Long-context Reasoning

Aaron Mueller
CS599 B1: Advanced Natural Language Processing
Oct. 15, 2026

Administrivia

- **HW2** is due tonight at 11:59pm!

- Quiz 2 will be next Tuesday, Oct. 20
  - *Not* cumulative
  - Very similar format to quiz 1
  - Topics: reinforcement learning; GRPO; Dr. GRPO; PPO; DPO; RLHF; RLVR; instruction fine-tuning; common post-training pipelines (e.g., Olmo 3's recipe); tools and tool calling; harnesses; reward hacking/misspecification
    - Will only give questions that require deeper mathematical reasoning/computational work for things you've seen on the homework, like (Dr.) GRPO/RLVR/REINFORCE

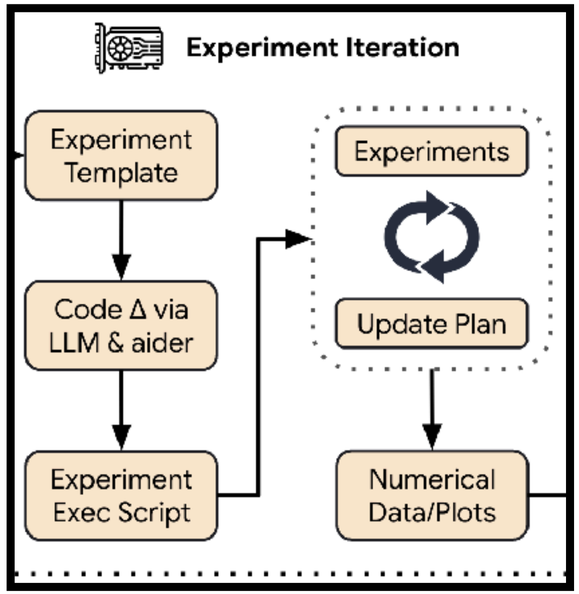
Reminder: Agents

We’ve now seen examples of a few types of agents:

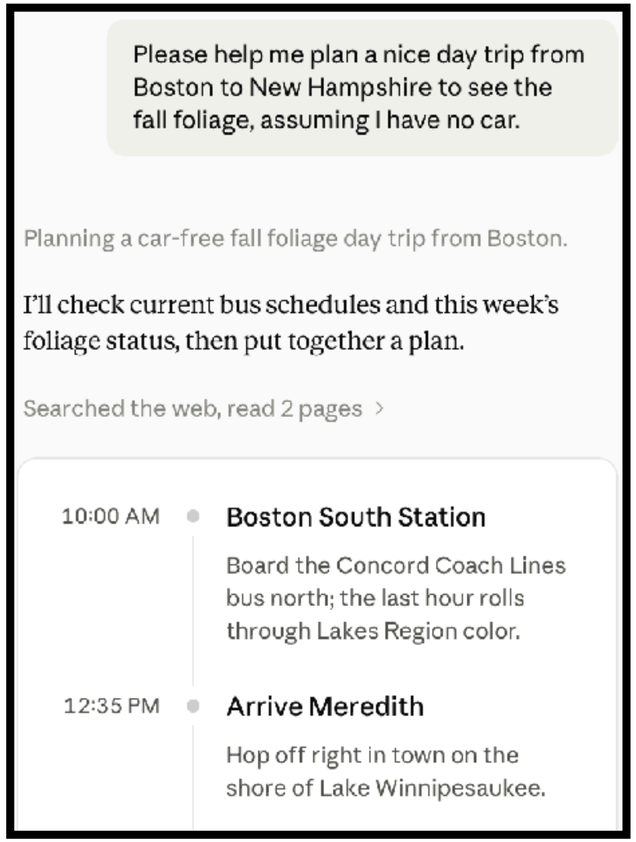
Coding



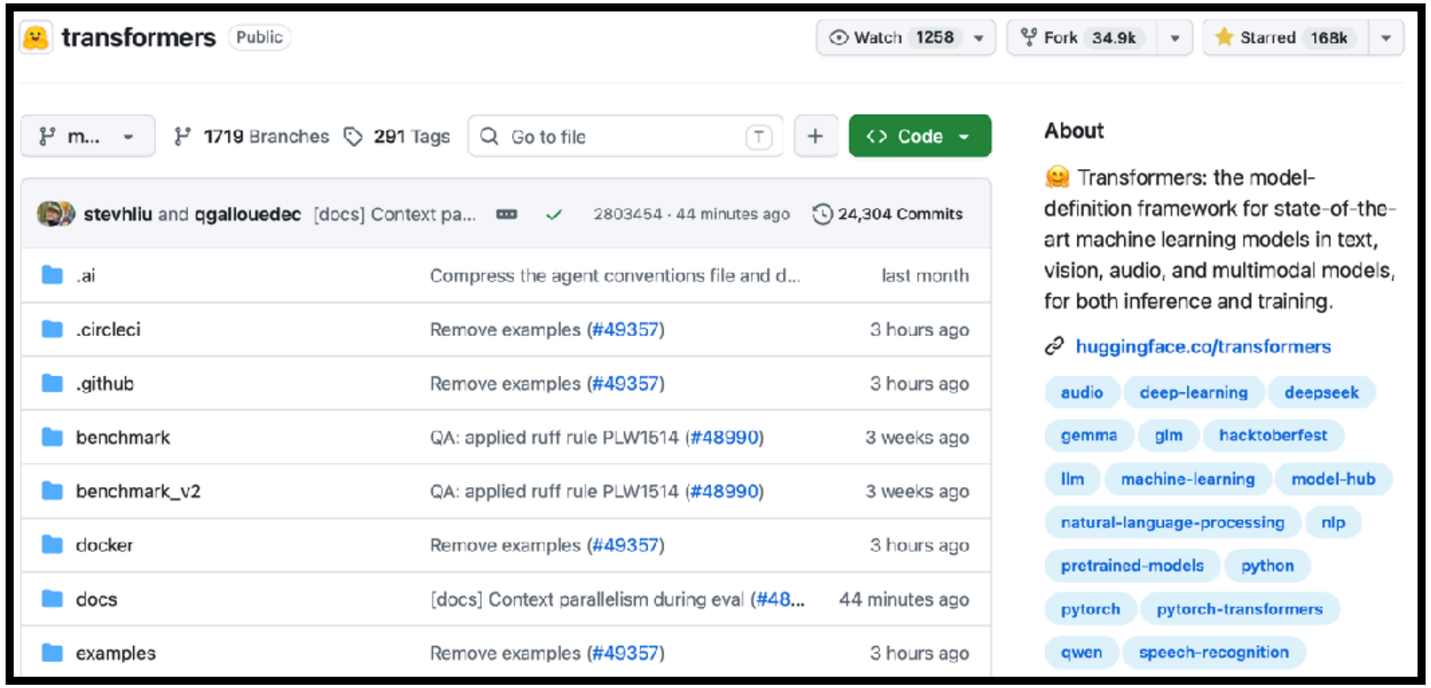
Scientific



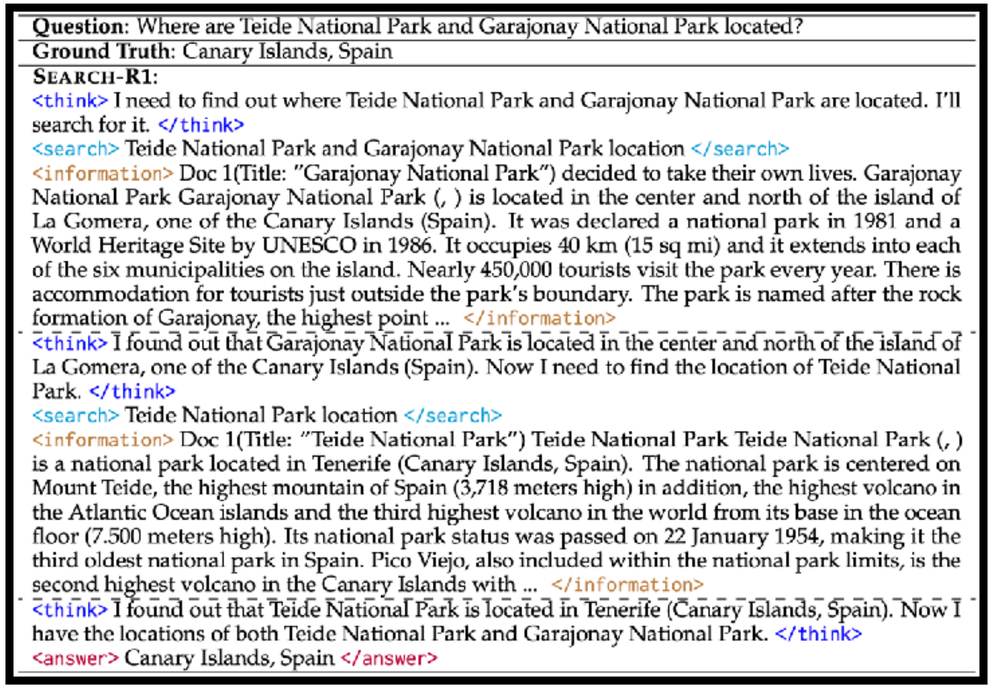
Conversational



Long Contexts



Huggingface Transformers repo: ≈5 – 15M tokens



Question: Where are Teide National Park and Garajonay National Park located?
Ground Truth: Canary Islands, Spain

SEARCH-R1:

<think> I need to find out where Teide National Park and Garajonay National Park are located. I'll search for it. </think>
<search> Teide National Park and Garajonay National Park location </search>
<information> Doc 1(Title: "Garajonay National Park") decided to take their own lives. Garajonay National Park Garajonay National Park (,) is located in the center and north of the island of La Gomera, one of the Canary Islands (Spain). It was declared a national park in 1981 and a World Heritage Site by UNESCO in 1986. It occupies 40 km (15 sq mi) and it extends into each of the six municipalities on the island. Nearly 450,000 tourists visit the park every year. There is accommodation for tourists just outside the park's boundary. The park is named after the rock formation of Garajonay, the highest point ... </information>
<think> I found out that Garajonay National Park is located in the center and north of the island of La Gomera, one of the Canary Islands (Spain). Now I need to find the location of Teide National Park. </think>
<search> Teide National Park location </search>
<information> Doc 1(Title: "Teide National Park") Teide National Park Teide National Park (,) is a national park located in Tenerife (Canary Islands, Spain). The national park is centered on Mount Teide, the highest mountain of Spain (3,718 meters high) in addition, the highest volcano in the Atlantic Ocean islands and the third highest volcano in the world from its base in the ocean floor (7.500 meters high). Its national park status was passed on 22 January 1954, making it the third oldest national park in Spain. Pico Viejo, also included within the national park limits, is the second highest volcano in the Canary Islands with ... </information>
<think> I found out that Teide National Park is located in Tenerife (Canary Islands, Spain). Now I have the locations of both Teide National Park and Garajonay National Park. </think>
<answer> Canary Islands, Spain </answer>

100 web pages: ≈100K – 1M tokens

Effective Context

- Just because a model has a long context window does not mean it can effectively leverage all of that context

- **Effective context:** how much context the model can reliably use for a particular task

- Long inputs and long outputs impose different demands:
  - Finding evidence in a document
  - Maintaining correctness and coherence throughout long-form generation

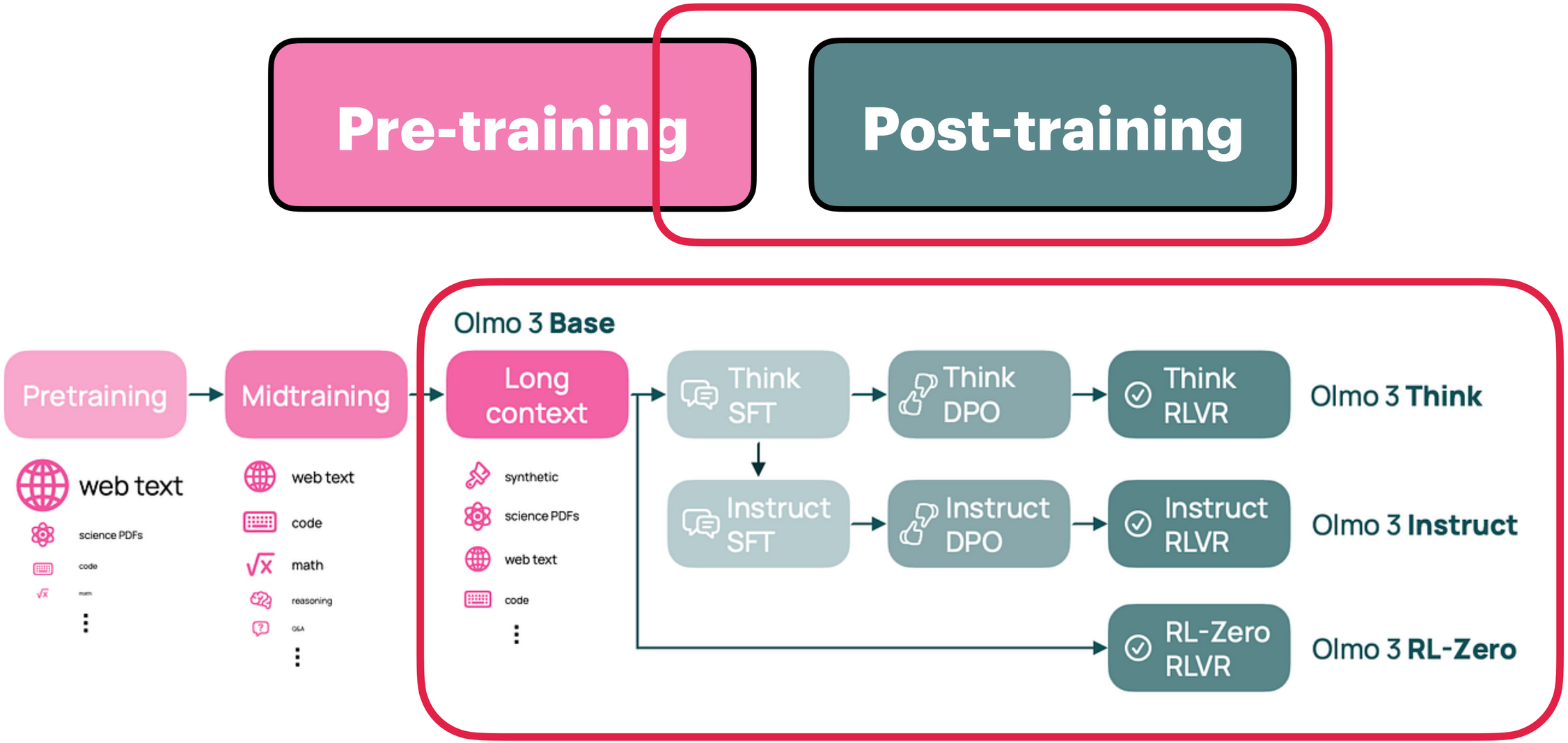
Getting LMs to Handle Long Contexts

1. (Post-)Training data and methods for long contexts

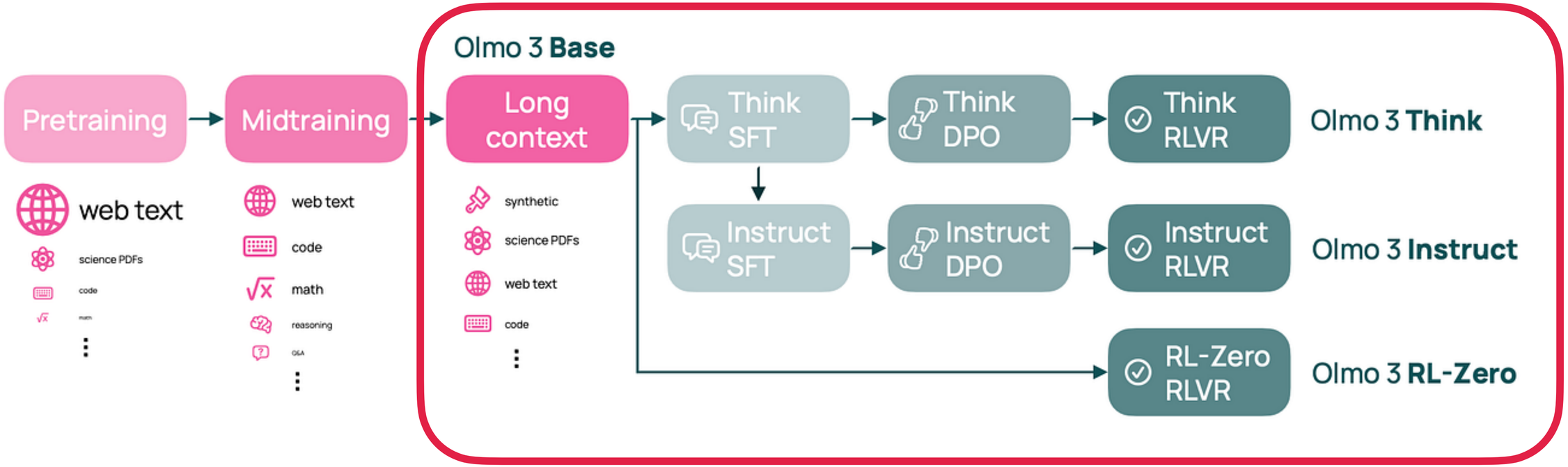
2. Model architectures for long contexts

3. Evaluation

When to train for long contexts?



When to train for long contexts?



- Train on short contexts first
- Expand to longer contexts later during pre-training
- Post-train the model to handle problems requiring reasoning over long contexts

Why? Harder to learn; it helps to have a good initialization first

Long-context Training

- The usual long-context training recipe is something like this:
    - Perform continued pre-training on long-context data
        - Mix of real and synthetic data
    - Adjust the positional embeddings to handle longer contexts
    - Perform post-training
        - Allow models to generate lots of thinking tokens during RLVR

Long-context Training Data

Data	#Long tokens
Code Repos	**98.8B**
SP/Books	**33.2B**
SP/CC	15.3B
SP/Arxiv	5.2B
SP/GitHub	2.8B
SP/Wiki	0.1B
SP/StackEx	<0.1B
SP/C4	<0.1B

Source	Length bucket	600B Pool		50B Mix	
		Tokens	Docs	Tokens	Docs
olmOCR PDFs	8K-16K	144B (22.5%)	12.7M	2.27B (4.55%)	235K
olmOCR PDFs	16K-32K	115B (18.0%)	5.06M	1.85B (3.70%)	110K
olmOCR PDFs	32K-64K	106B (16.6%)	2.30M	4.81B (9.63%)	177K
olmOCR PDFs	64K-128K	96.0B (15.0%)	1.05M	–	–
olmOCR PDFs	128K-256K	60.8B (9.5%)	342K	–	–
olmOCR PDFs	256K-512K	35.1B (5.49%)	97.1K	–	–
olmOCR PDFs	512K-1M	21.5B (3.36%)	30.2K	–	–
olmOCR PDFs	1M+	26.9B (4.21%)	12.2K	–	–
olmOCR PDFs + synth **CWE**	32K-64K	8.77B (1.37%)	189K	1.94B (3.88%)	71.3K
olmOCR PDFs + synth **REX**	32K-64K	24.1B (3.77%)	492K	6.08B (12.2%)	217K
Midtraining data mix	Variable	–	–	33.0B (66.1%)	79.2M
Total		**639B**	**22.3M**	**50.0B (100%)**	**80.0M**

Usually consists mostly of scientific article PDFs, code, and/or some synthetic data.

Olmo 3 uses a mix of 66% shorter-context data, 33% long-context data during the long-context training part of the pipeline.

Synthetic Long-context Training Data

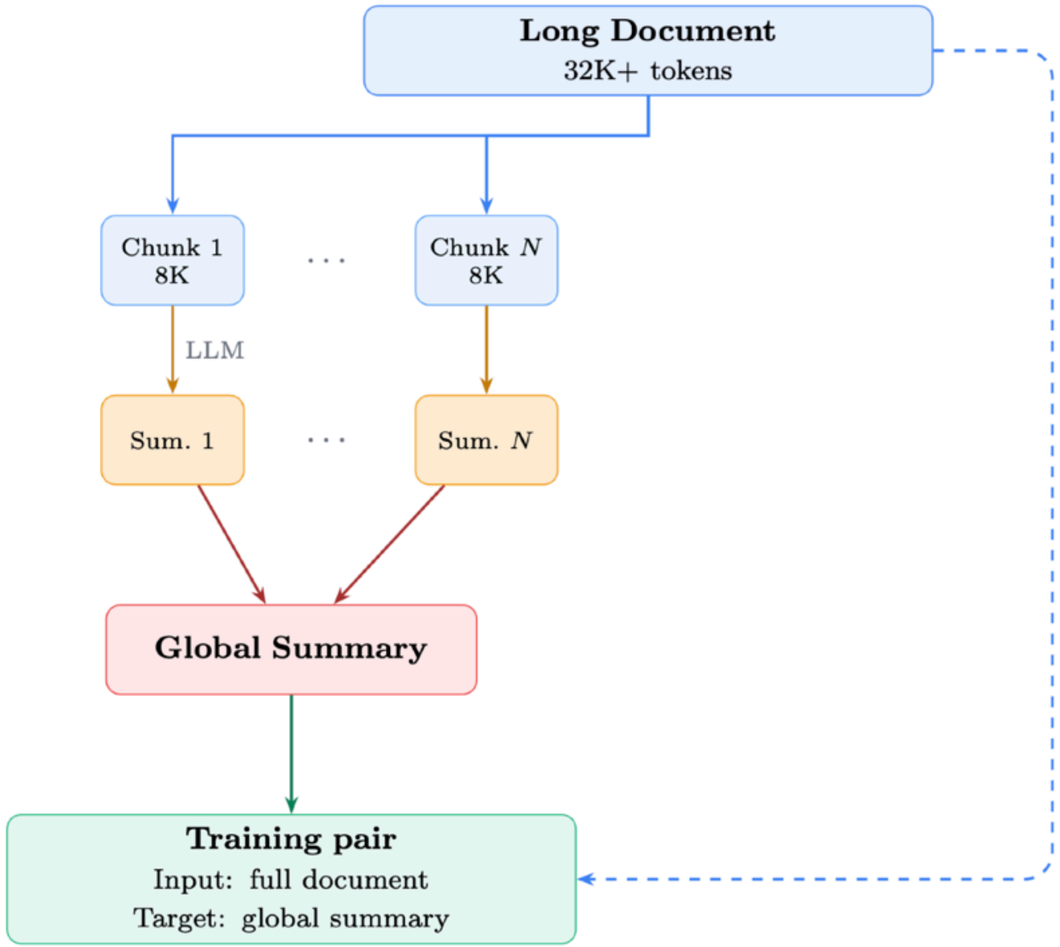
- Synthetic data is usually in an *instruction* style, rather than narrative (as in natural text)
- Tasks include fill-in-the-middle, summarization, question answering, among others

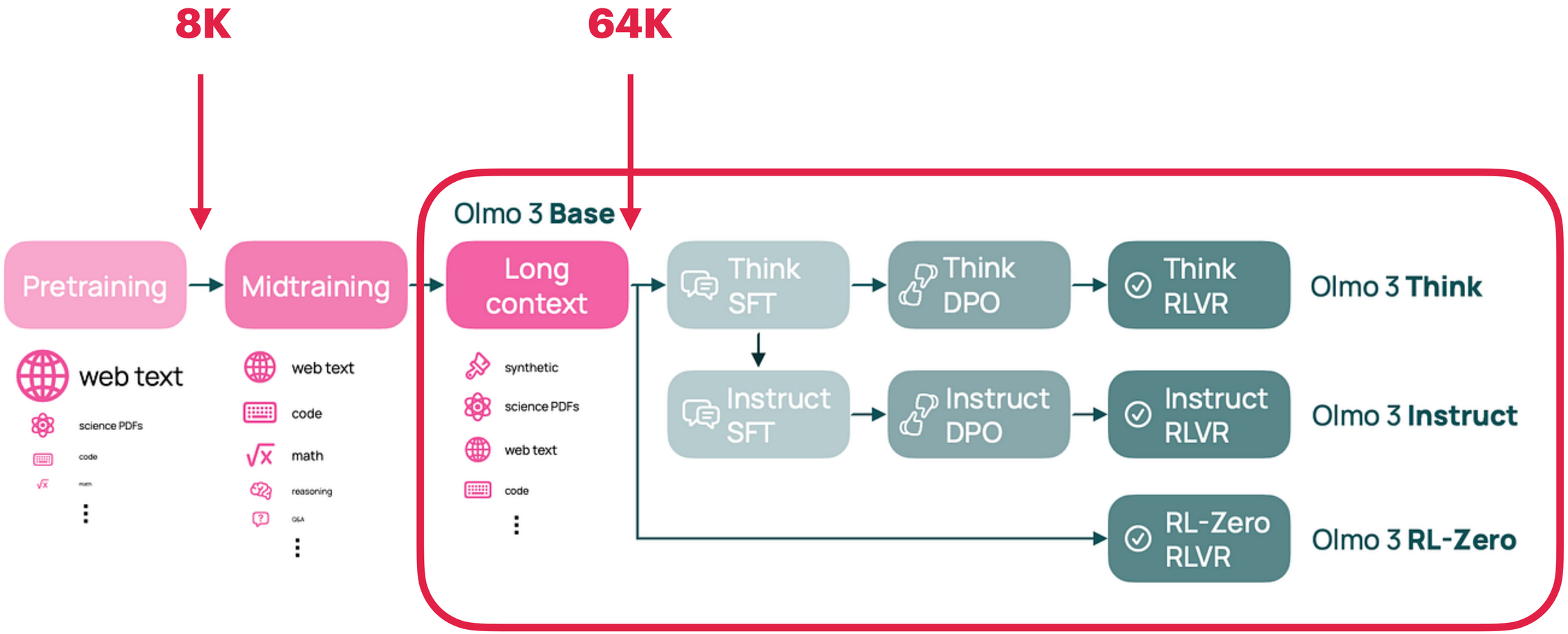
Fill-in-the-middle

```
<context>
<missing part>
<context>

Q: Fill in the missing part
```

Summarization

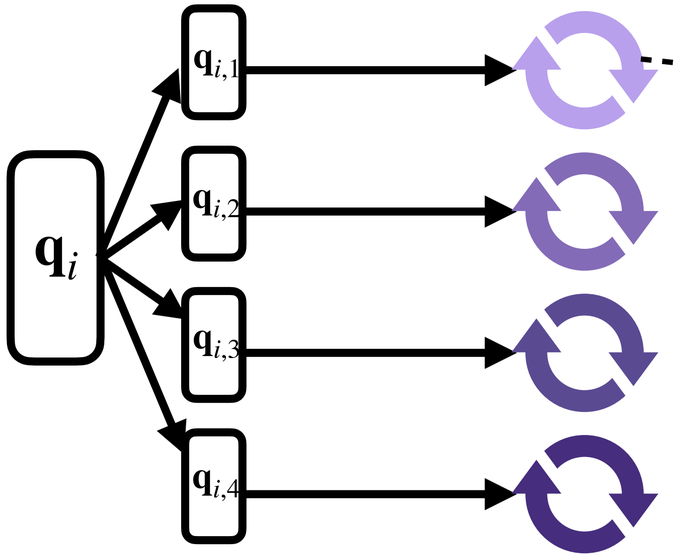




This gets us to a much longer context window size. But we're not finished yet!

To go even further, we'll need to change parts of the model itself.

Review: Rotary Position Embedding (RoPE)



$$\theta_{i,k} = \frac{i}{\Theta^{(2k-2)/d}}$$

$$R_{i,k} = \begin{bmatrix} \cos(\theta_{i,k}) & -\sin(\theta_{i,k}) \\ \sin(\theta_{i,k}) & \cos(\theta_{i,k}) \end{bmatrix}$$

Rotate each $\mathbf{q}_{i,k}$ by $\theta_{i,k}$, where $\theta_{i,k}$ depends on the vector index and the dimensionality of the embeddings d

After training with RoPE, we can later adjust the hyperparameters to support long contexts.

Yet another RoPE extensioN method (YaRN)

- What if we'd like to extend the size of the context window after training?
- If RoPE is trained with encodings up to context window length L, you can expand to L' via:

$$\theta_{i,k} = \frac{iL/L'}{\Theta^{(2k-2)/d}}$$

- At index k, we can define a notion of "wavelength", which is how many tokens are needed to do a full rotation:

$$\lambda_k = \frac{2\pi}{\theta_k} = 2\pi\Theta^{\frac{2k}{|D|}}$$

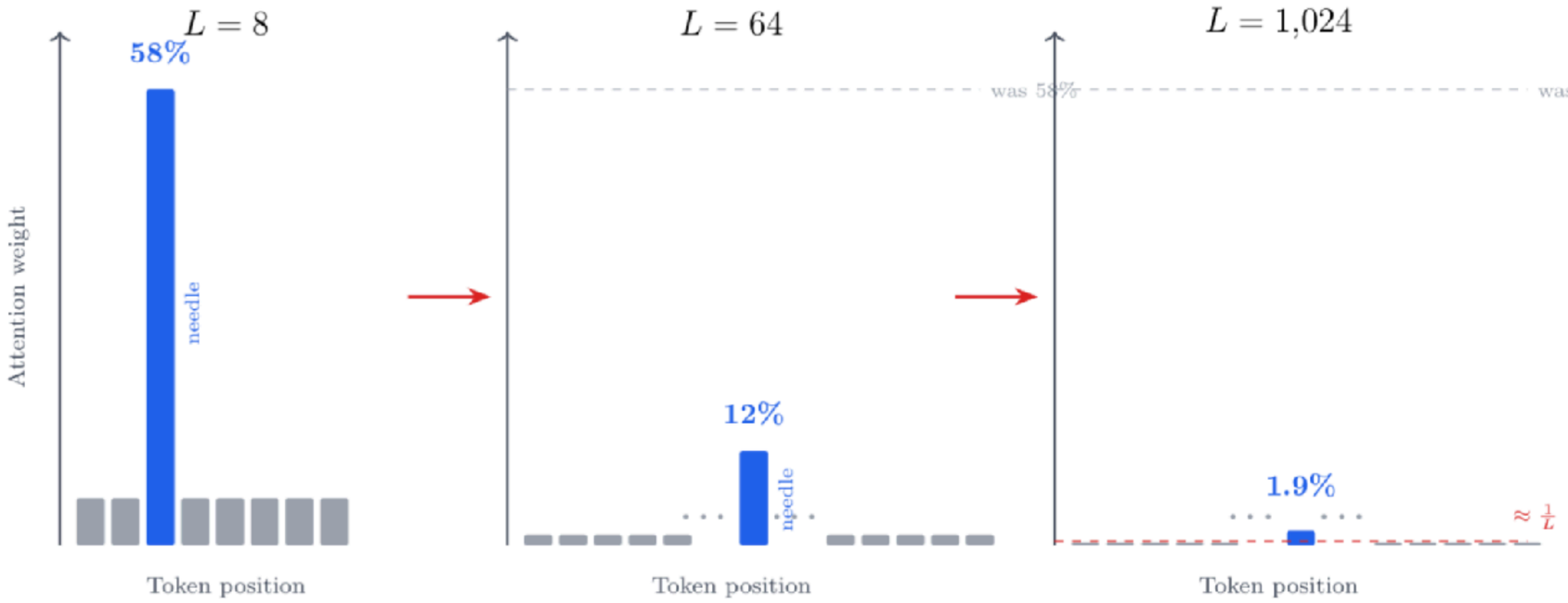
- When λ_k is small, don't use position interpolation; if large, use it
- We can also introduce temperature t on attention to further rescale it

YaRN

- We can also introduce temperature t on attention to further rescale it. *Why?*
- **Problem:** longer sequences mean more attention keys, which means attention weights will be more diffuse

- **Fix:** scale attention scores by temperature $\sqrt{1/t}$

$$\alpha = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}\cdot\sqrt{t}}\right)V$$



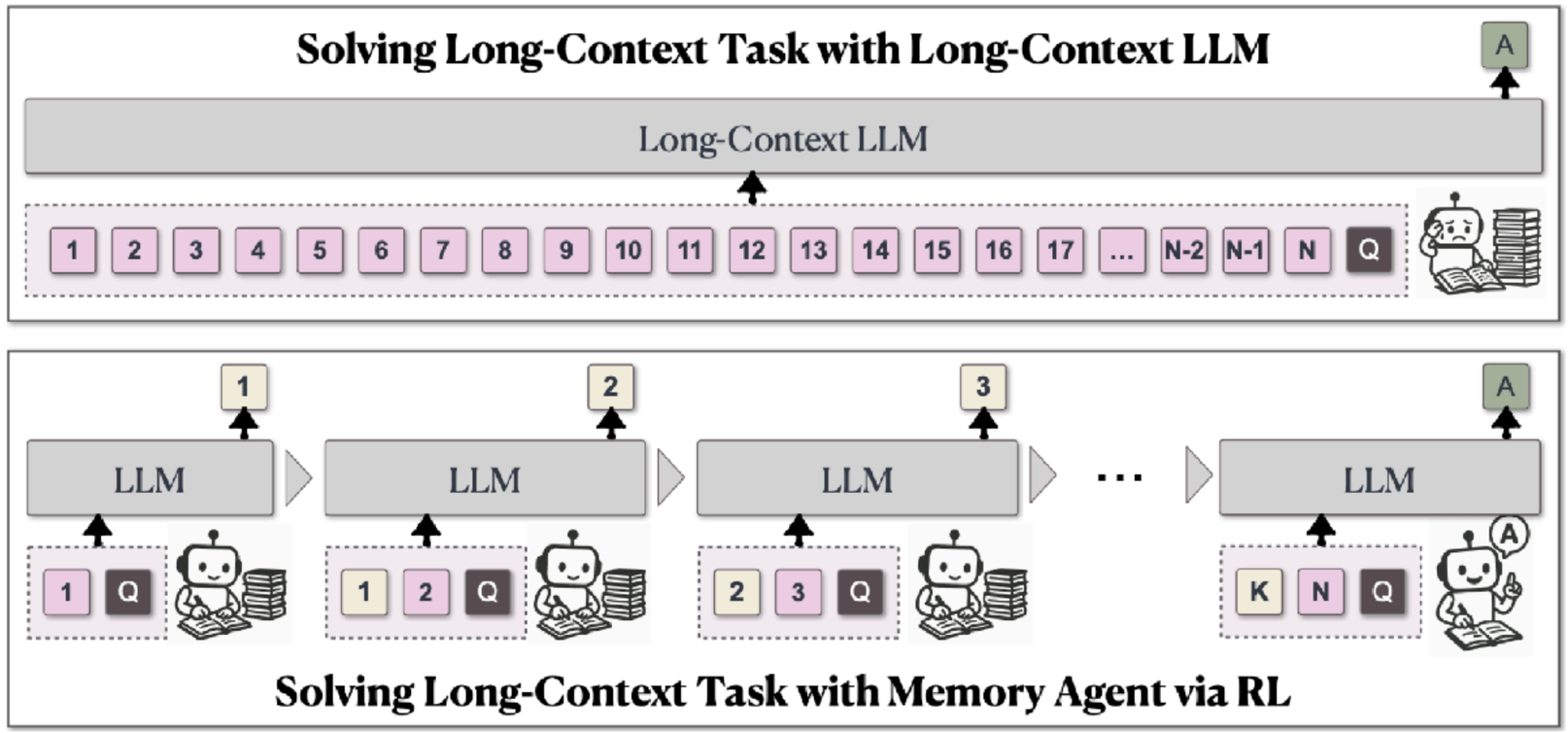
Long-context *Post*-training Data

- Agentic RL post-training naturally gives LMs opportunities to learn how to process longer contexts more effectively
- *Extremely* compute-intensive
- Data is hard to collect.
  Most still use synthetic data

Please read the following text.
Document 0:
…
Document 3:
Who's Who? is a studio album by American jazz musician John Scofield. It features two different bands, one acoustic and one electric. The acoustic group, featuring Scofield's then-employer Dave Liebman on saxophones, Eddie G\u00f3mez on bass, and Billy Hart on drums, recorded “The Beatles” and “How the West Was Won”. …
{“bdd640fb-0667-4ad1-9c80-317fa3b1799d”: “23b8c1e9-3924-46de-beb1-3b9046685257”}.
…
Document 10:
…
The university is one of the smallest of the 23 CSU campuses in California. Sonoma State offers 92 Bachelor's degrees, 19 Master's degrees, one Doctoral degree (Doctor of Education), and 11 teaching credentials. {“972a8469-1641-4f82-8b9d-2434e465e150”: “Musician and satirist Allie Goertz wrote a song about the “The Simpsons” character Milhouse, who Matt Groening named after who?”}.
…
Document 47:
Neil Affleck
{“bd9c66b3-ad3c-4d6d-9a3d-1fa7bc8960a9”: “972a8469-1641-4f82-8b9d-2434e465e150”}.
Neil Affleck (born 1953) is a Canadian animator, director, and former actor. He has worked as an animator on “The Simpsons” and “Family Guy”, and as an actor appeared in a leading role in the 1981 film “My Bloody Valentine”. {“9a1de644-815e-46d1-bb8f-aa1837f8a88b”: “b74d0fb1-32e7-4629-8fad-c1a606cb0fb3”}.
…
In the context above, there is one correct question to answer. The correct question can only be found by following the correct consecutive chain of key:value pairs encoded with UUID strings (e.g., f81d4fae-7dec-11d0-a765-00a0c91e6bf6), starting from "bdd640fb-0667-4ad1-9c80-317fa3b1799d".
Find the correct question first, then answer it.

Long-context *Post*-training Setups

- How can we design agentic systems that allow LMs to interact with context?
- One idea: **memory systems**



[Yu et al., 2026]

Q: The director of the romantic comedy ‘Big Stone Gap’ is based in what New York city?

CHUNK1

...

Document 20: Ghost (production team)

Ghost is a Swedish record producing and songwriting team, composed of Ulf Lindström and Johan Ekhé, based in *New York City*.

...

MEMORY1

The provided section does not contain any information about the director of the romantic comedy ”Big Stone Gap” or their base in New York City. **However, it does mention a production team called ”Ghost” which is based in New York City.** Since the problem is specifically asking about the director of ”Big Stone Gap,” and there is no direct information about the director in the given section, we cannot infer anything directly related to the problem from the provided text.

Updated memory:

- Ghost is a Swedish record producing and songwriting team, composed of Ulf Lindström and Johan Ekhé, based in New York City. They are perhaps best known for writing and producing Swedish singer Robyn’s three first studio albums.

The problem remains unanswered based on the provided section. Additional information would be needed to determine the director of ”Big Stone Gap” and their base in New York City.

CHUNK2
...(No relevant document)

MEMORY2
The provided section does not contain any information about the director of the romantic comedy "Big Stone Gap" or their base in New York City. 2 Therefore, the updated memory remains unchanged with respect to the problem at hand. The memory still notes that the Ghost record producing and songwriting team is based in New York City, but it does not provide any information about the director of "Big Stone Gap."
Updated memory:

- Ghost is a Swedish record producing and songwriting team, composed of Ulf Lindström and Johan Ekhé, based in New York City. They are perhaps best known for writing and producing Swedish singer Robyn's three first studio albums.
- The problem remains unanswered based on the provided section. Additional information would be needed to determine the director of "Big Stone Gap" and their base in New York City.

CHUNK3

...

Document 81: Big Stone Gap (film)

Big Stone Gap is a 2014 American drama romantic comedy film written and directed by **Adriana Trigiani** and produced by Donna Gigliotti for Altar Identity Studios, a subsidiary of Media Society.

...

Document 96: Adriana Trigiani

Adriana Trigiani is an Italian American best-selling author of sixteen books, television writer, film director, and entrepreneur based in **Greenwich Village, New York City**.

...

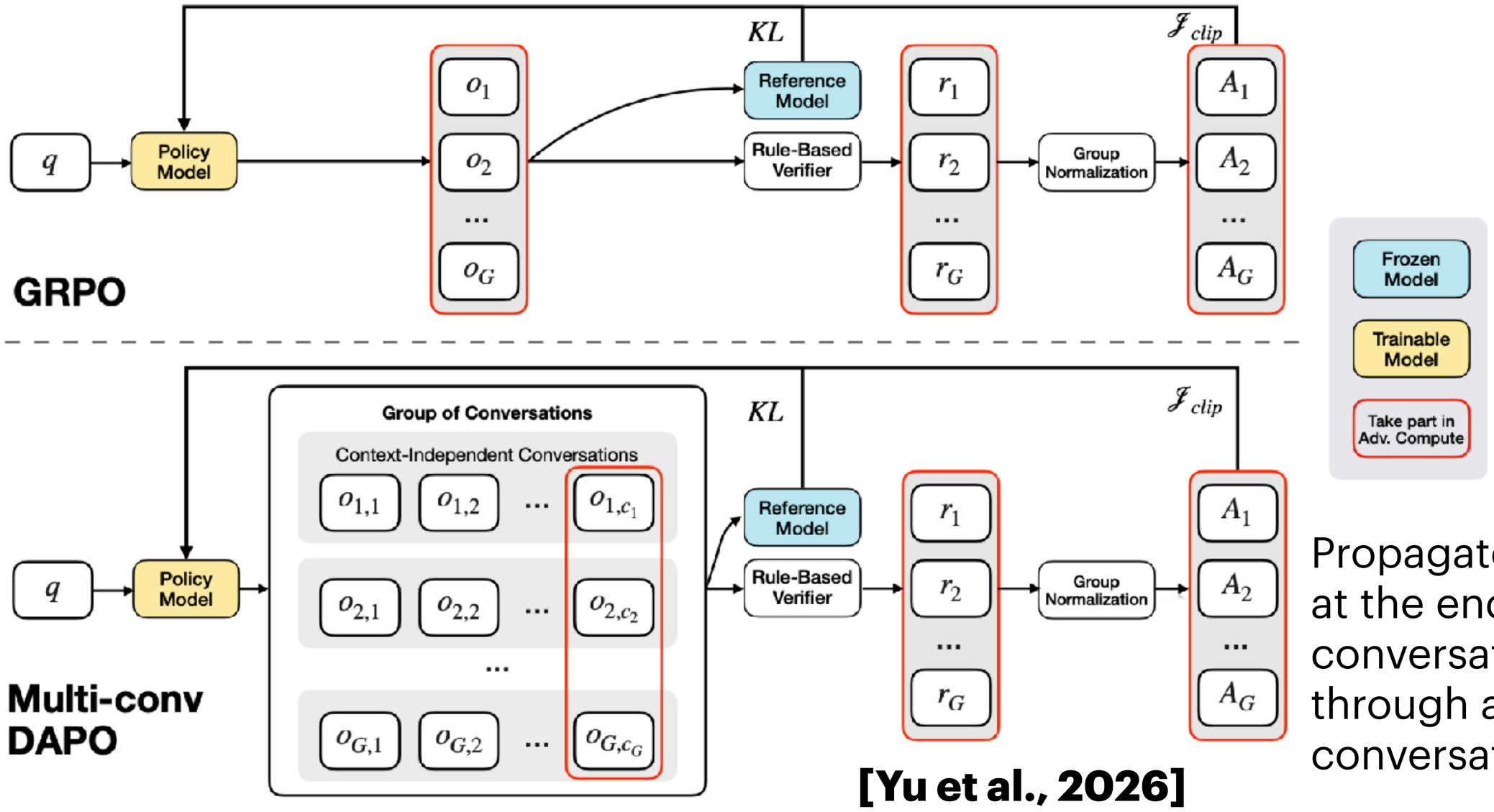
MEMORY3

Updated memory:

- Ghost is a Swedish record producing and songwriting team, composed of Ulf Lindström and Johan Ekhé, based in New York City. They are perhaps best known for writing and producing Swedish singer Robyn's three first studio albums.
- **Adriana Trigiani** is an Italian American best-selling author of sixteen books, television writer, film director, and entrepreneur based in Greenwich Village, New York City. Trigiani has published a novel a year since 2000.
- **Big Stone Gap** is a 2014 American drama romantic comedy film written and directed by Adriana Trigiani and produced by Donna Gigliotti for Altar Identity Studios, a subsidiary of Media Society. Based on Trigiani's 2000 best-selling novel of the same name, the story is set in the actual Virginia town of Big Stone Gap circa 1970s. The film had its world premiere at the Virginia Film Festival on November 6, 2014.

  The director of the romantic comedy "Big Stone Gap" is Adriana Trigiani, who is based in **Greenwich Village, New York City**.

Long-context *Post*-training Setups

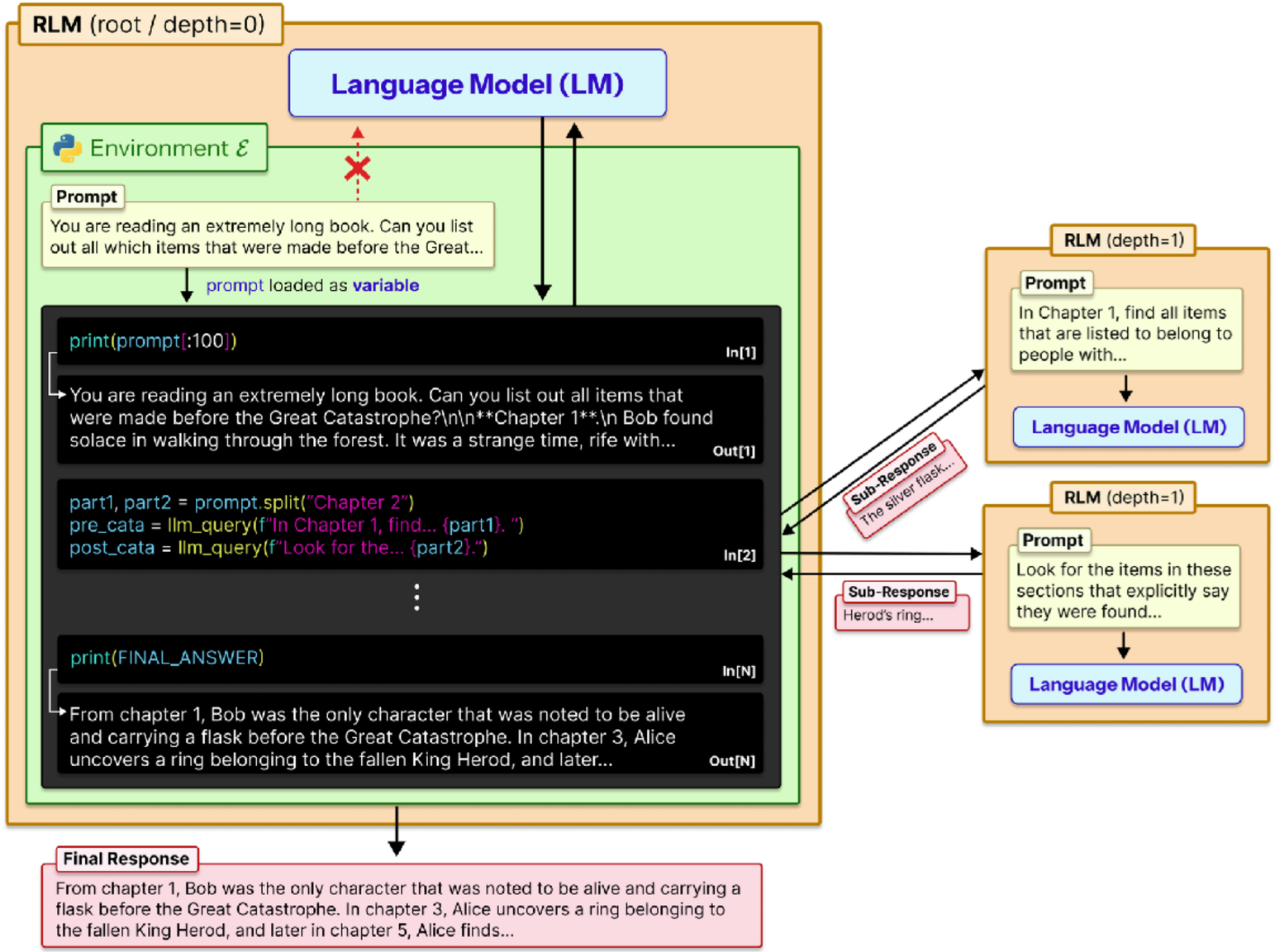


Propagate the reward at the end of last conversation through all previous conversations

[Yu et al., 2026]

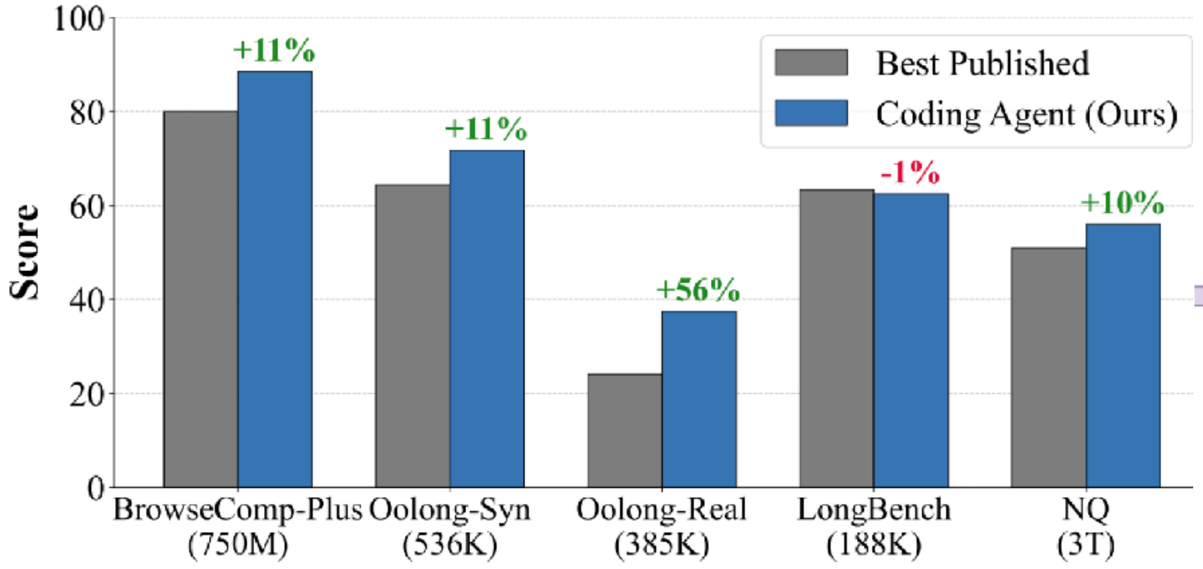
Long-context *Post*-training Setups

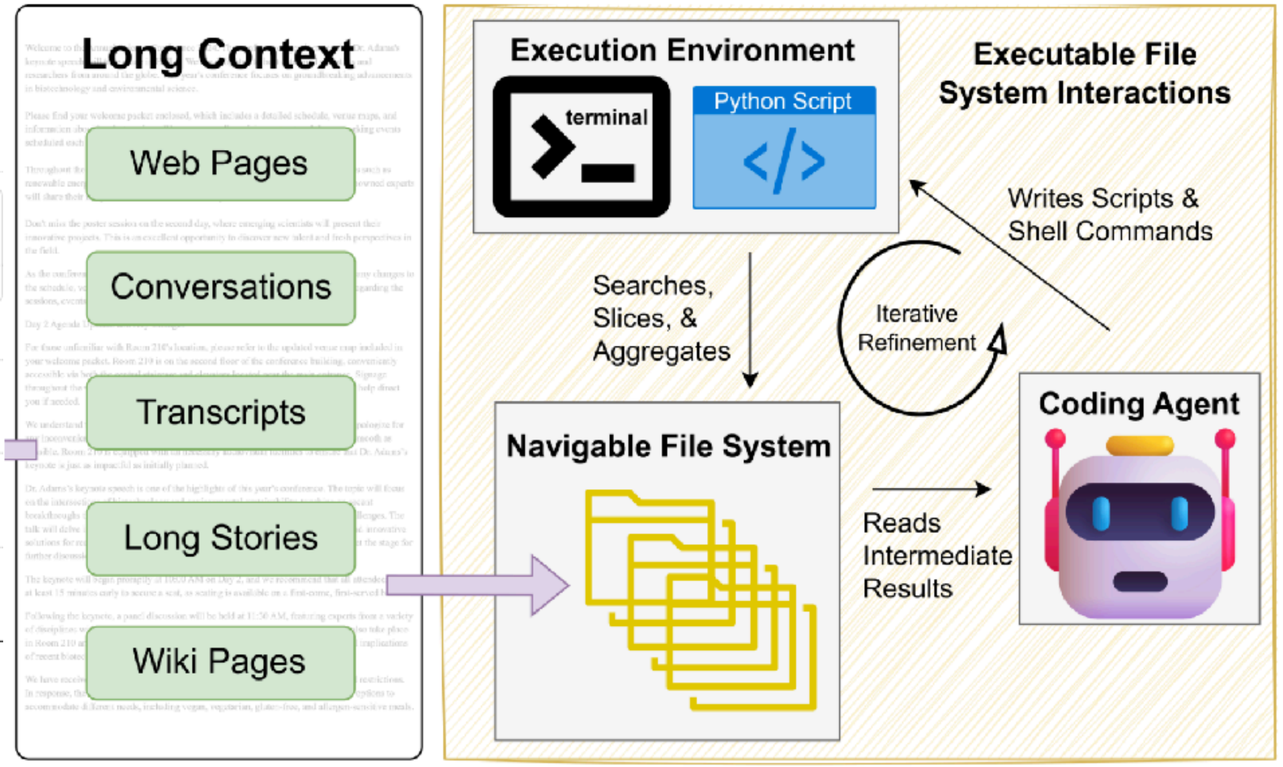
- How can we design agentic systems that allow LMs to interact with context?
- **Coding environments**



Long-context *Post*-training Setups

- How can we design agentic systems that allow LMs to interact with context?
- **General coding agents**

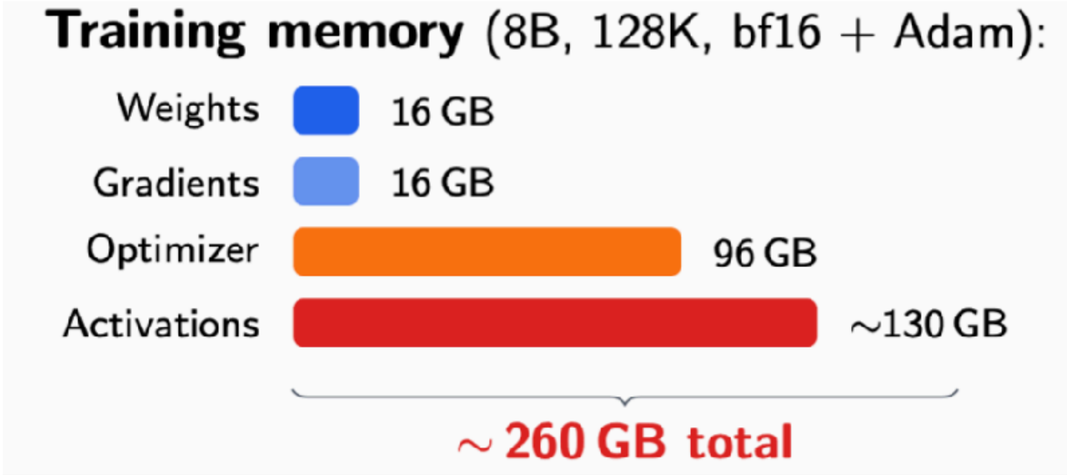




Long-context reasoning is highly expensive.

- Recall that FLOPs and (in naive implementations) memory for self-attention scale quadratically with context window size

Context	KV Cache	vs. Weights
4K	0.5 GB	3%
32K	4 GB	25%
128K	**16 GB**	100%
512K	**64 GB**	400%

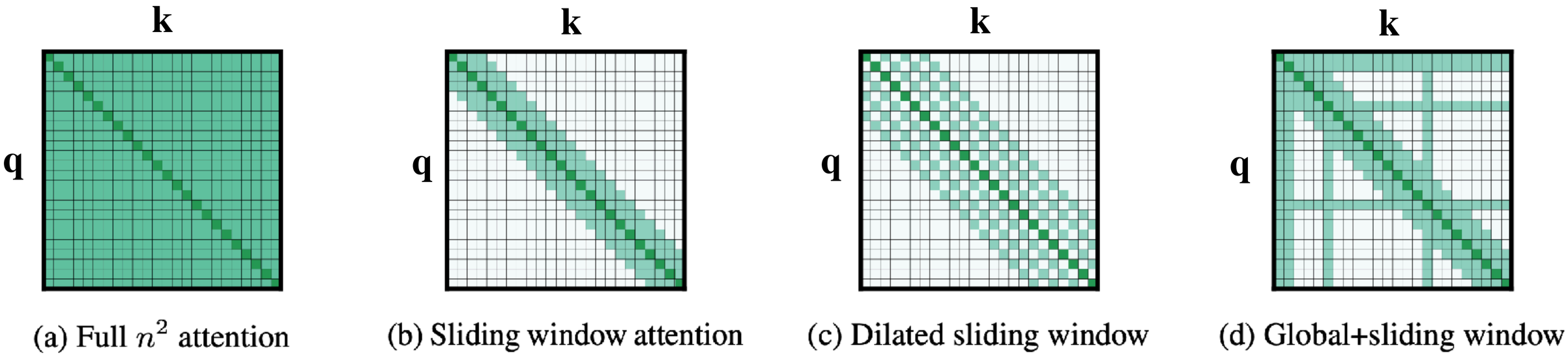


- (FlashAttention can take care of quadratic memory dependencies.)

Efficient Architectures

Longformer

- Some have proposed variants of self-attention that scale linearly w.r.t. context length

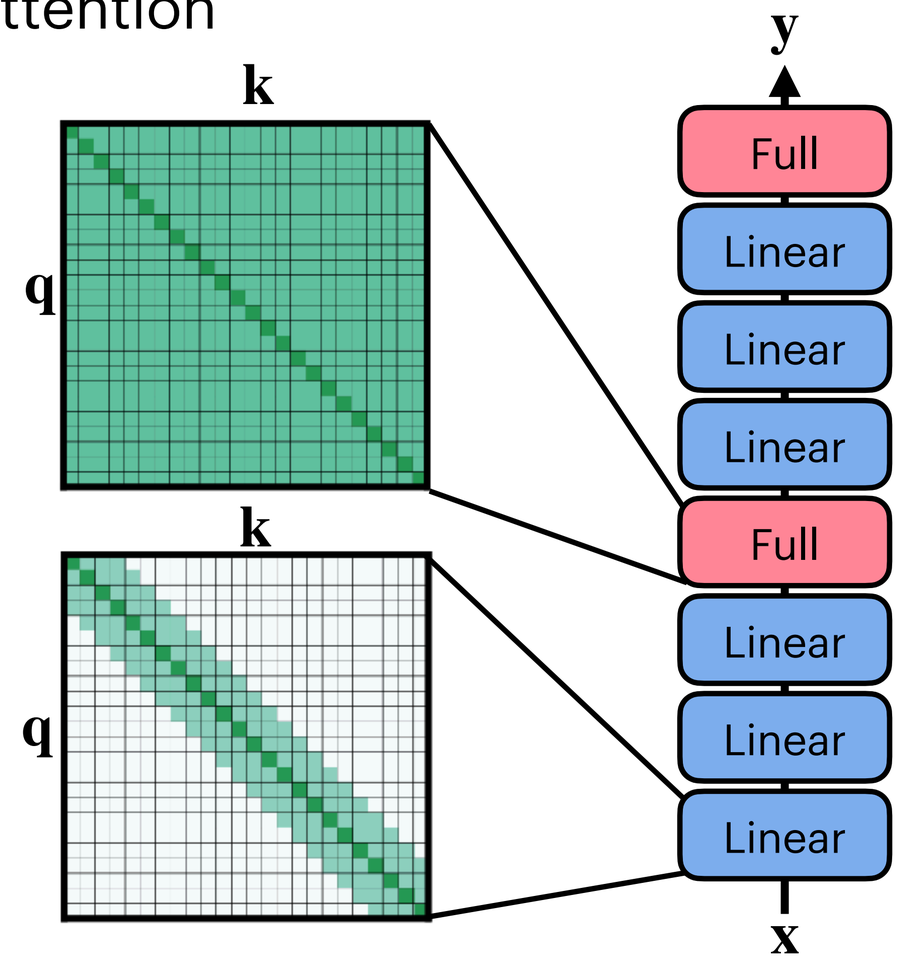


(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

- The **Longformer** architecture uses dilated sliding window or global + sliding window attention (depending on the task)

Efficient Architectures

Current models like Qwen3.5 and Olmo 3 *interleave* layers of full attention and layers of linear attention



3:1 ratio of linear to full attention

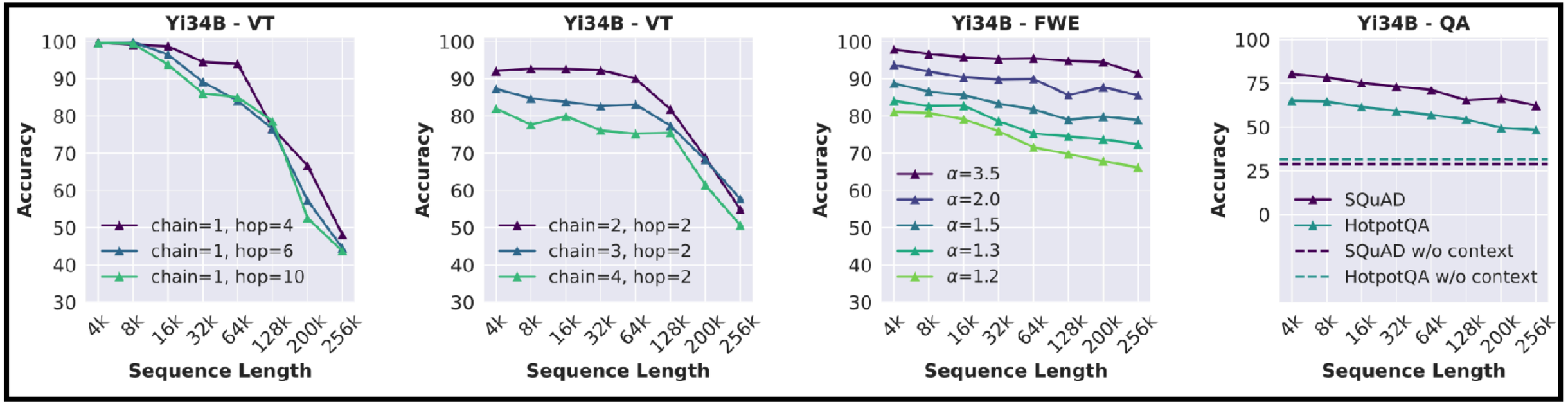
Evaluation

RULER

Variable Tracking (VT)	num_chains = 2 num_hops = 2 size_noises ∝ context length	(noises) ..... VAR X1 = 12345 ..... VAR Y1 = 54321 ..... VAR X2 = X1 ..... VAR Y2 = Y1 ..... VAR X3 = X2 ..... VAR Y3 = Y2 ..... Find all variables that are assigned the value 12345. Answer: X1 X2 X3
Common Words Extraction (CWE)	freq_cw = 2, freq_ucw = 1 num_cw = 10 num_ucw ∝ context length	aaa bbb ccc aaa ddd eee ccc fff ggg hhh iii iii ..... What are the 10 most common words in the above list? Answer: aaa ccc iii .....
Frequent Words Extraction (FWE)	$\alpha = 2$ num_word ∝ context length	aaa bbb ccc aaa ddd eee ccc fff ggg aaa hhh aaa ccc iii iii ..... What are the 3 most frequently appeared words in the above coded text? Answer: aaa ccc iii
Question Answering (QA)	dataset = SQuAD num_document ∝ context length	Document 1: ..... aaa ..... Document 2: ..... bbb ..... Document 3: ..... ccc ..... Question: question Answer: bbb

Evaluation

RULER



Evaluation

HELMET

	Type of tasks							Benchmark features		
	Cite	RAG	Re-rank	Long-QA	Summ	ICL	Synthetic Recall	Robust Eval.	$L \geq 128$k	Controllable L
ZeroSCROLLS	✗	✗	✗	✓	✓	✗	✗	✗	✗†	✗
LongBench	✗	✓	✗	✓	✓	✓	✓	✗	✗†	✗
L-Eval	✗	✓	✗	✓	✓	✗	✗	✓‡	✗†	✗
RULER	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓
∞BENCH	✗	✗	✗	✓	✓	✗	✓	✗	✓	✓
HELMET (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Early document: “The launch is scheduled for May 10.”

Later document: “Security approval is required before launch.”

Final document: “Security approval is delayed until May 17.”

Retrieval: *What was the original launch date?* **Updating:** *Which earlier plan needs revision?*

Integration: *Can the project launch on May 10 under these requirements?*

Synthesis: *Summarize the launch constraints and unresolved decisions.*

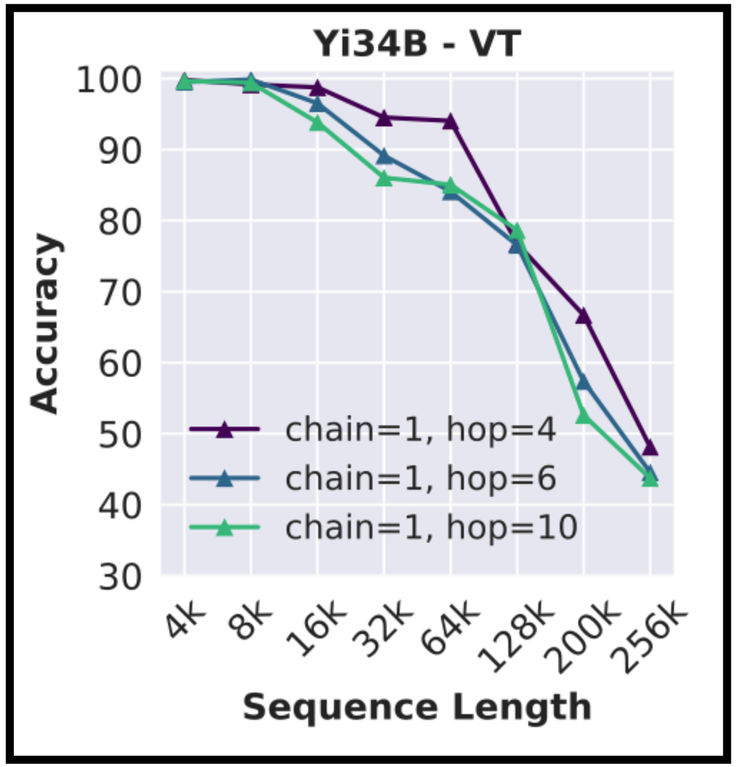
Evaluation

HELMET

	RULER (dev suite)					HELMET (held-out eval)			
Model	**4K**	**8K**	**16K**	**32K**	**65K**	**8K**	**16K**	**32K**	**65K**
7B scale									
Llama 3.1 8B	95.56	92.76	93.13	91.43	86.88	45.00	43.48	42.44	40.18
Qwen 2.5 7B	94.63	90.87	88.68	87.26	67.30	49.26	46.25	42.99	30.47
IBM Granite 3.3 8B	91.98	85.69	82.70	78.13	67.62	43.19	41.63	39.31	35.74
Qwen 3 8B	95.58	94.10	93.78	90.29	-	51.62	49.90	47.71	-
Xiaomi MiMo 7B	94.33	93.45	92.53	89.28	-	50.57	49.68	46.01	-
Nemotron Nano 9B	95.31	93.09	91.58	89.01	85.13	41.78	42.90	41.82	41.48
Apertus 8B	90.47	82.48	74.43	69.05	59.89	46.09	43.71	41.26	35.12
Olmo 3 7B	94.89	91.21	84.14	78.79	67.96	45.66	43.62	41.15	36.80
32B scale									
Qwen 2.5 32B	96.03	94.52	95.07	92.67	80.73	57.61	56.06	54.01	41.73
Gemma 3 27B	84.48	84.20	85.36	87.06	84.59	49.37	49.92	50.31	48.60
Mistral Small 3.1 24B	96.05	95.06	93.77	92.42	88.80	49.41	49.71	47.46	43.34
Apertus 70B	91.52	84.26	80.54	76.82	60.33	44.72	44.60	41.07	35.67
Olmo 3 32B	96.10	94.57	90.42	86.22	79.70	52.11	49.36	48.60	43.15

Long-context Decay

- There is a notion of **long-context decay**: the degradation of model performance as its context grows, even when everything fits in its context window
- This is an umbrella term for many things:
  - Position sensitivity
  - Distractor interference
  - Conversational/turn-taking degradation



Needle-in-a-haystack Tasks

- The goal is to find the answer in otherwise unrelated text

- Usually some direct signal as to what to retrieve: answer has high lexical overlap with question here, for example

[Question]

"What was the best writing advice I got from my college classmate?"

[Needle]

[Haystack]

I've discovered a handy test for figuring out what you're addicted to.

Imagine you were going to spend the weekend at a friend's house on a little island off the coast of Maine. There are no shops on the island and you won't be able to leave while you're there. Also, you've never been to this house before, so you can't assume it will have more than any house might.

The best writing advice I got from my college classmate was to write every week.

What, besides clothes and toiletries, do you make a point of packing? That's what you're addicted to.

For example, if you find yourself packing a bottle of vodka (just in case), you may want to stop and think about that. For me the list is four things: books, earplugs, a notebook, and a pen. There are other things I might bring if I thought of it, like music, or tea, but I can live without them.

I'm not so addicted to caffeine that I wouldn't risk the house not having any tea, just for a weekend.

LLMs get "lost in the middle."

Input Context

```
Write a high-quality answer for the given question using only the provided search
results (some of which might be irrelevant).

Document [1](Title: Asian Americans in science and technology) Prize in physics for
discovery of the subatomic particle J/ψ. Subrahmanyan Chandrasekhar shared...
Document [2](Title: List of Nobel laureates in Physics) The first Nobel Prize in
Physics was awarded in 1901 to Wilhelm Conrad Röntgen, of Germany, who received...
Document [3](Title: Scientist) and pursued through a unique method, was essentially
in place. Ramón y Cajal won the Nobel Prize in 1906 for his remarkable...

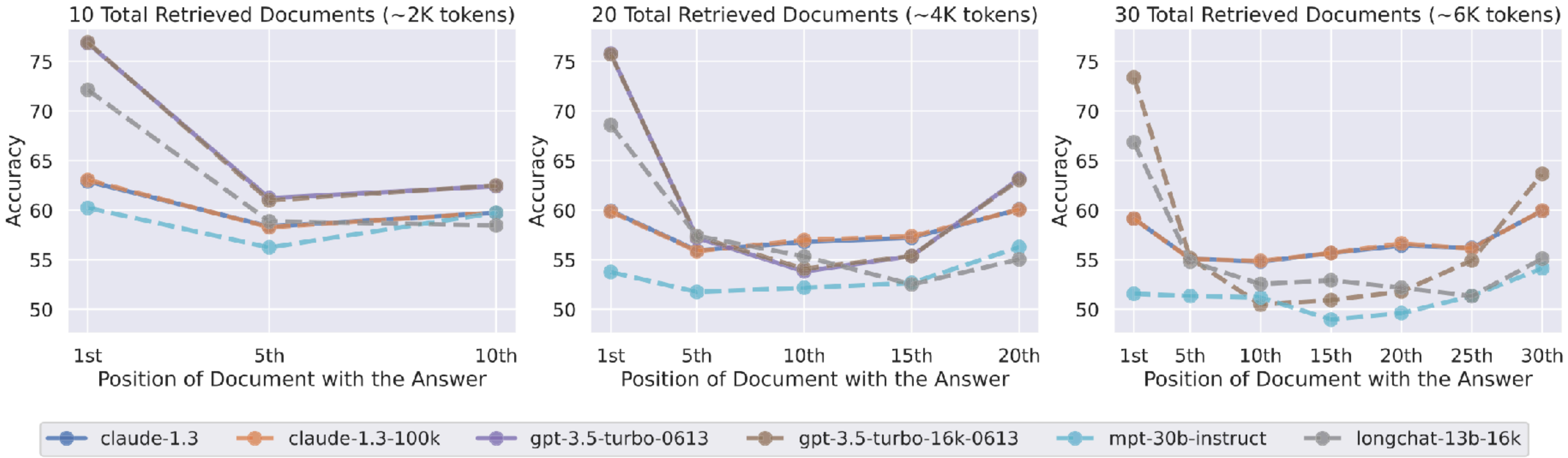
Question: who got the first nobel prize in physics
Answer:
```

Desired Answer

```
Wilhelm Conrad Röntgen
```

- Multi-document QA. Like needle-in-a-haystack, but with some related and some unrelated texts
- We can control the position of relevant information and the total length of the input

LLMs get "lost in the middle."



We see a "lost in the middle" phenomenon: performance is highest when relevant info appears near the beginning or end of the context

Multi-turn Conversation

Fully-Specified Instruction (original)

Jay is making snowballs to prepare for a snowball fight with his sister. He can build 20 snowballs in an hour, but 2 melt every 15 minutes. How long will it take before he has 60 snowballs?

Sharded Instruction (based on original)

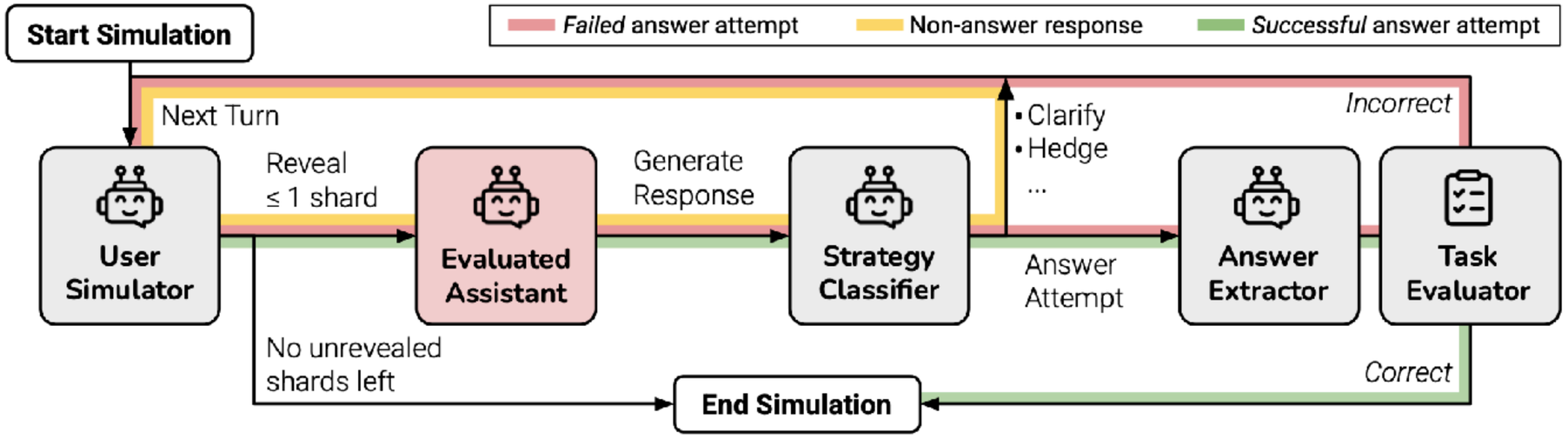
Shard 1: How long before Jay's ready for the snowball fight?

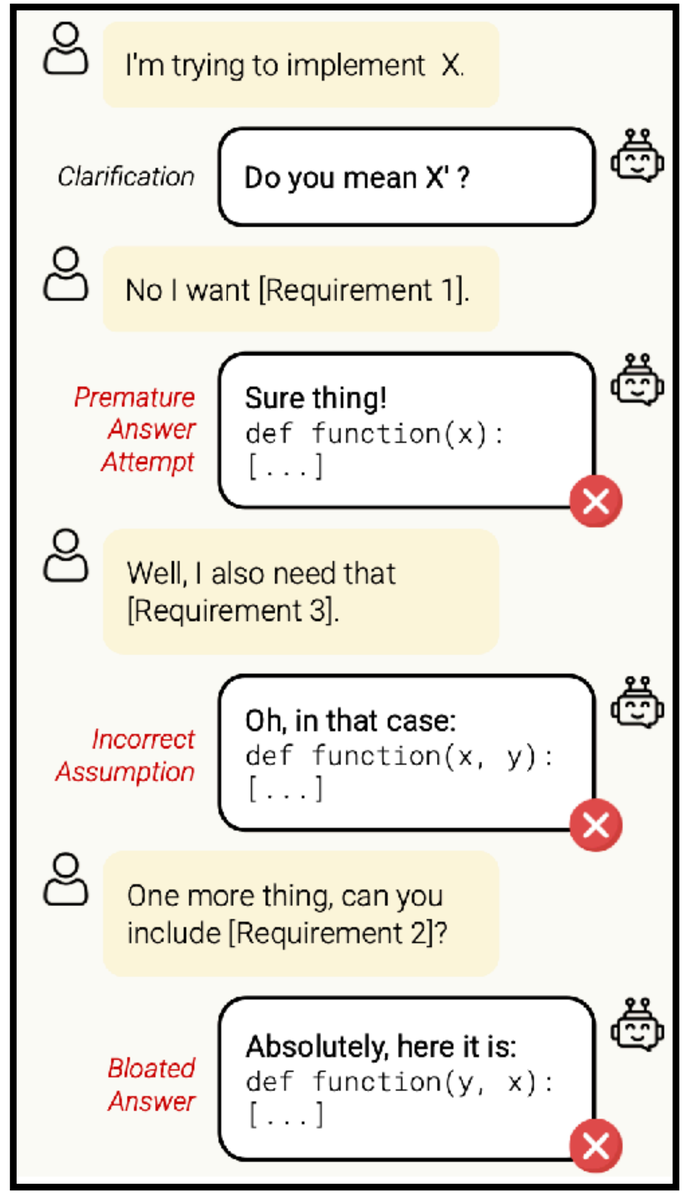
Shard 2: He's preparing for a snowball fight with his sister.

Shard 3: He can make 20 snowballs per hour.

Shard 4: He's trying to get to 60 total.

Shard 5: The problem is that 2 melt every 15 minutes.





Multi-turn Conversation

Fully-Specified Instruction (original)

Jay is making snowballs to prepare for a snowball fight with his sister. He can build 20 snowballs in an hour, but 2 melt every 15 minutes. How long will it take before he has 60 snowballs?

Sharded Instruction (based on original)

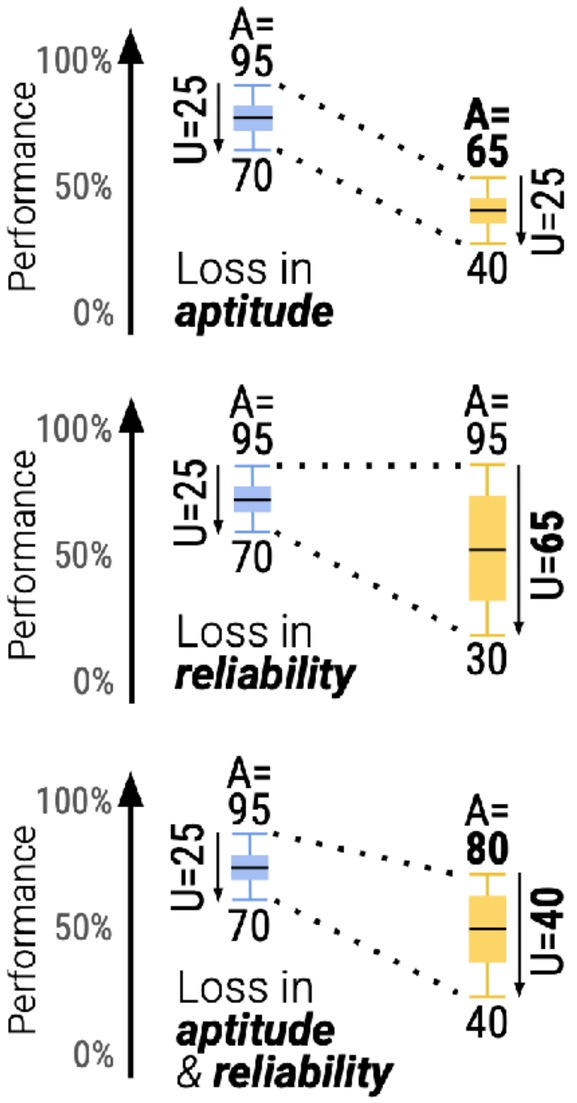
Shard 1: How long before Jay's ready for the snowball fight?

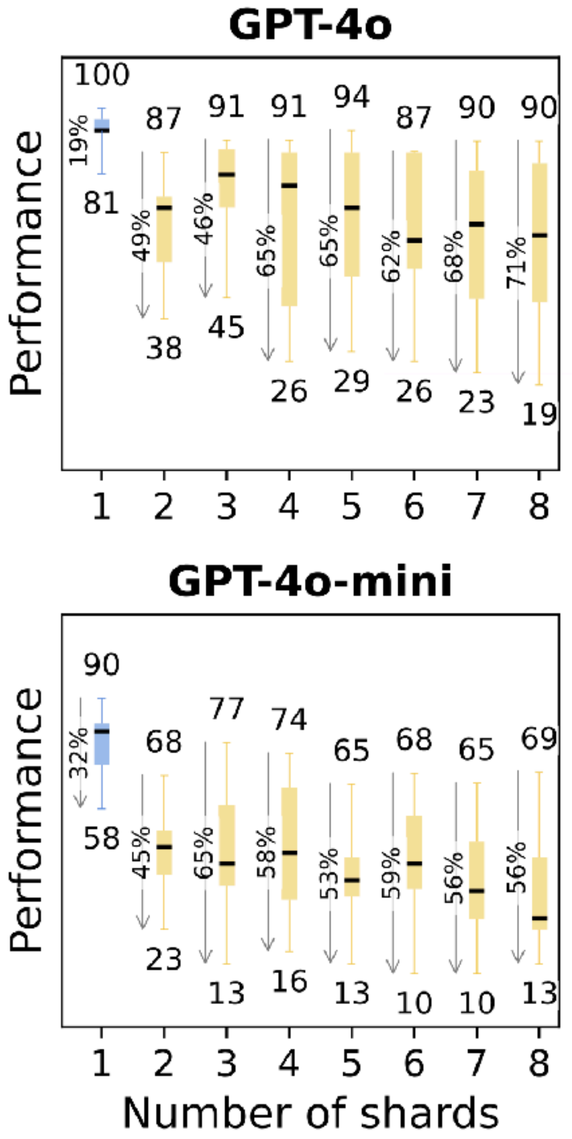
Shard 2: He's preparing for a snowball fight with his sister.

Shard 3: He can make 20 snowballs per hour.

Shard 4: He's trying to get to 60 total.

Shard 5: The problem is that 2 melt every 15 minutes.





Managing Context in an Agent

Strategy	Pros	Cons
Retrieve relevant passages	Reduces unnecessary context	Misses necessary evidence
Summarize older history	Preserves a compact record	Loses details, preserves mistakes
Maintain a task state	Tracks constraints and decisions	State can become stale or incorrect
Start a fresh context with some handoff procedure	Removes built-up conversational info	Can omit important dependencies

Takeaways

- Getting models to handle long contexts requires specialized data and (post-)training
- One can modify elements of the Transformer to make context scaling more efficient
- Performance can depend on many factors, including position, distractors, interaction history, etc.
    - We need diverse evaluations! Not enough to just evaluate retrieval
- Agents need to manage their contexts effectively. Each strategy has trade-offs