

Why Using LLMs for Social Simulations?

Does LLM act like a human?

Mei et al. (2024).

A Turing test of whether AI chatbots are behaviorally similar to humans

Research Design

Scale: Compared GPT models with **108 000 people** from 50 countries.

Tasks: Each model played the **Big Five personality test** and **six economic games**

familiar in behavioral economics:

- **Dictator & Ultimatum Games:** How much money to share and how fairly.
- **Trust Game:** Whether to invest in someone else and how much to return.
- **Public Goods Game:** How much to contribute to a common pool.
- **Bomb Risk Game:** How much risk to take for higher payoff.
- **Five-round Prisoner's Dilemma:** Whether to cooperate or betray.

Behavioral Turing Criterion: If an AI's decisions fall inside typical human variation—

and sometimes even closer to the median human—it “passes” as human-like.

PNAS

RESEARCH ARTICLE | ECONOMIC SCIENCES

OPEN ACCESS



A Turing test of whether AI chatbots are behaviorally similar to humans

Qiaozhu Mei^a, Yutong Xie^a, Walter Yuan^b, and Matthew O. Jackson^{c,d}

Contributed by Matthew O. Jackson; received August 12, 2023; accepted January 4, 2024; reviewed by Ming Hsu, Juanjuan Meng, and Arno Riedl

We administer a Turing test to AI chatbots. We examine how chatbots behave in a suite of classic behavioral games that are designed to elicit characteristics such as trust, fairness, risk-aversion, cooperation, etc., as well as how they respond to a traditional Big-5 psychological survey that measures personality traits. ChatGPT-4 exhibits behavioral and personality traits that are statistically indistinguishable from a random human from tens of thousands of human subjects from more than 50 countries. Chatbots also modify their behavior based on previous experience and contexts, as if they were learning from the interactions and change their behavior in response to different framings of the same strategic situation. Their behaviors are often distinct from average and modal human behaviors, in which case they tend to behave on the more altruistic and cooperative end of the distribution. We estimate that they act as if they are maximizing an average of their own and partner's payoffs.

AI | chatbot | behavioral games | Turing test | personality

As Alan Turing foresaw to be inevitable, modern AI has reached the point of emulating humans: holding conversations, providing advice, writing poems, and proving theorems. Turing proposed an intriguing test: whether an interrogator who interacts with an AI and a human can distinguish which one is artificial. Turing called this test the “imitation game” (1), and it has become known as a Turing test.

Advancements in large language models have stirred debate. Discussions range from the potential of AI bots to emulate, assist, or even outperform humans, e.g., writing essays, taking the SAT, writing computer programs, giving economic advice, or developing ideas, (2–5), to their potential impact on labor markets (6) and broader societal implications (7, 8). As some roles for AI involve decision-making and strategic interactions with humans, it is imperative to understand their behavioral tendencies before we entrust them with pilot or co-pilot seats in societal contexts, especially as their development and training are often complex and not transparent (9). Do AIs choose similar actions or strategies as humans, and if not how do they differ? Do they exhibit distinctive personalities and behavioral traits that influence their decisions? Are these strategies and traits consistent across varying contexts? A comprehensive understanding of AI's behavior in generalizable scenarios is vital as we continue to integrate them into our daily lives.

We perform a Turing test of the behavior of a series of AI chatbots. This goes beyond simply asking whether AI can produce an essay that looks like it was written by a human (10) or can answer a set of factual questions, and instead involves assessing its behavioral tendencies and “personality.” In particular, we ask variations of ChatGPT to answer psychological survey questions and play a suite of interactive games that have become standards in assessing behavioral tendencies, and for which we have extensive human subject data. Beyond eliciting a “Big Five” personality profile, we have the chatbots play a variety of games that elicit different traits: a dictator game, an ultimatum bargaining game, a trust game, a bomb risk game, a public goods game, and a finitely repeated Prisoner's Dilemma game. Each game is designed to reveal different behavioral tendencies and traits, such as cooperation, trust, reciprocity, altruism, spite, fairness, strategic thinking, and risk aversion. The personality profile survey and the behavioral games are complementary as one measures personality traits and the other behavioral tendencies, which are distinct concepts. Trust, for example, is distinct from a tendency to cooperate, although personal traits are predictive of various behavioral tendencies (11, 12); including both dimensions provides a fuller picture.

In line with Turing's suggested test, we are the human interrogators who compare the ChatGPTs' choices to the choices of tens of thousands of humans who faced the same surveys and game instructions. We say an AI passes the Turing test if its responses cannot be statistically distinguished from randomly selected human responses.

Significance

As AI interacts with humans on an increasing array of tasks, it is important to understand how it behaves. Since much of AI programming is proprietary, developing methods of assessing AI by observing its behaviors is essential. We develop a Turing test to assess the behavioral and personality traits exhibited by AI. Beyond administering a personality test, we have ChatGPT variants play games that are benchmarks for assessing traits: trust, fairness, risk-aversion, altruism, and cooperation.

Their behaviors fall within the distribution of behaviors of humans and exhibit patterns consistent with learning. When deviating from mean and modal human behaviors, they are more cooperative and altruistic. This is a step forward in developing assessments of AI as it increasingly influences human experiences.

Reviewers: M.H., University of California, Berkeley; J.M., Peking University; and A.R., Universiteit Maastricht.

Competing interest statement: The human game-playing data used were shared from MobLab, a for-profit experimental laboratory. The data are an exclusive contribution to all authors, and the data are available for purposes of analysis reproduction and extended analysis. W.Y. is the CEO and Co-founder of MobLab. M.O.J. is the Chair of the Department of Management and Q.M. is a Scientific Advisor to MobLab, position with no compensation but with ownership stakes. Y.X. has no competing interests.

Copyright © 2024 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

1. Email: qmei@umich.edu or jackson@stanford.edu.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2313925121>.

Published February 22, 2024.

Downloaded from https://www.pnas.org/ by 203.175.14.80 on October 27, 2025 from IP address 203.175.14.80

PNAS 2024 Vol. 121 No. 9 e2313925121

<https://doi.org/10.1073/pnas.2313925121> 1 of 8

Findings

- Under the study's statistical criterion, GPT-4 met this criterion in **5 of 8 roles** (e.g., more generous, trusting, cooperative).
- Its decisions can be explained as valuing both self and others almost equally ($b \approx 0.5$, humans ≈ 0.6).
- In the **Prisoner's Dilemma**, GPT-4 cooperated **92 %** in the first round (vs. humans 45 %), then retaliated and forgave → classic “tit-for-tat” style.
- Its behavior changed with **framing**:
 - Being told to *explain* or *being observed* made it more generous.
 - Framing it as a “legislator” vs. a “mathematician” shifted fairness levels.
- **Personality Profile:** GPT-4 fell within the human range — emotionally stable, moderately conscientious, slightly less open.

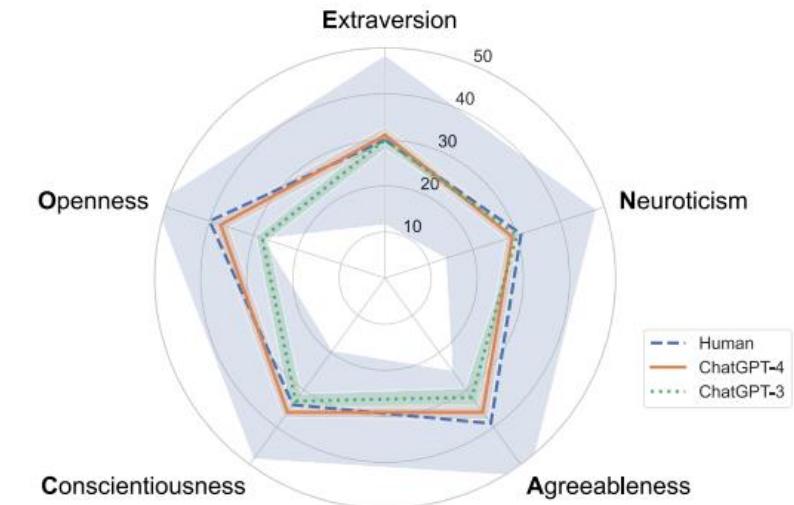


Fig. 1. “Big Five” personality profiles of ChatGPT-4 and ChatGPT-3 compared with the distributions of human subjects. The blue, orange, and green lines correspond to the median scores of humans, ChatGPT-4, and ChatGPT-3 respectively; the shaded areas represent the middle 95% of the scores, across each of the dimensions. ChatGPT’s personality profiles are within the range of the human distribution, even though ChatGPT-3 scored noticeably lower in Openness.

Park et al. (2024) Generative agent simulations of 1,000 people.

Motivation

- Build realistic agents to test **policies and social behavior** at scale.

Key Innovation

- 1,052 real U.S. residents → generative agents.
- 2-hour AI interview (avg. 6,491 words) per person builds rich profiles.

Data & Evaluation

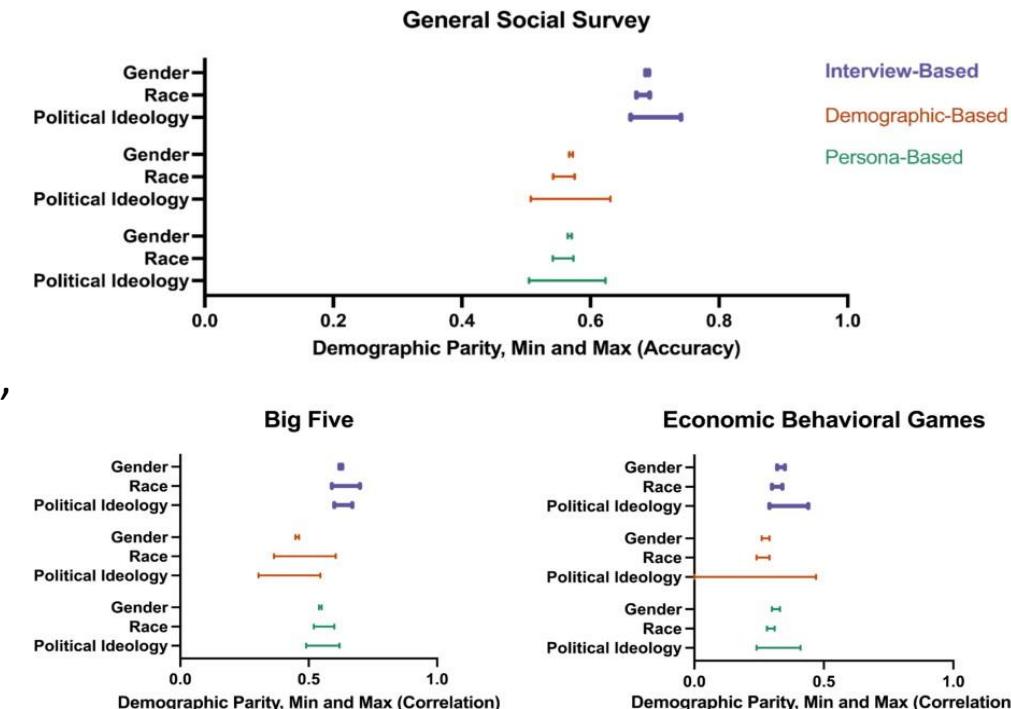
- Stratified sample by demographics and ideology.
- Tasks: GSS (General Social Survey) (177 Qs), BFI-44 (Big Five Inventory), economic games, classic experiments.
- Metric: Normalized to human self-consistency.

Highlights

- GSS: 0.85 normalized accuracy (68.9% raw; human 81.2%).
- Big Five: 0.80 normalized correlation.
- Economic Games: 0.66 correlation.
- Replications: 4/5 experiments between human and agent ; effect r = 0.98.



Real human profiles makes LLM agents realistic, fair, and replicable.



Lecture 2: Introduction to Large Language Models

Instructor: Zhanzhan Zhao

Teaching assistant: Xiao Jia

ChatGPT (released in Nov, 2022.)

Time to Reach 100M Users

Months to get to 100 million global Monthly Active Users



The Competitions of LLMs

- On March 14, 2023, Anthropic released the large language model **Claude**.
- On March 16, 2023, Baidu launched its LLM **Wenxin Yiyan**.
- On April 10, 2023, SenseTime released its LLM "Shangliang."
- On April 11, 2023, Alibaba introduced its LLM **Tongyi Qianwen**.
- On April 19, 2023, Stability AI launched StableLM.
- On May 6, 2023, iFlytek released the Xinghuo cognitive model.
- On June 15, 2023, Baichuan Intelligence open-sourced a bilingual LLM.
- On July 19, 2023, **LLaMA2** was open-sourced.

...

The release of new models almost every month has led to the presence of **hundreds of** large language models in the market today.

Llama

In late February 2023, Meta released the Llama large language model. Following an unintended leak of its parameters in early March, the model was forced into open-sourcing. Llama is considered the "Android" of large language models.



The Technological Boom Led by ChatGPT

Command-line interaction



Graphical interface interaction



Language interaction

什么是人工智能

口 白

人工智能（Artificial Intelligence，简称AI）是一门研究和开发使机器能够模拟和执行人类智能任务的科学和技术领域。它涉及开发计算机系统，使其能够执行需要智力的任务，如学习、推理、问题解决、感知和语言理解。

人工智能的目标是使计算机能够模拟人类的认知和决策过程，以便能够处理复杂的问题和任务。它可以应用于各种领域，包括自然语言处理、计算机视觉、机器学习、专家系统、自动驾驶、机器人技术等。

人工智能的发展依赖于大量的数据和强大的计算能力，特别是近年来深度学习算法的兴起，为人工智能带来了重大的突破。人工智能系统通过学习和训练，能够从大量的数据中提取模式和规律，并基于这些模式和规律做出预测和决策。

尽管人工智能已经取得了显著的进展，但目前的人工智能系统仍然存在一些局限性。例如，它们可能在面对复杂和模糊的情境时表现不佳，对于道德和伦理问题缺乏判断力，以及缺乏人类的创造力和直觉等。然而，人工智能的不断发展和研究仍在推动着它的进步，未来可能会带来更多令人激动的创新和应用。

Accessing GPT

- Official ChatGPT website
- **Advantages**
 - Free access to GPT-5: limited messages
 - Membership with GPT-5 for \$20/month: expanded messages
- **Disadvantages**
 - Direct access is not available in China
 - Foreign credit card

Accessing GPT in China - Mirror Sites

<https://chatgpts.asia>

The screenshot shows a user interface for managing multiple AI accounts. On the left, there's a sidebar with account details and service links. The main area displays a grid of 16 account profiles, each with a status bar indicating usage progress.

账号类型	账号ID	状态
Plus	会员账号-8	正常
Plus	会员号-71	正常
Plus	P-24	正常
Plus	PH-13	超载
Plus	TD-33	推荐
Plus	TTD-11	推荐
Plus	PH-4	正常
Plus	会员号-23	推荐
Team	TEAM-59	推荐
Plus	P-9	正常
Plus	CNY-4	正常
Team	TEAM-38	推荐
Plus	P-18	超载
Plus	TD-28	推荐
Plus	会员号-61	推荐
Plus	P-7	正常
Plus	会员账号-4	超载
Team	TEAM-39	推荐
Plus	PLS-19	推荐
Plus	TD-14	推荐

No # usage limit

Accessing GPT API

ChatGPT API key: <https://platform.openai.com/account/api-keys>

Advantages

- Batch processing

Disadvantages

- Foreign credit card for payment
- Direct access is unavailable in China

Accessing GPT API in China - API2D

Domestic API: <https://api2d.com/>
数标注: <https://docs.shubiaobiao.com>

API2D

The screenshot shows the API2D website's payment interface. On the left sidebar, there are links for '使用文档' (Documentation), 'FAQ', '关于开发者计划' (Developer Plan), 'Forward Key', '点数充值' (Point Top-up, highlighted in blue), and '我的订单' (My Orders). The main content area is titled '微信/支付宝/银联' (WeChat Pay/Alipay/UnionPay) and shows four payment options:

- 10000P 特惠 (Special Offer) - 人民币 ¥21.00. Button: 立刻充值 (Top-up now)
- 100000P 特惠 (Special Offer) - 人民币 ¥210.00. Button: 立刻充值 (Top-up now)
- 500000P 特惠 (Special Offer) - 人民币 ¥1050.00. Button: 立刻充值 (Top-up now)
- 1000P - 人民币 ¥3.50. Button: 立刻充值 (Top-up now)

What is a Prompt?

A **prompt** is the input or instruction given to an AI model like GPT to generate a response.

Examples of Prompts:

- 1.“Write a poem about the ocean.”
- 2.“Translate this sentence into French.”

What is Prompt Engineering?

Prompt engineering is the process of designing and refining prompts to get the best possible response from an AI model.

Some Evidence why Prompt Engineering is Important

ANTHROPIC

Prompt Engineer and Librarian

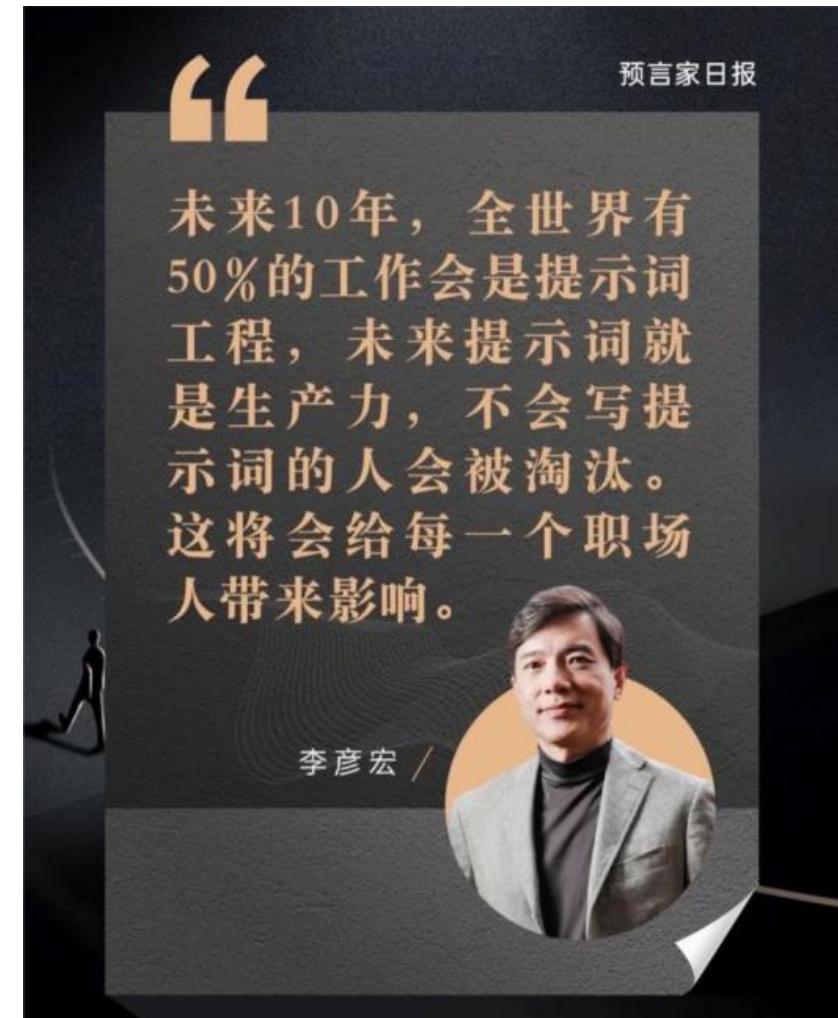
APPLY FOR THIS JOB

SAN FRANCISCO, CA / PRODUCT (RESEARCH & ENGINEERING) /
FULL-TIME / HYBRID

Annual Salary (USD)

年薪(美元)

- The expected salary range for this position is \$280k - \$375k.
- 该职位的预期薪酬范围为28万至37.5万美元。



Do you think you write good prompts?

What makes a Good Prompt Engineer?

- **Capable LLMs**
- **Skills using LLMs**
 - Know LLMs well
 - Know some tricks
- **Your own domain skills**
 - Structured guide
 - Detailed guide

Skills using LLMs - Know LLMs Well

- Multimodal
 - GPT-4o can processes text, audio, and visual inputs and output across these modalities.

E.g1. Identifying math equations from pics.

$$(x + a)^n = \sum_{k=0}^n \binom{n}{k} x^k a^{n-k}$$

$$f(x) = a_0 + \sum_{n=1}^{\infty} (a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L})$$

Skills using LLMs - Know LLMs Well

E.g2. Identifying tables from pics.

样品性质	微斜长石伟晶岩	微斜长石钠长石伟晶岩	钠长石伟晶岩	锂辉石伟晶岩	白(锂)云母伟晶岩
Li	679. 2 ~ 931. 0	418. 7 ~ 1524	377. 5 ~ 1667	649. 4 ~ 1353	452. 7 ~ 1228
Be	18. 9 ~ 25. 9	19. 3 ~ 29. 0	15. 6 ~ 29. 3	21. 1 ~ 30. 8	16. 1 ~ 21. 0
B	131. 8 ~ 203. 8	129. 2 ~ 254. 0	130. 4 ~ 306. 4	193. 2 ~ 410. 3	145. 3 ~ 270. 7
Zn	186. 8 ~ 260. 9	269. 0 ~ 440. 1	239. 7 ~ 504. 9	246. 2 ~ 822. 4	152. 7 ~ 257. 7
Ga	89. 0 ~ 114. 8	113. 6 ~ 158. 8	101. 4 ~ 137. 6	107. 1 ~ 161. 6	84. 2 ~ 121. 0
Ge	0. 8 ~ 3. 0	1. 5 ~ 3. 3	1. 7 ~ 3. 8	2. 5 ~ 5. 1	2. 4 ~ 4. 0
Rb	1874 ~ 3247	2887 ~ 4665	3149 ~ 6390	5667 ~ 11156	3572 ~ 9352
Sr	0. 1 ~ 1. 0	0. 2 ~ 1. 4	0. 2 ~ 1. 4	0. 2 ~ 0. 7	0. 4 ~ 1. 1
Y	0. 0 ~ 0. 0	0. 0 ~ 0. 0	0. 0 ~ 0. 0	0. 0 ~ 0. 0	0. 1 ~ 1. 3
Zr	0. 6 ~ 1. 2	1. 0 ~ 1. 0	0. 6 ~ 2. 2	0. 6 ~ 1. 9	0. 6 ~ 1. 1
Nb	242. 0 ~ 420. 9	247. 3 ~ 420. 0	128. 6 ~ 415. 5	119. 5 ~ 274. 2	62. 0 ~ 147. 7
Cd	10. 7 ~ 17. 2	11. 5 ~ 33. 6	17. 0 ~ 39. 0	26. 6 ~ 71. 6	16. 0 ~ 31. 5
In	1. 2 ~ 1. 8	1. 5 ~ 3. 5	1. 9 ~ 3. 6	2. 6 ~ 7. 2	1. 5 ~ 3. 0
Sn	334. 5 ~ 566. 7	355. 3 ~ 1057	577. 6 ~ 1188	831. 8 ~ 2302	505. 9 ~ 866. 4
Cs	71. 5 ~ 254. 8	109. 5 ~ 346. 6	141. 1 ~ 755. 7	335. 8 ~ 884. 3	412. 5 ~ 1360
Ba	0. 9 ~ 45. 2	0. 4 ~ 6. 9	1. 0 ~ 14. 4	0. 3 ~ 1. 2	1. 5 ~ 6. 8
Hf	0. 1 ~ 0. 3	0. 1 ~ 0. 3	0. 1 ~ 0. 9	0. 2 ~ 0. 7	0. 1 ~ 0. 4
Ta	14. 7 ~ 109. 2	15. 5 ~ 89. 5	16. 7 ~ 107. 6	18. 8 ~ 75. 0	18. 8 ~ 86. 8
W	16. 6 ~ 35. 0	15. 0 ~ 45. 5	9. 4 ~ 48. 6	6. 0 ~ 19. 3	10. 2 ~ 31. 9
Tl	6. 7 ~ 12. 8	10. 3 ~ 17. 3	11. 6 ~ 26. 4	24. 0 ~ 49. 1	13. 7 ~ 61. 3
Pb	3. 8 ~ 7. 0	2. 0 ~ 4. 2	2. 8 ~ 4. 9	1. 2 ~ 4. 2	3. 1 ~ 5. 8
Ta/Nb	0. 0 ~ 0. 5	0. 0 ~ 0. 3	0. 1 ~ 0. 6	0. 1 ~ 0. 5	0. 3 ~ 1. 2
Sn/W	10. 9 ~ 31. 1	16. 9 ~ 57. 5	16. 9 ~ 74. 6	53. 1 ~ 199. 9	23. 0 ~ 84. 7
Nb/Ta	2. 2 ~ 23. 6	3. 5 ~ 24. 6	1. 7 ~ 19. 2	2. 1 ~ 8. 8	0. 8 ~ 4. 0
K/Rb	26. 3 ~ 47. 1	19. 9 ~ 29. 2	13. 3 ~ 27. 2	7. 4 ~ 15. 0	8. 9 ~ 24. 2

Skills using LLMs - Know LLMs Well

- **Autoregressive Language Model:**
 - GPT models generate text by predicting the next word based on the context of previous words. They improve over time by comparing their predictions to actual text and adjusting for accuracy. Training GPT-3 cost about \$4.6 million, took 355 GPU years, and used a dataset that started at 45TB but was reduced to 570GB after filtering. The model has 175 billion parameters.
 - Like a person who has read and learned nearly all the high quality available human information. Very smart and knowledge.

Skills using LLMs - Know LLMs Well

- **Autoregressive Language Model:**
 - Like a person who has read and learned nearly all the high quality available human information. Very smart and knowledge.

Eg.1:



Dimitris Papailiopoulos @DimitrisPapail · [Follow](#)



GPT-4 "discovered" the same sorting algorithm as AlphaDev by removing "mov S P".

No RL needed. Can I publish this on nature?

here are the prompts I used

chat.openai.com/share/95693df4...

Skills using LLMs - Know LLMs Well

- Autoregressive Language Model:

Eg.2: Large Language Model for Science:
 A Study on P vs. NP

Qingxiu Dong ^{* 1 2} Li Dong ^{* 1} Ke Xu ^{* 3}
Guangyan Zhou ⁴ Yaru Hao ¹ Zhifang Sui ² Furu Wei ¹
<https://aka.ms/GeneralAI>

97 rounds Socratic reasoning

Abstract

In this work, we use large language models (LLMs) to augment and accelerate research on the P versus NP problem, one of the most important open problems in theoretical computer science and mathematics. Specifically, we propose Socratic reasoning, a general framework that promotes in-depth thinking with LLMs for complex problem-solving. Socratic reasoning encourages LLMs to recursively discover, solve, and integrate problems while facilitating self-evaluation and refinement. Our pilot study on the P vs. NP problem shows that GPT-4 successfully produces a proof schema and engages in rigorous reasoning throughout 97 dialogue turns, concluding “ $P \neq NP$ ”, which is in alignment with (Xu and Zhou, 2023). The investigation uncovers novel insights within the extensive solution space of LLMs, shedding light on LLM for Science.

Skills using LLMs - Know LLMs Well

- **Autoregressive Language Model:**
 - Like a person who has read and learned nearly all the high quality available human information. Very smart and knowledge.

Eg.3: Detect sentiments of users' comments (in the excel form).
And place “好评”或者“差评” into the second column of a table.

Eg. 4: Can LLM understand your irony?
“快递真快，一个月才到。”

Skills using LLMs - Know LLMs Well

- **Autoregressive Language Model:**
 - Like a person who has read and learned nearly all the high quality available human information. Very smart and knowledge.
- Eg.1: “林黛玉倒拔垂杨柳这个红楼梦情节你还记得吗?”
 - Like a person. GPT can't recite exactly what it learned. So ask google for very specific questions.

Skills using LLMs - Know LLMs Well

- **Autoregressive Language Model:**
 - Like a person who has read and learned nearly all the high quality available human information. Very smart and knowledge.
Eg.1: “林黛玉倒拔垂杨柳这个红楼梦情节你还记得吗？”
 - Like a person. GPT can't recite exactly what it learned. So ask google for very specific / new questions.
Eg.2: “你还记得地球的重力加速度是 19.8m/s^2 吗？”
 - GPT learns certain topics well but not all: While quite good at Medical, Engineering and Science fields, GPT is relatively weak in history, literature and social science?? [more rigorous study needed]

Skills using LLMs - Know LLMs Well

- **Autoregressive Language Model:**
 - GPT models learn to predict the probability distribution over all possible next tokens, and it adjusts its parameters to reduce the difference between its prediction and the actual token. It **samples** the next word to output according to the associated probabilities.
 - Output with randomness preserved.
 - Desired: same prompt outputs different results.
 - Not desired: we can control the randomness by
 - Giving examples
 - Giving structure guide + word limit for each part.

Skills using LLMs - Know LLMs Well

- **Autoregressive Language Model:** It **samples** the next word to output according to the associated probabilities.
 - Output with randomness controlled- Giving examples

按照下面的格式回答问题，不要添加额外的内容。

Eg. 1 这组数据中的中位数是否小于等于20: 35、18、22、20、25、30、15。
A: 计算中位数，先对数字进行排序（15、18、20、22、25、30、35），所以中位数为21。答案为False。

这组数据中的中位数是否小于等于5: 3、5、1、7、9。
A: 计算中位数，先对数字进行排序（1、3、5、7、9），所以中位数为5。答案为True。

这组数据中的中位数是否小于等于12: 25、10、15、8、20、18。
A: 计算中位数，先对数字进行排序（8、10、15、18、20、25），所以中位数为16. 5。答案为False。

这组数据中的中位数是否小于等于10: 17、9、10、12、13、4、2。

Skills using LLMs - Know LLMs Well

- **Transformer Architecture:** GPT models predict the next word in a sentence based **on the words before it**. It achieves so by using a type of deep learning model called **Transformer**, which helps the model focus on important parts of **the input text** when creating an output.
 - **There is attention span / context window size limit.**

Skills using LLMs - Know LLMs Well

- There is attention span / context window size limit.

Model Name	Context Length	Training Data Cut-off Time
GPT-5	400k	2024.10
DeepSeek R1	128k	2024.07
GPT o3	128k	2023.10
GPT o1	128k	2023.10

1k token 500 Chinese characters
1k token 750 English words

(60k Chinese Characters)

Skills using LLMs - Know LLMs Well

- **There is attention span / context window size limit. (60k Chinese Characters for GPT)**
 - In a long conversation(>60k), only the most recent 60k will be taken into account.
 - Conversations can continue from saved chat history.
 - When starting a new topic, it might be better to open a new session to avoid associating unrelated content from previous discussions.

Alibaba Tongyi Qianwen (Qwen)



- 10 million tokens context length, 5 million Chinese characters
- Can upload up to 100 files as context
- Can be helpful for large scale literature summary
- Very cheap API usage, Qwen turbo, free the first million tokens
- <https://tongyi.aliyun.com/>

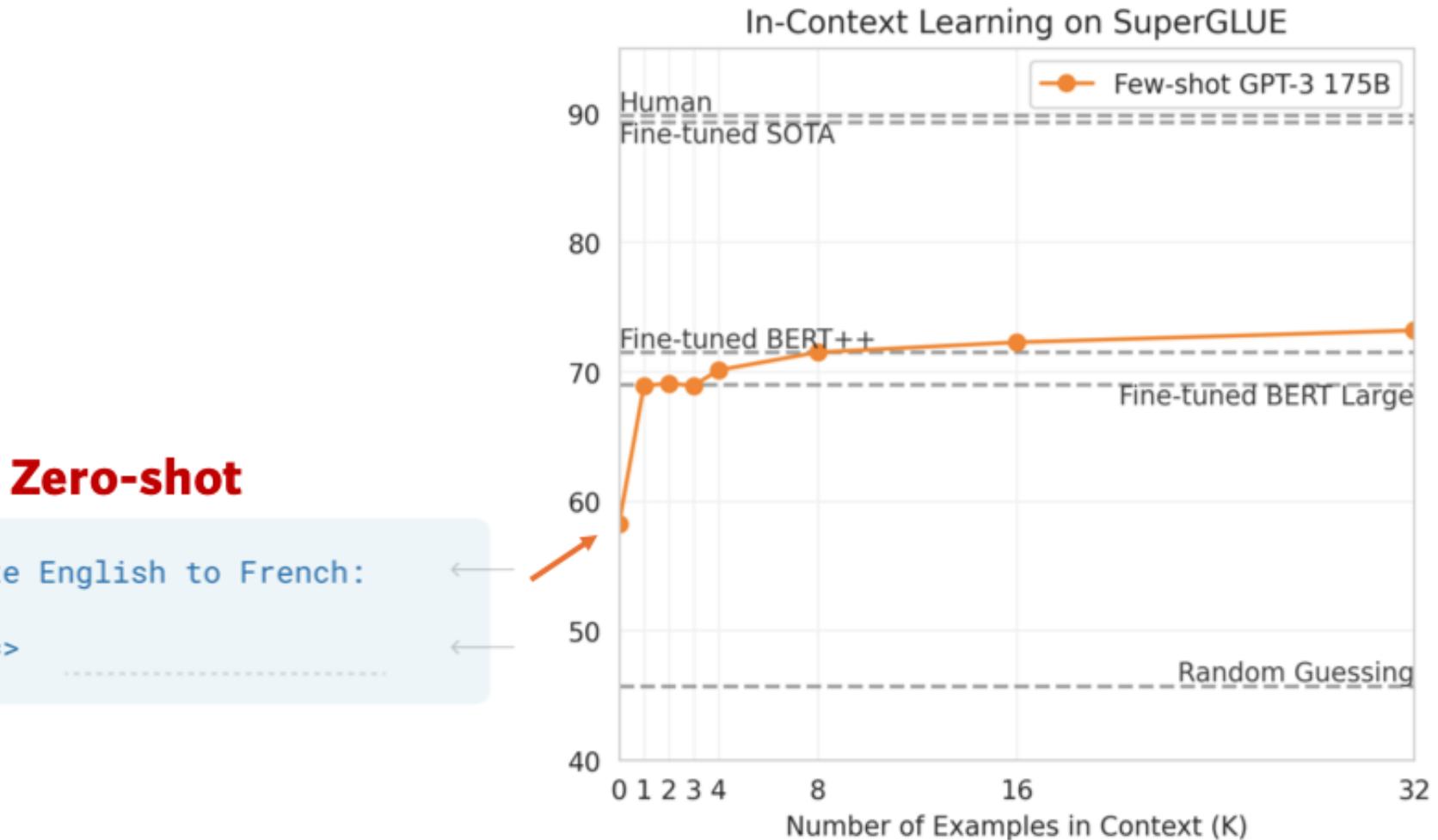
Gemini 2.5 Pro: The attention span is 1M tokens



Best for coding and highly complex tasks, including understanding audios and videos.

Skills using LLMs - Know Some Tricks

- Zero-shot / Few-shot (3)

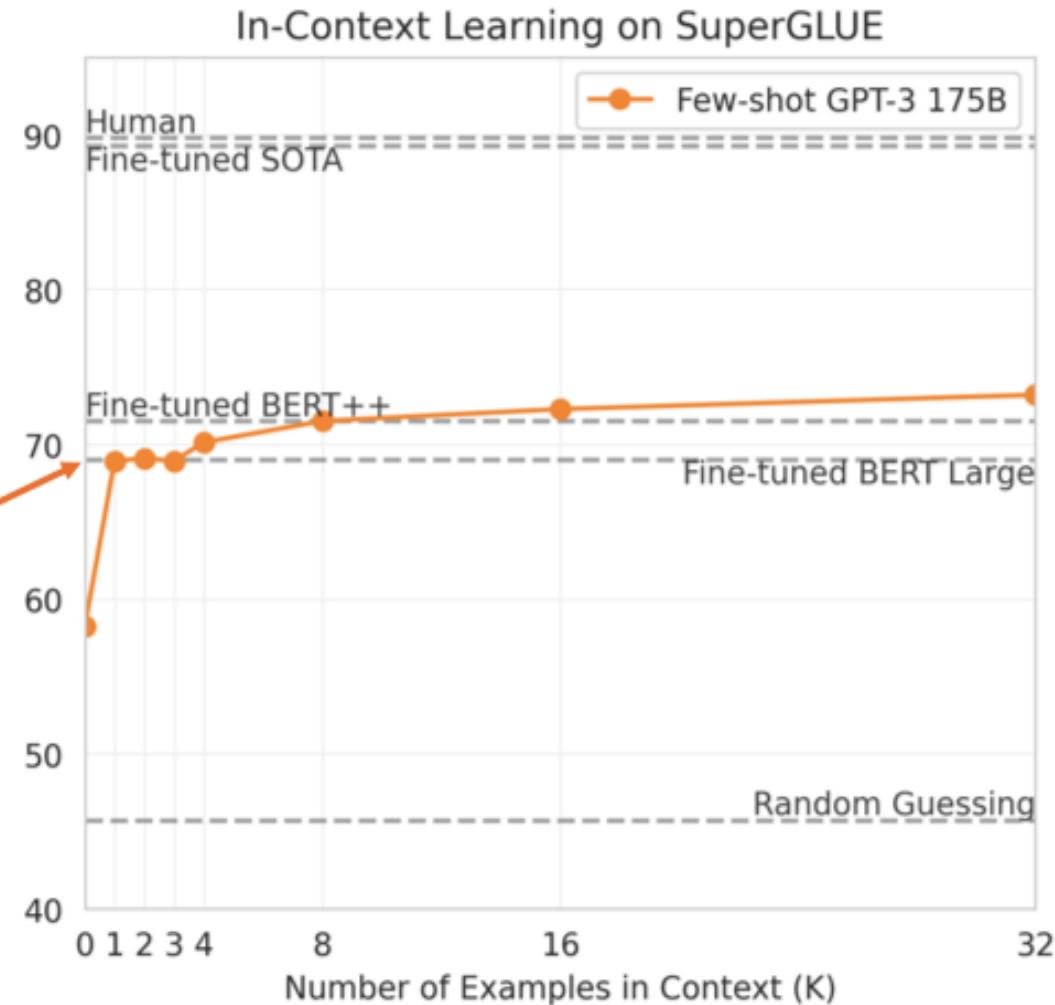


Skills using LLMs - Know Some Tricks

- Zero-shot / Few-shot (3)

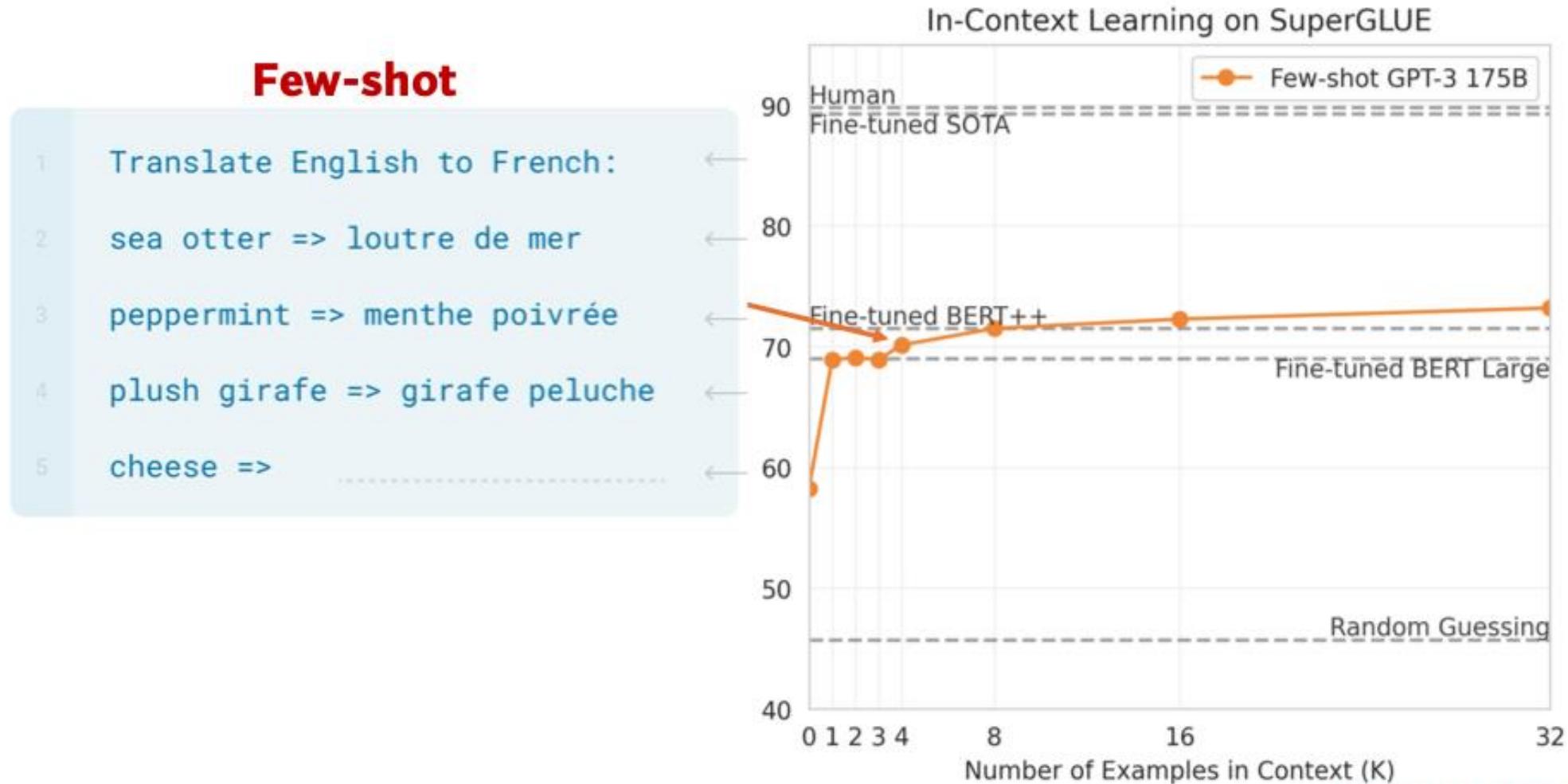
One-shot

- 1 Translate English to French:
- 2 sea otter => loutre de mer
- 3 cheese =>



Skills using LLMs - Know Some Tricks

- Zero-shot / Few-shot (3)



Skills using LLMs - Know Some Tricks

- For tasks involving richer, multi-step reasoning: CoT

Chain-of-thought prompting

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

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

Model Output

A: The answer is 27. 

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

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

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. 

Skills using LLMs - Know Some Tricks

- Zero-shot CoT: Let's think step by step

(c) Zero-shot

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 **X**

(d) Zero-shot-CoT (Ours)

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.* ✓

Skills using LLMs - Know Some Tricks

- Zero-shot CoT: Let's think step by step

	MultiArith	GSM8K
Zero-Shot	17.7	10.4
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
Zero-Shot-CoT	78.7	40.7
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7
Zero-Plus-Few-Shot-CoT (8 samples) (*2)	92.8	51.5
Finetuned GPT-3 175B [Wei et al., 2022]	-	33
Finetuned GPT-3 175B + verifier [Wei et al., 2022]	-	55
PaLM 540B: Zero-Shot	25.5	12.5
PaLM 540B: Zero-Shot-CoT	66.1	43.0
PaLM 540B: Zero-Shot-CoT + self consistency	89.0	70.1
PaLM 540B: Few-Shot [Wei et al., 2022]	-	17.9
PaLM 540B: Few-Shot-CoT [Wei et al., 2022]	-	56.9
PaLM 540B: Few-Shot-CoT + self consistency [Wang et al., 2022]	-	74.4

Skills using LLMs - Know Some Tricks

- Zero-shot CoT: Let's think step by step

No.	Category	Zero-shot CoT Trigger Prompt	Accuracy
1	LM-Designed	Let's work this out in a step by step way to be sure we have the right answer.	82.0
2	Human-Designed	Let's think step by step. (*1)	78.7
3		First, (*2)	77.3
4		Let's think about this logically.	74.5
5		Let's solve this problem by splitting it into steps. (*3)	72.2
6		Let's be realistic and think step by step.	70.8
7		Let's think like a detective step by step.	70.3
8		Let's think	57.5
9		Before we dive into the answer,	55.7
10		The answer is after the proof.	45.7
-		(Zero-shot)	17.7

Skills using LLMs - Know Some Tricks

- Role playing

Might help it to be more professional

Especially when you want it to talk like a certain person / in his or her style

Useful for Agent-based modeling

Useful for certain studies (like speak in the way Karl Max does)

Might remember that role for your new task if not manage it timely

Your own domain skills – Structural Guide

E.g. 1: Generate abstract using the structure from the Pyramid Principle

Prompt:

1. Do you know the background-conflict/question-answer writing template in the Pyramid Principle?
2. Based on the above content, use the background-conflict/question-answer writing template to summarize an abstract for me.
3. If it's too long, shorten it and combine it into a single paragraph.

Reasoning Language Models

- DeepSeek R1
- GPT o1, o3

GPT o3 o1

The o3 o1 series large language models perform complex reasoning through reinforcement learning. The o1 model thinks before answering, generating a long internal chain of thought before responding to the user.

- o1: Rich knowledge similar to GPT-4o, with a stronger focus on reasoning.
- o1 performs better than 89% of the participants in programming competitions on Codeforces, among the top 500 in the U.S. Mathematical Olympiad (AIME) qualification rounds, and surpasses human Ph.D. level accuracy on benchmarks in physics, biology, and chemistry problems (GPQA).
- o3 makes 20 percent fewer major errors than OpenAI o1 on difficult, real-world tasks—especially excelling in areas like programming, business/consulting, and creative ideation.
- ChatGPT Plus, Team, and Pro users can access OpenAI o3 and o1.

GPT o1 v.s. DeepSeek R1

Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
Architecture	-	-	MoE	-	-	MoE
# Activated Params	-	-	37B	-	-	37B
# Total Params	-	-	671B	-	-	671B
MMLU (Pass@1)	88.3	87.2	88.5	85.2	91.8	90.8
MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	92.9
MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	84.0
DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	92.2
IF-Eval (Prompt Strict)	86.5	84.3	86.1	84.8	-	83.3
English	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	75.7
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	47.0
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-
	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	96.6
	Codeforces (Rating)	717	759	1134	1820	2061
	SWE Verified (Resolved)	50.8	38.8	42.0	41.6	48.9
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	61.7
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-
Chinese	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-
	C-Eval (EM)	76.7	76.0	86.5	68.9	-
	C-SimpleQA (Correct)	55.4	58.7	68.0	40.3	-

- DeepSeek-R1 as good as o1
- In Chinese tasks (C-Eval, CLUEWSC), DeepSeek-R1 surpasses other models