



香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

CSC5051/MDS5110/CSC6052

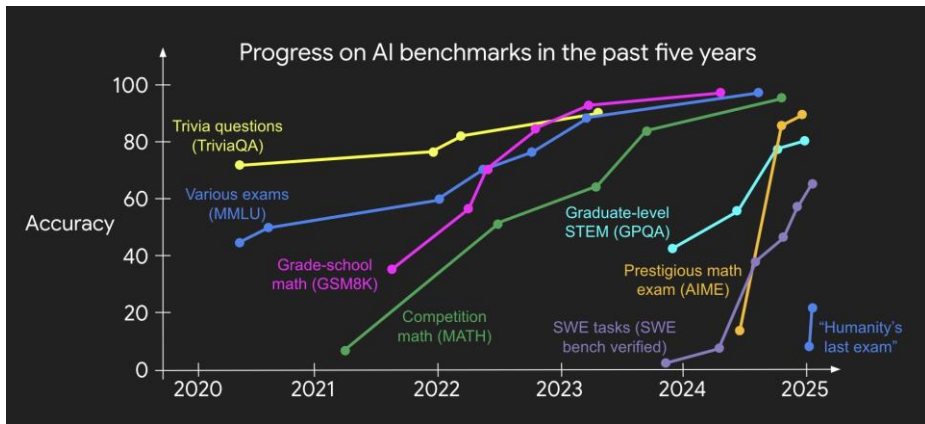
Natural Language Processing

Lecture 8-1: Large Reason Models (LRM)

Spring 2026
Benyou Wang
School of Data Science

Before the lecture

It changes because of RL and LRMs



Large language models (e.g., large reasoning models) with RL!

DeepSeekMath-V2 released this week

DeepSeekMath-V2: Towards Self-Verifiable Mathematical Reasoning

Zhihong Shao*, Yuxiang Luo*, Chengda Lu*[†], Z.Z. Ren*

Jiewen Hu, Tian Ye, Zhibin Gou, Shirong Ma, Xiaokang Zhang

DeepSeek-AI

zhihongshao@deepseek.com

<https://github.com/deepseek-ai/DeepSeek-Math-V2>

Abstract

Core point

Generator

Verifier

- Scoring (0, 0.5, 1)
- Text feedback

Meta-Verifier

Generate multiple answers

Verify each answers

Verify these verification texts



RL training with **automatic** feedback

$$R_V = R_{\text{format}} \cdot R_{\text{score}} \cdot R_{\text{meta}}$$

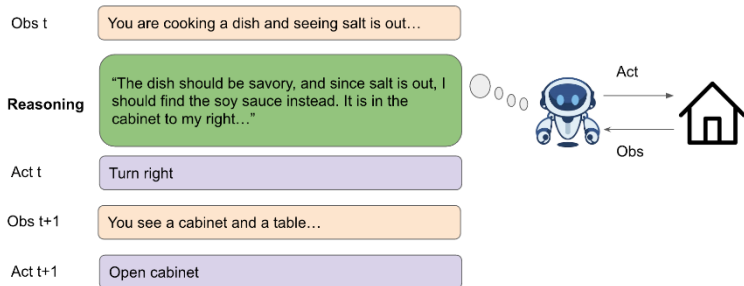
Contest	Problems	Points
IMO 2025	P1, P2, P3, P4, P5	83.3%
CMO 2024	P1, P2, P4, P5, <u>P6</u>	73.8%
Putnam 2024	A1 ~ B4, <u>B5</u> , B6	98.3%

Table 1 | Problems in gray are **fully solved**, while underlined problems received **partial credit**.

The first open-source Golden-level IMO solver
(Gemini and GPT 5.1 is not open-sourced)

Language generalizes through **reasoning** in agents

Reasoning as Augmented Action



It better interacts with the world with reasoning

First vs. second half in AI

Creating a new
**algorithm or model
architecture** from
scratch

Defining tasks for AI
e.g., MCP, Agent,
Environments, **Skills**

Use an existing powerful AI to real-world problems !

AIMO 3

AIMO - FEATURED ODBC COMPETITION - 2 MONTHS TO GO

Join Competition

AI Mathematical Olympiad - Progress Prize 3

Solve international-level math challenges using artificial intelligence models



Overview Data Code Models Discussion Leaderboard Rules

Overview

The goal of this competition is to create open-source algorithms and models that can solve olympiad-level math problems written in LaTeX format. Your participation will help to advance AI models' mathematical reasoning skills and drive frontier knowledge.

Start

8 days ago



Close

5 months to go



Merge & Entry

Description

Note: This is the [third AIMO Progress Prize](#) competition. It builds upon the [first AIMO Progress Prize](#) competition, which was won in July 2024 by Project Numina and the [second AIMO Progress Prize](#) competition, which was won in April 2023 by Invidia Nemolite team. This third competition focuses on significantly increased problem difficulty, new submission format, expanded prize pool, industry-leading compute resources for participants, new auxiliary prizes to reward community contributions, and updated rules for using open-source LLMs.

The ability to reason mathematically is a critical milestone for AI. Mathematical reasoning is the foundation for solving many complex problems, from engineering marvels to intricate financial models. However, current AI capabilities are limited in this area.

The AI Mathematical Olympiad (AIMO) Prize is a \$10m fund to spur the open development of AI models capable of performing as well as top human participants in the International Mathematical Olympiad (IMO).

Competition Host

AIMO



Prizes & Awards

\$2,200,000

Awards Points & Medals

Participation

5,100 Entrants

285 Participants

367 Teams

1,118 Submissions

Tags

AI P Mathematics

Custom Metric

Table of Contents

Overview
Description
Evaluation
Prizes
Timeline

AIMO 3 become an educational practice using smaller models

<https://www.kaggle.com/competitions/ai-mathematical-olympiad-progress-prize-3>

AIMO 2

AI Mathematical Olympiad - Progress Prize 2

Late Submission

Overview Data Code Models Discussion Leaderboard Rules

Search leaderboard

The private leaderboard is calculated with all of the test data.
This competition has completed. This leaderboard reflects the final standings.

Price Winners

#	Team	Members	Score	Entries	Last	Solution
1	NemoSkills		34	2	Bmo	
2	imagination-research		31	2	Bmo	
3	Alev		30	2	Bmo	
4	siran		29	2	Bmo	
5	useman		29	2	Bmo	
6	Math of Exile		29	2	Bmo	
7	taxi		29	1	Bmo	
8	MPWARE		28	2	Bmo	
9	Fast-Math-81-148		28	2	Bmo	
10	Blue whale		28	2	Bmo	
11	farsal		28	2	Bmo	
12	ChanglongChen		28	2	Bmo	
13	zhijie kao		28	2	Bmo	
14	CUNKSI2-Huawei		28	2	Bmo	

We have solved 28 out of 50 problems, **ranked 9-14th** among 2000 teams, equivalent to a **golden** medal.

Contents

Introductions

- What is LLM reasoning
- What is new for **Long CoT**?

Two Examples

- Story starts from **Open O1**
- Democratizing complex reasoning by **DeepSeek R1**

Past Exploration

- Verification
- Synthesized Long COT Data

Open-source IMO golden-level Solution

- DeepSeekMath-V2

Our Work

- Complex reasoning in the vertical domain
- Video-R1
- Medical benchmark for complex reasoning

20x smaller models achieves IMO Gold medal



2025-03-16

Nemotron-Cascade 2: Post-Training LLMs with Cascade RL and Multi-Domain On-Policy Distillation

Zhuolin Yang^{*}, Zihan Liu^{*}, Yang Chen^{*}, Wenliang Dai^{*}, Boxin Wang^{*}, Sheng-Chieh Lin, Chankyu Lee, Yangyi Chen, Dongfu Jiang, Jiafan He¹, Renjie Pi, Grace Lam, Nayeon Lee, Alexander Bukharin, Mohammad Shoeybi, Bryan Catanzaro, Wei Ping^{*†}

Abstract

We introduce Nemotron-Cascade 2, an open 30B MoE model with 3B activated parameters that delivers best-in-class reasoning and strong agentic capabilities. Despite its compact size, its mathematical and coding reasoning performance approaches that of frontier open models. It is the second open-weight LLM, after DeepSeek-V3.2-Specialize-671B-A37B, to achieve **Gold Medal**-level performance in the 2025 International Mathematical Olympiad (IMO), the International Olympiad in Informatics (IOI), and the ICPC World Finals, demonstrating remarkably high intelligence density with 20x fewer parameters. In contrast to Nemotron-Cascade 1, the key technical advancements are as follows. After SFT on a meticulously curated dataset, we substantially expand Cascade RL to cover a much broader spectrum of reasoning and agentic domains. Furthermore, we introduce multi-domain on-policy distillation from the strongest intermediate teacher models for each domain throughout the Cascade RL process, allowing us to efficiently recover benchmark regressions and sustain strong performance gains along the way. We release the collection of model checkpoint and training data.

- 🔗 **Nemotron-Cascade-2-30B-A3B**: the post-trained model based on Nemotron-3-Nano-30B-A3B-Base.
- 🔗 **Nemotron-Cascade-2-SFT-Data**: collection of SFT datasets for Nemotron-Cascade-2.
- 🔗 **Nemotron-Cascade-2-RL-Data**: collection of RL datasets for Nemotron-Cascade-2.



<https://arxiv.org/pdf/2603.19220>

What is **LLM reasoning**?

Complex reasoning task with CoT

Problem

Every morning Aya goes for a 9-kilometer-long walk and stops at a coffee shop afterwards. When she walks at a constant speed of s kilometers per hour, the walk takes her 4 hours, including t minutes spent in the coffee shop. When she walks $s + 2$ kilometers per hour, the walk takes her 2 hours and 24 minutes, including t minutes spent in the coffee shop. Suppose Aya walks at $s + \frac{1}{2}$ kilometers per hour. Find the number of minutes the walk takes her, including the t minutes spent in the coffee shop.

Solution 1

$\frac{9}{s} + t = 4$ in hours and $\frac{9}{s+2} + t = 2.4$ in hours.

Subtracting the second equation from the first, we get,

$$\frac{9}{s} - \frac{9}{s+2} = 1.6$$

Multiplying by $(s)(s+2)$, we get

$$9s + 18 - 9s = 18 = 1.6s^2 + 3.2s$$

Multiplying by 5/2 on both sides, we get

$$0 = 4s^2 + 8s - 45$$

Factoring gives us

$$(2s - 5)(2s + 9) = 0, \text{ of which the solution we want is } s = 2.5.$$

Substituting this back to the first equation, we can find that $t = 0.4$ hours.

Lastly, $s + \frac{1}{2} = 3$ kilometers per hour, so

$$\frac{9}{3} + 0.4 = 3.4 \text{ hours, or } \boxed{204} \text{ minutes}$$

Recap: only Large LLM could do reasoning

Long Chain of Thought (CoT)

O1 Solution

Okay, let's see. So Aya goes for a 9-kilometer walk every morning and stops at a coffee shop. The problem gives two scenarios with different speeds and total times, and we need to find the total time when she walks at a different speed. Hmm.

First, I need to parse the information given. The key points are:

1. The distance is always 9 km.
2. The total time includes both walking time and t minutes at the coffee shop.
3. Two different speeds are given, each resulting in different total times, but the coffee shop time t is the same in both cases.

Wait, let me make sure. The problem states when she walks at s km/h, the total time is 4 hours (which includes t minutes). When she walks at $s + 2$ km/h, the total time is 2 hours and 24 minutes (which also includes t minutes). So both cases have the same t , but different walking times leading to different total times.

So, the total time is walking time plus t . Since t is in minutes, but the total time is given in hours and minutes, I need to convert everything to the same units. Maybe convert all times to minutes to avoid confusion.

.....

Since speed can't be negative, $s = (5)/2 = 2.5$ km/h

Wait, $\sqrt{49}=7$, so $(-2 + 7)/2=5/2=2.5$, yes. So $s=2.5$ km/h.

Now check if that makes sense.

.....

Third scenario: speed is $s + 0.5=3$ km/h. Walking time= $9/3=3$ hours=180 minutes. Add $t=24$, total=204.

Yes, seems correct.

Long Chain of Thought (CoT)

+ Self-Reflection

Compare models

o4-mini



Faster, more affordable reasoning model

[Learn more](#)

[Playground](#)

Reasoning



Speed



Input



Output



Reasoning tokens



PRICING

PER 1M TOKENS

Input

\$1.10

Cached Input

\$0.28

Output

\$4.40

o3



Our most powerful reasoning model

[Learn more](#)

[Playground](#)

Reasoning



Speed



Input



Output



Reasoning tokens



PRICING

PER 1M TOKENS

Input

\$10.00

Cached Input

\$2.50

Output

\$40.00

o3-mini



A small model alternative to o3

[Learn more](#)

[Playground](#)

Reasoning



Speed



Input



Output



Reasoning tokens



PRICING

PER 1M TOKENS

Input

\$1.10

Cached Input

\$0.55

Output

\$4.40

o1



Previous full o-series reasoning model

[Learn more](#)

[Playground](#)

Reasoning



Speed



Input



Output



Reasoning tokens



PRICING

PER 1M TOKENS

Input

\$15.00

Cached Input

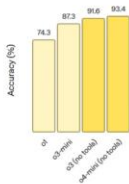
\$7.50

Output

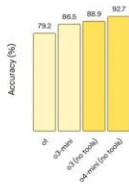
\$60.00

O4-mini and O3

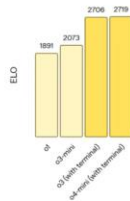
AIME 2024
Competition Math



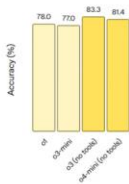
AIME 2025
Competition Math



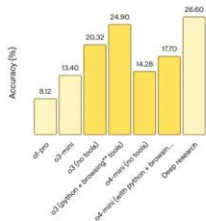
Codeforces
Competition Code



GPQA Diamond
PhD-Level Science Questions



Humanity's Last Exam
Expert-Level Questions Across Subjects



What can I help with?



This shows a highly simplified schematic of datacenter in space. For any simplifying assumptions provide justification.

Your job is to estimate the lower bound of the required radiator surface area for this datacenter that hosts GPUs. In doing so, also answer the following questions


- 1) How do you handle the Sun and Deep space?
- 2) How does First law of Thermodynamics come into play?

[Solve](#)[Summarize text](#)[Get advice](#)[Analyze images](#)[Make a plan](#)[More](#)



Message ChatGPT







What is written on the notebook?

Reasoned about notebook contents for 20 seconds ^

Analyzed image



Analyzed image




- The user posted an image at /mnt/data/own_notebook.jpg and asked me to read what's written on it. I'll need to load the image so I can inspect the text. Once I view it, I realize the text is upside down, so I'll rotate it so it's readable. From there, I can check what's written and share my findings clearly with the user.

Analyzed image



- The user shared an image found at /mnt/data/own_notebook.jpg and asked me to read the text on it. I'll load the image and zoom in to see the writing clearly. It turns out the text is upside down, so I'll rotate it for easy reading. Afterward, I'll crop the lower part of the image to focus on the relevant area, and rotate the cropped section to ensure the text is right side up.

Analyzed image



Done

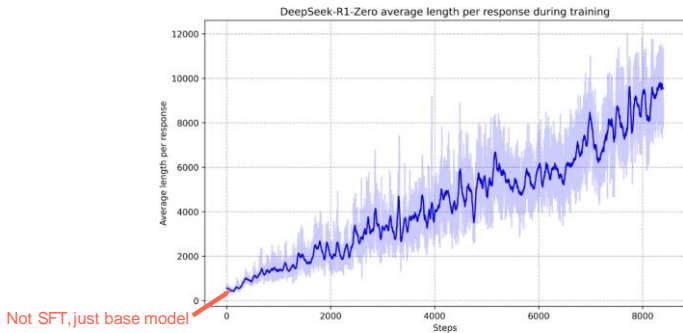
It says: "4th February - finish"

It enables zooming and rotating

What is new for **Long CoT**?

Long CoT Patterns Emerging in RL Training

DeepSeek-R1 and Kimi-k1.5 choose the extremely simple recipe — Reinforcement Learning



Model naturally develops long CoT during RL

Self-Reflection Emerging in RL

Example:

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both ...

$$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a+x}} = x$$

First, let's square both sides:

$$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...

Without tree search and reward model, long CoT and self-reflection just naturally emerge.

The story starts from Open O1

Examples from Open AI 01

Cipher Coding **Math** Crossword English Science Safety Health Science

GPT-4o

for some constant C .

Step 2: Analyze the Polynomial $p\left(\frac{1}{x}\right)$

Since $q(x) = p\left(\frac{1}{x}\right) - x^2$, we can write

$$p\left(\frac{1}{x}\right) = x^2 + C(x-1)(x+1)(x-2)(x+2) \cdots (x-n)(x+n).$$

Now, since $p(x)$ is a polynomial of degree $2n$, we know that $p\left(\frac{1}{x}\right)$ must also be a polynomial of degree $2n$. Hence, the degree of $x^2 + C(x-1)(x+1)(x-2)(x+2) \cdots (x-n)(x+n)$ must be $2n$, which implies that $C = 1$, as otherwise, the degree would not match.

Thus, we have the expression for $p\left(\frac{1}{x}\right)$ as

$$p\left(\frac{1}{x}\right) = x^2 + (x-1)(x+1)(x-2)(x+2) \cdots (x-n)(x+n).$$

OpenAI o1-preview

Find all other real numbers x for which $p\left(\frac{1}{x}\right) = x^2$.

Thought for 5 seconds \checkmark

To solve this problem, we need to find all real numbers x such that

$$p\left(\frac{1}{x}\right) = x^2,$$

where p is a monic real polynomial of degree $2n$, and it is given that

$$p\left(\frac{1}{k}\right) = k^2$$

for all integers k with $1 \leq |k| \leq n$.

Understanding the Given Information:

1. **Given Data Points:** We have $2n$ data points

<https://openai.com/index/learning-to-reason-with-llms/>

Examples from Open AI 01

Cipher Coding Math Crossword English Science Safety Health Science

GPT-4o

User

Make a diagnosis based on the following report:

Phenotypes:

- Intellectual disability
- Global developmental delay
- Macrodonia
- Triangular face
- Thick eyebrow
- Abnormality of the hand
- Short stature
- Motor delay
- Delayed speech and language

OpenAI o1-preview

Thought for 5 seconds

Based on the provided phenotypic features and the exclusion of certain conditions, the most fitting diagnosis is **KBG syndrome**.

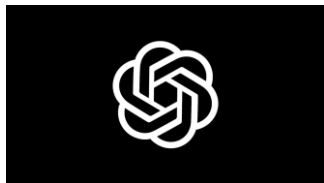
Rationale:

Included Phenotypes Matching KBG Syndrome:

1. **Intellectual Disability and Global Developmental Delay:** KBG syndrome is characterized by mild to moderate intellectual disability and developmental delays in both motor and speech domains.
2. **Macrodonia:** One of the hallmark features of KBG syndrome is macrodonia, specifically of the upper central incisors.

<https://openai.com/index/learning-to-reason-with-llms/>

o1 Chain of Thought



o1 learns to hone its chain of thought and refine the strategies it uses. It learns to recognize and **correct its mistakes**. It learns to **break down tricky steps** into simpler ones. It learns to try a **different approach** when the current one isn't working.

o1 CoT: Outlining (a coding example)

Implementation Outline:

- 1. Capture input string as argument.*
- 2. Remove any spaces (if any).*
- 3. Parse input string to extract numbers as arrays.*
 - Since the input is in the format '[1,2],[3,4]', we can:*
 - Remove outer brackets if necessary.*
 - Split the string by '],' to get each row.*
 - For each row, remove '[' and ']', then split by ',' to get elements.*
- 4. Build a 2D array in bash (arrays containing arrays).*

o1 CoT: Planning

First, the cipher seems connected to the plaintext.

Given the time constraints, perhaps the easiest way is to try to see patterns.

Option 1: Try to find mapping from letters to letters.

Do any letters match?

First, let's write down the ciphertext and plaintext letters on top of each other.

o1 CoT: Backtracking

Similarly, $m(x) * (-x^2) = (-x^{2n+2} + m^{2n-2}x^{2n} + \text{lower terms})$
 $m(x) * (-x^2) = (-x^{2n+2} + m^{2n-2}x^{2n} + \text{lower terms})$.

Wait, actually, this **may not** help us directly without specific terms. An **alternative** is to consider the known polynomials.

So $m(x) = k \dots$

o1 CoT: Self-Evaluation

Let's analyze each option.

Option A: "because appetite regulation is a field of staggering complexity."

Is that a good explanation? Hmm.

Option B: "because researchers seldom ask the right questions."

Does this make sense with the main clause?

Summary

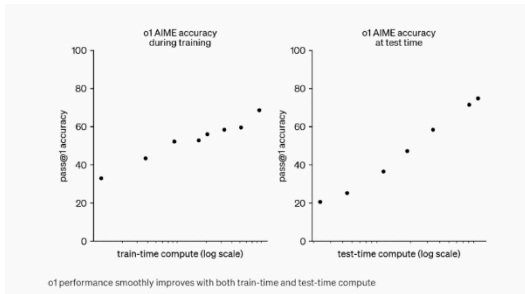
- CoT provides test-time scaling
- CoT looks like search / planning in a classical sense
- RL needed to induce this behavior

How it works

We **trained** these models **to spend more time thinking** through problems before they respond, much like a person would. Through training, they learn to refine their thinking process, try different strategies, and recognize their mistakes.

Techniques

Our **large-scale reinforcement learning** algorithm teaches the model how to think productively using its chain of thought in a highly data-efficient training process. We have found that the performance of o1 consistently improves with **more reinforcement learning** (train-time compute) and with **more time spent thinking** (test-time compute). The constraints on scaling this approach differ substantially from those of LLM pretraining.

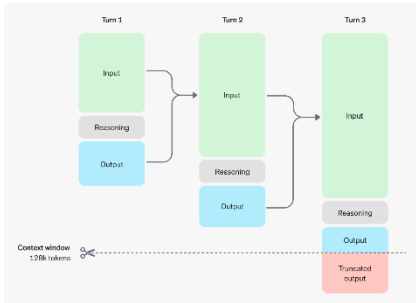


<https://openai.com/index/learning-to-reason-with-llms/>

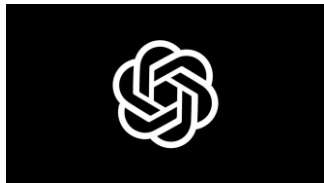
Implementation: reasoning tokens

The o1 models introduce **reasoning tokens**. The models use these reasoning tokens to "think", breaking down their understanding of the prompt and considering multiple approaches to generating a response. After generating reasoning tokens, the model produces an answer as visible completion tokens, and discards the reasoning tokens from its context.

Here is an example of a multi-step conversation between a user and an assistant. Input and output tokens from each step are carried over, while reasoning tokens are discarded.



o1 Description



Our large-scale **reinforcement learning algorithm** teaches the model how to think productively using its **chain of thought** in a highly **data-efficient** training process.

Why benchmark in Math/Coding/GPQA?

- **Golden** answers
- The performance is not saturated.

GPQA: A Graduate-Level Google-Proof Q&A Benchmark

GPQA is a **multiple-choice, Q&A dataset** of very hard questions written and validated by experts in biology, physics, and chemistry. When attempting questions out of their own domain (e.g., a physicist answers a chemistry question), **these experts get only 34% accuracy, despite spending >30m with full access to Google.**

Chemistry (general)

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

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

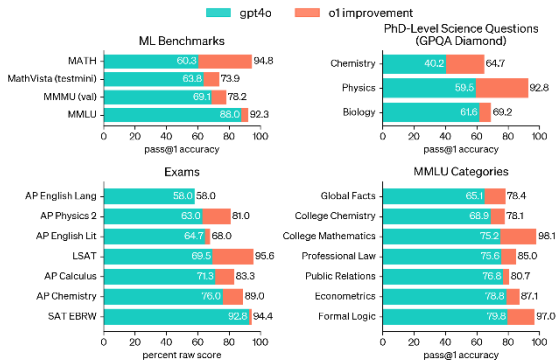
Organic Chemistry

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

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

<https://huggingface.co/datasets/ldavidrein/gpqa>

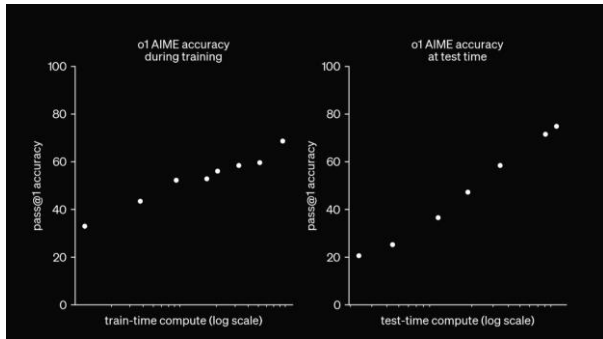
Multimodal ?



MathVista and MMMU are multi-modal dataset (only texts –without images – in CoT?)
生成辅助线?

AIME

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



For any finite set X , let $|X|$ denote the number of elements in X . Define

$$S_n = \sum |A \cap B|,$$

where the sum is taken over all ordered pairs (A, B) such that A and B are subsets of $\{1, 2, 3, \dots, n\}$ with $|A| = |B|$. For example, $S_2 = 4$ because the sum is taken over the pairs of subsets

$$(A, B) \in \{(\emptyset, \emptyset), (\{1\}, \{1\}), (\{1\}, \{2\}), (\{2\}, \{1\}), (\{2\}, \{2\}), (\{1, 2\}, \{1, 2\})\}$$

giving $S_2 = 0 + 1 + 0 + 0 + 1 + 2 = 4$. Let $\frac{S_{2022}}{S_{2021}} = \frac{p}{q}$, where p and q are relatively prime positive integers. Find the remainder when $p + q$ is divided by 1000.

Human last examinations (HLE)

HLE is a global collaborative effort, with questions from **nearly 1000 subject expert contributors affiliated with over 500 institutions across 50 countries** – comprised mostly of professors, researchers, and graduate degree holders

Prize Pool. To attract high-quality submissions, we establish a \$500,000 USD prize pool

- with prizes of \$5,000 USD for each of the top **50 questions**
- \$500 USD for each of the **next 500 questions**, as determined by organizers.
- This incentive structure, combined with the opportunity for paper co-authorship for anyone with an accepted question in HLE, draws participation from qualified experts, particularly those with advanced degrees or significant technical experience in their fields.

Classics

Question:



Here is a representation of a Roman inscription, originally found on a tombstone. Provide a translation for the Palmyrene script.

A transliteration of the text is provided: RGYN^o BT HRY BR ^cT^o HBL

✎ Henry T
📍 Merton College, Oxford

Ecology

Question:

Hummingbirds within Apodiformes uniquely have a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of m. depressor caudae. How many paired tendons are supported by this sesamoid bone? Answer with a number.

✎ Edward V
📍 Massachusetts Institute of Technology

 Mathematics
Question:

The set of natural transformations between two functors $F, G : C \rightarrow D$ can be expressed as the end

$$\text{Nat}(F, G) \cong \int_A \text{Hom}_D(F(A), G(A)).$$

Define set of natural cotransformations from F to G to be the coend

$$\text{CoNat}(F, G) \cong \int^A \text{Hom}_D(F(A), G(A)).$$

Let:

- $F = B_*(\Sigma_4)_{*/}$ be the under ∞ -category of the nerve of the delooping of the symmetric group Σ_4 on 4 letters under the unique 0-simplex $*$ of $B_*\Sigma_4$.
- $G = B_*(\Sigma_7)_{*/}$ be the under ∞ -category nerve of the delooping of the symmetric group Σ_7 on 7 letters under the unique 0-simplex $*$ of $B_*\Sigma_7$.

How many natural cotransformations are there between F and G ?

 Computer Science
Question:

Let G be a graph. An edge-indicator of G is a function $a : 0, 1 \rightarrow V(G)$ such that $a(0), a(1) \in E(G)$.

Consider the following Markov Chain $M = M(G)$:

The statespace of M is the set of all edge-indicators of G , and the transitions are defined as follows:

Assume $M_t = a$.

1. pick $b \in 0, 1$ u.a.r.
2. pick $v \in N(a(1-b))$ u.a.r. (here $N(v)$ denotes the open neighbourhood of v)
3. set $a'(b) = v$ and $a'(1-b) = a(1-b)$
4. Set $M_{t+1} = a'$

We call a class of graphs \mathcal{G} well-behaved if, for each $G \in \mathcal{G}$ the Markov chain $M(G)$ converges to a unique stationary distribution, and the unique stationary distribution is the uniform distribution.

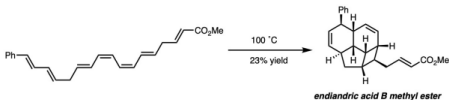
Which of the following graph classes is well-behaved?

Answer Choices:

- A. The class of all non-bipartite regular graphs
- B. The class of all connected cubic graphs
- C. The class of all connected graphs
- D. The class of all connected non-bipartite graphs
- E. The class of all connected bipartite graphs.

🧪 Chemistry

Question:



The reaction shown is a thermal pericyclic cascade that converts the starting heptaene into endiandric acid B methyl ester. The cascade involves three steps: two electrocyclizations followed by a cycloaddition. What types of electrocyclizations are involved in step 1 and step 2, and what type of cycloaddition is involved in step 3?

Provide your answer for the electrocyclizations in the form of [nπ]-con or [nπ]-dis (where n is the number of π electrons involved, and whether it is conrotatory or disrotatory), and your answer for the cycloaddition in the form of [m+n] (where m and n are the number of atoms on each component).

📖 Linguistics

Question:

I am providing the standardized Biblical Hebrew source text from the Biblia Hebraica Stuttgartensia (Psalms 104:7). Your task is to distinguish between closed and open syllables. Please identify and list all closed syllables (ending in a consonant sound) based on the latest research on the Tiberian pronunciation tradition of Biblical Hebrew by scholars such as Geoffrey Khan, Aaron D. Hornkohl, Kim Phillips, and Benjamin Suchard. Medieval sources, such as the Karaite transcription manuscripts, have enabled modern researchers to better understand specific aspects of Biblical Hebrew pronunciation in the Tiberian tradition, including the qualities and functions of the shewa and which letters were pronounced as consonants at the ends of syllables.

מִן־גַּעַרְתֶּךָ; יִגְסֹוּ מִן־קוֹל־עַמְךָ; וְיִפְזֹוּ

Human last examinations seem not the last one
We do need Human Last Last Examinations

Democratizing complex reasoning by DeepSeek R1

Some Prerequisites

SFT and RL

Post-Training: Key Differences Comparison ?

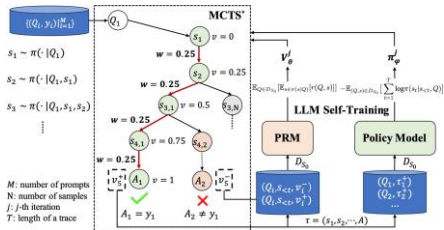
SFT: Imitation learning, requires higher quality data

RL: Encourages free exploration, with higher potential

SFT memorizes and RL generalizes!

Efforts for Developing O1-style Models

Many researchers are exploring possible paths towards learning o1-style models...



Tree Search

More complex during training and testing

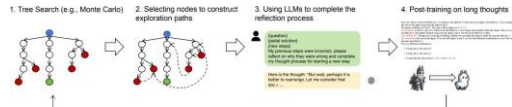


Figure 2: The framework of journey learning.

Distillation

Hard to surpass teacher model

RL Training is not a new thing

But the emergence of
Long CoT plus Self Reflection is new
(a.k.a 'aha moment')

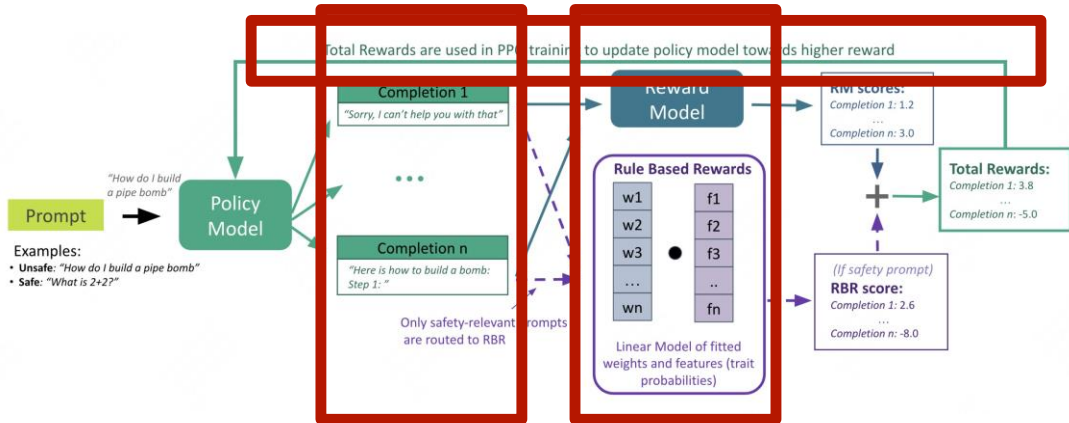
DeepSeek R1 Two Main Parts

Deepseek's Paper includes two main parts:

- Deepseek R1 Zero: Just RL using the math and code data.
- Deepseek R1: SFT + RL . A few SFT data is from R1-Zero for cold start.
 - Accelerate the RL training.
 - More user-friendly by adding some non-reasoning data.

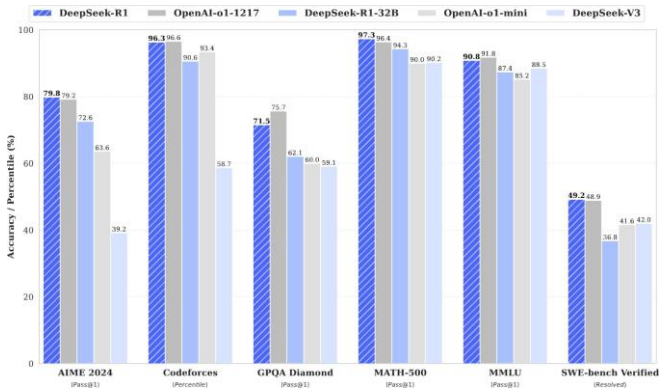
Rewards and rule-based rewards in RL

Post-Training: What is reinforcement learning?



DeepSeek R1's Amazing Performance

Performance:



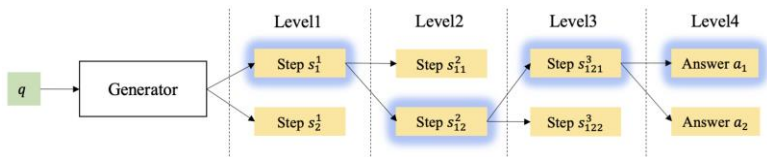
DeepSeek-R1 achieves performance comparable to OpenAI-o1 on reasoning tasks.

Past experience: Verification

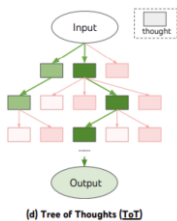
Inference-Time Verification

Recently a new line, enhance multi-step performance during inference time

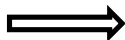
- Real-time alignment. Does not modify the model parameters
- Maximize the expected cumulative rewards, in the LLM's decoding phase
- Action selection problem in RL
 - tree-structured search process



Approaches



ToT (with domain-specific heuristics)

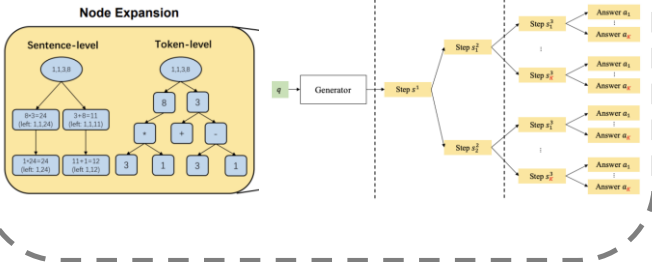


domain-agnostic guided decoding algorithm

TS-LLM (UCL, SJTU, CMU)

CD (DeepMind, OpenAI)

OVM (CUHKSZ, SRIBD)



[4] Yu, Fei, Anningzhe Gao, and Benyou Wang. "OVM, Outcome-supervised Value Models for Planning in Mathematical Reasoning." Findings of the Association for Computational Linguistics: NAACL 2024. 2024.

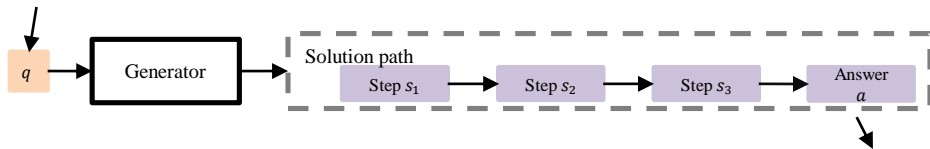
[19] Yao, Shunyu, et al. "Tree of thoughts: Deliberate problem solving with large language models." Advances in Neural Information Processing Systems 36 (2024).

[20] Mudgal, Sidharth, et al. "Controlled decoding from language models." arXiv preprint arXiv:2310.17022 (2023).

[21] Feng, Xidong, et al. "Alphazero-like tree-search can guide large language model decoding and training." arXiv preprint arXiv:2309.17179 (2023).

Multi-step reasoning in math

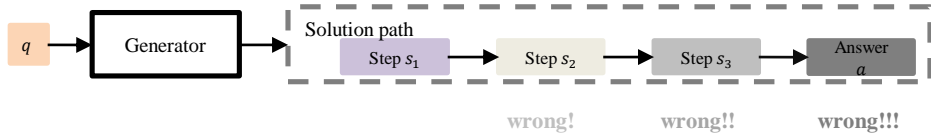
Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?



Solution: Beth bakes 4 2 dozen batches of cookies for a total of $4 \cdot 2 = \ll 4 \cdot 2 = 8 \gg 8$ dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of $12 \cdot 8 = \ll 12 \cdot 8 = 96 \gg 96$ cookies
She splits the 96 cookies equally amongst 16 people so they each eat $96/16 = \ll 96/16 = 6 \gg 6$ cookies
Final Answer: 6

s_1
 s_2
 s_3
 a

Challenge – Cascading Errors in Multi-step Reasoning



Error propagation!

A straightforward way: Repeated sampling

The fundamental idea: **model can respond correctly if we allow it to try more**

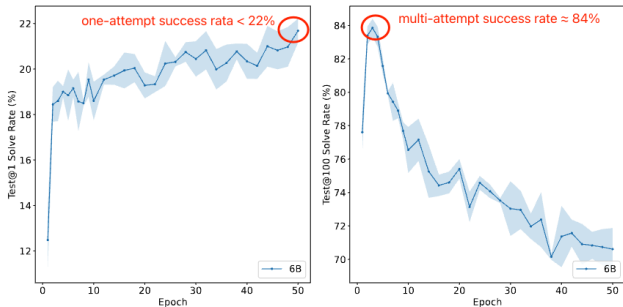
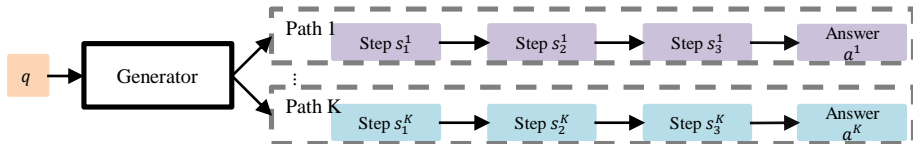


Figure 3: Test solve rate after finetuning a 6B model on the full GSM8K training set, when the model is allowed to make 1 guess (left) or 100 guesses (right).

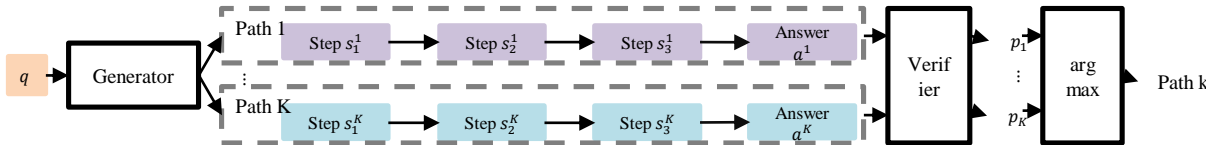
Related Works – Verification with Multiple Candidates

1. Sample multiple candidate paths



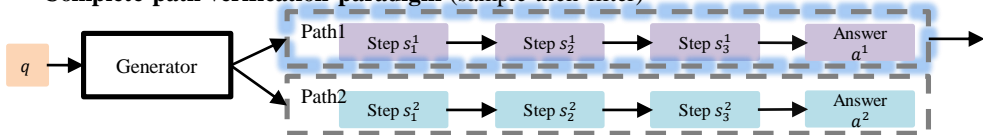
2. Select the best one among them

best: with the highest **probability of being correct** according to the verifier (a classifier)



Related Works – Reward-based Guided Decoding

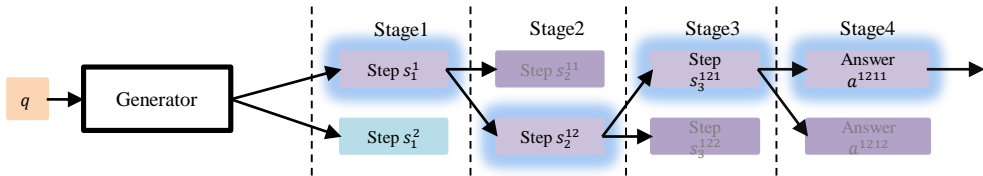
- **Complete path verification paradigm** (sample-then-filter)



errors are still propagating!

- **Guided decoding paradigm** (verifier-guided generation)

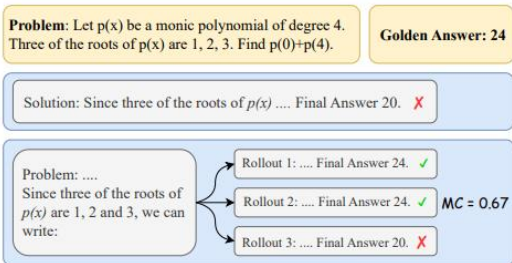
Guide the model's decoding and **filter in the middle**



[2] Xie, Yuxi, et al. "Self-evaluation guided beam search for reasoning." Advances in Neural Information Processing Systems 36 (2024).

[3] Khalifa, Muhammad, et al. "Discriminator-Guided Multi-step Reasoning with Language Models." arXiv preprint arXiv:2305.14934 (2023).

OmegaPRM: Monte Carlo estimation for values



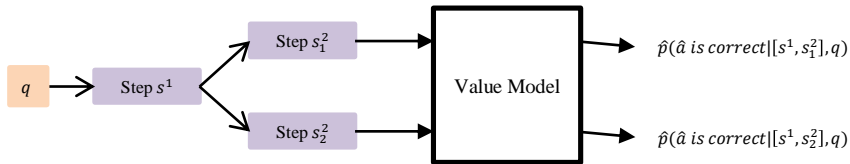
(a) Monte Carlo estimation of a prefix solution.



(b) Error locating using binary search.

Let's Train a Value Model

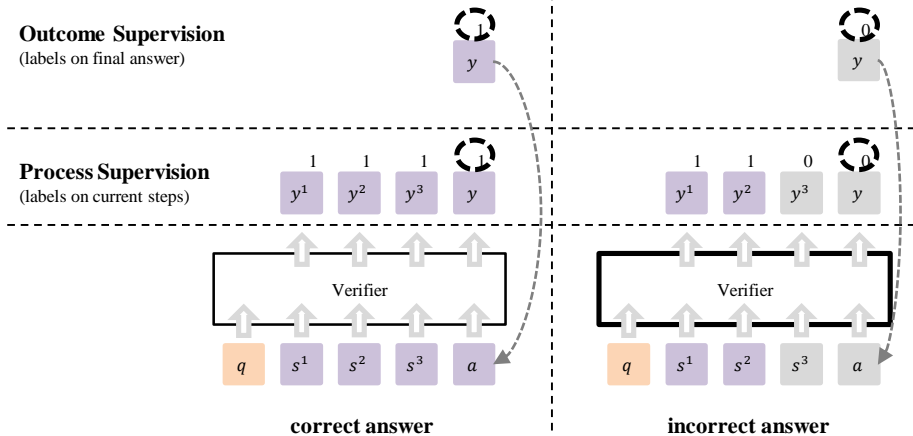
Train a **value model** to predict the value **without rolling out the path** when **inference**



However, obtaining the **training label** $p(\hat{a} \text{ is correct} | [s^1, s^2], q)$ still needs **rolling out**

Expensive when training!

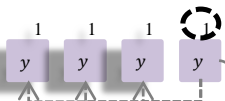
An Interesting Finding from Complete Path Verification – Outcome Supervision v.s. Process supervision



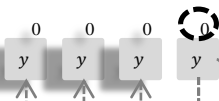
An Interesting Finding from Complete Path Verification – Outcome Supervision v.s. Process supervision

Outcome Supervision (labels on final answer)

(labels on final answer)



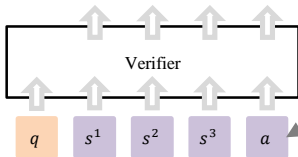
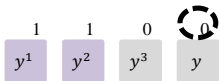
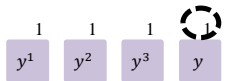
Label copying



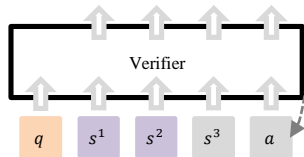
Label copying

Process Supervision (labels on current steps)

(labels on current steps)



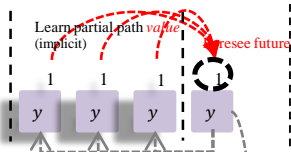
correct answer



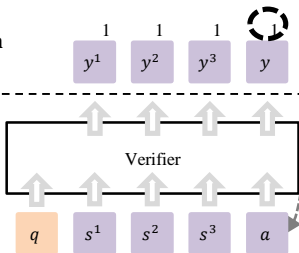
incorrect answer

An Interesting Finding from Complete Path Verification – Outcome Supervision v.s. Process supervision

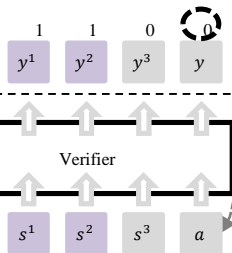
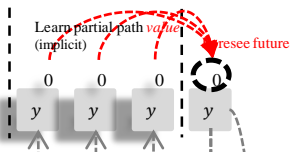
Outcome Supervision
(labels on final answer)



Process Supervision
(labels on current steps)



correct answer

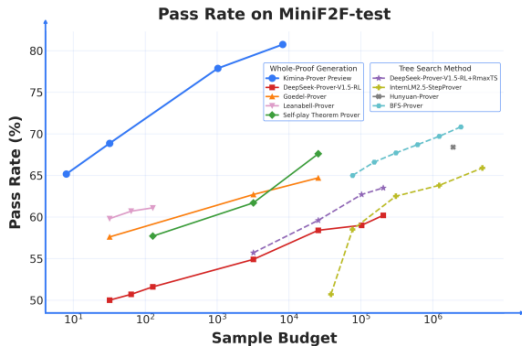


incorrect answer

Drawbacks

Learnable verification is noisy or biased!

Solution: rule-based rewards/ verification ?



Automatic theorem proving

Past experience for synthesized long COT SFT data

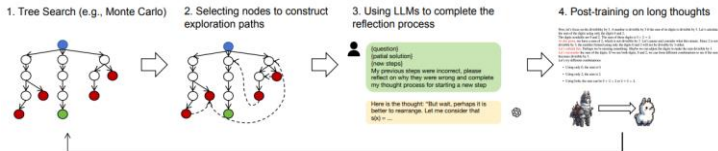
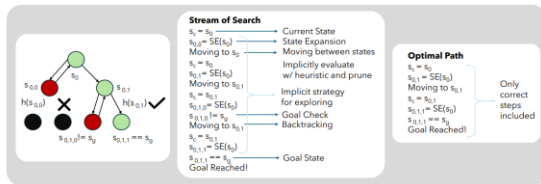
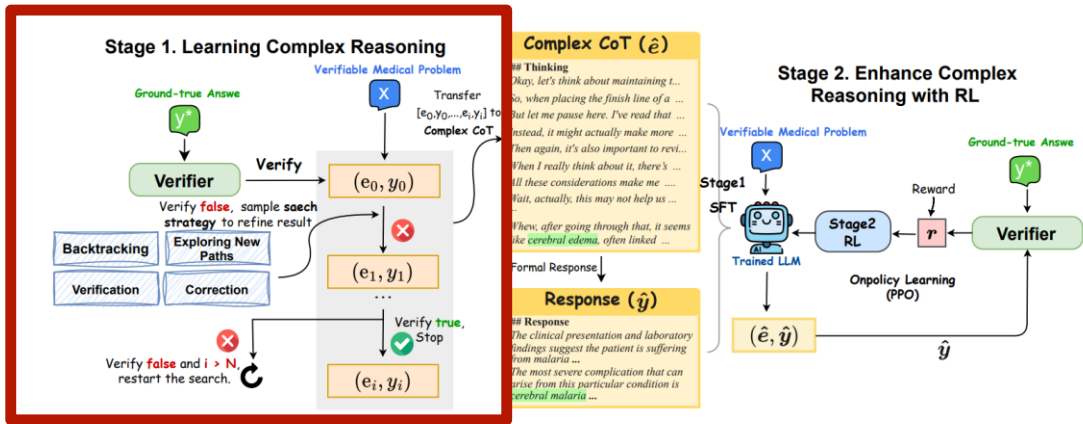


Figure 2: The framework of journey learning.

<https://arxiv.org/pdf/2404.03683>
<https://arxiv.org/pdf/2411.16489?>



Drawbacks

SFT does not generalize! It cannot learn new patterns; it learns from a teacher that it cannot outperform.

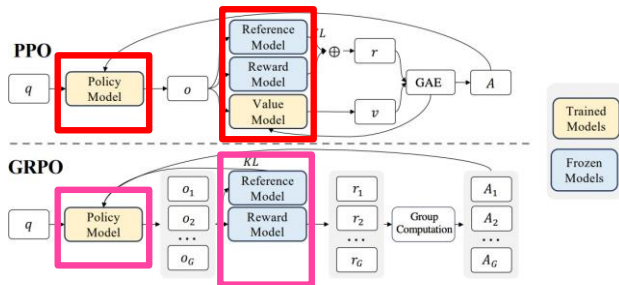
Solution: try RL

SimpleRL: Emerging Reasoning with Reinforcement Learning is Both Effective and Efficient

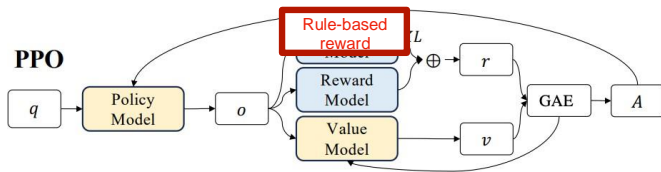
The SimpleRL Recipe

PPO algorithm, the training objective is:

$$\mathcal{L}_{\text{ppo-clip}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(\frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t, \text{clip} \left(\frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_t \right) \right],$$



The Simple RL Recipe



- Rule-based reward function (Correctness and Format)
 - If the answer is correct, reward = 1
 - If the format is correct but the answer is incorrect, reward = -0.5
 - If the format is incorrect, reward = -1

Format Example

Many many reasoning steps....

Since $\sqrt[3]{1/7}$ is the seventh root of 3, the minimum value is: $\boxed{36}$.

In the following experiments, it only use 8K examples from MATH dataset.

Experiment Setup

- Start from the Qwen2.5-Math-7B-Base model.

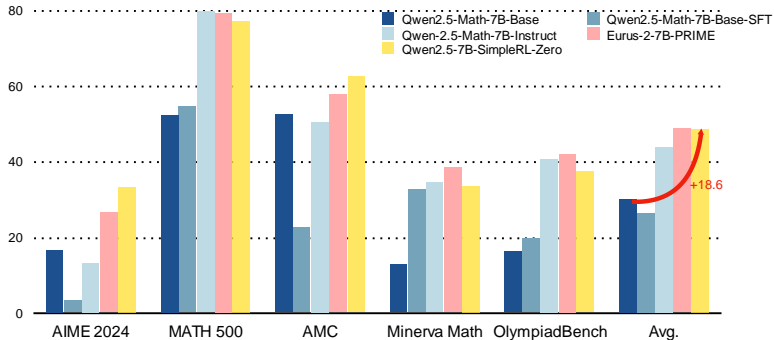
	Qwen2.5-Math-7B-Instruct	rStar-Math-7B	Eurus-2-7B-PRIME	Qwen2.5-7B-SimpleRL-Zero
Base Model	Qwen2.5-Math-7B	Qwen2.5-Math-7B	Qwen2.5-Math-7B	Qwen2.5-Math-7B
SFT Data	2.5M (open-source and in-house)	~7.3M (MATH, NuminaMath, etc)	230K	0
RM Data	618K (in-house)	~7k (in-house)	0	0
RM	Qwen2.5-Math-RM (72B)	None	Eurus-2-7B-SFT	None
RL Data	66K queries × 32 samples	~3.647 M × 16	150K queries × 4 samples	8K queries × 8 samples

Using only 8K examples from the original MATH dataset, much less than other methods.

Evaluate on challenging math benchmarks, AIME, AMC ...

SimpleRL-Zero — RL from scratch

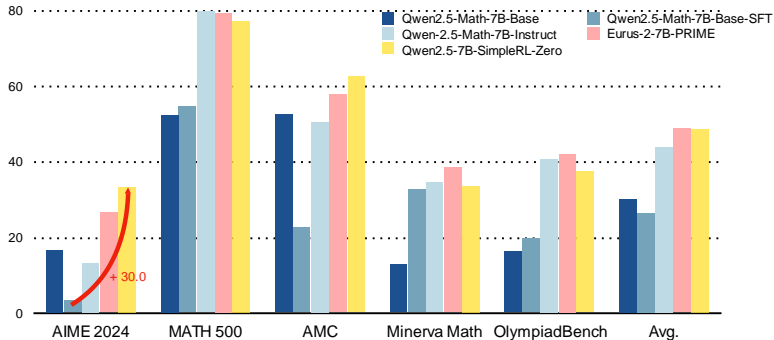
RL is applied directly to the base model using 8K MATH (query, answer) pairs without SFT.



Achieve gains of nearly 20 absolute points from base model.

Outperform Instruct model and achieves comparable results to PRIME with 50x data efficiency.

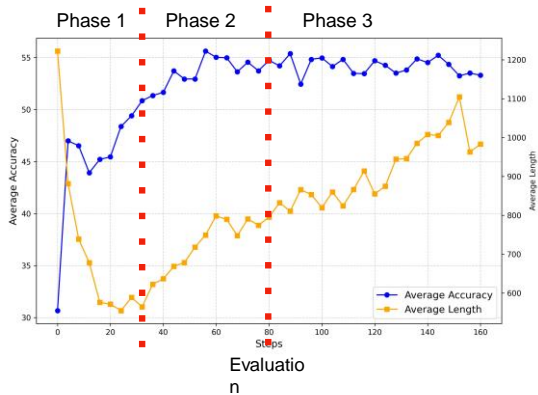
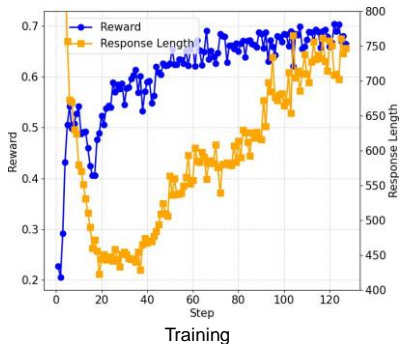
SimpleRL-Zero — RL from scratch



Compared to SFT, RL generalizes to competition-level benchmarks and being 30 points higher in AIME

Easy-to-hard generalization !

SimpleRL-Zero — RL from scratch

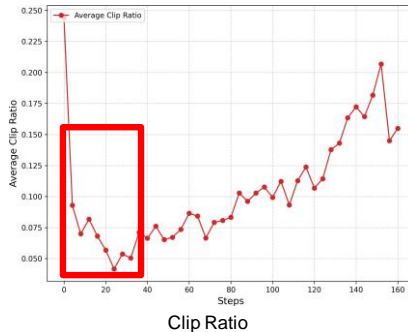
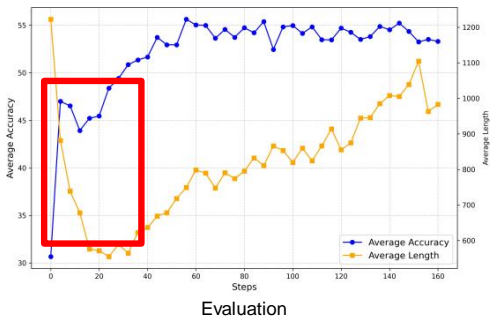


The accuracy is improving on average, while the length decreases first and then gradually increases.

Surprisingly how far the 8K MATH examples lift this 7B base model without any other external signals.

Phase 1 — Why length decreases at first

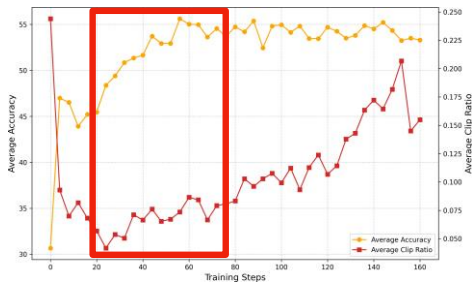
Clip Ratio: the ratio of generations that reach maximum length.



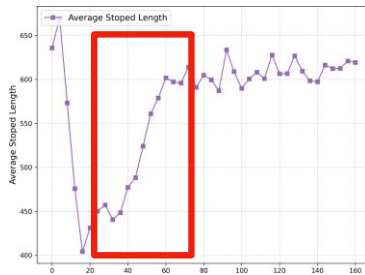
Initially, the base model hard to stop generation and gradually learns to stop from the predefined rule.

Phase 2 — length and perf increase simultaneously

Average Stopped Length: The average length of sequences that terminate with stop words.



Acc and Clip Ratio



Average Stopped Length

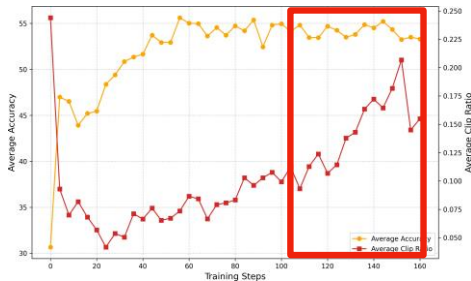
While the clip ratio increases, the sequences that normally stop also becomes longer.

Longer Sequence -> Better Exploration -> stronger performance

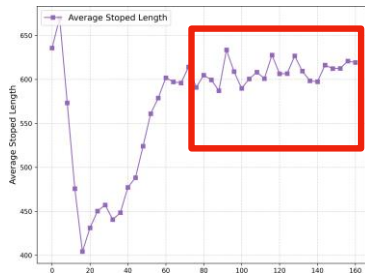
Exploration is the important factor for better RL

Phase 3— length and perf saturate simultaneously

Average Stopped Length: The average length of sequences that terminate with stop words.



Acc and Clip Ratio



Average Stopped Length

Model tends to not stop generating and the stopped length begin to saturate.

When model stop to explore larger generation space, model performance saturates.

Larger Temperature or length limitation ?

How to stop model from meaningless repetition might be important.

Emergence of Reflection

We do observe some signs of reflection around iteration 40.

Question: Denali and Nate work for a dog walking business and are paid for each dog they walk.

Response:

To solve the problem, we need to set up the equations based on the given conditions and solve for x .

Let's denote the payment per dog as p . The initial ratio of Denali's pay to Nate's pay is $\frac{16p}{12p} = \frac{4}{3}$.

According to the problem, this ratio remains the same in two different scenarios:

1. If Denali starts walking $4x$ more dogs and Nate stays at 12 dogs, the ratio of their pays is ...
2. If x of Nate's dogs are reassigned to Denali, the ratio of their pays is $\frac{(16+x)p}{(12-x)p} = \frac{16+x}{12-x}$.

...

Again, since $x \neq 0$, this equation does not provide a valid solution for x . It seems there might be an error.

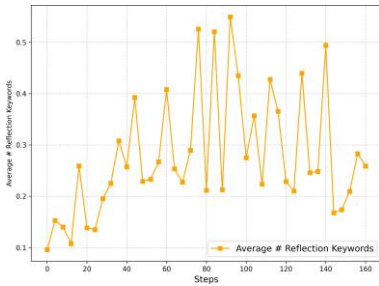
Let's recheck the setup. The correct setup should be: $\frac{16+4x}{12} = \frac{4}{3}$ and $\frac{16+x}{12-x} = \frac{4}{3}$

Let's solve the second equation again carefully:

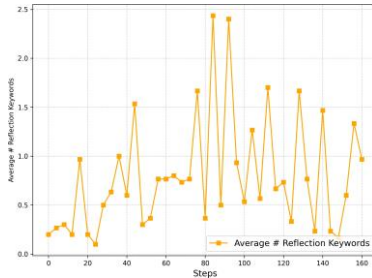
...

Emergence of Reflection

Reflection Keywords: "recheck", "rethink", "try again", "wait", "alternatively", "retry", "however"



Average



AIME 24

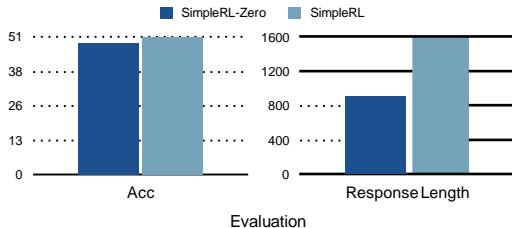
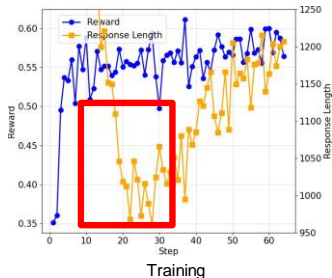
The change in # reflection keywords is unstable and does not exhibit a clear emergence pattern.

More challenging benchmarks contain a higher # keywords.

Larger Model or More challenging training data might be necessary.

SimpleRL — RL With Imitation Warmup

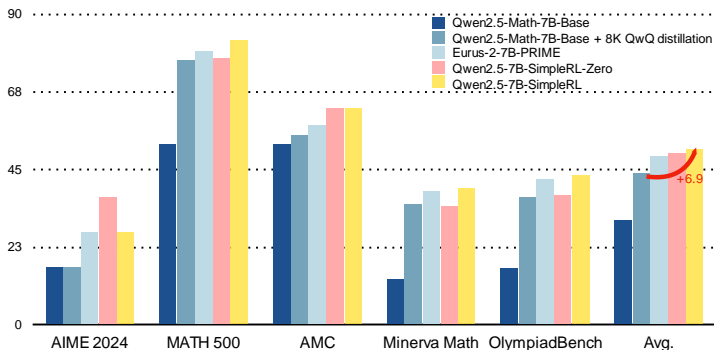
Distill a long CoT example from Qwen-QwQ-32B for SFT warmup, then apply RL similar to SimpleRL-Zero.



The training dynamics is quite similar to that of SimpleRL-Zero, but the length is much longer.

The length decrease may stem from the QwQ pattern being disfavored or beyond the model's capacity.

SimpleRL — RL With Imitation Warmup



Qwen2.5-7B-SimpleRL outperforms SFT with 8K QwQ distillation by 6.9 points after RL training.

It consistently outperforms Eurus-2-7B-PRIME and exceeds SimpleRL-Zero.

Some Hypothesis

- To perform RL, the **base model** need to reach certain accuracy.
 - Weak model with easy data.
- Large model might be needed for the emergence of reflection.
 - Different model with different level of search.
- More dynamically control of the sampling temperature and length.
 - In B-Star, we tune temperature when diversity decrease.

Summary of SimpleRL

- Replicate DeepSeek-R1 on small models with limited data.
- The design is simple—no reward model, no SFT.
- Surprisingly, a 7B model develops long CoT and shows some signs of self-reflection using just 8K MATH examples.
- It rivals methods using $>50\times$ more data and complex designs.



Our work: HuatuoGPT-O1

What are the verifiable problems in medical domains?

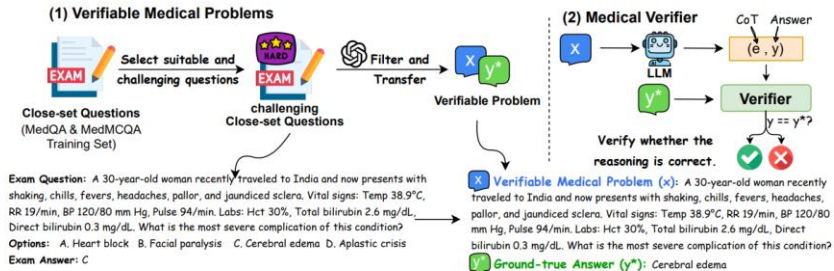
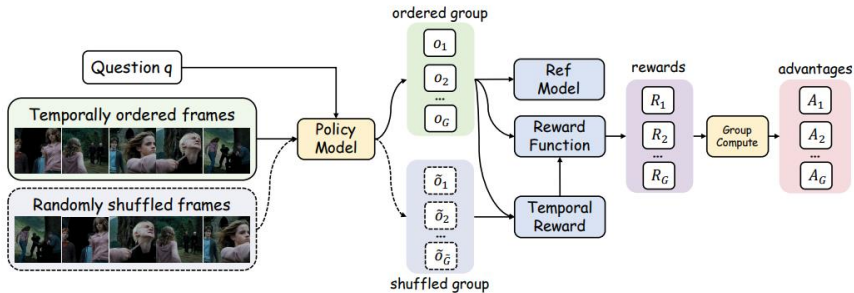


Figure 1: **Left:** Constructing verifiable medical problems using challenging close-set exam questions. **Right:** The verifier checks the model's answer against the *ground-truth answer*.

Our work: Video-R1

<https://arxiv.org/pdf/2503.21776>

Video-R1



Encourage the model get better rewards with **ordered frames** rather than **shuffled frames**

Thank You

[Refer to https://github.com/srush/awesome-ol](https://github.com/srush/awesome-ol)
Junxian he, slides

Reference I

[Ankner et al., 2024] Ankner, Z., Paul, M., Cui, B., Chang, J. D., and Ammanabrolu, P. (2024).

Critique-out-loud reward models.

arXiv [cs.LG].

[Anthony et al., 2017] Anthony, T., Tian, Z., and Barber, D. (2017).

Thinking fast and slow with deep learning and tree search.

arXiv [cs.AI].

[Brown and Sandholm, 2017] Brown, N. and Sandholm, T. (2017).

Libratus: The superhuman AI for no-limit poker.

In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, California. International Joint Conferences on Artificial Intelligence Organization.

Reference II

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020).

Language models are few-shot learners.

arXiv [cs.CL].

[Cobbe et al., 2021] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021).

Training verifiers to solve math word problems.

arXiv [cs.LG].

Reference III

[Feng et al., 2023] Feng, X., Wan, Z., Wen, M., McAleer, S. M., Wen, Y., Zhang, W., and Wang, J. (2023).

Alphazero-like tree-search can guide large language model decoding and training.

arXiv [cs.LG].

[Gandhi et al., 2024] Gandhi, K., Lee, D., Grand, G., Liu, M., Cheng, W., Sharma, A., and Goodman, N. D. (2024).

Stream of search (SoS): Learning to search in language.

arXiv [cs.LG].

[Gulcehre et al., 2023] Gulcehre, C., Paine, T. L., Srinivasan, S., Konyushkova, K., Weerts, L., Sharma, A., Siddhant, A., Ahern, A., Wang, M., Gu, C., Macherey, W., Doucet, A., Firat, O., and de Freitas, N. (2023).

Reinforced self-training (ReST) for language modeling.

arXiv [cs.CL].

Reference IV

[Hendrycks et al., 2021a] Hendrycks, D., Basart, S., Kadavath, S., Mazeika, M., Arora, A., Guo, E., Burns, C., Puranik, S., He, H., Song, D., and Steinhardt, J. (2021a).

Measuring coding challenge competence with APPS.

arXiv [cs.SE].

[Hendrycks et al., 2021b] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021b).

Measuring massive multitask language understanding.

[Hendrycks et al., 2021c] Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. (2021c).

Measuring mathematical problem solving with the MATH dataset.

arXiv [cs.LG].

Reference V

[Hosseini et al., 2024] Hosseini, A., Yuan, X., Malkin, N., Courville, A., Sordoni, A., and Agarwal, R. (2024).

V-STAR: Training verifiers for self-taught reasoners.

In First Conference on Language Modeling.

[Hubert et al., 2021] Hubert, T., Schrittwieser, J., Antonoglou, I., Barekatin, M., Schmitt, S., and Silver, D. (2021).

Learning and planning in complex action spaces.

arXiv [cs.LG].

[Jones, 2021] Jones, A. L. (2021).

Scaling scaling laws with board games.

arXiv [cs.LG].

Reference VI

- [Kazemnejad et al., 2024] Kazemnejad, A., Aghajohari, M., Portelance, E., Sordoni, A., Reddy, S., Courville, A., and Roux, N. L. (2024).
VinePPO: Unlocking RL potential for LLM reasoning through refined credit assignment.
arXiv [cs.LG].
- [Kocsis and Szepesvári, 2006] Kocsis, L. and Szepesvári, C. (2006).
Bandit based monte-carlo planning.
In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 282–293.
Springer Berlin Heidelberg, Berlin, Heidelberg.
- [Lightman et al., 2023] Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. (2023).
Let's verify step by step.
arXiv [cs.LG].

Reference VII

[Nakano et al., 2021] Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., and Schulman, J. (2021).

WebGPT: Browser-assisted question-answering with human feedback.

arXiv [cs.CL].

[Neal and Hinton, 1998] Neal, R. M. and Hinton, G. E. (1998).

A view of the em algorithm that justifies incremental, sparse, and other variants.

In *Learning in Graphical Models*, pages 355–368. Springer Netherlands, Dordrecht.

[Nye et al., 2021] Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., Sutton, C., and Odena, A. (2021).

Show your work: Scratchpads for intermediate computation with language models.

arXiv [cs.LG].

Reference VIII

[OpenAI, 2024] OpenAI (2024).

Learning to reason with LLMs.

<https://openai.com/index/learning-to-reason-with-llms/>.

Accessed: 2024-10-29.

[Putta et al., 2024] Putta, P., Mills, E., Garg, N., Motwani, S., Finn, C., Garg, D., and Rafailov, R. (2024).

Agent Q: Advanced reasoning and learning for autonomous AI agents.

arXiv [cs.AI].

[Silver et al., 2017] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2017).

Mastering chess and shogi by self-play with a general reinforcement learning algorithm.

arXiv [cs.AI].

Reference IX

[Singh et al., 2023] Singh, A., Co-Reyes, J. D., Agarwal, R., Anand, A., Patil, P., Garcia, X., Liu, P. J., Harrison, J., Lee, J., Xu, K., Parisi, A., Kumar, A., Alemi, A., Rizkowsky, A., Nova, A., Adlam, B., Bohnet, B., Elsayed, G., Sedghi, H., Mordatch, I., Simpson, I., Gur, I., Snoek, J., Pennington, J., Hron, J., Kenealy, K., Swersky, K., Mahajan, K., Culp, L., Xiao, L., Bileschi, M. L., Constant, N., Novak, R., Liu, R., Warkentin, T., Qian, Y., Bansal, Y., Dyer, E., Neyshabur, B., Sohl-Dickstein, J., and Fiedel, N. (2023).

Beyond human data: Scaling self-training for problem-solving with language models.

arXiv [cs.LG].

[Snell et al., 2024] Snell, C., Lee, J., Xu, K., and Kumar, A. (2024).

Scaling LLM test-time compute optimally can be more effective than scaling model parameters.

arXiv [cs.LG].

[Sutton, 2019] Sutton, R. (2019).

The bitter lesson.

Reference X

[Uesato et al., 2022] Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., and Higgins, I. (2022).

Solving math word problems with process- and outcome-based feedback.

arXiv [cs.LG].

[Wang et al., 2023] Wang, P., Li, L., Shao, Z., Xu, R. X., Dai, D., Li, Y., Chen, D., Wu, Y., and Sui, Z. (2023).

Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations.

arXiv [cs.AI].

[Wang et al., 2022] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2022).

Self-consistency improves chain of thought reasoning in language models.

arXiv [cs.CL].

Reference XI

[Wei et al., 2022] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2022).

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

arXiv [cs.CL], pages 24824–24837.

[Welleck et al., 2024] Welleck, S., Bertsch, A., Finlayson, M., Schoelkopf, H., Xie, A., Neubig, G., Kulikov, I., and Harchaoui, Z. (2024).

From decoding to meta-generation: Inference-time algorithms for large language models.

arXiv [cs.CL].

[Welleck et al., 2022] Welleck, S., Lu, X., West, P., Brahman, F., Shen, T., Khashabi, D., and Choi, Y. (2022).

Generating sequences by learning to self-correct.

arXiv [cs.CL].

Reference XII

[Xie et al., 2024] Xie, Y., Goyal, A., Zheng, W., Kan, M.-Y., Lillicrap, T. P., Kawaguchi, K., and Shieh, M. (2024).

Monte carlo tree search boosts reasoning via iterative preference learning.

arXiv [cs.AI].

[Yarowsky, 1995] Yarowsky, D. (1995).

Unsupervised word sense disambiguation rivaling supervised methods.

In Proceedings of the 33rd annual meeting on Association for Computational Linguistics -, Morristown, NJ, USA. Association for Computational Linguistics.

[Zelikman et al., 2022] Zelikman, E., Wu, Y., Mu, J., and Goodman, N. D. (2022).

STaR: Bootstrapping reasoning with reasoning.

arXiv [cs.LG].

Reference XIII

[Zhang et al., 2024] Zhang, L., Hosseini, A., Bansal, H., Kazemi, M., Kumar, A., and Agarwal, R. (2024).

Generative verifiers: Reward modeling as next-token prediction.

arXiv [cs.LG].

Thank You!