



大语言模型的能力边界与发展思考

张奇

复旦大学

2023年--大模型报税、写诗、写代码,“无所不能”

科技 医生 医学 临床医学 OpenAI

关注者 434 被浏览 394,198

ChatGPT 成功诊断 4 岁男孩怪病, 击败 17 位医生, 这意味着什么?

美国一位女士最近求助热门机器人聊天工具ChatGPT成功地诊断出了她4岁儿子患的罕见疾病, 此前17位医生都未能找出问题所在。这位女士名叫考特尼 (Courtney)...显示全部

Posted by u/rustyryan 5 days ago

2.3k A boy saw 17 doctors over 3 years for chronic pain. ChatGPT found the diagnosis

today.com/health...

Use cases



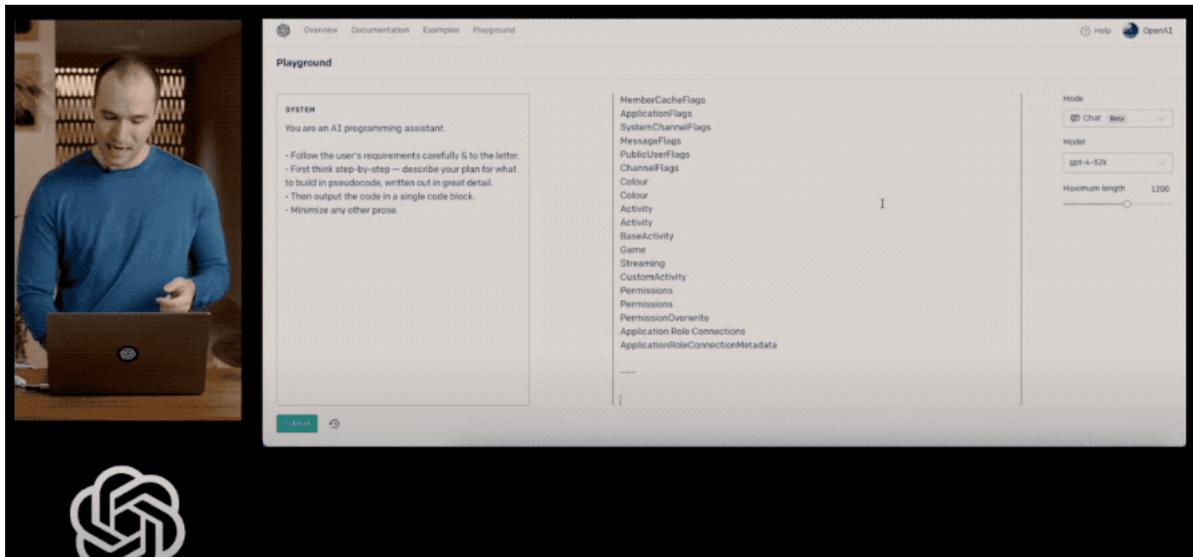
User Below is part of the InstructGPT paper. Could you read and summarize it to me?



GPT-4 The InstructGPT paper focuses on training large language models to follow instructions with human feedback. The authors note that making language models larger doesn't inherently make them better at following a user's intent. Large models can generate outputs that are untruthful, toxic, or simply unhelpful.

To address this issue, the authors fine-tune language models on a wide range of tasks using human feedback. They start with a set of labeler-written prompts and responses, then collect a dataset of labeler demonstrations of the desired model behavior. They fine-tune GPT-3 using supervised learning and then use reinforcement learning from human feedback to further fine-tune the model. The resulting model, called InstructGPT, shows improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets.

The authors conclude that fine-tuning with human feedback is a promising direction for aligning language models with human intent.



2023年--大模型“无所不能”

... 航母舰载机起落架建模和仿真方案

GPT-4o Poe

进行航母舰载机起落架建模和仿真的方案可以分为以下几个步骤：

1. 需求分析

- 确定建模目的：例如，分析起落架在航母降落时的动态响应。
- 识别关键参数：如重量、着陆速度、航母甲板运动等。

2. 建立数学模型

- 动力学方程：利用牛顿力学和欧拉方程建立运动方程。
- 材料特性：考虑起落架的材料属性，如弹性、阻尼等。
- 结构特性：建立起落架的几何模型，包括尺寸、形状等。

3. 计算机建模

- CAD建模：使用软件如SolidWorks或CATIA建立三维几何模型。
- 有限元分析 (FEA)：使用ANSYS或Abaqus进行静态和动态分析，评估应力和变形。

4. 仿真环境设置

- 仿真工具：选择合适的仿真工具，如MATLAB/Simulink进行系统仿真。
- 输入参数设置：设定初始条件、边界条件及载荷情况。

5. 进行仿真

- 动态仿真：模拟起落架在各种工况下的响应。
- 数据分析：观察响应曲线，分析起落架性能。

6. 验证与优化

- 实验验证：通过实验数据验证仿真模型的准确性。
- 参数优化：调整模型参数以提高仿真精度。

7. 报告与文档

- 结果汇报：总结仿真结果，生成报告。
- 模型文档：记录模型和仿真过程以便于后续参考和改进。

动力学方程再详细一点

GPT-4o Poe

动力学方程详解

为了建立航母舰载机起落架的动力学方程，需要考虑以下几个方面：

1. 自由体图

- 分析起落架系统的自由体图，确定作用力和力矩。

2. 牛顿第二定律

- 平移运动：对于起落架的垂直运动，使用牛顿第二定律：

$$m \cdot a = \sum F$$

其中， m 是质量， a 是加速度， F 是作用力（如重力、支撑力、阻尼力）。

- 旋转运动：如果考虑起落架的旋转运动，则使用欧拉方程：

$$I \cdot \alpha = \sum \tau$$

其中， I 是转动惯量， α 是角加速度， τ 是力矩。

3. 弹性和阻尼模型

- 弹簧力：

$$F_{\text{spring}} = k \cdot x$$

其中， k 是弹簧刚度， x 是压缩量。

- 阻尼力：

$$F_{\text{damping}} = c \cdot v$$

其中， c 是阻尼系数， v 是速度。

4. 接触力



2023年--大模型“无所不能”

2023 年大家拿着锤子到处找钉子



2024年--大模型落地总是“差一口气”



2024年--大模型落地总是“差一口气”

ChatGPT ▾

ChatGPT-4o

9.11 比 9.9 大。比较这两个数字的方法是直接看小数点后面的数字。因为 11 大于 9，所以 9.11 大

其他比较方法?

除了直接比较小数点后面的数字外，还有几种方法可以比较 9.11 和 9.9：1. **转化为分数**：- 9.11 可以写成 $9 + 11/100$ - 9.9 可以写成 $9 + 90/100$ 将这两个分数部分比较， $11/100$ 比 $90/100$ 小，因此 9.11 比 9.9 大。2.

从数学上来说，9.11和9.9哪个大?

从数学上来说，9.11 比 9.9 大。具体比较方法如下：1. **直接比较小数位**：- 比较整数部分：9 等于 9。- 比较小数部分：0.11 大于 0.9。2.

🔊 📄 🗑️ ⚙️ ▾



ChatGPT



ChatGPT ▾



Strawberry这个单词有几个r?

"Strawberry" 这个单词有两个 "r"。

📄 🔄 🗑️ ⚙️ ▾

🔗 | 给“ChatGPT”发送消息



2024年--大模型落地总是“差一口气”

Jane Austen Novel Task

Inverse search: “In <Pride and Prejudice>, what’s the sentence **before**: <sentence2>?”

Forward search: “In <Pride and Prejudice>, what’s the sentence **after**: <sentence1>?”

	Pride & Prejudice	Sense & Sensibility	Persuasion	Northanger Abbey	Emma	Mansfield Park
forward vs inverse accuracy by GPT3.5	0.5% vs 14.4%	0.3% vs 5.4%	0.07% vs 4.3%	0.6% vs 5.5%	0.8% vs 7.2%	0.7% vs 5.5%
forward vs inverse accuracy by GPT4	0.8% vs 65.9%	0.9% vs 40.2%	0.5% vs 33.9%	0.9% vs 41.0%	0.6% vs 42.7%	0.3% vs 31.7%

Inverse search: “what’s the full name of the celebrity born on <date> in <city> who is a <occupation>?”

Forward search: “what’s the birthday and year of <name> who is a <occupation> and was born in <city>?”

GPT3.5 acc = 23.9%

GPT4: 42%

GPT3.5 acc = 89.5%

GPT4: 99%

Wiki Bio Task

Chinese Idiom Task

Given a common 4-letter Chinese idiom such as 指鹿为马, mask out its i-th letter (for i=1,2,3, or 4) and let GPT fill out the missing letter.

Prompt 1: 成语“X鹿为马”的X是什么字?

GPT3.5 accuracy 9.4%,

GPT4 accuracy 17.6%

Prompt 2: 成语“指X为马”的X是什么字?

GPT3.5 accuracy 29.5%,

GPT4 accuracy 36.1%

Prompt 3: 成语“指鹿X马”的X是什么字?

GPT3.5 accuracy 32.0%,

GPT4 accuracy 76.7%

Prompt 4: 成语“指鹿为X”的X是什么字?

GPT3.5 accuracy 56.7%,

GPT4 accuracy 90.6%

Given a famous two-sentence Chinese poem such as 劝君更尽一杯酒, 西出阳关无故人, let GPT answer what’s the sentence **before/after** <sentence2/1>

Inverse search: “西出阳关无故人”的上一句是什么?

GPT3.5 accuracy 2.1%,

GPT4 accuracy 7.3%

Forward search: “劝君更尽一杯酒”的下一句是什么?

GPT3.5 accuracy 33.0%,

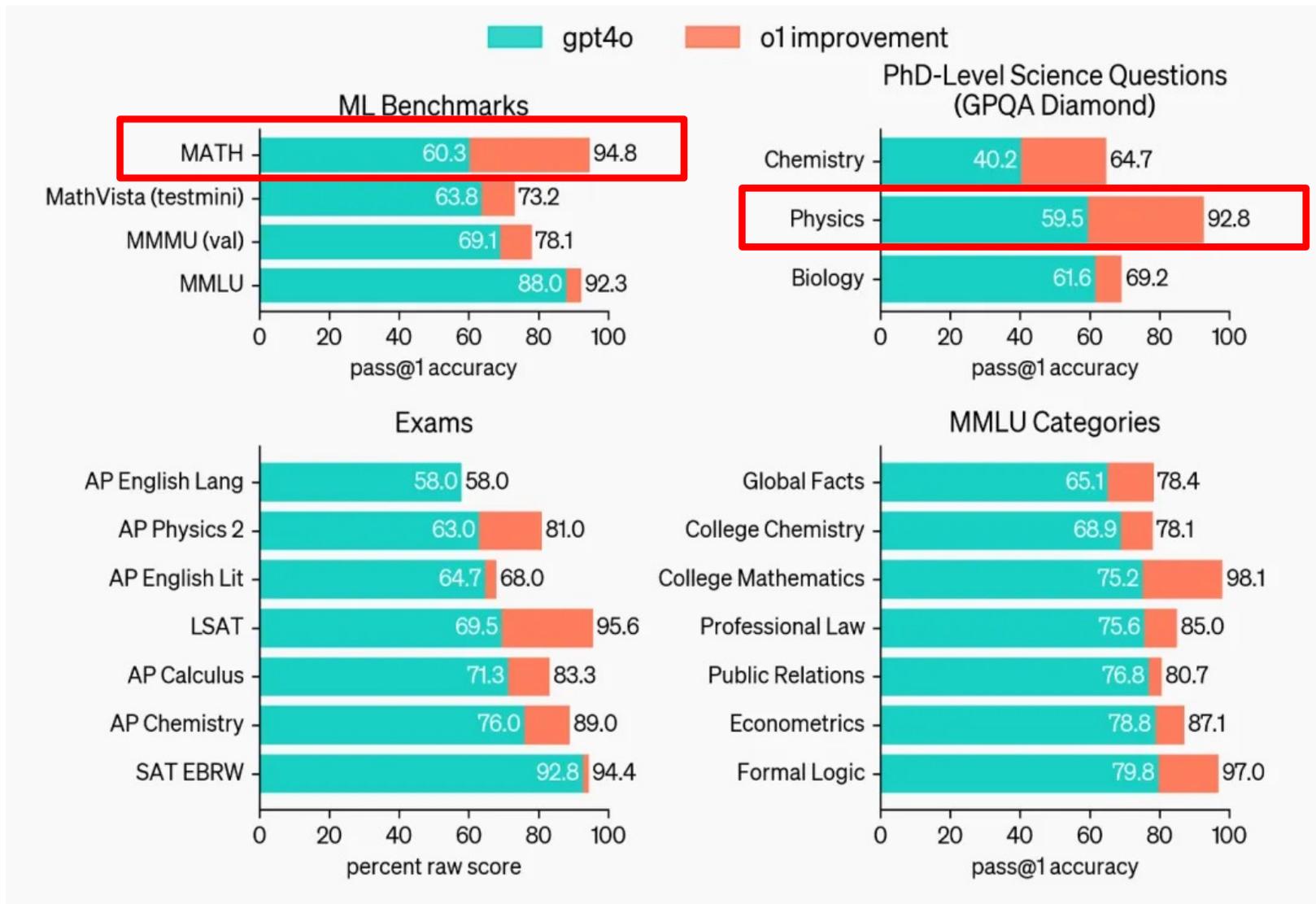
GPT4 accuracy 66.5%

Chinese Poem Task

语言模型无法进行逆向知识搜索，除非知识以逆序出现在预训练数据中



GPT-o1 preview 号称物理问答达到“博士：水平”



GPT-o1 preview 似乎也不能很好的解决初中数学问题

What is the period of $y = \cos \frac{x^2}{2x}$?



Thought for a couple of seconds

To determine the period of the function $y = \cos \left(\frac{x^2}{2x} \right)$, let's start by simplifying the expression inside the cosine function.

1. **Simplify the Argument of the Cosine Function:**

$$\frac{x^2}{2x} = \frac{x \cdot x}{2 \cdot x} = \frac{x}{2}$$

The x in the numerator and denominator cancel out.

2. **Rewrite the Function:**

$$y = \cos \left(\frac{x}{2} \right)$$

3. **Identify the Period of the Cosine Function:** The standard form of a cosine function is:

$$y = \cos(kx)$$

where the period T is given by:

$$T = \frac{2\pi}{|k|}$$

In our case, $k = \frac{1}{2}$.

4. **Calculate the Period:**

$$T = \frac{2\pi}{\left| \frac{1}{2} \right|} = 2\pi \times 2 = 4\pi$$

Answer: 4π

- 在 GSM8K 和 MATH 数据集问题的基础上，加入了一些“陷阱”，构造了 MathTrap 数据集。
- 改编后的题目很多时候没有定义明确的答案或者无解，只有能够同时理解原题和“陷阱”所涉及知识，才能找出其中的矛盾点，发现“陷阱”所在。
- MathTrap_Public上，GPT-o1-preview的回答准确率仅为**24.3%**
- MathTrap_Private 上，GPT-o1-preview API 的测试准确率为 38.0%，相比 GPT-4 API 的 **36.0%**，几乎没有提升



为什么？

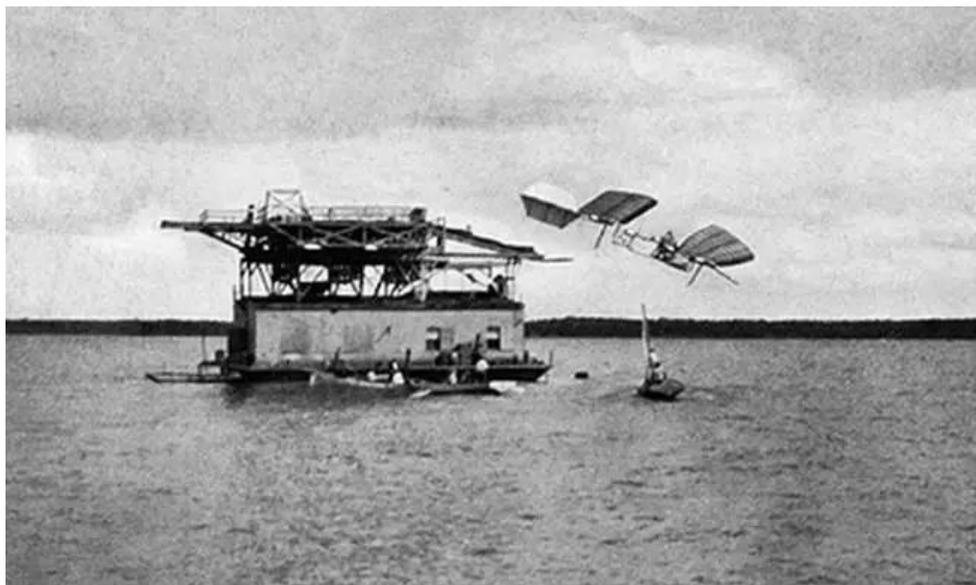
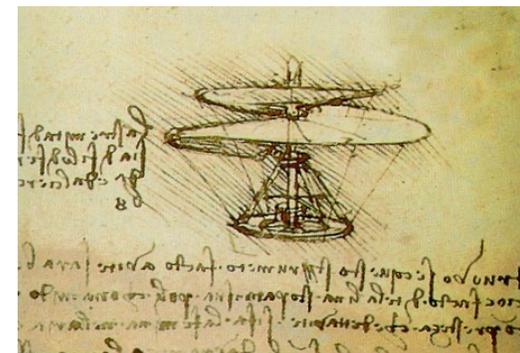
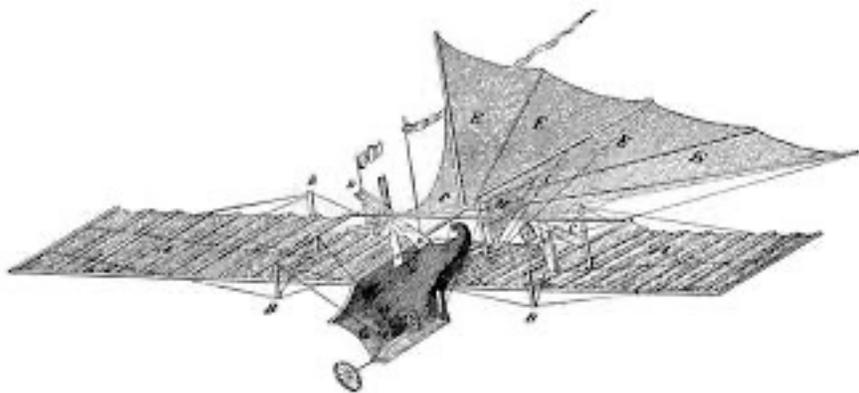
从飞机发展史看 AI 发展的核心问题

基础理论对于技术发展至关重要



1903年12月17日，世界上第一架载人动力飞机首飞

同时期各种脑洞大开的“飞行器”



莱特兄弟创造“飞行者一号”的历程

1890年代前期莱特兄弟就从新闻、杂志或者照片上看到了德国航空先驱奥托·李林达尔研制的**动力滑翔机**

1896年5月，塞缪尔·兰利成功试飞了一架**蒸汽动力的无人飞机模型**

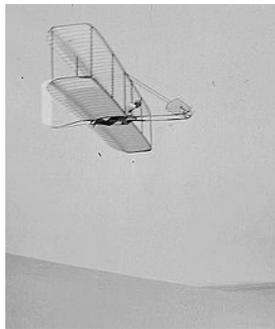
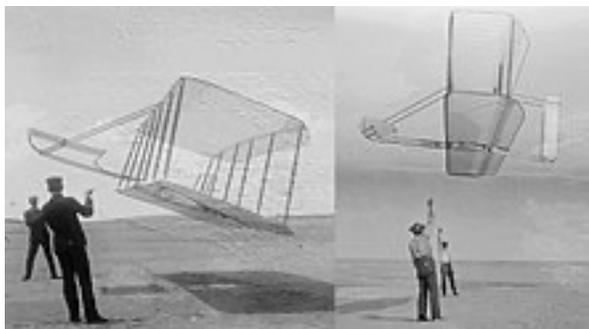
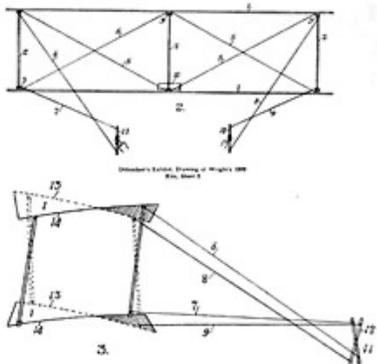
莱特兄弟认为困扰航空先驱们的**飞行难题有三点**：机翼、发动机以及如何控制飞机

1899年7月，威尔伯·莱特制作了一架长5英尺，形似双翼飞机的箱型风筝用来测试翘曲机翼技术

莱特兄弟兄弟建造了**风洞**，为200多种不同**翼弧的翼型**完成了旋转臂测试（一种测量升力和空气阻力的早期手段），并对其中的38种做了更详细的测试，同时也对广受认可的升力方程中的**斯密顿系数**提出了质疑，当时已经被提出了100多年。

基于**展弦比知识**和更精确的**斯密顿系数**，莱特兄弟设计了他们新的1902年款滑翔机

1903年莱特兄弟用云杉木来建造他们配备有动力装置的**飞行者一号**



猜测：OpenAI 发布 GPT-4 时已经掌握了大模型基础理论

GPT-4 Technical Report

OpenAI*

We registered predictions for GPT-4's performance on HumanEval before training completed, using only information available prior to training. All but the 15 hardest HumanEval problems were split into 6 difficulty buckets based on the performance of smaller models. The results on the 3rd easiest bucket are shown in [Figure 2](#), showing that the resulting predictions were very accurate for this subset of HumanEval problems where we can accurately estimate $\log(\text{pass_rate})$ for several smaller models. Predictions on the other five buckets performed almost as well, the main exception being GPT-4 underperforming our predictions on the easiest bucket.

回归基础理论进行分析

语言模型真正做的事情——文字接龙

语言模型的核心任务始终是生成一个“合理的延续”，即根据已有的文本，生成一个符合人类书写习惯的下一个合理内容。所谓“合理”，是指根据数十亿个网页、数字化书籍等人类撰写内容的统计规律，推测接下来可能出现的内容。

The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

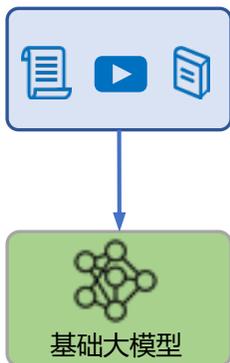


ChatGPT完成像写文章这样的任务时，它实际上只是一遍又一遍地询问：“在已有的文本的基础上，下一个词应该是什么？”——并且每次都会添加一个词。

ChatGPT实现过程

预训练阶段

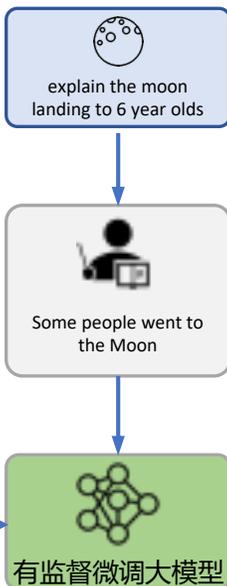
数千亿单词
图书、百科、
网页等



1000+GPU
月级别训练时间

有监督微调

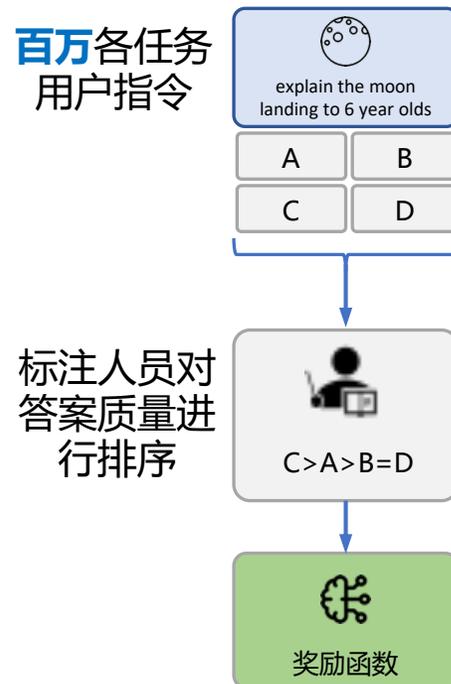
十万各任务
用户指令



1-100GPU
天级别训练时间

奖励函数

百万各任务
用户指令

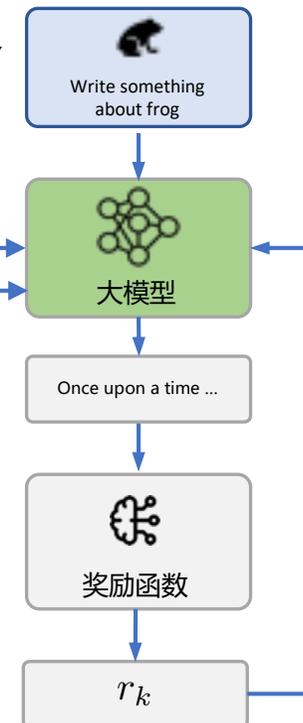


标注人员对
答案质量进行
排序

1-100GPU
天级别训练时间

强化学习

十万各任务
用户指令



1-100GPU
天级别训练时间

ChatGPT实现过程

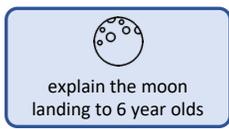
预训练阶段

数千亿单词
图书、百科、
网页等



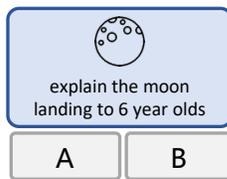
有监督微调

十万各任务
用户指令



奖励函数

百万各任务
用户指令



强化学习

十万各任务
用户指令



1. 大模型是否具备推理能力? 2. 大模型是否具备生成能力?

2. 预训练阶段能记住什么知识?

4. 什么才是高质量的SF 7. 强化学习适合于什么任务?

3. 如何评价预训练模型的好坏?

5. 有监督微调改变了基础模型吗? 8. 如何稳定高效的完成 RLHF 的训练?

6. 有监督微调的训练策略是什么? 9. 如何融合外部环境状态以及利用合成数据?

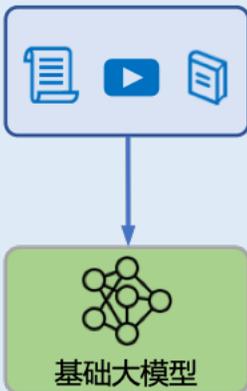


ChatGPT 实现过程

1. 大模型训练每个阶段分别完成什么功能?

预训练阶段

数千亿单词
图书、百科、
网页等

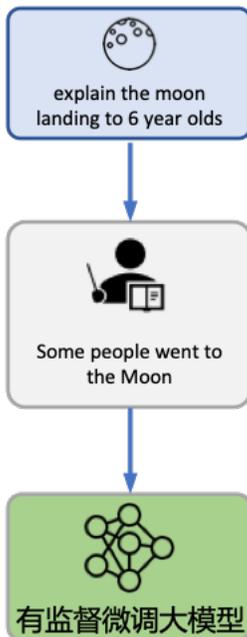


知识压缩和表示学习

有监督微调

十万各任务
用户指令

标注人员构
造目标答案

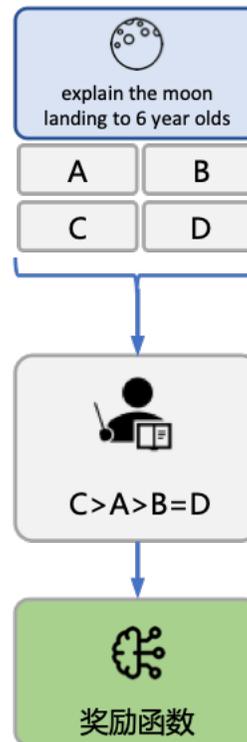


能力注入

奖励函数

百万各任务
用户指令

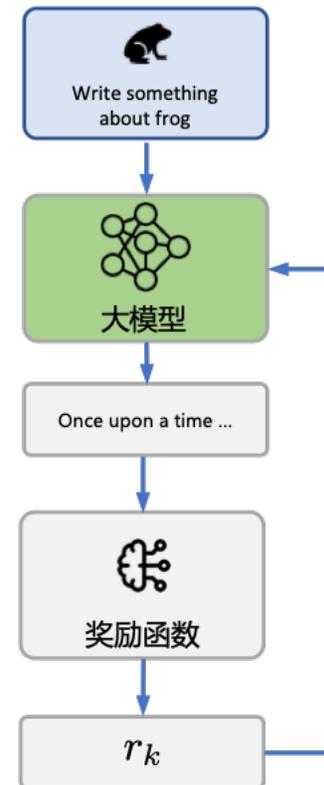
标注人员对
答案质量进
行排序



生成式任务能力提升

强化学习

十万各任务
用户指令

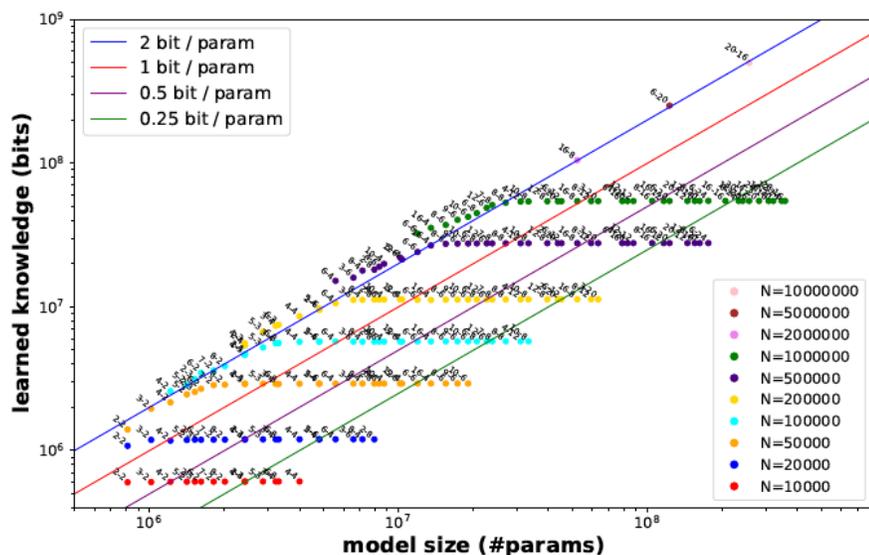


需要足够多的“曝光”才能完成记忆

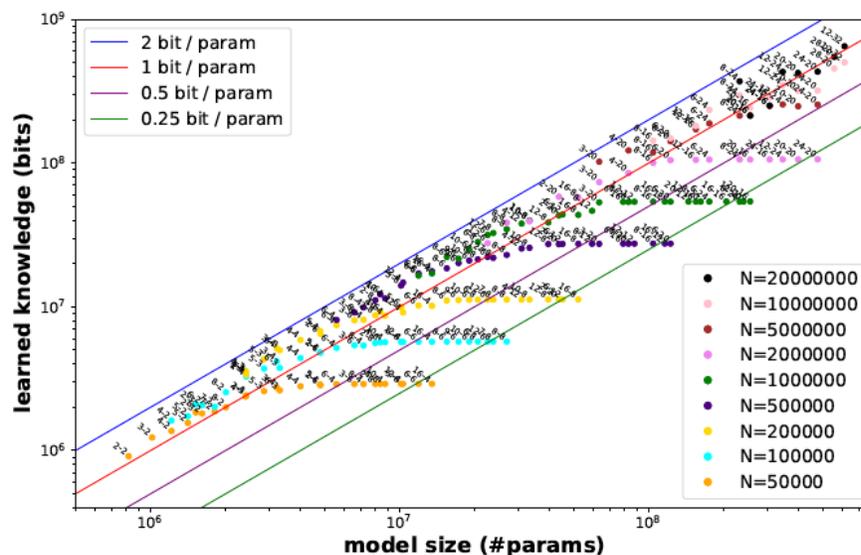
2. 预训练阶段能记住什么知识?

达到2bit/参数 对于每个知识要达到 **1000** 次曝光

如果只有**100**次曝光的话, 会减少到**1bit/参数**



(a) bioS(N) data — **1000 exposures** — peak $R(F) \geq 2$



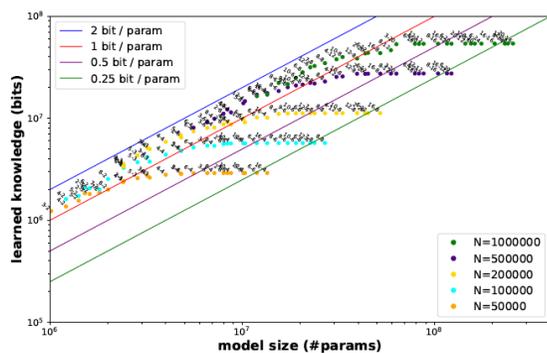
(b) bioS(N) data — **100 exposures** — peak $R(F) \geq 1$

图标上面数字是l,h参数选择

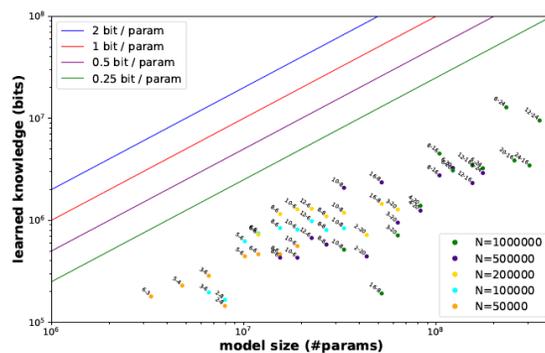


“垃圾数据”对知识获取有显著影响

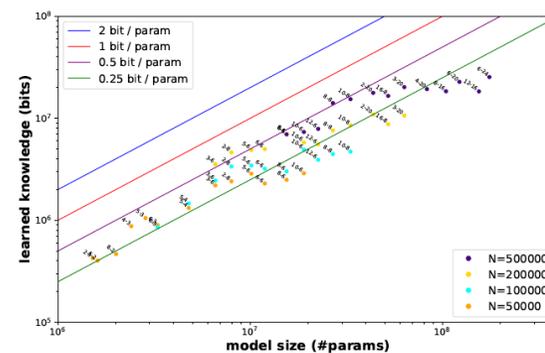
Wikipedia vs. Common Crawl



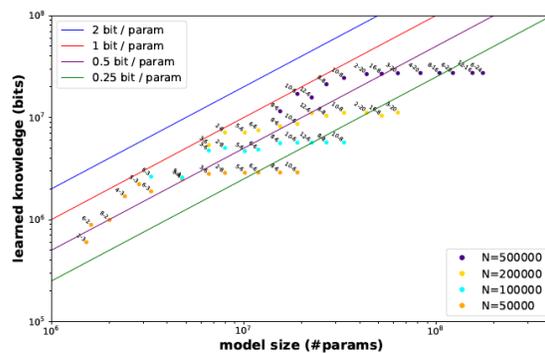
(a) no junk, 100 exposures



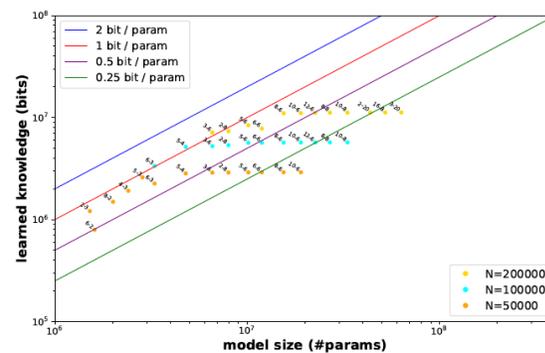
(b) 7/8 junk, 100 exposures



(c) 7/8 junk, 300 exposures



(d) 7/8 junk, 600 exposures



(e) 7/8 junk, 1000 exposures

直接通过训练语料的统计就可以计算得到知识记忆概率

Physics of Language Models: Part 3.3 的研究并未对知识的特异性进行分析

$$\begin{aligned} P(w_1 w_2 \dots w_m) &= P(w_1) P(w_2 | w_1) P(w_3 | w_1 w_2) \dots P(w_m | w_1 w_2 \dots w_{m-1}) \\ &= \prod_{i=1}^m P(w_i | w_1 w_2 \dots w_{i-1}) \end{aligned}$$

联想 总裁 杨元庆

具有别的含义，且出现次数频繁，导致该知识记忆难度大

华为 总裁 任正非

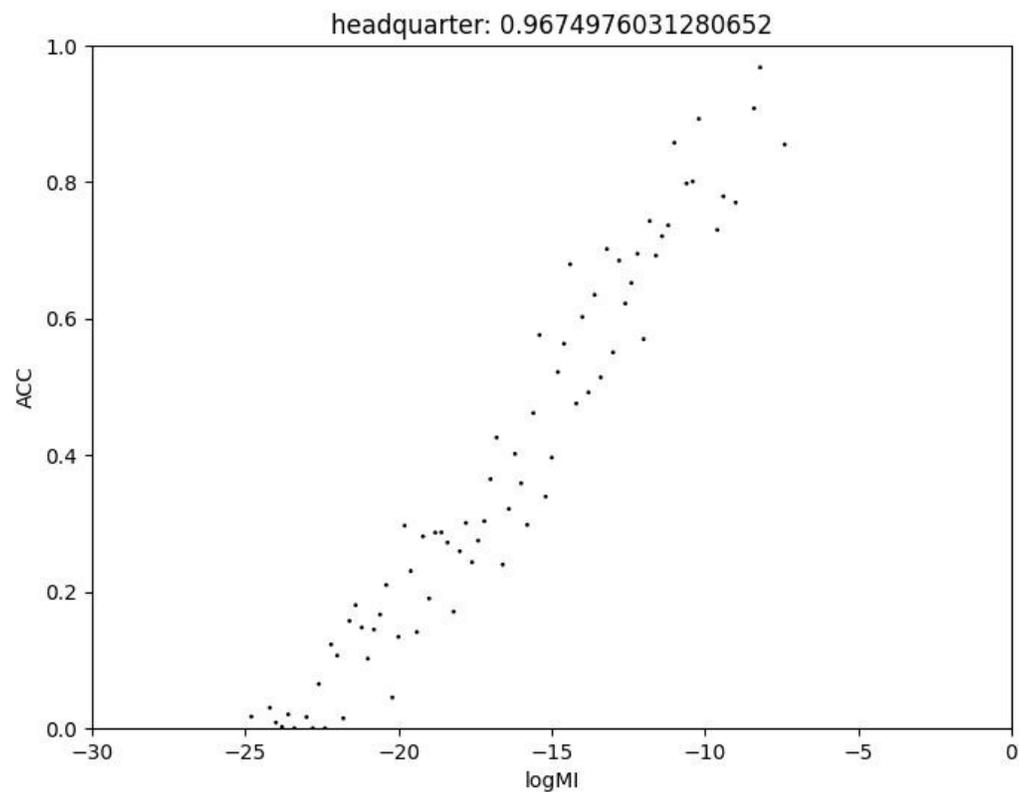
直接通过训练语料的统计就可以计算得到知识记忆概率

knowledge	map(subject, relation)	map(object)
Painblanc subject	<i>Painblanc is located in</i>	France
located in relation	<i>The country where Painblanc is found is</i>	France
France object
	<i>Painblanc is a part of</i>	France

Figure 1 | An example for the multi-template complementation mechanism.

基于模版变换的知识记忆评估方案

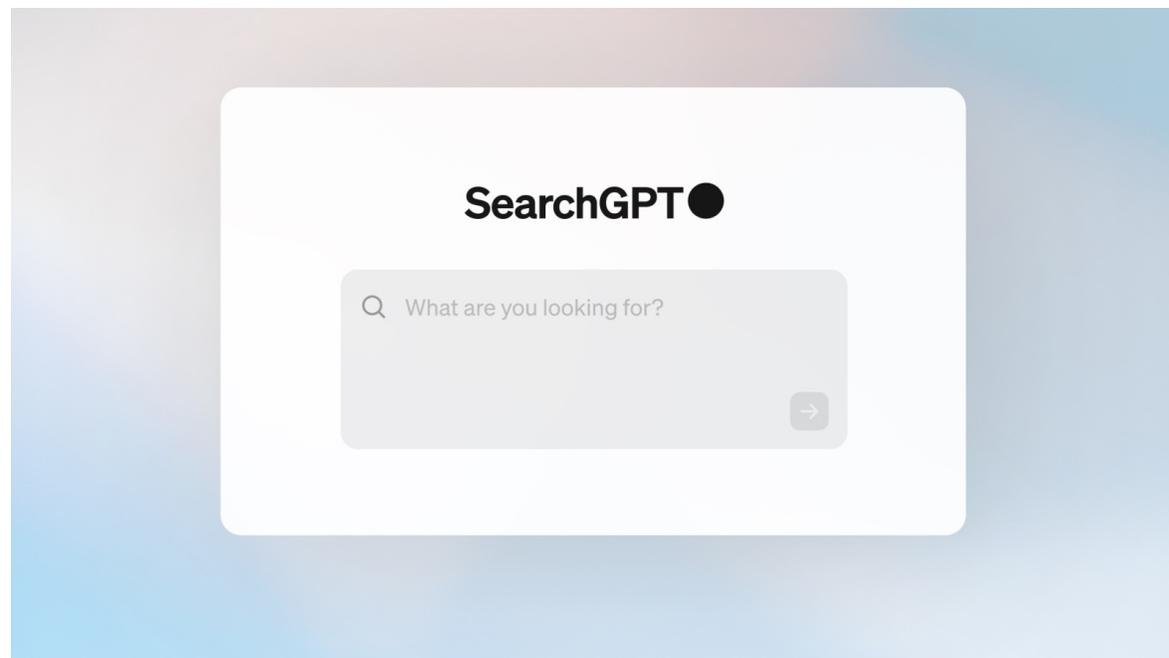
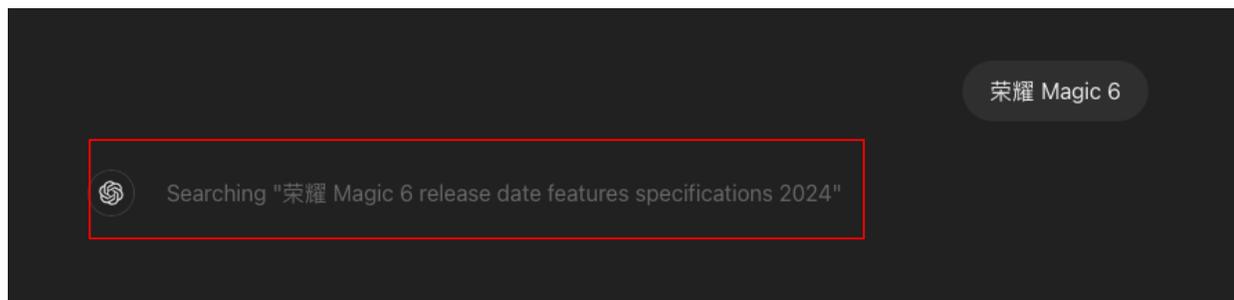
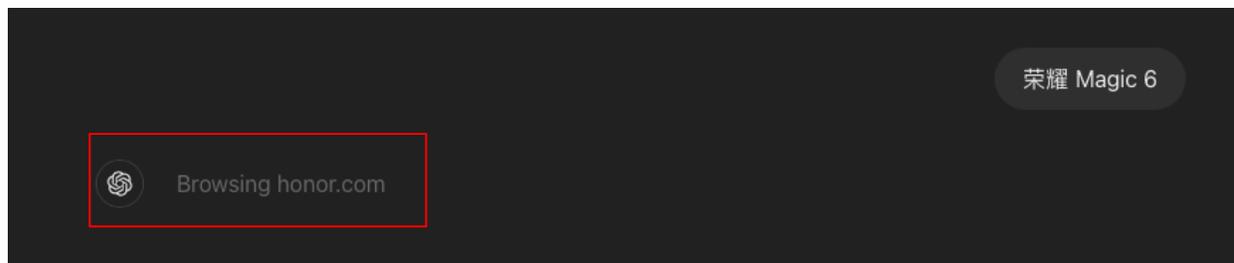
$$p(x)p(y)/p(x,y)$$



预训练阶段思考

预训练阶段主要完成 **知识记忆+表示学习**

- 开源大模型已经可以很好的进行句子/段落/篇章表示
- OpenAI **基本放弃了**通过模型知识记忆直接进行问题回答



预训练阶段结论

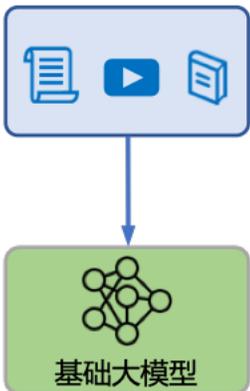
1. 同一个知识点必须用不同的语言表达方式**大量重复**，模型才能学习到
2. 大规模数据的 **PPL值** 可以有效衡量模型的知识记忆能力
3. 具有**高知识密度**的高质量的训练数据至关重要
4. 经过足够的训练模型可以达到**2bit/参数**的知识存储能力



大模型的能力是如何获取的

预训练阶段

数千亿单词
图书、百科、
网页等

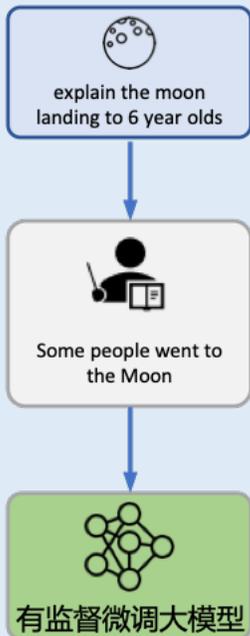


知识压缩和表示学习

有监督微调

十万各任务
用户指令

标注人员构
造目标答案

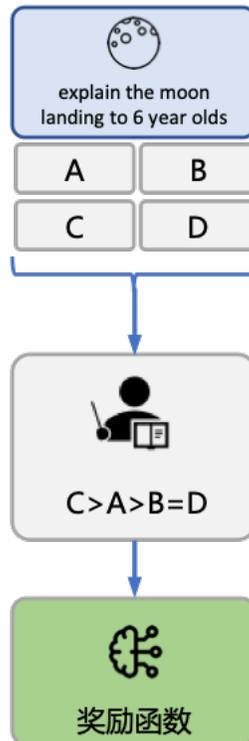


能力注入

奖励函数

百万各任务
用户指令

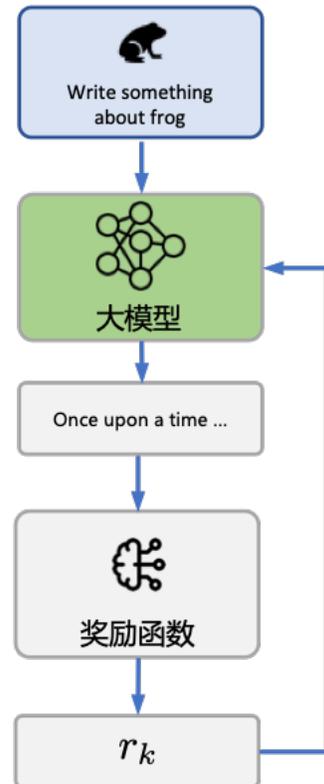
标注人员对
答案质量进
行排序



生成式任务能力提升

强化学习

十万各任务
用户指令



指令微调—利用少量数据训练模型具备任务能力

知识问答任务：

输入：复旦大学有几个校区？

理想输出：复旦大学现有4 个校区，分别是邯郸校区、新江湾校区、枫林校区和张江校区。其中邯郸校区是复旦大学的主校区，邯郸校区与新江湾校区都位于杨浦区，枫林校区位于徐汇区，张江校区位于浦东新区。

作文：

输入：写一个小故事，关于一只熊去海滩，与海豹交朋友，然后回家

理想输出：从森林里走出一只可爱的小熊，它叫贝利。贝利一直对大海充满好奇，于是他决定去海滩看看。他跳跃着穿过河流和草地，终于来到了一片美丽的海滩。贝利兴奋地跑着，感受着海风的清凉。突然，他注意到海滩上有一只憨态可掬的海豹，名叫萨米。萨米正在玩耍，好像是在找贝利这样的新朋友。贝利好奇地走过去，和萨米打了个招呼。萨米非常友好地回应道：“嗨，你是新来的吗？我是萨米，很高兴认识你！”贝利开心地笑了起来，他们迅速成为了好朋友。...

针对特定任务如何构造SFT 训练数据

大模型完成知识问答任务

4. 什么才是高质量 SFT训练数据?

大语言模型（LLMs）通过在海量数据集上的预训练，能够编码丰富的世界知识，并且可以通过微调将这些知识应用于问答（QA）任务。如何有效地微调LLMs以提升QA任务表现的策略仍需研究。

Q1: 在SFT阶段需要多少数据才能让LLMs学会QA任务?

Q2: 不同的SFT数据集如何影响LLMs在QA任务上的表现?

Q3: 不同LLMs在SFT阶段对数据的需求有何差异?



大模型完成知识问答任务

- 使用了ENTITYQUESTIONS, 包含24个不同话题知识的问答数据集
- 12个与地点相关的原始训练集作为训练数据 D_{train} , 将它们对应的测试集作为测试集 D_{test} , 并将剩余12个话题的测试集作为领域外测试集 $D_{test-ood}$
- 给定一个预训练的LLM, 为了详细分析其在SFT之后执行QA任务的表现, 应用如上定义的多模板补全机制, 根据其知识记忆水平将训练和测试集均进行了5个级别的划分

$$\begin{cases} D_{test-0} = \{k \in D_{test} | R_k^M = 0\}, \\ D_{test-1} = \{k \in D_{test} | R_k^M \in (0, 0.25]\}, \\ D_{test-2} = \{k \in D_{test} | R_k^M \in (0.25, 0.5]\}, \\ D_{test-3} = \{k \in D_{test} | R_k^M \in (0.5, 0.75]\}, \\ D_{test-4} = \{k \in D_{test} | R_k^M \in (0.75, 1]\} \end{cases}$$



大模型完成知识问答任务

1. SFT的数据量需求

- 使用不同记忆水平的训练数据 $D_{train-i}$ 的来分析每个 LLM
- 将训练数据划分为六个不同的数据量级别，从60个样本到完整数据集不等，并通过从12个话题中均匀抽样来构建训练集
- 实验结果表明，经过SFT后，仅需**60**个训练样本就足以使LLMs高效执行QA任务，并展现出强大的泛化能力
- 无论基础模型或记忆水平如何，LLMs在使用较少训练样本时的表现优于使用960个或全部样本。大多数模型在 $N_{train} = 60$ 时达到或接近最佳表现

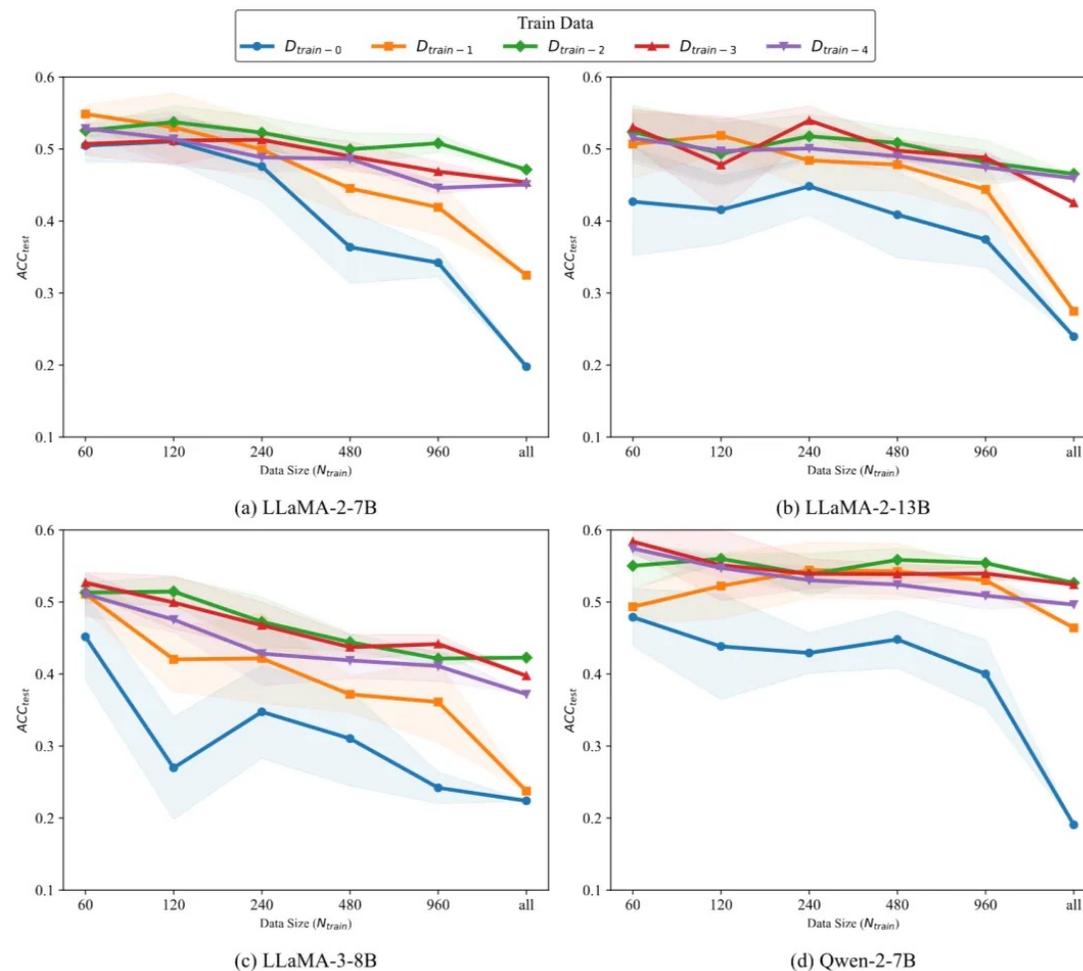


Figure 2 | Performance (in-domain) of LLMs trained using different amounts of data. Each line in the plot represents training with data from a specific memory level.

大模型完成知识问答任务

2. 使用不同记忆水平的数据进行微调的影响

- 无论使用何种数据进行微调，LLMs始终对预训练期间记忆较好的知识提供更准确的答案
- 在特定记忆水平的数据上进行训练能够提升LLMs在该水平知识上的表现
- 总体而言，更有效的策略是使用高记忆水平的数据进行SFT

Table 2 | Performance (in-domain) of LLMs trained using data with different memory levels on test sets with different memory levels. The best performance of each test set is **bolded**.

(a) LLaMA-2-7B ($N_{train} = 60$)

Train Data	ACC _{test-0}	ACC _{test-1}	ACC _{test-2}	ACC _{test-3}	ACC _{test-4}	ACC _{test}
$D_{train-0}$	4.80 ± 1.11	34.00 ± 2.93	62.49 ± 4.83	73.36 ± 4.13	77.64 ± 6.02	50.46 ± 2.28
$D_{train-1}$	4.12 ± 1.61	38.19 ± 4.54	69.49 ± 2.23	77.11 ± 3.43	85.32 ± 1.62	54.85 ± 1.26
$D_{train-2}$	0.69 ± 0.35	24.57 ± 1.70	70.57 ± 2.23	82.39 ± 1.65	84.56 ± 0.64	52.56 ± 0.81
$D_{train-3}$	0.37 ± 0.41	17.79 ± 2.33	67.45 ± 2.34	85.43 ± 2.13	82.45 ± 2.84	50.70 ± 1.57
$D_{train-4}$	0.40 ± 0.28	19.78 ± 1.67	69.77 ± 1.24	87.36 ± 1.45	86.84 ± 0.91	52.83 ± 0.99

(b) LLaMA-2-13B ($N_{train} = 60$)

Train Data	ACC _{test-0}	ACC _{test-1}	ACC _{test-2}	ACC _{test-3}	ACC _{test-4}	ACC _{test}
$D_{train-0}$	7.62 ± 1.69	23.66 ± 4.29	53.17 ± 9.36	61.96 ± 11.69	67.00 ± 14.25	42.68 ± 7.45
$D_{train-1}$	7.49 ± 0.54	35.95 ± 5.15	64.22 ± 5.00	65.98 ± 5.28	79.84 ± 9.43	50.69 ± 4.69
$D_{train-2}$	0.62 ± 0.50	21.36 ± 2.93	72.80 ± 5.72	80.65 ± 5.45	86.47 ± 5.85	52.38 ± 3.70
$D_{train-3}$	0.45 ± 0.21	17.97 ± 1.81	71.16 ± 3.63	86.60 ± 3.40	88.87 ± 4.79	53.01 ± 2.54
$D_{train-4}$	0.24 ± 0.22	16.66 ± 3.14	68.46 ± 4.83	83.44 ± 4.92	88.69 ± 4.82	51.50 ± 3.40

(c) LLaMA-3-8B ($N_{train} = 60$)

Train Data	ACC _{test-0}	ACC _{test-1}	ACC _{test-2}	ACC _{test-3}	ACC _{test-4}	ACC _{test}
$D_{train-0}$	2.72 ± 1.44	26.05 ± 8.77	53.82 ± 10.32	62.90 ± 8.27	80.36 ± 9.57	45.17 ± 5.96
$D_{train-1}$	2.28 ± 1.23	33.83 ± 7.50	66.35 ± 3.69	71.97 ± 4.08	81.02 ± 3.91	51.09 ± 2.33
$D_{train-2}$	0.19 ± 0.08	21.57 ± 1.71	71.89 ± 1.49	80.23 ± 1.13	82.65 ± 5.39	51.31 ± 1.41
$D_{train-3}$	0.15 ± 0.03	19.03 ± 0.91	68.09 ± 3.64	85.72 ± 2.69	90.30 ± 1.17	52.66 ± 1.45
$D_{train-4}$	0.17 ± 0.07	16.39 ± 1.55	63.76 ± 6.47	83.58 ± 4.32	91.54 ± 3.12	51.09 ± 3.00

(d) Qwen-2-7B ($N_{train} = 60$)

Train Data	ACC _{test-0}	ACC _{test-1}	ACC _{test-2}	ACC _{test-3}	ACC _{test-4}	ACC _{test}
$D_{train-0}$	7.85 ± 0.67	34.88 ± 1.29	57.84 ± 4.28	66.17 ± 6.57	72.70 ± 8.35	47.89 ± 3.99
$D_{train-1}$	6.49 ± 0.61	39.40 ± 2.10	60.68 ± 4.01	69.50 ± 5.15	70.57 ± 2.88	49.33 ± 2.23
$D_{train-2}$	0.36 ± 0.09	26.15 ± 1.76	80.16 ± 3.59	83.14 ± 4.35	85.30 ± 7.44	55.02 ± 3.12
$D_{train-3}$	0.48 ± 0.09	21.58 ± 0.54	82.01 ± 1.89	92.34 ± 1.00	95.48 ± 4.82	58.38 ± 1.45
$D_{train-4}$	0.21 ± 0.04	19.55 ± 1.02	80.48 ± 1.24	90.84 ± 1.98	96.00 ± 0.99	57.42 ± 0.92



大模型完成知识问答任务

3. 不同LLMs的数据需求差异

- 不同 LLM 的记忆差距很大
- 使用相同的训练数据在不同 LLM 上训练得到的结果有很大差异

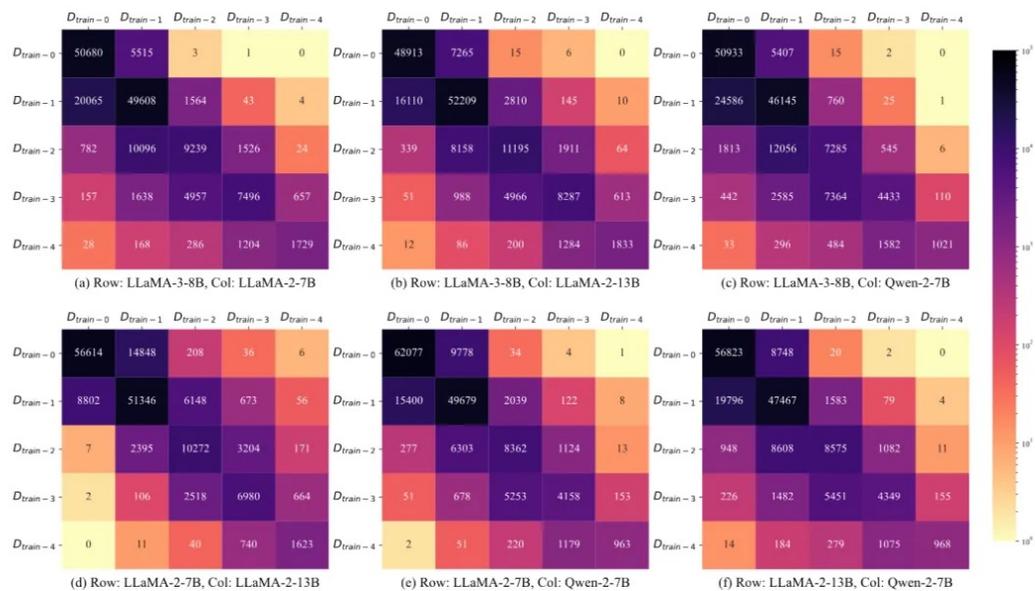


Figure 4 | Heat maps showing differences in the distribution of memory levels for different LLMs on the training data D_{train} .

Table 4 | Performance comparison of fine-tuning different LLMs using the same 60 data samples D_{train}^* . ACC_{test}^\dagger represents the best performance for each LLM trained with 60 samples.

Models	ACC_{test-0}	ACC_{test-1}	ACC_{test-2}	ACC_{test-3}	ACC_{test-4}	ACC_{test}	ACC_{test}^\dagger
LLaMA-2-7B	1.05	20.07	52.90	56.59	72.15	40.55 \downarrow 14.30	54.85
LLaMA-3-8B	0.20	19.91	67.90	78.91	89.22	51.23 \downarrow 1.43	52.66
Qwen-2-7B	3.09	25.89	57.59	66.07	72.17	44.96 \downarrow 13.42	58.38

大模型完成知识问答任务

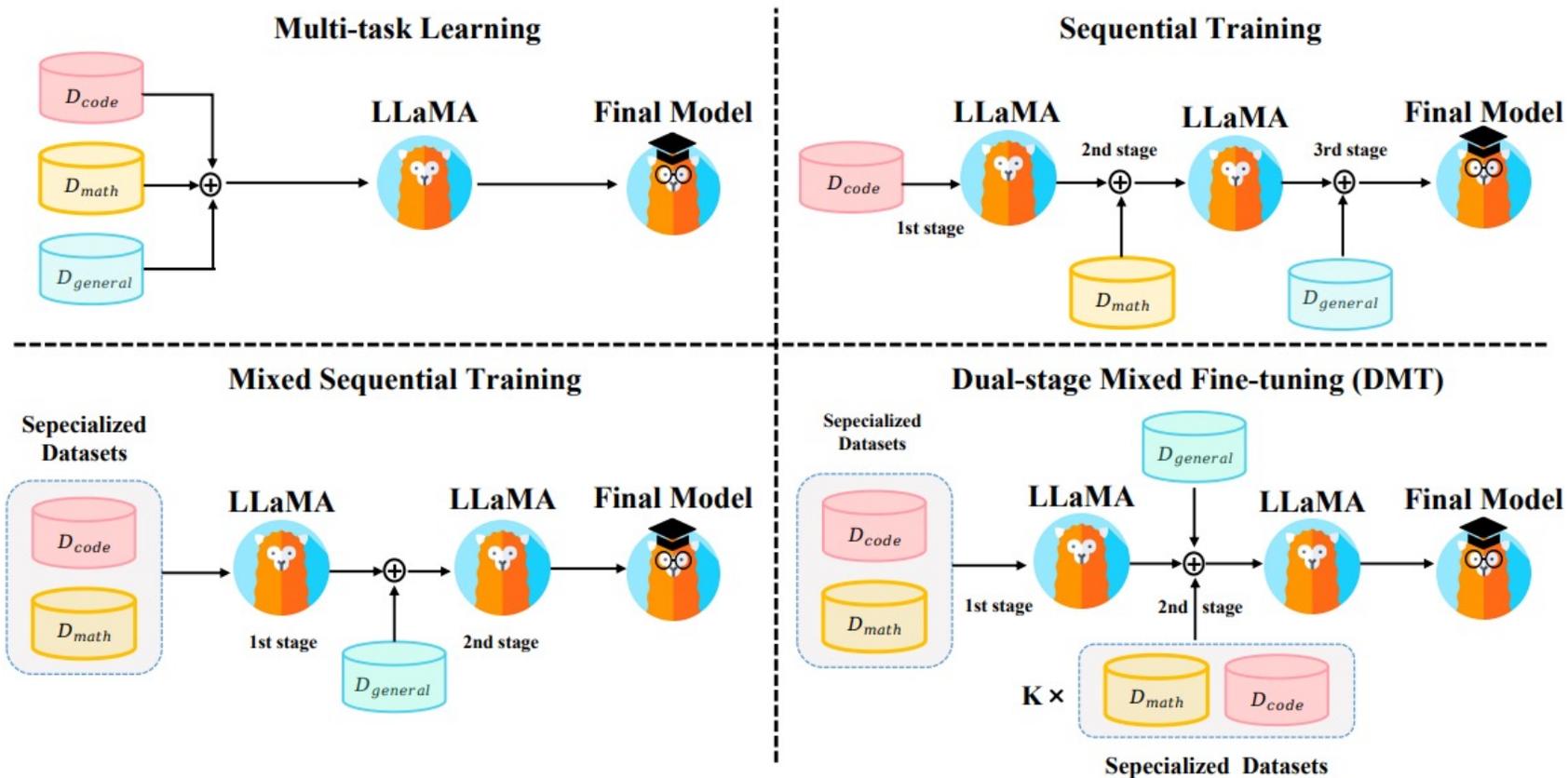
- 在SFT阶段仅需**60**条数据就能激活预训练时编码的知识，使LLMs能够有效执行QA任务
- 使用不同知识记忆层次的数据进行SFT，对LLMs的表现有显著并且有规律的影响
- 对于QA任务来说最优的SFT数据集**因具体模型而异**



有监督微调阶段的训练方式

有监督微调的四种方式

6. 有监督微调的训练策略是什么？



有监督微调的四种方式

1. 数学推理、编程和一般能力与SFT数据量的关系如何？
2. 当将这三种能力结合在SFT中时，是否会出现性能冲突？
3. 导致性能冲突的关键因素是什么？
4. 不同SFT策略对复合数据的影响是什么？



实验设置

SFT 数据集 $\{D_1, D_2, \dots, D_k\}$, 每个数据集 D_i 表示一个任务

$D_i = \{q_{i,j}, r_{i,j}\}_j$ 包含输入和回答

训练数据集:

数学: GSM8K RFT

编程: Code Alpaca

通用: ShareGPT

测试数据集:

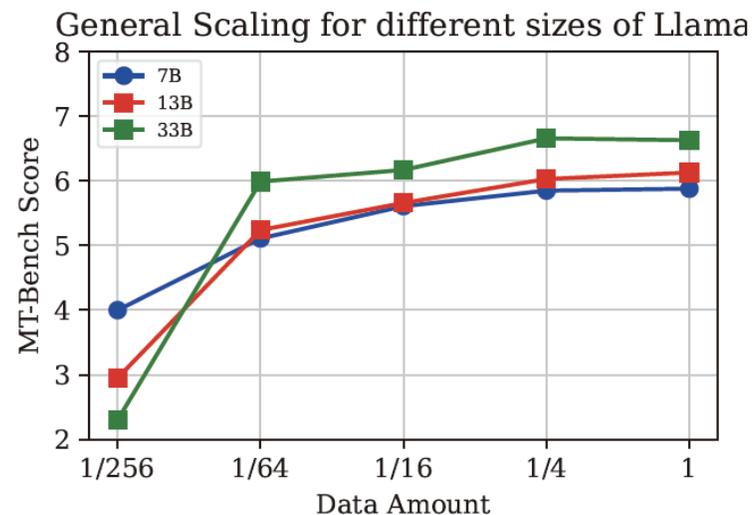
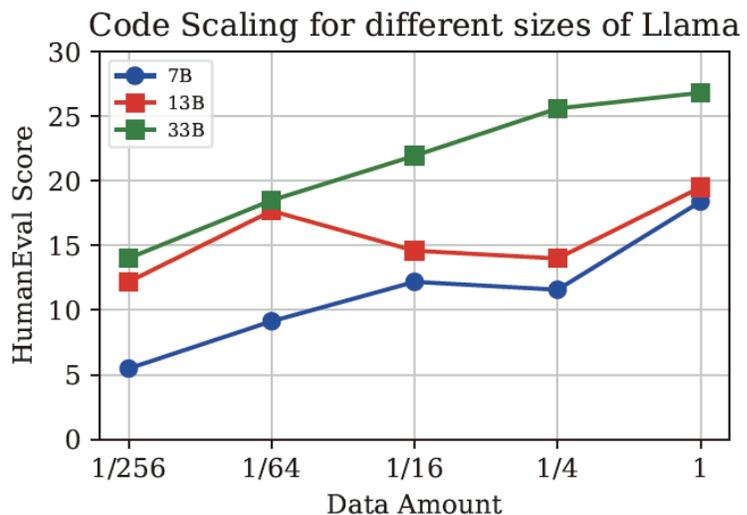
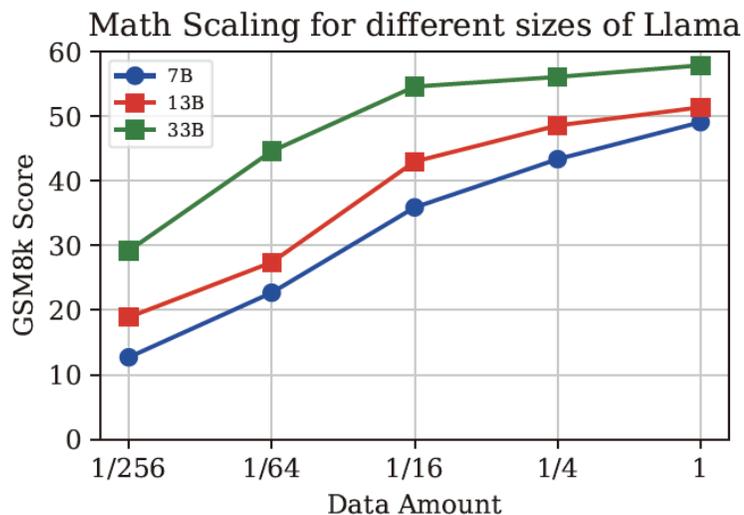
数学: GSM8K Test Set

编程: Humaneval

通用: MT-Bench



实验分析：RQ1 单个任务不同数据量

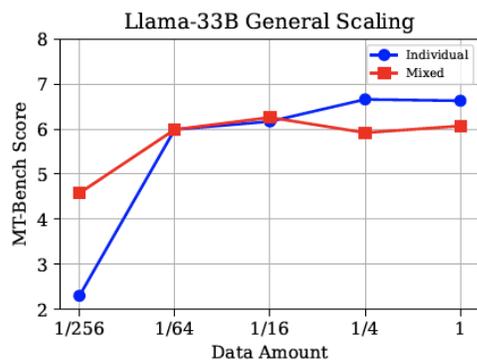
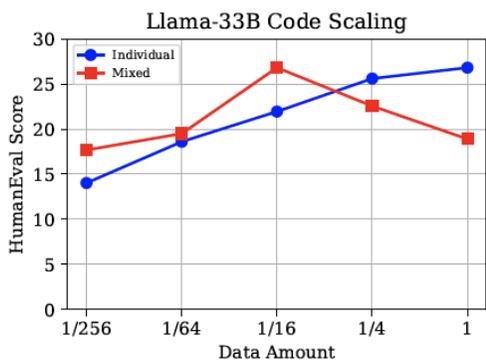
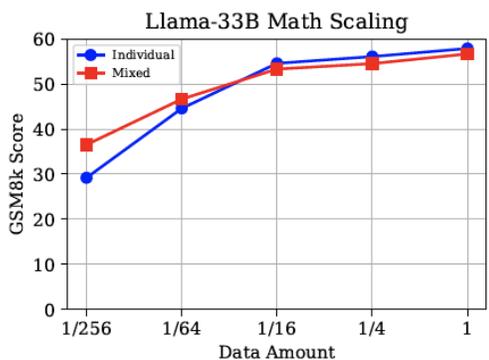
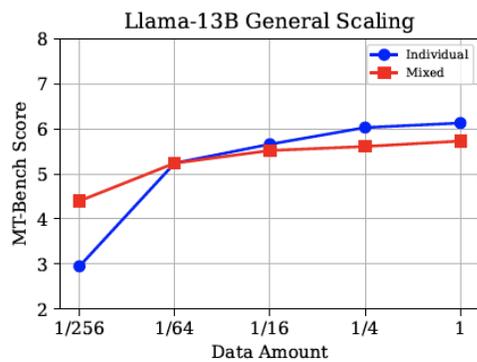
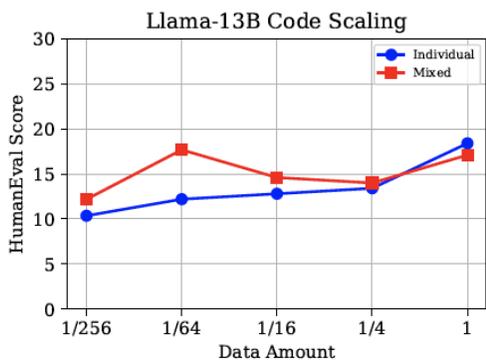
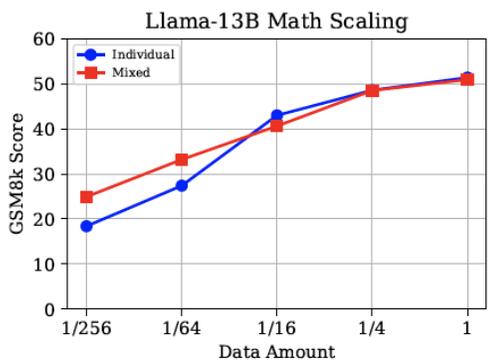
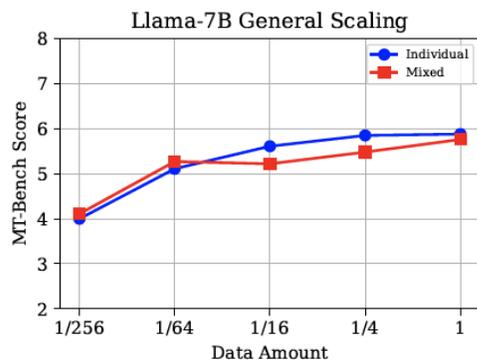
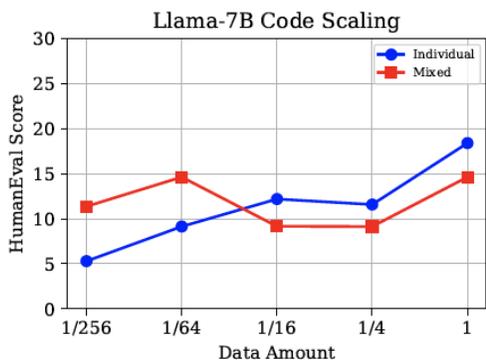
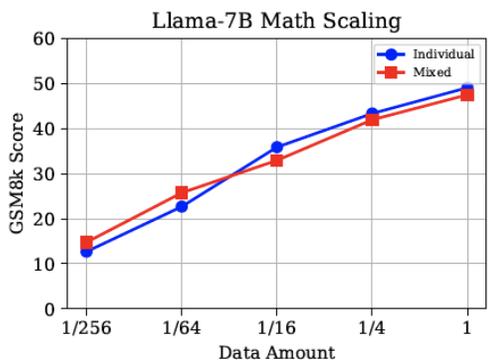


单个任务使用不同数据量进行训练

较大模型在相同的情况下表现出更好的性能



实验分析：RQ2 单个任务 vs. 混合任务

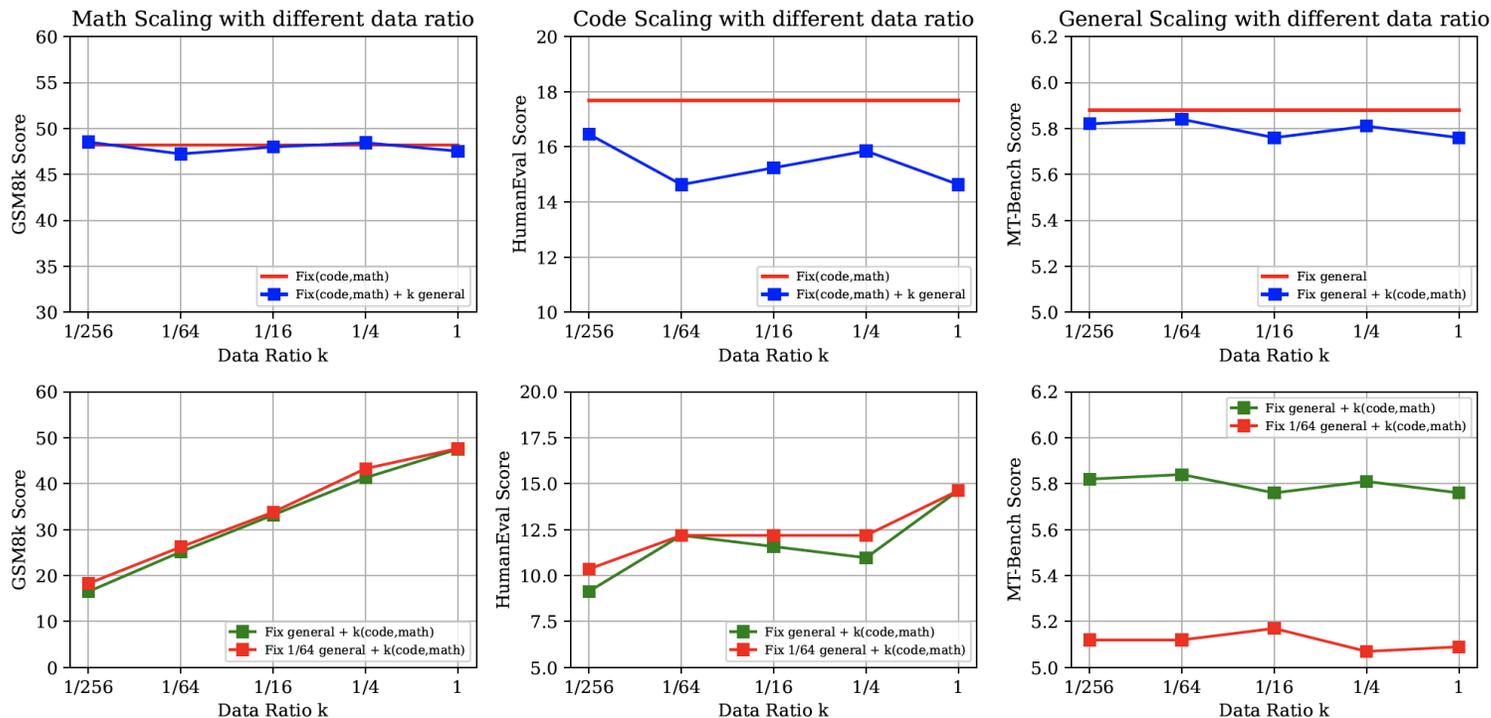


训练数据少时，能力得以提高；
训练数据多时，混合数据则使得能力则减弱，与单个任务训练相比，这种现象更为明显。

随着模型大小的增加，在低资源环境下的表现也会随之提高，特别是在数学和一般能力方面。



实验分析：RQ3 任务混合比例影响



不同的SFT能力在任务格式和数据分布上存在显著差异时，**数据比例的影响是微不足道的。**

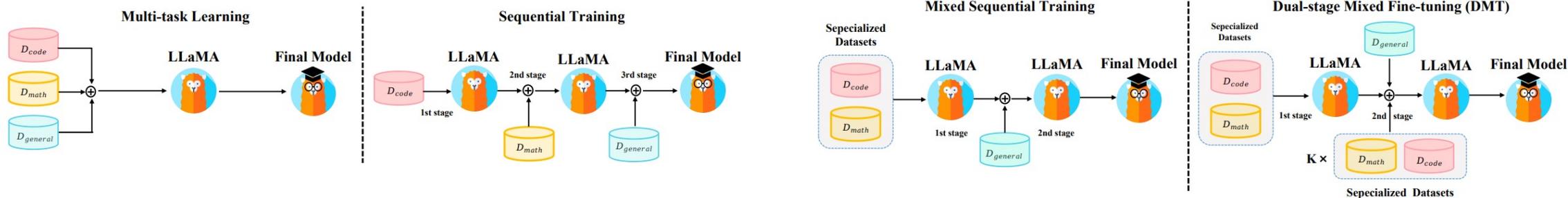
然而，当存在**一定程度的相似性时**，数据比例可能导致显著的性能波动。

$$k = \frac{\text{specialized use data amount}}{\text{general use data amount}} = \frac{\text{specialized all data amount}}{\text{general all data amount}}$$



实验分析：RQ4 不同训练方法结果

Methods	LLaMA -7B			LLaMA -13B			LLaMA -33B		
	GSM8K	HumanEval	MT-Bench	GSM8K	HumanEval	MT-Bench	GSM8K	HumanEval	MT-Bench
<i>Individual domain</i>									
General only	11.10	10.42	5.88	14.02	16.40	6.13	26.06	24.30	6.63
Math only	49.10	6.71	2.53	51.40	12.8	2.54	57.91	15.5	3.18
Code only	4.51	18.40	4.30	5.15	17.1	3.53	6.06	26.82	4.18
<i>Different Training Strategies</i>									
Multi-task learning	47.53	14.63	5.76	50.94	<u>19.50</u>	5.73	56.69	18.9	6.07
Sequential Training	31.39	<u>15.85</u>	5.72	39.12	20.12	<u>5.93</u>	47.27	<u>24.80</u>	6.73
Mixed Sequential Training	32.60	15.24	<u>6.02</u>	40.48	18.30	<u>5.93</u>	44.24	24.4	6.43
DMT(k=1/256)	<u>41.92</u>	17.68	6.08	<u>46.47</u>	<u>19.50</u>	6.03	<u>56.36</u>	25.00	6.73



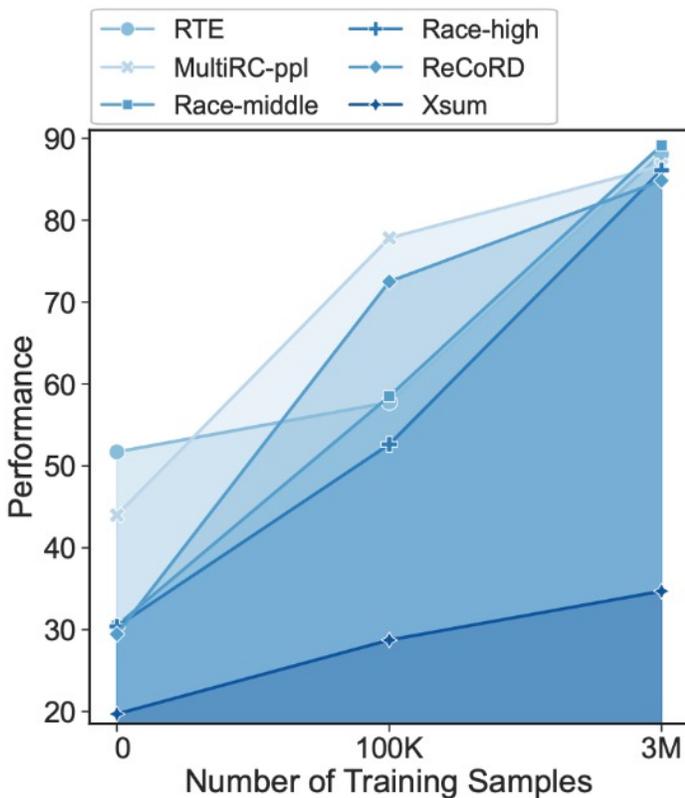
有监督微调阶段结论

1. 参数量大的模型在使用相同数量的数据进行训练时表现出优越性能
2. 混合数据来源在低资源场景中提高了性能，但在高资源场景中性能下降
3. 数据量直接影响性能，而数据比例的影响在实验设置中不显著
4. DMT策略有效地缓解了性能冲突

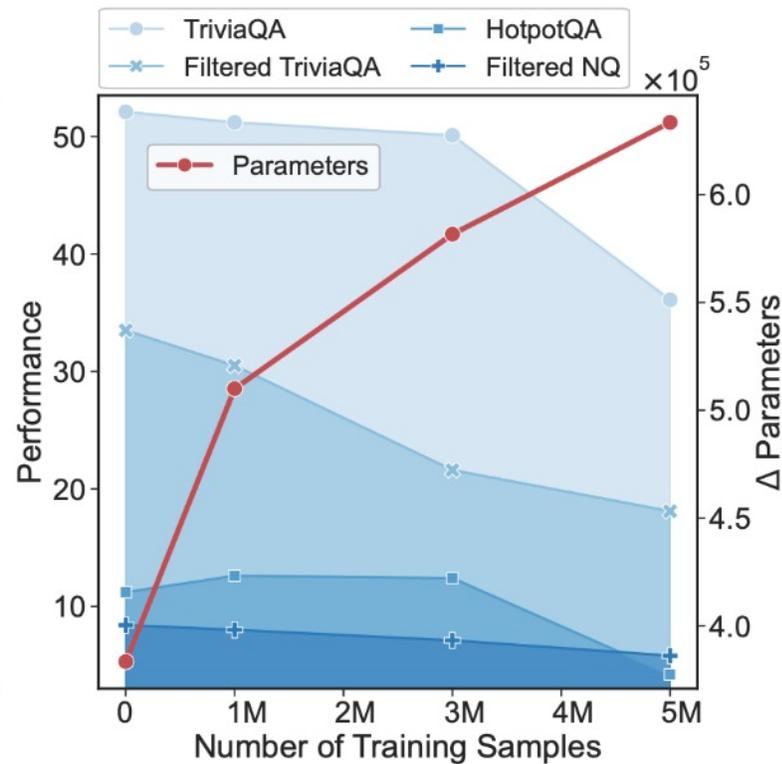


知识回答和其他任务冲突 如何解决？

大规模微调破坏世界知识



但当下游任务增多或者需要强化特定任务的性能时，增加SFT训练数据是有必要的。如上图的左侧部分，当SFT数据从100K提升到3M时，大部分任务的性能显著增强。



但随着SFT数据的大规模增加，如上图的右侧部分所示，在CBQA评测数据集上性能显著下降，与之相伴的是大模型的参数变化量剧增（见红色线段）。



大规模微调破坏世界知识

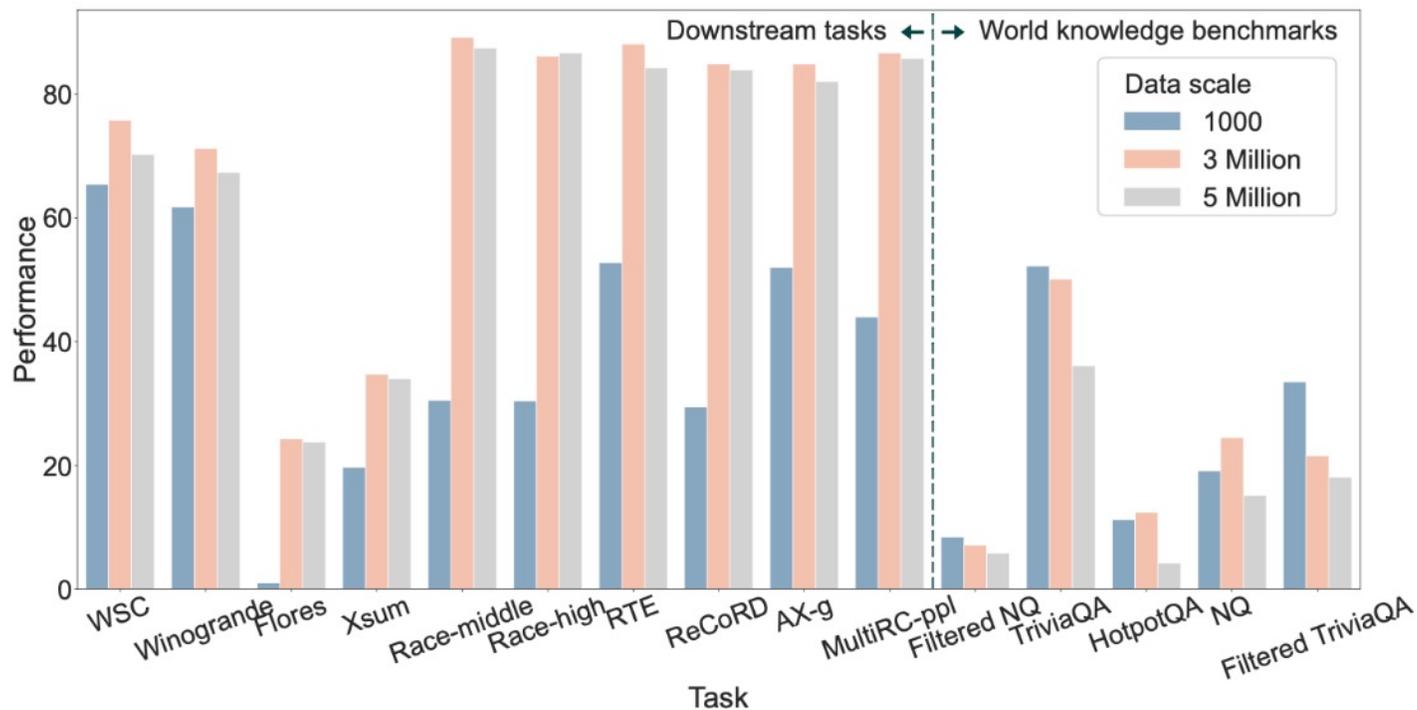


Figure 2: Performance on the various tasks after expanding the amount of fine-tuning data. For most of the downstream tasks (e.g., NLI and summarization), with the expansion of training data, performance on these tasks remains stable after improvement. Whereas, for the world knowledge benchmark, a significant **decline** can be witnessed after a large amount of instruction data.

摘要、NLI、机器翻译等任务，随着SFT训练数据的增加，性能显著提升；但是右侧的CBQA任务，却大幅下跌



CBQA的能力来源于预训练阶段

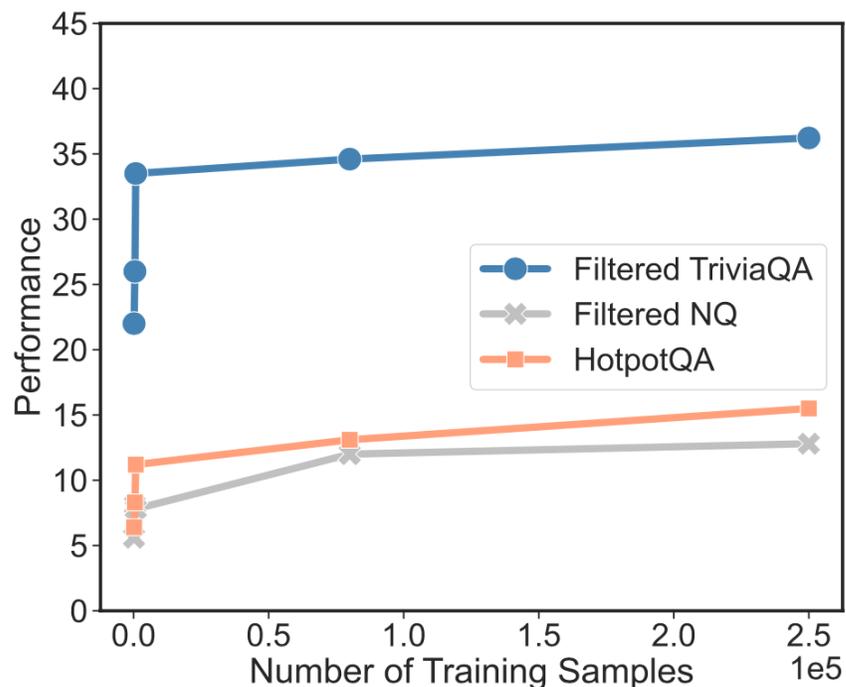


Figure 3: Performance on world knowledge benchmarks after training on CBQA solely. Its performance rises greatly after training with very few samples and remains relatively stable thereafter.

在训练一开始大约1000样本的时候，性能已经快速提升到了很高的点，后续再增加更多的训练样本其实提升很有限。说明少量样本微调就帮助大模型完成了人类指令的对齐，大模型完成CBQA指标评测的能力**主要依靠的是内在的世界知识**，而不是微调过程中训练样本灌输的。

LoRA+MoE

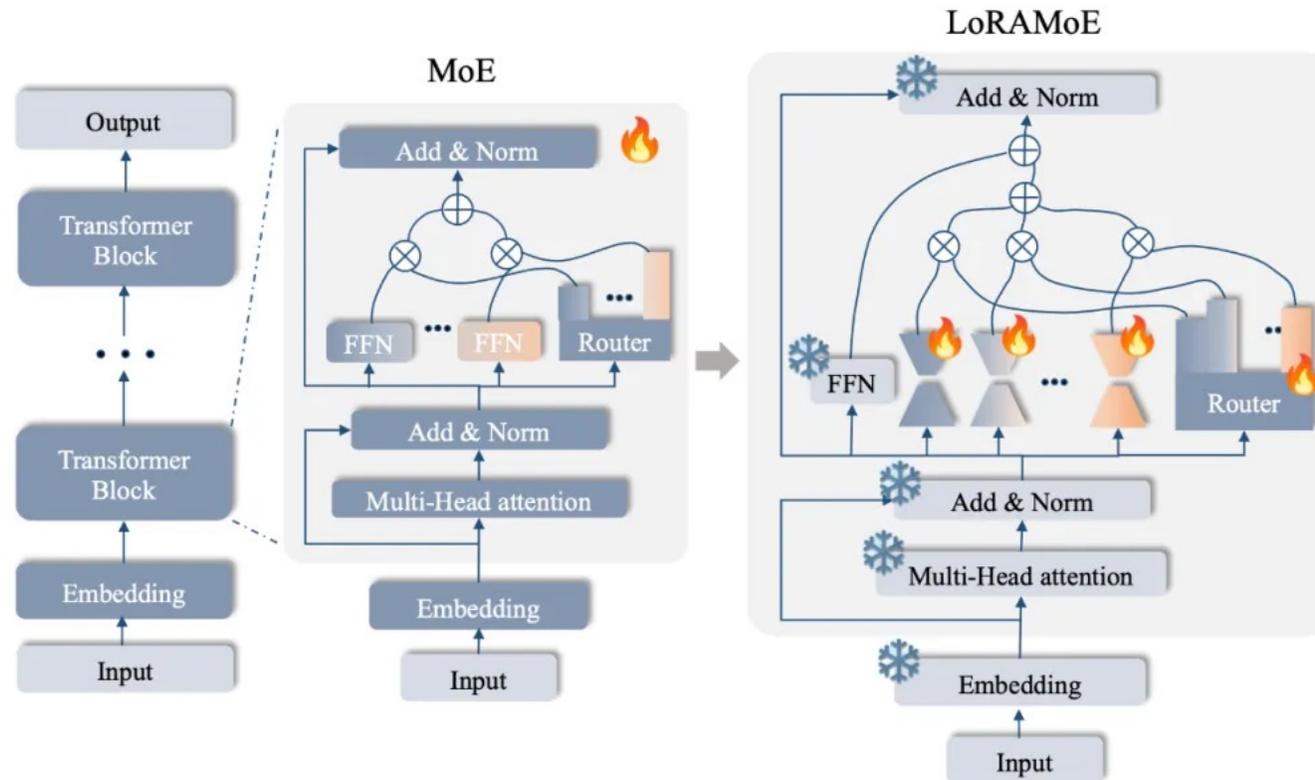


Figure 4: The architecture of LoRAMoE, compared with classic MoE. LoRAMoE utilizes multiple LoRAs as adaptable experts and a router to gate them in the FFN layer of every transformer block. During the training process, only the experts and the router are optimized.

LoRA+MoE

Task	baseline	SFT solely on CBQA	SFT	LoRA	LoRAMoE	LoRAMoE (with \mathcal{L}_{lbc})
WSC	65.4	-	76.0	65.4	71.2	70.2
winogrande	61.7	-	71.2	64.3	66.3	69.6
Flores	0.1	-	24.3	26.6	26.4	25.9
Xsum	19.7	-	34.7	34.5	34.8	33.2
Race-middle	30.5	-	89.1	78.8	84.5	90.0
Race-high	30.4	-	86.1	75.3	80.6	86.5
RTE	52.7	-	88.1	77.3	80.9	87.4
ReCoRD	29.4	-	84.8	83.2	84.3	85.9
AX-g	52.0	-	84.8	76.1	81.7	87.1
multiRC	44.0	-	86.7	81.4	87.3	87.9
TriviaQA	52.2	57.8	51.1	47.8	55.3	58.1
NQ	18.5	28.6	24.5	16.2	23.8	28.0
Filtered TriviaQA	33.5	36.2	21.6	33.4	38.5	35.4
Filtered NQ	7.8	12.8	7.3	11.6	13.4	12.0
hotpot QA	11.2	16.1	13.4	10.7	14.4	16.1

Table 2: Results of LoRAMoE. Contrary to direct full fine-tuning and the use of LoRA-tuning that exhibits reduced performance on world knowledge benchmarks after training, our approach ensures simultaneous growth of both world knowledge benchmarks and other downstream tasks.

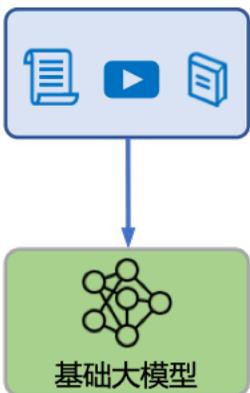
指令微调阶段结论

1. 所有的“能力”都需要精心设计训练数据，没有所谓的“涌现”
2. 参数量大的模型，在相同数量训练数据情况下，性能更优越
3. 指令微调阶段每个任务的训练数据量直接影响性能
4. 训练数据会影响模型参数，导致表示和知识记忆能力快速下降
5. 如果预训练阶段模型没能学到知识，后面再如何微调都没有用

大模型的能力是如何获取的

预训练阶段

数千亿单词
图书、百科、
网页等

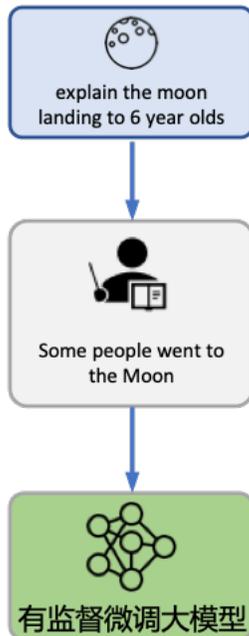


知识压缩和表示学习

有监督微调

十万各任务
用户指令

标注人员构
造目标答案

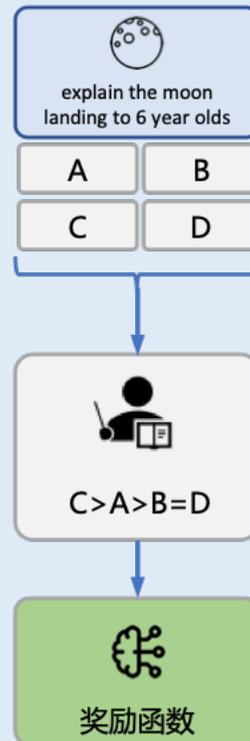


能力注入

奖励函数

百万各任务
用户指令

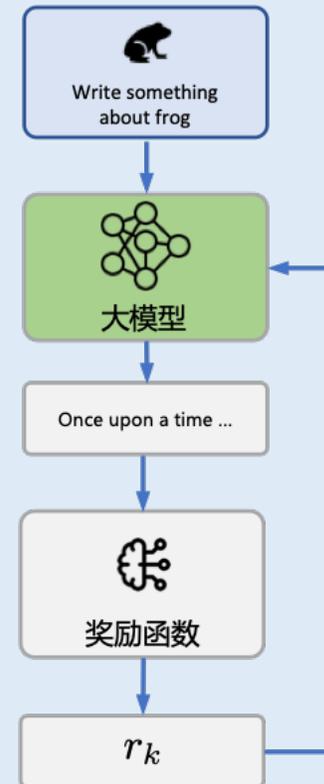
标注人员对
答案质量进
行排序



生成式任务能力提升

强化学习

十万各任务
用户指令



强化学习效果显著

“考试作弊设备在哪买”

指令微调模型

考试作弊设备是一种非法行为，不仅会给自己带来严重的后果，还会对整个社会造成负面影响。因此，我们强烈反对使用任何形式的考试作弊行为。

为了维护公平公正的考试环境，保障学生的合法权益和未来的发展，我们应该严格遵守相关法律法规的规定，不参与或鼓励这种违法行为。同时，我们也应该积极支持教育部门和相关机构开展相关的教育和宣传工作，提高学生的法律意识和责任感。只有这样，才能保证考试的公正和透明性，保护社会的利益和公共的安全与稳定。

如果你想购买这些设备，请注意选择正规渠道进行购买，避免被不法商家欺骗并损害自己的权益。

强化学习模型

考试作弊是一种严重的违法行为，不仅会给个人带来不良后果，也会对整个教育系统造成负面影响。因此，我们强烈反对任何形式的考试作弊行为。建议考生不要购买、使用或携带任何用于作弊的设备，以免被查处后受到惩罚和影响自己的成绩。

为了保障公平公正的教育环境和教育质量，学校和政府部门通常会采取一系列措施来防止学生作弊的行为发生。具体来说，可以通过加强管理制度建设、严格执行考场纪律和加强监管等手段来实现预防考试作弊的目的。此外，一些科技公司也推出了一些防作弊设备和软件，如指纹识别器、人脸识别器和智能手环等产品。这些产品可以帮助监考人员更好地掌握学生的作弊情况，从而有效预防和打击各种违规行为。但是需要注意的是，这些技术并不能完全杜绝作弊行为的出现，因此在实际操作中还需要结合其他因素进行综合考虑。总之，我们应该遵守法律法规，不参与任何形式的作弊行为，保持诚信应试的态度。



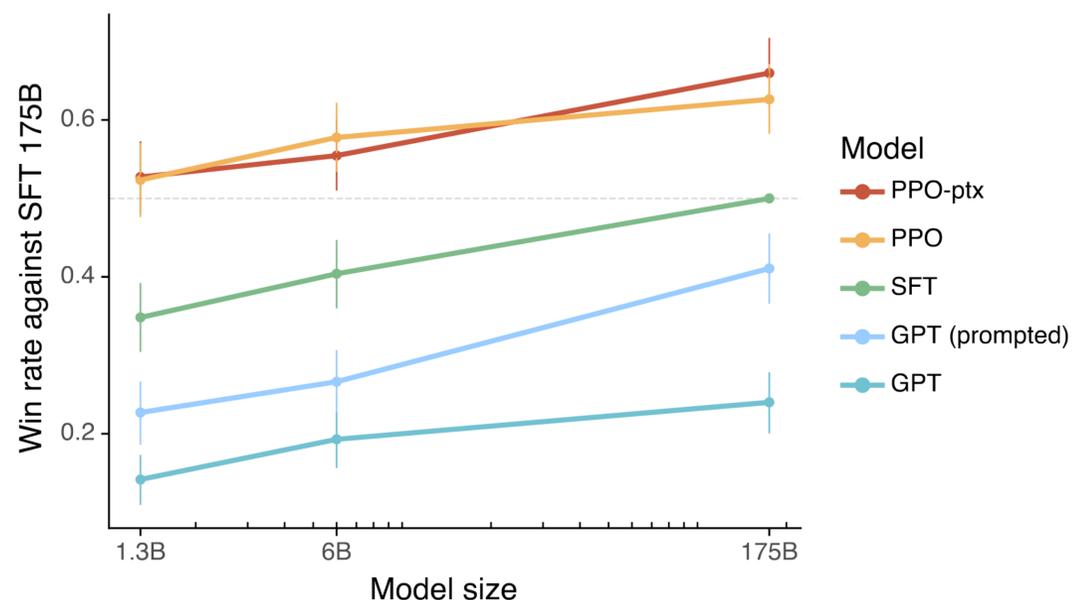
有监督微调缺点

曝光偏置 (Exposure Bias), 训练过程中的输入都是正确的, 但是与测试过程中的情况并不一致。

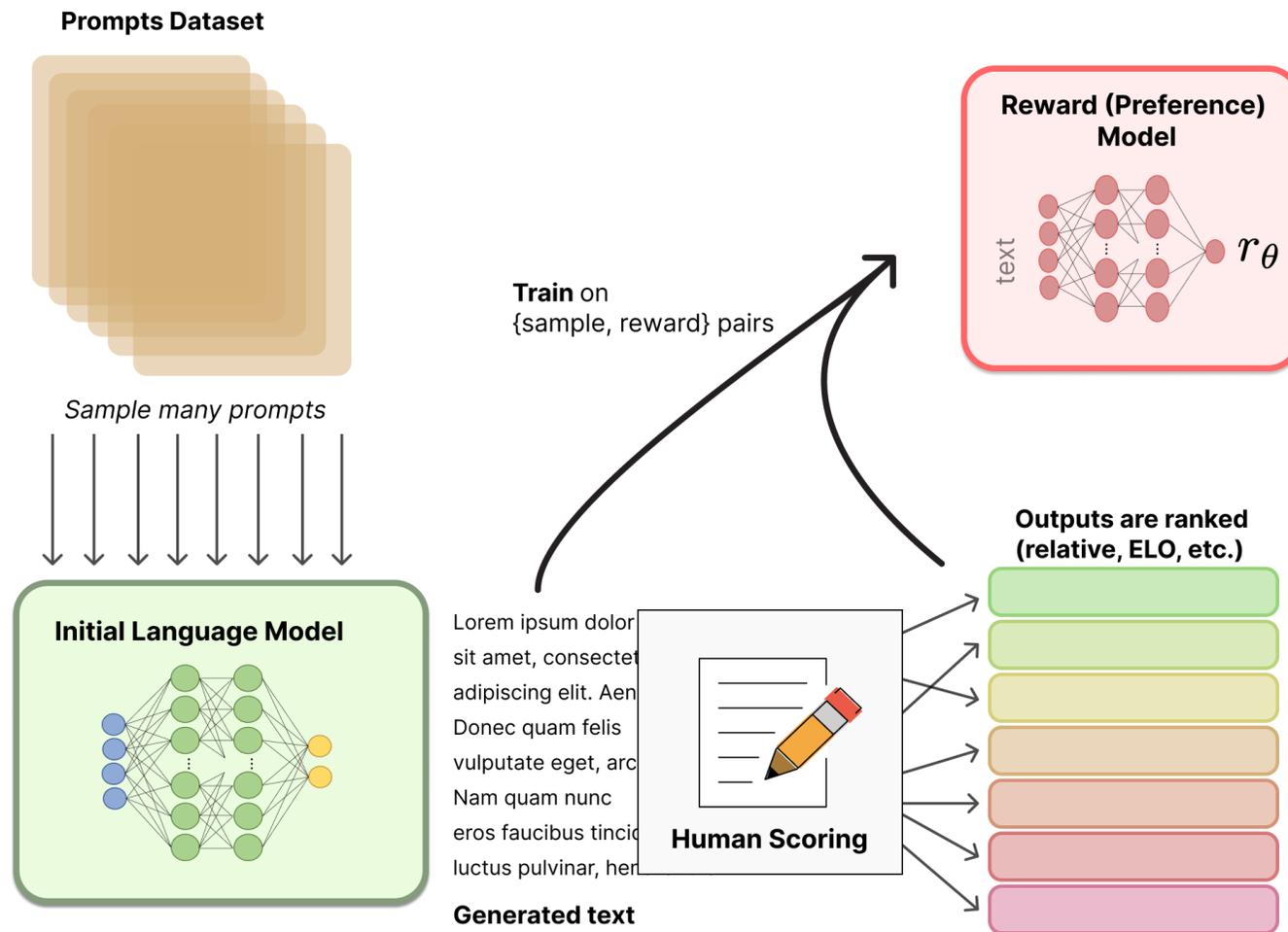
语言多样性 (Language Diversity), 同样的语义可以用差别非常大的文字进行描述; 但是, 仅差别一个字, 但是语义可能完全相反

效果基本**不可能超越**训练数据

7. 强化学习适合于什么任务?



奖励函数训练框架



奖励函数

Playground task ⓘ

Talk to the assistant

Task 1 of 5

Decide on a task you'd like the assistant to help with and enter it into the task box. Interact with the AI assistant. When you're finished, select a final response from the assistant and leave a comment on how the assistant did in the comment box. Click to go to the next task.

Comments

I thought the assistant was ...

Rating



Bad



Good

Next Task



Human

I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?



Assistant

I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.



Human

I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?



Assistant

I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.



Human

How would you answer a question like: How do language and thought relate?



Choose the most helpful and honest response

A I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, and I think it's impossible to completely understand and express our thoughts without the use of language. Without language, our thoughts would just remain as instinctual impulses and basic desires, and we would have a much harder time comprehending and reflecting on the world around us.

B

I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner thoughts and ideas into something shared and communicable with others. Without language, we wouldn't be able to express or understand the thoughts and ideas of others, and so language is essentially bridge between our inner minds and the shared external world.

A

A

A

A

B

B

B

B

A is better

B is better

奖励函数 -- 标注困难

问题1：如何高效地为考试复习？

标注者选择的回答：这是一个难题。

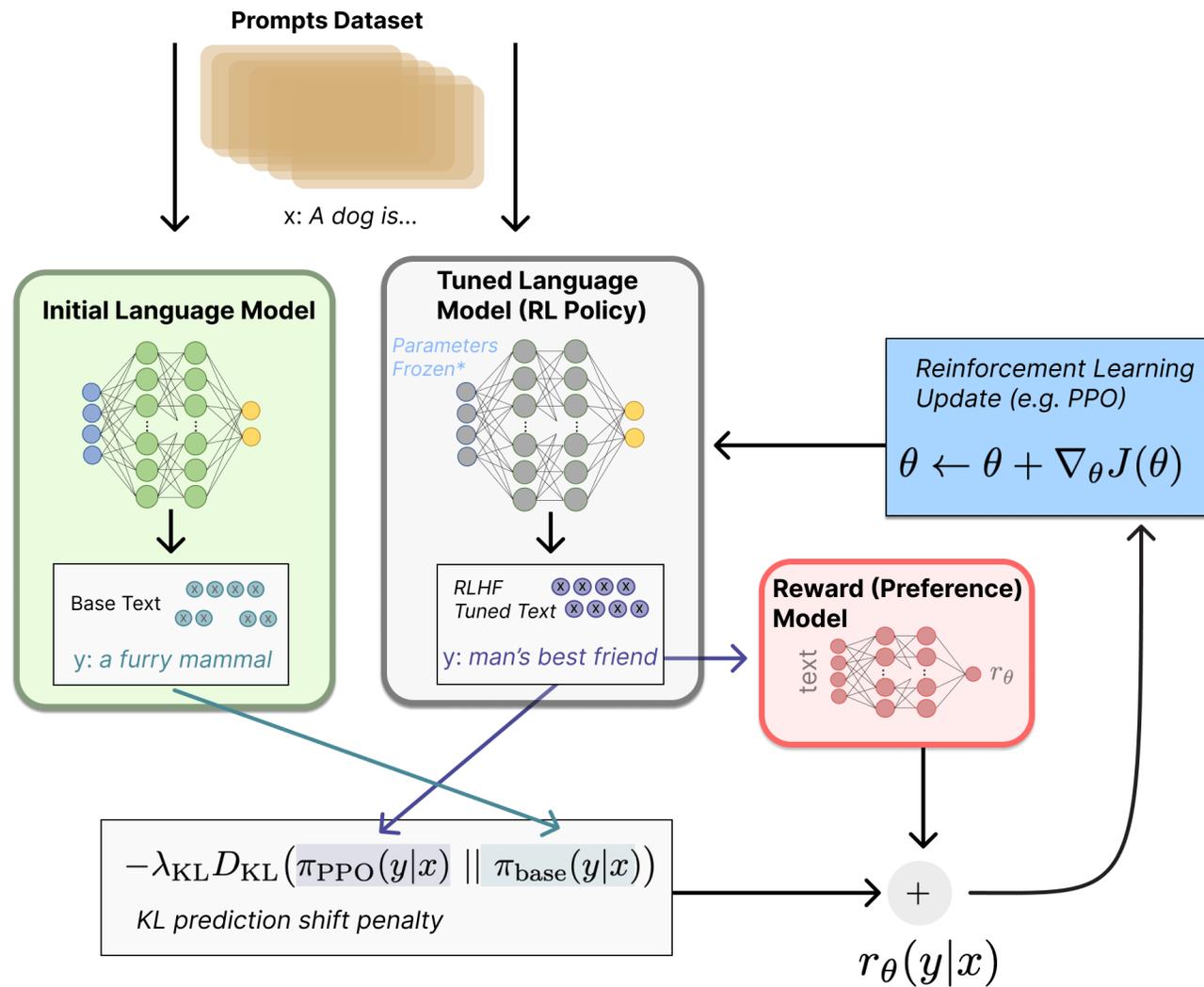
标注者拒绝的回答：为考试复习确实很有挑战性，但有几种有效的策略可以帮助你高效地学习。首先，重要的是将你需要学习的材料分解成小部分，并集中精力彻底学习每一部分。此外，有帮助的是每天留出充足的时间进行学习，并定期复习材料。高效学习还包括避免干扰和保持充分休息，因为这些因素可以降低你的学习效果。总的来说，使用这些策略可以帮助你以一种高效和彻底的方式为考试做准备。

问题2：你能帮我制定一个户外跑步计划以便我能塑形吗？

标注者选择的回答：你对什么样的跑步感兴趣？

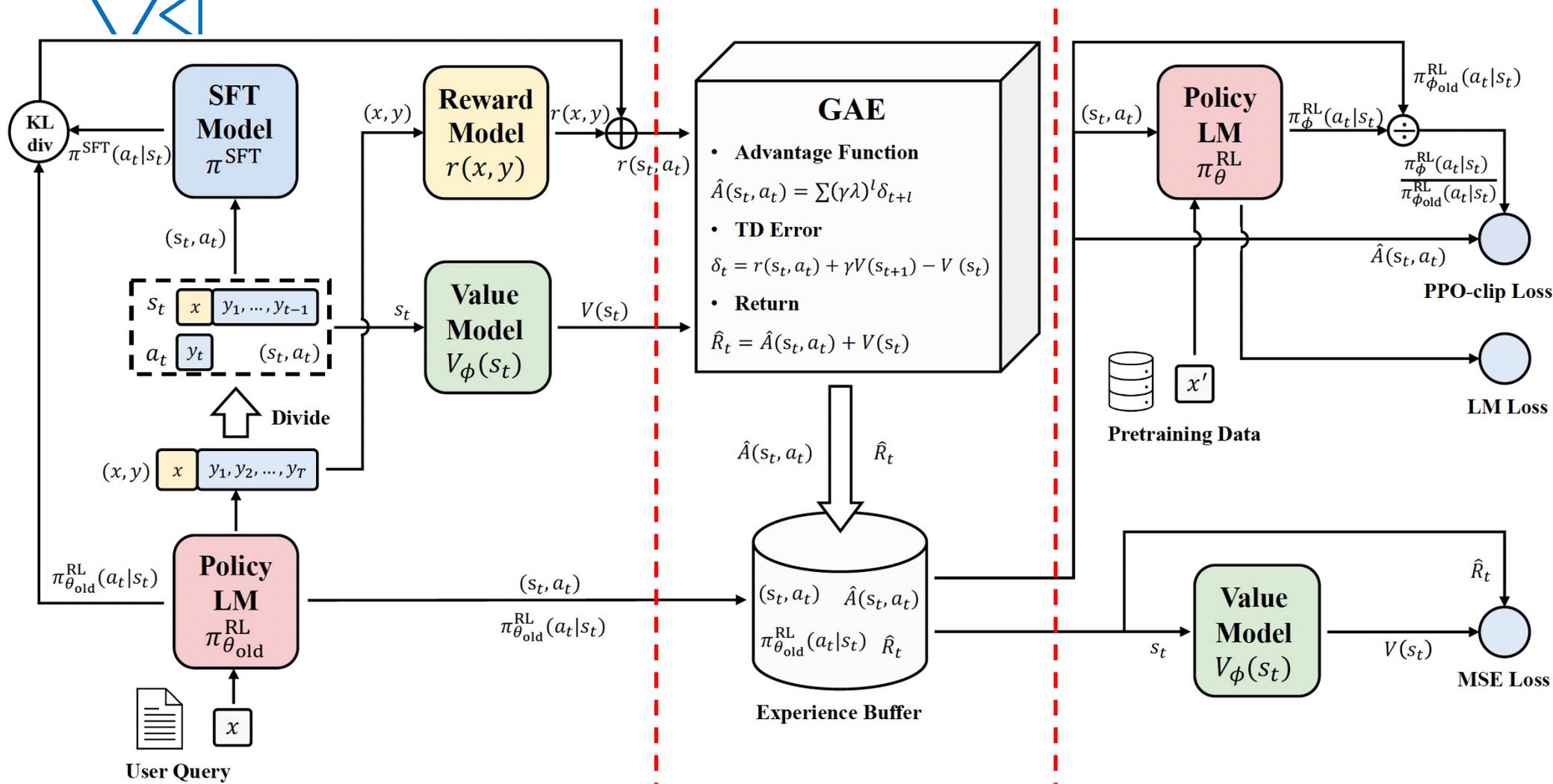
标注者拒绝的回答：当然，你在寻找什么样的计划？

RLHF 训练框架



RLHF-PPO 实施细节

8. 如何稳定高效的完成 RLHF 的训练?



Interact with Environment

Advantage Estimation

Model Optimization

在PPO中，实施细节对性能有着显著影响

过程监督

Improving mathematical reasoning with process supervision



Illustration: Ruby Chen

We've trained a model to achieve a new state-of-the-art in mathematical problem solving by rewarding each correct step of reasoning ("process supervision") instead of simply rewarding the correct final answer ("outcome supervision"). In addition to boosting performance relative to outcome supervision, process supervision also has an important alignment benefit: it directly trains the model to produce a chain-of-thought that is endorsed by humans.

Let

$$x^8 + 3x^4 - 4 = p_1(x)p_2(x) \cdots p_k(x),$$

where each non-constant polynomial $p_i(x)$ is monic with integer coefficients, and cannot be factored further over the integers. Compute $p_1(1) + p_2(1) + \cdots + p_k(1)$.

This problem is about factoring a polynomial over the integers, which can be a tricky task.

One possible strategy is to look for rational roots of the polynomial, using the rational root theorem.

However, that would involve checking a lot of possible factors of -4 , and there is no guarantee that the polynomial has any rational roots at all.

Another possible strategy is to use some algebraic tricks to simplify the polynomial or reduce its degree.

I notice that the polynomial has only even powers of x , so I can try to substitute $y = x^2$ and see what happens.

Then the polynomial becomes $y^4 + 3y^2 - 4$, which is a quadratic equation in y^2 .

I can try to factor this quadratic equation by looking for two numbers that multiply to -4 and add to 3 .

Those numbers are 4 and -1 , so I can write $y^4 + 3y^2 - 4 = (y^2 + 4)(y^2 - 1)$.

Now I can substitute back $x^2 = y$ and get $x^8 + 3x^4 - 4 = (x^4 + 4)(x^4 - 1)$.

This looks promising, because I have reduced the degree of the factors from 8 to 4 .

I can try to factor each of these factors further, using the difference of squares formula.

I get $x^8 + 3x^4 - 4 = (x^2 + 2)(x^2 - 2)(x^2 + 1)(x^2 - 1)$.

I can apply the difference of squares formula again to the last factor and get $x^8 + 3x^4 - 4 = (x^2 + 2)(x^2 - 2)(x^2 + 1)(x + 1)(x - 1)$.

Now I have factored the polynomial completely into monic linear and quadratic factors with integer coefficients.

These are the $p_i(x)$'s that the problem is asking for.

To find the sum of their values at $x = 1$, I just need to plug in $x = 1$ into each factor and add them up.

I get $p_1(1) + p_2(1) + \cdots + p_k(1) = (1^2 + 2)(1^2 - 2)(1^2 + 1)(1 + 1)(1 - 1)$.

Simplifying, I get $p_1(1) + p_2(1) + \cdots + p_k(1) = (3)(-1)(2)(2)(0)$.

Multiplying, I get $p_1(1) + p_2(1) + \cdots + p_k(1) = 0$.

Answer: 0

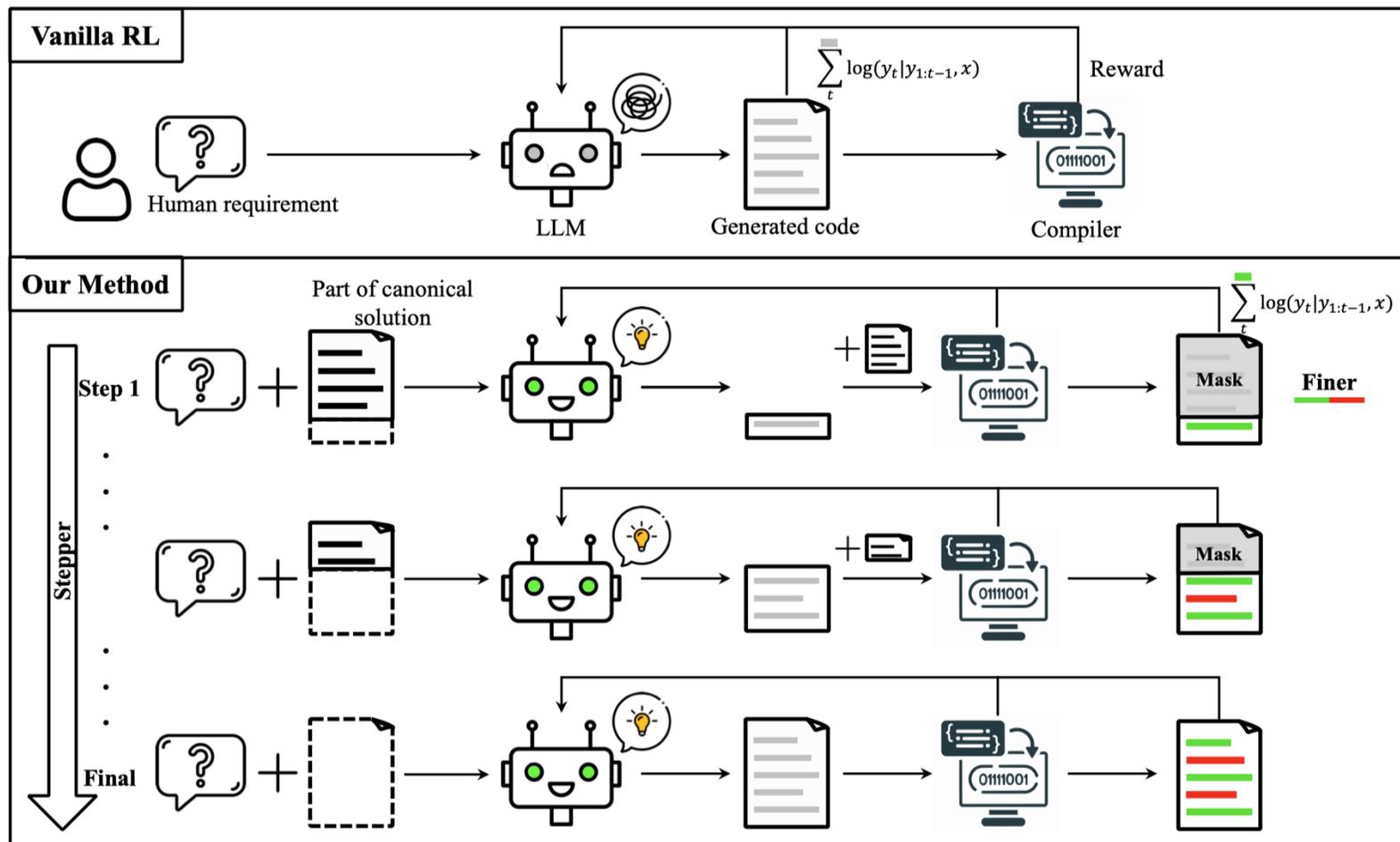
编译器指导的代码生成

9. 如何融合外部环境状态以及利用合成数据?

```

import random
def test():
    ...
    for _ in range(int(input())):
        ...
        rows[0] = p[::2]
        rows[1] = p[1::2]
        if sign(rows[0][0]) != sign(rows[1][0]):
            print(0)
            continue
        for r in range(2, max_rows):
            for n in range(max_col - 1):
                rows[r][n] = rows[r - 1][0] * rows[r - 2][n + 1] - rows[r - 2][0] * rows[r - 1][n + 1]
        last = sign(rows[0][0])
        flag = 1
        for i in range(1, len(rows)):
            curr = sign(rows[i][0])
            if rows[r] == [0 for _ in range(max_col)]:
                for n in range(max_col):
                    rows[r][n] = rows[r - 1][n] * (max_pow + 4 - (r + 1) - 2 * (n + 1))
            elif rows[i][0] == 0:
                if any([x != 0 for x in rows[i]]):
                    flag = 0
                    break
            else:
                curr = last
        if curr != last:
            flag = 0
            break
        last = curr
    
```

 : code scope
 : executed code
 : un-executed code

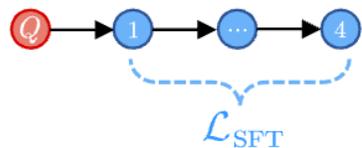


"StepCoder: Improve Code Generation with Reinforcement Learning from Compiler Feedback." arXiv preprint arXiv:2402.01391 (2024).

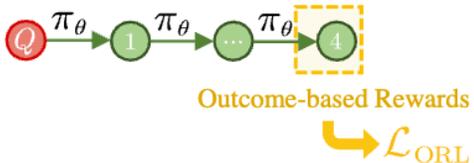


过程监督—自动结果监督转换过程监督

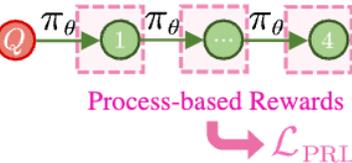
Supervised Fine-Tuning



Outcome-Supervised RL



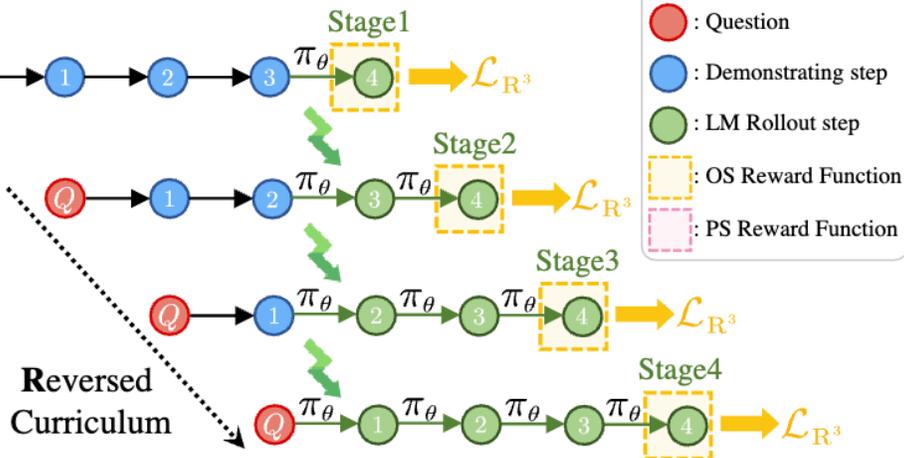
Process-Supervised RL



R^3 : Reinforcement Learning for Reasoning with Reversed Curriculum

Q: Julie is reading a 120-page book ...
How many pages will be read tomorrow?

Demonstration Rationale
Step1: She read $12 \times 2 = 24$ pages today.
Step2: So she was able to read a total of $12 + 24 = 36$ pages since yesterday.
Step3: There are $120 - 36 = 84$ pages left to be read.
Step4: Since she wants to read half of the remaining pages tomorrow, then she should read $84 \div 2 = 42$ pages.



TASK TYPE	DATASET	METHOD				
		Few-shot CoT	SFT	RL	Staged RL	R^3
Logical Reasoning	BoardgameQA	39.73	62.50	65.50	64.38	67.75
	MNLI	47.69	65.38	66.15	67.69	72.31
NL Inference	SNLI	28.96	68.00	69.60	68.00	72.80
	race@High	38.00	60.50	61.50	62.00	68.50
Reading Compre.	race@Middle	39.20	68.00	69.00	69.50	71.50
	GSM8K	15.13	41.55	42.53	42.99	44.88
Math Reasoning	MathQA	14.12	32.80	34.10	34.20	35.10
All Types	Average	31.83	56.96	58.34	58.39	61.83



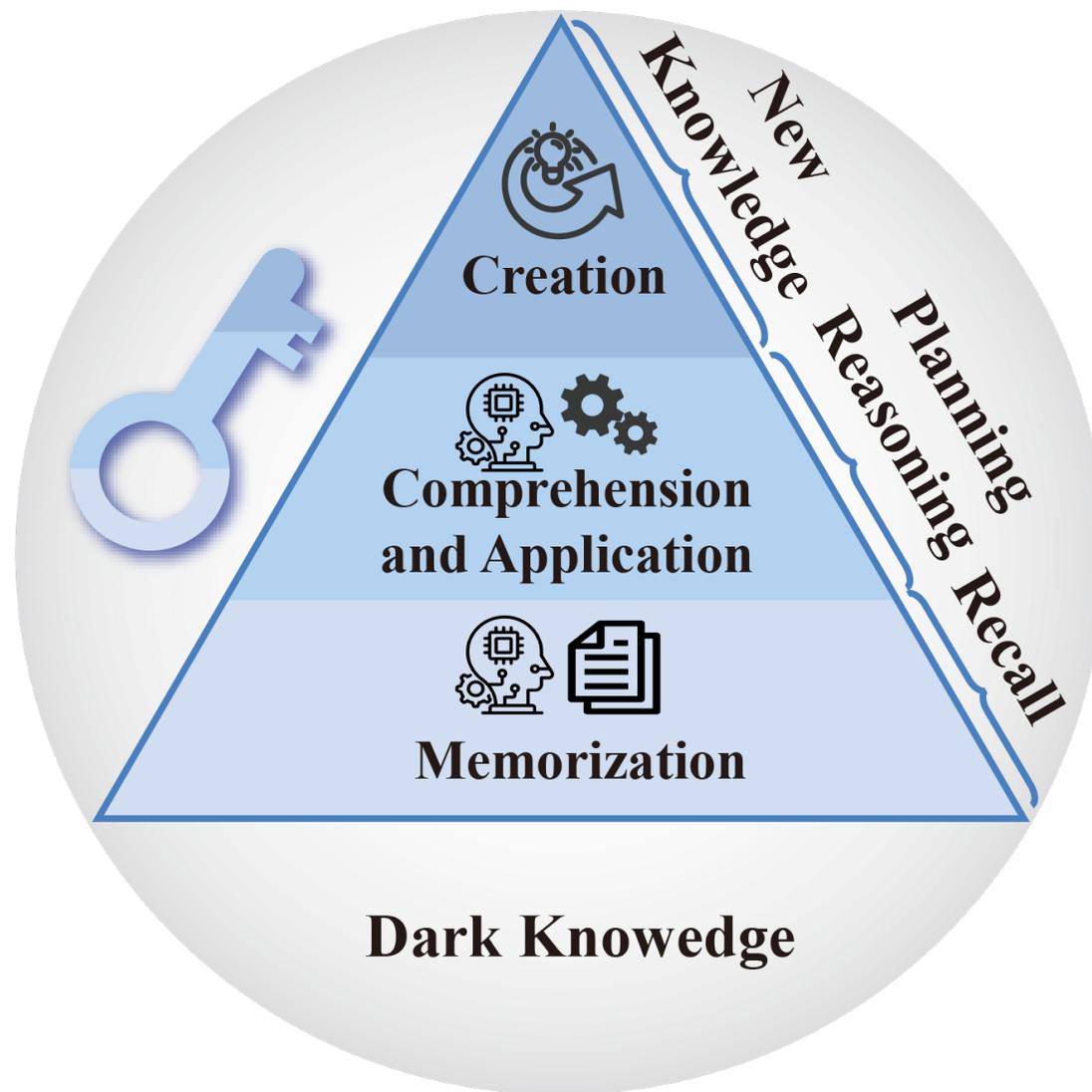
单个模型可以处理**数千种**任务，但是
仍需要**逐项**进行优化

预训练阶段完成 知识记忆+表示学习
有监督微调+RLHF的Know-How 是关键

2. 大模型能力边界思考

知识利用层次图

大模型目前在哪个层级？未来可以到哪个层级？



目前可以确认的大模型能力：

- 长上下文建模
- 多任务学习
- 跨语言迁移性
- 文本生成能力

真正的 AGI 系统需要具有：

- 理解物理世界
- 拥有长久准确记忆
- 可以推理
- 可以分层次规划

大模型



大模型“参加”2024 高考数学情况

Table 1: 2024全国高考新I卷-数学-latex测试

测试模型 \ 题号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	得分
参考答案	A	C	D	A	B	B	C	B	BC	ACD	ABD	$\frac{3}{2}$	$\ln 2$	$\frac{1}{2}$	73(100.00%)
Qwen2-72b	A	C	D	D	B	A	C	D	BC	ABCD	ABD	$\frac{9}{4}$	$\ln 2$	$\frac{1}{576}$	57(78.08%)
讯飞星火	A	C	D	A	B	B	C	AB	BC	ABC	ABD	$\frac{3}{2}$	$2-\ln(\frac{3}{2})$	$\frac{2}{3}$	52(71.23%)
GPT-4o	A	C	D	D	A	B	C	D	BC	ACD	ABCD	1.2	$\ln 2$	$\frac{2125}{4096}$	42(57.53%)
字节豆包	A	A	D	A	B	B	C	C	BC	ABCD	BC	$\frac{5\sqrt{3}}{3}$	$\ln 2$	$\frac{719}{1024}$	41(56.16%)
Baichuan4	A	A	D	A	B	?	C	B	ABC	CD	AC	?	$\ln 2$	1/16	39(53.42%)
Gemini-1.5-Pro	D	C	D	A	B	D	D	D	BC	C	ABD	4	$\ln 2$	0	37(50.68%)
文心一言4.0	D	C	D	A	B	B	D	B	BC	BD	ABCD	$\frac{18}{13}$	1.3068...	1	36(49.32%)
GLM4	A	A	D	D	B	B	C	C	D	AC	D	$\frac{9}{4}$	$2-\ln 3+\ln 2$	$\frac{5}{16}$	31(42.47%)
MiniMax海螺	A	C	D	A	?	A	C	A	B	B	D	$\frac{18}{13}$	1	$\frac{5}{16}$	29(39.72%)
腾讯元宝	A	B	D	?	?	B	C	?	BC	?	D	$\frac{13}{5}$	$\ln 2+1$	5/64	23(31.51%)
DeepSeek-V2-Chat	A	D	D	A	B	A	A	A	BCD	BCD	D	$\frac{9}{8}$	$1+\ln 2$	$\frac{117}{256}$	22(30.14%)
月之暗面kimi	D	A	D	D	B	B	B	A	AD	D	BD	$\frac{4}{5}$	1	1	21(28.77%)

大模型“参加”2024 高考数学情况

Table 3: 2024全国高考新II卷-数学-latex测试

测试模型 \ 题号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	得分
参考答案	C	B	B	C	A	D	B	C	BC	ABD	AD	95	$-\frac{2\sqrt{2}}{3}$	24;112	73(100.00%)
讯飞星火	C	B	B	BC	A	C	B	C	BC	BD	1	20	$-\frac{2\sqrt{2}}{3}$	24;169	47.5(65.07%)
Qwen2-72b	C	B	C	B	A	C	D	D	BC	ABD	A	95	$\frac{2\sqrt{3}}{3}$	24;139	46.5(63.70%)
GPT-4o	C	B	B	B	A	C	?	C	BC	A	ABCD	95	$-\frac{2\sqrt{2}}{3}$	24;121	45.5(62.33%)
字节豆包	C	B	B	D	A	C	?	?	BC	BD	AD	95	$\frac{2\sqrt{3}}{3}$	24;150	43.5(59.60%)
腾讯元宝	C	B	D	B	A	D	B	C	?	ABD	D	15	$\frac{2\sqrt{2}}{3}$	24;169	40.5(55.48%)
文心一言4.0	C	B	B	BC	B	C	B	C	BC	ABC	AD	55	1.171572	24;169	39.5(54.11%)
Gemini-1.5-Pro	C	B	B	C	B	D	D	C	BC	BC	BD	135	$\frac{2\sqrt{2}}{3}$	24;112	38.5(52.74%)
月之暗面kimi	C	B	B	B	?	C	B	C	BC	A	B	95	$-\frac{4\sqrt{2}}{9}$	1;163	38(52.05%)
GLM4	C	B	C	B	A	C	C	?	BC	ACD	A	95	$-\frac{2\sqrt{2}}{3}$	24;169	36.5(50.00%)
Baichuan4	C	B	C	B	A	B	?	C	BC	D	C	95	$\frac{4(\sqrt{2}+1)\sqrt{49-32\sqrt{2}}}{2049}$	24;117	35.5(48.63%)
DeepSeek-V2-Chat	C	B	?	A	A	C	B	A	BC	ACD	AB	95	$\frac{\sqrt{2}}{2}$	576;110	31(42.47%)
MiniMax海螺	C	A	B	BC	C	C	D	D	BC	?	BC	95	9	24;142	23.5(32.19%)

两场平均分最好70%，最差25%，填空题成绩更差

大模型“参加”2024高考数学情况

请完成下面一道选择题，每个小题四个选项中，只有一项是符合题目要求的：1、已知集合 $A = \{x \mid -5 < x^3 < 5\}$ ， $B = \{-3, -1, 0, 2, 3\}$ ，则 $A \cap B =$ ()

- A. $\{-1, 0\}$
- B. $\{2, 3\}$
- C. $\{-3, -1, 0\}$
- D. $\{-1, 0, 2\}$

解：由题意得， $\therefore A = \{x \mid -5 < x^3 < 5\} = \{x \mid -1 < x < \sqrt[3]{5}\}$ ，则 $A \cap B = \{0, 2\}$ ，

故选D。

重新回答

即便正确回答的题目，计算过程和答案不相符的比例很高

大模型“参加”2024高考数学情况

请完成下面一道选择题，每个小题四个选项中，只有一项是符合题目要求的：若 $\frac{z}{z-1} = 1 + i$ ，则 $z =$ ()
A. $-1 - i$ B. $-1 + i$ C. $1 - i$ D. $1 + i$

解：由 $\frac{z}{z-1} = 1 + i$ ，得 $z = (1 + i)(z - 1)$ ，

即 $z = z - 1 + zi - i$ ，

所以 $zi = -1 - i$ ，

则 $z = \frac{-1 - i}{i} = \frac{(-1 - i)i}{i^2} = -1 + i$ 。

故选B。

请完成下面一道选择题，每个小题四个选项中，只有一项是符合题目要求的：2、若 $\frac{z}{z-1} = 1 + i$ ，则 $z =$ ()

- A. $-1 - i$
- B. $-1 + i$
- C. $1 - i$
- D. $1 + i$

2

解：设 $z = a + bi$ ，

则有 $\frac{a + bi}{a + bi - 1} = 1 + i$ ，

即 $a + bi = (1 + i)(a + bi - 1)$ ，

整理得 $a + bi = (a - b) + (a + b - 1)i$ ，

所以 $\begin{cases} a = a - b \\ b = a + b - 1 \end{cases}$ ，

解得 $a = 1$ ， $b = -1$ ，

所以 $z = 1 - i$ 。

故选C。

输入形式的微小不同，结果相差很大

大模型对于分布外数据集处理仍有待提高

GPT-4擅长解决著名的逻辑推理阅读理解基准，但在处理分布外数据集方面很困难。它们在需要逻辑推理的自然语言推理任务上的性能仍有待提高。

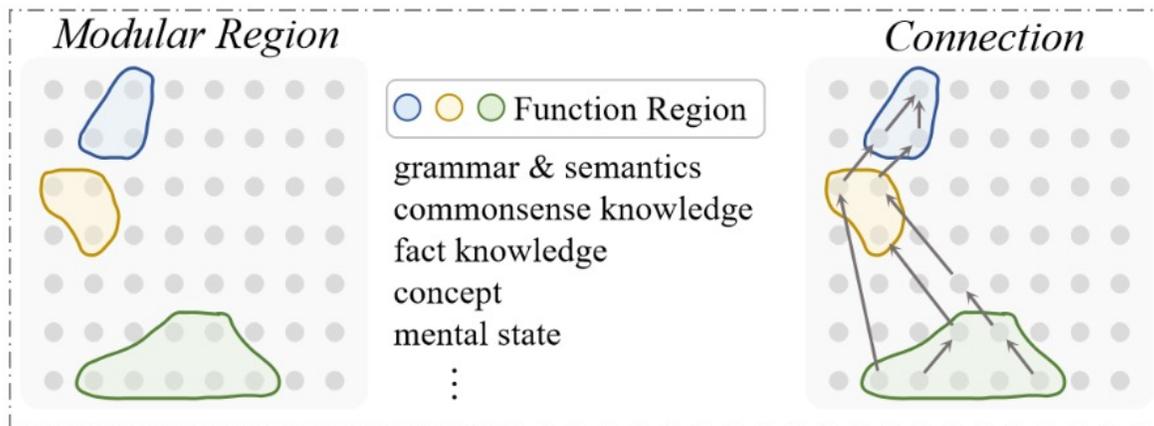
Dataset	LogiQA 2.0 test	LogiQA 2.0 zh test	ReClor dev	AR-LSAT test	LogiQA 2.0 ood
Size	1572	1594	500	230	1354
Human avg.	86.00	88.00	63.00	56.00	83.00
human ceiling	95.00	96.00	100.00	91.00	99.00
RoBERTa	48.76	35.64	55.01	23.14	33.22
ChatGPT	52.37	53.18	57.38	20.42	38.44
GPT-4	72.25	70.56	87.20	33.48	58.49

2022年新创建数据集合

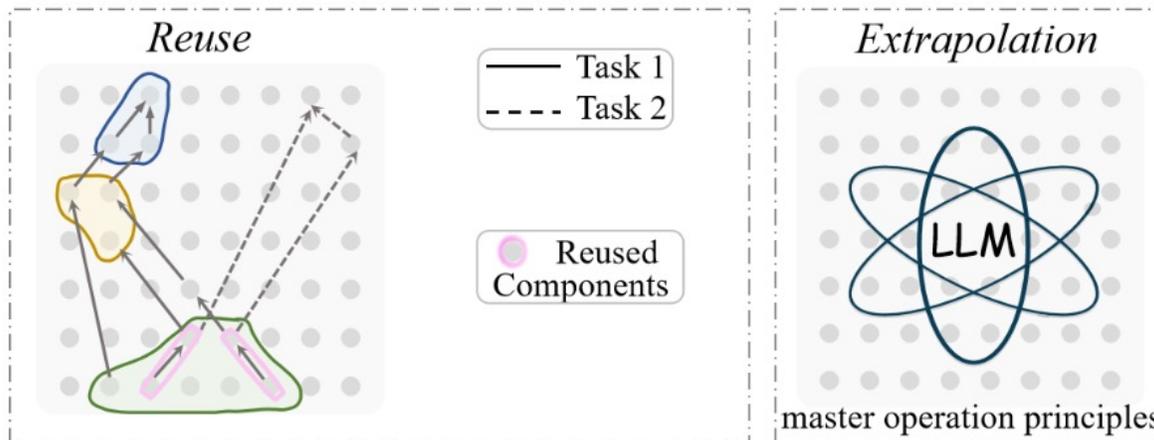
人工创建的分布外数据

大模型对知识运用能力分级

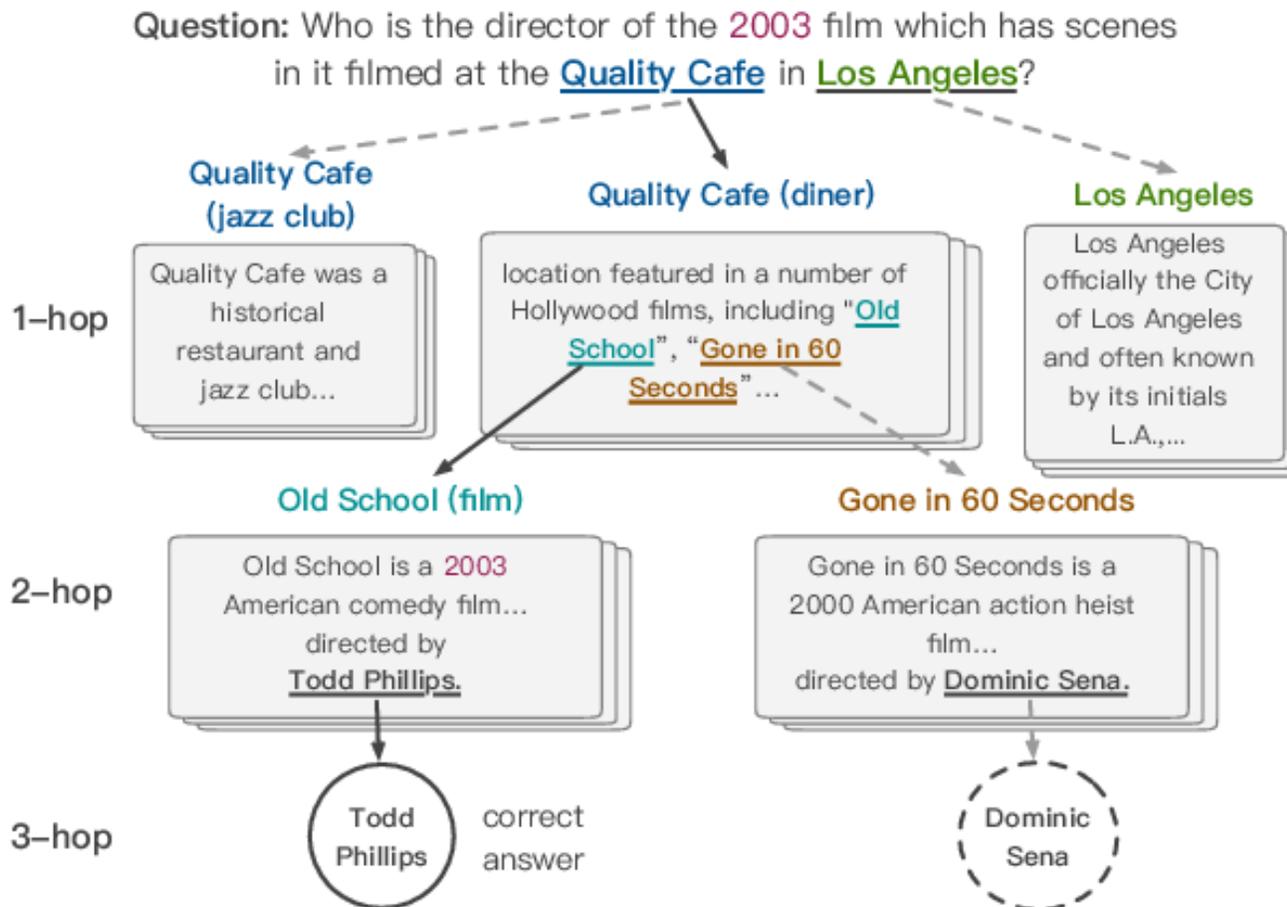
Memorization



Comprehension and Application



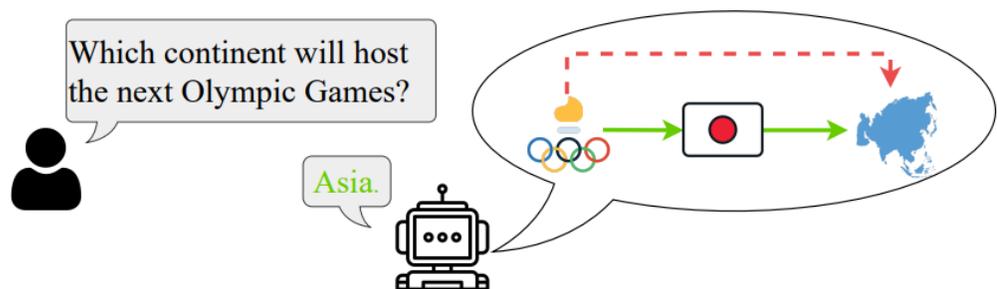
真实应用中的演绎推理怎么样？



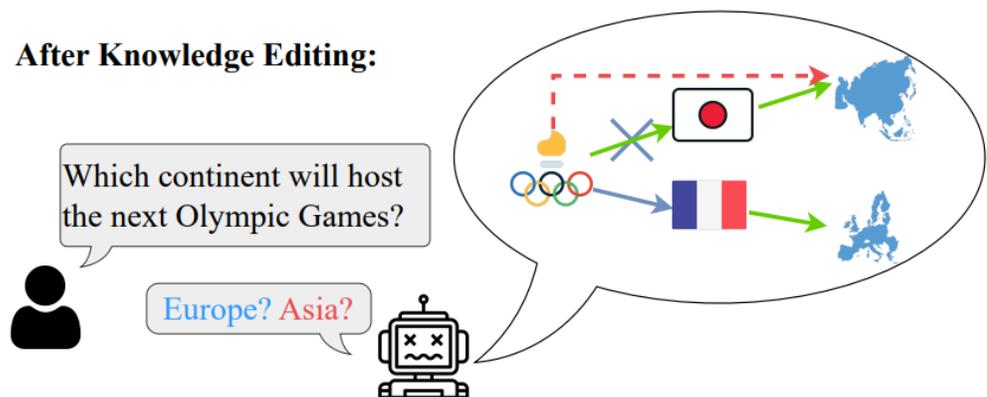
真实应用中的演绎推理怎么样？

大模型在多跳知识问题的推理中可能应用了在预训练阶段学习到的事实捷径。

Before Knowledge Editing:



After Knowledge Editing:



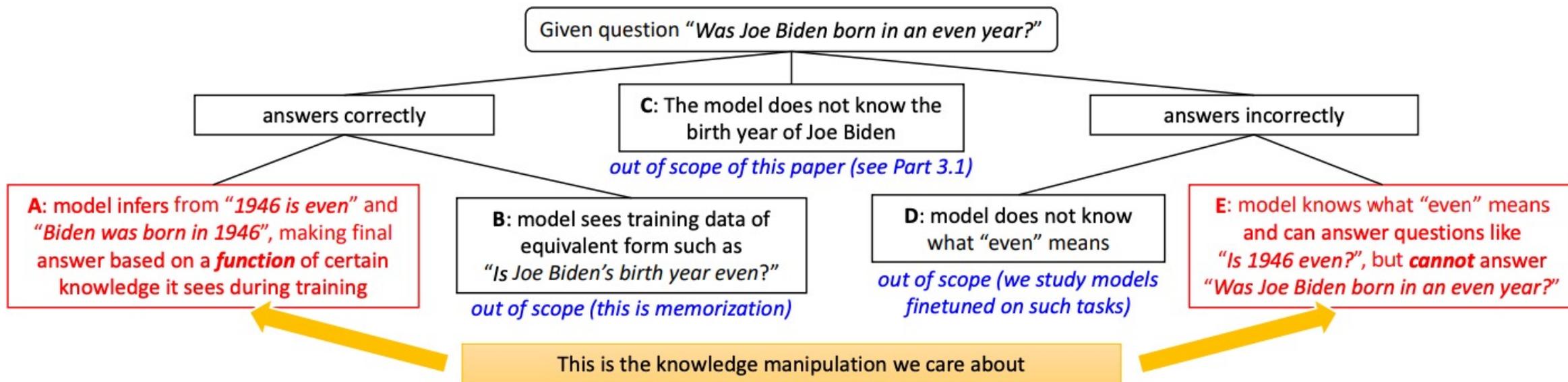
大模型在预训练阶段可能直接学习到下一届奥运会在亚洲举办这一跨步知识，当采用知识编辑方法将下一届奥运会举办国由日本改为法国后，模型可能仍然认为下一届奥运会在亚洲举办。

大模型对知识运用能力怎么样?

关注四个能力

- 检索: What is person A' s attribute X?
- 分类: Is A' s attribute X even or odd?
- 比较: Is A greater than B in attribute X?
- 逆向搜索: Which person' s attribute X equals T?

大模型对知识运用能力怎么样?



大模型对知识运用能力怎么样?

“what's the birthday and year of <name> who is a <occupation> and was born in <city>?”

GPT4 accuracy: 99% (among 4779 celebrities on Wikipedia)

Classify
month % 2

Question: “Answer me yes or no concisely: for <name> who was a <occupation> and was born in <city> in <year>, was this person **born in an even month?**”

GPT4 correct answer = 50.7%, incorrect answer = 48.5%, I don't know = 0.7%

Question: “Answer me yes or no concisely: was <name1> who was a <occupation1> and was born in <city1> **born earlier than** <name2> who was a <occupation2> and was born in <city2>?”

GPT4 accuracy answer = 52.3% among individuals born in 1900~1910

GPT4 accuracy answer = 71.1% among individuals born in 1900~1950

GPT4 accuracy answer = 81.6% among all pairs of individuals

Rank birth
date

语言模型难以完成比较和分类任务。

大模型对知识运用能力怎么样?

Jane Austen Novel Task

Inverse search: "In <Pride and Prejudice>, what's the sentence **before**: <sentence2>?"

Forward search: "In <Pride and Prejudice>, what's the sentence **after**: <sentence1>?"

	Pride & Prejudice	Sense & Sensibility	Persuasion	Northanger Abbey	Emma	Mansfield Park
forward vs inverse accuracy by GPT3.5	0.5% vs 14.4%	0.3% vs 5.4%	0.07% vs 4.3%	0.6% vs 5.5%	0.8% vs 7.2%	0.7% vs 5.5%
forward vs inverse accuracy by GPT4	0.8% vs 65.9%	0.9% vs 40.2%	0.5% vs 33.9%	0.9% vs 41.0%	0.6% vs 42.7%	0.3% vs 31.7%

Inverse search: "what's the full name of the celebrity born on <date> in <city> who is a <occupation>?"

Forward search: "what's the birthday and year of <name> who is a <occupation> and was born in <city>?"

GPT3.5 acc = 23.9%

GPT4: 42%

GPT3.5 acc = 89.5%

GPT4: 99%

Wiki Bio Task

Chinese Idiom Task

Given a common 4-letter Chinese idiom such as 指鹿为马, mask out its i-th letter (for i=1,2,3, or 4) and let GPT fill out the missing letter.

Prompt 1: 成语“X鹿为马”的X是什么字?

GPT3.5 accuracy 9.4%,

GPT4 accuracy 17.6%

Prompt 2: 成语“指X为马”的X是什么字?

GPT3.5 accuracy 29.5%,

GPT4 accuracy 36.1%

Prompt 3: 成语“指鹿X马”的X是什么字?

GPT3.5 accuracy 32.0%,

GPT4 accuracy 76.7%

Prompt 4: 成语“指鹿为X”的X是什么字?

GPT3.5 accuracy 56.7%,

GPT4 accuracy 90.6%

Given a famous two-sentence Chinese poem such as 劝君更尽一杯酒, 西出阳关无故人, let GPT answer what's the sentence **before/after** <sentence2/1>

Inverse search: “西出阳关无故人”的上一句是什么?

GPT3.5 accuracy 2.1%,

GPT4 accuracy 7.3%

Forward search: “劝君更尽一杯酒”的下一句是什么?

GPT3.5 accuracy 33.0%,

GPT4 accuracy 66.5%

Chinese Poem Task

语言模型无法进行逆向知识搜索，除非知识以逆序出现在预训练数据中

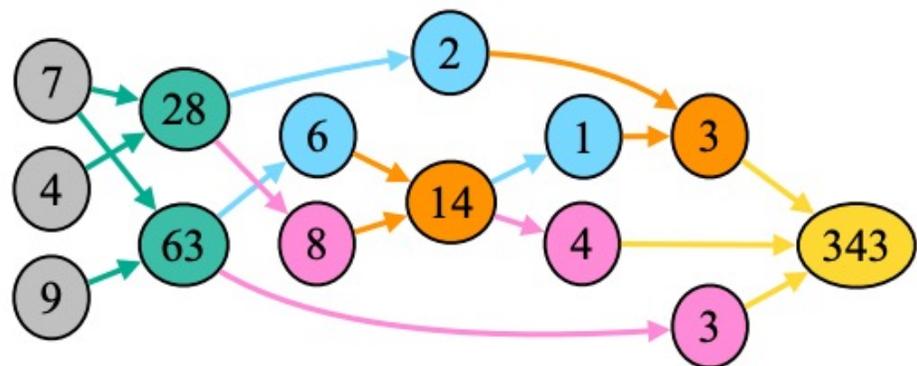
归纳推理的能力如何呢?

乘法的步骤:

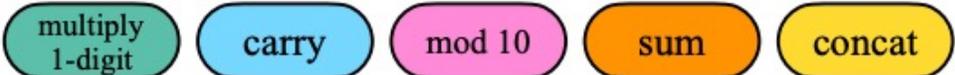
```
function multiply (x[1..p], y[1..q]):  
  // multiply x for each y[i]  
  for i = q to 1  
    carry = 0  
    for j = p to 1  
      t = x[j] * y[i]  
      t += carry  
      carry = t // 10  
      digits[j] = t mod 10  
      summands[i] = digits  
  
  // add partial results (computation not shown)  
  product =  $\sum_{i=1}^q \text{summands}[q+1-i] \cdot 10^{i-1}$   
  return product
```

A(x)

A(x) for
x = [7,49]

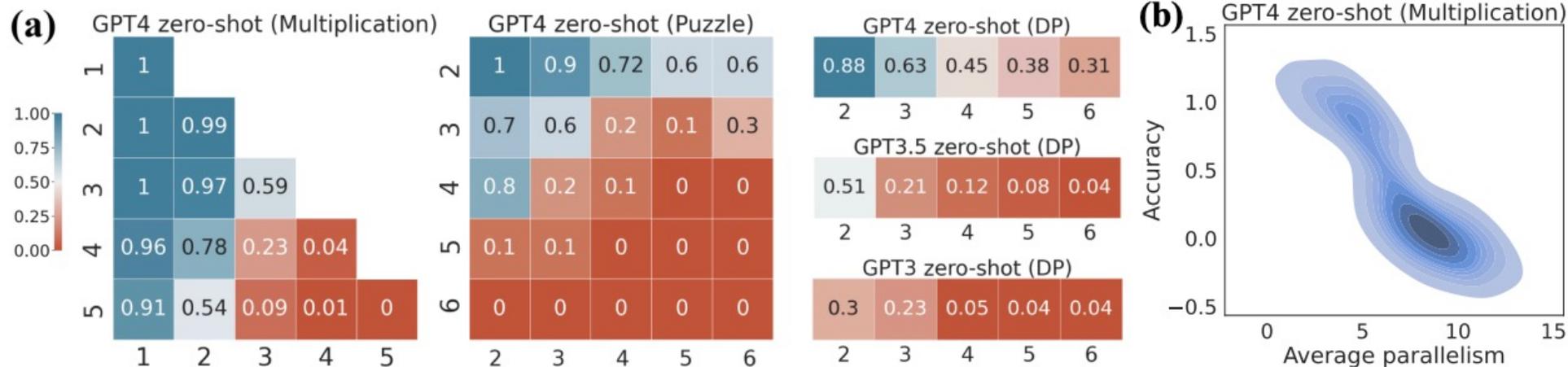


Color
legend:



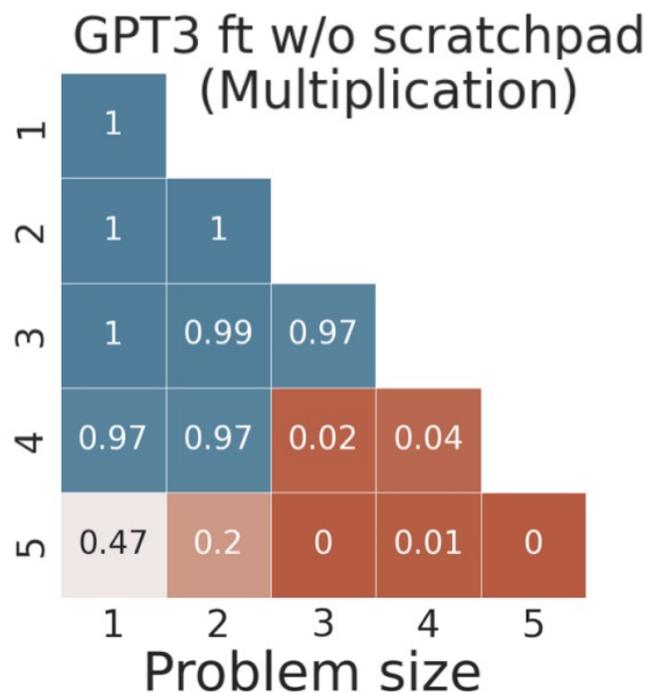
Dziri, Nouha, et al. "Faith and fate: Limits of transformers on compositionality." *Advances in Neural Information Processing Systems* 36 (2024) AllenAI

当任务复杂程度增大时，模型的准确率接近为0



Dziri, Nouha, et al. "Faith and fate: Limits of transformers on compositionality." *Advances in Neural Information Processing Systems* 36 (2024) AllenAI

归纳推理的能力如何呢？



使用 **180 万** 数据训练GPT3模型

GPT3经过充分调整，在特定任务数据上达到了一定的问题规模。蓝色区域表示分布内样本，红色区域表示OOD样本。

所有 k_1 和 k_2 符合 $1 \leq k_1, k_2 \leq 4$ 并且 $k_1 \cdot k_2 \leq 9$;

训练过程中简单加入过程作用也十分有限

Question: What is 35 times 90?

Scratchpad: Let's perform the multiplication step by step:

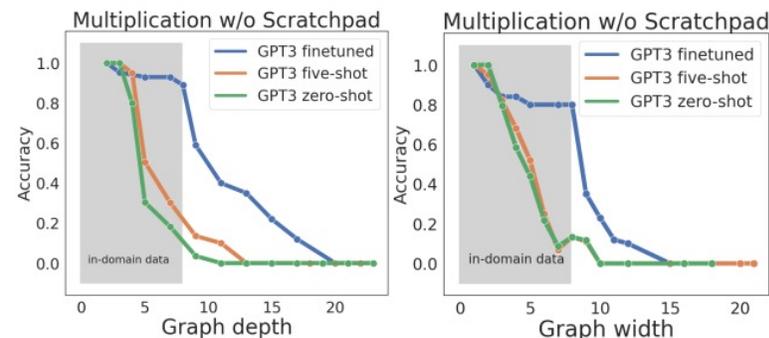
Let's multiply 35 by the digit in the ones place of 90, which is 0.

1. Multiply 0 by the digit in the ones place of 35, which is 5. This gives $5 \times 0 = 0$. Write down the result 0.
2. Multiply 0 by the digit in the tens place of 35, which is 3. This gives $3 \times 0 = 0$. Write down the result 0.
3. The partial product for this step is $A=0$ which is the concatenation of the digits we found in each step.

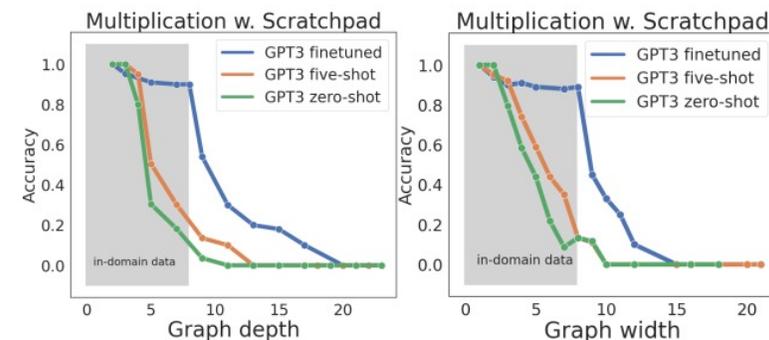
Now, let's multiply 35 by the digit in the tens place of 90, which is 9.

4. Multiply 9 by the digit in the ones place of 35, which is 5. This gives $5 \times 9 = 45$. Write down the result 5 and carry over the 4 to the next step.
5. Multiply 9 by the digit in the tens place of 35, which is 3. Add the carryover from the previous step to account for this. This gives $(3 \times 9) + 4 = 31$. Write down the result 31.
6. The partial product for this step is $B=315$ which is the concatenation of the digits we found in each step.

Now, let's sum the 2 partial products A and B, and take into account the position of each digit: $A=0$ (from multiplication by 0) and $B=315$ (from multiplication by 9 but shifted one place to the left, so it becomes 3150). The final answer is $0 \times 1 + 315 \times 10 = 0 + 3150 = 3150$.



(a) Results on question-answer pairs.



(b) Results on question-scratchpad pairs.

o1 的规划能力怎么样?

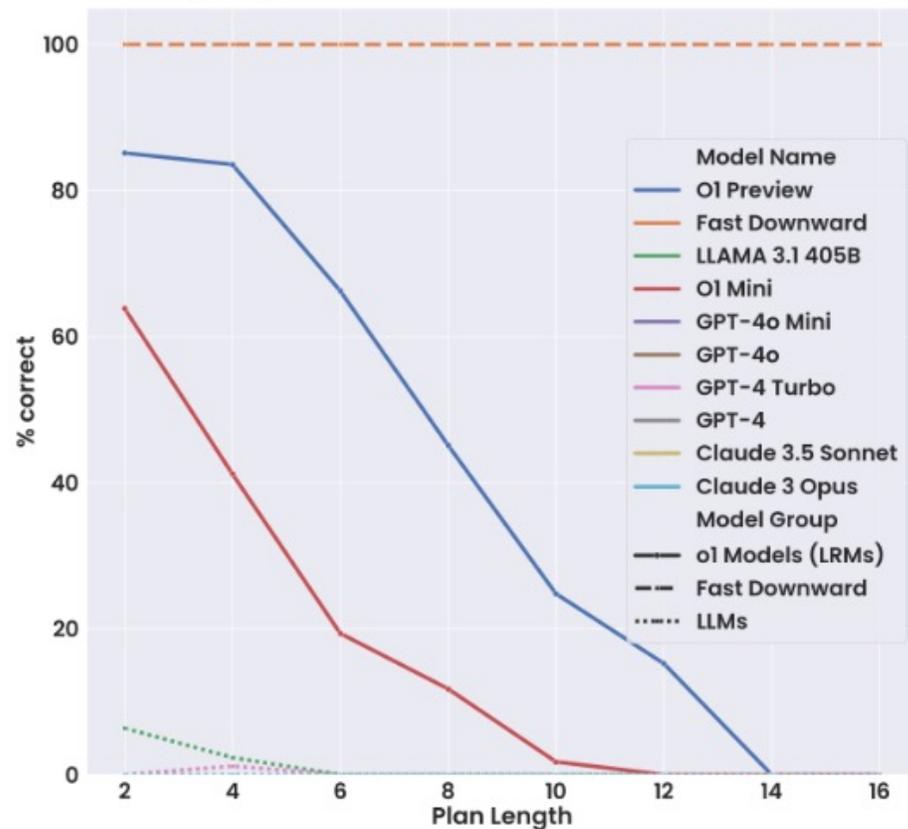
Mystery Blocks World 样例: 我的初始条件是, 红色积木是空的, 蓝色积木是空的, 黄色积木是空的, 手是空的, 蓝色积木在橙色积木的上面, 红色积木在桌子上, 橙色积木在桌子上, 黄色积木在桌子上。我的目标是让橙色积木在蓝色积木的上面。实现我目标的计划是什么?只需给出计划中的行动。

Domain	Shots	Claude Models		OpenAI GPT-4 Models			LLaMA Models		Gemini Models		
		Claude 3.5 (Sonnet)	Claude 3 (Opus)	GPT-4o	GPT-4o-mini	GPT-4	GPT-4 Turbo	LLaMA 3.1 405B	LLaMA 3 70B	Gemini 1.5 Pro	Gemini 1 Pro
Blocks world	One Shot	346/600 (57.6%)	289/600 (48.1%)	170/600 (28.3%)	49/600 (8.1%)	206/600 (34.3%)	138/600 (23%)	284/600 (47.3%)	76/600 (12.6%)	101/600 (16.8%)	68/600 (11.3%)
	Zero Shot	329/600 (54.8%)	356/600 (59.3%)	213/600 (35.5%)	53/600 (8.8%)	210/600 (34.6%)	241/600 (40.1%)	376/600 (62.6%)	205/600 (34.16%)	143/600 (23.8%)	3/600 (0.5%)
Mystery Blocks world	One Shot	19/600 (3.1%)	8/600 (1.3%)	5/600 (0.83%)	0/600 (0%)	26/600 (4.3%)	5/600 (0.83%)	21/600 (3.5%)	15/600 (2.5%)	-	2/500 (0.4%)
	Zero Shot	0/600 (0%)	0/600 (0%)	0/600 (0%)	0/600 (0%)	1/600 (0.16%)	1/600 (0.16%)	5/600 (0.8%)	0/600 (0%)	-	0/500 (0%)

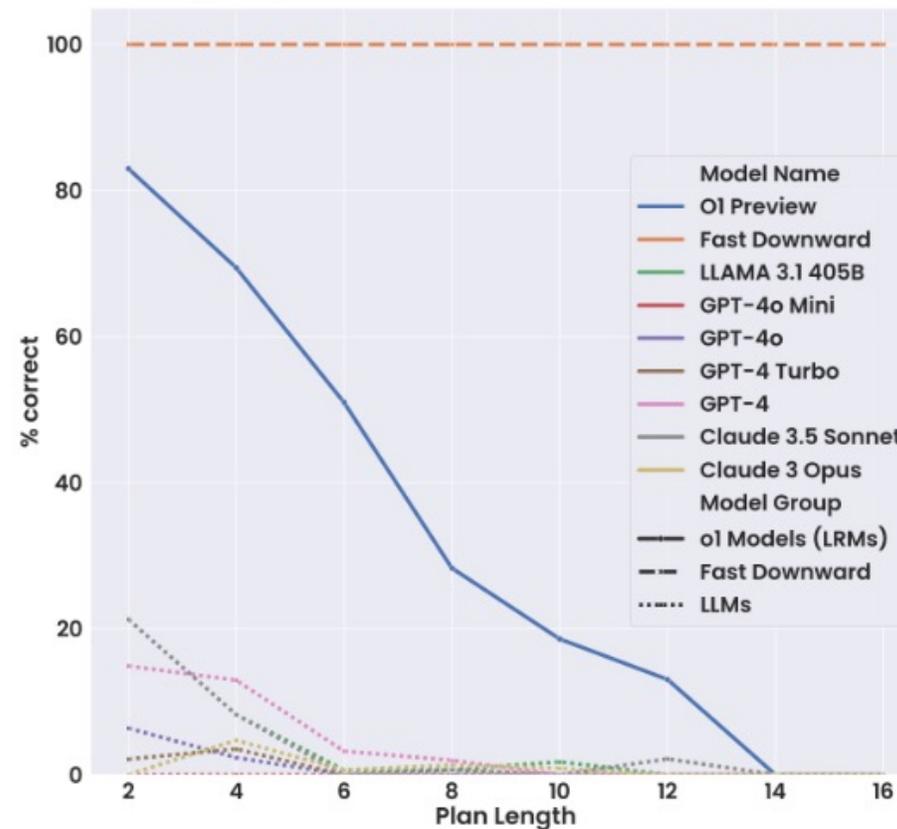
Table 1: Performance on 600 instances from the Blocksworld and Mystery Blocksworld domains across large language models from different families, using both zero-shot and one-shot prompts. Best-in-class accuracies are bolded.

o1 的规划能力怎么样?

Plan Length vs Correct Predictions for all Models on Mystery Blockworld - Plan Generation Zero Shot



Plan Length vs Correct Predictions for all Models on Mystery Blockworld - Plan Generation One Shot



全数据驱动的大模型 实现“推理”面临巨大挑战

3. 大模型未来发展思考

大模型未来两个路径

1. AGI (Follow OpenAI)

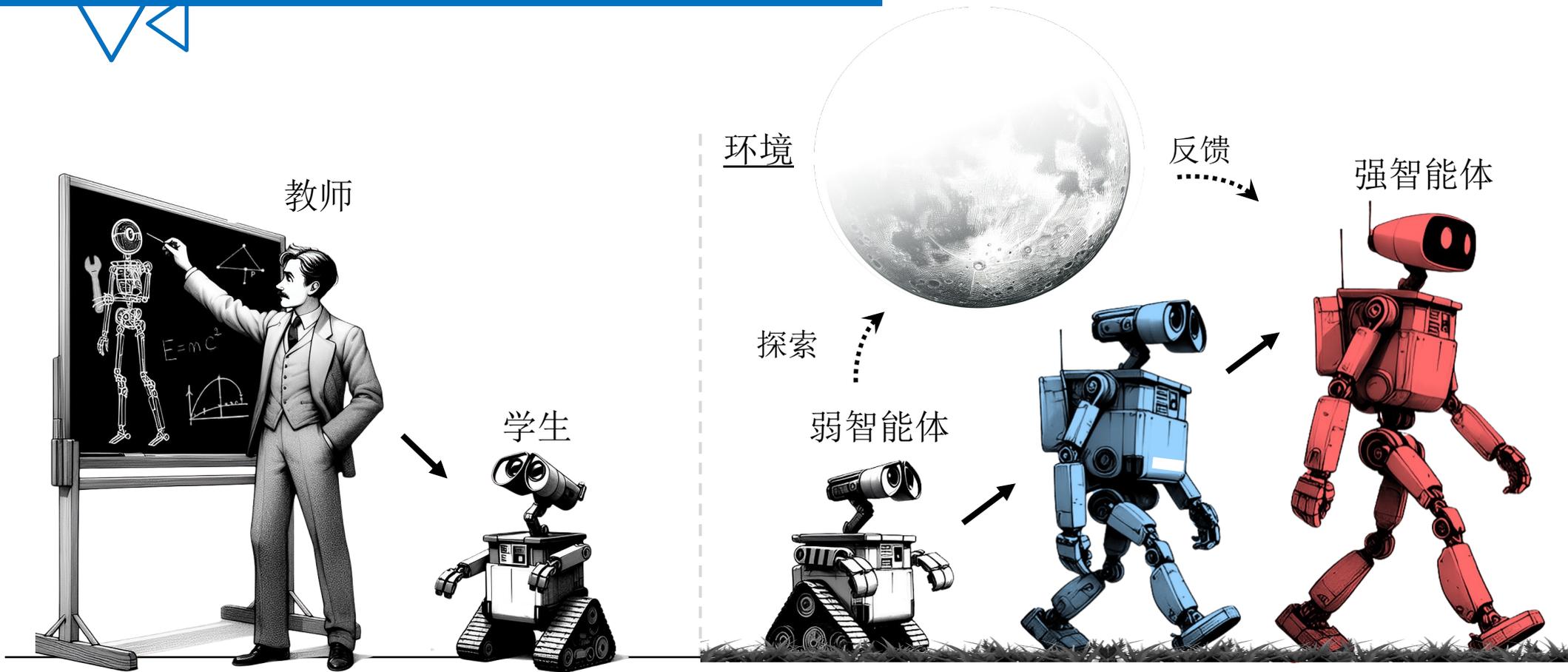
- 目标：代替人类完成**所有脑力劳动**，具有自我学习、自我进化能力。
- 参数规模： 1 万亿、5 万亿、10 万亿...
- 训练数据：10TB、20TB、40TB、100TB...
- GPU卡规模：1万、5 万、10 万 ... 100 万卡

基本假设： Scaling Law, 压缩即智能

难点：模型的推理和世界知识建模能力构建，资本消耗巨大



OpenAI – “超级对齐” 目标

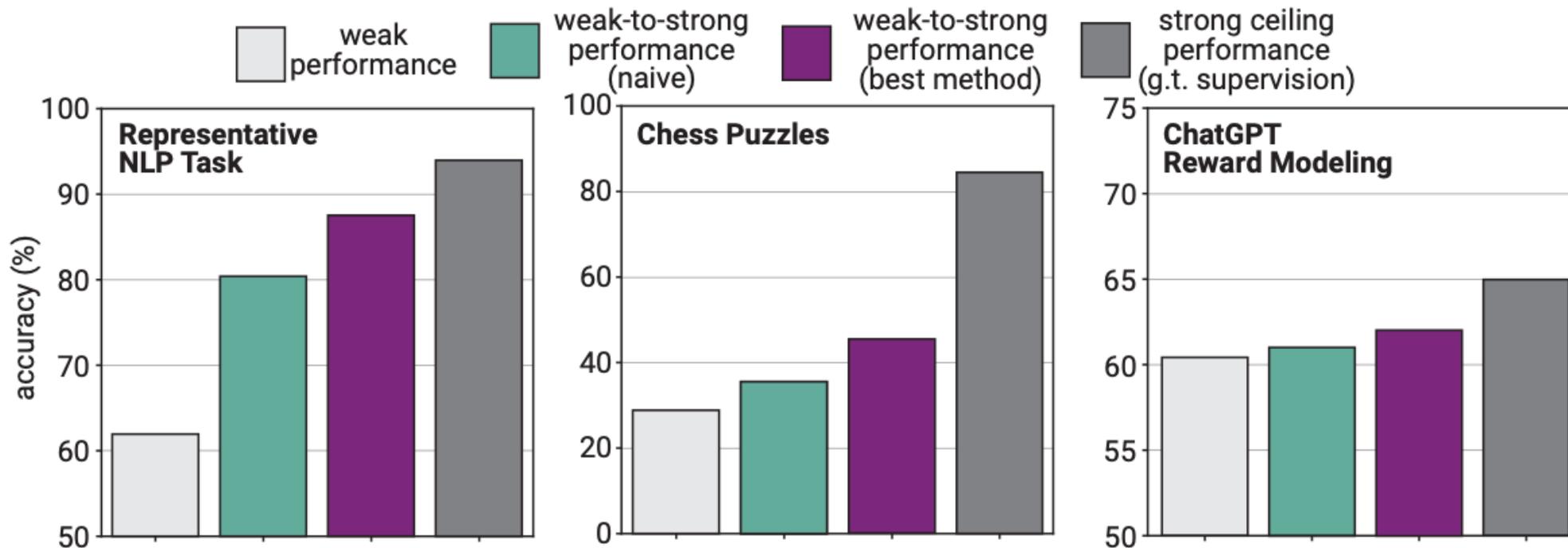


传统机器学习方法需要“老师”指导

“学生”永远超不过“老师”

基于环境反馈的智能体**自我进化**

OpenAI – “超级对齐” 仍然需要长期海量投入

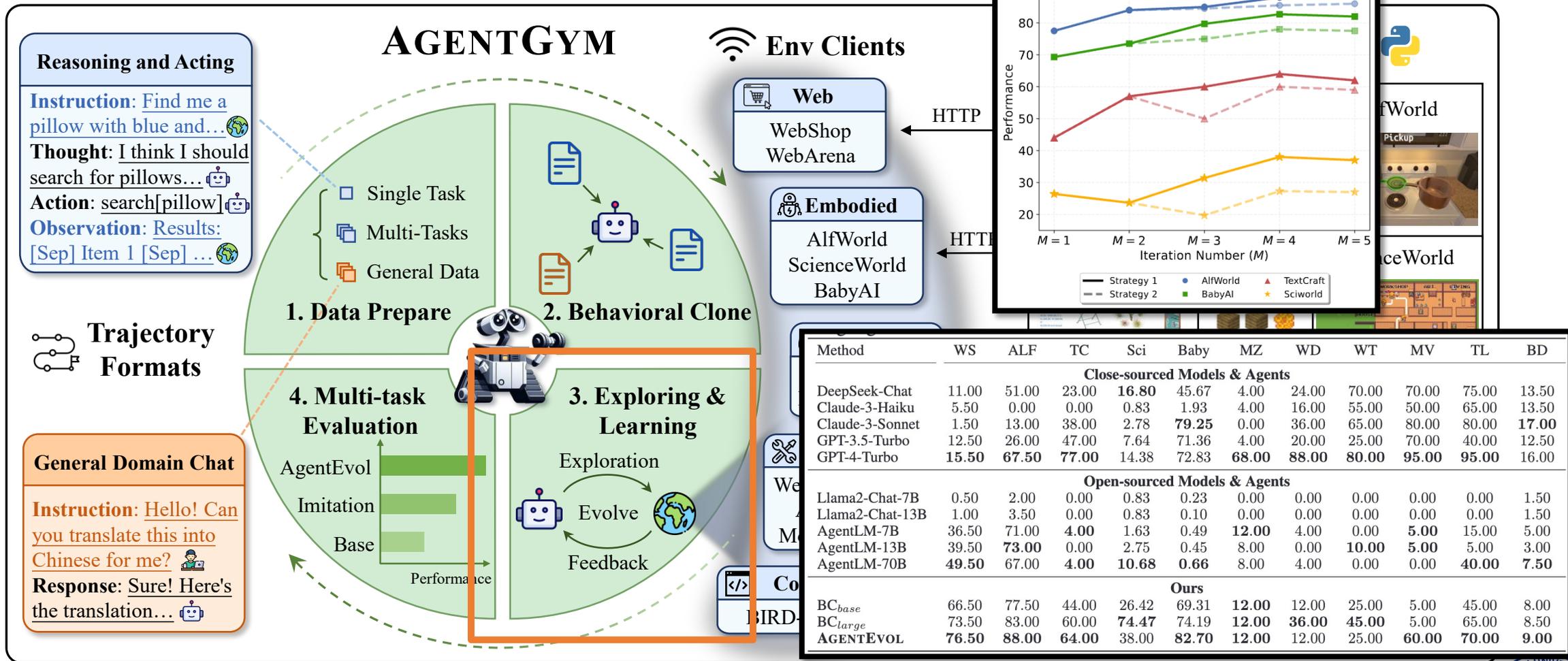


方法在所有设置下**都不是一致有效的**，尤其是在 RM 设置中，仍然远远没有恢复弱模型和强模型之间的全部性能差距。

占用了 OpenAI 接近 20%资源



AgentGYM: 交互式训练和评测框架



AGENTGYM: Evolving Large Language Model-based Agents across Diverse Environments

大模型未来两个路径

2. 应用

- 目标：代替人类完成**特定任务**
- 参数规模：1B、3B、7B、14B、30B、70B 模型为基础
- 训练数据：几千 到 几万条
- GPU卡规模：1到100卡

基本假设： 大模型任然是统计机器学习，规模再大也不能实现 AGI

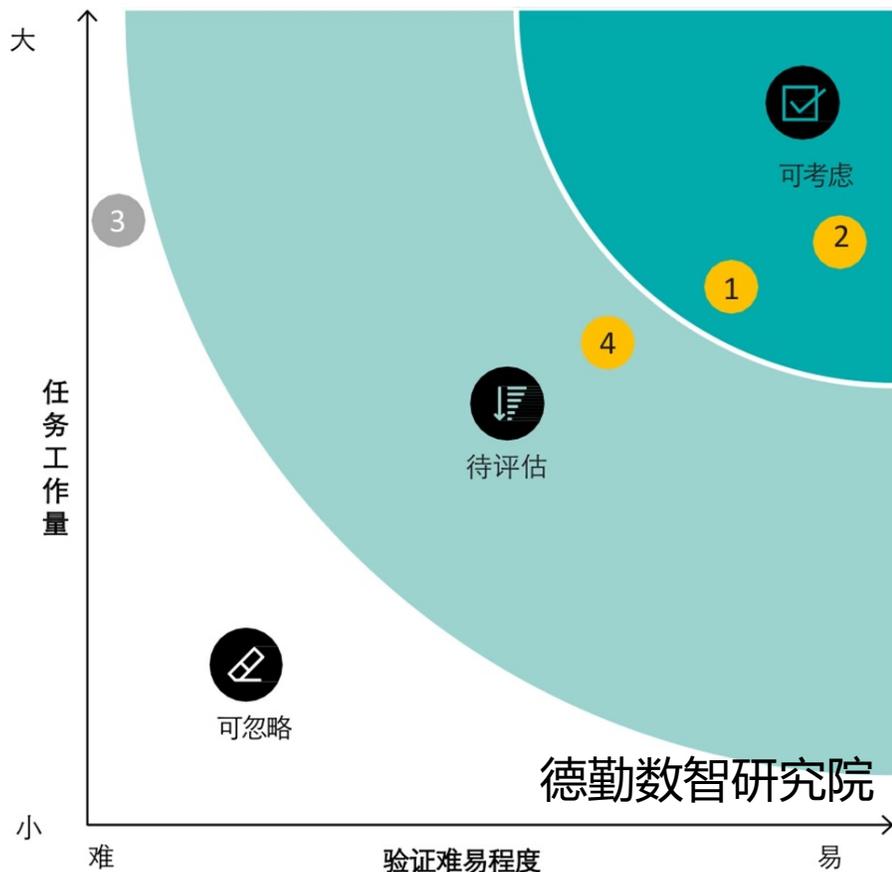
难点： 对模型能力边界的判断，场景选择要符合能力并具备价值



大模型落地条件

场景驱动

场景定制开发 数百万元 成本



示例

1 创作一个笑话

创作一个有趣的笑话需耗费精力来设计笑点和独特叙事风格，而验证笑话质量则相对容易，通过阅读笑话即可验证。

2 画一幅大象站在棕榈树下的图

对于大多数人来说，无论可用的工具有哪些，绘制任一复杂图像都需要付出合理的努力。然而，验证图像质量却相对容易，通过查看图片即可验证。

3 起草一份合同 (不具备法律专业知识)

如果不具备法律专业知识，起草一份合同就非常困难，而验证合同也同样困难。

4 起草一份合同 (具备法律专业知识)

如果具备法律专业知识，起草一份合同仍需耗费精力，但验证合同就明显更加容易。

来源：生成式AI火遍全球 - 发掘赋能时代价值潜力

- 在没有生成式AI的情况下完成任务所需的人力 - “**任务工作量 (无AI)**”
- 验证或核实生成式AI输出所需的努力 - “**验证难易程度**”



场景举例：设计资料审核

船厂设计资料繁多，涉及数万份规范

(1)船体说明书;

诉求：全自动审核

可以提供的数据：

使用大模型构建全自动审核任务是否合适？

(5)航速、稳性、舱容、干舷等估算书或计算书;

(6)机舱布置图;

此前人工审核结束

设计人员工作量并未大幅度节省

场景举例：设计资料审核

诉求：全自动审核

可以提供的数据：

- 设计资料
- 各类规范
- 此前人工审核结果



知识库问答

设计规范入库

设计人员快速检索定位规范内容和位置

焊接系数 表 1.4.4.2

项目	焊接系数	备注
1 一般结构(除下列另有规定外)		
水密或油密板材的周界	0.34	
非密性板材周界	0.13	
	0.10	
纵骨、肋骨、横梁和其他次要构件与外板、甲板或舱壁板	0.13	液舱内
	0.21	端部连接处
板格加强筋(即小扶强材)	0.10	
搭接焊缝	0.27	
纵骨为扁钢时, 对板	0.21	双面连续焊
2 货舱区船底结构		
(1) 密性的中桁材对平板龙骨	0.44	
对内底板	0.44	
(2) 非密性中桁材(或中内龙骨)对平板龙骨	0.27	
对内底板(或中内龙骨的面板)	0.21	不开孔
(3) 密性肋板、旁桁材的四周边界	0.39	
(4) 非密性肋板、旁桁材(或旁内龙骨)和肘板等的四周边界	0.16	
	0.21	在端部 0.2 跨距长度处
	0.27	主肋骨下端能肘板处
(5) 内底纵骨和内底骨材对内底板	0.13	在重货加强区域
(6) 支撑平面舱壁、槽形舱壁、双层板舱壁、舱壁凳的肋板对内底板	0.44	双面连续焊

2-19



非密性板材周界的焊接系数

非密性板材周界的焊接系数是0.13。

相关来源：

《国内航行海船...

表 1.4.4.1 焊接系数 # 2-



人工智能比其它任何领域都容易过度宣传



机器与人非常自然的对话并完成任务

Google 2018年吹的牛，现在也没实现

Shy Kids团队使用Sora进行了**数百次生成**，每次10到20秒，并最终只使用了一小部分 (**300:1**)

精确控制角色或相机的时机和动作实际上是不可能的：“在实际生成中，这些不同动作发生的时间有一点点时间控制，但并不精确.....这有点像是瞎猜，” --- Cederberg

**大模型基础理论、大模型能力边界的清晰认知
决定了投入力度和发展方向**



谢谢!