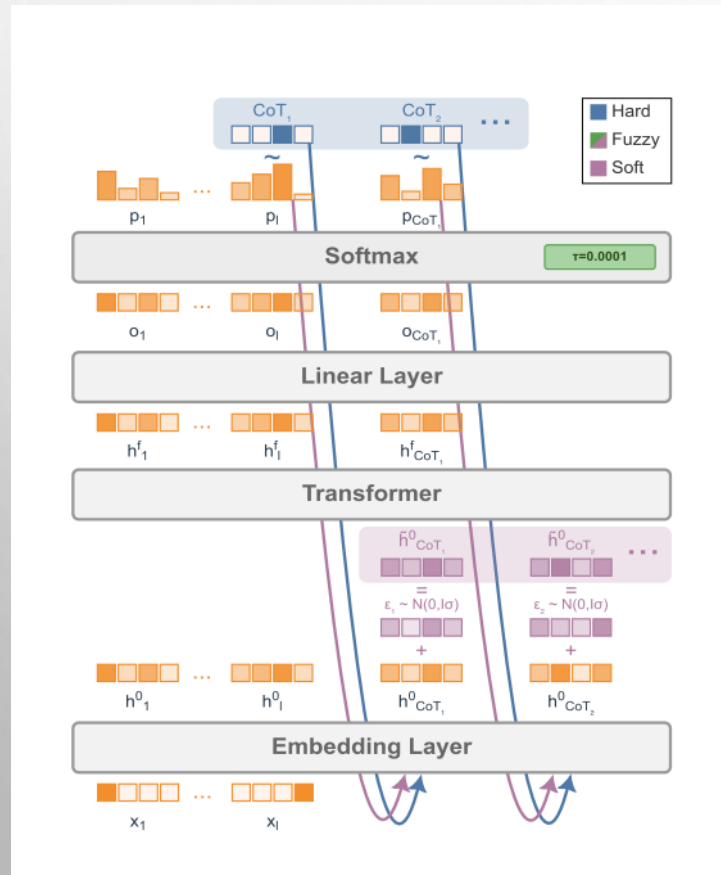
¹University of Amsterdam, ²Meta FAIR, ³New York University

*Work done during an internship at Meta, [†]Joint senior authors

EXPLORE IN LATENT SPACE

Token rollouts are very costly (long)
Diversity collapses, exploration is limited

Can we explore in latent (representation) space (not token space)?



Soft Tokens, Hard Truths

ICLR'26

Natasha Butt^{1,*}, Ariel Kwiatkowski², Ismail Labiad², Julia Kempe^{2,3,†}, Yann Ollivier^{2,†}

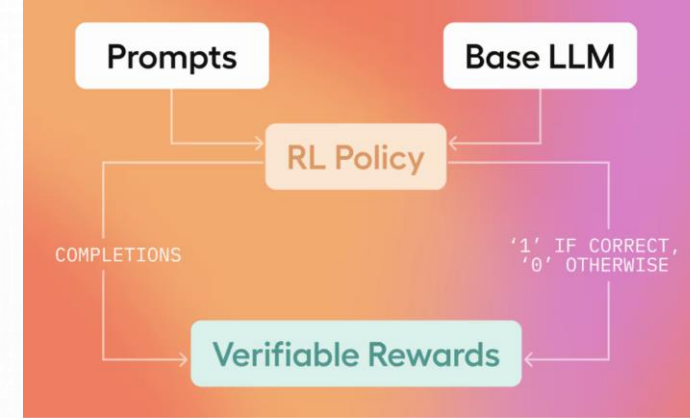
¹University of Amsterdam, ²Meta FAIR, ³New York University

*Work done during an internship at Meta, [†]Joint senior authors

Results (roughly):

- We can **scale latent inference** and RL
- Training with latent chain-of-thought gives **better diversity** (pass@32) → more exploration
- Better performance on **out-of-domain** tasks → better generalization

RLVR'S "INVISIBLE LEASH"*:



"CRITICAL CONSENSUS":

- **THE RLVR RECIPE PRIMARILY PRESERVES RATHER THAN EXPANDS THE BASE MODEL'S SOLUTION COVERAGE.**
 - Learning process is largely confined to re-weighting probabilities within the base model's pre-existing trajectories
 - Cannot discover fundamentally new reasoning trajectories outside this initial capability.
 - Diversity of answers declines.
- **HARD PROBLEMS CANNOT BE LEARNED**
 - What happens when a model gets stuck on problems it fails 100% of the time? There is no learning signal to begin. "Reasoning plateau".

*Wu et al.'25, "THE INVISIBLE LEASH? WHY RLVR MAY OR MAY NOT ESCAPE ITS ORIGIN"

Image credit: <https://labelstud.io/blog/reinforcement-learning-from-verifiable-rewards/>

META-RL: HOW TO SOLVE HARD PROBLEMS

Edge of Learnability:

- RL training often **plateaus**, “hard” problems in the training set remain unsolved
- Low initial success rate (e.g. fail@128) means **no training signal** and no learning
- Recall: RLVF learns in a sparse reward world , “wanders indefinitely” without positive reward

How to escape the “All-Zero” region?

META-RL: HOW TO SOLVE HARD PROBLEMS

Edge of Learnability:

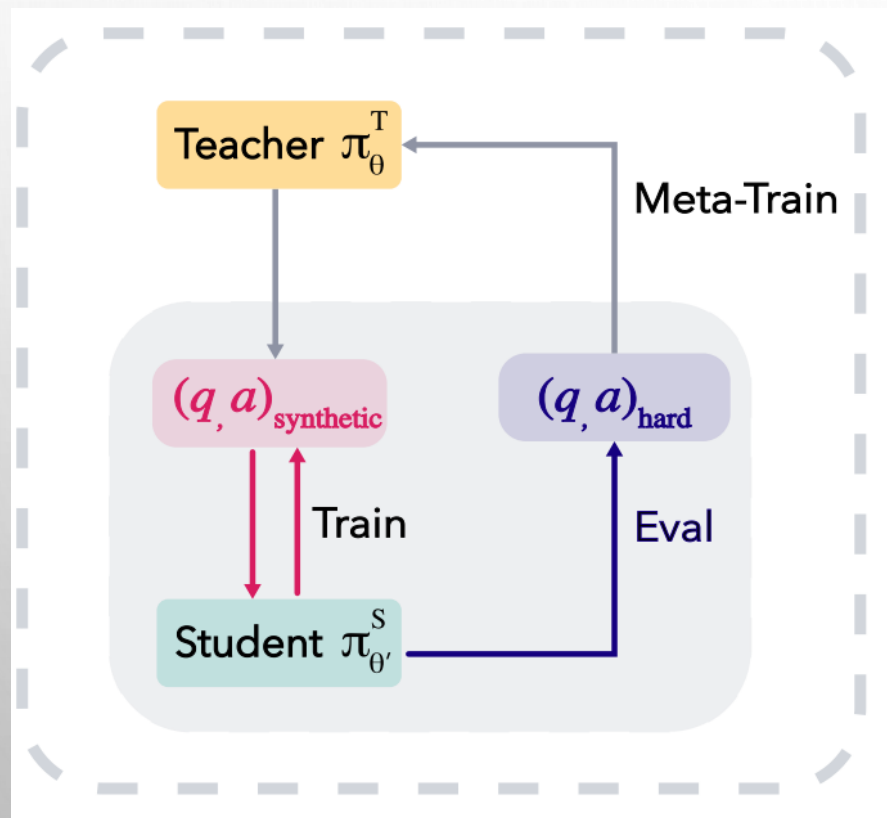
- RL training often **plateaus**, “hard” problems in the training set remain unsolved
- Low initial success rate (e.g. fail@128) means **no training signal** and no learning
- Recall: RLVF learns in a sparse reward world , “wanders indefinitely” without positive reward

How to escape the “All-Zero” region?

- **RL training on simpler questions can kickstart learning on hard datasets**
- **But how to design a “stepping stone curriculum” ?**
 - Curated additional data (expert-driven process, doesn’t scale)
 - Filtering of existing data (but what criteria? How do we know what easier questions transfer learning to hard ones? Also doesn’t scale.)

Pretrained LLMs themselves are an untapped resource (teacher) for learnable questions to unlock learning on hard datasets. RL on the teacher can “mine” pretraining corpus.

CAN A MODEL TEACH ITSELF TO SOLVE PROBLEMS THAT ARE, INITIALLY, COMPLETELY OUTSIDE ITS CAPABILITIES?



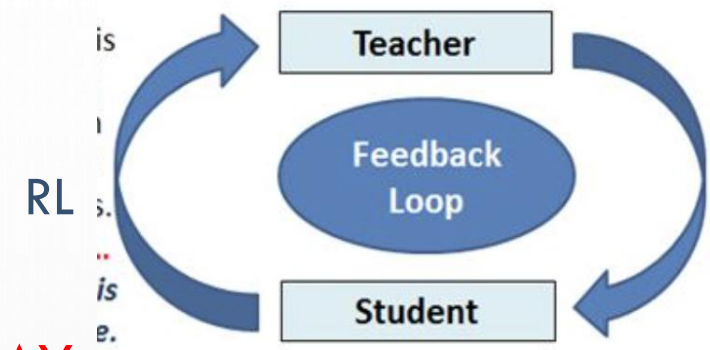
Teaching Models to Teach Themselves: Reasoning at the Edge of Learnability

ICML'26
Spotlight

Shobhita Sundaram^{1,*}, John Quan², Ariel Kwiatkowski², Kartik Ahuja², Yann Ollivier², Julia Kempe^{2,3}

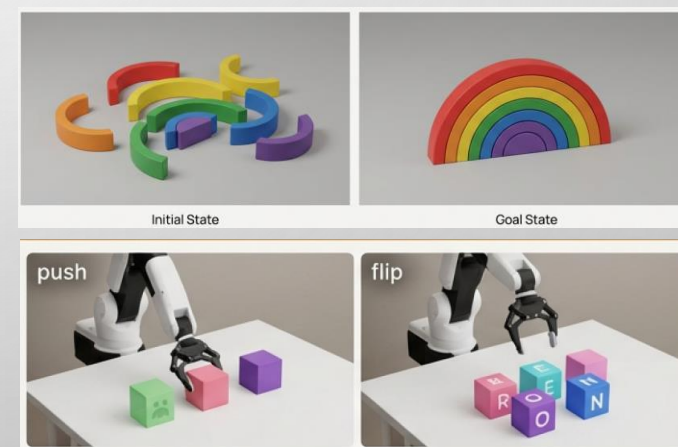
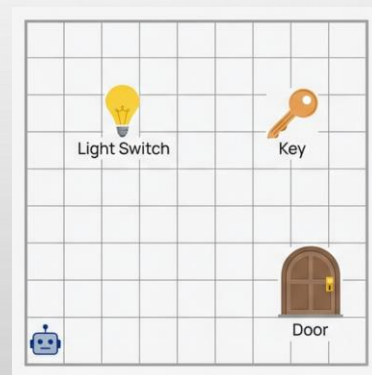
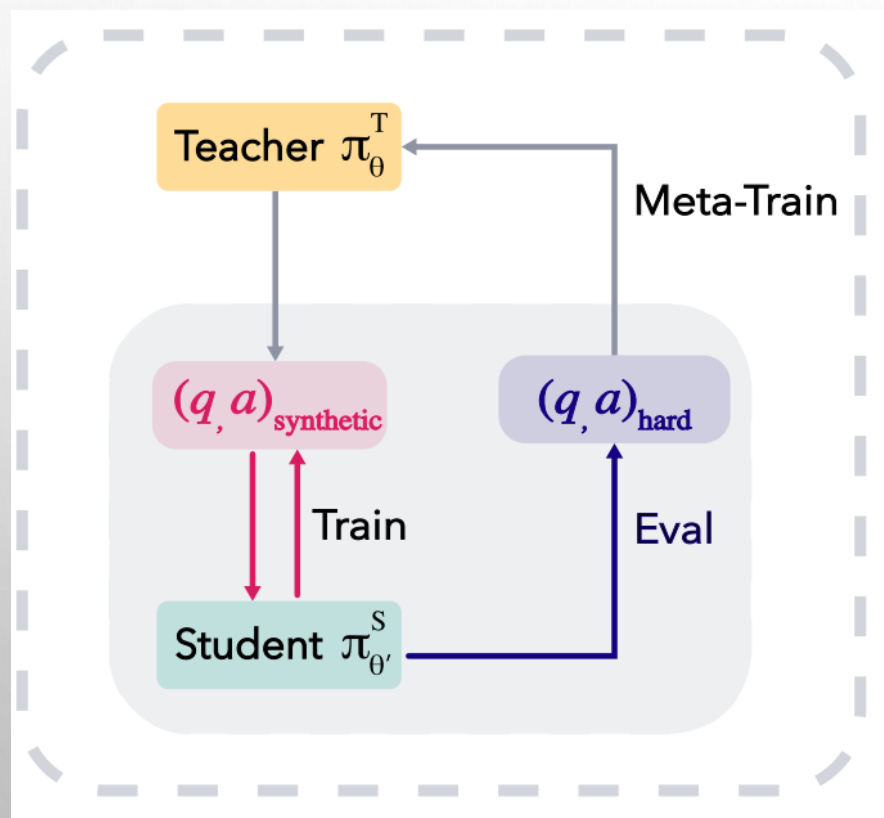
¹MIT, ²Meta FAIR, ³New York University

META-RL: HOW TO TEACH MODELS TO TEACH THEMSELVES



ASYMMETRIC SELF-PLAY:

Has proven to be a powerful engine for **skill-discovery** and **exploration** in physical domains (mazes, robotics).



Sukhbaatar et al.'17: Intrinsic motivation and automatic curricula via asymmetric self-play

OpenAI'21: "Asymmetric Self-Play for Automatic Goal Discovery..."

GROUNDING SELF-PLAY IN MEASURABLE PROGRESS

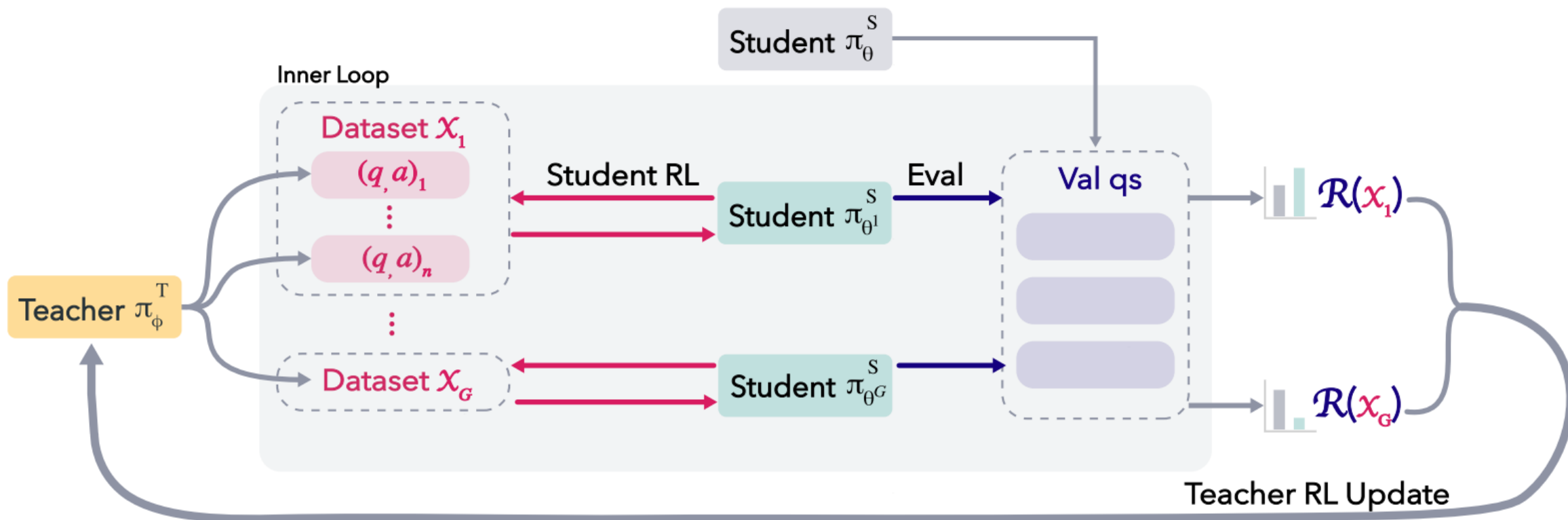
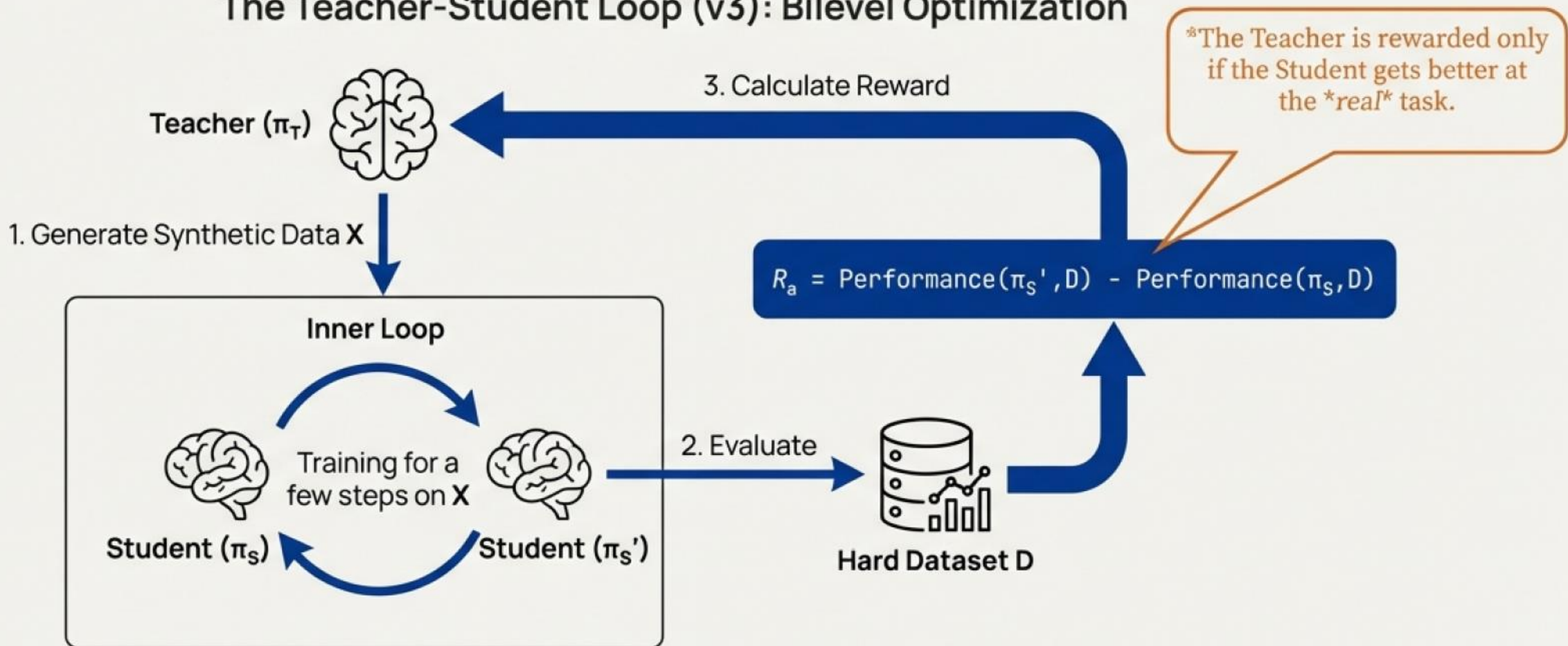


Figure 2 The soar Meta-RL Loop. The teacher and student are initialized from the same model. In the **outer RL loop** the teacher generates candidate question-answer pairs that are partitioned into datasets. In the **inner RL loop**, the student is trained for 10 steps on the candidate problems and evaluated on sampled hard problems. The teacher is rewarded based on the resulting student improvement over the student baseline, grounding the synthetic curriculum in real learning progress.

SOAR (SELF-OPTIMIZATION VIA ASYMMETRIC RL): GROUNDING SELF-PLAY IN MEASURABLE PROGRESS

The Teacher-Student Loop (v3): Bilevel Optimization



TEACHING MODELS TO TEACH THEMSELVES: TARGETED EXPERIMENTS

Technical details:

Model: LLAMA-3.2-3B-INSTRUCT

Create Hard Dataset: Filter from MATH or HARP for failure to pass @128

Setup: inner student loop – 10 steps of GRPO
outer teacher loop – sample 4 groups of 64 (q,a) pairs (GRPO with group size 4),
reward is student improvement over untrained student

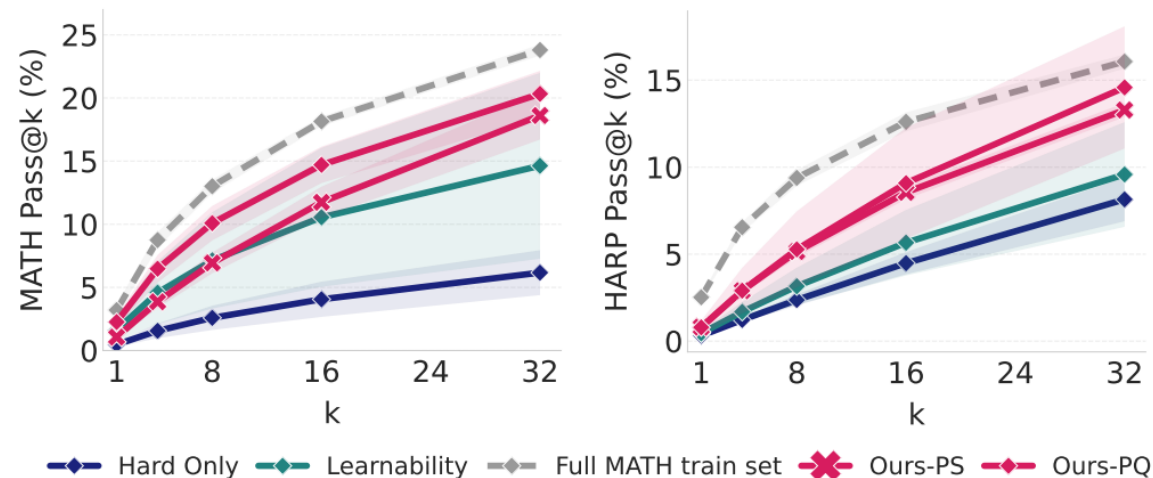


Figure 3 Performance on MATH/HARP fail@128 subsets. Synthetic curricula generated with SOAR (PQ) and inference with the promoted student (PS) improve performance on hard problems where the base model plateaus, outperforming direct training on the hard problems and teacher-training with intrinsic learnability rewards.

The ability to generate a useful curriculum does not require the pre-existing ability to solve the target problem!

TEACHING MODELS TO TEACH THEMSELVES: TARGETED EXPERIMENTS

The learned stepping-stone curriculum transfers **out of domain**.

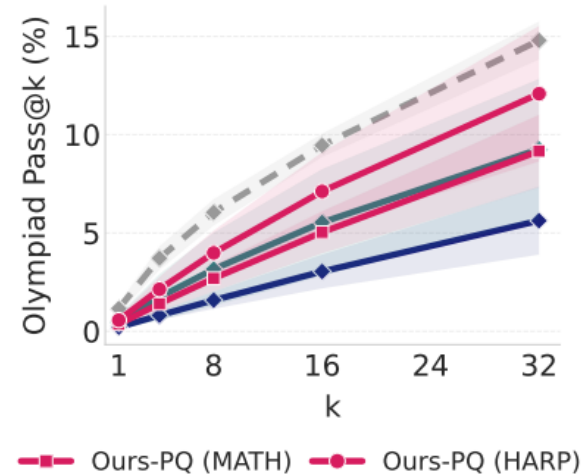


Figure 4 Transfer performance to OlympiadBench fail@128 subset. Questions optimized for MATH and HARP transfer to improving student performance on an external dataset.

Learned curricula transfer out of domain!

WHAT MAKES A GOOD “STEPPING-STONE” PROBLEM?

Stage 0

Q: Compute the value of x for the equation:
 $f(x) = (x^2 + 1) / (x - 1)$?
A: 2

Stage 1

Q: What is the probability of randomly picking a blue or green item from a bag containing 8 red, 5 blue, and 5 green items?
A: 5/9

Stage 2

Q: Given the polynomial $f(x) = x^3 - 6x^2 + 11x - 6$, factorize and solve for its roots.
A: (2, 3, 1)

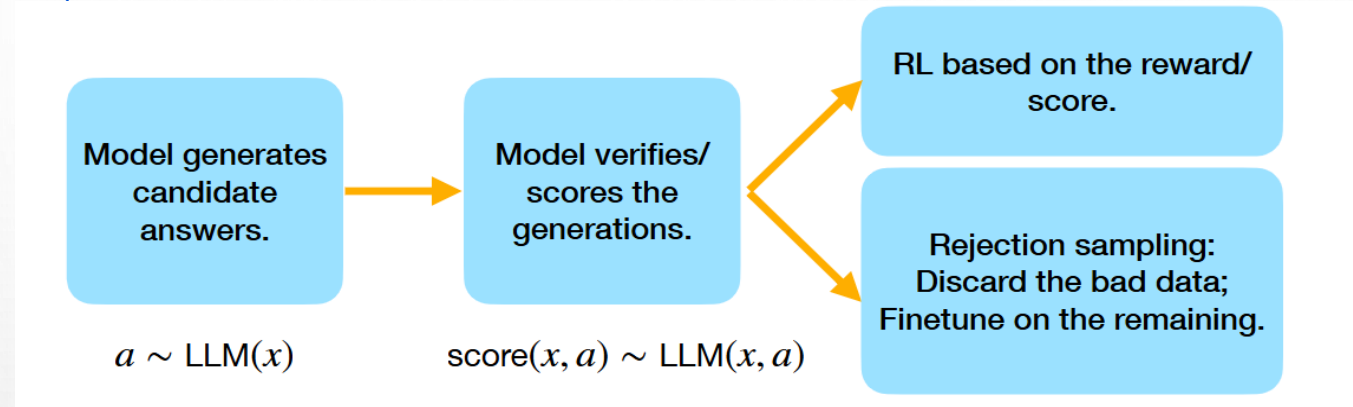
A "good" stepping-stone problem doesn't have to be perfectly correct or well-posed.

Analysis shows only **~36%** of the most effective generated problems (PQ) have correct solutions.

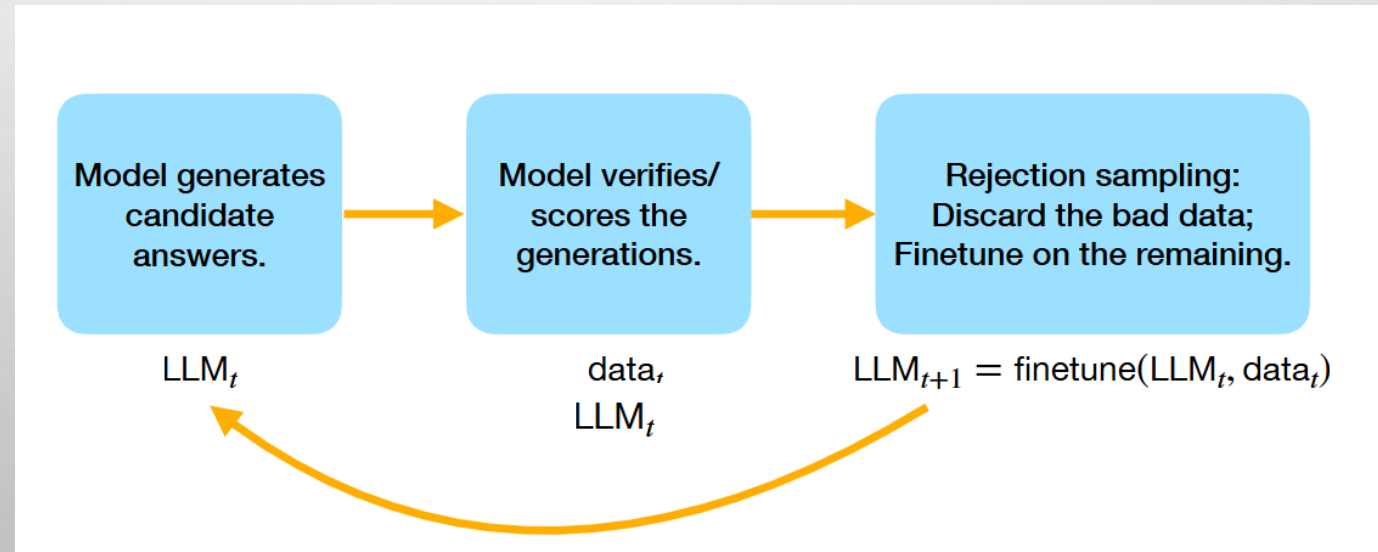
The structural and conceptual content of the question provides the crucial learning signal, even if the answer itself is noisy.

REASONING IMPROVEMENTS REQUIRE GROUNDING!

INTRINSIC (NON-GROUNDED) REWARDS



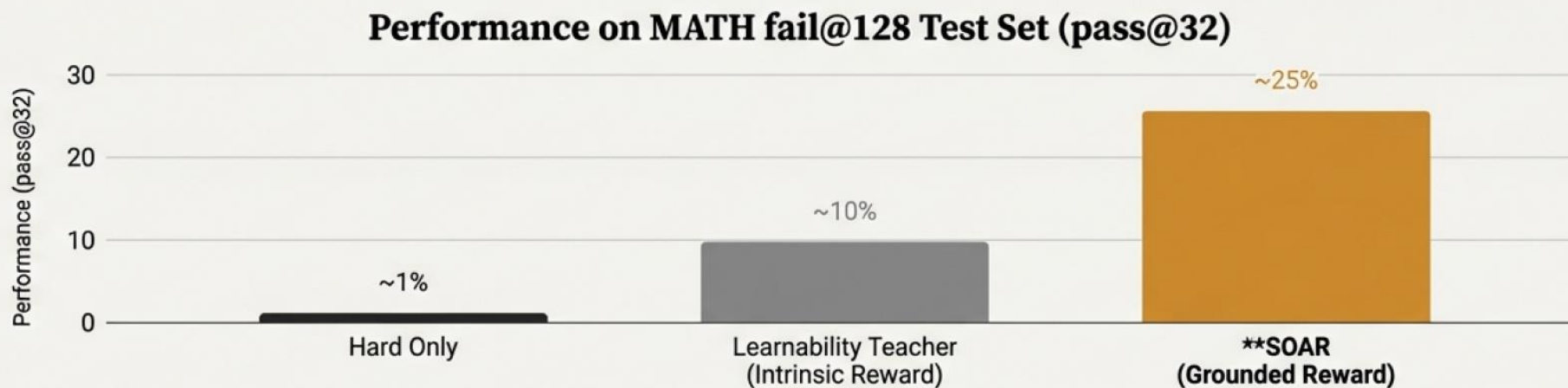
ITERATIVE SELF-IMPROVEMENT WITHOUT GROUNDING (REWARDS NOT USING GROUND TRUTH)



SELF-PLAY: INTRINSIC VERSUS GROUNDED TEACHER REWARDS

- Contemporary “zero-data” LLM self-play studies (R-Zero, AbsoluteZero, Language-Self-Play (LSP), Self-questioning LLMs (SQLM), Self-Play RL (SeRL) ...) use intrinsic teacher rewards (learnability, uncertainty,...).
- Lead to “Reward Hacking”: Curriculum can “drift”, generating interesting but irrelevant tasks, stalling student progress or collapsing it completely

Grounding the reward in real progress is critical to generating a useful curriculum.



ITERATIVE REASONING WITHOUT GROUNDING LEADS TO COLLAPSE

REWARD HACKING

DIVERSITY COLLAPSE

REMINISCENT OF MODEL COLLAPSE WITH SYNTHETIC DATA

MODEL COLLAPSE ...

Shumailov et al. '23 (Nature '24)

THE CURSE OF RECURSION: TRAINING ON GENERATED DATA MAKES MODELS FORGET

Ilia Shumailov*
University of Oxford

Zakhar Shumaylov*
University of Cambridge

Yiren Zhao
Imperial College London

Yarin Gal
University of Oxford

Nicolas Papernot
University of Toronto & Vector Institute

Ross Anderson
University of Cambridge & University of Edinburgh

nature

Explore content ▾ About the journal ▾ Publish with us ▾

nature > articles > article

Article | [Open access](#) | Published: 24 July 2024

AI models collapse when trained on recursively generated data

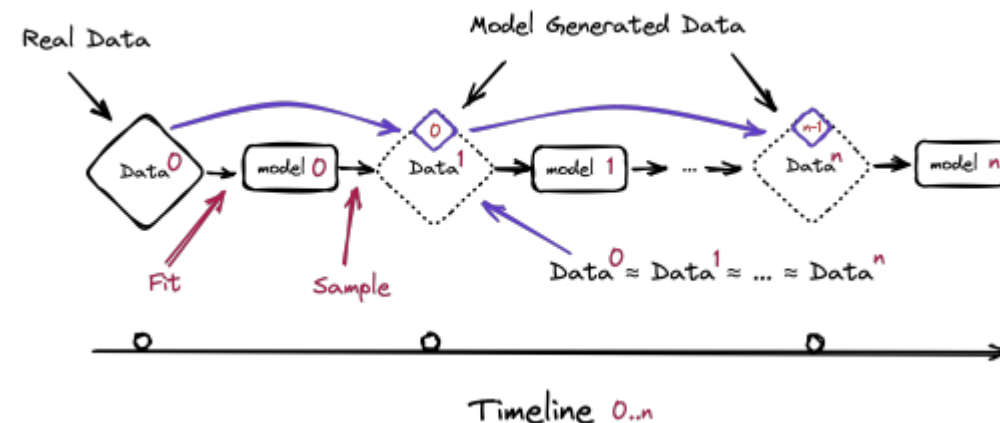
[Ilia Shumailov](#) , [Zakhar Shumaylov](#) , [Yiren Zhao](#), [Nicolas Papernot](#), [Ross Anderson](#) & [Yarin Gal](#) 

Nature **631**, 755–759 (2024) | [Cite this article](#)

268k Accesses | **4** Citations | **2558** Altmetric | [Metrics](#)

Abstract

Stable diffusion revolutionized image creation from descriptive text. GPT-2 (ref. ¹), GPT-3(⁵) (ref. ²) and GPT-4 (ref. ³) demonstrated high performance across a variety of language tasks. ChatGPT introduced such language models to the public. It is now clear that generative artificial intelligence (AI) such as large language models (LLMs) is here to stay and will substantially change the ecosystem of online text and images. Here we consider what may happen to GPT- $\{n\}$ once LLMs contribute much of the text found online. We find that indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails of the original content distribution disappear. We refer to this effect as ‘model collapse’ and show that it can occur in LLMs as well as in variational autoencoders (VAEs) and Gaussian mixture models (GMMs). We build theoretical intuition behind the phenomenon and portray its ubiquity among all learned generative models. We demonstrate that it must be taken seriously if we are to sustain the benefits of training from large-scale data scraped from the web. Indeed, the value of data collected about genuine human interactions with systems will be increasingly valuable in the presence of LLM-generated content in data crawled from the Internet.



MODEL COLLAPSE ...



Example 1

Example of text outputs of an OPT-125m model affected by model collapse—models degrade over generations, for which each new generation is trained on data produced by the previous generation.

- **Input:** some started before 1360 — was typically accomplished by a master mason and a small team of itinerant masons, supplemented by local parish labourers, according to Poyntz Wright. But other authors reject this model, suggesting instead that leading architects designed the parish church towers based on early examples of Perpendicular.
- **Outputs:**
 - **Gen 0:** Revival architecture such as St. John's Cathedral in London. The earliest surviving example of Perpendicular Revival architecture is found in the 18th @-@ century Church of Our Lady of Guernsey, which dates from the late 19th century. There are two types of perpendicular churches : those.
 - **Gen 1:** architecture such as St. Peter's Basilica in Rome or St. Peter's Basilica in Buenos Aires. There is no evidence that any of these buildings were built during the reign of Pope Innocent III, but it is possible that they may have been built during the reign of his successor, Pope Innocent.
 - **Gen 5:** ism, which had been translated into more than 100 languages including English, French, German, Italian, Spanish, Portuguese, Dutch, Swedish, Norwegian, Polish, Hungarian, Slovak, Lithuanian, Estonian, Finnish, Romanian, Bulgarian, Turkish, Croatian, Serbian, Ukrainian, Russian, Kazakh, Kyrgyz.
 - **Gen 9:** architecture. In addition to being home to some of the world's largest populations of black @-@ tailed jackrabbits, white @-@ tailed jackrabbits, blue @-@ tailed jackrabbits, red @-@ tailed jackrabbits, yellow @-.

The New York Times

When A.I.'s Output Is a Threat to A.I. Itself

As A.I.-generated data becomes harder to detect, it's increasingly likely to be ingested by future A.I., leading to worse results.

By Aatish Bhatia

Aatish Bhatia interviewed A.I. researchers, studied research papers and fed an A.I. system its own output.

Aug. 25, 2024

The New York Times

When A.I.'s Output Is a Threat to A.I. Itself

As A.I.-generated data becomes harder to detect, it's increasingly likely to be ingested by future A.I., leading to worse results.

By [Aatish](#)

Aatish

own o

Aug. 25

am its

MODEL COLLAPSE ...

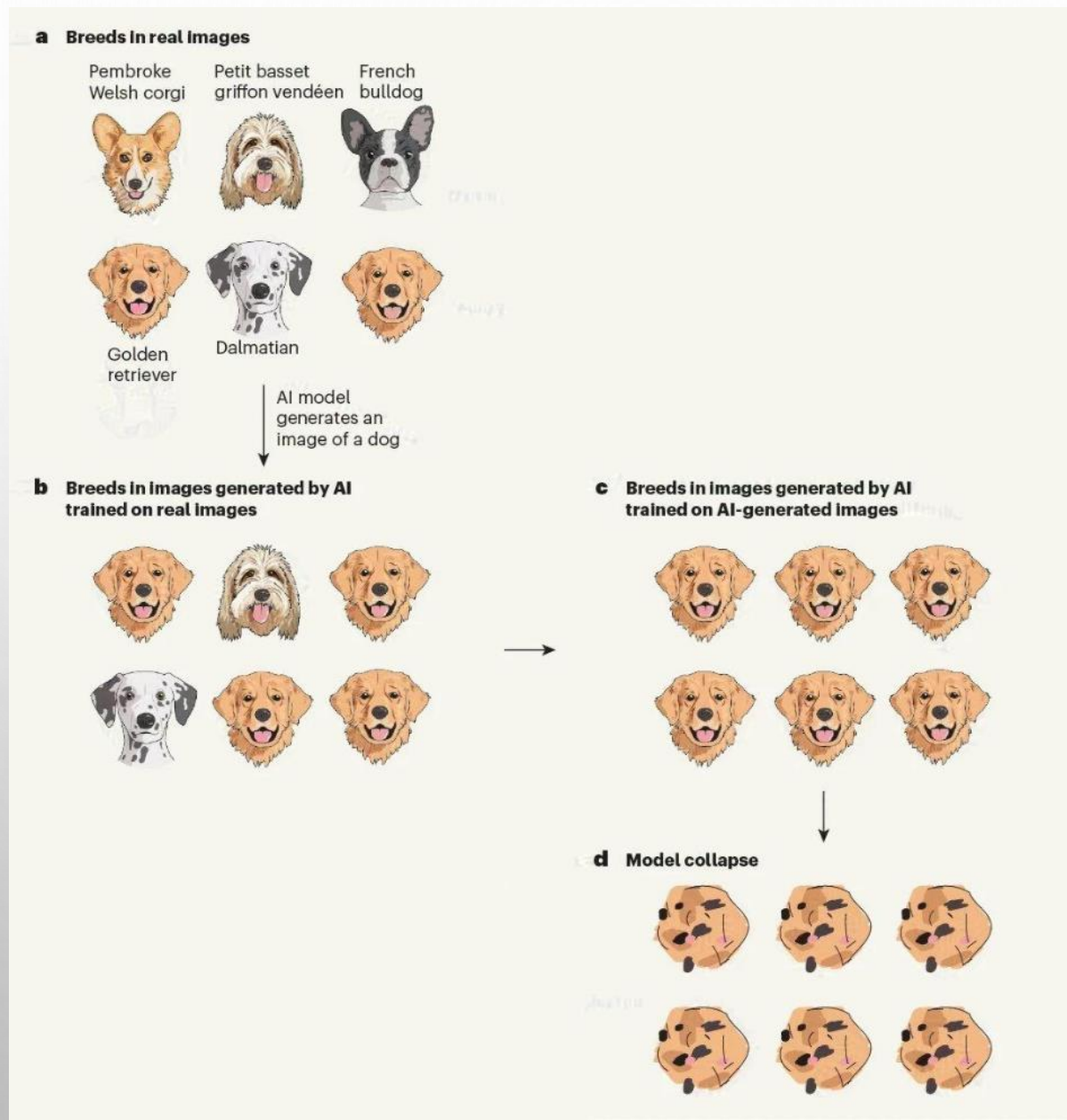


Image from: https://www.reddit.com/r/ArtistHate/comments/1ecx32v/this_chart_explains_the_ml_model_collapse_in_a/#lightbox

SELF-CONSUMING FEEDBACK LOOP...

What to Know About Tech Companies Using A.I. to Teach Their Own A.I.

As artificial intelligence developers run out of data to train their models, they are turning to “synthetic data” — data made by the A.I. itself.



By Cade Metz and Stuart A. Thompson

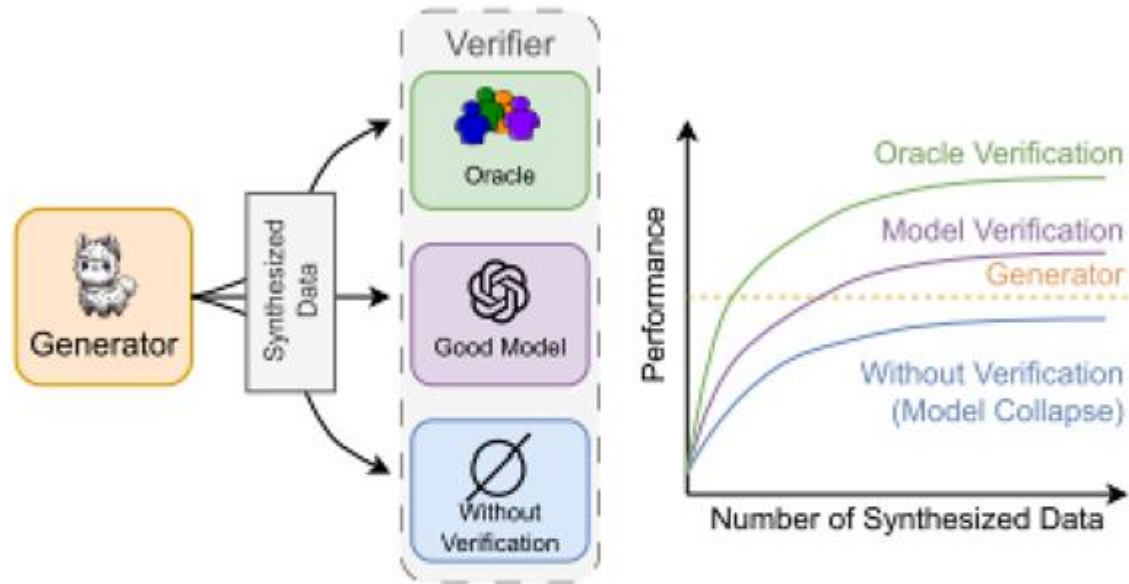
Cade Metz reports on artificial intelligence. Stuart Thompson reports on disinformation.

April 6, 2024

Does synthetic data work?

Not exactly. A.I. models get things wrong and make stuff up. They have also shown that they pick up on the biases that appear in the internet data from which they have been trained. So if companies use A.I. to train A.I., they can end up amplifying their own flaws.

CAN WE SALVAGE SYNTHETIC DATA?



(a) The generator is capable of producing high-quality data, and verifiers can extract this data through selection to prevent model collapse, which occurs in the absence of such selection.

Checking the data and throwing out the “bad” helps !!!
Which is what GRPO/RL does.

“A Tail of Tails: Model Collapse as a change of scaling laws”, Dohmatob, Feng, Yang, Charton, JK, ICML’24

“Model Collapse Demystified”, Dohmatob, Feng, JK, NeurIPS’24

“Strong Model Collapse”, Dohmatob, Feng, Subramanian, JK, ICLR’25 Spotlight

“Beyond Model Collapse”, Feng, Dohmatob, Yang, Charton, JK, ICLR’25

SYNTHETIC DATA ALONE – NO
SYNTHETIC DATA WITH CHECKING – YES!

BUT WHO IS GOING TO CHECK???

SYNTHETIC DATA ALONE – NO SYNTHETIC DATA WITH CHECKING – YES!

BUT WHO IS GOING TO CHECK???

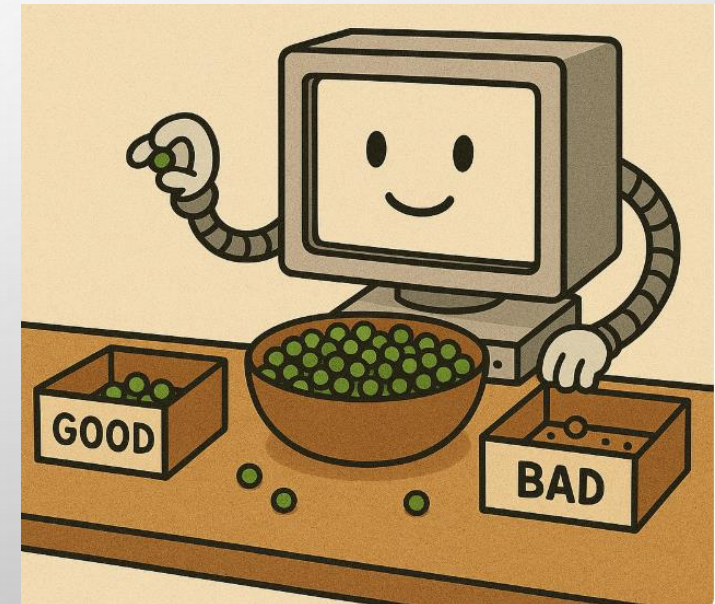
- HUMANS? (RLHF)



SYNTHETIC DATA ALONE – NO SYNTHETIC DATA WITH CHECKING – YES!

BUT WHO IS GOING TO CHECK???

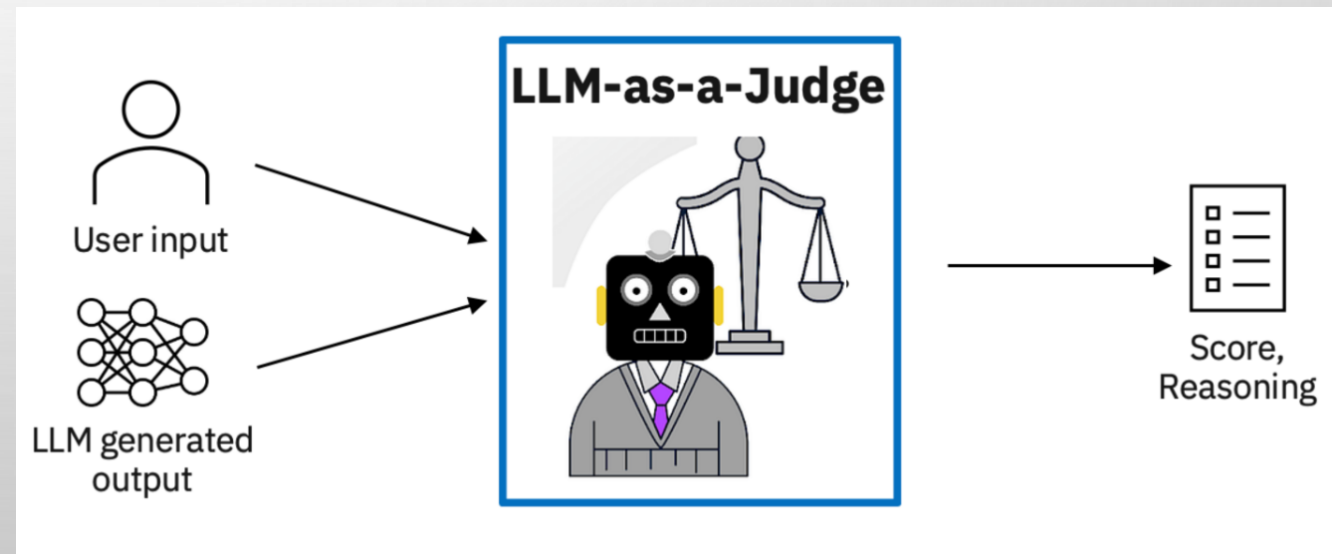
- HUMANS? (RLHF)
- COMPUTERS?



SYNTHETIC DATA ALONE – NO SYNTHETIC DATA WITH CHECKING – YES!

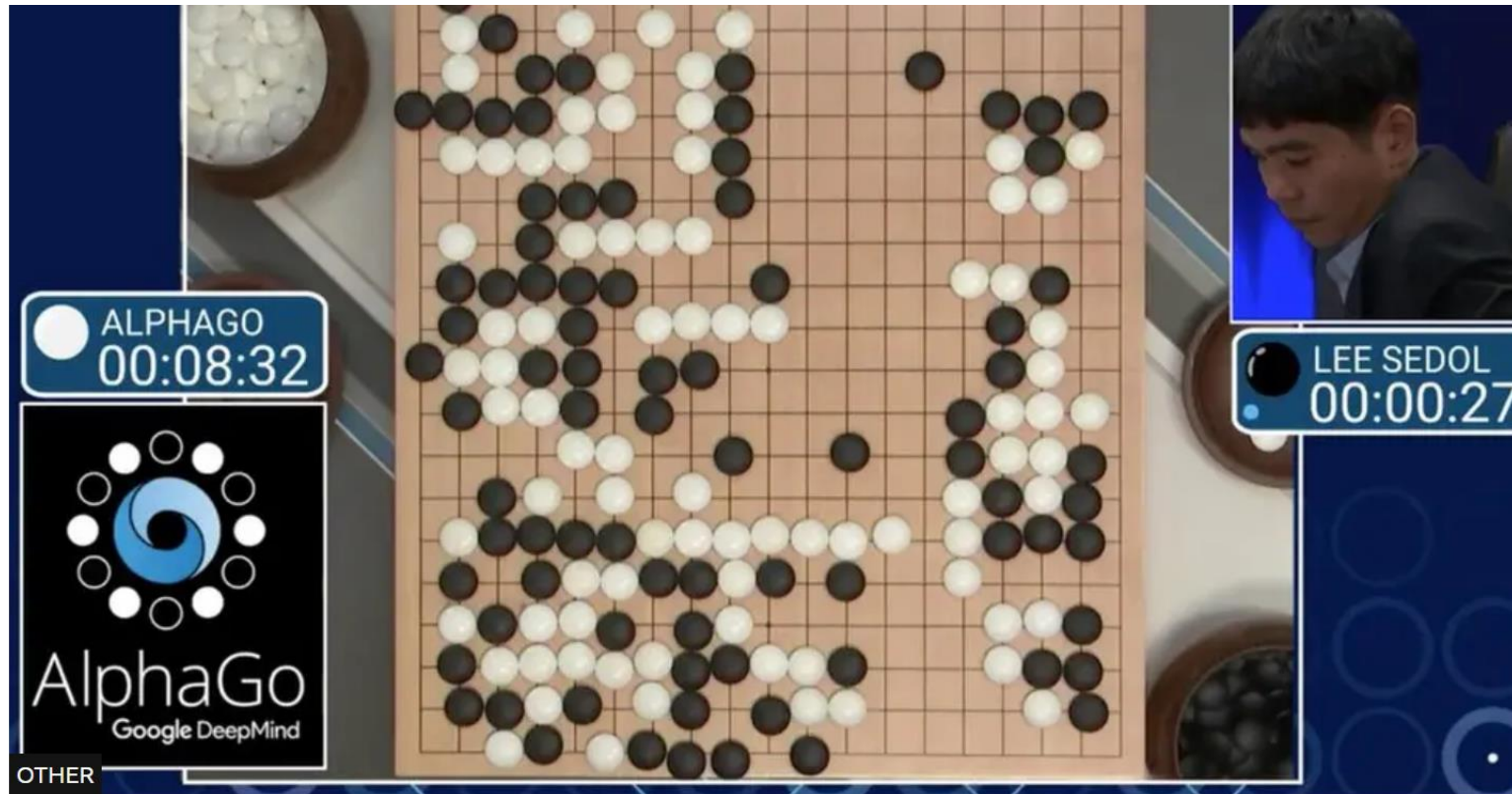
BUT WHO IS GOING TO CHECK???

- HUMANS? (RLHF)
- COMPUTERS?
- OTHER LLMS ???



Artificial intelligence: Google's AlphaGo beats Go master Lee Se-dol

🕒 12 March 2016



A computer program has beaten a master Go player 3-0 in a best-of-five competition, in what is seen as a landmark moment for artificial intelligence.

ALPHAGO – HOW WAS IT TRAINED?

- ON A LOT OF RECORDED GAMES?

ALPHAGO – HOW WAS IT TRAINED?

- ON A LOT OF RECORDED GAMES?
- ON SELF-PLAY! SELF-GENERATED GAMES !!!!

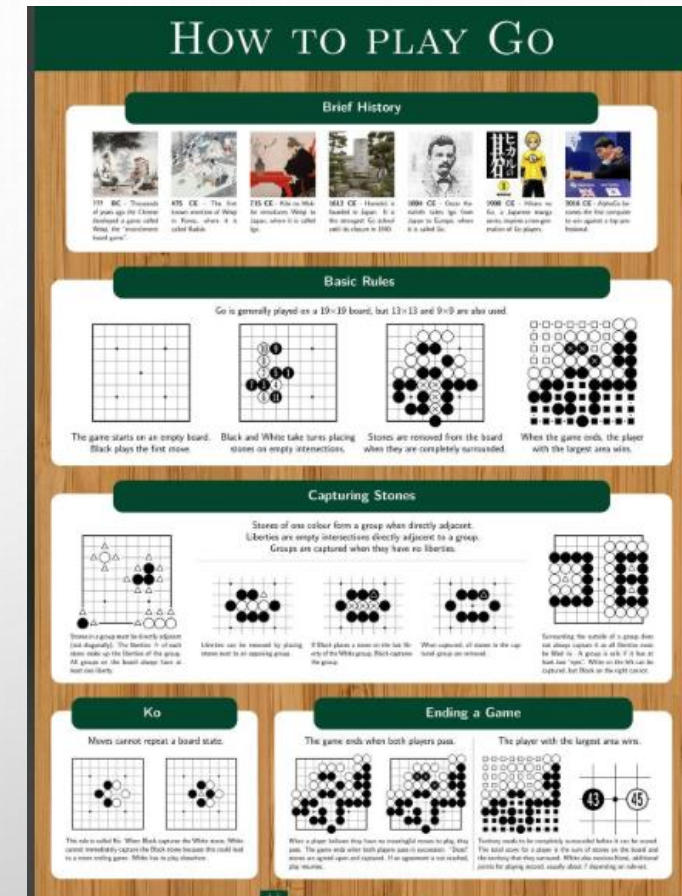
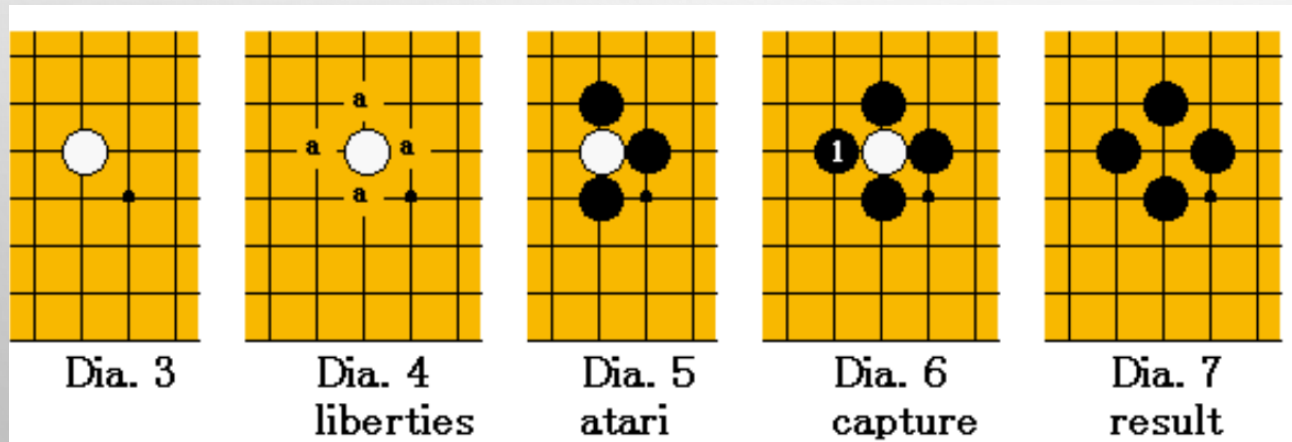
ALPHAGO – HOW WAS IT TRAINED?

- ON A LOT OF RECORDED GAMES?
- ON SELF-PLAY! SELF-GENERATED GAMES !!!!
- WAIT A MINUTE! WHAT ABOUT MODEL COLLAPSE ????



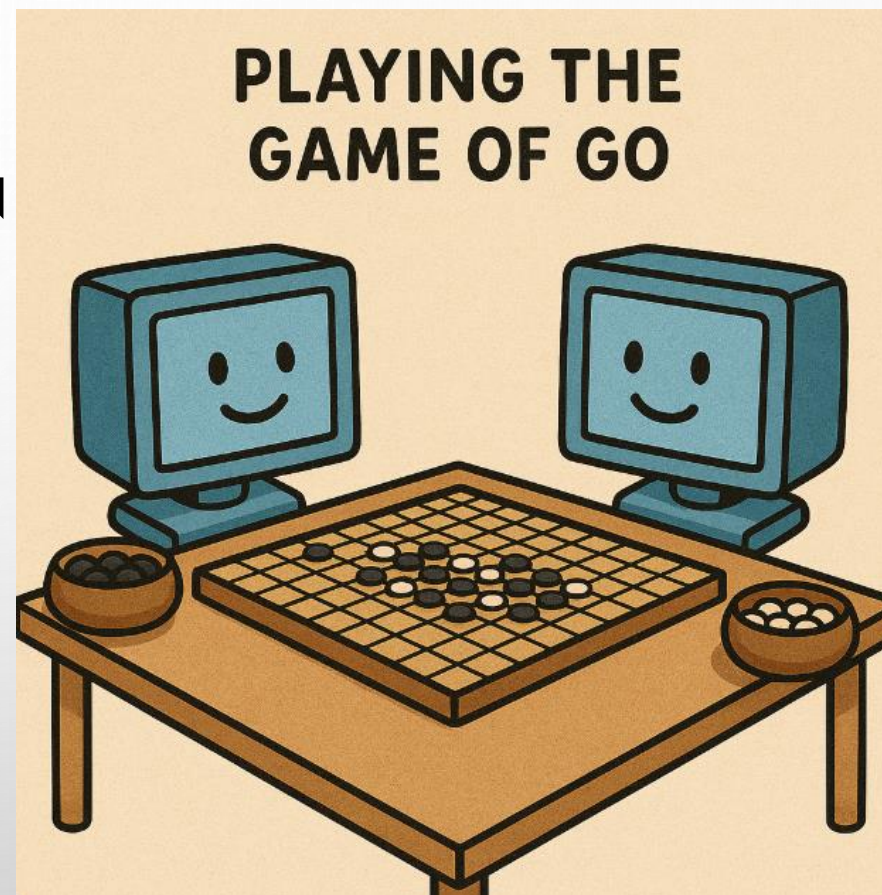
ALPHAGO – HOW WAS IT TRAINED?

- GO HAS RULES
- THEY CAN BE CHECKED



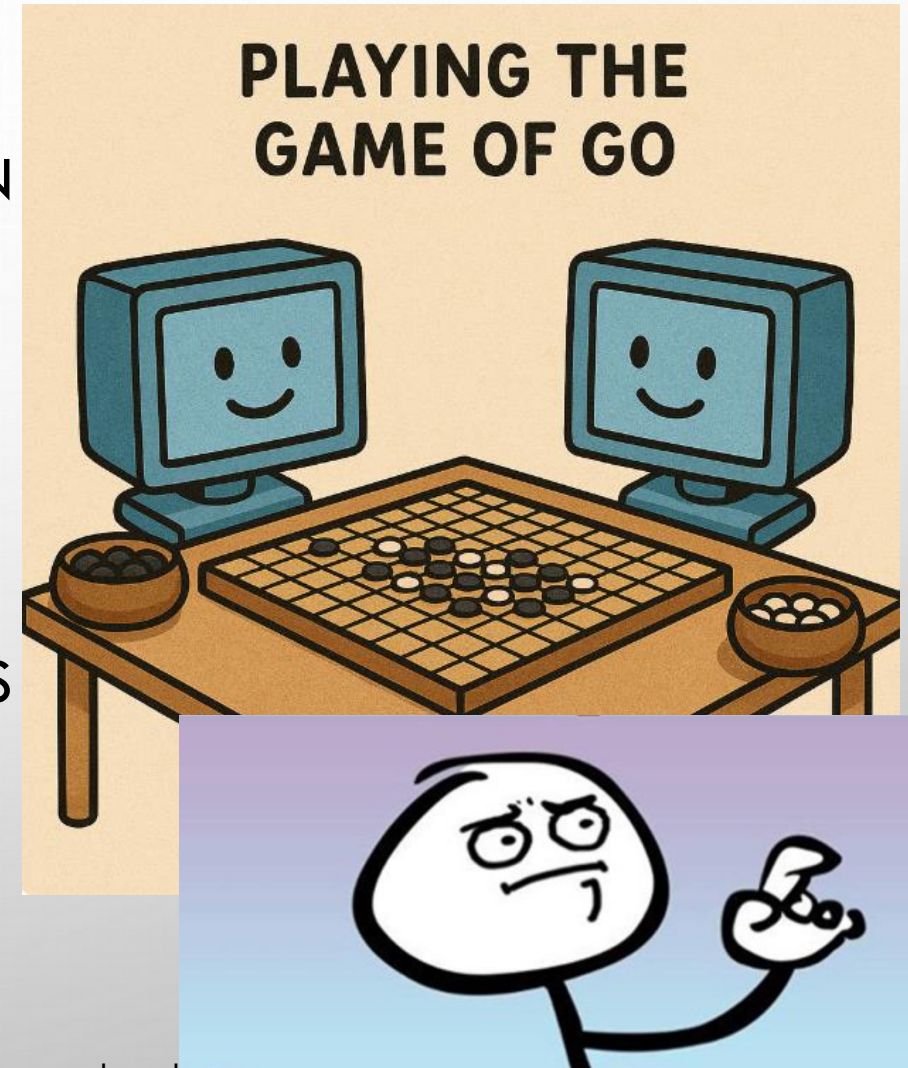
ALPHAGO – NO HUMAN DATA AT ALL!

- GO HAS RULES
- THEY CAN BE CHECKED **AUTOMATICALLY** BY AN AUTOMATED RULE CHECKER
- GAMES CAN BE **SCORED AUTOMATICALLY**
- THIS WAY, WE CAN GET LOTS OF CHECKED **SYNTHETIC DATA**



ALPHAGO – NO HUMAN DATA AT ALL!

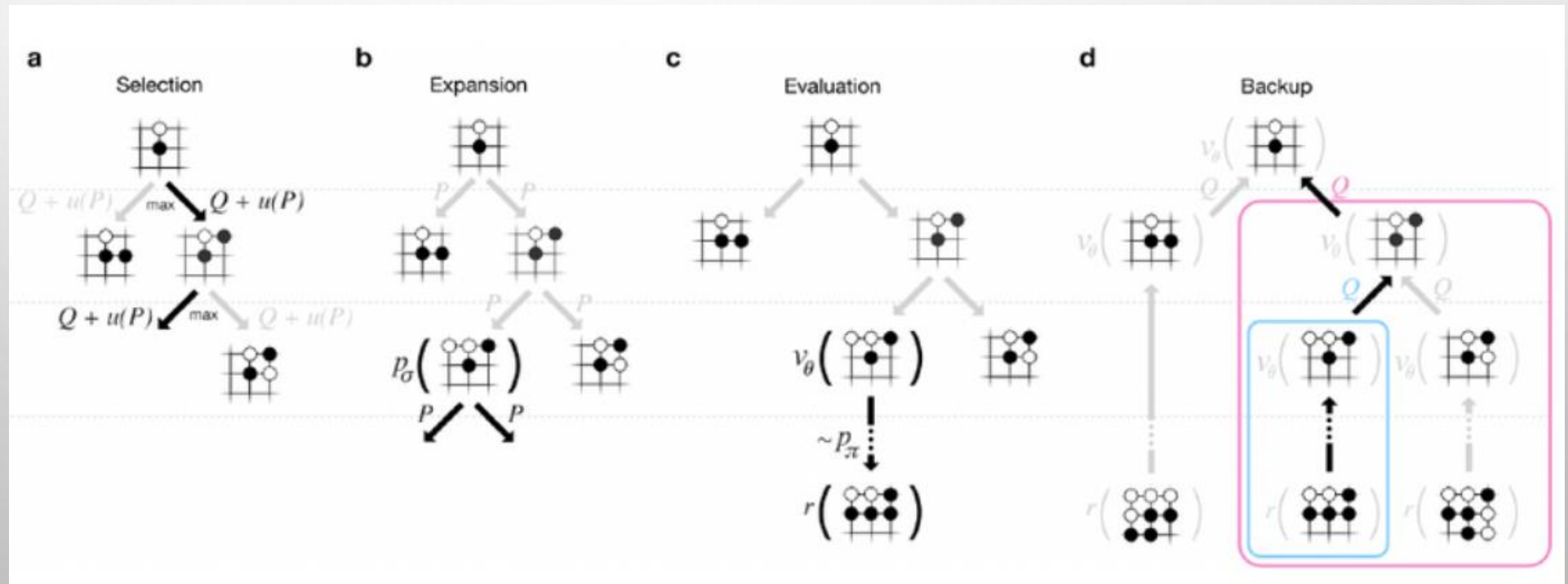
- GO HAS RULES
- THEY CAN BE CHECKED **AUTOMATICALLY** BY AN AUTOMATED RULE CHECKER
- GAMES CAN BE **SCORED AUTOMATICALLY**
- THIS WAY, WE CAN GET LOTS OF CHECKED **SYNTHETIC DATA**
- ...AND TRAIN OUR MODEL ON THE BEST GAMES
- AND HAVE IT PLAY AGAIN
- AND REPEAT....



ALPHAGO – HOW WAS IT TRAINED?

AND MANY OTHER INGREDIENTS....

- MONTE-CARLO TREE SEARCH
 - TO LOOK FOR STRATEGIES



ALPHAGO – HOW WAS IT TRAINED?

AND MANY OTHER INGREDIENTS....

- MONTE-CARLO TREE SEARCH
 - TO LOOK FOR STRATEGIES
- REINFORCEMENT LEARNING
 - TO LEARN FROM GOOD GAMES AND “UNLEARN” FROM BAD ONES
- ...

Artificial intelligence: Google's AlphaGo beats Go master Lee Se-dol

🕒 12 March 2016

“

I thought AlphaGo was based on probability calculation and that it was merely a machine. But when I saw this move, I changed my mind. Surely, AlphaGo is creative.

LEE SEDOL
WINNER OF 18 WORLD GO TITLES

HOW LLMS WON GOLD MEDAL AT THE 2025 IMO

The New York Times

July 21, 2025

Google A.I. System Wins Gold Medal in International Math Olympiad

OpenAI said it, too, had built a system that achieved similar results.

July 20, 2026 JACOBIAN CONJECTURE

Harvard mathematician uses AI to crack 87-year-old riddle

A 33-year-old junior fellow appears to have disproven the Jacobian conjecture

May 29, 2026 ERDOS UNIT DISTANCE CONJECTURE REFUTED

THE WALL STREET JOURNAL.

TECHNOLOGY • ARTIFICIAL INTELLIGENCE • SCIENCE OF SUCCESS

A Famous Math Problem Stumped Humans for 80 Years. AI Just Cracked It.

The math world is losing its mind over the new solution to an Erdős problem. This is what AI found, how we missed it—and why it matters.

August 1, 2026 Publication

Ten advances in mathematics and theoretical computer science

Read the paper ↗

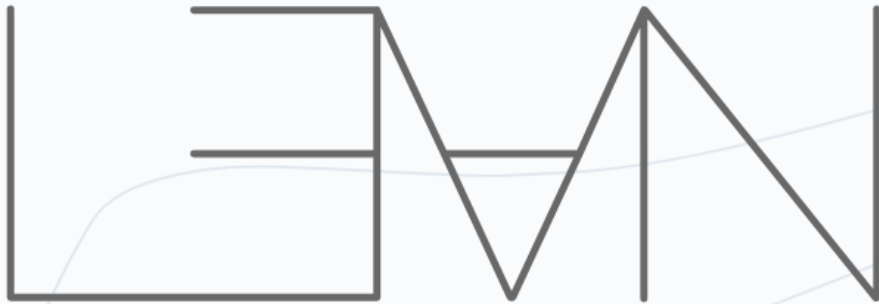
Read the reasoning walkthroughs ↗

HOW LLMS WON GOLD MEDAL AT THE 2025 IMO ... AND STARTED TO SOLVE RESEARCH MATH IN 2026

- THE “TRICK” ?
- MATH CAN BE VERIFIED!!!
- IT IS ARGUABLY EASIER TO VERIFY A PROOF THEN TO FIND IT....

HOW LLMS WON GOLD MEDAL AT THE 2025 IMO ... AND STARTED TO SOLVE RESEARCH MATH IN 2026

- MATH CAN EVEN BE AUTOMATICALLY VERIFIED... WITH LEAN



Lean is a theorem prover and programming language that enables correct, maintainable, and formally verified code

```
example (x : Nat) : 0 < match x with
| 0    => 1
| n+1 => x + n := by
  grind

-- Automatically solves systems of linear inequalities.

example (x y : Int) :
  27 ≤ 11*x + 13*y → 11*x + 13*y ≤ 45
  → -10 ≤ 7*x - 9*y → 7*x - 9*y > 4 := by
  grind
```

AUTOMATED PROOF VERIFICATION = GROUNDING IN “TRUTH”

AUTO-FORMBOT AND ATLAS

CREATING A LARGE LEAN CORPUS AND AN AUTO-FORMALIZATION HARNESS

Formalizing Mathematics at Scale

Ahmad Rammal^{1,2}, Niket Patel^{1,3}, Fabian Goeckle^{1,†,2}, Amaury Hayat^{2,4}, Julia Kempe^{1,3}, Remi Munos¹, Charles Arnal^{1,*}, Vivien Cabannes^{1,*}

¹FAIR at Meta, ²CERMICS, ENPC, Institut Polytechnique de Paris, ³New York University, ⁴Korea Institute for Advanced Study

*Equal contribution, [†]Work done while at Meta

We present AUTOFORMBOT, a multi-agent system for building an **Autoformalized Textbook Library At Scale (ATLAS)** in Lean 4. **AUTOFORMBOT** orchestrates thousands of LLM agents, equipped with formal verification tools, dependency-aware task scheduling, and collaborative version control, to translate informal textbook prose into machine-checked definitions and proofs. We apply our methods to a **corpus of 26 open-access textbooks** spanning analysis, algebra, topology, combinatorics, and probability, producing ATLAS: a verified library of over **45,000** Lean 4 declarations and **500 thousand** lines of code. We release two artifacts: (i) AUTOFORMBOT, the open-source multi-agent framework; and (ii) ATLAS, the resulting formal library. Our results suggest that autoformalizing the core content of graduate-level mathematics at scale is now economically and technically feasible. This opens the door to the automated verification of both human- and machine-generated mathematics at a research level.

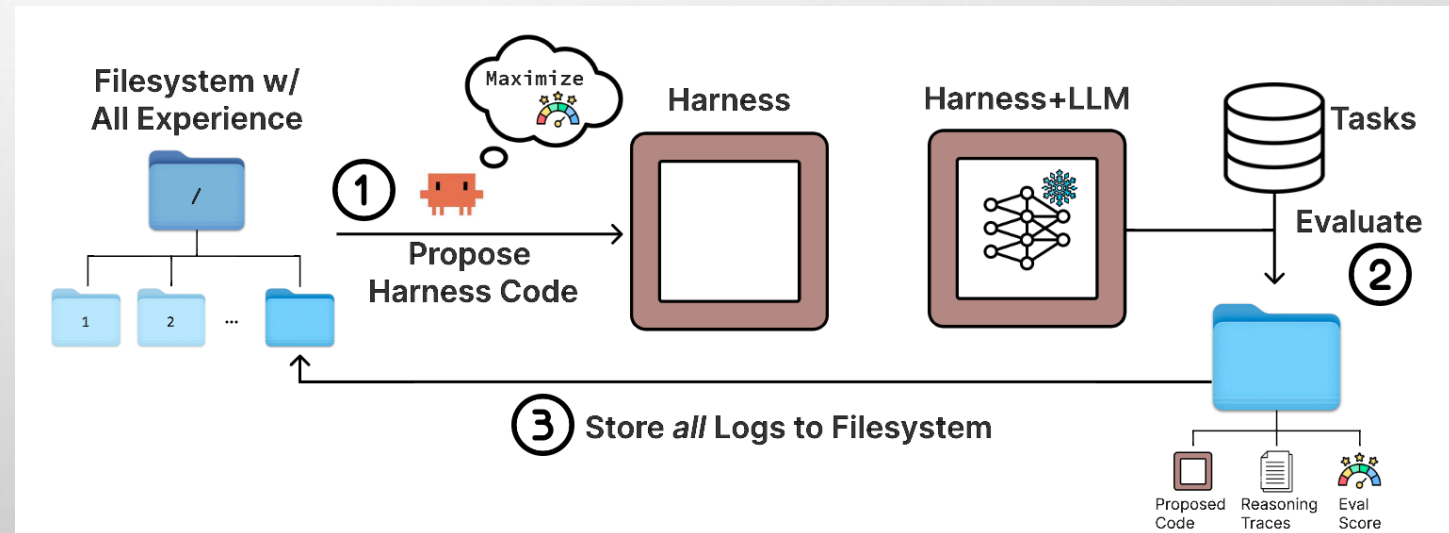
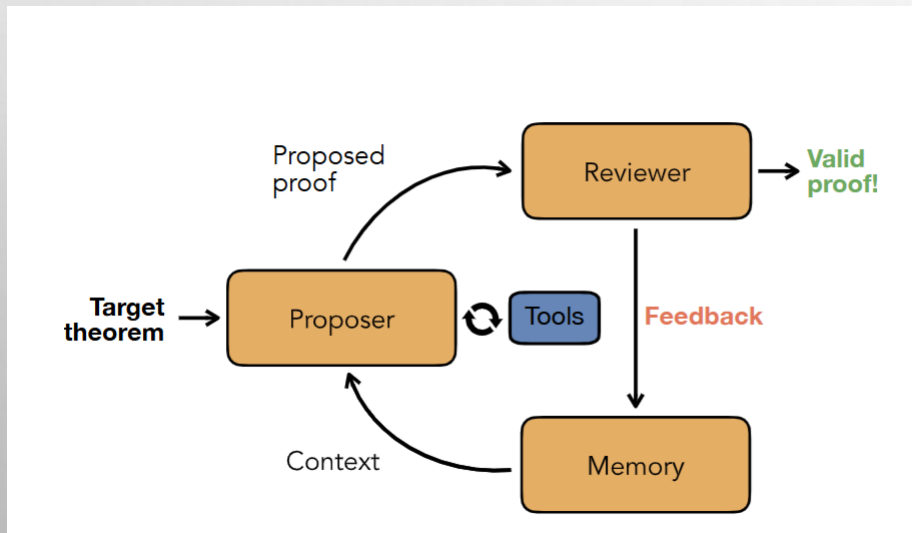
Date: May 29, 2026



LLMS NOW

2026: Agents, Harnesses, “Self-Improvement”

- Harnesses and multi-agent workflows
- (Some form of) continual learning and differentiation through contexts and skill.md files
- Allows for “meta-RL” frameworks



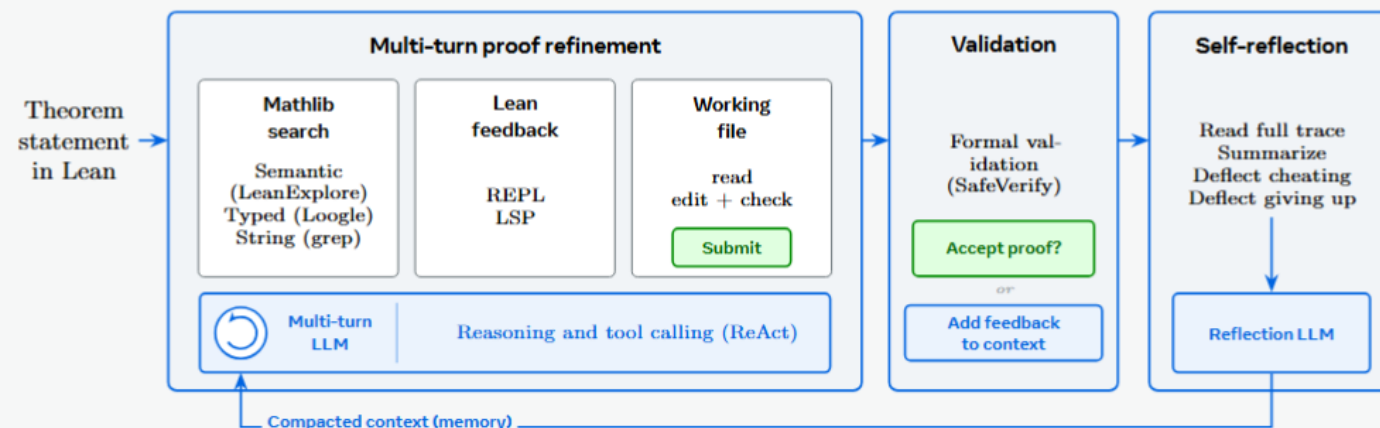
FROM PROOF SEARCH TO PROVER AGENTS

HARNESSES AND PROVER AGENTS

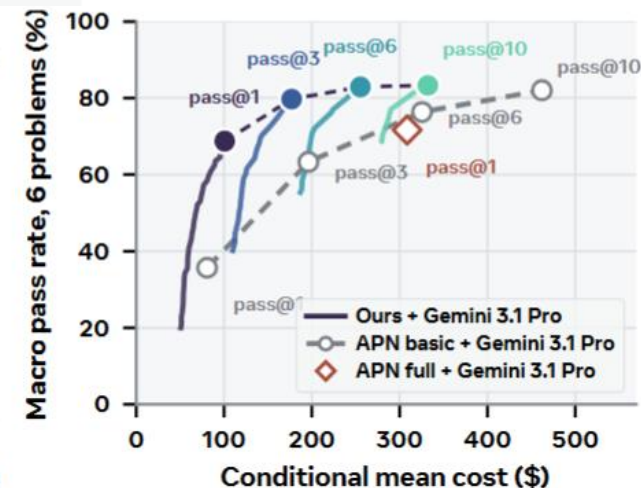
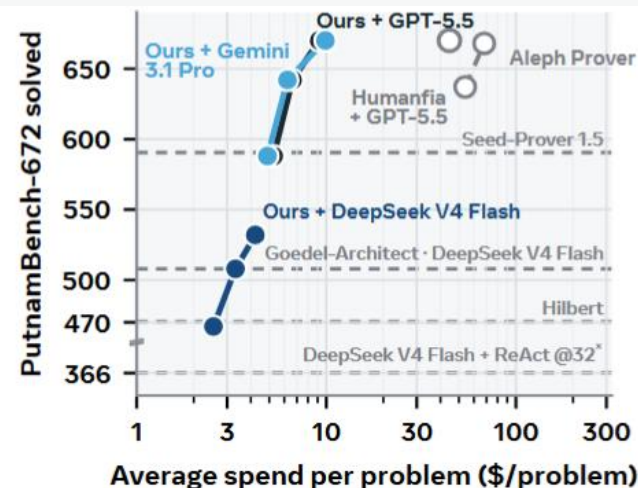
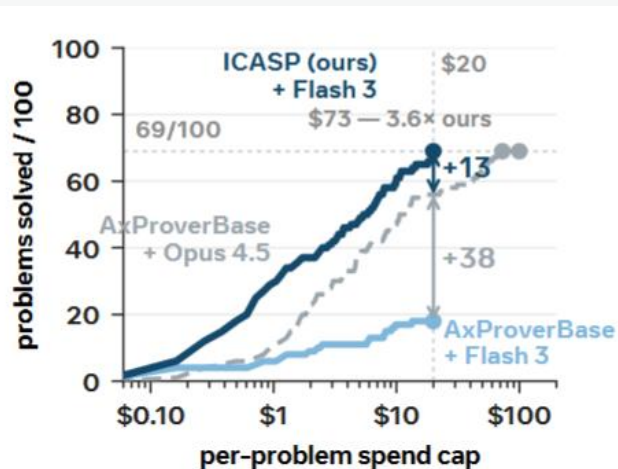
Harnessing Simple Test-Time Scaling: A Taste of the Bitter Lesson in Automated Theorem Proving

Skander Moalla^{1,2,†}, Ahmad Rammal¹, Niket Patel^{1,3,†}, Rémi Munos¹, Julia Kempe^{1,3,†}

FAIR at Meta, ²EPFL, ³New York University



A good harness can achieve SOTA on math benchmarks, beating special purpose models like AlphaProof.



(a) Putnam100, one model in two harnesses

(b) PutnamBench-672

(c) Six open Erdős problems

AUTOMATIC THEOREM DISCOVERY

LLMS CAN NOW STATE, PROVE AND VERIFY MATHEMATICAL THEOREMS

→ CREATE NEW MATH ? SCIENCE?

“Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house.”

Science and Hypothesis, Poincaré (1905)



AUTOMATIC (INTERESTING) THEOREM DISCOVERY

START WITH KNOWN PREMISES

HAVE THE MODEL CONJECTURE AND PROVE
NEW THEOREMS

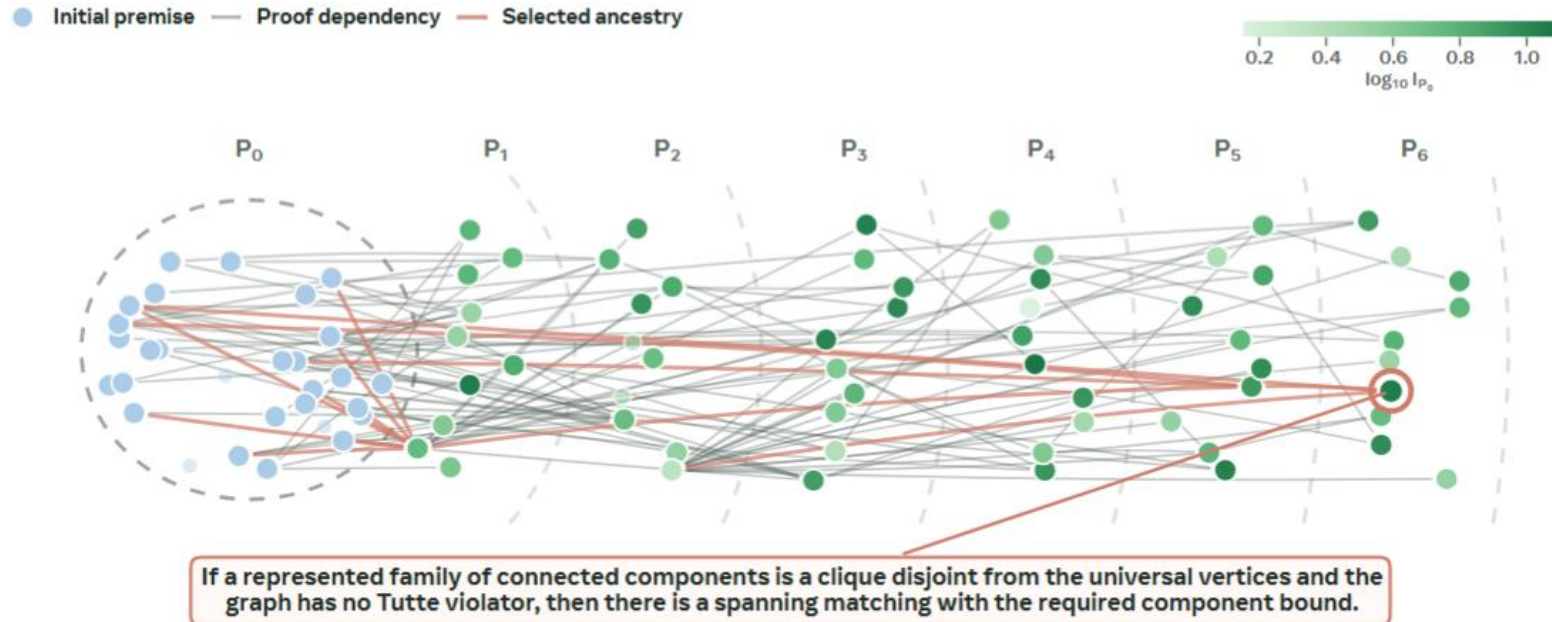
PRUNE FOR “INTERESTINGNESS” (A LEARNED
VALUE FUNCTION)

RESTART...

Learning to Discover Interesting Mathematics

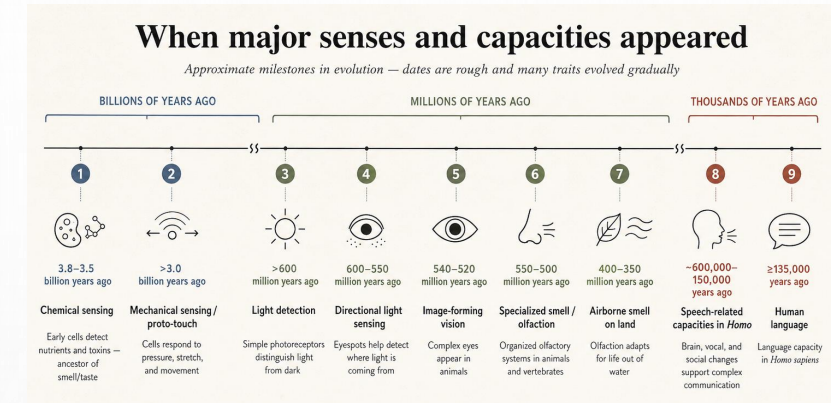
Niket Patel^{1,2*}, Ahmad Rammal^{1,3}, Amaury Hayat³, Remi Munos¹, Julia Kempe^{1,2}

¹FAIR @ Meta, ²New York University, ³CERMICS, ENPC, Institut Polytechnique de Paris



Lean is a theorem prover and programming language that enables correct, maintainable, and formally verified code

BACK TO MORAVEC'S PARADOX



“It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility.”

Hans Moravec (1988)

“The main lesson of thirty-five years of AI research is that the hard problems are easy and the easy problems are hard“

Stephen Pinker (1994)

GROUNDING IS KEY

LLMs are not grounded in the physical world

Physical Feedback Loop: observe → act → get feedback → update

Exploration: Babies drop things on purpose to practice new motor skills, learn about gravity, and test cause and effect in a tight signal-feedback loop

“RLVF (on LLMs) is like sucking bits of supervision through a straw.”

Andrej Karpathy, 2025

Bottlenecked by inefficiency compared to humans



LEARNING SEMANTICS FROM THE PHYSICAL WORLD

PLANNING IN LATENT SPACE

World Model Architectures

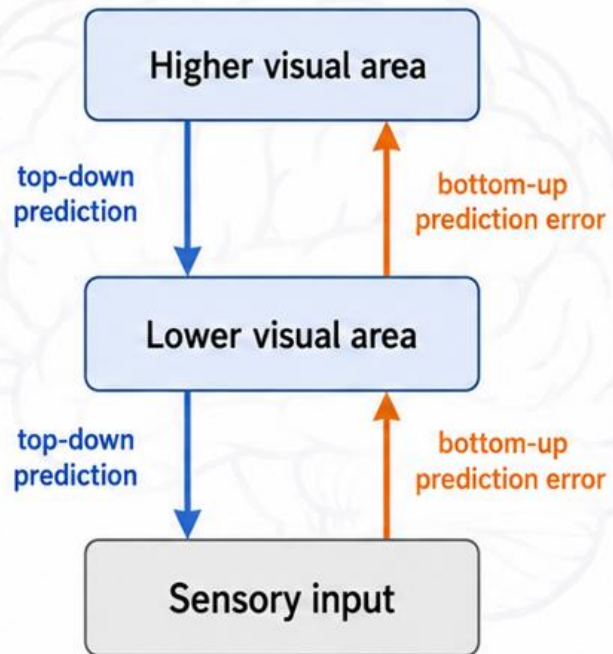
PREDICTIVE CODING IN THE BRAIN

- Cognitive learning theories suggest that humans **learn an internal model of the world** by integrating low-level sensory inputs to represent and predict future states
- This **world model shapes our perception** at any given moment
- The ability to **predict the effects of our actions on future states** of the world is also essential for **goal-oriented planning**

For reasoning and planning we do not need to reconstruct/render every pixel (or every token)!

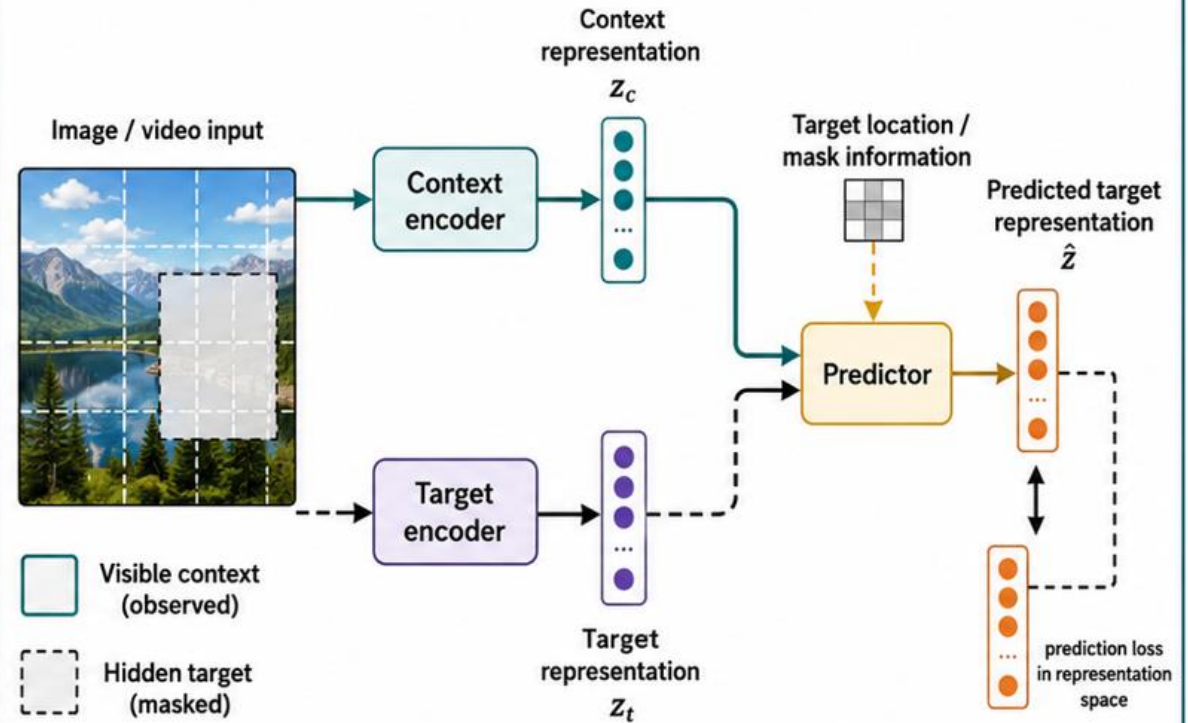
Predictive coding vs. JEPA

Predictive coding in the brain (after Rao & Ballard)



*Higher areas predict lower-level activity;
lower areas send residual error upward.*

JEPA (Joint-Embedding Predictive Architecture)



Predict latent representations of missing content, not raw pixels.

Both are predictive ideas. Predictive coding emphasizes top-down predictions and bottom-up errors in a hierarchy; JEPA predicts abstract embeddings of missing content.

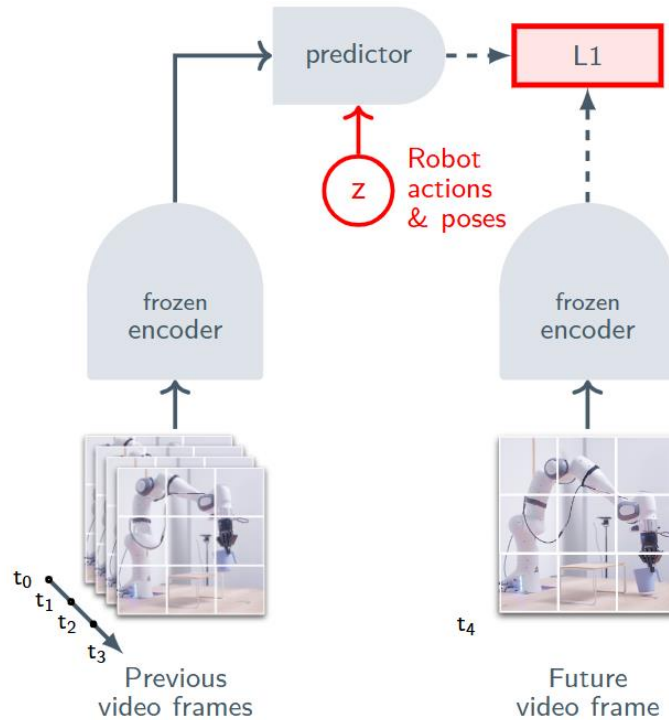
- “Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture”, Assran et al., CVPR 2023
- “Revisiting Feature Prediction for Learning Visual Representations from Video”, Bardes et al., ICLR 2025
- “V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning, Assran et al., arXiv 2506.09985
- “V-JEPA 2.1: Unlocking Dense Features in Video Self-Supervised Learning”, Mur-Labadia et al., arXiv 2603.14482

from Meta,
(not my works)

PLANNING IN LATENT SPACE

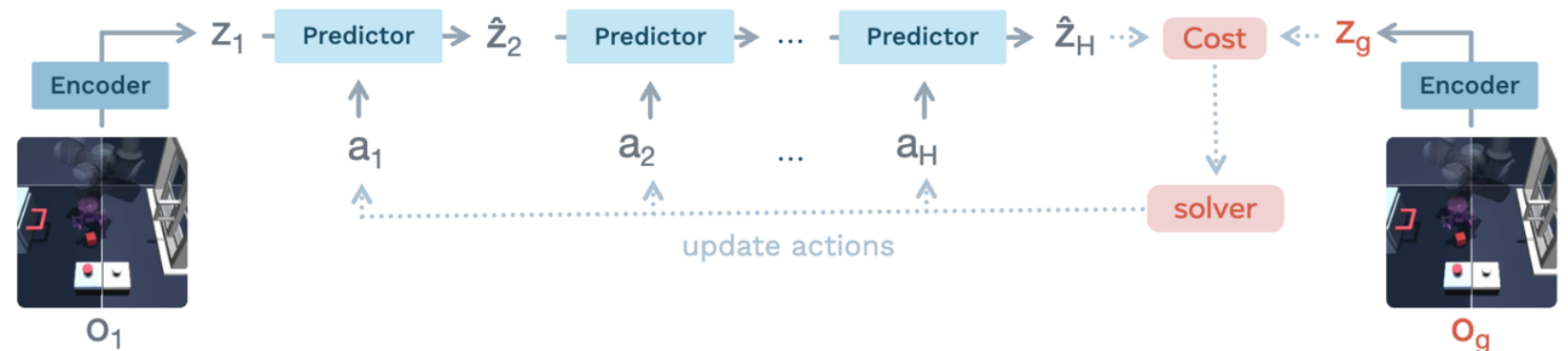
Joint Embedding Predictive Architecture (JEPA)

V-JEPA 2-AC

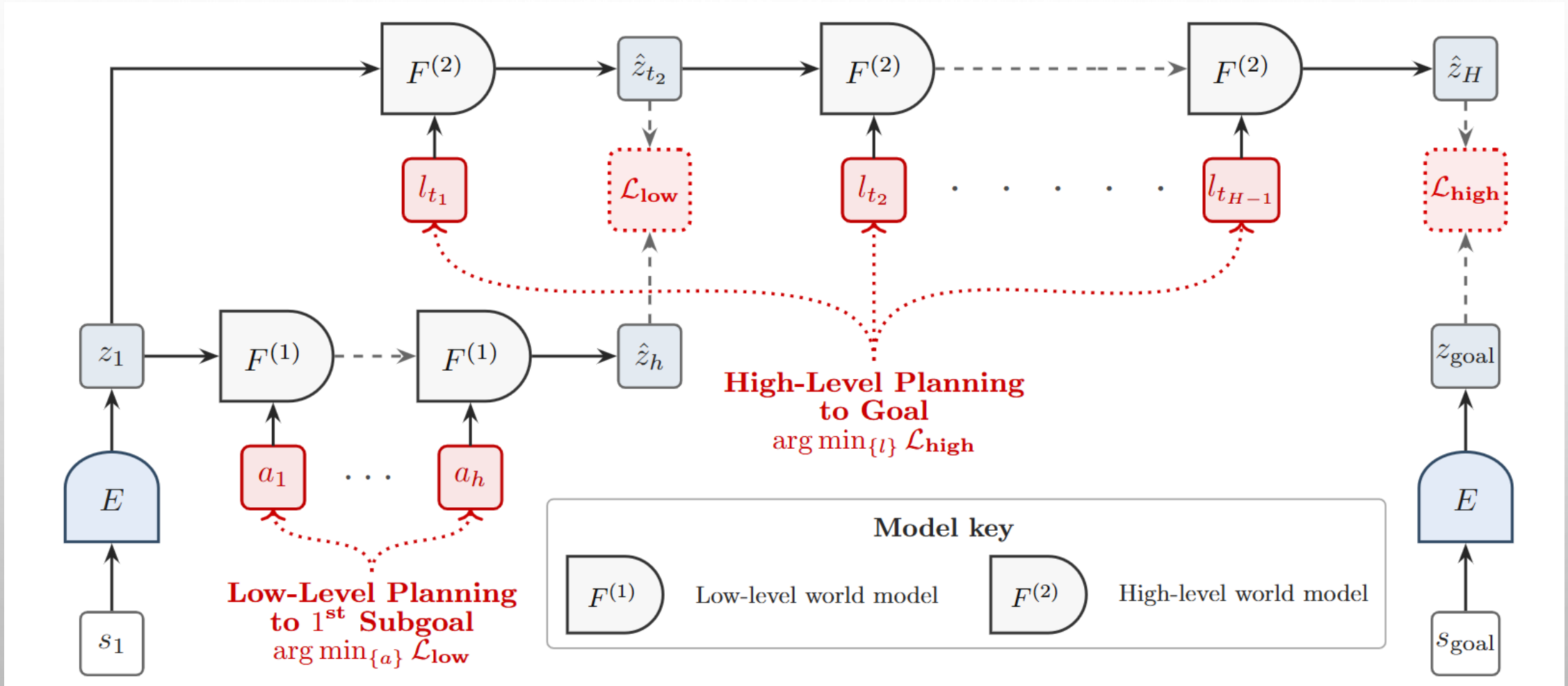


TRAIN A PREDICTOR IN LATENT SPACE

**FOR PLANNING (REASONING) SIMULATE AND OPTIMIZE
ACTIONS IN LATENT SPACE (WORLD MODEL)**



Advanced
Machine
Intelligence



THANKS !

