南京大学
NANJING UNIVERSITY

人工智能导论

大模型时代的推理与规划

郭兰哲

南京大学 智能科学与技术学院

https://www.lamda.nju.edu.cn/guolz

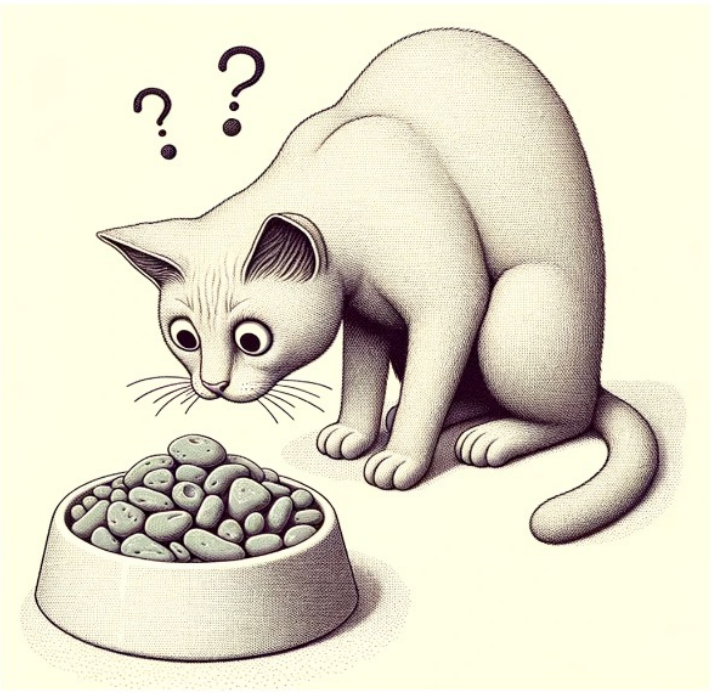
Email: guolz@nju.edu.cn

大语言模型简介

什么是语言模型？

The cat drank a bowl of diet coke.

The cat drank a bowl of milk.



The cat drank a bowl of rocks.

语言模型的任务：哪句话出现的概率大

什么是语言模型？

- 给定文字序列 (word sequence)
  - $w_1, w_2, \cdots, w_n$
  - 语言模型旨在估计一个句子在语言中出现的概率

$$P(w_1, w_2, \cdots, w_n)$$

- 可以直接帮助提升各种自然语言处理任务的性能，例如：
  - 机器翻译：英语：You go first ->中文：你先走 vs 你走先
  - 内容补全：我今天想吃 _____（苹果、可乐、火锅、石头）
  - ……

语言模型学习的挑战

- 组合爆炸：假设我们有5000个常用词，一句话的长度是5，那么可能的句子数量：5000^5

- **数据稀疏**：很多合理句子在训练语料中可能一次都没出现过
  - 南京大学的学生今天想吃苹果

统计语言模型

- N-Gram模型：统计语言模型的代表
  - 出发点：序列越长，计算和存储多个词共同出现概率的复杂度指数增加
  - 解决方案：引入马尔科夫假设，认为当前词出现的概率仅仅依赖前面N个词

我 今天 想

今天 想 吃

想 吃 苹果

Chain rule

$$P(w_1, w_2, \cdots, w_n) = P(w_n|w_1, \cdots, w_{n-1})P(w_1, \cdots, w_{n-1})$$

一阶：P(南京大学) = P(南)P(京|南)P(大|京)P(学|大)

二阶：P(南京大学) = P(南)P(京|南)P(大|南京)P(学|京大)

统计语言模型

- 统计语言模型的局限
  - 无法理解词语之间的相似性

我想吃苹果

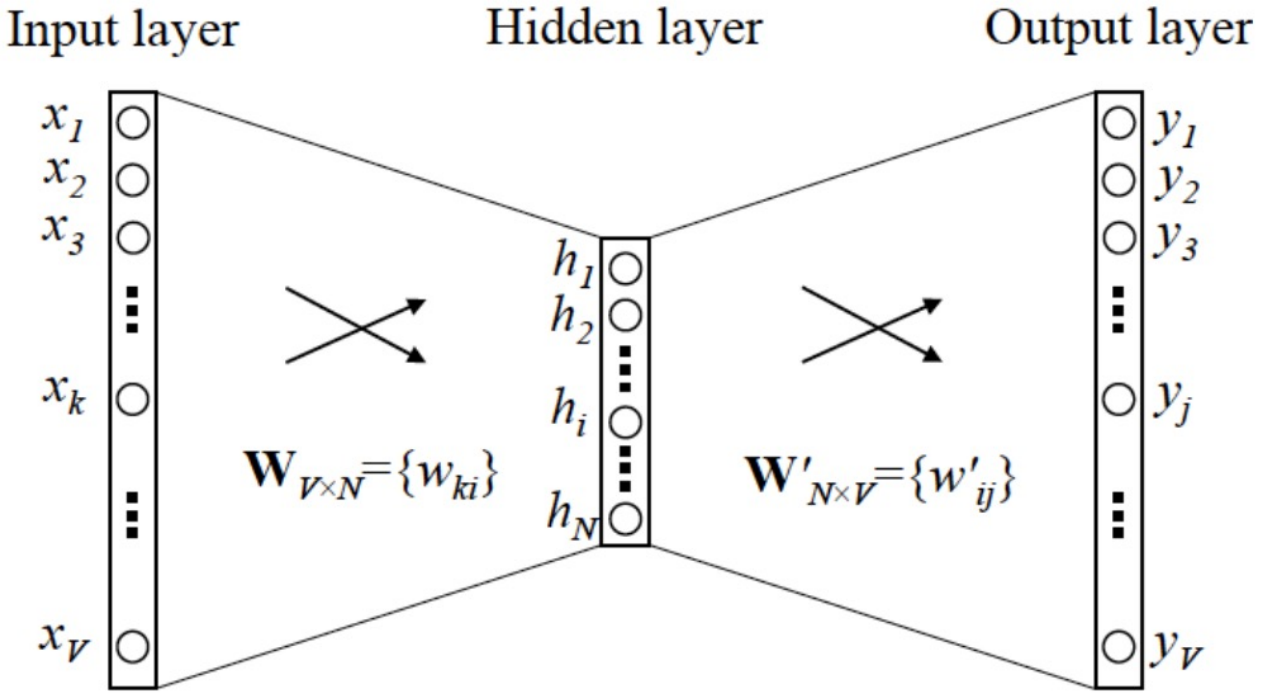
我想吃梨

统计语言模型仅基于频率统计，很难知道“苹果”和“梨”都是水果，

它们在很多语境中是相似的

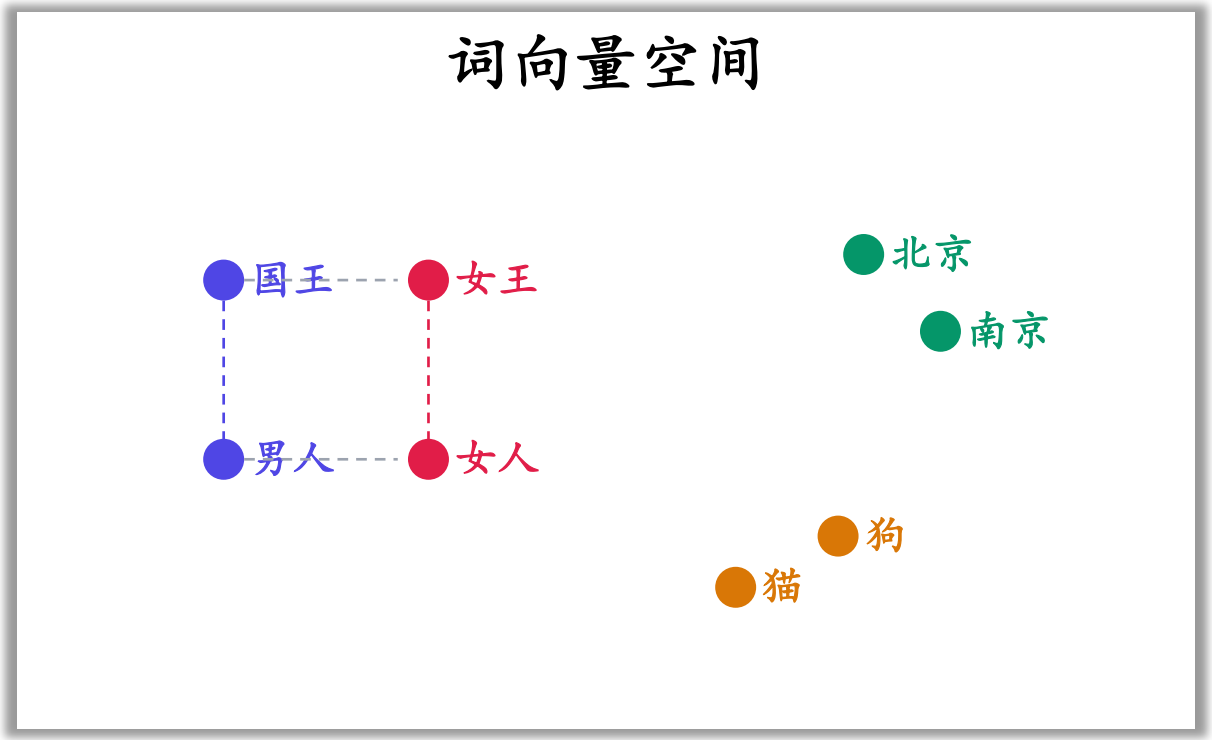
神经网络语言模型

- 神经语言模型(Neural Language Model)
  - 基于神经网络，将单词预测成词向量，再基于词向量完成诸如翻译等自然处理任务



神经网络语言模型

- 词向量：语义相似的词，向量空间的相似度也相似



传统神经语言模型的不足

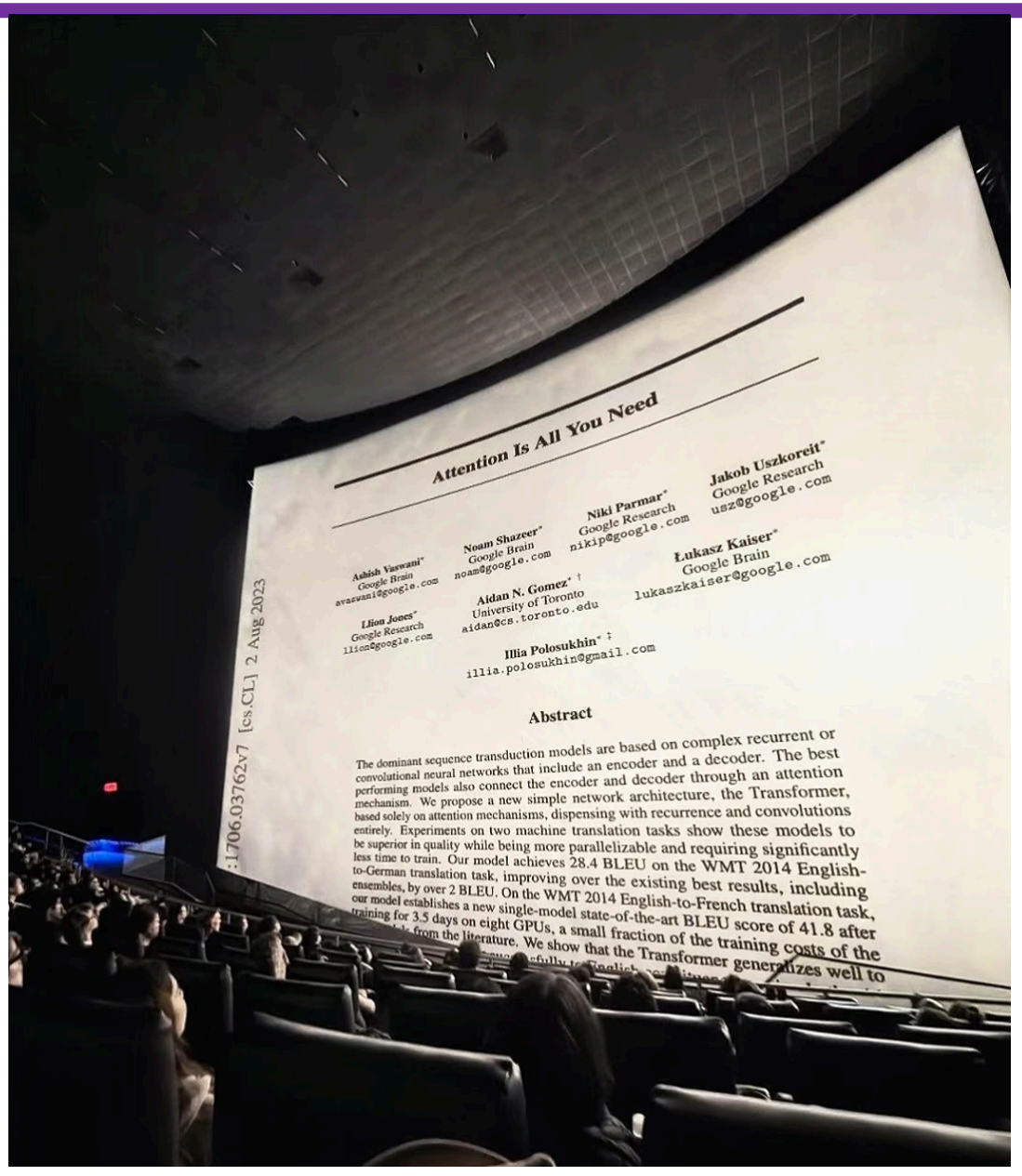
- 上下文窗口太短

小明昨天在图书馆借了一本关于人工智能的书，

今天他把这本书还给了 ____

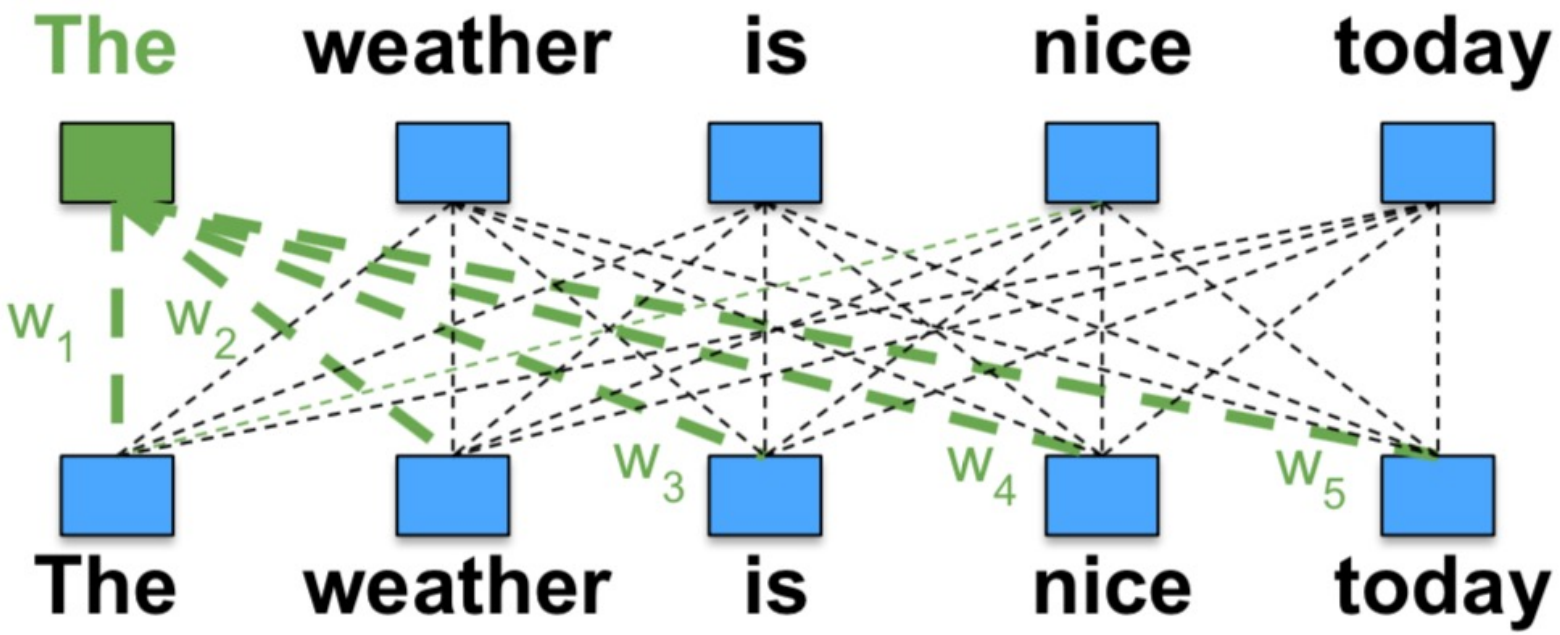
难以处理长距离依赖，当模型读到后面时，前面的内容已经忘了

现代语言模型的基石：Transformer



现代语言模型的基石：Transformer

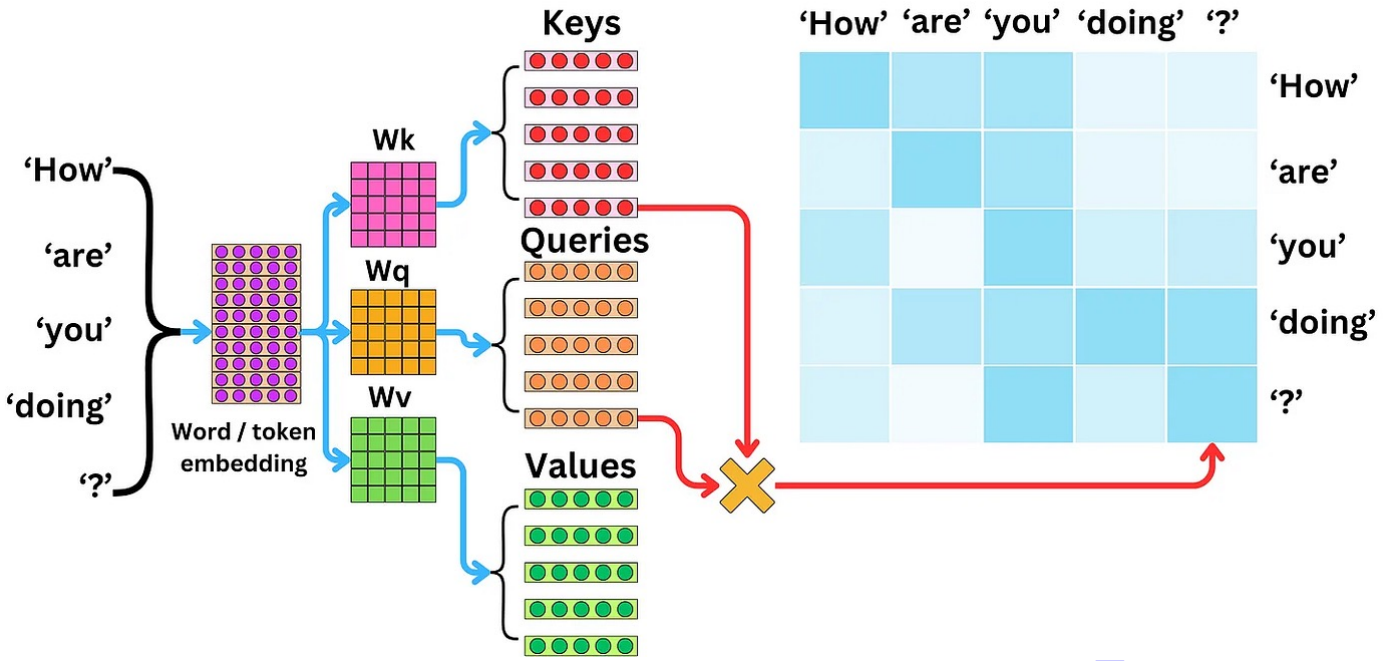
- Transformer：采用**自注意力机制(Self-Attention)**处理远程依赖关系，直接建模全局相关性，通过计算注意力权重捕获中远距离元素之间的关系
- **注意力就是权重，衡量词与词之间的相关性**



现代语言模型的基石：Transformer

每个 token 都会提出三个向量：

Query	Key	Value
我在找什么	我能提供什么	我的实际信息



$$Attention(Q,K,V) = softmax\left(\frac{QK^{\top}}{\sqrt{d}}\right)V$$

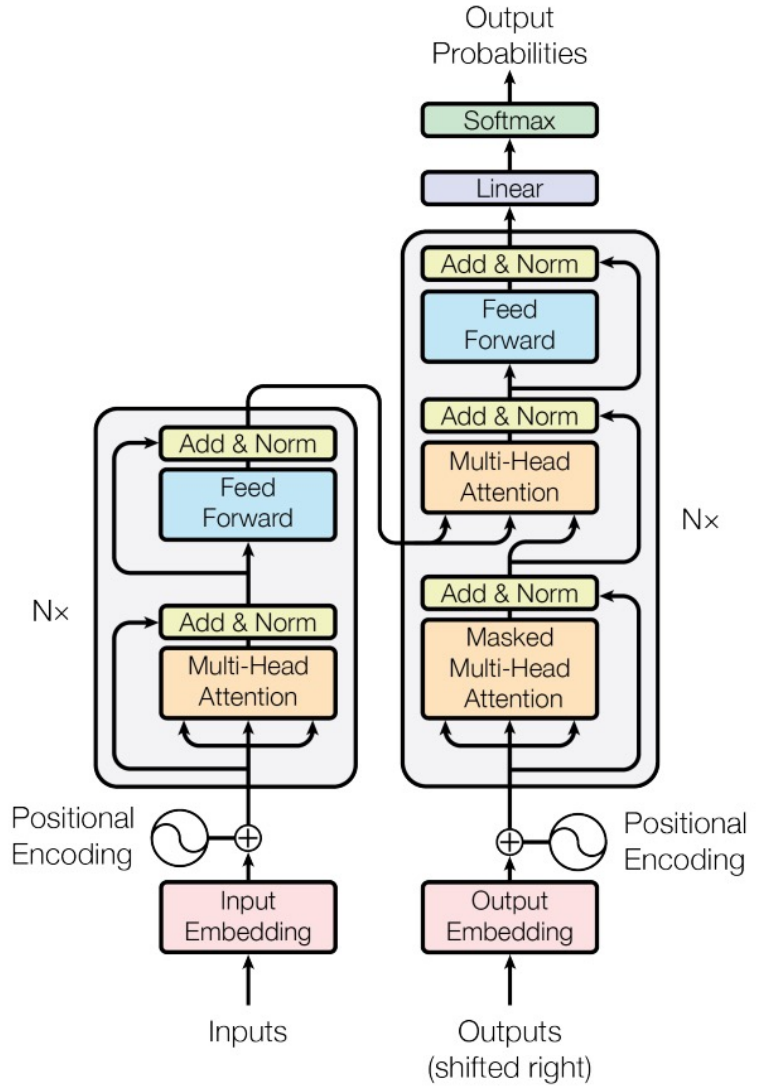
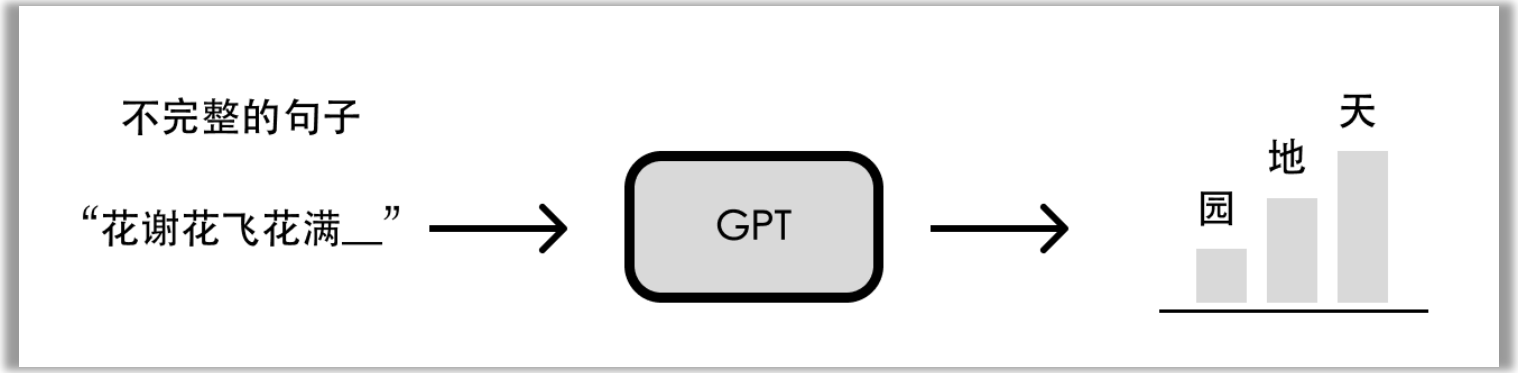


Figure 1: The Transformer - model architecture.

从Transformer到GPT 1.0

- 2018年，OpenAI推出第一代GPT (Generative Pre-training Transformer)
  - 结论：无监督预训练 + 有监督微调，可以在多种自然语言处理任务上取得效果
- 无监督预训练：给定前文，让模型补全下文



- 有监督微调：预训练完成后，针对具体下游任务（分类、问答、翻译等）用少量标注数据做微调，只需改动输出层，主体参数不变

Improving Language Understanding by Generative Pre-Training. OpenAI, 2018.

全面开花的GPT 2.0

- 基本想法：
  - 增加更多任务
    - 句子打乱再排序、选择题、判断题、问答题、寻找文中的错别字等
  - 扩大模型规模
    - 从1.17亿参数扩大到15亿参数、40G文本语料

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Language Models are Unsupervised Multitask Learners. OpenAI, 2019.

持续扩展的GPT 3.0

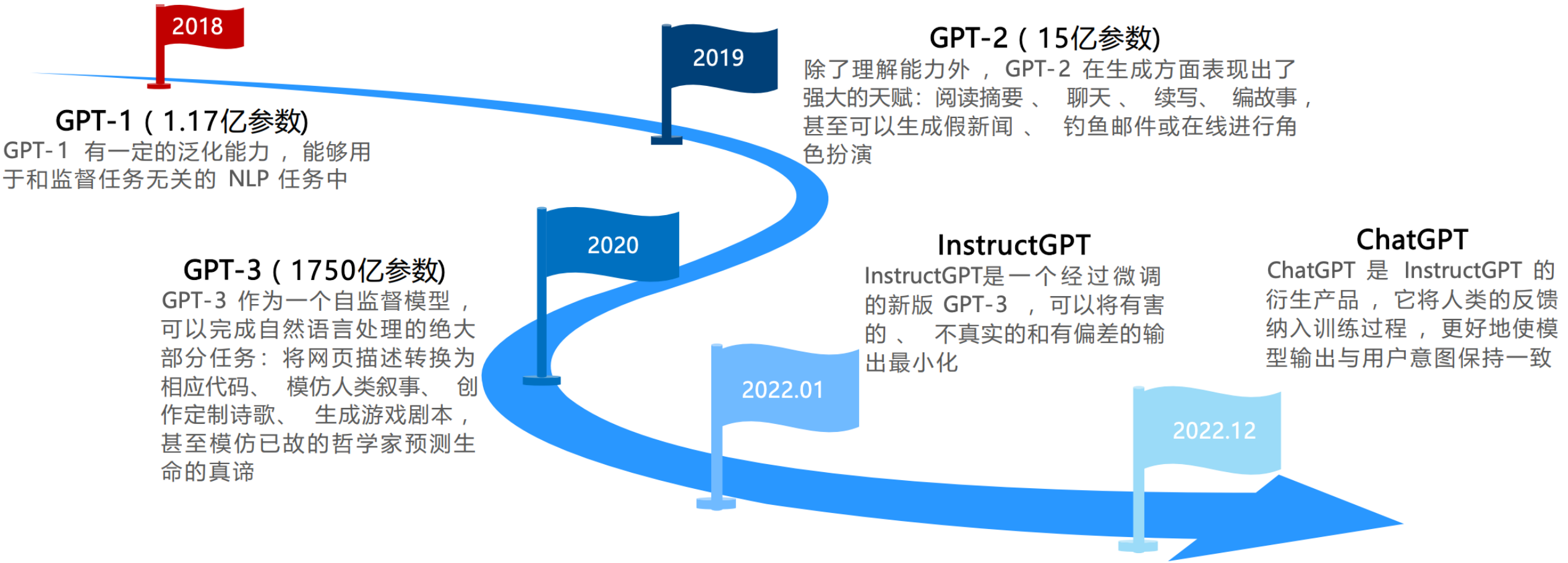
继续扩展：增加数据、参数量、计算力

模型名	参数量	训练数据
Original GPT (GPT-1)	1.17 亿	4.5 GB文本
GPT-2	15 亿	40 GB文本
GPT-3	1750亿	570 GB文本

GPT-3: Language Models are Few-Shot Learners. OpenAI, 2020. https://arxiv.org/abs/2005.14165

ChatGPT

2022年11月30日，ChatGPT正式发布

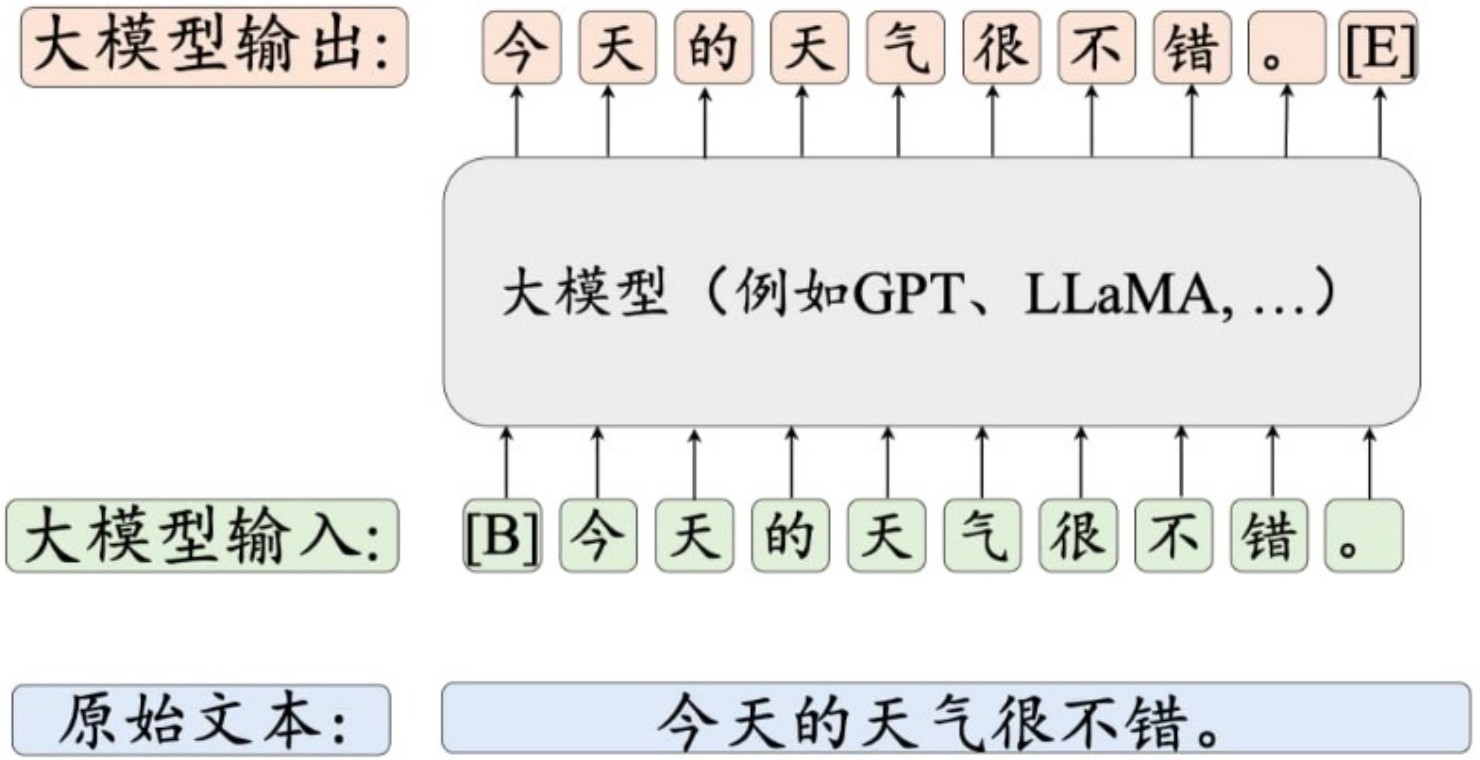


ChatGPT

- Phase 1: 预训练(Pre-Training)
  - 掌握通用知识
- Phase 2: 后训练：有监督微调(Supervised Fine-Tuning)
  - 与特定任务对齐
- Phase 3: 后训练：强化微调(RLHF)
  - 与人类偏好对齐

预训练大模型

预训练任务：预测下一个Token

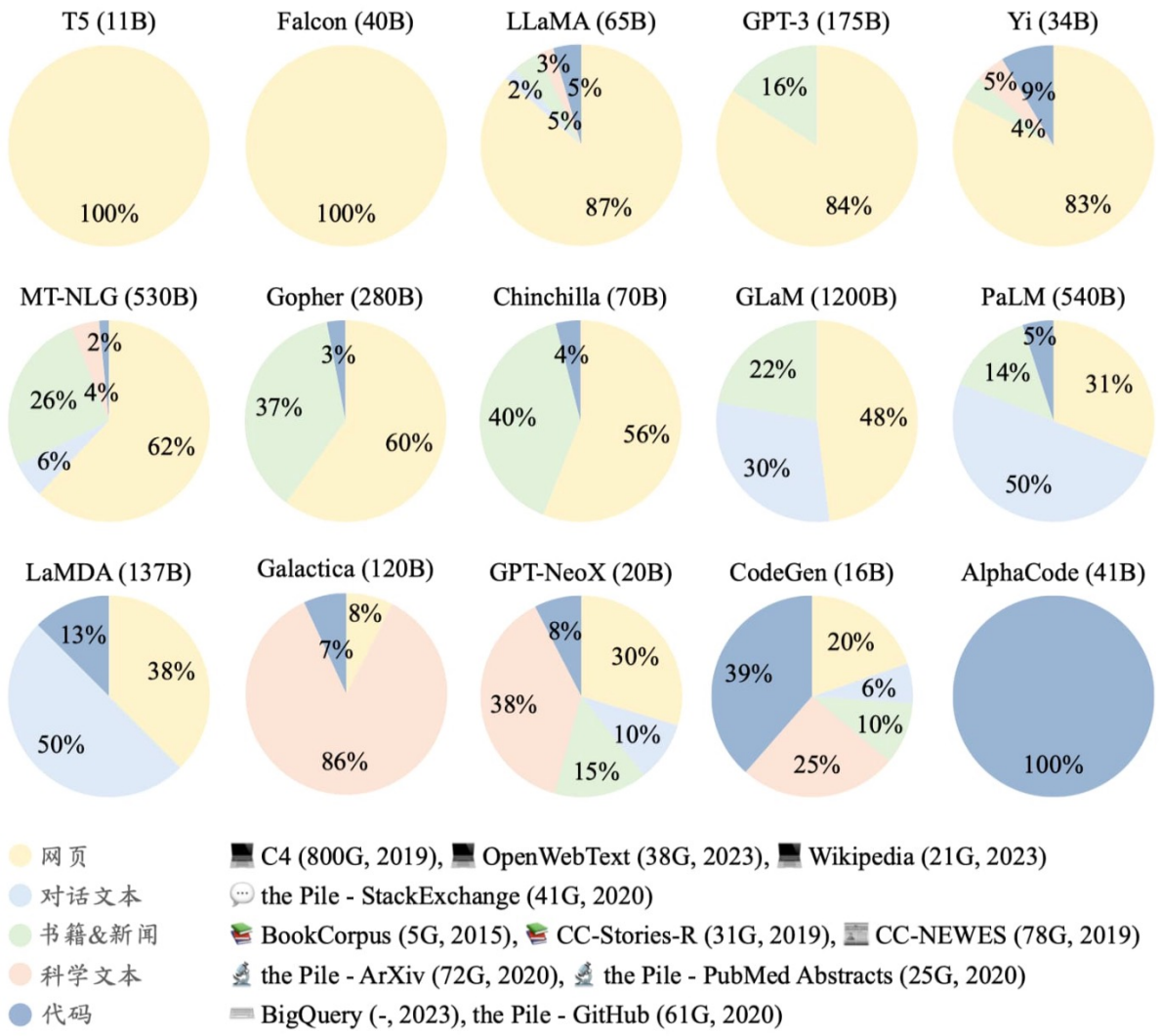


预训练数据来源

预训练数据来源：

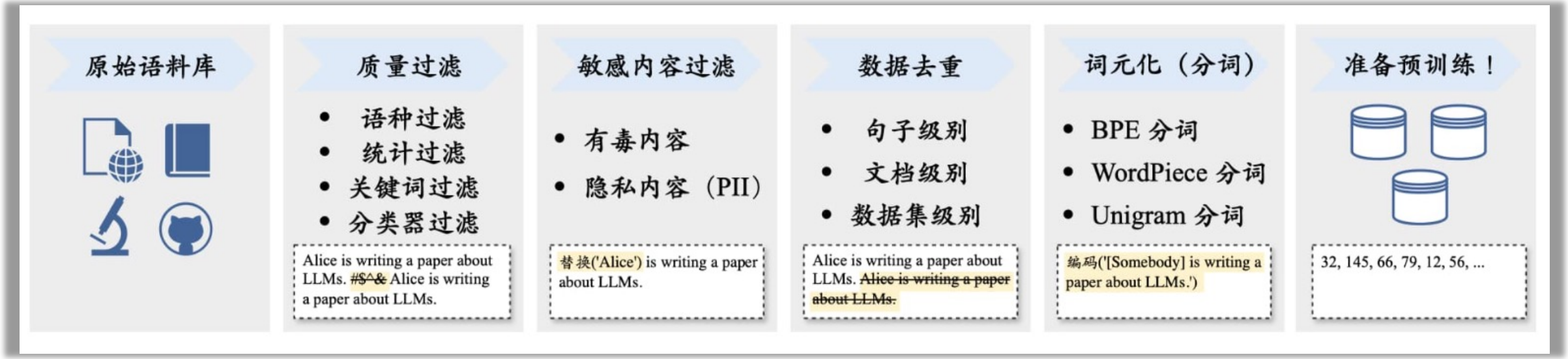
- **通用文本数据：**网页、书籍和对话等，规模较大，提升模型的语言建模能力
- **专用文本数据：**多语言数据、科学数据、代码数据等

代表性大语言模型的预训练数据分布



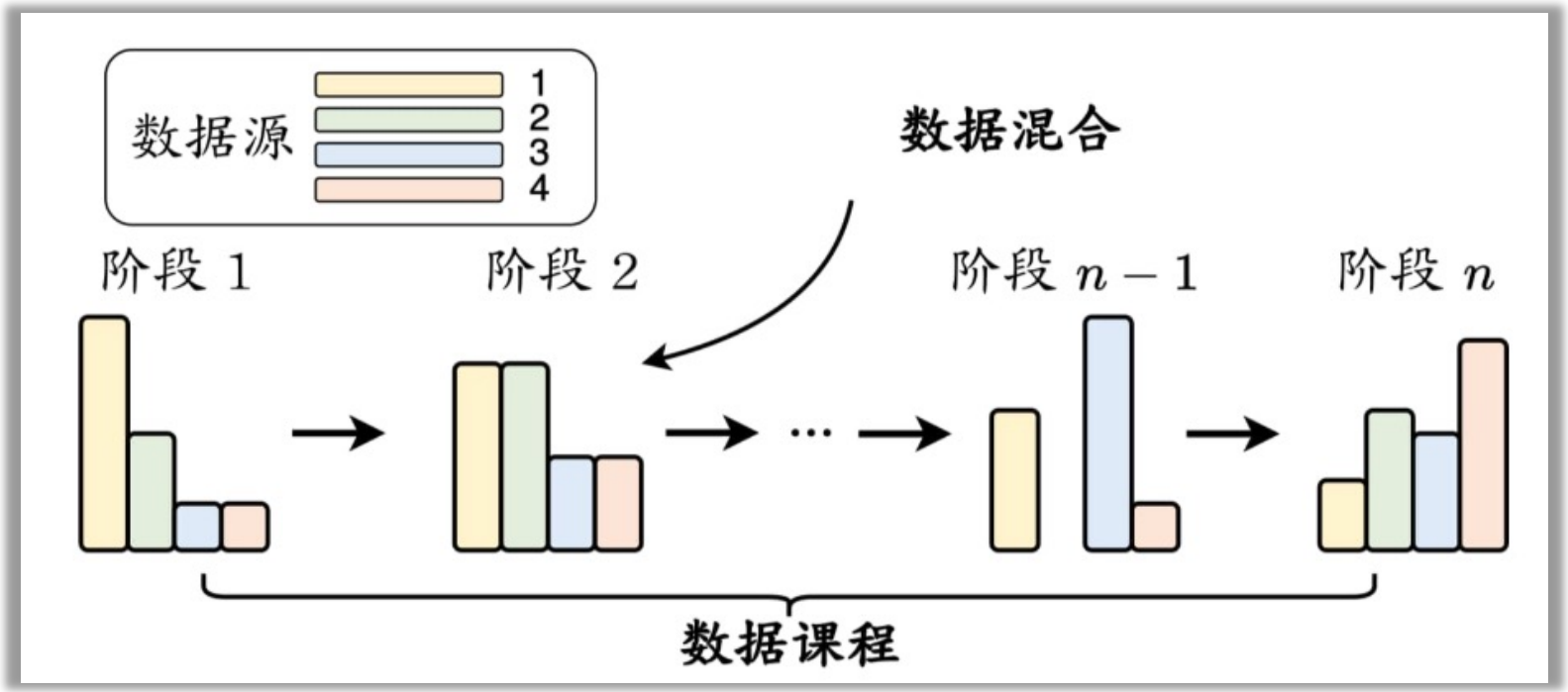
预训练数据处理

- 在收集了丰富的文本数据之后，就需要对数据进行预处理，消除低质量、冗余、无关甚可能有害的数据
- 典型的数据预处理流程包括**质量过滤**、**敏感内容过滤**、**数据去重**等步骤



预训练数据调度

- 完成数据预处理之后，需要设计合适的调度策略来安排这些多来源的数据
- 数据调度主要关注两个方面：各个数据源的混合比例、各数据源用于训练的顺序



有监督微调(Supervised Fine-Tuning, SFT)

- **等价概念：有监督微调(Supervised Fine-Tuning)、指令微调(Instruction Tuning)**
- **预训练让模型学会说话、有监督微调让模型学会回答问题**

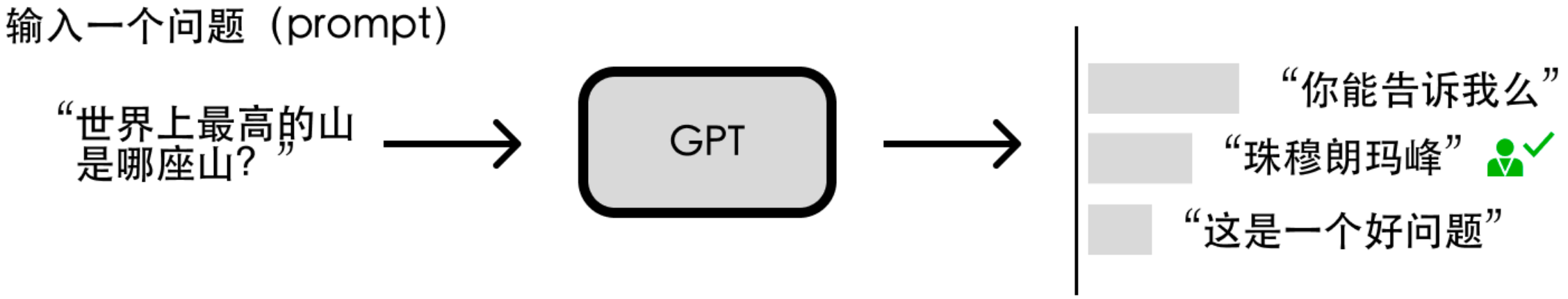
- Instruction: 输入的用户指令
- Input: 执行该指令可能需要的补充输入
- output: 即模型应该给出的回复

- Instruction:将下列文本翻译成英文：
- Input: 今天天气真好
- output: Today is a nice day

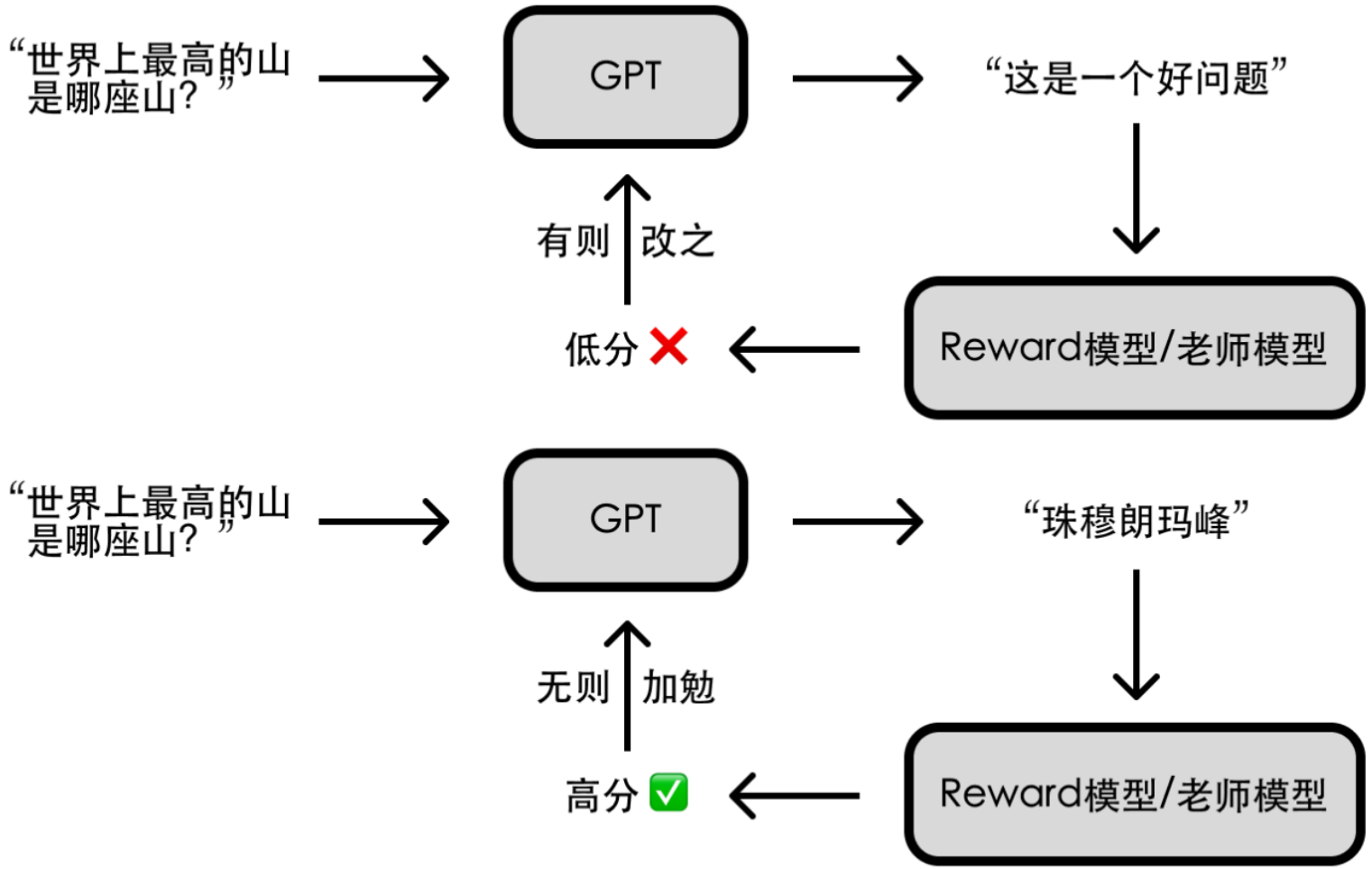
指令微调数据格式示例

人类反馈的强化学习(RLHF)

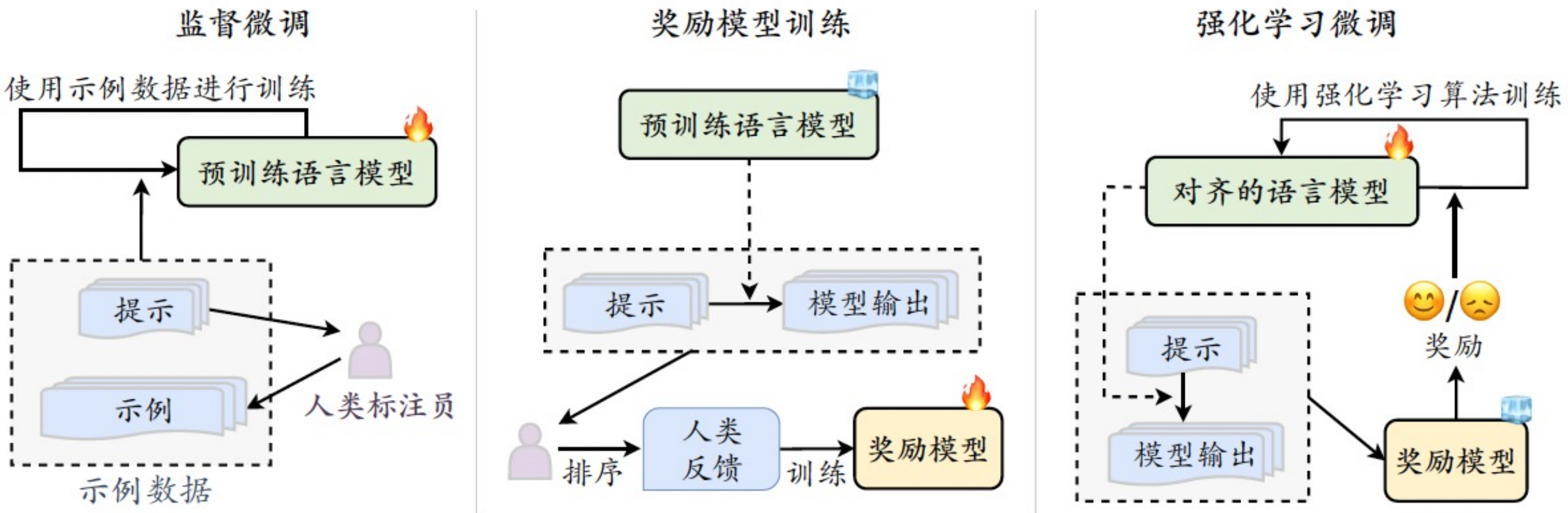
- 人类反馈强化学习： Reinforcement Learning from Human Feedback
- 目的：其实就是让模型与人类价值观一致，从而输出人类希望其输出的内容



人类反馈的强化学习(RLHF)

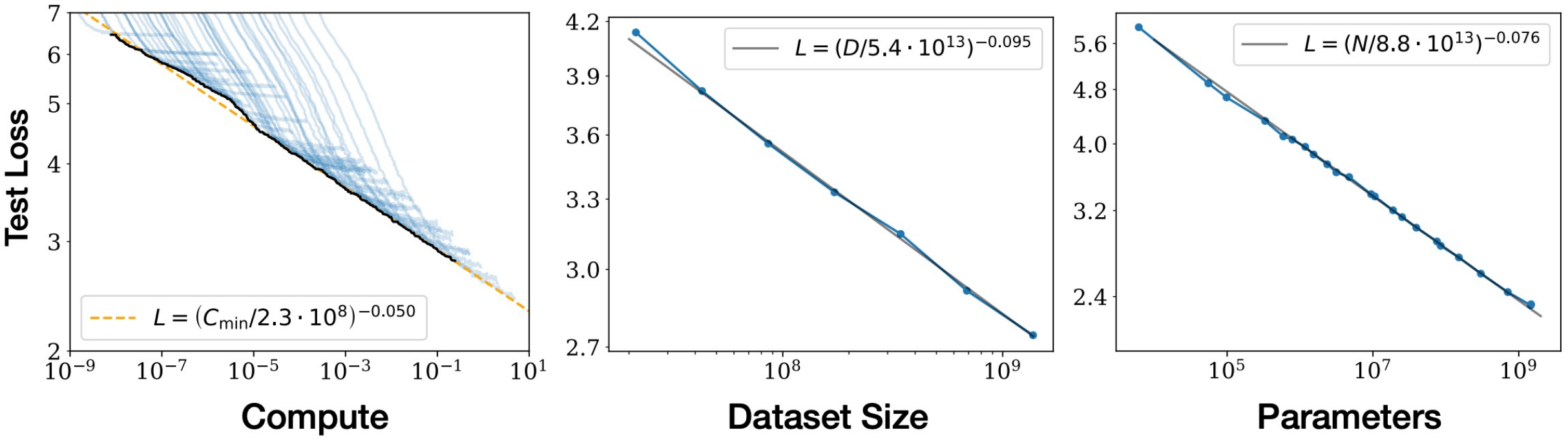


人类反馈的强化学习(RLHF)



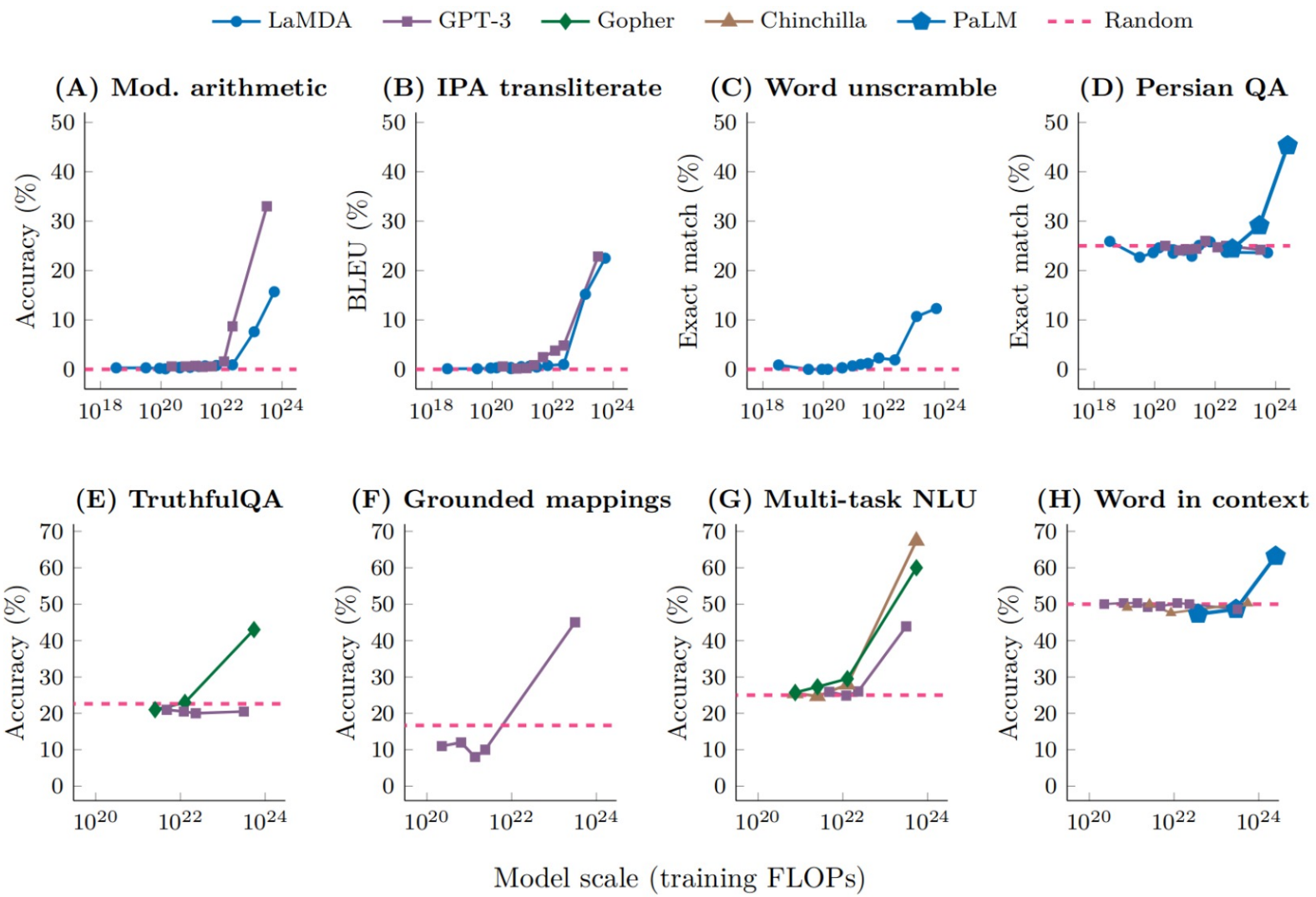
两个有意思的事情：Scaling Law

扩展法则(Scaling Law)：参数越大、数据越多、算力越强 -> 模型越强



两个有意思的事情：涌现能力(Emergent ability)

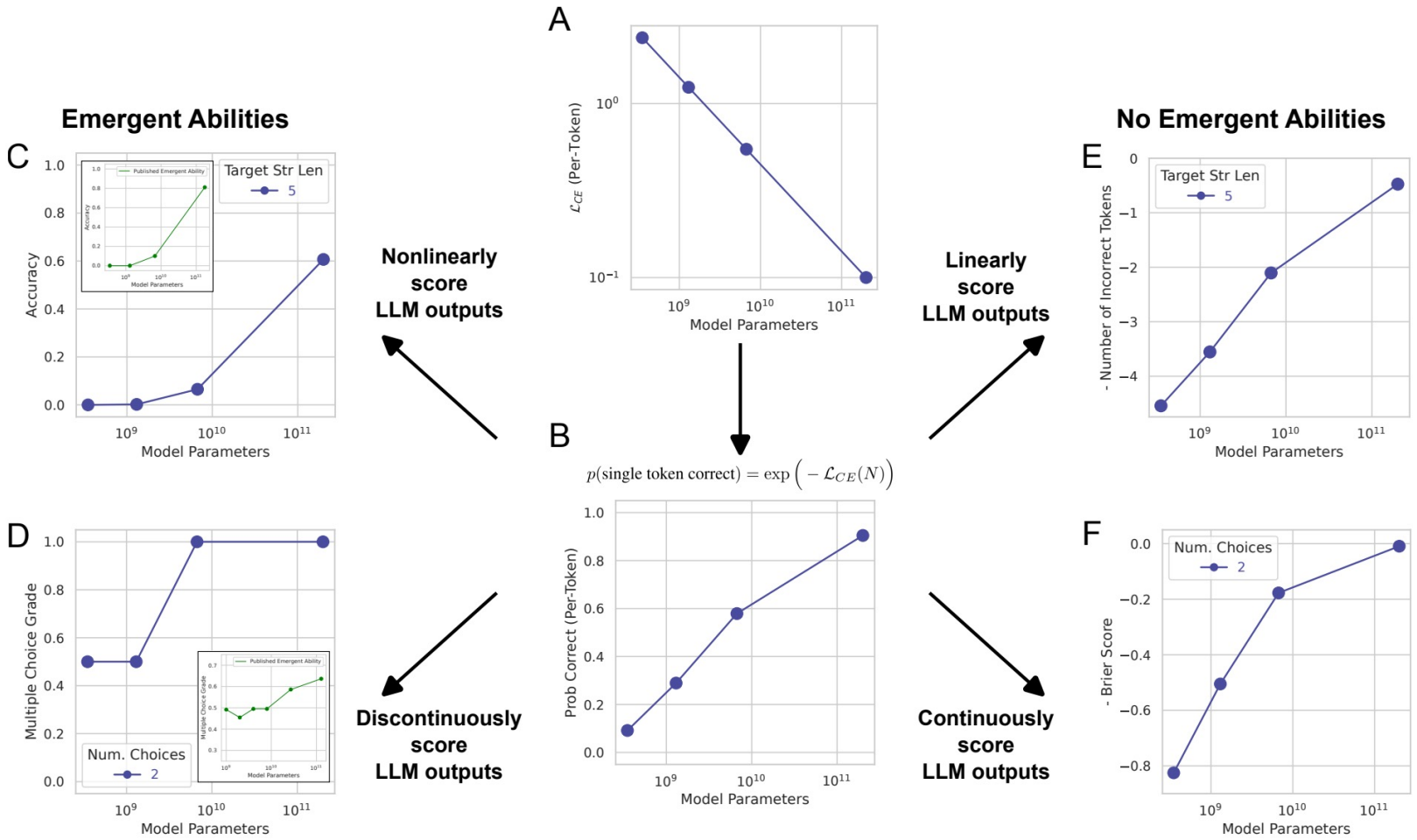
随着模型的增大，突然“涌现”了全新的能力



涌现能力(Emergent ability)

- 目前，关于涌现的机制尚不明确，围绕涌现仍有许多争议

- 2023年4月，斯坦福团队发布研究报告，认为所谓“涌现”，仅仅是评价指标的选择不同所导致



Are Emergent Abilities of Large Language Models a Mirage? NeurIPS 2024. https://arxiv.org/pdf/2304.15004

从聊天机器人到推理大模型

大模型提示工程

“Talk is cheap, show me the code”

-> “Code is cheap, show me the talk”

提示(Prompt)、上下文学习(In-Context Learning)

上下文学习（In-context Learning），旨在利用大语言模型所涌现的能力，以少量示例结合自身知识泛化至所需目标任务

Instruction: Solve the following math problem

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: *The answer is 5 cars.*

Prompt

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

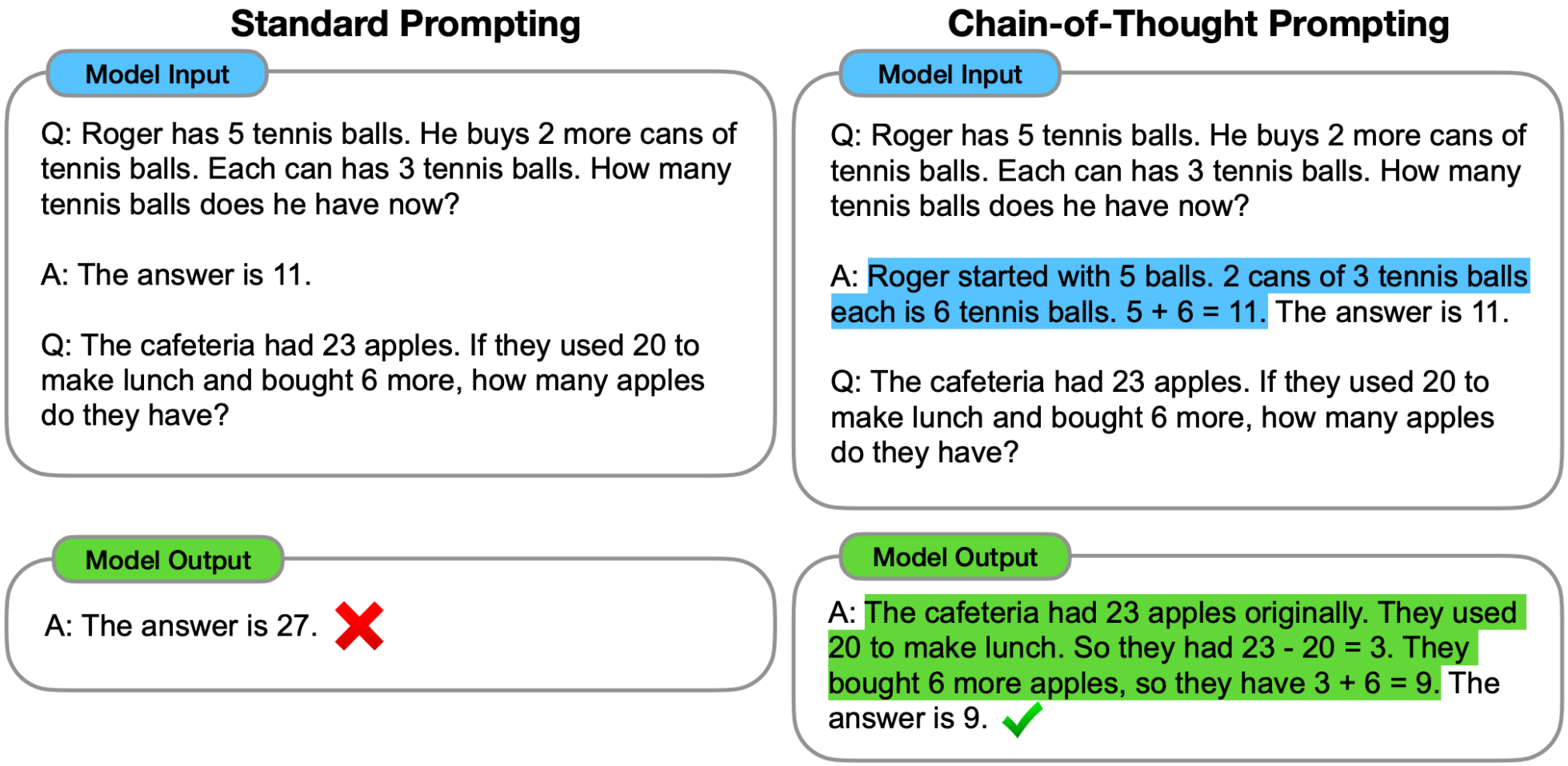
A:

Test Example

零样本/小样本提示(Zero-Shot 提示)

	No Prompt	Prompt
Zero-shot (0s)	skicts = sticks	Please unscramble the letters into a word, and write that word: skicts = sticks
1-shot (1s)	chiar = chair skicts = sticks	Please unscramble the letters into a word, and write that word: chiar = chair skicts = sticks
Few-shot (FS)	chiar = chair [...] pciinc = picnic skicts = sticks	Please unscramble the letters into a word, and write that word: chiar = chair [...] pciinc = picnic skicts = sticks

思维链(Chain-of-Thought Prompting)



Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS 2022. https://arxiv.org/pdf/2201.11903

思维链(Chain-of-Thought Prompting)

Direct Prompt

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: *The answer is 5 cars.*

Chain-of-Thought Prompt

Standard prompt:

- Q → A

Chain-of-thought prompt:

- Q → Reasoning Process, A

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

Thought (T): There are originally 3 cars. 2 more cars arrive. 3 + 2 = 5.

A: *The answer is 5 cars.*

Zero-Shot CoT Prompting

(a) Few-shot

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: 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:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

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: 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:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

(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 ✗

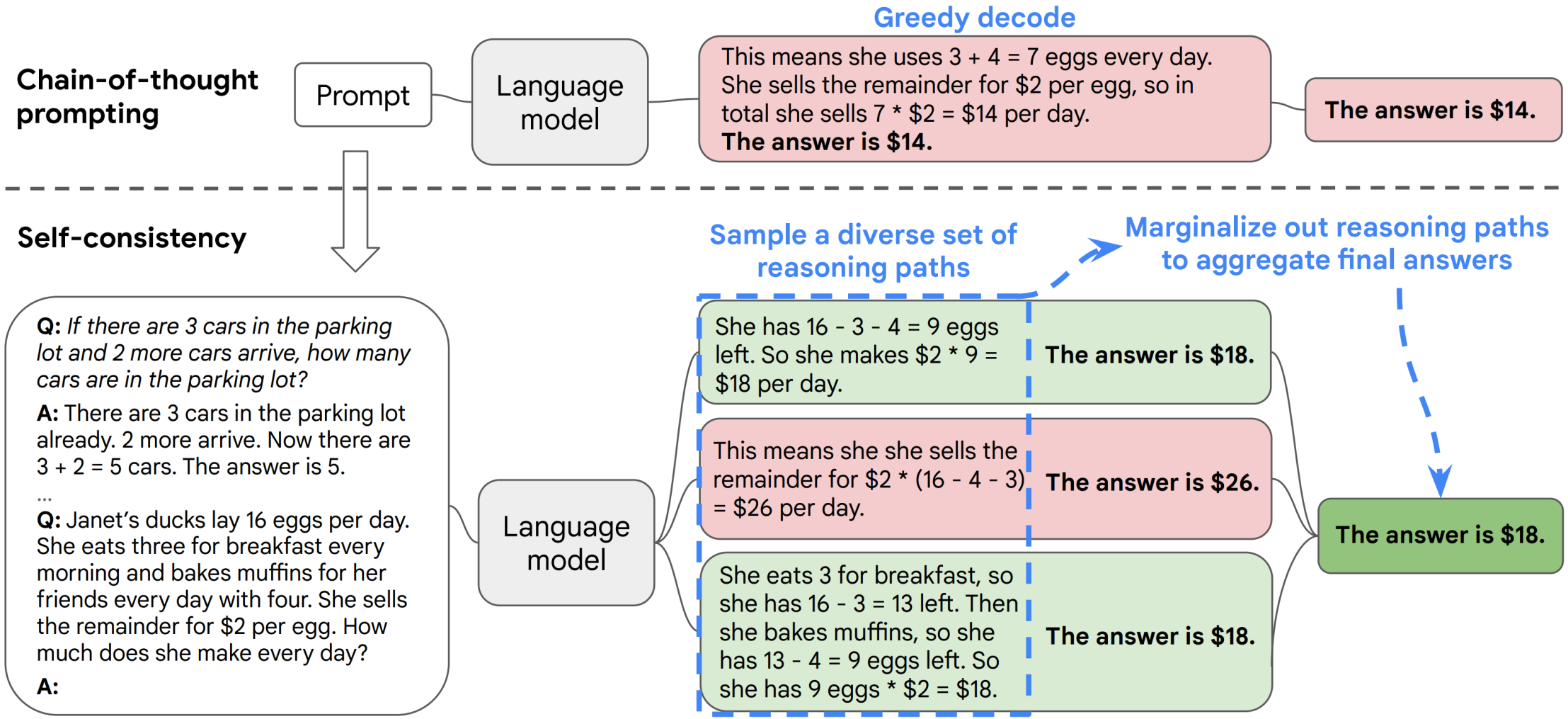
(d) Zero-shot-CoT (Ours)

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. ✓

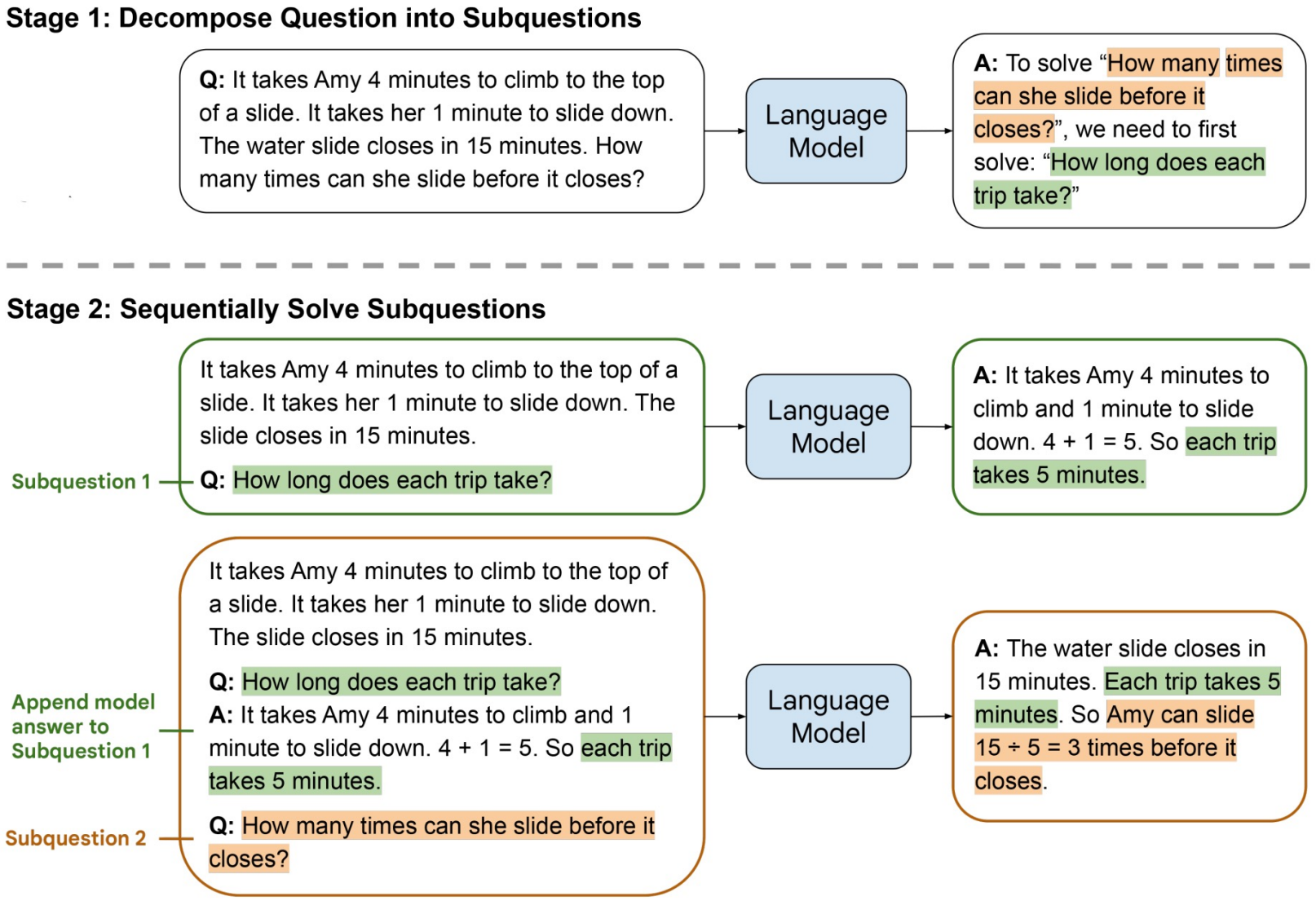
Large Language Models are Zero-Shot Reasoners. NeurIPS 2022. https://arxiv.org/pdf/2205.11916

Self-Consistency Prompting



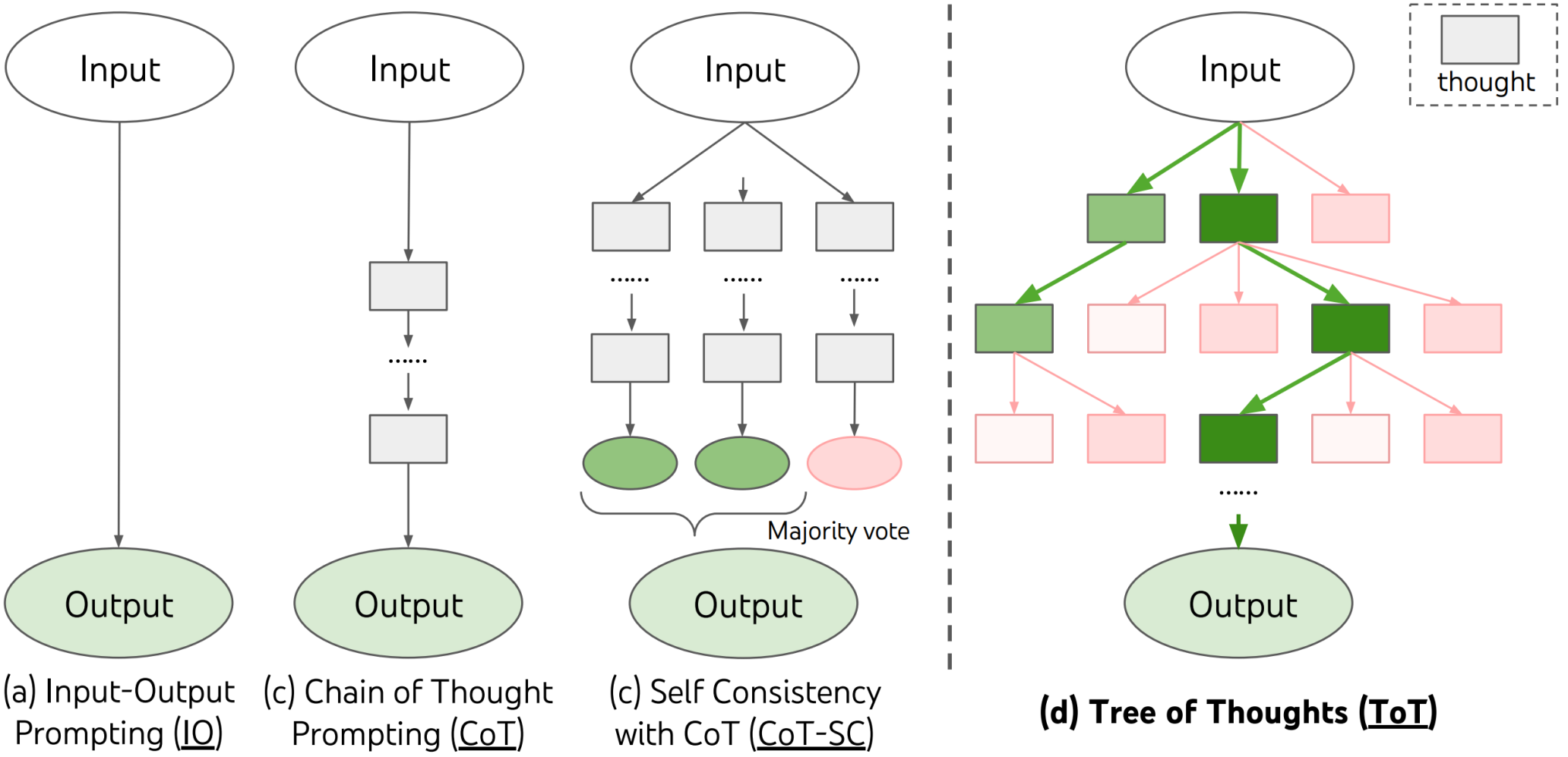
Self-Consistency Improves Chain of Thought Reasoning in Language Models. ICLR 2023. https://arxiv.org/pdf/2203.11171.pdf

Least-to-Most Prompting



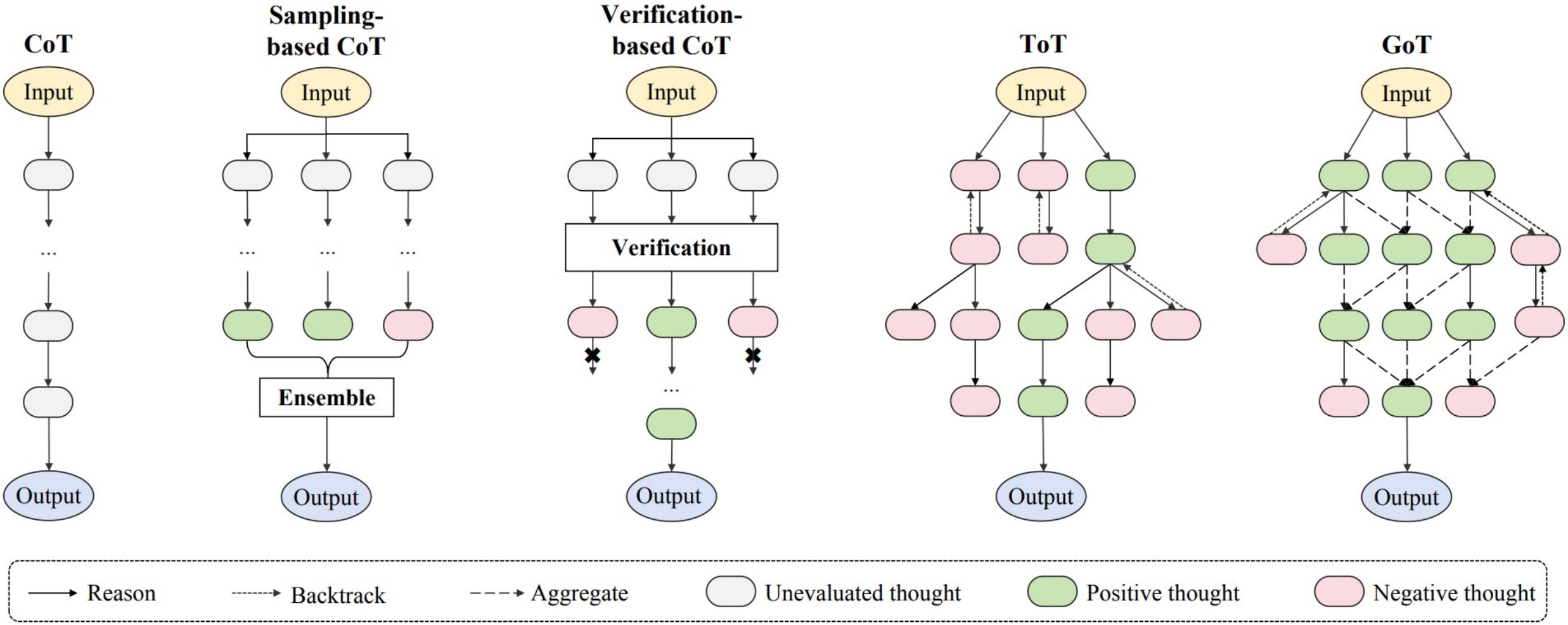
Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. ICLR 2023. https://arxiv.org/abs/2205.10625

Tree of Thought



Tree of Thoughts: Deliberate Problem Solving with Large Language Models. NeurIPS 2023. https://arxiv.org/pdf/2305.10601

Prompting Techniques



更多提示工程技巧

https://www.promptingguide.ai/zh

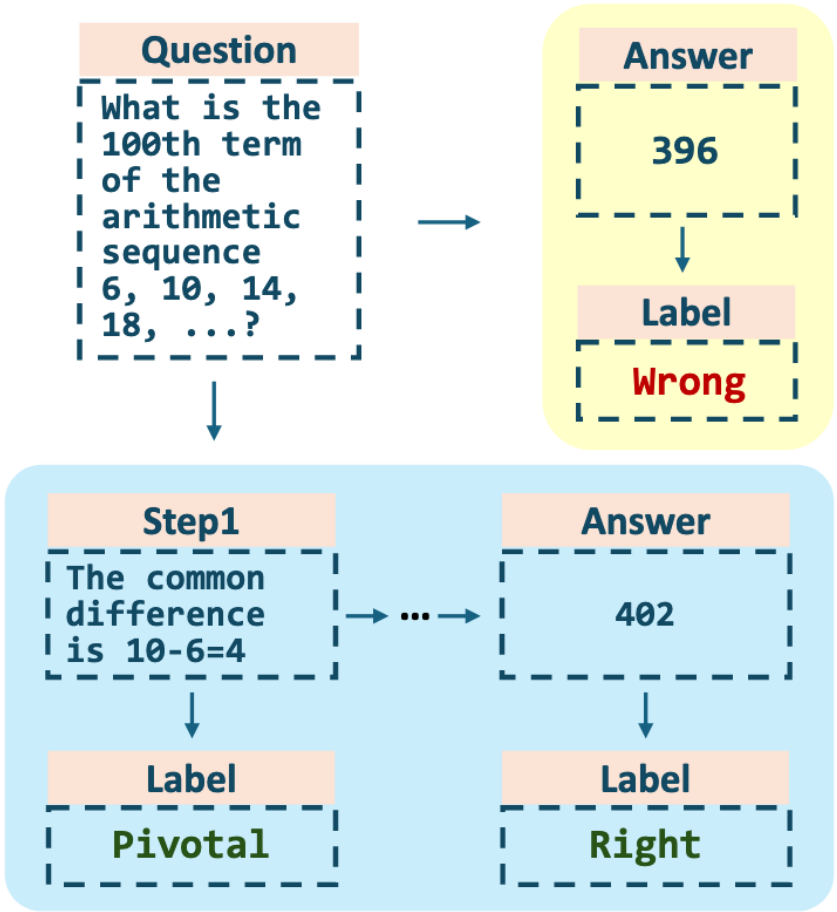
思维链(CoT)引爆了大模型社区，如何提升大模型的推理能力得到了大量的关注

Large Language Models

-> Large Reasoning Models

OpenAI: 过程标注(Process Label)

过程标注：对推理链的每个步骤进行打分



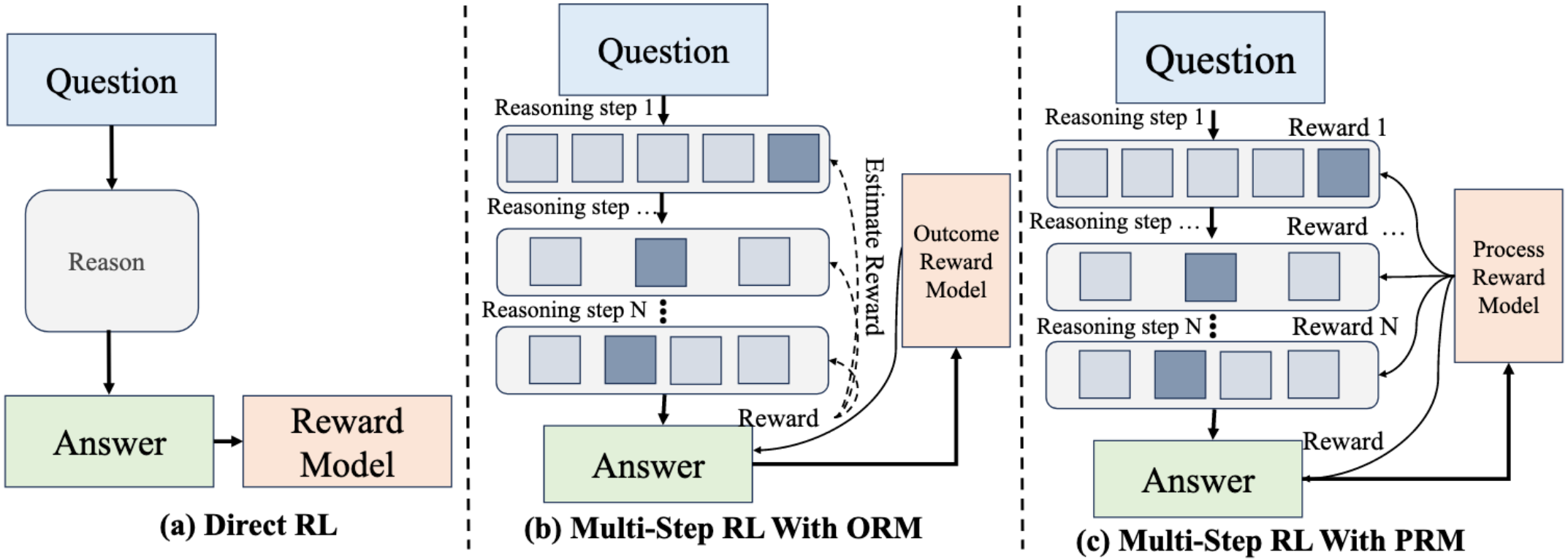
The denominator of a fraction is 7 less than 3 times the numerator. If the fraction is equivalent to 2/5, what is the numerator of the fraction? (Answer: 14)

- Let's call the numerator x.
- So the denominator is 3x-7.
- We know that x/(3x-7) = 2/5.
- So 5x = 2(3x-7).
- 5x = 6x - 14.
- So x = 7.

Let's Verify Step by Step. OpenAI. https://arxiv.org/abs/2305.20050

过程奖励模型(Processing Reward Model)

训练过程奖励模型(PRM)，基于过程奖励模型进行强化微调



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

DeepSeek-AI

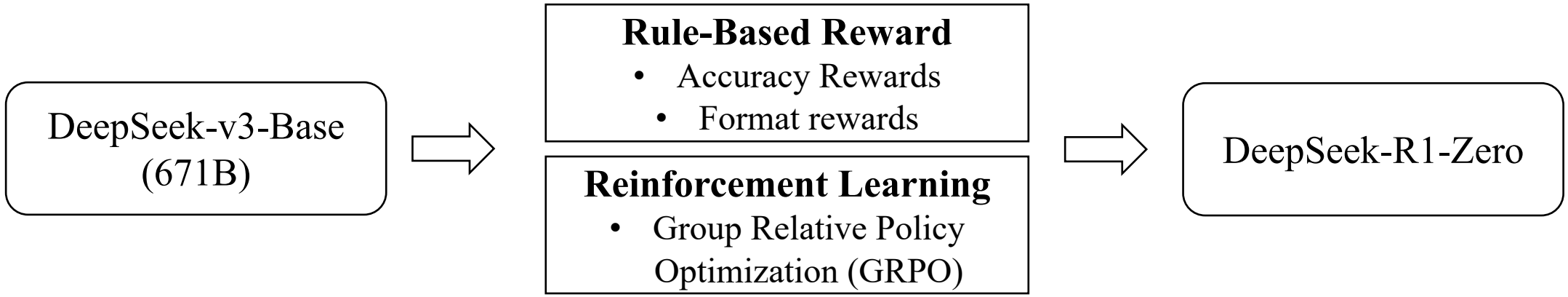
research@deepseek.com

https://arxiv.org/pdf/2501.12948

DeepSeek-R1 Zero

DeepSeek-R1 Zero: 无需SFT，纯RL驱动

不靠人类示范，只靠自己做题、自己对答案，把推理能力练出来



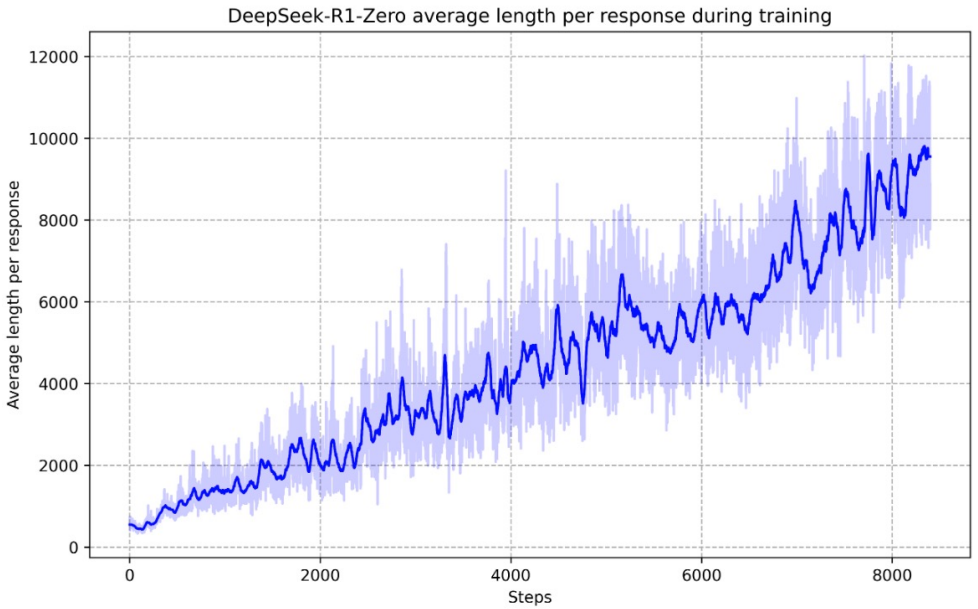
一句话介绍GRPO：通过**同一问题生成多条候选回答，并用组内相对奖励作为优势估计**来更新模型，从而减少对价值模型的依赖、提升训练效率。

DeepSeek-R1 Zero

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

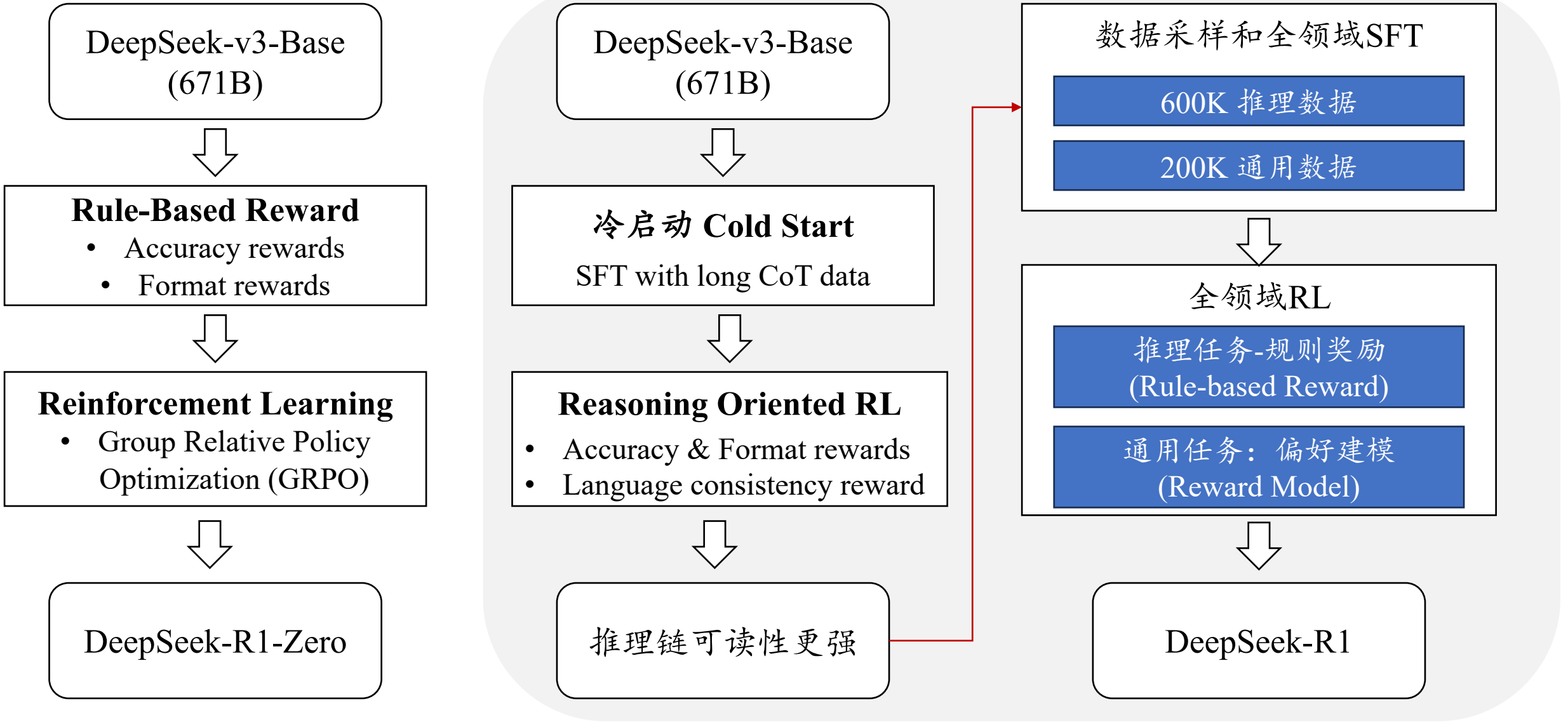
Response: <think>
To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both $\cdots$
$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2$.
Rearrange to isolate the inner square root term:
$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$
...
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 $\cdots$
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: $\cdots$
...

Aha Moment



RL驱动下自然涌现出Long-COT推理能力

DeepSeek-R1


DeepSeek-v3-Base
(671B)
Rule-Based Reward
• Accuracy rewards
• Format rewards
Reinforcement Learning
• Group Relative Policy Optimization (GRPO)
DeepSeek-R1-Zero
DeepSeek-v3-Base
(671B)
冷启动 Cold Start
SFT with long CoT data
Reasoning Oriented RL
• Accuracy & Format rewards
• Language consistency reward
推理链可读性更强
数据采样和全领域SFT
600K 推理数据
200K 通用数据
全领域RL
推理任务-规则奖励
(Rule-based Reward)
通用任务：偏好建模
(Reward Model)
DeepSeek-R1

在多种LLM评测基准中取得SOTA性能

	Benchmark (Metric)	R1-Zero	R1-Dev1	R1-Dev2	R1-Dev3	R1
English	MMLU (EM)	88.8	89.1	**91.2**	91.0	90.8
	MMLU-Redux (EM)	85.6	90.0	93.0	93.1	92.9
	MMLU-Pro (EM)	68.9	74.1	83.8	83.1	**84.0**
	DROP (3-shot F1)	89.1	89.8	91.1	88.7	**92.2**
	IF-Eval (Prompt Strict)	46.6	71.7	72.0	78.1	**83.3**
	GPQA Diamond (Pass@1)	**75.8**	66.1	70.7	71.2	71.5
	SimpleQA (Correct)	30.3	17.8	28.2	24.9	30.1
	FRAMES (Acc.)	82.3	78.5	81.8	81.9	**82.5**
	AlpacaEval2.0 (LC-winrate)	24.7	50.1	55.8	62.1	**87.6**
	ArenaHard (GPT-4-1106)	53.6	77.0	73.2	75.6	**92.3**
Code	LiveCodeBench (Pass@1-COT)	50.0	57.5	63.5	64.6	**65.9**
	Codeforces (Percentile)	80.4	84.5	90.5	92.1	**96.3**
	Codeforces (Rating)	1444	1534	1687	1746	**2029**
	SWE Verified (Resolved)	43.2	39.6	44.6	45.6	**49.2**
	Aider-Polyglot (Acc.)	12.2	6.7	25.6	44.8	**53.3**
Math	AIME 2024 (Pass@1)	77.9	59.0	74.0	78.1	**79.8**
	MATH-500 (Pass@1)	95.9	94.2	95.9	95.4	**97.3**
	CNMO 2024 (Pass@1)	**88.1**	58.0	73.9	77.3	78.8
Chinese	CLUEWSC (EM)	93.1	92.8	92.6	91.6	92.8
	C-Eval (EM)	**92.8**	85.7	91.9	86.4	91.8
	C-SimpleQA (Correct)	66.4	58.8	64.2	66.9	63.7

雨后春笋：XX-R1

VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model

Haozhan Shen[1], Peng Liu[2], Jingcheng Li[2], Chunxin Fang[2], Yibo Ma[2], Jiajia Liao[2], Qiaoli Shen[2], Zilun Zhang[1], Kangjia Zhao[1], Qianqian Zhang[2], Ruochen Xu[2], Tiancheng Zhao[2,3]✉

[1] Zhejiang University [2] Om AI Research [3] Binjiang Institute of Zhejiang University

{tianchez}@zju-bj.com

WEBAGENT-R1: Training Web Agents via End-to-End Multi-Turn Reinforcement Learning

Zhepei Wei[†*], Wenlin Yao[‡], Yao Liu[‡], Weizhi Zhang[‡], Qin Lu[‡], Liang Qiu[‡]
Changlong Yu[‡], Puyang Xu[‡], Chao Zhang[§], Bing Yin[‡], Hyokun Yun[‡], Lihong Li[‡]
[†]University of Virginia [‡]Amazon [§]Georgia Institute of Technology
zhepei.wei@virginia.edu, chaozhang@gatech.edu
{ywenlin, yaoliuai, zhweizhi, luqn, liangqxx, changlyu
puyax, alexbyin, yunhyoku, llh}@amazon.com

Abstract

While reinforcement learning (RL) has demonstrated remarkable success in enhancing large language models (LLMs), it has primarily focused on single-turn tasks such as solving math problems. Training effective web agents for multi-turn interactions remains challenging due to the complexity of long-horizon decision-making across dynamic web interfaces. In this work, we present WEBAGENT-R1, a simple yet effective end-to-end multi-turn RL framework for training web agents. It learns directly from online interactions with web environments by generating diverse trajectories in parallel, entirely guided by binary rewards depending on task success. Experiments on the WebArena-Lite benchmark demonstrate the effectiveness of WEBAGENT-R1, boosting the task success rate of Qwen-2.5-3B from 6.1% to 33.9% and Llama-3.1-8B from 8.5% to 44.8%, signifi-

single-turn, non-interactive tasks such as mathematical reasoning (Shao et al., 2024; Zeng et al., 2025). Their effectiveness in multi-turn, interactive environments—particularly in complex scenarios requiring long-horizon decision-making and domain-specific skills, such as web browsing (Zhou et al., 2024a; He et al., 2024a; Chae et al., 2025)—still remains underexplored.

Unlike static environments, web tasks pose unique challenges for LLM agents due to their dynamic nature and diverse solution spaces. Early works on web agents primarily relied on prompting-based methods (Wang et al., 2024b; Sodhi et al., 2024; Fu et al., 2024; Zhang et al., 2025; Yang et al., 2025b) or behavior cloning (BC), which imitates demonstrated trajectories via supervised fine-tuning (Yin et al., 2024; Hong et al., 2024; Lai et al., 2024; He et al., 2024b; Putta et al., 2024). Despite

Published as a conference paper at COLM 2025

Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning

Bowen Jin[1], Hansi Zeng[2], Zhenrui Yue[1], Jinsung Yoon[3], Sercan Ö. Arık[3], Dong Wang[1], Hamed Zamani[2], Jiawei Han[1]
[1] Department of Computer Science, University of Illinois at Urbana-Champaign
[2] Center for Intelligent Information Retrieval, University of Massachusetts Amherst
[3] Google Cloud AI Research
{bowenj4,zhenrui3,dwang24,hanj}@illinois.edu, {hzeng, zamani}@cs.umass.edu
{jinsungyoon,soarik}@google.com

Abstract

Efficiently acquiring external knowledge and up-to-date information is essential for effective reasoning and text generation in large language models (LLMs). Prompting advanced LLMs with reasoning capabilities to use search engines during inference is often suboptimal, as the LLM might not fully possess the capability on how to interact optimally with the search engine. This paper introduces SEARCH-R1, an extension of reinforcement learning (RL) for reasoning frameworks where the LLM learns to autonomously generate (multiple) search queries during step-by-step reasoning with real-time retrieval. SEARCH-R1 opti-

Agent-R1: Training Powerful LLM Agents with End-to-End Reinforcement Learning

Mingyue Cheng, Jie Ouyang, Shuo Yu, Ruiran Yan, Yucong Luo,
Zirui Liu, Daoyu Wang, Qi Liu, Enhong Chen
State Key Laboratory of Cognitive Intelligence,
University of Science and Technology of China, Hefei, China

Abstract

Large Language Models (LLMs) are increasingly being explored for building Agents capable of active environmental interaction (e.g., via tool use) to solve complex problems. Reinforcement Learning (RL) is considered a key technology with significant potential for training such Agents; however, the effective application of RL to LLM Agents is still in its nascent stages and faces considerable challenges. Currently, this emerging field lacks in-depth exploration into RL approaches specifically tailored for the LLM Agent context, alongside a scarcity of flexible and easily extensible training frameworks designed for this purpose. To help advance this area, this paper first revisits and clarifies Reinforcement Learning methodologies for LLM Agents by systematically extending the Markov Decision Process (MDP) framework to comprehensively define the key components of an LLM Agent. Secondly, we introduce Agent-R1, a modular, flexible, and user-friendly training framework for RL-based LLM Agents, designed for straightforward adaptation across diverse task scenarios and interactive environments. We conducted experiments on Multihop QA benchmark tasks, providing initial validation for the effectiveness of our proposed methods and framework.

https://github.com/0russwest0/Agent-R1. [1]

VISION-R1: INCENTIVIZING REASONING CAPABILITY IN MULTIMODAL LARGE LANGUAGE MODELS

Wenxuan Huang[1,2*†] Bohan Jia[1*] Zijie Zhai[1] Shaosheng Cao[3✉]
Zheyu Ye[3] Fei Zhao[3] Zhe Xu[3] Xu Tang[3] Yao Hu[3] Shaohui Lin[1✉]
[1]East China Normal University [2]The Chinese University of Hong Kong [3]Xiaohongshu Inc.
wxhuang0616@gmail.com (Wenxuan Huang)
*: Equal Contribution †: Project Leader ✉: Corresponding Author

ABSTRACT

DeepSeek-R1-Zero has successfully demonstrated the emergence of reasoning capabilities in LLMs purely through Reinforcement Learning (RL). Inspired by this breakthrough, we explore how RL can be utilized to enhance the reasoning capability of MLLMs. However, direct training with RL struggles to activate complex reasoning capabilities such as questioning and reflection in MLLMs, due to the absence of substantial high-quality multimodal reasoning data. To address this issue, we propose the reasoning MLLM, ***Vision-R1***, to improve multimodal reasoning capability. Specifically, we first construct a high-quality multimodal CoT dataset without human annotations by leveraging an existing MLLM and DeepSeek-R1 through modality bridging and data filtering to obtain a 200K multimodal CoT dataset, ***Vision-R1-cold dataset***. It serves as cold-start initialization data for Vision-R1. To mitigate the optimization challenges caused by overthinking after cold start, we propose ***Progressive Thinking Suppression Training (PTST)*** strategy and employ Group Relative Policy Optimization (GRPO) with the hard formatting result reward function to gradually refine the model's ability to learn correct and complex reasoning processes on the multimodal math dataset. Comprehensive experiments show our model achieves an average improvement of ∼6% across various multimodal math reasoning benchmarks using only a 10K multimodal math data during RL training. Vision-R1-7B achieves a 73.5% accuracy on the widely used MathVista benchmark, which is only 0.4% lower than the leading reasoning model, OpenAI O1. Scaling up the amount of multimodal math data in the RL training, Vision-R1-32B and Vison-R1-72B achieves 76.4% and 78.2% MathVista benchmark scores, respectively. The datasets, weight and code will be released in: https://github.com/Osilly/Vision-R1.

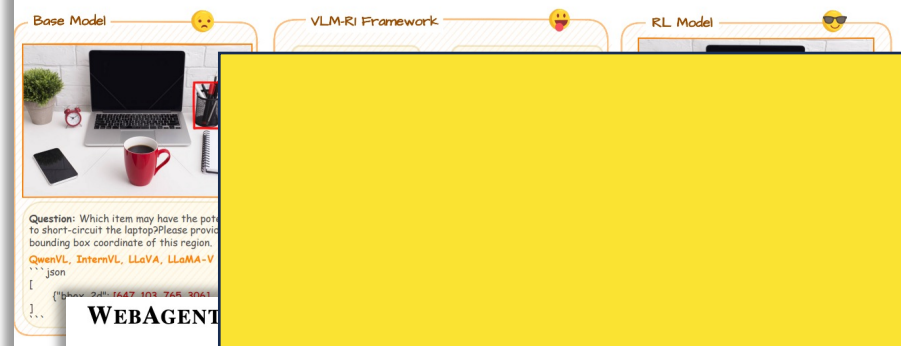
雨后春笋：XX-R1

VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model

Haozhan Shen[1], Peng Liu[2], Jingcheng Li[2], Chunxin Fang[2], Yibo Ma[2], Jiajia Liao[2], Qiaoli Shen[2], Zilun Zhang[1], Kangjia Zhao[1], Qianqian Zhang[2], Ruochen Xu[2], Tiancheng Zhao[2,3]✉

[1] Zhejiang University [2] Om AI Research [3] Binjiang Institute of Zhejiang University

{tianchez}@zju-bj.com



WEBAGENT

Zhepei Wei
Changlong Yu
[†]Univers
{ywen

While reinforcement learning (RL) has demonstrated remarkable success in enhancing large language models (LLMs), it has primarily focused on single-turn tasks such as solving math problems. Training effective web agents for multi-turn interactions remains challenging due to the complexity of long-horizon decision-making across dynamic web interfaces. In this work, we present WEBAGENT-R1, a simple yet effective end-to-end multi-turn RL framework for training web agents. It learns directly from online interactions with web environments by generating diverse trajectories in parallel, entirely guided by binary rewards depending on task success. Experiments on the WebArena-Lite benchmark demonstrate the effectiveness of WEBAGENT-R1, boosting the task success rate of Qwen-2.5-3B from 6.1% to 33.9% and Llama-3.1-8B from 8.5% to 44.8%, signifi-

2025). Their effectiveness in multi-turn, interactive environments—particularly in complex scenarios requiring long-horizon decision-making and domain-specific skills, such as web browsing (Zhou et al., 2024a; He et al., 2024a; Chae et al., 2025)—still remains underexplored.

Unlike static environments, web tasks pose unique challenges for LLM agents due to their dynamic nature and diverse solution spaces. Early works on web agents primarily relied on prompting-based methods (Wang et al., 2024b; Sodhi et al., 2024; Fu et al., 2024; Zhang et al., 2025; Yang et al., 2025b) or behavior cloning (BC), which imitates demonstrated trajectories via supervised fine-tuning (Yin et al., 2024; Hong et al., 2024; Lai et al., 2024; He et al., 2024b; Putta et al., 2024). Despite

Published as a conference paper at COLM 2025

Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning

Bowen Jin[1], **Hansi Zeng**[2], **Zhenrui Yue**[1], **Jinsung Yoon**[3], **Sercan Ö. Arık**[3], **Dong Wang**[1], **Hamed Zamani**[2], **Jiawei Han**[1]
[1] Department of Computer Science, University of Illinois at Urbana-Champaign
[2] Center for Intelligent Information Retrieval, University of Massachusetts Amherst

Abstract

Large Language Models (LLMs) are increasingly being explored for building Agents capable of active environmental interaction (e.g., via tool use) to solve complex problems. Reinforcement Learning (RL) is considered a key technology with significant potential for training such Agents; however, the effective application of RL to LLM Agents is still in its nascent stages and faces considerable challenges. Currently, this emerging field lacks in-depth exploration into RL approaches specifically tailored for the LLM Agent context, alongside a scarcity of flexible and easily extensible training frameworks designed for this purpose. To help advance this area, this paper first revisits and clarifies Reinforcement Learning methodologies for LLM Agents by systematically extending the Markov Decision Process (MDP) framework to comprehensively define the key components of an LLM Agent. Secondly, we introduce Agent-R1, a modular, flexible, and user-friendly training framework for RL-based LLM Agents, designed for straightforward adaptation across diverse task scenarios and interactive environments. We conducted experiments on Multihop QA benchmark tasks, providing initial validation for the effectiveness of our proposed methods and framework.

https://github.com/0russwest0/Agent-R1. [1]

VISION-R1: INCENTIVIZING REASONING CAPABILITY IN MULTIMODAL LARGE LANGUAGE MODELS

Wenxuan Huang[1,2*†] Bohan Jia[1*] Zijie Zhai[1] Shaosheng Cao[3✉]
[u][3] Shaohui Lin[1✉]
ong [3]Xiaohongshu Inc.

ence of reasoning
(RL). Inspired by
nce the reasoning
uggles to activate
n in MLLMs, due
g data. To address
prove multimodal
y multimodal CoT
LM and DeepSeek-
K multimodal CoT
on data for Vision-
hinking after cold
g (PTST) strategy
he hard formatting
learn correct and
. Comprehensive
nt of ∼6% across
K multimodal math
data during RL training. Vision-R1-7B achieves a 73.5% accuracy on the widely used MathVista benchmark, which is only 0.4% lower than the leading reasoning model, OpenAI O1. Scaling up the amount of multimodal math data in the RL training, Vision-R1-32B and Vison-R1-72B achieves 76.4% and 78.2% MathVista benchmark scores, respectively. The datasets, weight and code will be released in: https://github.com/Osilly/Vision-R1.

标准三板斧：

数据合成 -> 监督微调SFT -> 强化微调GRPO

阅读材料

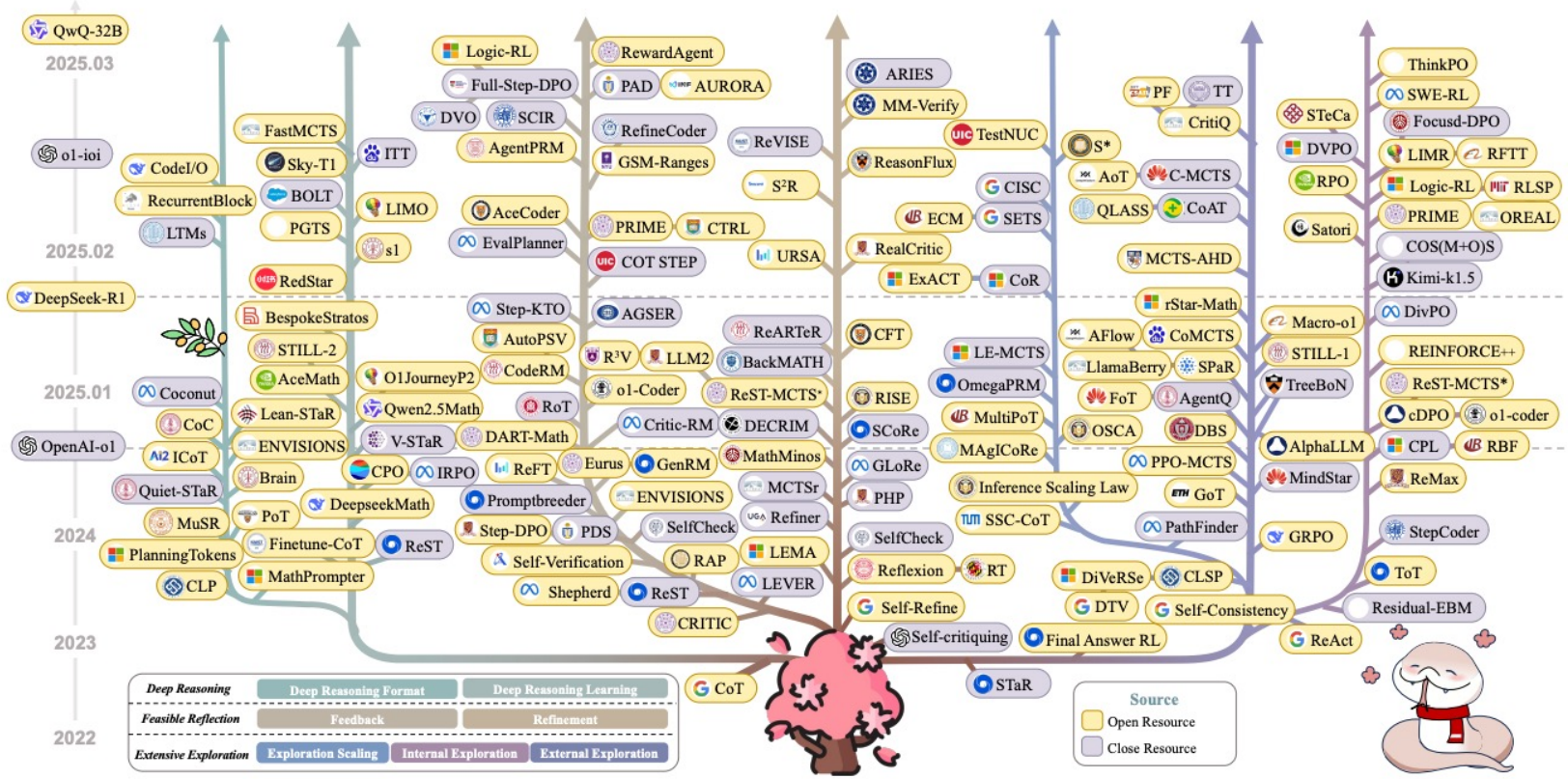
Towards Reasoning Era: A Survey of Long Chain-of-Thought for Reasoning Large Language Models

Qiguang Chen† Libo Qin‡ Jinhao Liu† Dengyun Peng† Jiannan Guan†
Peng Wang‡ Mengkang Hu◇ Yuhang Zhou♡ Te Gao‡ Wanxiang Che†
† LARG,
† Research Center for Social Computing and Interactive Robotics,
† Harbin Institute of Technology
‡ School of Computer Science and Engineering, Central South University
◇ The University of Hong Kong
♡ Fudan University
{qgchen,car}@ir.hit.edu.cn, lbqin@csu.edu.cn

Project: https://long-cot.github.io/

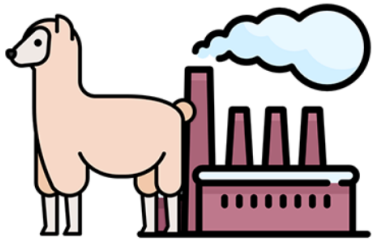
Github: LightChen233/Awesome-Long-Chain-of-Thought-Reasoning

https://arxiv.org/pdf/2503.09567



LLaMA Factory

https://llamafactory.readthedocs.io/zh-cn/latest/index.html



LLaMA-Factory

Easy and Efficient LLM Fine-Tuning

LLaMA Factory 是一个简单易用且高效的大型语言模型（Large Language Model）训练与微调平台。通过 LLaMA Factory，可以在无需编写任何代码的前提下，在本地完成上百种预训练模型的微调，框架特性包括：

- 模型种类：LLaMA、LLaVA、Mistral、Mixtral-MoE、Qwen、Yi、Gemma、Baichuan、ChatGLM、Phi 等等。
- 训练算法：（增量）预训练、（多模态）指令监督微调、奖励模型训练、PPO 训练、DPO 训练、KTO 训练、ORPO 训练等等。
- 运算精度：16 比特全参数微调、冻结微调、LoRA 微调和基于 AQLM/AWQ/GPTQ/LLM.int8/HQQ/EETQ 的 2/3/4/5/6/8 比特 QLoRA 微调。
- 优化算法：GaLore、BAdam、DoRA、LongLoRA、LLaMA Pro、Mixture-of-Depths、LoRA+、LoftQ 和 PiSSA。
- 加速算子：FlashAttention-2 和 Unsloth。
- 推理引擎：Transformers 和 vLLM。
- 实验监控：LlamaBoard、TensorBoard、Wandb、MLflow、SwanLab 等等。

一些观点

SFT vs RLFT

- 对于形式化推理任务，例如math、code，更容易设计Rule-based reward

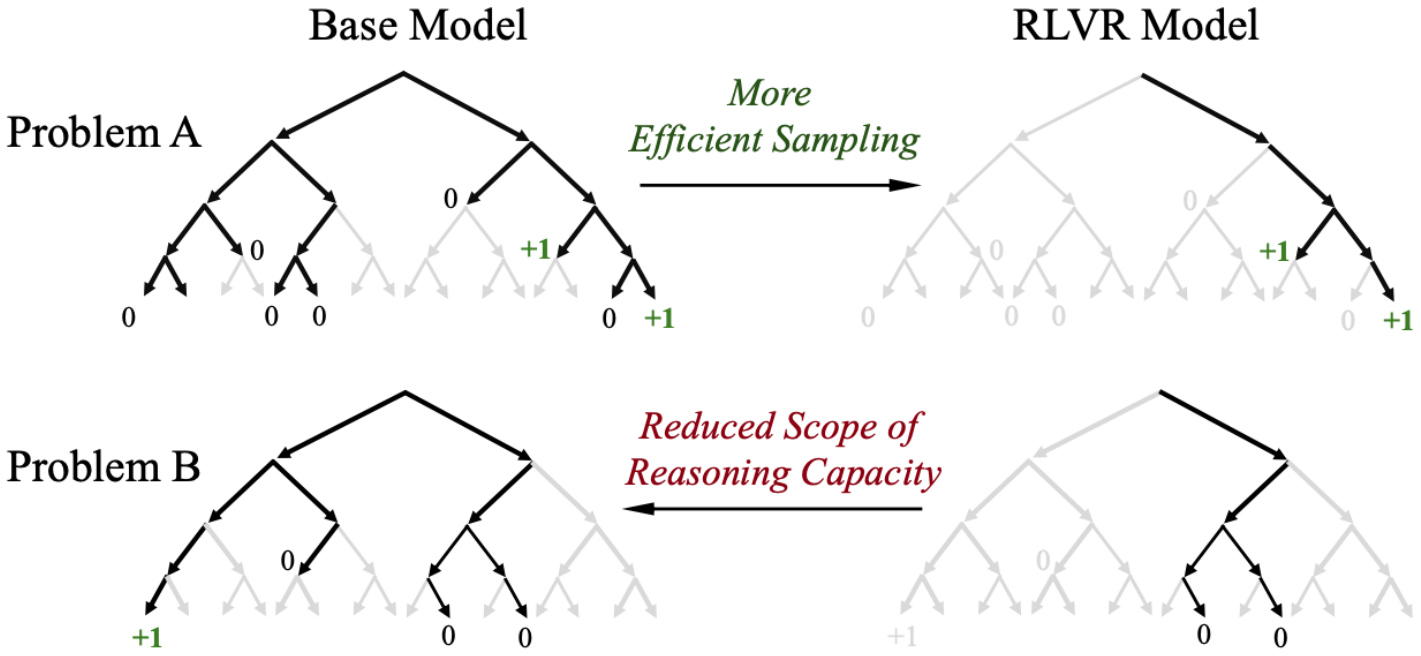
  - **SFT vs RL**

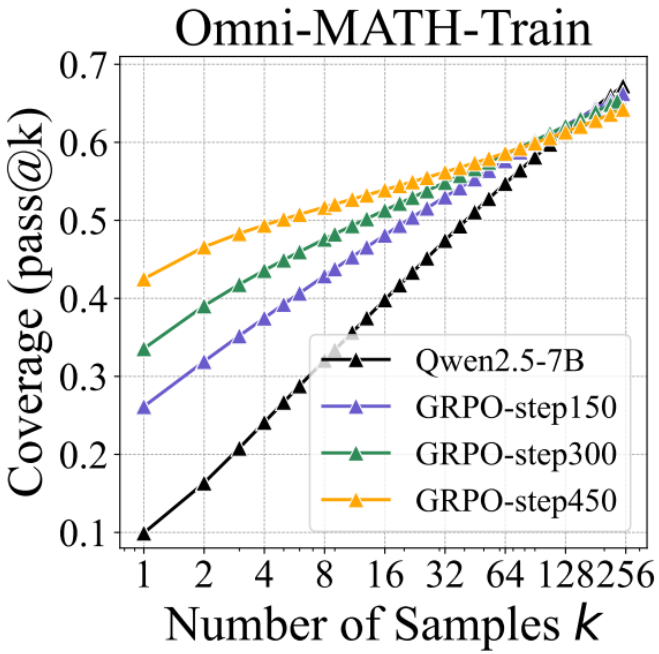
    - 直接利用SFT更多是去拟合数据中的pattern，难以学习到背后的规律；
    - RL则是通过鼓励模型在最大化reward的过程中学习到推理背后的规律，获得的泛化性和推理表现上界更高
    - SFT规范模型输出格式，使得后续的RL可以获得更高收益 [1]
    - 小模型RL不容易出现Long CoT [2]

[1] SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. https://arxiv.org/pdf/2501.17161
[2] Demystifying Long Chain-of-Thought Reasoning in LLMs. https://arxiv.org/pdf/2502.03373

RL能否带来推理能力？

观点：RL并不一定真正扩展了大模型的推理能力，它更多是在提升采样效率、强化 base model 已经会的高奖励推理路径





Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? https://arxiv.org/pdf/2504.13837

LLM是记忆还是推理?

LLM是记忆还是推理?

GSM-NoOp

Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

---

o1-mini: Oliver picks kiwis over three days as follows:
- Friday: 44 kiwis
- Saturday: 58 kiwis
- Sunday: Double the number he picked on Friday, which is $2 \times 44 = 88$ kiwis

However, on Sunday, 5 of these kiwis were smaller than average. We need to subtract them from the Sunday total: 88 (Sunday's kiwis) - 5 (smaller kiwis) = 83 kiwis Now, summing up the kiwis from all three days: 44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis Oliver has a total of 185 kiwis.

Llama3-8B: Let's break it down step by step:
Oliver picks 44 kiwis on Friday. He picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, which is 2 x 44 = 88 kiwis.
Five of the kiwis picked on Sunday are a bit smaller than average, so we subtract 5 from the total number of kiwis picked on Sunday: 88 - 5 = 83 kiwis. Now, let's add up the total number of kiwis Oliver has:
44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis
So, Oliver has 185 kiwis in total.

[1] GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. ICLR 2024.

观点：大模型+工具

如果能够赋予大模型工具调用的能力，让模型学会调用大量已有的工具处理任务，就可以进一步提升大模型能力边界