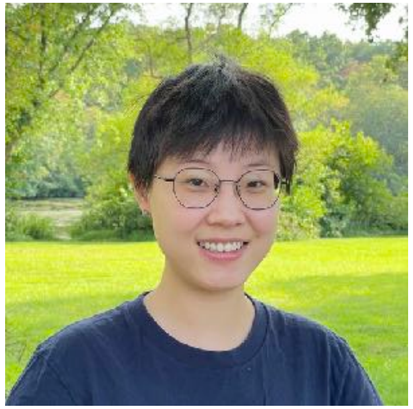
Synthetic Continued Pretraining

Zitong Yang

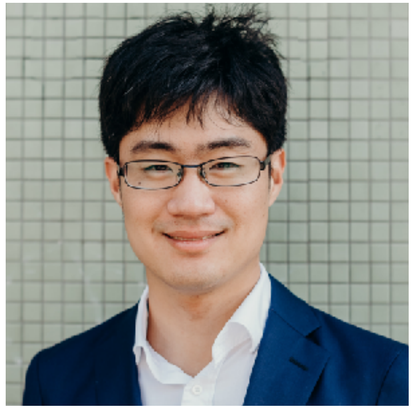
Stanford Statistics

Neil Band*



Shuangping Li

Emmanuel Candès



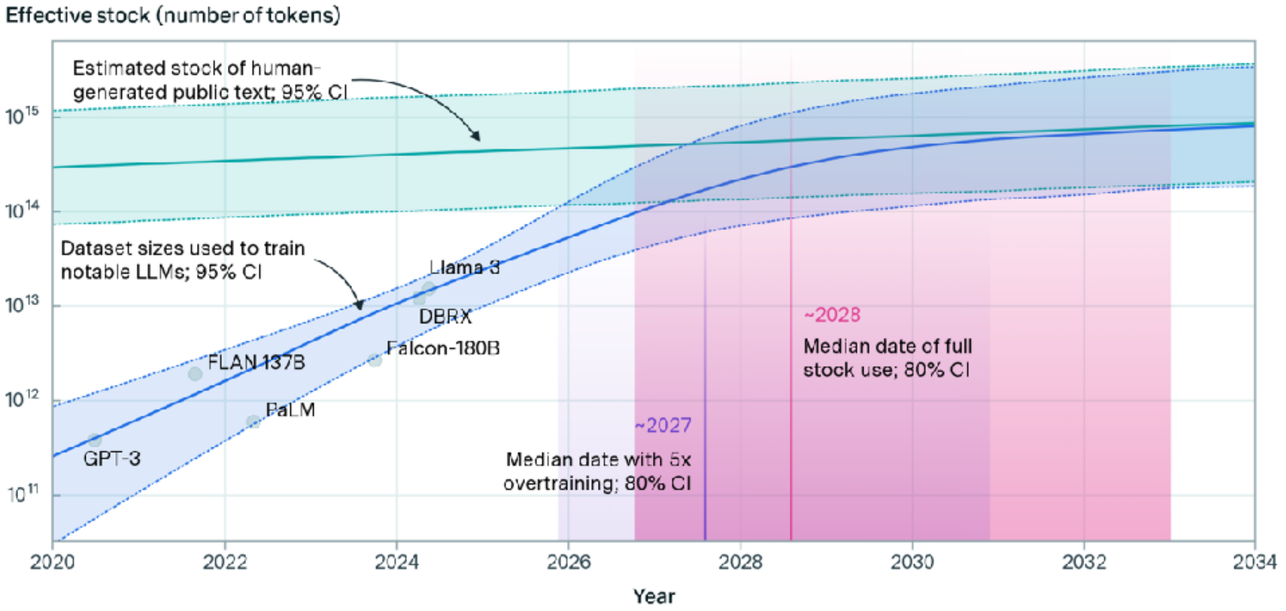
Tatsunori Hashimoto

ICLR 2025, Oral, https://arxiv.org/abs/2409.07431

**Equal contribution*

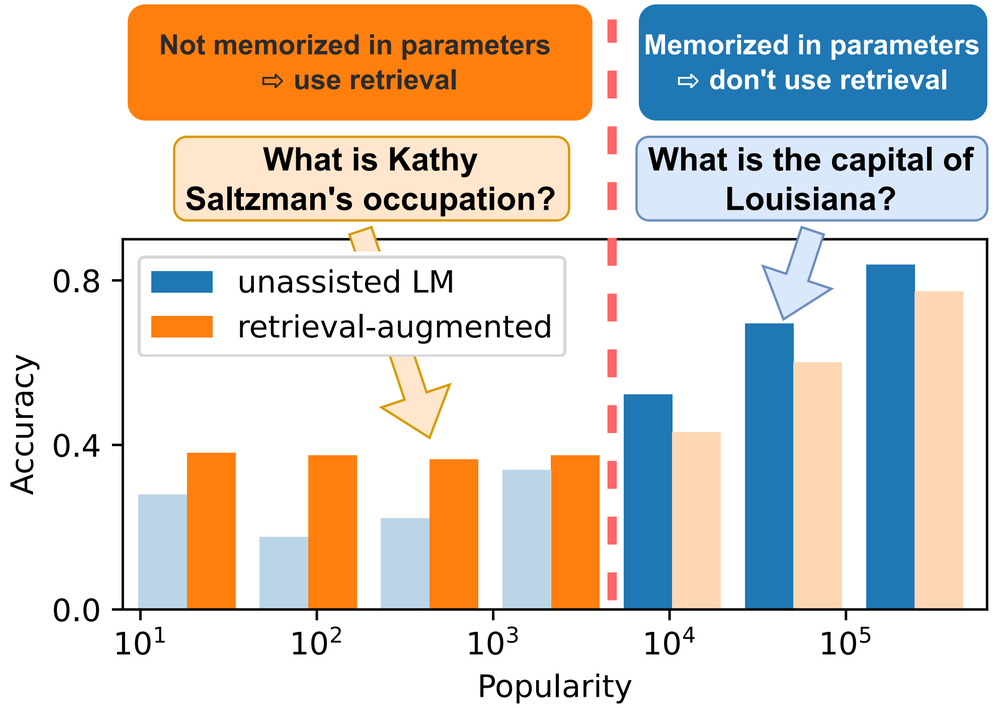
Approaching data constraints for LLM pretraining

- Public internet text is the “fossil fuel” powering AI
- The growth of compute is outpacing the growth of internet data
- “We have but one internet” (Sutskever)
- We will enter a **data-constrained regime** as early as 2028

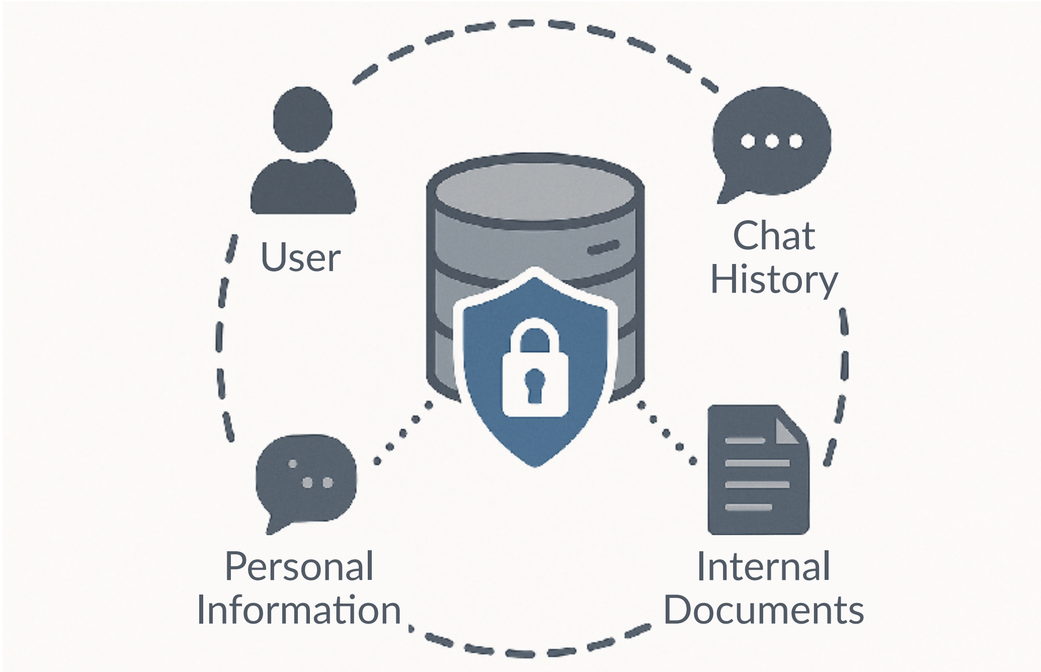


“Pretraining as we know it will end”

Current pretraining is remarkably data-inefficient



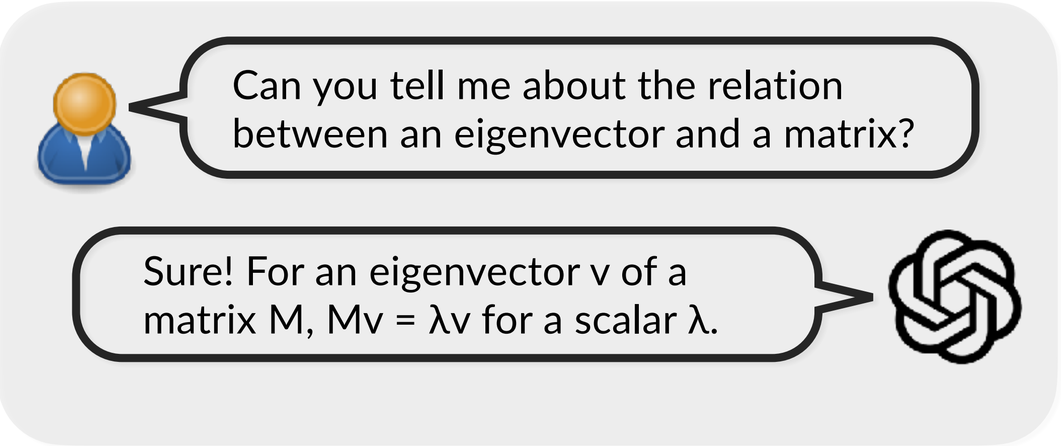
Models fail to learn long-tailed facts appearing <1K times in a corpus...

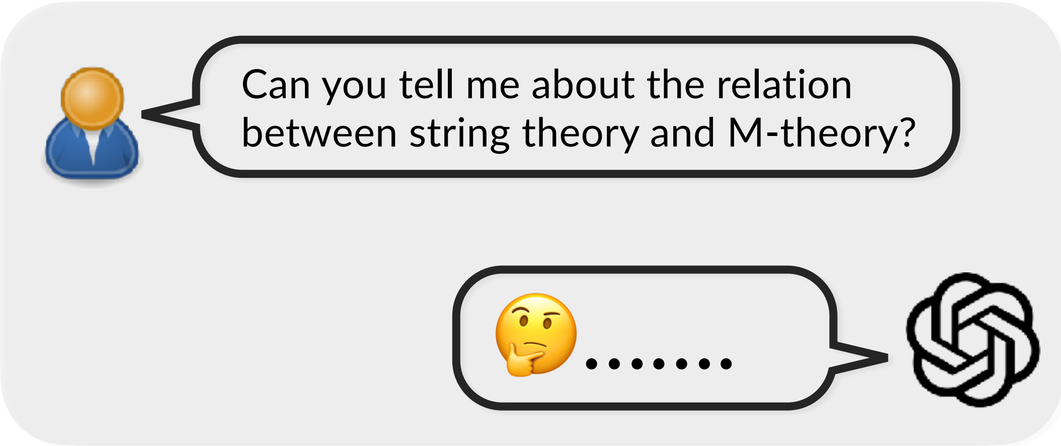


... so they struggle to learn from the unique knowledge in siloed, proprietary datasets

A concrete example of data inefficiency

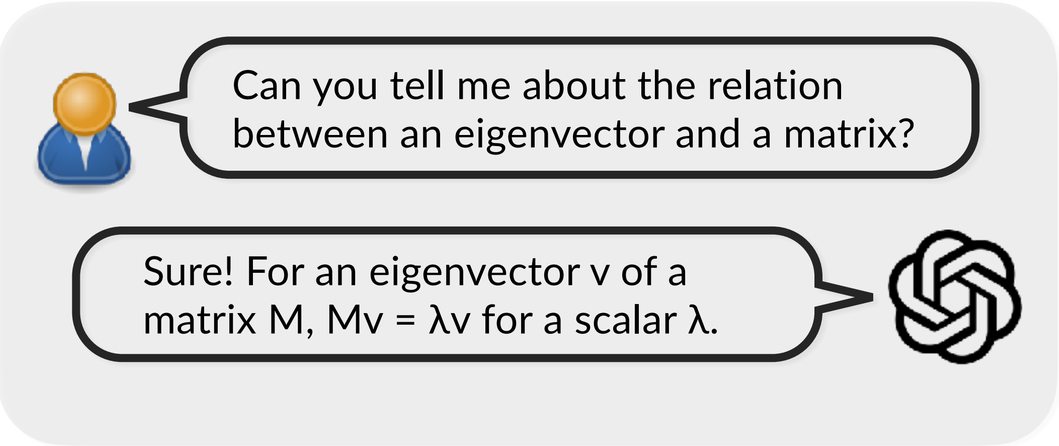
- LLMs know linear algebra, but not niche domains like quantum gravity, or user personalization

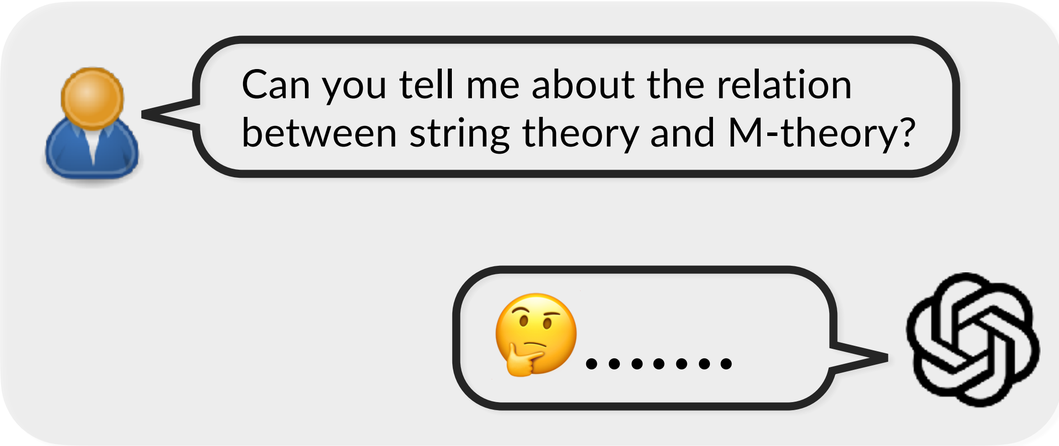




A concrete example of data inefficiency

- LLMs know linear algebra, but not niche domains like quantum gravity, or user personalization



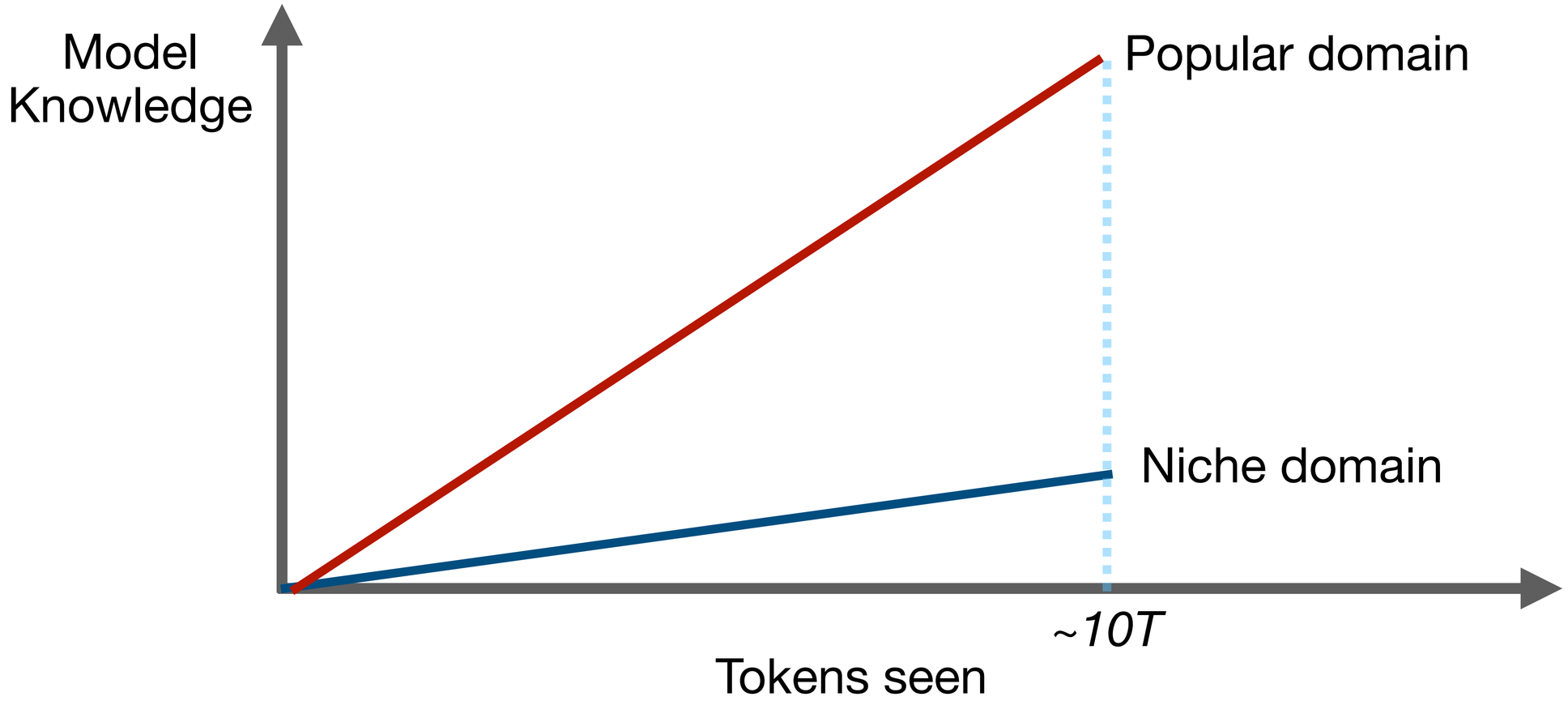


- LLMs learn linear algebra by acquiring knowledge from **diversely presented** web data

- Many textbooks, lecture notes about linear algebra
- Online discussions of linear algebra exercises
- GitHub implementations of the SVD…

Data efficiency from the scaling perspective

- Consequence of poor pretraining data efficiency: performance gap in popular vs. niche domains
- Can we continually pretrain a model to bridge this gap?



Synthetic continued pretraining

Goal: teach an LM the knowledge of a niche domain with few “source” documents.

Step 1: Generate synthetic text based on source documents

Step 2: Pretrain/finetune the model on generated text

Synthetic continued pretraining

Goal: teach an LM the knowledge of a niche domain with few “source” documents.

Step 1: Generate synthetic text based on source documents

Step 2: Pretrain/finetune the model on generated text

Experiment setup

- **A corpus of niche documents (not something the base LM knows)**
- **A task that tests the LM’s knowledge about the source documents**

A dataset and a benchmark

QuALITY Books

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

QuALITY [Pang+ ’21]

- 265 *specialized* books, 1.8M tokens (infrequent in pretraining corpus)
- High-quality multiple-choice Q&As
- Prompt LM to answer without the book in-context (closed-book)

Synthetic continued pretraining

Goal: teach an LM the knowledge of a niche domain with few “source” documents.

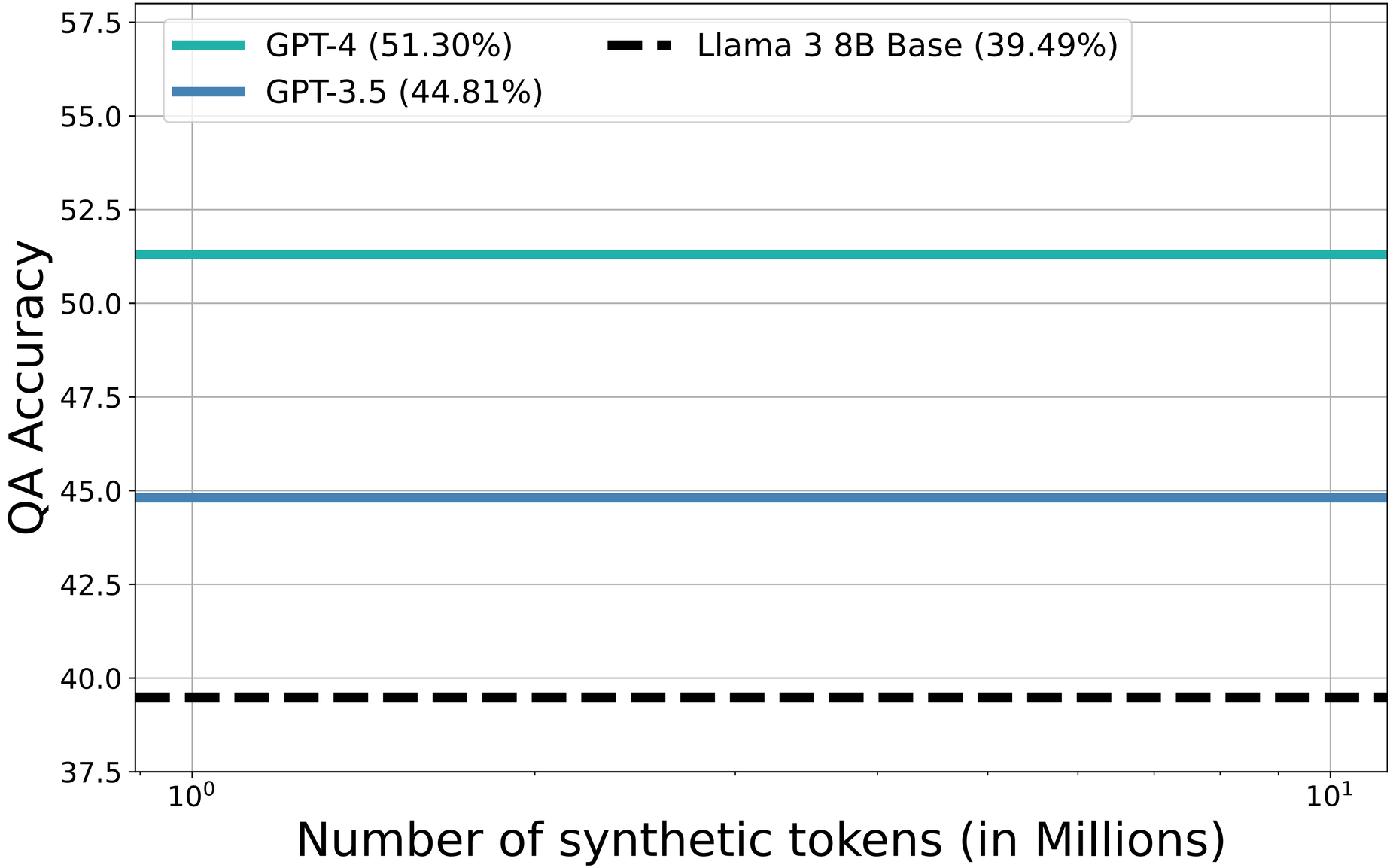
Step 1: Generate synthetic text based on source documents

Step 2: Pretrain/finetune the model on generated text

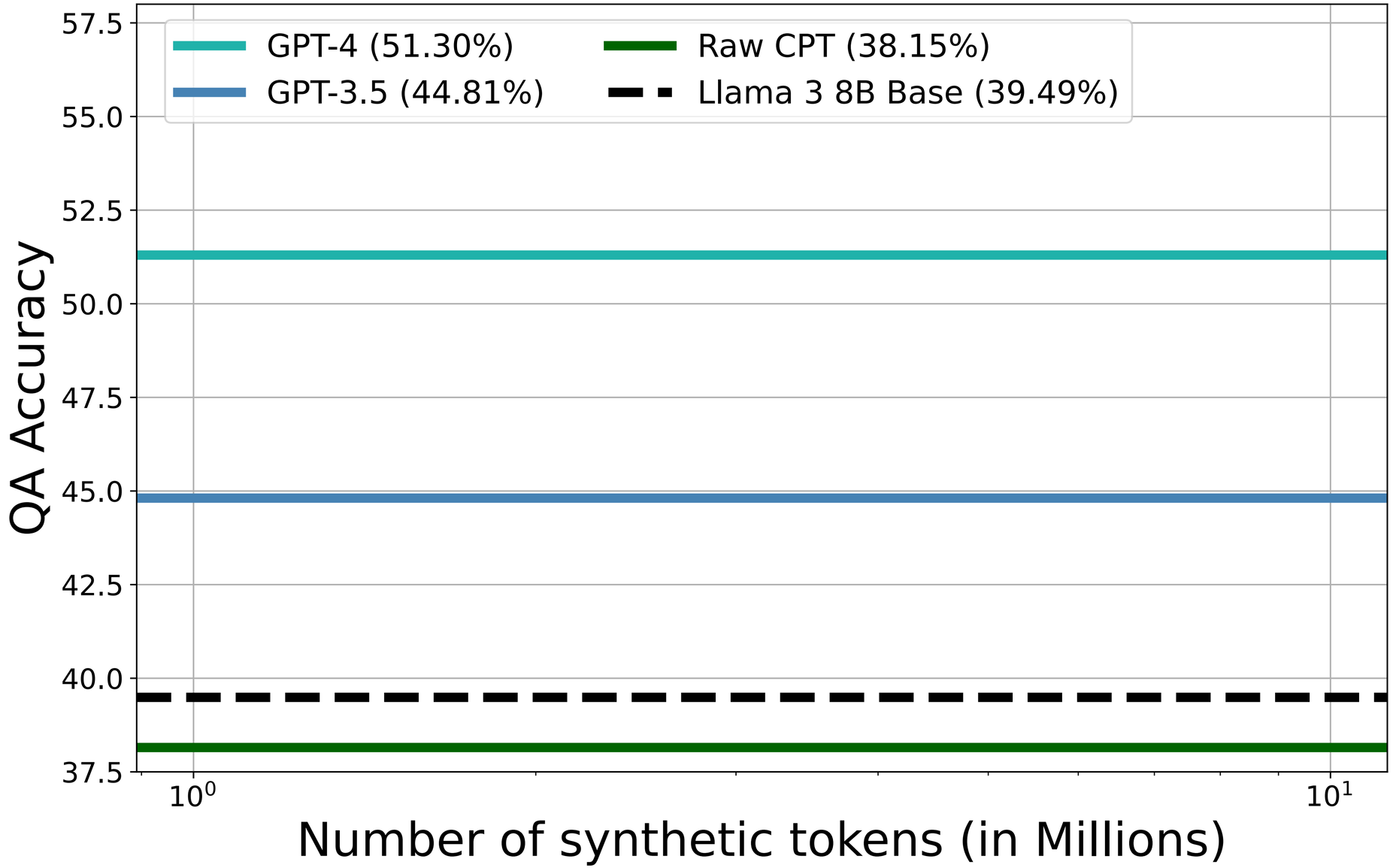
Experiment setup

- A corpus of niche documents (not something the base LM knows): **QuALITY books**
- A task that tests the LM’s knowledge about the source documents: **Closed-book QA**

Base models perform poorly on the niche corpus



Simply repeating the data fails to learn



Synthetic continued pretraining

Goal: teach an LM the knowledge of a niche domain with few “source” documents.

Step 1: Generate synthetic text based on source documents

Step 2: Pretrain/finetune the model on generated text

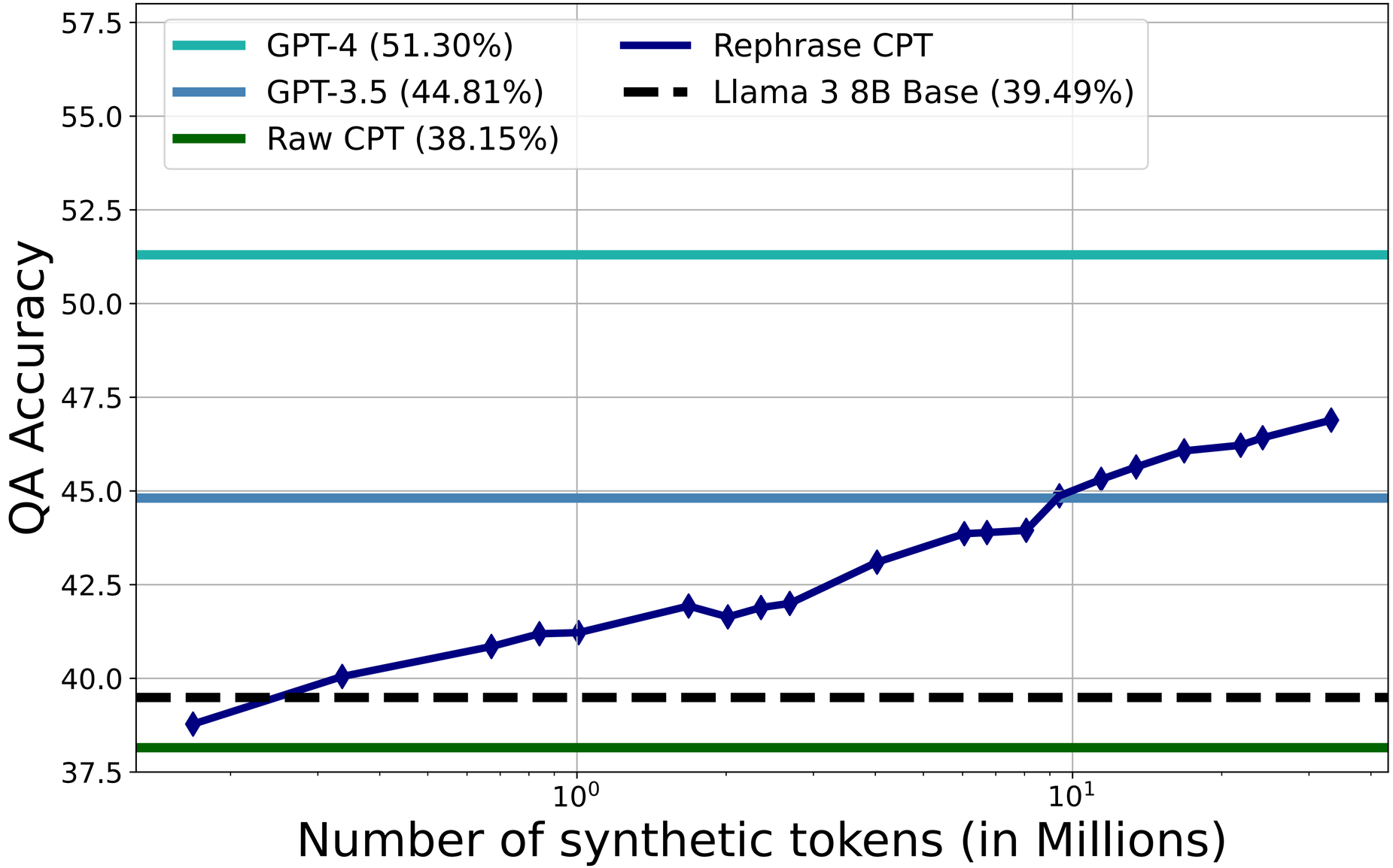
Experiment setup

- A corpus of niche documents (not something the base LM knows): QuALITY books
- A task that tests the LM’s knowledge about the source documents: Closed-book QA

How to generate synthetic data?

- **Baseline: simply rephrase the document [Maini et al. 2024]**

Rephrase baseline, closed-book evaluation



Synthetic continued pretraining

Goal: teach an LM the knowledge of a niche domain with few “source” documents.

Step 1: Generate synthetic text based on source documents

Step 2: Pretrain/finetune the model on generated text

Experiment setup

- A corpus of niche documents (not something the base LM knows): QuALITY books
- A task that tests the LM’s knowledge about the source documents: Closed-book QA

How to generate synthetic data?

- Baseline: simply rephrase the document [Maini et al. 2024]: **Limited diversity**

Synthetic continued pretraining

Goal: teach an LM the knowledge of a niche domain with few “source” documents.

Step 1: Generate synthetic text based on source documents

Step 2: Pretrain/finetune the model on generated text

Experiment setup

- A corpus of niche documents (not something the base LM knows): QuALITY books
- A task that tests the LM’s knowledge about the source documents: Closed-book QA

How to generate synthetic data?

- Baseline: simply rephrase the document [Maini et al. 2024]: Limited diversity
- **EntiGraph: Entity graph-based synthetic data generation**

EntiGraph: seed data diversity with an entity graph

...

Title: The Blue Behemoth

Title: Cosmic Yo-Yo

Title: Defining Decay Down
Author: David Plotz

If you haven't visited a dentist in the past few years, first of all, that's gross. (Checkups are every six months, and don't pretend you…

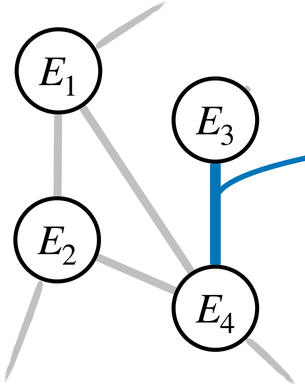
Input: small, niche corpus of documents

(1) Entity Extraction
For each document, extract a list of entities

E_1 Dentist E_2 Fluoride

E_3 Enamel ...

(2) Relation Analysis
Form a knowledge graph and prompt an LM to describe its edges

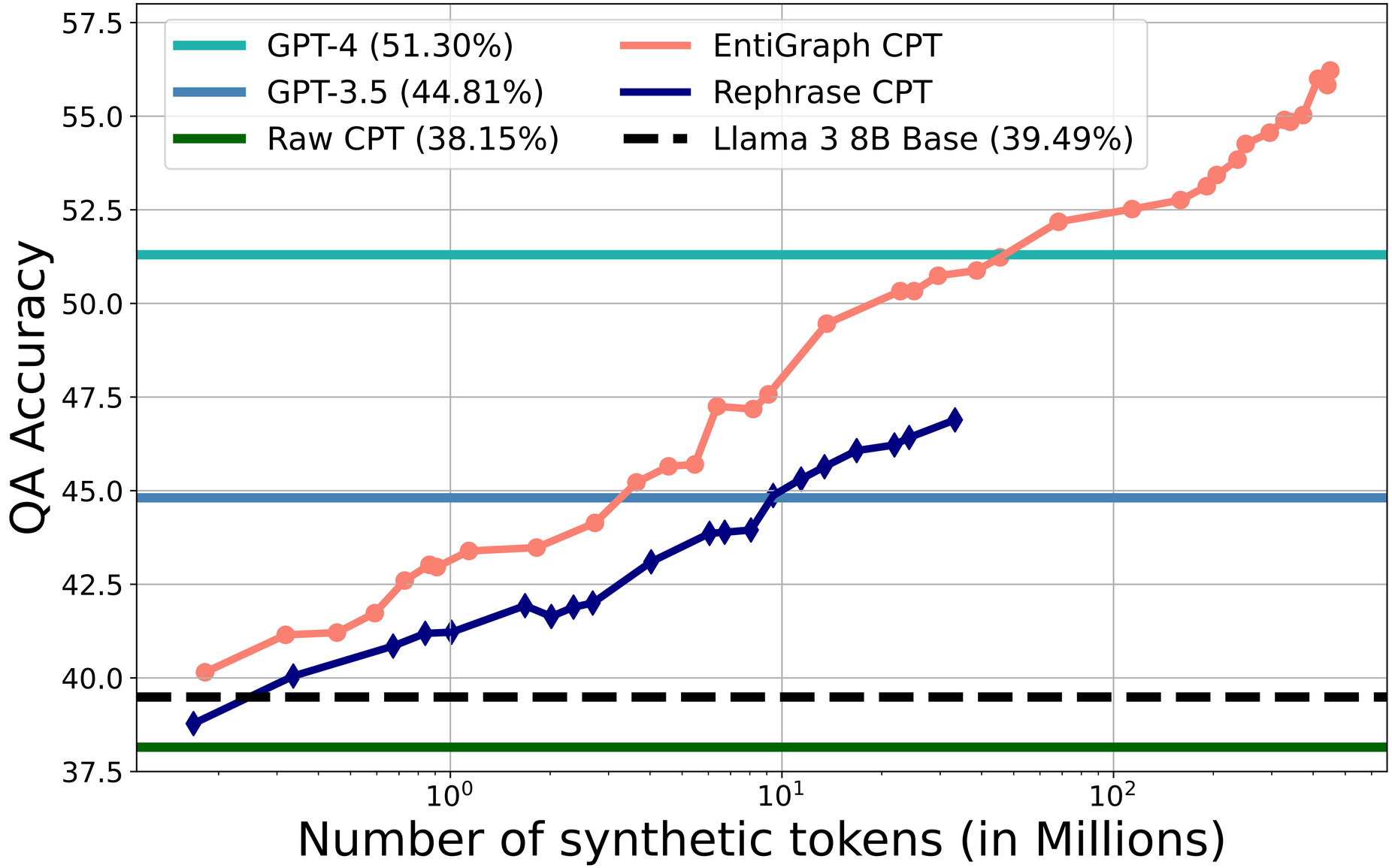


User: Analyze relations among given entities in the provided text.
[…]
Document {Defining Decay Down}
Entities {E_2 = Fluoride, E_3 = Enamel}

LM: The interplay between enamel and fluoride within the context of "Defining Decay Down" is a telling one, as it underpins the significant shift […]

Output: diverse synthetic corpus for continued pretraining

EntiGraph improves the scaling trend (closed-book)



Synthetic continued pretraining

Goal: teach an LM the knowledge of a niche domain with few “source” documents.

Step 1: Generate synthetic text based on source documents

Step 2: Pretrain/finetune the model on generated text

Experiment setup

- A corpus of niche documents (not something the base LM knows): QuALITY books
- A task that tests the LM’s knowledge about the source documents: Closed-book QA

How to generate synthetic data?

- Baseline: simply rephrase the document [Maini et al. 2024]: Lacks diversity
- EntiGraph: Entity graph synthetic data generation

Tasks beyond closed-book QA?

Closed-book vs. open-book evaluations

Closed-book

Question

In the context of “Defining Decay Down” written by David Plotz, what does the author think about American dentistry?

Tests the LM’s “parametric knowledge” about the book chapters

Open-book

Book Chapter
Title: Defining Decay Down
Author: David Plotz
If you haven’t visited a dentist in the past few years…

Question
Based on the article above, answer the following question:
….

Tests the LM’s in-context reasoning over book content

Can we approach the open-book limit?

EntiGraph CPT: model continually pretrained from Llama 3 8B Base

Model	Llama 3 8B Base	Llama 3 8B Base
Eval Format	Closed	Open
Accuracy	39.49	60.35

EntiGraph CPT at 455M tokens

- Base model: open-book access improves on closed-book by 20.8%

Can we approach the open-book limit?

EntiGraph CPT: model continually pretrained from Llama 3 8B Base

Model	Llama 3 8B Base	Llama 3 8B Base	EntiGraph CPT
Eval Format	Closed	Open	Closed
Accuracy	39.49	60.35	56.22

EntiGraph CPT at 455M tokens

- Base model: open-book access improves on closed-book by 20.8%
- Closed-book EntiGraph improves on the base model by 16.7%, about 80% of the gain achieved with in-context knowledge

Can we approach the open-book limit?

EntiGraph CPT: model continually pretrained from Llama 3 8B Base

Model	Llama 3 8B Base	Llama 3 8B Base	EntiGraph CPT	EntiGraph CPT
Eval Format	Closed	Open	Closed	Open
Accuracy	39.49	60.35	56.22	62.60

EntiGraph CPT at 455M tokens

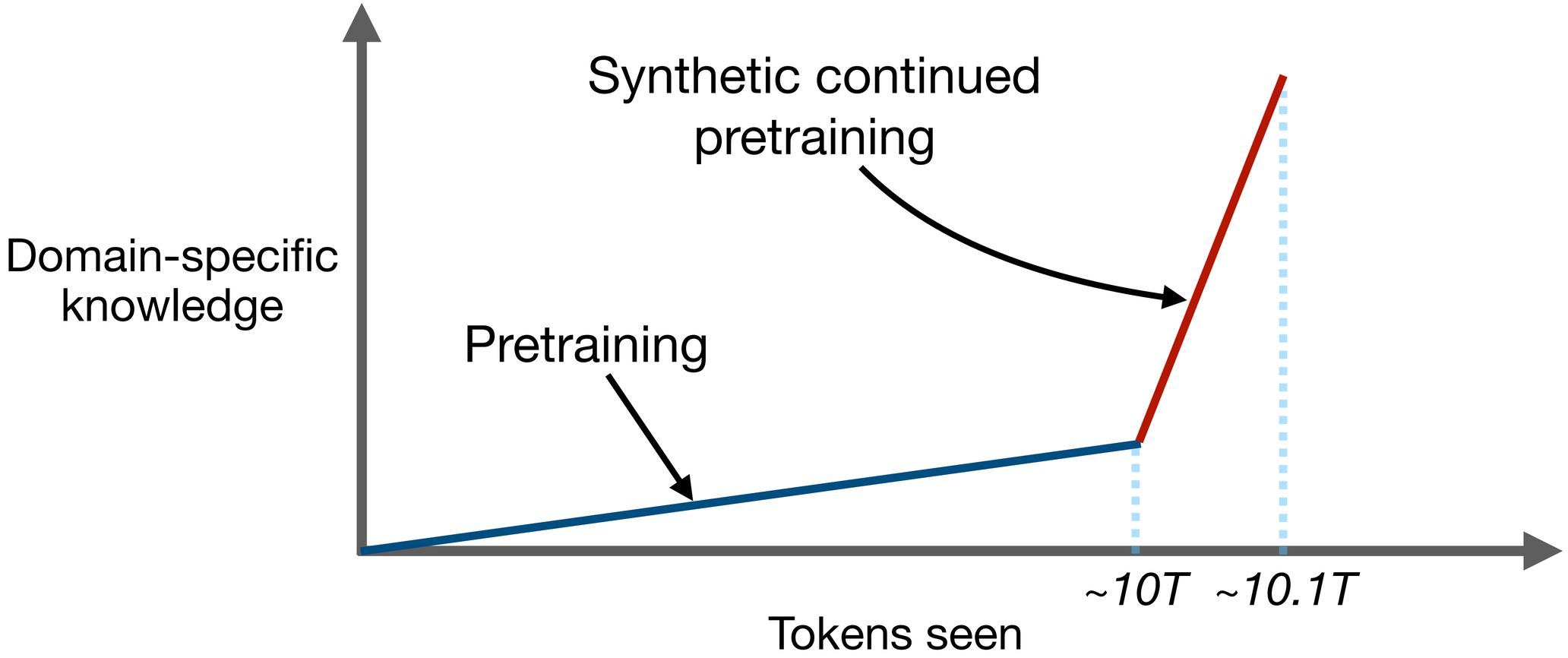
- Base model: open-book access improves on closed-book by 20.8%
- Closed-book EntiGraph improves on the base model by 16.7%, about 80% of the gain achieved with in-context knowledge
- EntiGraph CPT knowledge composes with RAG: open-book EntiGraph CPT performs best

Direct generalization after 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: [...]

Synthetic data scaling with continued pretraining

- Pretraining isn't totally over
- Opportunities to leverage small high-quality datasets more effectively



EntiGraph on 1,000 ICLR 2025 accetped papers

How does Synthetic Continued Pretraining work?

Synthetic Continued Pretraining involves generating synthetic data to adapt pretrained language models to small, domain-specific corpora.

The process includes creating a synthetic corpus from a small source corpus using an entity-centric augmentation algorithm, followed by continued pretraining on this synthetic data.

This approach allows the model to learn parametric knowledge from limited data, improving performance on downstream tasks.

Chenglei Si*