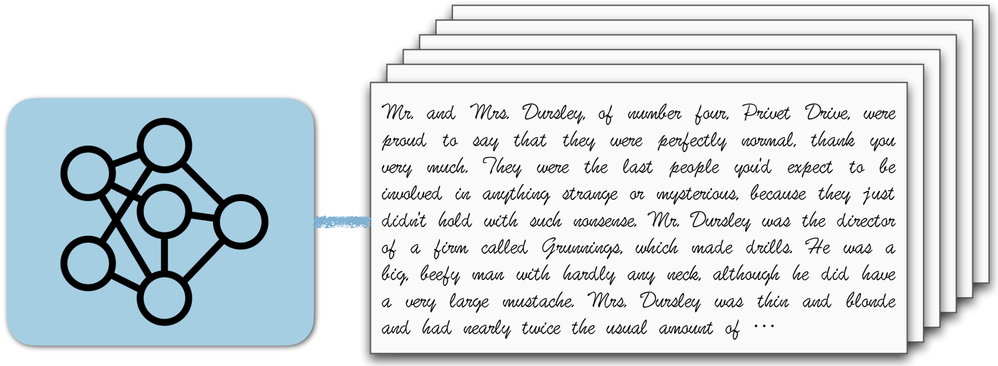
What's Next for Scaling, and How Should Pre-Training Change?

Sewon Min

sewonmin.com

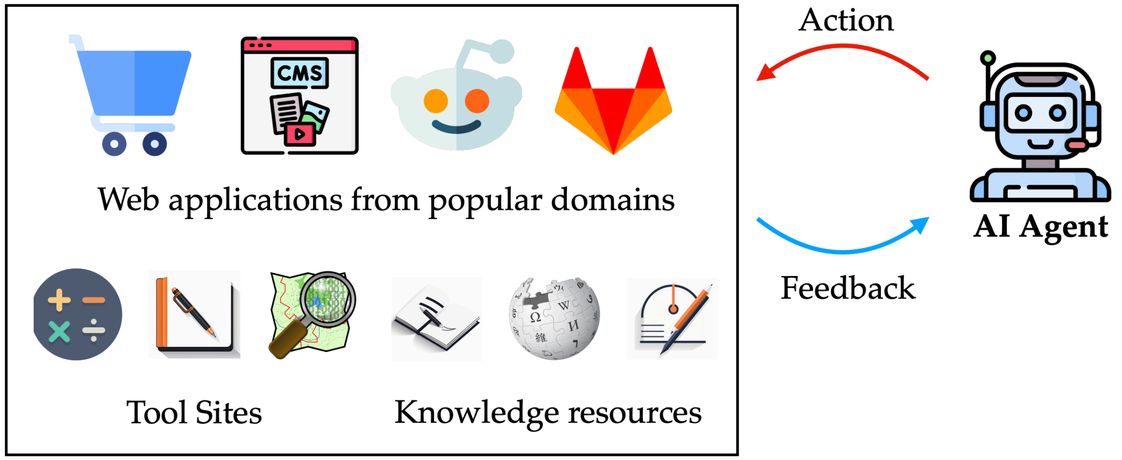
Berkeley UNIVERSITY OF CALIFORNIA BAIR Ai2

~2023: What can't scaling alone achieve?



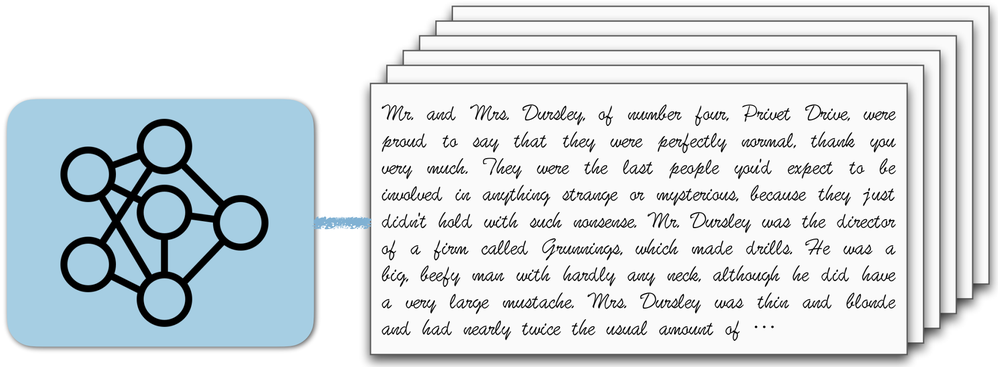
A retrieval-based LM

corpus can grow, update, or remove



WebArena — Zhou, Xu et al., ICLR 2024

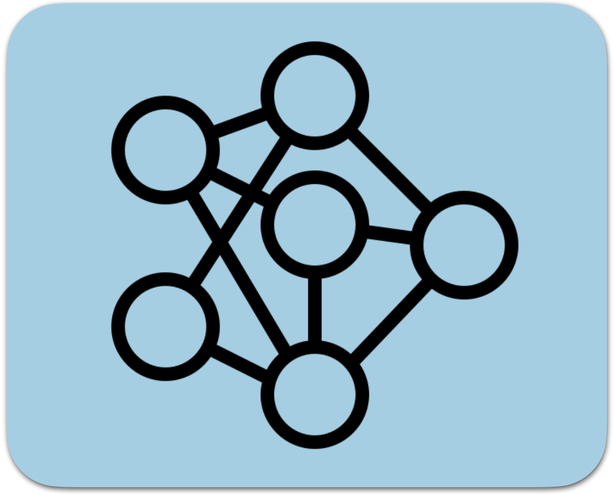
~2023: What can't scaling alone achieve?



A retrieval-based LM

corpus can grow, update, or remove

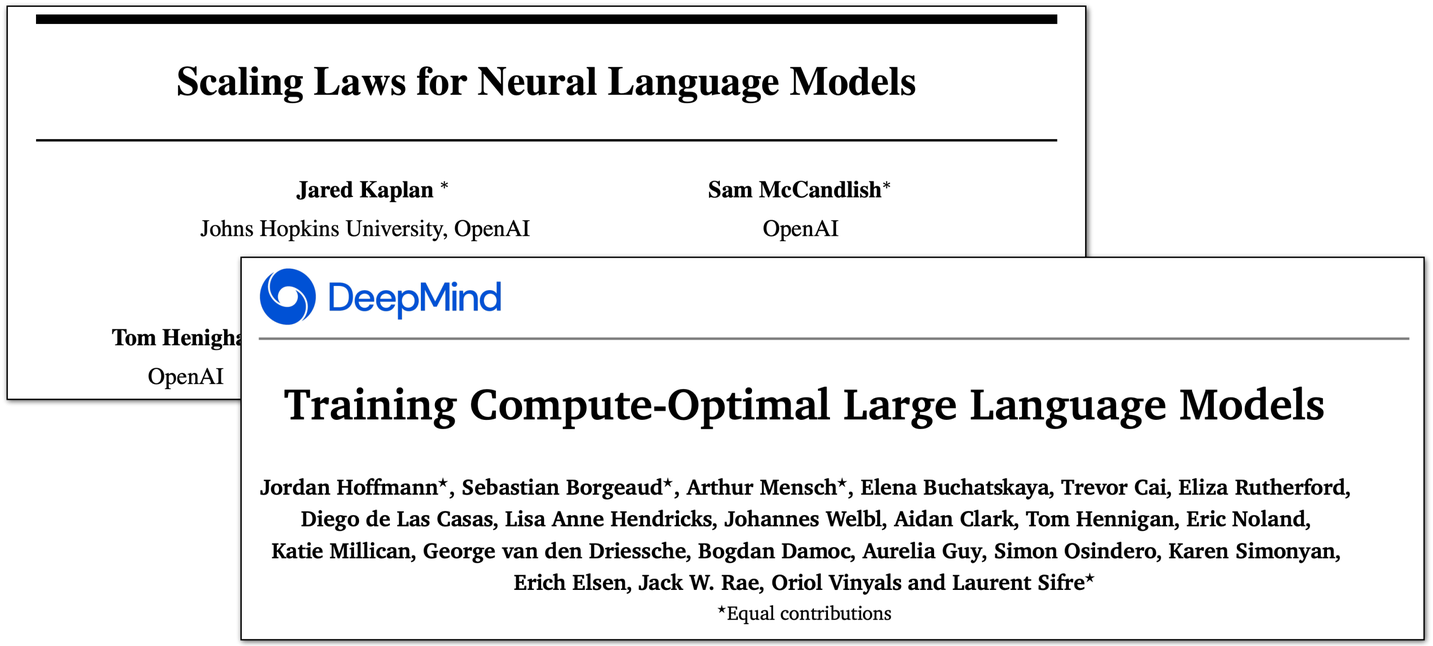
VS



How do we train more capable models?
What's next in scaling?

Pre-training

- So far: Compute has been the binding resource



Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI

Sam McCandlish*
OpenAI

Tom Henigha
OpenAI

DeepMind

Training Compute-Optimal Large Language Models

Jordan Hoffmann*, Sebastian Borgeaud*, Arthur Mensch*, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals and Laurent Sifre*

*Equal contributions

Pre-training

- So far: Compute has been the binding resource
- Now: Constraints are becoming more complex
  - Data is the binding resource — data is “running out”!
  - Compute ≠ FLOPs
    - # of GPU nodes we can bring together (+ the cost of communication) is becoming a bottleneck
    - Memory is increasingly expensive

CNBC

AI memory is sold out, causing an unprecedented surge in prices

AI memory is sold out, causing an unprecedented surge in prices ... All computing



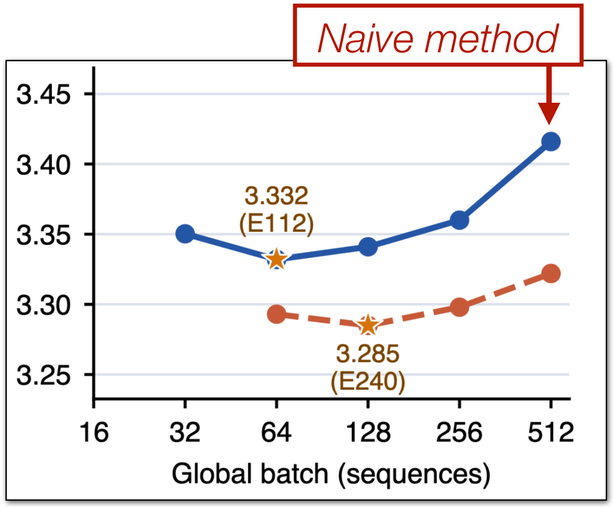
www.cnbc.com/2026/01/10/micron-ai-memory-shortage-hbm-nvidia-samsung.html

Pre-training

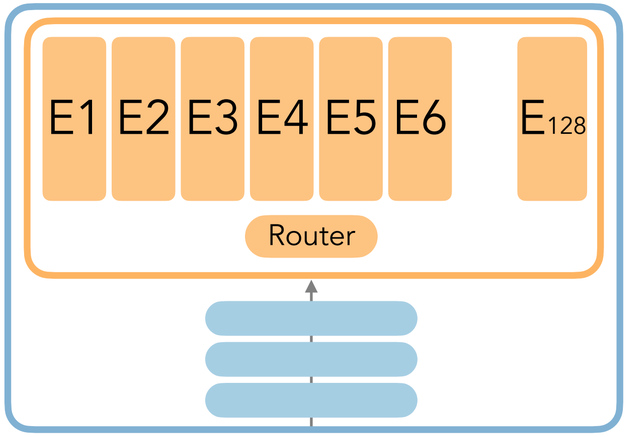
- So far: Compute has been the binding resource
- Now: Constraints are becoming more complex
  - Data is the binding resource — data is “running out”!
  - Compute ≠ FLOPs
    - # of GPU nodes we can bring together (+ the cost of communication) is becoming a bottleneck
    - Memory is increasingly expensive

Punchline: When the constraints change, pre-training needs to change too.

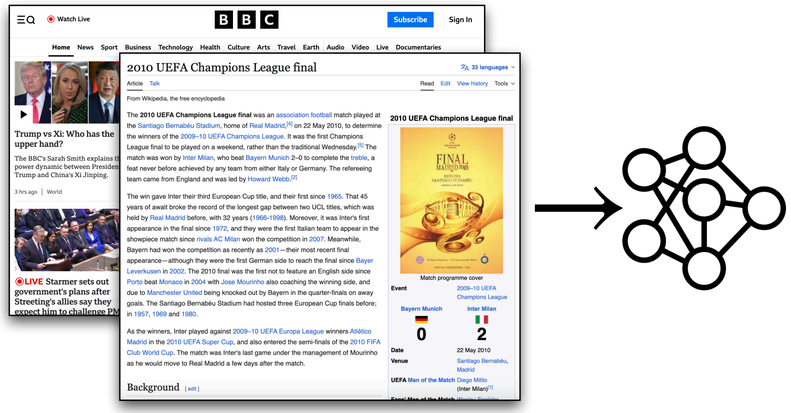
Today's talk



1. How to do repeated-data training *correctly*

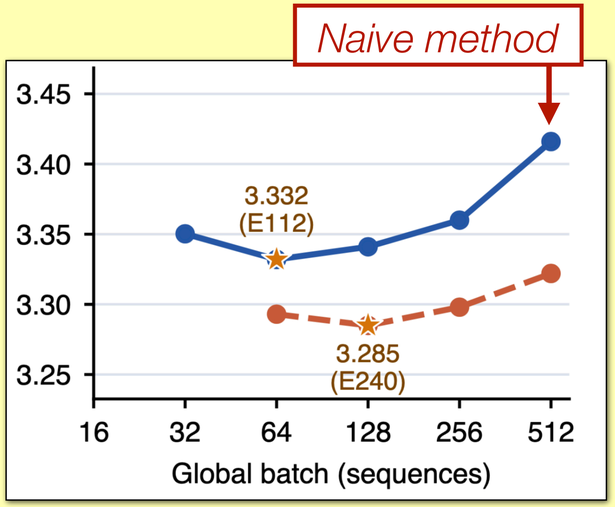


2. New ways of training Mixture-of-Experts

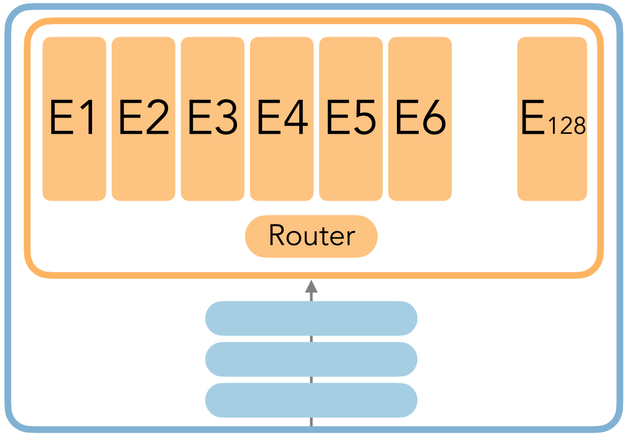


3. Bypassing text parsing

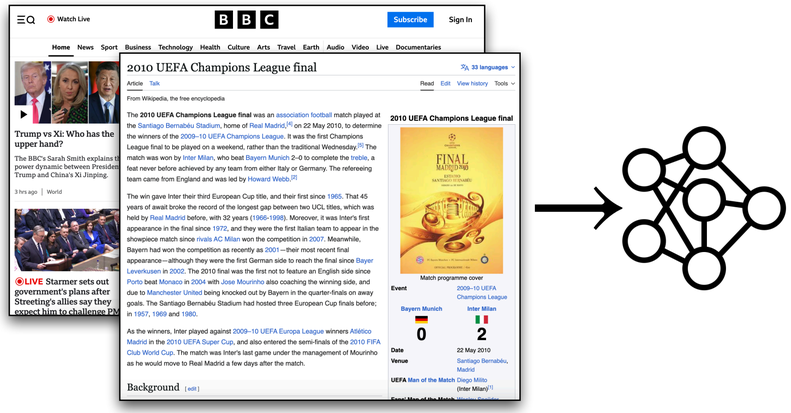
Today's talk



1. How to do repeated-data training *correctly*



2. New ways of training Mixture-of-Experts

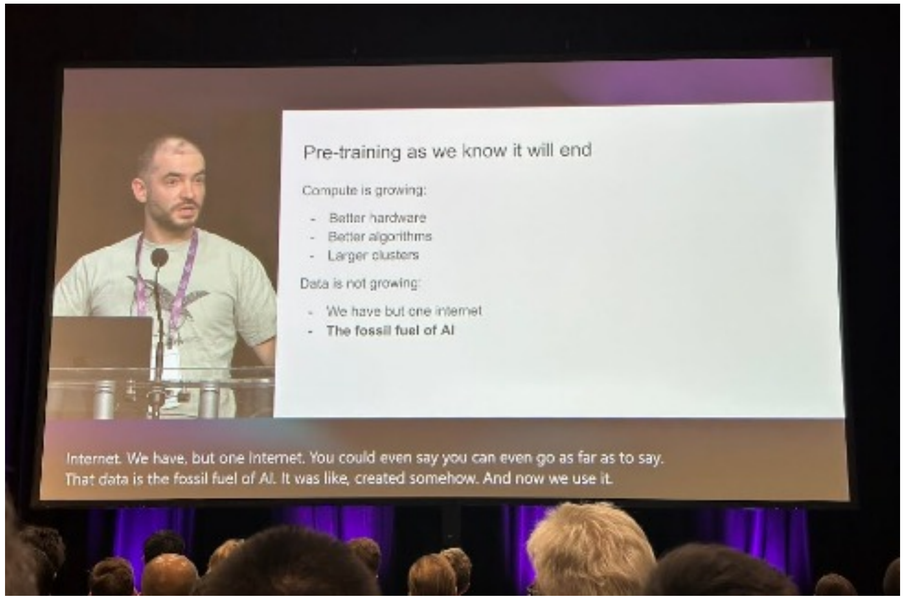


3. Bypassing text parsing

"Scaling Repeated-Data Pretraining: Batch Size, Regularization, and GPU-Time Tradeoffs"

arxiv.org/abs/2610.15965

Data is “running out”



“We have but one Internet.”

Ilya Sutskever @ NeurIPS 2024

Will we run out of data? Limits of LLM scaling based on human-generated data

Pablo Villalobos [1] **Anson Ho** [1] **Jaime Sevilla** [1 2] **Tamay Besiroglu** [1 3] **Lennart Heim** [1 4] **Marius Hobbhahn** [1 5]

Abstract

We investigate the potential constraints on LLM scaling posed by the availability of public human-generated text data. We forecast the growing demand for training data based on current trends and estimate the total stock of public human text data. Our findings indicate that if current LLM development trends continue, models will be trained on datasets roughly equal in size to the available stock of public human text data between 2026 and 2032, or slightly earlier if models are overtrained. We explore how progress in language modeling can continue when human-generated text datasets cannot be scaled any further. We argue that synthetic data generation, transfer learning from data-rich domains, and data efficiency improvements might support further progress.

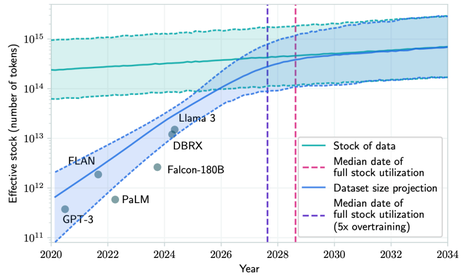


Figure 1. Projections of the effective stock of human-generated public text and dataset sizes used to train notable LLMs. The intersection of the stock and dataset size projection lines indicates the median year (2028) in which the stock is expected to be fully utilized if current LLM development trends continue. At

Forecast: Human text may be exhausted in 2028.

Data is “running out” — why?

1. The absolute quantity is limited physically.

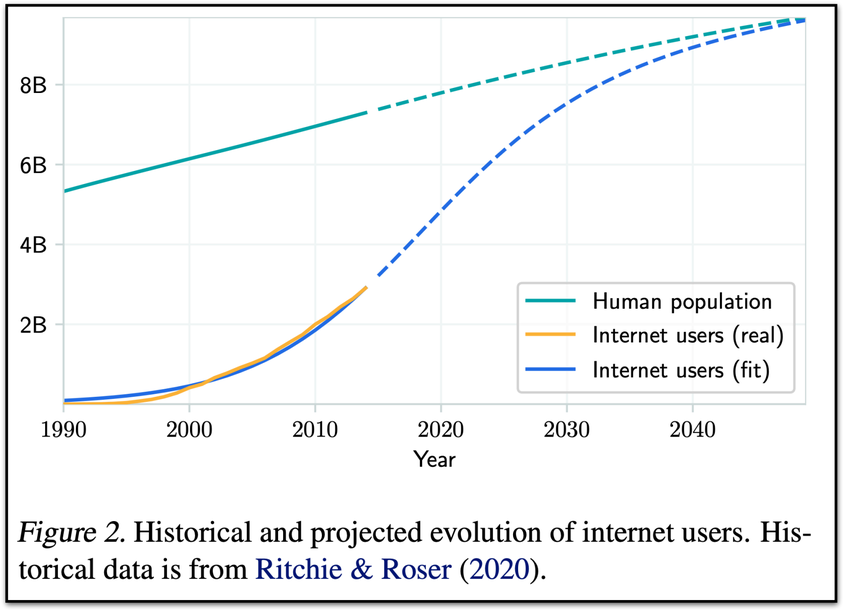


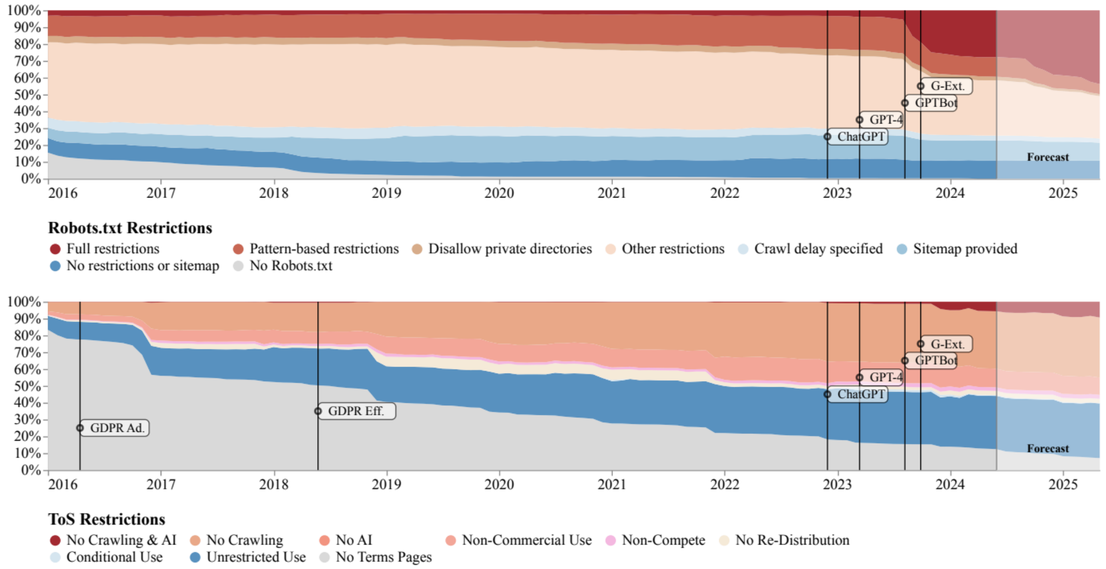
Figure 2. Historical and projected evolution of internet users. Historical data is from Ritchie & Roser (2020).



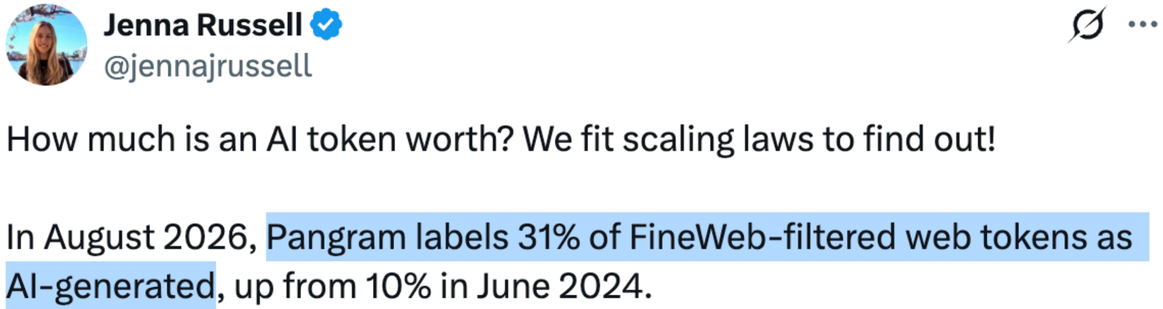
Finite Internet users, finite token generation rate per user

Data is “running out” — why?

2. The growth in data supply is declining



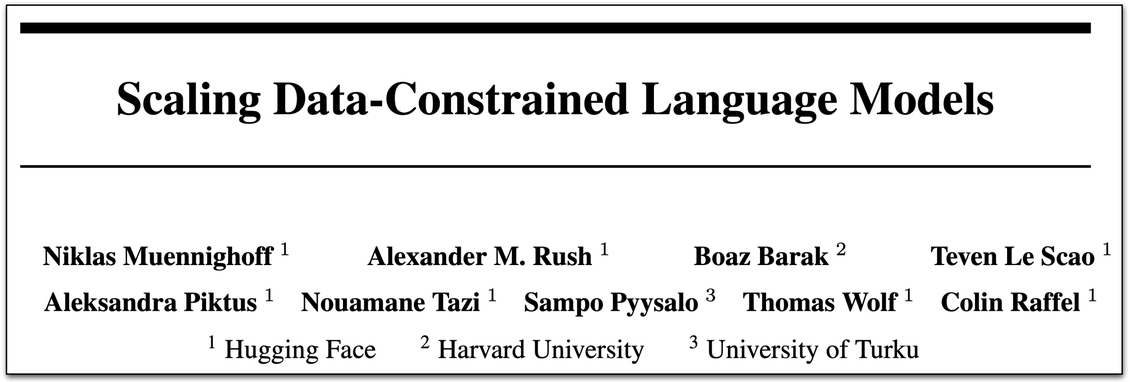
Longpre et al. 2024. "Consent in Crisis: The Rapid Decline of the AI Data Commons"



Russell et al. 2026. "How Much Is an AI Token Worth? Scaling Laws for Wild AI-Generated Web Text"

We should repeat the data!

Prior repeated-data training



Scaling Data-Constrained Language Models

Niklas Muennighoff [1] **Alexander M. Rush** [1] **Boaz Barak** [2] **Teven Le Scao** [1]
Aleksandra Piktus [1] **Nouamane Tazi** [1] **Sampo Pyysalo** [3] **Thomas Wolf** [1] **Colin Raffel** [1]
[1] Hugging Face [2] Harvard University [3] University of Turku



Pre-training under infinite compute

Konwoo Kim∞, Suhas Kotha∞, Percy Liang, Tatsunori Hashimoto
Stanford University

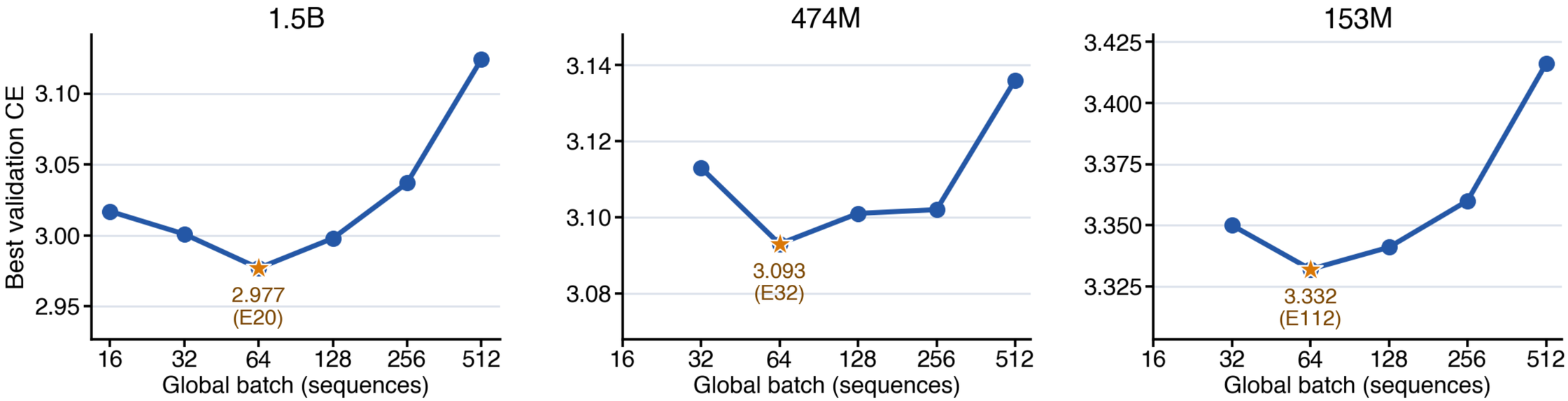
"Up to 4 epochs is similar to fresh data"

Training recipe remains unchanged

"Tuning weight decay improves repeated-data frontier"

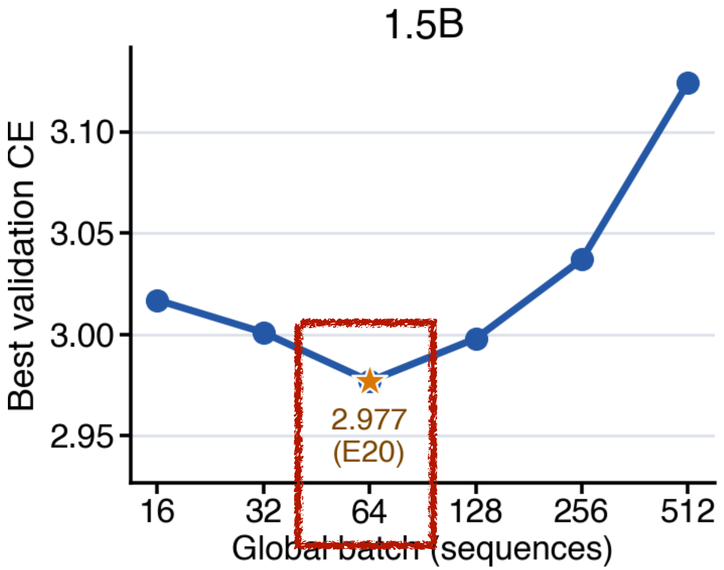
Repeated-data training calls for a different recipe.

Finding 1: Infinite-compute optimal batch size is much smaller than what's normally used



(batch size, learning rate, weight decay, number of repetitions) jointly tuned; 1B-token unique-data pool

Finding 1: Infinite-compute optimal batch size is much smaller than what's normally used

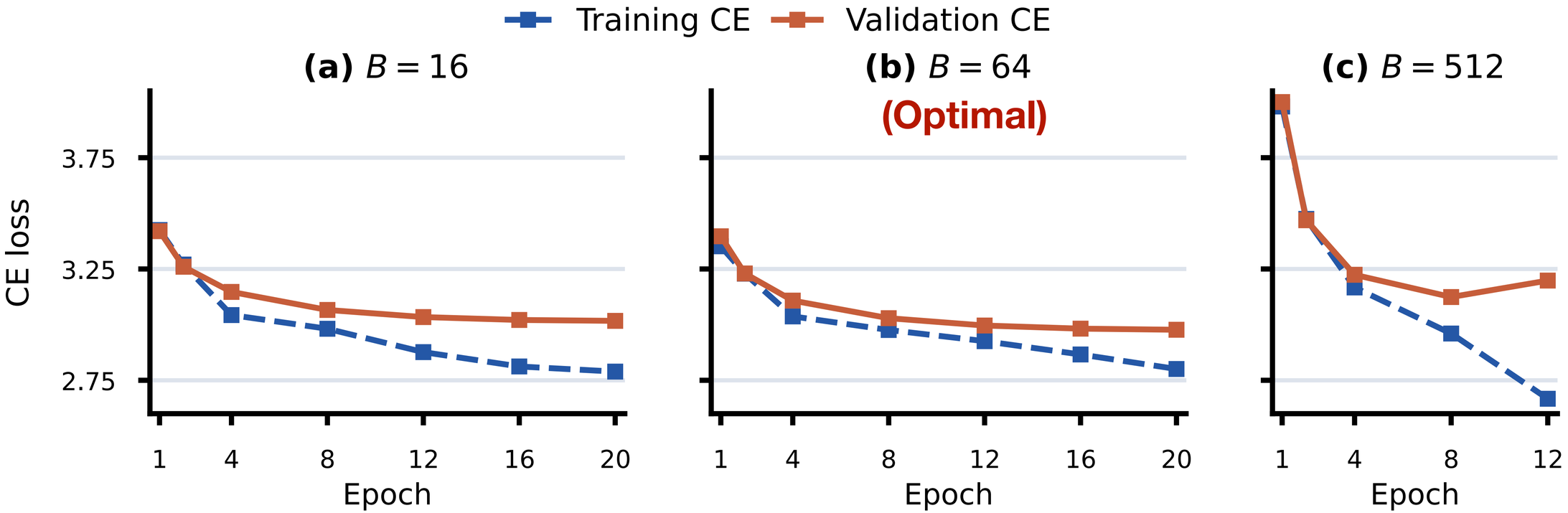


- Much smaller than what's used for public artifacts — GPT-2, Cerebras-GPT, Pythia, and Olmo all used global batches that are 8x to 32x larger.

- Much smaller than critical batch size identified by prior work

	C_{F} (equiv. E)	Batch size (million tokens)
Bi et al.	0.90×10^{19} ($E = 1$)	0.463
Bi et al.	7.20×10^{19} ($E = 8$)	0.914
Bi et al.	18.0×10^{19} ($E = 20$)	1.233
Shuai et al.	0.90×10^{19} ($E = 1$)	0.551
Shuai et al.	7.20×10^{19} ($E = 8$)	0.681
Shuai et al.	18.0×10^{19} ($E = 20$)	0.747
Ours	∞ (C_{F}: $E = 20$)	0.262

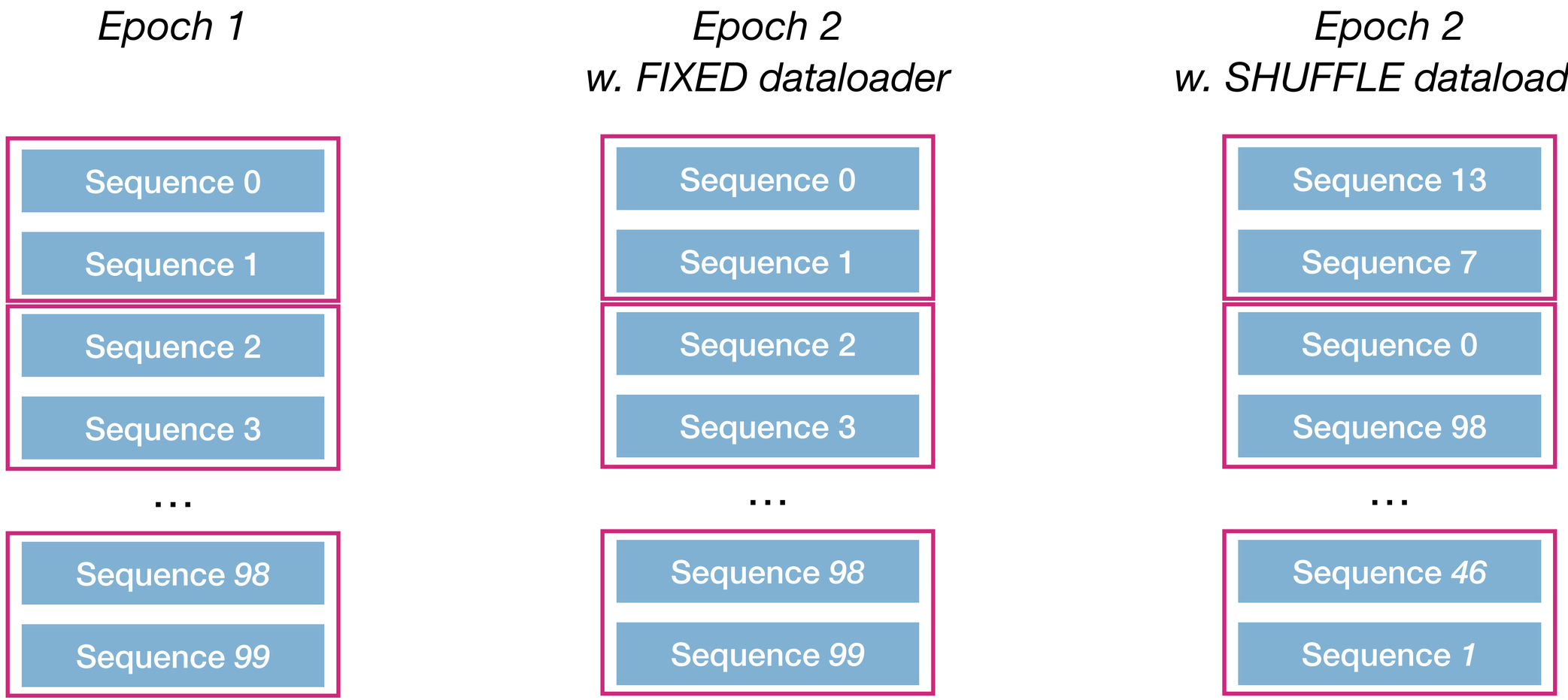
Finding 1: Infinite-compute optimal batch size is much smaller than what's normally used



Noisy gradients from small-batch training provide useful regularization!

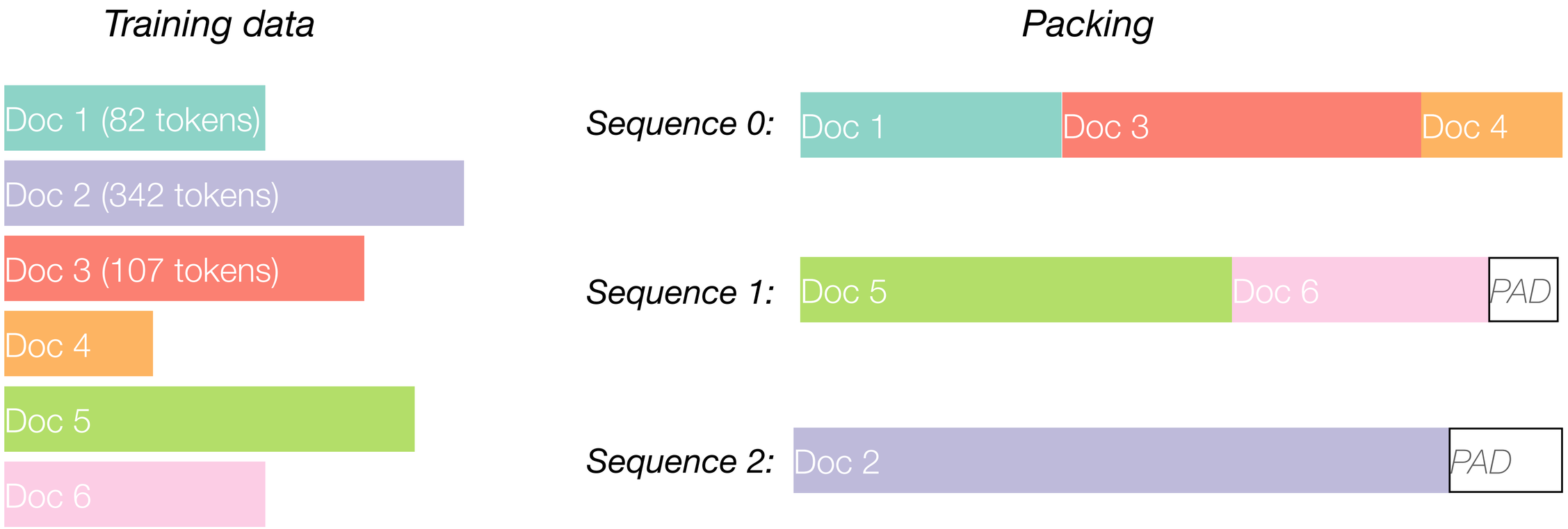
Finding 2: Recipe itself can be tuned

1. Data loader



Finding 2: Recipe itself can be tuned

1. Data loader — REPACK



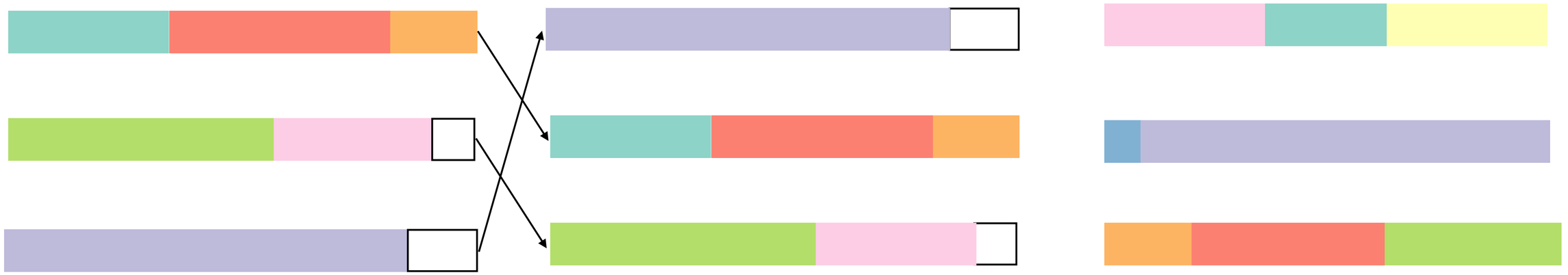
Finding 2: Recipe itself can be tuned

1. Data loader — REPACK

Epoch 1

Epoch 2
w. SHUFFLE dataloader

Epoch 2
w. REPACK dataloader

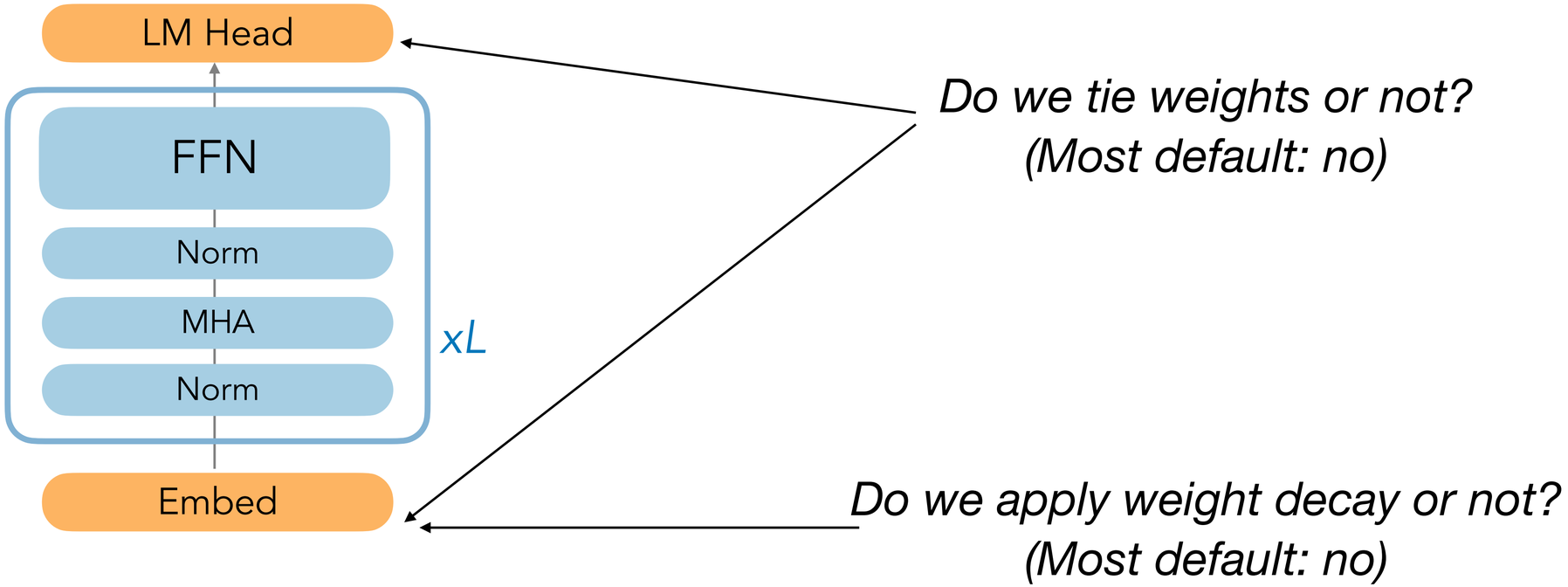


Order is shuffled, but the same sequences are seen

Every sequence is literally fresh

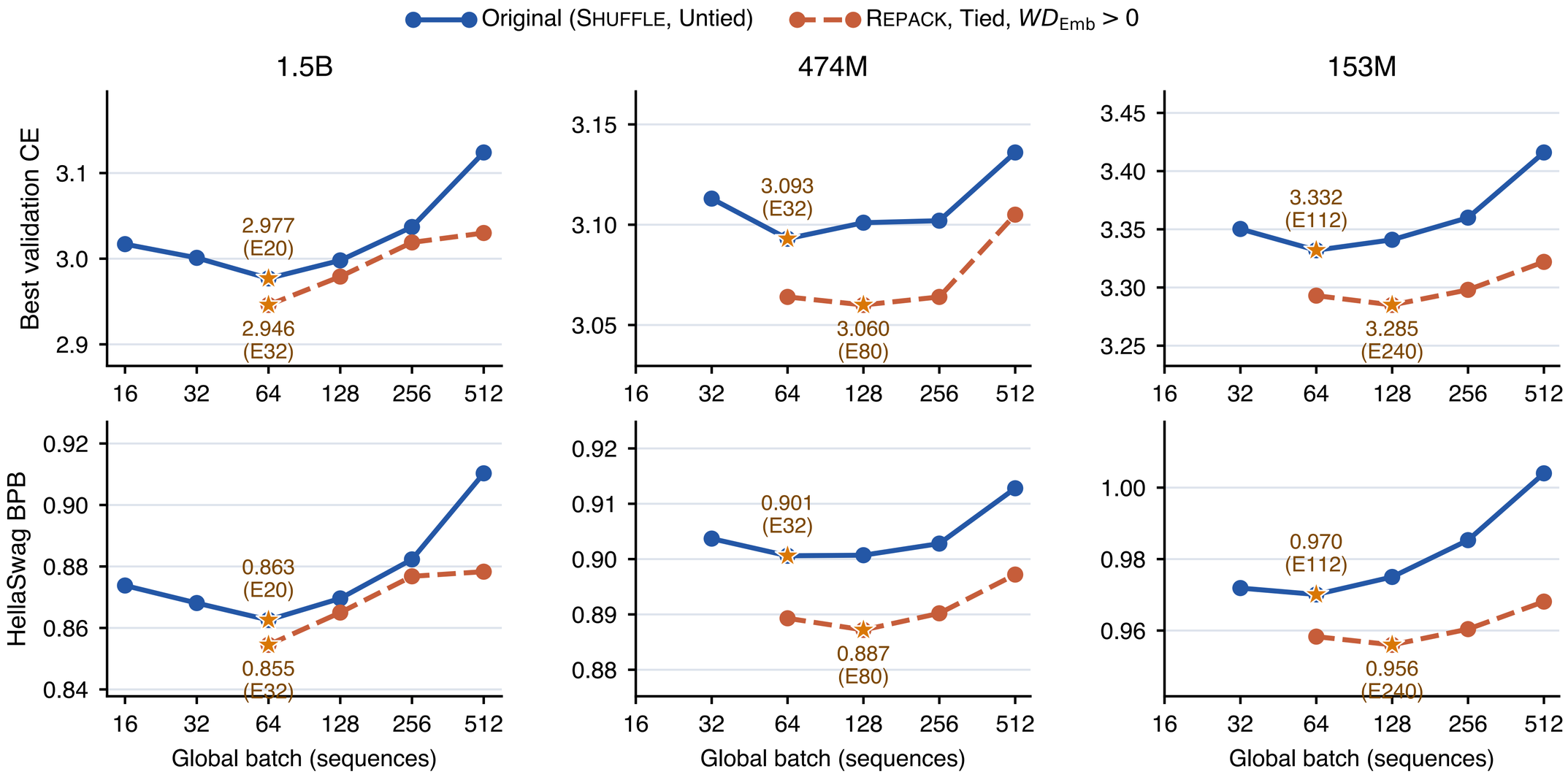
Finding 2: Recipe itself can be tuned

2. Embeddings (tying, decay)

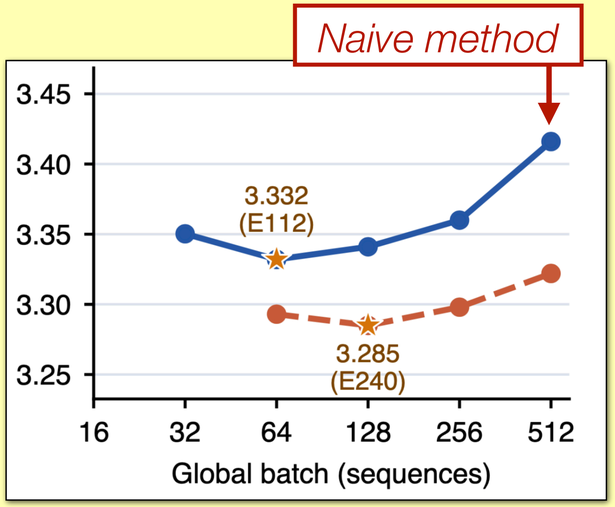


Turns out: weight tying & weight decay to Embed helps

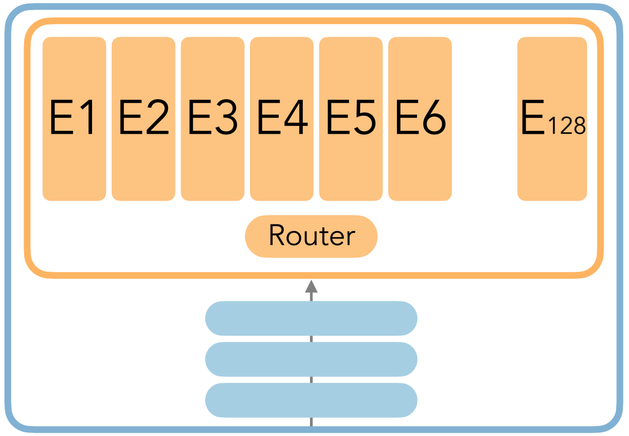
Finding 2: Recipe itself can be tuned



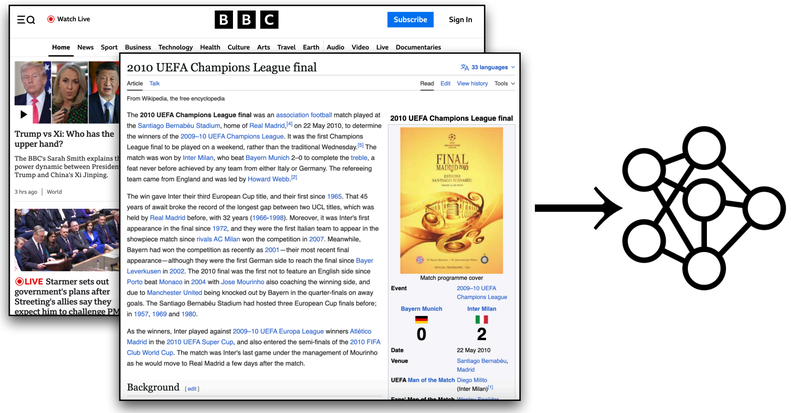
Today's talk



1. How to do repeated-data training *correctly*



2. New ways of training Mixture-of-Experts

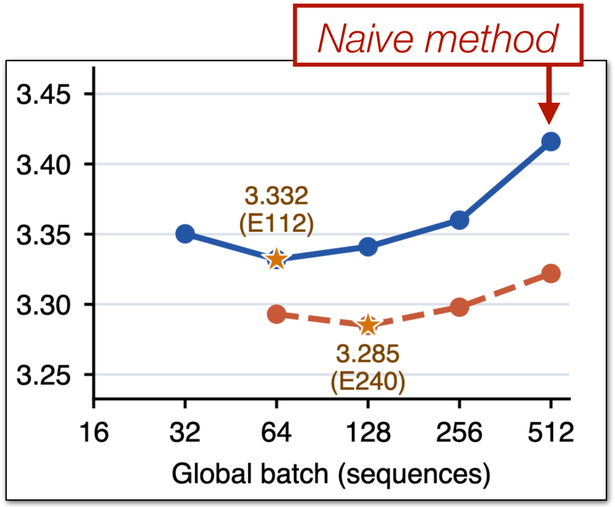


3. Bypassing text parsing

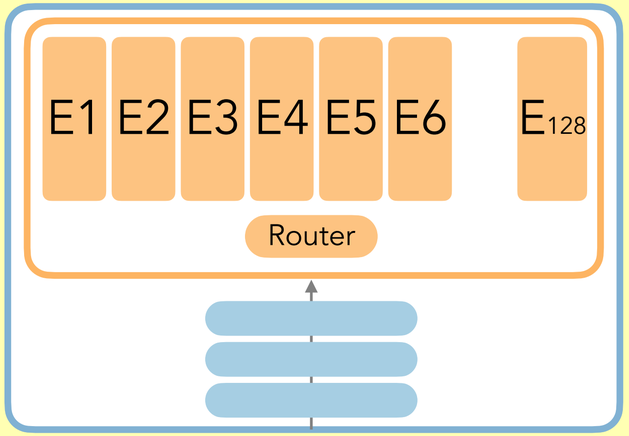
"Scaling Repeated-Data Pretraining: Batch Size, Regularization, and GPU-Time Tradeoffs"

arxiv.org/abs/2610.15965

Today's talk



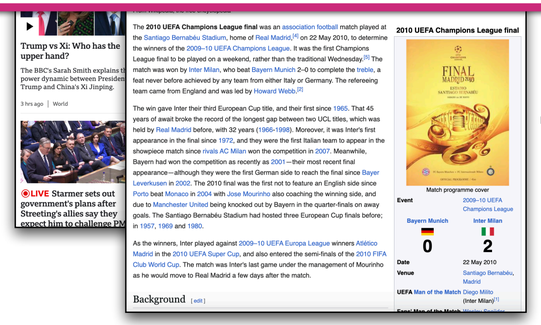
1. How to do repeated-data training *correctly*

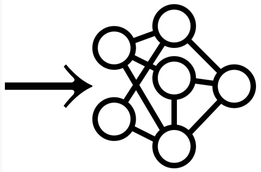


2. New ways of training Mixture-of-Experts



w/ Ryan Wang, Akshita Bhagia
+ Many others for scaling it up

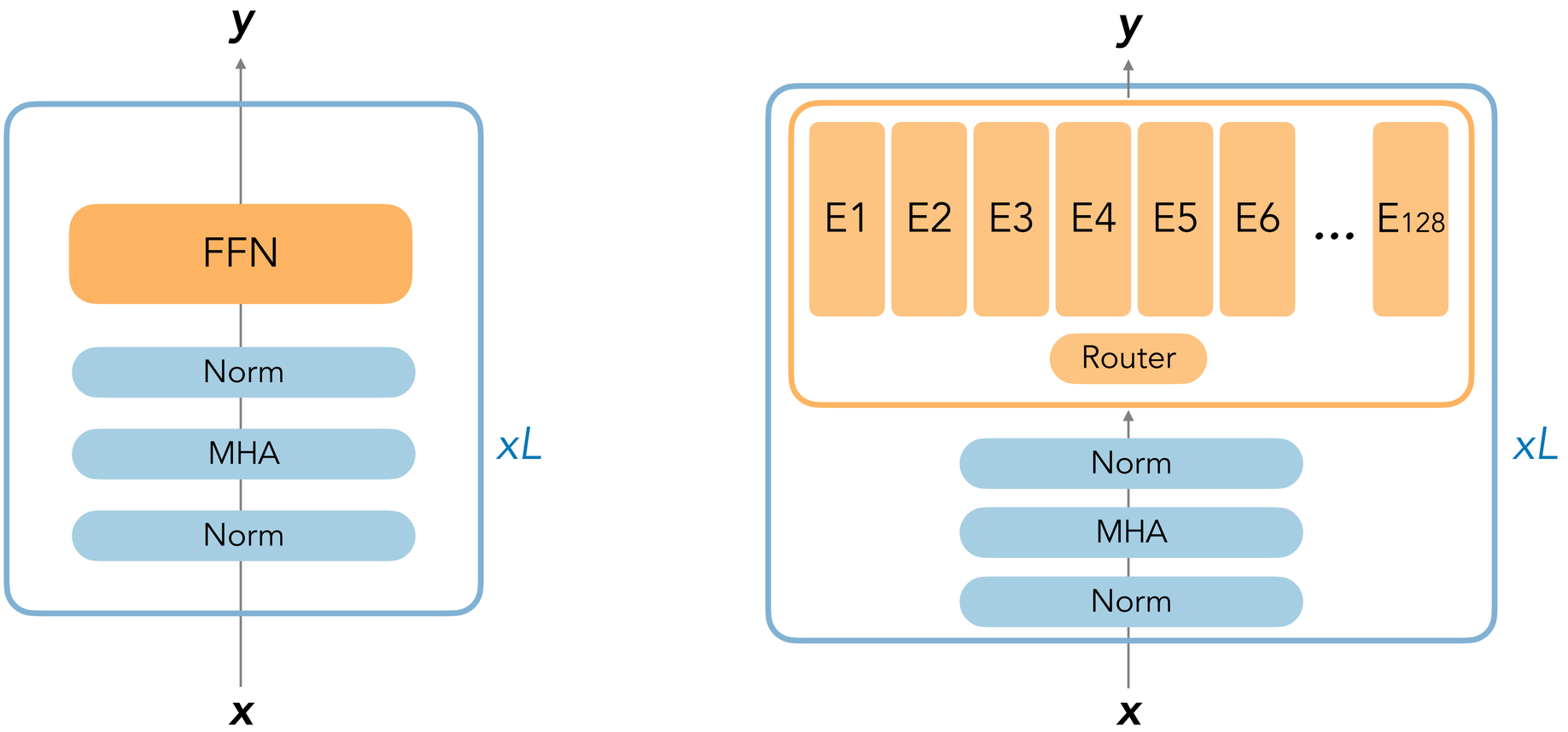




3. Bypassing text parsing

"EMO: Pretraining Mixture of Experts for Emergent Modularity" arxiv.org/abs/2605.06663
+ Upcoming work

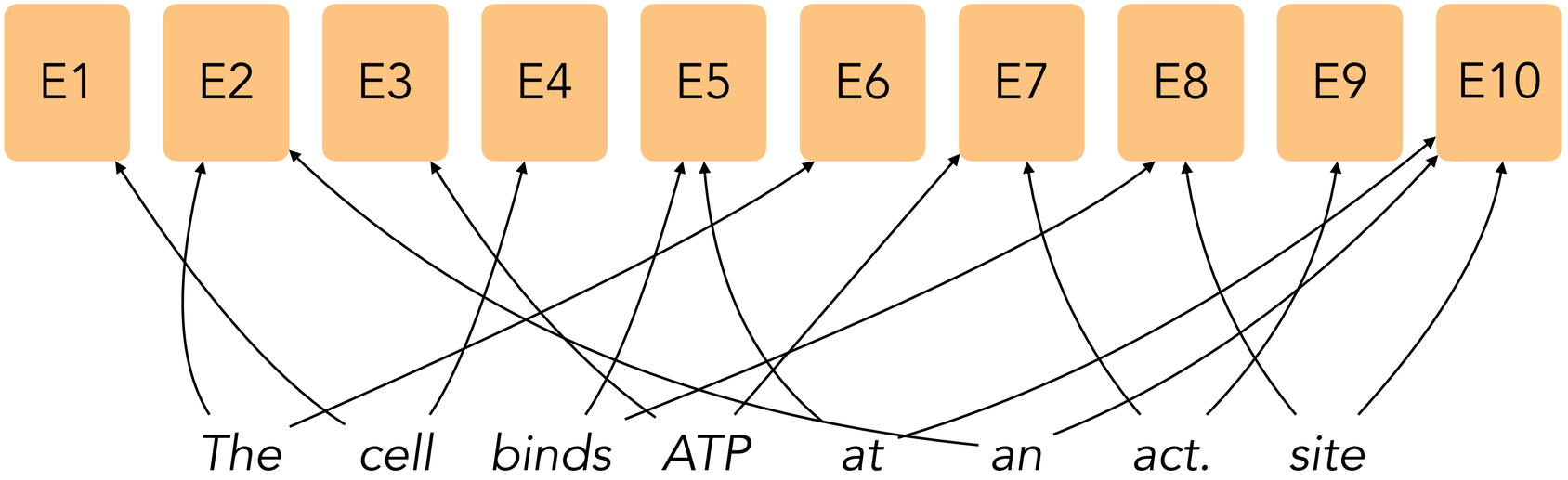
Mixture-of-Experts (MoE) 101



*Residual connection omitted

Mixture-of-Experts (MoE) 101

N=10, K=2



Each token chooses K activated experts from a pool of N

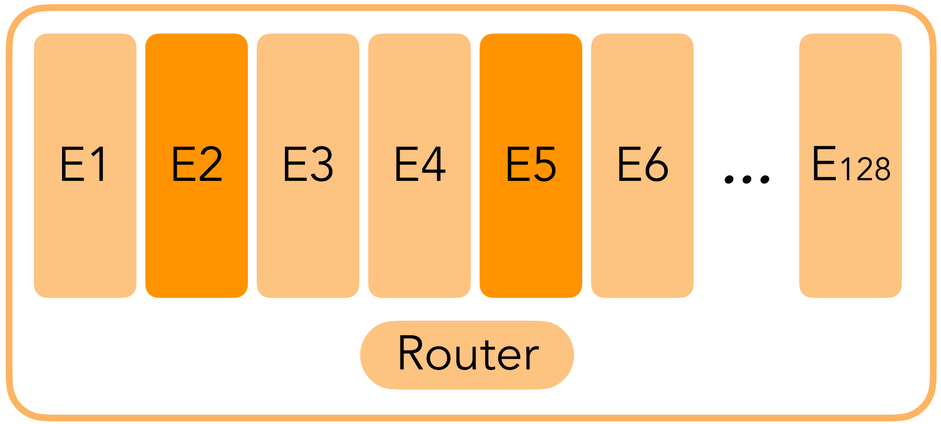
Most of Today's LLMs are MoEs!

		Parameters (Active/Total)	Experts
DeepSeekMoE	Jan 2024	2.8B Active, 16.4B Total	2 Shared + 6 out of 64
OLMoE	Sep 2024	1.3B Active, 6.9B Total	8 out of 64
DeepSeek-V3	Dec 2024	37B Active, 671B Total	1 Shared + 8 out of 256
GPT-OSS	Aug 2025	5.1B Active, 120B Total	4 out of 128 (No shared)
Qwen3-235B-A22B	Aug 2025	22B Active, 235B Total	8 out of 128 (No shared)
Nemotron-3	Dec 2025	3.2B to 3.5B Active, 30B Total	2 Shared + 4 out of 128
GLM-5	Feb 2026	40B to 44B Active, 744B Total	1 Shared + 8 out of 256
DeepSeek-V4 Pro	Apr 2026	49B Active, 1.6T Total	1 Shared + 6 out of 384
Nemotron 3.5 Super VL	Oct 2026	12B Active, 120B Total	1 Shared + 22 out of 512

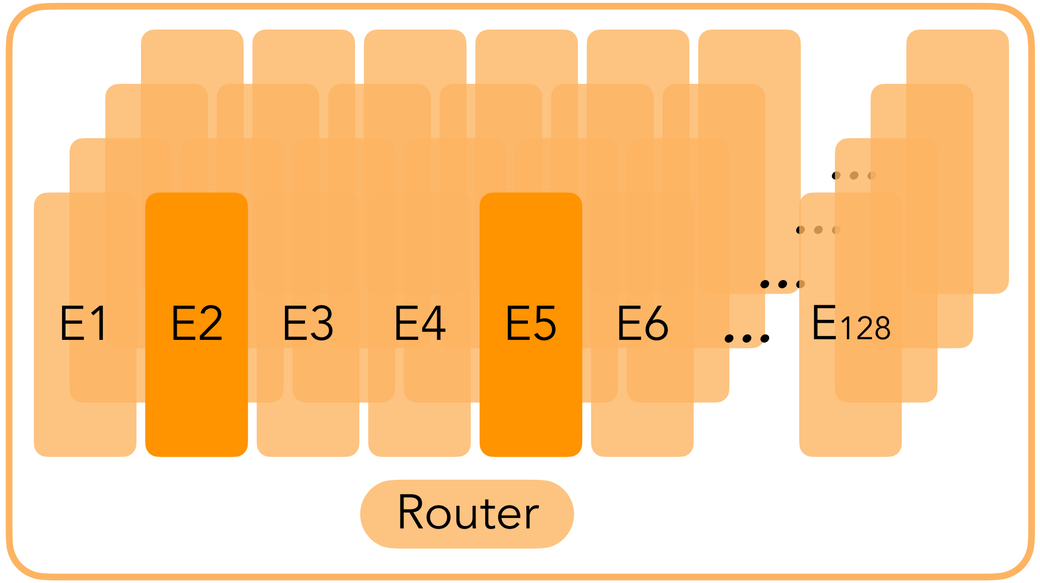
MoEs are getting bigger and sparser!

		Parameters (Active/Total)	Experts	
DeepSeekMoE	Jan 2024	2.8B Active, 16.4B Total	2 Shared + 6 out of 64	~17%
OLMoE	Sep 2024	1.3B Active, 6.9B Total	8 out of 64	
DeepSeek-V3	Dec 2024	37B Active, 671B Total	1 Shared + 8 out of 256	
GPT-OSS	Aug 2025	5.1B Active, 120B Total	4 out of 128 (No shared)	
Qwen3-235B-A22B	Aug 2025	22B Active, 235B Total	8 out of 128 (No shared)	
Nemotron-3	Dec 2025	3.2B to 3.5B Active, 30B Total	2 Shared + 4 out of 128	
GLM-5	Feb 2026	40B to 44B Active, 744B Total	1 Shared + 8 out of 256	
DeepSeek-V4 Pro	Apr 2026	49B Active, 1.6T Total	1 Shared + 6 out of 384	~3%
Nemotron 3.5 Super VL	Oct 2026	12B Active, 120B Total	1 Shared + 22 out of 512	

But why not bigger (by making it sparser)?



But why not bigger (by making it sparser)?

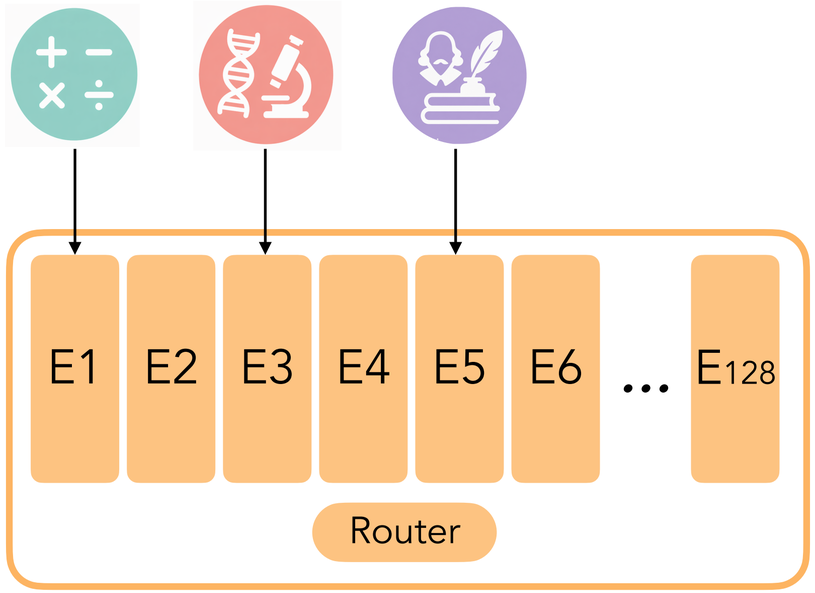


Two challenges

1. Training is more difficult
   (communication overhead, smaller per-expert batches, load-balancing challenges)

2. Inference is memory-intensive
   (weight storage scales with total parameters, making deployment harder on limited hardwares)

Q: Are experts actually “experts”?



What we expect

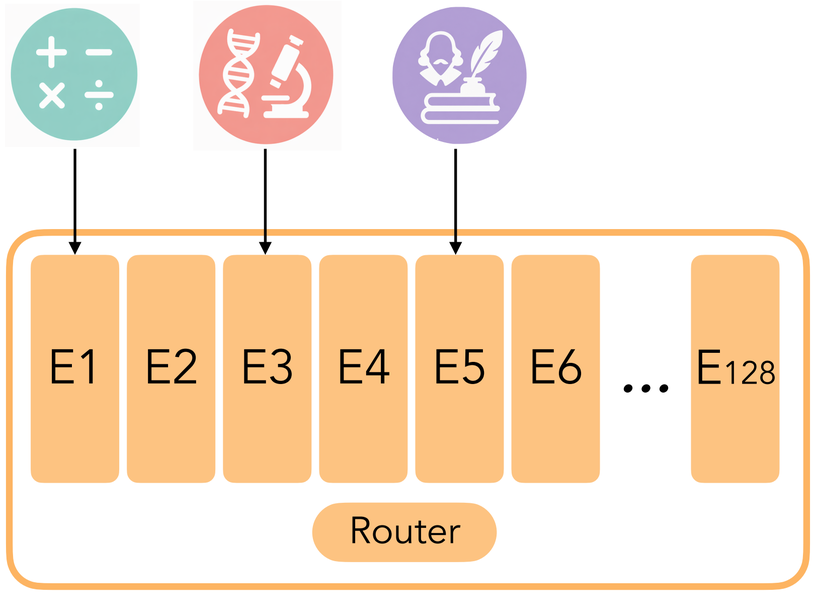
Standard MoE

TOP CLUSTERS · BY TOKEN COUNT

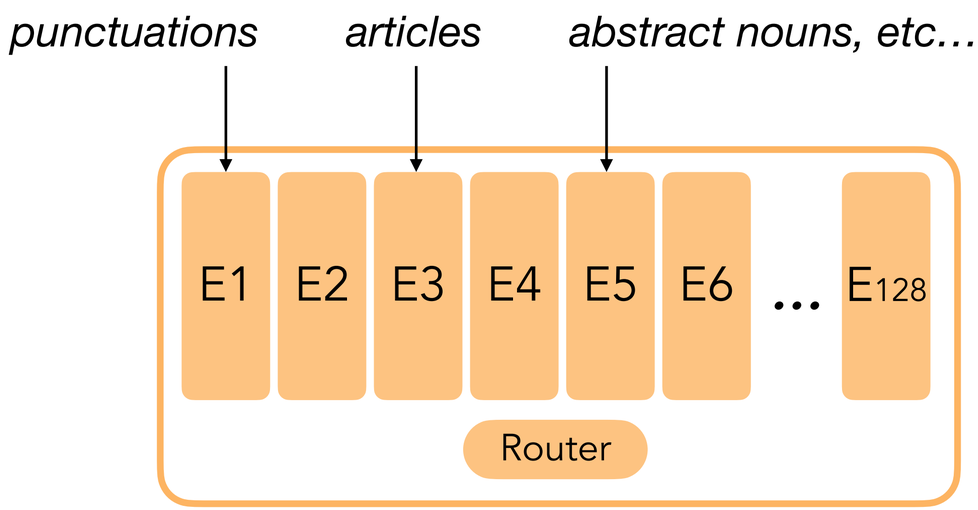
Word-internal subword fragments	6.0%
Prepositions (of/in/for/to/on/with)	5.1%
Abstract nouns & domain topics	4.8%
Section breaks & boilerplate punctua…	4.6%
Web page metadata & forum boilerplate	4.5%
Society, people & institutions	4.2%
Coordinating conjunctions (and/or/but)	4.0%
Proper names & attributions	3.9%
Copula verbs (is/are/was/be)	3.8%
Possessives & definite articles (the…	3.7%
Action verbs & general commentary	3.7%
Physical objects & household items	3.4%
Small numbers, counts & ordinals	3.3%
Prices, years & date strings	3.2%
Topic keywords & title nouns	3.1%
Source code	3.1%
Descriptive adjectives & modifiers	3.0%
Evaluative & reflective prose	2.9%

Top 18 of 32 clusters

Q: Are experts actually "experts"?

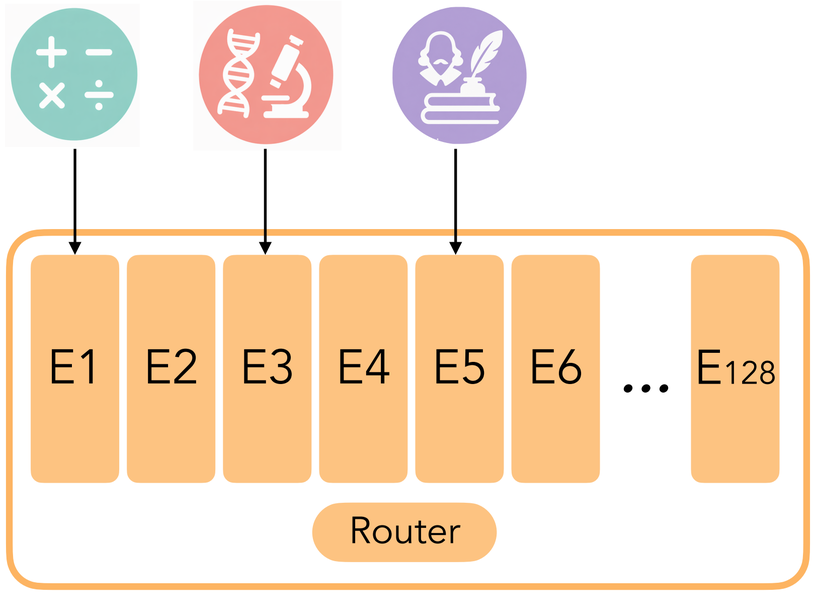


What we expect

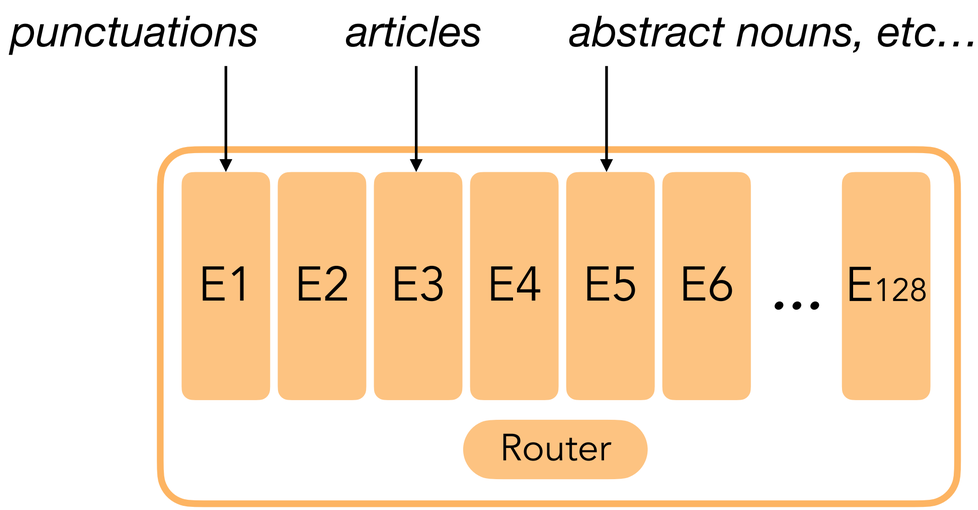


What actually happens

Q: Can we make experts specialize at domain/capability level?

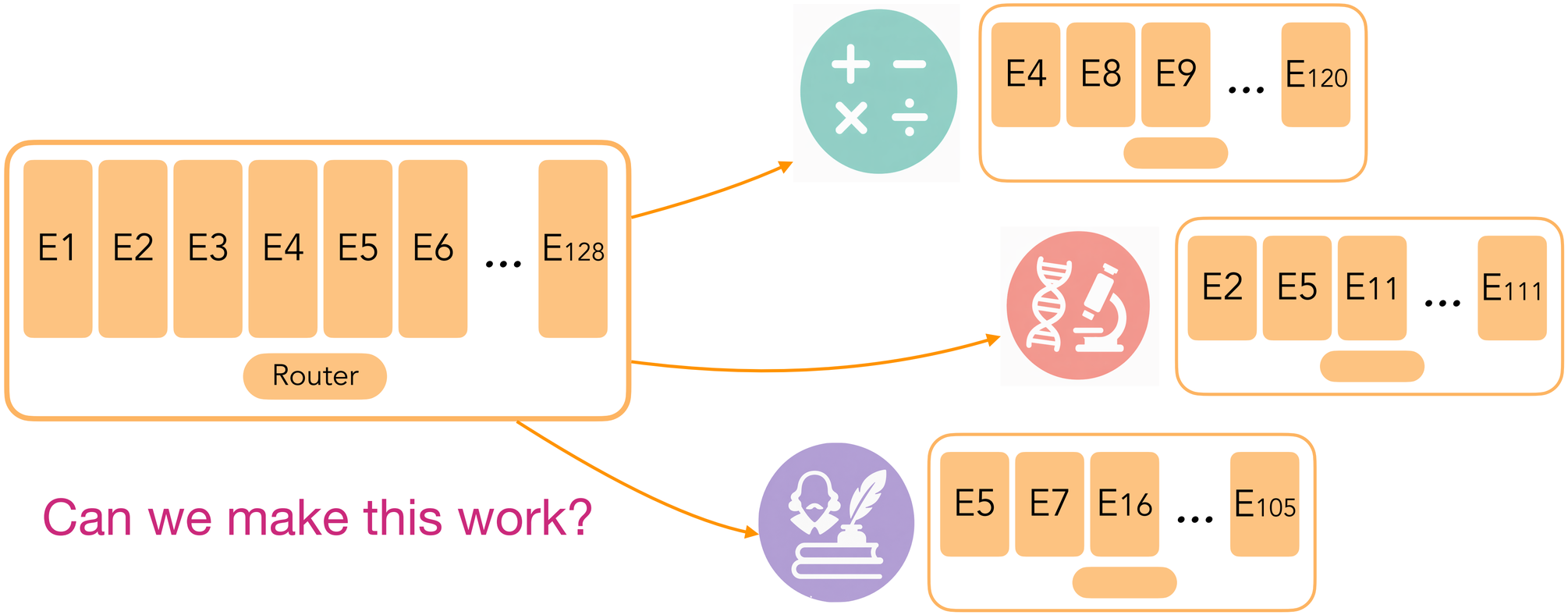


Can we make this work?

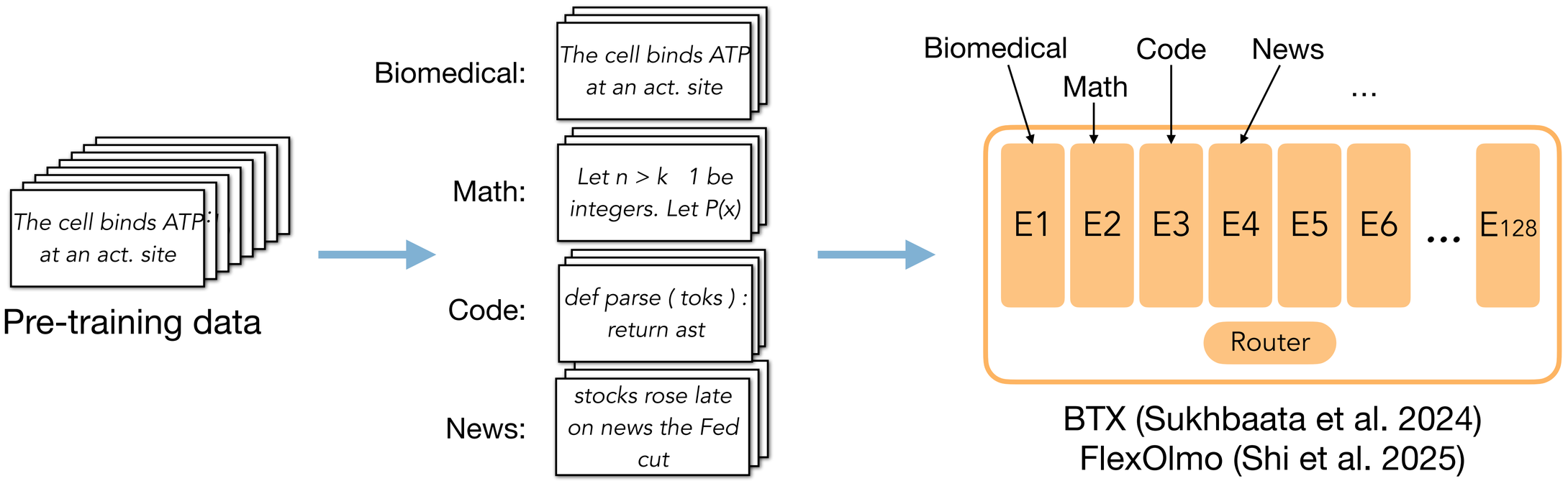


What actually happens

Q: Can we make experts specialize at domain/capability level?

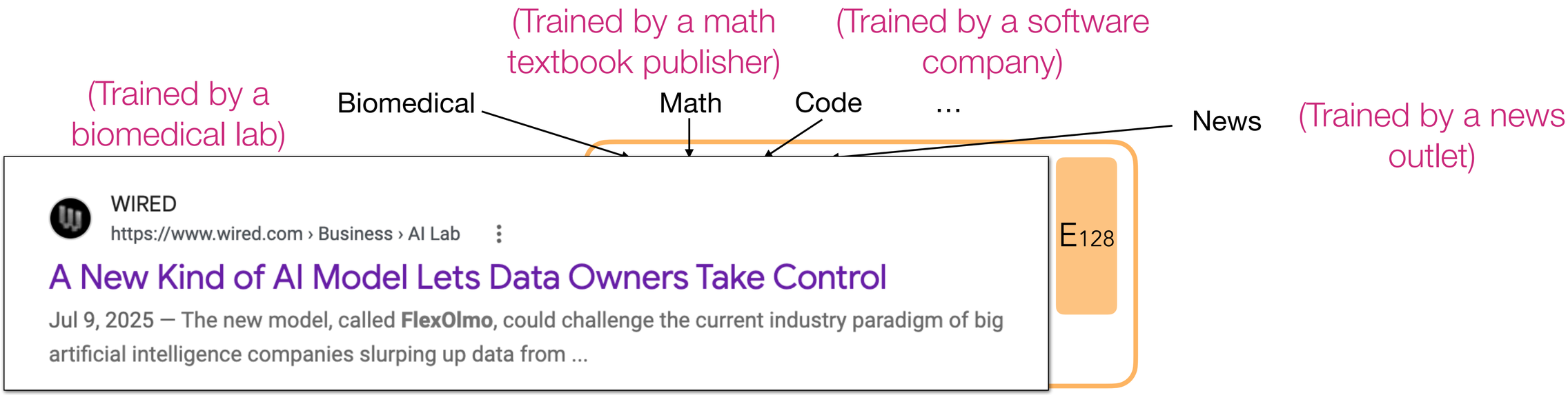


Prior work: assign domain labels



BTX (Sukhbaata et al. 2024)
FlexOlmo (Shi et al. 2025)

Prior work: assign domain labels



FlexOlmo (Shi et al. 2025)

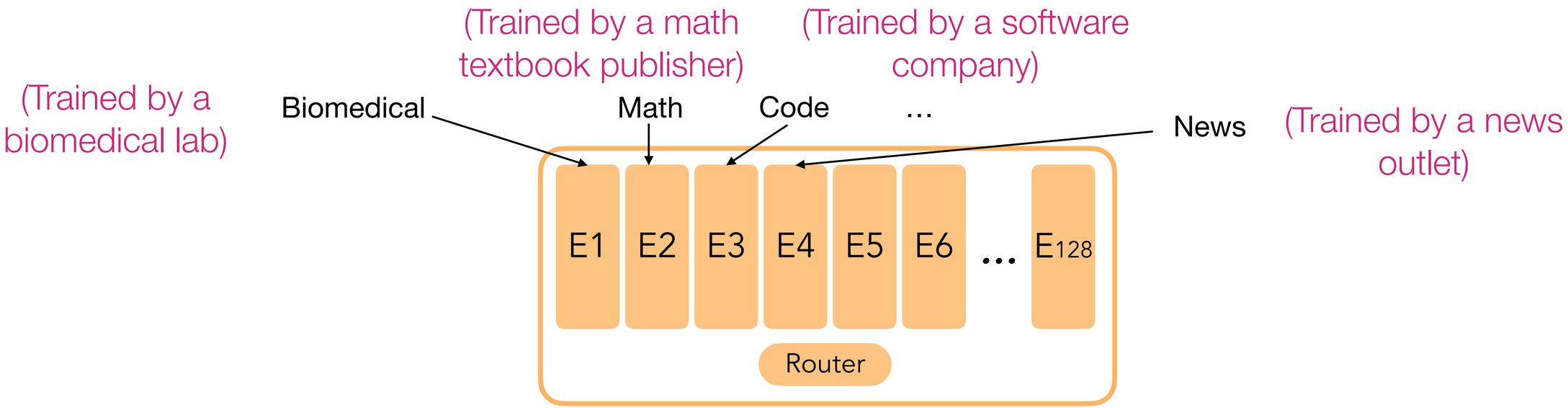




w/ Weijia Shi, Akshita Bhagia, Kevin Farhat + Other collaborators

- Domain-level specialization → solves inference memory problem
- Decentralized training → solves training synchronization problem
- Enabling use of private, proprietary data → solves data problem

Prior work: assign domain labels

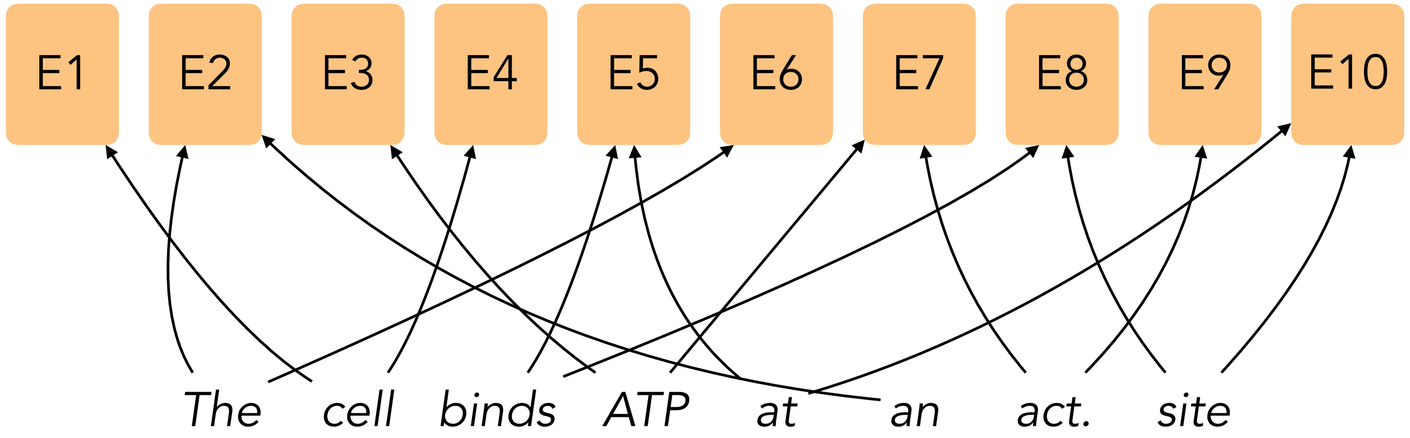


FlexOlmo (Shi et al. 2025)

- Doesn't really scale: domain labels are ambiguous and difficult to obtain at scale
- Imposing human-defined structure is not good!

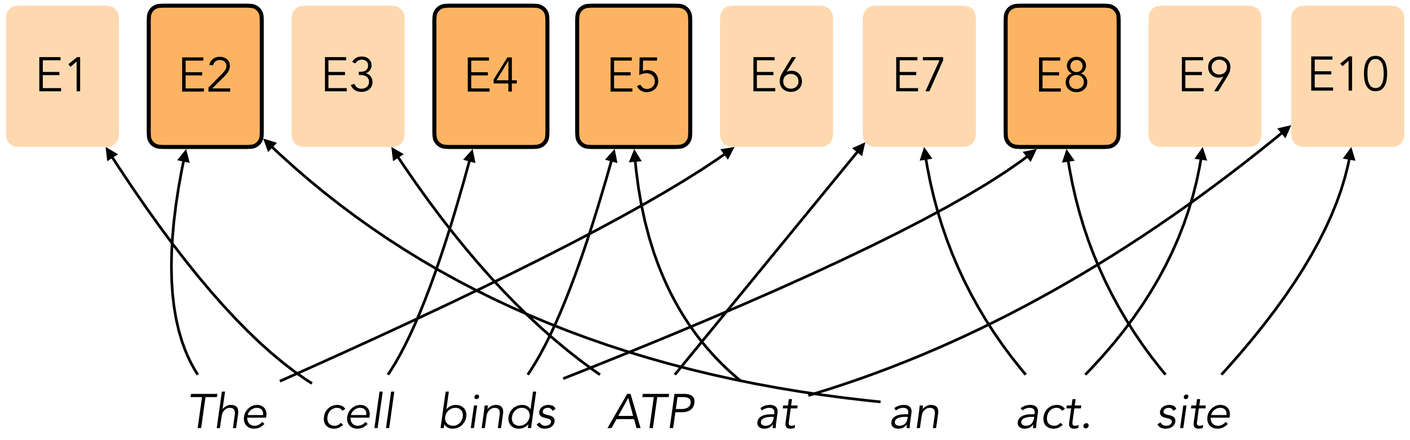
Our work: EMO

- Intuition: Tokens from the same domain should activate a similar subset of experts
- Tokens within the same document are likely to share a domain
- Let's encourage **tokens from the same document** to **activate a shared pool of experts!**



Our work: EMO

- Intuition: Tokens from the same domain should activate a similar subset of experts
- Tokens within the same document are likely to share a domain
- Let's encourage **tokens from the same document** to **activate a shared pool of experts!**

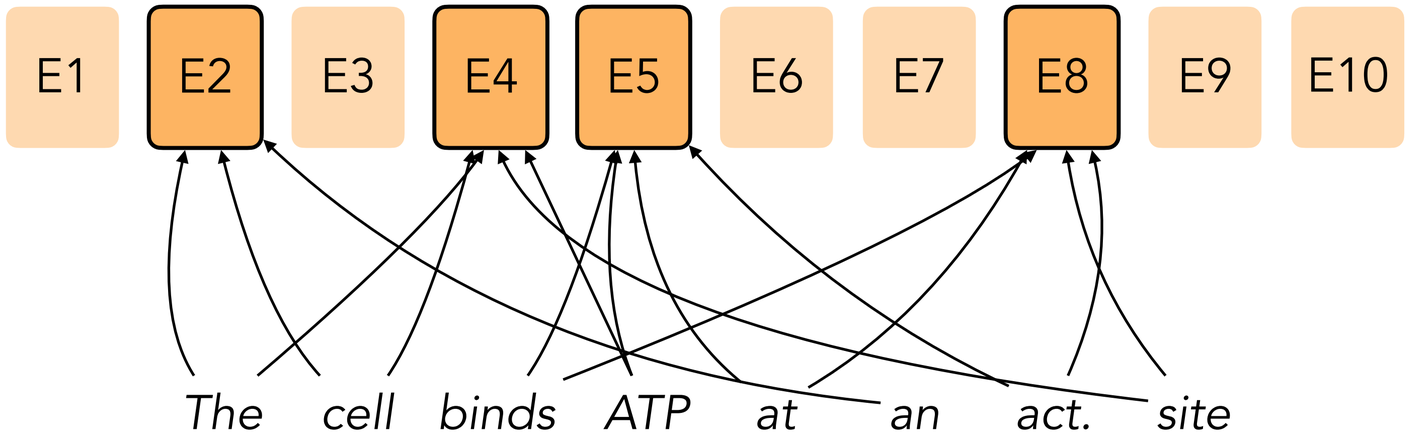


Step 1: Identify a pool of D experts

(We sample D ∈ [K, D] for every document)

Our work: EMO

- Intuition: Tokens from the same domain should activate a similar subset of experts
- Tokens within the same document are likely to share a domain
- Let's encourage **tokens from the same document** to **activate a shared pool of experts!**

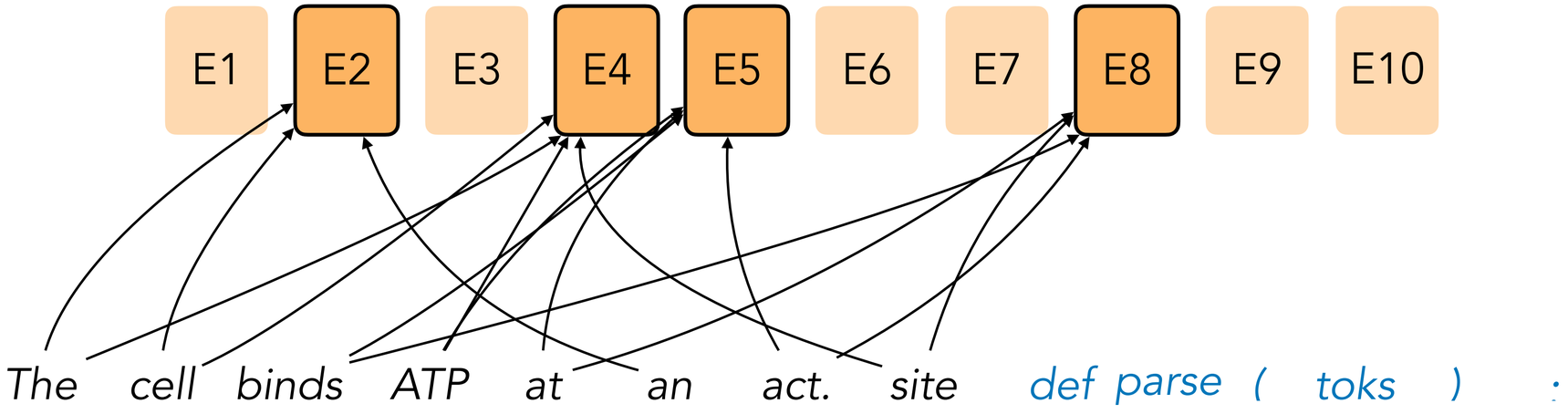


Step 2: Each token chooses K activated experts from a pool of D

(We sample D ∈ [K, D] for every document)

Our work: EMO

- Intuition: Tokens from the same domain should activate a similar subset of experts
- Tokens within the same document are likely to share a domain
- Let's encourage **tokens from the same document** to **activate a shared pool of experts!**

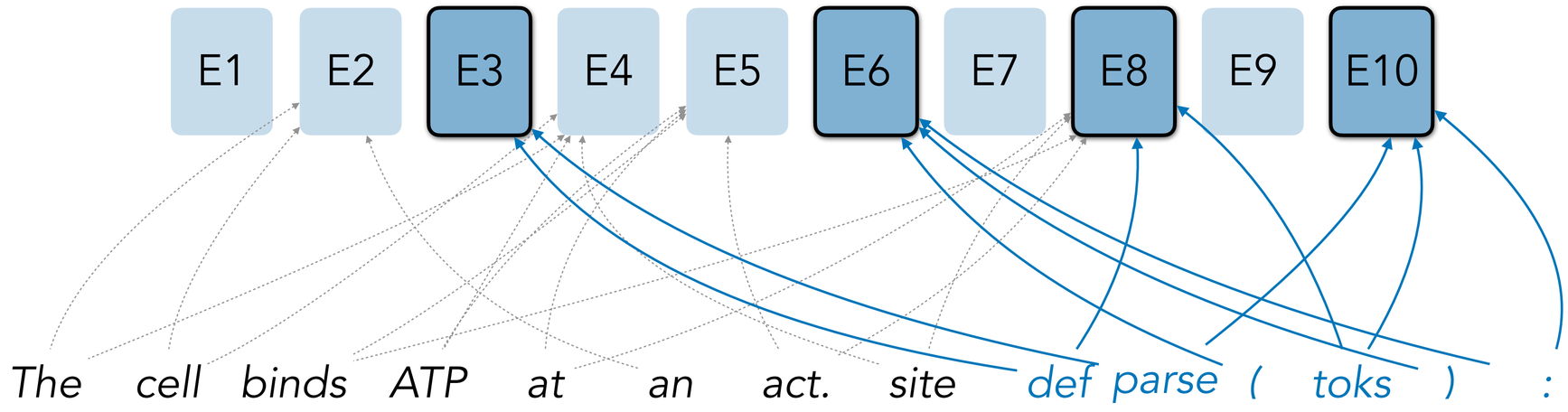


Step 3: Each document within a sequence has its own pool of D

(We sample D ∈ [K, D] for every document)

Our work: EMO

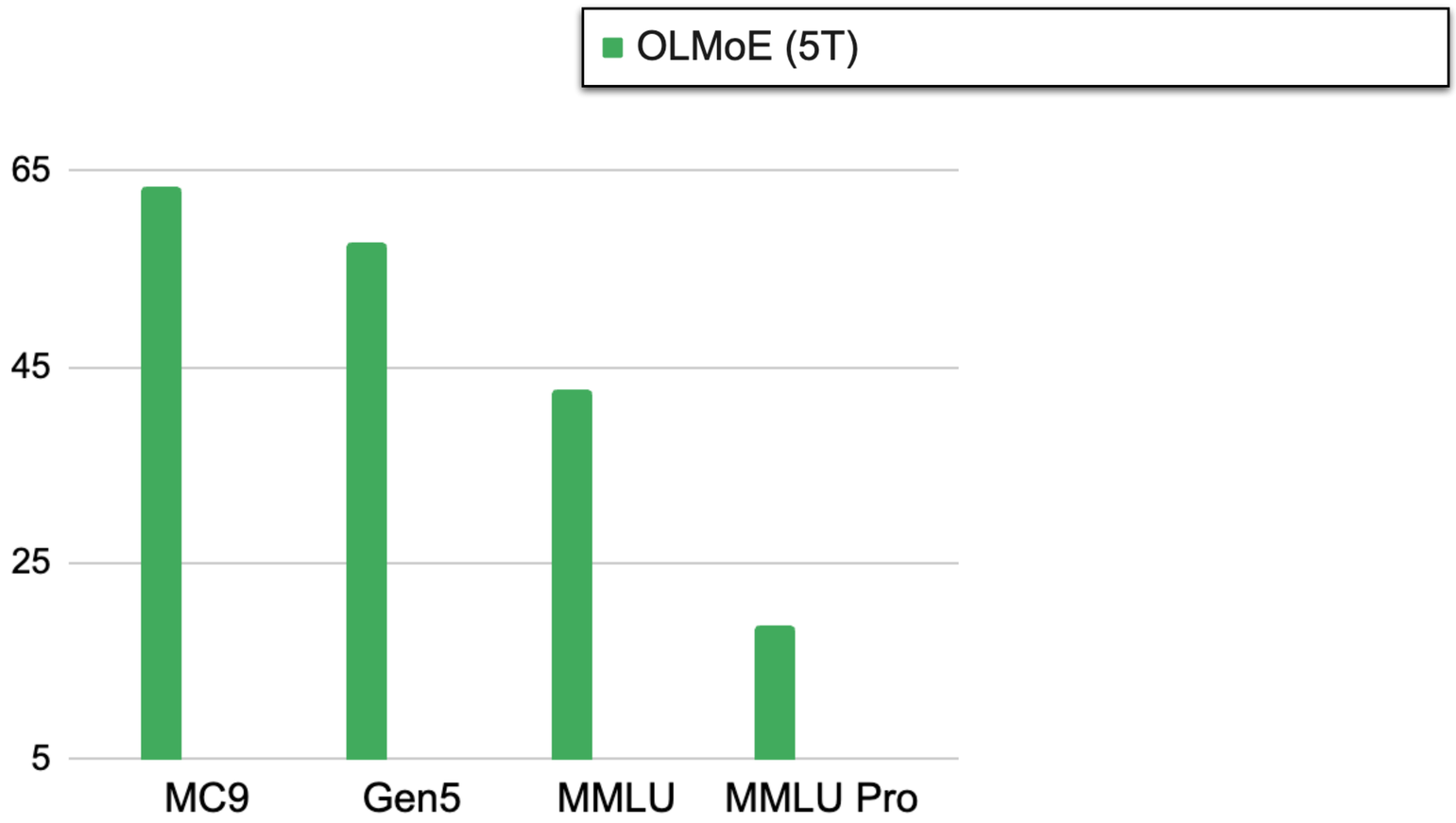
- Intuition: Tokens from the same domain should activate a similar subset of experts
- Tokens within the same document are likely to share a domain
- Let's encourage **tokens from the same document** to **activate a shared pool of experts!**



Step 3: Each document within a sequence has its own pool of D

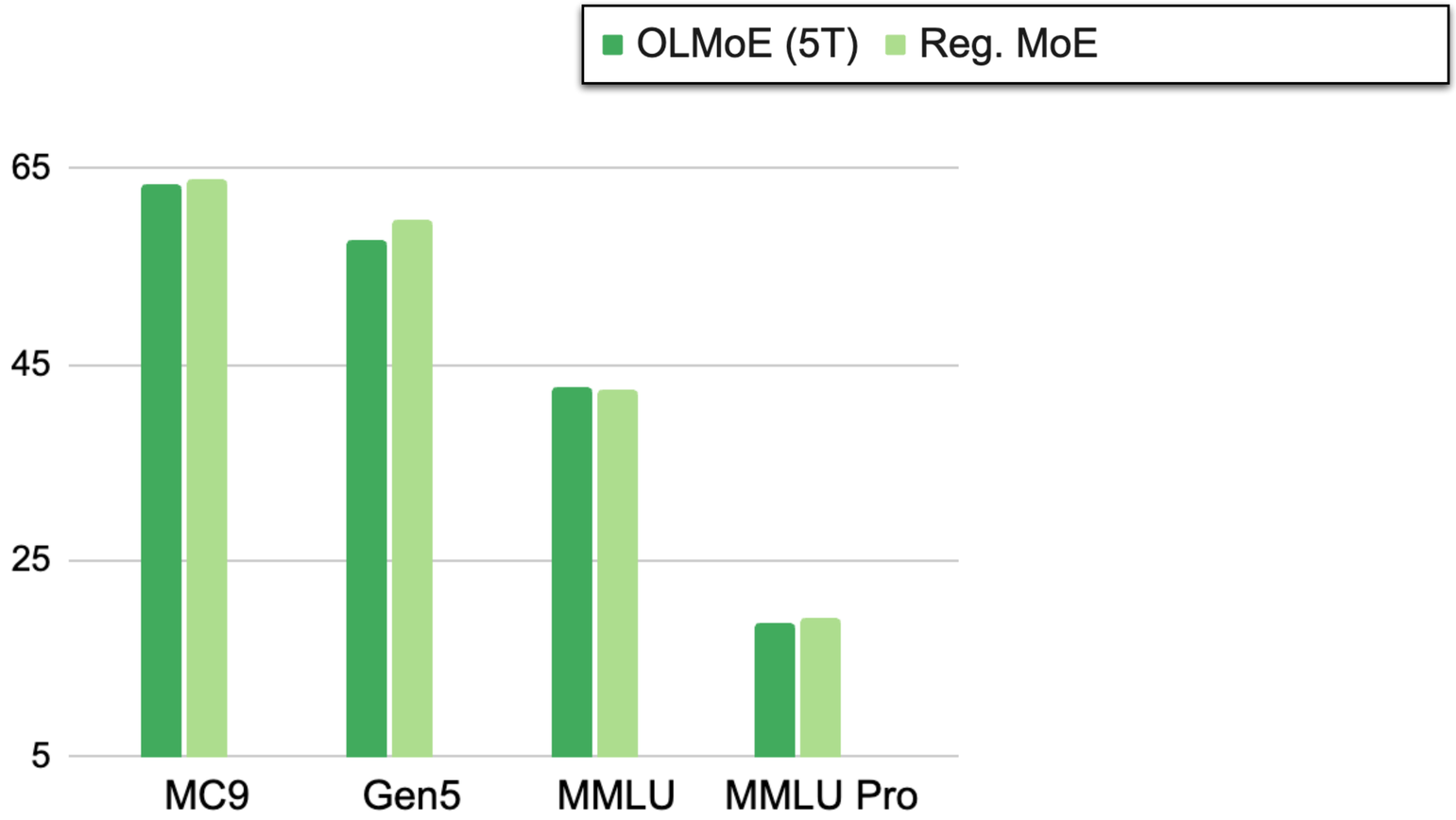
(We sample D ∈ [K, D] for every document)

Results 1: Overall performance



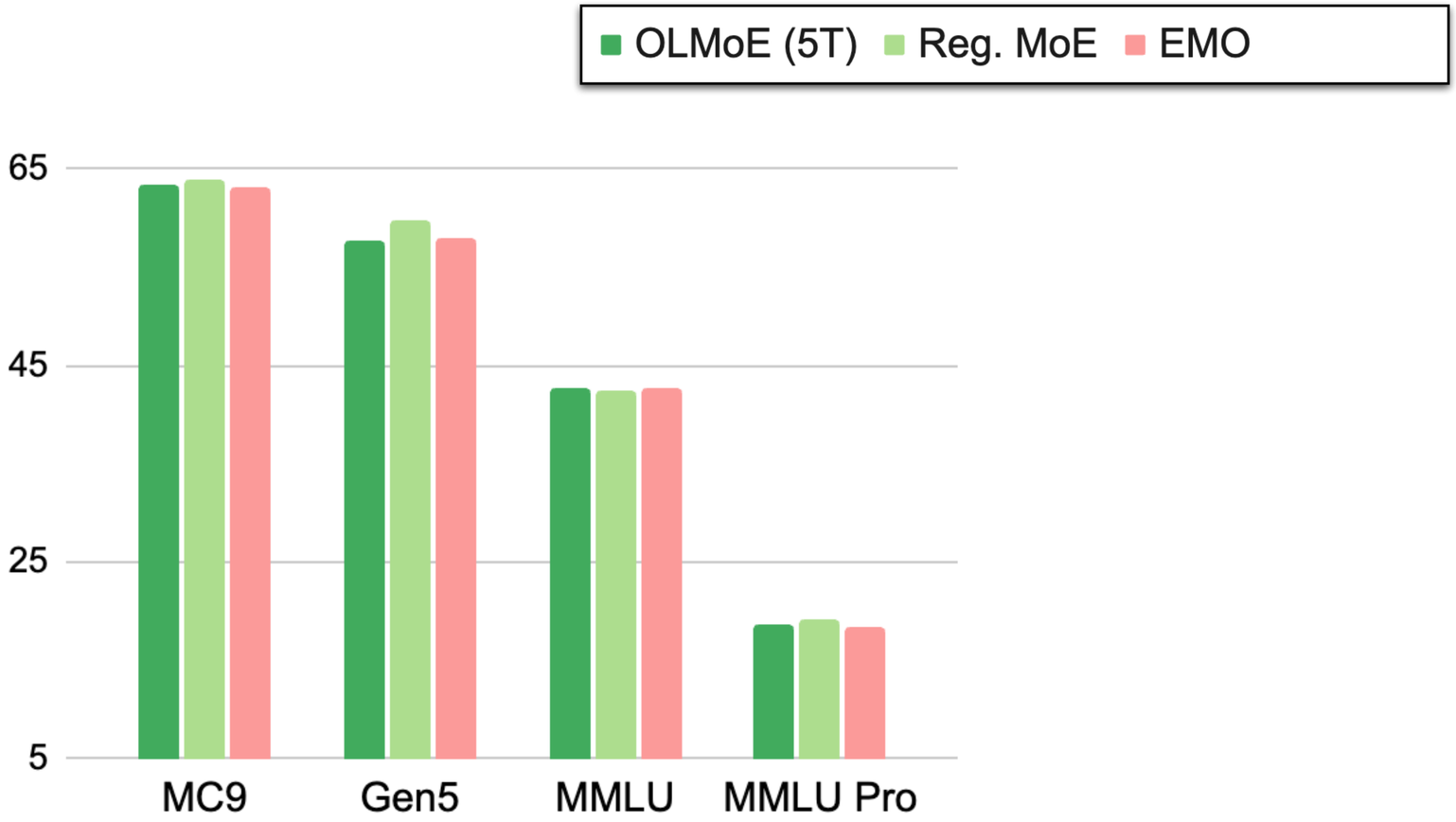
Trained for 1T tokens except for OLMoE (5T)

Results 1: Overall performance



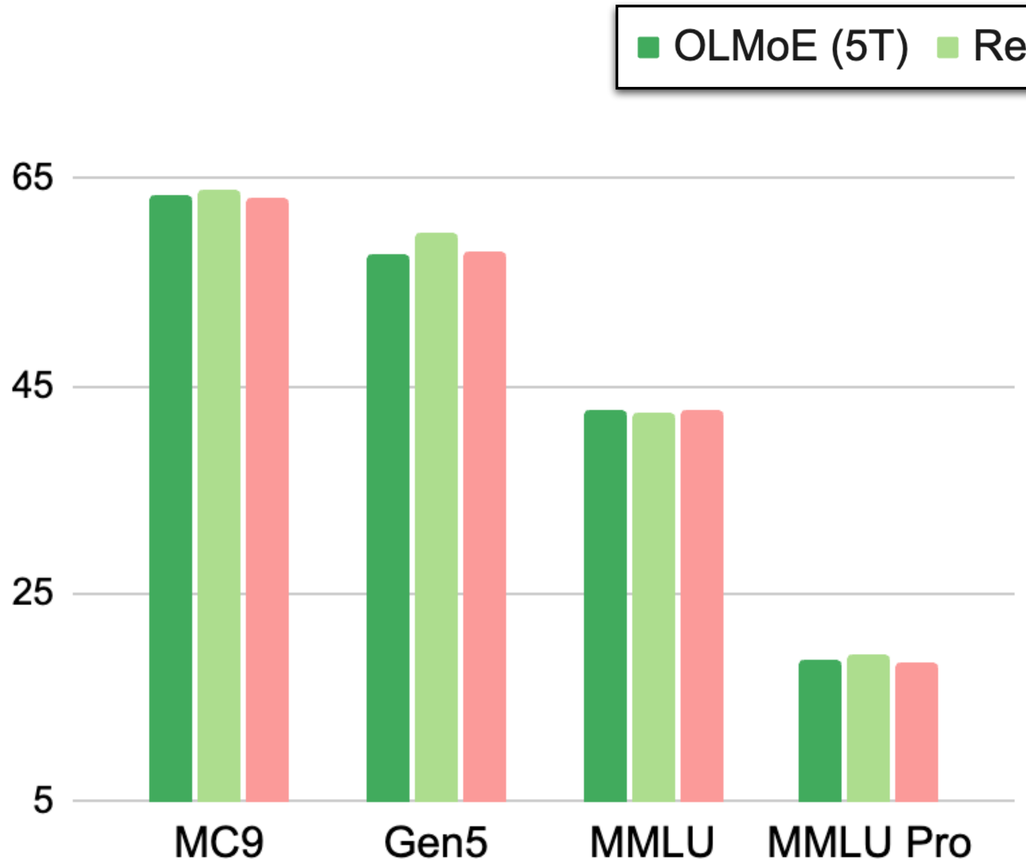
Trained for 1T tokens except for OLMoE (5T)

Results 1: Overall performance

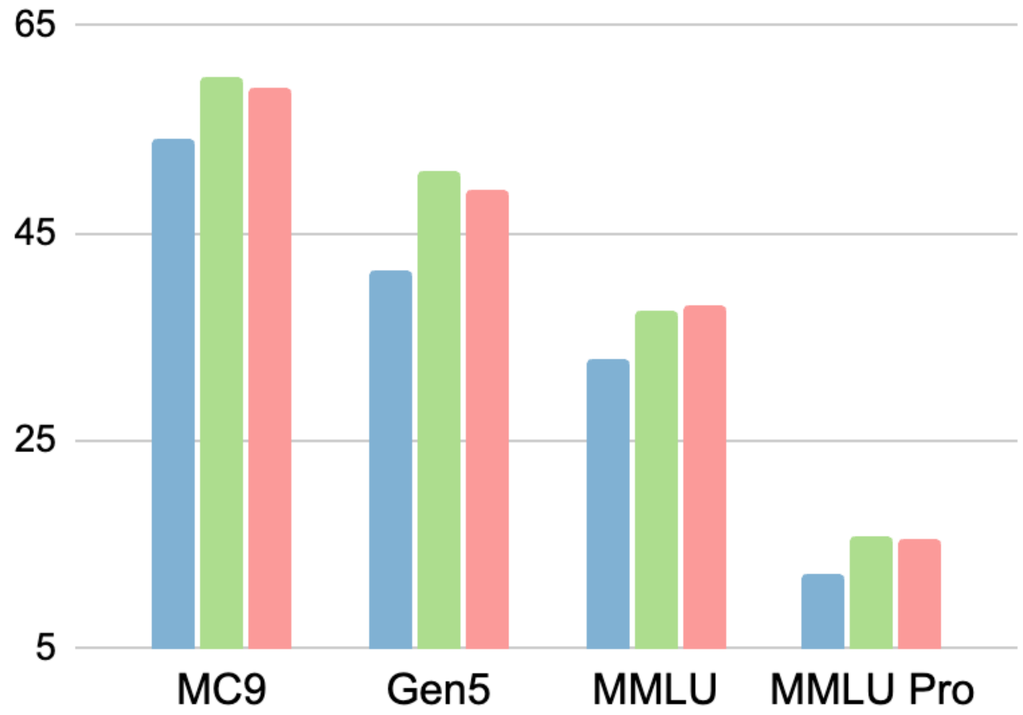


Trained for 1T tokens except for OLMoE (5T)

Results 1: Overall performance

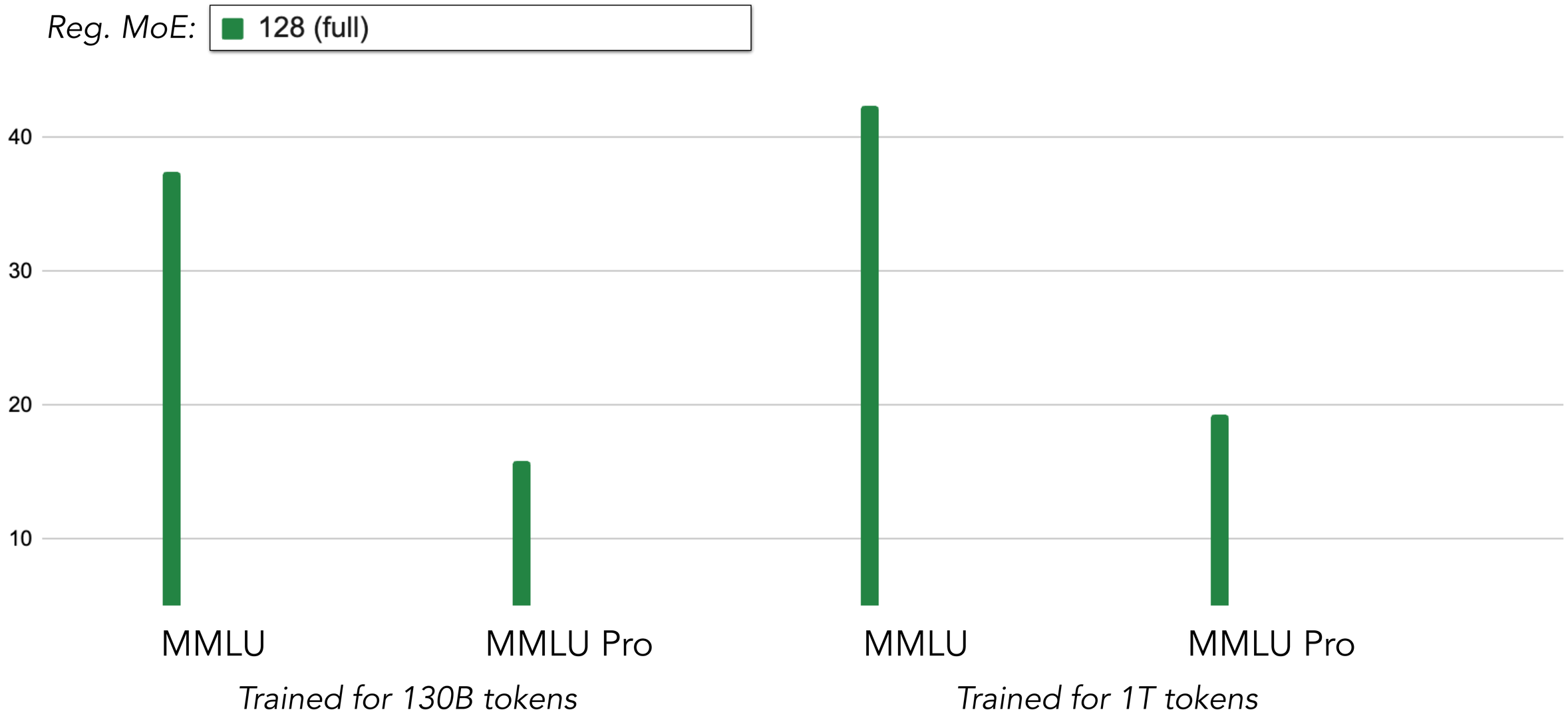


Trained for 1T tokens except for OLMoE (5T)

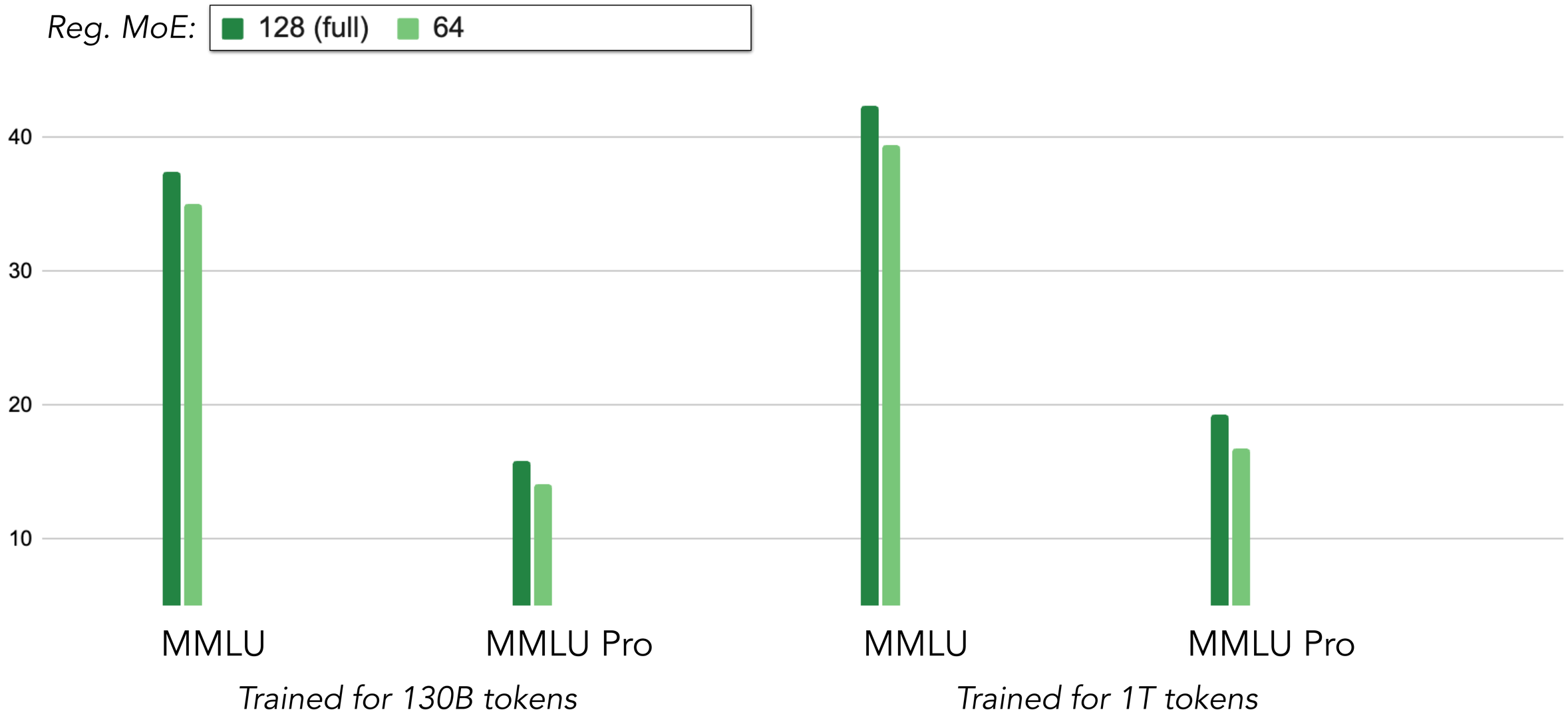


Trained for 130B tokens

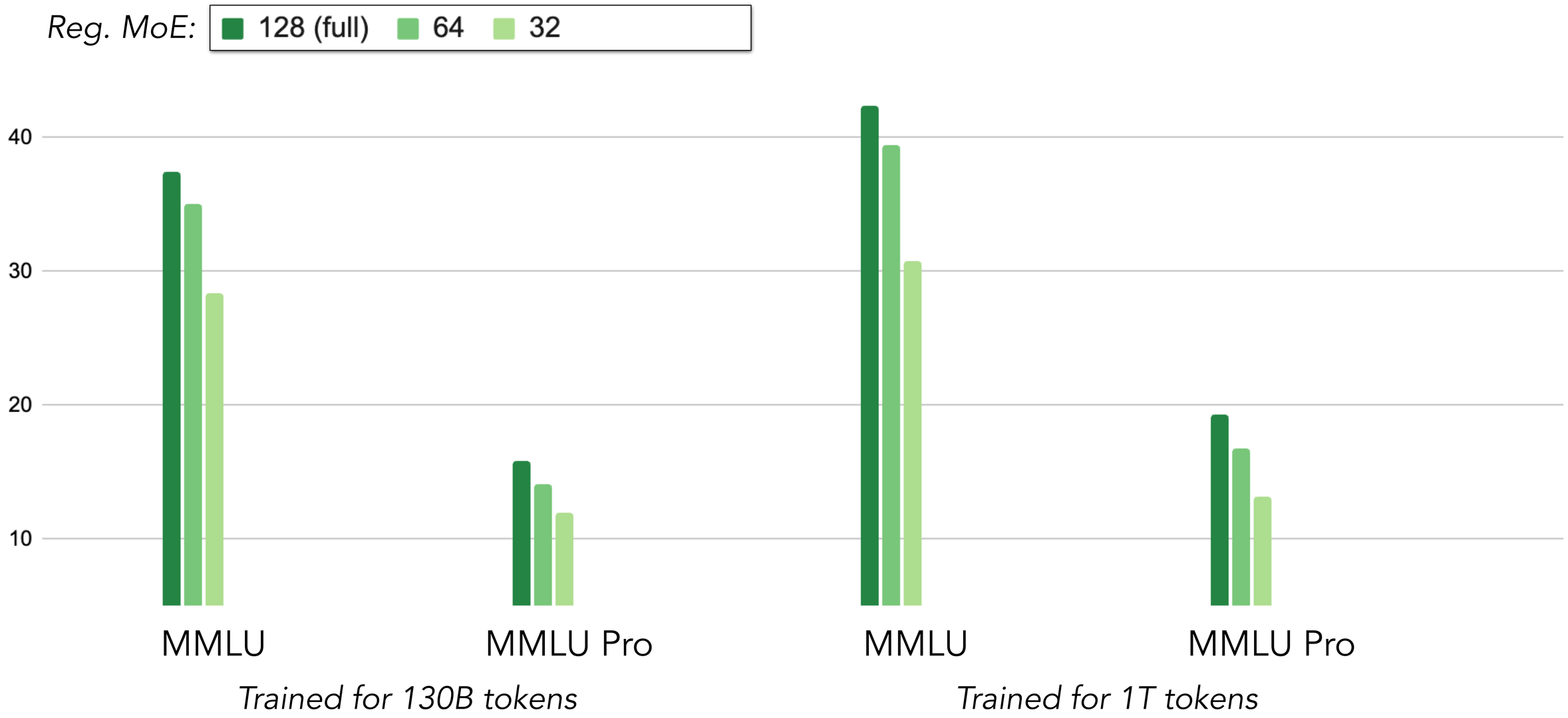
Results 2: Selective expert use



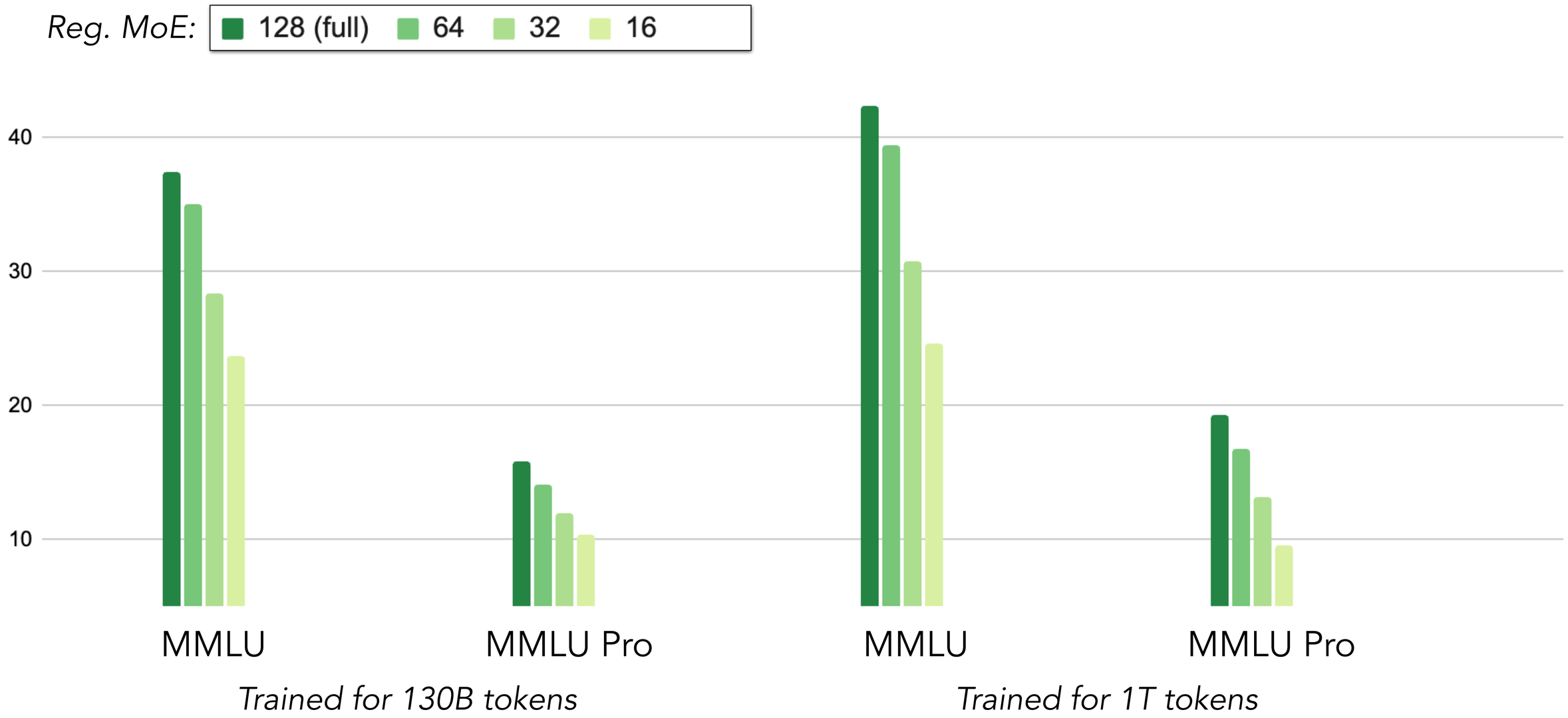
Results 2: Selective expert use



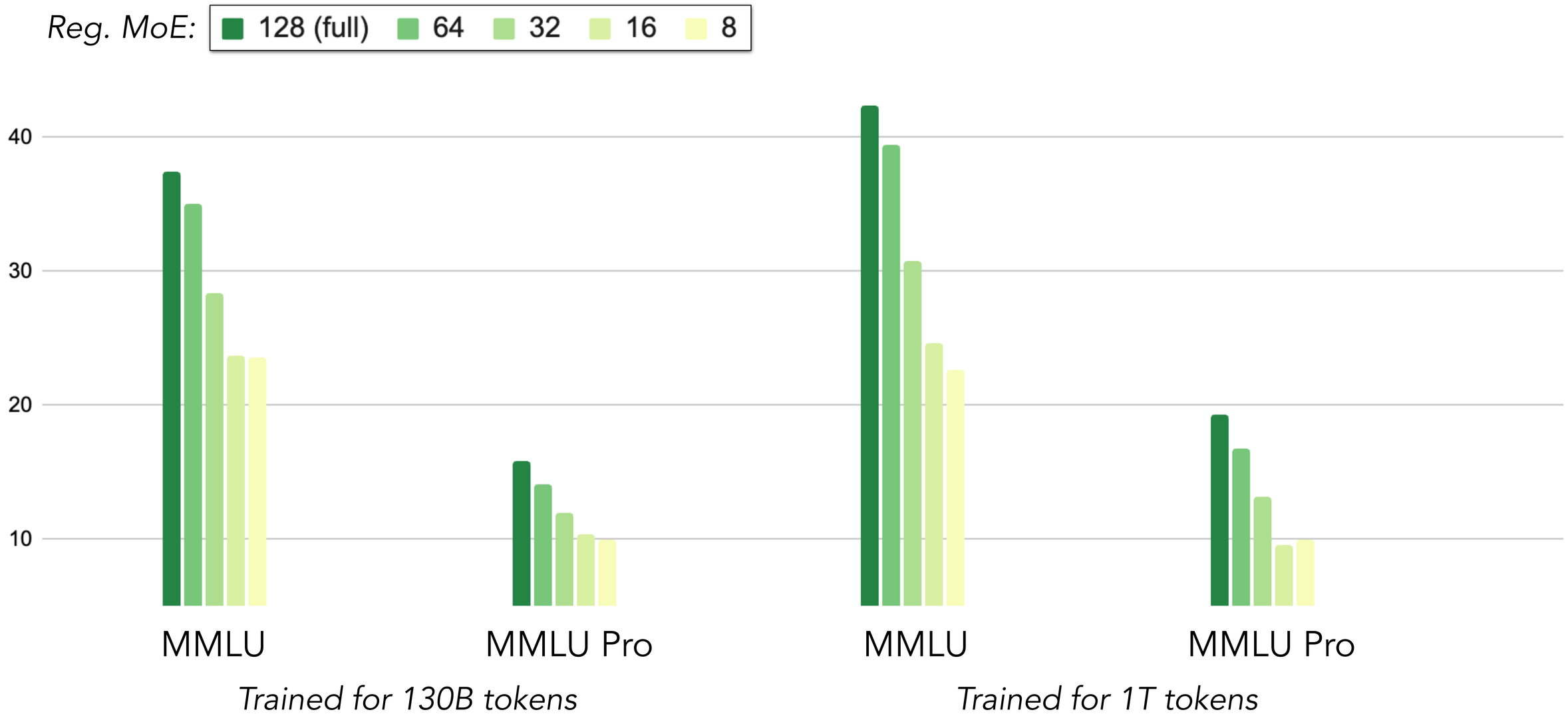
Results 2: Selective expert use



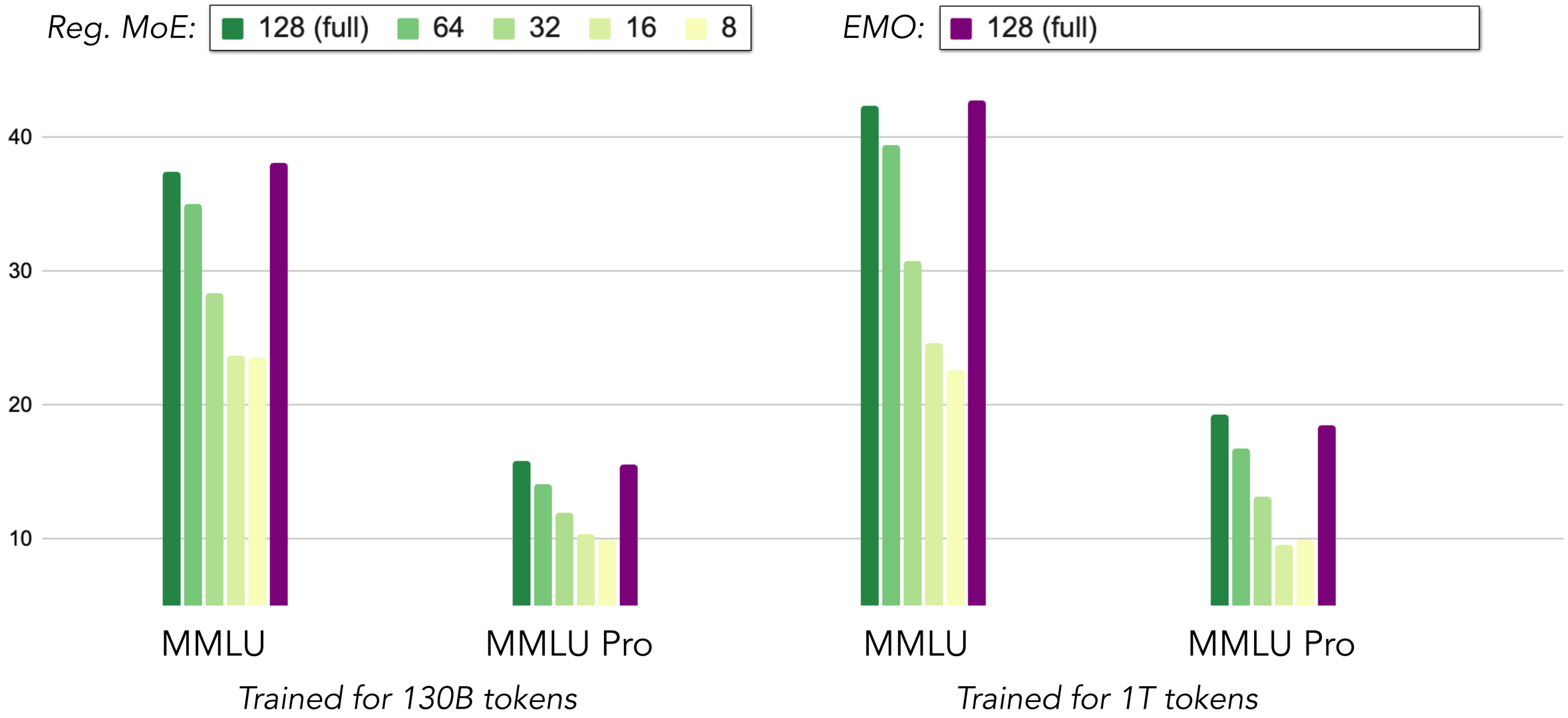
Results 2: Selective expert use



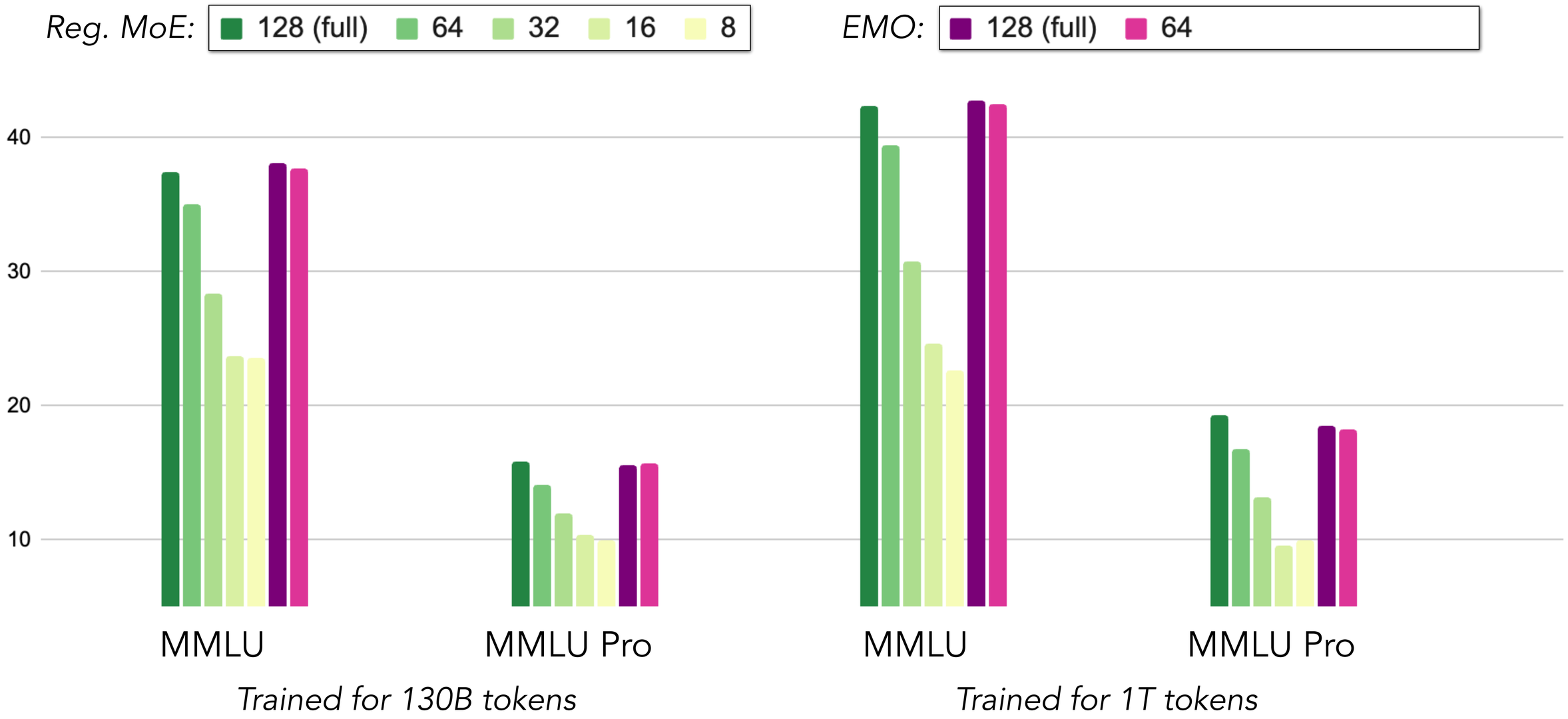
Results 2: Selective expert use



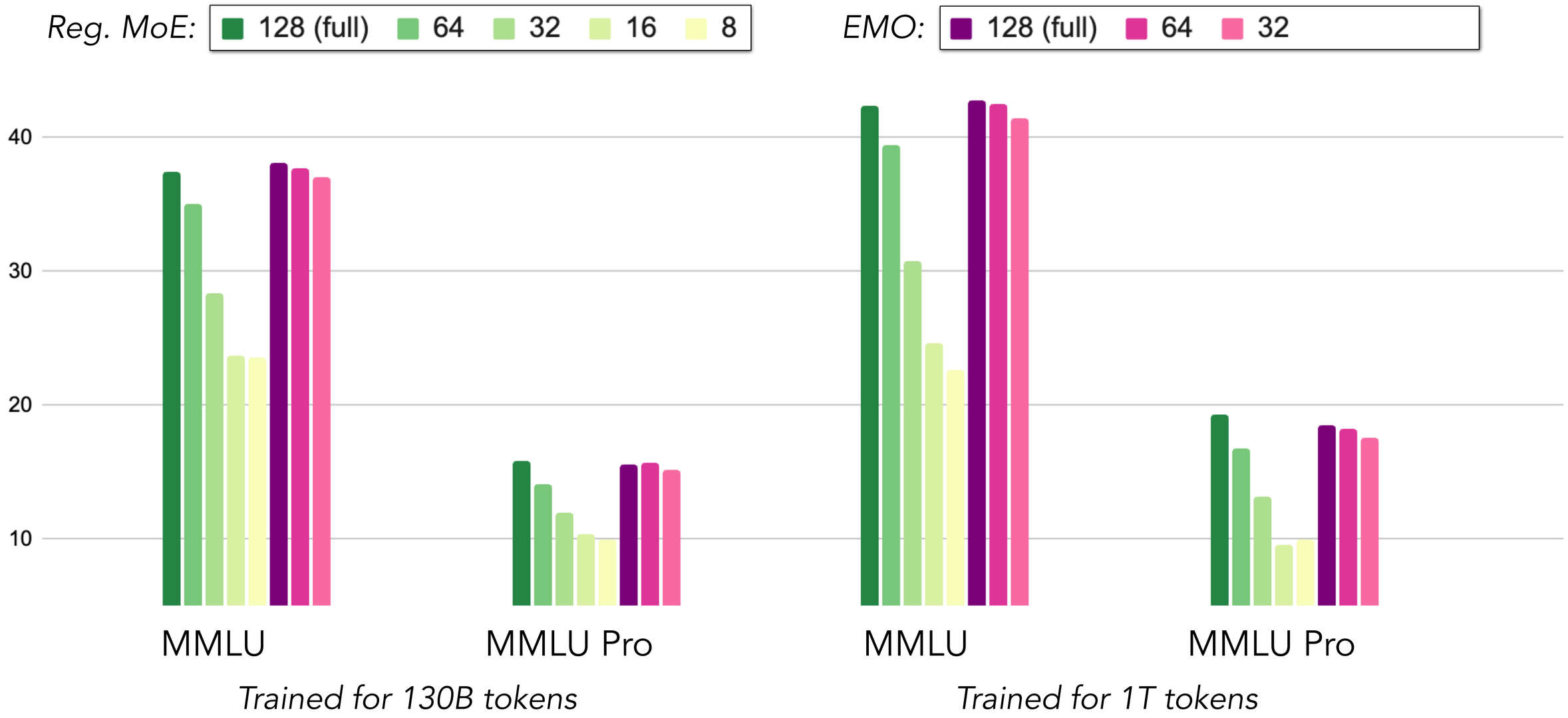
Results 2: Selective expert use



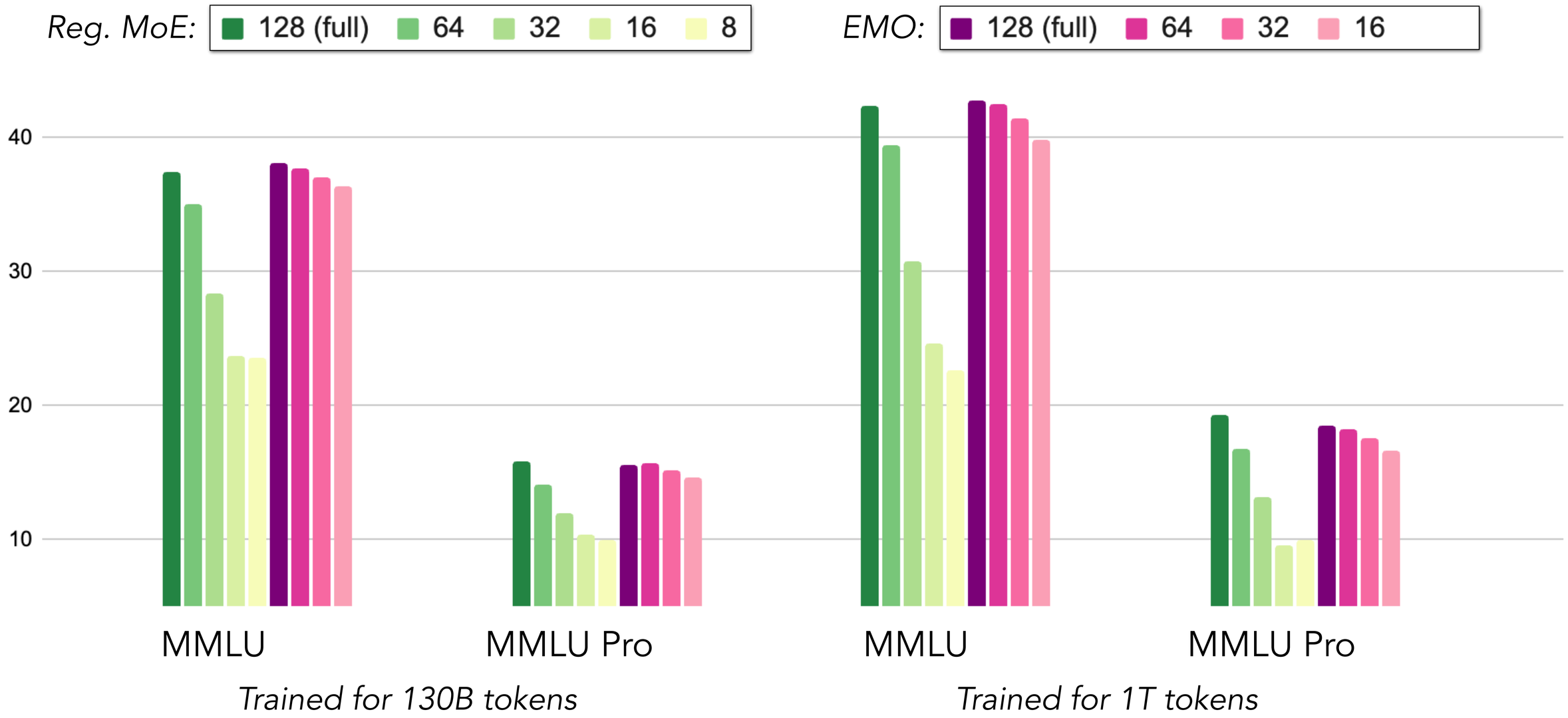
Results 2: Selective expert use



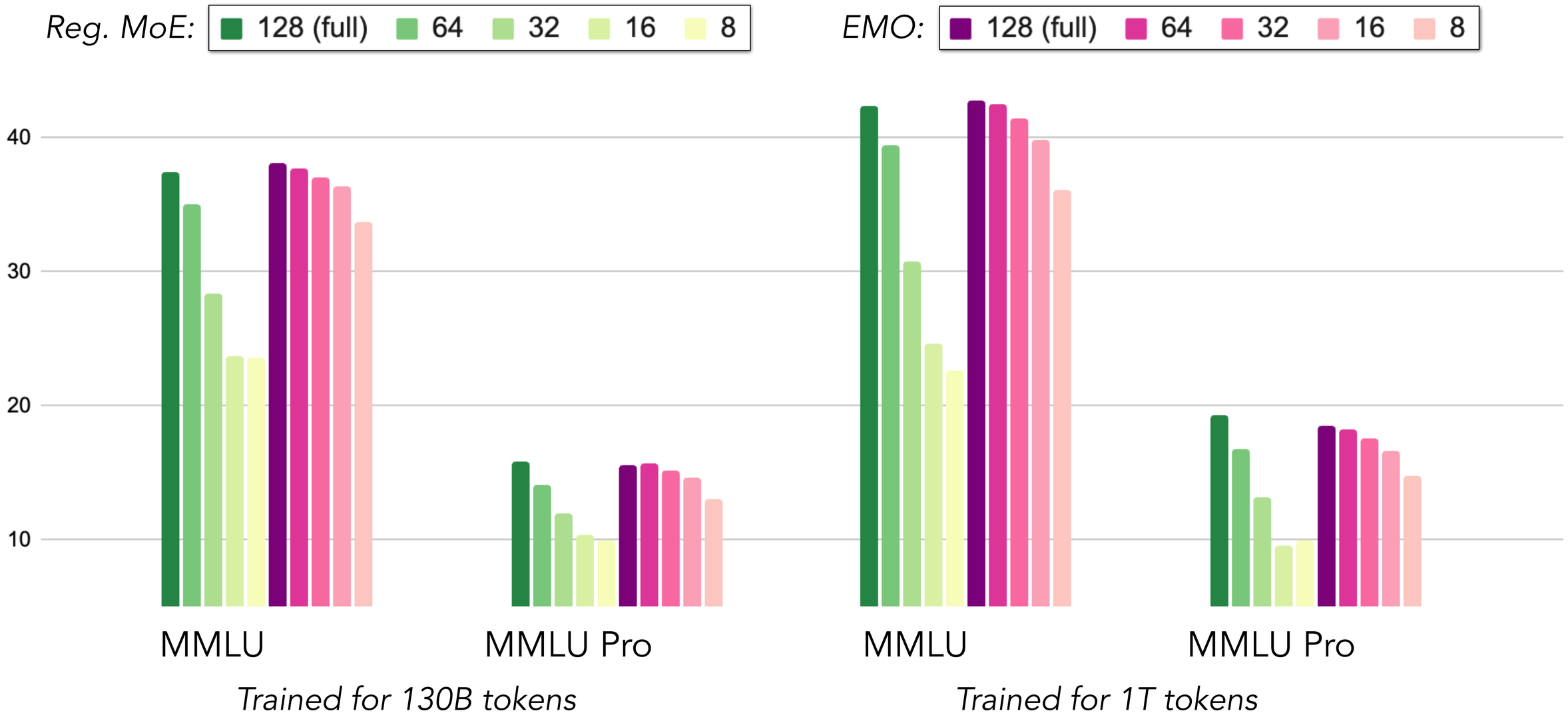
Results 2: Selective expert use



Results 2: Selective expert use



Results 2: Selective expert use



Specialization emerges naturally (!)

Standard MoE

TOP CLUSTERS · BY TOKEN COUNT

Word-internal subword fragments	6.0%
Prepositions (of/in/for/to/on/with)	5.1%
Abstract nouns & domain topics	4.8%
Section breaks & boilerplate punctua…	4.6%
Web page metadata & forum boilerplate	4.5%
Society, people & institutions	4.2%
Coordinating conjunctions (and/or/but)	4.0%
Proper names & attributions	3.9%
Copula verbs (is/are/was/be)	3.8%
Possessives & definite articles (the…	3.7%
Action verbs & general commentary	3.7%
Physical objects & household items	3.4%
Small numbers, counts & ordinals	3.3%
Prices, years & date strings	3.2%
Topic keywords & title nouns	3.1%
Source code	3.1%
Descriptive adjectives & modifiers	3.0%
Evaluative & reflective prose	2.9%

Top 18 of 32 clusters

EMO

TOP CLUSTERS · BY TOKEN COUNT

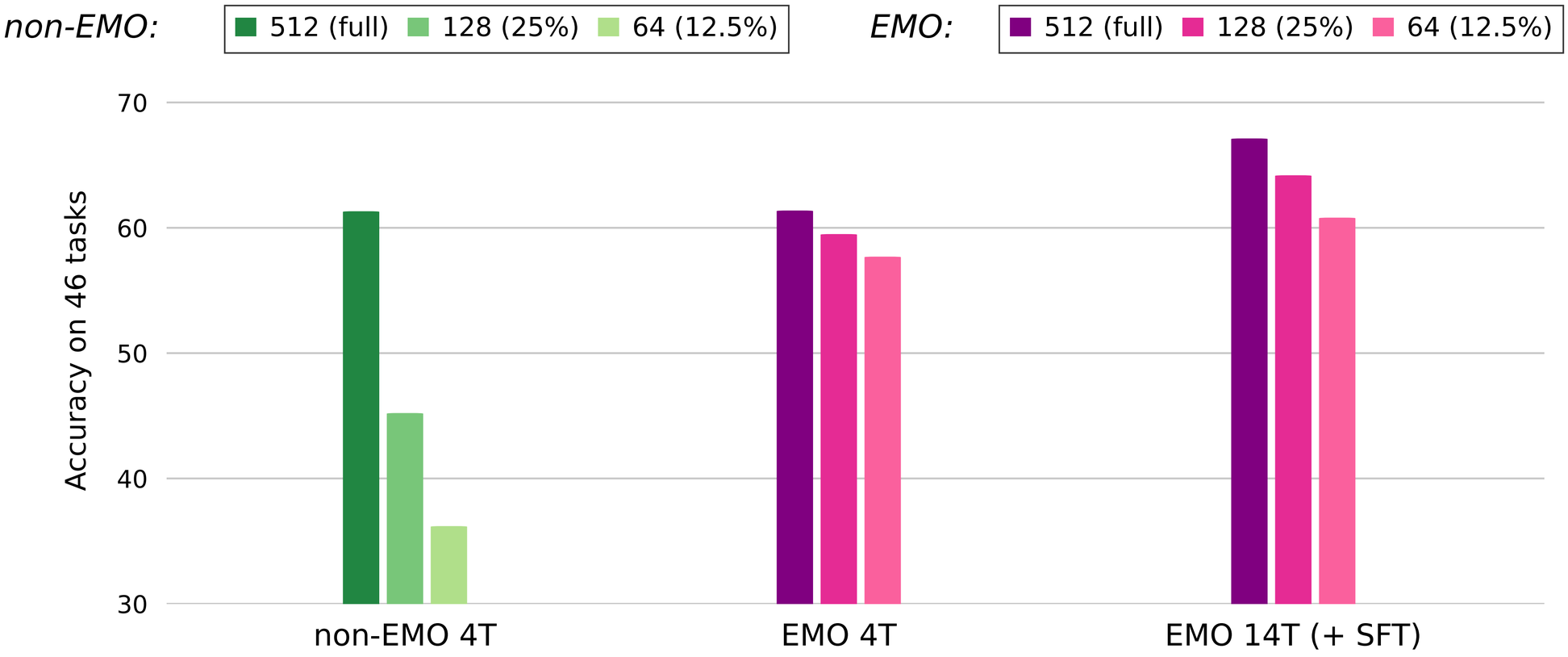
Film, music, TV & book reviews	5.1%
Personal blogs & casual commentary	4.4%
Web boilerplate: forms, listings & nav	4.2%
Health, medical & wellness	4.1%
Spam, adult, gambling & low-quality …	4.1%
Tech support, IT ops & web dev	4.1%
Personal narrative, emotions & relat…	3.9%
Business, marketing & e-commerce	3.9%
News reporting & wire service stories	3.7%
Sentence & paragraph endings	3.7%
Forum metadata & timestamps	3.5%
Source code	3.4%
U.S. politics & elections	3.2%
Engineering, automotive & hardware	3.2%
Consumer tech, gadgets & networking	3.2%
Proper names & foreign-language frag…	3.1%
Biomedical & life-sciences research	3.1%
Religion, philosophy & spirituality	3.1%

Top 18 of 32 clusters

emovisualization.netlify.app

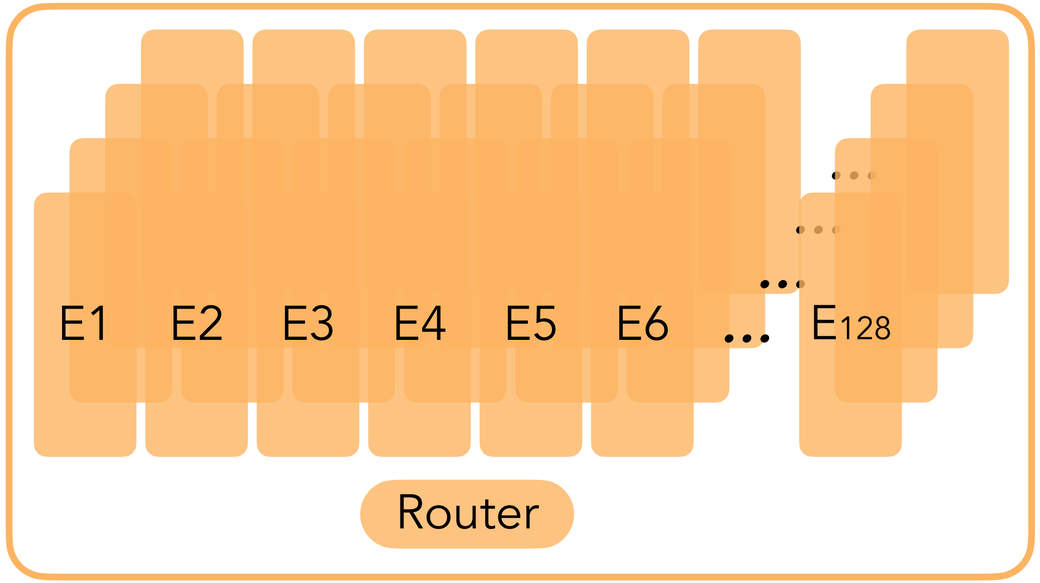
After the paper… Scaling up?

14T + 600B decay + 100B mid-training + 45B SFT —- 51,800 B300 hours



Huge thanks to the Ai2 Olmo team, including Akshita Bhagia, Jacob Morrison, Will Merrill, Yashas Samaga, Iz Beltage, Noah Smith, and many others…

But why not bigger (by making it sparser)?



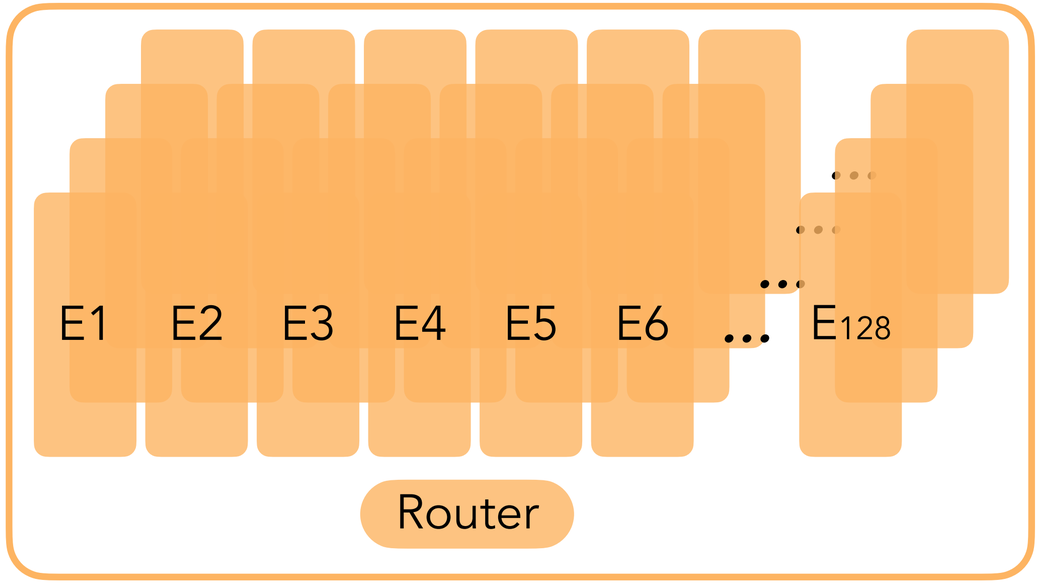
(Activate 8 out of 512, 1024, ….)

Two challenges

1. Training is more difficult
   (communication overhead, smaller per-expert batches, load-balancing challenges)
2. Inference is memory-intensive
   (weight storage scales with total parameters, making deployment harder on limited hardwares)

EMO!

But why not bigger (by making it sparser)?



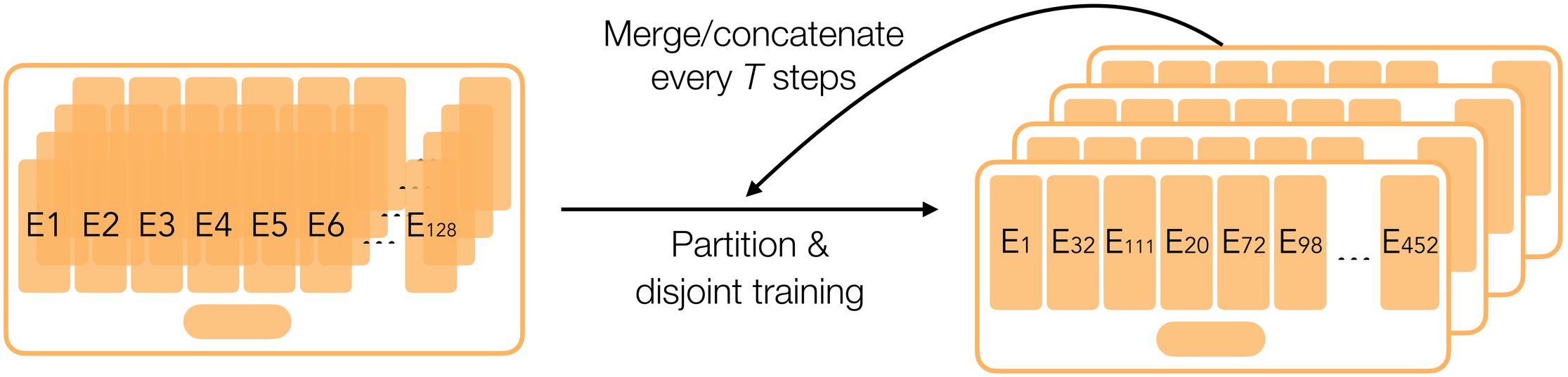
(Activate 8 out of 512, 1024, ….)

Two challenges ??

1. Training is more difficult
   (communication overhead, smaller per-expert batches, load-balancing challenges)
2. Inference is memory-intensive
   (weight storage scales with total parameters, making deployment harder on limited hardwares)

Split-and-Concat MoE

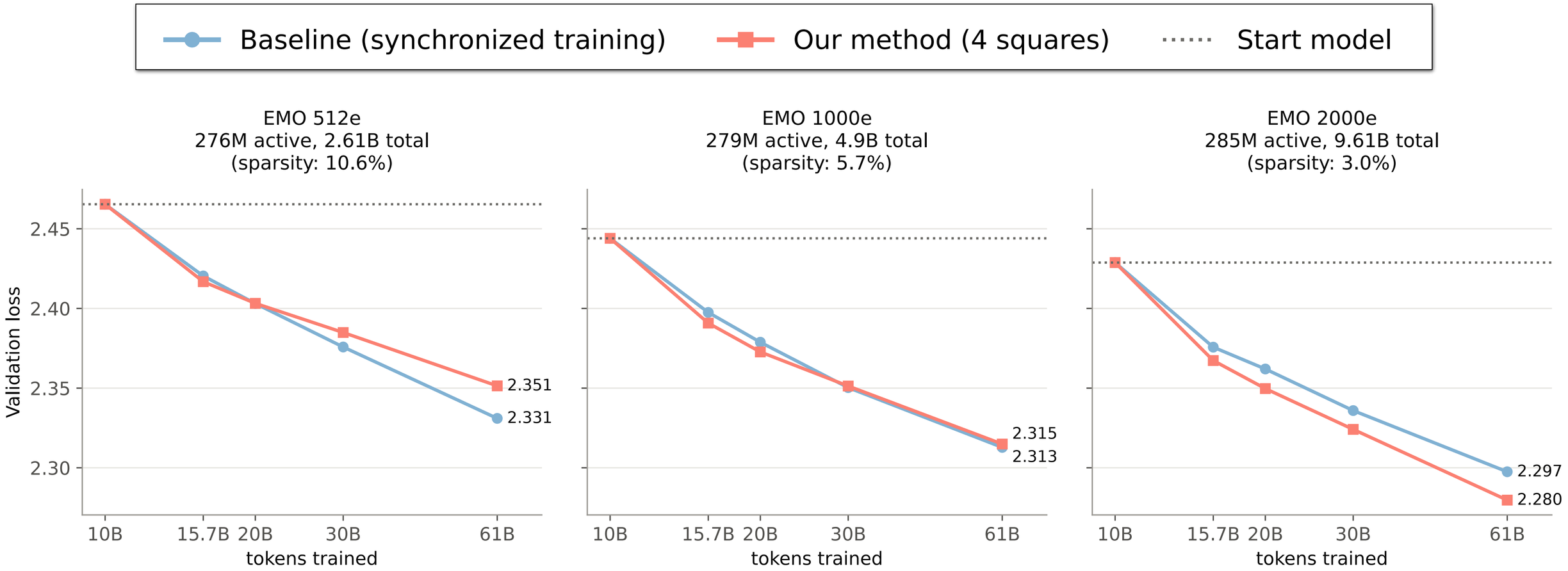
- Baseline: Training ***N*** total, ***K*** active for ***M*** total tokens
- Ours: Training ***4x*** (***N/4*** total, ***K*** active for ***M/4*** total tokens)



No synchronized training across all *N* experts after the seed stage

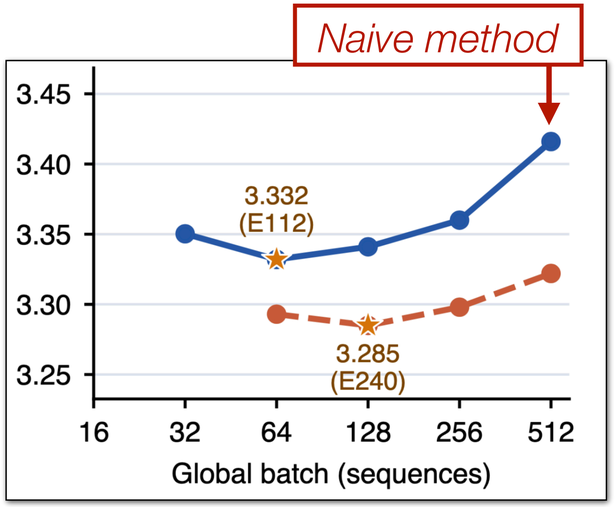
Ryan Wang et al. Upcoming Work

Split-and-Concat MoE: Results

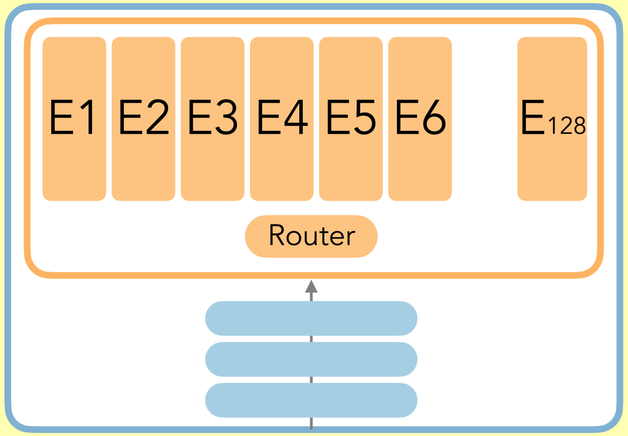


Ryan Wang et al. Upcoming Work

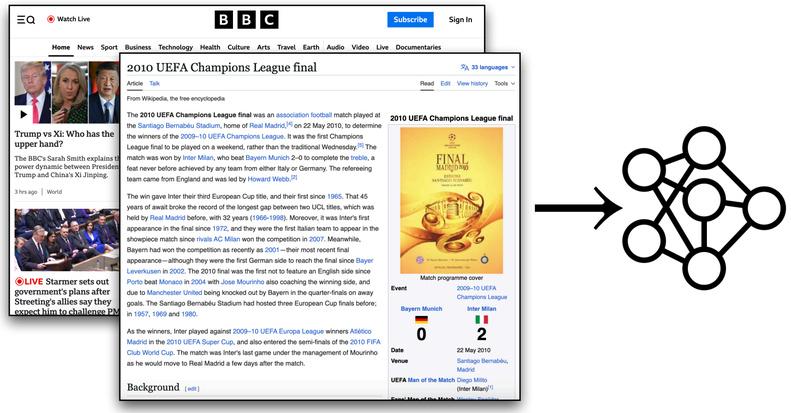
Today's talk



1. How to do repeated-data training *correctly*



2. New ways of training Mixture-of-Experts

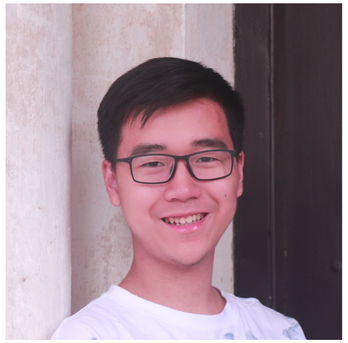
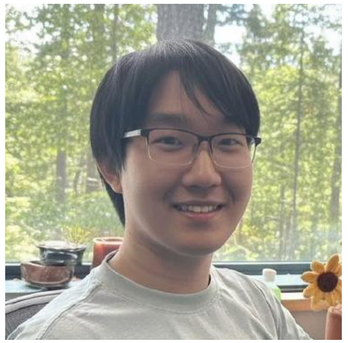


3. Bypassing text parsing

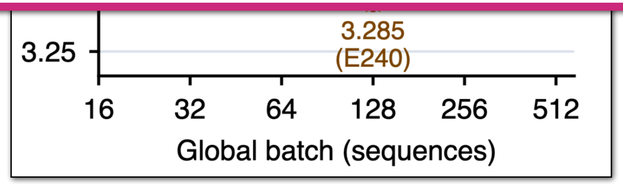
"EMO: Pretraining Mixture of Experts for Emergent Modularity" arxiv.org/abs/2605.06663

+ Upcoming work

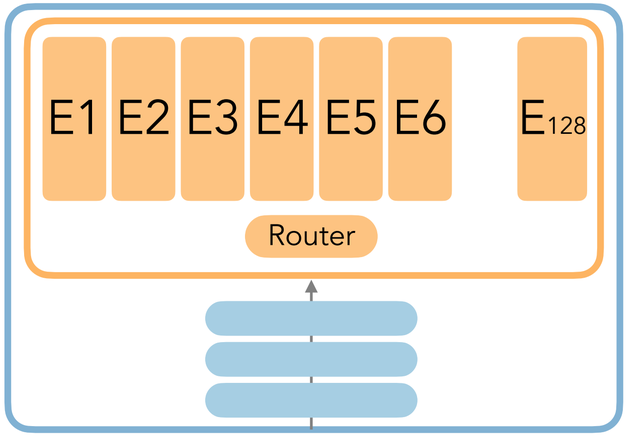
Today's talk

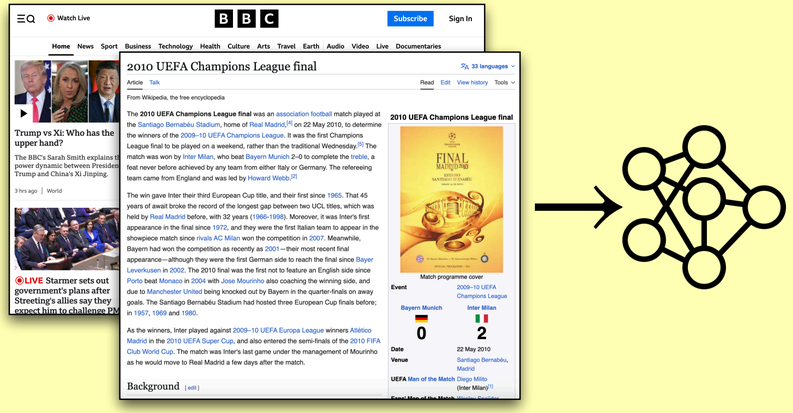
Yichuan Wang, Zhifei Li,
Zirui Wang, Paul Teiletche, Lesheng Jin, Matei Zaharia, Joey Gonzalez



1. How to do repeated-data training *correctly*



2. New ways of training Mixture-of-Experts

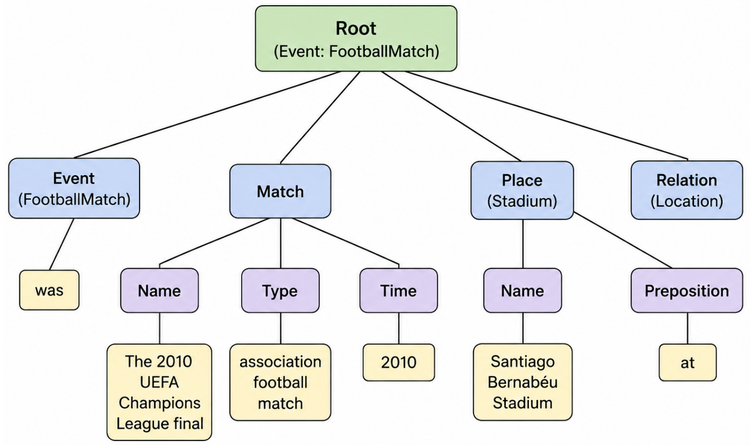


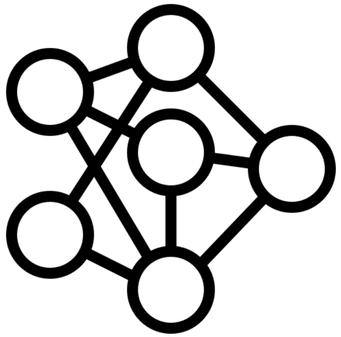
3. Bypassing text parsing

PixelRAG: Web Screenshots Beat Text for Retrieval-Augmented Generation; pixelrag.ai

We love E2E!

The 2010 UEFA Champions League final was an association football match played at the Santiago Bernabéu Stadium

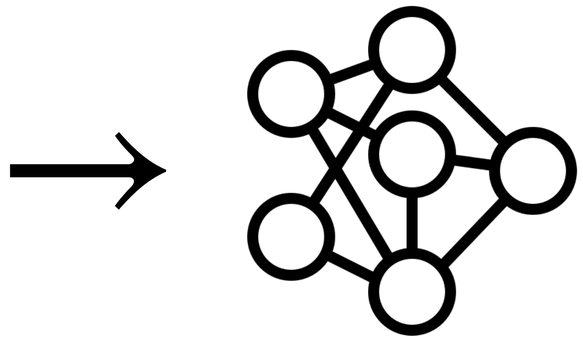




(Created by ChatGPT)

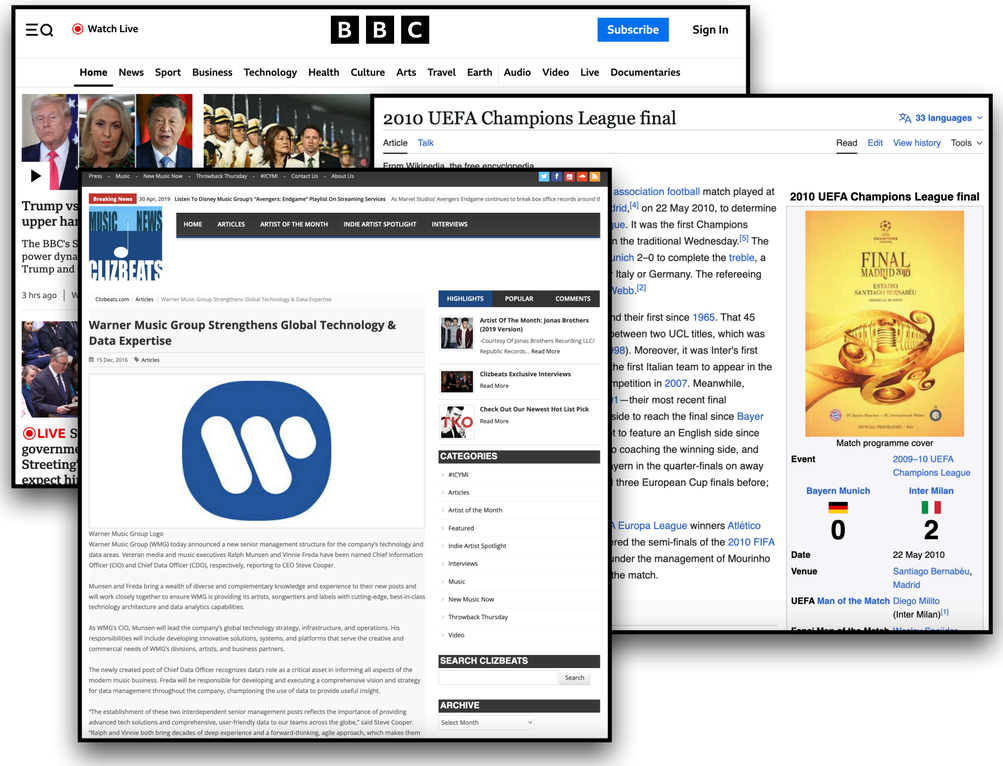
We love E2E!

The 2010 UEFA Champions League final was an association football match played at the Santiago Bernabéu Stadium



Wait…. Is text actually the most native form of Internet data?

Hidden Stage: HTML-to-text parsing



The 2010 UEFA Champions League final was an association football match played at the Santiago Bernabéu Stadium, home of Real Madrid, on 22 May 2010, to determine the winners of the 2009–10 UEFA Champions League. It was the first Champions League final to be played on a weekend, rather than the traditional Wednesday.[5] The match was won by Inter Milan, who beat Bayern Munich 2–0 to complete the treble, a feat never before achieved by any team from either Italy or Germany. The refereeing team came from England and was led by Howard Webb.[2]

Background== Prior to the 2010 final, …

Date	Stadium	Teams	Score	Key Scorer
22 May 2010	Santiago Bernabéu, Madrid	Bayern Munich vs Inter Milan	0–2	Diego Milito (35', 70')

Parsing is the bottleneck!

[WET file]

A Guide To Markets - The New York Times
NYTimes.com no longer supports Internet Explorer 9 or earlier. Please upgrade your browser. LEARN |
Sections
Home
Search
Skip to content Skip to navigation View mobile version
The New York Times
Archives|A Guide To Markets
Search
Subscribe Now
Log In
0
Settings
Close search
Site Search Navigation
Search NYTimes.com
Clear this text input
Go
https://nyti.ms/29nVV3Q
Loading...
See next articles
See previous articles
Site Navigation
Site Mobile Navigation
Advertisement
Archives | 1990
A Guide To Markets
MAY 10, 1990
Continue reading the main story Share This Page
Continue reading the main story

Choppy. Sentences split to many newlines. A lot of undesirable website content.

[Resiliparse]

This is a digitized version of an article from The Times's print archive, before the start of online publication in 1996. To preserve these articles as they originally appeared, The Times does not alter, edit or update them.

Occasionally the digitization process introduces transcription errors or other problems. Please send reports of such problems to archive_feedback@nytimes.com.

May 10, 1990, Page 00006 The New York Times Archives

HERE is a sampling of some of the better antiques and flea markets around the United States.

Two or Three Times a Year

BRIMFIELD Route 20, Brimfield, Mass. 01010; 413-245-3436. Second weekend of May and July, and the second weekend after Labor Day.

RENNINGER'S OUTDOOR EXTRAVAGANZA Noble Street, Kutztown, Pa.; 717-385-0104. Thursday, Friday and Saturday of the last weekend of April, June, September.

FARMINGTON ANTIQUES WEEKEND Farmington Polo Grounds, Town Farm Road, Farmington, Conn. 06032; 508-839-9735. Starting Wednesday before shows open; 203-677-7862. June 9-10 and Sept. 1-2.

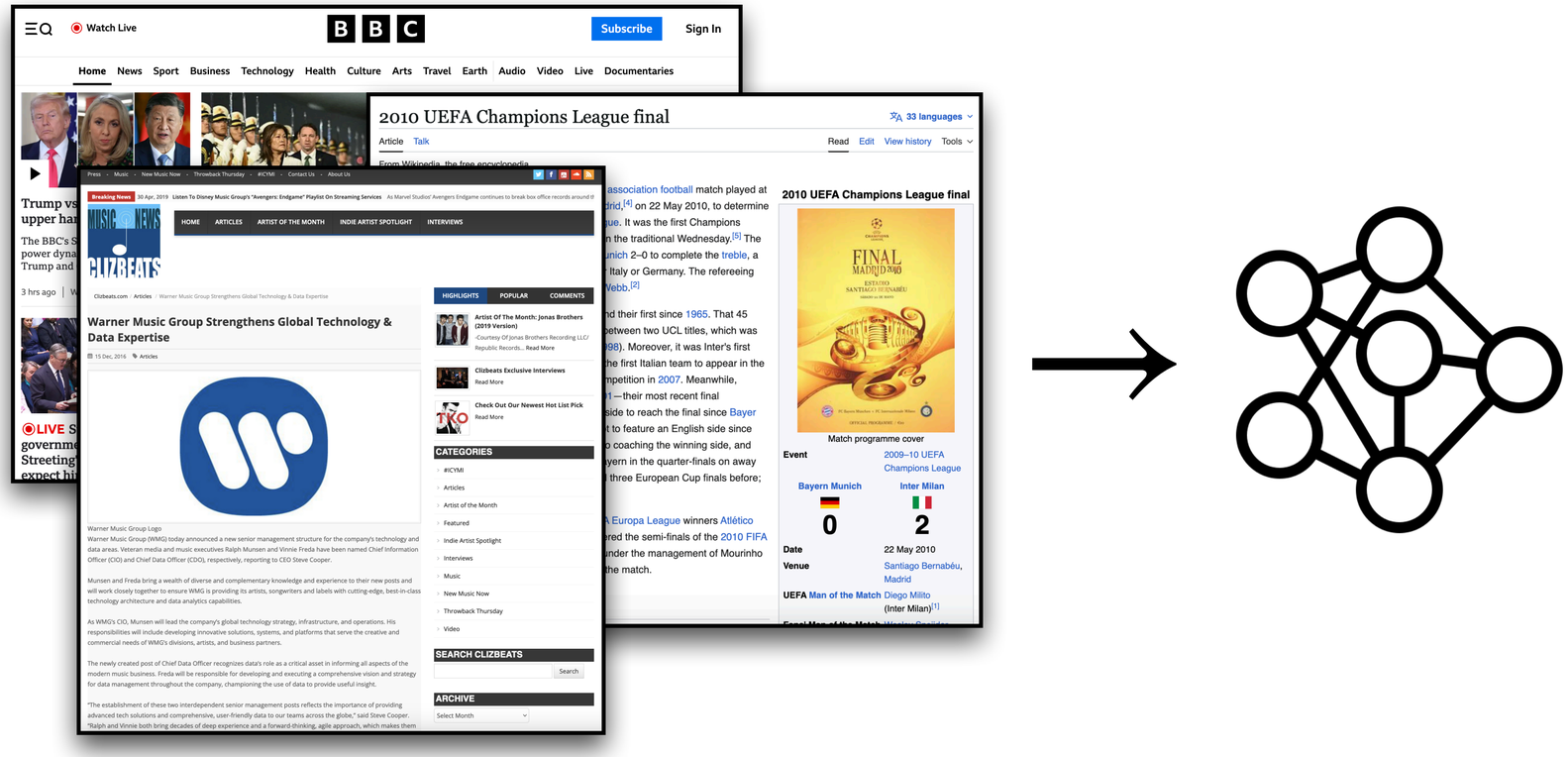
Monthly

ANN ARBOR ANTIQUES MARKET, P.O. Box 1512, Ann Arbor, Mich. 48106; 313-662-9453. May through October, third Sunday.

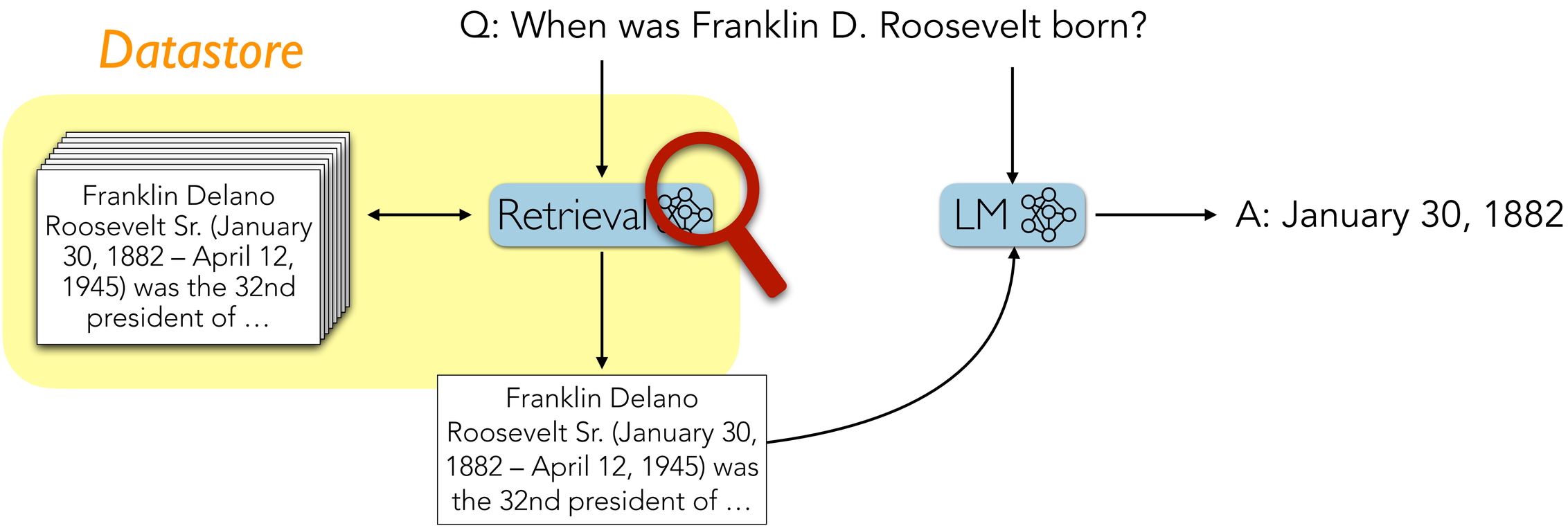
- Li et al. (2026): A single HTML parser can discard up to 40% of website content
- Li et al., 2024; Penedo et al., 2024; Fang et al., 2025: Parser choice significantly impacts downstream performance
- Wang et al., 2025; Liu et al., 2025: Uses VLM-based parsers

Slide credit: Kyle Lo

Parsing is the bottleneck!



Retrieval-Augmented Generation (RAG)

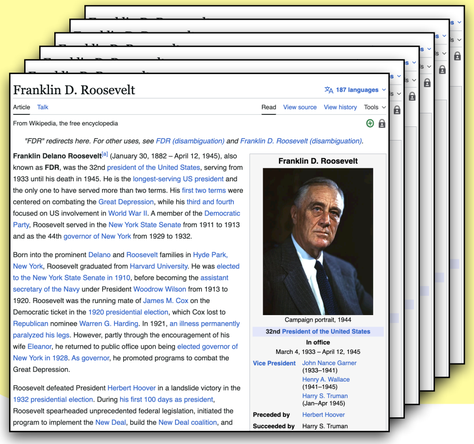


Chen et al. 2017. Guu et al. 2018. Lee et al. 2019. Karpukhin et al. 2020. Lewis et al. 2020. …

PixelRAG: RAG over pixel space

Q: When was Franklin D. Roosevelt born?

Datastore



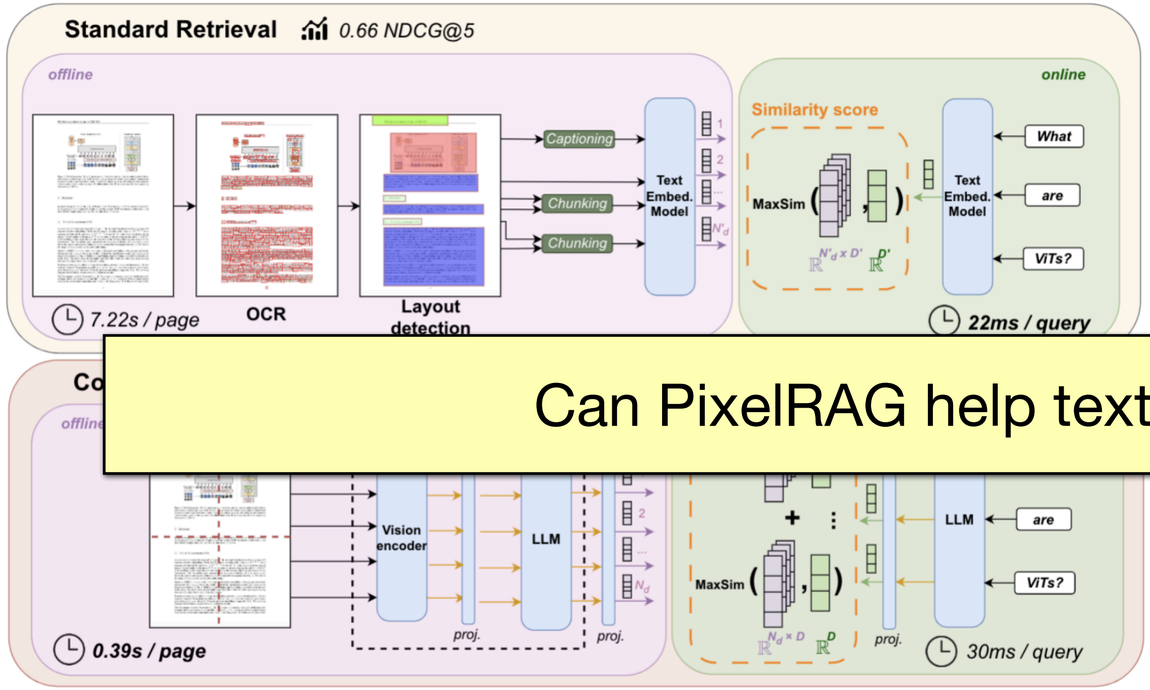
Retrieval

VLM

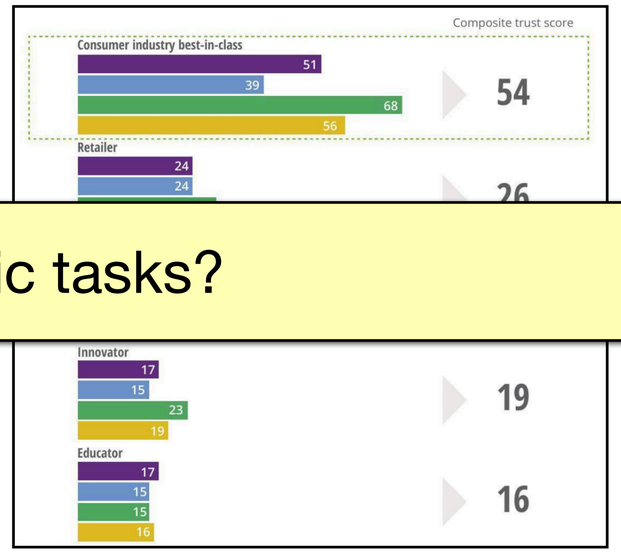
A: January 30, 1882



Prior work on visual tasks



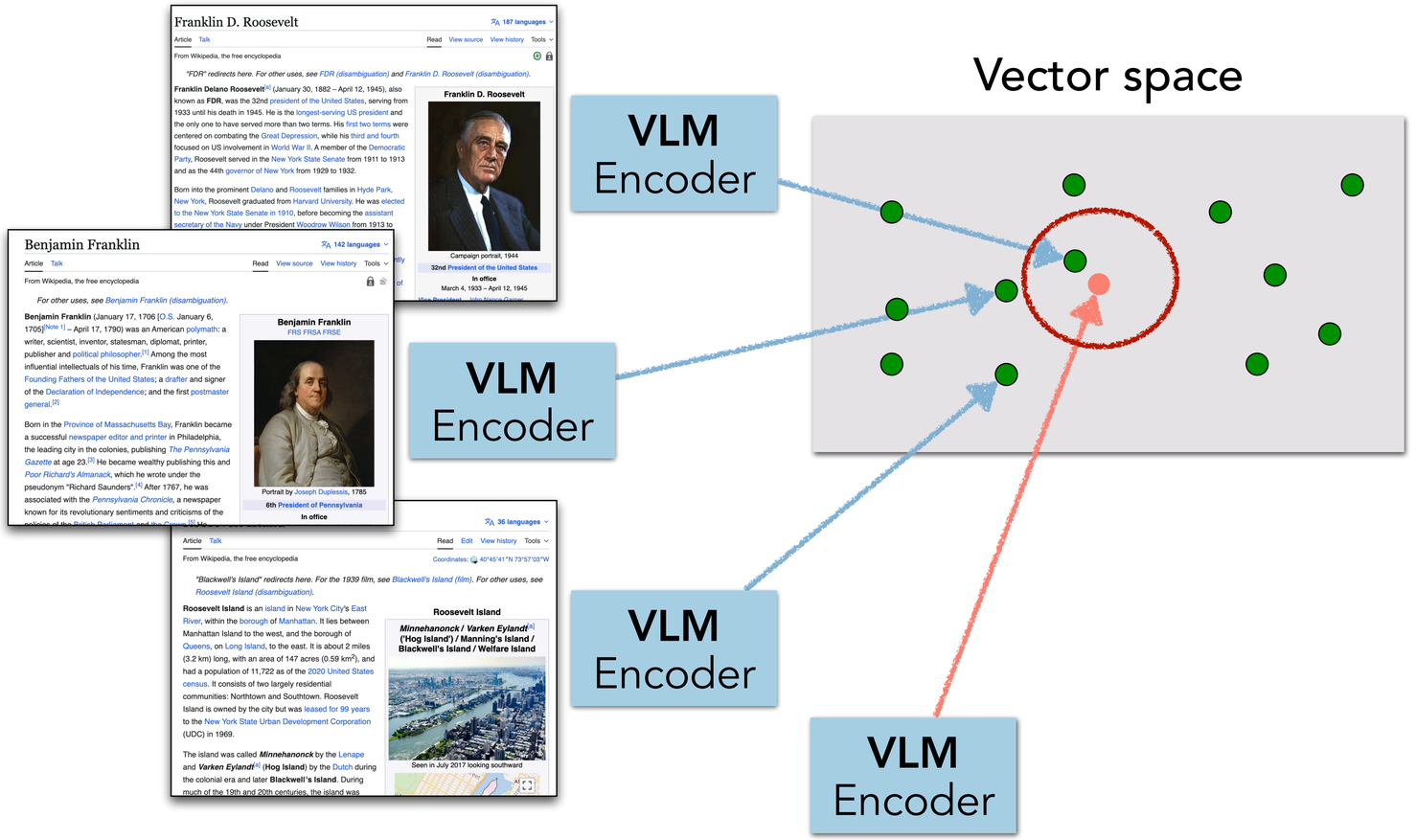
Q: What is the trust signal score for the consumer industry best-in-class archetype?



Can PixelRAG help text-centric tasks?

Deepseek-OCR [Wei et al. 2025], ColPali [Faysse et al. 2025], ColQwen [Zhou et al. 2025]

PixelRAG stage 1: Retrieval



$$\mathbf{z}_i = \mathrm{Enc}(z_i) \in \mathbb{R}^h \ (1 \le i \le N)$$

$$\mathbf{x} = \mathrm{Enc}(x) \in \mathbb{R}^h$$

Q: When was Franklin D. Roosevelt born?

PixelRAG stage 2: Reading

A: January 30, 1882





Franklin D. Roosevelt

187 languages

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

"FDR" redirects here. For other uses, see FDR (disambiguation) and Franklin D. Roosevelt (disambiguation).

Franklin Delano Roosevelt[a] (January 30, 1882 – April 12, 1945), also known as **FDR**, was the 32nd president of the United States, serving from 1933 until his death in 1945. He is the longest-serving US president and the only one to have served more than two terms. His first two terms were centered on combating the Great Depression, while his third and fourth focused on US involvement in World War II. A member of the Democratic Party, Roosevelt served in the New York State Senate from 1911 to 1913 and as the 44th governor of New York from 1929 to 1932.

Born into the prominent Delano and Roosevelt families in Hyde Park, New York, Roosevelt graduated from Harvard University. He was elected to the New York State Senate in 1910, before becoming the assistant secretary of the Navy under President Woodrow Wilson from 1913 to 1920. Roosevelt was the running mate of James M. Cox on the Democratic ticket in the 1920 presidential election, which Cox lost to Republican nominee Warren G. Harding. In 1921, an illness permanently paralyzed his legs. However, partly through the encouragement of his wife Eleanor, he returned to public office upon being elected governor of New York in 1928. As governor, he promoted programs to combat the Great Depression.

Roosevelt defeated President Herbert Hoover in a landslide victory in the 1932 presidential election. During his first 100 days as president, Roosevelt spearheaded unprecedented federal legislation, initiated the program to implement the New Deal, build the New Deal coalition, and

Franklin D. Roosevelt

Campaign portrait, 1944

32nd President of the United States

In office

March 4, 1933 – April 12, 1945

Vice President John Nance Garner (1933–1941)
Henry A. Wallace (1941–1945)
Harry S. Truman (Jan–Apr 1945)

Preceded by Herbert Hoover

Succeeded by Harry S. Truman

Q: When was Franklin D. Roosevelt born?

PixelRAG stage 2: Reading

A: January 30, 1882

VLM



Q: When was Franklin D. Roosevelt born?

PixelRAG Training

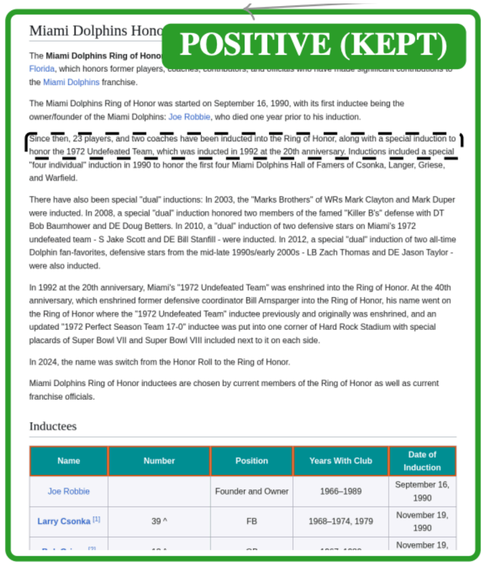
Input

Positive

$$-\log \frac{e^{\text{sim}(x,\ z^+)}}{e^{\text{sim}(x,\ z^+)} + \sum_{i=1}^{n} e^{\text{sim}(x,\ z_i^-)}}$$

Hard negatives

Q: When was the ‘1972 Undefeated Team’ inducted into the Miami Dolphins Ring of Honor?



(Paper for more details…)

- 3 hours on a single H100
- Using LoRA applied to both the ViT and LM backbones

Evaluation

Text-centric benchmarks

- SimpleQA [Wei et al. 2024]
- NQ [Kwiatkowski et al. 2019]
- NQ-Table [Herzig et al. 2021]

Very widely studied in the NLP/text community!!

Most (probably all) questions are answerable from text alone.

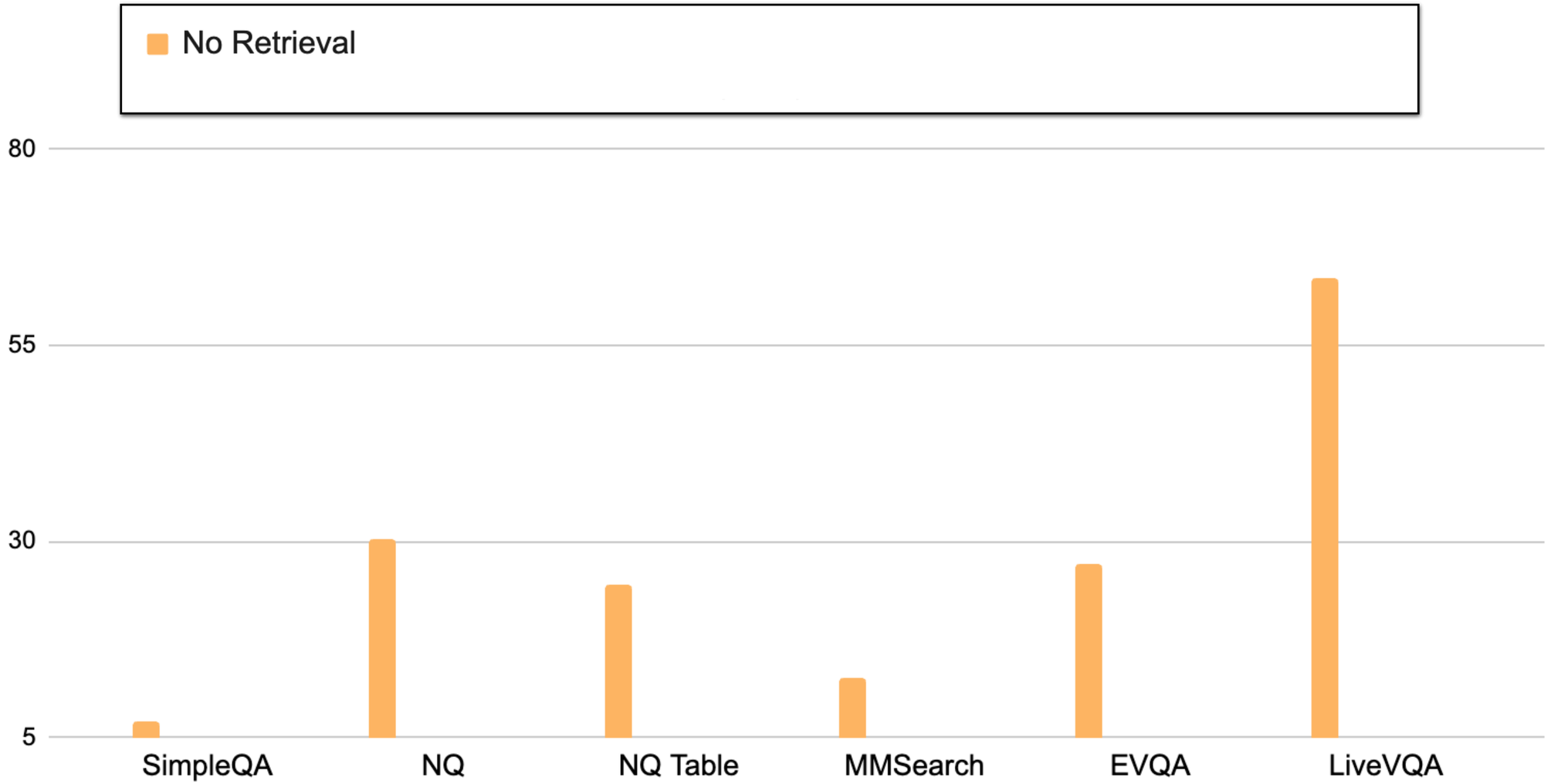
Multimodal benchmarks

- MMSearch [Jiang et al. 2024]
- Encyclopedic VQA [Mensink et al. 2023]

QA over noisy news corpora

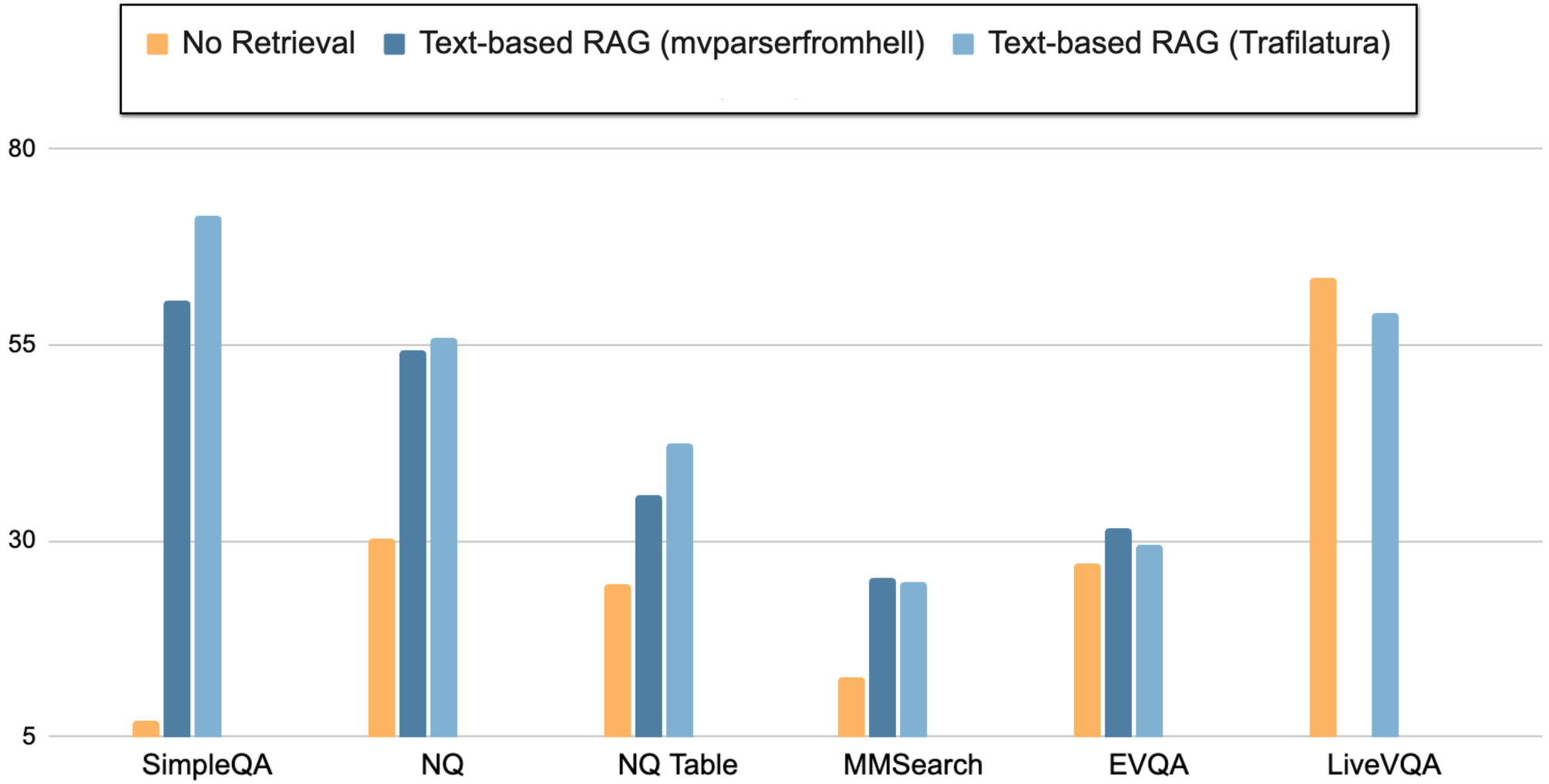
- LiveVQA [Fu et al. 2025]

PixelRAG: Results



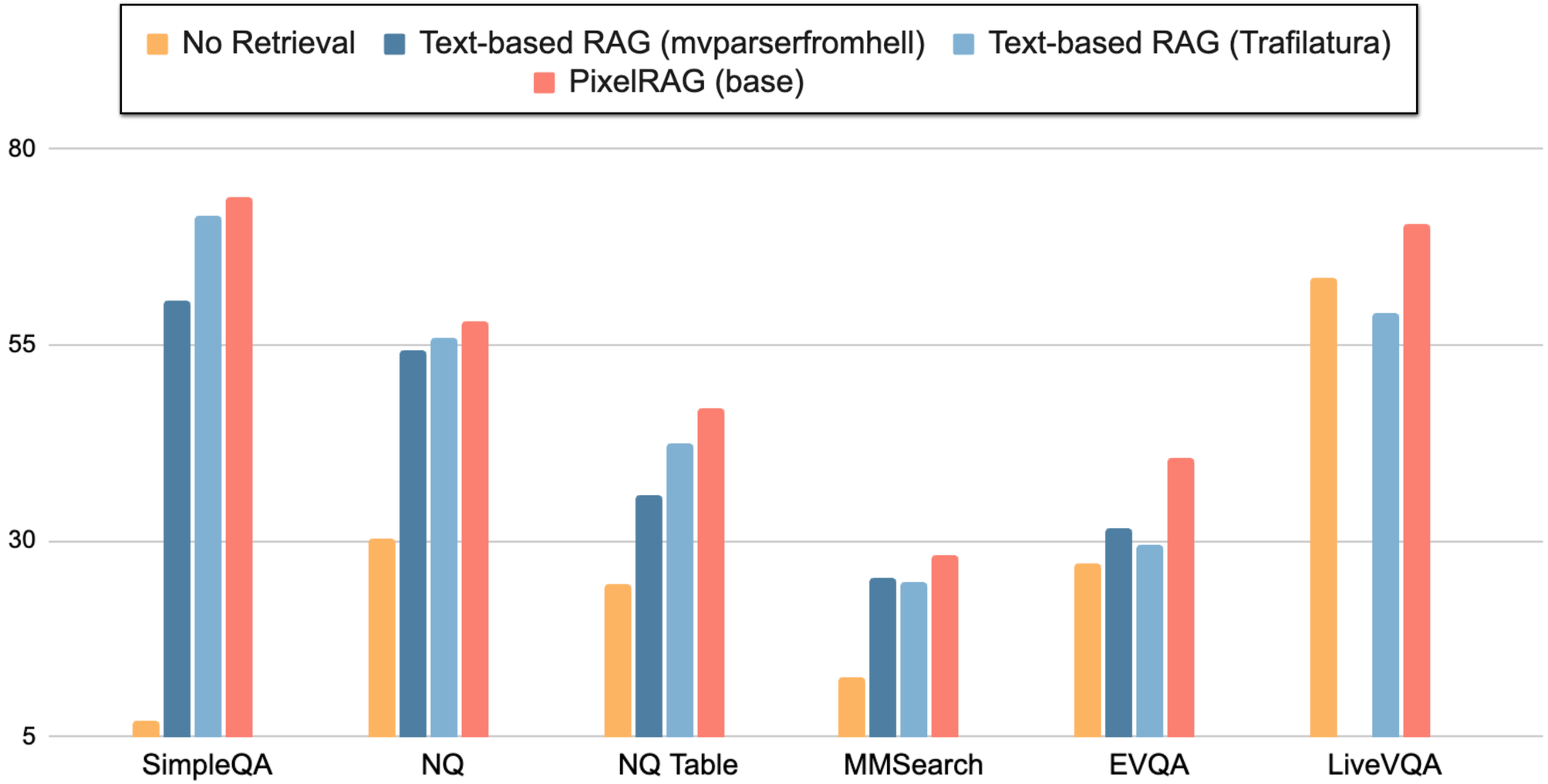
[Qwen3.5-4B]

PixelRAG: Results



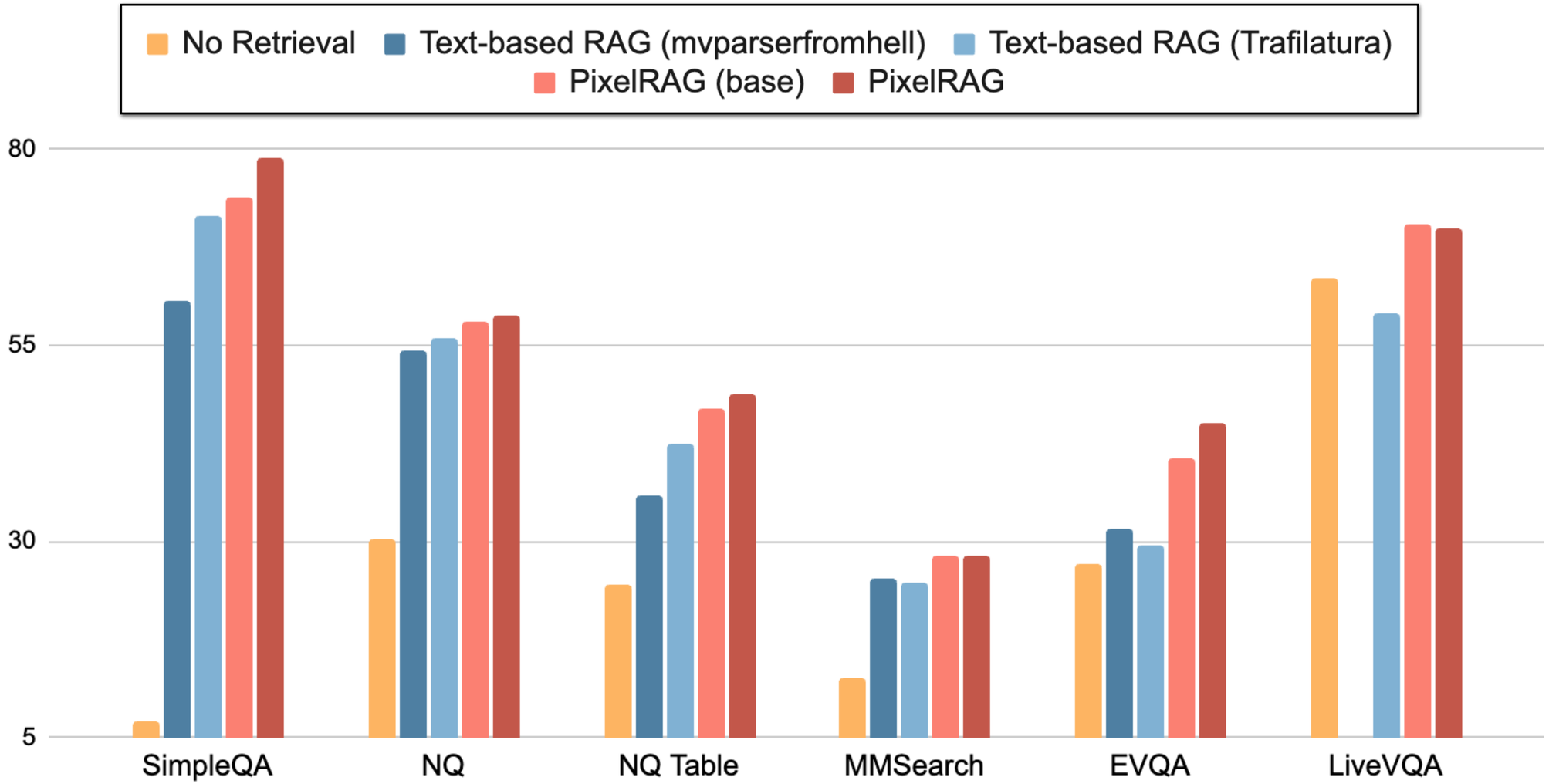
[Qwen3.5-4B]

PixelRAG: Results



[Qwen3.5-4B]

PixelRAG: Results



[Qwen3.5-4B]

Consistent across readers

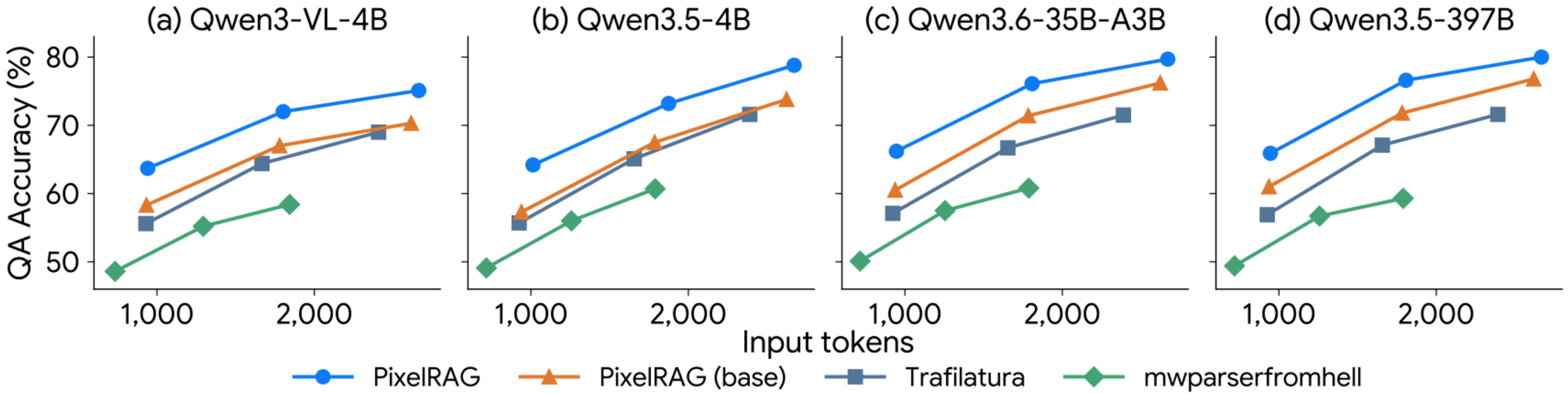


Figure 3: SimpleQA accuracy versus average input tokens across four reader models ($k \in \{1, 2, 3\}$).

Q: “How many shots did Inter attempt on target in the Champions League Final . . . May 23, 2010?” *A:* **7**

Trafilatura text top-3

rank 1, score 0.694 — *2010 UEFA Champions League final* (chunk 0)

```
2010 UEFA Champions League final | Date | 22 May
2010 | Venue | Santiago Bernabéu, Madrid | ...
```

↪ *Infobox metadata; no shots-on-target stats.*

rank 2, score 0.670 — *2009–10 UEFA Champions League* (chunk 0)

```
2009-10 UEFA Champions League | Dates | 30 June ...
22 May 2010 | Champion | Inter Milan | ...
```

↪ *Season overview infobox, not match stats.*

rank 3, score 0.667 — *Wesley Sneijder* (chunk 13)

```
"Bayern Munich 0-2 Internazionale".  ESPN Soccernet.
22 May 2010.  Archived from ...
```

↪ *Bibliography, not data.*

Q: "How many shots did Inter attempt on target in the Champions League Final . . . May 23, 2010?" *A:* **7**

Trafilatura text top-3

rank 1, score 0.694 — *2010 UEFA Champions League final* (chunk 0)

```
2010 UEFA Champions League final | Date | 22 May
2010 | Venue | Santiago Bernabéu, Madrid | ...
```

↪ *Infobox metadata; no shots-on-target stats.*

rank 2, score 0.670 — *2009–10 UEFA Champions League* (chunk 0)

```
2009-10 UEFA Champions League | Dates | 30 June ...
22 May 2010 | Champion | Inter Milan | ...
```

↪ *Season overview infobox, not match stats.*

rank 3, score 0.667 — *Wesley Sneijder* (chunk 13)

```
"Bayern Munich 0-2 Internazionale".  ESPN Soccernet.
22 May 2010.  Archived from ...
```

↪ *Bibliography, not data.*

⋮

rank 26, score 0.614 — *2010 UEFA Champions League final* (chunk 4)

```
Statistics | | | | Post-match As a result of
Inter's victory, Italy held onto its position ...
```

↪ *Statistics table destroyed by linearization; header survived, table body reduced to empty pipes.*

PIXELRAG screenshot top-3

rank 1, score 0.601 — *2010 UEFA CL final* (tile 7)

Martin Atkinson (England)[c]
Reserve assistant referee:
Peter Kirkup (England)[2]

Statistics

First half[41]			Second half[41]			Overall[41]		
Statistic	Bayern Munich	Inter Milan	Statistic	Bayern Munich	Inter Milan	Statistic	Bayern Munich	Inter Milan
Goals scored	0	1	Goals scored	0	1	Goals scored	0	2
Total shots	10	7	Total shots	11	4	Total shots	21	11
Shots on target	1	4	Shots on target	5	3	Shots on target	6	7
Saves	3	1	Saves	2	5	Saves	5	6
Ball possession	67%	33%	Ball possession	69%	31%	Ball possession	68%	32%
Corner kicks	2	0	Corner kicks	4	2	Corner kicks	6	2
Fouls committed	8	6	Fouls committed	8	7	Fouls committed	16	13
Offsides	0	0	Offsides	0	0	Offsides	0	0
Yellow cards	1	1	Yellow cards	1	0	Yellow cards	2	1
Red cards	0	0	Red cards	0	0	Red cards	0	0

Post-match

As a result of Inter's victory, Italy held onto its position in the top three of the UEFA country coefficient rankings and would therefore retain its fourth berth in the 2010–11 UEFA Champions League. If Bayern had won or the match had gone to penalties, Germany would have overtaken Italy and received an extra entry in 2010–11.[42] Having beaten Roma both in the 2010 Coppa Italia Final and to the 2009–10 Serie A title, Inter won the treble for the first time in Italian football history.[43]



Match winner Diego Milito's jersey and memorabilia from the final displayed at the San Siro museum

Despite winning the treble, José Mourinho resigned as manager of Inter on 28 May to take over from Manuel Pellegrini as manager of Real Madrid, citing a desire to become the first manager to win the Champions League with three clubs.[44] However, he was unable to accomplish this feat, being eliminated in

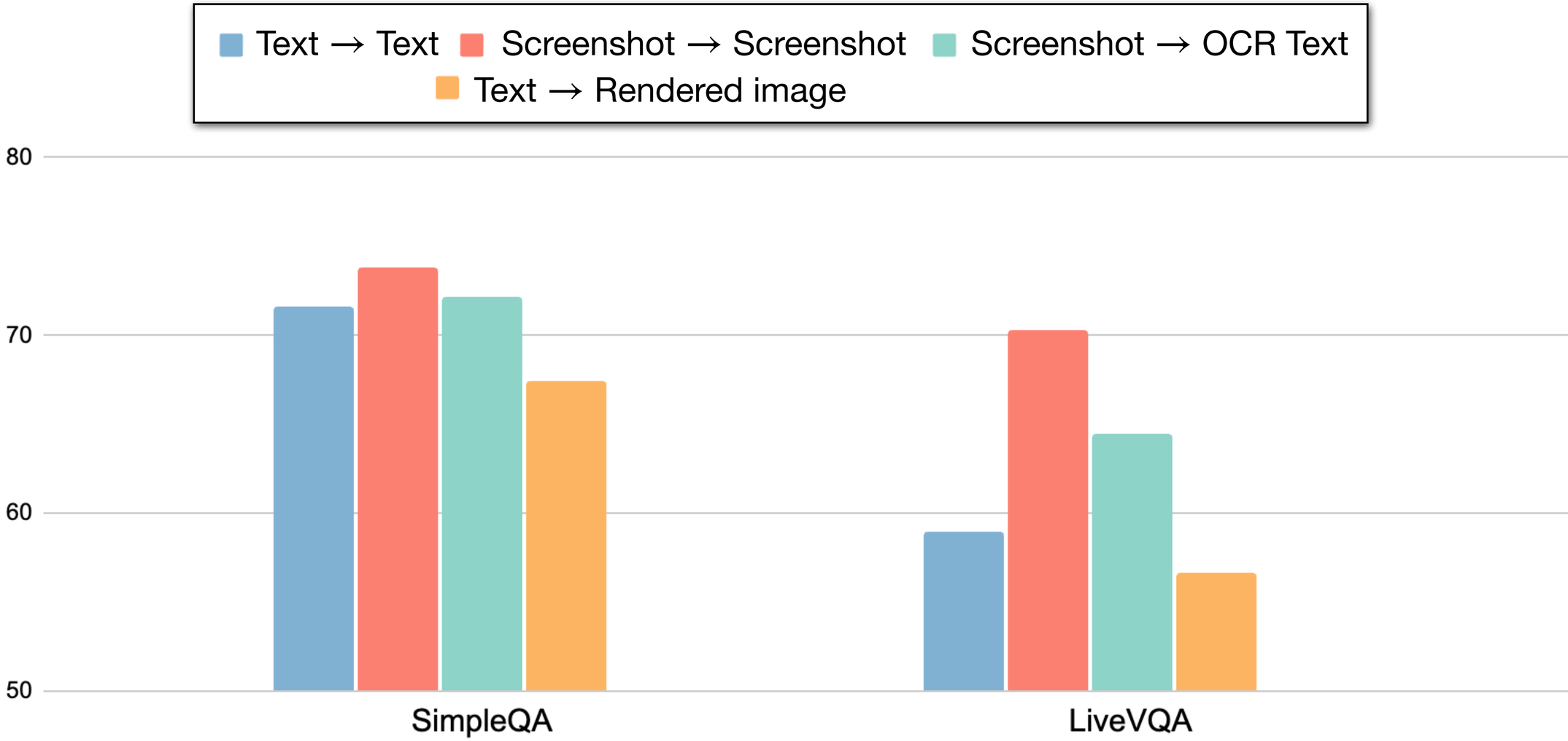
Modality ablations



Modality ablations



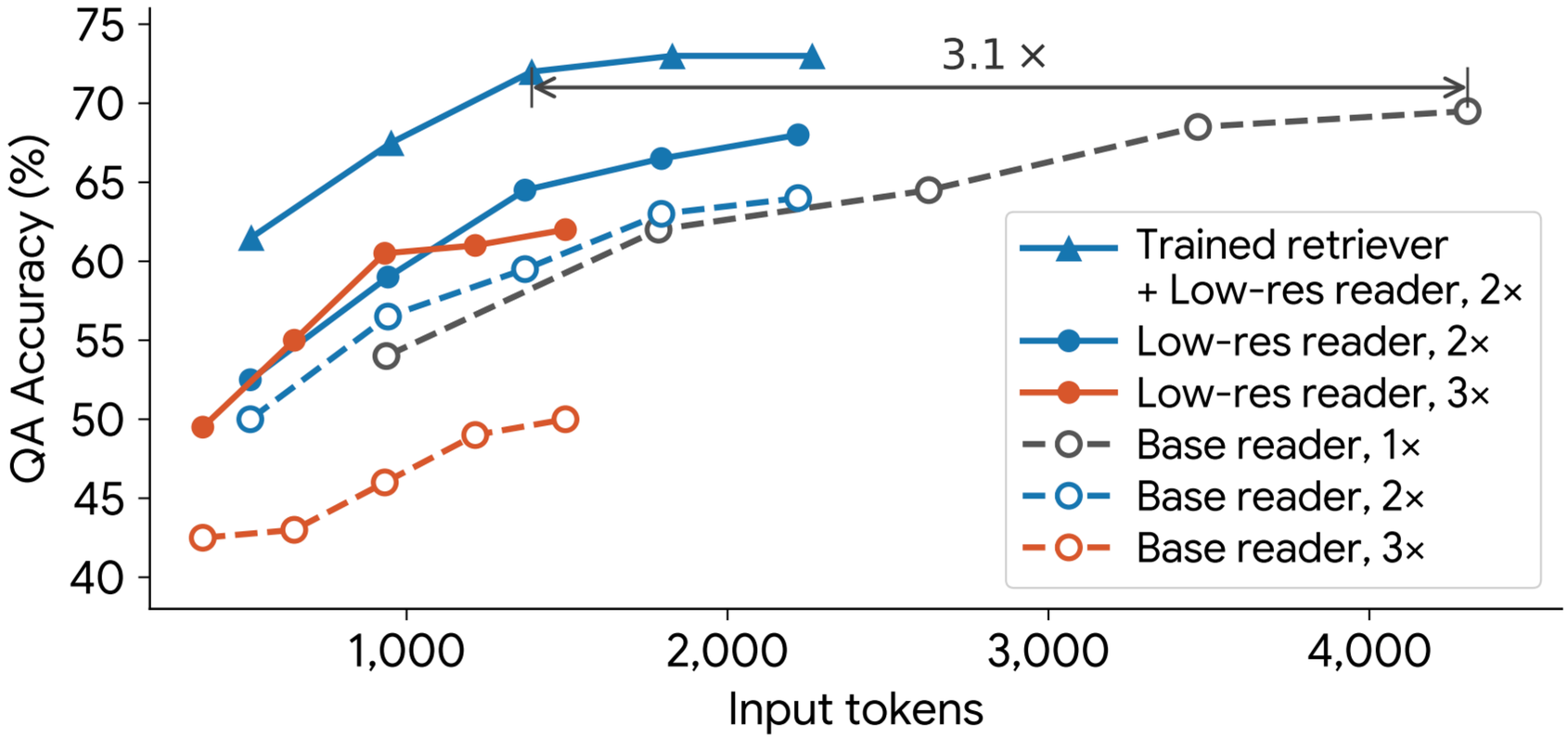
Modality ablations



Modality ablations



Image compression as an efficiency lever



Qwen3-VL-4B, SimpleQA

Reader model	PixelRAG Acc	Text-based RAG Acc
LLaVA-1.5-7B [63]‡	0.092	0.504
Llama-3.2-11B-Vision [64]†	0.534	0.700
Llama-3.2-90B-Vision [64]†	0.574	0.670
Llama-4-Scout-17B-16E [65]	0.618	0.667
Llama-4-Maverick-17B-128E [65]	0.756	0.719
Qwen2-VL-2B [66]	0.439	0.564
Qwen2-VL-7B [66]	0.628	0.663
Qwen2-VL-72B [66]	0.745	0.718
Qwen2.5-VL-3B [67]	0.460	0.661
Qwen2.5-VL-7B [67]	0.713	0.678
Qwen2.5-VL-32B [67]	0.717	0.691
Qwen2.5-VL-72B [67]	0.745	0.694
Qwen3-VL-235B-A22B [16]	0.759	0.720
Qwen3.5-0.8B [55]	0.602	0.593
Qwen3.5-2B [55]	0.667	0.667
Qwen3.5-4B [55]	0.738	0.716
Qwen3.5-9B [55]	0.750	0.701
Qwen3.5-27B [55]	0.759	0.711
Qwen3.5-35B-A3B [55]	0.737	0.713
Qwen3.6-27B [68]	0.754	0.715
Qwen3.6-35B-A3B [68]	0.751	0.705
Qwen3.5-4B (reasoning) [55]	0.745	0.697
Qwen3.5-397B-A17B (reasoning) [55]	0.769	0.710
Qwen3.6-35B-A3B (reasoning) [68]	0.763	0.717

SimpleQA Performance

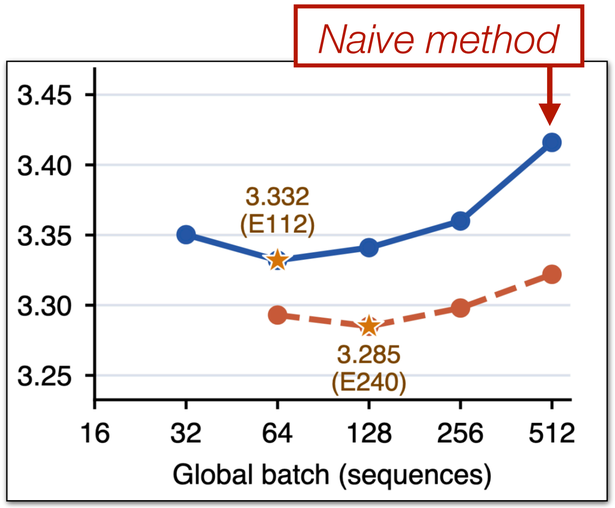
2024: PixelRAG: **53-57%**, Text-based RAG: **70%**

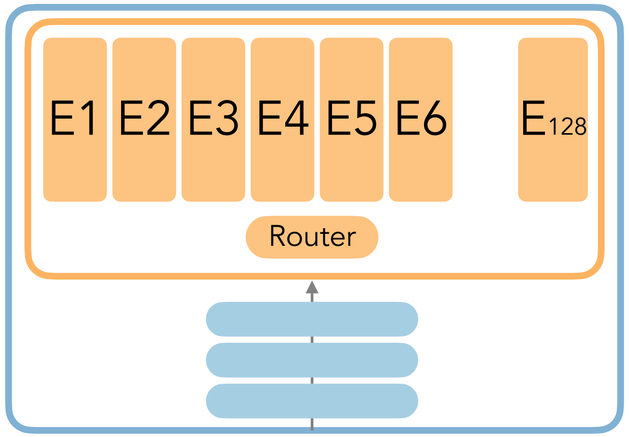
Parsing may be the bottleneck than anything else.
Should all foundation models bypass parsing by default?

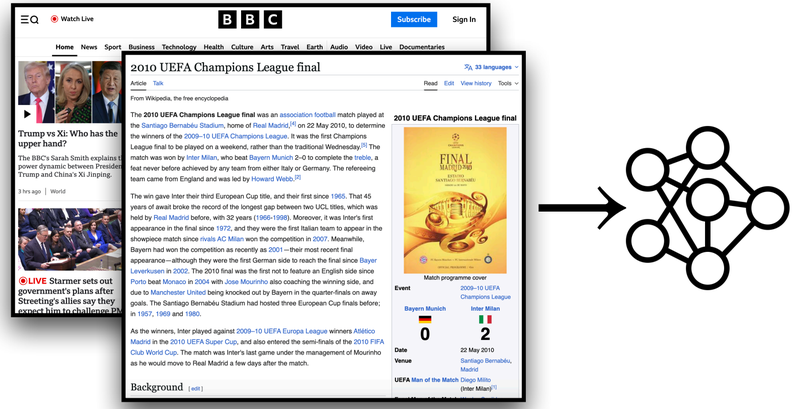
2026: PixelRAG: **77%**, Text-based RAG: **71%**

Wrapping up

When the constraints change, pre-training needs to change too.







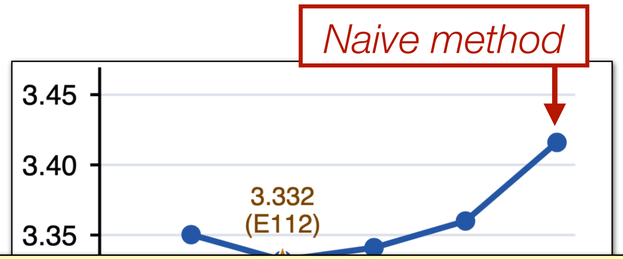
1. How to do repeated-data training *correctly*

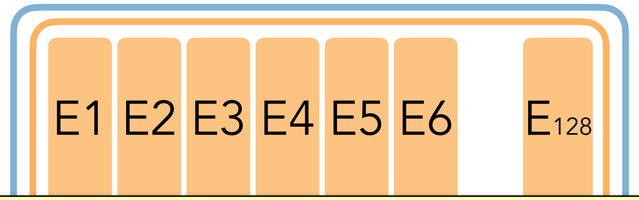
2. New ways of training Mixture-of-Experts

3. Bypassing text parsing

Hot takes (?)

Q: But aren't these problems better tackled in industry?







1. We can't train frontier models, but we can tackle frontier problems.
2. Industry approaches problems in certain ways (e.g., B300s instead of H100s, GB300 NVL72 instead of B300s). But that's not the only way.
3. If I don't know the solution, the problem isn't solved.

Thank you for listening!

sewonmin.com

sewonm@berkeley.edu

Please leave feedback at tinyurl.com/sewonm-talk