Lightweight and general-purpose language model memories via self-study

Sabri Eyuboglu

Joint work with: Ryan Ehrlich, Simran Arora, Neel Guha, Dylan Zinsley, Emily Liu, Atri Rudra, James Zou, Azalia Mirhoseini, Christopher Ré

tl;dr

The promise of **long context** language models is…

- LLMs that understand your personal documents.

- Chatbots with long-term memory.

- Coding assistants with full-repository understanding.

Problem: Long contexts are extremely costly to serve to users due to the *KV cache*!

Observation: Contexts are often static and shared across many requests.

Idea: Spend compute training the KV cache on the context offline!

tl;dr

Training a smaller KV cache enables **39x memory reduction** and **26x higher tok/s** while maintaining the capabilities of the full KV cache.

Talk outline

Background. *Why is it so costly to serve long contexts?*

Space-quality tradeoffs. *Why don't existing approaches cut it?*

Training the KV-cache. Trading compute for space.

Self-study. Synthetic data generation enables generalization.

Results. Expanding the space-quality frontier and more!

Talk outline

Background. *Why is it so costly to serve long contexts?*

Space-quality tradeoffs. *Why don't existing approaches cut it?*

Cartridges. Trading compute for space with offline training.

Self-study. Synthetic data generation enables generalization.

Results. Expanding the space-quality frontier and more!

Quiz: Why are long contexts so costly?

With **1k** tokens of context per user, we can output **10,000** tokens per second
With **128k** tokens of context per user, we can output **?,???** tokens per second

*numbers computed for Llama 8B running at peak throughput (*i.e.* excluding prefill) on a single H100

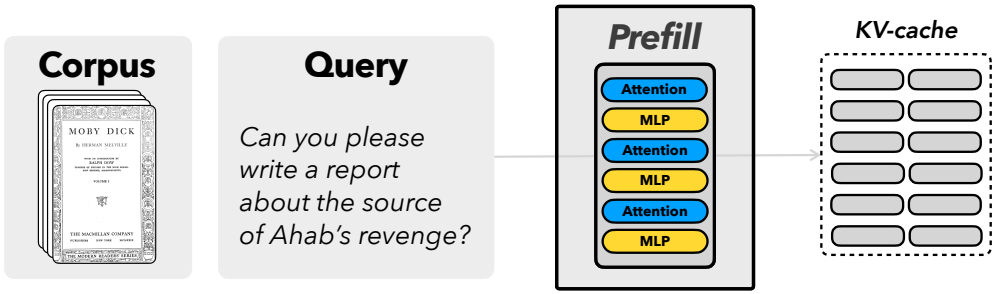
Quiz: Why are long contexts so costly?

With **1k** tokens of context per user, we can output **10,000** tokens per second
With **128k** tokens of context per user, we can output **130** tokens per second

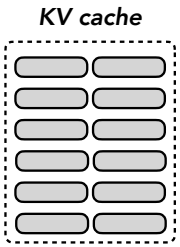
*numbers computed for Llama 8B running at peak throughput (*i.e.* excluding prefill) on a single H100

Why are long contexts so costly?

When we provide a large corpus of text to a Transformer, we first need to cache the **key** and **value** vectors!

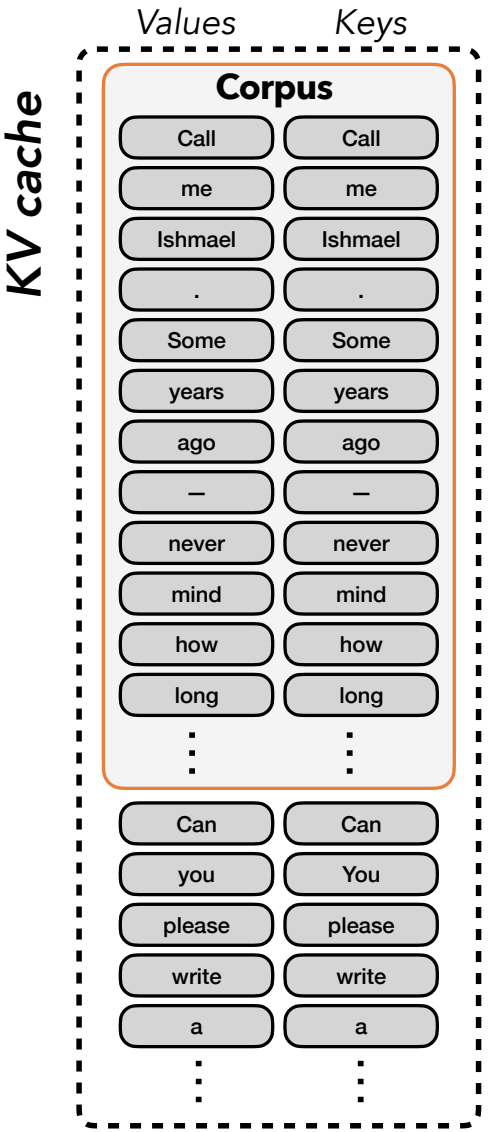


For long contexts this cache is enormous!



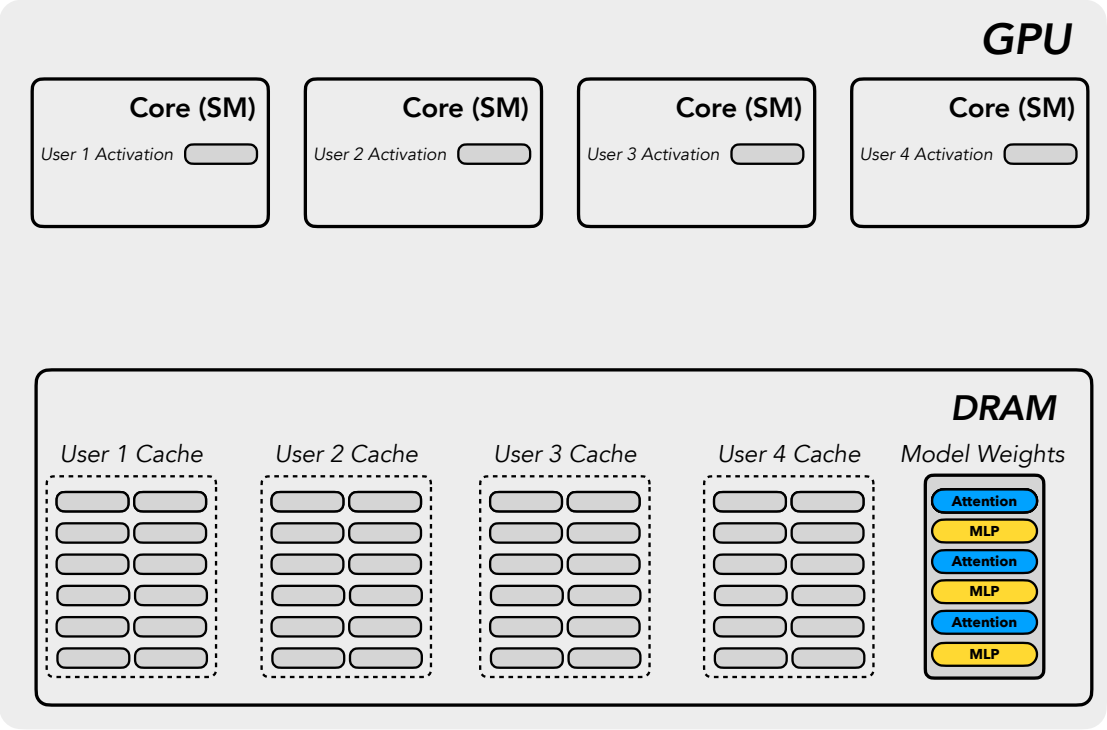
KV cache for 128k tokens of context is **~84 GB**

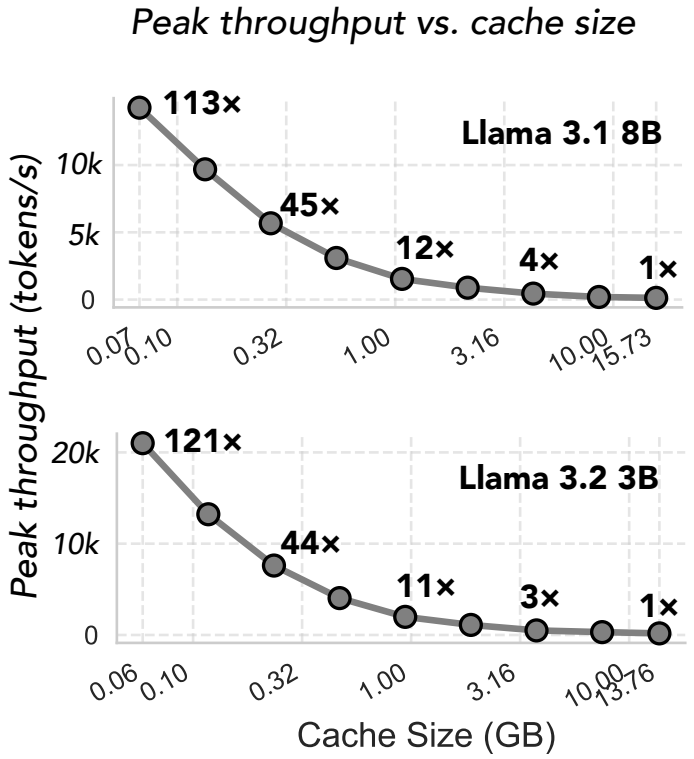
Llama 70B parameters are **~140 GB**



Why are long contexts so costly?

Large KV caches can make decoding I/O bound.

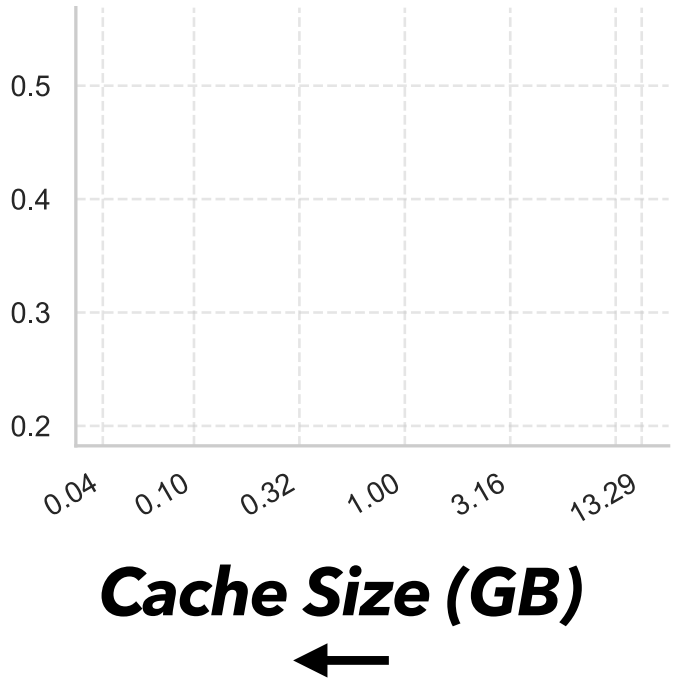




The **problem** is that long contexts are very costly to serve due to size of the model's KV cache.

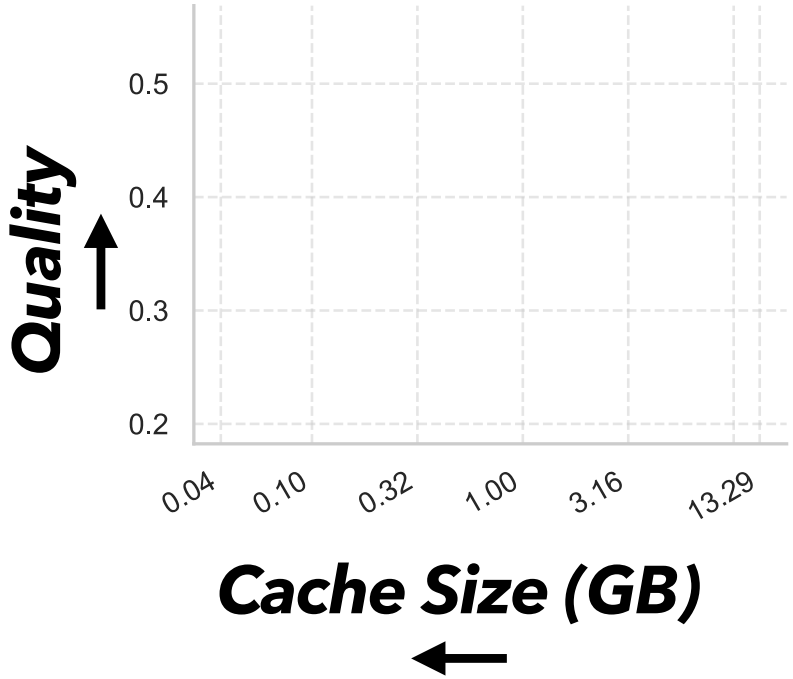
The **problem** is that long contexts are very costly to serve due to size of the model's KV cache.

Our **objective** is to *reduce* ***cache size***



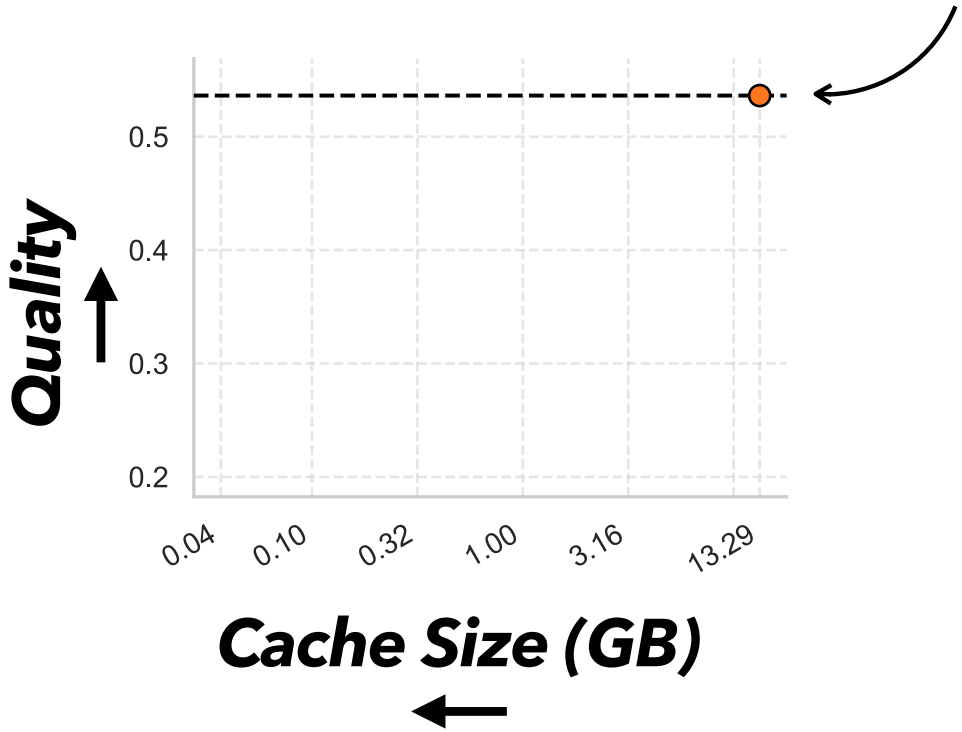
The **problem** is that long contexts are very costly to serve due to size of the model's KV cache.

Our **objective** is to *reduce* ***cache size*** *while* maintaining or improving ***quality***



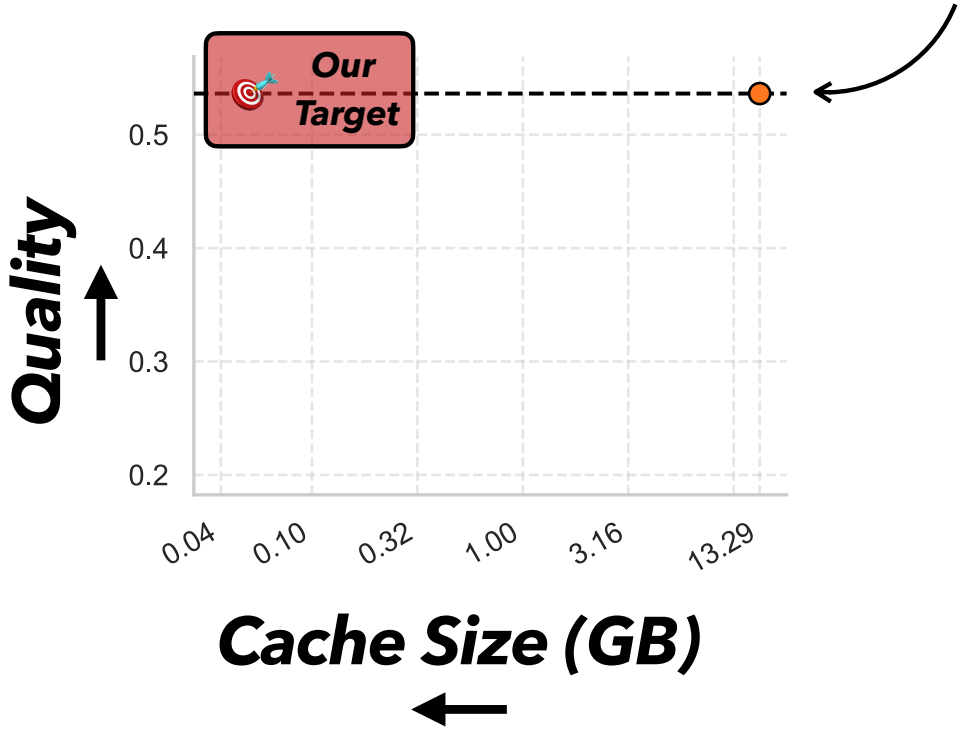
The **problem** is that long contexts are very costly to serve due to size of the model's KV cache.

Our **objective** is to *reduce* ***cache size*** *while* maintaining or improving ***quality***, *relative to a* ***full KV cache***



The **problem** is that long contexts are very costly to serve due to size of the model's KV cache.

Our **objective** is to *reduce* ***cache size*** *while* maintaining or improving ***quality***, *relative to a* ***full KV cache***



Talk outline

Background. *Why is it so costly to serve long contexts?*

Space-quality tradeoffs. *Why don't existing approaches suffice?*

Cartridges. Trading compute for space with offline training.

Self-study. Synthetic data generation enables generalization.

Results. Expanding the space-quality frontier and more!

Disclaimer

*Due to time, the **related work** that follows is incomplete!*

Disclaimer

*Due to time, the **related work** that follows is incomplete!*

*In the **Q/A**, ask me about how Cartridges relates to the efficient architectures below!*

Introduce Sparsity

Longformer [Beltagy et al. 2020]
BigBird [Ainslie et al. 2020]
Sparse Transformers [Child et al. 2019]
…and many others.

Reduce # of key-value heads

Multi-query Attention [Shazeer 2019]
Grouped-query Attention [Ainslie et al. 2023]

Use low rank key-value heads

Multi-head latent attention [Liu 2024]

Linearize attention

[Katharopoulos et al. 2020]
Performer [Choromanski et al. 2022]
Based [Arora, Eyuboglu et al. 2024]
…and many others.

+ modified update rules

Mamba [Gu and Dao 2024]
DeltaNet [Yang et al. 2024]
Titans [Behrouz et al. 2024]
…and many others.

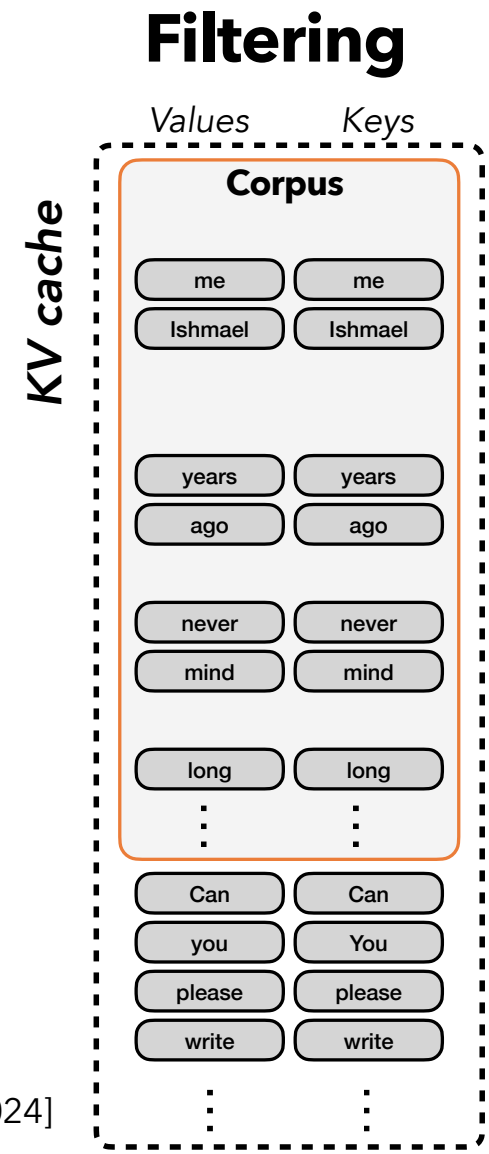
Prompt compression

Hard-token prompt compression methods reduce the number of tokens in the context.

Examples: LLMLingua [Jiang *et al.* 2023], Adacomp [Zhang *et al.* 2024]

Prompt compression

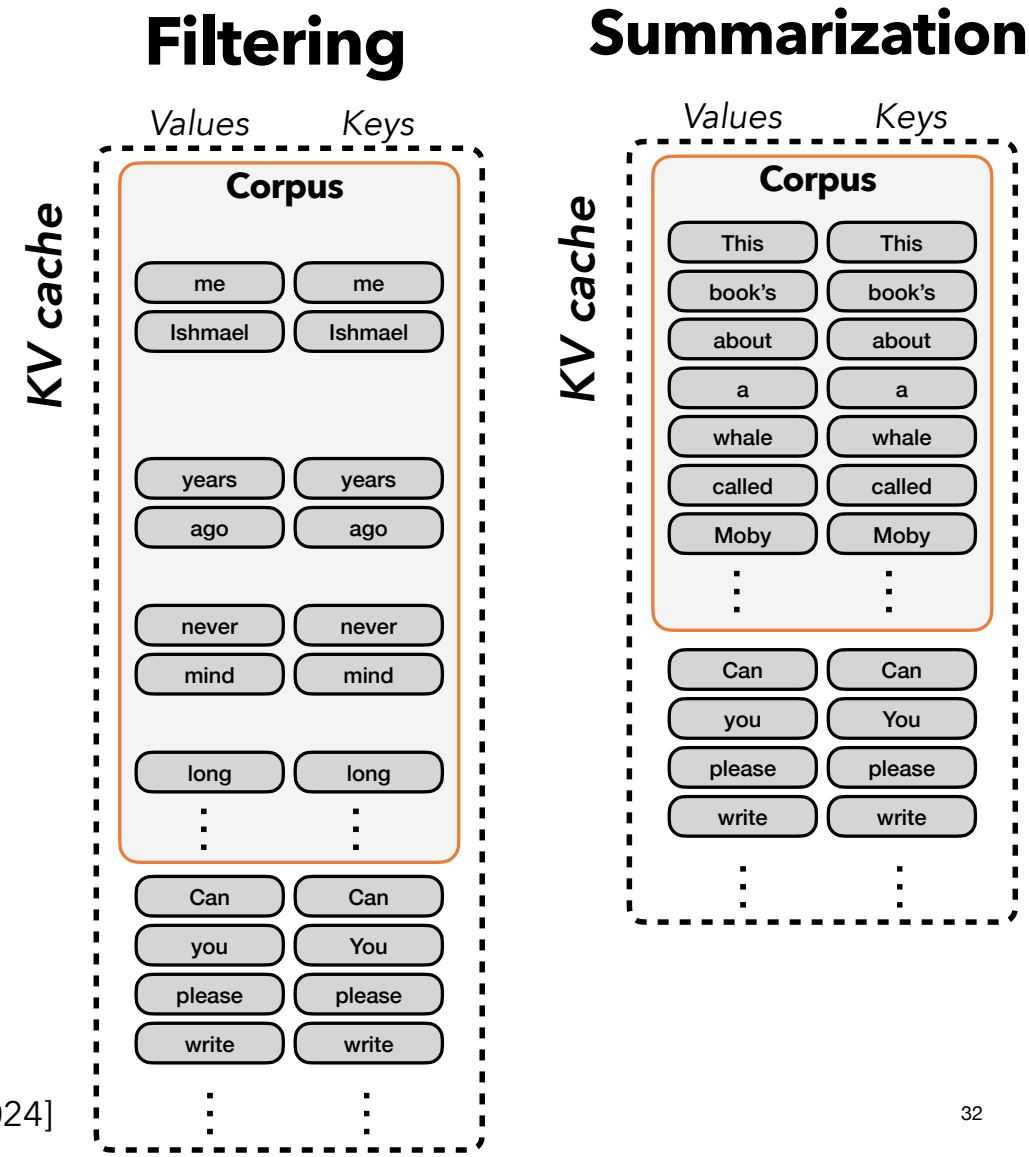
Hard-token prompt compression methods reduce the number of tokens in the context.



Examples: LLMLingua [Jiang *et al.* 2023], Adacomp [Zhang *et al.* 2024]

Prompt compression

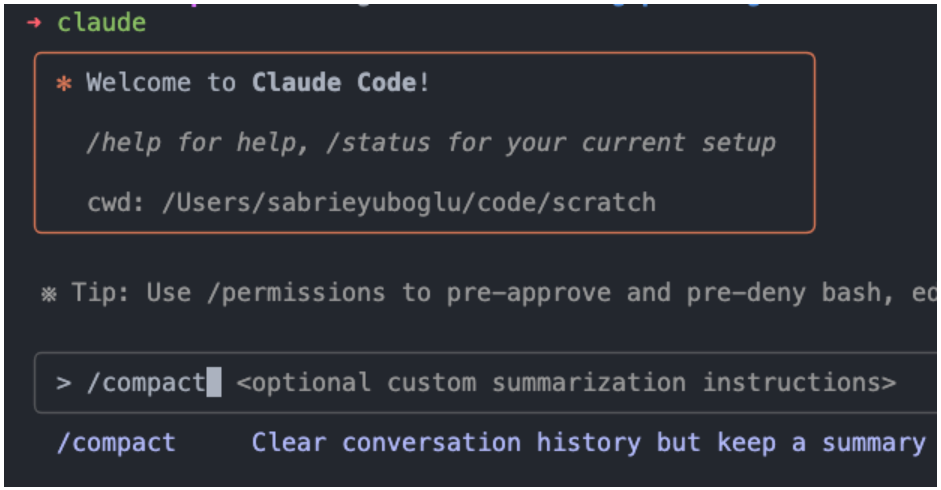
Hard-token prompt compression methods reduce the number of tokens in the context.

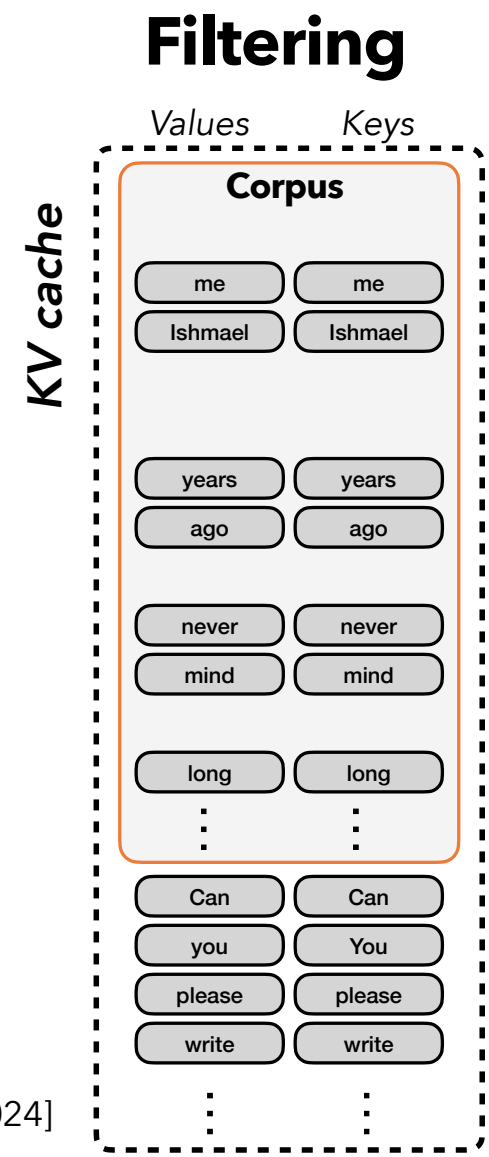


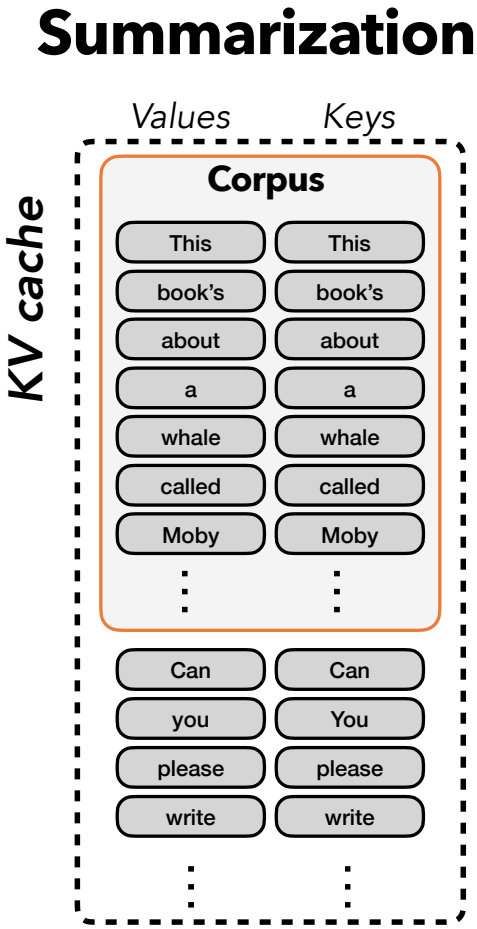
Examples: LLMLingua [Jiang *et al.* 2023], Adacomp [Zhang *et al.* 2024]

Prompt compression

Hard-token prompt compression methods reduce the number of tokens in the context.



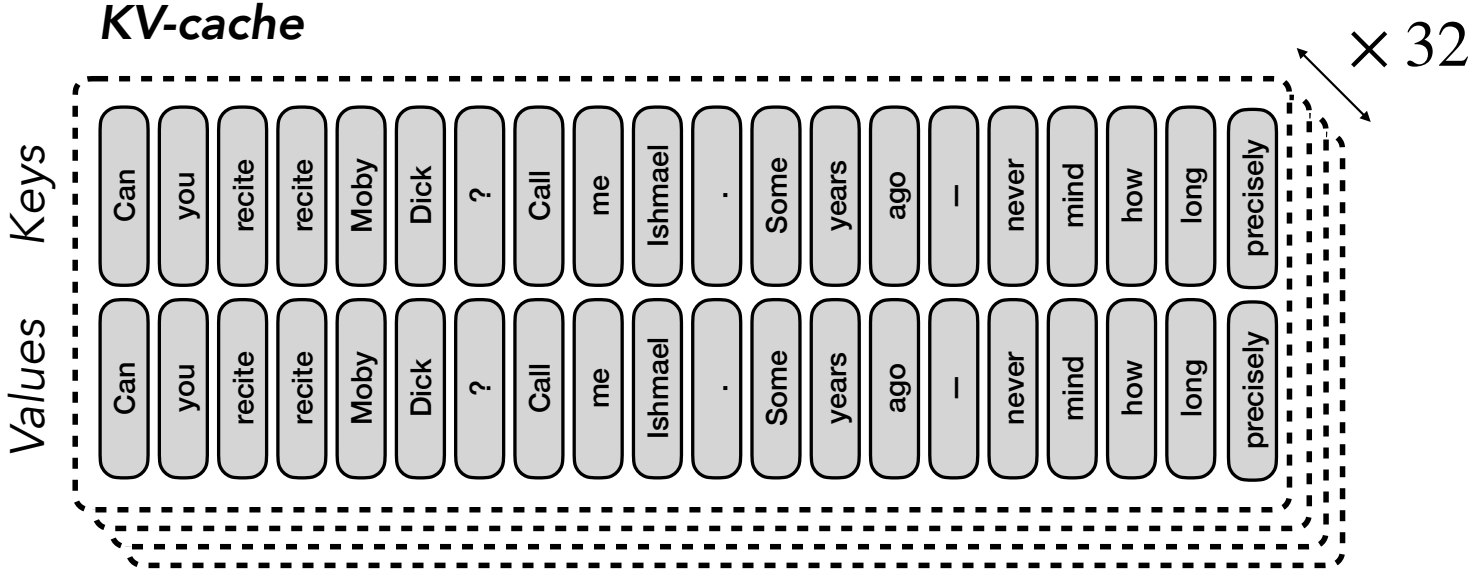




Examples: LLMLingua [Jiang *et al.* 2023], Adacomp [Zhang *et al.* 2024]

KV cache compression

Every layer maintains it's own keys and values!



Idea: drop different tokens at different layers!

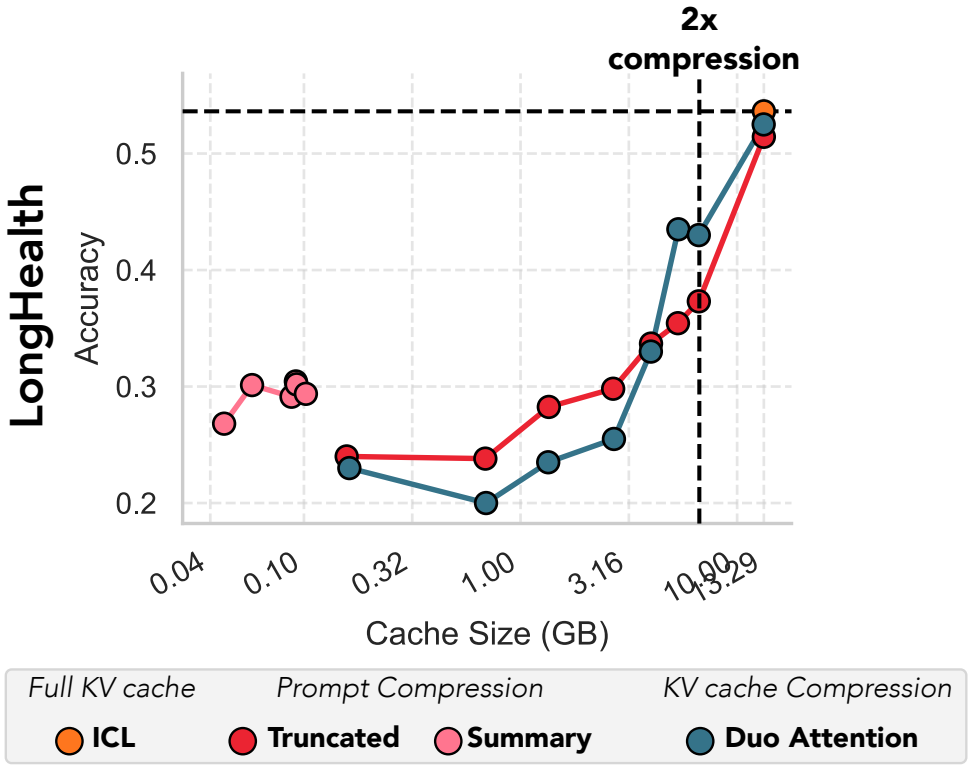
For example: DuoAttention [Xiao *et al.* 2024], H2O [Zhang *et al.* 2023]

Space-quality tradeoffs

Problem: these methods struggle to maintain quality at large compression ratios.

Space-quality tradeoffs

Problem: these methods struggle to maintain quality at large compression ratios.



Space-quality tradeoffs

There is a tradeoff between space consumed by KV cache and quality of LLM responses.

Other tradeoffs?

There is a tradeoff between space consumed by KV cache and quality of LLM responses.

*Is there something else, other than quality, that we can **trade off** for reduced space consumption?*

Write compute

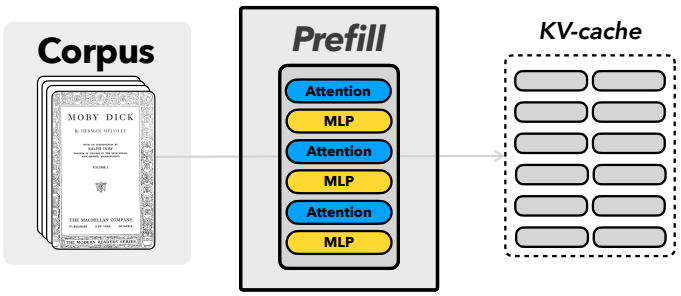
*Is there something else, other than quality, that we can **trade off** for reduced space consumption?*

Write compute: the amount of compute used to construct the KV cache.

Write compute

*Is there something else, other than quality, that we can **trade off** for reduced space consumption?*

Write compute: the amount of compute used to construct the KV cache.



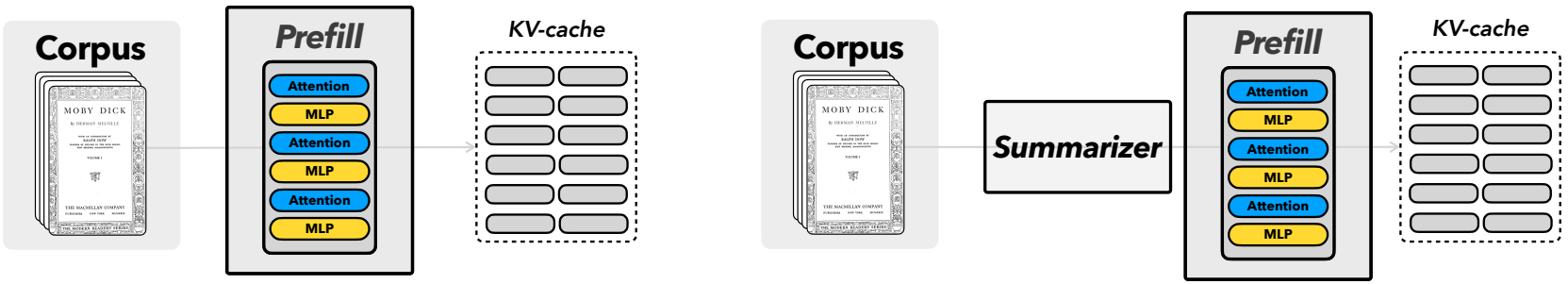
Standard KV Cache

x

Write compute

*Is there something else, other than quality, that we can **trade off** for reduced space consumption?*

Write compute: the amount of compute used to construct the KV cache.



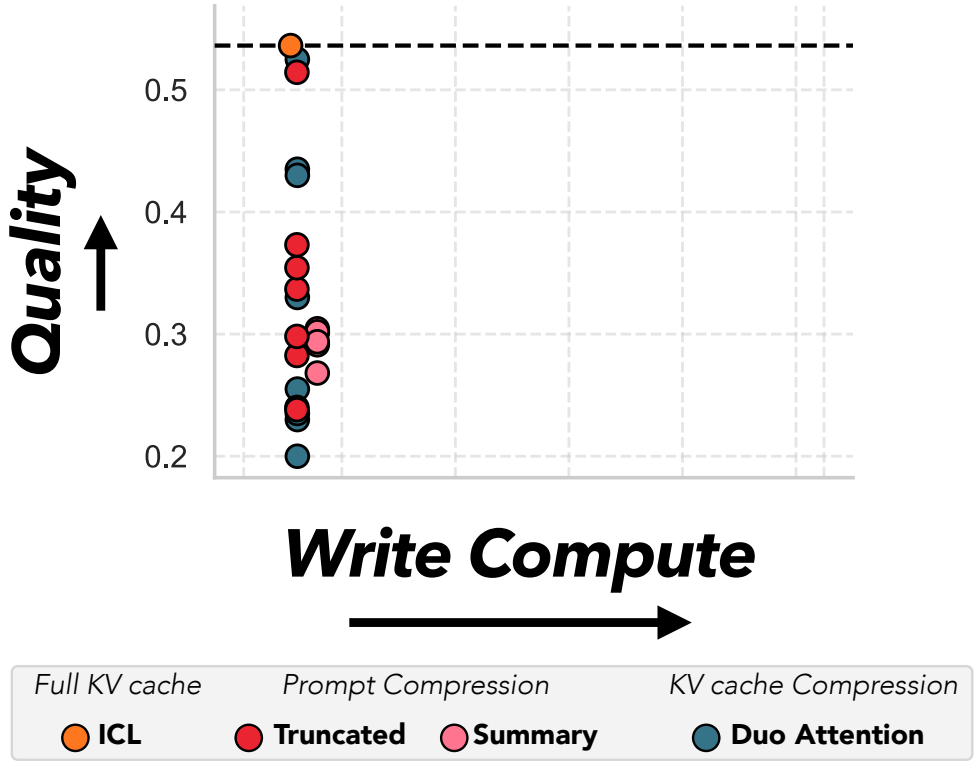
Standard KV Cache

Prompt compression

x

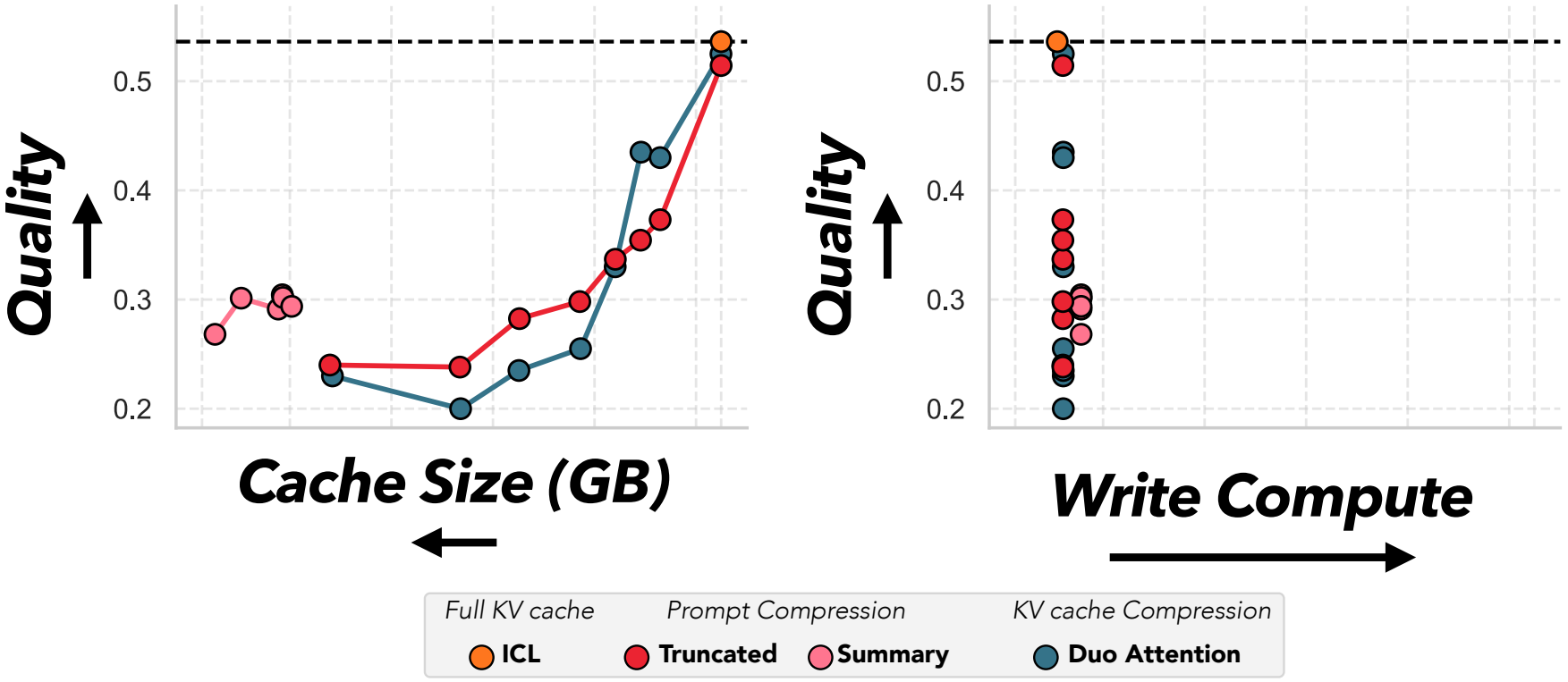
Write compute

Write compute: the amount of compute used to construct the KV cache.



Write compute

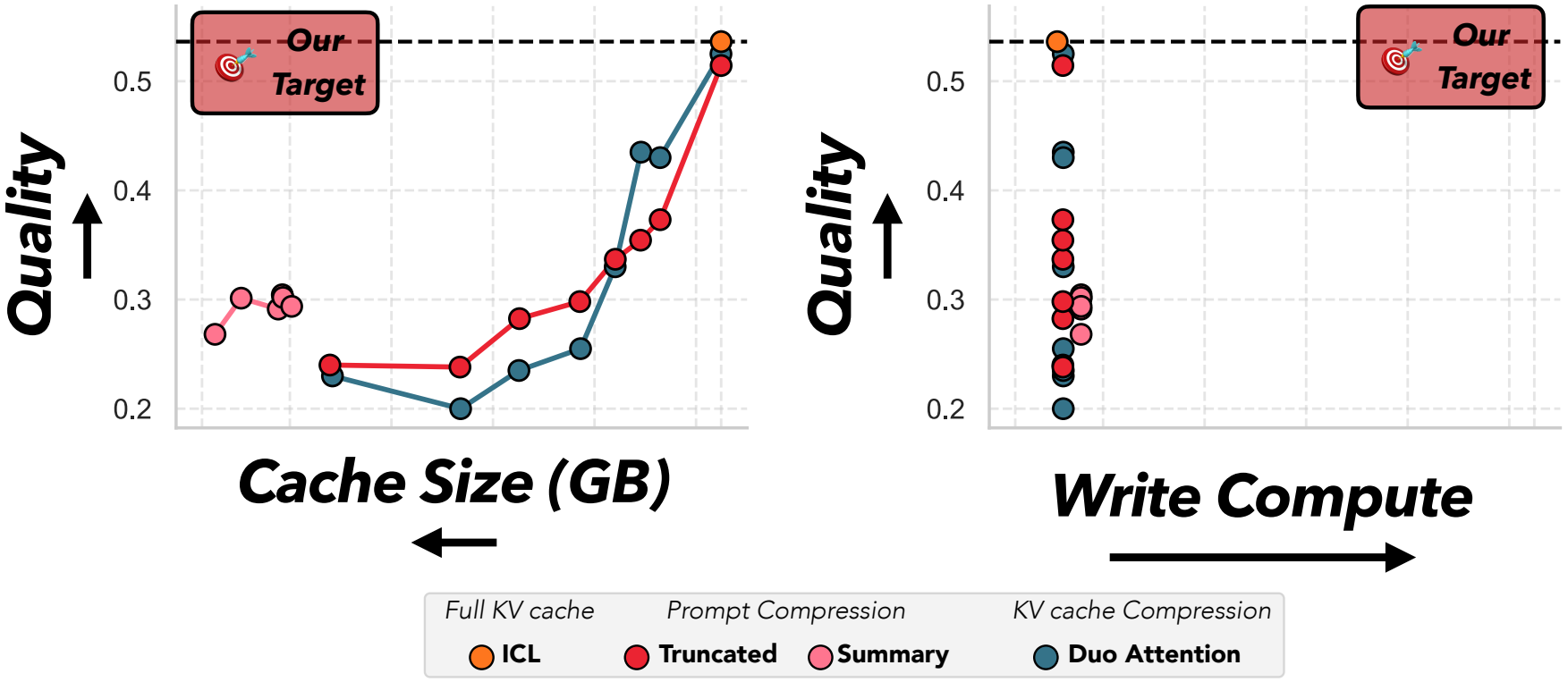
Write compute: the amount of compute used to construct the KV cache.



x

Write compute

*What if we allow ourselves to **increase** the write compute?*

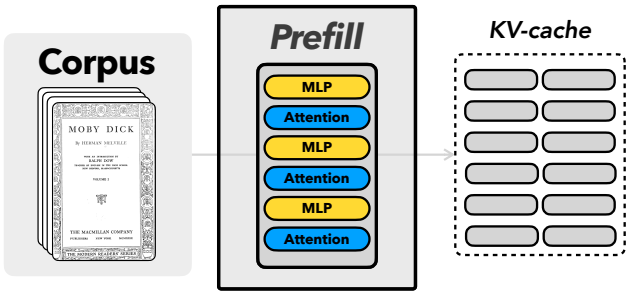


Trading compute for space

These methods use the same or less compute to construct the KV cache/state.

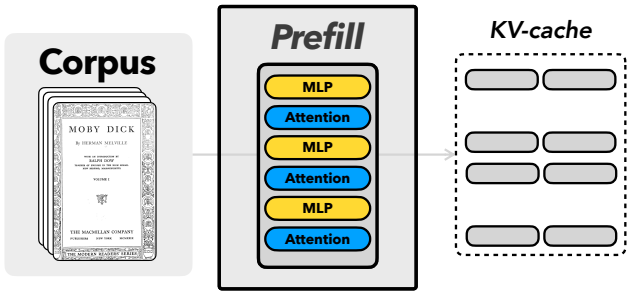
Regular attention

1 forward pass with complexity $O(n^2d + nd^2)$ to construct the KV cache.



Duo attention

1 forward pass with complexity $O(n^2d + nd^2)$ and a cheap pruning step to construct the KV cache.



Background X

Trading write compute for space

Trading write compute for space

The compute used to construct the KV cache can be **amortized across queries** referencing the same context!

Talk outline

Background. *Why is it so costly to serve long contexts?*

Space-quality tradeoffs. *Why don't existing approaches cut it?*

Cartridges. Trading compute for space with offline training.

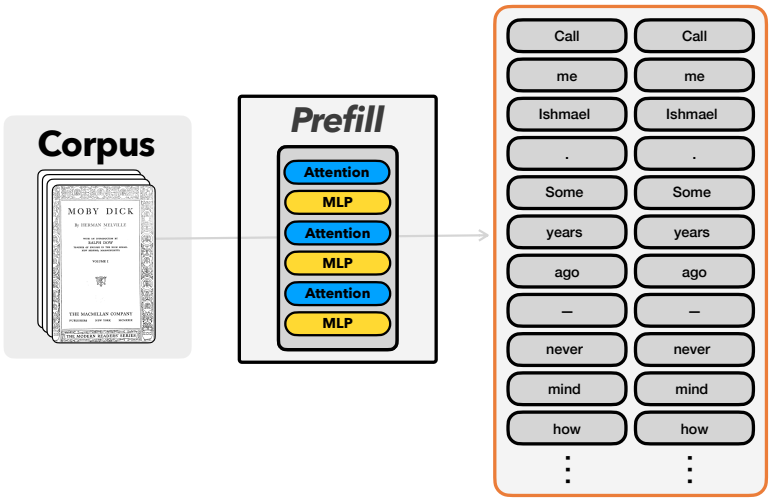
Self-study. Synthetic data generation enables generalization.

Results. Expanding the space-quality frontier and more!

Training the KV cache

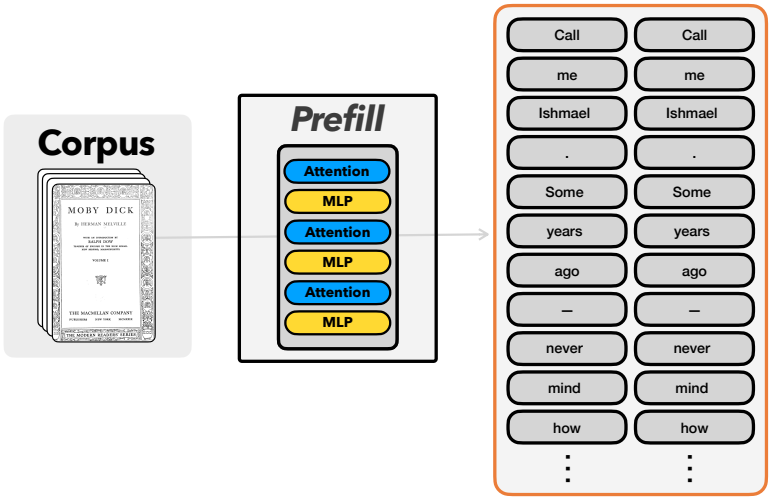
What does it mean to train the KV Cache?

Training the KV cache

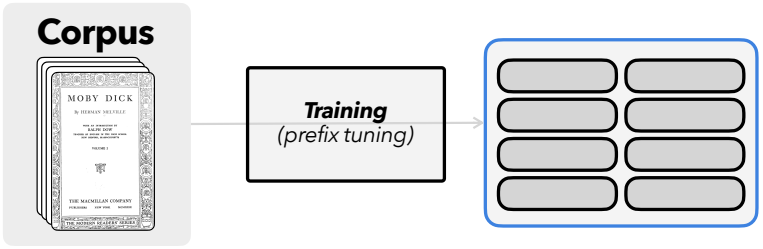


Keys and values
produced with **prefill**

Training the KV cache



Keys and values produced with **prefill**

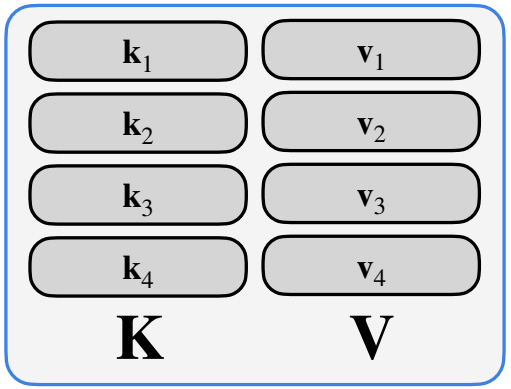


Keys and values trained with **SGD**

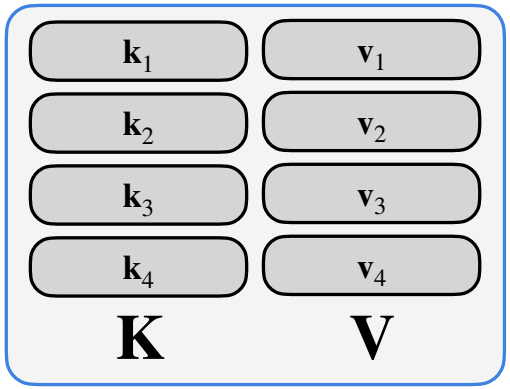
(prefix tuning)

Training the KV cache

We treat the keys and values like they are weights!



Training the KV cache



Given a loss function ℓ we can compute the gradient with respect to $\mathbf{K}$ and $\mathbf{V}$ by back-propagating into the KV cache and updating them gradient descent!

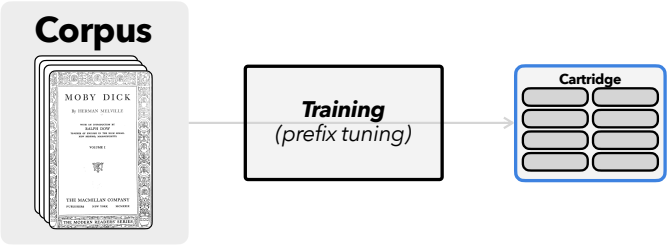
$$\mathbf{K} = \mathbf{K} - \nabla_{\mathbf{K}} \ell(\mathbf{K}, \mathbf{V})$$

$$\mathbf{V} = \mathbf{V} - \nabla_{\mathbf{V}} \ell(\mathbf{K}, \mathbf{V})$$

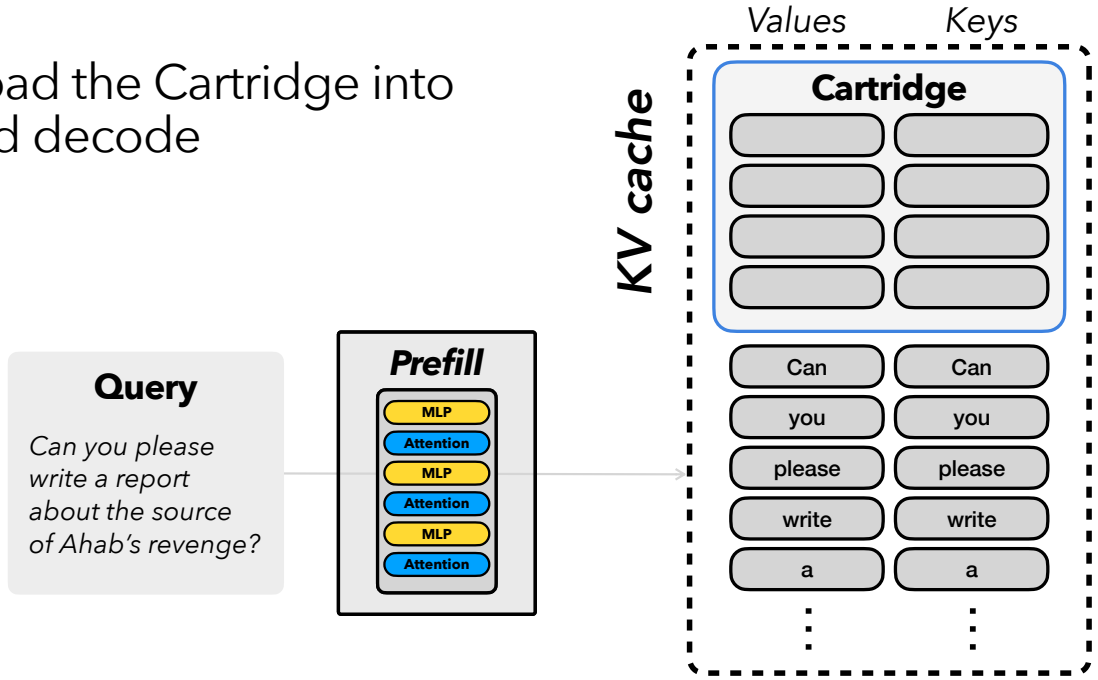
Note: Training the KV cache is equivalent to *prefix tuning*, without the reparameterization.
Li, Xiang Lisa, and Percy Liang (2021)

Cartridges

1 *Train the Cartridge on the corpus* ***offline***



2 At query-time, load the Cartridge into the KV cache and decode



Cartridges

What objective ℓ should we use?

Cartridges

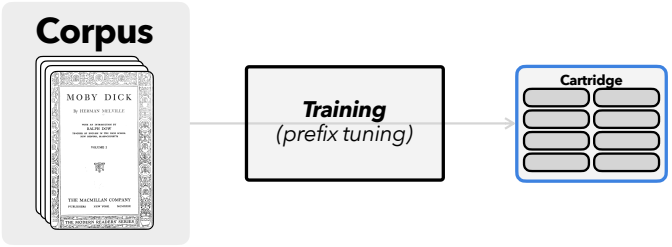
What objective ℓ should we use?

Challenge: choose ℓ so as to maintain generality!

A first attempt: next-token prediction

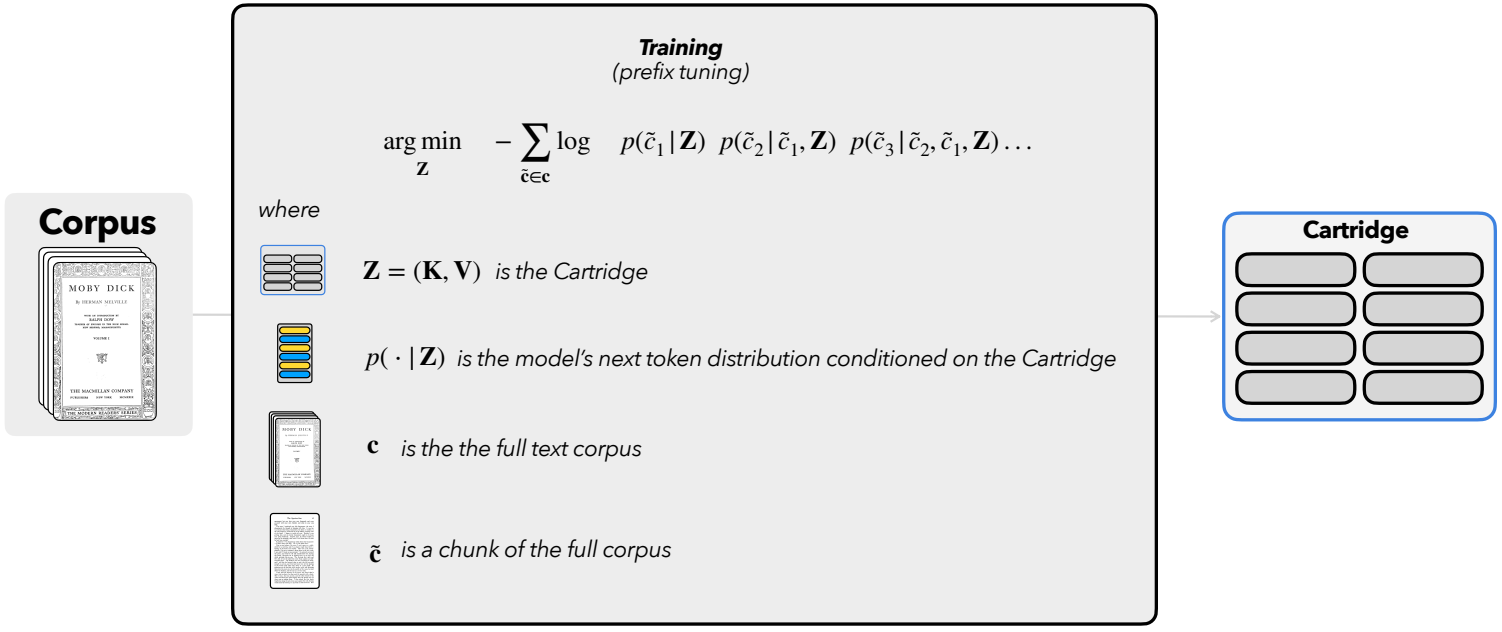
A simple first attempt:

We train the KV cache by computing **next-token prediction loss** on chunks of the corpus.



A first attempt: next-token prediction

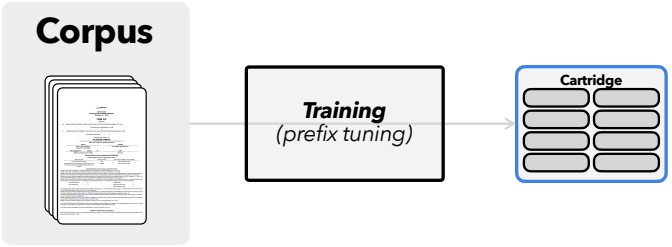
We train the KV cache by computing next-token prediction loss on chunks of the corpus.



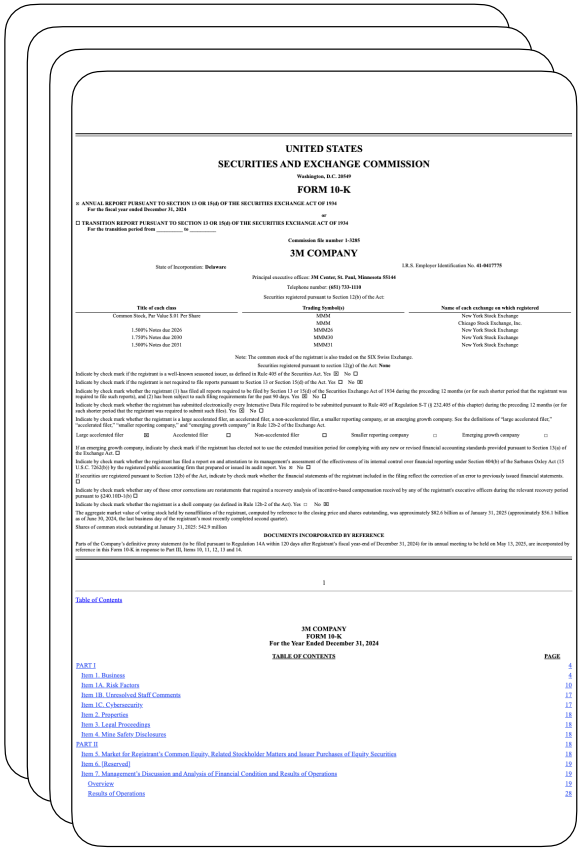
Let’s try it out!

Enough with Moby Dick….

Let’s train a Cartridge on 3M’s 10-K (a ~100 page financial report filed by public companies).



3M Science. Applied to Life.™ **10-K**



UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-K

☒ ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the fiscal year ended December 31, 2024

or

☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the transition period from _____ to _____

Commission file number 1-3285

3M COMPANY

State of Incorporation: Delaware — I.R.S. Employer Identification No. 41-0417775

Principal executive offices: 3M Center, St. Paul, Minnesota 55144

Telephone number: (651) 733-1110

Securities registered pursuant to Section 12(b) of the Act:

Title of each class	Trading Symbol(s)	Name of each exchange on which registered
Common Stock, Par Value $.01 Per Share	MMM	New York Stock Exchange
	MMM	Chicago Stock Exchange, Inc.
1.500% Notes due 2026	MMM26	New York Stock Exchange
1.750% Notes due 2030	MMM30	New York Stock Exchange
1.500% Notes due 2031	MMM31	New York Stock Exchange

Note: The common stock of the registrant is also traded on the SIX Swiss Exchange.

Securities registered pursuant to section 12(g) of the Act: None

Indicate by check mark if the registrant is a well-known seasoned issuer, as defined in Rule 405 of the Securities Act. Yes ☒ No ☐

Indicate by check mark if the registrant is not required to file reports pursuant to Section 13 or Section 15(d) of the Act. Yes ☐ No ☒

Indicate by check mark whether the registrant (1) has filed all reports required to be filed by Section 13 or 15(d) of the Securities Exchange Act of 1934 during the preceding 12 months (or for such shorter period that the registrant was required to file such reports), and (2) has been subject to such filing requirements for the past 90 days. Yes ☒ No ☐

Indicate by check mark whether the registrant has submitted electronically every Interactive Data File required to be submitted pursuant to Rule 405 of Regulation S-T (§ 232.405 of this chapter) during the preceding 12 months (or for such shorter period that the registrant was required to submit such files). Yes ☒ No ☐

Indicate by check mark whether the registrant is a large accelerated filer, an accelerated filer, a non-accelerated filer, a smaller reporting company, or an emerging growth company. See the definitions of "large accelerated filer," "accelerated filer," "smaller reporting company," and "emerging growth company" in Rule 12b-2 of the Exchange Act.

Large accelerated filer ☒ Accelerated filer ☐ Non-accelerated filer ☐ Smaller reporting company ☐ Emerging growth company ☐

If an emerging growth company, indicate by check mark if the registrant has elected not to use the extended transition period for complying with any new or revised financial accounting standards provided pursuant to Section 13(a) of the Exchange Act. ☐

Indicate by check mark whether the registrant has filed a report on and attestation to its management's assessment of the effectiveness of its internal control over financial reporting under Section 404(b) of the Sarbanes-Oxley Act (15 U.S.C. 7262(b)) by the registered public accounting firm that prepared or issued its audit report. Yes ☒ No ☐

If securities are registered pursuant to Section 12(b) of the Act, indicate by check mark whether the financial statements of the registrant included in the filing reflect the correction of an error to previously issued financial statements. ☐

Indicate by check mark whether any of those error corrections are restatements that required a recovery analysis of incentive-based compensation received by any of the registrant's executive officers during the relevant recovery period pursuant to §240.10D-1(b). ☐

Indicate by check mark whether the registrant is a shell company (as defined in Rule 12b-2 of the Act). Yes ☐ No ☒

The aggregate market value of voting stock held by nonaffiliates of the registrant, computed by reference to the closing price and shares outstanding, was approximately $82.6 billion as of January 31, 2025 (approximately $56.1 billion as of June 30, 2024, the last business day of the registrant's most recently completed second quarter).

Shares of common stock outstanding at January 31, 2025: 542.9 million

DOCUMENTS INCORPORATED BY REFERENCE

Parts of the Company's definitive proxy statement (to be filed pursuant to Regulation 14A within 120 days after Registrant's fiscal year-end of December 31, 2024) for its annual meeting to be held on May 13, 2025, are incorporated by reference in this Form 10-K in response to Part III, Items 10, 11, 12, 13 and 14.

1

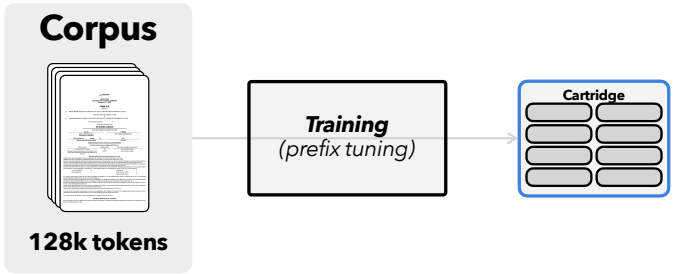
Table of Contents

3M COMPANY
FORM 10-K
For the Year Ended December 31, 2024

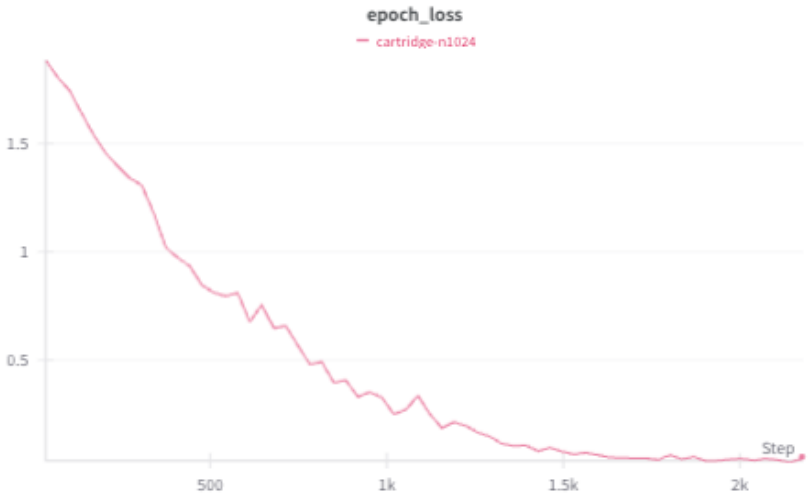
TABLE OF CONTENTS — PAGE

Let's try it out!

Let's train a Cartridge on 3M's 10-K (a ~100 page financial report filed by public companies).



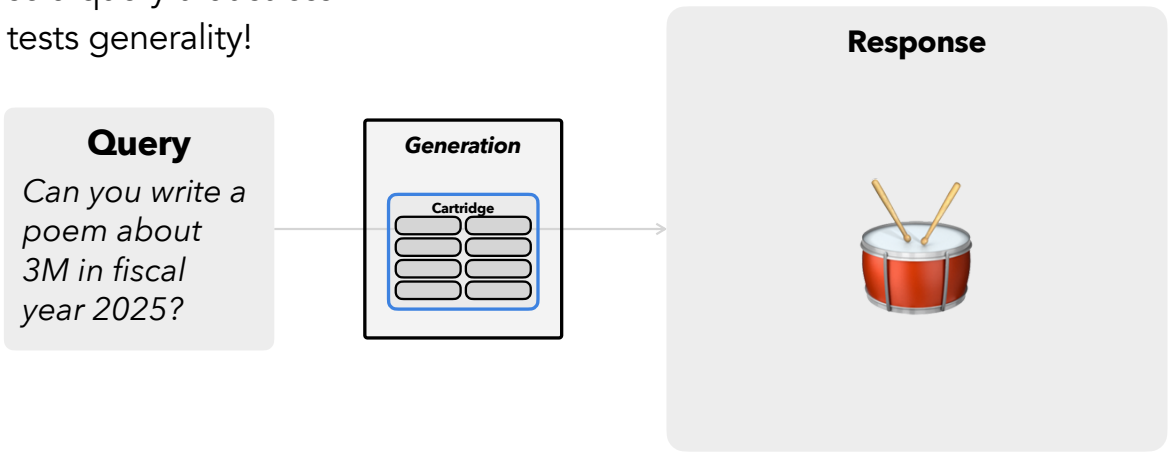
Cartridge 1024 tokens can achieve **zero** next token prediction loss over 128k token corpus (125x compression)!



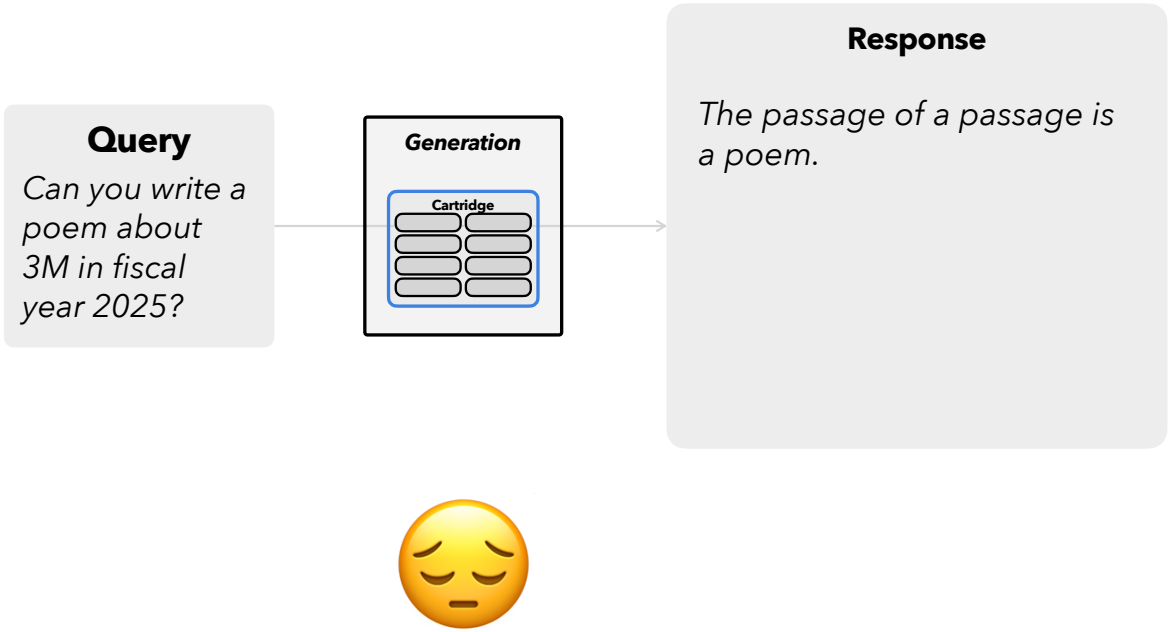
Note: Recent work has also found compression ratios like this (up to 1500x). Kuratov *et al.* *(2025)*

Let’s try it out!

We use a query that stress tests generality!



Let’s try it out!



Problem: Poor generalization

***Memorization**
e.g. "Please complete the rest of the passage..."

Data structuring tasks
e.g. "Please list AMD's customers in JSON format"

Synthesis tasks
e.g. "Please summarize the AMD's FY20 10K."

Creative tasks
e.g. "Write a poem about AMD's Q3 performance."

Mathematical reasoning
e.g. "Compute AMD FY20 days payable outstanding."

Disjoint reasoning
e.g. "List all the tables in AMD's FY20 10K document."

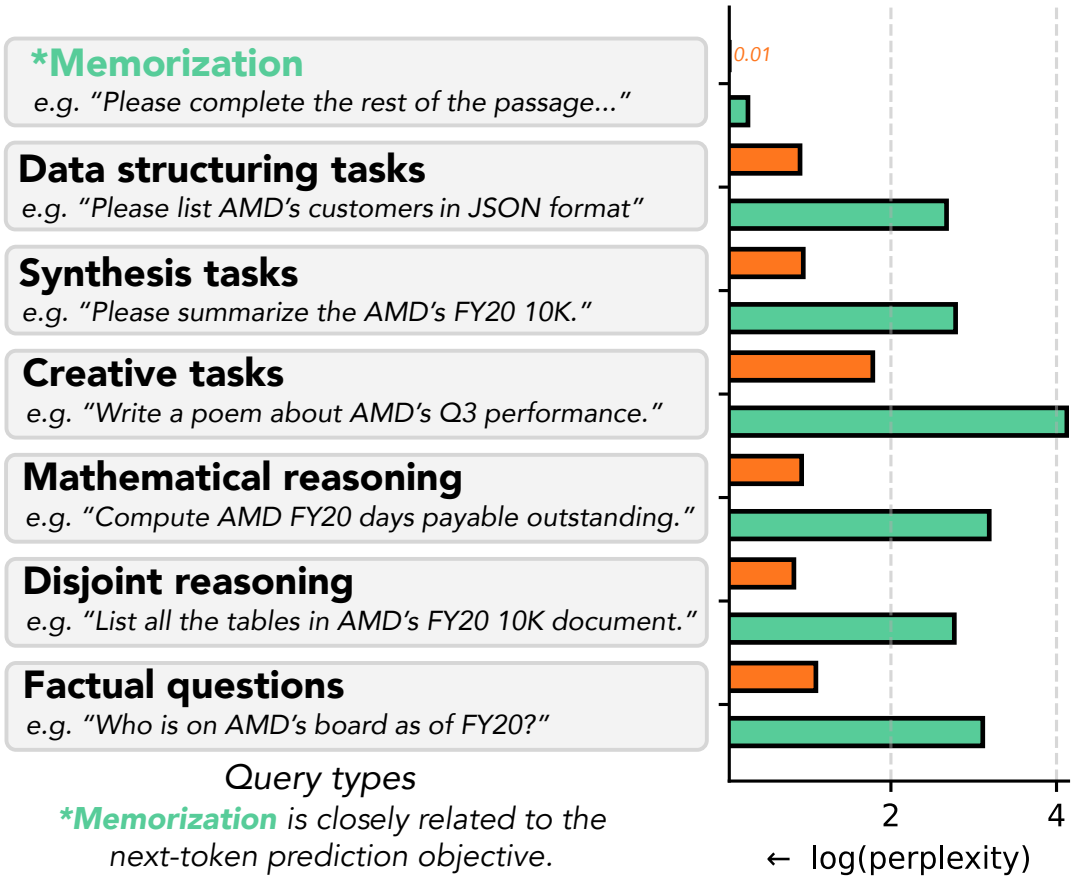
Factual questions
e.g. "Who is on AMD's board as of FY20?"

Query types

****Memorization** is closely related to the next-token prediction objective.*

Problem: Poor generalization

Generalization to diverse queries



Query types

***Memorization** is closely related to the next-token prediction objective.*

Problem: Poor generalization with next token prediction

Idea: Replace the next token prediction objective

Self-study

Talk outline

Background. *Why is it so costly to serve long contexts?*

Space-quality tradeoffs. *Why don't existing approaches cut it?*

Cartridges. Trading compute for space with offline training.

Self-study. Synthetic data generation enables generalization.

Results. Expanding the space-quality frontier and more!

Self-study

Self-study proceeds in two stages:

1. ***Synthesize*** *conversations about the corpus by sampling from an LLM with a chunk of the corpus in context.*

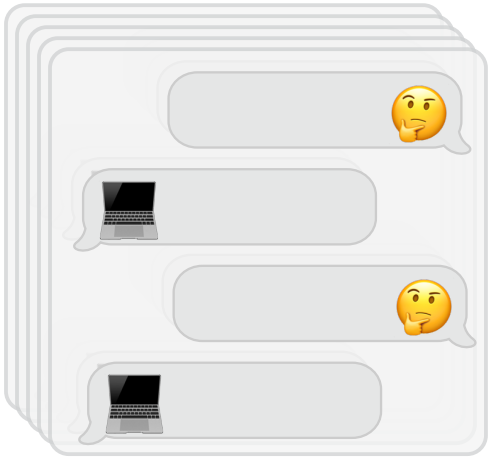
2. ***Train*** *the KV-cache with a context distillation objective on the synthesized conversations.*

Self-study

Self-study proceeds in two stages:

1. ***Synthesize*** *conversations about the corpus by sampling from an LLM with a chunk of the corpus in context.*

2. ***Train*** *the KV-cache with a context distillation objective on the synthesized conversations.*



Self-study

Self-study proceeds in two stages:

1. ***Synthesize*** *conversations about the corpus by sampling from an LLM with a chunk of the corpus in context.*

$$\arg\min_{\mathbf{Z}} \sum_{(\mathbf{x},\tilde{\mathbf{c}})\in\mathbf{C}} D_{\mathrm{KL}}(p(\cdot\,|\,\mathbf{Z})\,||\,p(\cdot\,|\,\tilde{\mathbf{c}})) + D_{\mathrm{KL}}(p(\cdot\,|\,x_1,\mathbf{Z})\,||\,p(\cdot\,|\,x_1,\tilde{\mathbf{c}})) + \ldots$$

2. ***Train*** *the KV-cache with a context distillation objective on the synthesized conversations.*

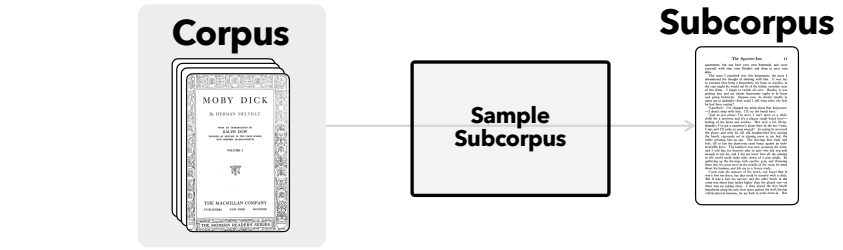
1 **Self-study:** Data synthesis

A Sample a **subcorpus**

B Sample a **seed prompt**

C Sample a **conversation**

conditioned on subcorpus and seed prompt

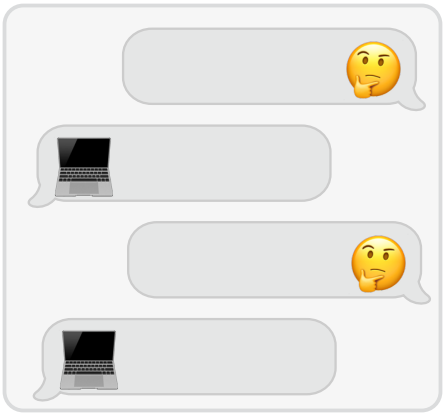


Please generate a single chat message instructing an LLM to **structure the information in JSON.**

Please generate a single **question** that tests an LLM's knowledge of the information above.

Please generate a single chat message instructing an LLM to **summarize the context**.

Think of a practical, real-world task related to context , and start a conversation with a chatbot about it.



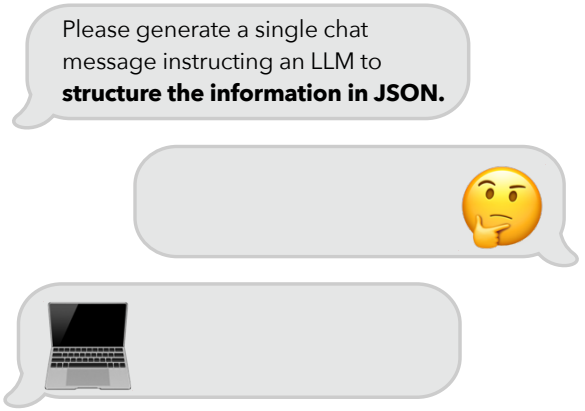
Self-study: Data synthesis

A Sample a **subcorpus**

B Sample a **seed prompt**

C Sample a **conversation**

conditioned on subcorpus and seed prompt



Slide is WIP

Self-study

1. Sample a **chunk** of corpus

2. Sample a **seed prompt**

3. Sample a **conversation**

   conditioned on chunk and prompt

Background X

2 **Self-study:** Context distillation

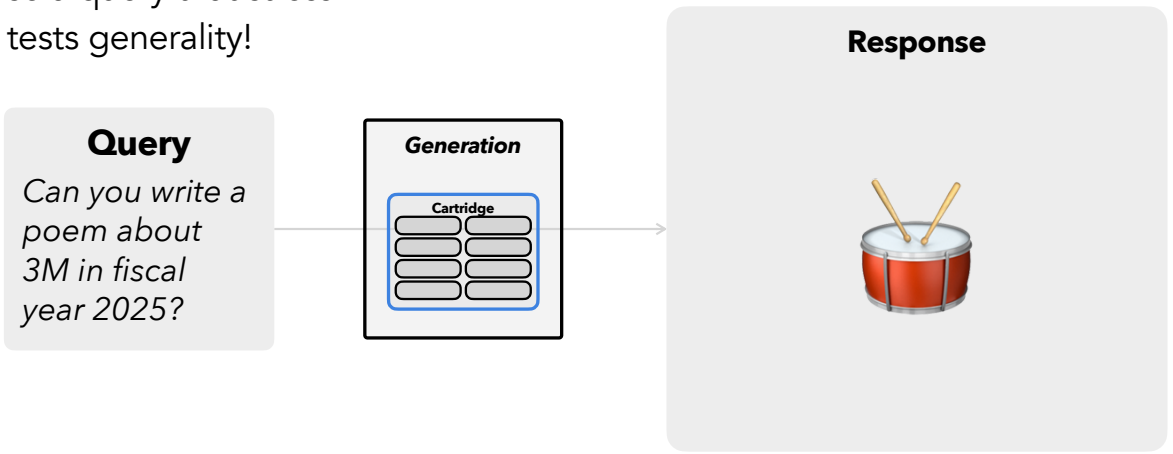
Minimize the KL divergence between the next token distributions $p(\cdot \mid \mathbf{K}, \mathbf{V})$ (model conditioned on the Cartridge) and $p(\cdot \mid \tilde{\mathbf{c}})$ (model conditioned on the text chunk)!

Note: this is related to the standard model distillation objective [Hinton *et al.* 2015, Kim and Rush 2016]. Typically, this is applied between a different models (*i.e.* a teacher and a student).

More recently, this idea of distilling *context* into parameters has been proposed. [Snell *et al.* 2022, Kujanpaa *et al.* 2024]

Let’s try it out!

We use a query that stress tests generality!

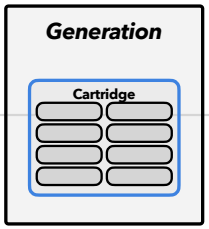


Let’s try it out!

We use a query that stress tests generality!

Query

Can you write a poem about 3M in fiscal year 2025?



Response

ThreeM, a company of might,
With innovations in sight,
Their products a marvel to behold,
In various markets, stories unfold.

Industrial, Safety and Graphics too,
Health Care, Electronics, and Energy anew,
Consumer, with products for all,
A diverse range, for one and all.

Their business segments, a tale to tell,
Of growth and profit, of challenges to sell,
The company's performance, a story to share,
Of successes and failures, with lessons to spare.

The board of directors, a group of wise men,
Guiding the company, with vision and men,
Their decisions, a path to follow,
To drive growth, and profits to grow.

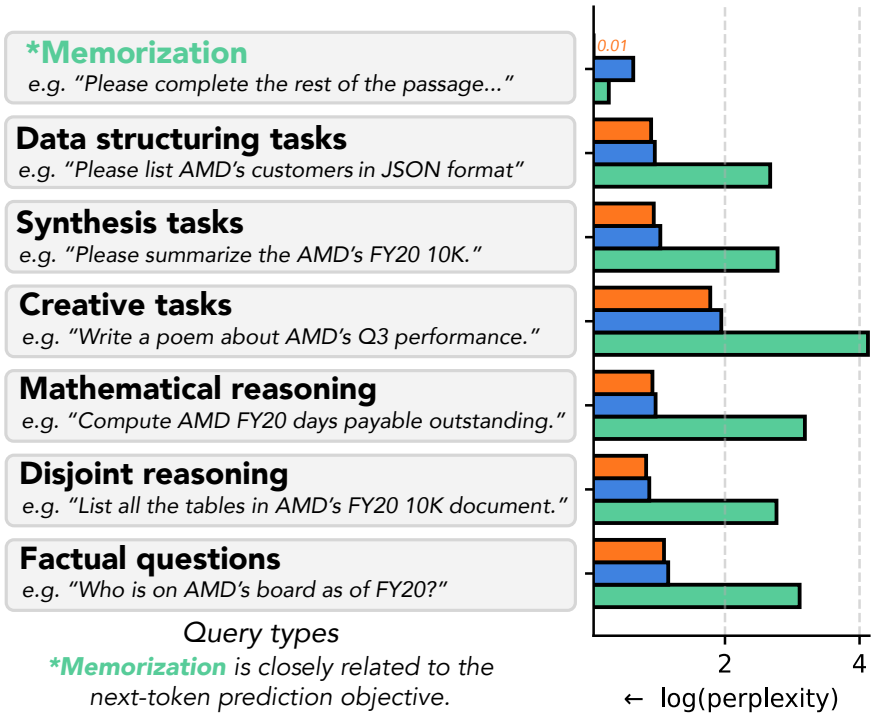
The compensation plan, a system to abide,
To attract and retain, with rewards to provide,
The stock option plan, a way to share,
The company's success, with employees to spare.

The executive compensation, a measure to take,
To ensure fairness, and a fair shake,
The bonus plan, a way to reward,
The company's performance, with a bonus to be had.

The accounting standards, a set of rules to follow,
To ensure transparency, and a fair show,
The financial statements, a report to share,
With information to help

Self-study enables generalization

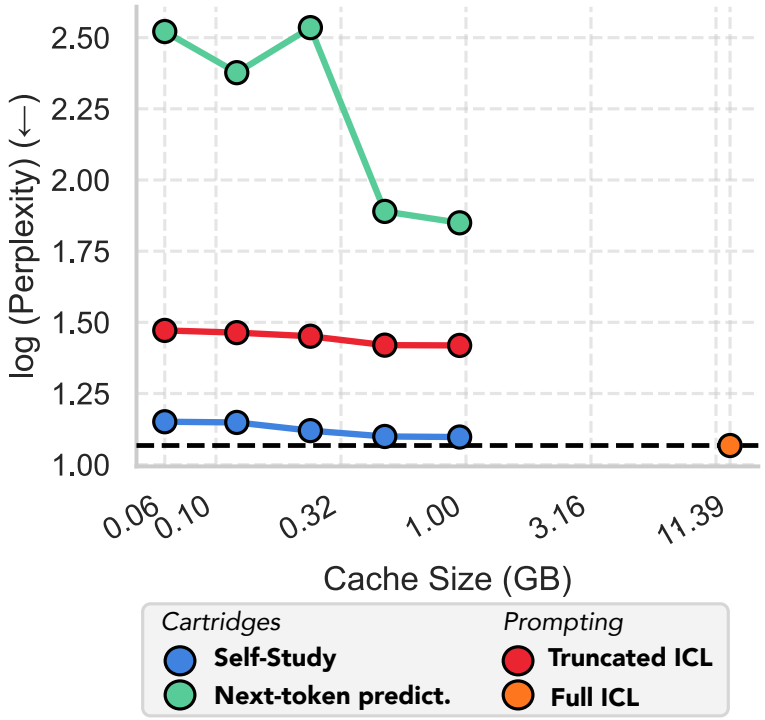
Generalization to diverse queries



Query types

***Memorization** is closely related to the next-token prediction objective.*

Quality-memory tradeoff



Talk outline

Background. *Why is it so costly to serve long contexts?*

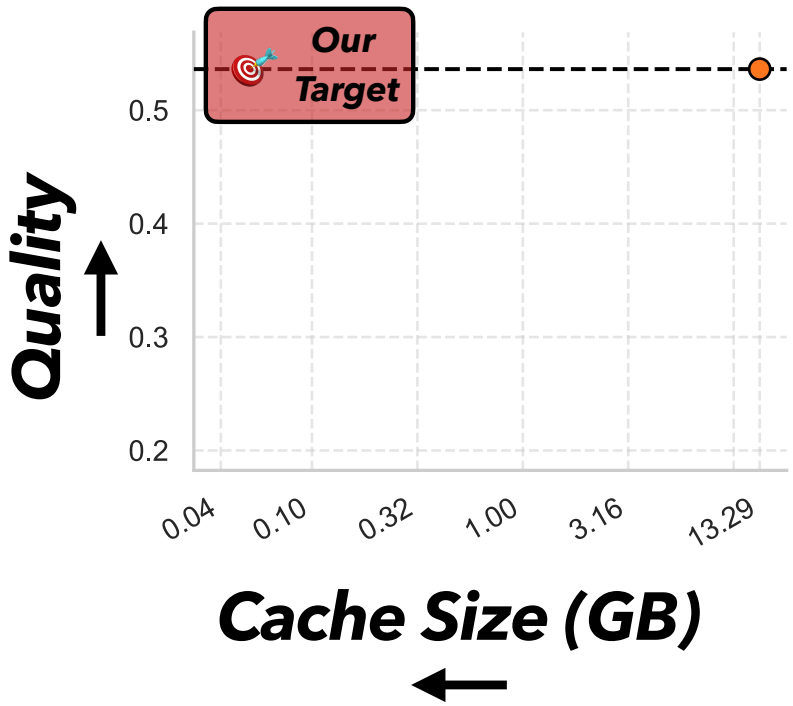
Space-quality tradeoffs. *Why don't existing approaches cut it?*

Cartridges. Trading compute for space with offline training.

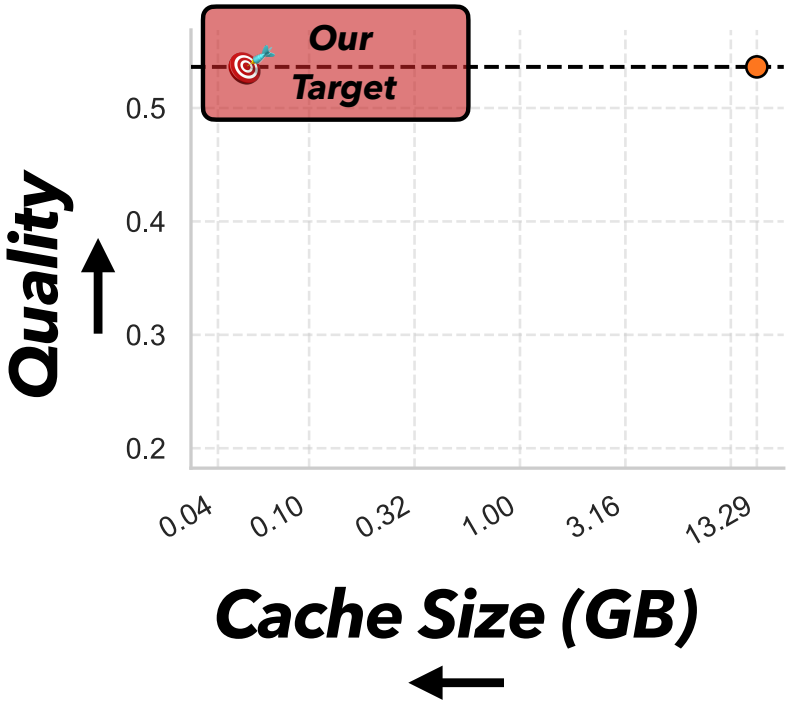
Self-study. Synthetic data generation enables generalization.

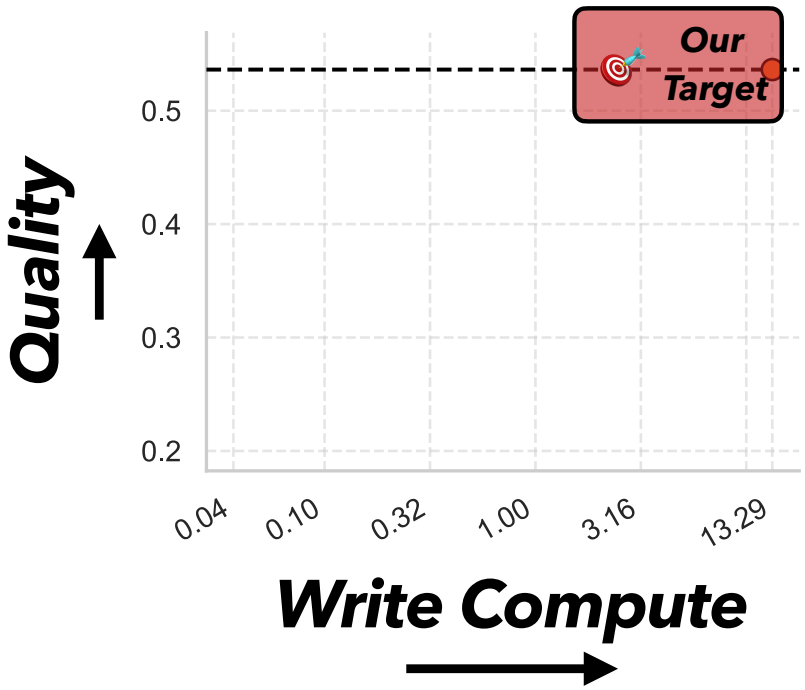
Results. Expanding the space-quality frontier and more!

Our **objective** is to *reduce* ***cache size*** *while* maintaining or improving ***quality***, *relative to a* ***full KV cache***

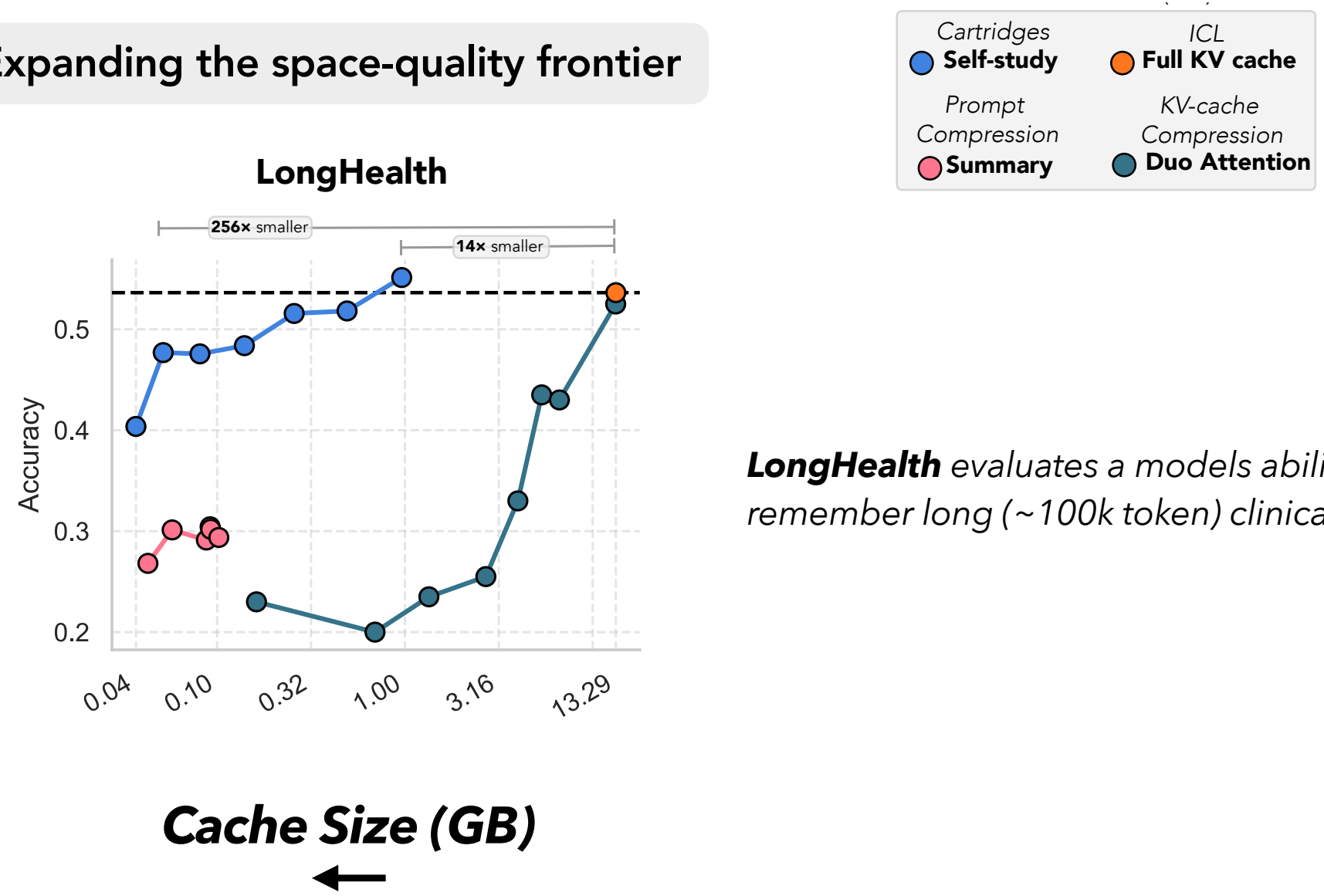


Our **idea** is to achieve this by increasing the compute we use *to* create our KV cache



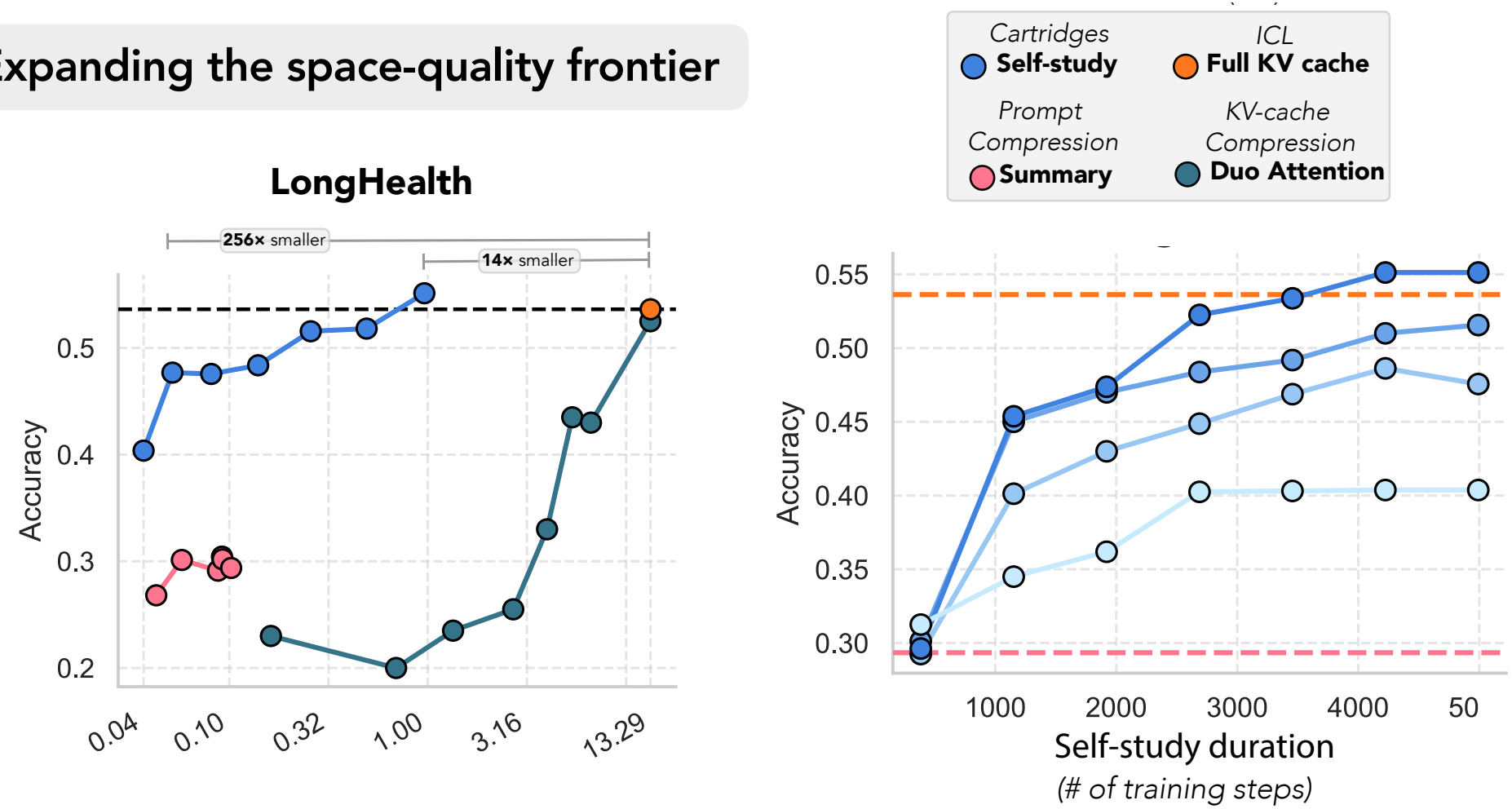


Expanding the space-quality frontier



LongHealth *evaluates a models ability to remember long (~100k token) clinical texts.*

Expanding the space-quality frontier



Expanding the space-quality frontier

LongHealth *evaluates a models ability to remember long clinical texts.*

Space-Quality Tradeoff with Cartridges

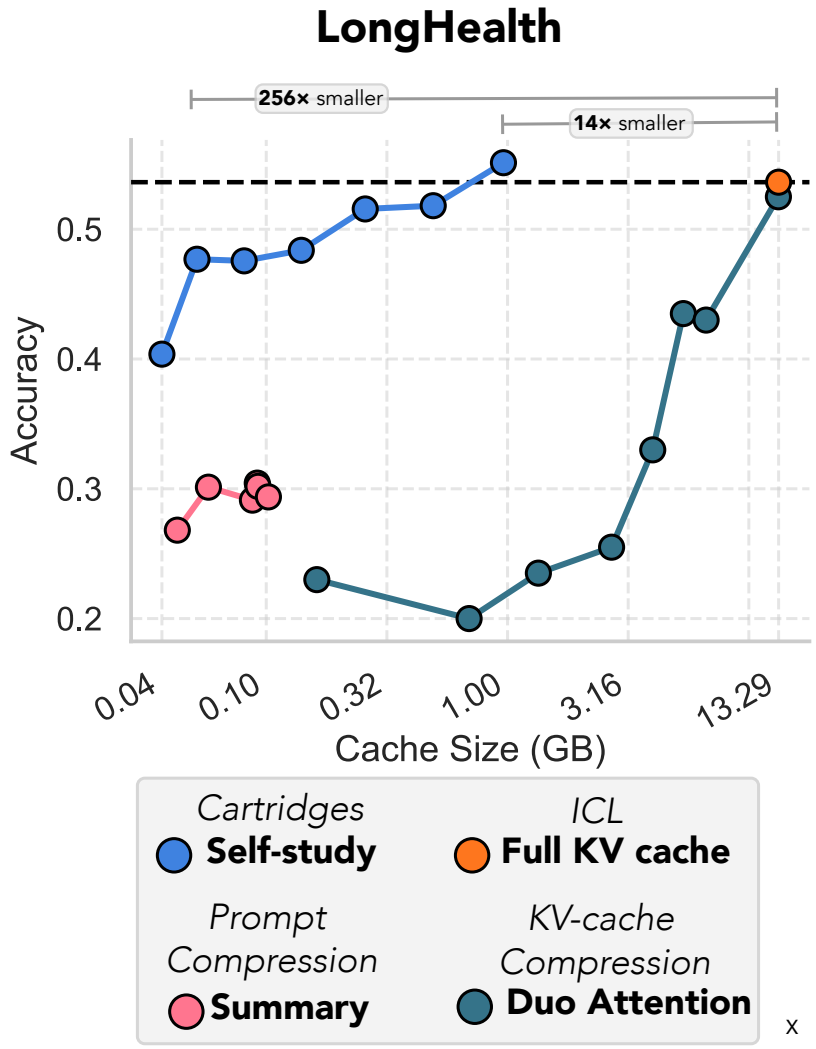
On the LongHealth benchmark, a 0.96 GB cartridge reached 55.1% accuracy, exceeding the full ICL baseline. This corresponds to a 13.8× reduction in cache size.

Compression and Throughput Results

A 52 MB cartridge (256× compression, 121× throughput increase) achieved 47.7% accuracy.

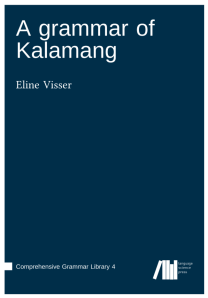
Comparison with prompt and cache compression

Duo Attention, a recent cache compression algorithm, achieved 43.0% accuracy at 6.77 GB (2× compression).



Extending the context length

Machine-translation from one book (MTOB) *evaluates a model's ability to translate a rare* language *given only a textbook of its grammar.*

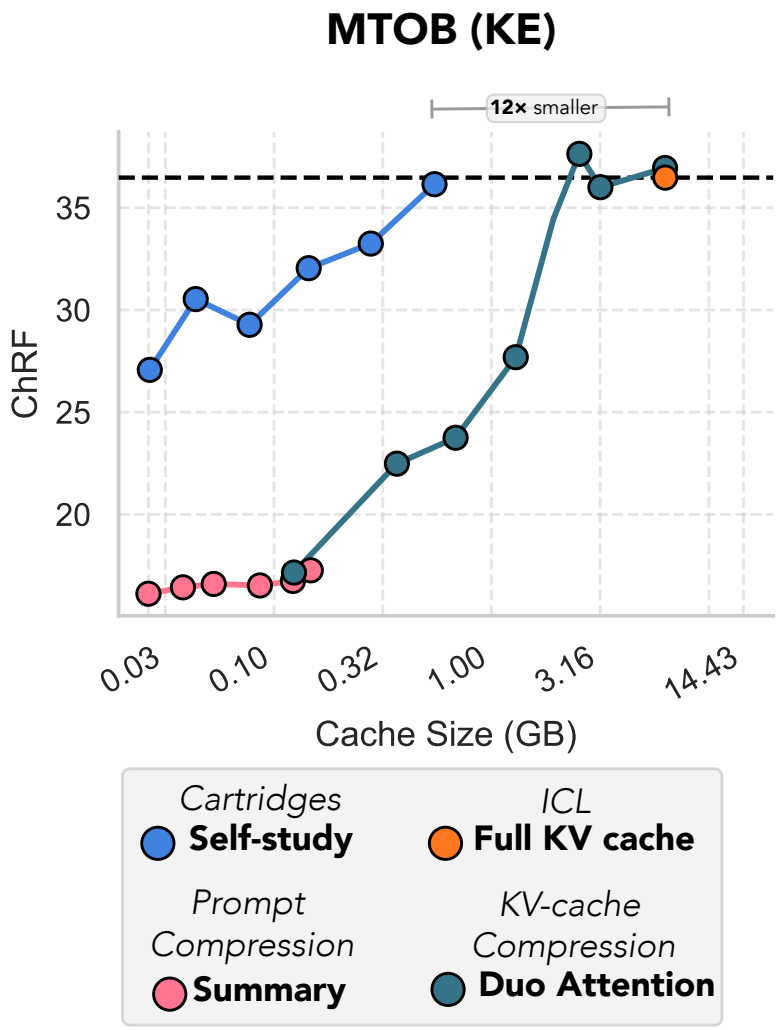


572 Pages

Key challenge: *the full textbook is 484k tokens, but Llama3.1 8B only has 128k token context.*

Self-study on Llama 3.1 8B achieves 36.1 ChRF!

Llama 4 Scout (with 10 million token context length, and 109B parameters) achieves a ChRF of ***36.3***



CARTRDIGE trained on entire book, while ICL, prompt compression, and KV-cache compression used 60k token expert-constructed subset of textbook as context. ICL on first 60k tokens of entire book achieves a ChRF of ~25.1.

Future directions

Theoretical underpinnings. It isn't obvious why Cartridges are so much more memory efficient than standard KV caches.

Relationship to new architectures. Offline training and self-study are related to new architectures that incorporate gradient descent-like memory updates (*e.g.* Test-time training, Titans, Atlas, DeltaNet).

Applications. Cartridges for *agentic lifelong learning, coding assistants, or personalization!*

CARTRIDGES

Lightweight and general-purpose language model memories via self-study

tl;dr *Store huge contexts in tiny KV caches with offline* ***self-study****!*

Sabri Eyuboglu*

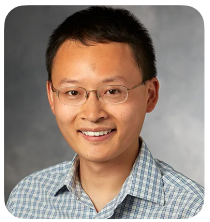
Ryan Ehrlich*

Simran Arora*

Neel Guha

Dylan Zinsley

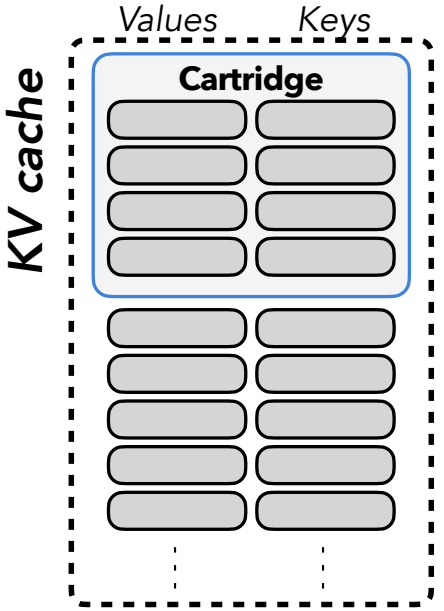
Emily Liu

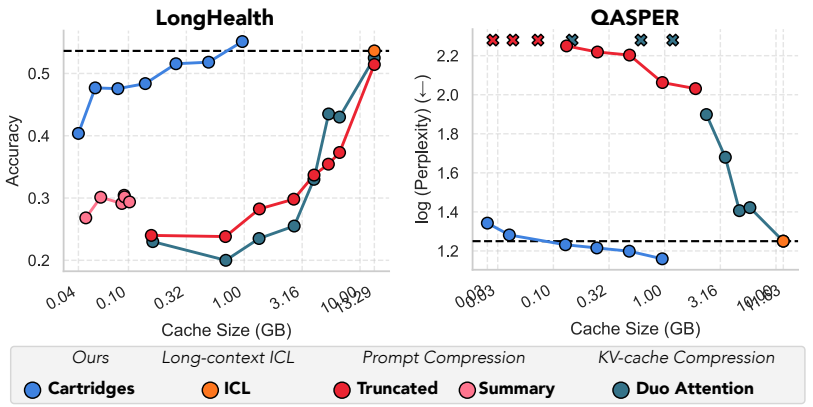


James Zou

Azalia Mirhoseini

Chris Ré





CARTRIDGES

Lightweight and general-purpose language model memories via self-study

Ask me about…

- Why parameterize Cartridges with *prefix tuning* instead of ***LoRA***?
- Can Cartridges be **composed**?
- How do Cartridges relate to architectures with fixed state-size (*e.g.* ***linear attention*** *and* ***state-space models****)*?
- How do you pick ***seed prompts*** in self-study*?*
- How do you ***initialize*** *Cartridges?*

Architectural changes

Introduce Sparsity

Longformer [Beltagy et al. 2020]
BigBird [Ainslie et al. 2020]
Sparse Transformers [Child et al. 2019]
…and many others.

Reduce # of key-value heads

Multi-query Attention [Shazeer 2019]
Grouped-query Attention [Ainslie et al. 2023]

Use low rank key-value heads

Multi-head latent attention [Liu 2024]

Linearize attention

[Katharopoulos et al. 2020]
Performer [Choromanski et al. 2022]
Based [Arora, Eyuboglu et al. 2024]
…and many others.

+ modified update rules

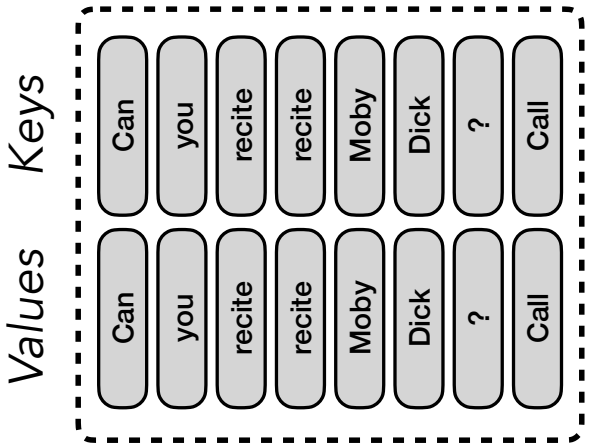
Mamba [Gu and Dao 2024]
DeltaNet [Yang et al. 2024]
Titans [Behrouz et al. 2024]
…and many others.

The *advantage* of Linear Attention

Attention

$$\text{softmax}(\mathbf{QK}^\top)\mathbf{V}$$

KV-cache

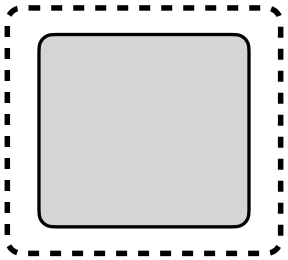


$O(Nd)$

Linear Attention

$$\mathbf{Q}(\mathbf{K}^\top\mathbf{V})$$

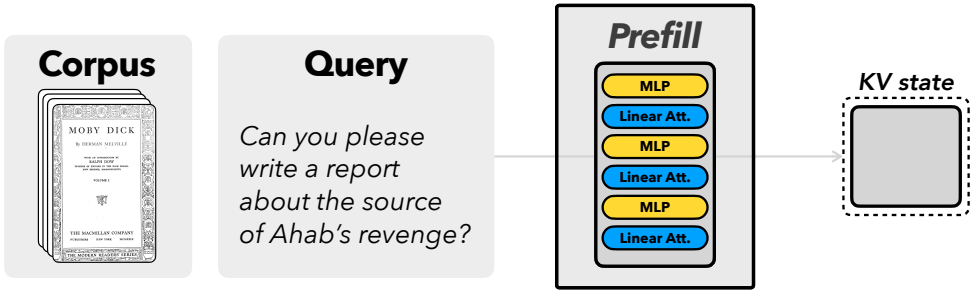
KV-state

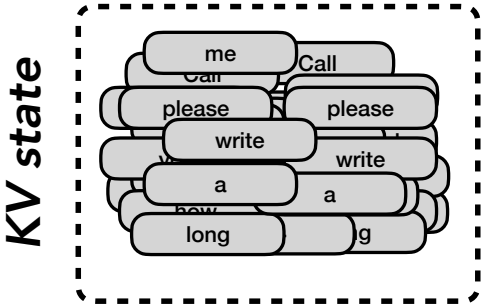


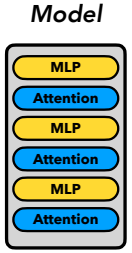
$O(d^2)$

The *advantage* of Linear Attention

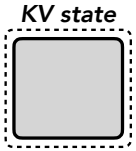
Linear attention consumes constant-sized space!







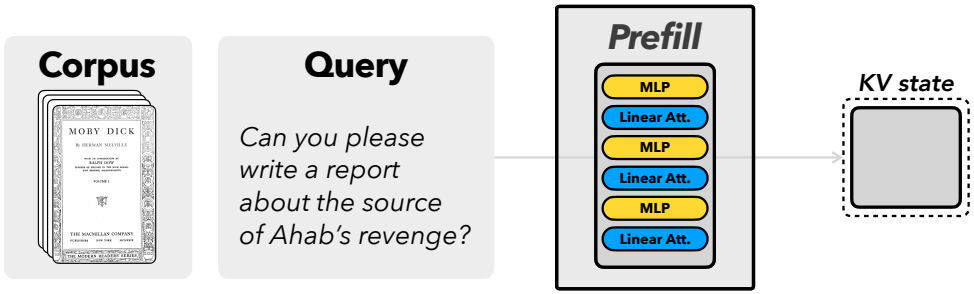
Llama 70B parameters are **~140 GB**



KV state for 128k tokens of context is whatever we want it to be!

The *advantage* of Linear Attention

Linear attention consumes constant-sized space!



Linearize attention

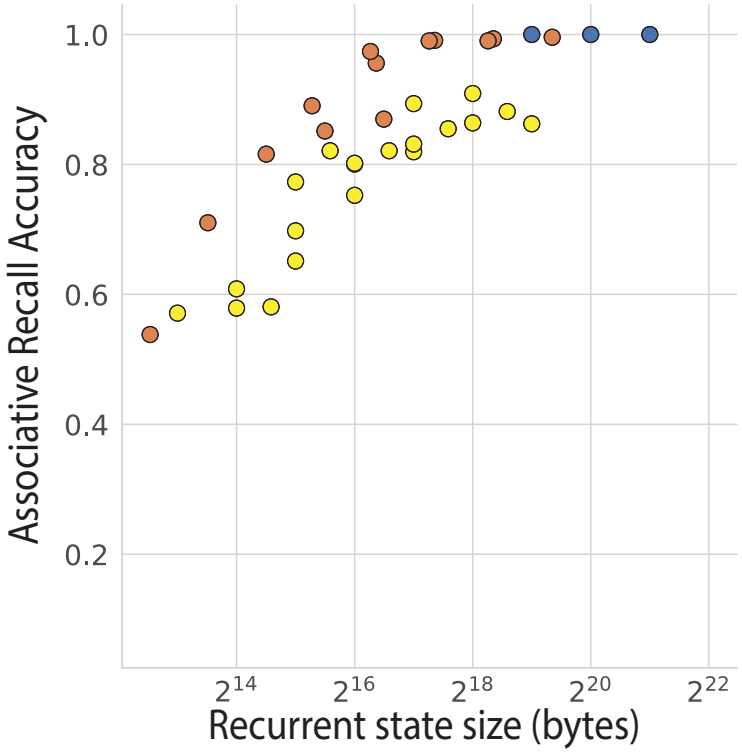
[Katharopoulos et al. 2020]
Performer [Choromanski et al. 2022]
Based [Arora, Eyuboglu et al. 2024]
…and many others.

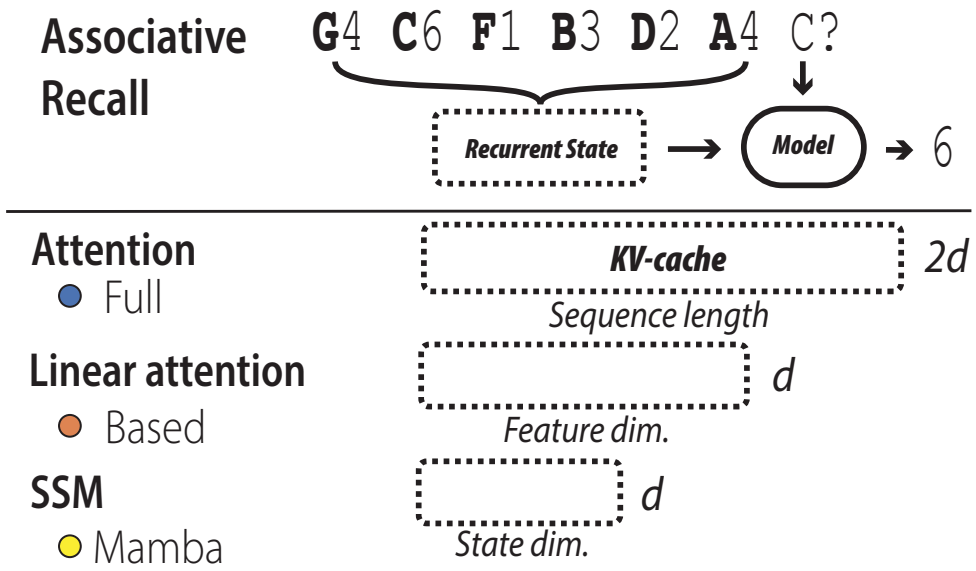
+ modified update rules

Mamba [Gu and Dao 2024]
DeltaNet [Yang et al. 2024]
Titans [Behrouz et al. 2024]
…and many others.

No free lunch!

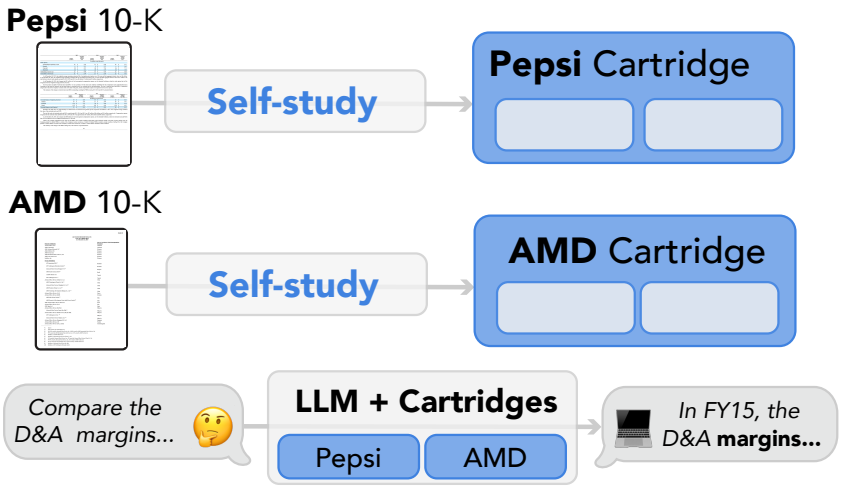
We study the **tradeoff** between **space** and **recall**



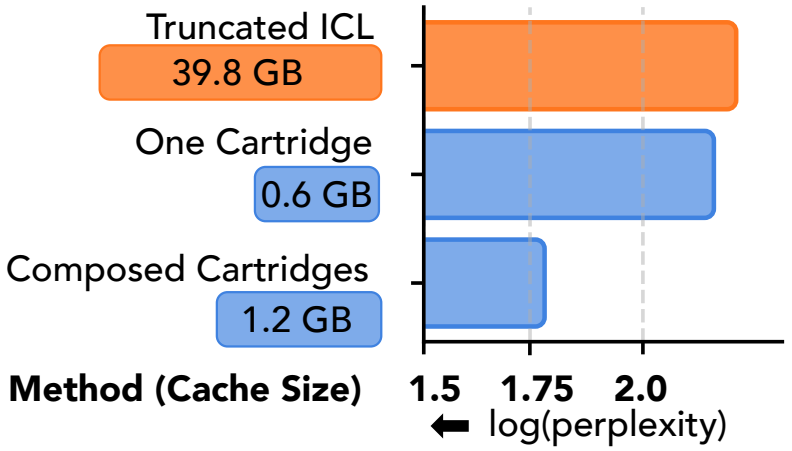


Composition

Cartridge Composition



Multi-doc Question Answering



Who audited the Boeing and AMD statements, respectively?

The audit of the consolidated financial statements of AMD was performed by Ernst & Young LLP, while the audit of the consolidated financial statements of Boeing was performed by Deloitte & Touche LLP.

List a few competitors for each of PepsiCo and AMD as stated in each10K."

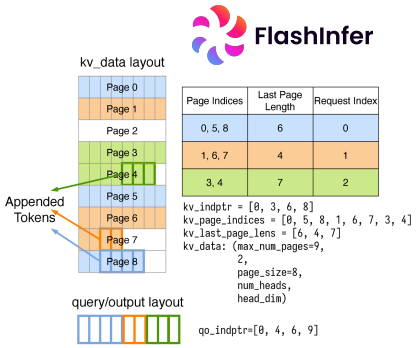
Here are some competitors for PepsiCo and AMD:
** Unilever (as a competitor) ...*
** Red Bull (as a competitor in the energy drink market)*
AMD:
** Intel (as a competitor in the ...*

Ablation: Cartridge *parameterization*

Why parameterize Cartridges with prefix tuning?
What about LoRA?

Two reasons:

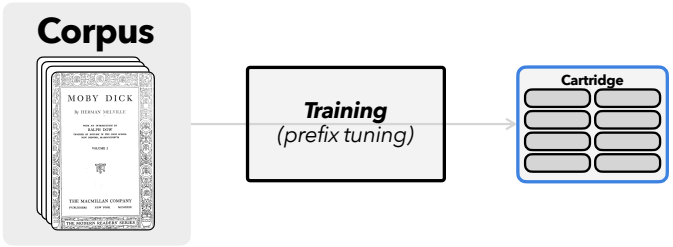
1 ***Easier to serve*** *– only need optimizations for per-user KV caches, which is already supported by inference engines. No multi-LoRA!*



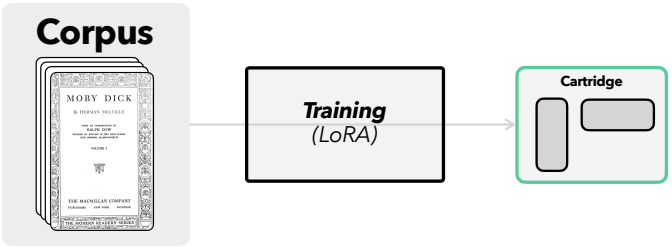
2 ***Improved quality -*** see next slide...

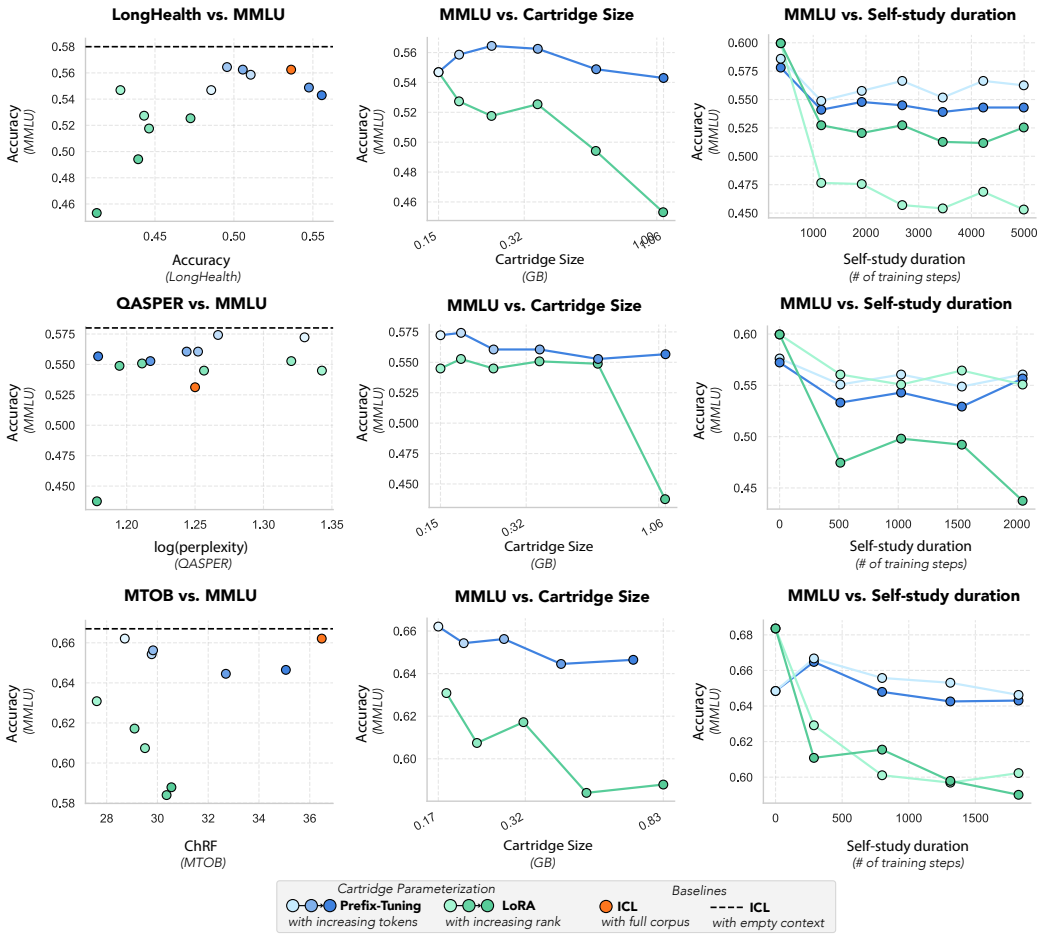
Ablation: *Cartridge Parameterization*

Prefix tuning

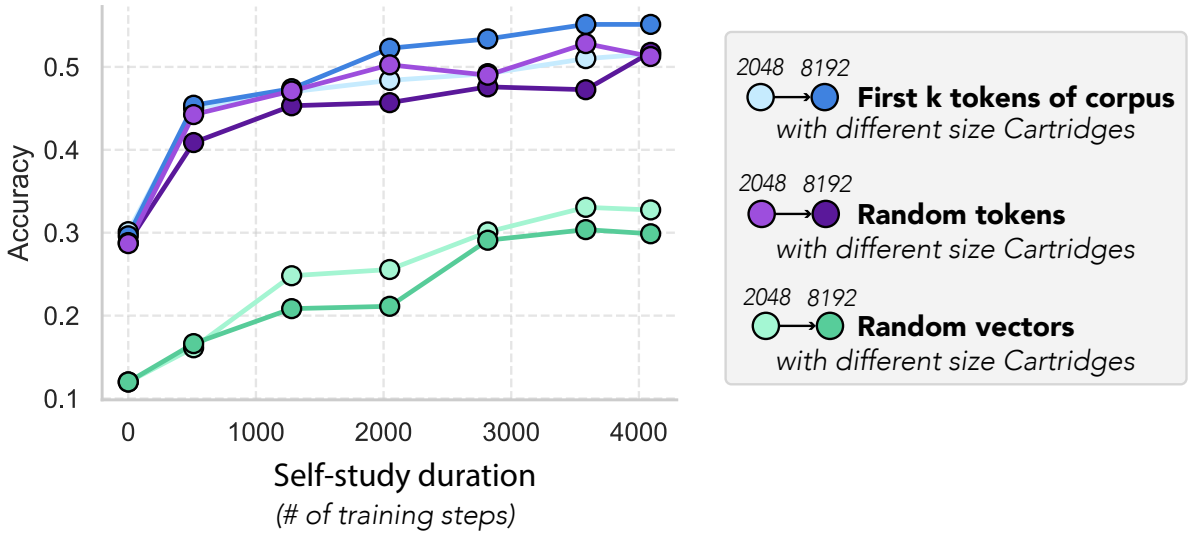


Low-rank Adaptation

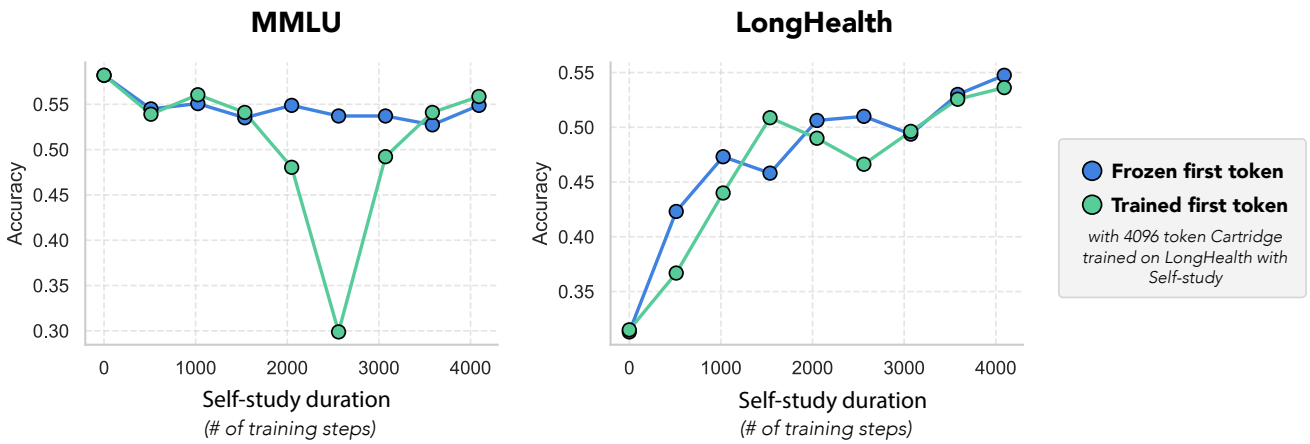




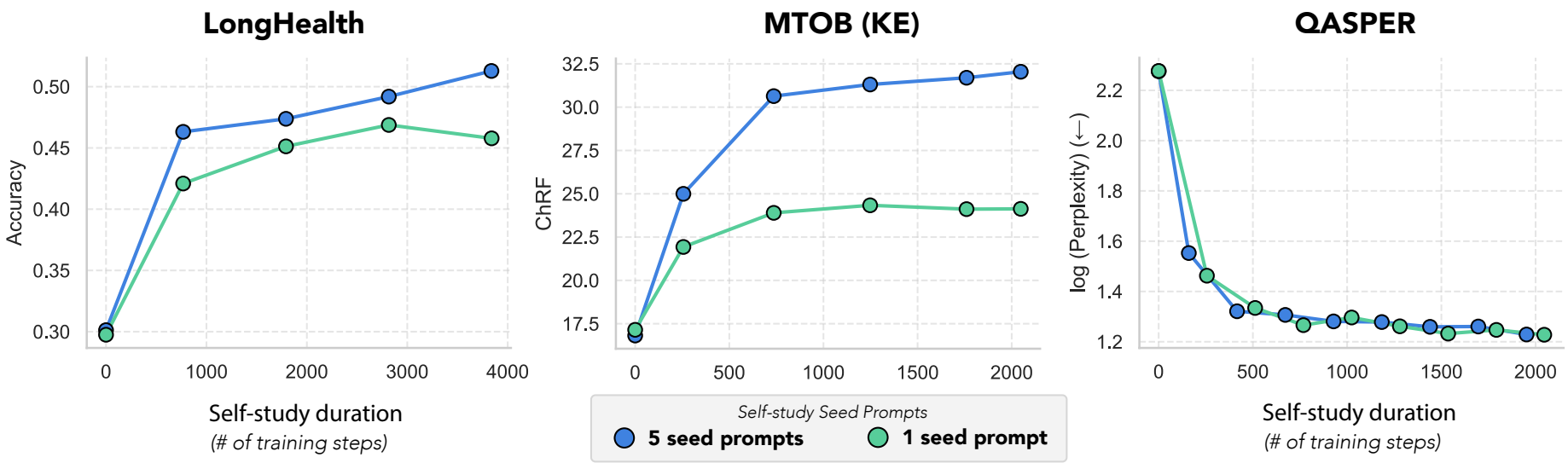
Ablation: Cartridge Initialization



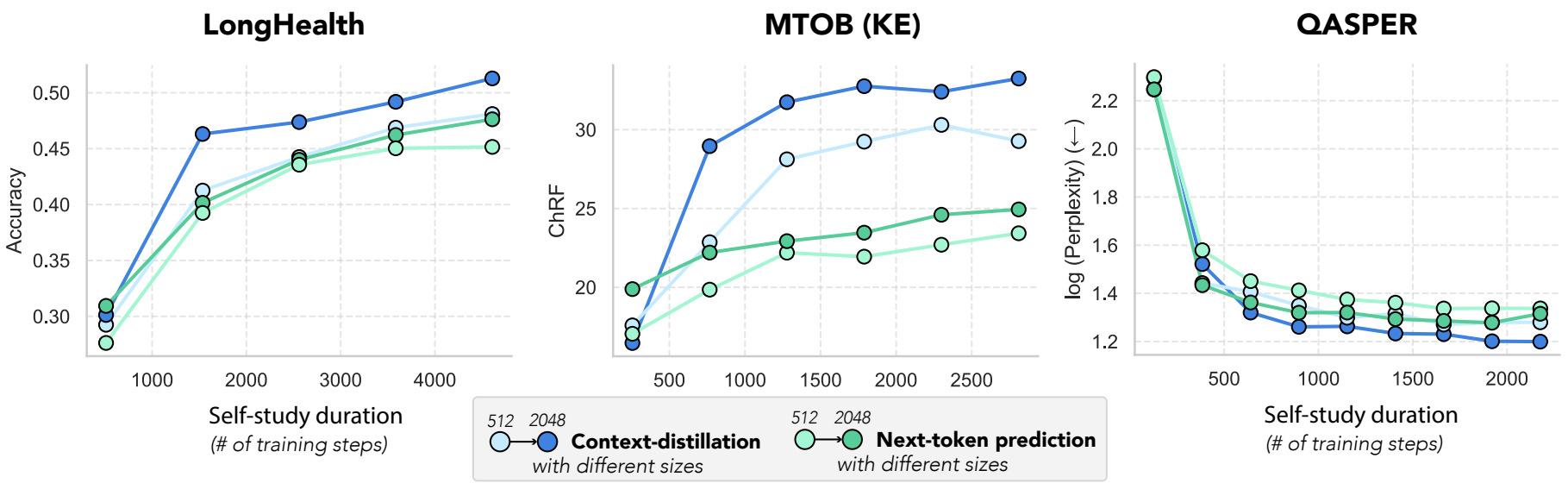
Ablation: Freezing attention sink



Ablation: Seed prompts

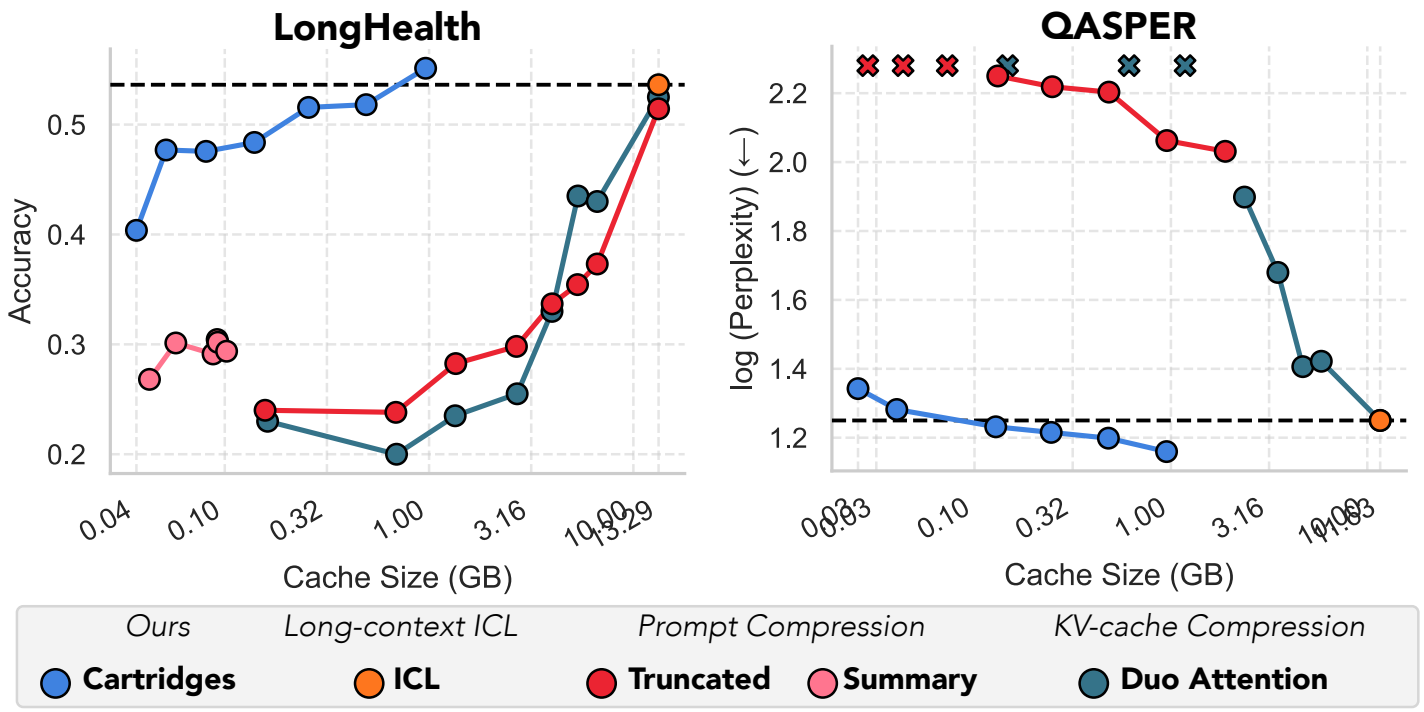


Ablation: Context distillation



Expanding the space-quality frontier

Other datasets



2 **Self-study:** Context distillation

Context Distillation

$$\arg\min_{\mathbf{Z}} \sum_{(\mathbf{x},\tilde{\mathbf{c}})\in\mathbf{C}} D_{\mathrm{KL}}(p(\cdot\,|\,\mathbf{Z})\,||\,p(\cdot\,|\,\tilde{\mathbf{c}})) + D_{\mathrm{KL}}(p(\cdot\,|\,x_1,\mathbf{Z})\,||\,p(\cdot\,|\,x_1,\tilde{\mathbf{c}})) + \ldots$$

where

$\mathbf{Z}$ *is the Cartridge*

$p(\cdot\,|\,\mathbf{Z})$ *is the model's next token distribution conditioned on the Cartridge*

$p(\cdot\,|\,\tilde{\mathbf{c}})$ *is the model's next token distribution conditioned on text*

S *is the synthetic dataset generated in Step 1*

$\mathbf{x}$ *a single synthetic conversation generated in Step 1*

$\tilde{\mathbf{c}}$ *is the text chunk on which the conversation was conditioned in Step 1*

Minimize the KL divergence between the next token distributions $p(\cdot\,|\,\mathbf{Z})$ (model conditioned on the Cartridge) and $p(\cdot\,|\,\tilde{\mathbf{c}})$ (model conditioned on the text chunk)!

This is related to the standard model distillation objective [Hinton *et al.* 2015, Kim and Rush 2016]. Typically, this is applied between a different models (*i.e.* a teacher and a student).

More recently, this idea of distilling *context* into parameters has been proposed. [Snell *et al.* 2022, Kujanpaa *et al.* 2024]