LLM Reasoning

Denny Zhou

Google DeepMind

The Future of Language Models and Transformers, March 31-April 4, 2025, Simons Institute

What is LLM Reasoning

Ling et al. Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems. ACL 2017

Chen et al. Compositional generalization via neural-symbolic stack machines. NeurIPS 2020.

What is the output when concatenating the last letter of each word in “artificial intelligence”?

No reasoning

The answer is “le”.

Reasoning

The last letter of “artificial” is “l”. The last letter of “intelligence” is “e”. Concatenating “l” and “e” leads to “le”. So the answer is “le”.

Parisotto, Emilio, Abdel-rahman Mohamed, Rishabh Singh, Lihong Li, Denny Zhou, and Pushmeet Kohli. Neuro-symbolic program synthesis. arXiv preprint arXiv:1611.01855 (2016).

Why “Intermediate Tokens” / “Reasoning” Matters

- For any problems solvable by boolean circuits of size T, **constant-size transformers** can solve it by generating O(T) intermediate tokens
- If directly generating final answers, either requires a huge depth or cannot solve at all

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of Thought Empowers Transformers to Solve Inherently Serial Problems. ICLR 2024.

Common Belief

Pretrained LLMs cannot reason without further prompting engineering or finetuning

WRONG

Pretrained LLMs are ready to reason

All we need is decoding

Xuezhi Wang and Denny Zhou. Chain-of-Thought Reasoning Without Prompting. NeurIPS 2024.

No reasoning? Check more generation candidates!

I have 3 apples. My dad has 2 more apples than me. How many apples do we have in total?

5 apples. (Greedy Decoding)

I have 3 apples, my dad has 2 more apples than me, so he has 5 apples. 3+5=8.

We have 8 apples in total.

You have 3 apples, your dad has 2 more apples than you, so he has 5 apples. 3+5=8.

The answer is 5.

Xuezhi Wang and Denny Zhou. Chain-of-Thought Reasoning Without Prompting. NeurIPS 2024.

How to identify reasoning generation? By length?

I have 3 apples. My dad has 2 more apples than me. How many apples do we have in total?

5 apples. (Greedy Decoding)

I have 3 apples, my dad has 2 more apples than me, so he has 5 apples. 3+5=8.

We have 8 apples in total.

You have 3 apples, your dad has 2 more apples than you, so he has 5 apples. 3+5=8.

The answer is 5.

Xuezhi Wang and Denny Zhou. [Chain-of-Thought Reasoning Without Prompting](). NeurIPS 2024.

Identify reasoning generation by confidence!

I have 3 apples. My dad has 2 more apples than me. How many apples do we have in total?

5 apples. (Greedy Decoding)

I have 3 apples, my dad has 2 more apples than me, so he has 5 apples. 3+5= **8**.

We have 8 apples in total.

You have 3 apples, your dad has 2 more apples than you, so he has 5 apples. 3+5= **8.**

The answer is 5.

Way higher confidence on reasoning-based answers!

Chain-of-Thought Decoding

1. Go beyond greedy decoding by checking more generation candidates
2. Choose candidates which have the highest confidence on the final answer

Xuezhi Wang and Denny Zhou. Chain-of-Thought Reasoning Without Prompting. NeurIPS 2024.

Any ways to push reasoning introduced generations to top positions such that they can be found by greedy decoding?

Chain-of-Thought Prompting

Q: Elsa has 3 apples. Anna has 2 more apples than Elsa. How many apples do they have together?

A: Anna has 2 more apples than Elsa. So Anna has 2 + 3 = 5 apples. So Elsa and Anna have 3 + 5 = 8 apples together.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. NeurIPS 2022

Let’s Think Step by Step

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

Let’s think step by step.

Kojima, T., Gu, S.S., Reid, M., Matsuo, Y. and Iwasawa, Y. Large language models are zero-shot reasoners. NeurIPS 2022.

Pros and Cons of Prompting

Pros: Simple and works

Cons:

CoT prompting needs task-specific examples

“Let’s think step by step” is generic, but performs much worse than few-shot

Pros and Cons of Prompting

Prompting approaches are actually weird

When asking someone a question —

will you first show similar problems/solutions before asking?

or, at the end of you question, will you have to say “let’s think step by step”?

Of course not!

Supervised Finetuning (SFT)

Step 1: collect a set of problems and their step-by-step solutions from human annotators

Step 2: finetune model with (problem, solution) pairs

Then apply the model everywhere

Ling et al. Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems. ACL 2017

Cobbe et al. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168. 2021

Nye et al. Show Your Work: Scratchpads for Intermediate Computation with Language Models. arXiv:2112.00114, 2021

Supervised Finetuning (SFT)

What is the output when concatenating the last letter of each word in “artificial intelligence”? The last letter of “artificial” is “l”. The last letter of “intelligence” is “e”. Concatenating “l” and “e” leads to “le”. So the answer is “le”.

Elsa has 3 apples. Anna has 2 more apples than Elsa. How many apples do they have together? Anna has 2 more apples than Elsa. So Anna has 2 + 3 = 5 apples. So Elsa and Anna have 3 + 5 = 8 apples together.

Training data

Finetuning

LLM

Test problem

How many “r”s in “strawberry”?

Pros and Cons of SFT

Pros: Generic

Cons:

Does not generalize well

Scaling does not help much

RL Finetuning

Step 1: collect a set of problems and let model generate step-by-step solutions

Step 2: finetune model with (problem, solution) pairs

Zelikman E, Wu Y, Mu J, Goodman N. Star: Bootstrapping reasoning with reasoning. NeurIPS 2022.

Huang J, Gu SS, Hou L, Wu Y, Wang X, Yu H, Han J. Large language models can self-improve. arXiv:2210.11610. 2022

RL Finetuning

Step 1: collect a set of problems and let model generate step-by-step solutions

Step 2: finetune model with (problem, solution) pairs

Zelikman E, Wu Y, Mu J, Goodman N. Star: Bootstrapping reasoning with reasoning. NeurIPS 2022.

Huang J, Gu SS, Hou L, Wu Y, Wang X, Yu H, Han J. Large language models can self-improve. arXiv:2210.11610. 2022

RL Finetuning

Repeat this process:

Step 1: collect a set of problems and let model generate step-by-step solutions

Step 2: finetune model with (problem, solution) pairs

Luong TQ, Zhang X, Jie Z, Sun P, Jin X, Li H. ReFT: **Reasoning with Reinforced Finetuning**. arXiv:2401.08967. **2024 Jan 17**.

Needless to say, verifier needed in RL finetuning

“Verification, the key to AI”

— Rich Sutton, 11/15/2001

Why “let model generate”?

First principle in ML

Directly optimize what we want

So what do we want to optimize?

Generation quality (e.g. accuracy, BLEU, …)

Then compute gradient and do backpropagation!

So what do we want to optimize?

$$\max_{\theta} \mathbb{E}[r(\text{response}|\text{problem}, \theta)]$$

$$r \text{ measuring the quality of a response}$$

$$\theta \text{ model parameters}$$

Then compute gradient and do backpropagation!

Scaling Reasoning Training: What to Scale?

- For any problems solvable by boolean circuits of size T, **constant-size transformers** can solve it by generating O(T) intermediate tokens — Scaling output length!
- If directly generating final answers, either requires a huge depth to solve or cannot solve at all — Scaling model depth!

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of Thought Empowers Transformers to Solve Inherently Serial Problems. ICLR 2024.

The Beauty of LLM Reasoning

Human-like reasoning emerges from token-to-token generation, rather than relying on exhaustive search as in classic AI

"Deep Blue was only intelligent the way your programmable alarm clock is intelligent"
—- Garry Kasparov (after losing to Deep Blue in 1997)

[Question] Using the numbers from 1 to 10 to make 2025. Using each number once and primary operations + and *

[Answer] (10 * 4 + 5) * (9 * 3 + 8 + 7 + 2 + 1)

Let's break it down:

- First part: 10 * 4 + 5 = 40 + 5 = 45
- Second part: 9 * 3 + 8 + 7 + 2 + 1 = 27 + 8 + 7 + 2 + 1 = 45
- Final Calculation: 45 * 45 = 2025

Therefore, the expression (10 * 4 + 5) * (9 * 3 + 8 + 7 + 2 + 1) uses the numbers 1 through 10 exactly once and only the primary operations of addition (+) and multiplication (*) to equal 2025.

Gemini 2.0 thinking mode (Dec, 2024)

The target number is 2025. This is a relatively large number, suggesting multiplication will be heavily involved. It's also worth noting that 2025 is 45 squared (45 *45). This might or might not be directly helpful, but it gives us a sense of scale. **Prioritize Multiplication:**Since the target is large, start thinking about how to get large intermediate products using multiplication. The largest numbers (10,9,8,7) are good candidates for multiplication early on ... This confirms multiplication is key, but simply multiplying everything together is too much. We need to combine additions. … Let's aim for products that get us closer to the square root of 2025 (which is 45) …

“We want AI agents that can discover like we can, not which contain what we have discovered. Building in our discoveries only makes it harder to see how the discovering process can be done.”

— Rich Sutton (“The Bitter Lesson”)

Pros and Cons of RL Finetuning

Pros: Generalize well

Cons: Many tasks are not automatically verifiable

LLM reasoning is powerful — but does the paradigm of “generating ‘intermediate thoughts’ and then final answer” introduce challenges?

Always keep in mind that LLMs are probabilistic models trained to predict next tokens. They are not humans!

What LLM does in decoding:

$$\arg\max \mathbb{P}(\text{reasoning, final answer}|\text{problem})$$

What we want:

$$\arg\max \mathbb{P}(\text{final answer}|\text{problem})$$

Not align!

Let's fix it!

Marginalization

$$\arg\max \mathbb{P}(\text{final answer}|\text{problem})$$

$$= \sum_{\text{reasoning}} \mathbb{P}(\text{reasoning, final answer}|\text{problem})$$

$$\approx \frac{\text{frequency of final answer}}{\text{total number of sampled responses}}$$

Self-Consistency

1. Generate multiple responses by randomly sampling
2. Choose the answer that appears most frequently

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. ICLR 2023.

[Question] Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for $2 per egg. How much does she make every day?

Sampled responses:

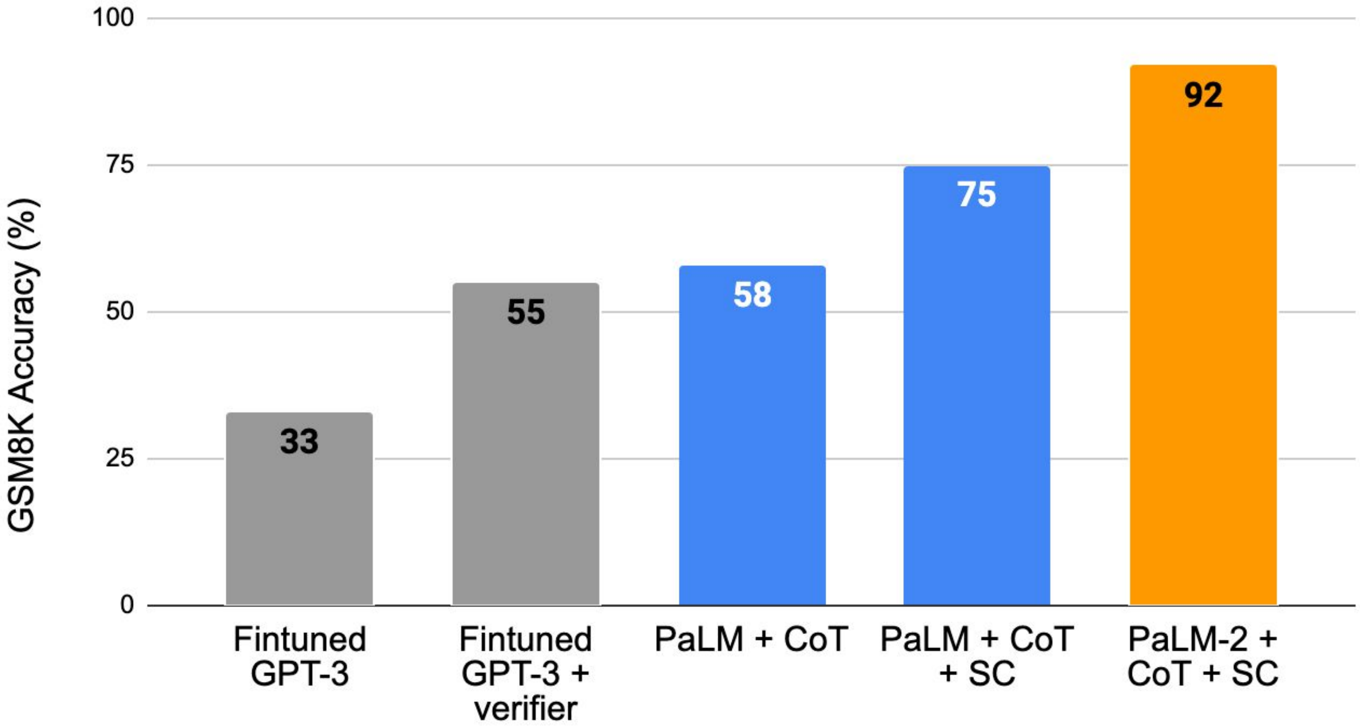
Response 1: She has 16 - 3 - 4 = 9 eggs left. So she makes $2 * 9 = $\boxed{\$\mathbf{18}}$ per day.

Response 2: This means she she sells the remainder for $2 * (16 - 4 - 3) = $\boxed{\$\mathbf{26}}$ per day.

Response 3: She eats 3 for breakfast, so she has 16 - 3 = 13 left. Then she bakes muffins, so she has 13 - 4 = 9 eggs left. So she has 9 eggs * $2 = $\boxed{\$\mathbf{18}.}$

Most frequent answer is: 18
(Not most frequent reasoning path!)

Results on GSM8K (8 shots, Jan 2022/3)

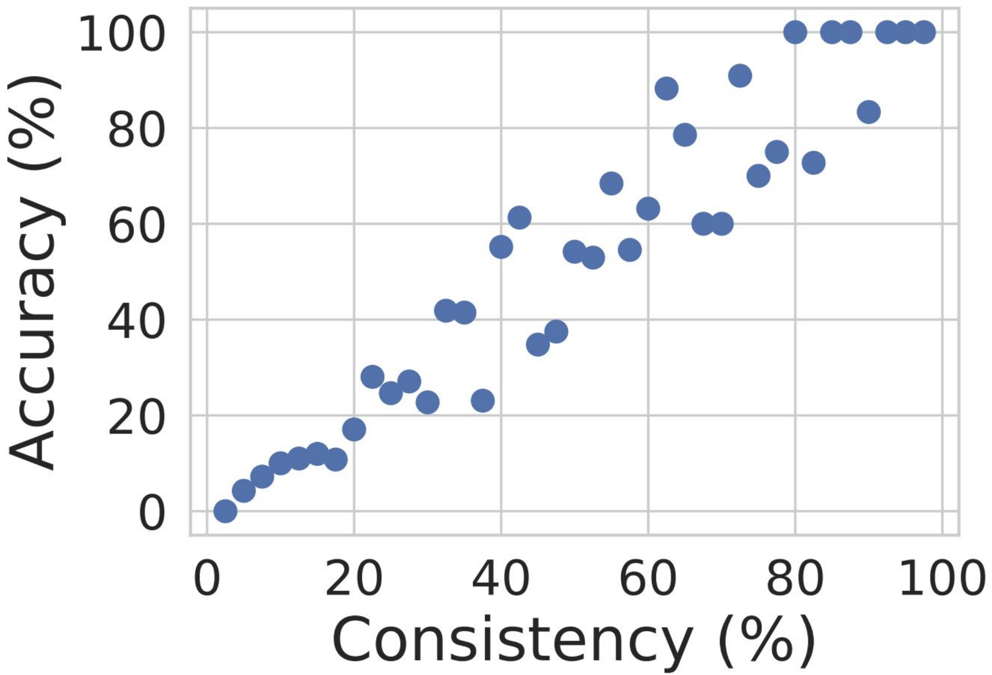


Appendix A

Dataset	Metric	gpt-4o	o1-preview	o1
Competition Math	cons@64	13.4	56.7	83.3
AIME (2024)	pass@1	9.3	44.6	74.4

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

Higher Consistency Indicates Higher Accuracy



QUIZ

[Q1] When the LLM outputs a direct answer without intermediate steps, will you still sample several times, and then choose the most common answer?

[Q2] Change self-consistency by letting LLM generate multiple responses, instead of sampling multiple times, and then choosing the most common answer. Does this make sense?

$$\arg\max \mathbb{P}(\text{final answer}|\text{problem})$$

How about free-from answers?

Universal Self-Consistency (USC)

Ask LLMs to self-select the most consistent answer

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, Denny Zhou. Universal Self-Consistency for Large Language Model Generation. arXiv:2311.17311 [cs.CL], 2023.

[Question] Where do people drink less coffee than they do in Mexico?

Response 1: ... Some examples include $\boxed{\text{Japan, China and the United Kingdom}}$.

It is important to note that coffee consumption can vary among individuals within these countries, and preferences can change depending on different factors such as…

Response 2: People in countries like $\boxed{\text{Japan, China, and India}}$ typically drink less coffee than they do in Mexico...

Response 3: There are several countries where people generally drink less coffee compared to Mexico. Some of these countries include:

1. Japan:...
2. China...
3. Saudi Arabia...
4. India...

...

The most consistent response: 2

Retrieval or reasoning?

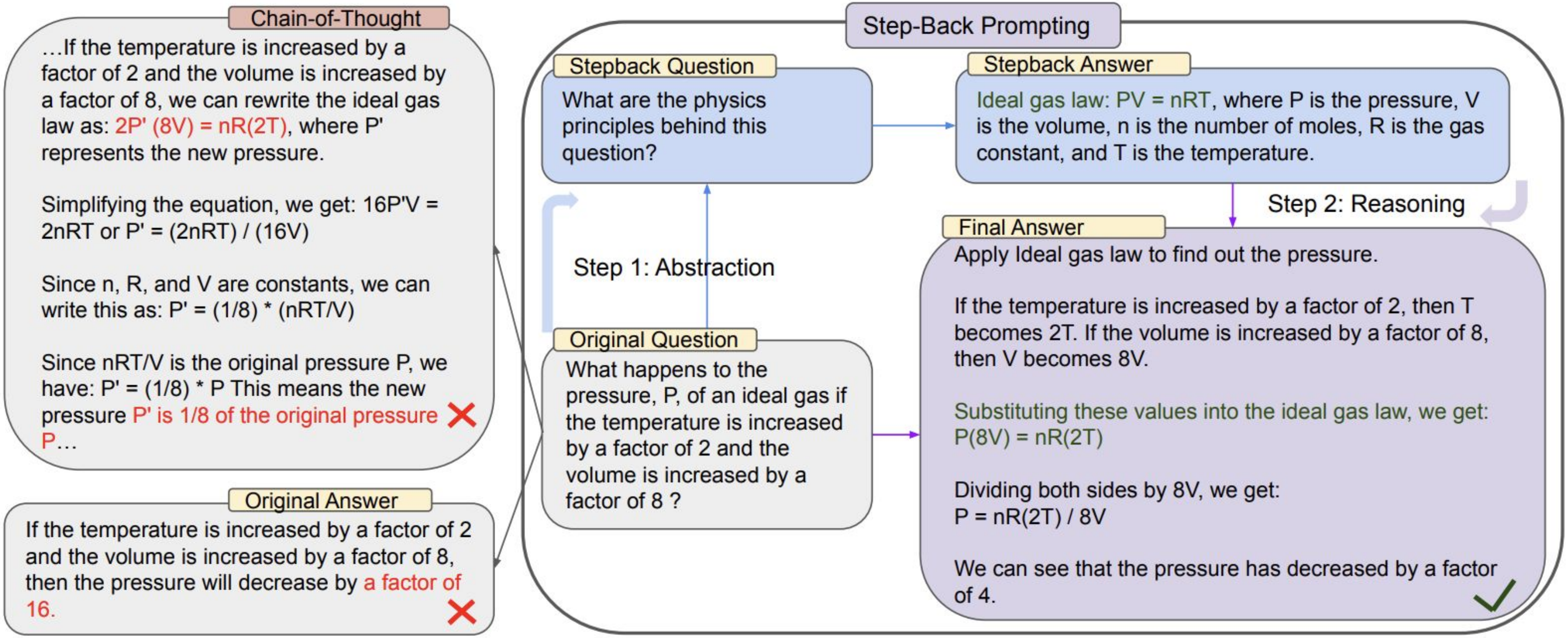
Do “retrieval + reasoning”!

What is the area of the square with the four vertices at (-2, 2), (2, -2), (-2, -6), and (-6, -2)? Recall a related problem, and then solve this one.

The related problem is finding the distance between two points on a coordinate plane. To find the distance between two points, we use the distance formula: d = √((x2-x1)^2 + (y2-y1)^2).

In this problem, we have four points: (-2, 2), (2, -2), (-2, -6), and (-6, -2). Let's choose (-2, 2) and (2, -2) to find the distance d between them: d = √((2-(-2))^2 + (-2-2)^2) = √(4^2 + (-4)^2) = √(16 + 16) = √32. We know that the distance between two opposite sides of a square is equal. So, the length of each side of the square is √32. To find the area of the square, we square the length of one side: Area = (√32)^2 = 32.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, Denny Zhou. [Large Language Models as Analogical Reasoners](). ICLR 2024.



Zheng, Huaixiu Steven, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V. Le, and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models. ICLR 2024.

Gemini Deep Research

Search

Deep research

perplexity

pro

Deep Research

Summary

- Reasoning > no reasoning
- RL finetuning > SFT
- Aggregating multiple answers > single answer
- Retrieval + reasoning > reasoning only

Next Big Breakthroughs

Go beyond problems with unique verifiable answers

Go beyond benchmarks to build real applications

THE END

“The truth always turns out to be simpler than you thought.”
— Richard P. Feynman