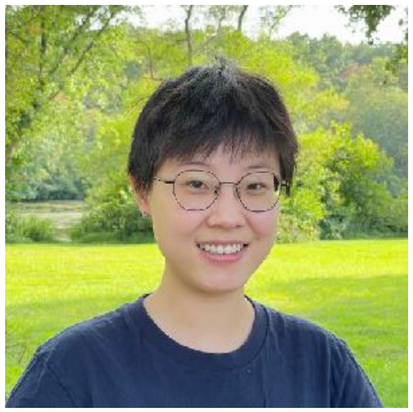
在合成数据上继续预训练：突破真实数据的局限

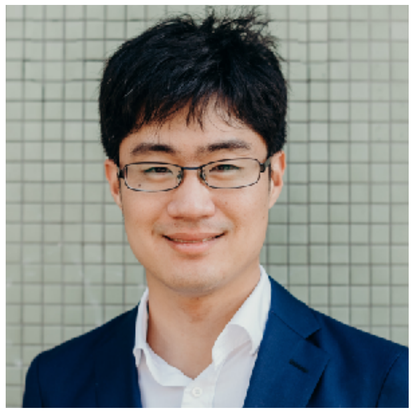
杨紫童
Zitong Yang

斯坦福大学


Neil Band*


李双平
Shuangping Li


Emmanuel Candès


橋本龍範
Tatsunori Hashimoto

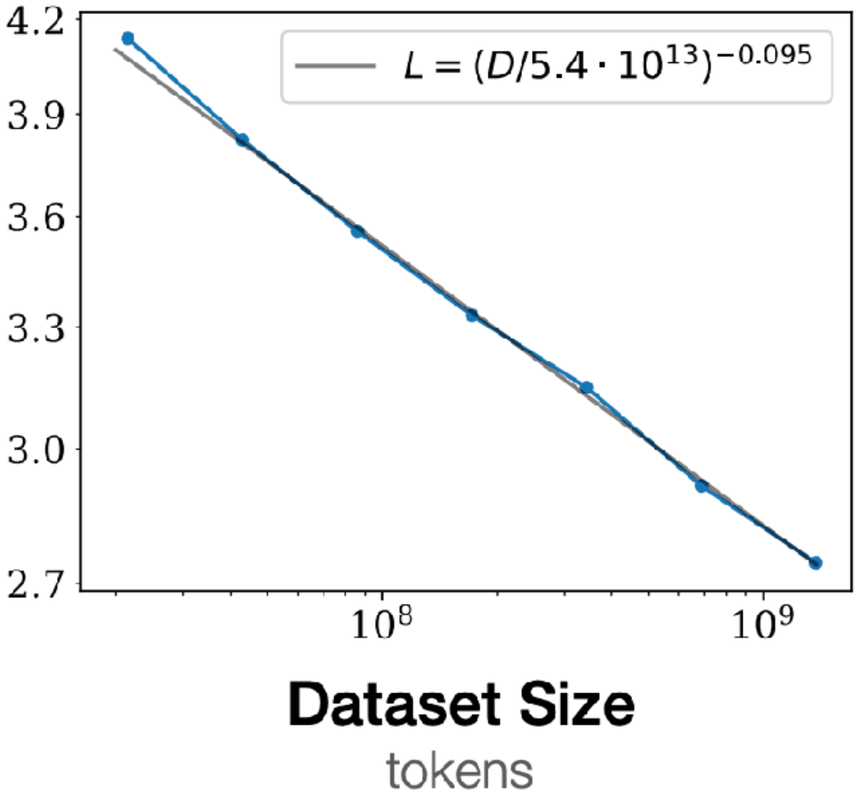
https://arxiv.org/pdf/2409.07431

大型语言模型

通过在大量互联网文本上进行预训练，模型拥有了丰富的世界知识。

大型语言模型

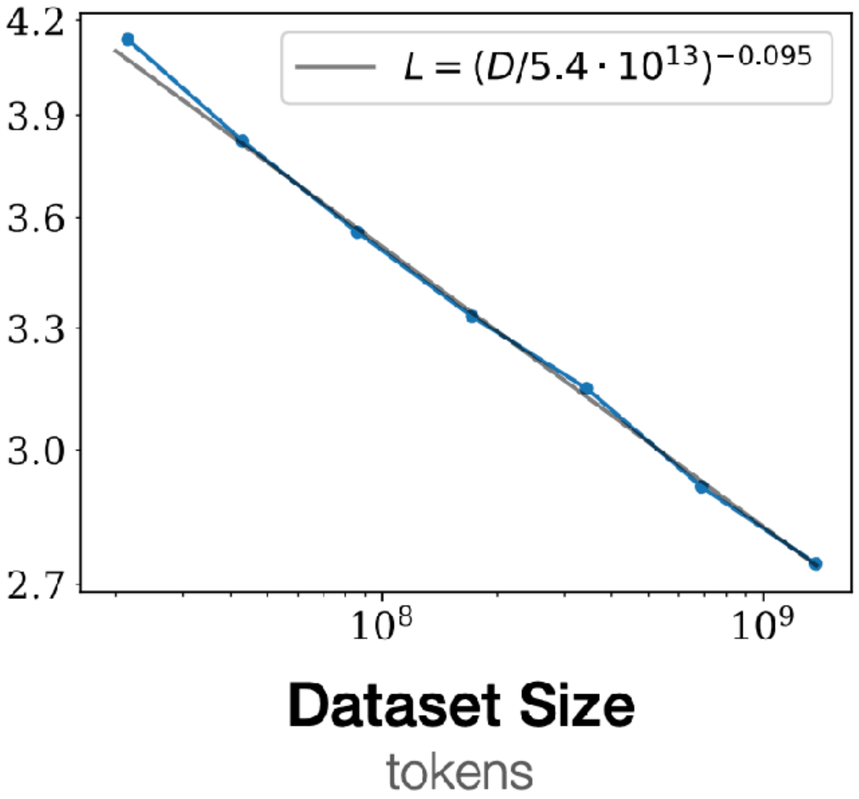
通过在大量互联网文本上进行预训练，模型拥有了丰富的世界知识。

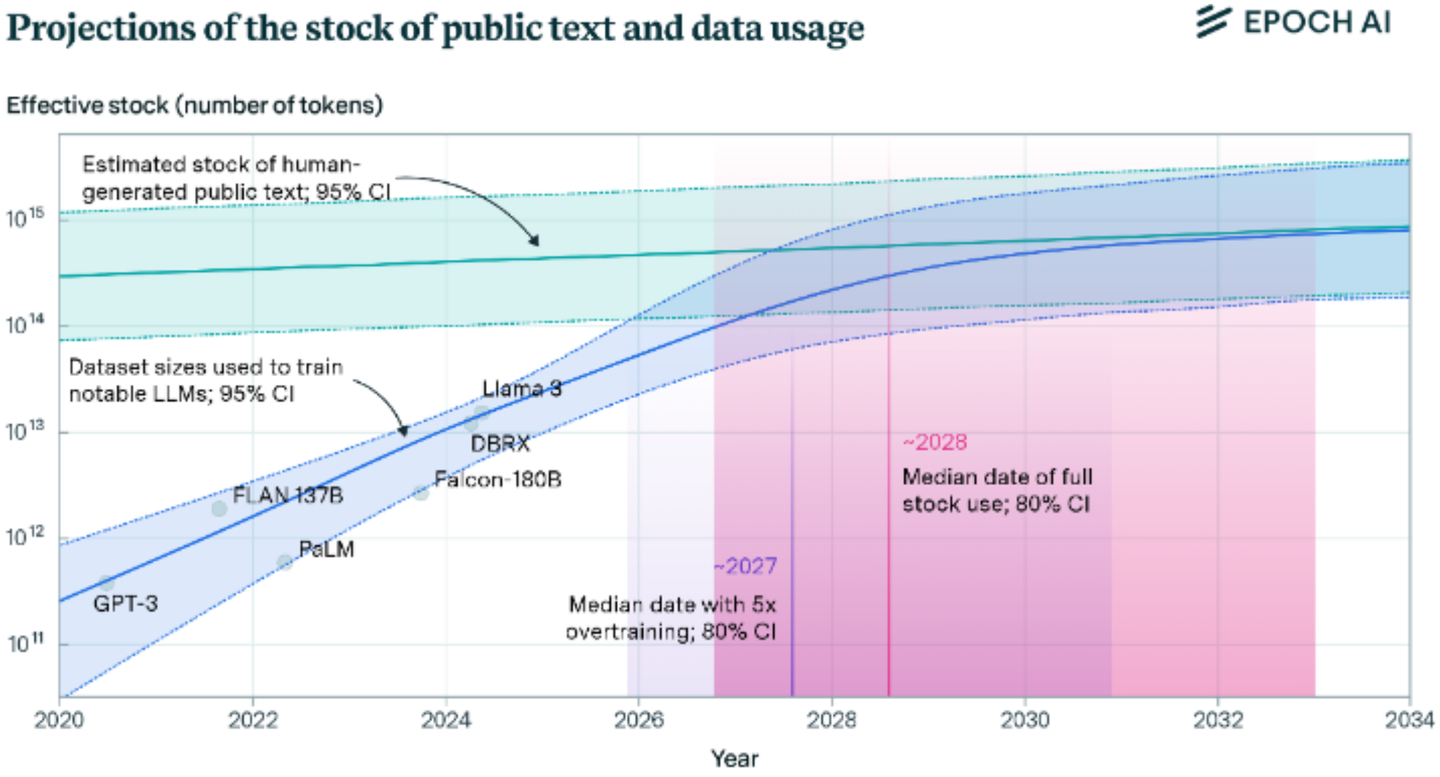


- 规模定律 (Scaling Law)：模型的能力被预训练的数据量所影响。

大型语言模型

通过在大量互联网文本上进行预训练，模型拥有了丰富的世界知识。

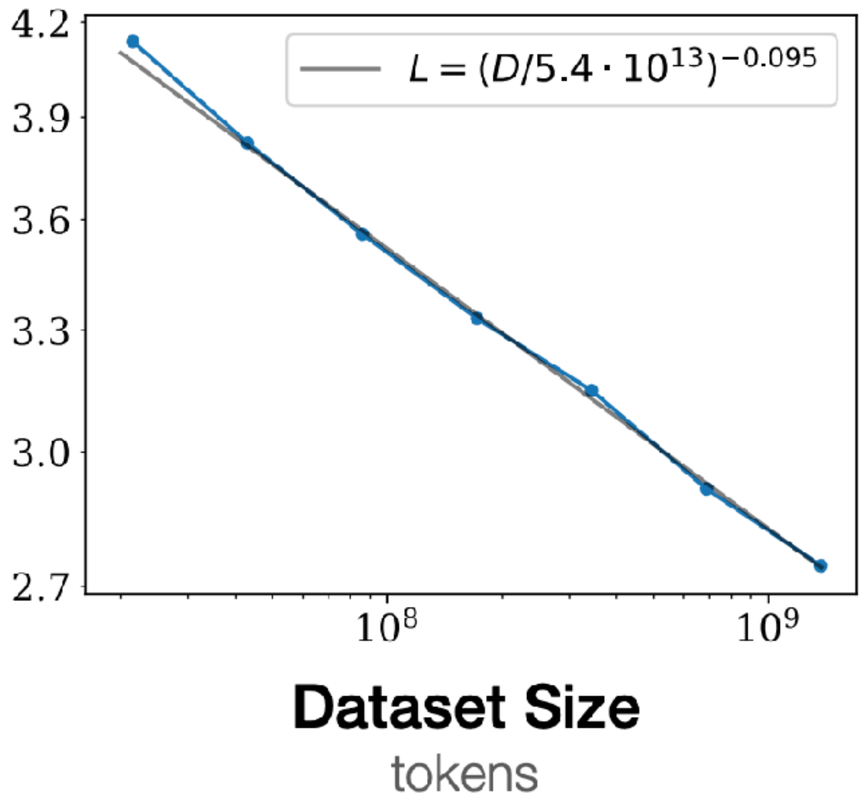


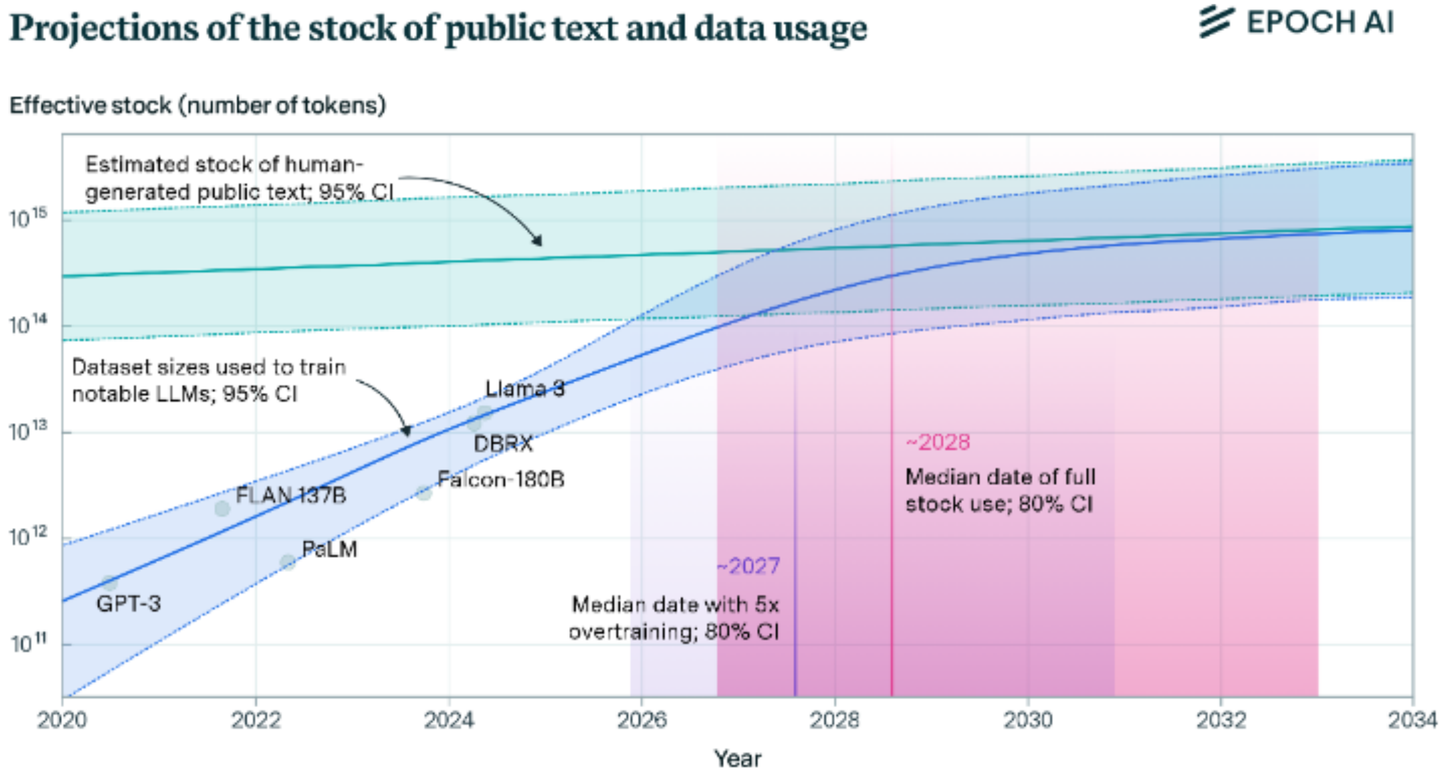


- 规模定律 (Scaling Law)：模型的能力被预训练的数据量所影响。
- 预计在2028年，最前沿的语言模型将耗尽所有公开的互联网文本。

大型语言模型

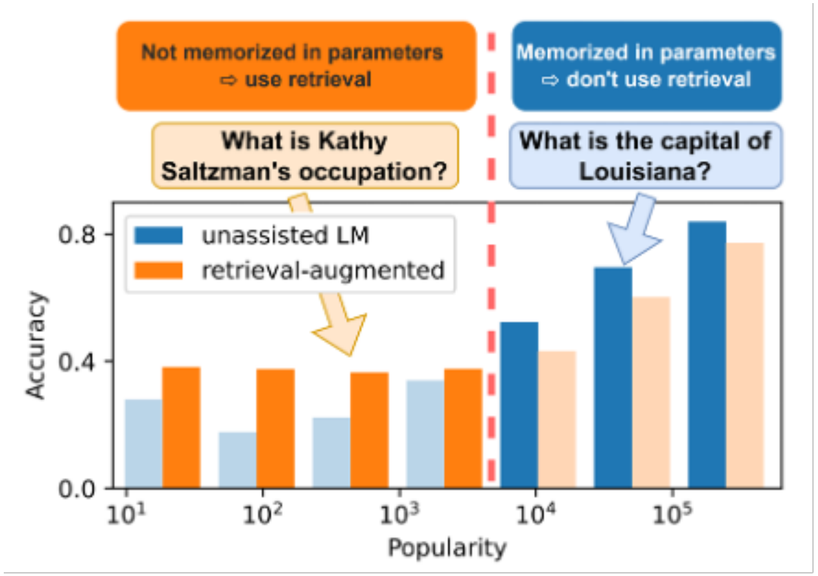
通过在大量互联网文本上进行预训练，模型拥有了丰富的世界知识。





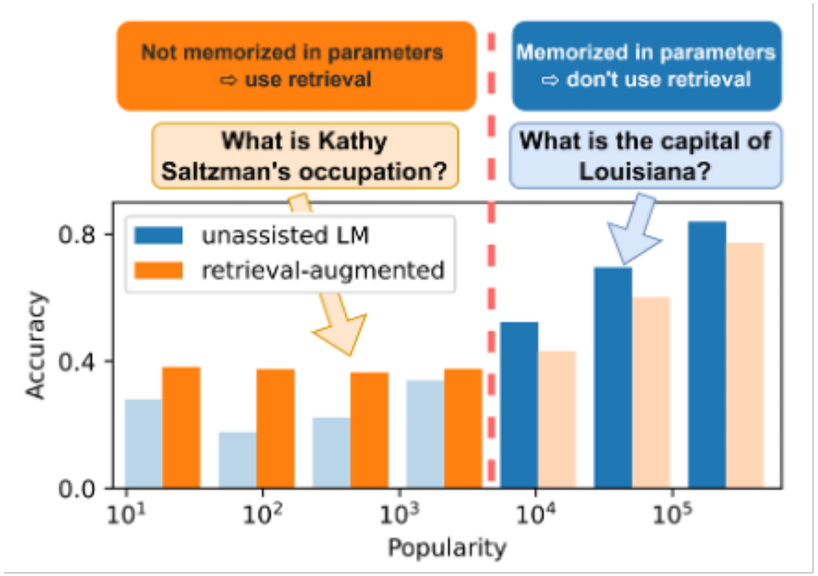
- 规模定律 (Scaling Law)：模型的能力被预训练的数据量所影响。
- 预计在2028年，最前沿的语言模型将耗尽所有公开的互联网文本。
- 如何在那之后继续提升模型的能力？

即使在现在，很多小领域也没有大量互联网数据记载



在预训练数据上不常见的知识

即使在现在，很多小领域也没有大量互联网数据记载



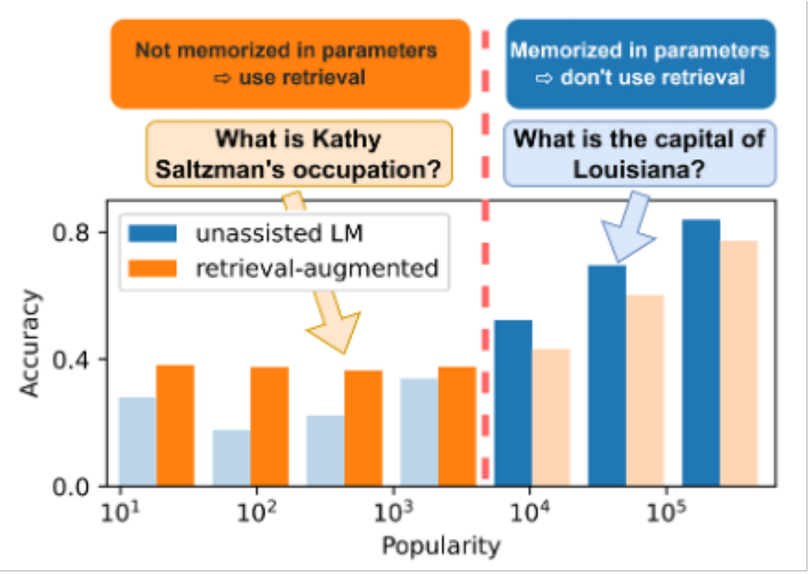
在预训练数据上不常见的知识

GPQA: A Graduate-Level Google-Proof Q&A Benchmark

David Rein[1,2] Betty Li Hou[1] Asa Cooper Stickland[1]
Jackson Petty[1] Richard Yuanzhe Pang[1] Julien Dirani[1]
Julian Michael[†1] Samuel R. Bowman[†1,3]
[1]New York University [2]Cohere [3]Anthropic, PBC

博士生级别的前沿问题

即使在现在，很多小领域也没有大量互联网数据记载

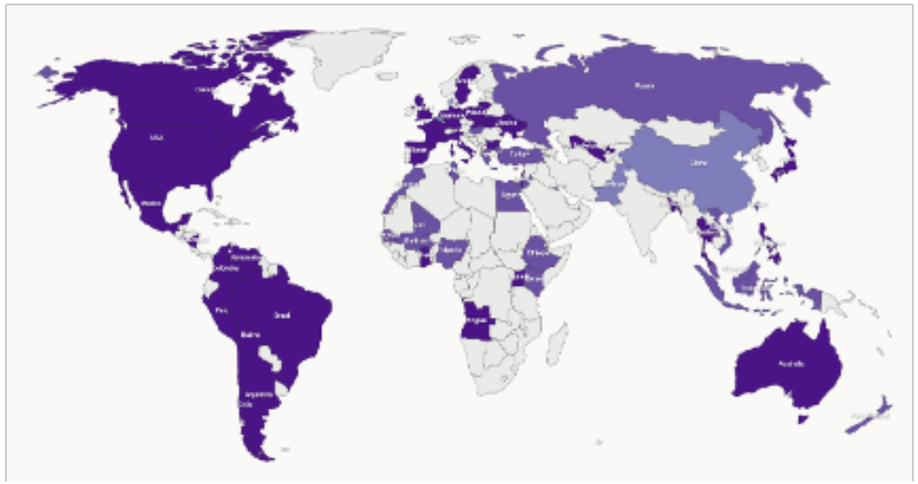


在预训练数据上不常见的知识

GPQA: A Graduate-Level Google-Proof Q&A Benchmark

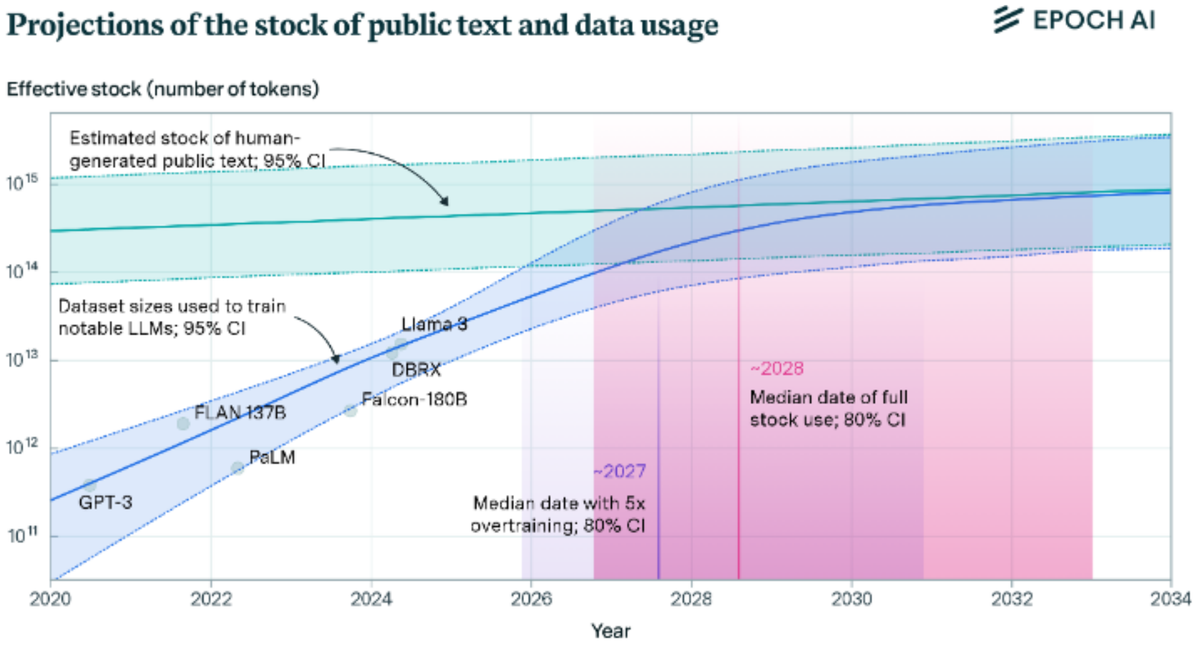
David Rein[1,2] Betty Li Hou[1] Asa Cooper Stickland[1]
Jackson Petty[1] Richard Yuanzhe Pang[1] Julien Dirani[1]
Julian Michael[†1] Samuel R. Bowman[†1,3]
[1]New York University [2]Cohere [3]Anthropic, PBC

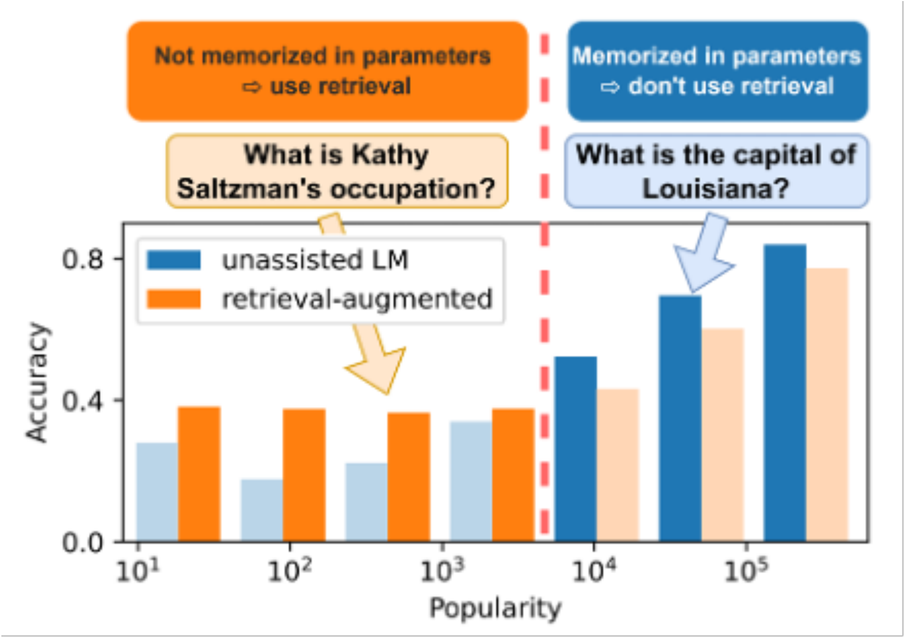
博士生级别的前沿问题



没有互联网主流语言记载的小领域知识

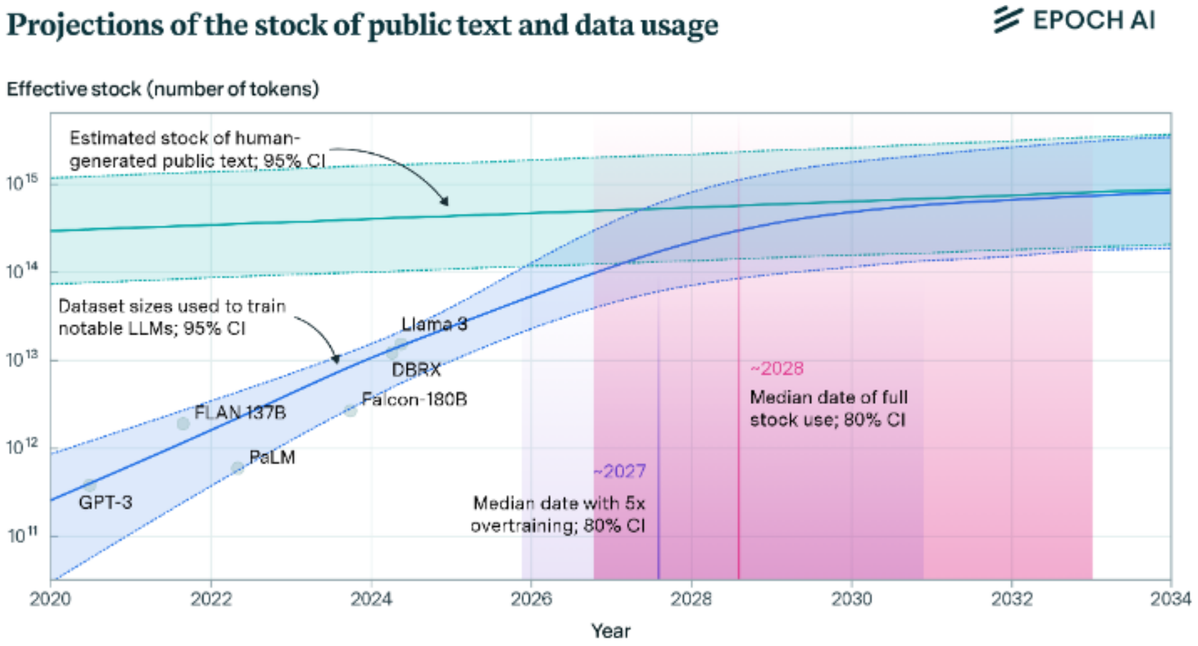
问题：如何提升利用数据的效率

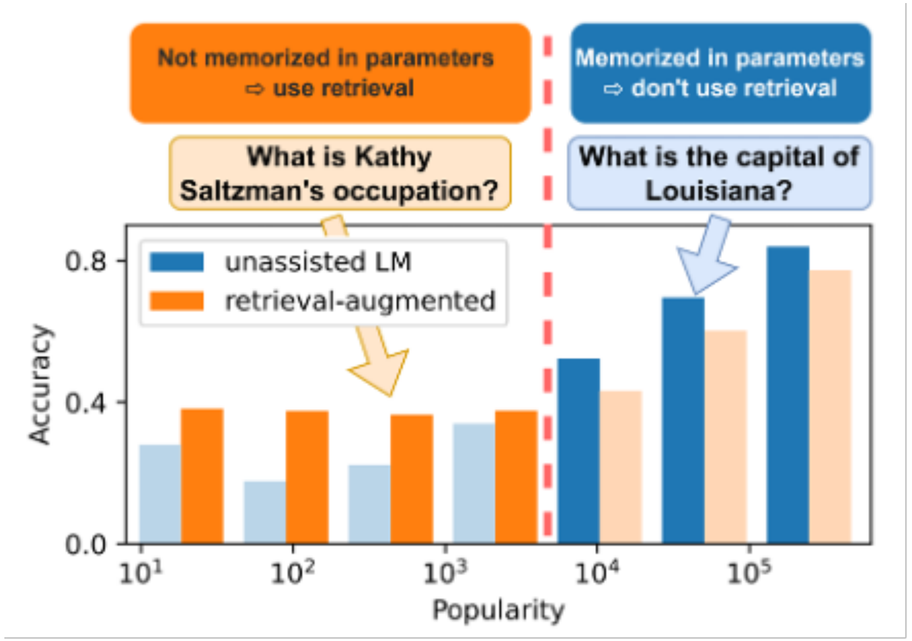




无论是面临耗尽与训练数据的挑战，亦或是学习数据非常稀缺的小领域，一个共同的问题是：

问题：如何提升利用数据的效率





无论是面临耗尽与训练数据的挑战，亦或是学习数据非常稀缺的小领域，一个共同的问题是：

如何更高效的利用已有的真实数据？

场景：根据少量文本数据学习小领域知识

场景：根据少量文本数据学习小领域知识

第一个简单的尝试是 Continued pretraining（继续预训练），也就是在领域相关的文本数据里预测下一个单词。

场景：根据少量文本数据学习小领域知识

第一个简单的尝试是 Continued pretraining（继续预训练），也就是在领域相关的文本数据里预测下一个单词。但这往往需要我们有关于目标领域的大量数据。

Study	Domain	Model Parameter Count	Total Unique CPT Tokens
Minerva (Lewkowycz et al., 2022)	STEM	8B, 62B, 540B	26B-38.5B
MediTron (Chen et al., 2023)	Medicine	7B, 70B	46.7B
Code Llama (Rozière et al., 2024)	Code	7B, 13B, 34B	520B-620B
Llemma (Azerbayev et al., 2024)	Math	7B, 34B	50B-55B
DeepSeekMath (Shao et al., 2024)	Math	7B	500B
SaulLM-7B (Colombo et al., 2024b)	Law	7B	30B
SaulLM-{54, 141}B (Colombo et al., 2024a)	Law	54B, 141B	520B
HEAL (Yuan et al., 2024a)	Medicine	13B	14.9B
Our setting	Articles & Books	7B	1.3M

- 已有的工作通过继续预训练让模型学会了医疗、数学、法律等领域知识。
- 这些领域不仅有大量的数据，并且这些数据的形式非常丰富。

场景：根据少量文本数据学习小领域知识

第一个简单的尝试是 Continued pretraining（继续预训练），也就是在领域相关的文本数据里预测下一个单词。但这往往需要我们有关于目标领域的大量数据。

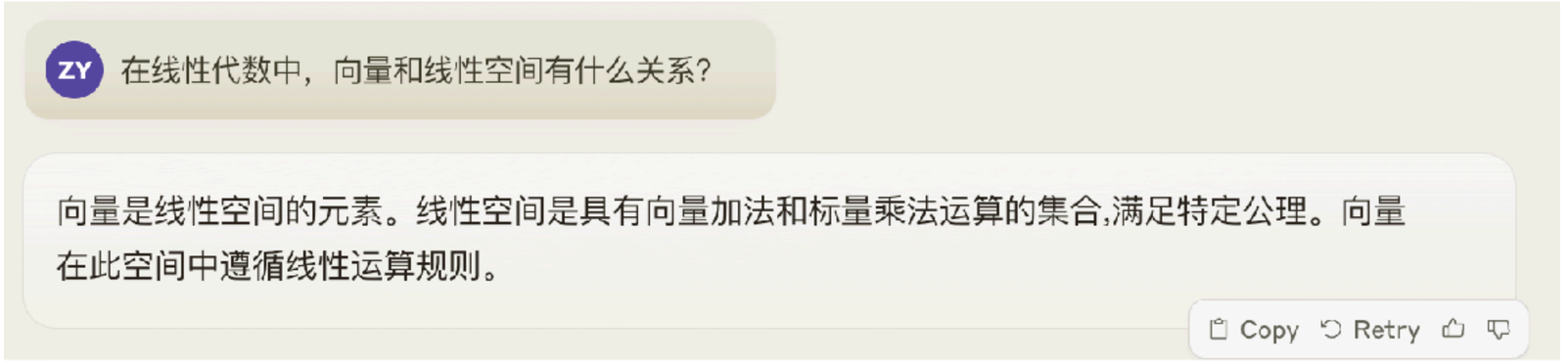
Study	Domain	Model Parameter Count	Total Unique CPT Tokens
Minerva (Lewkowycz et al., 2022)	STEM	8B, 62B, 540B	26B-38.5B
MediTron (Chen et al., 2023)	Medicine	7B, 70B	46.7B
Code Llama (Rozière et al., 2024)	Code	7B, 13B, 34B	520B-620B
Llemma (Azerbayev et al., 2024)	Math	7B, 34B	50B-55B
DeepSeekMath (Shao et al., 2024)	Math	7B	500B
SaulLM-7B (Colombo et al., 2024b)	Law	7B	30B
SaulLM-{54, 141}B (Colombo et al., 2024a)	Law	54B, 141B	520B
HEAL (Yuan et al., 2024a)	Medicine	13B	14.9B
Our setting	Articles & Books	7B	1.3M

- 已有的工作通过继续预训练让模型学会了医疗、数学、法律等领域知识。
- 这些领域不仅有大量的数据，并且这些数据的形式非常丰富。

如何让语言模型从小文本里学习到新知识？比如几篇最新的arXiv论文、公司内部的文件、用户过去几十年的个人信息。这些文本库往往只有1M-10M 左右的token数。

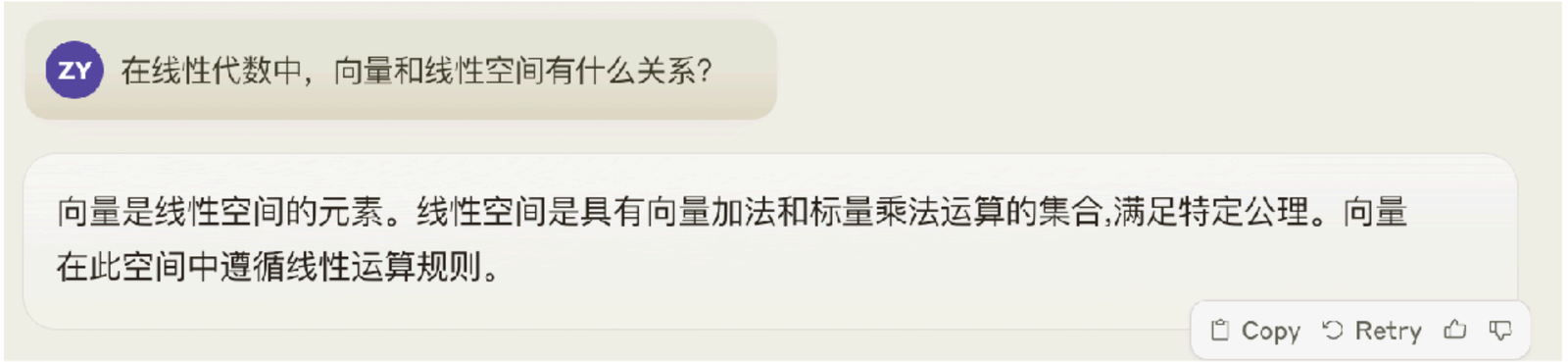
考虑一个具体的例子

如果问模型关于线性代数的问题，模型可以回答的很好



考虑一个具体的例子

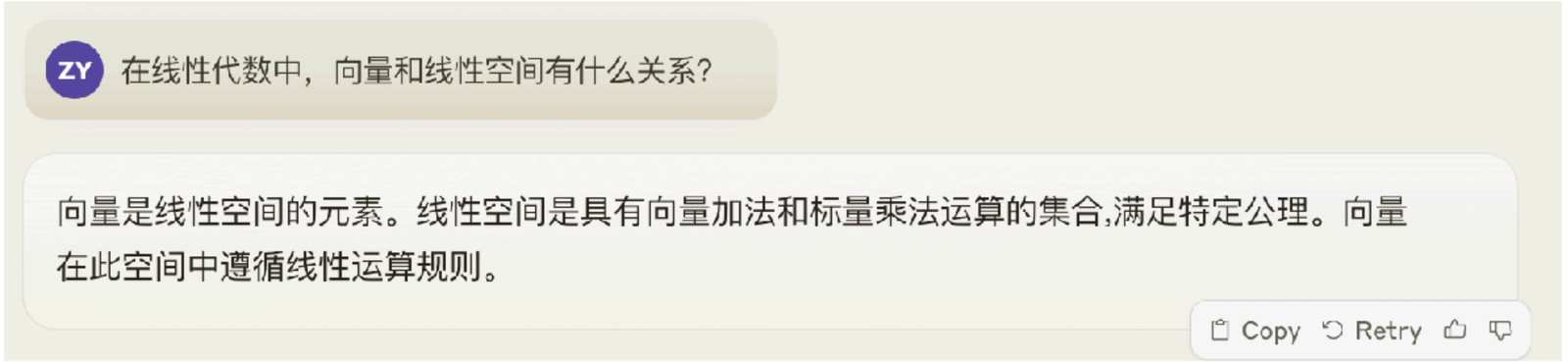
如果问模型关于线性代数的问题，模型可以回答的很好



但对于一个新的领域，可能只有几篇arXiv论文，模型很难学会其中的知识。

考虑一个具体的例子

如果问模型关于线性代数的问题，模型可以回答的很好

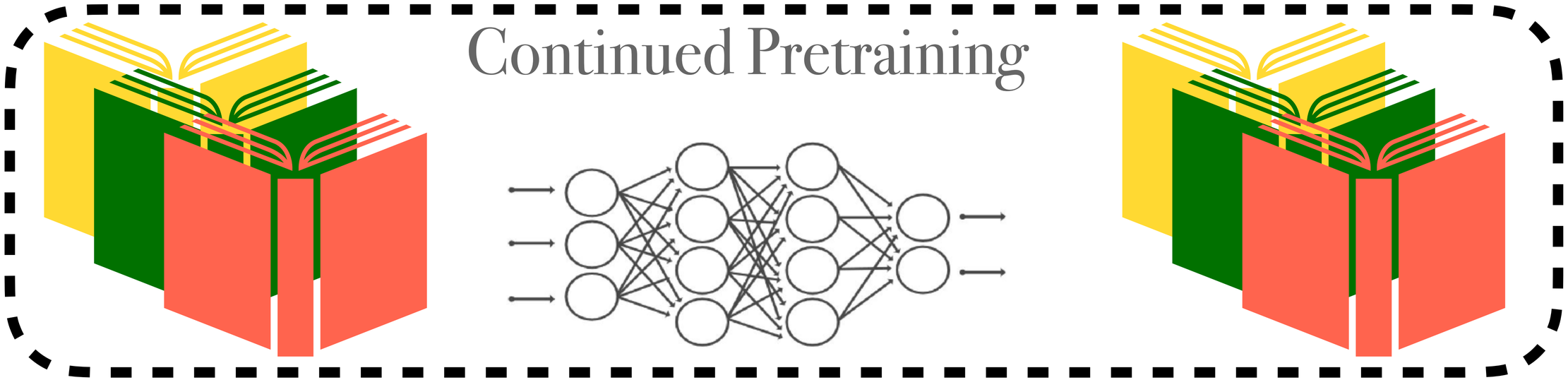


但对于一个新的领域，可能只有几篇arXiv论文，模型很难学会其中的知识。

想象一下线性代数相关的知识在预训练数据里出现的形式：

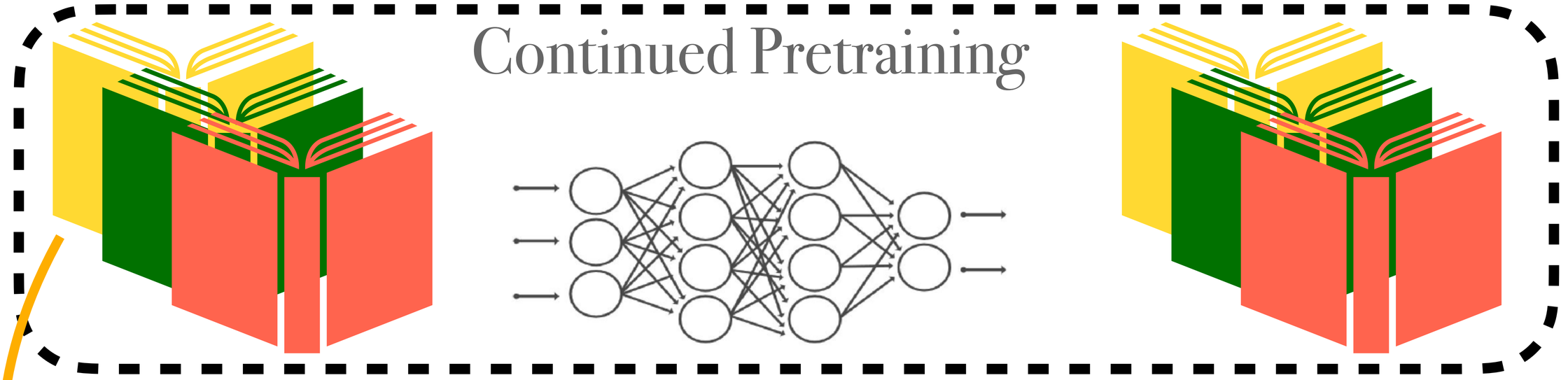
- 许多关于线性代数的教材，通过各种语言记载
- 互联网论坛上线性代数习题的讨论
- GitHub 里实现的实现奇异值分解的代码
- ...

Synthetic Continued Pretraining—在合成数据上继续预训练



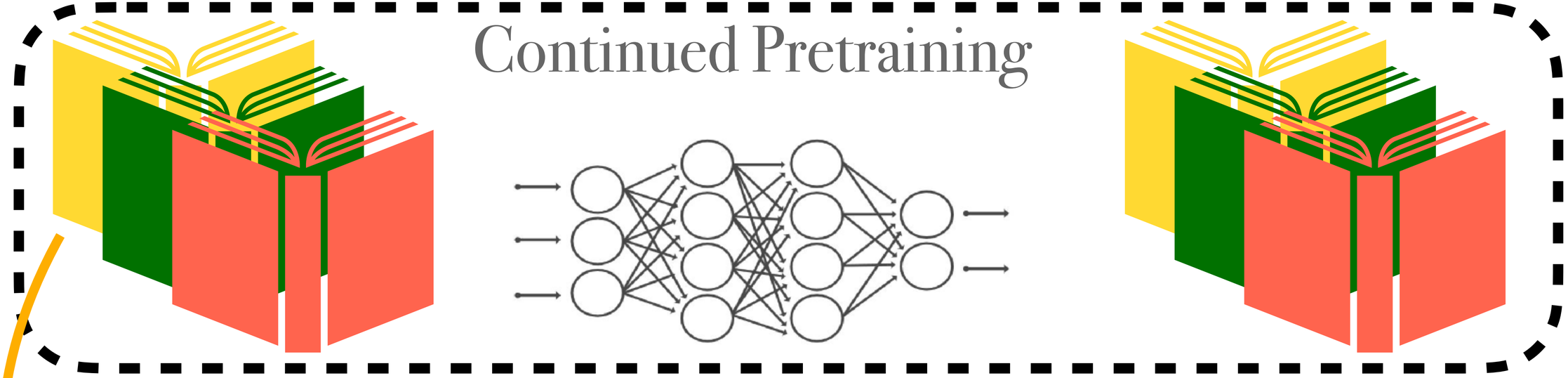
Synthetic Continued Pretraining—在合成数据上继续预训练

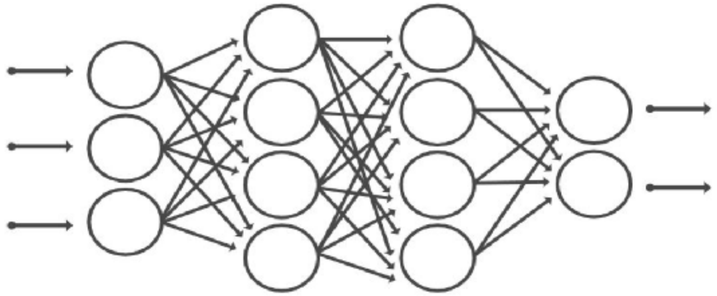
根据原文本合成新数据



Synthetic Continued Pretraining—在合成数据上继续预训练

根据原文本合成新数据

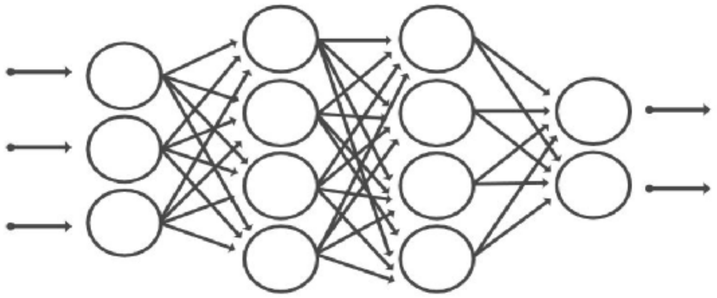




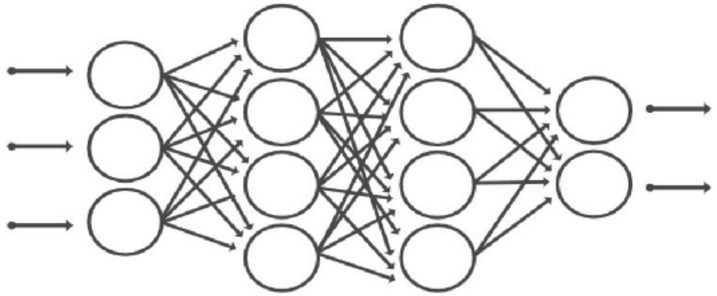
Synthetic Continued Pretraining—在合成数据上继续预训练

根据原文本合成新数据

Continued Pretraining



Synthetic Continued Pretraining



具体的实验文本以及测试数据

(i) 一些冷门书籍和文章的集合;(ii)高质量的问答题来检测模型在目标领域的知识。

具体的实验文本以及测试数据

(i) 一些冷门书籍和文章的集合; (ii)高质量的问答题来检测模型在目标领域的知识。

QuALITY 文章集

- Project Gutenberg 小说故事（以科幻小说为主）
- Slate 杂志文章，选自 Open American National（美国社会讨论）
- The Long and Short, Freesouls, 等刊物

QuALITY [Pang+ '21] 数据集

- 265本冷门书籍或杂着文章，总共1.8M token。

具体的实验文本以及测试数据

(i) 一些冷门书籍和文章的集合; (ii)高质量的问答题来检测模型在目标领域的知识。

QuALITY 文章集

- Project Gutenberg 小说故事（以科幻小说为主）
- Slate 杂志文章，选自 Open American National（美国社会讨论）
- The Long and Short, Freesouls, 等刊物

QuALITY [Pang+ '21] 数据集

- 265本冷门书籍或杂着文章，总共1.8M token。
- 高质量的 Q&A 多选题（形式类似MMLU）。

具体的实验文本以及测试数据

(i) 一些冷门书籍和文章的集合; (ii)高质量的问答题来检测模型在目标领域的知识。

QuALITY 文章集

- Project Gutenberg 小说故事（以科幻小说为主）
- Slate 杂志文章，选自 Open American National（美国社会讨论）
- The Long and Short, Freesouls, 等刊物

QuALITY [Pang+ '21] 数据集

- 265本冷门书籍或杂着文章，总共1.8M token。
- 高质量的 Q&A 多选题（形式类似MMLU）。
- 高质量的人工文章总结（Summarization）。

具体的实验文本以及测试数据

(i) 一些冷门书籍和文章的集合; (ii)高质量的问答题来检测模型在目标领域的知识。

QuALITY 文章集

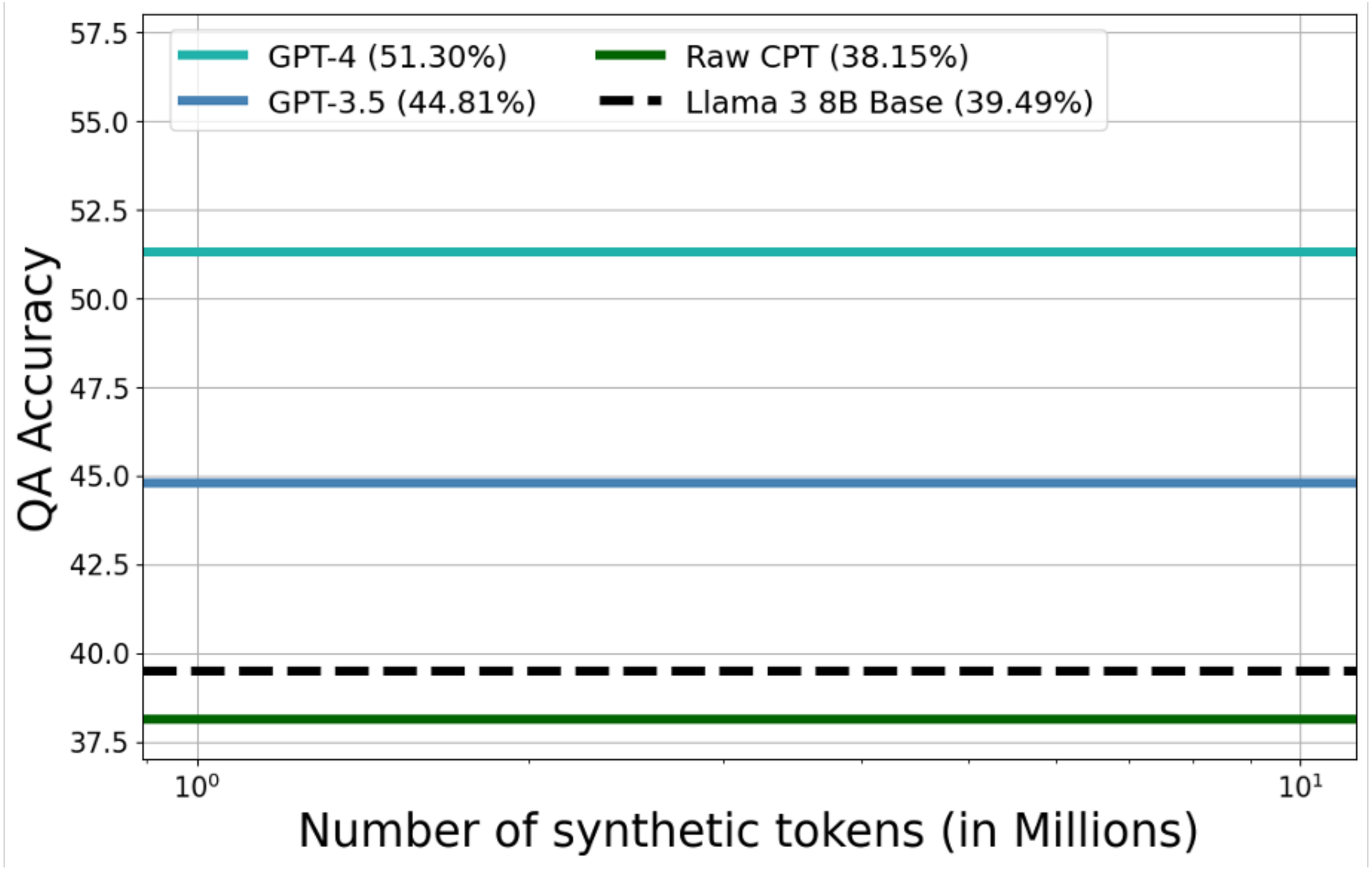
- Project Gutenberg 小说故事（以科幻小说为主）
- Slate 杂志文章，选自 Open American National（美国社会讨论）
- The Long and Short, Freesouls, 等刊物

QuALITY [Pang+ '21] 数据集

- 265本冷门书籍或杂着文章，总共1.8M token。
- 高质量的 Q&A 多选题（形式类似MMLU）。
- 高质量的人工文章总结（Summarization）。
- 在一般预训练数据里可能出现过，但网络上没有足够的讨论让模型学会这些文章里的内容。
- GPT-4在这个Q&A多选题上的准确率是51%，Llama 8B Base是39%。

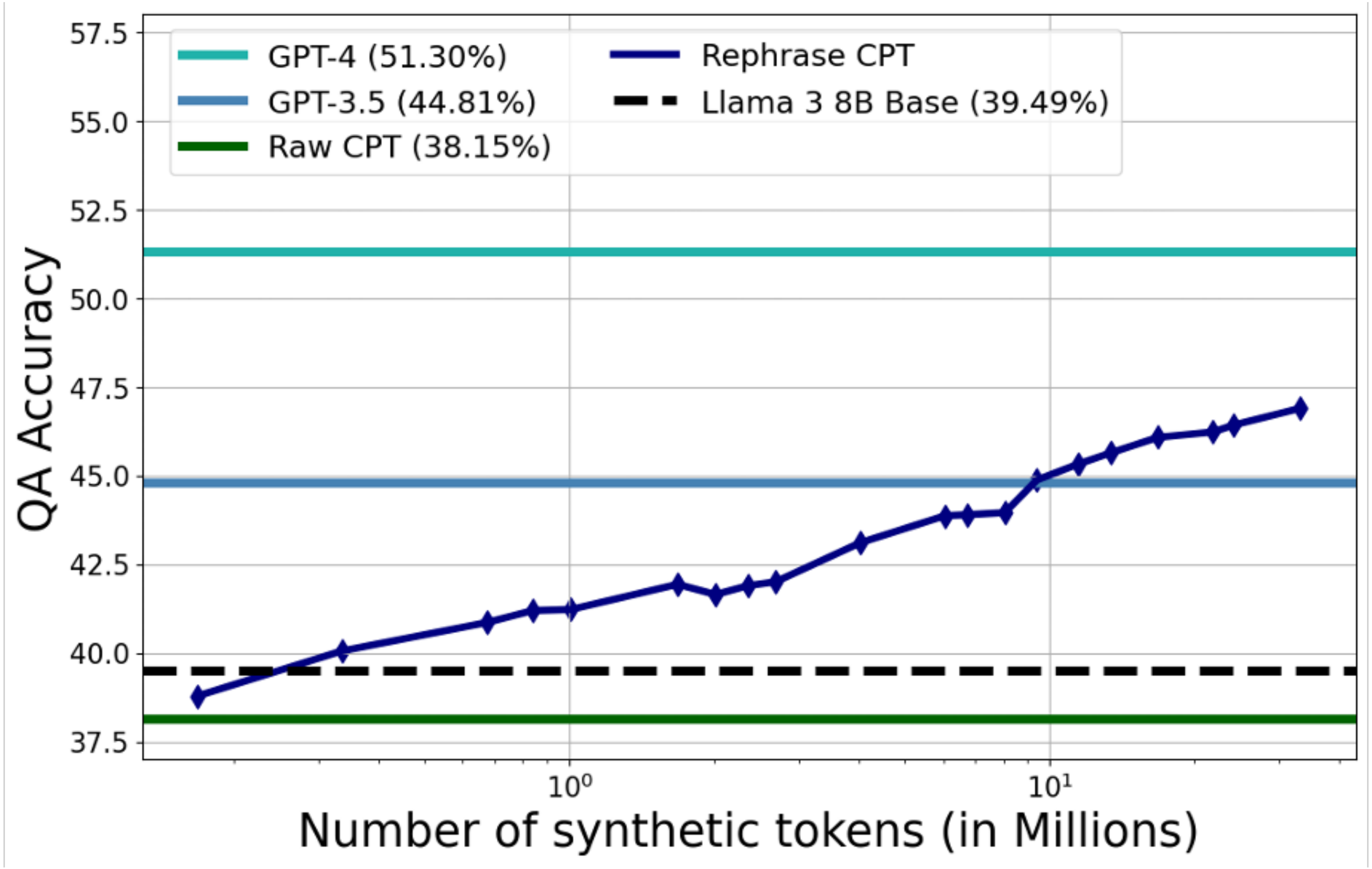
第一个尝试：直接在这1.8M token上训练

第一个尝试：直接在这1.8M token上训练



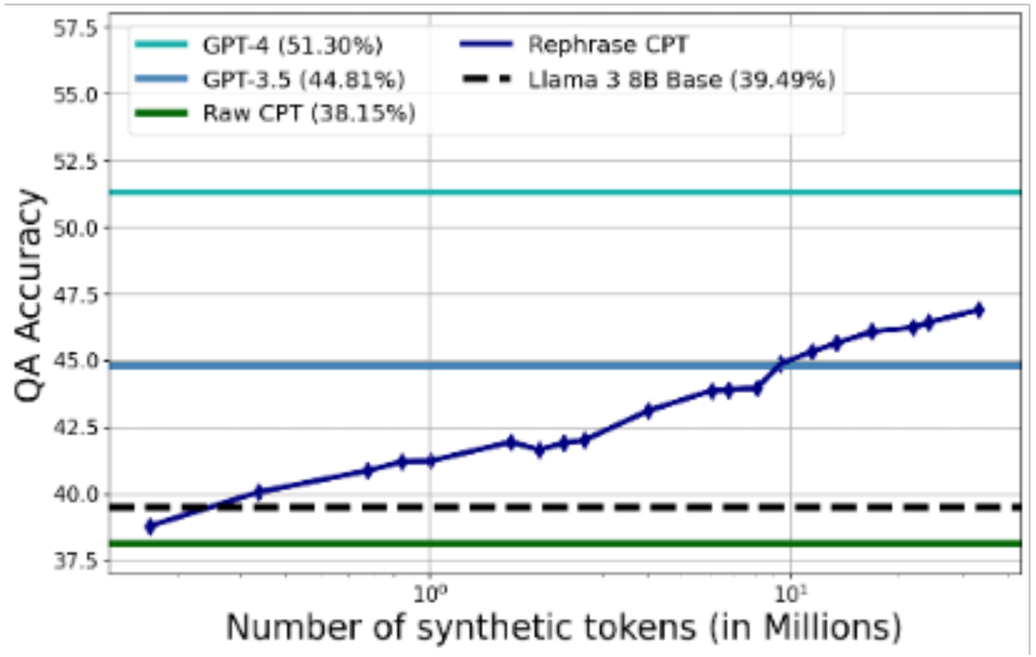
第二个尝试：简单的通过模型来重述原始数据

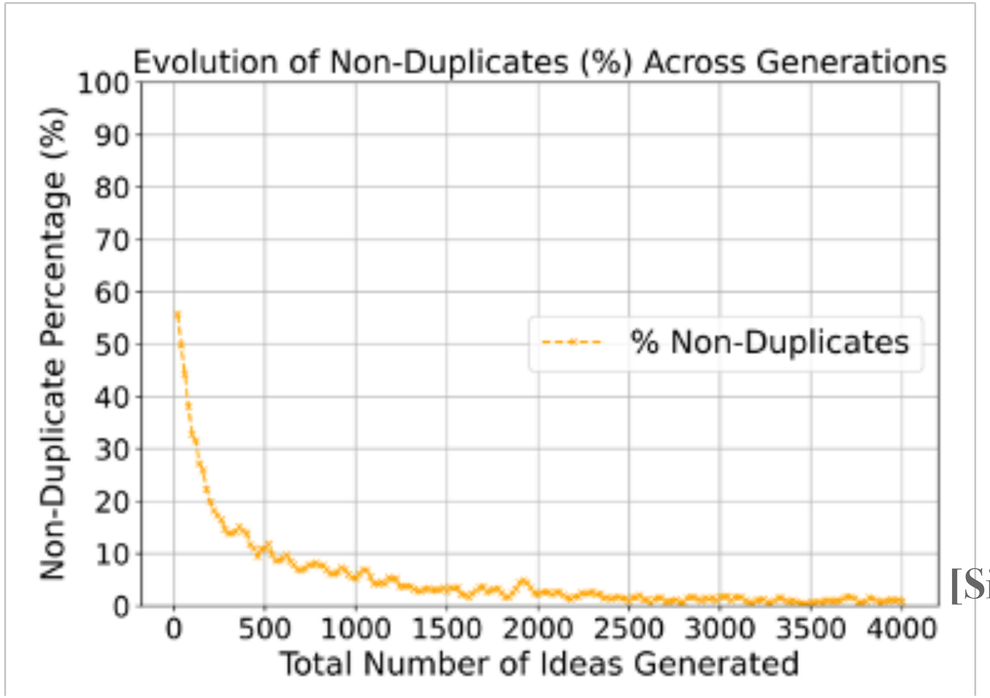
第二个尝试：简单的通过模型来重述原始数据



思路：引入外部的多样性

- 通过不断重述原始数据合成新数据的方法确实可以教会模型书里的知识。但是这个过程利用合成数据的效率很低。

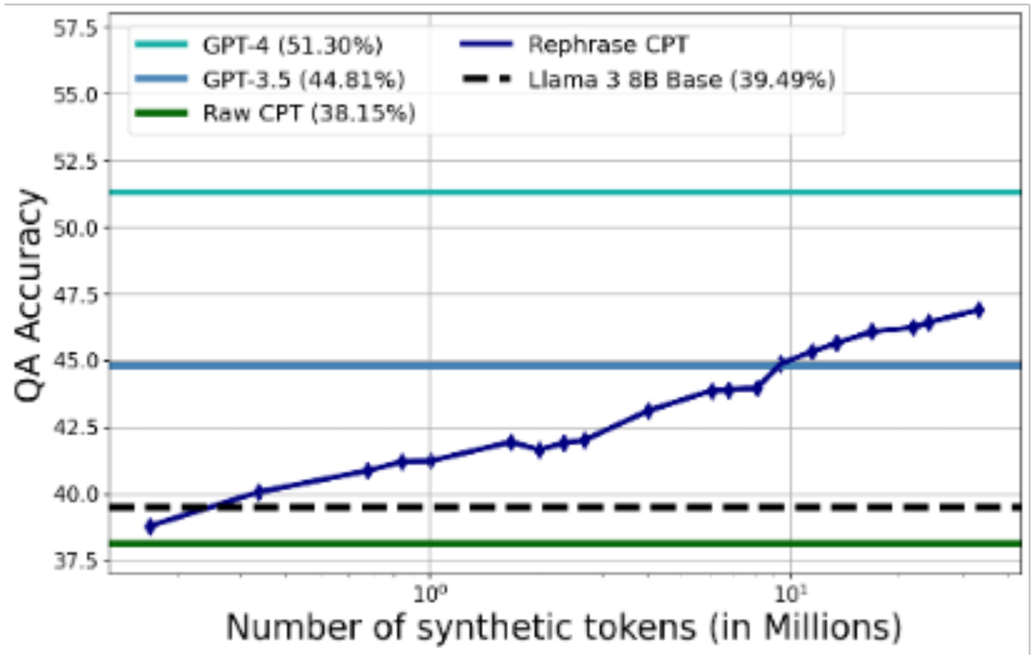


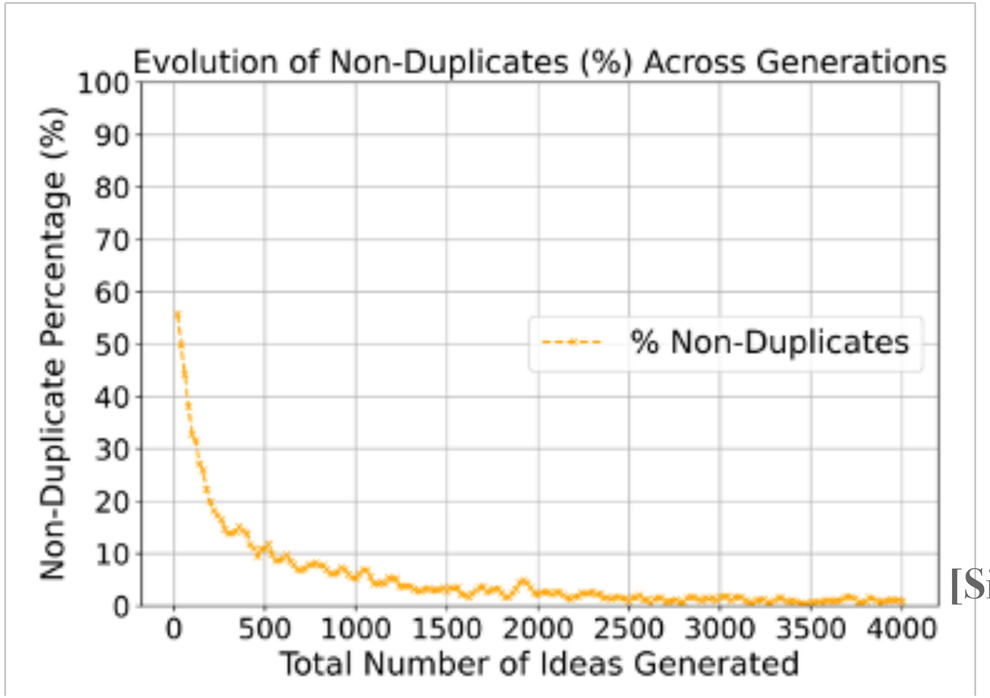


[Si+ '24]

思路：引入外部的多样性

- 通过不断重述原始数据合成新数据的方法确实可以教会模型书里的知识。但是这个过程利用合成数据的效率很低。





[Si+ '24]

- **核心问题是**：大语言模型没有很好的多样性。依赖于采样的温度来不断重复相同的提示词很难得到足够多样的数据。

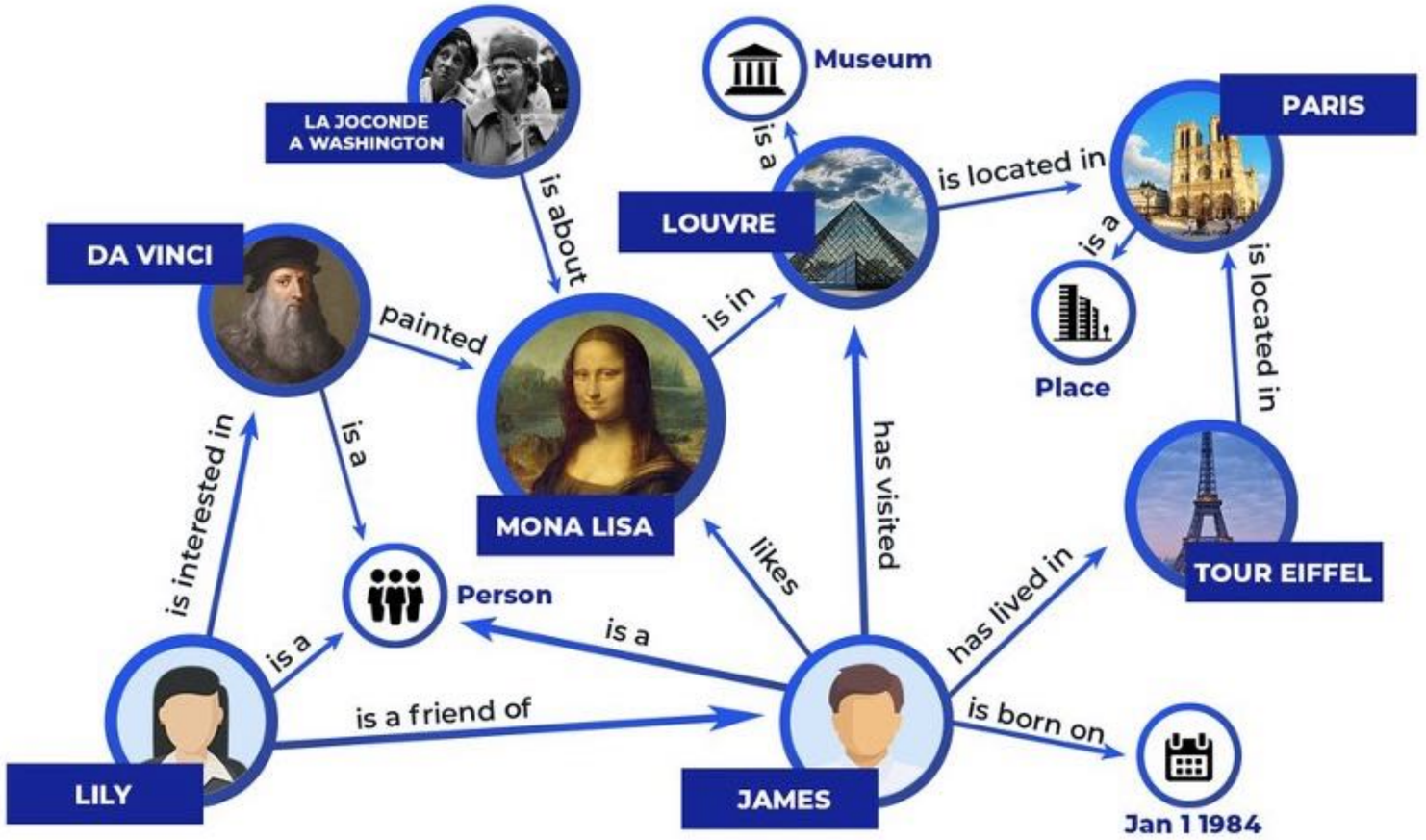
我们的方法：EntiGraph — 实体知识图谱

通过原文章所蕴含的知识图谱开增强提示词的多样性。

我们的方法：EntiGraph — 实体知识图谱

通过原文章所蕴含的知识图谱开增强提示词的多样性。

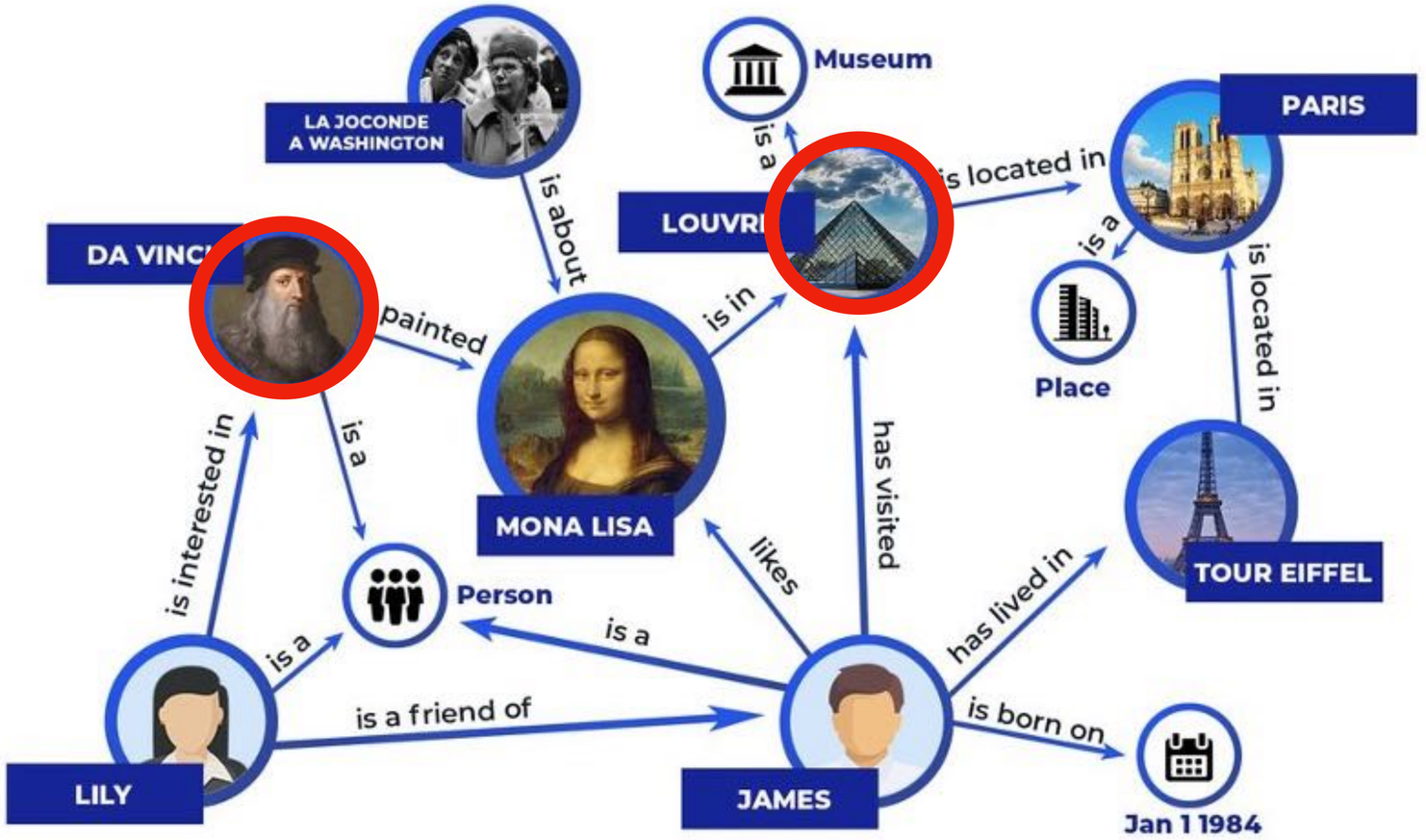
- 首先让模型提取出文章中的重要实体（Entities）。



我们的方法：EntiGraph — 实体知识图谱

通过原文章所蕴含的知识图谱开增强提示词的多样性。

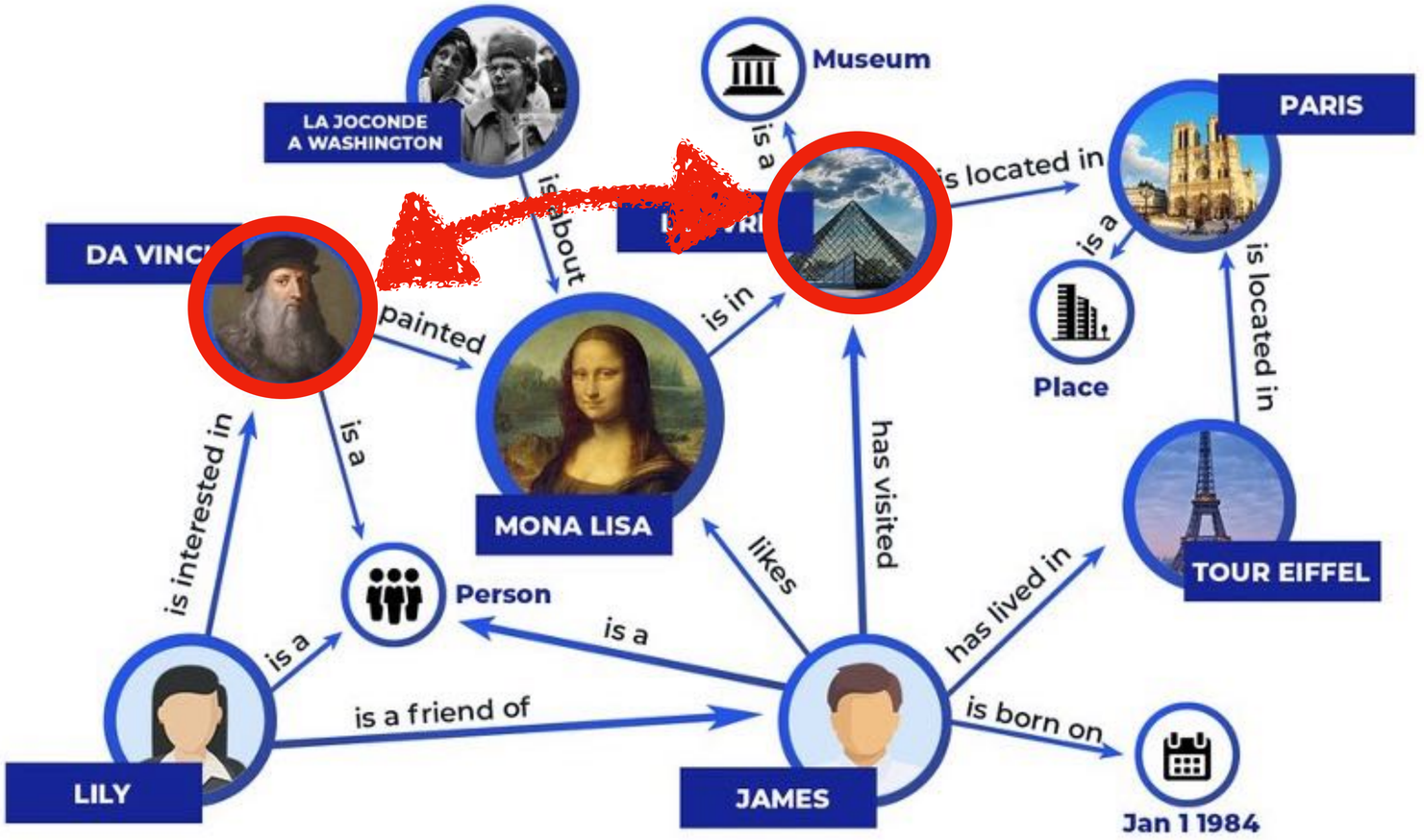
- 首先让模型提取出文章中的重要实体（Entities）。
- 从中采样任意 k 个实体。



我们的方法：EntiGraph — 实体知识图谱

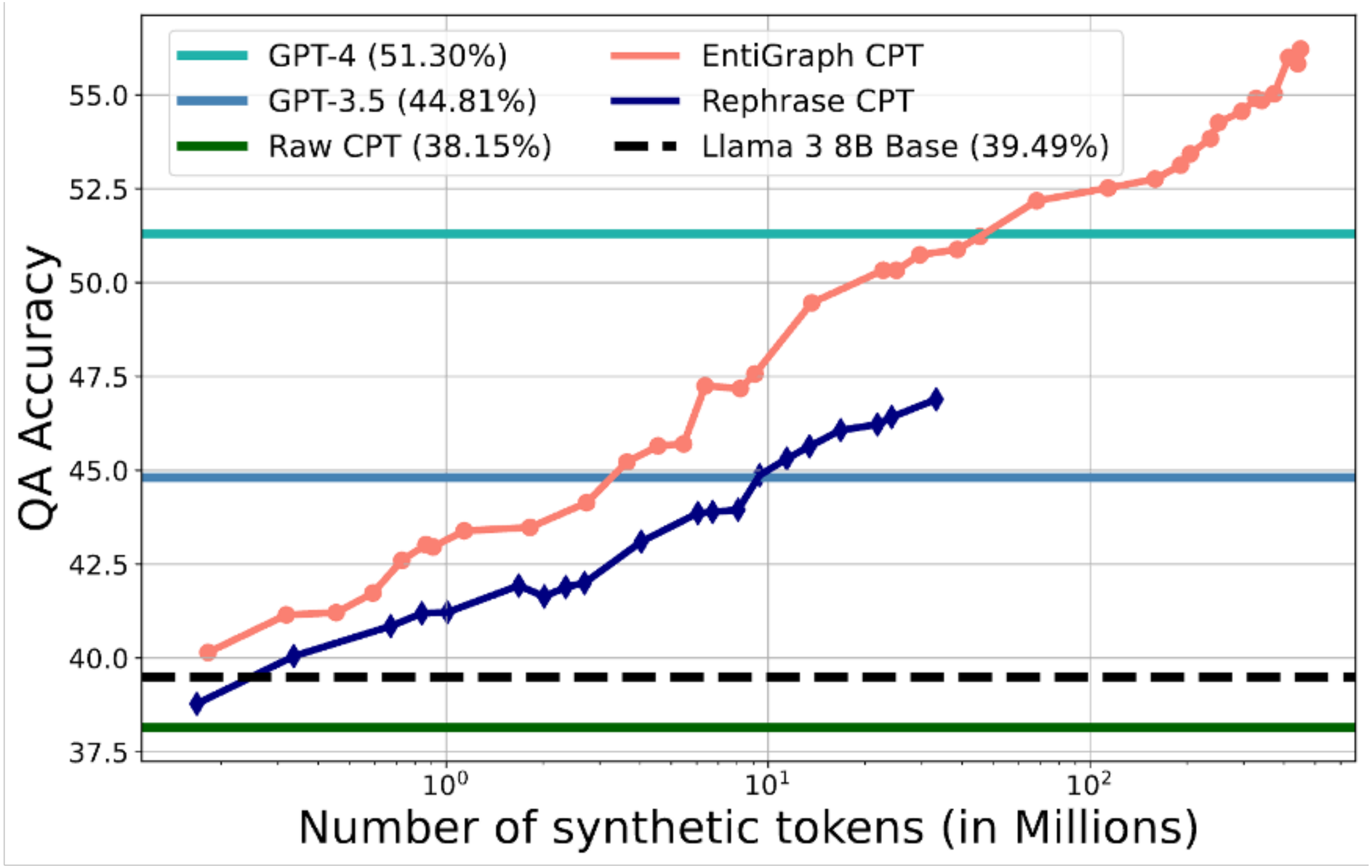
通过原文章所蕴含的知识图谱开增强提示词的多样性。

- 首先让模型提取出文章中的重要实体（Entities）。
- 从中采样任意 k个实体。
- 让模型描述生成关于这 k个实体的在原文中的关系，作为EntiGraph的合成数据。



EntiGraph合成的新数据：“达芬奇有很多作品被收藏在卢浮宫里…”

EntiGraph的闭卷Q&A问答表现



EntiGraph是否有预训练数据一样的泛化性质？

EntiGraph是否有预训练数据一样的泛化性质？

在EntiGraph数据上做完Synthetic continued pretraining之后，我们用指令跟随（Instruction following）数据来微调模型：

EntiGraph是否有预训练数据一样的泛化性质？

在EntiGraph数据上做完Synthetic continued pretraining之后，我们用指令跟随（Instruction following）数据来微调模型：

- 模型可以被用来总结文章

➤	**Explicit article reference:** Summarize the article "Defining Decay Down" by "David Plotz".
	The article "Defining Decay Down" by David Plotz discusses the evolution of dentistry in America. It highlights how dentists have adapted to changing times by shifting their focus from essential dental work to elective cosmetic procedures. [...]

在QuALITY数据集的例子

EntiGraph是否有预训练数据一样的泛化性质？

在EntiGraph数据上做完Synthetic continued pretraining之后，我们用指令跟随（Instruction following）数据来微调模型：

- 模型可以被用来总结文章
- 即使提示词中没有提到文章的标题，模型学到的知识也会影响其行为。

➤	**Explicit article reference:** Summarize the article "Defining Decay Down" by "David Plotz".
	The article "Defining Decay Down" by David Plotz discusses the evolution of dentistry in America. It highlights how dentists have adapted to changing times by shifting their focus from essential dental work to elective cosmetic procedures. [...]
➤	**Implicit article reference:** How has the practice of dentistry in the United States changed?
	The dentistry practice in the United States has undergone significant changes over the years. Here are some of the key changes: 1. Increased focus on cosmetic dentistry: In recent years, [...] 2. Improved dental technology: Advances in dental technology [...] 3. Increased emphasis on preventive care: Dentists have become more focused [...] 4. Increased competition: The dental industry has become more competitive, [...] 5. Improved patient experience: [...]

在QuALITY数据集的例子

EntiGraph是否有预训练数据一样的泛化性质？

在EntiGraph数据上做完Synthetic continued pretraining之后，我们用指令跟随（Instruction following）数据来微调模型：

- 模型可以被用来总结文章
- 即使提示词中没有提到文章的标题，模型学到的知识也会影响其行为。
- 模型可以同时讨论几篇不同的文章。

➤	**Explicit article reference:** Summarize the article "Defining Decay Down" by "David Plotz".
	The article "Defining Decay Down" by David Plotz discusses the evolution of dentistry in America. It highlights how dentists have adapted to changing times by shifting their focus from essential dental work to elective cosmetic procedures. [...]
➤	**Implicit article reference:** How has the practice of dentistry in the United States changed?
	The dentistry practice in the United States has undergone significant changes over the years. Here are some of the key changes: 1. Increased focus on cosmetic dentistry: In recent years, [...] 2. Improved dental technology: Advances in dental technology [...] 3. Increased emphasis on preventive care: Dentists have become more focused [...] 4. Increased competition: The dental industry has become more competitive, [...] 5. Improved patient experience: [...]
➤	**Cross article instruction:** Compare David Plotz's commentary style between how he analyzed American dentistry and how he discussed the movie Fight Club?
	David Plotz's commentary style is analytical and critical in both his analysis of American dentistry and his discussion of the movie Fight Club. In both cases, he uses a sharp, incisive tone to critique the subject matter, highlighting the absurdity and excesses of American dentistry and the nihilistic philosophy of Fight Club. However, there are some differences in his approach. [...]

在QuALITY数据集的例子

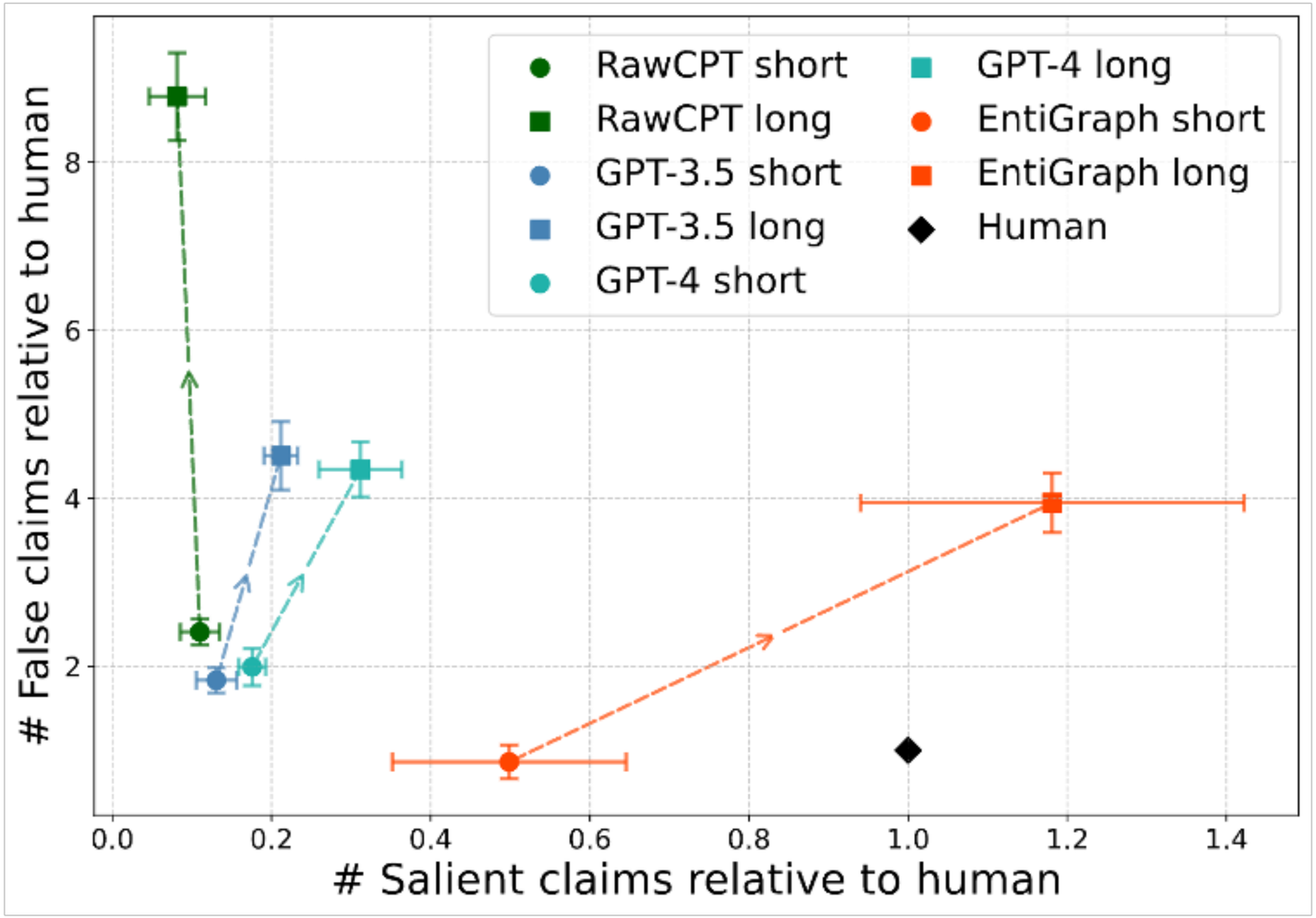
其他两个方法明显没有让模型学会文章的知识

EntiGraph Instruct with short prompt: Cosmic Yo-Yo is a science fiction story by Ross Rocklynne about Bob Parker and Queazy, two interplanetary haulers who are hired to deliver an asteroid to Earth for a wealthy client's backyard wedding. However, they face competition from their rivals, the Saylor brothers, who also want to deliver the asteroid. The story follows their race to deliver the asteroid before the deadline, using a yo-yo-like device to maneuver the asteroid and outsmart their rivals. The story is filled with action, adventure, and a touch of romance.

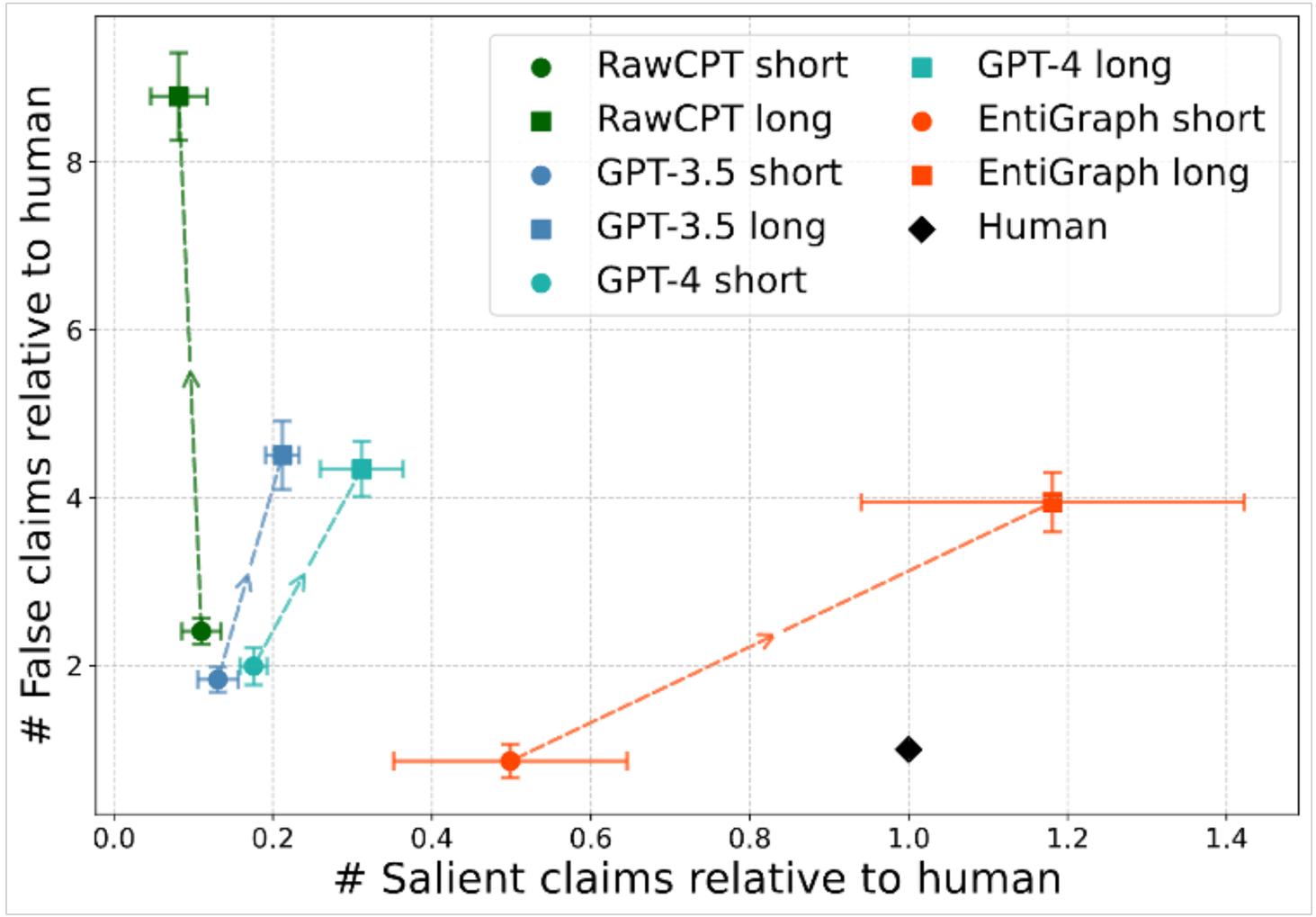
Raw Instruct with short prompt: Cosmic Yo-Yo by Rocklynne, Ross is a science fiction novel that follows the story of a young woman named Susan who discovers a mysterious device that can transport her through time and space. She uses the device to travel to different planets and meet different alien species, but soon realizes that her actions have consequences and she must find a way to return home. The novel explores themes of identity, responsibility, and the consequences of our actions.

Rephrase Instruct with short prompt: Cosmic Yo-Yo is a science fiction novel written by Ross Rocklynne in 1954. The story follows the adventures of a group of astronauts who are sent on a mission to explore a new planet. However, they soon discover that the planet is not as hospitable as they thought and must fight for their survival against the hostile environment and the indigenous creatures. The novel explores themes of exploration, survival, and the unknown.

在文本总结 (Summarization) 任务上的定量比较



在文本总结 (Summarization) 任务上的定量比较



EntiGraph在生成更多正确总结的同时，不会大量生成更多的错误总结。

在开卷的情况下，EntiGraph可以用来提升RAG表现

检索增强生成（RAG）是一种另一种让模型了解新知识的方法。

EntiGraph CPT + RAG		Llama 3 8B Base + RAG		GPT-4 + Oracle RAG		GPT-3.5 + Oracle RAG	
Accuracy	Recall@8	Accuracy	Recall@8	Accuracy	Recall@8	Accuracy	Recall@8
62.60	99.63	60.35	99.63	86.09	100.0	72.60	100.0

在开卷的情况下，EntiGraph可以用来提升RAG表现

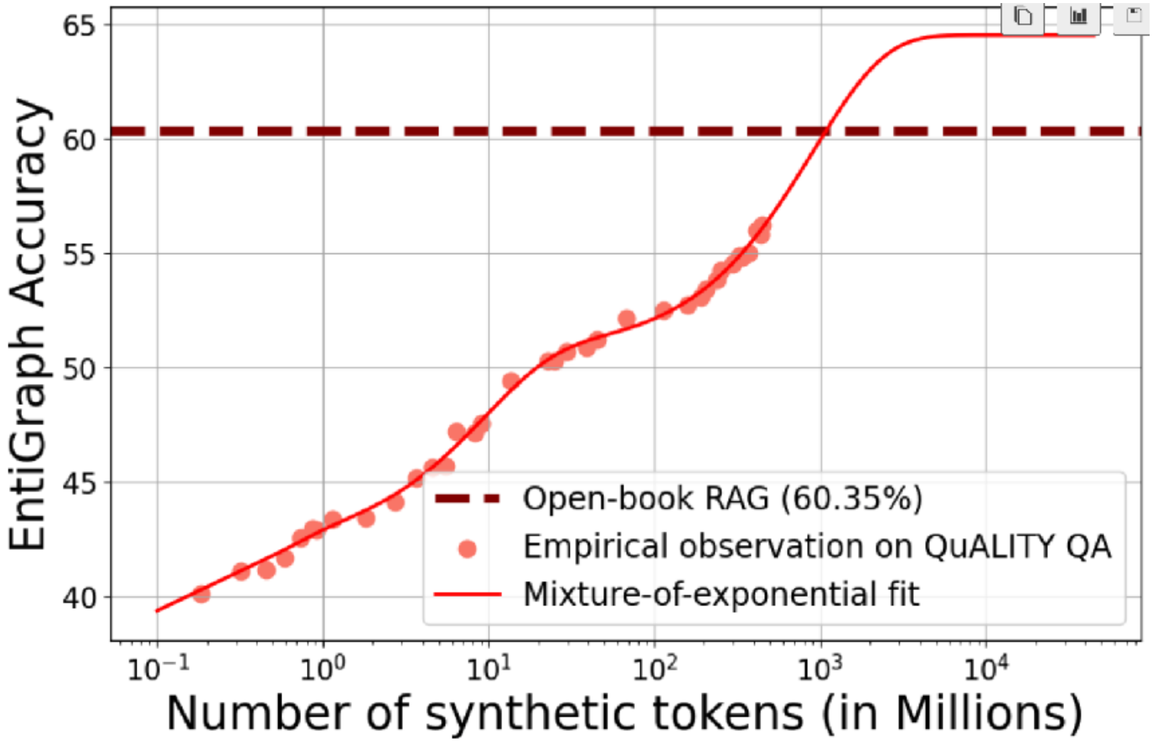
检索增强生成（RAG）是一种另一种让模型了解新知识的方法。

EntiGraph CPT + RAG		Llama 3 8B Base + RAG		GPT-4 + Oracle RAG		GPT-3.5 + Oracle RAG	
Accuracy	Recall@8	Accuracy	Recall@8	Accuracy	Recall@8	Accuracy	Recall@8
62.60	99.63	60.35	99.63	86.09	100.0	72.60	100.0

- 我们建立了一个物理模型来研究EntiGraph的性质。这个模型预测准确率的形状是一个Mixture-of- exponential 曲线。

$$A(t) \sim p + C\left[1 - \sum_{k=1}^{\infty} \mu(k)(1 - a_k)^t\right]$$

- 用该模型来拟合我们观察到的准确率，预测EntiGraph最终可以达到65%的准确性。

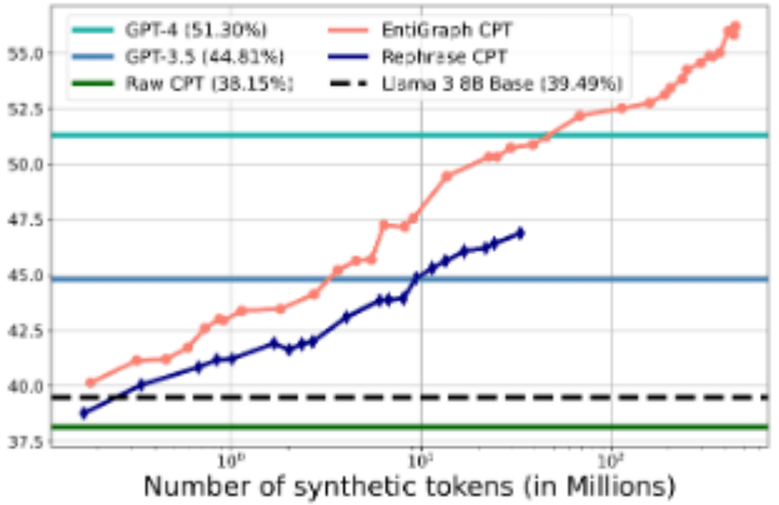


总结：在合成数据上继续预训练

- 我们提出了Synthetic continued pretraining这个概念，用来让模型从小文本上学习新的知识。同时我们提出了EntiGraph这个方法来有效的合成数据。

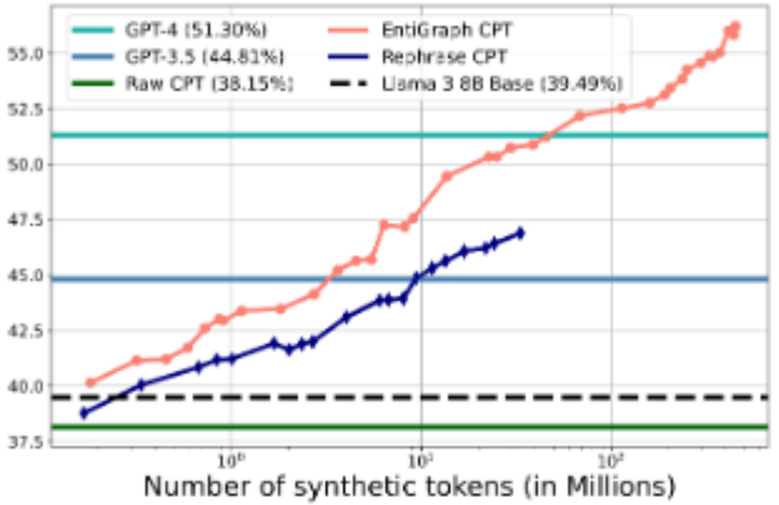
总结：在合成数据上继续预训练

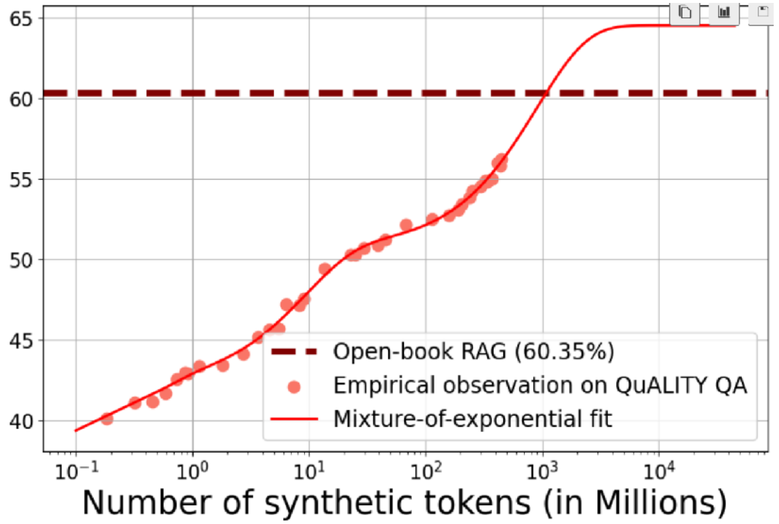
- 我们提出了Synthetic continued pretraining这个概念，用来让模型从小文本上学习新的知识。同时我们提出了EntiGraph这个方法来有效的合成数据。
- 实验上，EntiGraph有效的增加了模型关于目标文本的知识。与简单的重述文本，或者在原文本上训练相比，EntiGraph能更有效的更容易扩大规模。



总结：在合成数据上继续预训练

- 我们提出了Synthetic continued pretraining这个概念，用来让模型从小文本上学习新的知识。同时我们提出了EntiGraph这个方法来有效的合成数据。
- 实验上，EntiGraph有效的增加了模型关于目标文本的知识。与简单的重述文本，或者在原文本上训练相比，EntiGraph能更有效的更容易扩大规模。
- 我们简单讨论了检索加强生成（RAG）的情况，在这里发现EntiGraph所学到的知识也可以增加RAG的表现。并且提供了一个物理模型，对这个模型的理论分析很好的拟合了真实数据上的实验。



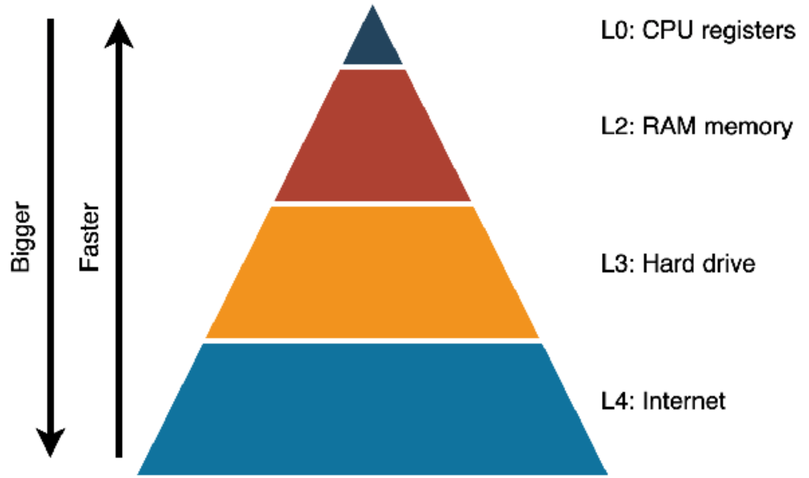


Synthetic continued pretraining 对我们有哪些启发？

Synthetic continued pretraining 对我们有哪些启发？

让模型学习非公开领域的新知识

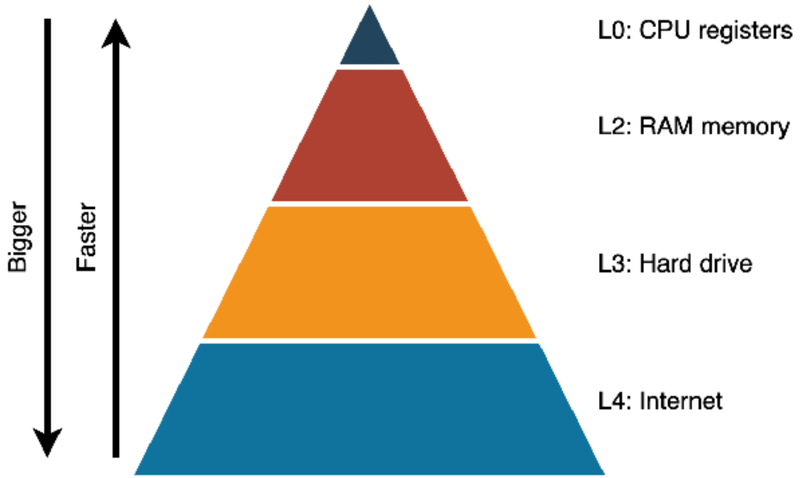
- 纵观计算机发展的历史，革命新的进步往往伴随着对于Memory（内存，记忆）这个概念的新理解。



Synthetic continued pretraining 对我们有哪些启发？

让模型学习非公开领域的新知识

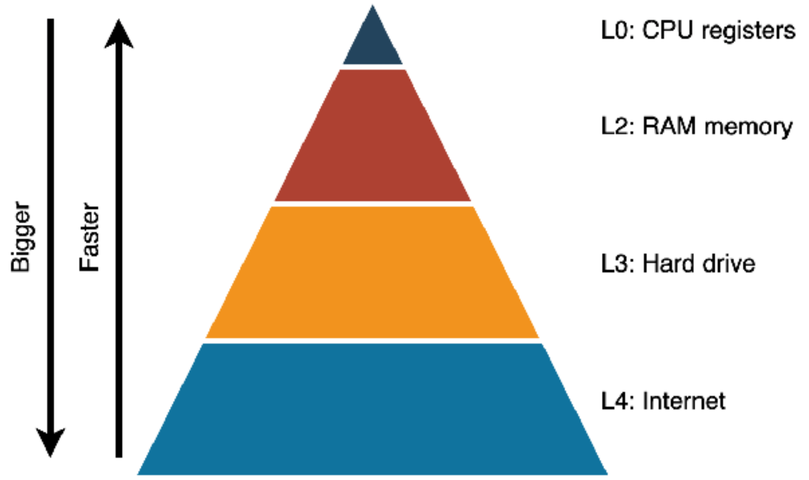
- 纵观计算机发展的历史，革命新的进步往往伴随着对于Memory（内存，记忆）这个概念的新理解。
- 人工智能也是一样，模型的最底层的知识储存在他的权重中。RAG提供了外部的浅层知识。Synthetic continued pretrianing可以用来增加最底层的知识。



Synthetic continued pretraining 对我们有哪些启发？

让模型学习非公开领域的新知识

- 纵观计算机发展的历史，革命新的进步往往伴随着对于Memory（内存，记忆）这个概念的新理解。
- 人工智能也是一样，模型的最底层的知识储存在他的权重中。RAG提供了外部的浅层知识。Synthetic continued pretrianing可以用来增加最底层的知识。

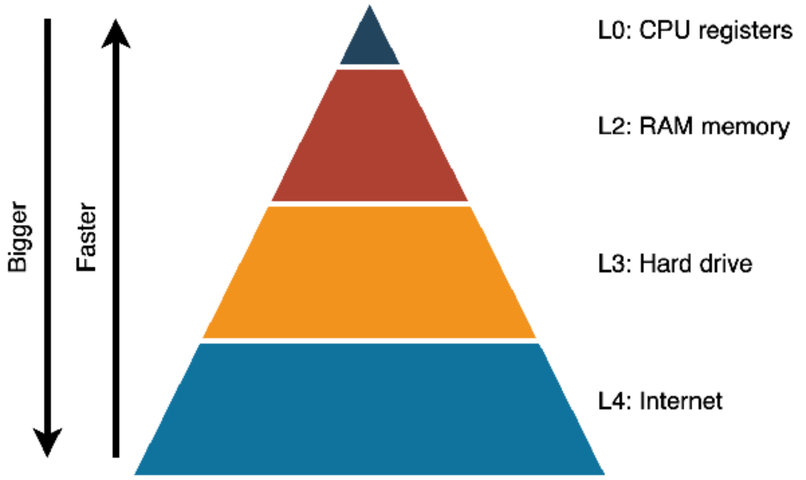


如果模型可以通过合成数据自我学习，人类的工作是什么？

Synthetic continued pretraining 对我们有哪些启发？

让模型学习非公开领域的新知识

- 纵观计算机发展的历史，革命新的进步往往伴随着对于Memory（内存，记忆）这个概念的新理解。
- 人工智能也是一样，模型的最底层的知识储存在他的权重中。RAG提供了外部的浅层知识。Synthetic continued pretrianing可以用来增加最底层的知识。



如果模型可以通过合成数据自我学习，人类的工作是什么？

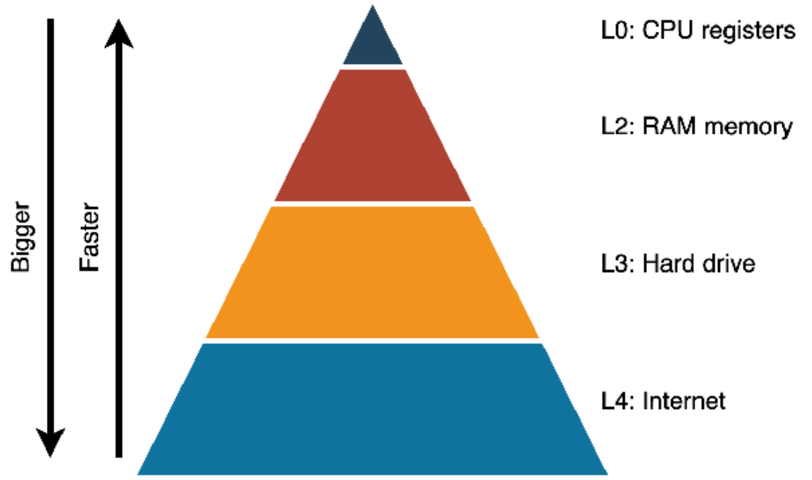
- 人类的仍然可以为模型提供“启蒙性”的数据。

$$P_{\text{LM}}(\texttt{ define } \sqrt{-1} \texttt{ as root of } x^2+1 \texttt{ | <all human text before 1572>}) = 10^{-60}$$

Synthetic continued pretraining 对我们有哪些启发？

让模型学习非公开领域的新知识

- 纵观计算机发展的历史，革命新的进步往往伴随着对于Memory（内存，记忆）这个概念的新理解。
- 人工智能也是一样，模型的最底层的知识储存在他的权重中。RAG提供了外部的浅层知识。Synthetic continued pretrianing可以用来增加最底层的知识。



如果模型可以通过合成数据自我学习，人类的工作是什么？

- 人类的仍然可以为模型提供“启蒙性”的数据。

$$P_{\text{LM}}(\texttt{ define } \sqrt{-1} \texttt{ as root of } x^2+1 \texttt{ | <all human text before 1572>}) = 10^{-60}$$

- 当人类提供了这些“启蒙性”数据以后，机器可以自我学习，推理出这些数据的推论。但是模型生成“启蒙性”数据的概率可能很低，因为这些数据和训练数据的分布非常不一样。