

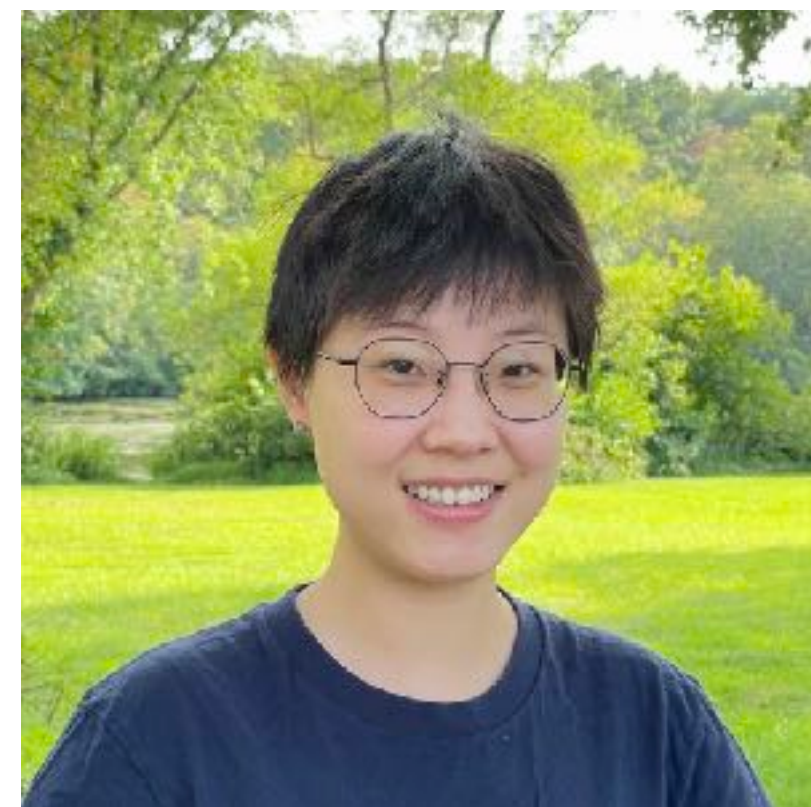


Synthetic Continued Pretraining

Zitong Yang*
Stanford Statistics



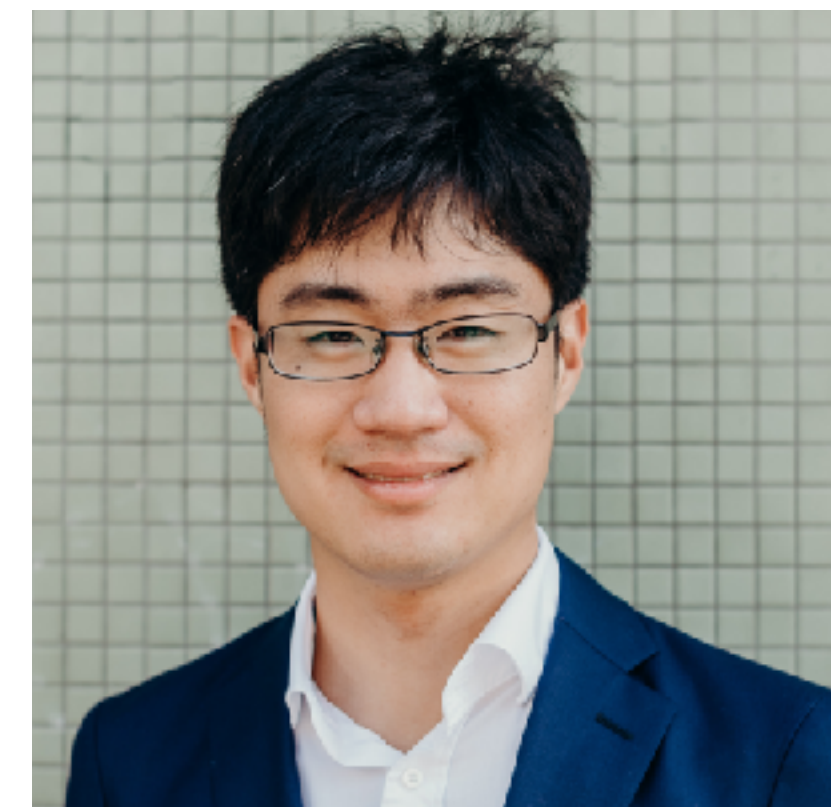
*Neil Band**



Shuangping Li



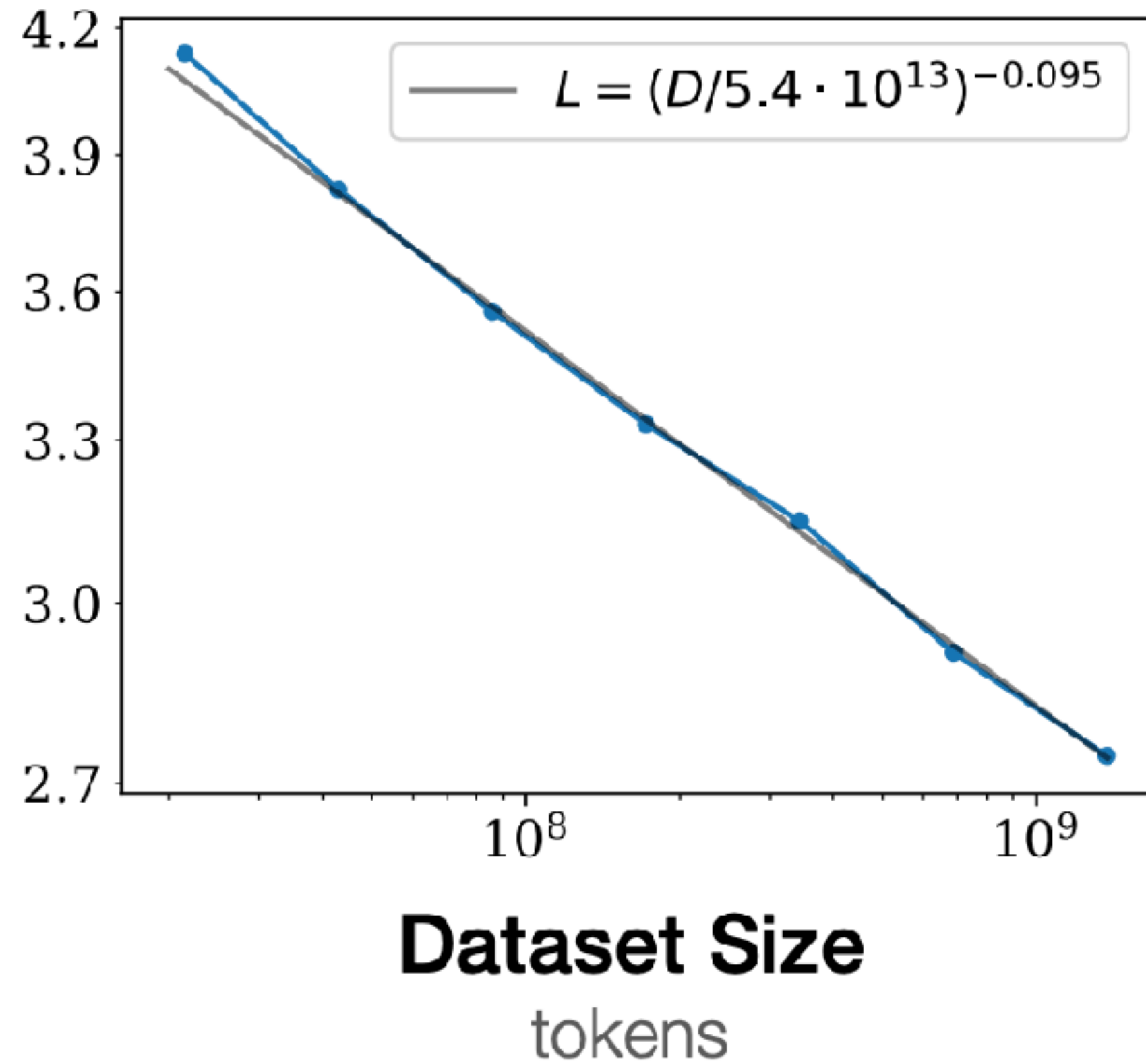
Emmanuel Candès



Tatsunori Hashimoto

Large language model

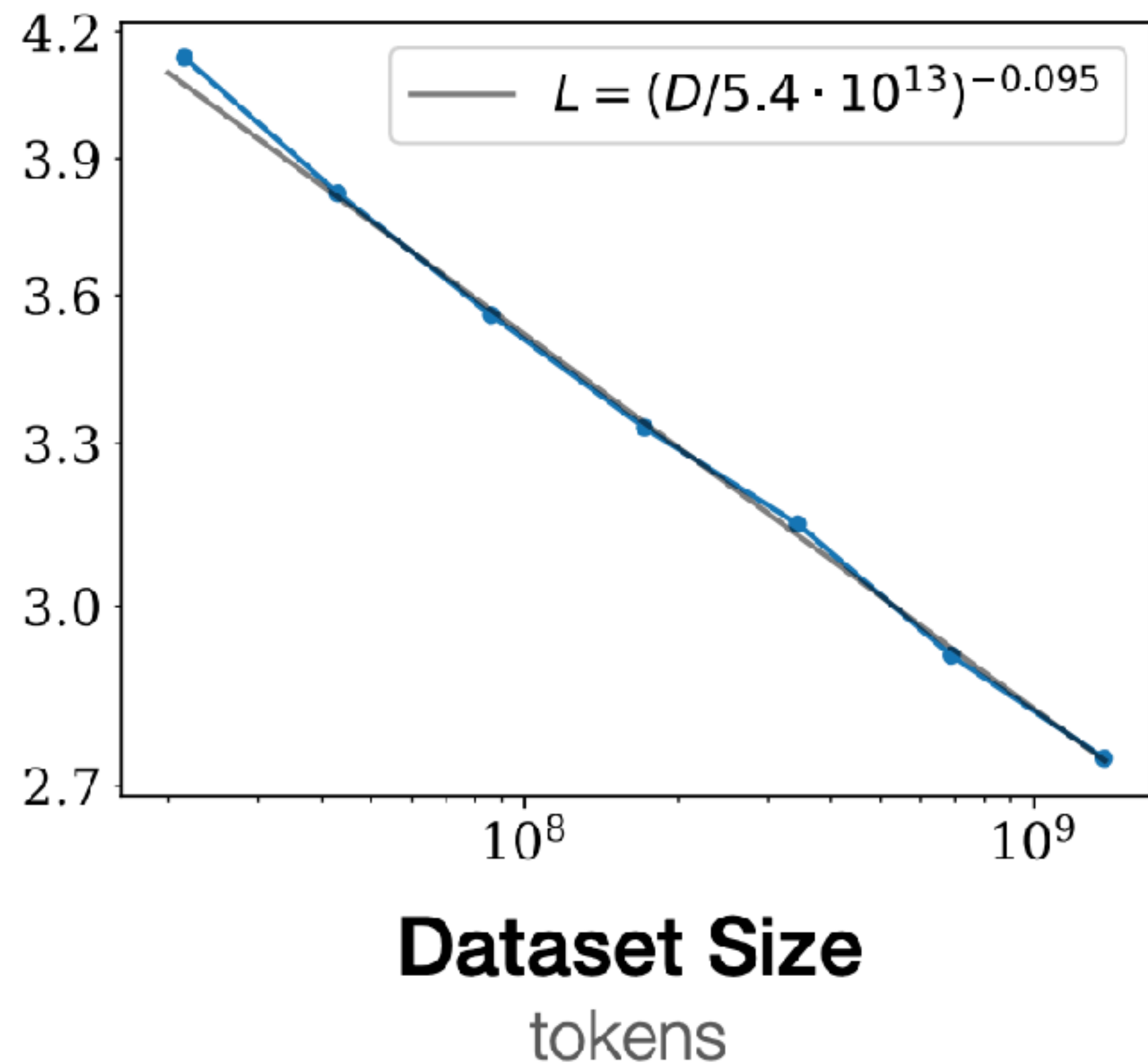
Pretraining on large volume of internet text teaches model rich world knowledge



- Scaling law: model capability improves with more pretraining data.

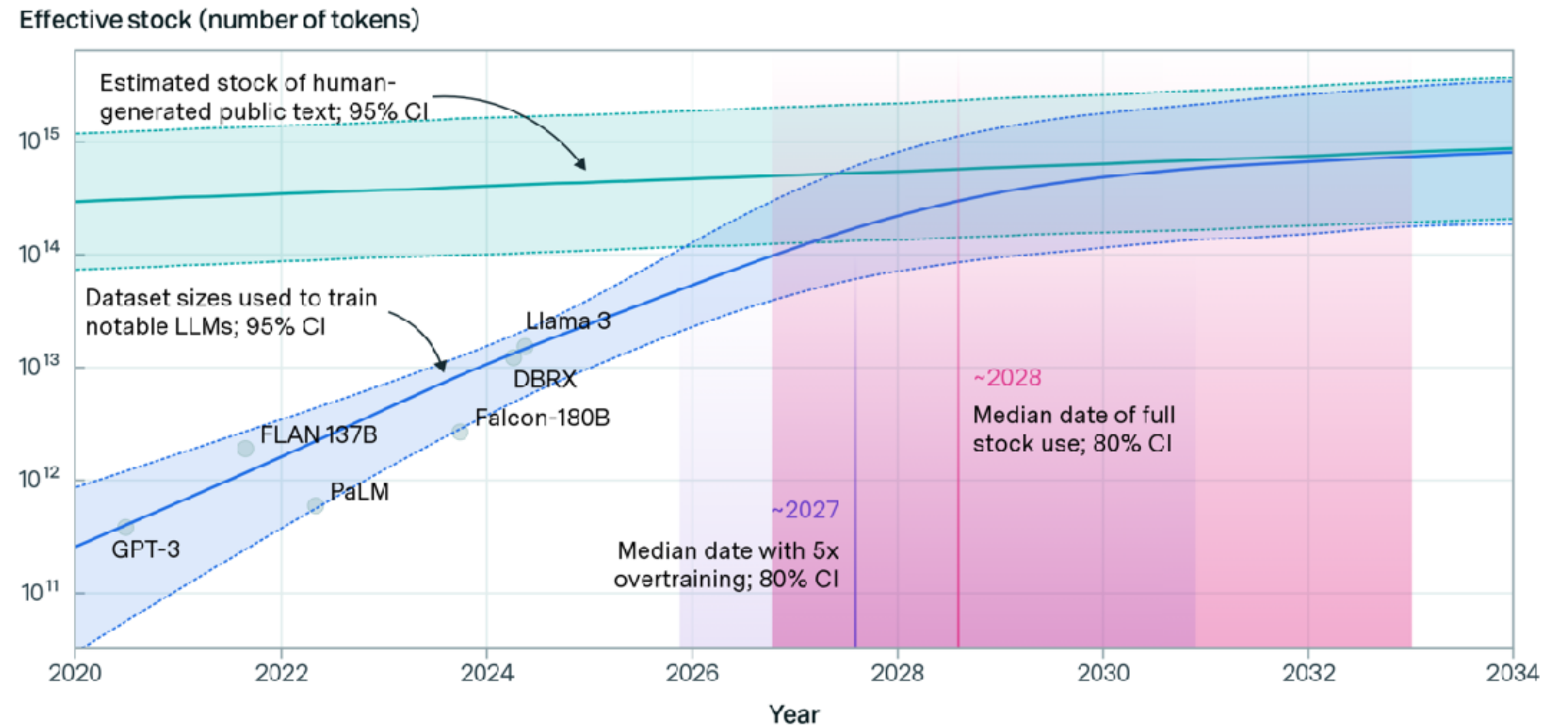
Large language model

Pretraining on large volume of internet text teaches model rich world knowledge



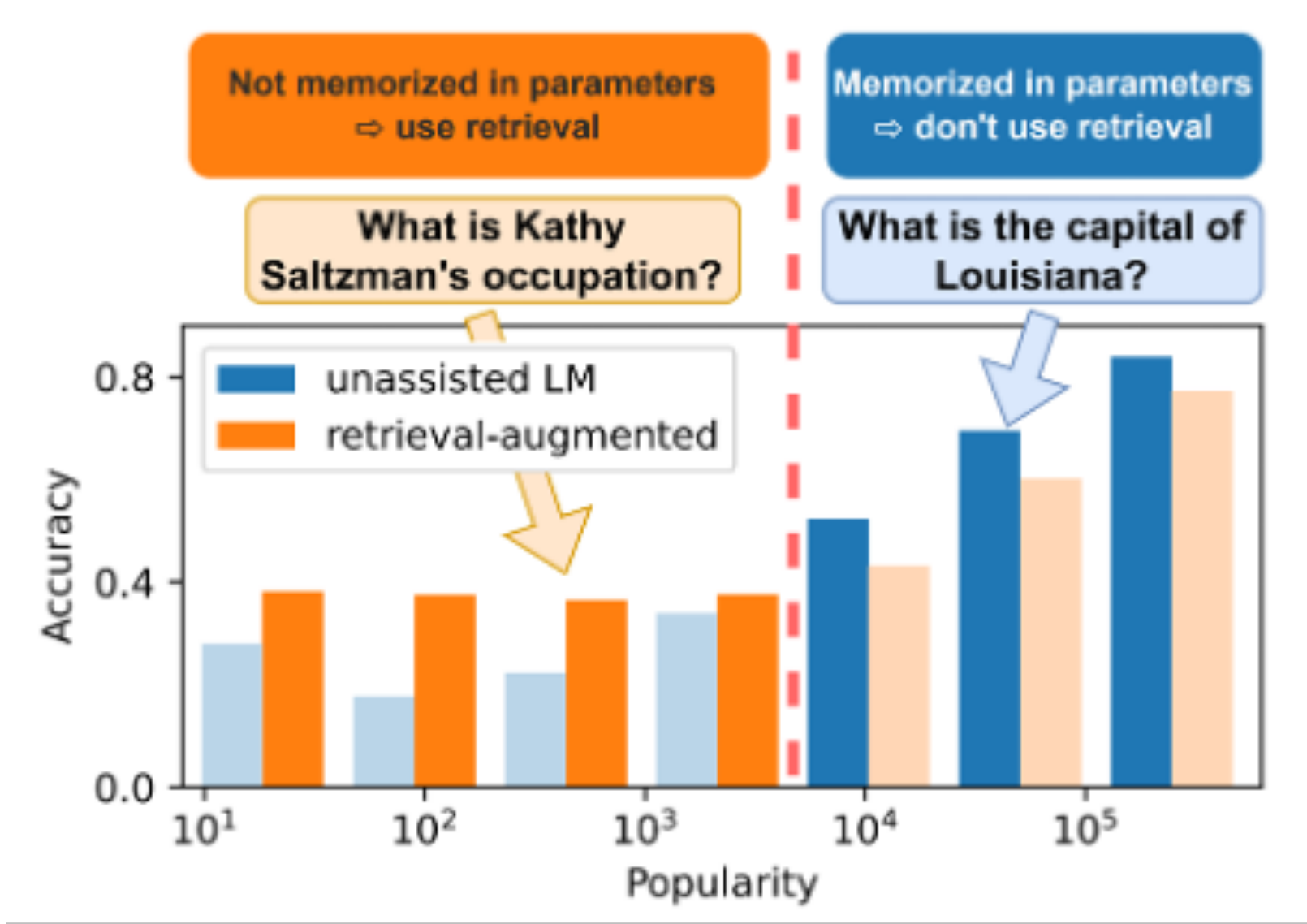
Projections of the stock of public text and data usage

EPOCH AI



- Scaling law: model capability improves with more pretraining data.
- Frontier models will run out of internet text as soon as 2028.
- How should we continued to improve model capability beyond that?

Even before 2028...

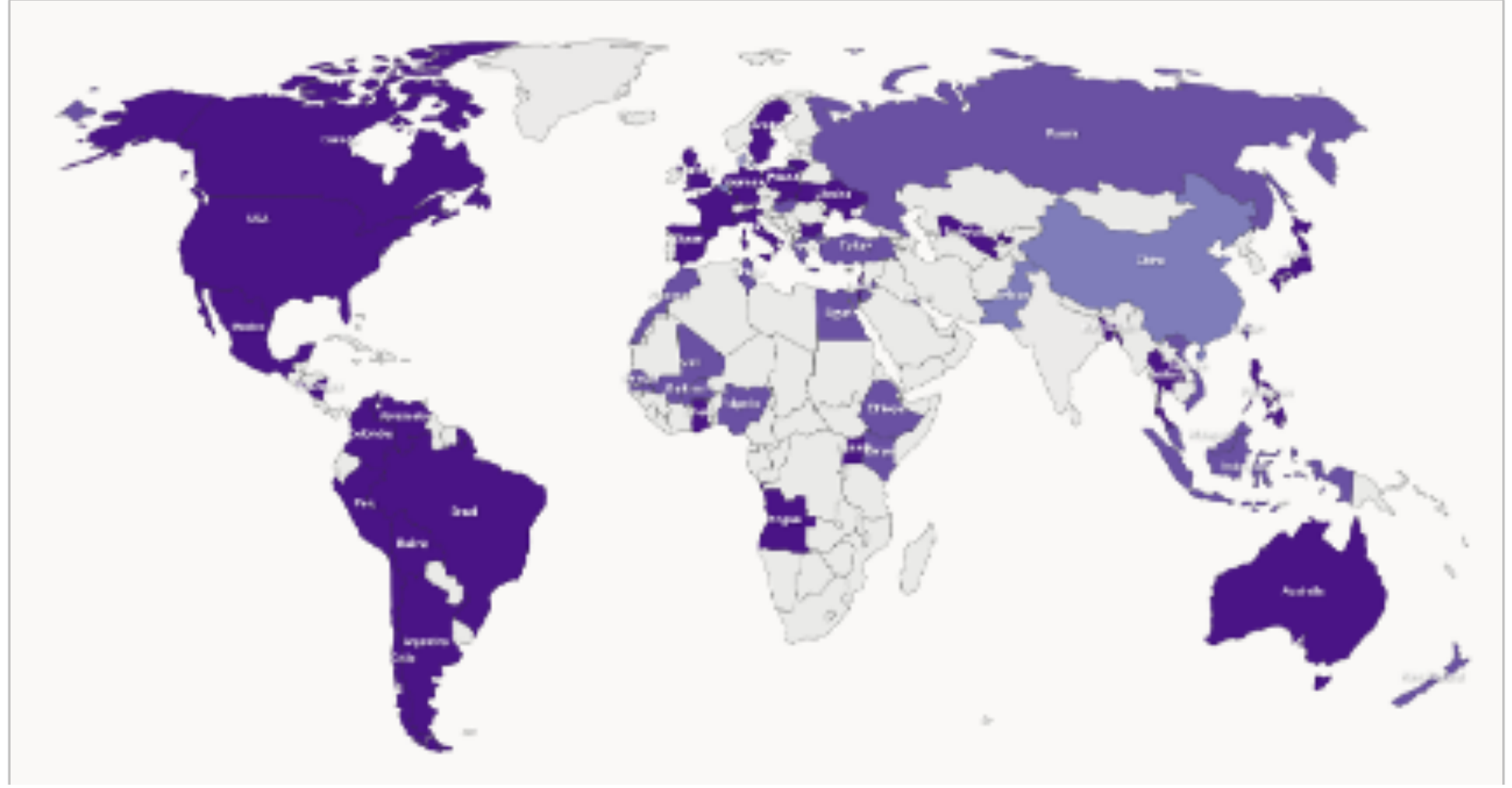


GPQA: A Graduate-Level Google-Proof Q&A Benchmark

David Rein^{1,2} Betty Li Hou¹ Asa Cooper Stickland¹
Jackson Petty¹ Richard Yuanzhe Pang¹ Julien Dirani¹
Julian Michael^{†1} Samuel R. Bowman^{†1,3}
¹New York University ²Cohere ³Anthropic, PBC

Rare knowledge in niche domains

Ph.D. level frontier questions

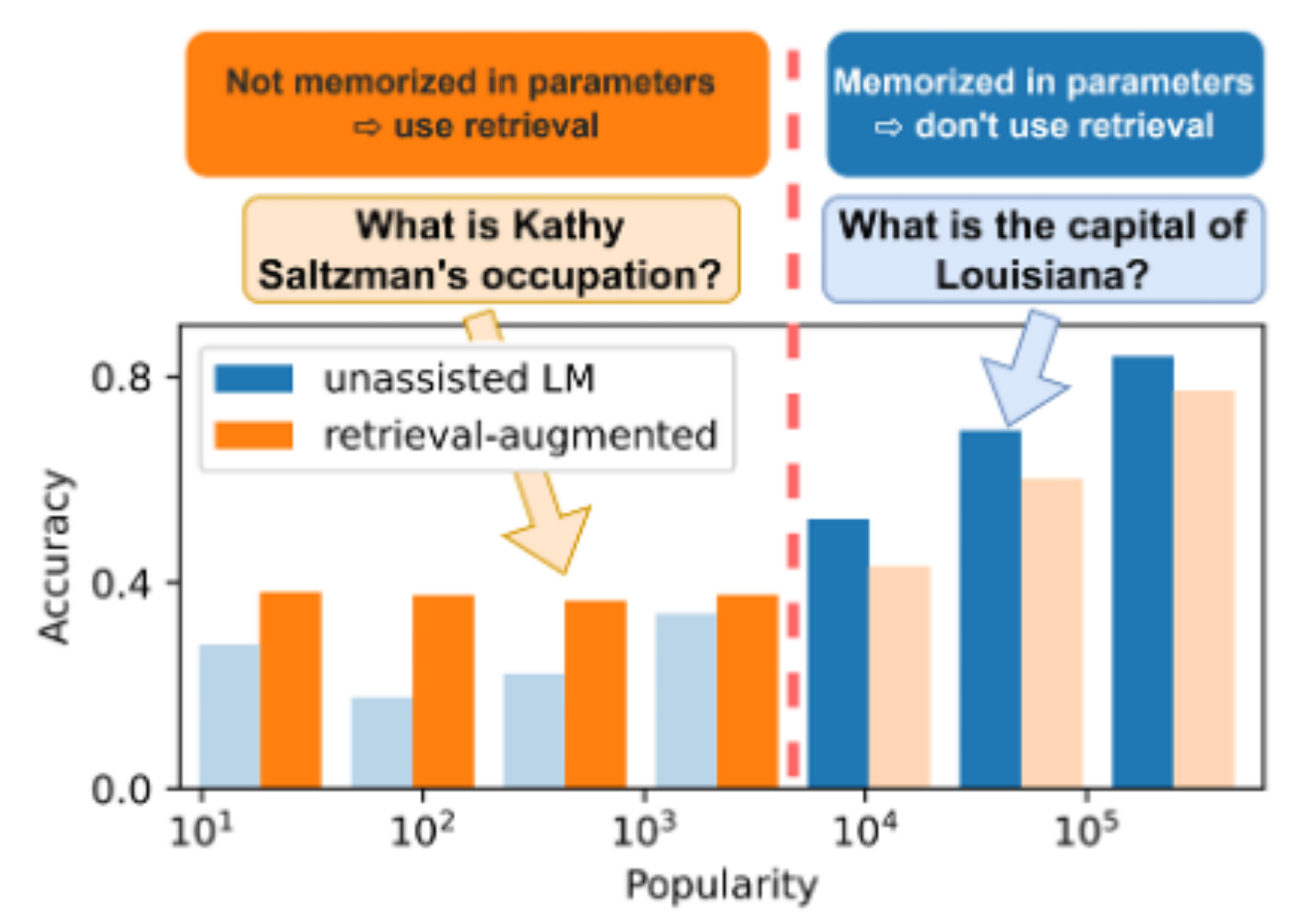
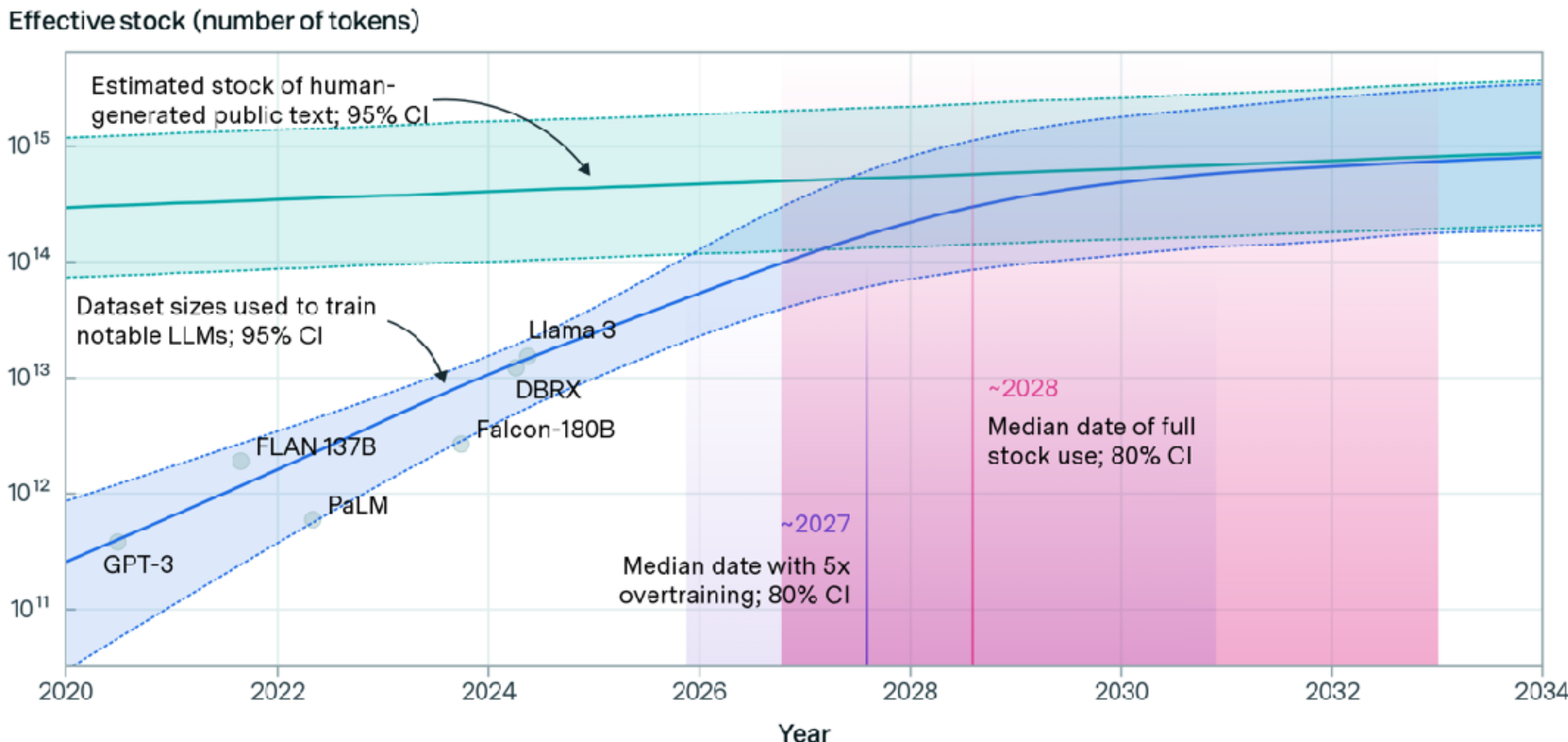


Knowledge recorded in rare languages

Goal: Improve data efficiency

Projections of the stock of public text and data usage

EPOCH AI



Whether it's running out of pretraining data, or niche domains, a common question:

How to improve the efficiency of using existing real data?

Setting: Learn knowledge from small domain

Teaching language model latest arXiv preprints, corporate proprietary documents, user's personal data (domains with only 1M-10M tokens of text).

Setting: Learn knowledge from small domain

Teaching language model latest arXiv preprints, corporate proprietary documents, user's personal data (domains with only 1M-10M tokens of text).

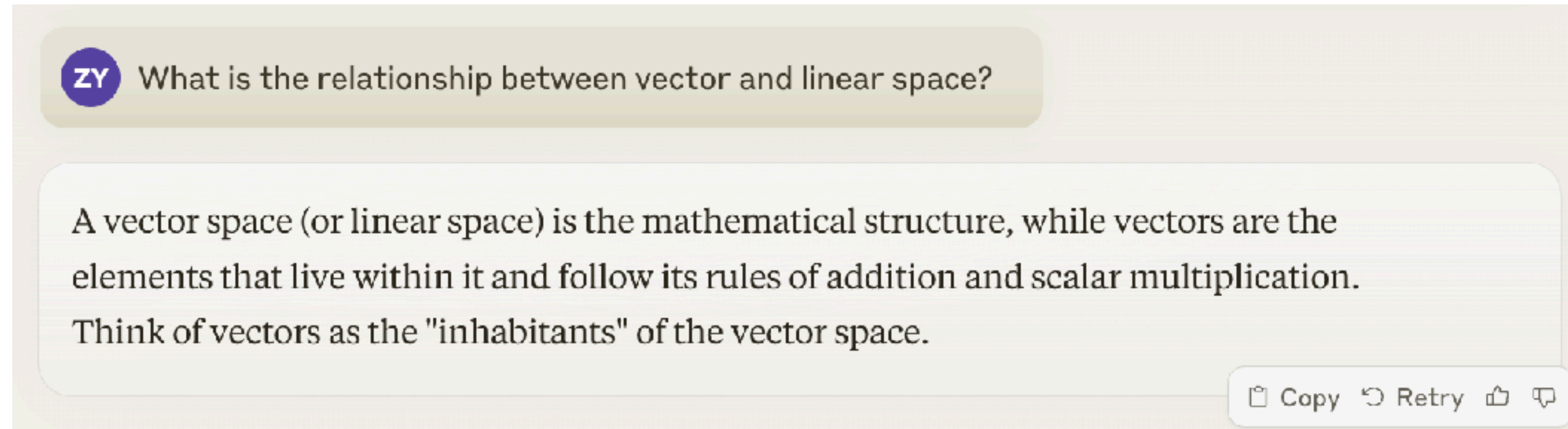
Natural attempt: continue pretraining with the available text in the target domain.

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

- Existing arts using continued pretraining in medicine, mathematics, law, etc.
- Domains with large corpus of text and diverse knowledge representation.

Consider a concrete example

Model knows about linear algebra



ZY What is the relationship between vector and linear space?

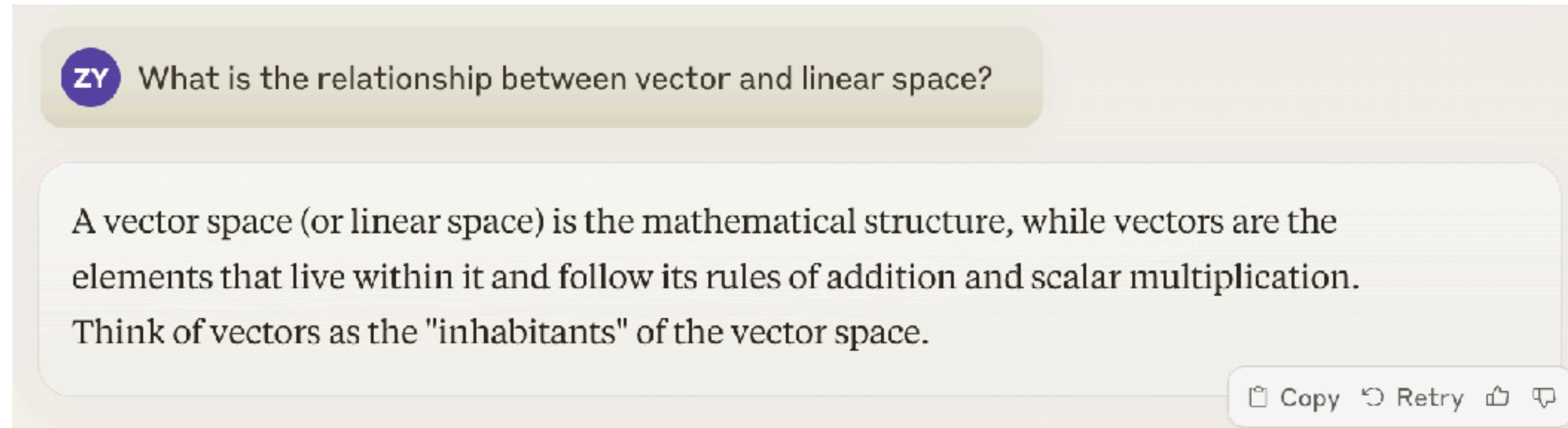
A vector space (or linear space) is the mathematical structure, while vectors are the elements that live within it and follow its rules of addition and scalar multiplication.
Think of vectors as the "inhabitants" of the vector space.

Copy Retry

For a new domain with a few arXiv papers, this is hard...

Consider a concrete example

Model knows about linear algebra



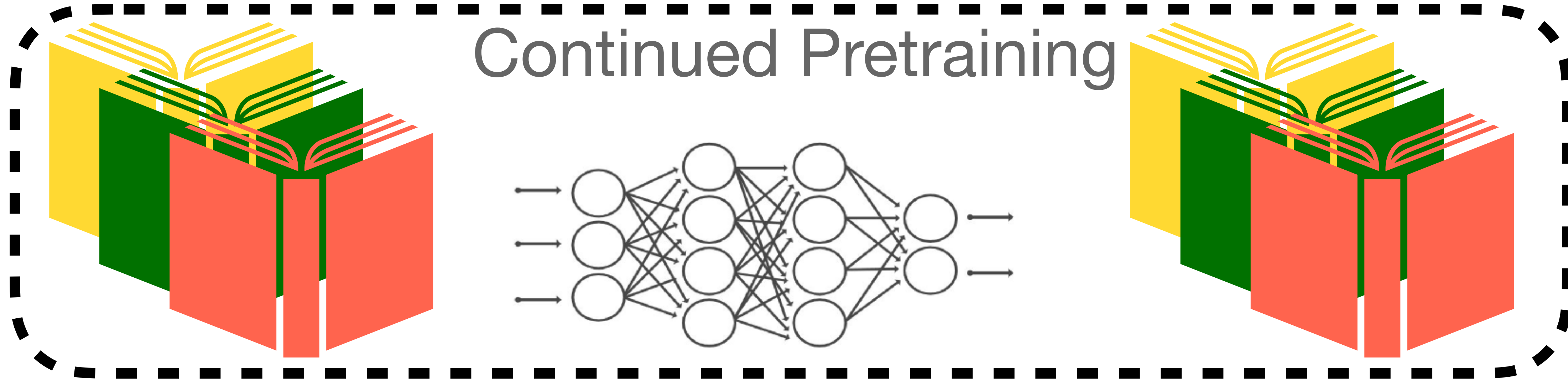
For a new domain with a few arXiv papers, this is hard...

Imagine how linear algebra related knowledge appear in pretraining.

- Many linear algebra textbook
- Reddit discussion of linear algebra exercise
- GitHub implementation of matrix completion :)
- ...

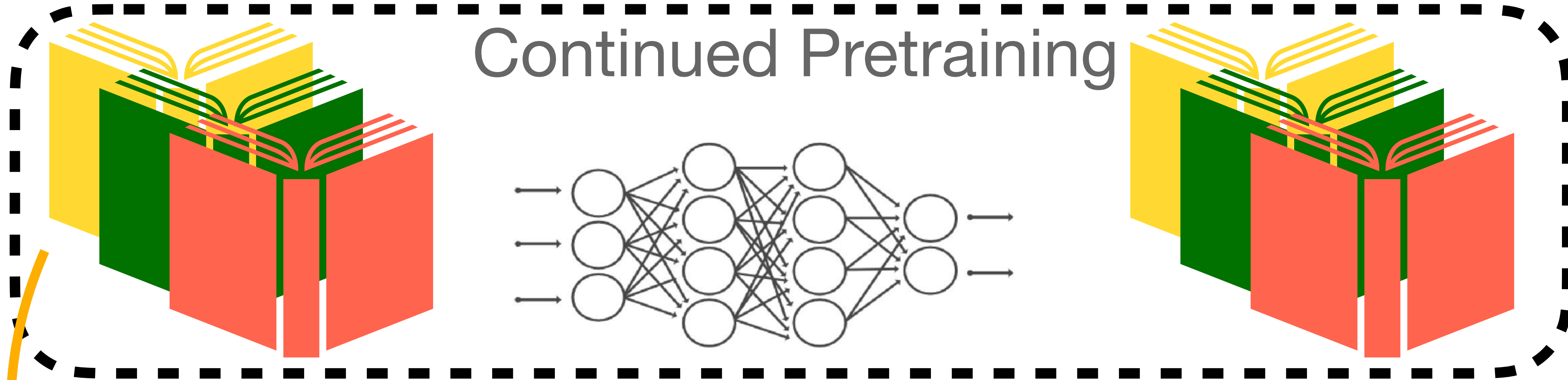


Synthetic Continued Pretraining



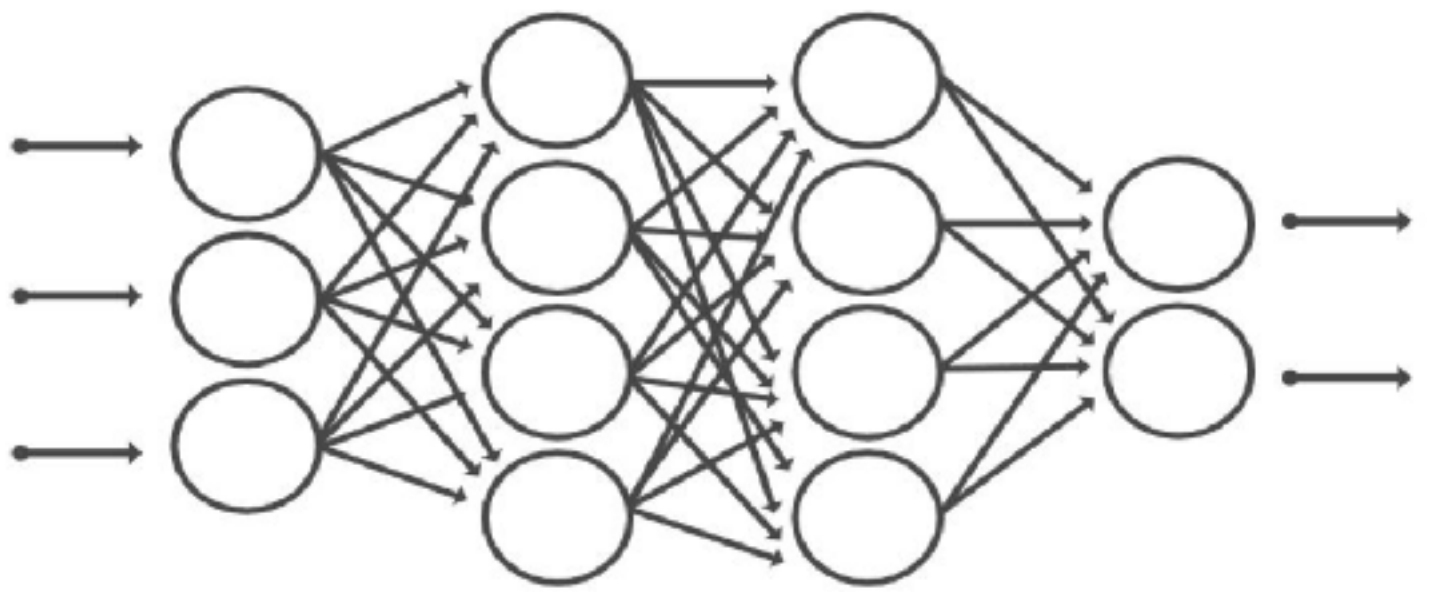
Synthetic Continued Pretraining

Synthetic data



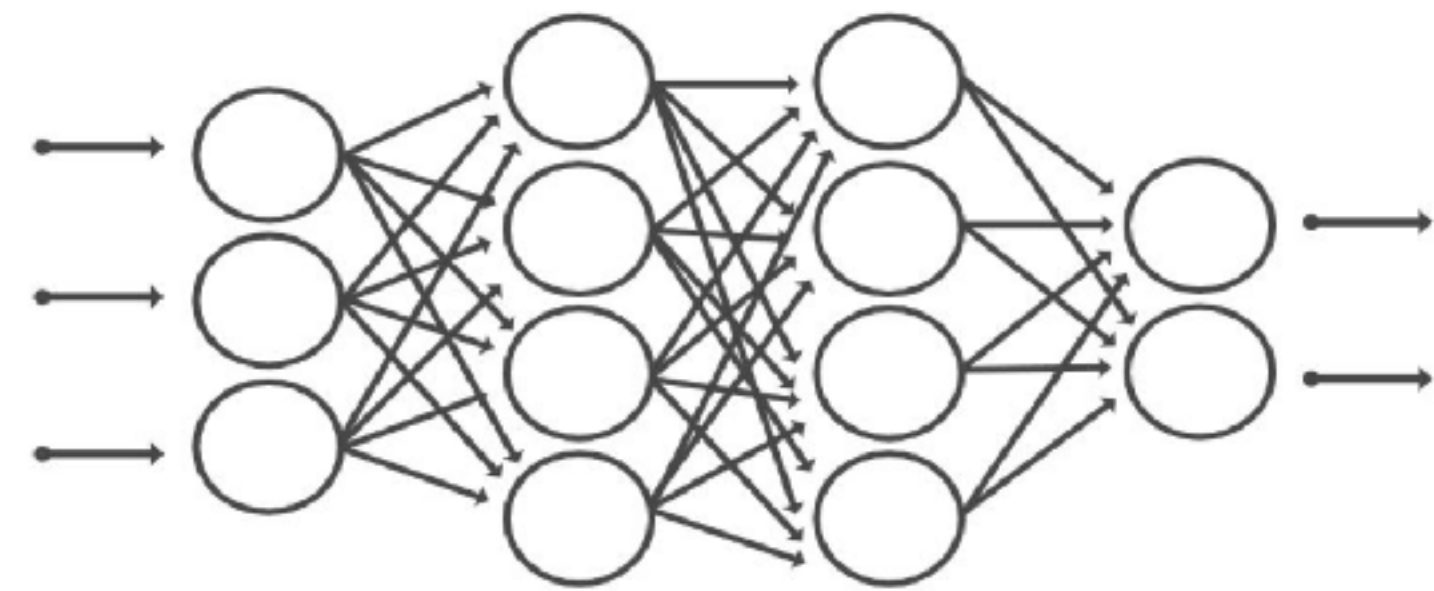
Synthetic Continued Pretraining

Continued Pretraining



Synthetic data

Synthetic Continued Pretraining



Concrete experiment data and benchmark

(i) Collections of niche books and articles; (ii) high quality Q&A for evaluation

Concrete experiment data and benchmark

(i) Collections of niche books and articles; (ii) high quality Q&A for evaluation

QuALITY Books

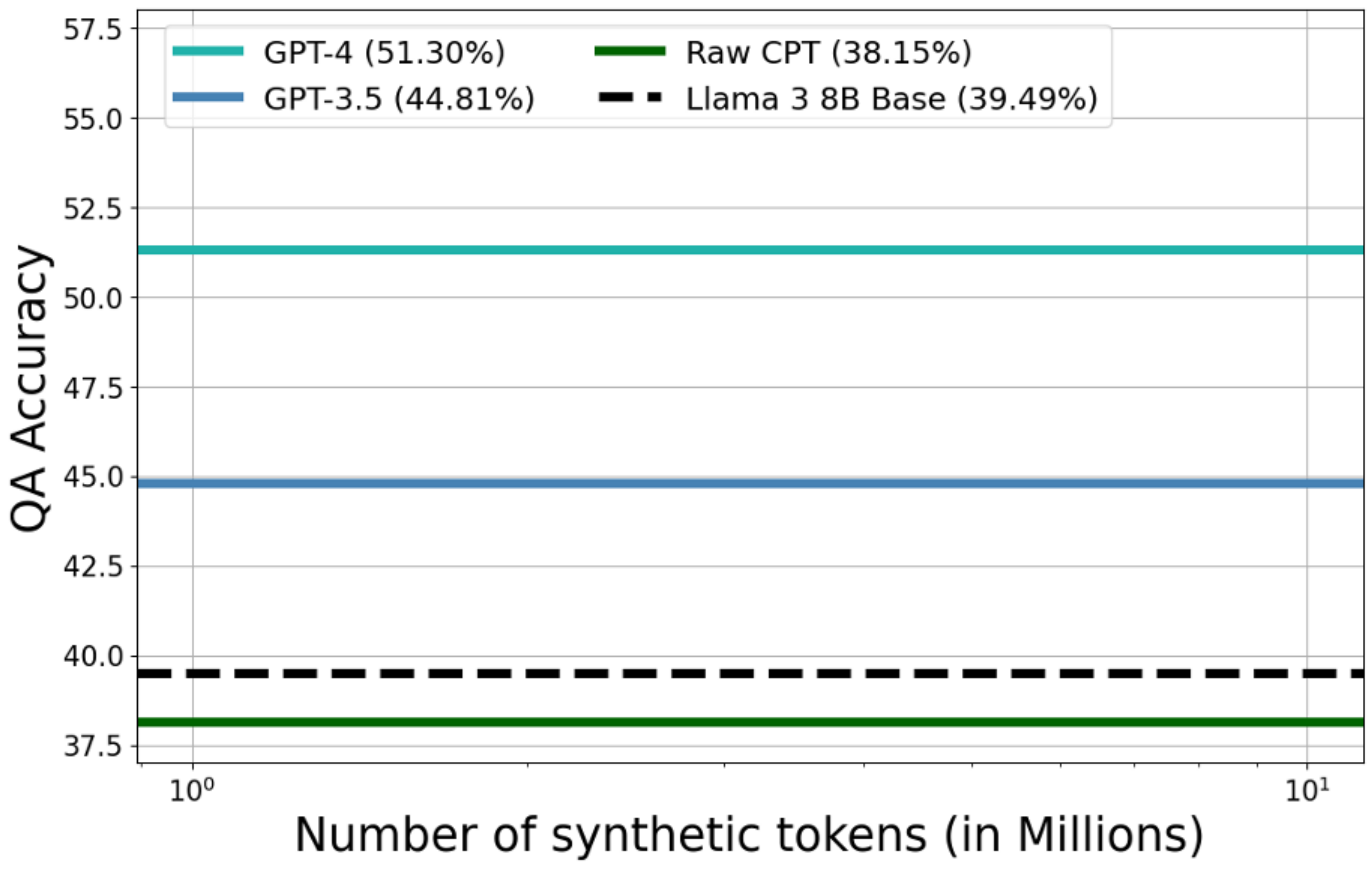


- Project Gutenberg fictions (mainly science fiction)
- Slate magazine articles
- The Long and Short, Freesouls, etc

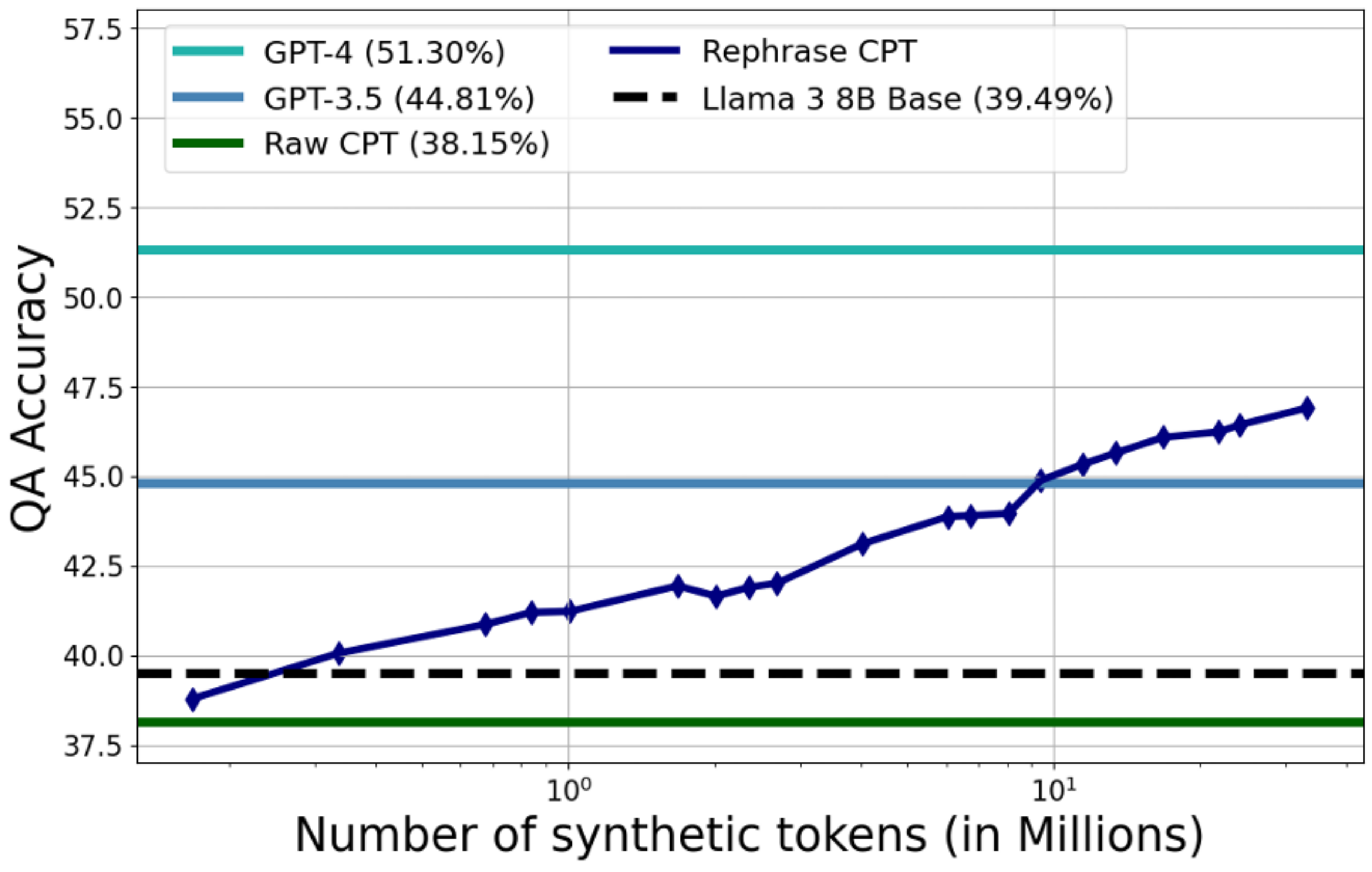
QuALITY [Pang+ '21]

- 265 niche books totaling 1.8M tokens.
- High quality multiple choice Q&A .
- Human written summarization.
- Infrequent appearance in pretraining corpus
- GPT-4 accuracy 51%. Llama 3 accuracy 49%.

First attempt: directly training on 1.8M tokens

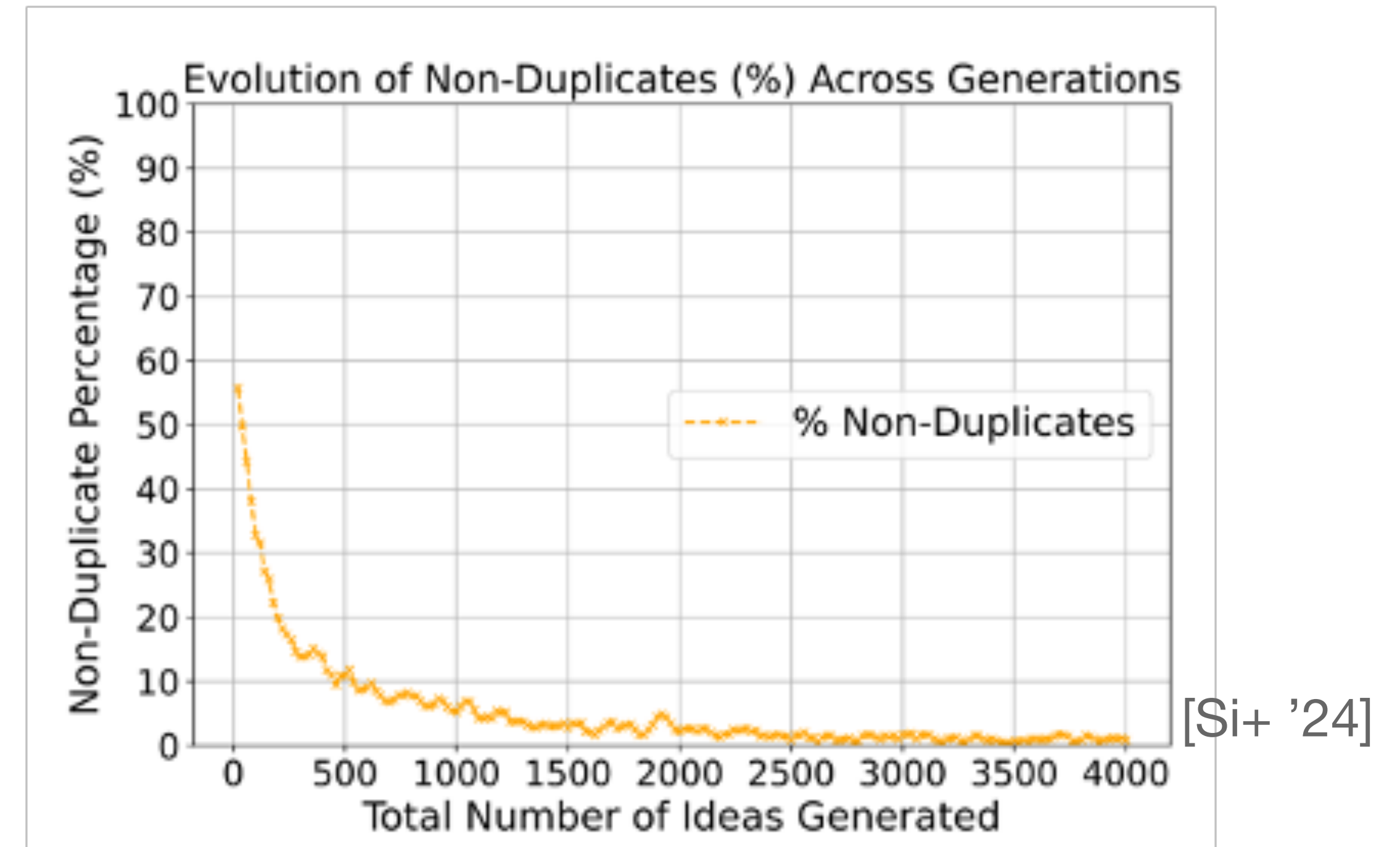
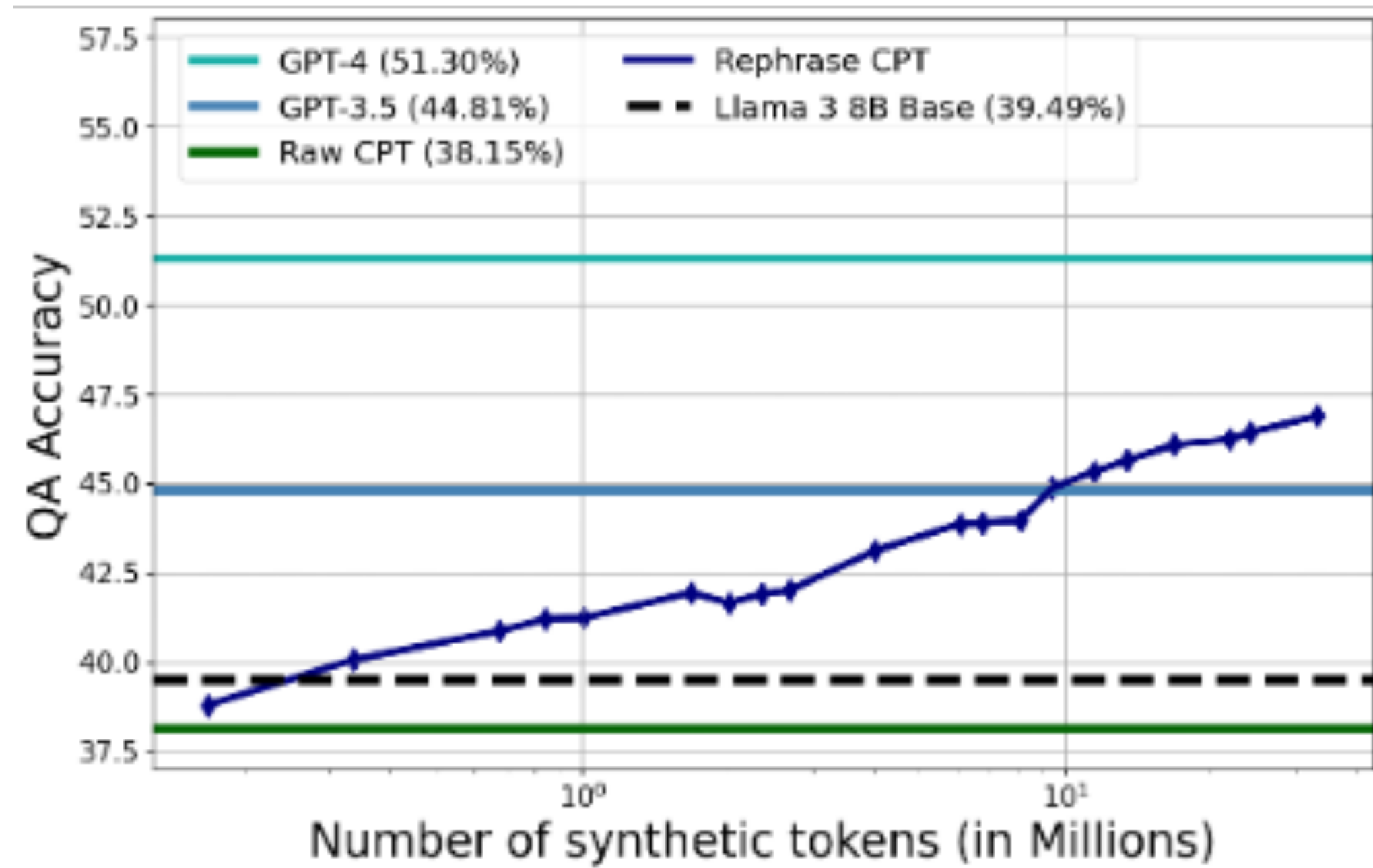


Second attempt: simply rephrase the data



Idea: Introducing external diversity

- Repetitively rephrasing the book content can teach model new knowledge, but efficiency is very low



- **Core limitation:** Language model do not have good built-in diversity.

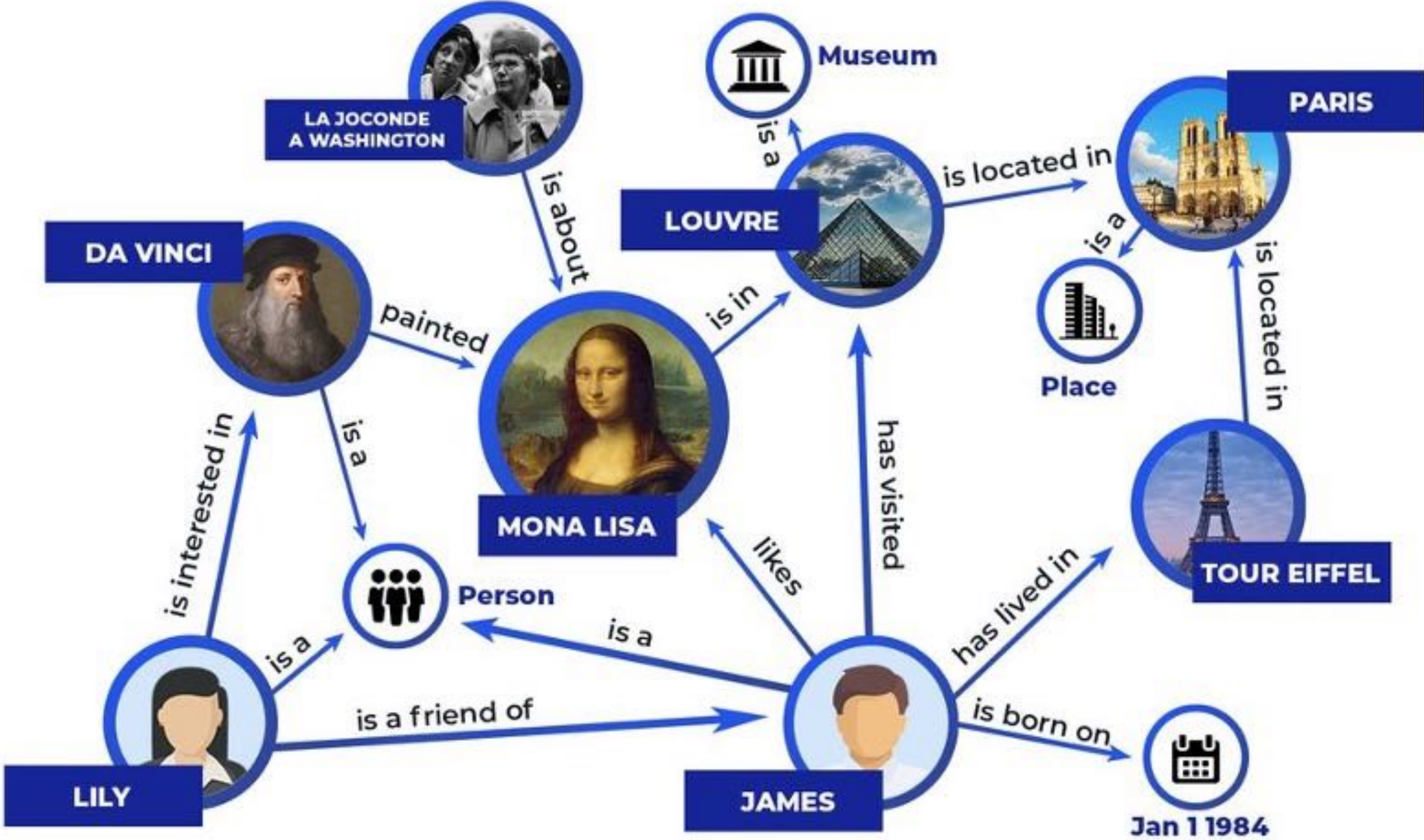
EntiGraph

Using the knowledge graph associated with the article to boost prompt diversity.

EntiGraph

Using the knowledge graph associated with the article to boost prompt diversity.

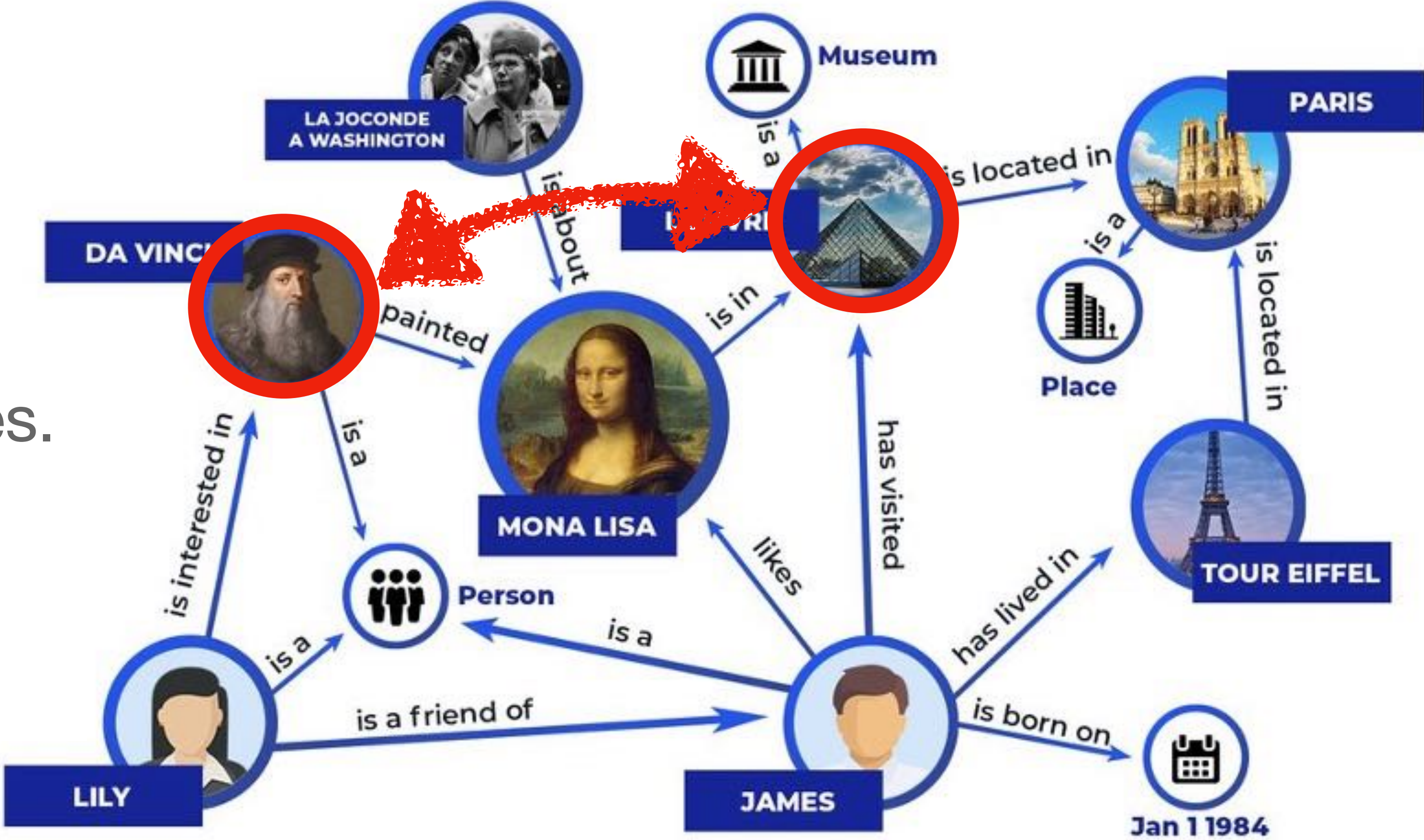
- First as the model to extract salient entities.



EntiGraph

Using the knowledge graph associated with the article to boost prompt diversity.

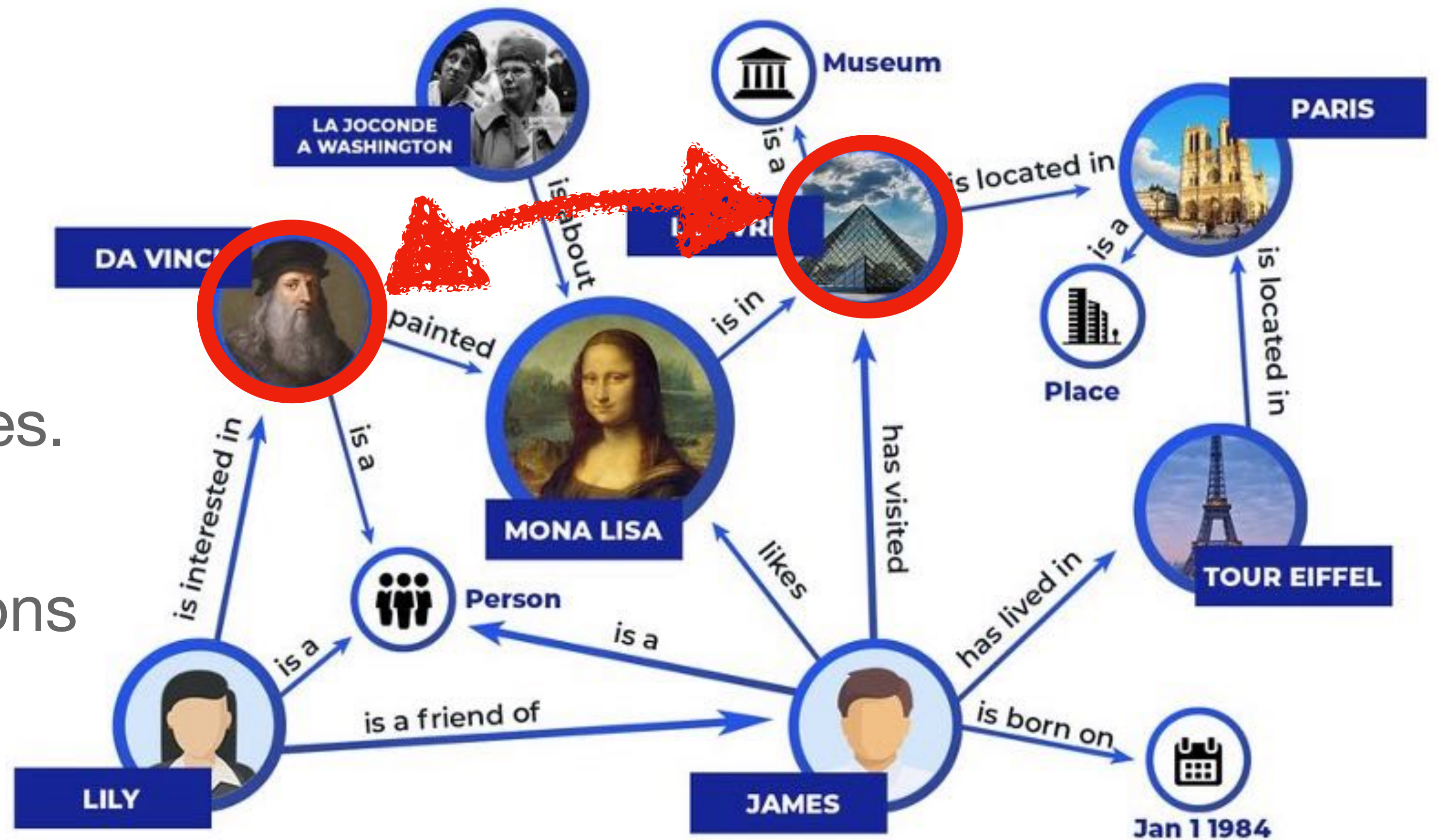
- First as the model to extract salient entities.
- Randomly subsample k entities.



EntiGraph

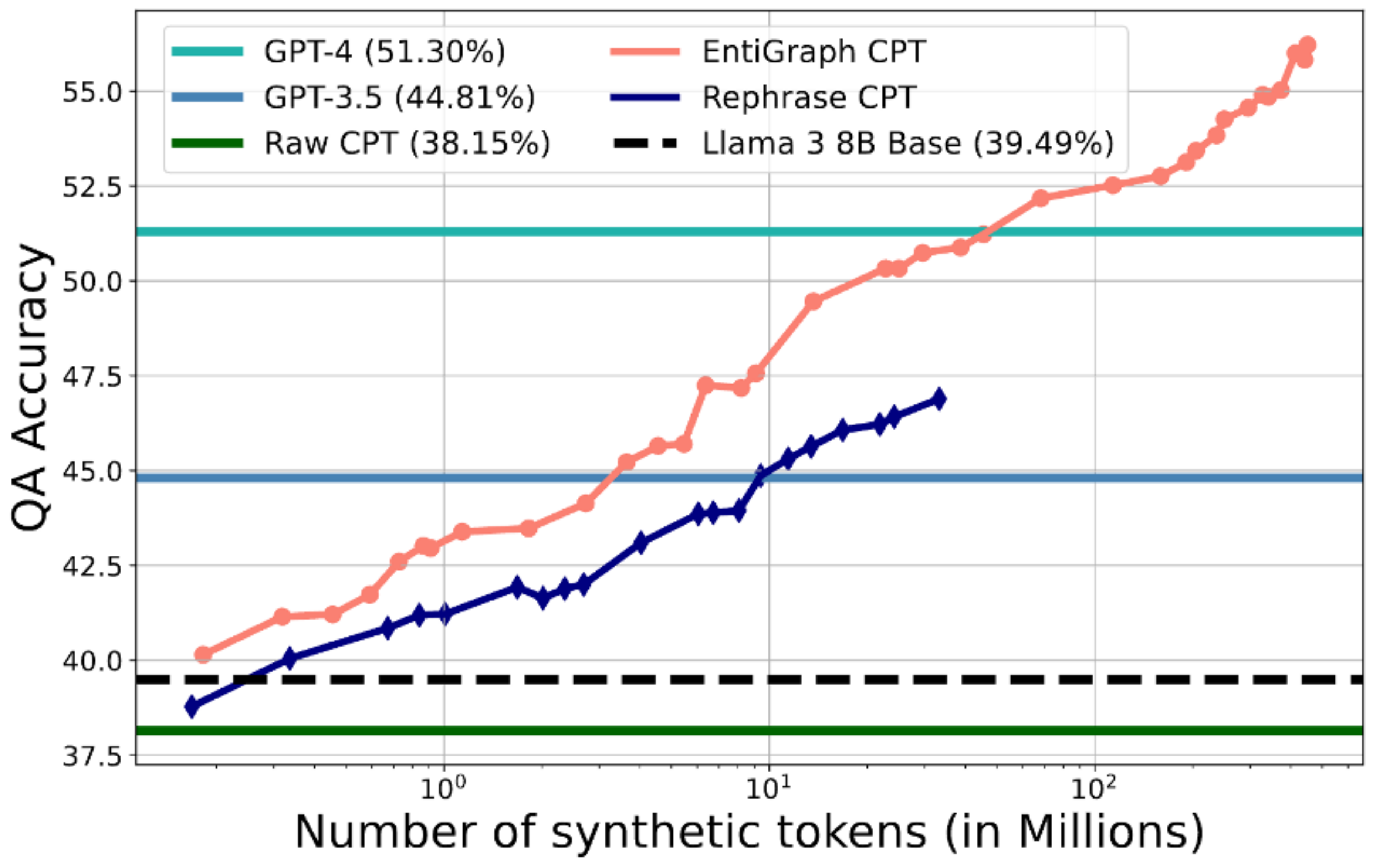
Using the knowledge graph associated with the article to boost prompt diversity.

- First as the model to extract salient entities.
- Randomly subsample k entities.
- Let model generate descriptions about sampled entities.



EntiGraph synthesized data: “Many of da Vinci's works are kept in the Louvre”

EntiGraph performance (no book access)



EntiGraph generalize to instruction tuning

-
- **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: [...]
-

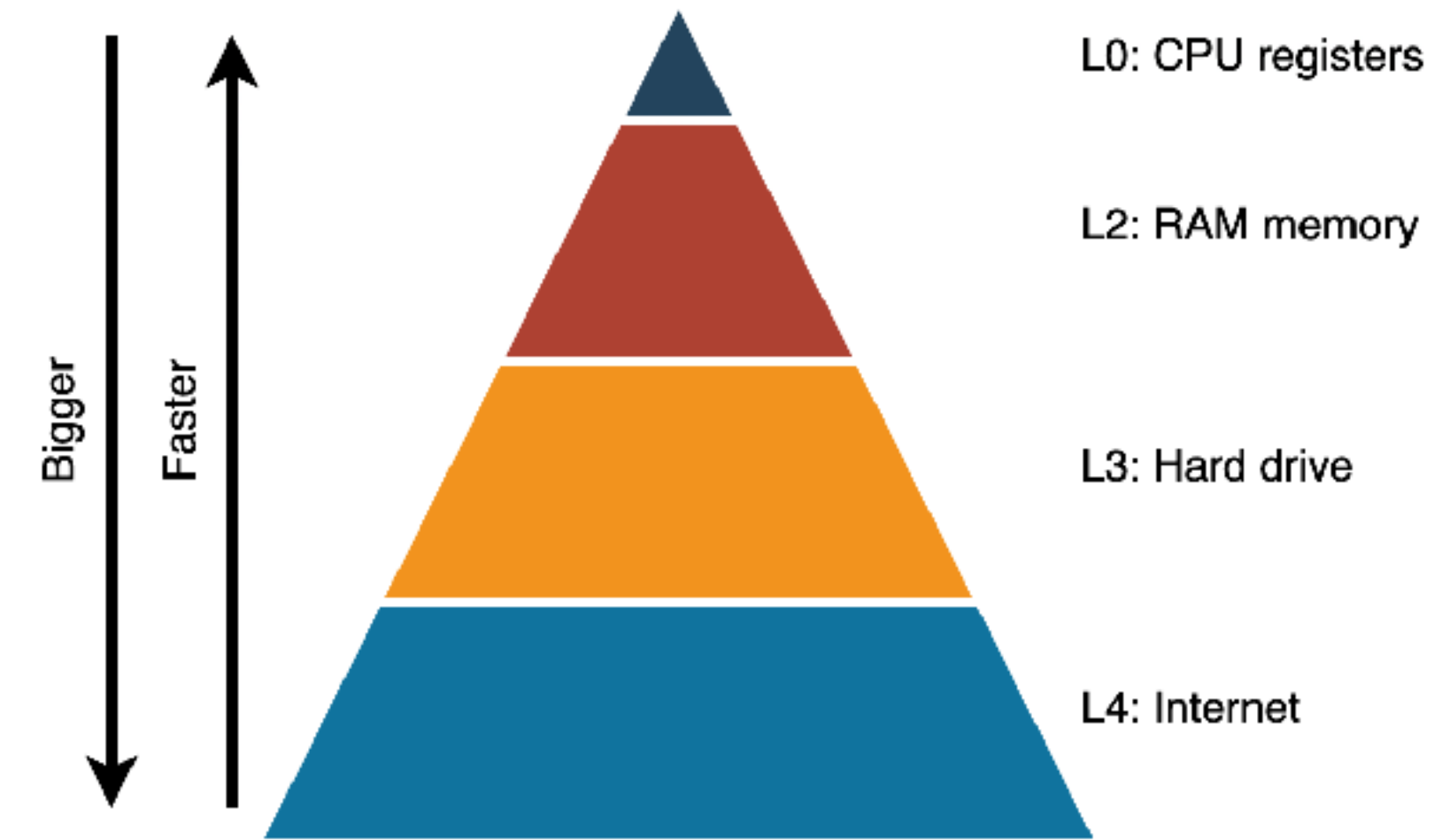
- **Cross article instruction:** Compare David Plotz’s commentary style between how he analyzed American dentistry and how he discussed the movie Fight Club?
-

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. [...]

Synthetic continued pretraining

Allowing models to learn new knowledge

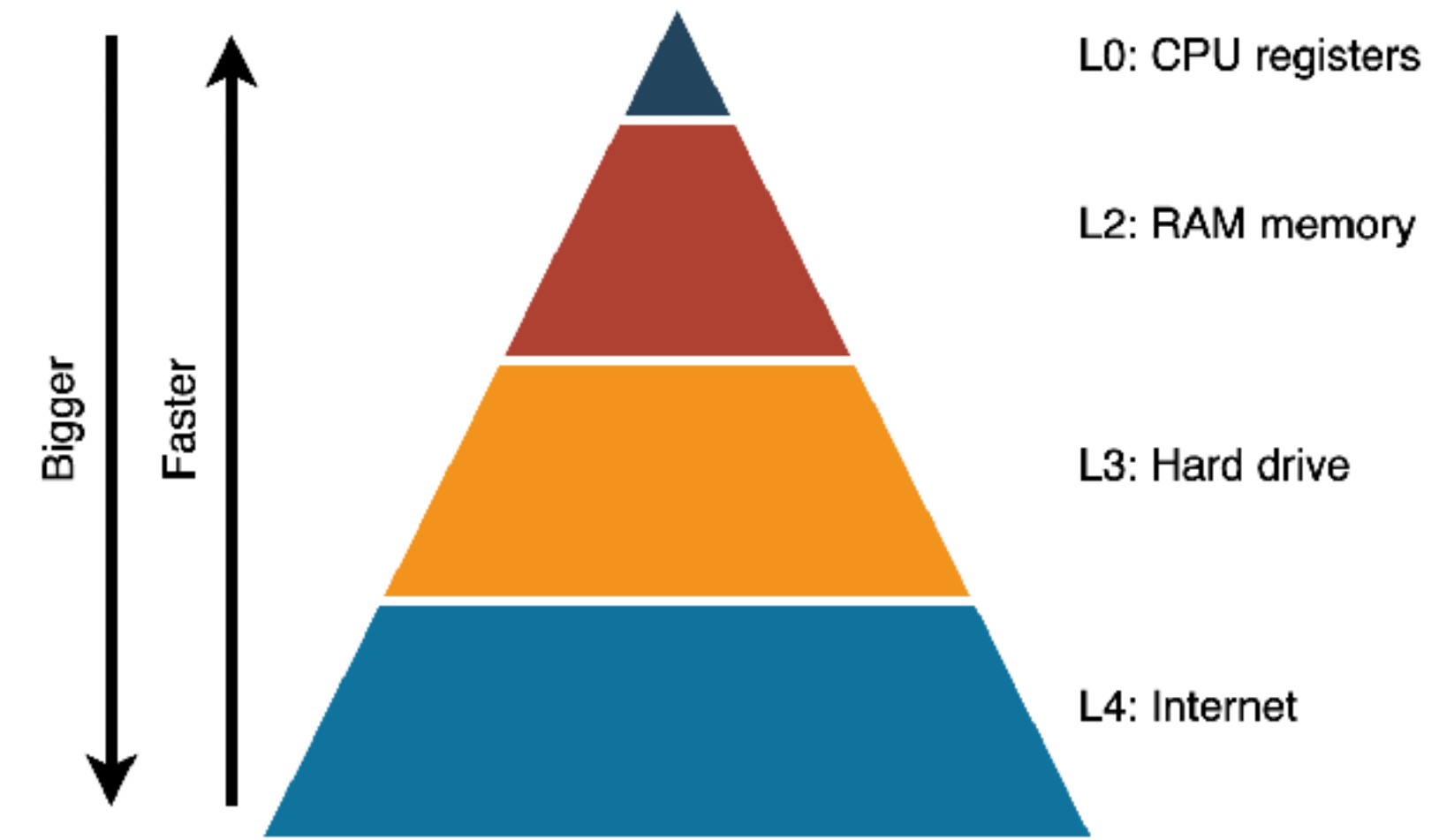
- Looking at history, many revolutionary ideas revolves around the concept “*memory*” (RAM, Internet, etc.)



Synthetic continued pretraining

Allowing models to learn new knowledge

- Looking at history, many revolutionary ideas revolve around the concept “*memory*” (RAM, Internet, etc.)
- Same in AI, Synthetic Continued Pretraining offers a way for model to modify its lower level knowledge.



Synthetic continued pretraining

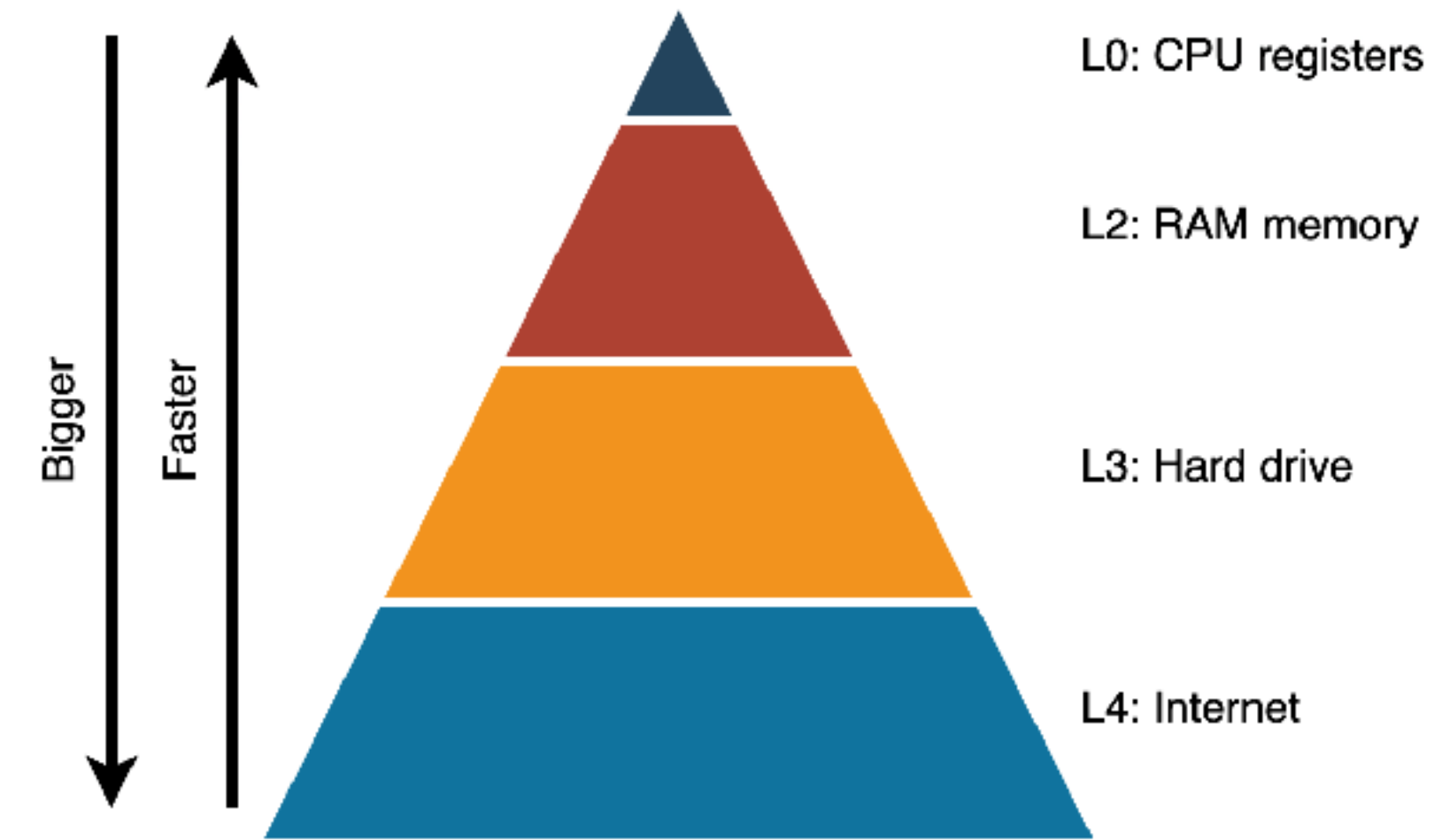
Allowing models to learn new knowledge

- Looking at history, many revolutionary ideas revolve around the concept “*memory*” (RAM, Internet, etc.)
- Same in AI, Synthetic Continued Pretraining offers a way for model to modify its lower level knowledge.

If models can learn on their own, what's our role?

- Human can offer “enlightenment text”

$$P_{\text{LM}}(\text{define } \sqrt{-1} \text{ as root of } x^2 + 1 \mid \langle \text{all human text before 1572} \rangle) = 10^{-60}$$



Synthetic continued pretraining

Allowing models to learn new knowledge

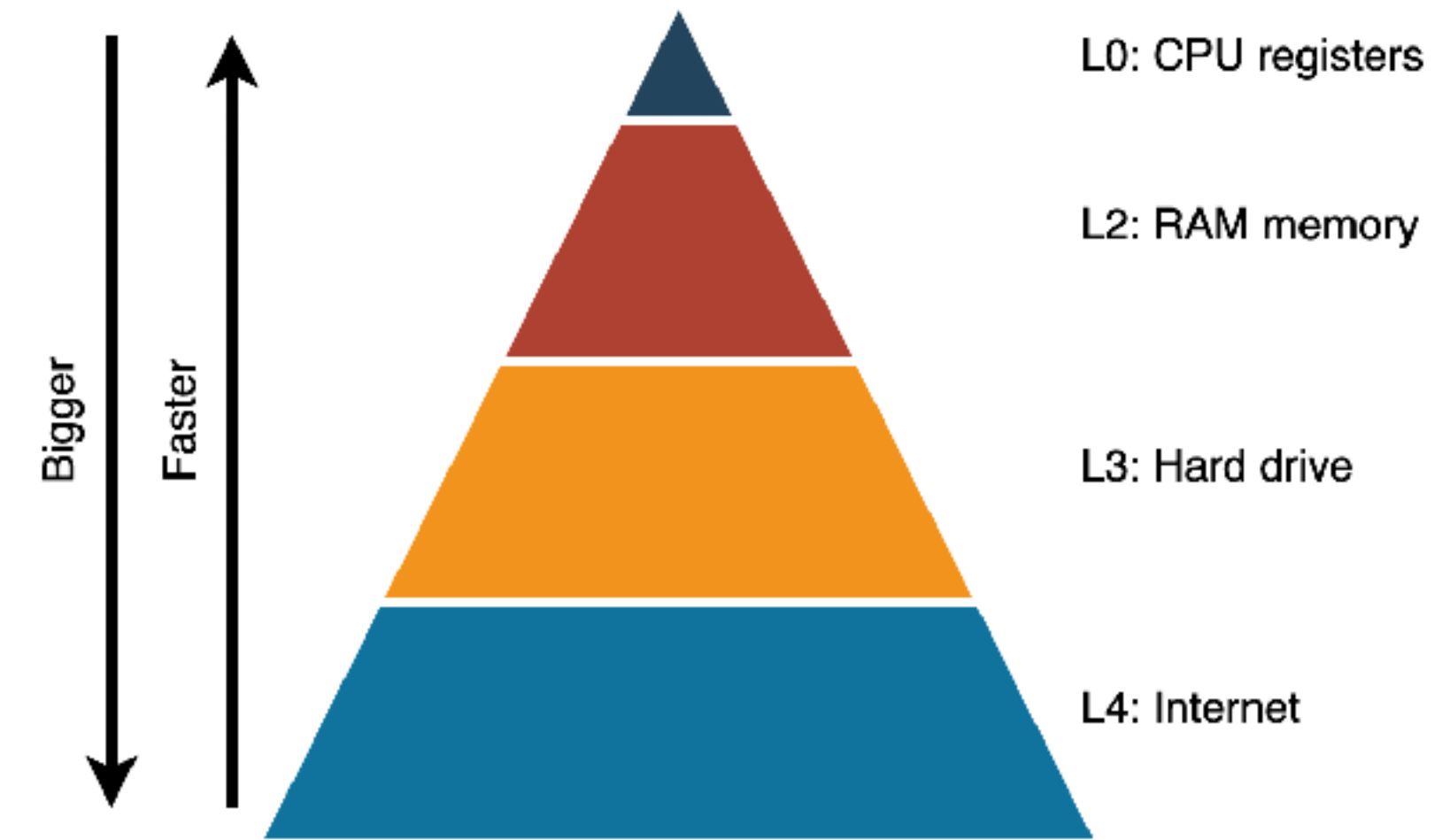
- Looking at history, many revolutionary ideas revolve around the concept “*memory*” (RAM, Internet, etc.)
- Same in AI, Synthetic Continued Pretraining offers a way for model to modify its lower level knowledge.

If models can learn on their own, what's our role?

- Human can offer “enlightenment text”

$$P_{\text{LM}}(\text{define } \sqrt{-1} \text{ as root of } x^2 + 1 \mid \langle \text{all human text before 1572} \rangle) = 10^{-60}$$

- Machines can infer the consequence of the “enlightenment text”, and learn on their own.



Synthetic continued pretraining

Allowing models to learn new knowledge

- Looking at history, many revolutionary ideas revolves around the concept “*memory*” (RAM, Internet, etc.)
- Same in AI, Synthetic Continued Pretraining offers a way for model to modify it’s lower level knowledge.

If models can learn on their own, what’s our role?

- Human can offer “enlightenment text”

$$P_{\text{LM}}(\text{define } \sqrt{-1} \text{ as root of } x^2 + 1 \mid \langle \text{all human text before 1572} \rangle) = 10^{-60}$$

- Machines can infer the consequence of the “enlightenment text”, and learn on their own.
- However, machine has little hope of generating the “enlightenment text” as their distribution are too different from pretraining.

