Synthetic Data Trilogy

Zitong Yang

CS525, Guest lecture, June 10th, 2025

Background

Language model training pipeline

Pretraining
Generic world knowledge

- Web data perplexity
- Code data perplexity
- High quality perplexity
- Few-shot QA in MMLU
- …

Continued Pretraining
Domain specific knowledge

- Some where in between

Post-training
Task oriented capability

- SWE-Bench
- AIME / HLE
- IFEval
- ….

Outline

- Synthetic pretraining
- Synthetic continued pretraining
- Synthetic post-training

Outline

- Synthetic post-training
- Synthetic continued pretraining
- Synthetic pretraining

Taxonomy of post-training capabilities

Instruction following

```
<goal>
Summarize the given text.
</goal>

<rules>
1. The "summary" field must be exactly 10 words long.
2. Negative Constraint: You must NOT use the word
"cat".
</rules>

<input>
Erwin Schrödinger disagreed with Niels Bohr's
interpretation. Schrödinger famously used a thought
experiment about a cat in a box to explain his
objections.
</input>
```

Taxonomy of post-training capabilities

Reasoning

```
Demonstrate rigorously that there are no positive
integers $x$, $y$, and $z$ such that:

$$x^n + y^n = z^n$$

for any integer $n > 2$.
```

Taxonomy of post-training capabilities

Multi-turn tool use

```
<tooldeclare>
1. `search(query: str, dir: str)`: Searches for a
string in a directory.
2. `bash(command: str)`: Executes a bash command in
the repository root.
Fix the issue described by the user.
</tooldeclare>

<user>
Issue #402: `process_data.py` crashes with an
`IndexError` when parsing empty text sequences in
the evaluation pipeline.
</user>
```

Synthetic post-training

Typically have the “type signature” of

$$\arg\max_{\theta} \sum_{i} p_{\theta}(y_i \mid x_i)$$

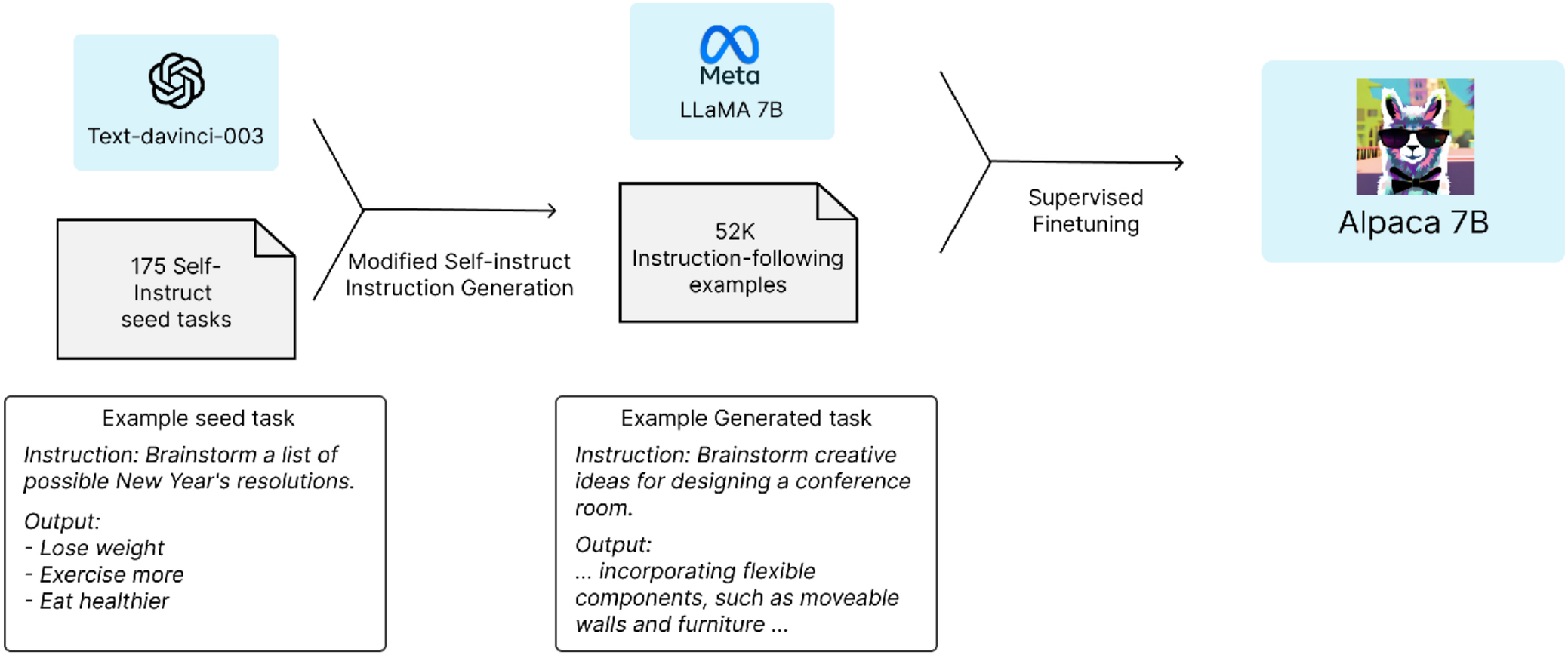
where θ is initialized from some checkpoint (as opposed to randomly initialized)

Template to examine synthetic post-training paper

- What is the target capability?
- What are the prompts — x_i
- What are the responses — y_i

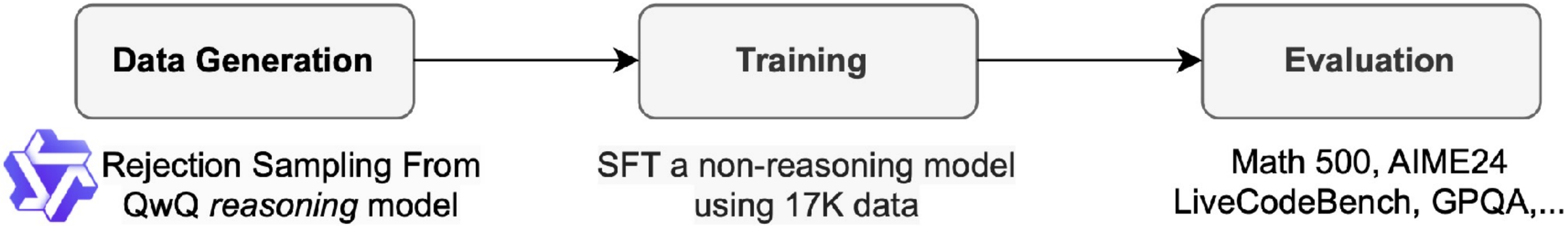
Synthetic post-training for instruction following

Alpaca: A Strong, Replicable Instruction-Following Model



Synthetic post-training for reasoning

Sky-T1: Train your own O1 preview model within $450



OpenThoughts3 - A new SOTA Reasoning Data Recipe

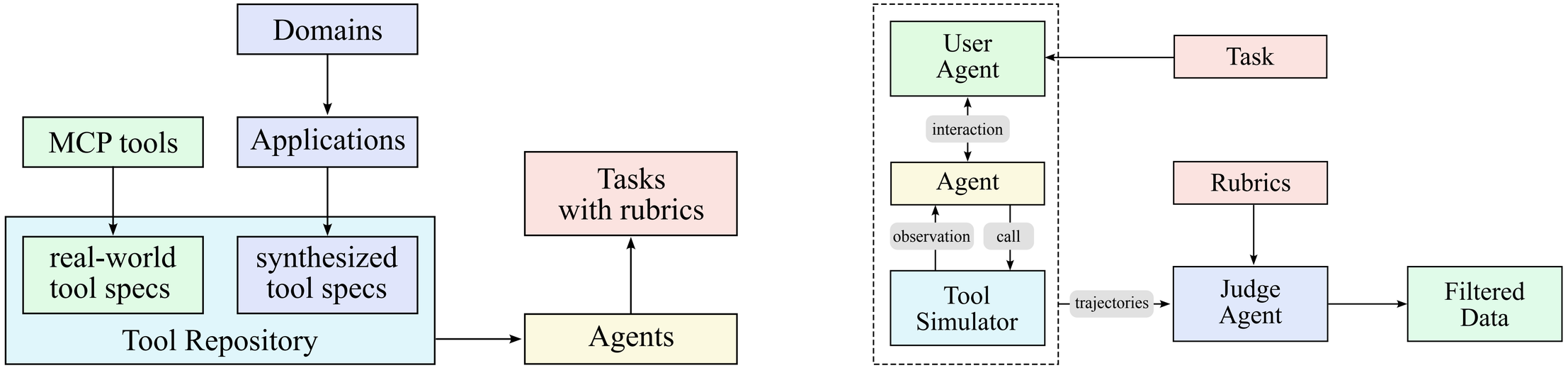
Sky-T1: Dacheng Li // Shiyi Cao // Shu Liu // Tyler Griggs // Simon Mo // Shishir G. Patil // Joseph E. Gonzalez // Ion Stoica (2025)

OpenThoughts3: Stanford // UW // BespokeLabs.ai // TRI // UCB // UT Austin // UCLA // JSC // LAION // NYU // UNC // ASU // Lila // Cornell Tech // TUM // Open-sci Collective(2025)

Synthetic post-training for tool-use

Kimi K2: Open agentic intelligence

- Crawl real world tools from online MCP
- Synthesize new tools from existing tools
- Combinatorially sample tools to create an agent with personal
- Environment simulator to simulate tool output



Kimi K2 team (2025)

Minimalist synthetic post-training

- LIMA: 1K (instruction, response) pairs suffice to allow base model to follow instructions
- S1: 1K (question, reasoning, answer) triplets suffices to grant instruct model internal CoT

	Instruction following	**Reasoning**	**Tool use**
Synthetic posttraining	Alpaca / etc.	T1 / Open thoughts / etc.	Kimi K2 / etc.
Minimalist version	LIMA / etc.	S1 / LIMR / etc.	?

LIMA: Chunting Zhou // Pengfei Liu // Puxin Xu // Srini Iyer // Jiao Sun // Yuning Mao // Xuezhe Ma // Avia Efrat // Ping Yu // Lili Yu // Susan Zhang // Gargi Ghosh // Mike Lewis // Luke Zettlemoyer // Omer Levy (2023)

S1: Niklas Muennighoff // Zitong Yang* // Weijia Shi* // Xiang Lisa Li* // Li Fei-Fei // Hannaneh Hajishirzi // Luke Zettlemoyer // Percy Liang // Emmanuel Candès // Tatsunori Hashimoto (2025)*

Discussion

Examples typically focused on light-weight capability distillation.

Minimalist version suggests those capability are already present in pretrained checkpoint.

This journey spawns two natural questions:

- Knowledge acquisition — what's a synthetic data paradigm for learning all knowledge from 2026 ?
- Self-improvement — would synthetic data enable genuine bootstrap of pretraining capability?

Outline

- ~~Synthetic post-training~~
- Synthetic continued pretraining
- Synthetic pretraining

Continued pretraining

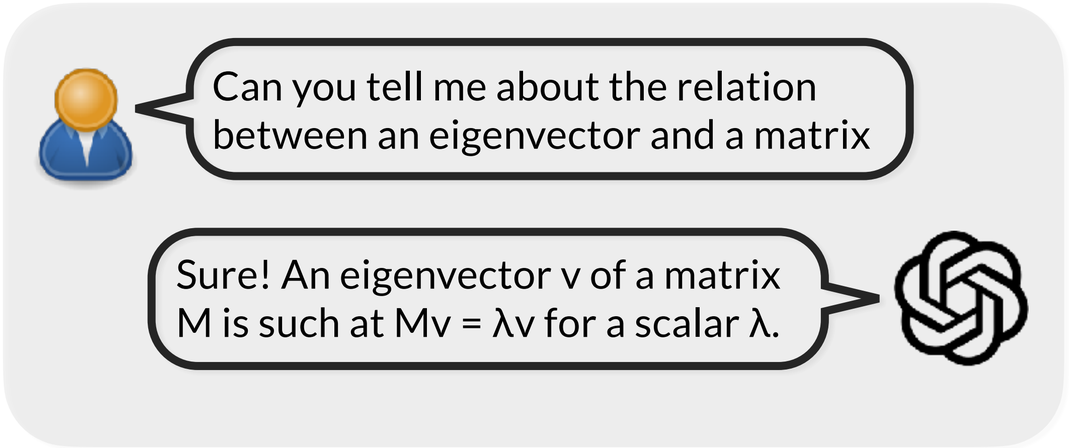
Language model training pipeline

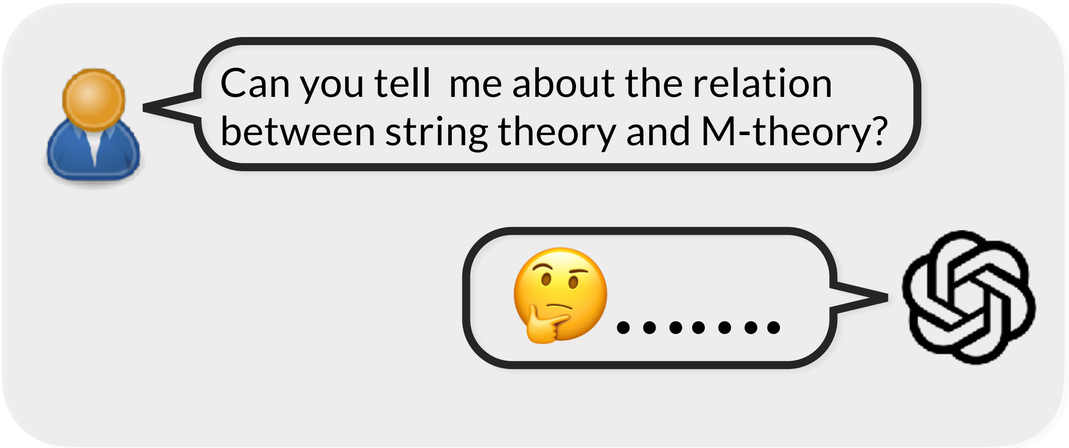
In general, continued pretraining aims to

- Reinforce under-represented data from pretraining corpus (e.g., transformer paper v.s. a niche paper)
- Whole datasets typically excluded from internet data: e.g. private data to a company or user
- Up-weight high-quality data such as mathematics and coding
- …

Knowledge can be sparse without synthetic data

- Model knows a bit about linear algebra, but not much about niche domains, e.g. quantum gravity





- Model acquires linear algebra knowledge from a wide range of internet data

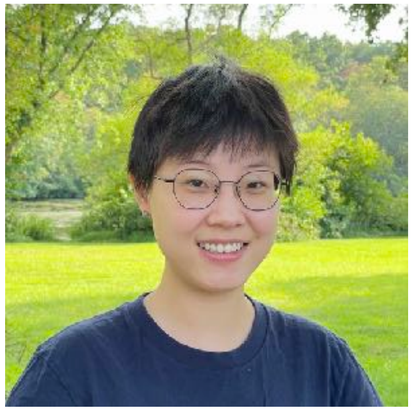
Many textbooks, lecture notes about linear algebra.

Online discussion of linear algebra exercise

GitHub implementation of SVD

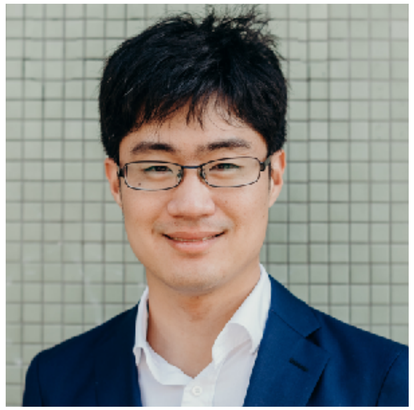
Synthetic continued pretraining

Neil Band*



Shuangping Li

Emmanuel Candès



Tatsunori Hashimoto

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

** Equal contribution*

Synthetic continued pretraining

Goal: teach model the knowledge from a niche domain consisted of a few “source documents"

Step 1: Generate synthetic text based on the source documents

Step 2: Continually pretrain (finetune) the model on generated text

Experiment setup

- A collection of niche (not something model already know) source documents
- A task that tests a model’s knowledge about the source documents

A dataset and a benchmark

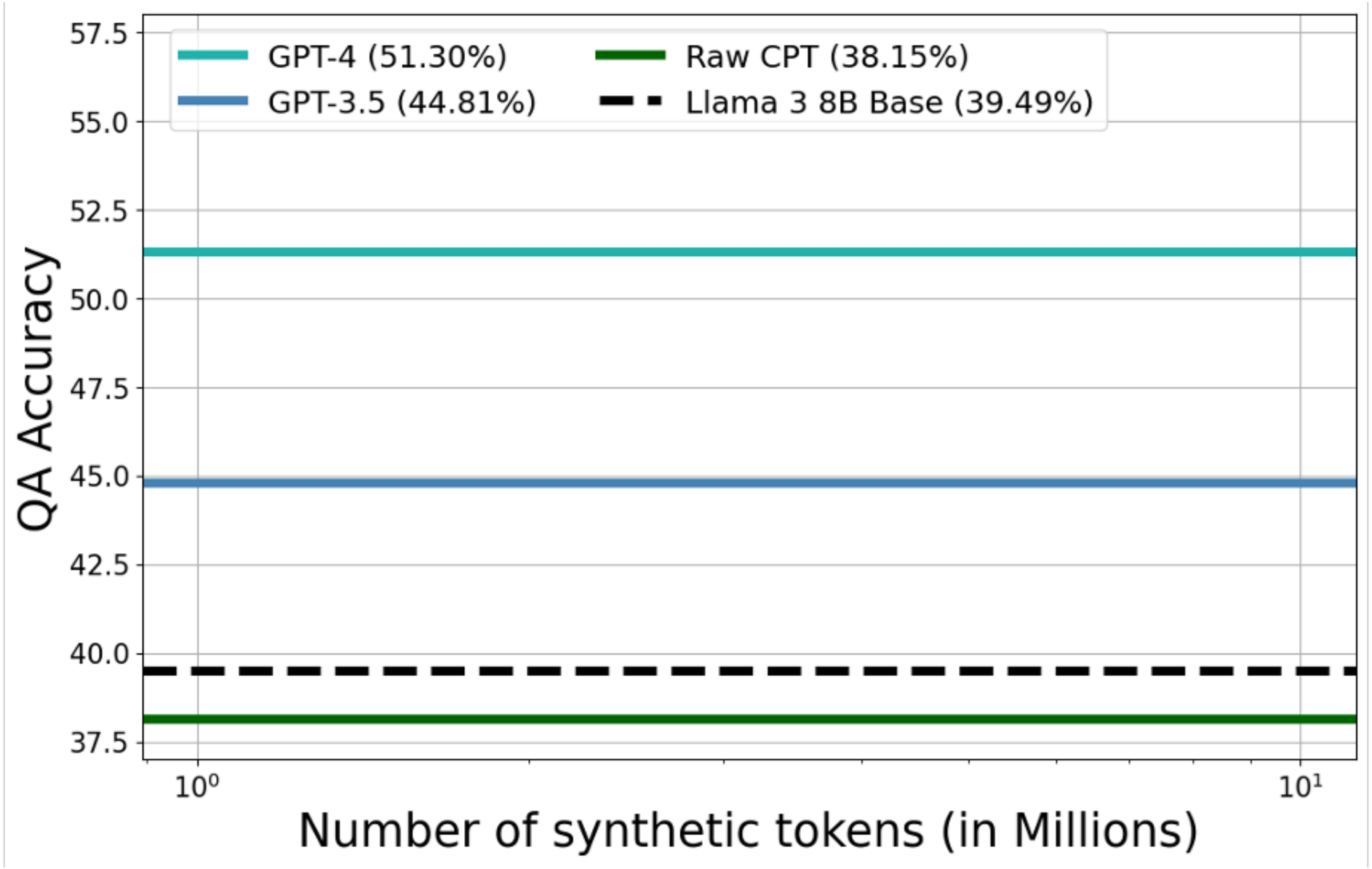
QuALITY Books

- Project Gutenberg fictions (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
- Want model to answer without book in-context (closed book)

Base models has poor performance



Synthetic continued pretraining

Goal: teach model the knowledge from a niche domain consisted of a few “source documents"

Step 1: Generate synthetic text based on the source documents

Step 2: Continually pretrain (finetune) the model on generated text

Experiment setup

- A collection of niche (not something model already know) source documents **—> QuALITY books**
- A task that tests a model’s knowledge about the source documents **—> Closed book QA**

Synthetic continued pretraining

Goal: teach model the knowledge from a niche domain consisted of a few “source documents"

Step 1: Generate synthetic text based on the source documents

Step 2: Continually pretrain (finetune) the model on generated text

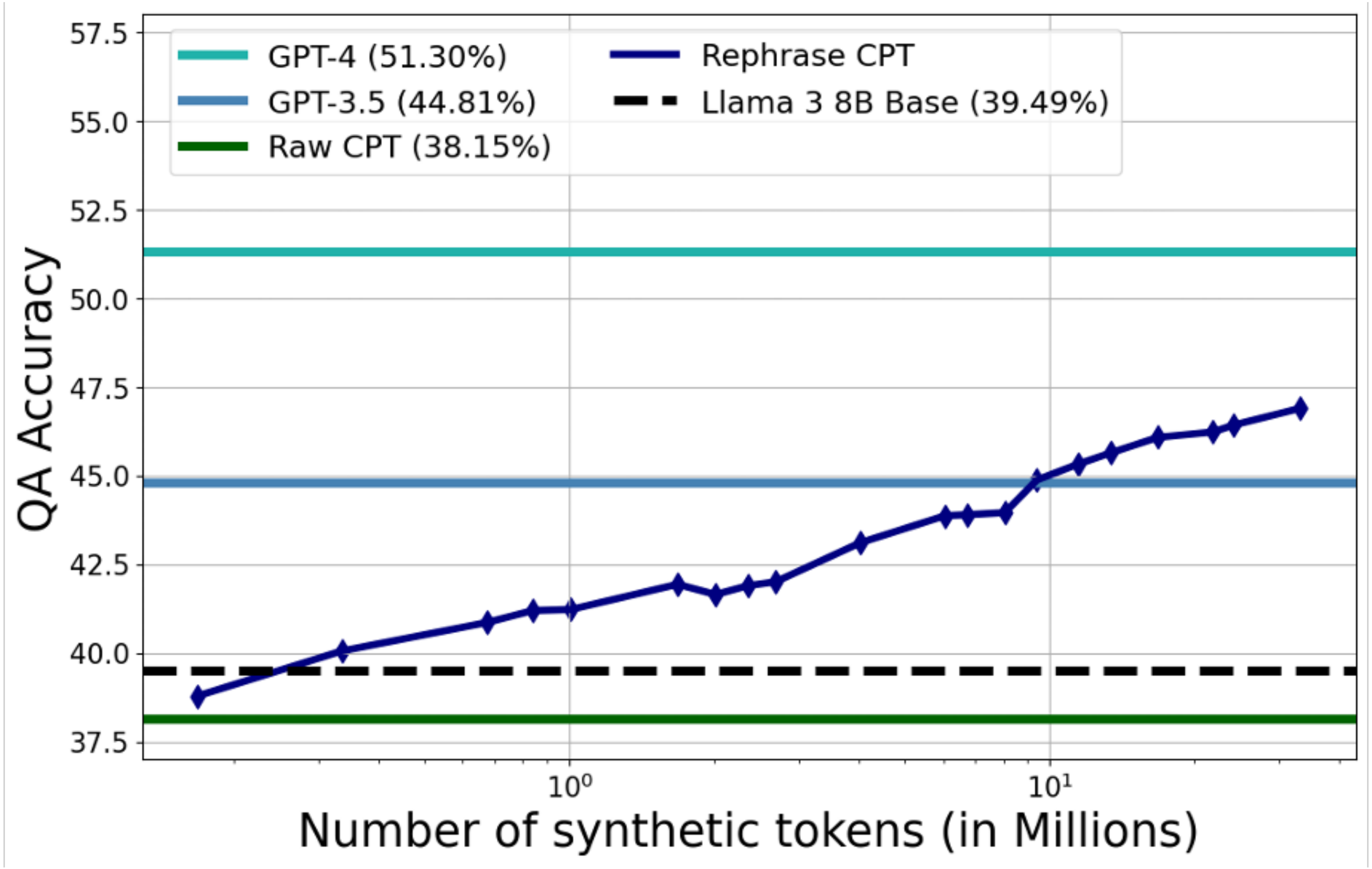
Experiment setup

- A collection of niche (not something model already know) source documents **—> QuALITY books**
- A task that tests a model’s knowledge about the source documents **—> Closed book QA**

How to generate synthetic data?

- Baseline: simply rephrase the document Pratyush et al. 2024

First attempt: simply rephrase (closed book)



Synthetic continued pretraining

Goal: teach model the knowledge from a niche domain consisted of a few “source documents"

Step 1: Generate synthetic text based on the source documents

Step 2: Continually pretrain (finetune) the model on generated text

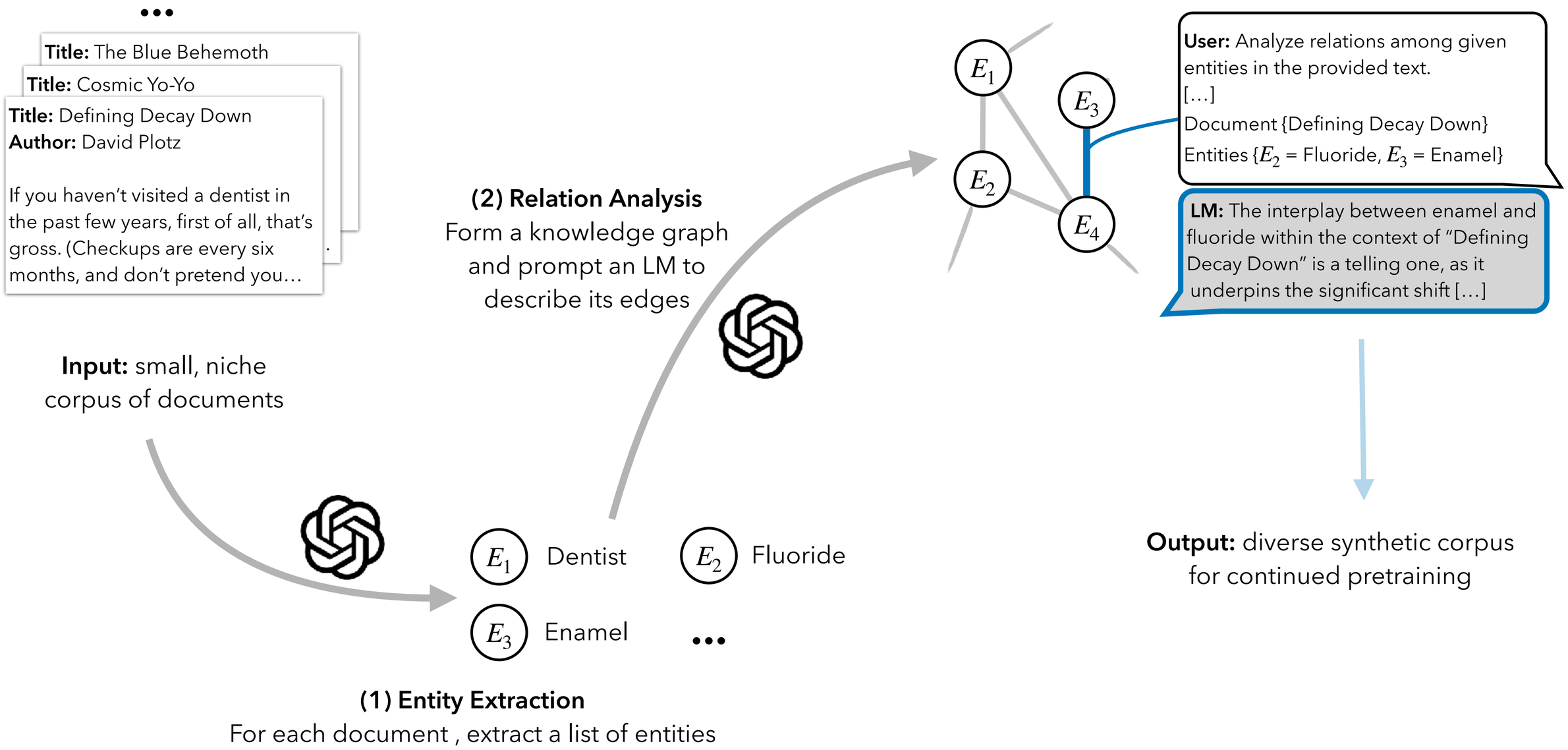
Experiment setup

- A collection of niche (not something model already know) source documents —> QuALITY books
- A task that tests a model’s knowledge about the source documents —> Closed book QA

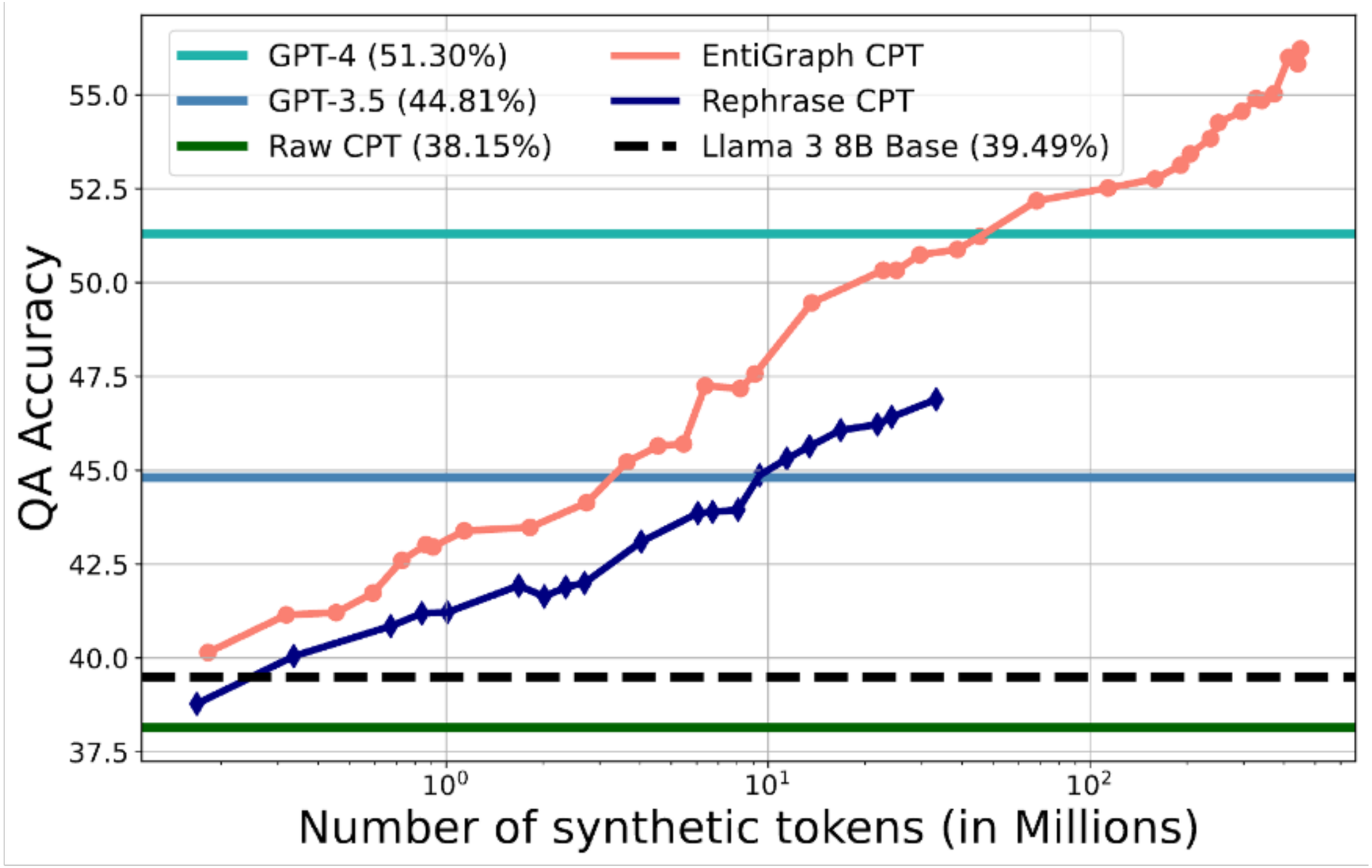
How to generate synthetic data?

- Baseline: simply rephrase the document Pratyush et al. 2024 **—> Lacks diversity**
- **EntiGraph: Entity graph synthetic data generation**

EntiGraph: scalable data generator

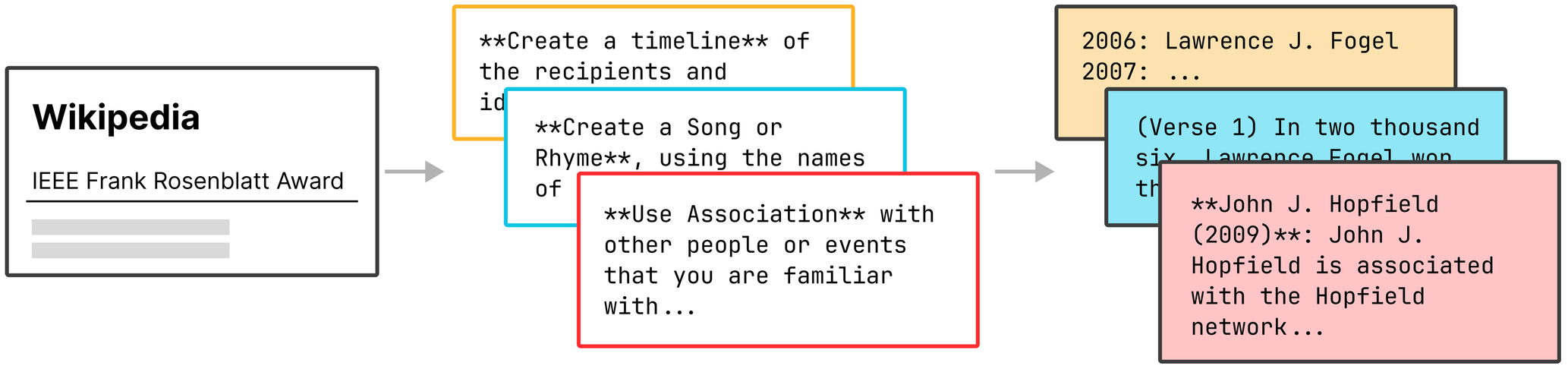


EntiGraph performance (closed book)



Similar two-stage data synthesis techniques

Learning Facts at Scale with Active Reading



Source Document

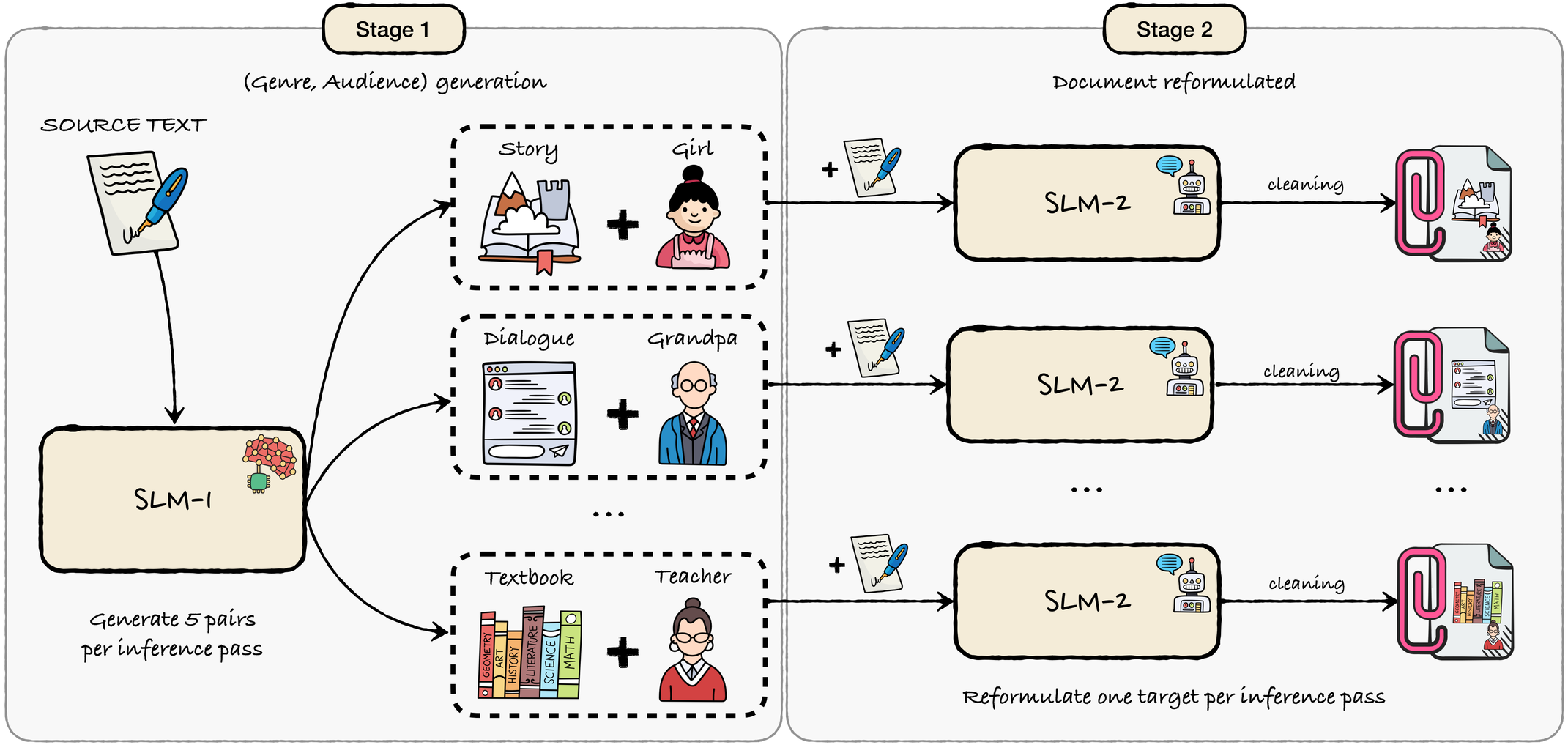
1. Generate learning strategies to seed data generation

2. Apply strategies to generate diverse synthetic documents

Active Reading: Jessy Lin // Vincent-Pierre Berges // Xilun Chen // Wen-Tau Yih // Gargi Ghosh // Barlas Oğuz (2025)

Similar two-stage data synthesis techniques

Reformulation for Pretraining Data Augmentation



Reformulation for Pretraining Data Augmentation: Xintong Hao // Ruijie Zhu // Ge Zhang // Ke Shen // Chenggang Li (2025)

Discussion

Examples typically focused on light-weight capability distillation.

Minimalist version suggests those capability are already present in pretrained checkpoint.

This journey spawns two natural questions:

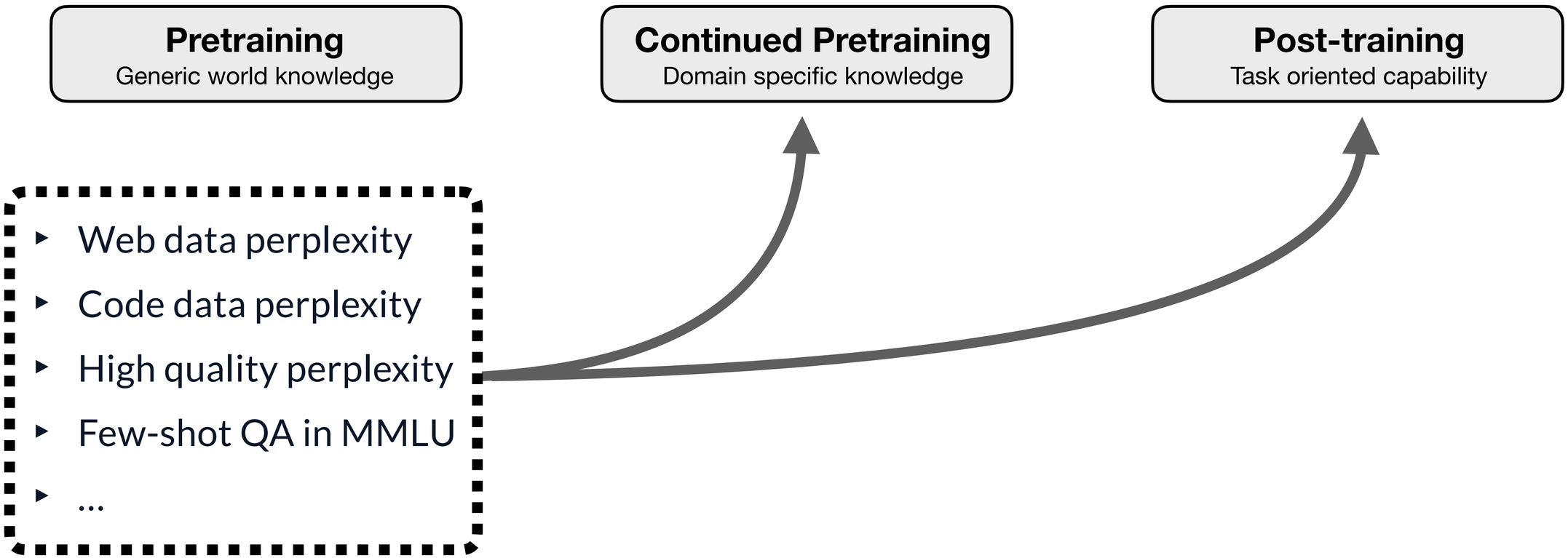
- ~~Knowledge acquisition — what's a synthetic data paradigm for learning all knowledge from 2025 ?~~
- Self-improvement — would synthetic data enable genuine bootstrap of pretraining capability?

Outline

- ~~Synthetic post-training~~
- ~~Synthetic continued pretraining~~
- Synthetic pretraining

Pretraining

Language model training pipeline



A typical pretraining dataset

Training horizon: 10T

- GitHub data: 20% -> 2T
- Web crawl: 50% -> 5T
- …

Repetition calculus

- GitHub data: 0.5T total + 2T horizon -> 4 epochs
- Web crawl: 5T total + 5T horizon -> 1 epoch
- …

For each data source

- Number of unique tokens
- Number of training horizon
- Repetition factor = Number of training horizon / Number of unique tokens

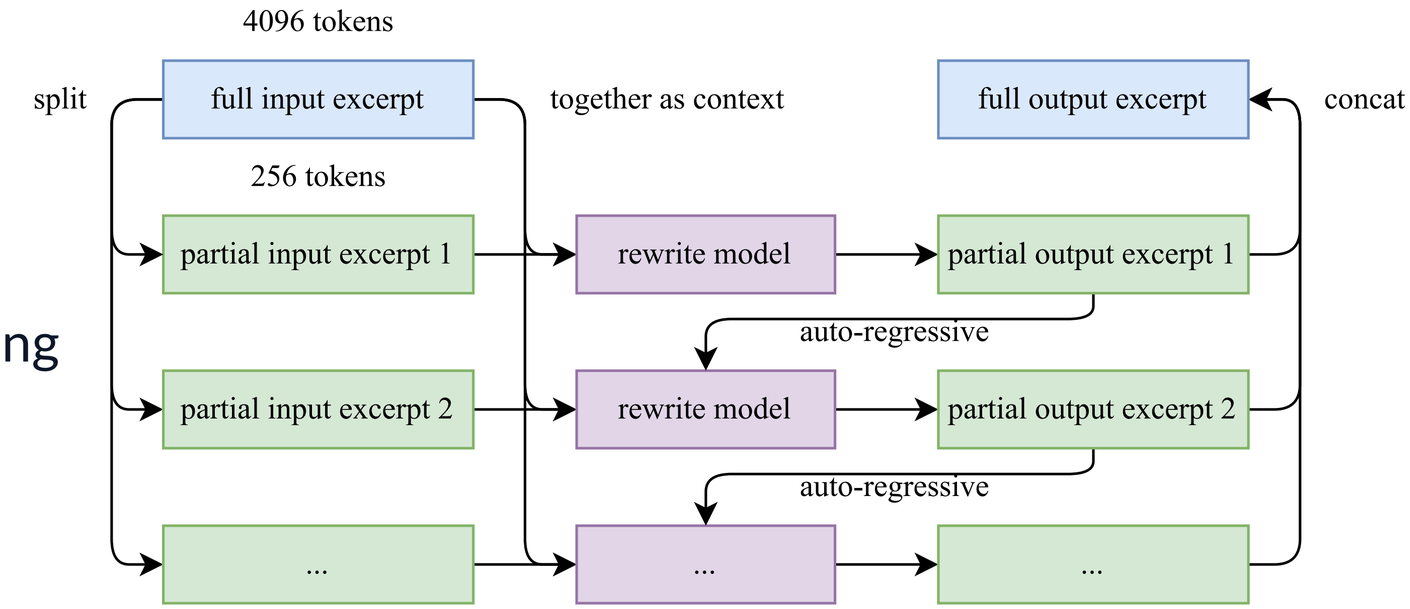
Repetition tradeoff

- Single epoch: insufficient for comprehensive knowledge absorption
- Multi-epoch: yields diminishing returns and increases the risk of overfitting

Kimi K2: Open agentic intelligence

Knowledge Data Rephrasing:

- Style- and perspective-diverse prompting
- Chunk-wise autoregressive generation
- Fidelity verification



More aggressive synthetic pretraining

Blurring boundary between pretraining and post-training

Phi-1 & 1.5: “Textbook” paradigm

- Prompting teacher LM to generate education content from scratch
- “Generating a Python tutorial on list comprehensions tailored for a high school student”

Phi-3: Seed-material rephrase with reasoning steps

- Prompting teacher LM to transform web snippets into rigorous, step-by-step reasoning chains
- “Rewrite Wikipedia about cellular mitosis to a college-level, logically sequenced exam prep guide.”

Phi-4: Pretraining scale post-training data

- Back translation: “Examine `[crawled_code.py]` and work backward and write a highly challenging, multi-constraint user prompt that this exact code perfectly solves.”
- Similar process as in synthetic post-training using back-translated prompt
- Generate -> Critic -> Revise loop

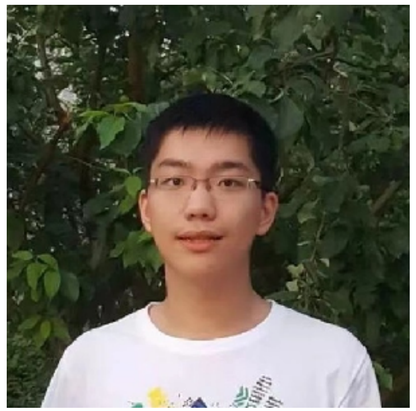
Can we get rid of the “teacher model”?

Genuine bootstrap of pretraining capability

- Pretrain a language model from scratch
- Finetune the model to be a synthetic data generator
- Pretrain a new model on the synthetic data and see improvement performance

Synthetic bootstrapped pretraining

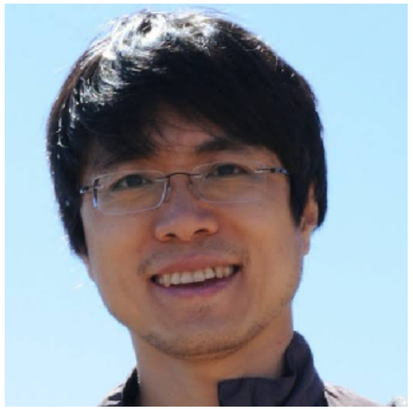

Aonan Zhang*


Hong Liu


Tatsunori Hashimoto


Emmanuel Candès


Chong Wang


Ruoming Pang

ICLR 2026, Oral, https://arxiv.org/abs/2409.07431

** Equal contribution*

Where does the knowledge come from in pretraining?

Thought experiment

- World with 5 tokens “A”, “B”, “C”, “D”, “E”
- Text documents with each token sampled u.a.r. [“BDECD…”, “ACEAC…”,]
- Perform next token prediction with transformer LM: No meaningful learning signal

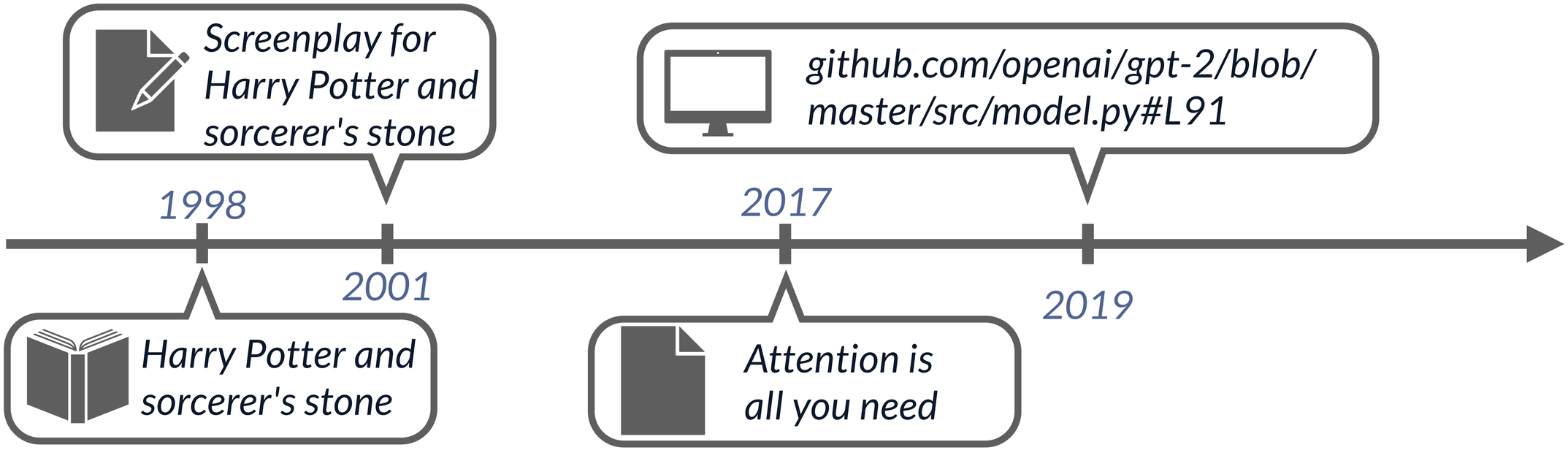
Two views

- Statistical view: Natural language tokens has *statistical* correlations
- Computational view: Natural language has *compressible* patterns

Views aside: **structural correlation** enables learning

Under-exploited sources of correlation

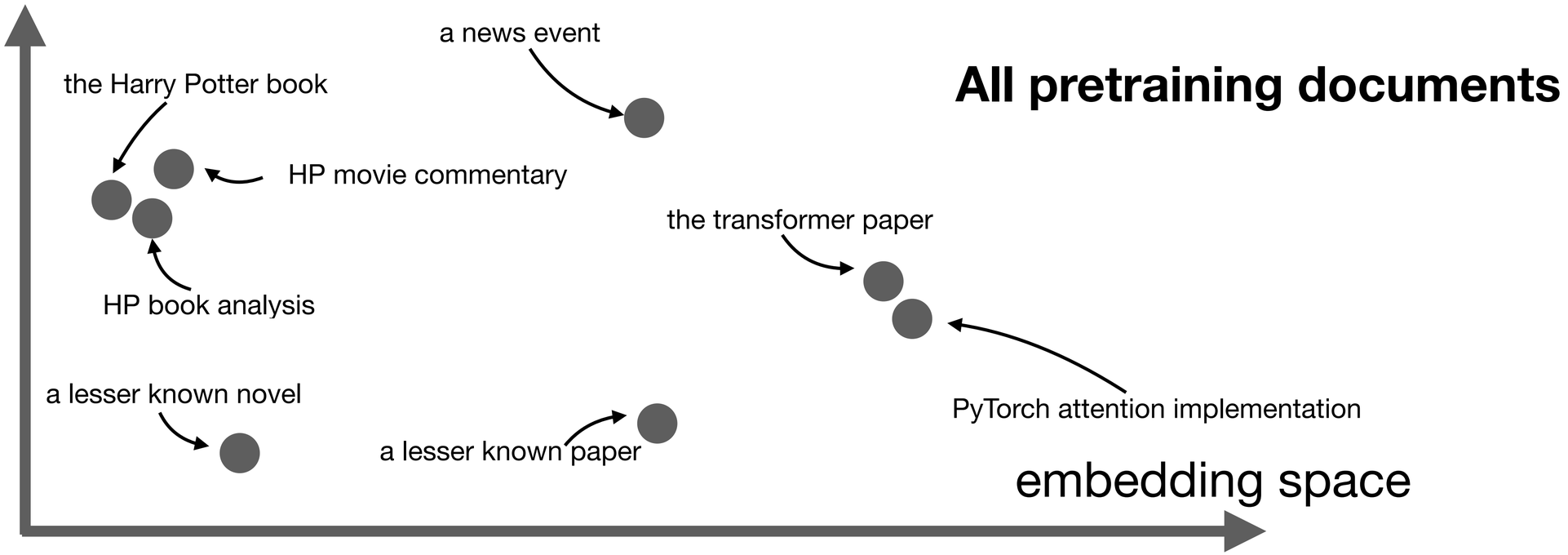
There exists *rich* correlations between documents



Technique: take-advantage of inter-document correlation

Synthetic bootstrapped pretraining

1. **Nearest-neighbor pairing:** we used DCLM subset and Qwen-0.6B-Embedding



Examples of related documents look like

doc1

The Cultural Sites of Iran

With 196 countries and countless exciting destinations worldwide, there is so much to see in a very limited time. Even the most well-traveled person hardly gets to visit all and has to be selective. So, why should you consider visiting a country like Iran, especially when it comes to all those negative news and stereotypes surrounding it?
Here we're here to give you the reasons and to help you overcome your doubts and even encourage you to consider your next trip to Iran, this mysterious land as soon as you return to your home country!

Beautiful cities, friendly people, fabulous food, glorious architecture, Iran has delighted visitors for centuries with its World Heritage Sites, friendly towns and inspiring desert landscapes.

Things to Do in Iran – Activities & Attractions

Iran is the land of four seasons, history and culture, souvenir and authenticity. This is not a tourism slogan, this is the reality inferred from the experience of visitors who have been impressed by Iran's beauties and amazing attractions.

Antiquity and richness of the Cultural Sites of Iran and civilization, the variety of natural and geographical attractions, four – season climate,
...

History of Iran

doc2

Query Text: Home > FAQ Login / Register

Why should we spend our holiday in Iran?

Iran is a country, located in the Middle East, which can meet the various needs of tourists and satisfy their different tastes, due to its rich civilization, historical sites, geographic location, nature of the four seasons and diverse tourist attractions. Therefore, considering the high security and low cost of travel to the country, it is introduced as one of the major tourist destinations to spend holidays in.

Is Iran a safe travel destination?

One of the wrong assumptions about the country of Iran is in terms of its security. Despite its location in Asia and the Middle East, and neighboring countries like Iraq, Afghanistan and Pakistan, Iran is considered as one of the safest countries in the region. According to the international data, security in Iran is much more than a touristic country such as Turkey.

To confirm the statements made above, refer to websites like www.travelriskmap.com.

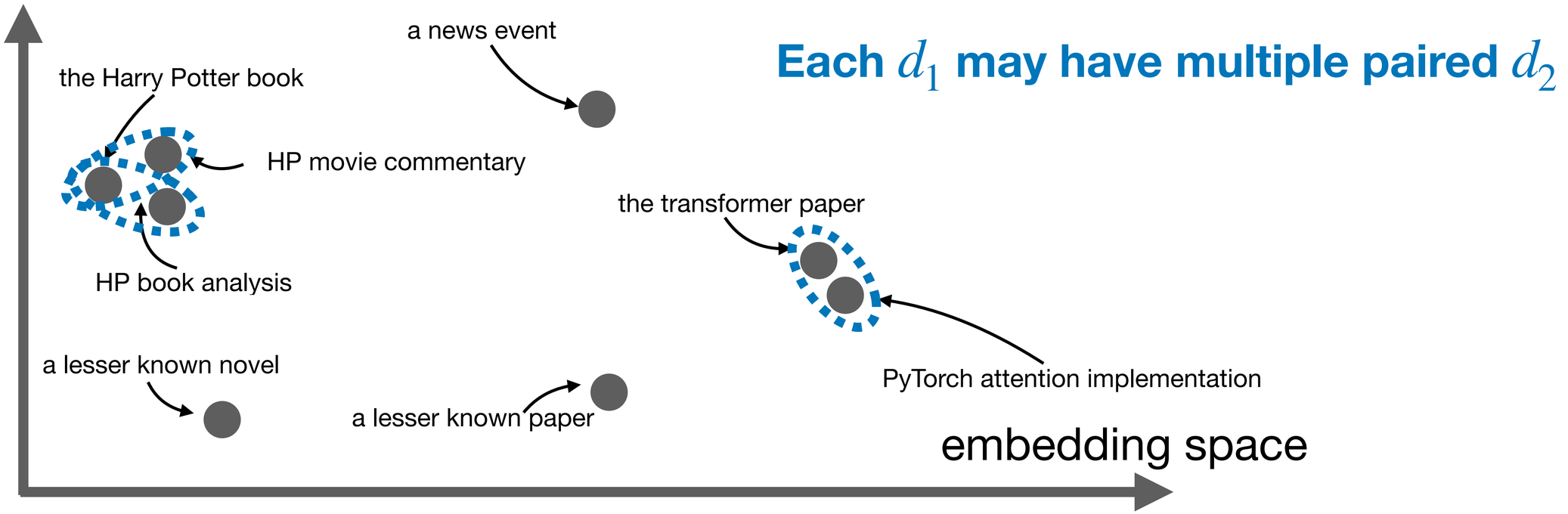
What does "the rich civilization" mean, as mentioned about Iran?

According to documentation in some of the world history references,
...

Travel guide to Iran

Synthetic bootstrapped pretraining

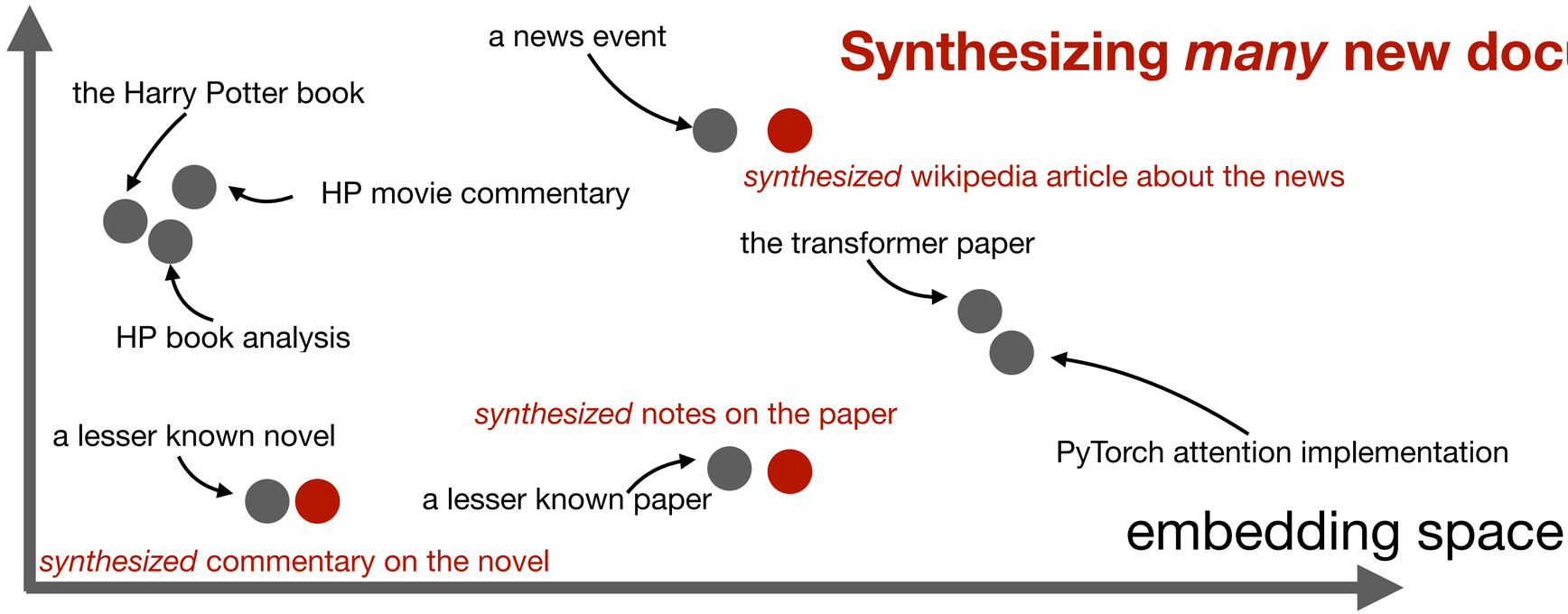
1. **Nearest-neighbor pairing:** we used DCLM subset and Qwen-0.6B-Embedding
2. **Synthesizer tuning:** SFT-like objective $p_\theta(d_1 \mid d_2)$ initialized at pretrained checkpoint



Each d_1 may have multiple paired d_2

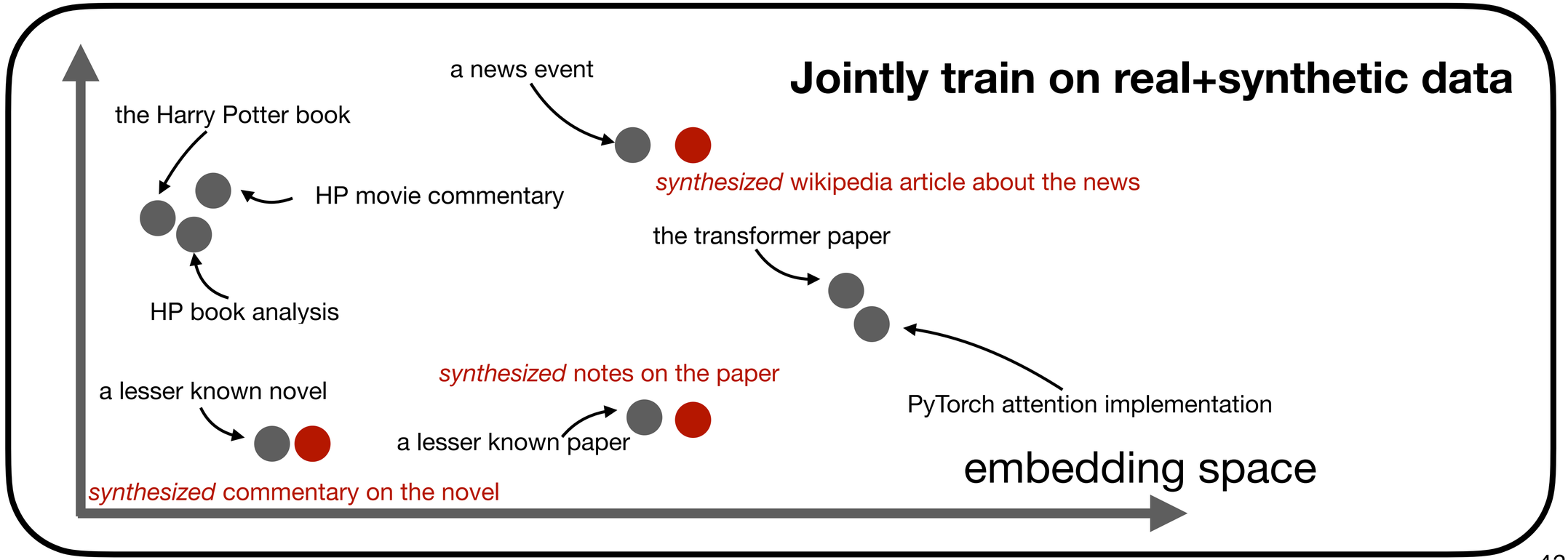
Synthetic bootstrapped pretraining

1. **Nearest-neighbor pairing:** we used DCLM subset and Qwen-0.6B-Embedding
2. **Synthesizer tuning:** SFT-like objective $p_\theta(d_1 \mid d_2)$ initialized at pretrained checkpoint
3. **Synthesis at scale:** Temperature=1 allows each document to have varied synthesis



Synthetic bootstrapped pretraining

1. **Nearest-neighbor pairing:** we used DCLM subset and Qwen-0.6B-Embedding
2. **Synthesizer tuning:** SFT-like objective $p_\theta(d_1 \mid d_2)$ initialized at pretrained checkpoint
3. **Synthesis at scale:** Temperature=1 allows each document to have varied synthesis



Visualization of synthesized document

Real document

HOME > Chowhound > California >

Coffeehouses in SD

I am always looking for good coffeehouses in San Diego. I am looking for coffeehouses with (obviously) excellent coffeebeans, baristas who know their business (good microfoam, good crema, they should know the different ratios for steamed milk, foamed milk, espresso for the cappuccino, latte etc., etc), good snacks (paninis, cakes etc), good (cold) flavored coffeedrinks and good ambience. So far I haven't found many who are good but would like to hear more suggestion.

Caffe Calabria: By far the best coffeehouse in SD. Excellent coffee and baristas (best cappuccino in SD,), great paninis. Only downside is that they close very early.

Café Lestat: Good example what happens with good coffee and average baristas. Lestat is also using coffeebeans from Caffé Calabria but it is sad to see that they don't know how to make excellent espresso, cappuccino.

E Street Café: Average espresso but strange flavored ...

Synthesis I

Coffee = Love

On a recent trip to San Diego, one thing I wanted to find was coffee. I was worried it might be difficult with the huge decrease in Asian flavors, but I was pleasantly surprised. Many of the Mexican, Italian, and Cuban flavors that are pretty common in California are now in the mix. While I did find a few highly recommended spots, the majority of the locations I checked out were nothing to write home about. At the least, they were not being sold out, which I consider a plus.

One major factor that really needs to be taken into account is the type of machine being used. There are two types of espresso machines: the cheap ones that spit out lots of pressure and don't have any control over the shot, and the expensive ones that create amazing shots and utilize the types of superfine grinds you usually find in Italy. The latter types are much harder to find and cost significantly more.

Another factor to consider is the type of beans being used. If they are done correctly, great espresso can be made from any type of bean, but it's a rare experience to find a place that does ...

Synthesis II

Discuss as:

Coffeehouses: San Diego. The best?

Café Lestat, by far, is a stand-out. (photo by Thomas Moriarty)

If the coffee scene in San Diego sounds like the one in New York, there's a reason. The upper echelons of NYC's drink scene are the same ones who have taken over the majority of the coffee houses in San Diego. This is probably partly due to the fact that both cities have large student populations, both are home to large coffee drinker bases, both have a large number of espresso shops, and both offer top-notch, hand-crafted cappuccinos. But if there is one city in America that deserves to have its name above that of New York, it's San Diego.

There are just under 100 coffee shops in San Diego, with almost half of them located on University Ave. alone. So finding the perfect coffee shop is crucial. We spent a whole day just roaming around the area, hunting for the best.

In terms of the coffee itself, it's hard to beat Café Lestat. The baristas are amazing and their methods are pristine ...

Experiment design

Data, model, and evaluation

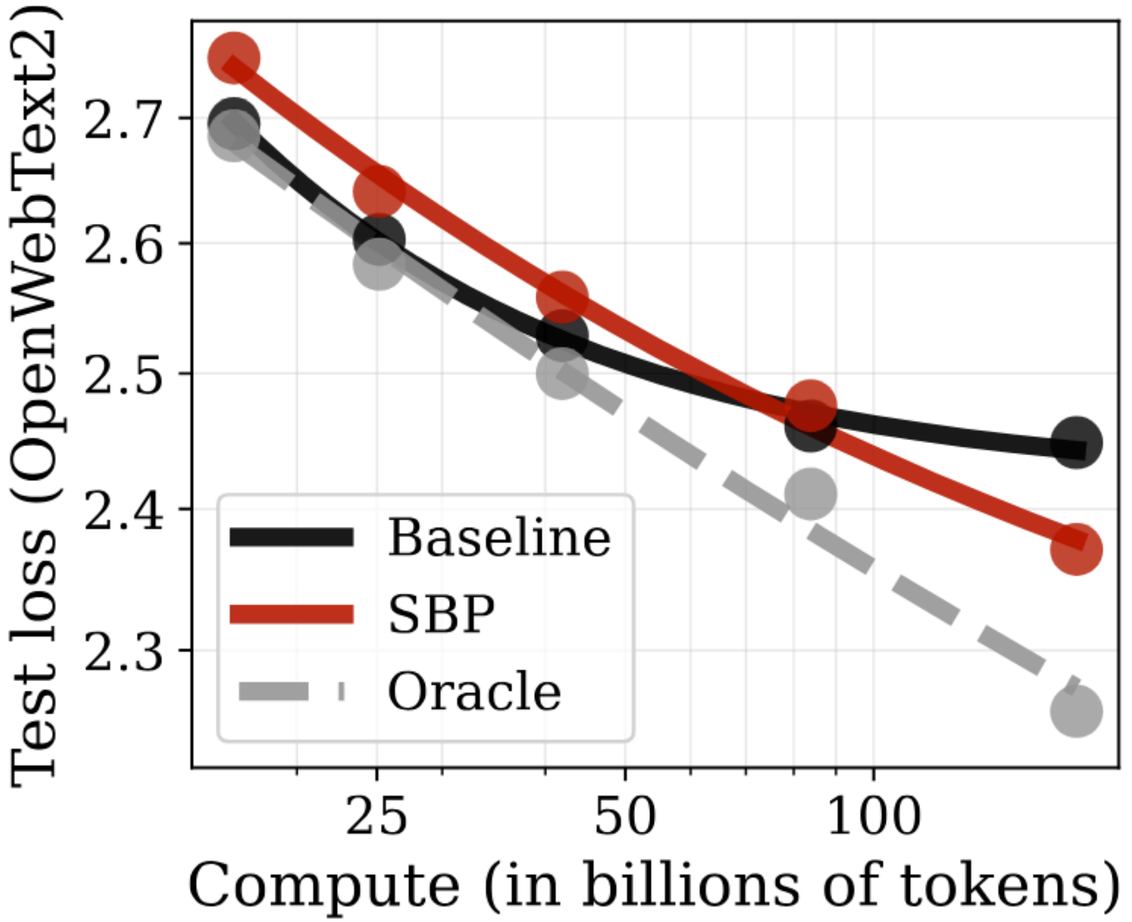
- Data: DCLM dataset
- Model: Llama 3 architecture with additional QK-norm
- Evaluation: 6 QA accuracies and 3 perplexity evaluation

Compute-matched comparison

- Baseline: 20 times repetition
- SBP: Same compute, same data source
- Oracle: 20x additional data, no repetition during pretraining

Training dynamics

- At first, oracle ~ baseline < SBP
- Later on, baseline saturates
- Finally, oracle + SBP continues to scale



Result: 40% improvement attained by the oracle

Benchmark	200B-scale			1T-scale		
	Baseline	SBP	Oracle	Baseline	SBP	Oracle
	Perplexity on held-out data ↓					
OpenWebText2	5.74	-0.53	-1.02	4.51	-0.02	-0.12
LAMBADA	6.87	-0.85	-1.86	4.33	-0.03	-0.22
Five-shot MMLU	3.83	-0.36	-0.51	3.17	-0.06	-0.05
	QA accuracy ↑					
ARC-Challenge (0-shot)	35.32	+1.28	+2.82	42.66	+1.62	+3.84
ARC-Easy (0-shot)	68.94	+2.65	+4.29	75.63	+0.42	+2.11
SciQ (0-shot)	90.50	+1.00	+2.40	93.20	+0.80	+0.50
Winogrande (0-shot)	60.14	+1.90	+5.53	65.19	+1.42	+2.92
TriviaQA (1-shot)	22.51	+3.36	+7.37	36.07	+0.25	+0.59
WebQS (1-shot)	8.56	+3.74	+10.83	19.34	+0.54	+0.44
Average QA accuracy	**47.66**	**+2.32**	**+5.54**	**55.35**	**+0.84**	**+1.73**

Synthetic data quality

	Repetition ↓	Duplicate@1M ↓	Non-factual ↓	Pair-irrelevance ↓	Pair-copying ↓
200B-scale	4.3%	0.8%	15.1%	25.6%	0.1%
1T-scale	3.9%	0.8%	8.7%	7.8%	0.9%
Real data	1.8%	0.7%	1.8%	n.a.	n.a.

- Better data quality with larger scale
- Synthesized data is not mere repetition

Summary

Synthetic data through all stages of LM life cycle

Pretraining
Generic world knowledge

- Better “regularization”

Continued Pretraining
Domain specific knowledge

- Diverse knowledge representation

Post-training
Task oriented capability

- Distillation in IF / reasoning / tool-use