



大语言模型是实现AGI的必由之路吗？

张奇

复旦大学

这近乎是个哲学问题

**很多结论都可以用一个统一的问题
进行“反驳”**

什么是AGI

AGI (Artificial General Intelligence), 中文翻译为“通用人工智能”, 亦被称为强AI, 任何可以想象的人类的专业领域内, 具备相当于人类智慧程度的AI, 一个 AGI 可以执行任何人类可以完成的智力任务。



围棋AI 的 五子棋 能力为0

AGI 要具备的能力

AGI 要具备哪些能力呢

1. 语言：使用自然语言交流
2. 知识：常识知识、专业知识
3. 推理：演绎、归纳、溯因
4. 学习：根据历史进行自我学习

“推理能力” 是AGI的核心

推理能力包含哪些

演绎推理 (Deductive reasoning)：就是从一般性的前提出发，通过推导，得出具体陈述或个别结论的过程。

例如经典的三段论推理：

前提：人皆会死

前提：苏格拉底是人

结论：苏格拉底会死

推理能力包含哪些

归纳推理 (Inductive reasoning)：是论证的前提**支持结论但不确保结论**的推理过程。它基于对特殊的代表 (token) 的有限观察，把归结到类型；或基于对反复再现的现象的模式 (pattern) 的有限观察，公式表达规律。

例如：

观察：我们每次看到的有翅膀的动物都是鸟

观察：我们又看到了一个有翅膀的动物

结论：这个动物应该是也是鸟

推理能力包含哪些

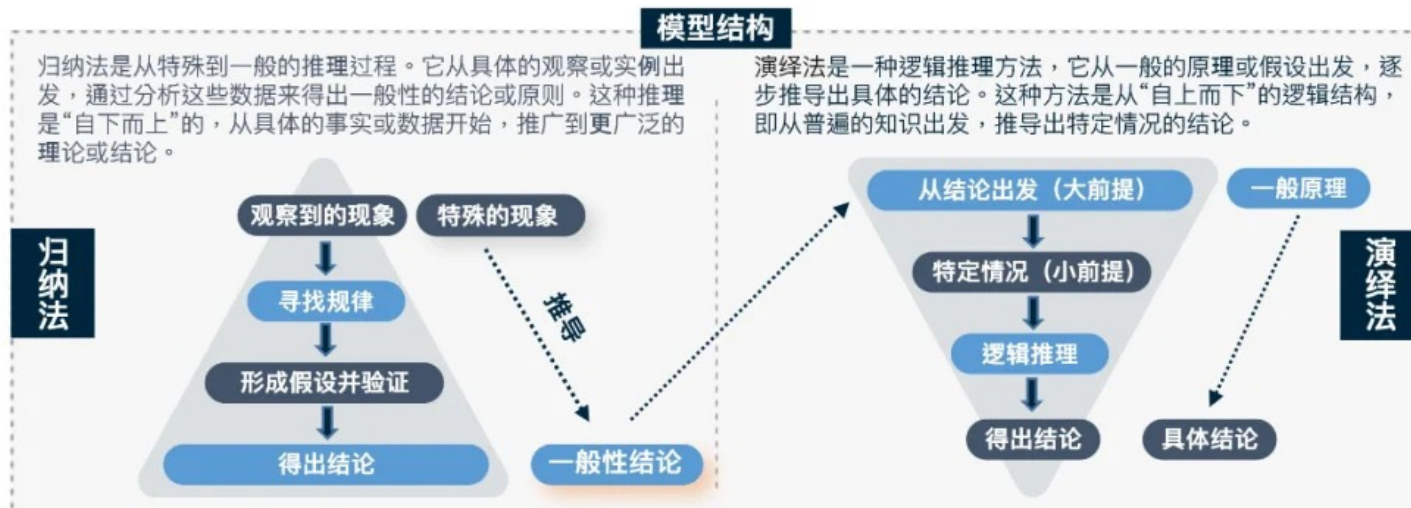
溯因推理 (Abductive reasoning, 也译作反绎推理)：是从事实推理到最佳解释的过程。换句话说，它是开始于事实的，并推导出其最佳解释的推理过程。

例如：

观察：这个车不能启动了，在发动机下看到了一滩液体。

结论：这个车不能启动的原因大概率是因为发动机有漏油。

归纳-演绎法在我们认识世界过程中，扮演重要作用



举个栗子

归纳法推理



边牧很快就能听懂命令
德牧很快就能找到藏起来的食物
金毛听到门铃声知道有人来访
经过这种观察，于是你得出所有的狗都很聪明的论。
直到遇到哈士奇。

归纳法是建立在有限的观察基础上，可能导致不完全或不准确的结论。因此，重要的是要认识到归纳推理的局限性，并在可能的情况下通过更多的证据和分析来支持我们的结论。

演绎法推理



“所有的鸟会飞”和“企鹅是鸟”是前提，“企鹅会飞”是结论。前面两个前提为真，那么“企鹅会飞”这个结论为真，反之这个结论为假。

演绎推理能力如何？

演绎推理有比较好的数学工具，目前大模型模型效果较好

QZ You

下雨路会湿，草坪也会湿；洒水器打开时草坪也会湿；阳光普照、天气又热时要打开洒水器给草浇水；阳光普照和下雨是不可能同时发生的。现在草坪湿了但是路没湿，现在的天气是什么？

 ChatGPT

根据描述，草坪湿了但是路没湿，这意味着洒水器被打开了，而且下雨没有发生。因此，当前的天气应该是阳光普照、天气炎热，需要用洒水器给草坪浇水。

QZ You

如果路湿了，但是草坪没湿呢

 ChatGPT

如果路湿了但草坪没湿，那么很可能是下过雨而草坪处于遮蔽下，没有被雨水直接淋湿。所以，当前的天气应该是下过雨，但是现在可能是晴天或者云开雾散，阳光普照，使得路面湿润，但草坪没有被雨水淋湿。

2023年5月版本并不能很好的解决演绎推理

QZ

A会导致B和C；D会导致C；在E条件下，会执行D；E和A不会同时出现，现在C出现了，B没有出现，是在A还是E条件



根据你提供的信息，我们无法确定C出现的条件是在A还是在E。因为A会导致B和C，D会导致C，在E条件下，会执行D，E和A不会同时出现，现在C出现了，B没有出现。这五个前提都只是告诉我们，C出现的原因可能是D。但是，这些前提并不能告诉我们C出现的条件是在A还是在E。

如果想要确定C出现的条件是在A还是在E，我们可能需要更多的信息。例如，如果我们能够确定E条件下D执行了，那么C出现的条件就是E。如果我们能够确定E条件下D没有执行，那么C出

通过强化学习，分步骤的推理评分，可以很好的解决演绎推理问题，但是仍然是当做单独的任务进行单独优化。

当前的版本复杂逻辑题仍存在问题



You

某地区发生了一个案件，警方在这次案件中，一共逮捕了3名嫌疑犯，他们的名字分别叫做：Devin、Felix和Grove。据警方的了解，这三名嫌疑犯非常的狡猾，所以，警方分别对他们进行了传讯。不过，在面对警方的询问时，他们三人都只回答了一句话（供词）：

- 1、Devin是无罪的。
- 2、Felix说的是真话
- 3、Grove说的是谎话

在这三人的供词中，供词1是最先提供的，而供词2和供词3则不一定是按照时间的先后顺序说的，但是它们都是在针对前面所做的供词，此外，经过警方的调查、取证，还发现了下面2条确切的线索。

- A、犯罪分子肯定是三人中的某一人，并且这个人是在说谎。
 - B、每个人所作的一条供词，都是针对另一个值得怀疑的对象的。
- 请问，你能否根据以上信息，推理出来谁才是真正的犯罪分子呢？



ChatGPT

根据提供的信息，我们可以进行逻辑推理来确定真正的犯罪分子。

1. 如果Devin是无罪的，那么Felix和Grove中至少有一个在说谎。
2. 如果Felix说的是真话，那么Grove说的也是真话，这与B条矛盾。
3. 所以，Felix说的一定是谎话，而Grove说的一定是真话。
4. 既然Felix在说谎，那么Devin就不是无罪的，所以Devin是犯罪分子。

综上所述，Devin是真正的犯罪分子。



大模型对于分布外数据集处理仍有待提高

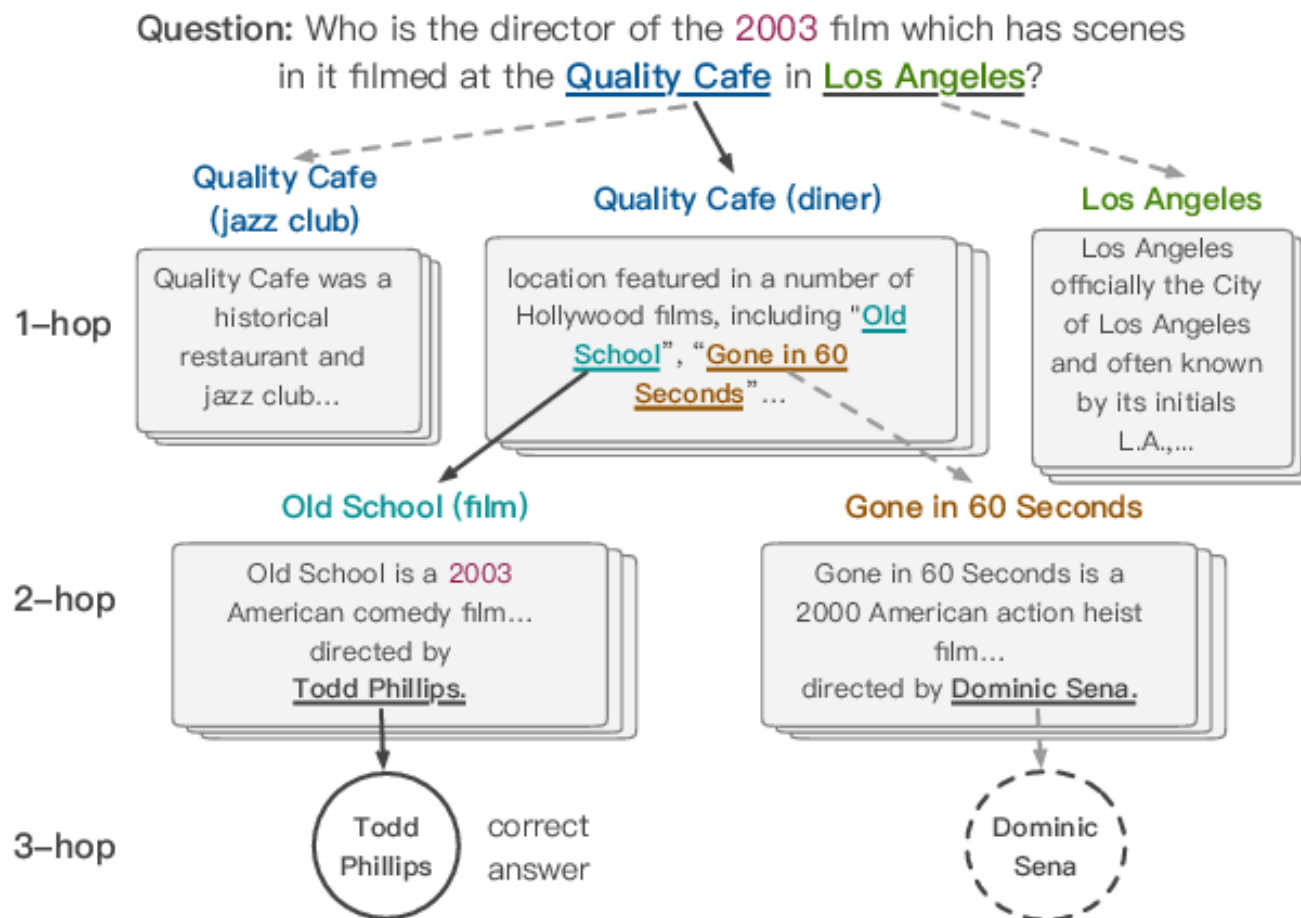
ChatGPT和GPT-4都擅长解决著名的逻辑推理阅读理解基准，但在处理分布外数据集方面很困难。它们在需要逻辑推理的自然语言推理任务上的性能仍有待提高。

| 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

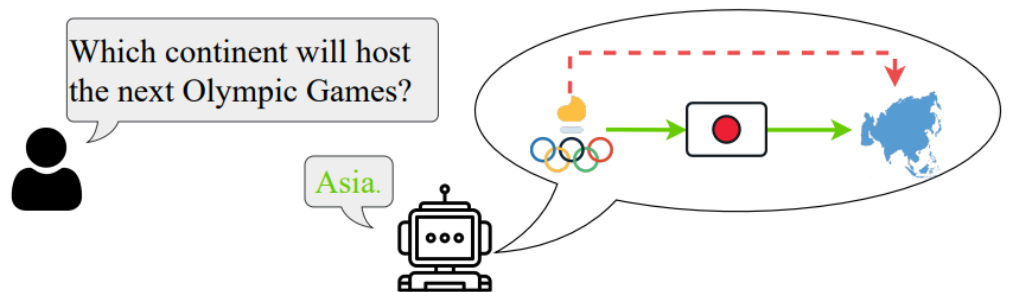
真实应用中的演绎推理怎么样？



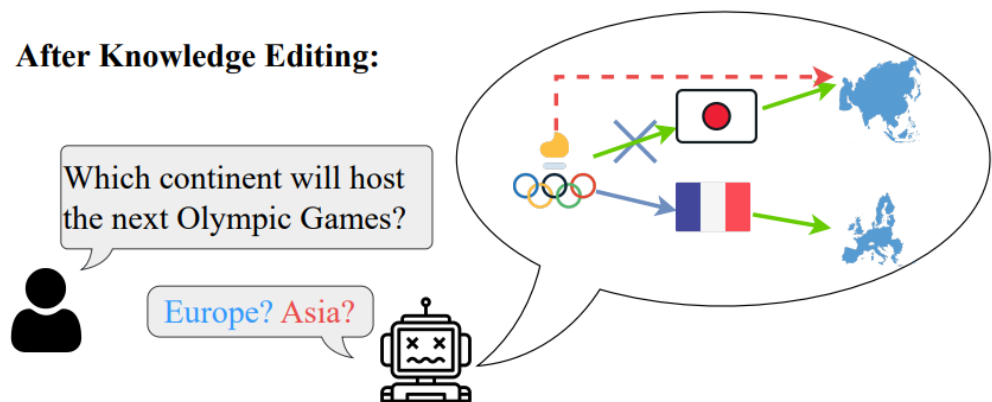
真实应用中的演绎推理怎么样？

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:



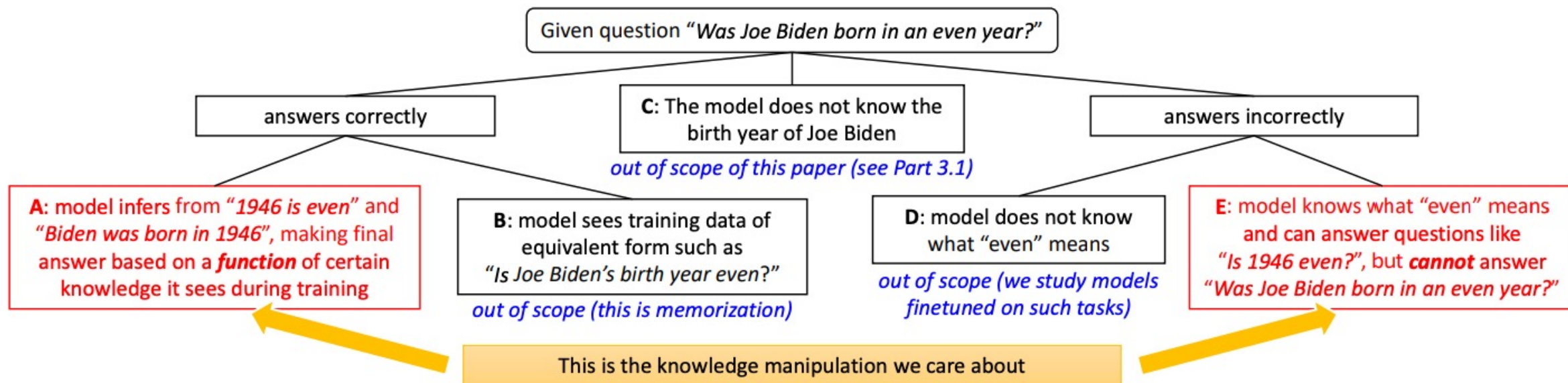
大模型在预训练阶段可能直接学习到下一届奥运会在亚洲举办这一跨步知识，当采用知识编辑方法将下一届奥运会举办国由日本改为法国后，模型可能仍然认为下一届奥运会在亚洲举办。

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>?"

GPT3.5 acc = 23.9%

GPT4: 42%

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

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

演绎推理做为单独任务可以优化

演绎推理能力在其他任务中并不能自动获得

是否具备归纳推理能力需要仔细思考

归纳推理能力如何？

归纳推理的能力如何呢？

乘法的步骤：

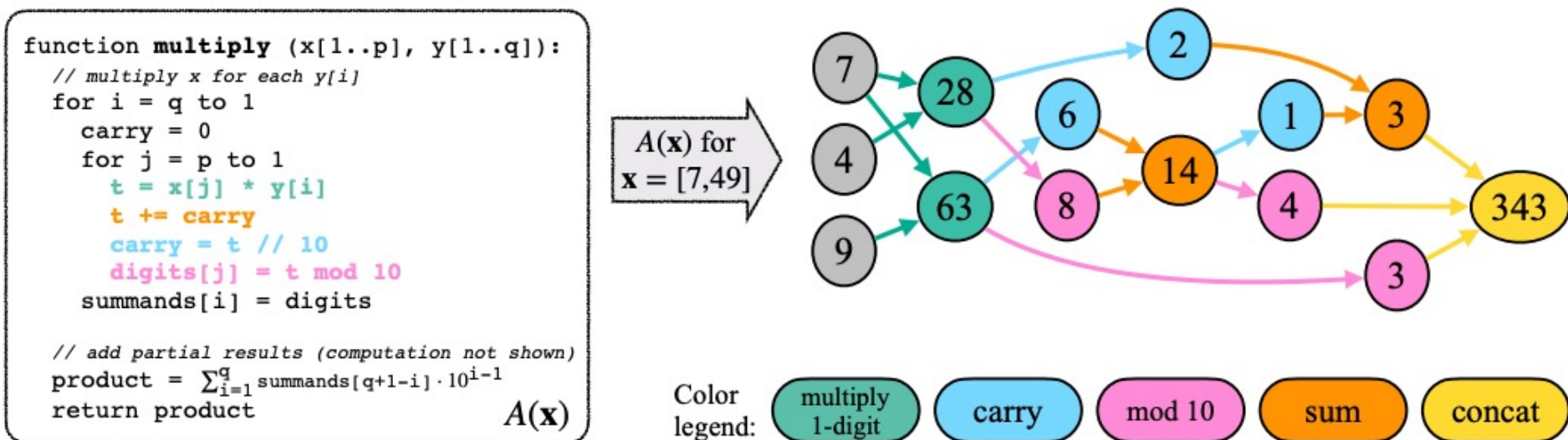


Figure 1: Transformation of an algorithm A to its computational graph $G_{A(\mathbf{x})}$. The depicted example is of long-form multiplication algorithm A , for inputs $\mathbf{x} = [7, 49]$ (i.e. computing 7×49).

当任务复杂程度增大时，模型的准确率接近为0

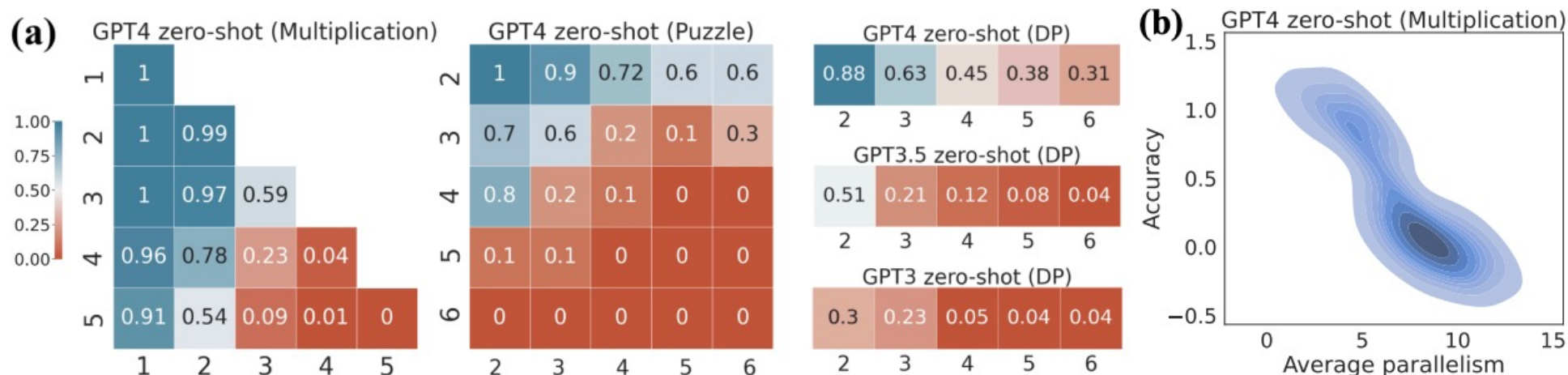
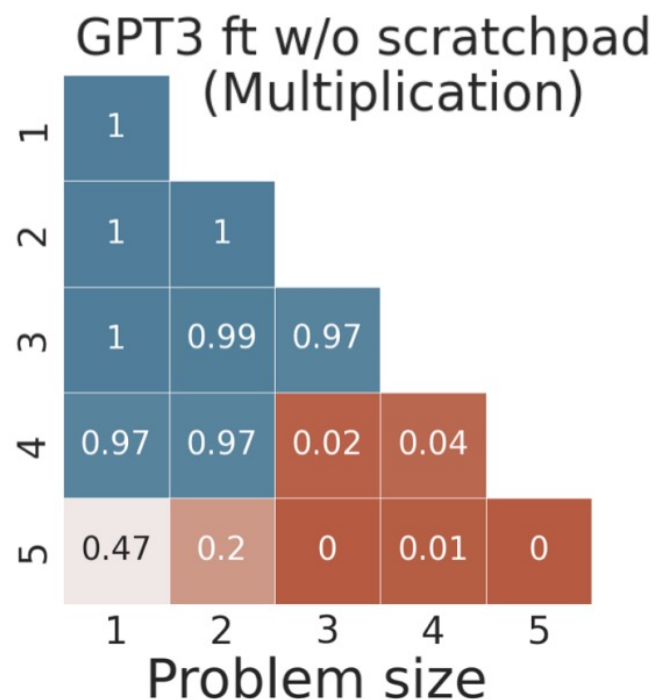


Figure 2: (a) **Zero-shot accuracy**. Axes refer to problem sizes (number of digits in multiplication, number of houses and attributes in puzzle, and sequence length in the DP task). Transformers' accuracy decreases to near zero as task complexity increases, measuring task complexity by the problem size. (b) **Average parallelism** negatively correlates with accuracy.

归纳推理的能力如何呢？



GPT3 models on **~1.8M** multiplication pairs

GPT3 finetuned exhaustively on task-specific data up to a certain problem size. The blue region represents the in-distribution examples and the red region refers to OOD examples.

They consider all k_1 -by- k_2 digit multiplications with $1 \leq k_1, k_2 \leq 4$ and $k_1 \cdot k_2 \leq 9$;

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:

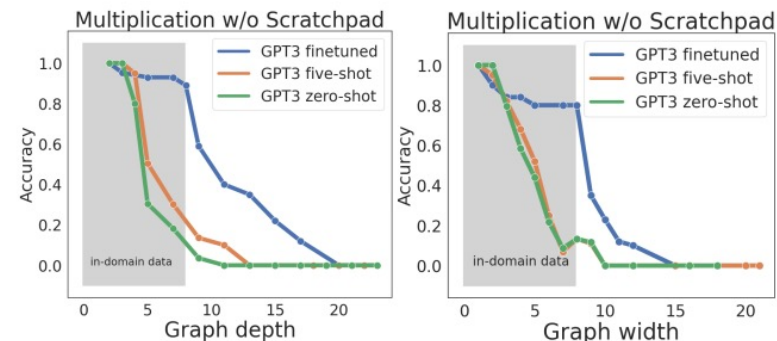
Let's multiply 35 by the digit in the ones place of 90, which is 0.

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.

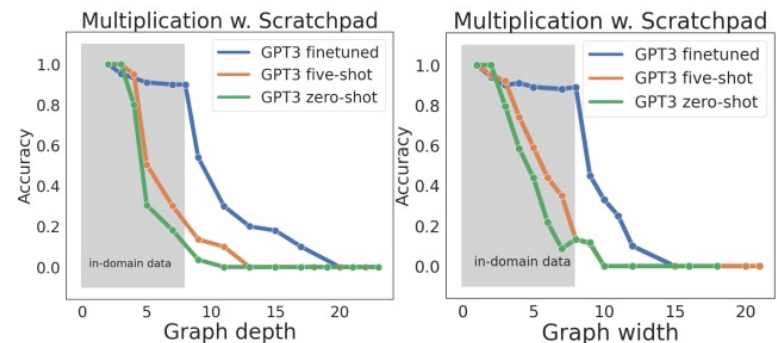
Now, let's multiply 35 by the digit in the tens place of 90, which is 9.

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.

Figure 5: GPT3 finetuning and prompting accuracy on different data splits. Although the in-distribution performance is almost perfect, GPT3 exhibits poor generalization with increasing graph depth and width. Refer to §B.3 and §B.4 for results on the puzzle and DP tasks.

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

数学计算评测

- Euler Equation.
- Add & Subtract of two integers within 10.
- Add & Subtract of two integers within 100.
- Add & Subtract of two integers within 1,000.
- Add & Subtract of two integers within 1,000,000,000,000.
- Add & Subtract of two integers within -10~10.
- Add & Subtract of two decimal numbers within -100~100.
- Multiply two integers within 100.
- Multiply two decimal numbers within 10.
- Multiply two integers within 100,000.
- Division of two integers within 100.
- Exponentiation of with integer base within 10 and integer exponent within 2~4.
- Exponentiation of with a decimal number within 10 as the base and a decimal number within 2~4 as the exponent.
- Add, Subtract & Multiply with one integer within 10 and a common irrational number (i.e. e or π).
- Long arithmetic expressions with brackets, involved integers are all within 100 and operators contain add, subtract, multiply, and division.
- Trigonometry functions including sin, cos, and tan. Inputs can be in the format of degrees and radians (π can also appear in the inputs).
- Logarithm of integers within 1000 of different bases: 2, e , 10.

| Model | Prompt | Acc \uparrow | RE \downarrow | NNR \downarrow |
|--------------------|-------------|----------------|-----------------|------------------|
| gpt-4 | Cal*4 | 83.54 | 0.07 | 0.00 |
| gpt-3.5-turbo-0301 | Cal* | 75.06 | 0.14 | 0.50 |
| text-davinci-003 | Cal | 56.61 | 0.76 | 2.99 |
| text-davinci-002 | Cal | 42.89 | 2.13 | 15.96 |
| text-curie-001 | Cal | 11.47 | 1.92 | 6.48 |
| text-babbage-001 | Eqa | 5.24 | 2.59 | 5.74 |
| code-davinci-002 | Eqa | 21.70 | 2.39 | 11.47 |
| galactica-120b | Eqa | 45.14 | 1.30 | 3.99 |
| galactica-30b | Eqa | 45.14 | 0.69 | 1.75 |
| galactica-6.7b | Cal | 34.41 | 2.61 | 8.73 |
| llama-65b | Eqa | 28.43 | 1.61 | 4.74 |
| llama-30b | Eqa | 30.17 | 1.72 | 3.74 |
| llama-13b | \$ | 27.68 | 2.40 | 9.73 |
| llama-7b | \$\$ | 21.95 | 2.11 | 7.48 |
| opt-175b | Cal | 21.70 | 3.18 | 21.70 |
| opt-66b | \emptyset | 20.70 | 2.66 | 18.70 |
| opt-impl-max-30b | Cal | 17.46 | 1.52 | 6.23 |
| opt-30b | \emptyset | 15.96 | 2.28 | 11.22 |
| opt-13b | \emptyset | 15.21 | 2.19 | 10.97 |
| opt-6.7b | Cal | 14.46 | 1.46 | 4.24 |
| gpt-neox-20b | Eqa | 35.41 | 1.19 | 4.49 |
| gpt-j-6b | Cal | 27.18 | 1.55 | 8.98 |
| bloomz-176b | \$\$ | 22.44 | 1.50 | 4.74 |
| bloom-176b | \$ | 20.2 | 2.60 | 18.45 |
| bloomz-7b1 | \$ | 12.72 | 2.56 | 15.46 |
| bloom-7b1 | Cal | 7.23 | 2.41 | 6.48 |
| bloomz-3b | \$\$ | 7.98 | 2.63 | 12.47 |
| bloom-3b | Cal | 4.24 | 2.41 | 8.73 |
| bloomz-1b7 | Eqa | 4.74 | 4.28 | 31.17 |
| bloom-1b7 | Cal | 5.24 | 2.54 | 11.22 |
| T0++-11b | Cal | 4.24 | 3.34 | 9.48 |

**如果不能解决推理问题
大模型成为AGI是不可能的**

现阶段降低对推理的诉求

可以确认的大模型能力

长上下文建模、更好的语义表示、多任务学习、跨语言迁移性



谢谢!