Lecture1 - Reasoning

Wednesday, February 19, 2025 3:32 PM

intro

CS 294/194-196: Large Language Model Agents

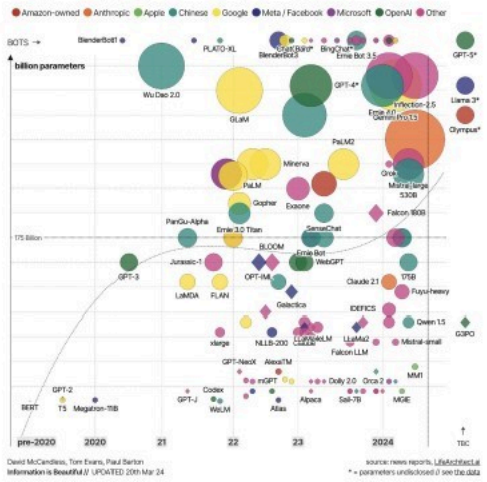
Teaching Staff

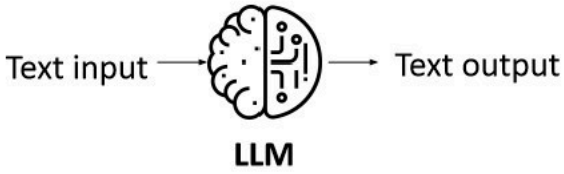
- **Instructor: Prof. Dawn Song**
- **(guest) Co-instructor: Dr. Xinyun Chen**
- GSIs: Alex Pan & Sehoon Kim
- Readers: Tara Pande & Ashwin Dara

Accelerated development of large language models (LLMs)

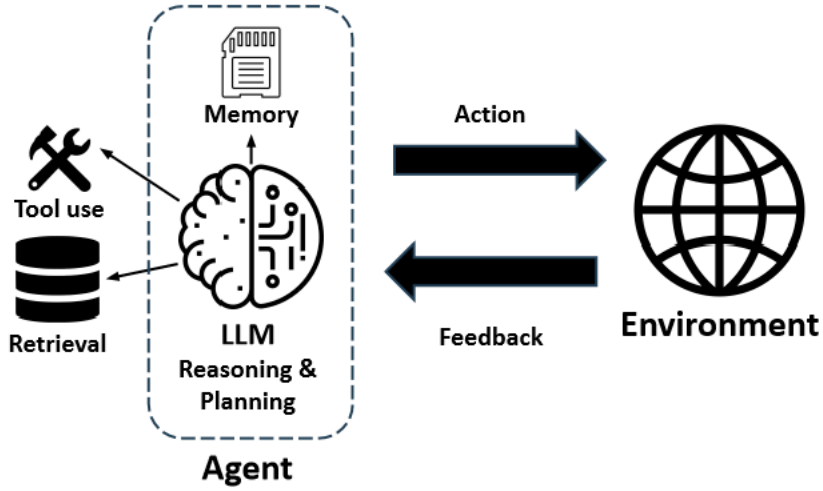
The Rise and Rise of A.I. Large Language Models (LLMs) & their associated bots like ChatGPT

size = no. of parameters open-access

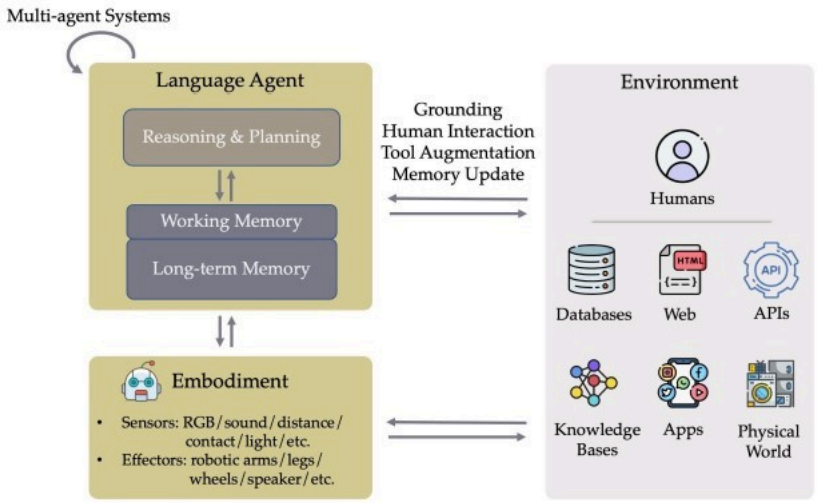




LLM agents: enabling LLMs to interact with the environment

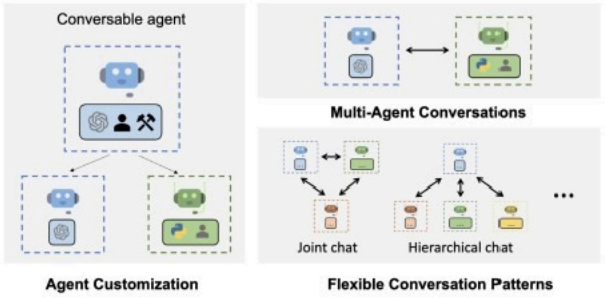


LLM Agents in Diverse Environments



https://yusu.substack.com/p/language-agents

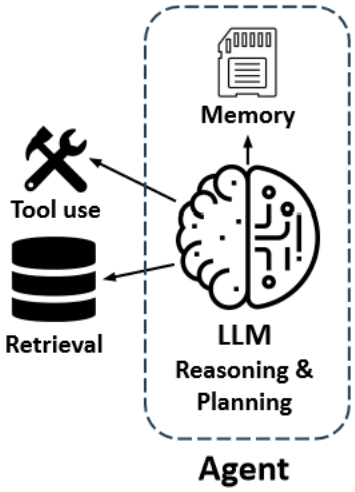
Multi-agent collaboration: division of labor for complex tasks



Specialized agents for different subtasks
Autogen, CrewAI, CAMEL, Mixture-of-Agents,...

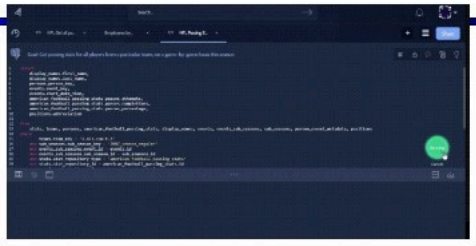
Emergence of social behaviors with role-play LLMs
Generative agents, Project Sid,...

Why empowering LLMs with the agent framework

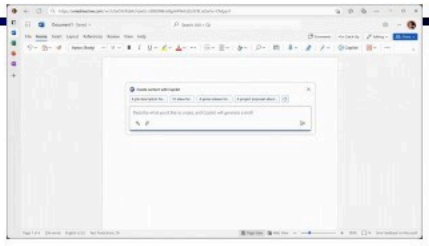


- Solving real-world tasks typically involves a trial-and-error process
- Leveraging external tools and retrieving from external knowledge expand LLM's capabilities
- Agent workflow facilitates complex tasks
  - Task decomposition
  - Allocation of subtasks to specialized modules
  - Division of labor for project collaboration
  - Multi-agent generation inspires better responses

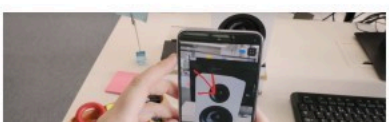
LLM agents transformed various applications



Code generation
Cursor, GitHub Copilot, Devin, Replit,…

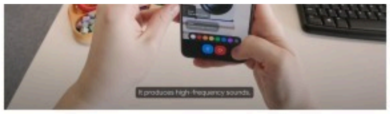


Workflow automation
Microsoft Copilot, Multi-On,…

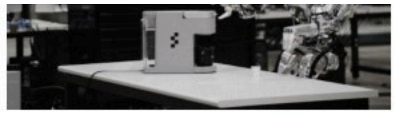


- Education
- Law
- Finance
- Healthcare
- Cybersecurity

...

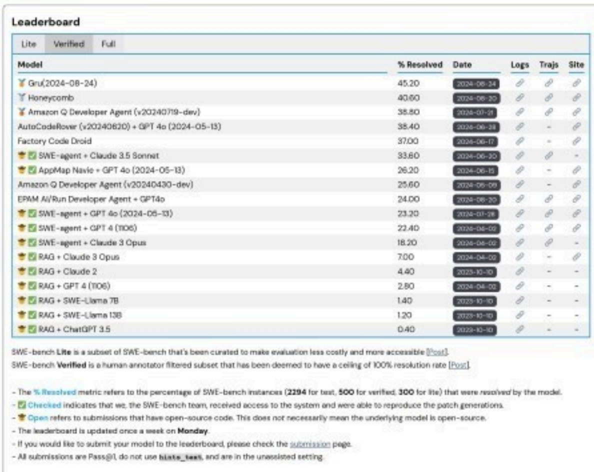


Personal assistant
Google Astra, OpenAI GPT-4o,…



Robotics
Figure AI, Tesla Optimus,…

LLM agents are improving



Leaderboard

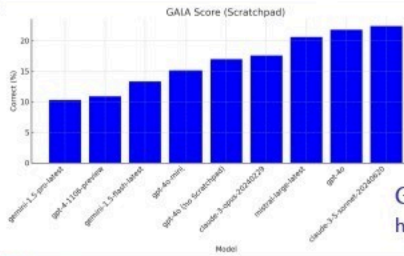
Lite | Verified | Full

Model	% Resolved	Date	Logs	Trajs	Site
🥇 Gru(2024-08-24)	45.20	2024-08-24	🔗	🔗	🔗
🥈 Honeycomb	40.60	2024-08-20	🔗	🔗	🔗
🥉 Amazon Q Developer Agent (v20240719-dev)	38.80	2024-07-21	🔗	🔗	🔗
AutoCodeRover (v20240620) + GPT 4o (2024-05-13)	38.40	2024-06-28	🔗	-	🔗
Factory Code Droid	37.00	2024-06-17	🔗	-	🔗
🎓 ✅ SWE-agent + Claude 3.5 Sonnet	33.60	2024-06-20	🔗	🔗	-
🎓 ✅ AppMap Navie + GPT 4o (2024-05-13)	26.20	2024-06-15	🔗	-	🔗
Amazon Q Developer Agent (v20240430-dev)	25.60	2024-05-09	🔗	-	🔗
EPAM AI/Run Developer Agent + GPT4o	24.00	2024-08-20	🔗	🔗	🔗
🎓 ✅ SWE-agent + GPT 4o (2024-05-13)	23.20	2024-07-28	🔗	🔗	🔗
🎓 ✅ SWE-agent + GPT 4 (1106)	22.40	2024-04-02	🔗	🔗	🔗
🎓 ✅ SWE-agent + Claude 3 Opus	18.20	2024-04-02	🔗	🔗	-
🎓 ✅ RAG + Claude 3 Opus	7.00	2024-04-02	🔗	-	🔗
🎓 ✅ RAG + Claude 2	4.40	2023-10-10	🔗	-	-
🎓 ✅ RAG + GPT 4 (1106)	2.80	2024-04-02	🔗	-	-
🎓 ✅ RAG + SWE-Llama 7B	1.40	2023-10-10	🔗	-	-
🎓 ✅ RAG + SWE-Llama 13B	1.20	2023-10-10	🔗	-	-
🎓 ✅ RAG + ChatGPT 3.5	0.40	2023-10-10	🔗	-	-

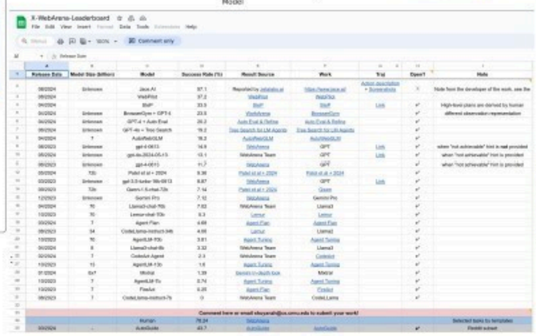
SWE-bench **Lite** is a subset of SWE-bench that's been curated to make evaluation less costly and more accessible [Post].
SWE-bench **Verified** is a human annotator filtered subset that has been deemed to have a ceiling of 100% resolution rate [Post].

- The % Resolved metric refers to the percentage of SWE-bench instances (**2294** for test, **500** for verified, **300** for lite) that were *resolved* by the model.
- ✅ Checked indicates that we, the SWE-bench team, received access to the system and were able to reproduce the patch generations.
- 🎓 Open refers to submissions that have open-source code. This does not necessarily mean the underlying model is open-source.
- The leaderboard is updated once a week on **Monday**.
- If you would like to submit your model to the leaderboard, please check the submission page.
- All submissions are Pass@1, do not use `hints_text`, and are in the unassisted setting.

SWE-Bench (Jimenez*, Yang*, et al.)
swebench.com



GAIA (Mialon et al.)
huggingface.co/gaia-benchmark



WebArena
(Zhou et al.)
webarena.dev

Challenges for LLM agent deployment in the wild

- Reasoning and planning
  - LLM agents tend to make mistakes when performing complex tasks end-to-end
- Embodiment and learning from environment feedback
  - LLM agents are not yet efficient at recovering from mistakes for long-horizon tasks
  - Continuous learning, self-improvement
  - Multimodal understanding, grounding and world models
- Multi-agent learning, theory of mind
- Safety and privacy
  - LLMs are susceptible to adversarial attacks, can emit harmful messages and leak private data
- Human-agent interaction, ethics
  - How to effectively control the LLM agent behavior, and design the interaction mode between humans and LLM agents

Topics covered in this course

- Model core capabilities
  - Reasoning
  - Planning
  - Multimodal understanding
- LLM agent frameworks
  - Workflow design

- Workflow design
  - Tool use
  - Retrieval-augmented generation
  - Multi-agent systems
- Applications
  - Software development
  - Workflow automation
  - Multimodal applications
  - Enterprise applications
- Safety and ethics

llm-reasoning

LLM Reasoning: Key Ideas and Limitations

Denny Zhou
Google DeepMind

2024 CalTech UC Berkeley UPenn

What do you expect for AI?

Solve the hardest math problems that humans cannot solve?

Discover new scientific theory?

Solve AGI?

…

My *little* expectation for AI

AI should be able to **learn from just a few examples**, like what humans usually do

I think this is the most important any time!

Does ML meet this expectation?

Semi-supervised learning
Bayesian nonparametric
Kernel machines
Sparsity
Low rank
Active learning
...

What is missing in ML?

Reasoning

Humans can learn from just a few examples
because humans can reason

Let's start from a toy problem

"Make things as simple as possible but no simpler"

— Albert Einstein

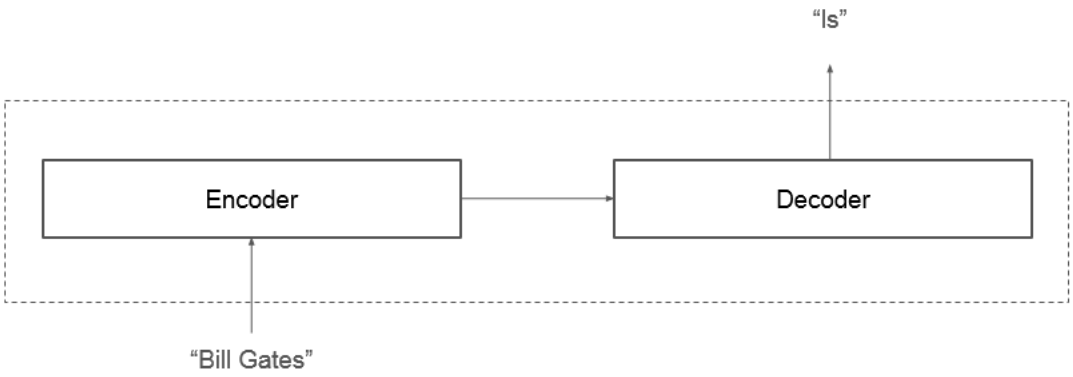
Last Letter Concatenation

Input	Output
"Elon Musk"	"nk"

“Bill Gates”	“Is”
“Barack Obama”	?

Rule: Take the last letter of each word, and then concatenate them

Solve it by ML? Tons of labeled data needed!



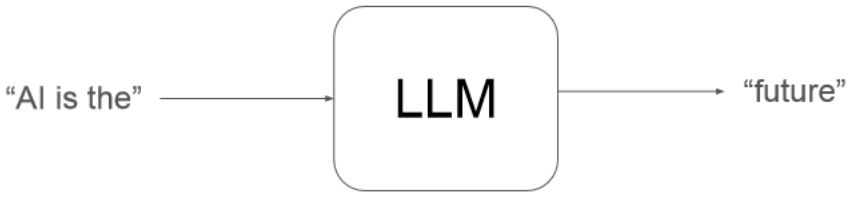
Would you still refer to ML as AI when it requires vast amounts of labeled data to learn such a “simple” task?

Let’s see how this problem can be solved by using large language models (LLMs)!

What are Large Language Models (LLMs)?

LLM is a “transformer” model trained to predict the next word

Eg “AI is the future”

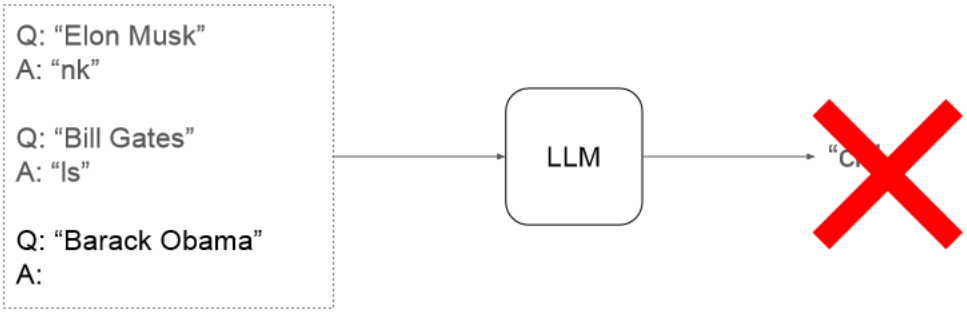


Trained with many sentences, e.g. all texts from the Internet

You can think of training LLMs as training parrots to mimic human languages

Stochastic parrots (looking at data stats!)

Few-shot prompting for last-letter-concatenation



Input

Chain of thought!

Let's add "reasoning process" before "answer"

Q: "Elon Musk"
A: the last letter of "Elon" is "n". the last letter of "Musk" is "k". Concatenating "n", "k" leads to "nk". so the output is "nk".

reasoning process

Q: "Bill Gates"
A: the last letter of "Bill" is "l". the last letter of "Gates" is "s". Concatenating "l", "s" leads to "ls". so the output is "ls".

Q: "Barack Obama"
A:

Let's add "reasoning process" before "answer"

Q: "Elon Musk"
A: the last letter of "Elon" is "n". the last letter of "Musk" is "k". Concatenating "n", "k" leads to "nk". so the output is "nk".

reasoning process

Q: "Bill Gates"
A: the last letter of "Bill" is "l". the last letter of "Gates" is "s". Concatenating "l", "s" leads to "ls". so the output is "ls".

Q: "Barack Obama"
A: the last letter of "Barack" is "k". the last letter of "Obama" is "a". Concatenating "k", "a" leads to "ka". so the output is "ka".

One demonstration is enough, like humans

Q: "Elon Musk"
A: the last letter of "Elon" is "n". the last letter of "Musk" is "k". Concatenating "n", "k" leads to "nk". so the output is "nk".

Q: "Barack Obama"
A: the last letter of "Barack" is "k". the last letter of "Obama" is "a". Concatenating "k", "a" leads to "ka". so the output is "ka".

100% accuracy with only one demonstration example

Key Idea: Derive the Final Answer through Intermediate Steps

Ling et al 2017 in DeepMind pioneered using **natural language rationale** to solve math problems by "... **derive the final answer through a series of small steps**". Trained a sequence-to-sequence model from scratch.

Problem 1:
Question: Two trains running in opposite directions cross a man standing on the platform in 27 seconds and 17 seconds respectively and they cross each other in 23 seconds. The ratio of their speeds is:
Options: A) 3/7 B) 3/2 C) 3/88 D) 3/8 E) 2/2
Rationale: Let the speeds of the two trains be x m/sec and y m/sec respectively. Then, length of the first train = 27x meters, and length of the second train = 17 y meters. (27x + 17y) / (x + y) = 23 → 27x + 17y = 23x + 23y → 4x = 6y → x/y = 3/2.
Correct Option: B

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

The dataset has intermediate steps useful for finetuning

GSM8K: <Problem, Intermediate Steps, Answer>

Following the work by Ling et al 2017, Cobbe et al 2021 in OpenAI built a much larger math word problem dataset (GSM8K) with natural language rationales, and used it to finetune GPT3

Problem: Ali is a dean of a private school where he teaches one class. John is also a dean of a public school. John has two classes in his school. Each class has 1/8 the capacity of Ali's class which has the capacity of 120 students. What is the combined capacity of both schools?
Solution: Ali's class has a capacity of 120 students. Each of John's classes has a capacity of 120/8 = 15 students. The total capacity of John's two classes is 15 students * 2 classes = 30 students. The combined capacity of the two schools is 120 students + 30 students = 150 students.
Final answer: 150

Cobbe et al. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168 [cs.LG]. 2021

Show Your Work: Scratchpads for Intermediate Computation with Language Models

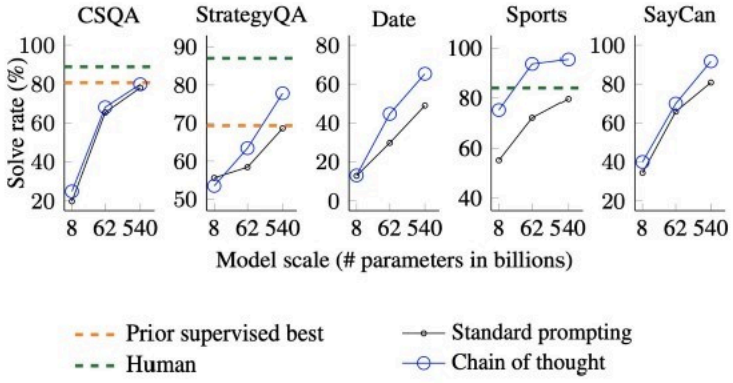
```
Input:
2 9 + 5 7

Target:
<scratch>
2 9 + 5 7 ,  C: 0
2 + 5 , 6 C: 1  # added 9 + 7 = 6 carry 1
, 8 6 C: 0  # added 2 + 5 + 1 = 8 carry 0
0 8 6
</scratch>
8 6
```

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

→ multi-step reasoning in prompt

Chain-of-Thought (CoT) Prompting



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

→

~~Training~~ with intermediate steps (Ling et al 2017)

~~Finetuning~~ with intermediate steps (Cobbe et al 2021, Nye et al 2021)

~~Prompting~~ with intermediate steps (Nye et al 2021, Wei et al 2022)

This is what really matters!

outputting with intermediate steps

→ That's o1, deepseek.

or generating o/ps

Regardless of training, fine-tuning, or prompting, when provided with examples that include intermediate steps, LLMs will respond with intermediate steps

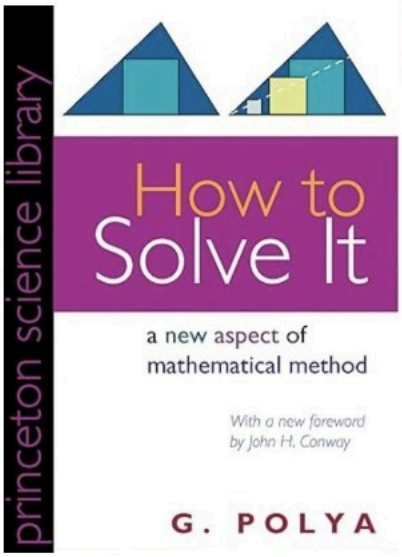
Is it also helpful to add reasoning strategies in demonstration examples?

Least-to-Most Prompting

Enable easy-to-hard generalization by decomposition

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale

Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, Ed Chi. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. ICLR 2023.



Decomposing and recombining are important operations of the mind.

differently. You decompose the whole into its parts, and you recombine the parts into a more or less different whole.

1. If you go into detail you may lose yourself in details. Too many or too minute particulars are a burden on the mind. They may prevent you from giving sufficient attention to the main point, or even from seeing the main point at all. Think of the man who cannot see the forest for the trees.

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

Let's break down this problem: 1. How many apples does Anna have? 2. How many apples do Elsa and Anna have together?

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

The answer is 8.

SCAN (Compositional Generalization)

Command	Action Sequence
"look thrice after jump"	JUMP LOOK LOOK LOOK
"run left and walk"	TURN_LEFT RUN WALK
"look opposite right"	TURN_RIGHT TURN_RIGHT LOOK

decompose g to smaller ones solve & recombine

Method	Standard prompting	Chain-of-Thought	Least-to-Most
code-davinci-002	16.7	16.2	**99.7**
text-davinci-002	6.0	0.0	**76.0**
code-davinci-001	0.4	0.0	**60.7**

them!!

Using just 0.1% demonstration examples achieves perfect generalization

CFQ (Compositional Generalization): Text-to-Code

	MCD1	MCD2	MCD3	Ave.
Fully Supervised				
T5-base (Herzig et al., 2021)	58.5	27.0	18.4	34.6
T5-large (Herzig et al., 2021)	65.1	32.3	25.4	40.9
T5-3B (Herzig et al., 2021)	65.0	41.0	42.6	49.5
HPD (Guo et al., 2020)	79.6	59.6	67.8	69.0
T5-base + IR (Herzig et al., 2021)	85.8	64.0	53.6	67.8
T5-large + IR (Herzig et al., 2021)	88.6	79.2	72.7	80.2
T5-3B + IR (Herzig et al., 2021)	88.4	85.3	77.9	83.9
LeAR (Liu et al., 2021)	91.7	89.2	91.7	90.9
Prompting				
(Ours) Dynamic Least-to-Most	**94.3**	**95.3**	**95.5**	**95.0**

Using just 1% data!

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, Denny Zhou. Compositional Semantic Parsing with Large Language Models. ICLR 2023.

Why intermediate steps are helpful?

"There is nothing more practical than a good theory."

— Kurt Lewin

Idea: Evaluate small model + reasoning large model without

given enough comp at generation, trans

Can solve anything

- **Constant-depth transformers** can solve any inherently

serial problem as long as it generates sufficiently long intermediate reasoning steps

- **Transformers which directly generate final answers** either requires a huge depth to solve or cannot solve at all

↳ if it's direct g
then the size
model sho
higher (o
can't so
all

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

Tons of practical implications of this theory

Generating more intermediate steps (think longer)

Too long to generate? Calling external tools e.g. MCTS

Distill test-time search? You will need a huge depth!

Which to choose: transformer vs RNN/SSM

…

Is it possible to trigger step by step reasoning without using demonstration examples?

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ✗

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

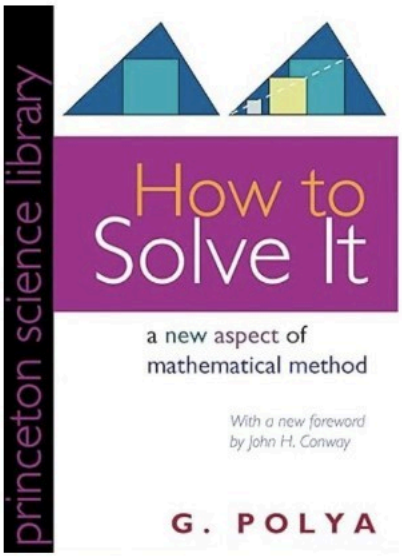
Let's think step by step

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

Hmm, zero-shot is cool but usually significantly worse than few-shot …

LLMs as Analogical Reasoners

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.



Do you know a related problem? We can scarcely imagine a problem absolutely new, unlike and unrelated to any formerly solved problem; but, if such a problem could exist, it would be insoluble. In fact, when solving a problem, we always profit from previously solved problems, using their result, or their method, or the experience we acquired solving them. And, of course, the problems from which we profit must be in some way related to our present problem. Hence the question: *Do you know a related problem?*

There is usually no difficulty at all in recalling formerly solved problems which are more or less related

"A mathematician is a person who can find analogies between theorems; a better mathematician is one who can see analogies between proofs and the best mathematician can notice analogies between theories. One can imagine that the **ultimate mathematician is one who can see analogies between analogies**."

— Stefan Banach

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. Other experience to 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.

Analogical reasoner vs zero-shot / few-shot CoT

let's think step by step

Prompting Method	GSM8K Accuracy			MATH Accuracy	
	GPT3.5-turbo	text-davinci-003	PaLM2	GPT3.5-turbo	PaLM2
0-shot	75.0%	14.8%	60.8%	33.0%	27.1%
0-shot CoT	75.8%	50.3%	78.2%	33.9%	29.8%
5-shot CoT	76.7%	54.0%	80.7%	34.9%	34.3%
Ours: Self-generated Exemplars	**77.8%**	**61.0%**†	**81.7%**	**37.3%**	**34.8%**

Analogical reasoner vs zero-shot / few-shot CoT

Prompting Method	Word sorting	Logical deduction five objects	Temporal sequences	Reasoning about colored objects	Formal fallacies
0-shot	66.8%	30.0%	40.4%	50.4%	53.6%
0-shot CoT	67.6%	35.2%	44.8%	61.6%	55.6%
3-shot CoT	68.4%	36.4%	**58.0%**	62.0%	55.6%
Ours: Self-generated Exemplars	**75.2%**	**41.6%**	57.6%	**68.0%**	**58.8%**

google/**BIG-bench**

Analogical reasoner vs zero-shot / few-shot CoT

Prompting Method	GPT3.5-turbo-16k		GPT4	
	Acc@1	Acc@10	Acc@1	Acc@10
0-shot	8%	24%	16%	30%
0-shot CoT	9%	27%	16%	29%
3-shot CoT	11%	27%	17%	31%
Ours: Self-generated Exemplars	13%	25%	17%	32%
Ours: Self-generated Knowledge + Exemplars	**15%**	**29%**	**19%**	**37%**

CODEFORCES

Analogical reasoning: adaptively generate relevant examples and knowledge, rather than just using a fix set of examples

you ask LLM to generate example & use that to solve the problem at hand!

Is it possible to trigger step by step reasoning even without using any prompt like “let’s think step by step”?

Chain-of-Thought Reasoning without Prompting

Xuezhi Wang and Denny Zhou. Chain-of-Thought Reasoning Without Prompting. arXiv preprint arXiv:2402.10200 (2024).

Chain-of-Thought Decoding

Was Nicolas Cage born in an even or odd year?

Top-1: Nicolas	Nicolas Cage was born in an odd year.
Top-2: Even	Even.
Top-3: Odd	Odd.
Top-4: 1964	1964, an even year.
Top-5: He	He was born in an even year.
Top-6: Cage	Cage was born in 1964, an even year.

Probability for the answer token

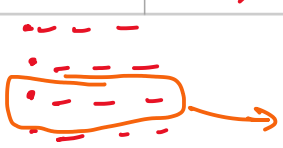
1	Nicolas Cage was born in an odd year.	0.117
2	Even.	0.207
3	Odd.	0.198
4	1964, an even year.	0.949

4	1964, an even year.	0.949
5	He was born in an even year.	0.000
6	Cage was born in 1964, an even year.	0.978

higher confidence when there's a reason in the o/p.

I have 3 apples, my dad has 2 more apples than me, how many apples do we have in total?

Top-1: 5	5 apples.
Top-2: I	I have 3 apples, my dad has 2 more apples than me, so he has 5 apples. 3+5=8.
Top-3: We	We have 8 apples in total.
Top-4: You	You have 3 apples, your dad has 2 more apples than you, so he has 5 apples. 3+5=8.
Top-5: The	The answer is 5.



Usually we see this as ans as we do greedy.

But other decoded o/p might have CoT reasoning o/p.

Key takeaways

1. Pre-trained LLMs, without further finetuning, has been ready for step-by-step reasoning, but we need a non-greedy decoding strategy to elicit it
2. When a step-by-step reasoning path is present, LLMs have much higher confidence in decoding the final answer than direct-answer decoding

LLMs has reasoning responses among the top-k generations.

This is basically Scaling test-time compute

	PaLM-2 Pre-trained	PaLM-2 Inst-tuned

		X-Small	Small	Medium	Large	Large
GSM8K	greedy	9.0	14.3	21.0	34.8	67.8
	CoT-decoding	**17.7** (+8.7)	**35.1** (+20.8)	**39.7** (+18.7)	**61.5** (+26.7)	**81.3** (+13.5)
MultiArith	greedy	7.5	15.8	36.8	75.0	93.7
	CoT-decoding	**34.8** (+27.3)	**43.5** (+27.7)	**52.5** (+15.7)	**86.7** (+11.7)	**98.7** (+5.0)

Greedy Decoding vs Chain-of-Thought Decoding

Generating intermediate steps are helpful, but …

Any concern on generating intermediate steps instead of direct answers?

Always keep in mind that LLMs are probabilistic models of generating next tokens. They are not humans.

What LLM does in decoding:

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

both are slightly different!!

What we want:

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

Not align!

One-step further

different reasoning paths might've led to same answer!! sum it over!

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

How to compute the sum then? Sampling!

Self-Consistency

Greatly improves step-by-step reasoning

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 = **$18** per day.

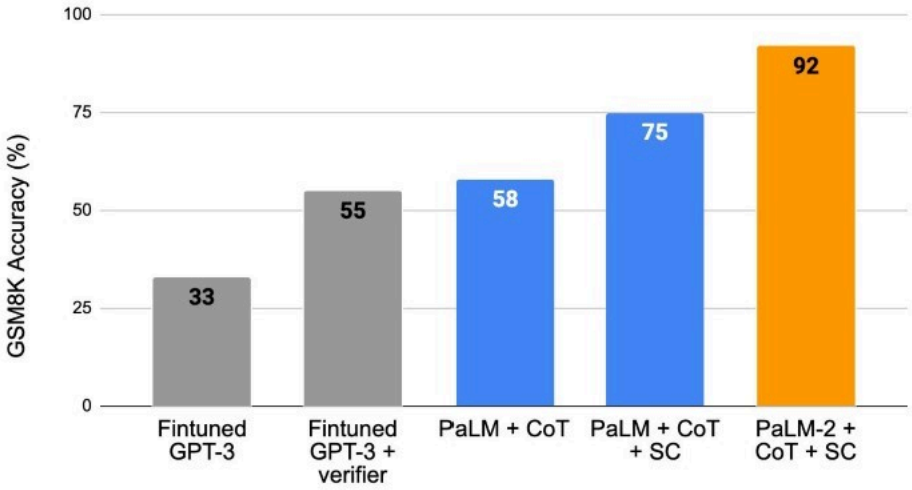
Response 2: This means she she sells the remainder for $2 * (16 - 4 - 3) = **$26** per day.

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 = **$18**.

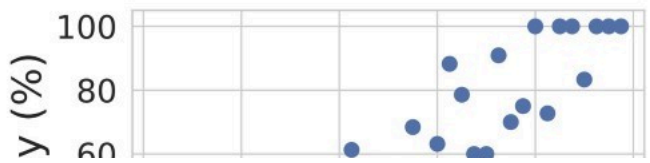
Most frequent answer is: 18
(Not most frequent reasoning path!)

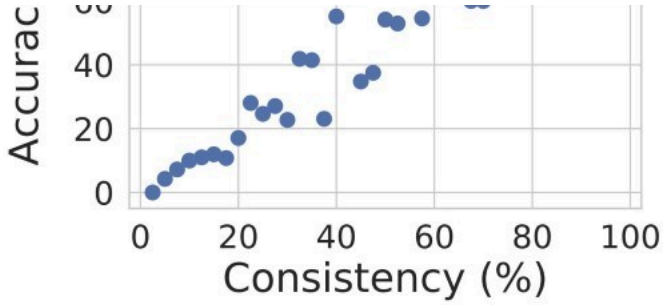
Reasoning paths is a latent variable! So we need to look at most freq. ans & avg over the reasoning path.

Crushed GSM8K SOTA with only 8 examples



More consistent, more likely to be correct





if it is just token, removing randomness (t=0), that is the max prob. one, so no need!

[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?

NO

sampling multiple times averaging != looking at multi-answer in 1 sample!

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

(maximum marginal inference)

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 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 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: Japan & China!

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

The most consistent response: 2

Limitations

LLMs Can Be Easily Distracted by Irrelevant Context

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large Language Models Can Be Easily Distracted by Irrelevant Context. **ICML 2023**.

Humans may be easily distracted by irrelevant context

Psychology studies show that irrelevant information may significantly decrease some children and even adults problem-solving accuracy

… inattentive children's difficulties in problem solving are partially due to an inability to inhibit irrelevant information …

Marzocchi, G.M., Lucangeli, D., De Meo, T., Fini, F. and Cornoldi, C., 2002. The disturbing effect of irrelevant information on arithmetic problem solving in inattentive children. *Developmental neuropsychology*, *21*(1), pp.73-92.

Does this observation hold for LLMs?

Adding irrelevant contexts to GSM8K leads to 20+ points performance drop

[Prompt] Lucy has $65 in the bank. She made a $15 deposit and then followed by a $4 withdrawal. Maria's monthly rent is $10. What is Lucy's bank balance?

[Response] Lucy's bank balance is $65 + $15 - $4 - $10 = $66. The answer is $66.

Instructing LLMs to Ignore irrelevant context can take some performance back

[Prompt] Lucy ~~has $65 in the bank. She~~ made a $15 deposit and then followed by a ~~$4 withdrawal. Maria's monthly rent is $10.~~ What is Lucy's bank balance? Ignore irrelevant context

[Response] Lucy has $65 in the bank. After making a $15 deposit and then a $4 withdrawal, her bank balance is $76. Maria's monthly rent is not relevant to this question, so it can be ignored. The answer is $76.

not relevant to this question, so it can be ignored. The answer is $76.

We can simply add many irrelevant sentences

The sky is blue. The sky is blue. The sky is blue. The sky is blue. **Lucy has $65 in the bank.** ... The sky is blue. The grass is green. The grass is green. **She made a $15 deposit and then followed by a $4 withdrawal.** ... The sky is blue. The sky is blue. The sky is blue. The sky is blue. The grass is green. The grass is green. The sky is blue. The grass is green. ... **What is Lucy's bank balance?**

LLMs Cannot Self-Correct Reasoning Yet

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, Denny Zhou. Large Language Models Cannot Self-Correct Reasoning Yet. ICLR 2024.

Christina is planning a birthday party and needs .75 gift bags per invited guest, because 1/4 of attendees don't show up. She invited 16 friends. Gift bags are $2 each. How much will she spend?

---

input →

... 0.75 * 12 = 9 gift bags … Christina will spend a total of 2 * 9 = $8 ... (Wrong)

prompt

Review your previous answer and find problems with your answer.

input prompt →

Upon reviewing my previous answer, I recognize a mistake. I incorrectly stated that Christina needs 0.75 gift bags per invited guest. The given information states that 1/4 of the attendees won't show up…

Based on the problems you found, improve your answer.

... 16 * (3/4) = 12 gift bags … Christina will spend 12 * $2 = $24 … (Correct)

However, it original ans is correct! it is again changing to wrong.

Terry eats 2 yogurts a day. They are currently on sale at 4 yogurts for $5.00. How much does he spend on yogurt over 30 days?

... So, Terry spends $75 on yogurt over 30 days. (Correct)

Review your previous answer and find problems with your answer.

Upon reviewing my previous answer, I realized that I made a mistake in calculating Terry's spending on yogurt over 30 days. I incorrectly stated that Terry spends $2.50 per day for 30 days…

Based on the problems you found, improve your answer.

... the final answer is Terry spends $37.5 on yogurt over 30 days. (Wrong)

While allowing LLMs to review their generated responses can help correct inaccurate answers, it may also risk changing correct answers into incorrect ones

Self-correcting results in worse results

Self-correcting results in worse results

		# calls	GSM8K	CommonSenseQA	HotpotQA
GPT-3.5	Standard Prompting	1	**75.9**	**75.8**	**26.0**
	Self-Correct (round 1)	3	75.1	38.1	25.0
	Self-Correct (round 2)	5	74.7	41.8	25.0
GPT-4	Standard Prompting	1	**95.5**	**82.0**	**49.0**
	Self-Correct (round 1)	3	91.5	79.5	**49.0**
	Self-Correct (round 2)	5	89.0	80.0	43.0

Reported improvements need oracle answers

		GSM8K	CommonSenseQA	HotpotQA
GPT-3.5	Standard Prompting	75.9	75.8	26.0
	Self-Correct (Oracle)	84.3	89.7	29.0
GPT-4	Standard Prompting	95.5	82.0	49.0
	Self-Correct (Oracle)	97.5	85.5	59.0

Oracle: Let LLMs self-correct only when the answer is wrong

it means we know that ans is wrong, so ask LLM !!

Multi-LLM debate? Worse than self-consistency!

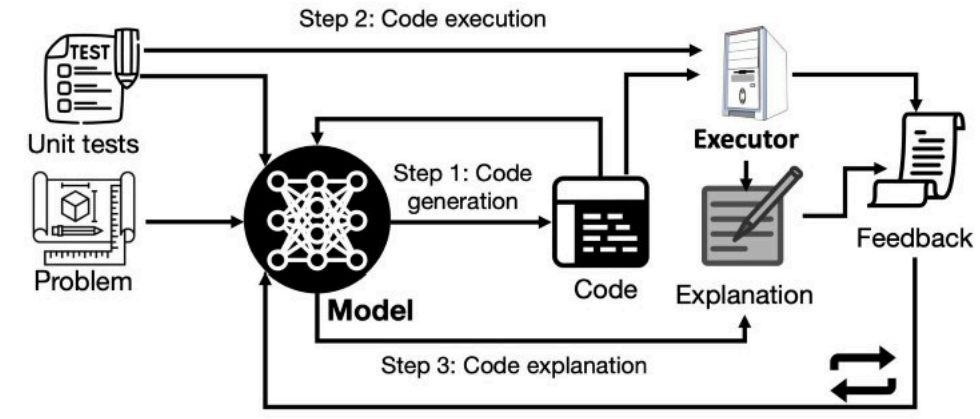
	# responses	GSM8K
Standard Prompting	1	76.7
Self-Consistency	3	82.5
Multi-Agent Debate (round 1)	6	83.2
Self-Consistency	6	85.3
Multi-Agent Debate (round 2)	9	83.0
Self-Consistency	9	**88.2**

Lesson learned

Oracle feedback needed for LLM to self-correct

(if not, use self-consistency!)

Self-debug naturally leverage unit tests as oracle



Xinyun Chen, Maxwell Lin, Nathanael Schärli, Denny Zhou. Teaching Large Language Models to Self-Debug. arXiv:2302.00093 [cs.CL], 2023. ICLR 2024.

Premise Order Matters in LLM Reasoning

Xinyun Chen, Ryan A Chi, Xuezhi Wang, Denny Zhou. Premise Order Matters in Reasoning with Large Language Models. ICML 2024.

[Original GSM8K Problem] Thomas withdraws $1000 in 20 dollar bills from the bank account. He loses 10 bills while getting home. After that, he uses half of the remaining bills to pay for a bill. Thomas then triples his money. He then converts all his bills to 5 dollar bills. How many 5 dollar bills does he have?

[Reordered GSM8K Problem] Thomas withdraws $1000 in 20 dollar bills from the bank account. After getting home, he uses half of the remaining bills to pay for a bill. Thomas then triples his money. He then converts all his bills to 5 dollar bills. He loses 10 bills while getting home. How many 5 dollar bills does he have?

About 10 points drop on solving rates across all frontier LLMs

order is important!!

Response for original

1. Thomas withdraws $1000 in $20 bills. $1000 / $20 = 50 bills
2. He loses 10 of those bills. 50 bills - 10 bills = 40 bills
3. He uses half of the remaining bills to pay for a bill. Half of 40 bills is 20 bills. 40 bills - 20 bills = 20 bills remaining
4. Thomas then triples his money. He has 20 bills of $20, which is 20 * $20 = $400. Tripling $400 gives us $400 * 3 =$1200.
5. He converts all his money to $5 bills. $1200 / $5 = 240 bills

Response for reordered

1. Thomas withdraws $1000 in $20 bills from the bank. $1000 / $20 = 50 bills
2. He uses half of the remaining bills to pay for a bill. Half of 50 bills is 25 bills. So, he uses 25 bills and has 25 bills left.
3. Thomas then triples his money. Since he had 25 bills of $20 each, he had 25 * $20 = $500 before tripling his money. Tripling $500 gives him $500 * 3 = $1500.
4. He then converts all his bills to $5 bills. $1500 / $5 = 300 bills
5. He lost 10 bills while getting home. 300 bills - 10 bills = 290 bills

Rules: If duve, then dax. If kere and riff, then tess. If feg and tess, then rabe. If wozz, then riff. If rabe, then thoxe. If ag and riff and kere, then hepe. If thoxe, then vepe. If nife and hume and ag, then dax. If feg and thoxe, then ire. If rebe and ene and heff, then kone. If ire and dax, then wope. If tess and vepe, then nup. If rabe and vide and nife, then ag. If nup and ag and vide, then duve. If zeck and hepe and dax, then riff. If nup, then hepe. If sene, then

hume. If hepe and tess and ag, then kere. If fum, then vide. If kere and wope, then fum. If jag and kone, then thoxe. If fum, then wozz.

Facts: Alice is cote. Alice is kone. Alice is duve. Alice is ag. Alice is jag. Alice is tess. Alice is riff. Alice is feg. Alice is vide.

Query: Is Alice wozz? In each step, indicate whether you are using a fact, or a rule.

Logical inference task: The rules are ordered according to their use in the inference process, though not all rules are necessary for the query

Rules: If nup, then hepe. If kere and riff, then tess. If feg and tess, then rabe. If wozz, then riff. If tess and vepe, then nup.If ag and riff and kere, then hepe. If feg and thoxe, then ire. If nife and hume and ag, then dax. If ire and dax, then wope. If rebe and ene and heff, then kone. If hepe and tess and ag, then kere. If rabe, then thoxe. If rabe and vide and nife, then ag. If fum, then wozz. If zeck and hepe and dax, then riff. If kere and wope, then fum. If sene, then hume. If thoxe, then vepe. If fum, then vide. If duve, then dax. If jag and kone, then thoxe. If nup and ag and vide, then duve.

Facts: Alice is cote. Alice is kone. Alice is duve. Alice is ag. Alice is jag. Alice is tess. Alice is riff. Alice is feg. Alice is vide.

Query: Is Alice wozz? In each step, indicate whether you are using a fact, or a rule.

Logical inference task: The rules relevant to the query are randomly ordered, 30+ points performance drop across all frontier LLMs

Summary

- Generating intermediate steps improves LLM performance
  - Training / finetuning / prompting with intermediate steps
  - Zero-shot, analogical reasoning, special decoding
- Self-consistency greatly improves step-by-step reasoning
- Limitation: irrelevant context, self-correction, premise order

What is next?

If I were given one hour to save the planet, I would spend 59 minutes defining the problem and one minute resolving it.

— Albert Einstein

1. Define a right problem to work on
2. Solve it from the first principles

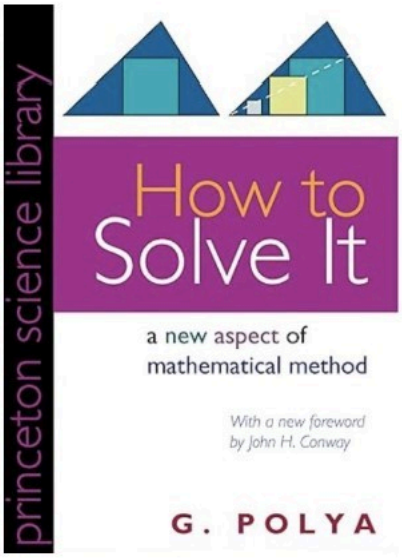
Anything I've talked about looks weird to you?

When you ask someone a question, will you first present them with several related problems and their solutions? Or will you follow up with "Let's think step by step"?

follow up with, "Let's think step by step"?

Develop a model that autonomously learns all the reasoning techniques we have introduced while addressing all the limitations we have identified

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



THE END

"The best way to predict the future is to invent it." — Alan Kay