Enabling Language Models to Process Information at Scale

Tianyu Gao

Department of Computer Science

Princeton University

The rapid progress of language models (LMs)

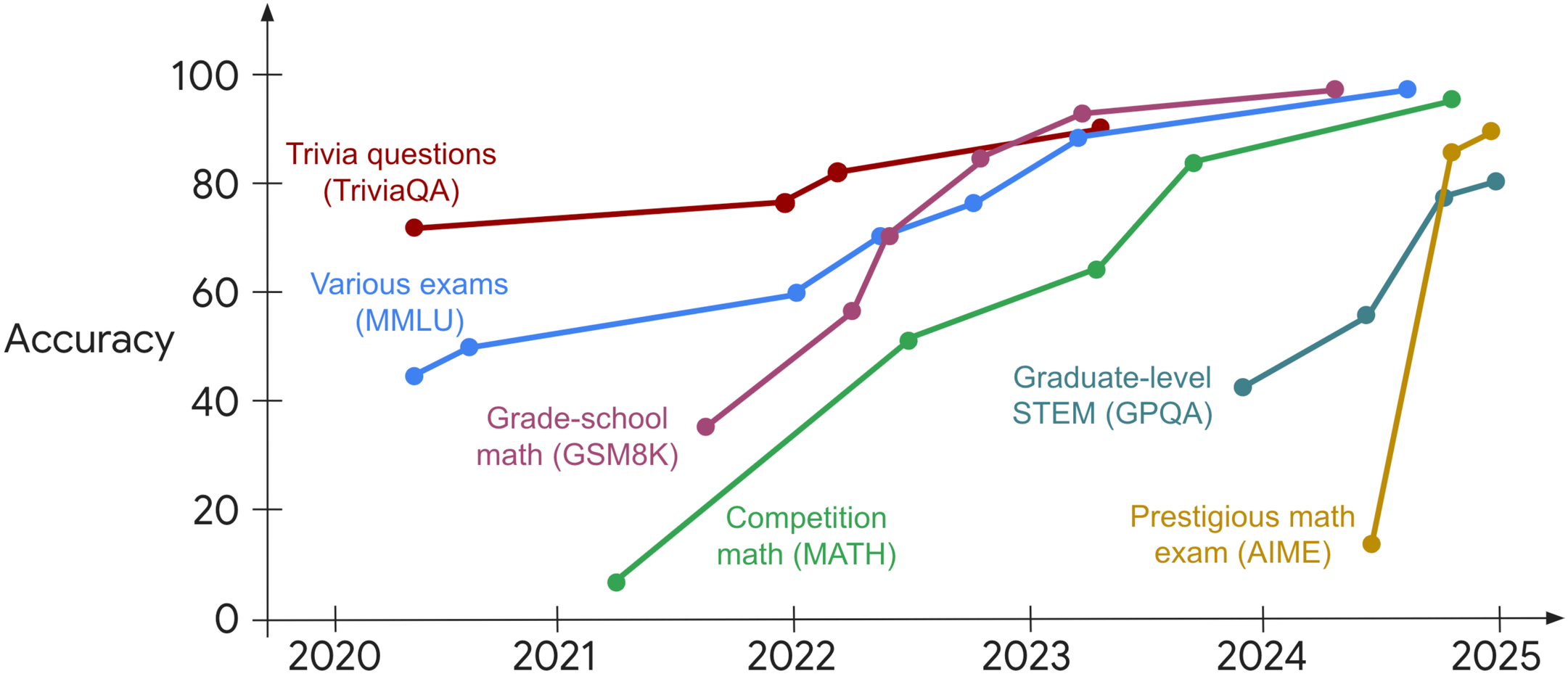
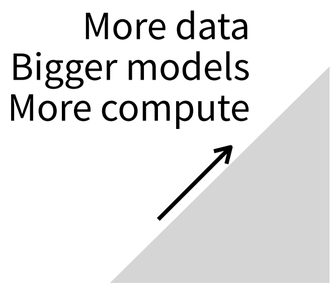


FIgure credit: https://x.com/_jasonwei/status/1889096555254456397

The current focus:



The current focus:

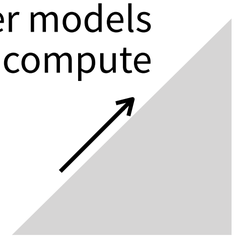
Pre-training data

↓

internalize

“knowledge”

More data
Bigger models
More compute

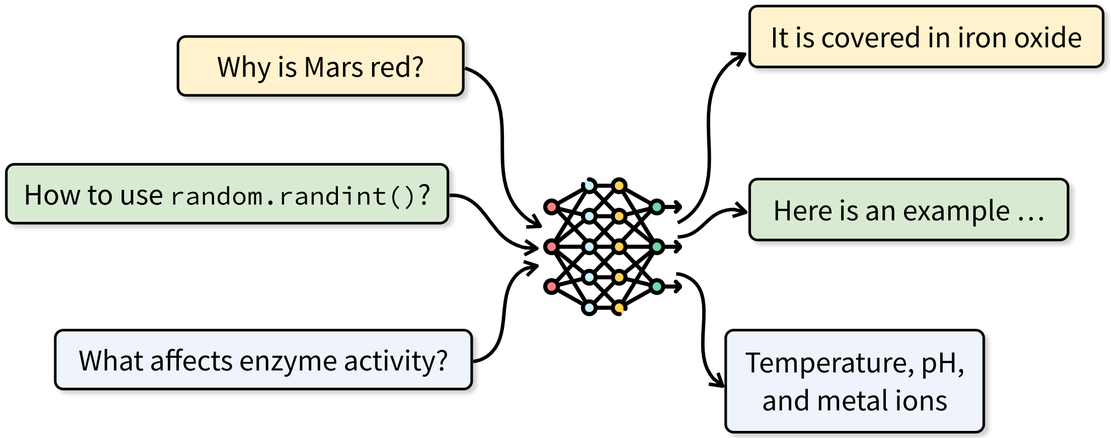


The current focus:

Pre-training data

↓

internalize “knowledge”



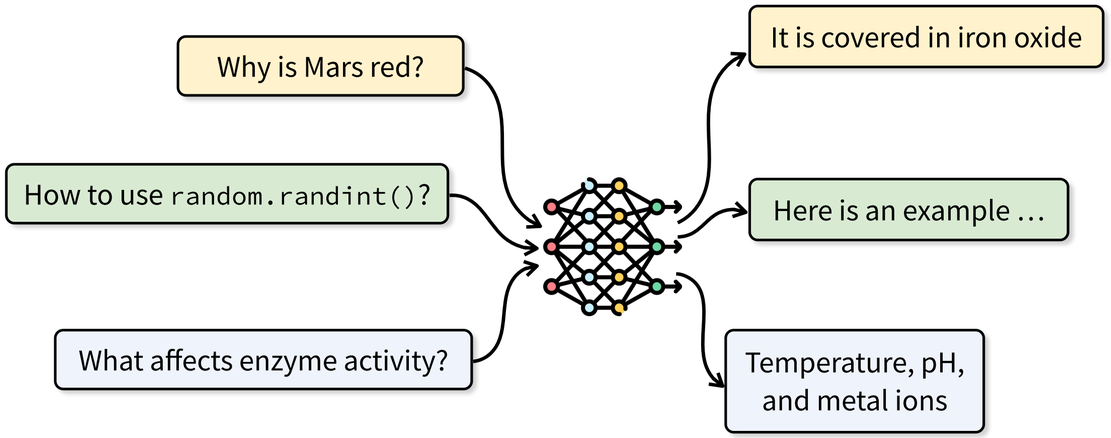
More data
Bigger models
More compute

The current focus:

Pre-training data

↓

internalize “knowledge”



More data
Bigger models
More compute

How to make LMs more useful beyond just passing exams?

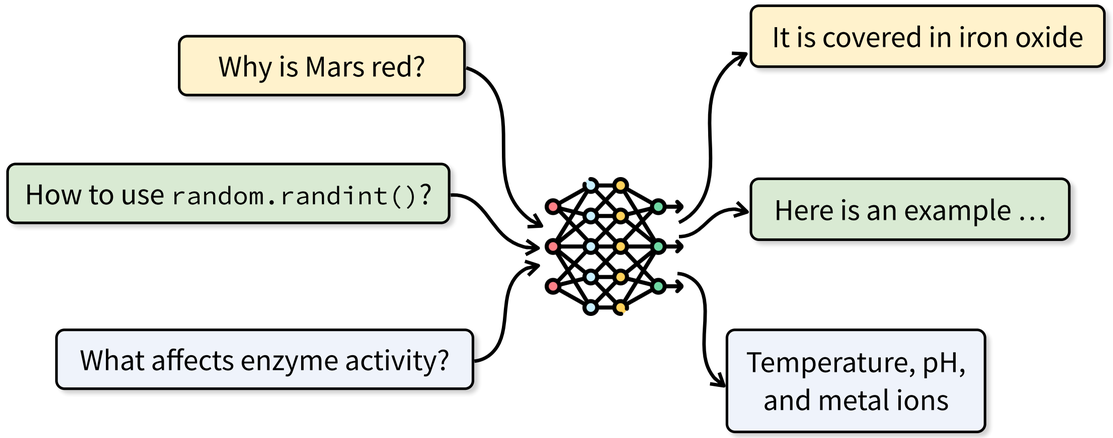
The current focus:

Pre-training data

↓

internalize

“knowledge”



More data
Bigger models
More compute

Next Frontier:

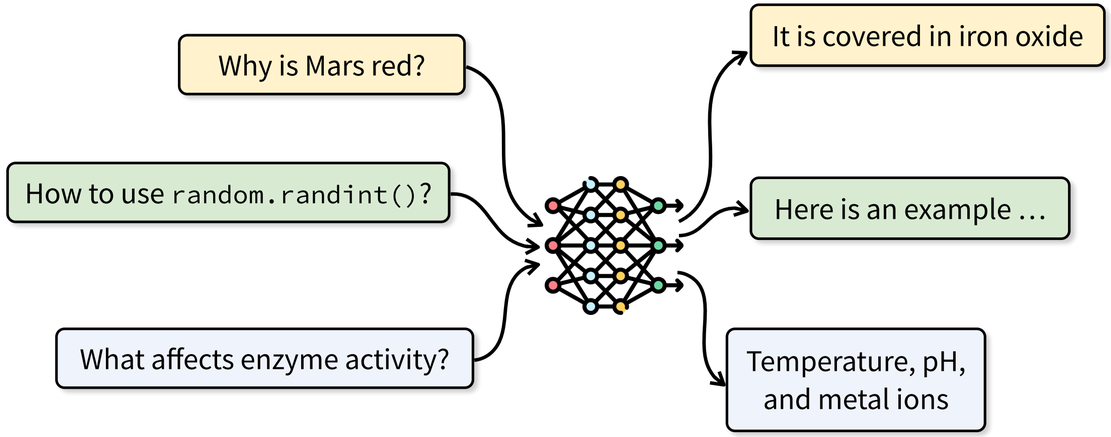
The current focus:

Pre-training data

↓

internalize

“knowledge”



More data
Bigger models
More compute

Next Frontier:

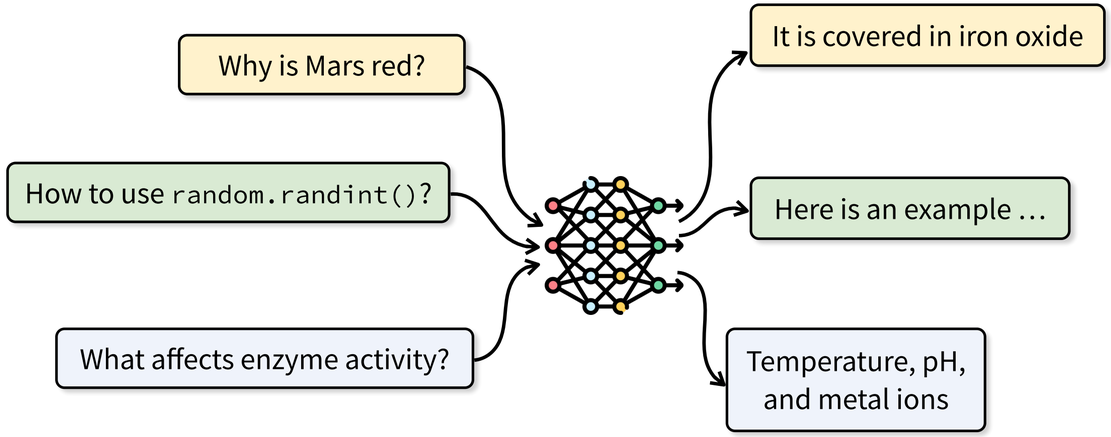
The current focus:

Pre-training data

↓

internalize

“knowledge”



More data
Bigger models
More compute

Next Frontier:

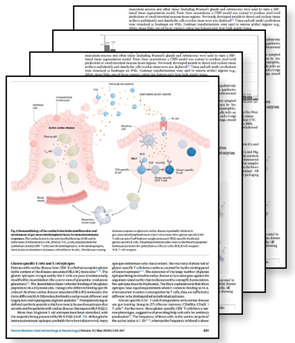
Input context

↓

derive new insights on the fly

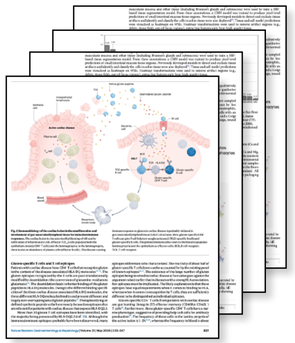
Next Frontier: Deriving new insights from input context on the fly

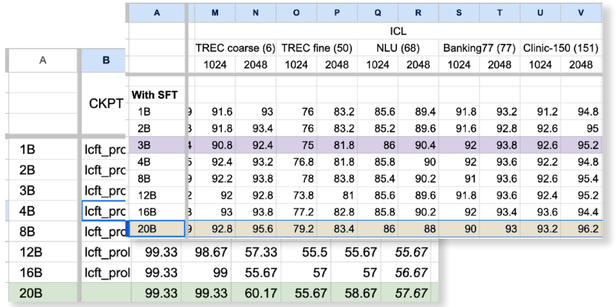
Next Frontier: Deriving new insights from input context on the fly



Summarizing a dozen papers into coherent **understanding**

Next Frontier: Deriving new insights from input context on the fly



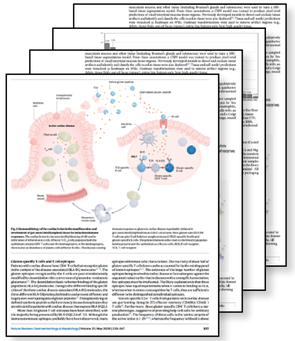


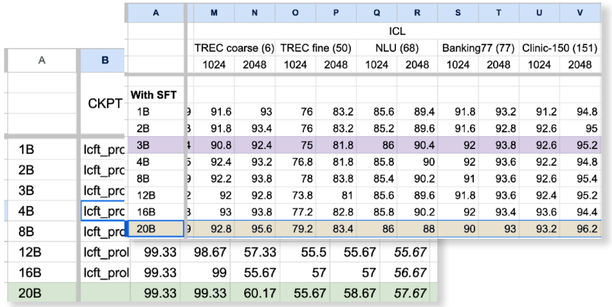
			ICL								
			TREC coarse (6)		TREC fine (50)		NLU (68)		Banking77 (77)		Clinic-150 (151)	
			1024	2048	1024	2048	1024	2048	1024	2048	1024	2048
	CKPT	With SFT										
		1B	91.6	93	76	83.2	85.6	89.4	91.8	93.2	91.2	94.8
		2B	91.8	93.4	76	83.2	85.2	89.6	91.6	92.8	92.6	95
1B	lcft_prc	3B	90.8	92.4	75	81.8	86	90.4	92	93.8	92.6	95.2
2B	lcft_prc	4B	92.4	93.2	76.8	81.8	85.8	90	92	93.6	92.2	94.8
3B	lcft_prc	8B	92.2	93.8	78	83.8	85.4	90.2	91	93.6	92.6	95.4
4B	lcft_prc	12B	92	92.8	73.8	81	85.6	89.6	91.8	93.6	92.4	95.2
8B	lcft_prc	16B	93	93.8	77.2	82.8	85.8	90.2	92	93.4	93.6	94.4
12B	lcft_prol	20B	92.8	95.6	79.2	83.4	86	88	90	93	93.2	96.2
16B	lcft_prol	99.33	98.67	57.33	55.5	55.67	*55.67*					
20B		99.33	99	55.67	57	57	*56.67*					
		99.33	99.33	60.17	55.67	58.67	*57.67*					

Summarizing a dozen papers into coherent **understanding**

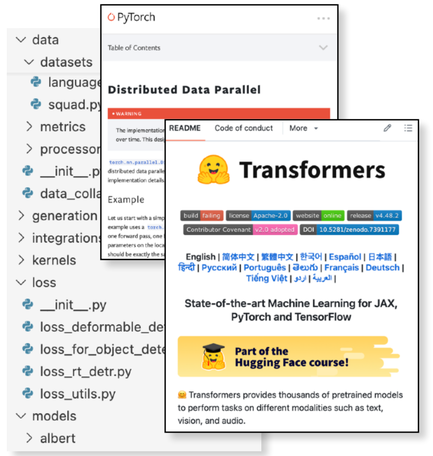
Condensing experiment data into **patterns**

Next Frontier: Deriving new insights from input context on the fly





A	B	A	M	N	O	P	Q	R	S	T	U	V
			ICL									
			TREC coarse (6)		TREC fine (50)		NLU (68)		Banking77 (77)		Clinic-150 (151)	
	CKPT		1024	2048	1024	2048	1024	2048	1024	2048	1024	2048
		With SFT										
		1B	91.6	93	76	83.2	85.6	89.4	91.8	93.2	91.2	94.8
		2B	91.8	93.4	76	83.2	85.2	89.6	91.6	92.8	92.6	95
1B	lcft_prc	3B	90.8	92.4	75	81.8	86	90.4	92	93.8	92.6	95.2
2B	lcft_prc	4B	92.4	93.2	76.8	81.8	85.8	90	92	93.6	92.2	94.8
3B	lcft_prc	8B	92.2	93.8	78	83.8	85.4	90.2	91	93.6	92.6	95.4
4B	lcft_prc	12B	92	92.8	73.8	81	85.6	89.6	91.8	93.6	92.4	95.2
8B	lcft_prc	16B	93	93.8	77.2	82.8	85.8	90.2	92	93.4	93.6	94.4
12B	lcft_prol	20B	92.8	95.6	79.2	83.4	86	88	90	93	93.2	96.2
16B	lcft_prol	99.33	98.67	57.33	55.5	55.67	*55.67*					
20B		99.33	99	55.67	57	57	*56.67*					
		99.33	99.33	60.17	55.67	58.67	*57.67*					

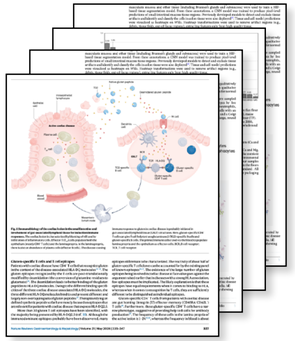


Summarizing a dozen papers into coherent **understanding**

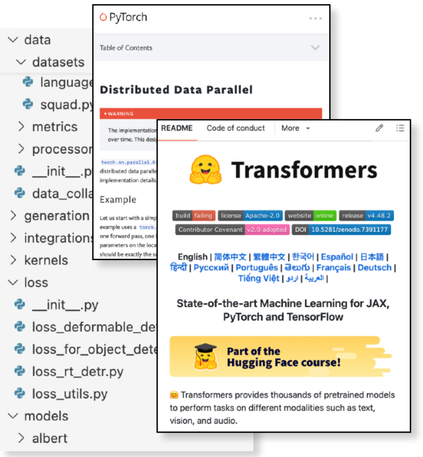
Condensing experiment data into **patterns**

Integrating API documents for implementing a new **feature**

Next Frontier: Deriving new insights from input context on the fly



	A	M	N	O	P	Q	R	S	T	U	V
		ICL									
		TREC coarse (6)		TREC fine (50)		NLU (68)		Banking77 (77)		Clinic-150 (151)	
		1024	2048	1024	2048	1024	2048	1024	2048	1024	2048
A	B										
	CKPT	With SFT									
		1B	91.6	93	76	83.2	85.6	89.4	91.8	93.2	91.2	94.8
		2B	91.8	93.4	76	83.2	85.2	89.6	91.6	92.8	92.6	95
1B	lcft_prc	3B	90.8	92.4	75	81.8	86	90.4	92	93.8	92.6	95.2
2B	lcft_prc	4B	92.4	93.2	76.8	81.8	85.8	90	92	93.6	92.2	94.8
3B	lcft_prc	8B	92.2	93.8	78	83.8	85.4	90.2	91	93.6	92.6	95.4
4B	lcft_prc	12B	92	92.8	73.8	81	85.6	89.6	91.8	93.6	92.4	95.2
8B	lcft_prc	16B	93	93.8	77.2	82.8	85.8	90.2	92	93.4	93.6	94.4
12B	lcft_prol	20B	92.8	95.6	79.2	83.4	86	88	90	93	93.2	96.2
16B	lcft_prol	99.33	98.67	57.33	55.5	55.67	*55.67*				
20B		99.33	99	55.67	57	57	*56.67*				
		99.33	99.33	60.17	55.67	58.67	*57.67*				



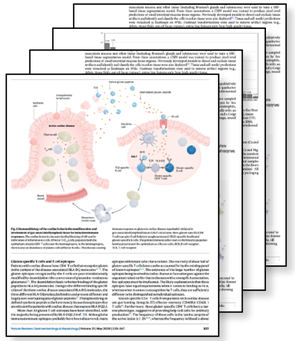
Use **LMs** to *compile*, *synthesize*, and *reason over* input data on the fly

Summarizing a dozen papers into coherent **understanding**

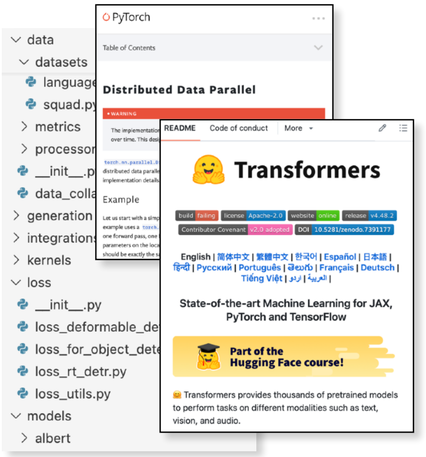
Condensing experiment data into **patterns**

Integrating API documents for implementing a new **feature**

Next Frontier: Deriving new insights from input context on the fly



	A	M	N	O	P	Q	R	S	T	U	V
		ICL									
		TREC coarse (6)		TREC fine (50)		NLU (68)		Banking77 (77)		Clinic-150 (151)	
		1024	2048	1024	2048	1024	2048	1024	2048	1024	2048
A	B										
	CKPT	With SFT									
		1B	91.6	93	76	83.2	85.6	89.4	91.8	93.2	91.2	94.8
		2B	91.8	93.4	76	83.2	85.2	89.6	91.6	92.8	92.6	95
1B	lcft_prc	3B	90.8	92.4	75	81.8	86	90.4	92	93.8	92.6	95.2
2B	lcft_prc	4B	92.4	93.2	76.8	81.8	85.8	90	92	93.6	92.2	94.8
3B	lcft_prc	8B	92.2	93.8	78	83.8	85.4	90.2	91	93.6	92.6	95.4
4B	lcft_prc	12B	92	92.8	73.8	81	85.6	89.6	91.8	93.6	92.4	95.2
8B	lcft_prc	16B	93	93.8	77.2	82.8	85.8	90.2	92	93.4	93.6	94.4
12B	lcft_prol	20B	92.8	95.6	79.2	83.4	86	88	90	93	93.2	96.2
16B	lcft_prol	99.33	98.67	57.33	55.5	55.67	*55.67*				
20B		99.33	99	55.67	57	57	*56.67*				
		99.33	99.33	60.17	55.67	58.67	*57.67*				



Use **LMs** to *compile*, *synthesize*, and *reason over* input data on the fly

Boost productivity

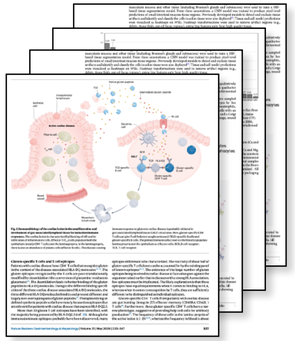
Accelerate scientific discovery

Summarizing a dozen papers into coherent **understanding**

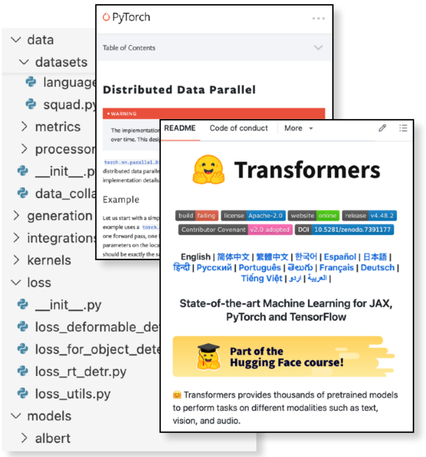
into **patterns**

Integrating API documents for implementing a new **feature**

Next Frontier: Deriving new insights from input context on the fly



			TREC coarse (6)		TREC fine (50)		NLU (68)		Banking77 (77)		Clinic-150 (151)	
			1024	2048	1024	2048	1024	2048	1024	2048	1024	2048
	CKPT	With SFT										
		1B	91.6	93	76	83.2	85.6	89.4	91.8	93.2	91.2	94.8
		2B	91.8	93.4	76	83.2	85.2	89.6	91.6	92.8	92.6	95
1B	lcft_prc	3B	90.8	92.4	75	81.8	86	90.4	92	93.8	92.6	95.2
2B	lcft_prc	4B	92.4	93.2	76.8	81.8	85.8	90	92	93.6	92.2	94.8
3B	lcft_prc	8B	92.2	93.8	78	83.8	85.4	90.2	91	93.6	92.6	95.4
4B	lcft_prc	12B	92	92.8	73.8	81	85.6	89.6	91.8	93.6	92.4	95.2
8B	lcft_prc	16B	93	93.8	77.2	82.8	85.8	90.2	92	93.4	93.6	94.4
12B	lcft_prol	20B	92.8	95.6	79.2	83.4	86	88	90	93	93.2	96.2
16B	lcft_prol	99.33	98.67	57.33	55.5	55.67	*55.67*					
20B		99.33	99	55.67	57	57	*56.67*					
			99.33	99.33	60.17	55.67	58.67	*57.67*				



Use **LMs** to **process** (=*compile*, *synthesize*, and *reason over*) input data on the fly

Boost productivity

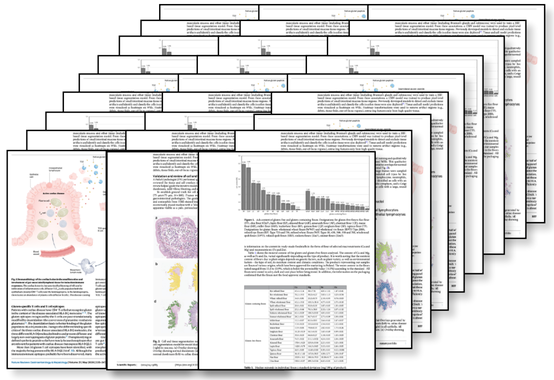
Accelerate scientific discovery

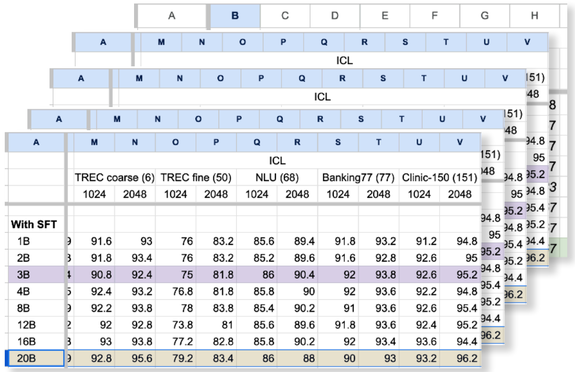
Summarizing a dozen papers into coherent **understanding**

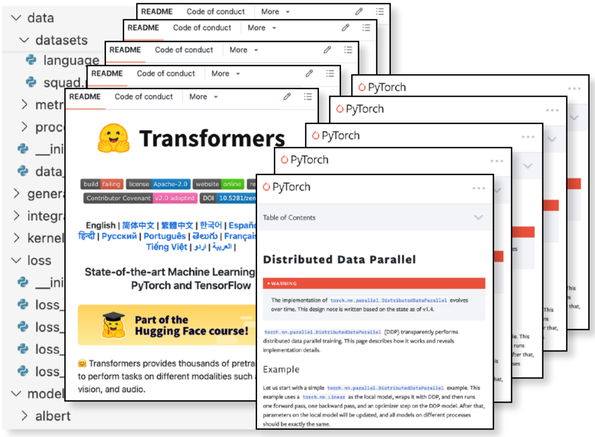
into **patterns**

Integrating API documents for implementing a new **feature**

Next Frontier: Deriving new insights from input context on the fly







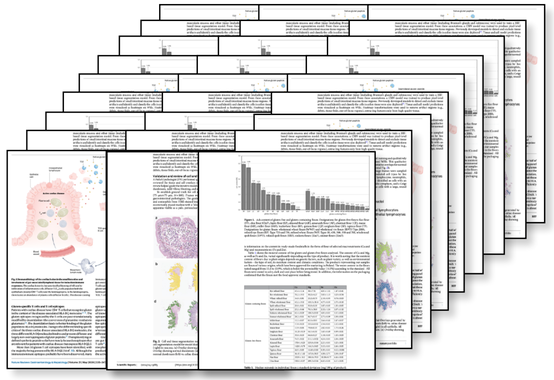
New insights often come from understanding **a large volume** of materials

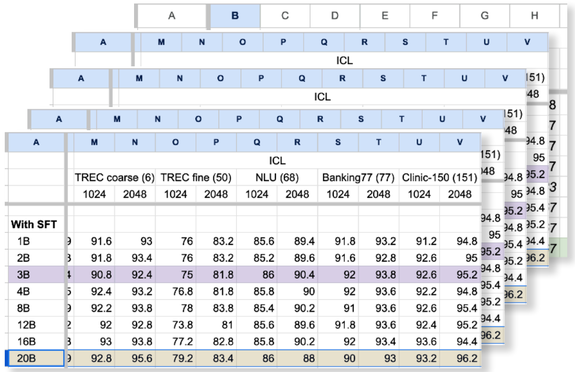
Summarizing a dozen papers into coherent **understanding**

Condensing experiment data into **patterns**

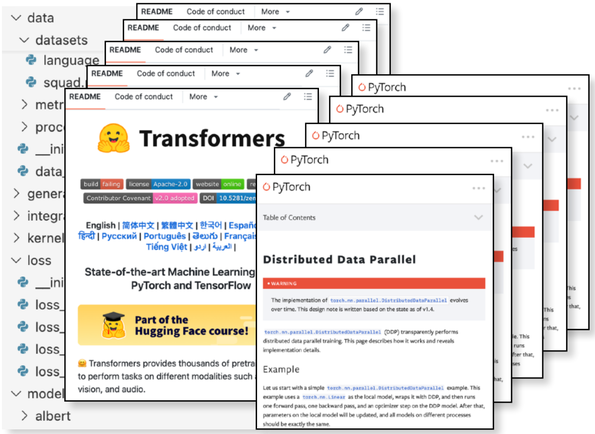
Integrating API documents for implementing a new **feature**

Next Frontier: Deriving new insights from input context on the fly





	TREC coarse (6)		TREC fine (50)		NLU (68)		Banking77 (77)		Clinic-150 (151)	
	1024	2048	1024	2048	1024	2048	1024	2048	1024	2048
With SFT										
1B	91.6	93	76	83.2	85.6	89.4	91.8	93.2	91.2	94.8
2B	91.8	93.4	76	83.2	85.2	89.6	91.6	92.8	92.6	95
3B	90.8	92.4	75	81.8	86	90.4	92	93.8	92.6	95.2
4B	92.4	93.2	76.8	81.8	85.8	90	92	93.6	92.2	94.8
8B	92.2	93.8	78	83.8	85.4	90.2	91	93.6	92.6	95.4
12B	92	92.8	73.8	81	85.6	89.6	91.8	93.6	92.4	95.2
16B	93	93.8	77.2	82.8	85.8	90.2	92	93.4	93.6	94.4
20B	92.8	95.6	79.2	83.4	86	88	90	93	93.2	96.2



New insights often come from understanding **a large volume** of materials

Challenge: **process** vast amounts of input information

Summarizing a c… into coherent **understanding** … **into patterns** … PI documents for **implementing** a new **feature**

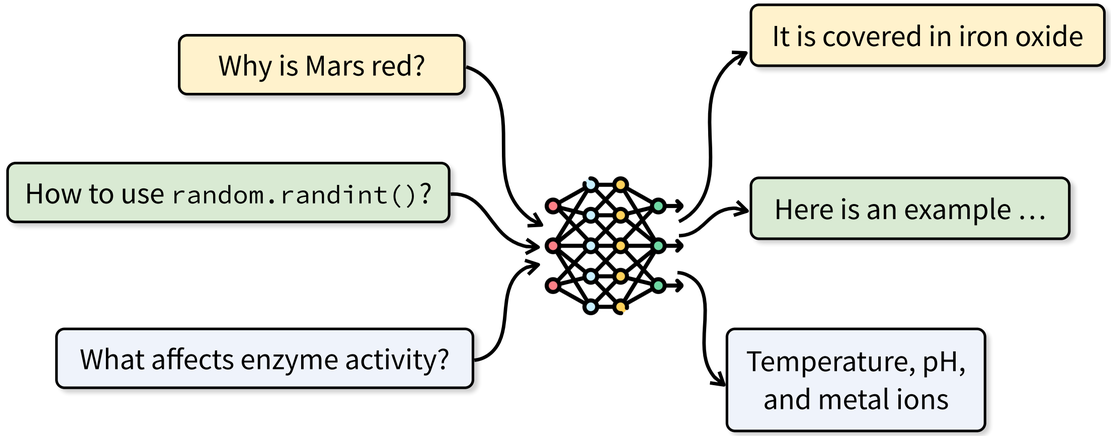
The current focus:

Pre-training data

↓

internalize

“knowledge”



More data
Bigger models
More compute

Next Frontier:

Input context

↓

derive new insights on the fly

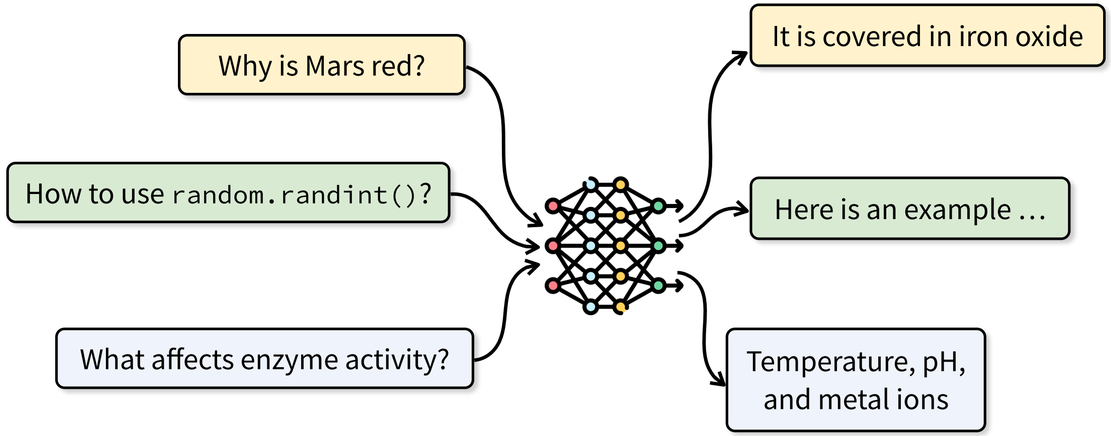
The current focus:

Pre-training data

↓

internalize

“knowledge”



More data
Bigger models
More compute

Next Frontier:

Input context

↓

derive new insights on the fly

?

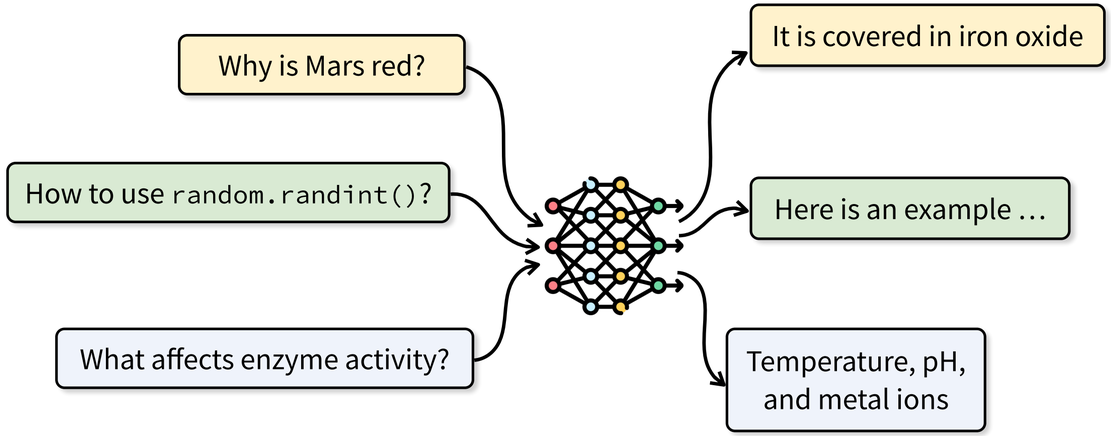
The current focus:

Pre-training data

↓

internalize

“knowledge”



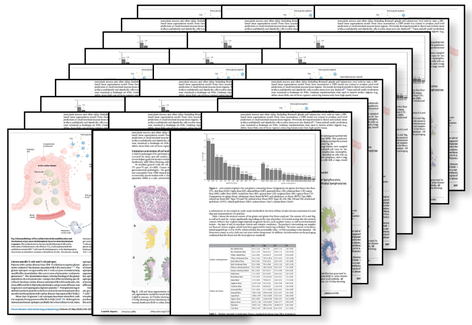
More data
Bigger models
More compute

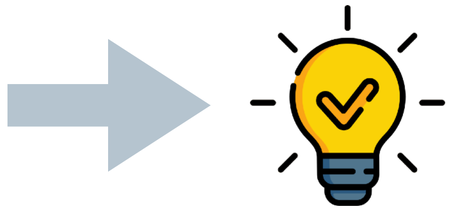
Next Frontier:

Input context

↓

derive new insights on the fly





?

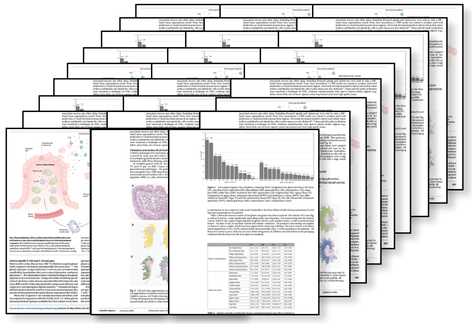
My work: **Enabling LMs to process information at scale**

My work: Enabling LMs to process information at scale

1. Effective long-context processing

My work: Enabling LMs to process information at scale

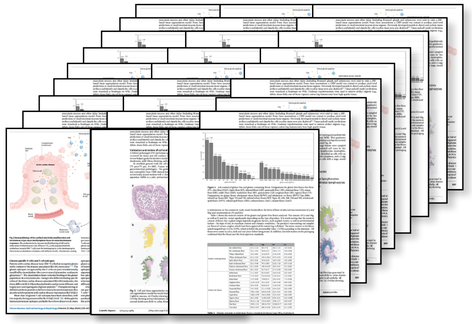
1. Effective long-context processing



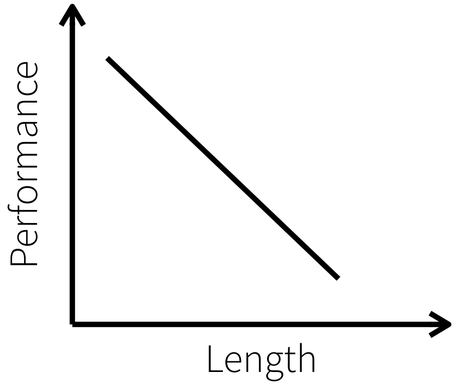
Applications require long input lengths

My work: Enabling LMs to process information at scale

1. Effective long-context processing



Applications require long input lengths



Current LMs degrade with longer inputs

My work: Enabling LMs to process information at scale

1. Effective long-context processing

GYYC EMNLP23; P**G**CC Findings of ACL23; Y**G**+ ICLR25; Y**G**C ACL24, **G***W*YC 2024

My work: Enabling LMs to process information at scale

1. Effective long-context processing

- Desiderata for long-context capabilities

GYYC EMNLP23; P**G**CC Findings of ACL23; Y**G**+ ICLR25; Y**G**C ACL24, **G***W*YC 2024

My work: Enabling LMs to process information at scale

1. Effective long-context processing

- Desiderata for long-context capabilities → building robust evaluation

GYYC EMNLP23; P**G**CC Findings of ACL23; Y**G**+ ICLR25; Y**G**C ACL24, **G***W*YC 2024

My work: Enabling LMs to process information at scale

1. Effective long-context processing

- Desiderata for long-context capabilities → building robust evaluation

Widely adopted by academia and industry researchers

GYYC EMNLP23; P**G**CC Findings of ACL23; Y**G**+ ICLR25; Y**G**C ACL24, **G***W*YC 2024

My work: Enabling LMs to process information at scale

1. Effective long-context processing

- Desiderata for long-context capabilities → building robust evaluation

Widely adopted by academia and industry researchers

- Developing state-of-the-art long-context models

GYYC EMNLP23; P**G**CC Findings of ACL23; Y**G**+ ICLR25; Y**G**C ACL24, **G***W*YC 2024

My work: Enabling LMs to process information at scale

1. Effective long-context processing

- Desiderata for long-context capabilities → building robust evaluation

Widely adopted by academia and industry researchers

- Developing state-of-the-art long-context models

Outperform industry models with a fraction of the compute

GYYC EMNLP23; P**G**CC Findings of ACL23; Y**G**+ ICLR25; Y**G**C ACL24, **G***W*YC 2024

My work: Enabling LMs to process information at scale

1. Effective long-context processing

2. Accurate search

G*Y*C EMNLP21; AX+**G** EMNLP24

My work: Enabling LMs to process information at scale

1. Effective long-context processing

2. Accurate search

G*Y*C EMNLP21; AX+**G** EMNLP24

My work: Enabling LMs to process information at scale

1. Effective long-context processing

2. Accurate search

Traditional keyword-based search

cannot accurately find relevant materials

G*Y*C EMNLP21; AX+**G** EMNLP24

My work: Enabling LMs to process information at scale

1. Effective long-context processing

2. Accurate search via text embeddings

G*Y*C EMNLP21; AX+**G** EMNLP24

My work: Enabling LMs to process information at scale

1. Effective long-context processing

2. Accurate search via text embeddings

= represent text in vector forms
better capture semantic meaning

G*Y*C EMNLP21; AX+**G** EMNLP24

My work: Enabling LMs to process information at scale

1. Effective long-context processing

2. Accurate search via text embeddings

= represent text in vector forms
better capture semantic meaning

- Building foundational techniques that turn LMs into text embeddings

G*Y*C EMNLP21; AX+**G** EMNLP24

My work: Enabling LMs to process information at scale

1. Effective long-context processing

2. Accurate search via text embeddings

= represent text in vector forms
better capture semantic meaning

- Building foundational techniques that turn LMs into text embeddings

Downloads ⏬
> 22M

G*Y*C EMNLP21; AX+**G** EMNLP24

My work: Enabling LMs to process information at scale

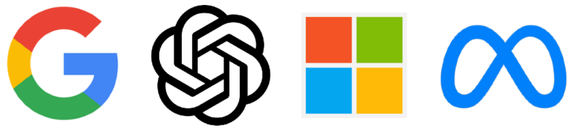
1. Effective long-context processing

2. Accurate search via text embeddings

= represent text in vector forms
better capture semantic meaning

- Building foundational techniques that turn LMs into text embeddings

Downloads ⬇️
> 22M



G*Y*C EMNLP21; AX+**G** EMNLP24

My work: Enabling LMs to process information at scale

1. Effective long-context processing
2. Accurate search via text embeddings
3. **Foundations: Efficient language models**

G*F*C ACL21; M***G***+ NeurIPS23 oral; W***G***ZC EACL23; X**G**ZC ICLR24; **G**W+ 2025

My work: Enabling LMs to process information at scale

1. Effective long-context processing
2. Accurate search via text embeddings
3. **Foundations: Efficient language models**

- Producing capable small models

G*F*C ACL21; M***G***+ NeurIPS23 oral; W***G***ZC EACL23; X**G**ZC ICLR24; **G**W+ 2025

My work: Enabling LMs to process information at scale

1. Effective long-context processing
2. Accurate search via text embeddings
3. **Foundations: Efficient language models**

- Producing capable small models
- Efficient training and customization methods

G*F*C ACL21; M***G***+ NeurIPS23 oral; W***G***ZC EACL23; X**G**ZC ICLR24; **G**W+ 2025

My work: Enabling LMs to process information at scale

1. Effective long-context processing
2. Accurate search via text embeddings
3. **Foundations: Efficient language models**

- Producing capable small models
- Efficient training and customization methods

Democratize LMs

G*F*C ACL21; M***G***+ NeurIPS23 oral; W***G***ZC EACL23; X**G**ZC ICLR24; **G**W+ 2025

My work: Enabling LMs to process information at scale

1. Effective long-context processing
2. Accurate search via text embeddings
3. **Foundations: Efficient language models**

- Producing capable small models
- Efficient training and customization methods

Democratize LMs

Enable broader and on-device use

G*F*C ACL21; M***G***+ NeurIPS23 oral; W***G***ZC EACL23; X**G**ZC ICLR24; **G**W+ 2025

Enabling LMs to process information at scale

1. Effective long-context processing

2. Accurate search via text embeddings

3. Foundations: Efficient language models

Enabling LMs to process information at scale

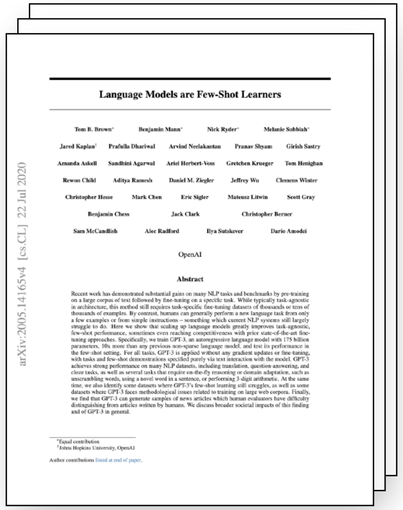
1. Effective long-context processing

2. Accurate search via text embeddings

3. Foundations: Efficient language models

Long-context LMs support numerous applications

Long-context LMs support numerous applications



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

*Equal contribution
†Johns Hopkins University, OpenAI

Author contributions listed at end of paper.

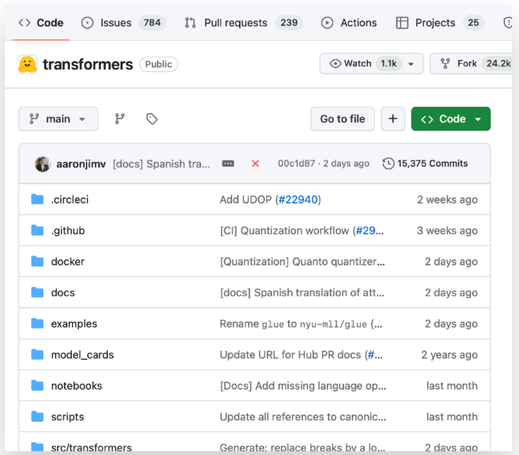
arXiv:2005.14165v4 [cs.CL] 22 Jul 2020

The GPT-4 paper

(~**300K** tokens)

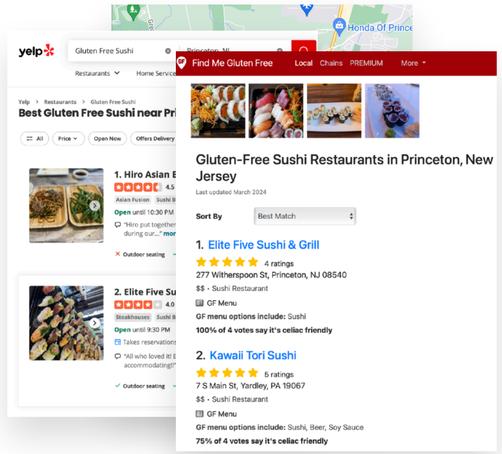
The Dune series

(~**2M** tokens)



The Transformers package

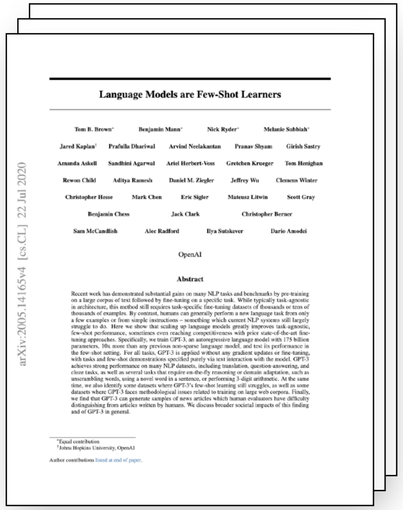
(~**10M** tokens)



100 web pages

(**~100K** tokens)

Long-context LMs support numerous applications



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

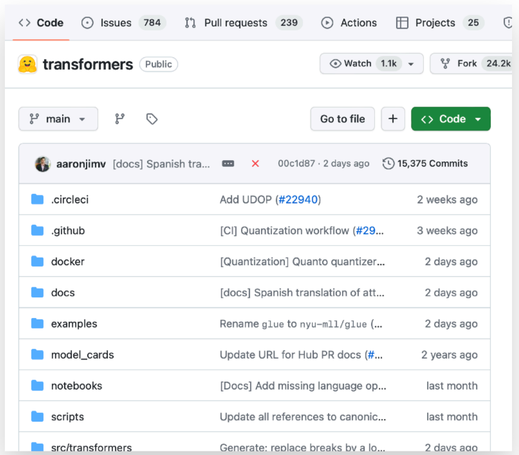
Abstract

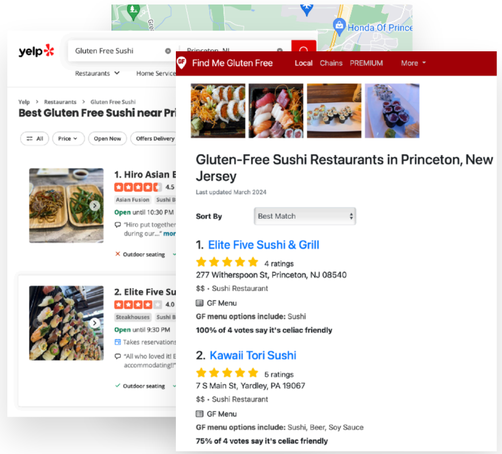
Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

*Equal contribution
†Johns Hopkins University, OpenAI

Author contributions listed at end of paper.

arXiv:2005.14165v4 [cs.CL] 22 Jul 2020





The GPT-4 paper (~**300K** tokens)

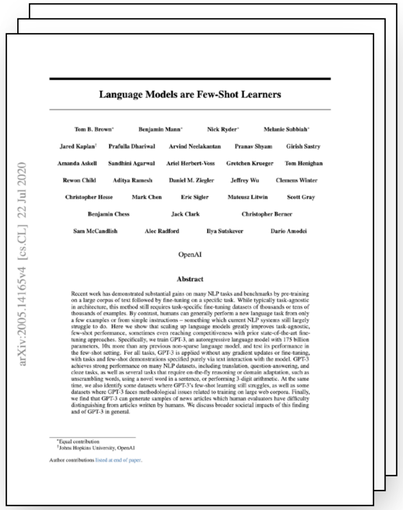
The Dune series (~**2M** tokens)

The Transformers package (~**10M** tokens)

100 web pages (**~100K** tokens)

Basic unit of LM input

Long-context LMs support numerous applications



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

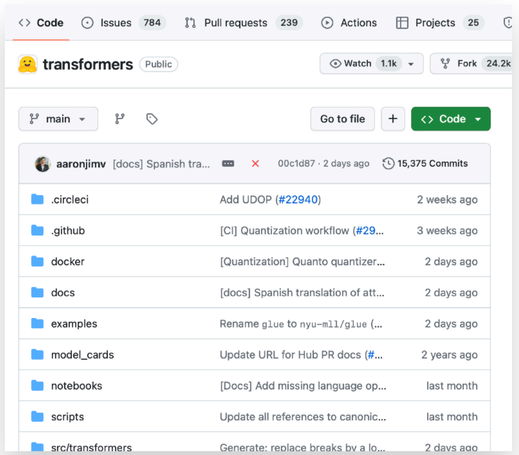
Abstract

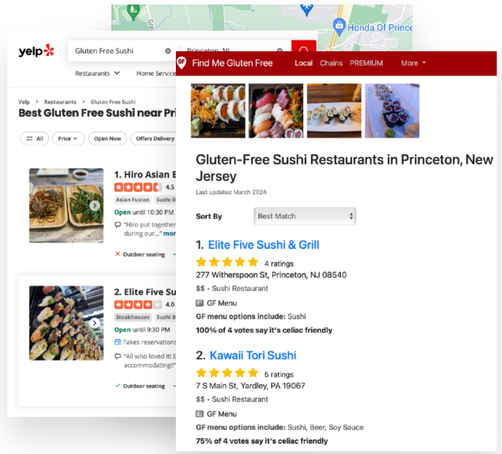
Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

*Equal contribution
†Johns Hopkins University, OpenAI

Author contributions listed at end of paper.

arXiv:2005.14165v4 [cs.CL] 22 Jul 2020





The GPT-4 paper (~**300K** tokens)

The Dune series (~**2M** tokens)

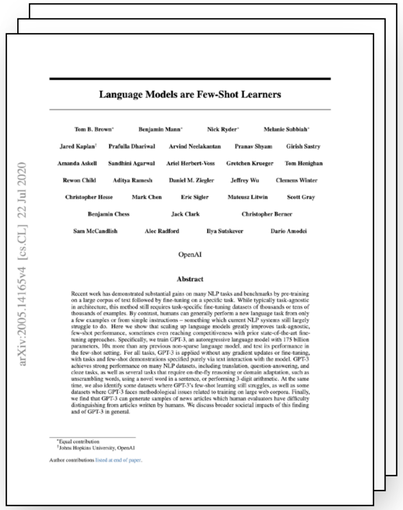
The Transformers package (~**10M** tokens)

100 web pages (**~100K** tokens)

→ *Basic unit of LM input*

In 2023, state-of-the-art language models have very limited context windows

Long-context LMs support numerous applications



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

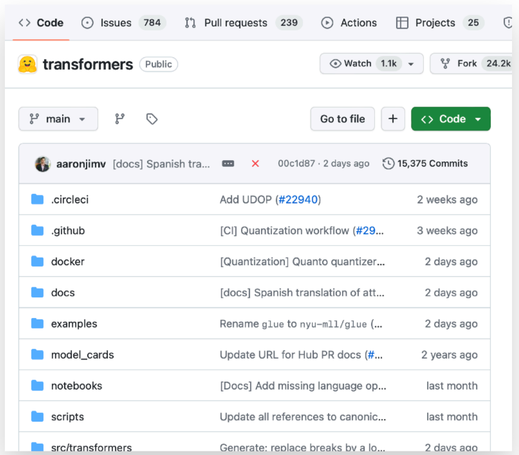
Abstract

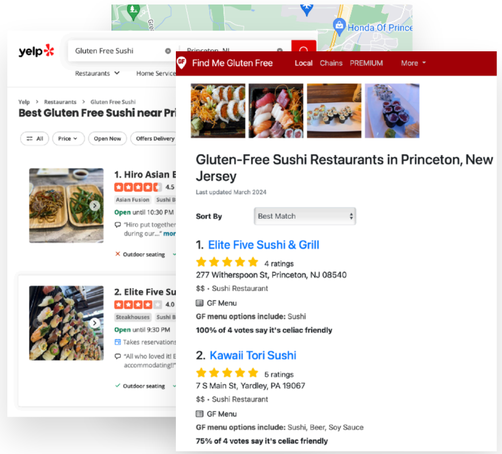
Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

*Equal contribution
†Johns Hopkins University, OpenAI

Author contributions listed at end of paper.

arXiv:2005.14165v4 [cs.CL] 22 Jul 2020





The GPT-4 paper (~**300K** tokens)

The Dune series (~**2M** tokens)

The Transformers package (~**10M** tokens)

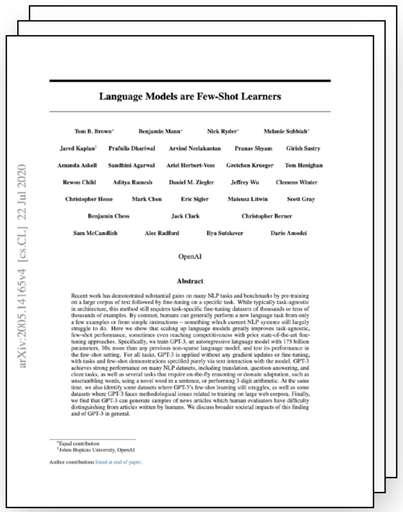
100 web pages (**~100K** tokens)

→ *Basic unit of LM input*

In 2023, state-of-the-art language models have very limited context windows

Meta 4K tokens

Long-context LMs support numerous applications



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

*Equal contribution
†Johns Hopkins University, OpenAI

Author contributions listed at end of paper.

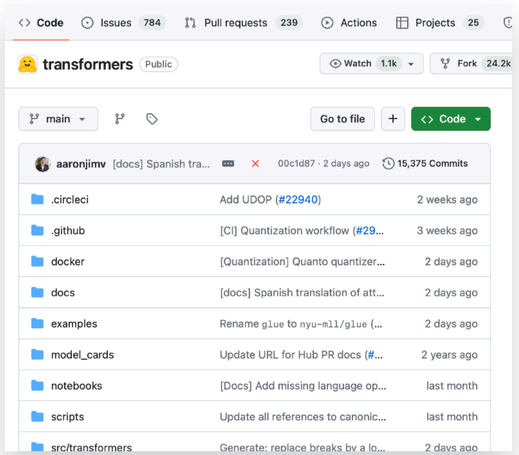
arXiv:2005.14165v4 [cs.CL] 22 Jul 2020

The GPT-4 paper

(~**300K** tokens)

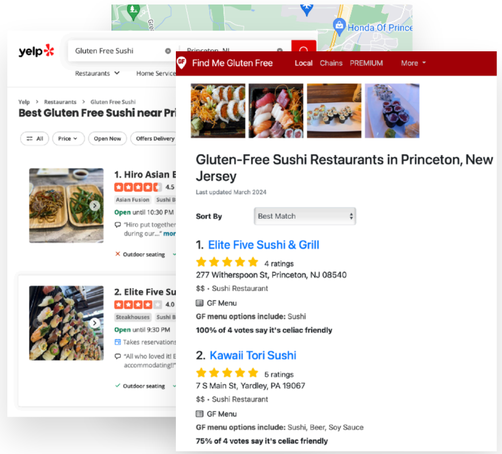
The Dune series

(~**2M** tokens)



The Transformers package

(~**10M** tokens)



100 web pages

(**~100K** tokens)

Basic unit of LM input

In 2023, state-of-the-art language models have very limited context windows

Meta 4K tokens

8K tokens

Why is long-context processing challenging?

Why is long-context processing challenging?

Usually context window = training input length

Why is long-context processing challenging?

Usually context window = training input length

Most LMs (Transformers):

- **Quadratic** computational complexity
- **Linear** memory complexity

w.r.t. **input length**

Why is long-context processing challenging?

Usually context window = training input length

Most LMs (Transformers):

- **Quadratic** computational complexity
- **Linear** memory complexity

w.r.t. **input length**

Cannot simply increase the input length in pre-training

Why is long-context processing challenging?

Usually context window = training input length

Most LMs (Transformers):

- **Quadratic** computational complexity
- **Linear** memory complexity

w.r.t. **input length**

Cannot simply increase the input length in pre-training

Context window ≠ effective length

Why is long-context processing challenging?

Usually context window = training input length

Most LMs (Transformers):

- **Quadratic** computational complexity
- **Linear** memory complexity

w.r.t. **input length**

Cannot simply increase the input length in pre-training

Context window ≠ effective length

Long-context ability doesn't come for free!

Why is long-context processing challenging?

Usually context window = training input length

Most LMs (Transformers):

- **Quadratic** computational complexity
- **Linear** memory complexity

w.r.t. **input length**

Cannot simply increase the input length in pre-training

Context window ≠ effective length

Long-context ability doesn't come for free!

But first, how do we **measure** a model's "long-context capability"?

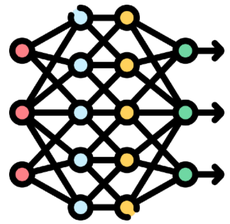
Needle in a haystack

(Liu et al., July 2023; Kamradt, Nov 2023)

Needle in a haystack

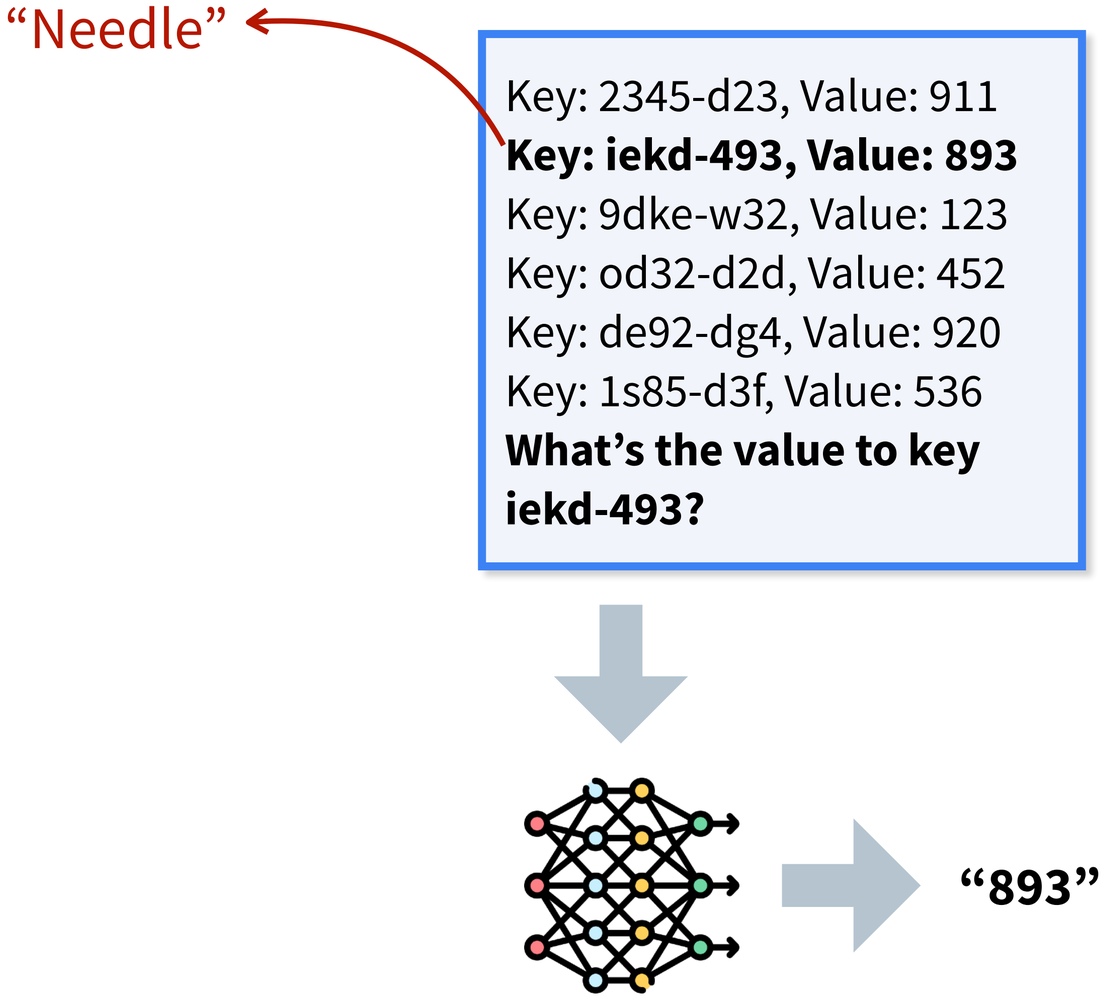
(Liu et al., July 2023; Kamradt, Nov 2023)

Key: 2345-d23, Value: 911
Key: iekd-493, Value: 893
Key: 9dke-w32, Value: 123
Key: od32-d2d, Value: 452
Key: de92-dg4, Value: 920
Key: 1s85-d3f, Value: 536
What's the value to key iekd-493?



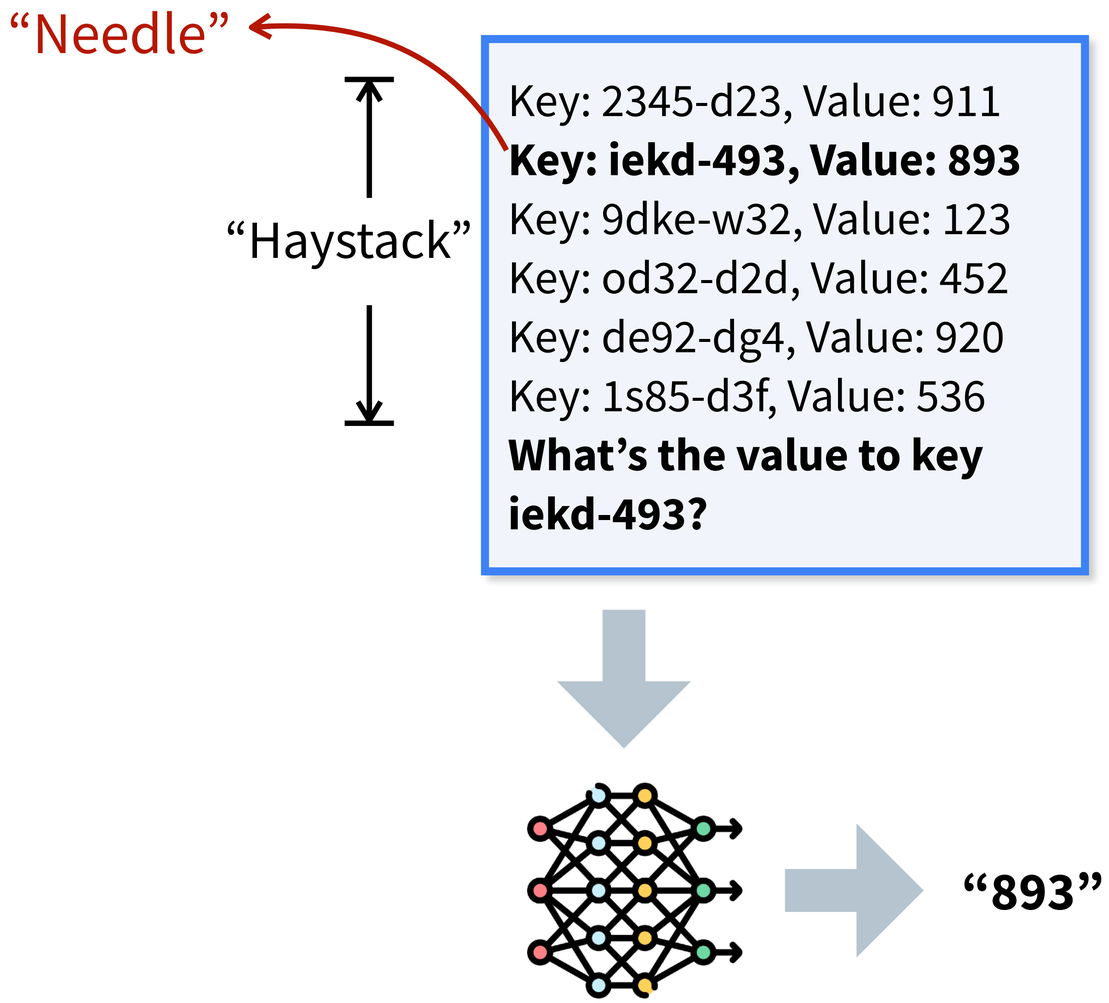
Needle in a haystack

(Liu et al., July 2023; Kamradt, Nov 2023)



Needle in a haystack

(Liu et al., July 2023; Kamradt, Nov 2023)



Needle in a haystack ● ▲
(Liu et al., July 2023; Kamradt, Nov 2023)

Recall ●

Controllable, synthetic evaluation ▲

Needle in a haystack ● ▲

Comprehensive long-context capabilities

Recall ●

Controllable, synthetic evaluation ▲

Needle in a haystack ● ▲

Comprehensive long-context capabilities

Recall ●

Reasoning and synthesis ●

Controllable, synthetic evaluation ▲

Needle in a haystack ● ▲

Comprehensive long-context capabilities

Recall ●

Reasoning and synthesis ●

Learning new tasks ●

Controllable, synthetic evaluation ▲

Needle in a haystack ● ▲

Comprehensive long-context capabilities

Recall ●

Reasoning and synthesis ●

Learning new tasks ●

Diverse task types

Controllable, synthetic evaluation ▲

Needle in a haystack ● ▲

Comprehensive long-context capabilities

Recall ●

Reasoning and synthesis ●

Learning new tasks ●

Diverse task types

Controllable, synthetic evaluation ▲

Real-world applications ▲

HELMET: Holistic long-context LM evaluation

Needle in a haystack ● ▲

Comprehensive long-context capabilities

Recall ●

Reasoning and synthesis ●

Learning new tasks ●

Diverse task types

Controllable, synthetic evaluation ▲

Real-world applications ▲

H. Yen, **T. Gao**, et al. ICLR 2025. **HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly**.

HELMET: Holistic long-context LM evaluation

Tasks

Needle in a haystack ● ▲

Retrieval-augmented generation ● ▲

Passage re-ranking ● ● ▲

Many-shot in-context learning ● ▲

Long-document QA ● ▲

Long-document summarization ● ▲

Generation with citations (ALCE) ● ● ▲

Comprehensive long-context capabilities

Recall ●

Reasoning and synthesis ●

Learning new tasks ●

Diverse task types

Controllable, synthetic evaluation ▲

Real-world applications ▲

H. Yen, **T. Gao**, et al. ICLR 2025. **HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly**.

HELMET: Holistic long-context LM evaluation

Tasks

Needle in a haystack ● ▲

Retrieval-augmented generation ● ▲

Passage re-ranking ● ● ▲

Many-shot in-context learning ● ▲

Long-document QA ● ▲

Long-document summarization ● ▲

Generation with citations (ALCE) ● ● ▲

Comprehensive long-context capabilities

Recall ●

Reasoning and synthesis ●

Learning new tasks ●

Diverse task types

Controllable, synthetic evaluation ▲

Real-world applications ▲

H. Yen, **T. Gao**, et al. ICLR 2025. **HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly**.

HELMET: Holistic long-context LM evaluation

Tasks

Needle in a haystack ● ▲

Retrieval-augmented generation ● ▲

Passage re-ranking ● ● ▲

Many-shot in-context learning ● ▲

Long-document QA ● ▲

Long-document summarization ● ▲

Generation with citations (ALCE) ● ● ▲

In-context learning: the ability to "**learn**" a new task on the fly

HELMET: Holistic long-context LM evaluation

Tasks

Needle in a haystack ● ▲

Retrieval-augmented generation ● ▲

Passage re-ranking ● ● ▲

Many-shot in-context learning ● ▲

Long-document QA ● ▲

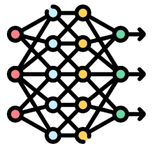
Long-document summarization ● ▲

Generation with citations (ALCE) ● ● ▲

Article: Nasdaq dropped 3% today
Topic: **Business**

Article: The 30 best TV shows to watch now
Topic: **Entertainment**
…

Article: The new Apple TV+ show “Severance”
Topic: _____



“Entertainment”

In-context learning: the ability to “**learn**” a new task on the fly

HELMET: Holistic long-context LM evaluation

Tasks

Needle in a haystack ● ▲

Retrieval-augmented generation ● ▲

Passage re-ranking ● ● ▲

Many-shot in-context learning ● ▲

Long-document QA ● ▲

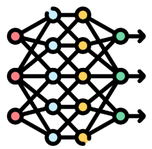
Long-document summarization ● ▲

Generation with citations (ALCE) ● ● ▲

Article: Nasdaq dropped 3% today
Topic: **Business**

Article: The 30 best TV shows to watch now
Topic: **Entertainment**
…

Article: The new Apple TV+ show “Severance”
Topic: _____



“Entertainment”

In-context learning: the ability to “**learn**” a new task on the fly

Many-shot: learn from **thousands of** examples

HELMET: Holistic long-context LM evaluation

Tasks

Needle in a haystack ● ▲

Retrieval-augmented generation ● ▲

Passage re-ranking ● ● ▲

Many-shot in-context learning ● ▲

Long-document QA ● ▲

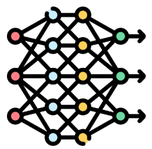
Long-document summarization ● ▲

Generation with citations (ALCE) ● ● ▲

Article: Nasdaq dropped 3% today
Label: ~~**Business**~~ **3**

Article: The 30 best TV shows to watch now
Label: ~~**Entertainment**~~ **5**
…

Article: The new Apple TV+ show “Severance”
Label: _____



~~**“Entertainment”**~~ **“5”**

Using **abstract** (e.g., numbers) labels in in-context learning better reflects “learning” [P**G**CC Findings of ACL23]

HELMET: Holistic long-context LM evaluation

Tasks

Needle in a haystack ● ▲

Retrieval-augmented generation ● ▲

Passage re-ranking ● ● ▲

Many-shot in-context learning ● ▲

Long-document QA ● ▲

Long-document summarization ● ▲

Generation with citations (ALCE) ● ● ▲

Comprehensive long-context capabilities

Recall ●

Reasoning and synthesis ●

Learning new tasks ●

Diverse task types

Controllable, synthetic evaluation ▲

Real-world applications ▲

H. Yen, **T. Gao**, et al. ICLR 2025. **HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly**.

Evaluation by new applications: Generation with citations

Evaluation by new applications: Generation with citations

- Given a **question** and **relevant documents**

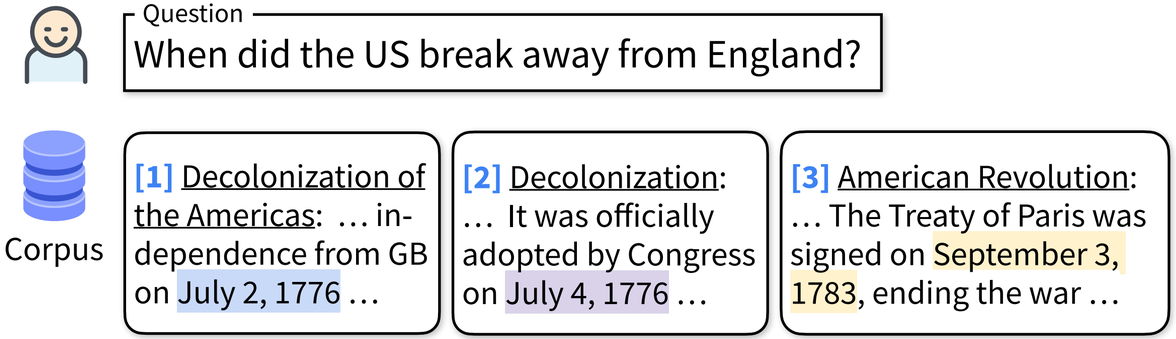
Evaluation by new applications: Generation with citations

Question

When did the US break away from England?

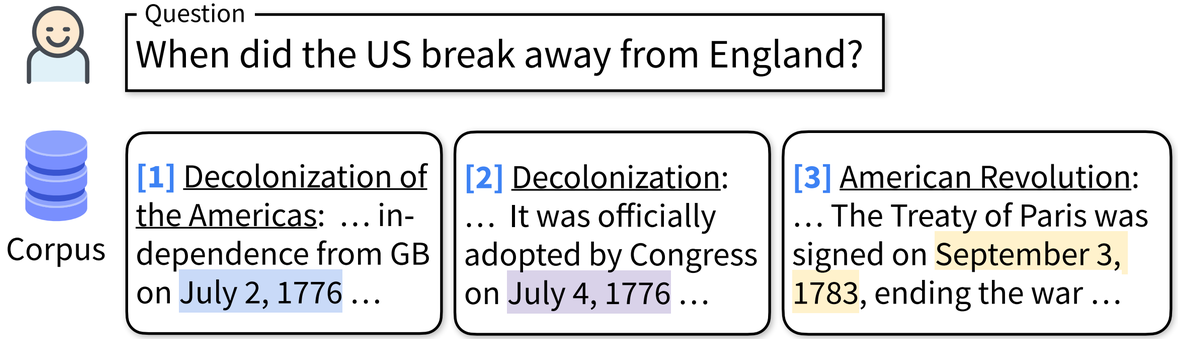
- Given a **question** and **relevant documents**

Evaluation by new applications: Generation with citations



- Given a **question** and **relevant documents**

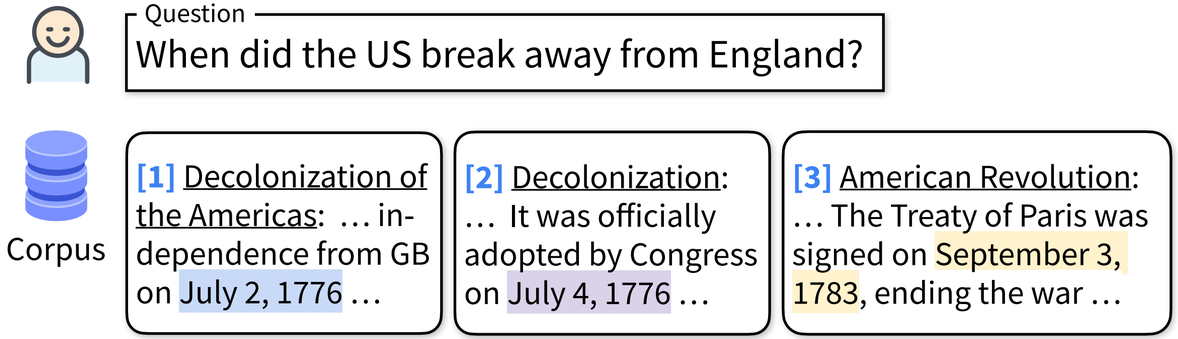
Evaluation by new applications: Generation with citations



Either provided by users or retrieved from Internet/databases

- Given a **question** and **relevant documents**

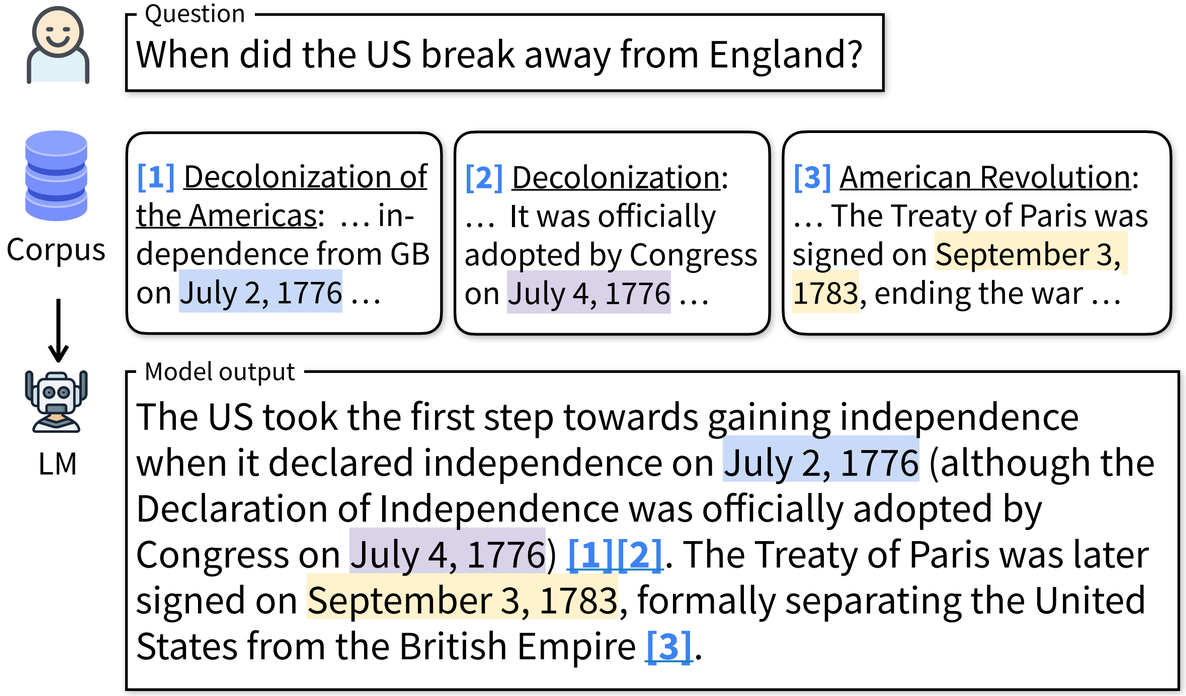
Evaluation by new applications: Generation with citations



Either provided by users or retrieved from Internet/databases

- Given a **question** and **relevant documents**
  - LMs to generate a concise, informative answer

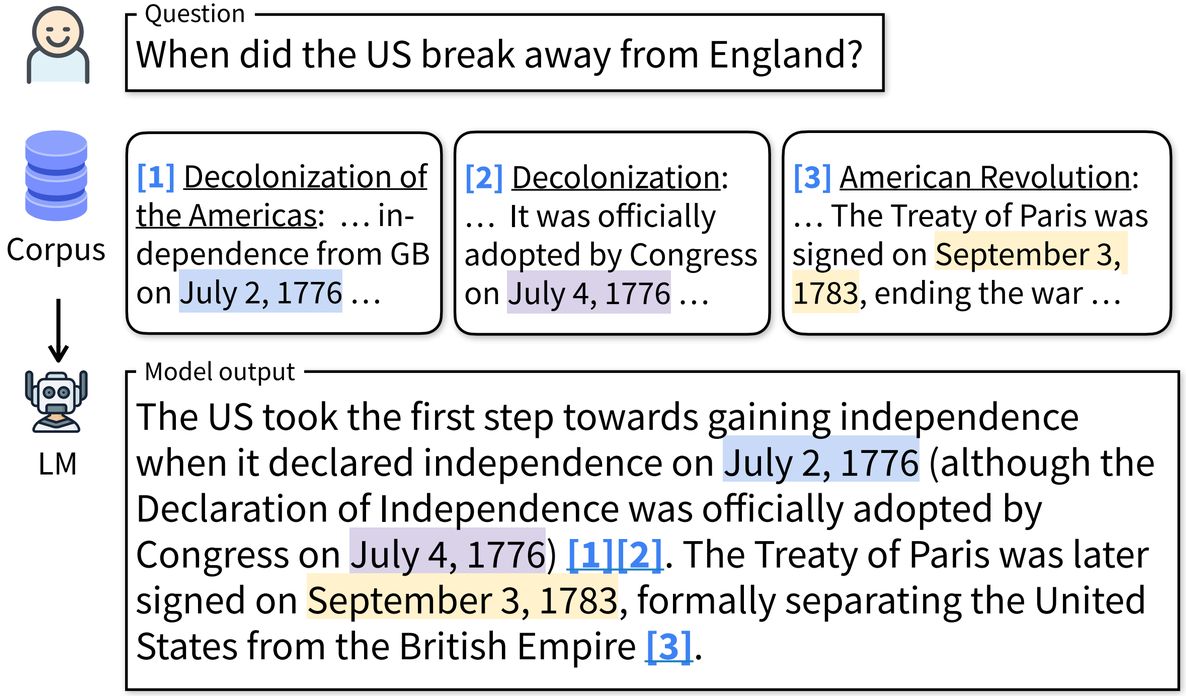
Evaluation by new applications: Generation with citations



Either provided by users or retrieved from Internet/databases

- Given a **question** and **relevant documents**
  - LMs to generate a concise, informative answer

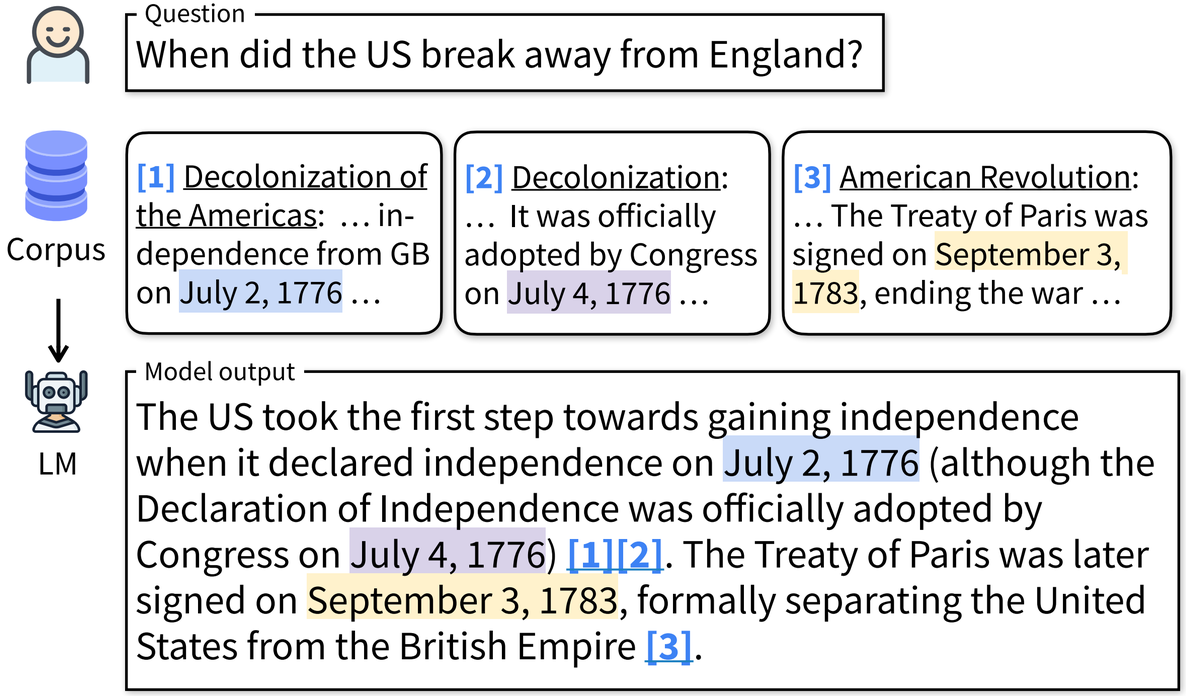
Evaluation by new applications: Generation with citations



Either provided by users or retrieved from Internet/databases

- Given a **question** and **relevant documents**
  - LMs to generate a concise, informative answer
  - **Desiderata**?

Evaluation by new applications: Generation with citations

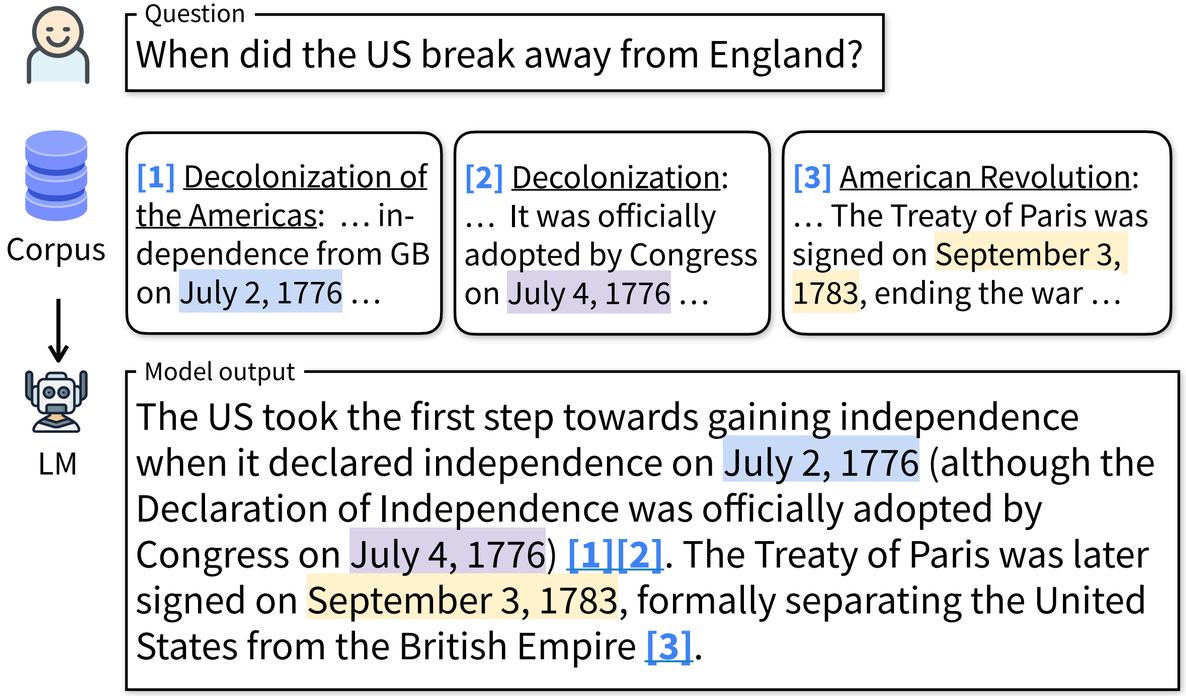


Either provided by users or retrieved from Internet/databases

- Given a **question** and **relevant documents**
  - LMs to generate a concise, informative answer
  - **Desiderata**?

Fluent

Evaluation by new applications: Generation with citations



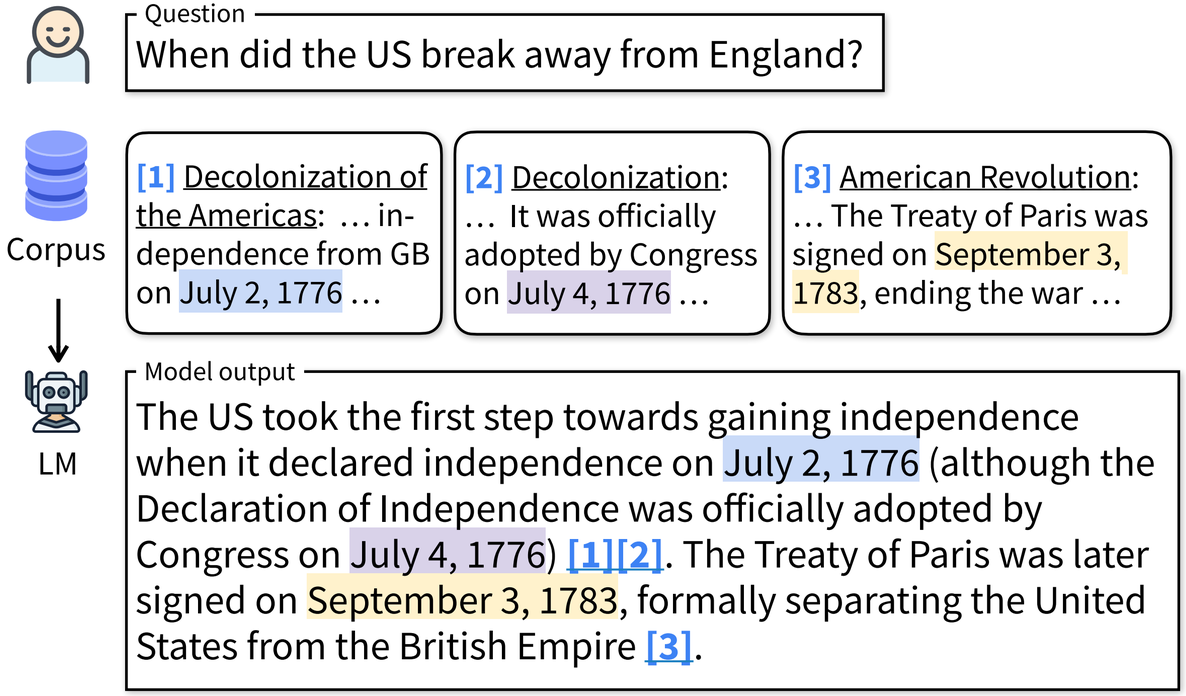
Either provided by users or retrieved from Internet/databases

- Given a **question** and **relevant documents**
  - LMs to generate a concise, informative answer
  - **Desiderata**?


Fluent


Accurate

Evaluation by new applications: Generation with citations



Either provided by users or retrieved from Internet/databases

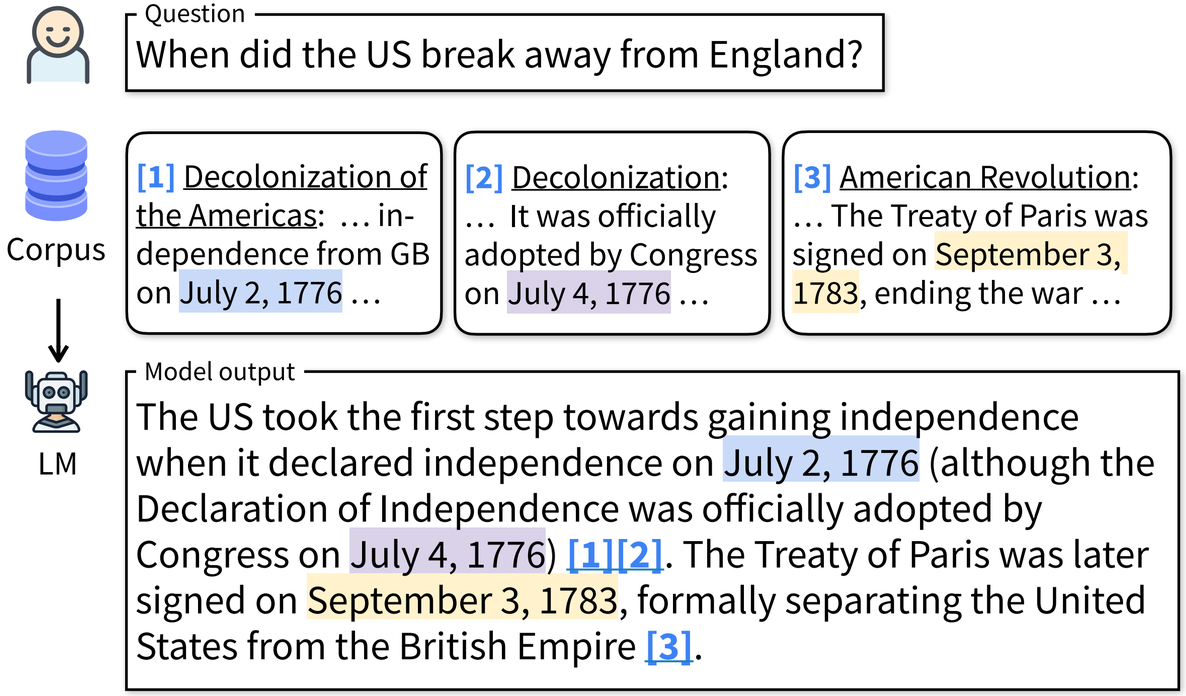
- Given a **question** and **relevant documents**
  - LMs to generate a concise, informative answer
  - **Desiderata**?


Fluent


Accurate

Multi-aspect answers require synthesizing multiple docs

Evaluation by new applications: Generation with citations



Either provided by users or retrieved from Internet/databases

- Given a **question** and **relevant documents**
  - LMs to generate a concise, informative answer
  - **Desiderata**?


Fluent

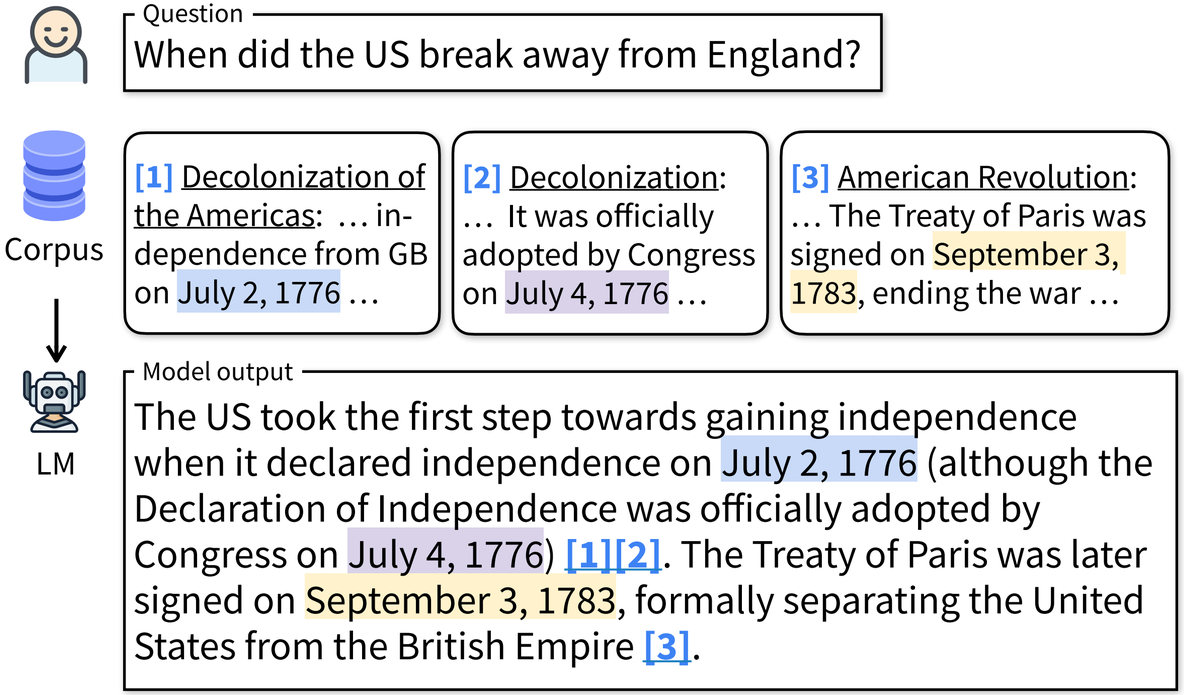

Accurate

Multi-aspect answers require synthesizing multiple docs


Verifiable

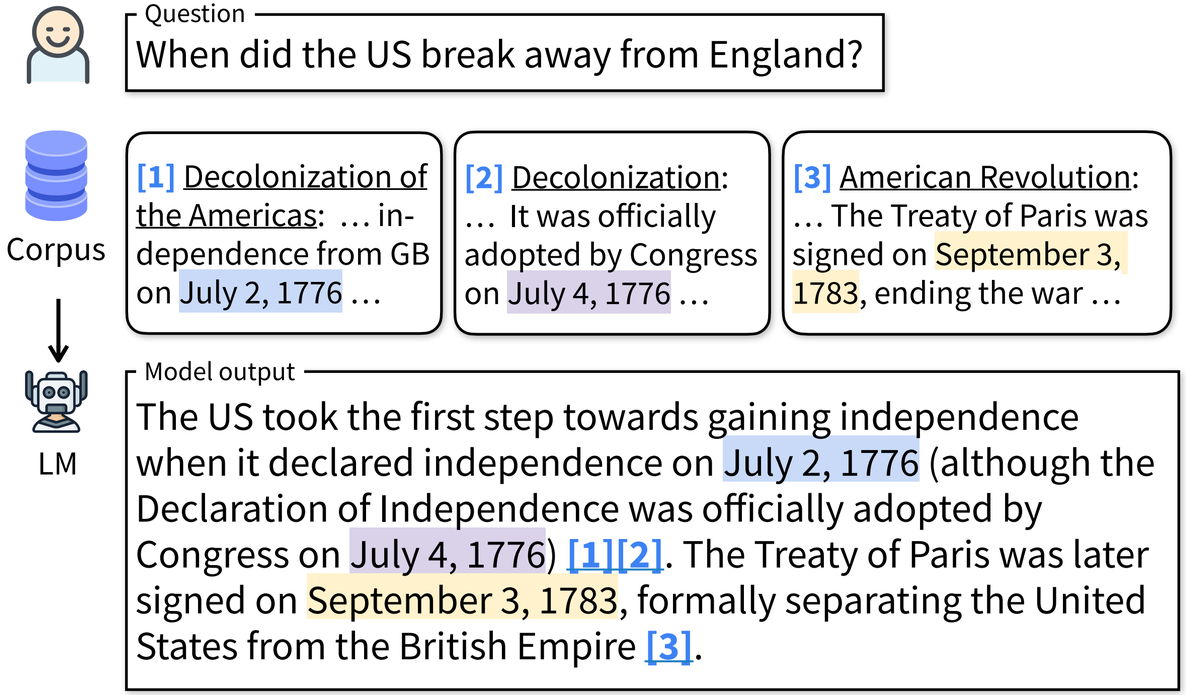
Evaluation by new applications: Generation with citations

Why do we need citations?



Evaluation by new applications: Generation with citations

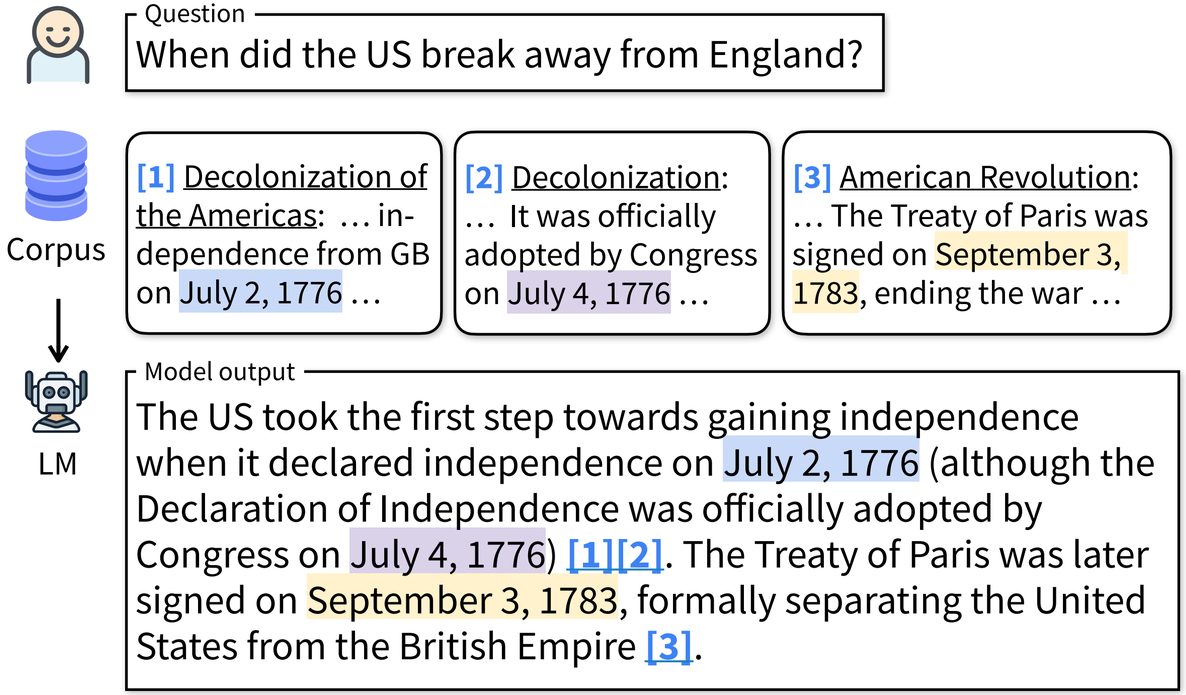
Why do we need citations?



- **Easy to verify**

Evaluation by new applications: Generation with citations

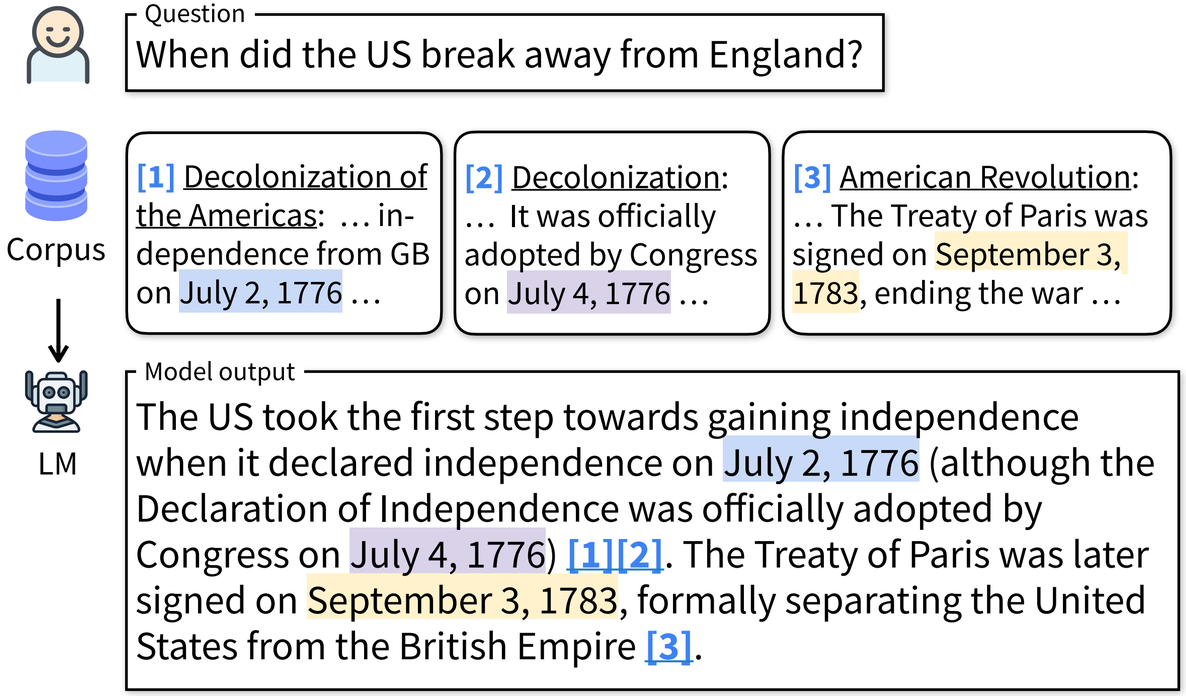
Why do we need citations?



- **Easy to verify**
  - We cannot trust LMs 100%

Evaluation by new applications: Generation with citations

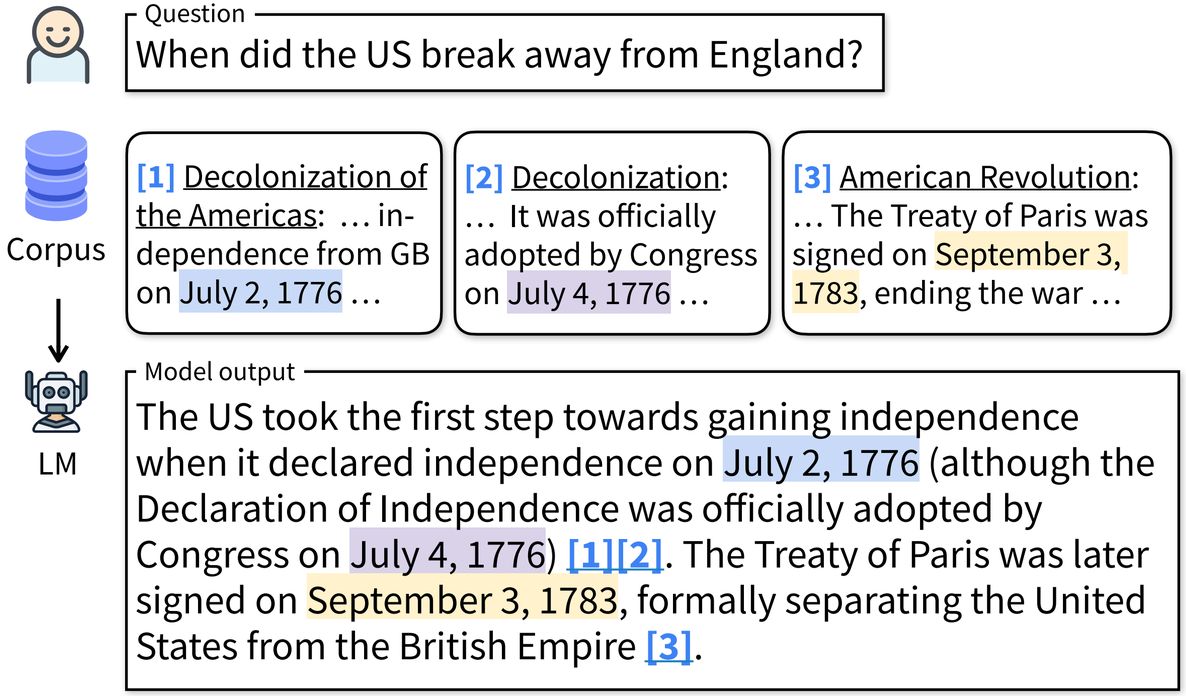
Why do we need citations?



- **Easy to verify**
  - We cannot trust LMs 100%
  - Important for high-stake applications (medicine, finance)

Evaluation by new applications: Generation with citations

Why do we need citations?



- **Easy to verify**
  - We cannot trust LMs 100%
  - Important for high-stake applications (medicine, finance)
- **Copyright issues**

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Evaluation by new applications: Generation with citations

ALCE: **A**utomatic **L**LM **C**itation **E**valuation

T. Gao, H. Yen, J. Yu, D. Chen. EMNLP 2023. **Enabling Large Language Models to Generate Text with Citations**.

Evaluation by new applications: Generation with citations

ALCE: **A**utomatic **L**LM **C**itation **E**valuation

```
Instruction: Write an accurate, engaging, and
concise answer for ...
```

T. Gao, H. Yen, J. Yu, D. Chen. EMNLP 2023. **Enabling Large Language Models to Generate Text with Citations**.

Evaluation by new applications: Generation with citations

ALCE: **A**utomatic **L**LM **C**itation **E**valuation

```
Instruction: Write an accurate, engaging, and
concise answer for ...

Document [1](Title: American Decolonization)
...
Document [2](Title: Decolonization) ...
Document [3](Title: American Revolution) ...
...
```

T. Gao, H. Yen, J. Yu, D. Chen. EMNLP 2023. **Enabling Large Language Models to Generate Text with Citations**.

Evaluation by new applications: Generation with citations

ALCE: **A**utomatic **L**LM **C**itation **E**valuation

```
Instruction: Write an accurate, engaging, and
concise answer for ...

Document [1](Title: American Decolonization)
...
Document [2](Title: Decolonization) ...
Document [3](Title: American Revolution) ...
...

Question: When did US break away from England?
```

T. Gao, H. Yen, J. Yu, D. Chen. EMNLP 2023. **Enabling Large Language Models to Generate Text with Citations**.

Evaluation by new applications: Generation with citations

ALCE: **A**utomatic **L**LM **C**itation **E**valuation

Instruction: Write an accurate, engaging, and concise answer for ...

Document [1](Title: American Decolonization)
...
Document [2](Title: Decolonization) ...
Document [3](Title: American Revolution) ...
...

Question: When did US break away from England?

Answer: The United States took the first step towards gaining independence ... **[1][2]**. The Treaty of Paris was later signed ... **[3]**.

T. Gao, H. Yen, J. Yu, D. Chen. EMNLP 2023. **Enabling Large Language Models to Generate Text with Citations**.

Evaluation by new applications: Generation with citations

ALCE: **A**utomatic **L**LM **C**itation **E**valuation

Instruction: Write an accurate, engaging, and concise answer for ...

Document [1](Title: American Decolonization)
...
Document [2](Title: Decolonization) ...
Document [3](Title: American Revolution) ...
...

Question: When did US break away from England?

Answer: The United States took the first step towards gaining independence ... **[1][2]**. The Treaty of Paris was later signed ... **[3]**.


Fluency


Correctness


Citation quality

Evaluation by new applications: Generation with citations

ALCE: **A**utomatic **L**LM **C**itation **E**valuation

```
Instruction: Write an accurate, engaging, and
concise answer for ...

Document [1](Title: American Decolonization)
...
Document [2](Title: Decolonization) ...
Document [3](Title: American Revolution) ...
...

Question: When did US break away from England?
Answer: The United States took the first step
towards gaining independence ... [1][2]. The
Treaty of Paris was later signed ... [3].
```

Fluency

Correctness

Citation quality

ALCE (May 2023) is the *first automatic evaluation*!

Evaluation by new applications: Generation with citations

ALCE: **A**utomatic **L**LM **C**itation **E**valuation

```
Instruction: Write an accurate, engaging, and
concise answer for ...

Document [1](Title: American Decolonization)
...
Document [2](Title: Decolonization) ...
Document [3](Title: American Revolution) ...
...

Question: When did US break away from England?
Answer: The United States took the first step
towards gaining independence ... [1][2]. The
Treaty of Paris was later signed ... [3].
```


Fluency


Correctness


Citation quality

ALCE (May 2023) is the *first automatic evaluation*!

Before: only human evaluation

Evaluation by new applications: Generation with citations

ALCE: **A**utomatic **L**LM **C**itation **E**valuation

```
Instruction: Write an accurate, engaging, and
concise answer for ...

Document [1](Title: American Decolonization)
...
Document [2](Title: Decolonization) ...
Document [3](Title: American Revolution) ...
...

Question: When did US break away from England?
Answer: The United States took the first step
towards gaining independence ... [1][2]. The
Treaty of Paris was later signed ... [3].
```

Fluency

Correctness

Citation quality

ALCE (May 2023) is the *first automatic evaluation*!

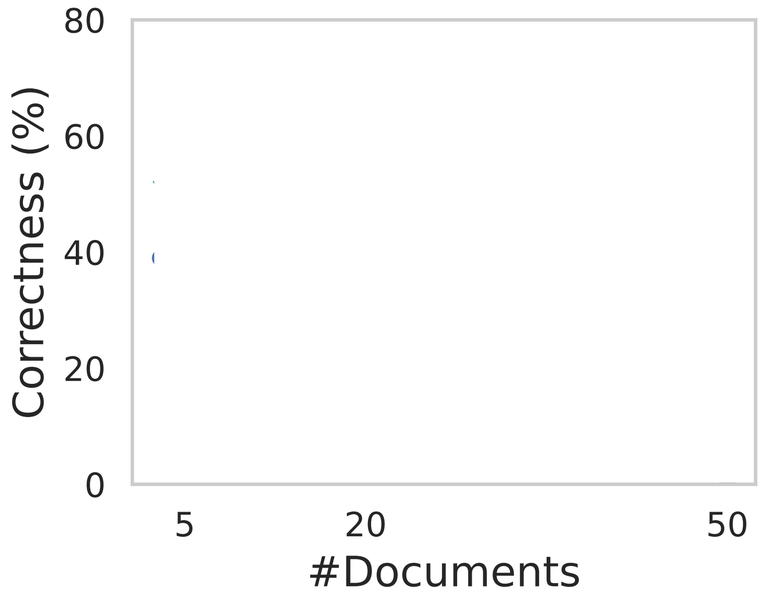
Before: only human evaluation

ALCE highly correlates with human evaluation

Generation with citations (ALCE): Results

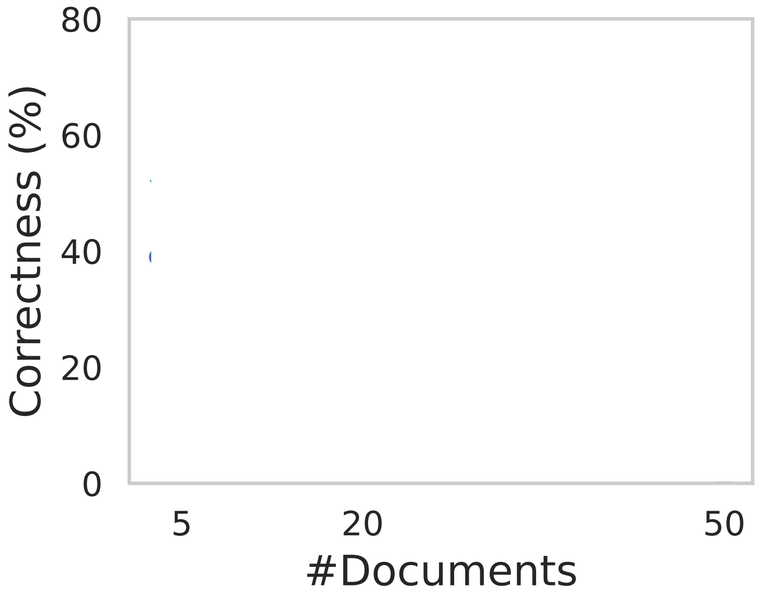
Using questions and answers from the ASQA dataset (Stelmakh et al., 2022). Using question-retrieved documents from Wikipedia.

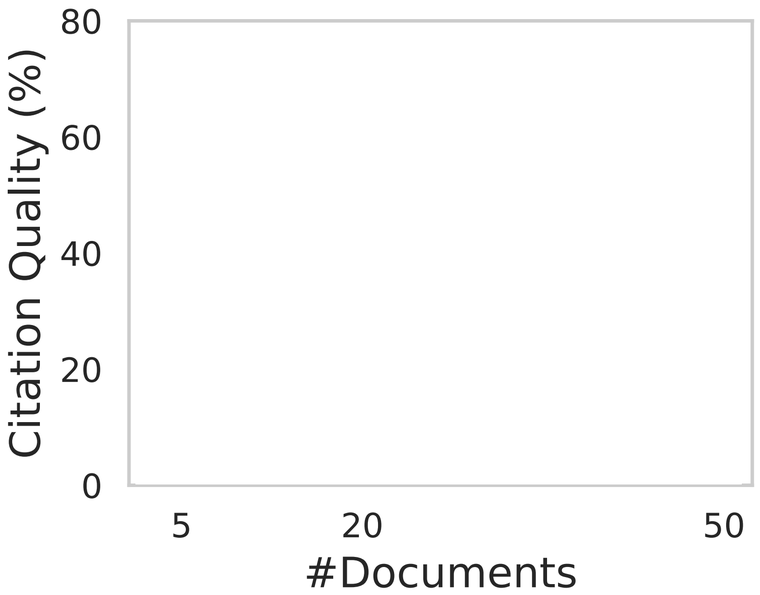
Generation with citations (ALCE): Results



Using questions and answers from the ASQA dataset (Stelmakh et al., 2022). Using question-retrieved documents from Wikipedia.

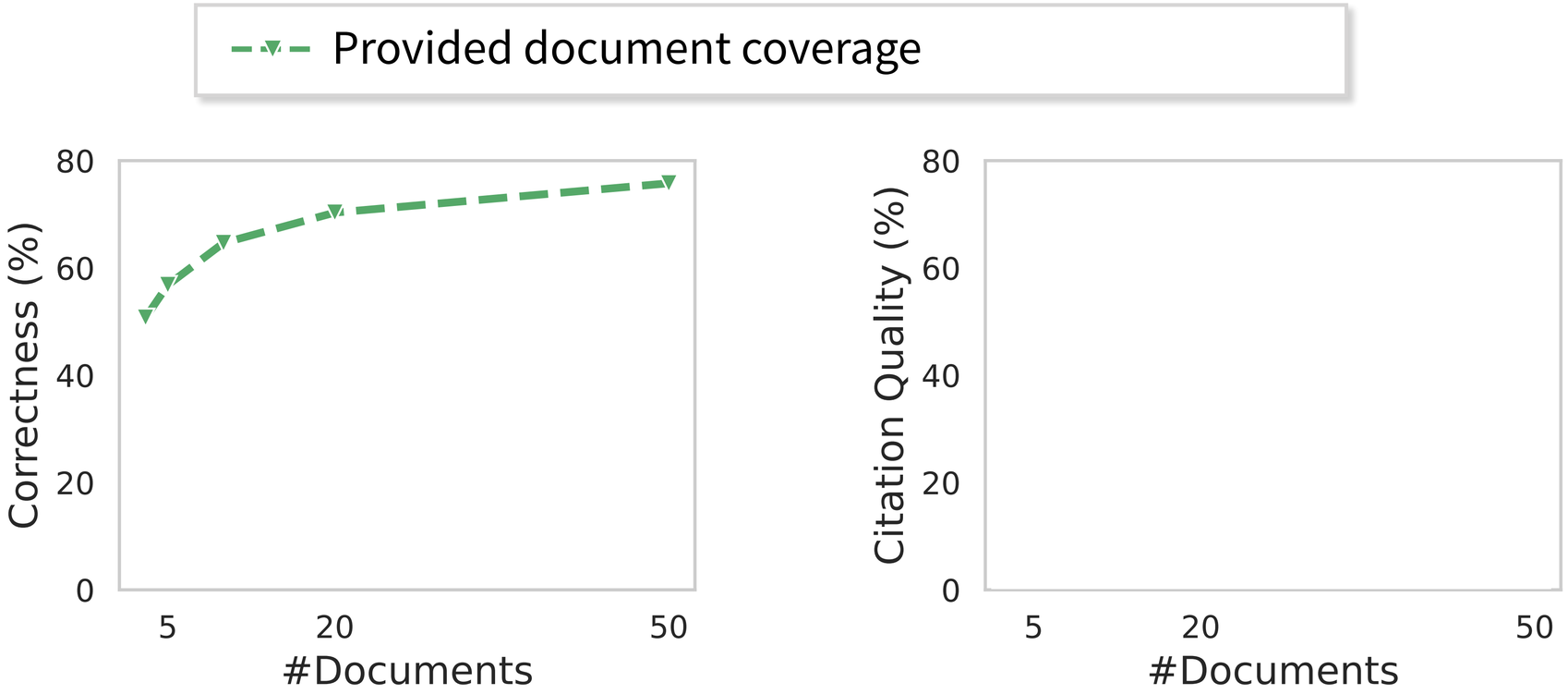
Generation with citations (ALCE): Results





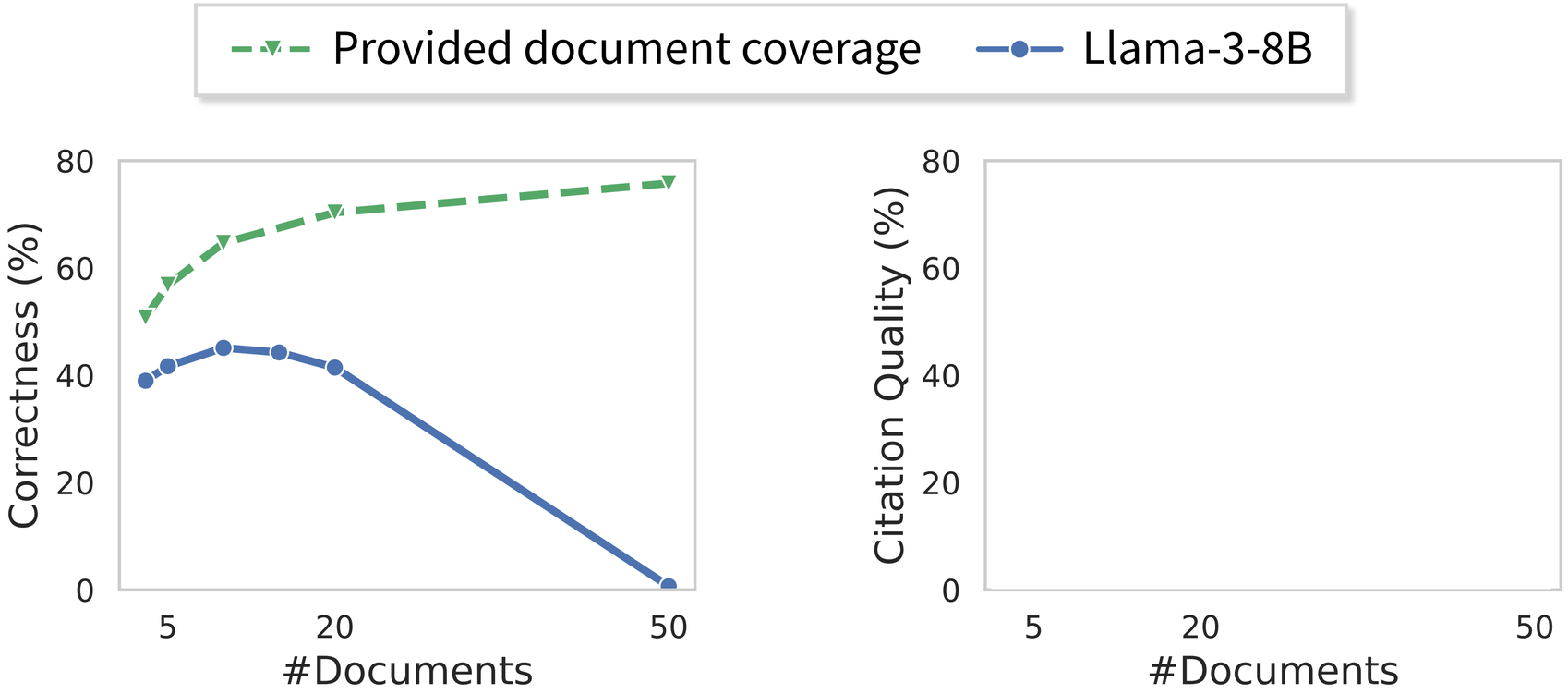
Using questions and answers from the ASQA dataset (Stelmakh et al., 2022). Using question-retrieved documents from Wikipedia.

Generation with citations (ALCE): Results



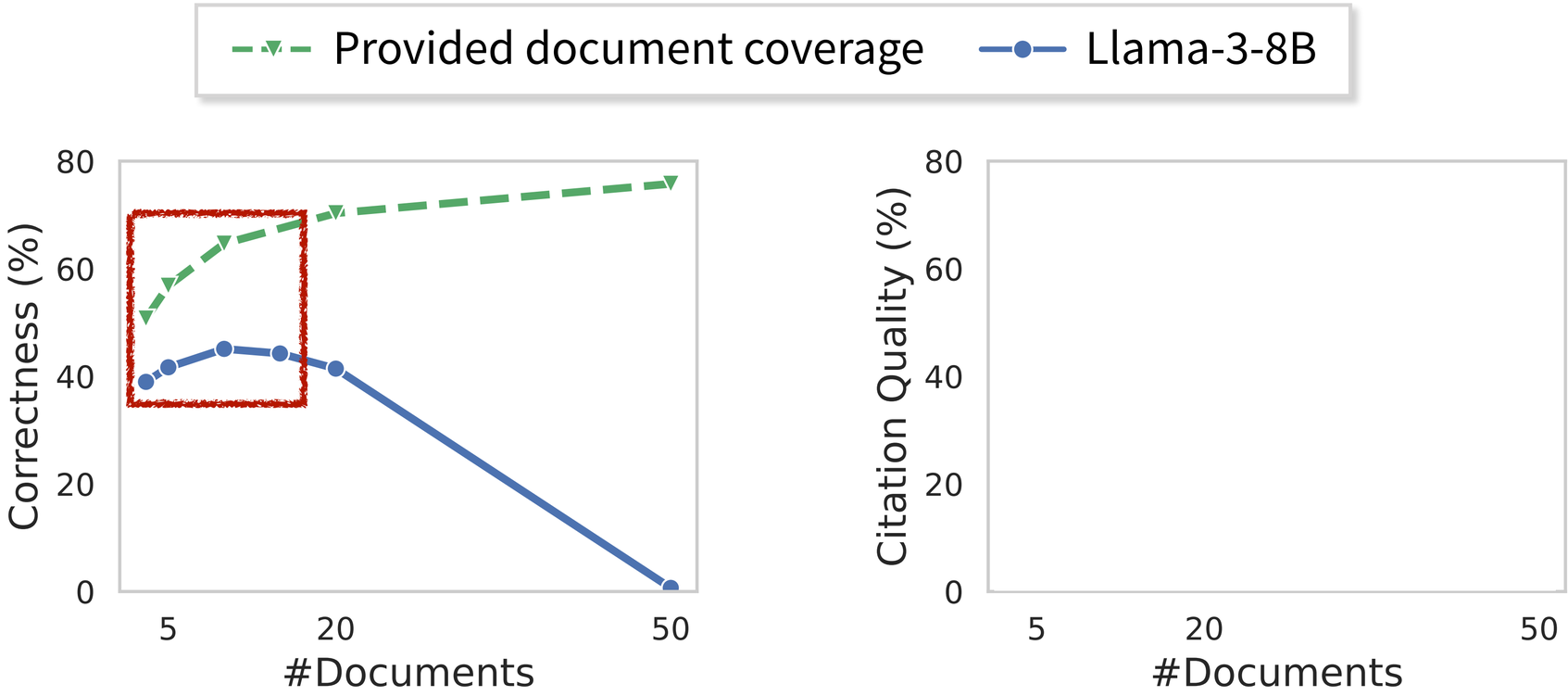
Using questions and answers from the ASQA dataset (Stelmakh et al., 2022). Using question-retrieved documents from Wikipedia.

Generation with citations (ALCE): Results



Using questions and answers from the ASQA dataset (Stelmakh et al., 2022). Using question-retrieved documents from Wikipedia.

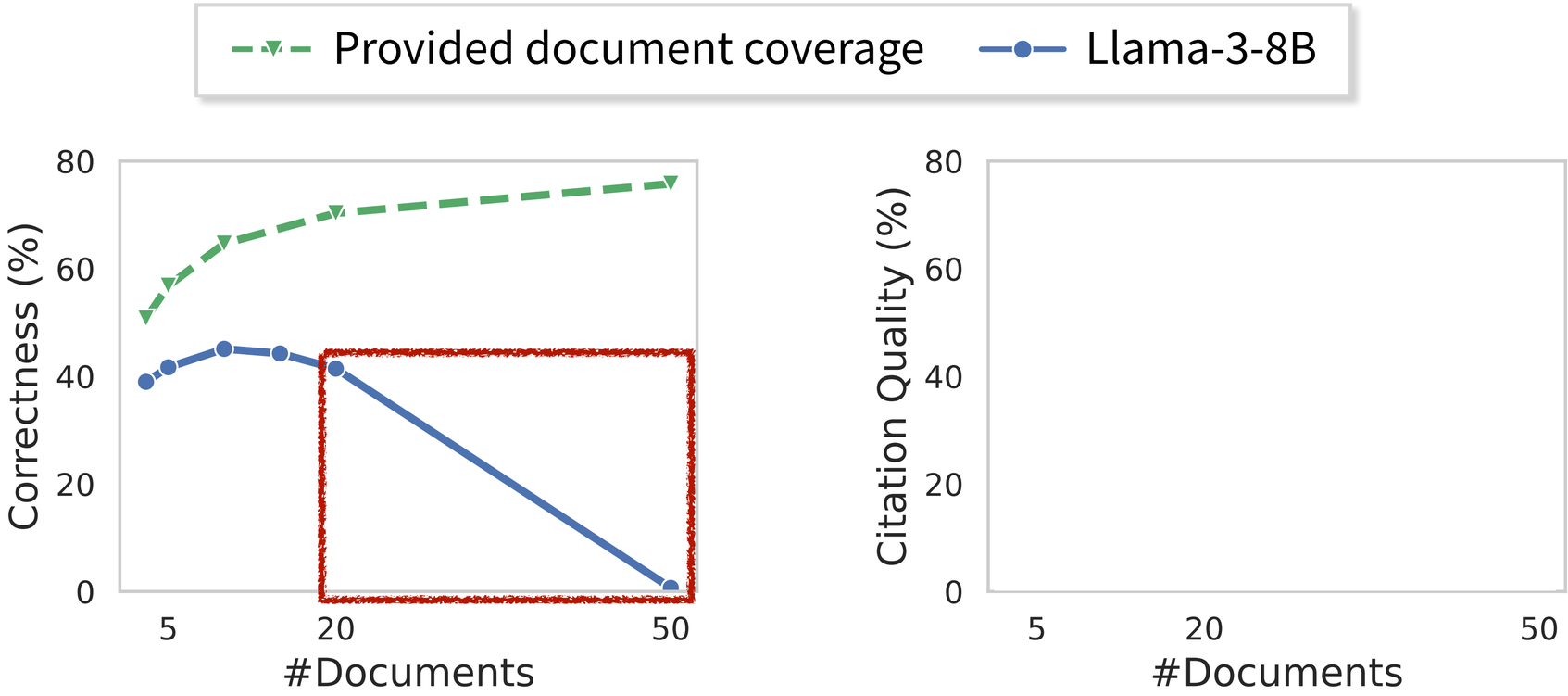
Generation with citations (ALCE): Results



- LMs **cannot** fully extract useful information **from context**

Using questions and answers from the ASQA dataset (Stelmakh et al., 2022). Using question-retrieved documents from Wikipedia.

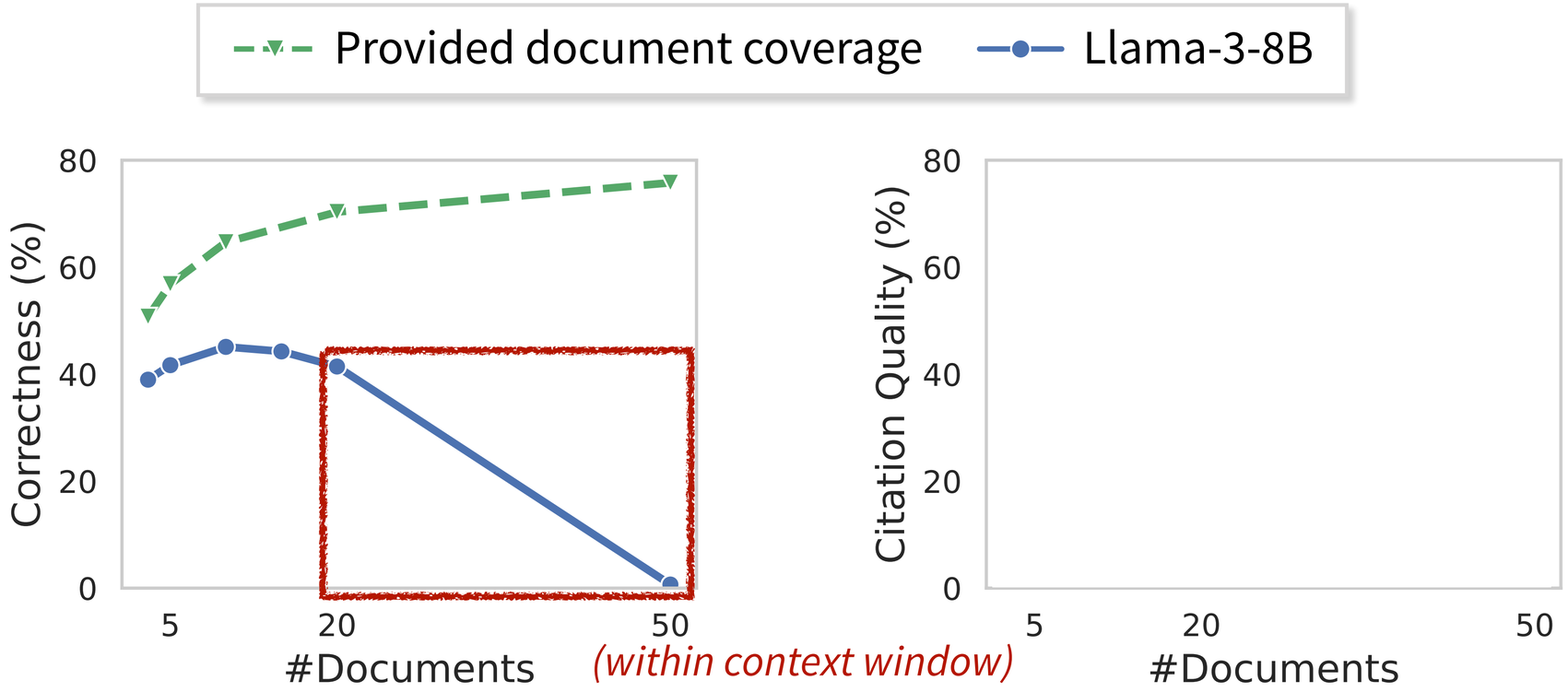
Generation with citations (ALCE): Results



- LMs **cannot** fully extract useful information **from context**
- LMs degenerate with longer inputs

Using questions and answers from the ASQA dataset (Stelmakh et al., 2022). Using question-retrieved documents from Wikipedia.

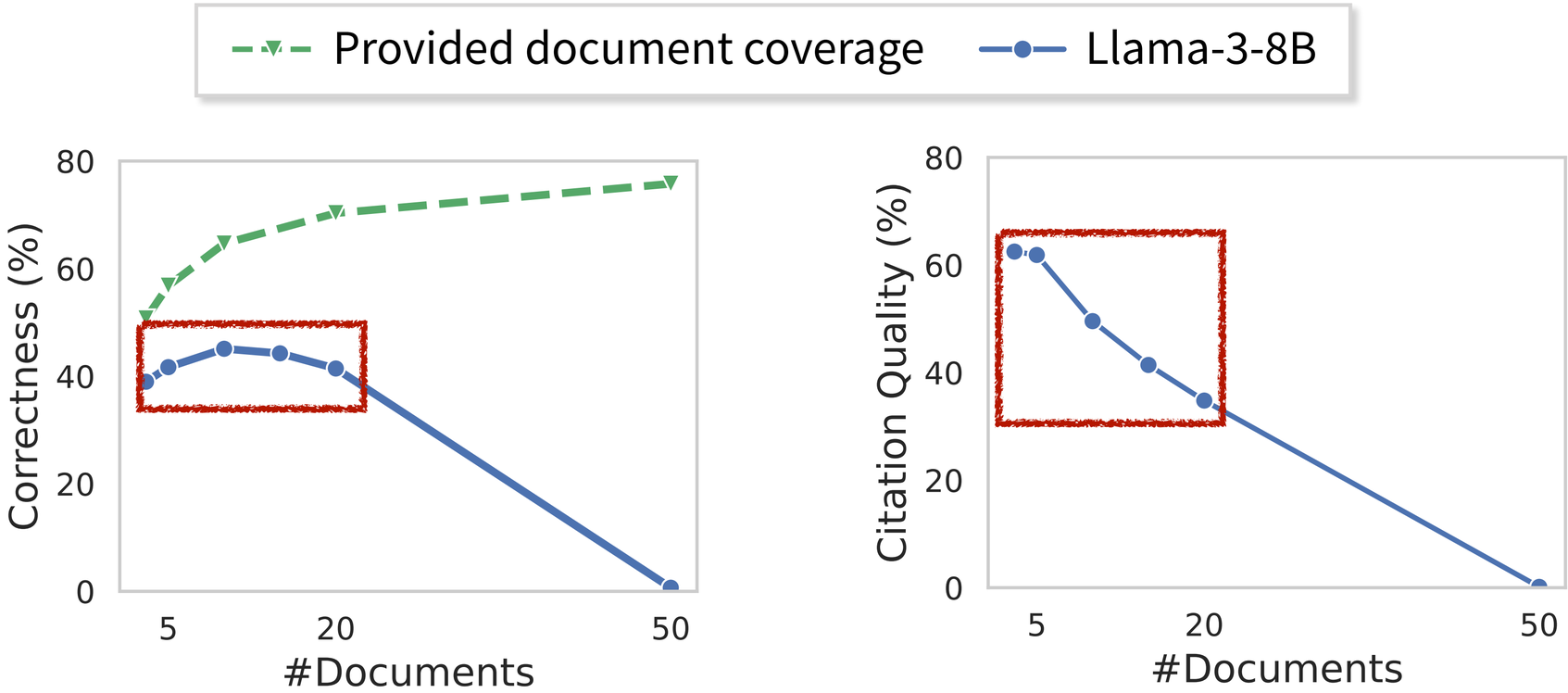
Generation with citations (ALCE): Results



- LMs **cannot** fully extract useful information **from context**
- LMs degenerate with longer inputs

Using questions and answers from the ASQA dataset (Stelmakh et al., 2022). Using question-retrieved documents from Wikipedia.

Generation with citations (ALCE): Results

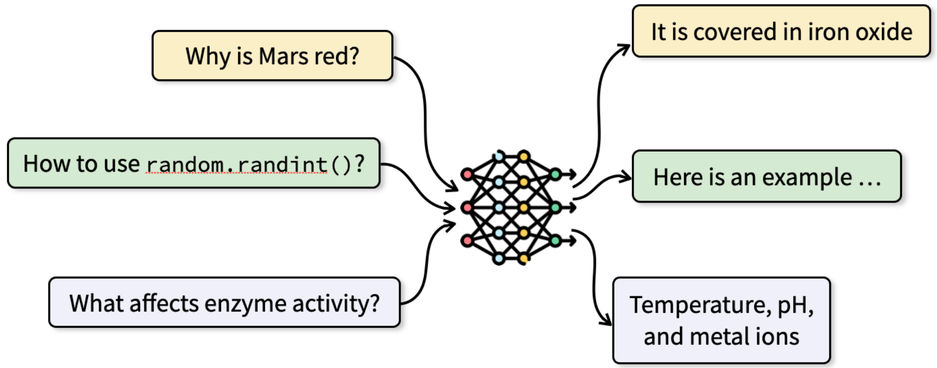


- LMs **cannot** fully extract useful information **from context**
- LMs degenerate with longer inputs
- Citation quality degrades faster than correctness

Using questions and answers from the ASQA dataset (Stelmakh et al., 2022). Using question-retrieved documents from Wikipedia.

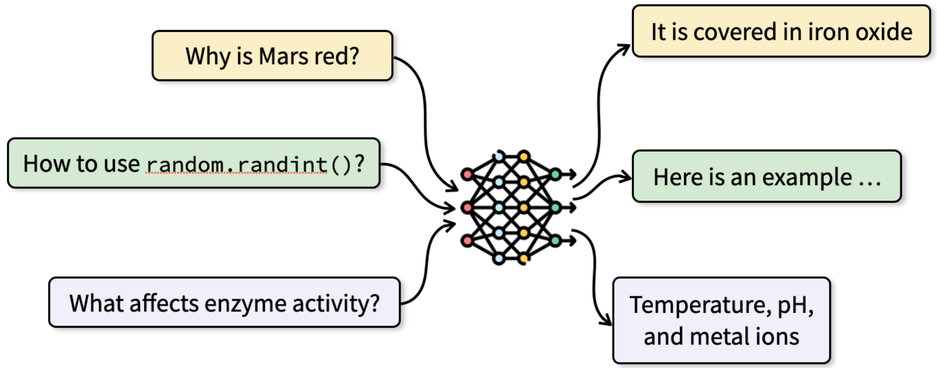
Generation with citations (ALCE): Impact

Generation with citations (ALCE): Impact



Existing LM evaluation

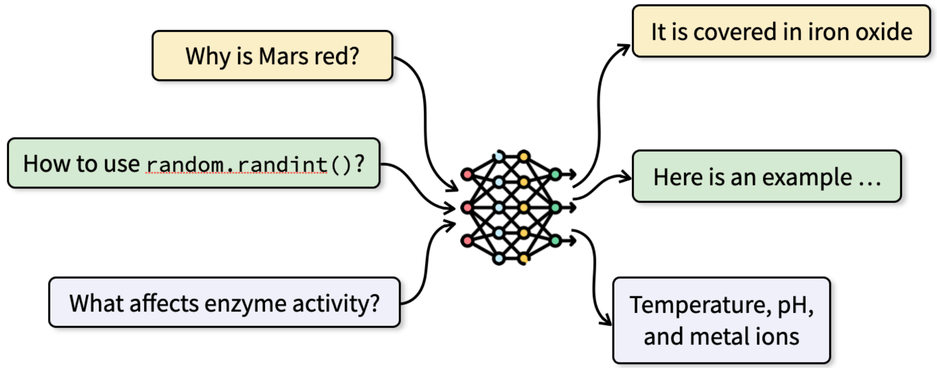
Generation with citations (ALCE): Impact



Existing LM evaluation

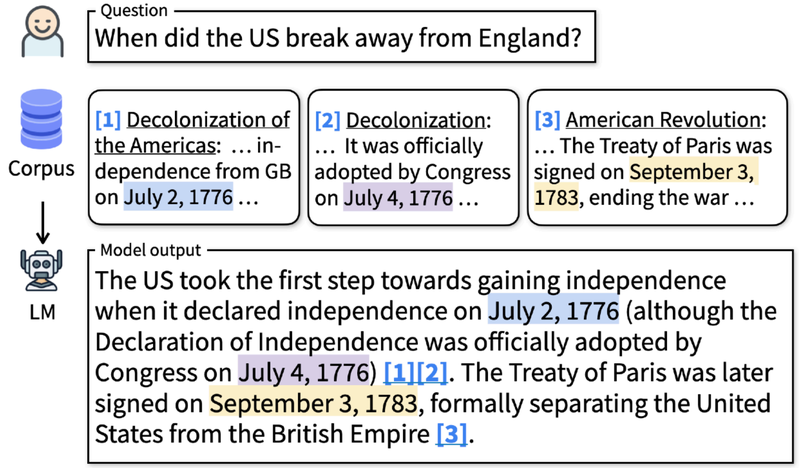
- Mostly Trivia Q&A style
- Testing knowledge capacity

Generation with citations (ALCE): Impact



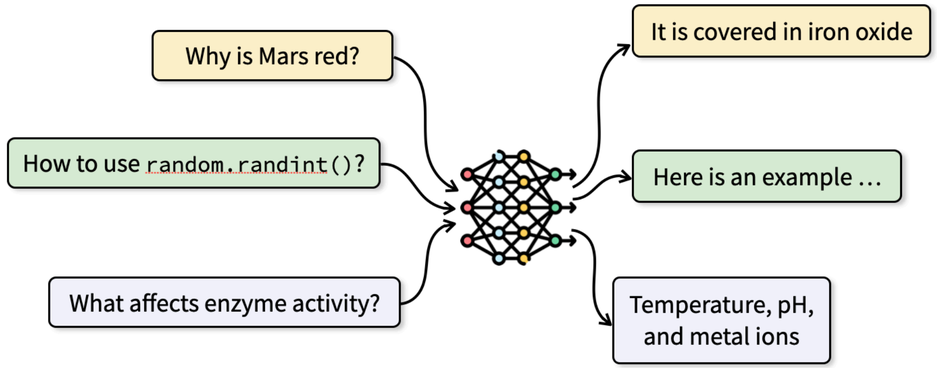
Existing LM evaluation

- Mostly Trivia Q&A style
- Testing knowledge capacity



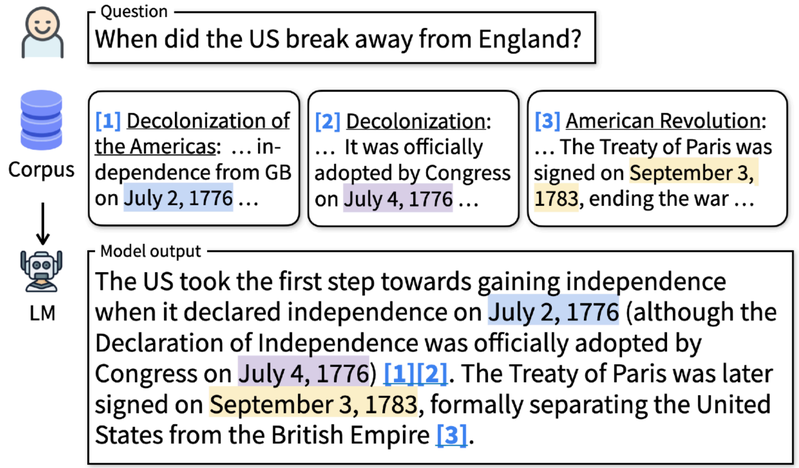
ALCE (May 2023)

Generation with citations (ALCE): Impact



Existing LM evaluation

- Mostly Trivia Q&A style
- Testing knowledge capacity



ALCE (May 2023)

- First automatic evaluation for synthesizing contexts to answer with citations

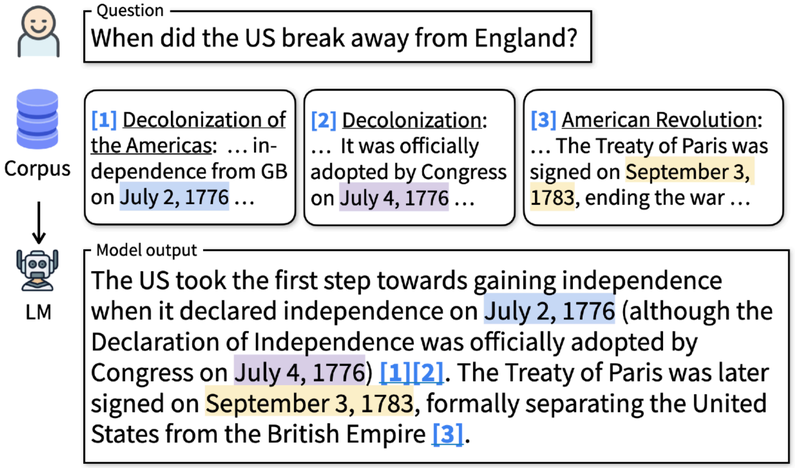
Generation with citations (ALCE): Impact

Enabling reproducible research on novel applications

Li et al., 2023; Asai et al., 2024; Li et al., 2024; Chuang et al., 2025

perplexity

SearchGPT



ALCE (May 2023)

First automatic evaluation for synthesizing contexts to answer with citations

Generation with citations (ALCE): Impact

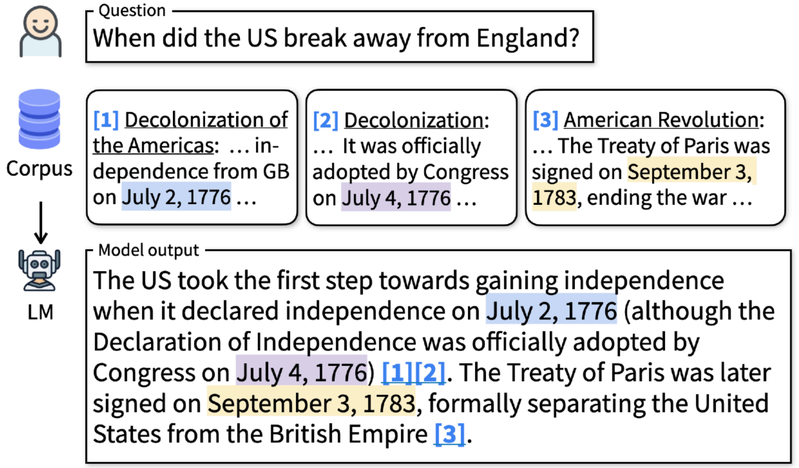
Enabling reproducible research on novel applications

Li et al., 2023; Asai et al., 2024; Li et al., 2024; Chuang et al., 2025

perplexity SearchGPT

Extended to more settings

Kamalloo et al., 2024; Cohen-Wang et al., 2024; Zhang et al., 2024



ALCE (May 2023)

First automatic evaluation for synthesizing contexts to answer with citations

Generation with citations (ALCE): Impact

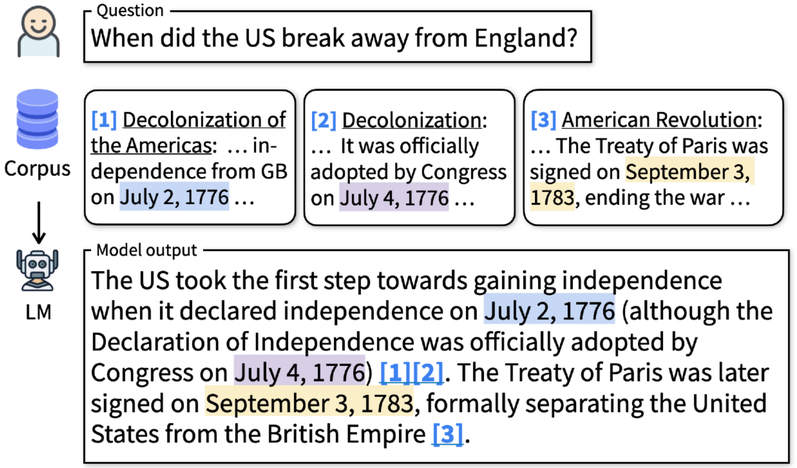
Enabling reproducible research on novel applications

Li et al., 2023; Asai et al., 2024; Li et al., 2024; Chuang et al., 2025

perplexity SearchGPT

Extended to more settings

Kamalloo et al., 2024; Cohen-Wang et al., 2024; Zhang et al., 2024



ALCE (May 2023)

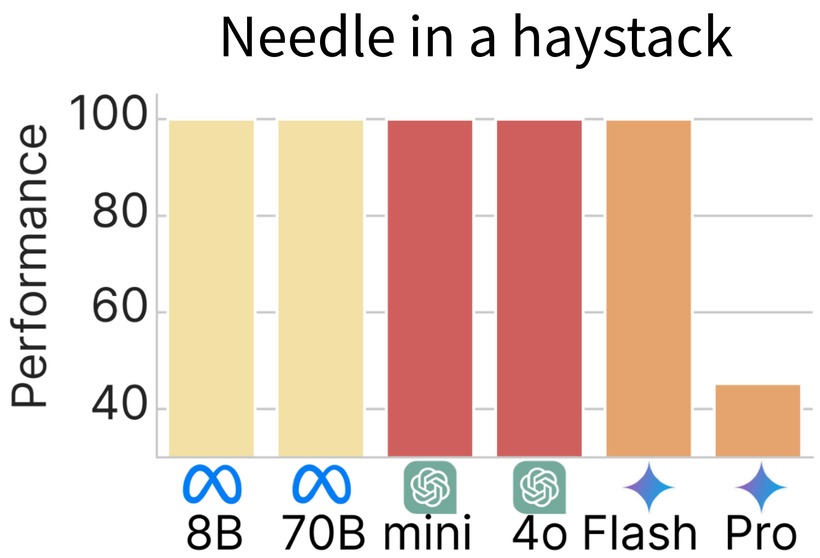
First automatic evaluation for synthesizing contexts to answer with citations

First to reveal LM deficiency in long-context processing

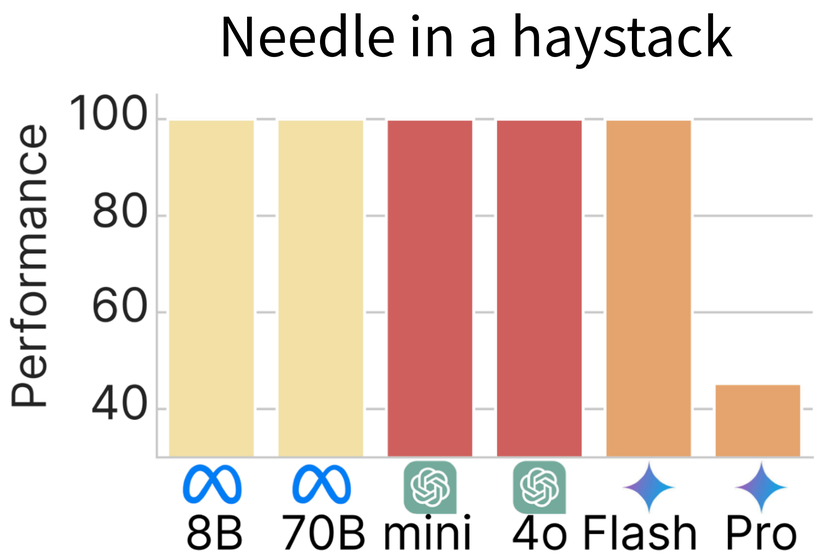
(Earlier than “needle in a haystack”)

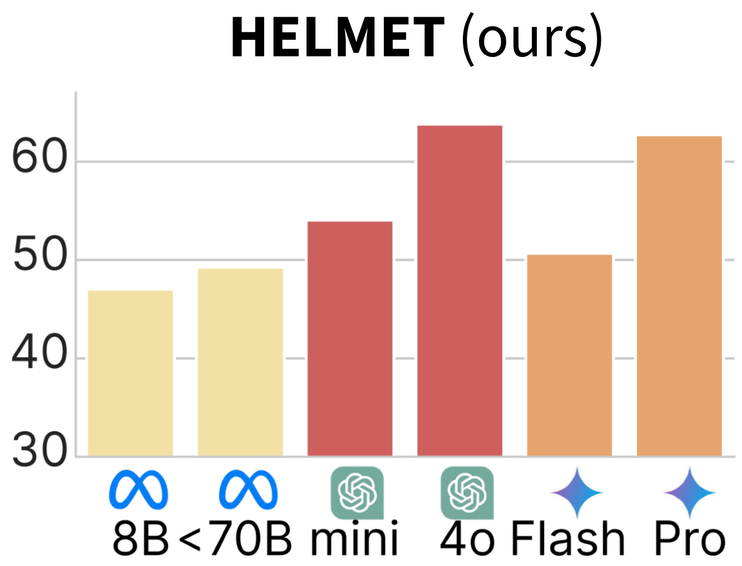
HELMET offers more reliable long-context evaluation

HELMET offers more reliable long-context evaluation

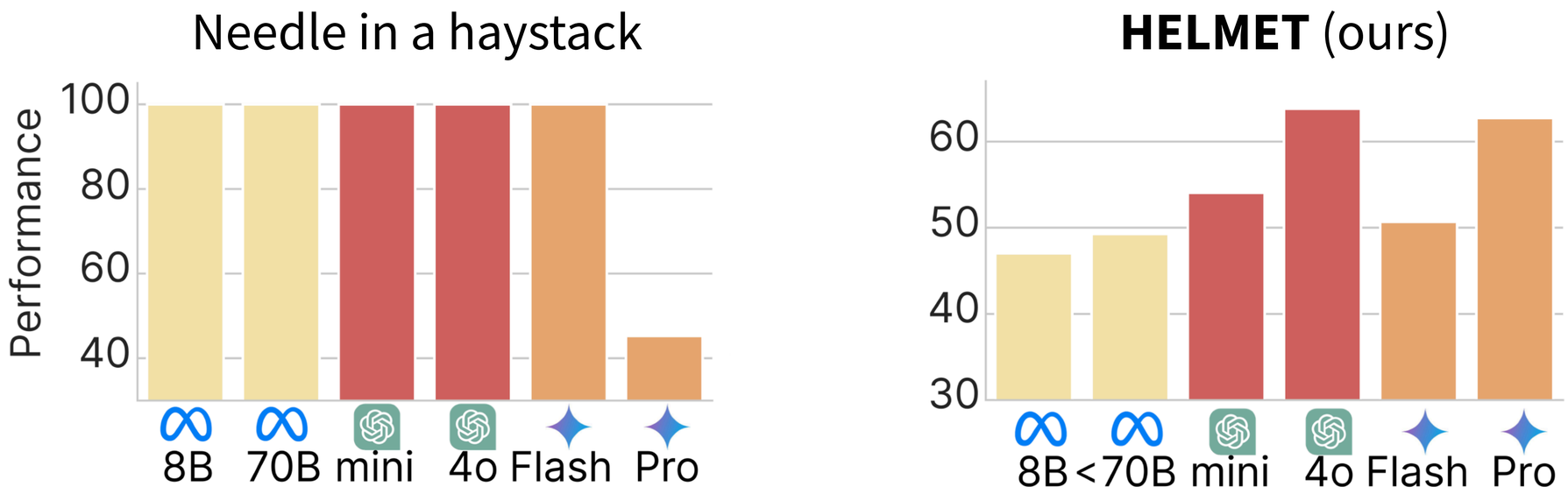


HELMET offers more reliable long-context evaluation



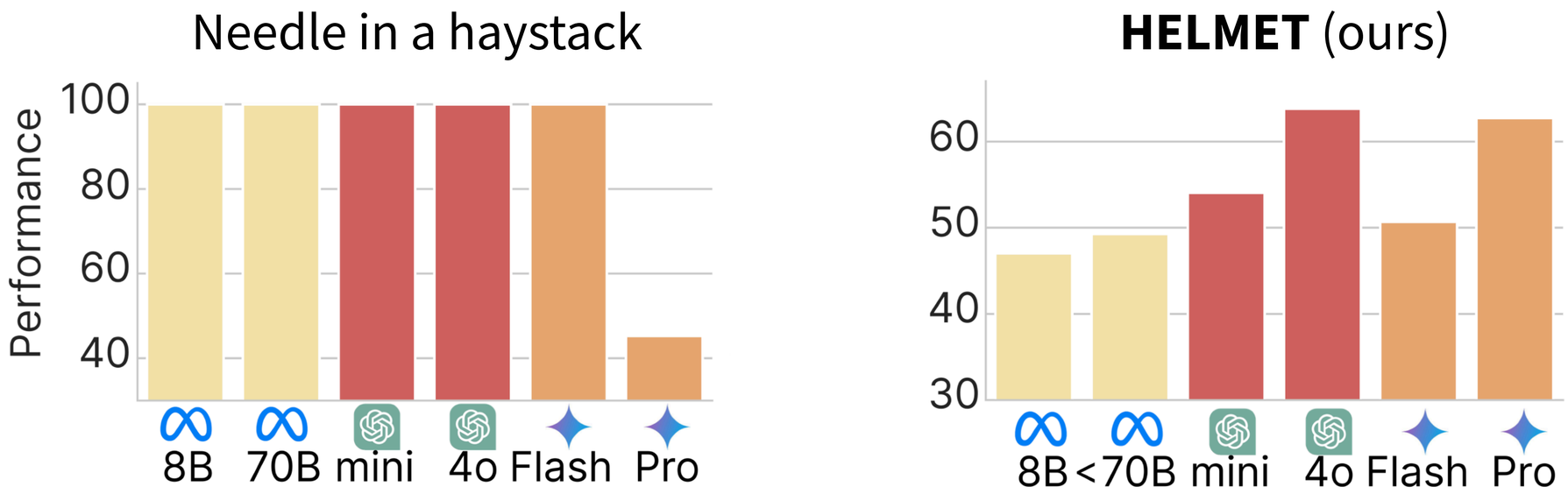


HELMET offers more reliable long-context evaluation



- HELMET provides more **informative** and **reliable** comparison

HELMET offers more reliable long-context evaluation



- HELMET provides more **informative** and **reliable** comparison
- Has already been adopted by industry (e.g., Phi-4)

Developing effective long-context language models

Developing effective long-context language models

Setting: starting from a **pre-trained LM** and continuing training

Developing effective long-context language models

Setting: starting from a **pre-trained LM** and continuing training

Question: What types of **data** leads to desirable **long-context capabilities**?

Developing effective long-context language models

Setting: starting from a **pre-trained LM** and continuing training

Question: What types of **data** leads to desirable **long-context capabilities**?

ProLong: a thoroughly-ablated long-context LM recipe

T. Gao*, A. Wettig*, H. Yen, D. Chen. Preprint 2024. **How to Train Long-Context Language Models (Effectively)**.

Developing effective long-context language models

Setting: starting from a **pre-trained LM** and continuing training

Question: What types of **data** leads to desirable **long-context capabilities**?

ProLong: a thoroughly-ablated long-context LM recipe

8K → 512K (**longest** context window for **open-source models**!)

ProLong: Effective long-context LM data recipe

Question #1: Is it enough to increase the visible sequence length?

The ablation's training and evaluation lengths are both 64K by default. The ablation uses Llama-3-8B-Base and trains on 5B tokens of data.

ProLong: Effective long-context LM data recipe

Question #1: Is it enough to increase the visible sequence length?

Training

Long-context recall

The ablation's training and evaluation lengths are both 64K by default. The ablation uses Llama-3-8B-Base and trains on 5B tokens of data.

ProLong: Effective long-context LM data recipe

Question #1: Is it enough to increase the visible sequence length?

Training

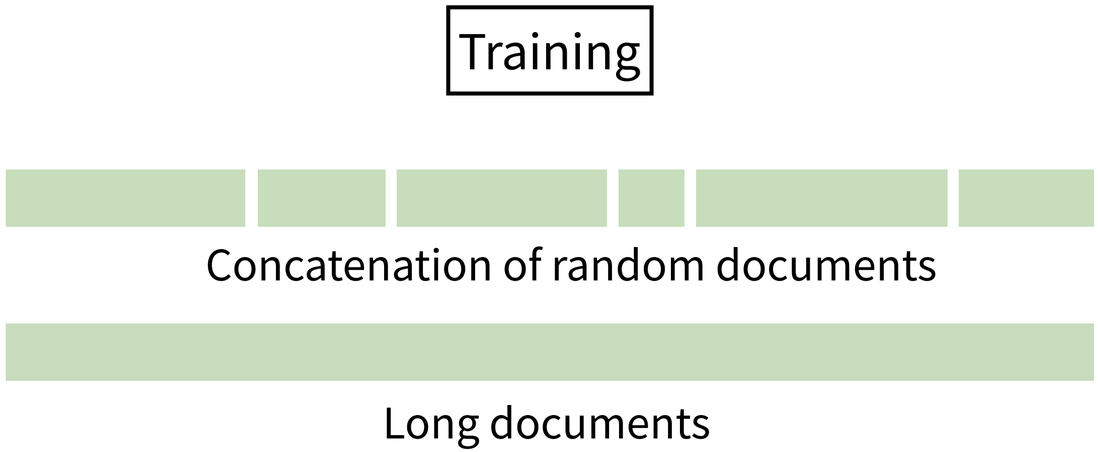
Long-context recall

Concatenation of random documents

The ablation's training and evaluation lengths are both 64K by default. The ablation uses Llama-3-8B-Base and trains on 5B tokens of data.

ProLong: Effective long-context LM data recipe

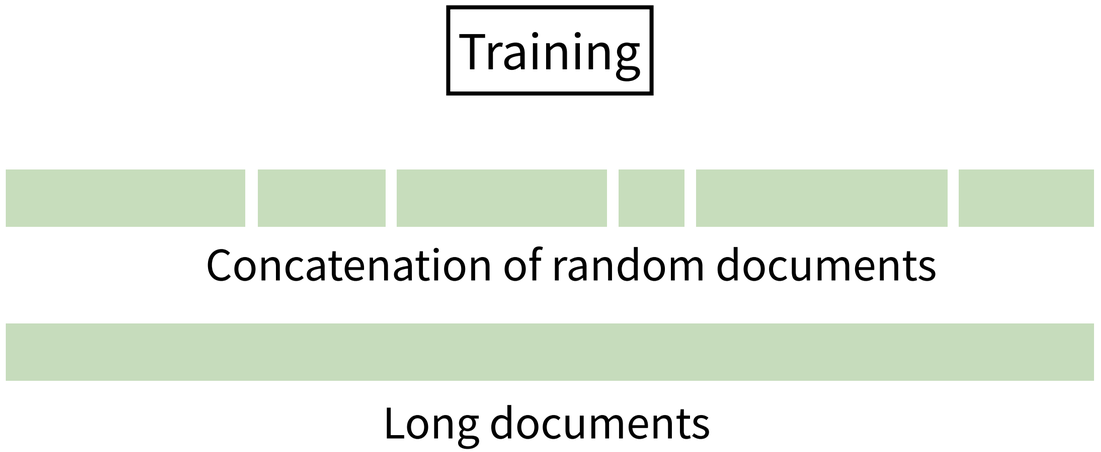
Question #1: Is it enough to increase the visible sequence length?



The ablation's training and evaluation lengths are both 64K by default. The ablation uses Llama-3-8B-Base and trains on 5B tokens of data.

ProLong: Effective long-context LM data recipe

Question #1: Is it enough to increase the visible sequence length?



Long-context recall	
39.0%	(Previous common practice)
92.7%	

Finding #1: using naturally-occurring long documents is important

The ablation's training and evaluation lengths are both 64K by default. The ablation uses Llama-3-8B-Base and trains on 5B tokens of data.

ProLong: Effective long-context LM data recipe

Question #2: Is it enough to train on documents within the target length?

Training

Evaluation

Long-context recall

ProLong: Effective long-context LM data recipe

Question #2: Is it enough to train on documents within the target length?

ProLong: Effective long-context LM data recipe

Question #2: Is it enough to train on documents within the target length?

ProLong: Effective long-context LM data recipe

Question #2: Is it enough to train on documents within the target length?

Training	Evaluation	Long-context recall
64K (same amount of data as below)	64K	95.0%
512K	64K	**98.5%**

Finding #2: training on longer sequences helps with performance at a shorter length

ProLong: Effective long-context LM data recipe

Question #3: Should we only train on long documents?

ProLong: Effective long-context LM data recipe

Question #3: Should we only train on long documents?

Long document data (%)

↘

Ablation on ratios of long documents vs. natural short data

ProLong: Effective long-context LM data recipe

Question #3: Should we only train on long documents?

Long Task Avg. Short Task Avg.

Long document data (%)

Ablation on ratios of long documents vs. natural short data

ProLong: Effective long-context LM data recipe

Question #3: Should we only train on long documents?

HELMET → Long Task Avg.

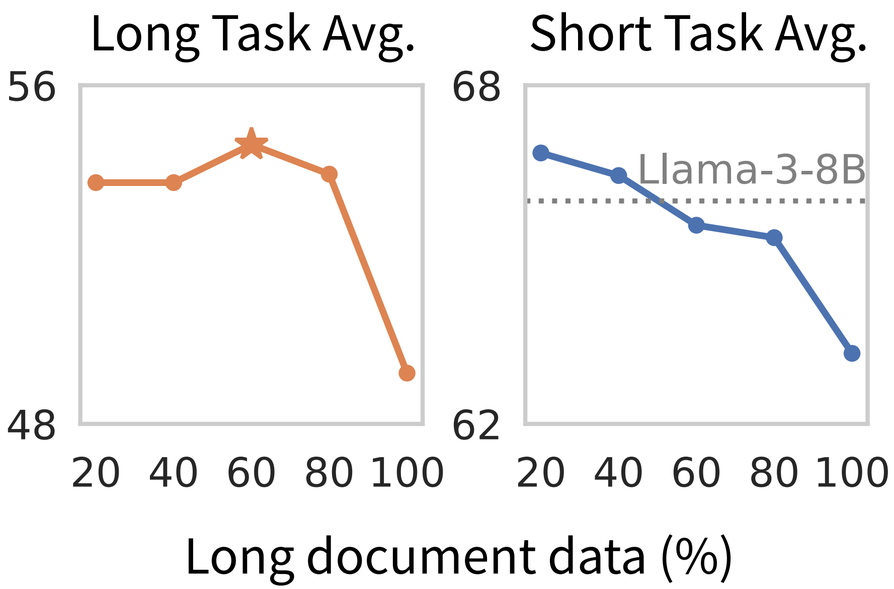
Standard LM evaluation (commonsense, exams) → Short Task Avg.

Long document data (%)

→ Ablation on ratios of long documents vs. natural short data

ProLong: Effective long-context LM data recipe

Question #3: Should we only train on long documents?

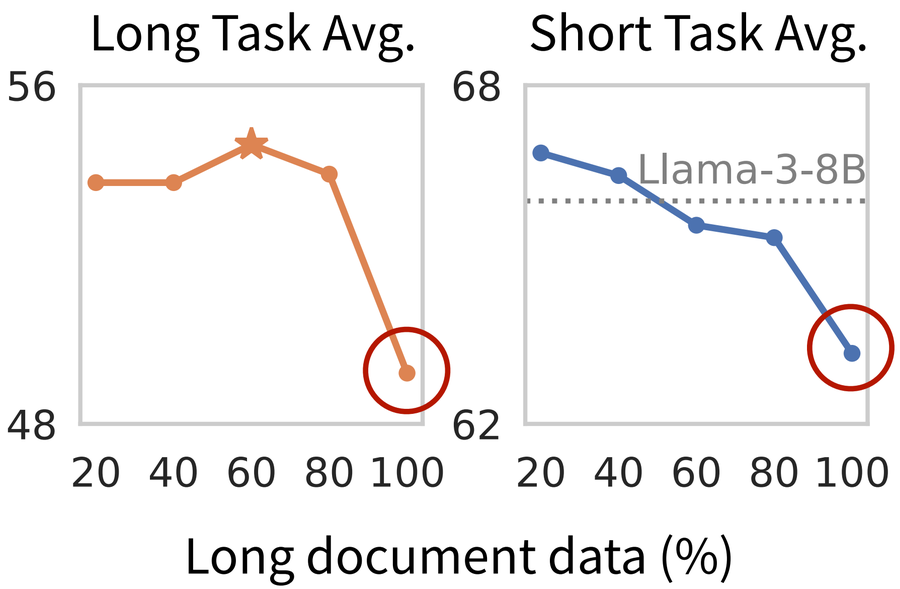


Finding #3: Only using long documents hurts both long-context and standard task performance

The ablation's training and evaluation lengths are both 64K by default. The ablation uses Llama-3-8B-Base and trains on 5B tokens of data.

ProLong: Effective long-context LM data recipe

Question #3: Should we only train on long documents?

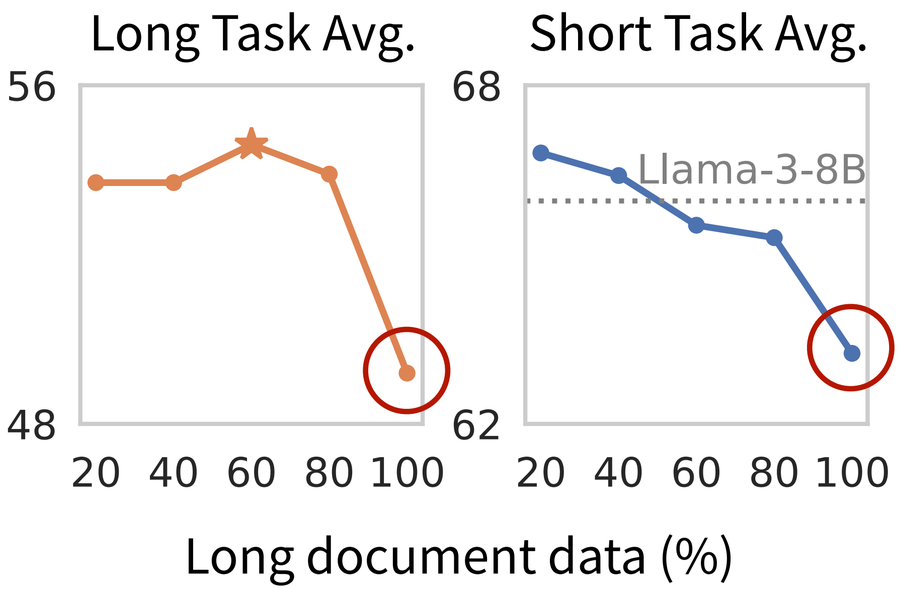


Finding #3: Only using long documents hurts both long-context and standard task performance

The ablation's training and evaluation lengths are both 64K by default. The ablation uses Llama-3-8B-Base and trains on 5B tokens of data.

ProLong: Effective long-context LM data recipe

Question #3: Should we only train on long documents?



Finding #3: Only using long documents hurts both long-context and standard task performance

→ In ProLong, we mix long documents with standard pre-training data (short)

The ablation's training and evaluation lengths are both 64K by default. The ablation uses Llama-3-8B-Base and trains on 5B tokens of data.

ProLong: Effective long-context LM data recipe

Question #4: What domains of long documents best induce long-context capabilities?

All experiment runs use 40% short data mix and 60% domain-specific long documents.

ProLong: Effective long-context LM data recipe

Question #4: What domains of long documents best induce long-context capabilities?

Data

Web-crawled data

ArXiv papers

Long books

Code repos

All experiment runs use 40% short data mix and 60% domain-specific long documents.

ProLong: Effective long-context LM data recipe

Question #4: What domains of long documents best induce long-context capabilities?

Data	Long-context recall	In-context learning
 Web-crawled data		
 ArXiv papers		
 Long books		
 Code repos		

All experiment runs use 40% short data mix and 60% domain-specific long documents.

ProLong: Effective long-context LM data recipe

Question #4: What domains of long documents best induce long-context capabilities?

Data	Long-context recall	In-context learning
 Web-crawled data	69.7%	66.4%
 ArXiv papers	83.0%	67.6%
 Long books	90.3%	**71.4%**
 Code repos	**98.7%**	62.4%

Finding #4: long books / code repos are most effective for long-context training

All experiment runs use 40% short data mix and 60% domain-specific long documents.

ProLong: Effective long-context LM data recipe

Question #4: What domains of long documents best induce long-context capabilities?

Context consistency

Data	Long-context recall	In-context learning
Web-crawled data	69.7%	66.4%
ArXiv papers	83.0%	67.6%
Long books	90.3%	**71.4%**
Code repos	**98.7%**	62.4%

Finding #4: long books / code repos are most effective for long-context training

All experiment runs use 40% short data mix and 60% domain-specific long documents.

ProLong: Effective long-context LM data recipe

Question #4: What domains of long documents best induce long-context capabilities?

	Data	Long-context recall	In-context learning
	Web-crawled data	69.7%	66.4%
	ArXiv papers	83.0%	67.6%
Context consistency	Long books	90.3%	**71.4%**
Long-term dependency	Code repos	**98.7%**	62.4%

Finding #4: long books / code repos are most effective for long-context training

All experiment runs use 40% short data mix and 60% domain-specific long documents.

ProLong: Effective long-context LM data recipe

Question #4: What domains of long documents best induce long-context capabilities?

Many more findings / best practice in the paper!

Long-context instruction tuning

Efficient data loading

Memory efficient long-context training

Sequence parallelism

-context learning

66.4%

67.6%

Context consistency **71.4%**

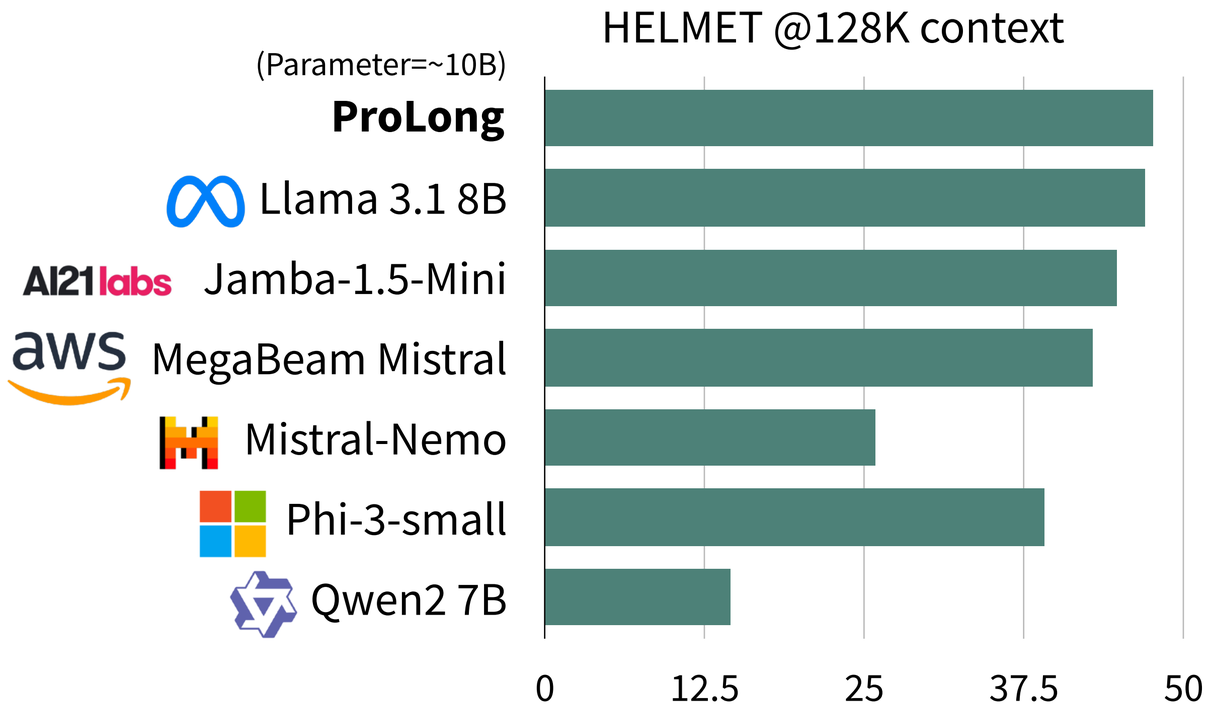
We use both!

Long-term dependency Code repos **98.7%** 62.4%

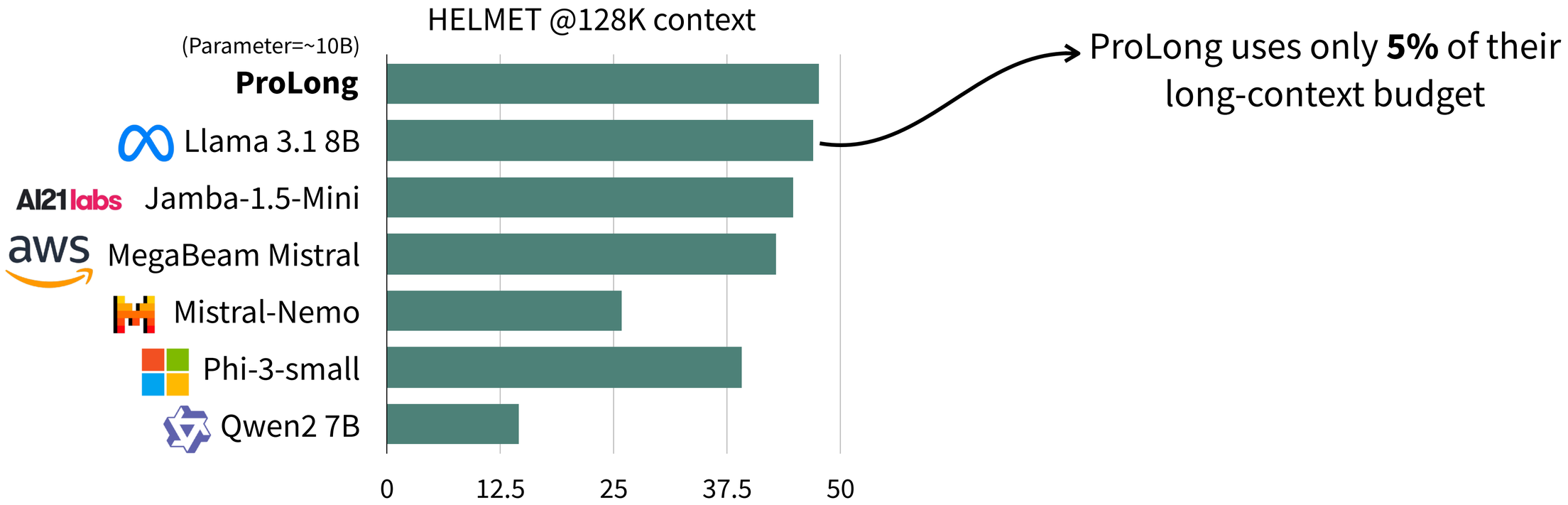
Finding #4: long books / code repos are most effective for long-context training

ProLong achieves state-of-the-art long-context performance

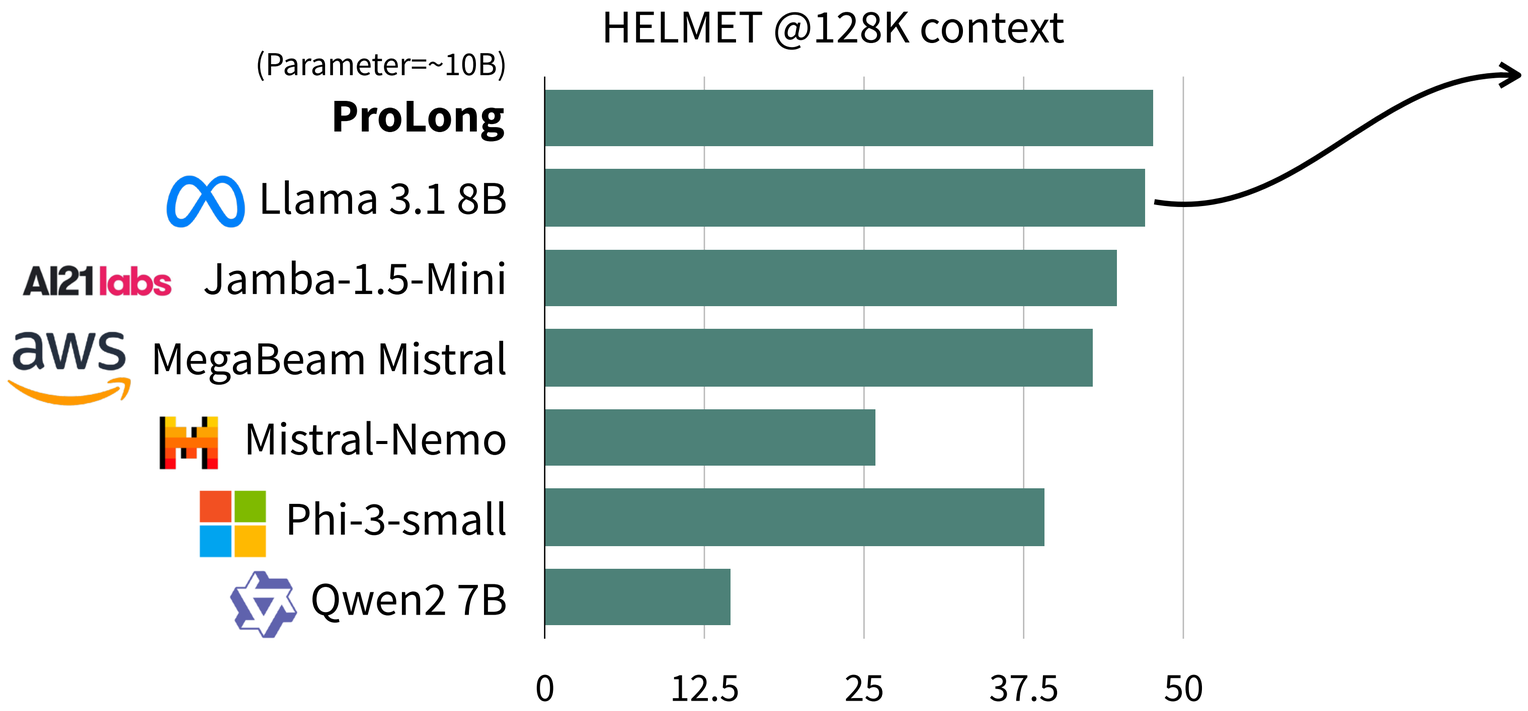
ProLong achieves state-of-the-art long-context performance



ProLong achieves state-of-the-art long-context performance



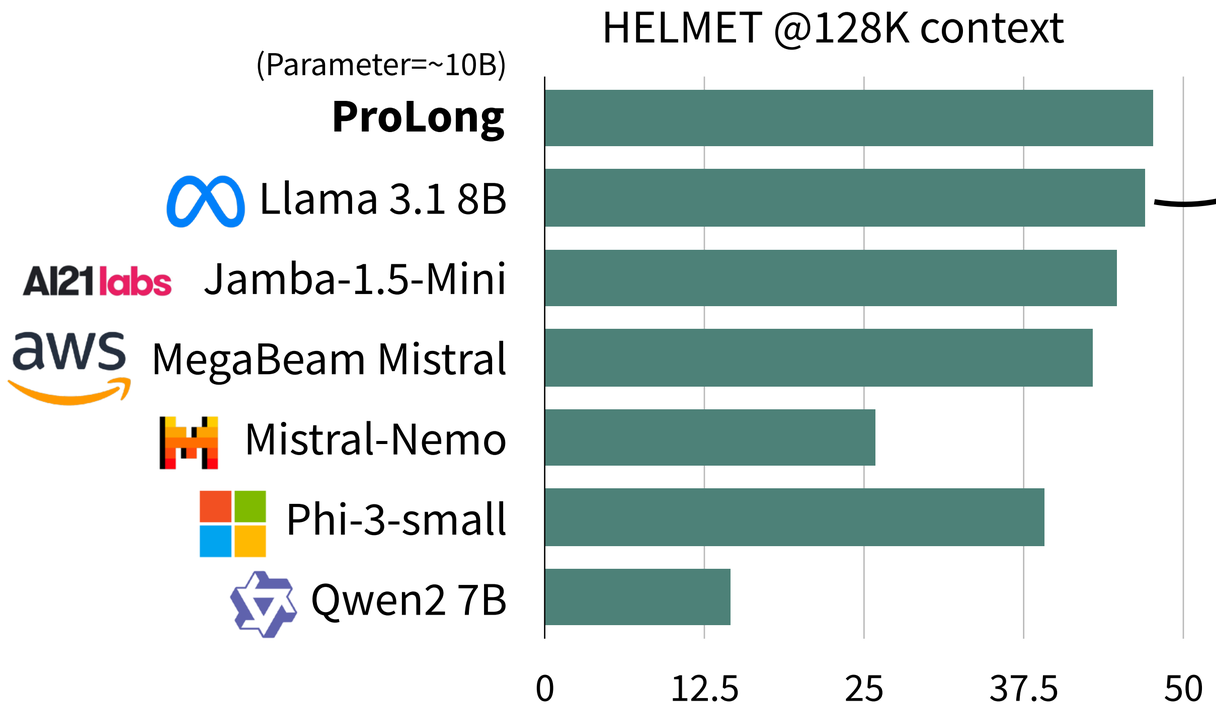
ProLong achieves state-of-the-art long-context performance



ProLong uses only **5%** of their long-context budget

Downloads ⬇️
> 120,000
within 4 months

ProLong achieves state-of-the-art long-context performance



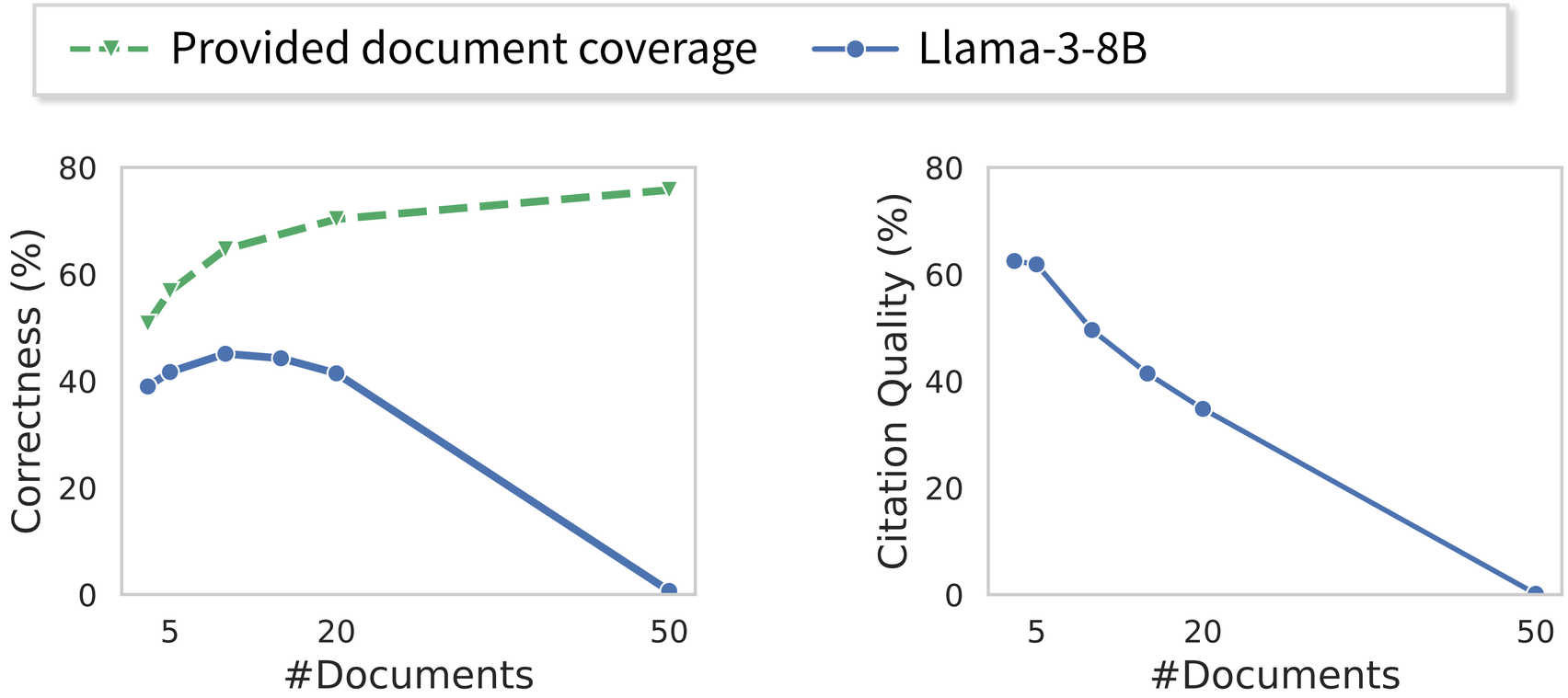
ProLong uses only **5%** of their long-context budget

Downloads ⬇️

> 120,000

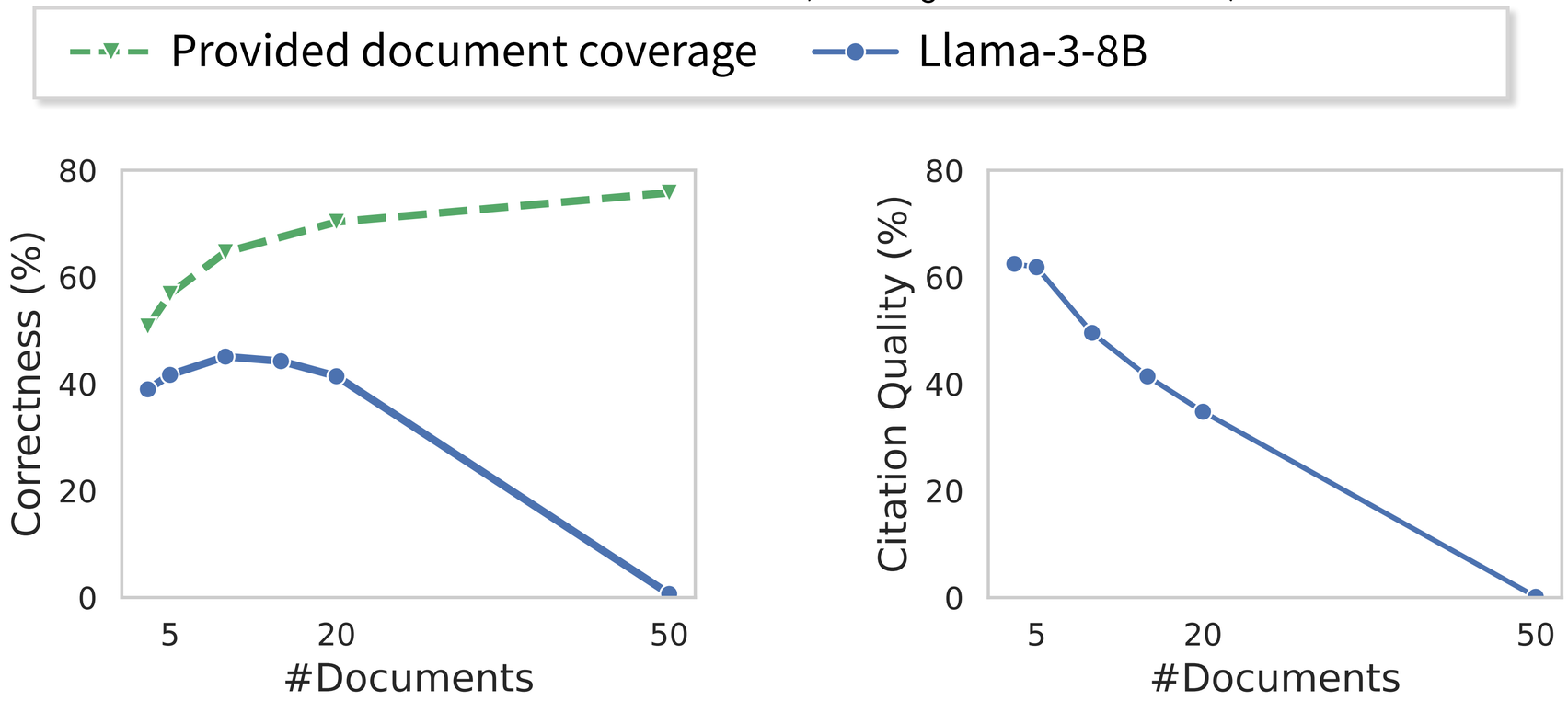
within 4 months

ProLong on generation with citations (ALCE)

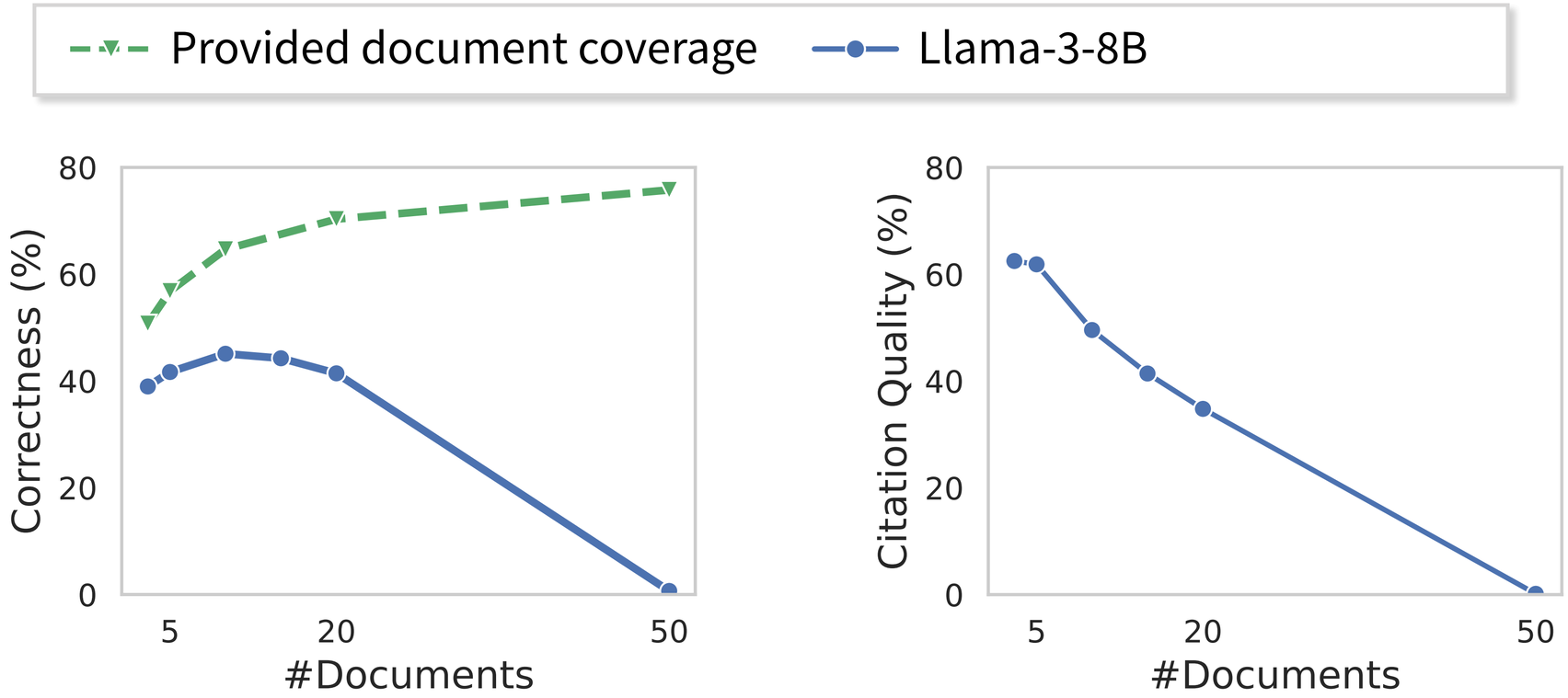


ProLong on generation with citations (ALCE)

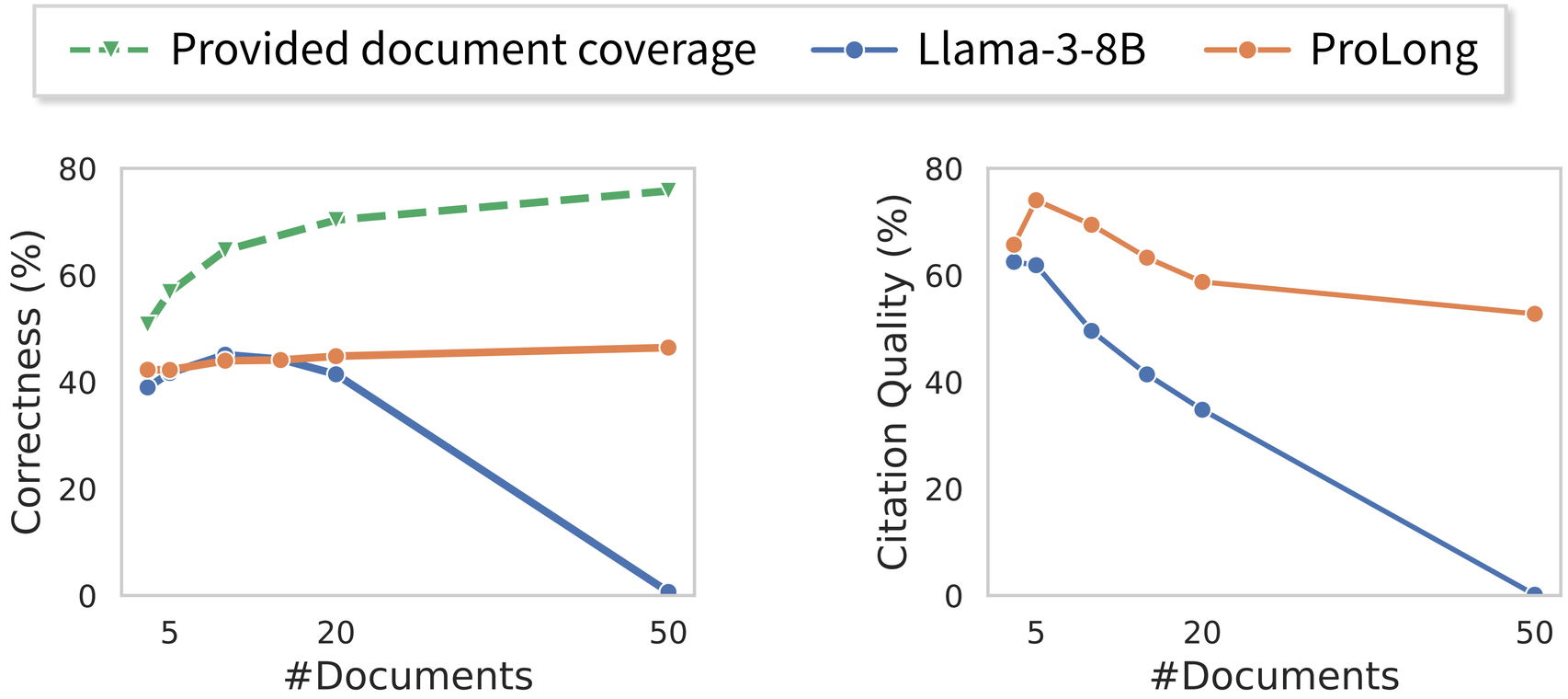
(ProLong is trained from this)



ProLong on generation with citations (ALCE)



ProLong on generation with citations (ALCE)



- **ProLong** can continue to utilize an increasing number of documents in the context
- **ProLong** can reliably cite supporting documents among diverse contexts

Summary: Effective long-context processing

GYYC EMNLP23; P**G**CC Findings of ACL23; Y**G**+ ICLR25; Y**G**C ACL24, **G***W*YC 2024

Summary: Effective long-context processing

Building robust evaluation

GYYC EMNLP23; P**G**CC Findings of ACL23; Y**G**+ ICLR25; Y**G**C ACL24, **G***W*YC 2024

Summary: Effective long-context processing

Building robust evaluation

- **ALCE**: generation with citations
  - Realistic and challenging application
  - First to identify LMs' long-context deficiency
- **HELMET**: comprehensive evaluation

GYYC EMNLP23; P**G**CC Findings of ACL23; Y**G**+ ICLR25; Y**G**C ACL24, **G***W*YC 2024

Summary: Effective long-context processing

Building robust evaluation ⟶ **Developing effective long-context models**

- **ALCE**: generation with citations
  - Realistic and challenging application
  - First to identify LMs' long-context deficiency
- **HELMET**: comprehensive evaluation

GYYC EMNLP23; P**G**CC Findings of ACL23; Y**G**+ ICLR25; Y**G**C ACL24, **G***W*YC 2024

Summary: Effective long-context processing

Building robust evaluation

- **ALCE**: generation with citations
  - Realistic and challenging application
  - First to identify LMs' long-context deficiency
- **HELMET**: comprehensive evaluation

Developing effective long-context models

- **ProLong**: thoroughly-ablated long-context recipe
  - Outperforming industry effort with only **5%** of its computational budget

GYYC EMNLP23; P**G**CC Findings of ACL23; Y**G**+ ICLR25; Y**G**C ACL24, **G***W*YC 2024

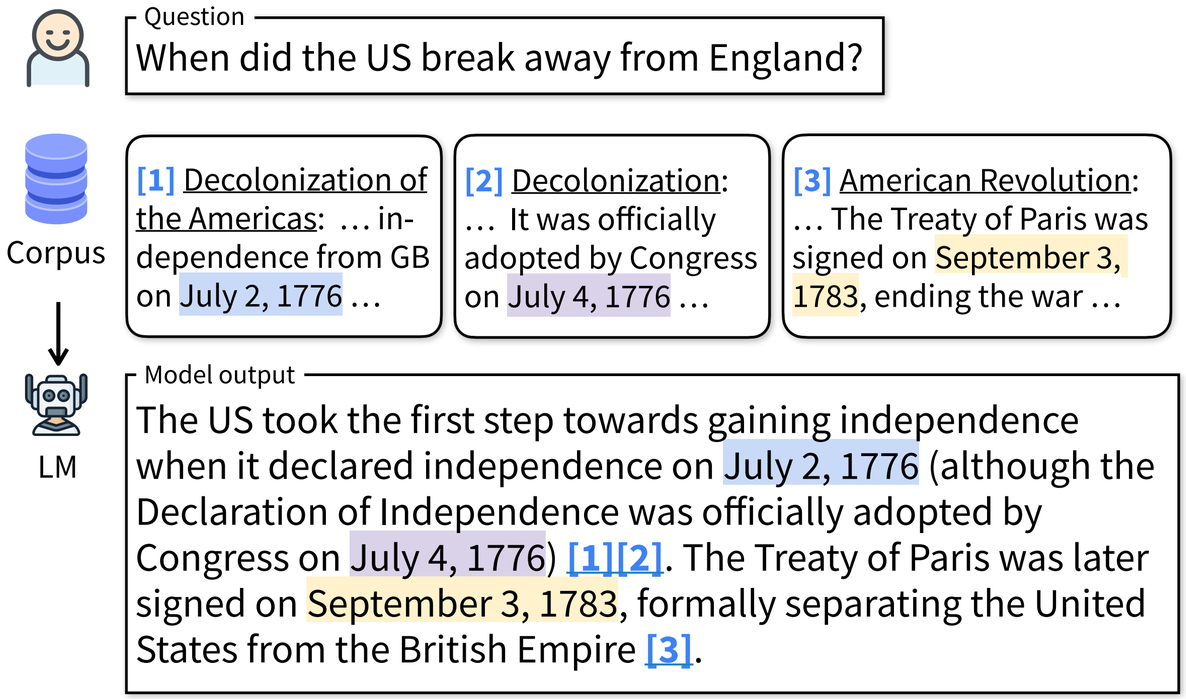
Enabling LMs to process information at scale

1. Effective long-context processing

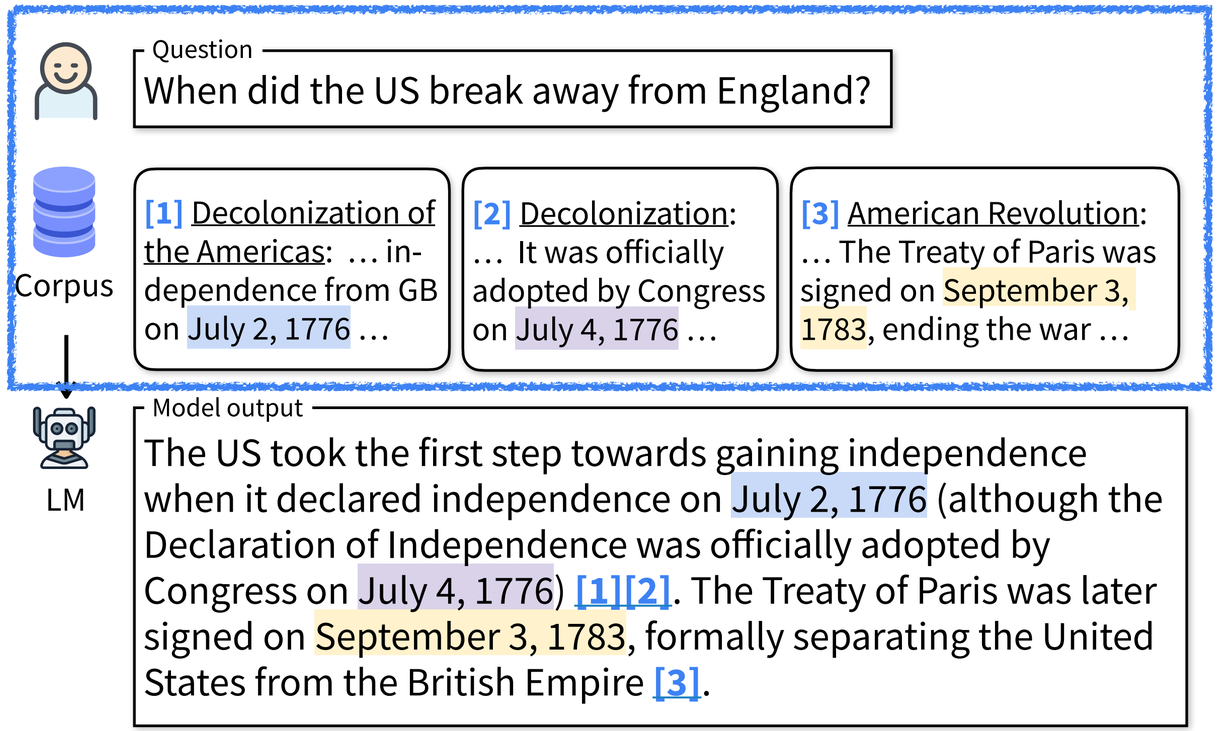
2. Accurate search via text embeddings

3. Foundations: Efficient language models

Retrieval is key in retrieval-augmented generation

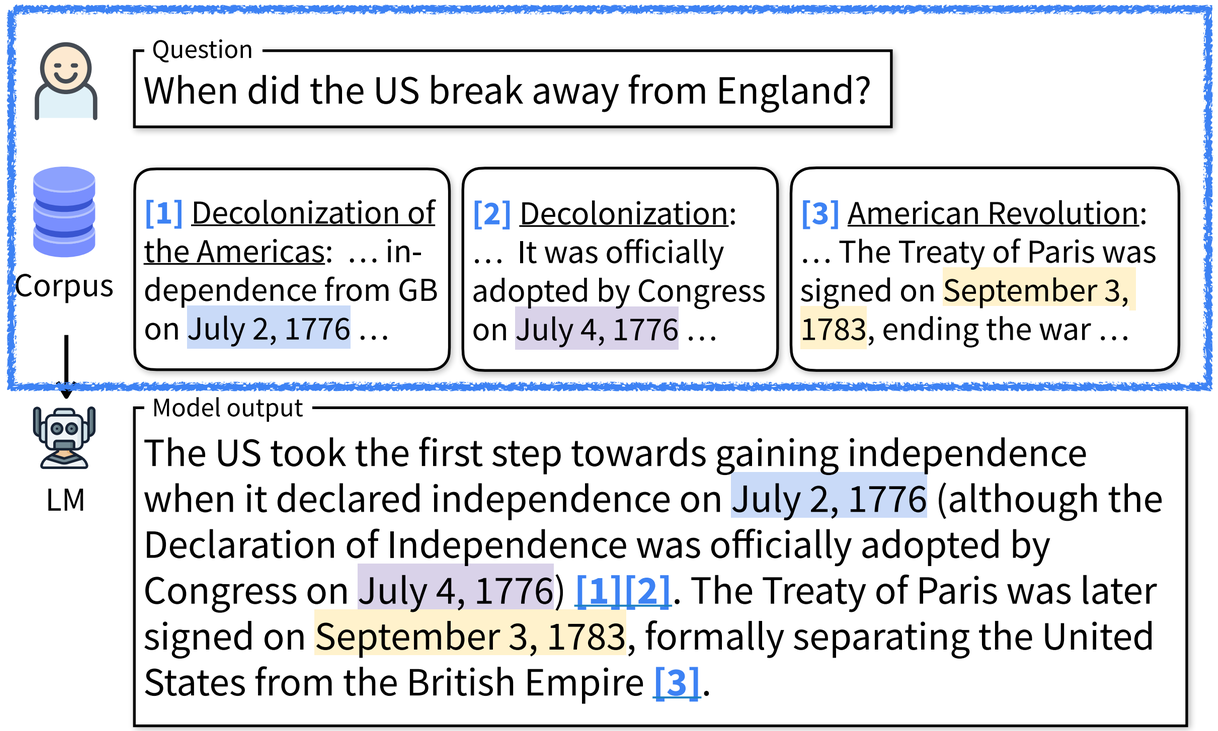


Retrieval is key in retrieval-augmented generation



Key component: retrieval

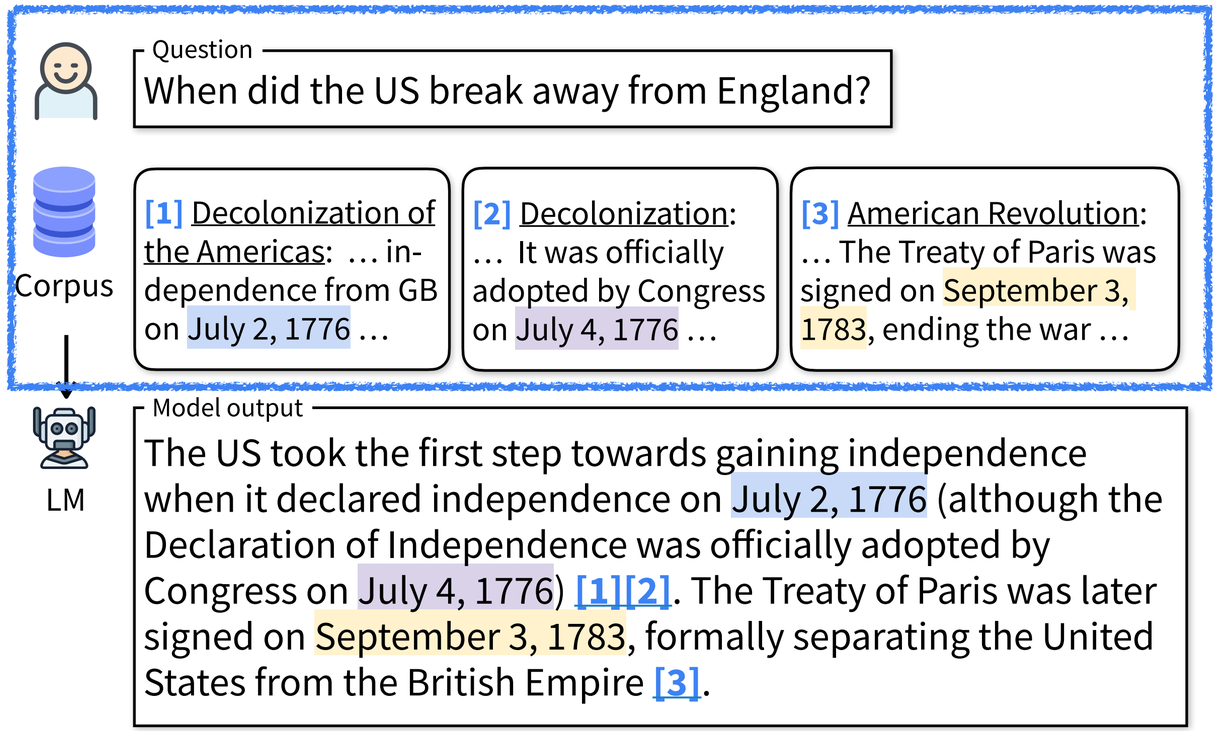
Retrieval is key in retrieval-augmented generation



Key component: retrieval

- Traditionally: keyword matching

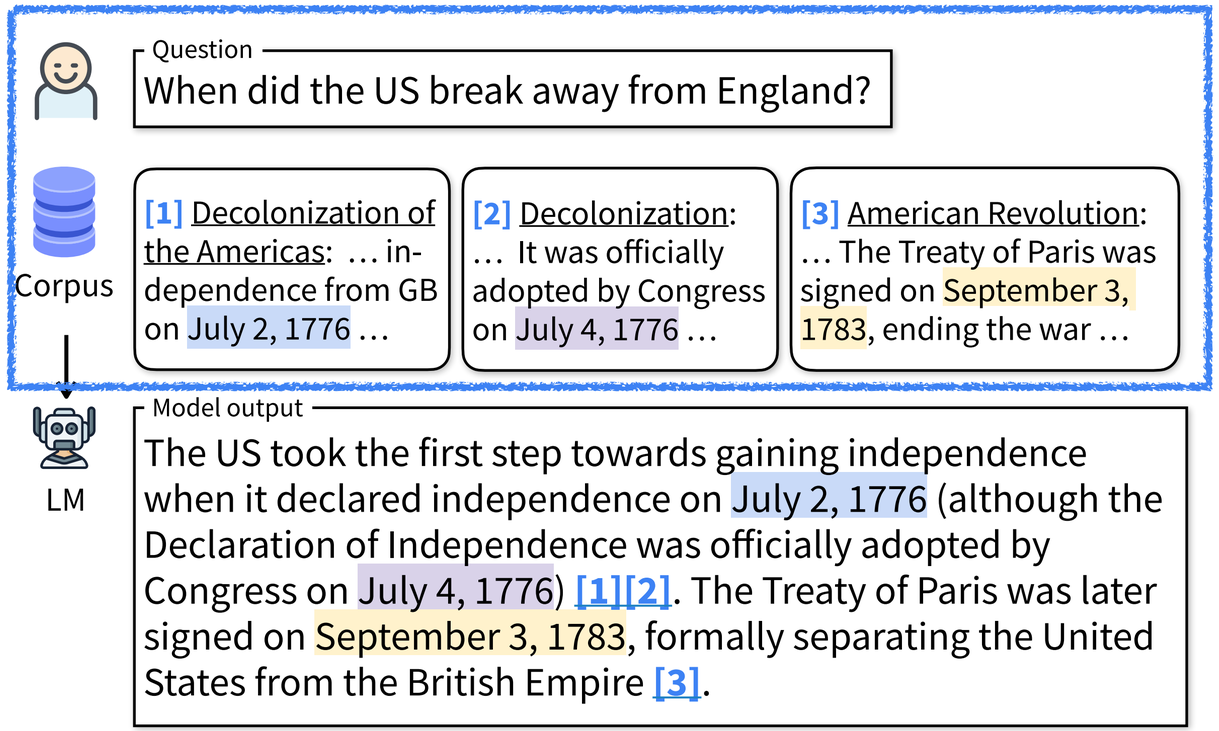
Retrieval is key in retrieval-augmented generation



Key component: retrieval

- Traditionally: keyword matching
- New: text embeddings

Retrieval is key in retrieval-augmented generation



Key component: retrieval

- Traditionally: keyword matching
- New: text embeddings

Better captures semantic meanings

Text embeddings

Text embeddings

BERT achieves state-of-the-art performance on …

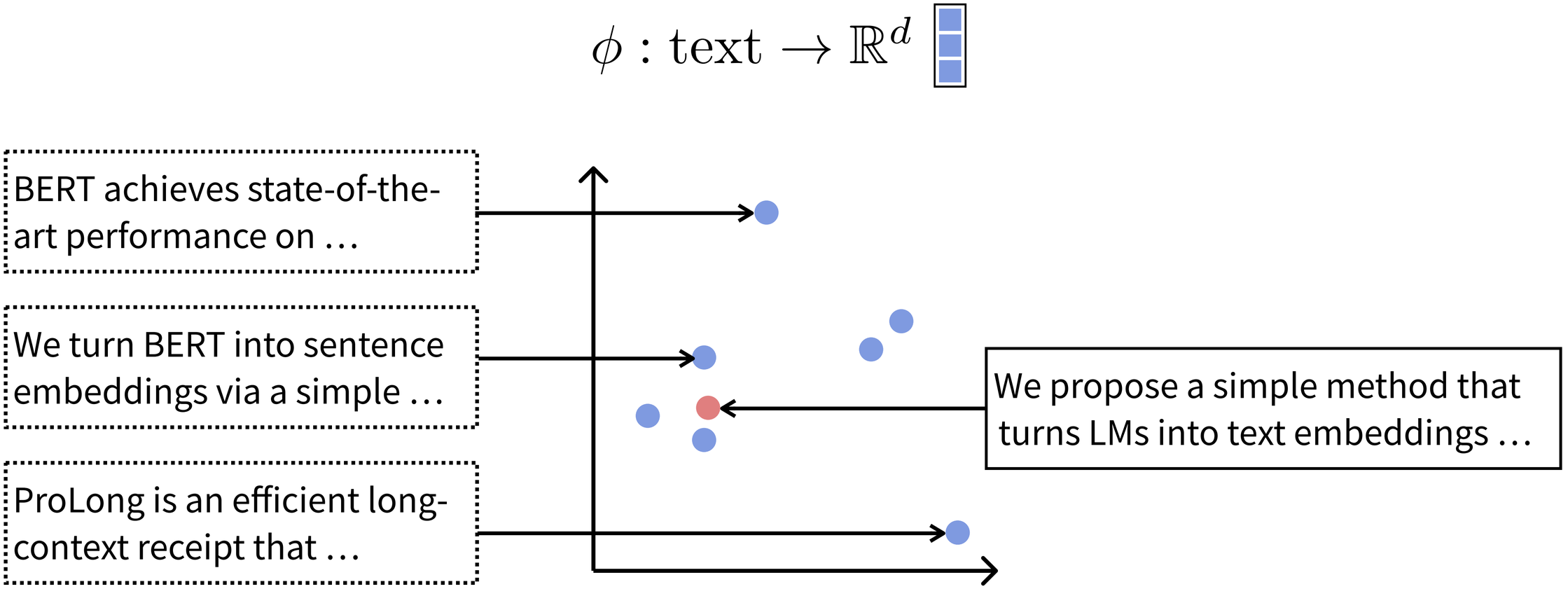
We turn BERT into sentence embeddings via a simple …

ProLong is an efficient long-context receipt that …

We propose a simple method that turns LMs into text embeddings …

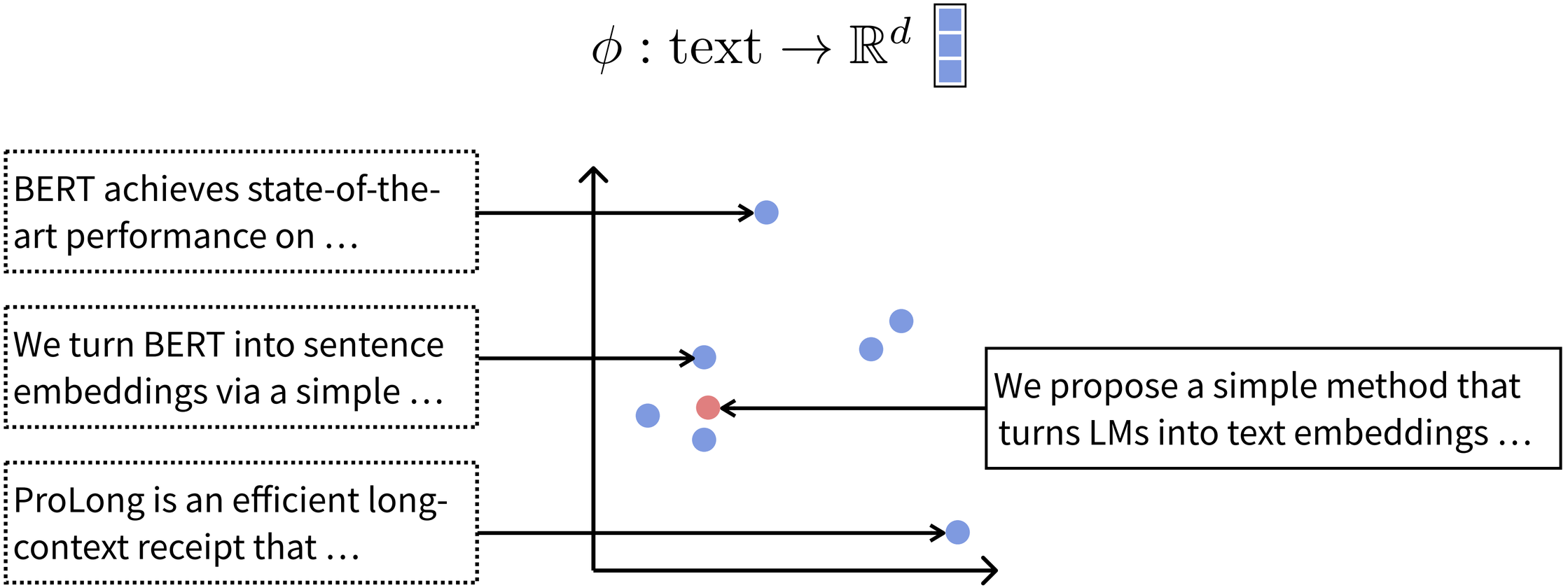
Text embeddings

$$\phi : \text{text} \rightarrow \mathbb{R}^d$$



Text embeddings

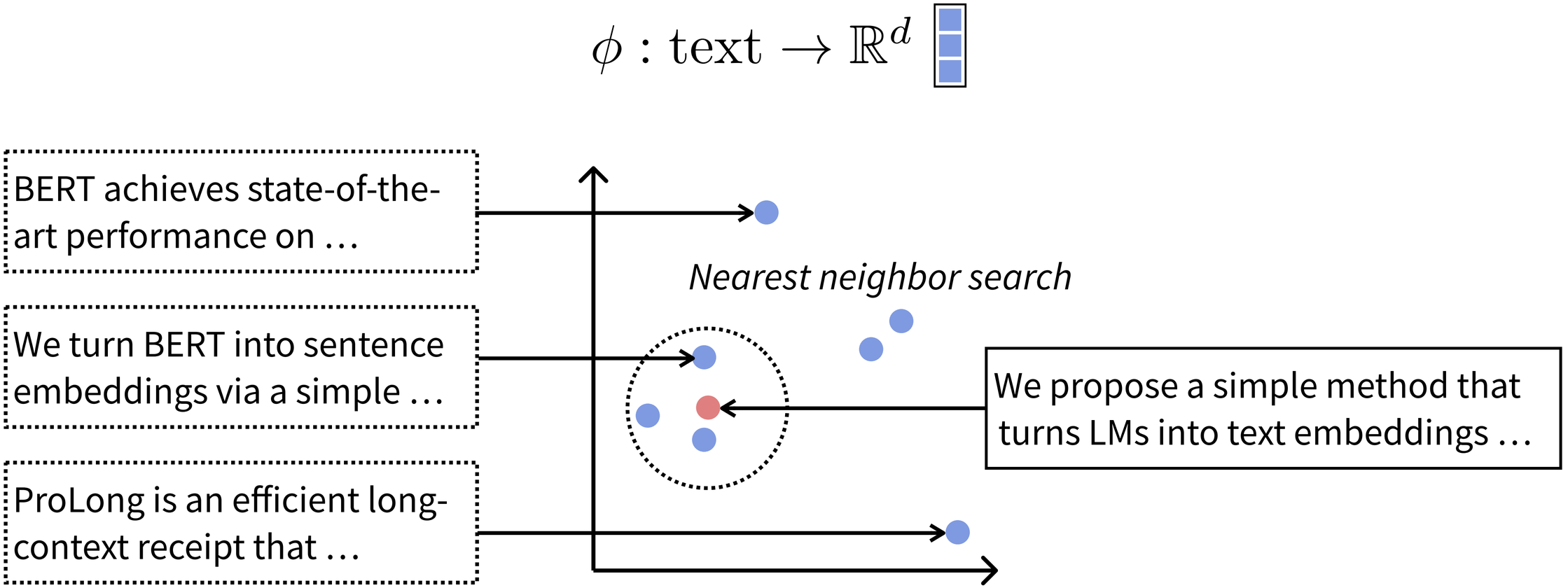
$$\phi : \text{text} \rightarrow \mathbb{R}^d$$



Applications: clustering, visualization, measuring similarities, retrieval

Text embeddings

$$\phi : \text{text} \rightarrow \mathbb{R}^d$$



Applications: clustering, visualization, measuring similarities, retrieval

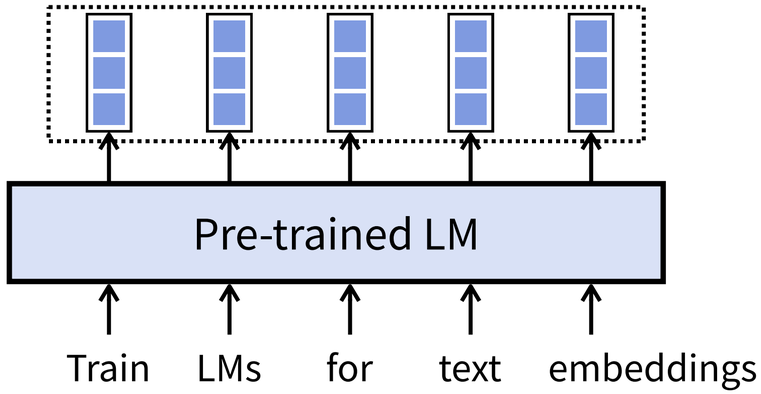
Text embeddings: Development

Text embeddings: Development

Before 2021

Text embeddings: Development

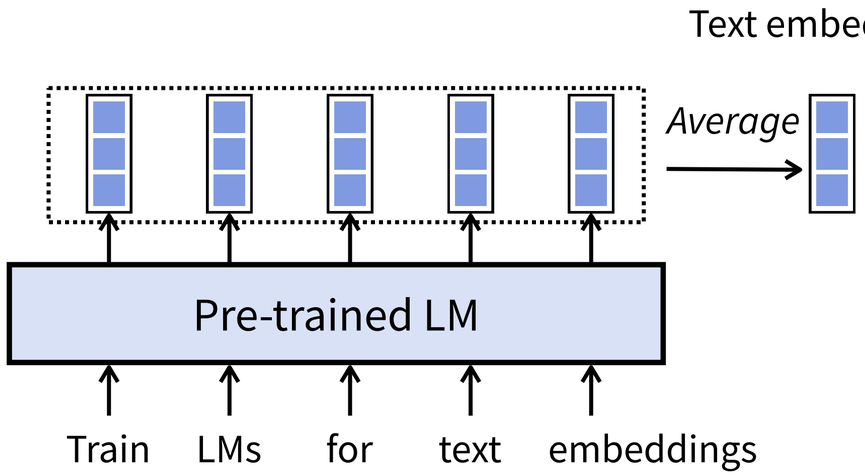
Before 2021



Directly use a pre-trained LM (e.g., BERT)’s embeddings

Text embeddings: Development

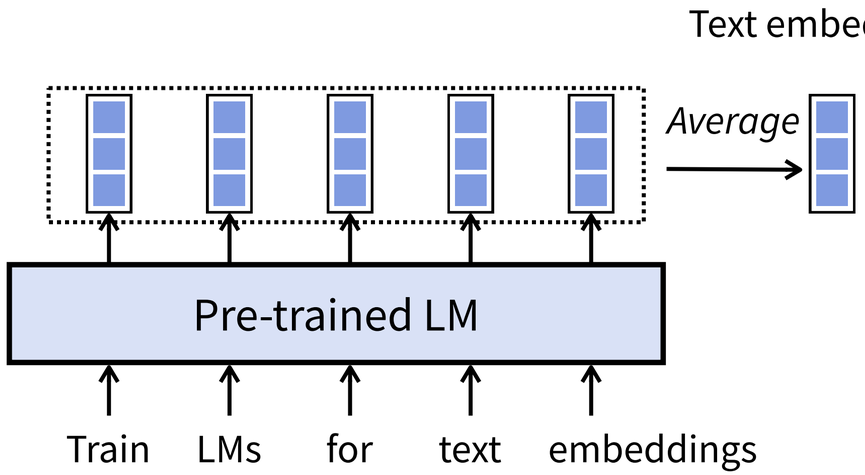
Before 2021



Directly use a pre-trained LM (e.g., BERT)'s embeddings

Text embeddings: Development

Before 2021

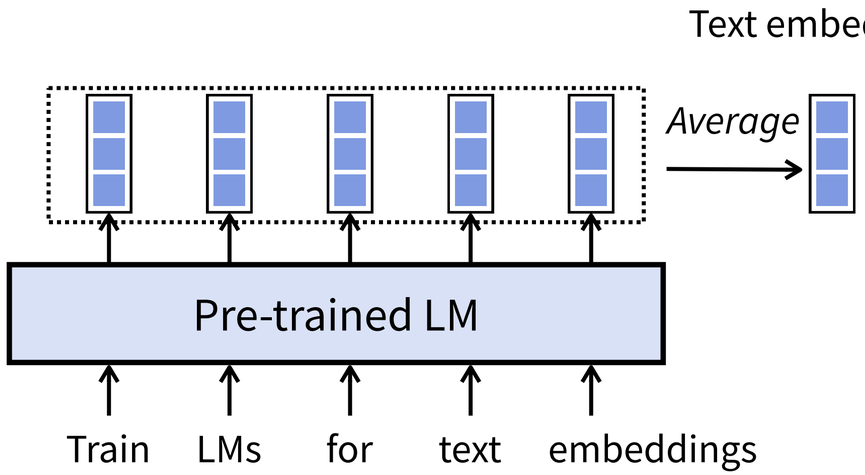


Directly use a pre-trained LM (e.g., BERT)'s embeddings

Train the embeddings on a classification task
(Reimers and Gurevych, 2019)

Text embeddings: Development

Before 2021



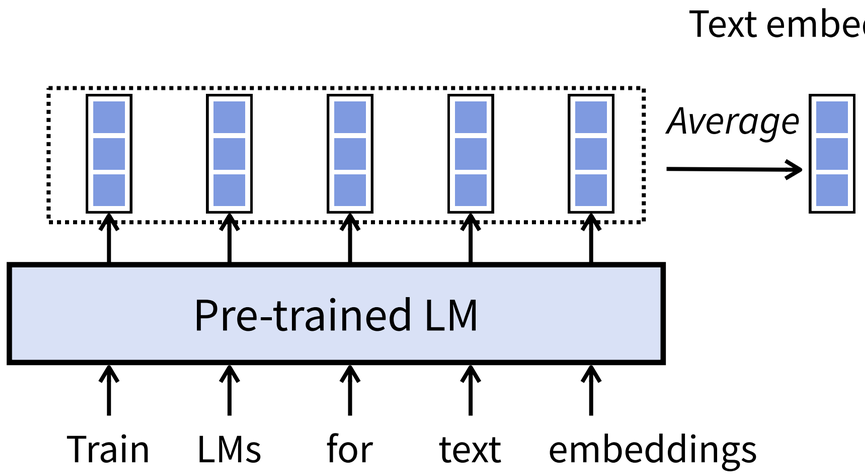
Directly use a pre-trained LM (e.g., BERT)'s embeddings

Train the embeddings on a classification task
(Reimers and Gurevych, 2019)

Do not always outperform keyword matching
(Thakur et al., 2021)

Text embeddings: Development

Before 2021



Directly use a pre-trained LM (e.g., BERT)'s embeddings

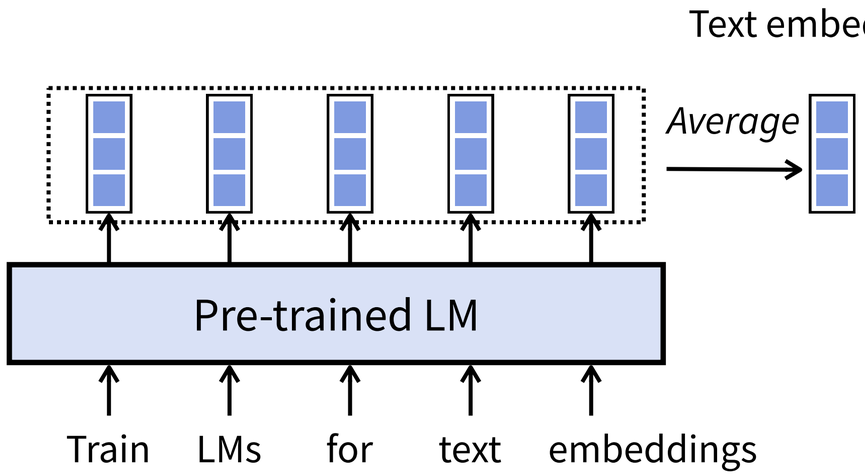
Train the embeddings on a classification task
(Reimers and Gurevych, 2019)

Do not always outperform keyword matching
(Thakur et al., 2021)

2021

Text embeddings: Development

Before 2021



Directly use a pre-trained LM (e.g., BERT)'s embeddings

Train the embeddings on a classification task
(Reimers and Gurevych, 2019)

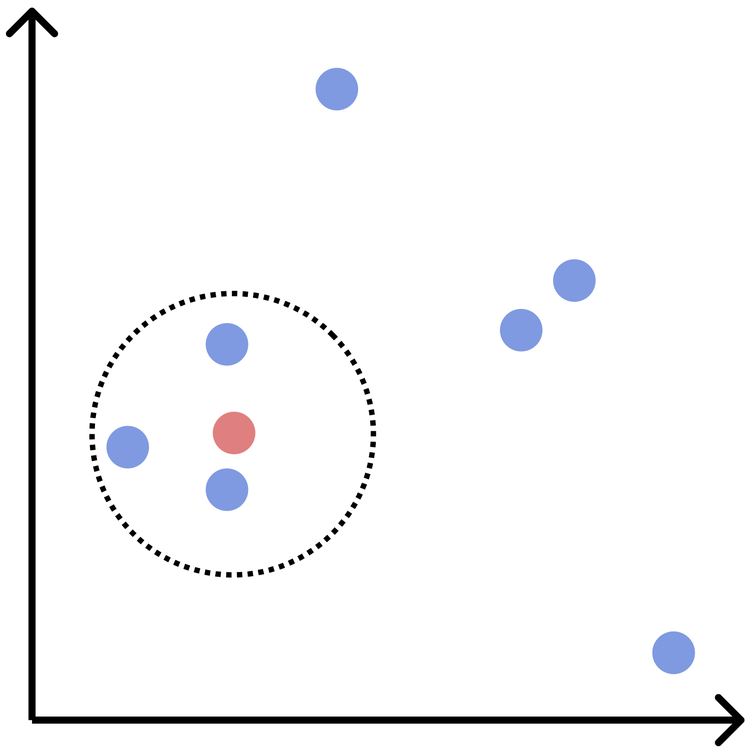
Do not always outperform keyword matching
(Thakur et al., 2021)

2021 **SimCSE**: a simple contrastive learning objective that can better **transform pre-trained LM** into **performant text embeddings**

most influential paper

T. Gao*, X. Yao*, D. Chen. EMNLP 2021. **SimCSE: Simple Contrastive Learning of Sentence Embeddings**.

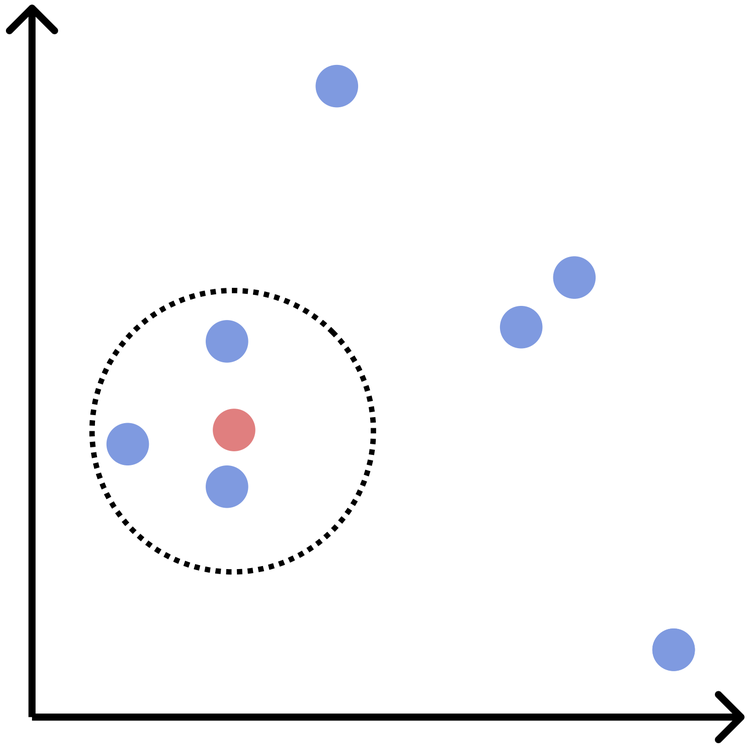
SimCSE: contrastive learning objective



most influential paper

T. Gao*, X. Yao*, D. Chen. EMNLP 2021. **SimCSE: Simple Contrastive Learning of Sentence Embeddings**.

SimCSE: contrastive learning objective

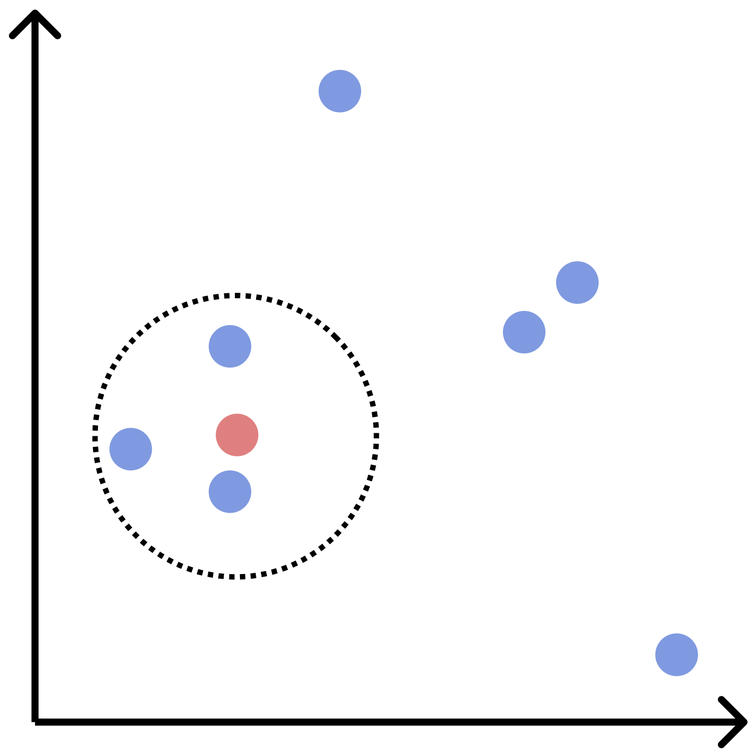


Contrastive learning

most influential paper

T. Gao*, X. Yao*, D. Chen. EMNLP 2021. **SimCSE: Simple Contrastive Learning of Sentence Embeddings**.

SimCSE: contrastive learning objective



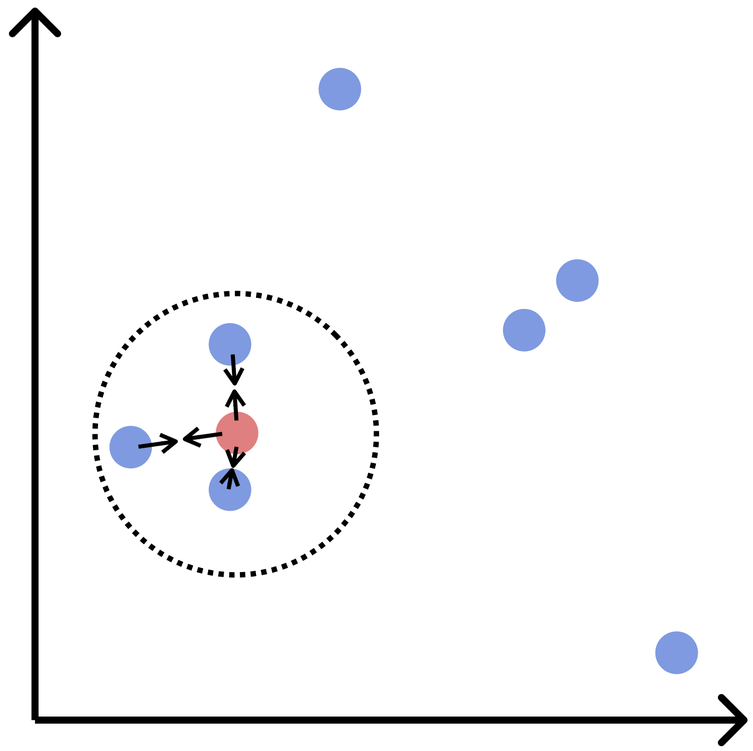
Contrastive learning

Idea: pulling together **positive pairs**, pushing apart **negative pairs**

most influential paper

T. Gao*, X. Yao*, D. Chen. EMNLP 2021. **SimCSE: Simple Contrastive Learning of Sentence Embeddings**.

SimCSE: contrastive learning objective



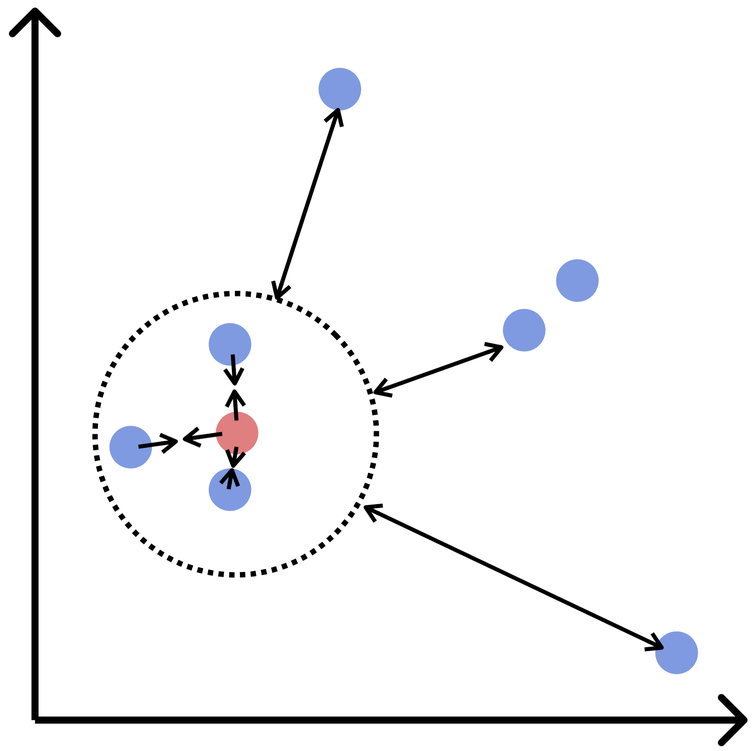
Contrastive learning

Idea: pulling together **positive pairs**, pushing apart **negative pairs**

most influential paper

T. Gao*, X. Yao*, D. Chen. EMNLP 2021. **SimCSE: Simple Contrastive Learning of Sentence Embeddings**.

SimCSE: contrastive learning objective



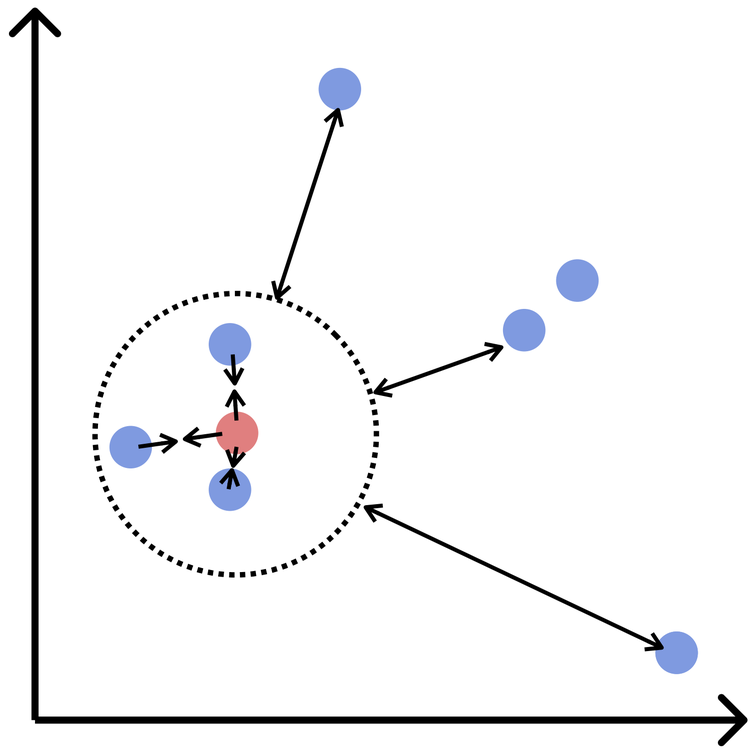
Contrastive learning

Idea: pulling together **positive pairs**, pushing apart **negative pairs**

most influential paper

T. Gao*, X. Yao*, D. Chen. EMNLP 2021. **SimCSE: Simple Contrastive Learning of Sentence Embeddings**.

SimCSE: contrastive learning objective



Contrastive learning

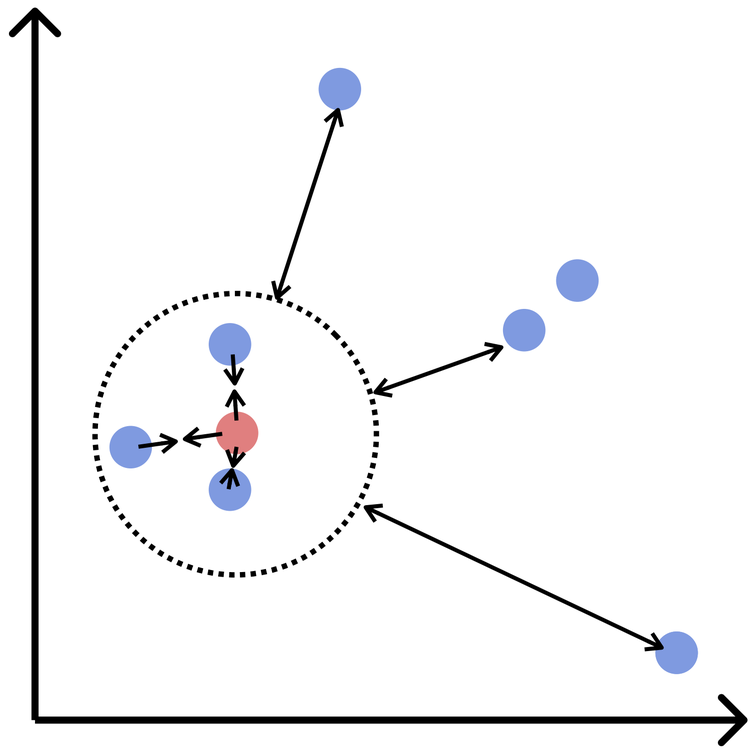
Idea: pulling together **positive pairs**, pushing apart **negative pairs**

InfoNCE loss (Oord et al., 2018; Chen et al., 2020)

most influential paper

T. Gao*, X. Yao*, D. Chen. EMNLP 2021. **SimCSE: Simple Contrastive Learning of Sentence Embeddings**.

SimCSE: contrastive learning objective



Contrastive learning

Idea: pulling together **positive pairs**, pushing apart **negative pairs**

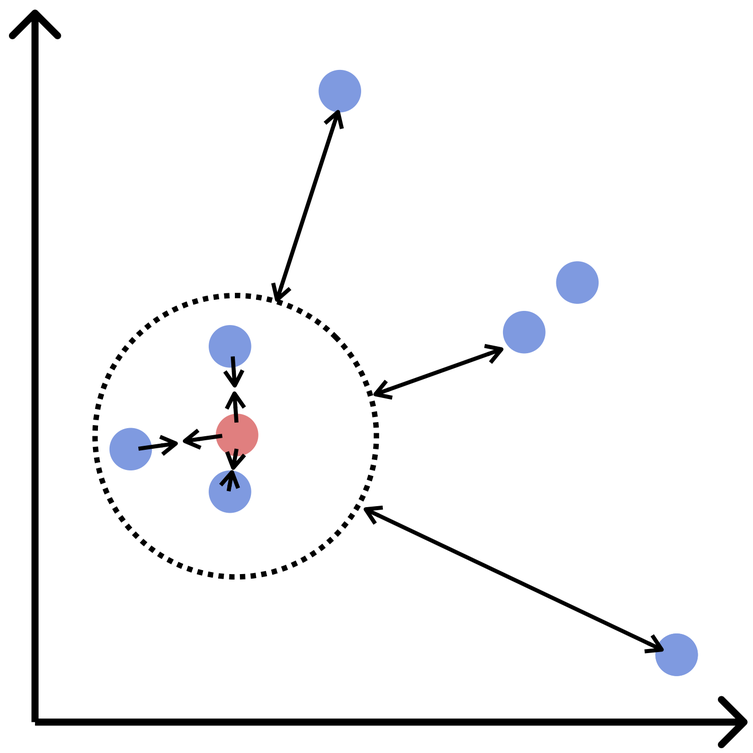
InfoNCE loss (Oord et al., 2018; Chen et al., 2020)

$$l_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_j)/\tau)}{\sum_{k \neq i} \exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_k)/\tau)}$$

most influential paper

T. Gao*, X. Yao*, D. Chen. EMNLP 2021. **SimCSE: Simple Contrastive Learning of Sentence Embeddings**.

SimCSE: contrastive learning objective



Contrastive learning

Idea: pulling together **positive pairs**, pushing apart **negative pairs**

InfoNCE loss (Oord et al., 2018; Chen et al., 2020)

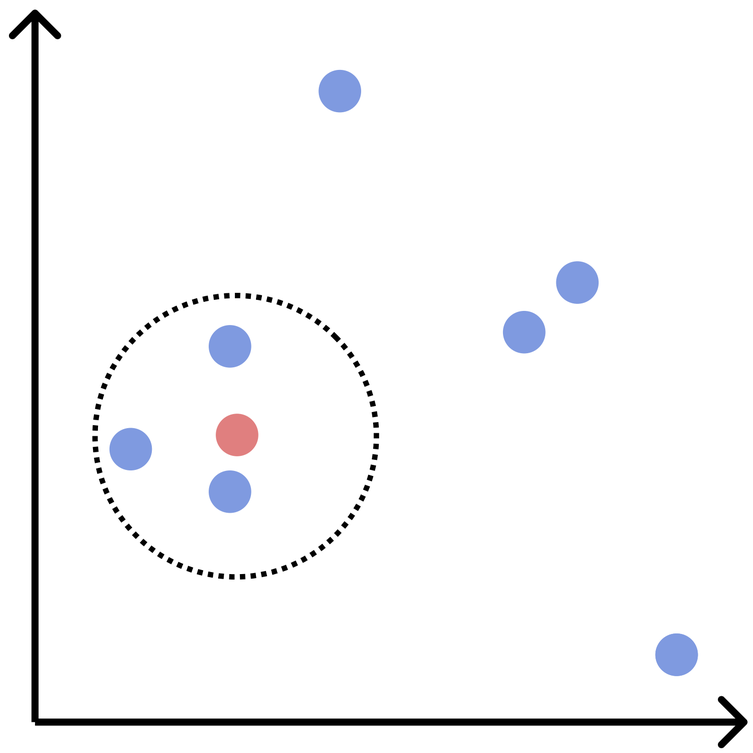
Similarity function (cosine) Sentence embedding

$$l_{i,j} = -\log \frac{\exp(\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_j)/\tau)}{\sum_{k \neq i} \exp(\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_k)/\tau)}$$

most influential paper

T. Gao*, X. Yao*, D. Chen. EMNLP 2021. **SimCSE: Simple Contrastive Learning of Sentence Embeddings**.

SimCSE: contrastive learning objective



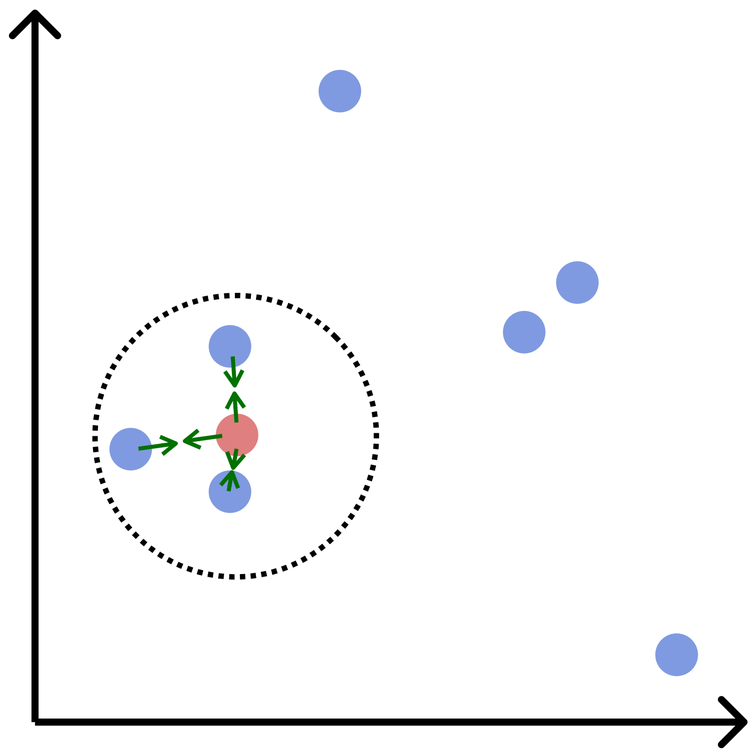
Contrastive learning

Idea: pulling together **positive pairs**, pushing apart **negative pairs**

InfoNCE loss (Oord et al., 2018; Chen et al., 2020)

$$l_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_j)/\tau)}{\sum_{k \neq i} \exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_k)/\tau)}$$

SimCSE: contrastive learning objective



Contrastive learning

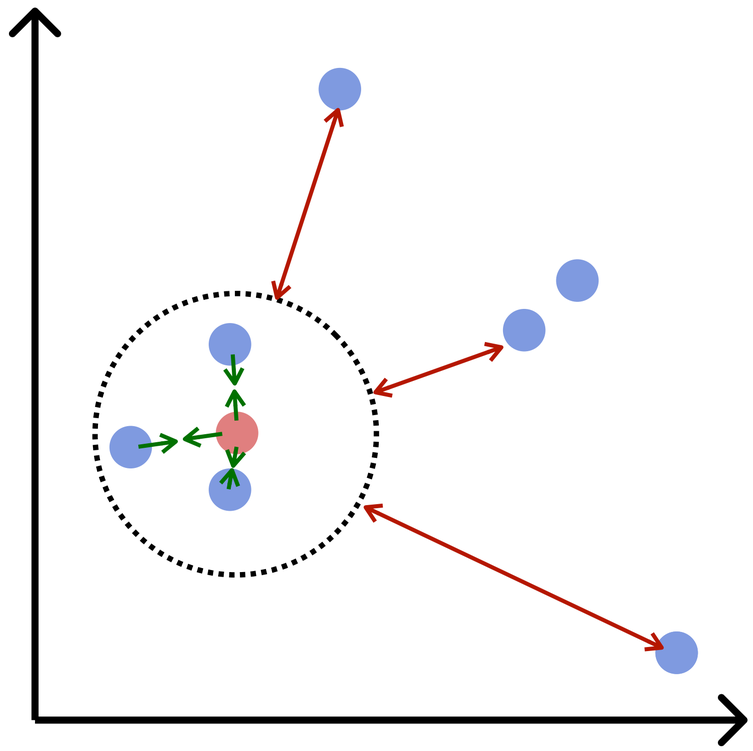
Idea: pulling together **positive pairs**, pushing apart **negative pairs**

InfoNCE loss (Oord et al., 2018; Chen et al., 2020)

Positive pairs ↑

$$l_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_j)/\tau)}{\sum_{k \neq i} \exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_k)/\tau)}$$

SimCSE: contrastive learning objective



Contrastive learning

Idea: pulling together **positive pairs**, pushing apart **negative pairs**

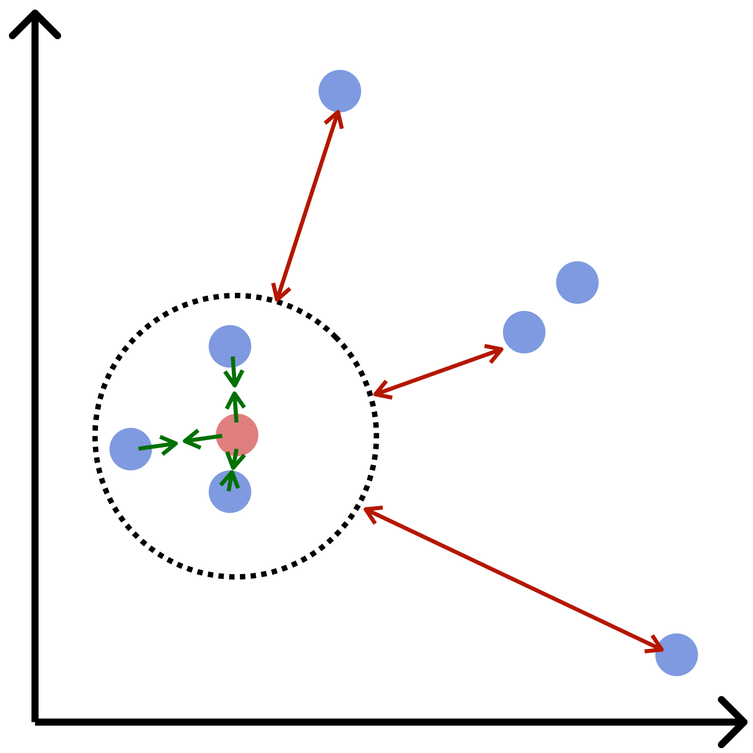
InfoNCE loss (Oord et al., 2018; Chen et al., 2020)

Positive pairs ↑

$$l_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_j)/\tau)}{\sum_{k \neq i} \exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_k)/\tau)}$$

Negative pairs ↓

SimCSE: contrastive learning objective



Contrastive learning

Idea: pulling together **positive pairs**, pushing apart **negative pairs**

InfoNCE loss (Oord et al., 2018; Chen et al., 2020)

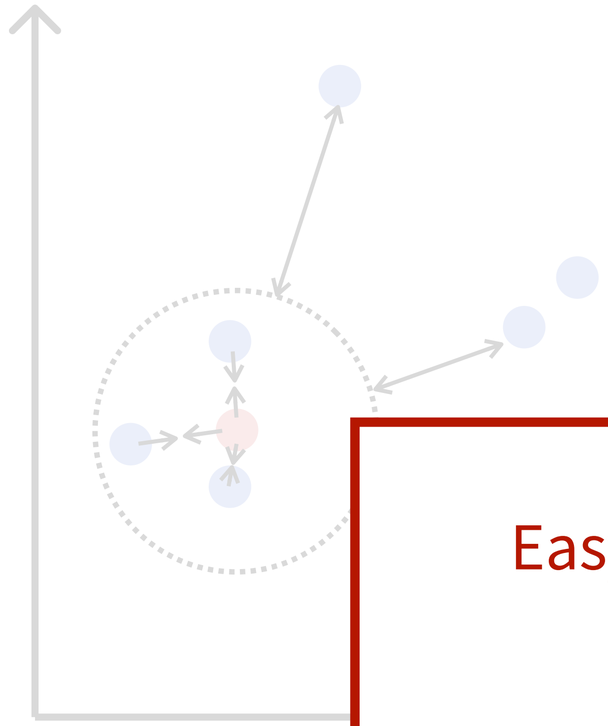
$$l_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_j)/\tau)}{\sum_{k \neq i} \exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_k)/\tau)}$$

Positive pairs ↑

Negative pairs ↓

*Question: what **positive**/**negative** pairs to choose?*

SimCSE: contrastive learning objective



Contrastive learning

Idea: pulling together **positive pairs**, pushing apart **negative pairs**

InfoNCE loss [Oord et al., 2018; Chen et al., 2020]

Positive pairs ↑

Easy negatives: other randomly-sampled instances in the same mini-batch (Chen et al., 2020)

*Question: what **positive**/**negative** pairs to choose?*

SimCSE: How to construct positive pairs

SimCSE: How to construct positive pairs

SimCSE: How to construct positive pairs

Data augmentation

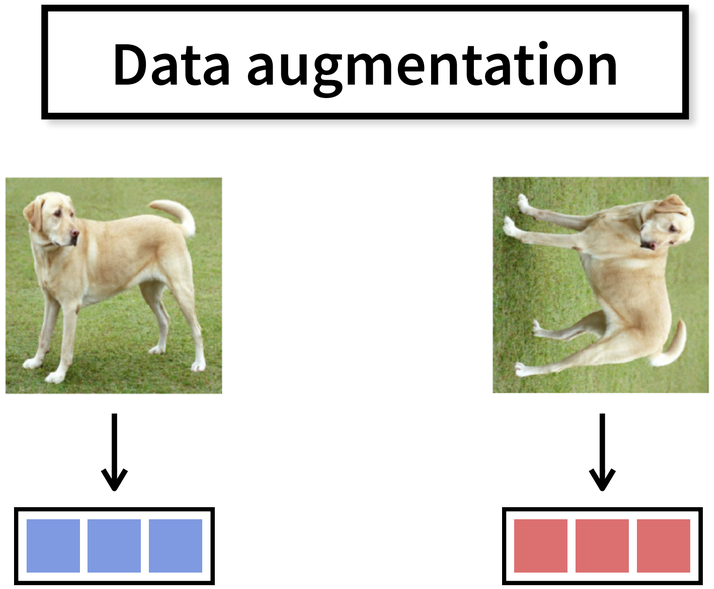
Computer vision:
image data augmentation
as positive pairs (Chen et al., 2020)

SimCSE: How to construct positive pairs

Data augmentation

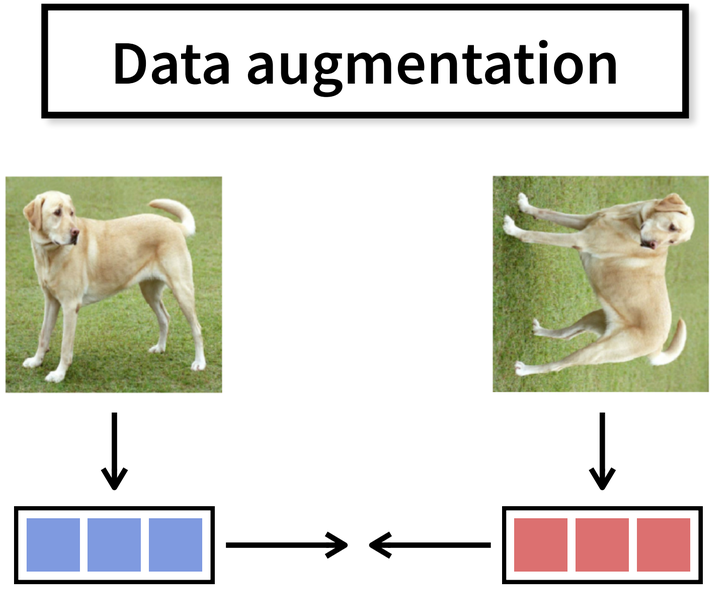
Computer vision:
image data augmentation
as positive pairs (Chen et al., 2020)

SimCSE: How to construct positive pairs



Computer vision:
image data augmentation
as positive pairs (Chen et al., 2020)

SimCSE: How to construct positive pairs



Computer vision:
image data augmentation
as positive pairs (Chen et al., 2020)

SimCSE: How to construct positive pairs

Data augmentation

SimCSE: How to construct positive pairs

Data augmentation

He was
running away

SimCSE: How to construct positive pairs

Data augmentation

He was running away

He was running ~~away~~

SimCSE: How to construct positive pairs

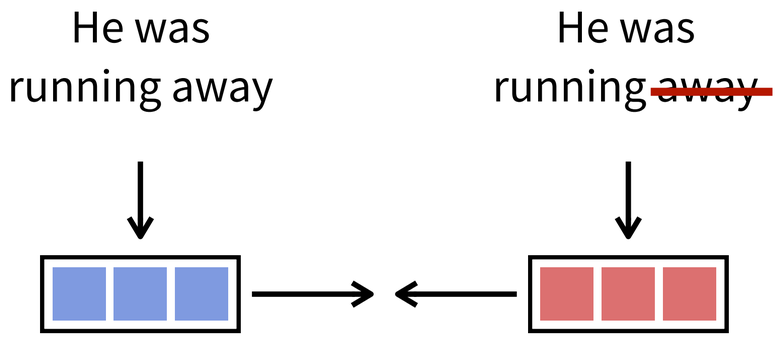
Data augmentation

He was running away	He was running ~~away~~

Even deleting one word could change the meaning of a sentence

SimCSE: How to construct positive pairs

Data augmentation



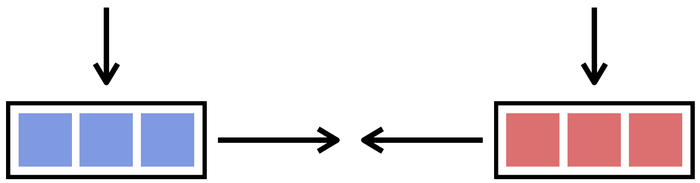
Even deleting one word could change the meaning of a sentence

SimCSE: How to construct positive pairs

Data augmentation

He was running away

He was running ~~away~~



(Semantic text similarity)

STS score: 63%

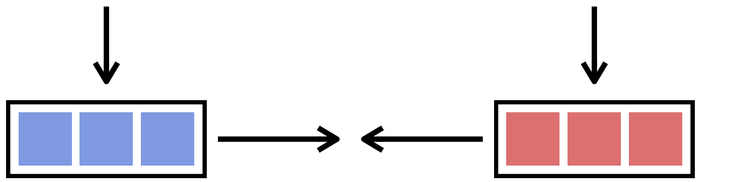
Even deleting one word could change the meaning of a sentence

SimCSE: How to construct positive pairs

Data augmentation

He was running away

He was running ~~away~~

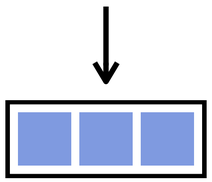


(Semantic text similarity)

STS score: 63%

Identical

He was running away



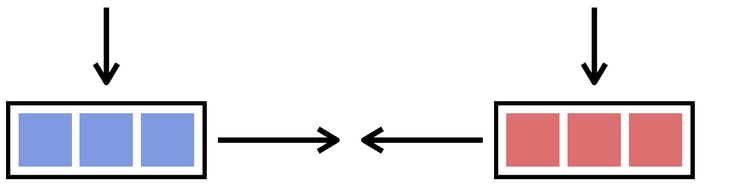
Even deleting one word could change the meaning of a sentence

SimCSE: How to construct positive pairs

Data augmentation

He was running away

He was running ~~away~~



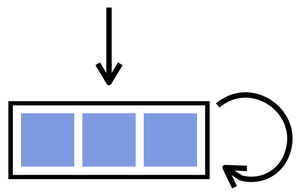
(Semantic text similarity)

STS score: 63%

Even deleting one word could change the meaning of a sentence

Identical

He was running away

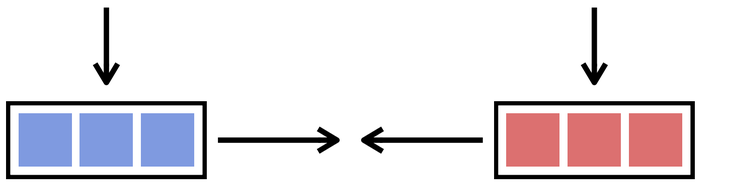


SimCSE: How to construct positive pairs

Data augmentation

He was running away

He was running ~~away~~



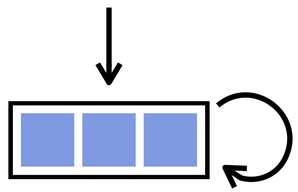
(Semantic text similarity)

STS score: 63%

Even deleting one word could change the meaning of a sentence

Identical

He was running away



STS score: 43%

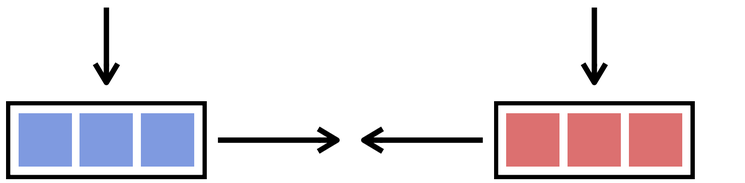
Embeddings collapse without positive-pair supervision

SimCSE: How to construct positive pairs

Data augmentation

He was running away

He was running ~~away~~



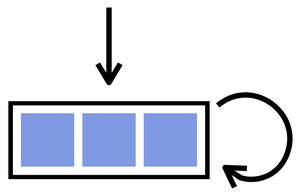
(Semantic text similarity)

STS score: 63%

Even deleting one word could change the meaning of a sentence

Identical

He was running away



STS score: 43%

Embeddings collapse without positive-pair supervision

We need some **minimal** **“data augmentation”**

SimCSE: How to construct positive pairs

Data augmentation

He was running away

He was running ~~away~~

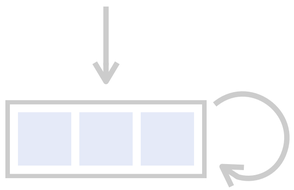
(Semantic text similarity)

STS score: 63%

Even deleting one word could change the meaning of a sentence

Identical

He was running away



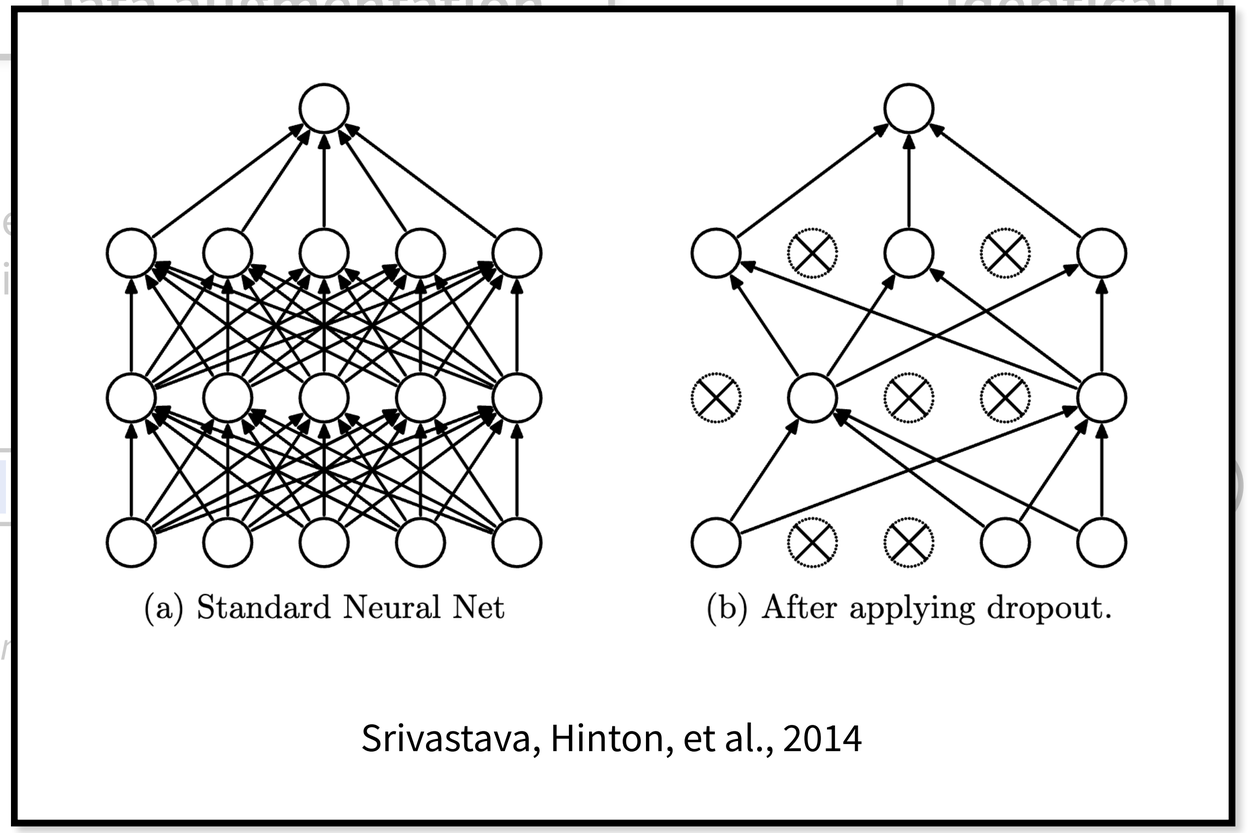
STS score: 43%

Embeddings collapse without positive-pair supervision

Dropout (SimCSE)

SimCSE: How to construct positive pairs

Dropout (SimCSE)



Srivastava, Hinton, et al., 2014

SimCSE: How to construct positive pairs

Data augmentation

He was running away

He was running ~~away~~

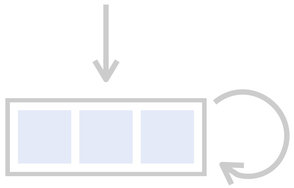
(Semantic text similarity)

STS score: 63%

Even deleting one word could change the meaning of a sentence

Identical

He was running away



STS score: 43%

Embeddings collapse without positive-pair supervision

Dropout (SimCSE)

He was running away

SimCSE: How to construct positive pairs

Data augmentation

He was running away

He was running ~~away~~

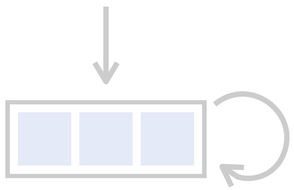
(Semantic text similarity)

STS score: 63%

Even deleting one word could change the meaning of a sentence

Identical

He was running away



STS score: 43%

Embeddings collapse without positive-pair supervision

Dropout (SimCSE)

He was running away

Encode twice

*Slightly different embeddings due to **dropout***

SimCSE: How to construct positive pairs

Data augmentation

He was running away

He was running ~~away~~

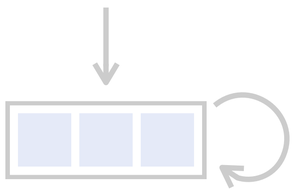
(Semantic text similarity)

STS score: 63%

Even deleting one word could change the meaning of a sentence

Identical

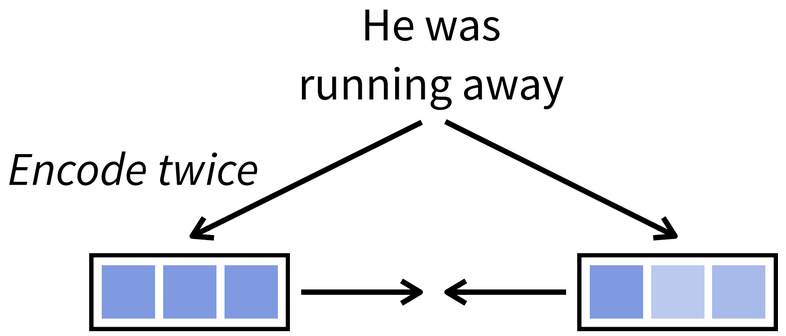
He was running away



STS score: 43%

Embeddings collapse without positive-pair supervision

Dropout (SimCSE)



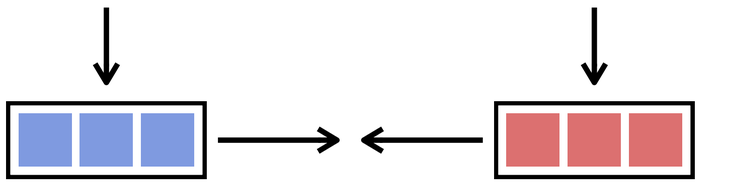
*Slightly different embeddings due to **dropout***

SimCSE: How to construct positive pairs

Data augmentation

He was running away

He was running ~~away~~



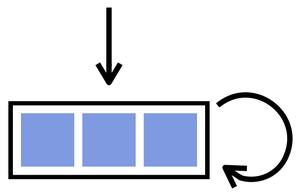
(Semantic text similarity)

STS score: 63%

Even deleting one word could change the meaning of a sentence

Identical

He was running away



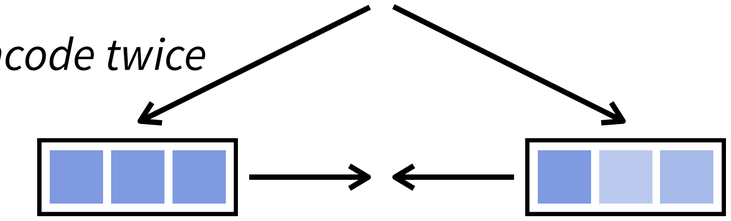
STS score: 43%

Embeddings collapse without positive-pair supervision

Dropout (SimCSE)

He was running away

Encode twice



*Slightly different embeddings due to **dropout***

STS score: **83%**

Dropout serves as minimal "data augmentation"

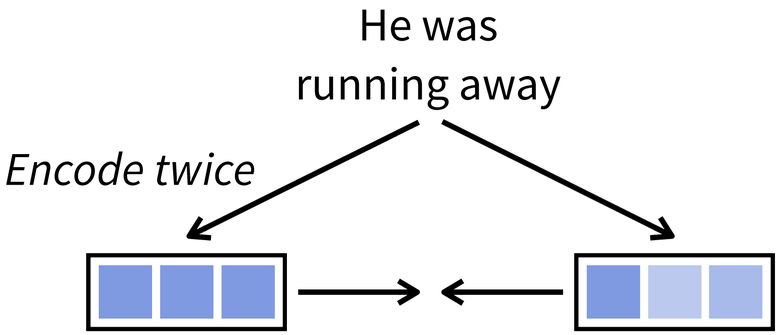
SimCSE: How to construct positive pairs

Data augmentation

Identical

Dropout (SimCSE)

- SimCSE (unsupervised) outperforms **previous SoTA (which relies on annotated data)**
- SimCSE + human-annotated pairs: even better!



*Slightly different embeddings due to **dropout***

(Semantic text similarity)

STS score: 63%

STS score: 43%

STS score: **83%**

Even deleting one word could change the meaning of a sentence

Embeddings collapse without positive-pair supervision

Dropout serves as minimal "data augmentation"

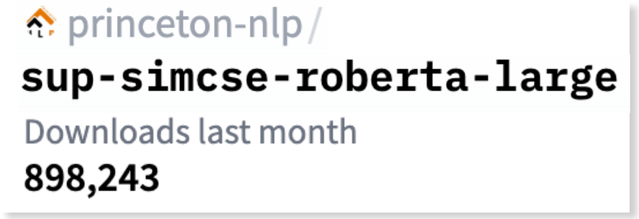
SimCSE: Impact

SimCSE: Impact

Popular artifacts

Downloads ⏬

> 22M



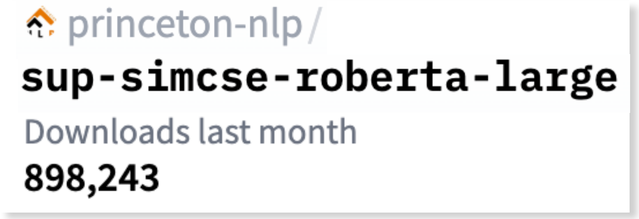
SimCSE: Impact

Popular artifacts

Numerous followups

Downloads ⏬

> 22M



Citations 🔍

> 3,500

Extended to

- Multimodal and multilingual use
- Graph and recommendation
- Code
- Biology
- Finance

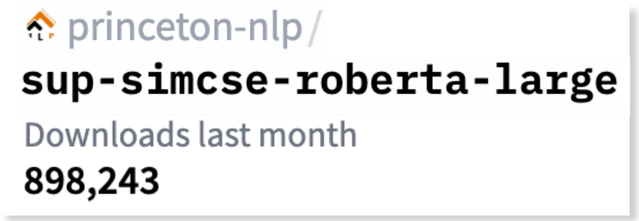
Zhang et al., 2022; Wang et al., 2022; Yu et al., 2021; Guo et al., 2022; Kanakarajan et al., 2022; Liu et al., 2024

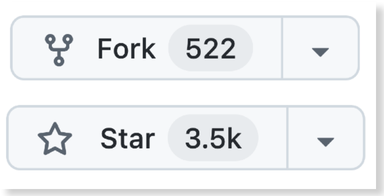
SimCSE: Impact

Popular artifacts

Downloads ⏬

> 22M





Numerous followups

Citations 🔍

> 3,500

Extended to

- Multimodal and multilingual use
- Graph and recommendation
- Code
- Biology
- Finance

Zhang et al., 2022; Wang et al., 2022; Yu et al., 2021; Guo et al., 2022; Kanakarajan et al., 2022; Liu et al., 2024

Foundations for modern embeddings

+ Better LMs and better data:

Izacard et al., 2021; Neelakantan et al., 2022; Wang et al., 2022; J. Lee et al., 2024; C. Lee et al., 2024

SimCSE: Impact

Popular artifacts

Downloads ⏬
> 22M

princeton-nlp /
sup-simcse-roberta-large
Downloads last month
898,243

Fork 522
Star 3.5k

Numerous followups

Citations 🔍
> 3,500

Extended to

- Multimodal and multilingual use
- Graph and recommendation
- Code
- Biology
- Finance

Zhang et al., 2022; Wang et al., 2022; Yu et al., 2021; Guo et al., 2022; Kanakarajan et al., 2022; Liu et al., 2024

Foundations for modern embeddings

+ Better LMs and better data:

Izacard et al., 2021; Neelakantan et al., 2022; Wang et al., 2022; J. Lee et al., 2024; C. Lee et al., 2024

Enabling new RAG businesses

Powerful embeddings enable new applications: Literature search

Powerful embeddings enable new applications: Literature search

LitSearch: first natural-language literature search benchmark

A. Ajith et al. and **T. Gao**. EMNLP 2024. **LitSearch: A Retrieval Benchmark for Scientific Literature Search.**

Powerful embeddings enable new applications: Literature search

LitSearch: first natural-language literature search benchmark

Literature search question

Can you find a research paper that uses structured pruning techniques to scale down billion-parameter LMs?

A. Ajith et al. and **T. Gao**. EMNLP 2024. **LitSearch: A Retrieval Benchmark for Scientific Literature Search.**

Powerful embeddings enable new applications: Literature search

LitSearch: first natural-language literature search benchmark

Literature search question

Can you find a research paper that uses structured pruning techniques to scale down billion-parameter LMs?

Semantic Scholar database

A. Ajith et al. and **T. Gao**. EMNLP 2024. **LitSearch: A Retrieval Benchmark for Scientific Literature Search.**

Powerful embeddings enable new applications: Literature search

LitSearch: first natural-language literature search benchmark

Literature search question

Can you find a research paper that uses structured pruning techniques to scale down billion-parameter LMs?

Semantic Scholar database

Target paper

Title: Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning
Author: M. Xia, T. Gao, Z. Zeng, D. Chen
Abstract: … structured pruning …
Full text: … prune a 7B-parameter model …

A. Ajith et al. and **T. Gao**. EMNLP 2024. **LitSearch: A Retrieval Benchmark for Scientific Literature Search.**

Powerful embeddings enable new applications: Literature search

LitSearch: first natural-language literature search benchmark

Literature search question

Can you find a research paper that uses structured pruning techniques to scale down billion-parameter LMs?

Semantic Scholar database

Target paper

Title: Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning
Author: M. Xia, T. Gao, Z. Zeng, D. Chen
Abstract: … structured pruning …
Full text: … prune a 7B-parameter model …

Keyword-based

SimCSE
+ better LM
+ better data

A. Ajith et al. and **T. Gao**. EMNLP 2024. **LitSearch: A Retrieval Benchmark for Scientific Literature Search.**

Powerful embeddings enable new applications: Literature search

LitSearch: first natural-language literature search benchmark

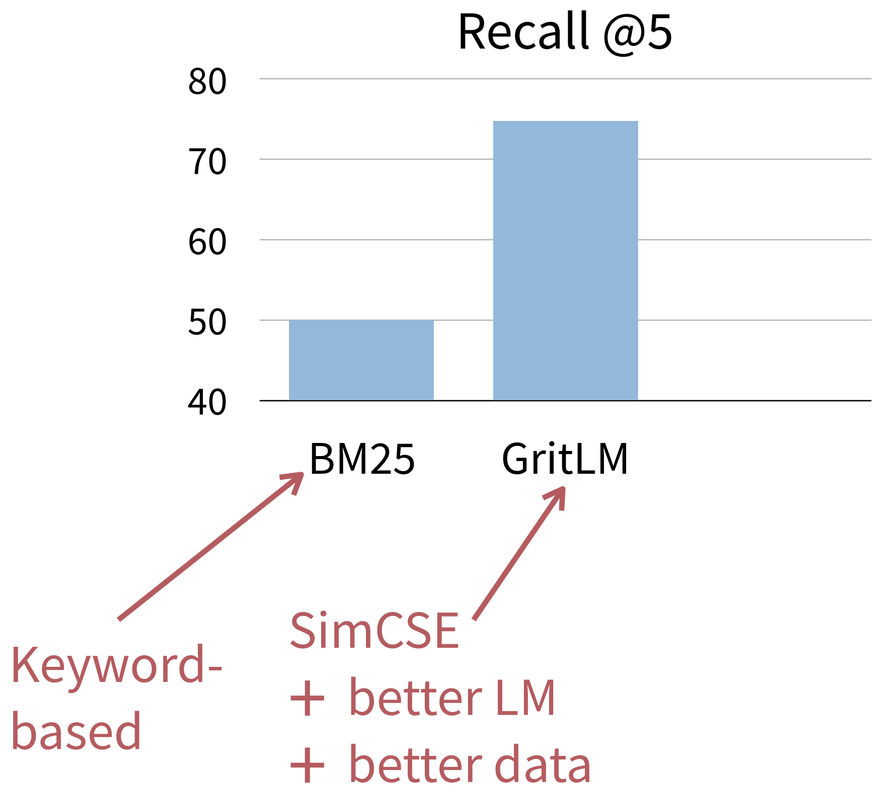
Literature search question

Can you find a research paper that uses structured pruning techniques to scale down billion-parameter LMs?

Semantic Scholar database

Target paper

Title: Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning
Author: M. Xia, T. Gao, Z. Zeng, D. Chen
Abstract: … structured pruning …
Full text: … prune a 7B-parameter model …



Powerful embeddings enable new applications: Literature search

LitSearch: first natural-language literature search benchmark

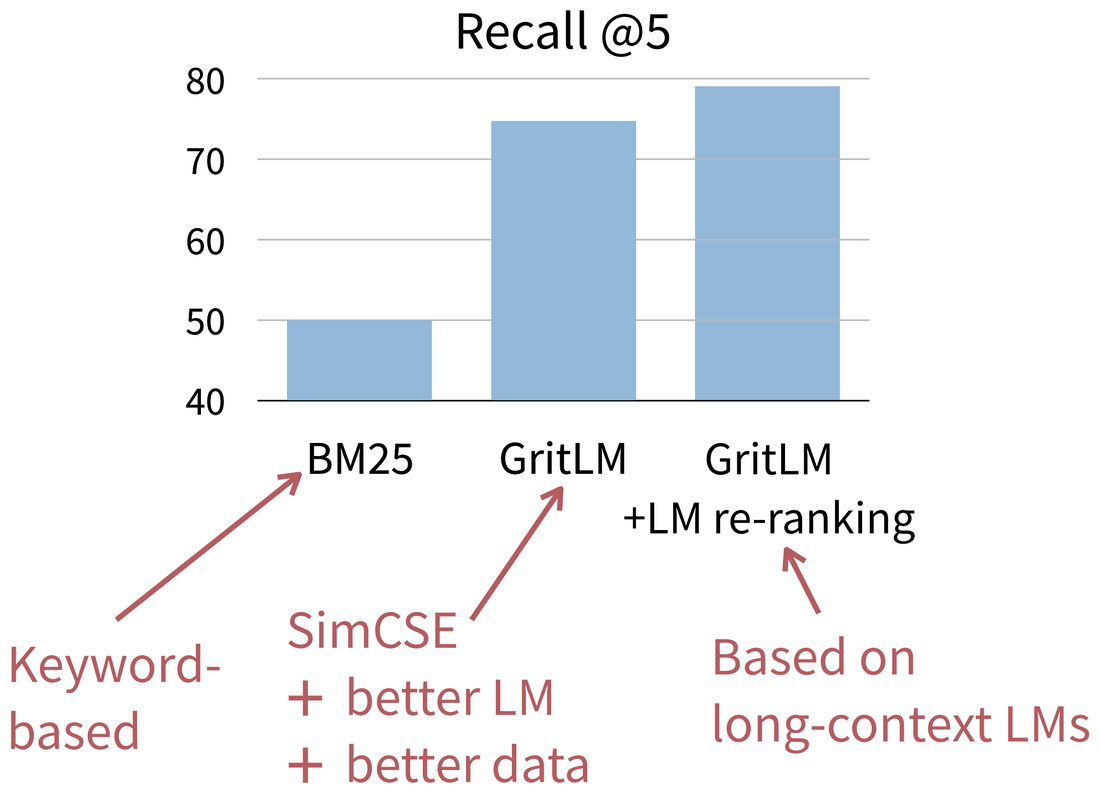
Literature search question

Can you find a research paper that uses structured pruning techniques to scale down billion-parameter LMs?

Semantic Scholar database

Target paper

Title: Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning
Author: M. Xia, T. Gao, Z. Zeng, D. Chen
Abstract: … structured pruning …
Full text: … prune a 7B-parameter model …



A. Ajith et al. and **T. Gao**. EMNLP 2024. **LitSearch: A Retrieval Benchmark for Scientific Literature Search.**

Summary: Accurate search via text embeddings

- **SimCSE**: pre-trained LM → text embeddings
  - Lasting impact: foundation of modern text embeddings
- **LitSearch**: a novel application and a retrieval benchmark

GYYC EMNLP23; P**G**CC Findings of ACL23; Y**G**+ ICLR25; Y**G**C ACL24, **G***W*YC 2024

Enabling LMs to process information at scale

1. Effective long-context processing

2. Accurate search via text embeddings

3. Foundations: Efficient language models

Broader applications demand more affordable LMs

Broader applications demand more affordable LMs

Scaling applications requires compact yet capable LMs

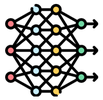
Deploying DeepSeek V3/R1 requires at least 32 x H100 GPUs*

*The estimation depends on many factors such as model architectures, quantization, optimizers, and sequence lengths

Broader applications demand more affordable LMs

Scaling applications requires compact yet capable LMs

Deploying DeepSeek V3/R1 requires at least 32 x H100 GPUs*



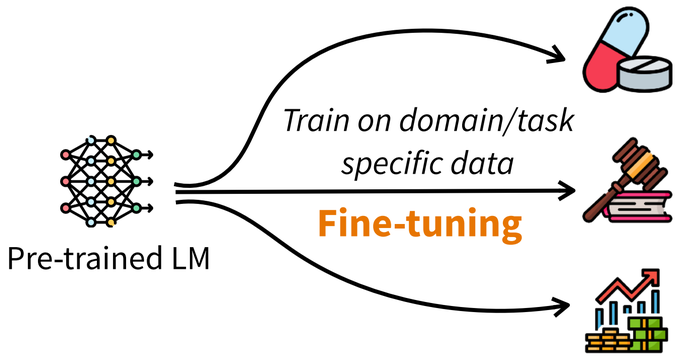
Pre-trained LM

*The estimation depends on many factors such as model architectures, quantization, optimizers, and sequence lengths

Broader applications demand more affordable LMs

Scaling applications requires compact yet capable LMs

Deploying DeepSeek V3/R1 requires at least 32 x H100 GPUs*



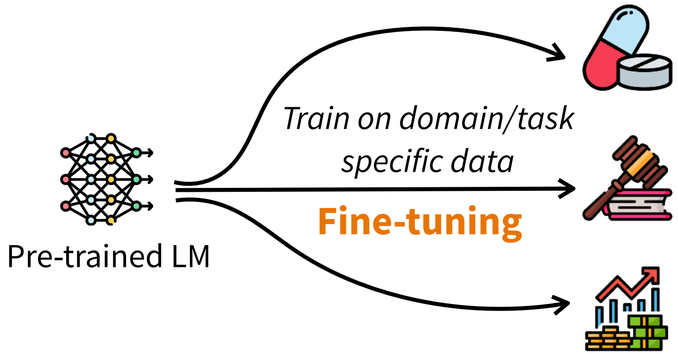
*The estimation depends on many factors such as model architectures, quantization, optimizers, and sequence lengths

Broader applications demand more affordable LMs

Scaling applications requires compact yet capable LMs

Deploying DeepSeek V3/R1 requires at least 32 x H100 GPUs*

Demand for LM customization is high, but so is the cost



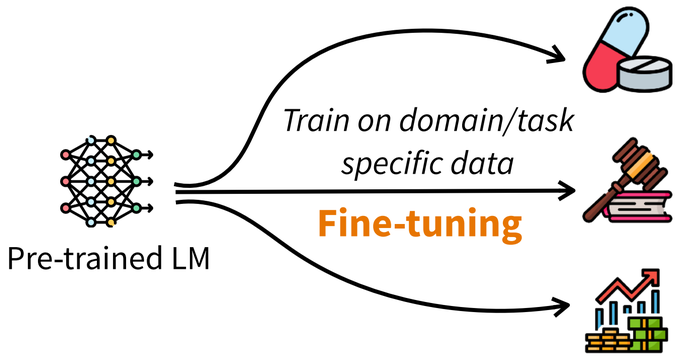
*The estimation depends on many factors such as model architectures, quantization, optimizers, and sequence lengths

Broader applications demand more affordable LMs

Scaling applications requires compact yet capable LMs

Deploying DeepSeek V3/R1 requires at least 32 x H100 GPUs*

Demand for LM customization is high, but so is the cost



- Often needs **thousands of** human-annotated data

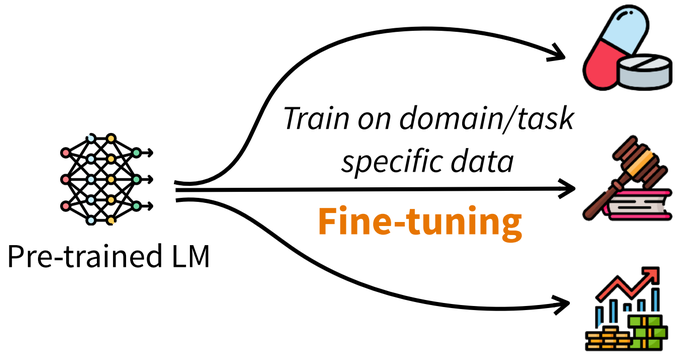
*The estimation depends on many factors such as model architectures, quantization, optimizers, and sequence lengths

Broader applications demand more affordable LMs

Scaling applications requires compact yet capable LMs

Deploying DeepSeek V3/R1 requires at least 32 x H100 GPUs*

Demand for LM customization is high, but so is the cost



- Often needs **thousands of** human-annotated data
- Requires **significant resources** to fine-tune

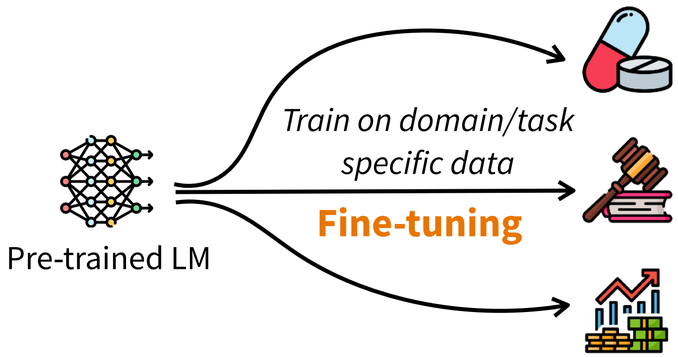
*The estimation depends on many factors such as model architectures, quantization, optimizers, and sequence lengths

Broader applications demand more affordable LMs

Scaling applications requires compact yet capable LMs

Deploying DeepSeek V3/R1 requires at least 32 x H100 GPUs*

Demand for LM customization is high, but so is the cost



- Often needs **thousands of** human-annotated data
- Requires **significant resources** to fine-tune

1 x 3B parameter*
4 x 8B parameter
8 x 13B parameter
16 x 30B parameter

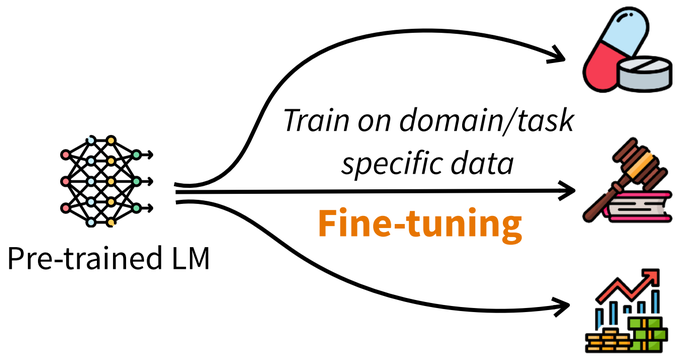
*The estimation depends on many factors such as model architectures, quantization, optimizers, and sequence lengths

Broader applications demand more affordable LMs

Scaling applications requires compact yet capable LMs

Deploying DeepSeek V3/R1 requires at least 32 x H100 GPUs*

Demand for LM customization is high, but so is the cost



- Often needs **thousands of** human-annotated data
- Requires **significant resources** to fine-tune

1 x 3B parameter*
4 x 8B parameter
8 x 13B parameter
16 x 30B parameter

We need cheaper models and efficient training methods for open science of language models

*The estimation depends on many factors such as model architectures, quantization, optimizers, and sequence lengths

My work: Cheaply producing small yet capable LMs

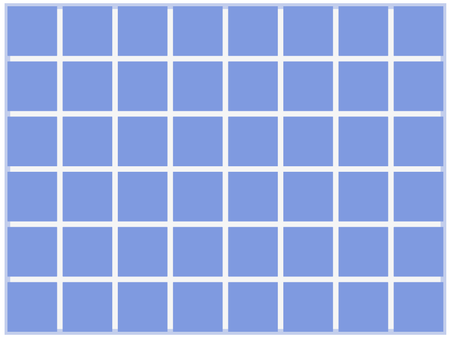
My work: Cheaply producing small yet capable LMs

Utilizing existing large models by Sheared Llama

M. Xia, **T. Gao**, Z. Zeng, D. Chen. ICLR 2024. **Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning**.

My work: Cheaply producing small yet capable LMs

Utilizing existing large models by Sheared Llama

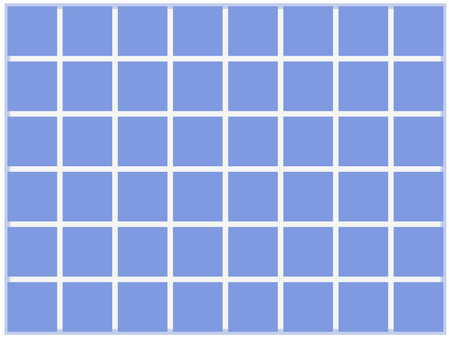


An existing large LM

M. Xia, **T. Gao**, Z. Zeng, D. Chen. ICLR 2024. **Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning**.

My work: Cheaply producing small yet capable LMs

Utilizing existing large models by Sheared Llama



An existing large LM

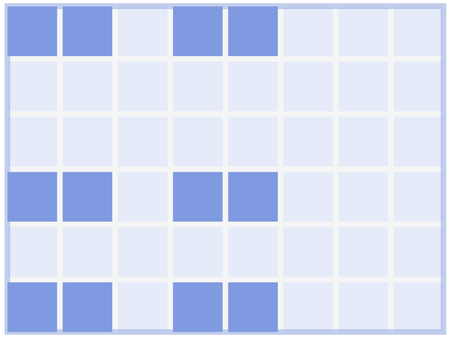
To reduce parameters

Structured pruning

M. Xia, **T. Gao**, Z. Zeng, D. Chen. ICLR 2024. **Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning**.

My work: Cheaply producing small yet capable LMs

Utilizing existing large models by Sheared Llama



An existing large LM

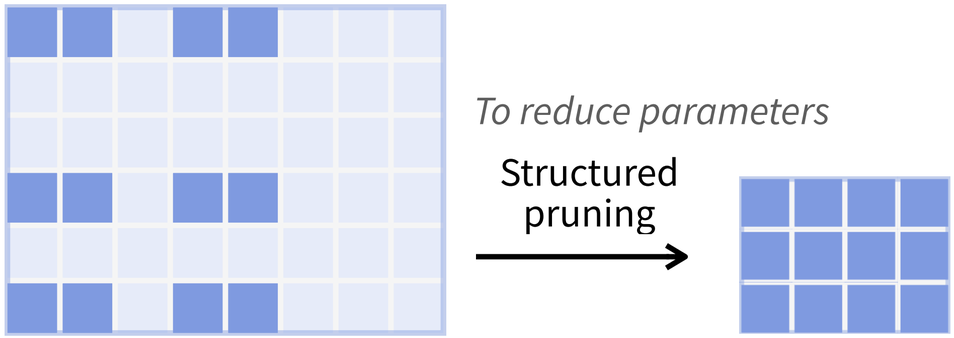
To reduce parameters

Structured pruning →

M. Xia, **T. Gao**, Z. Zeng, D. Chen. ICLR 2024. **Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning**.

My work: Cheaply producing small yet capable LMs

Utilizing existing large models by Sheared Llama



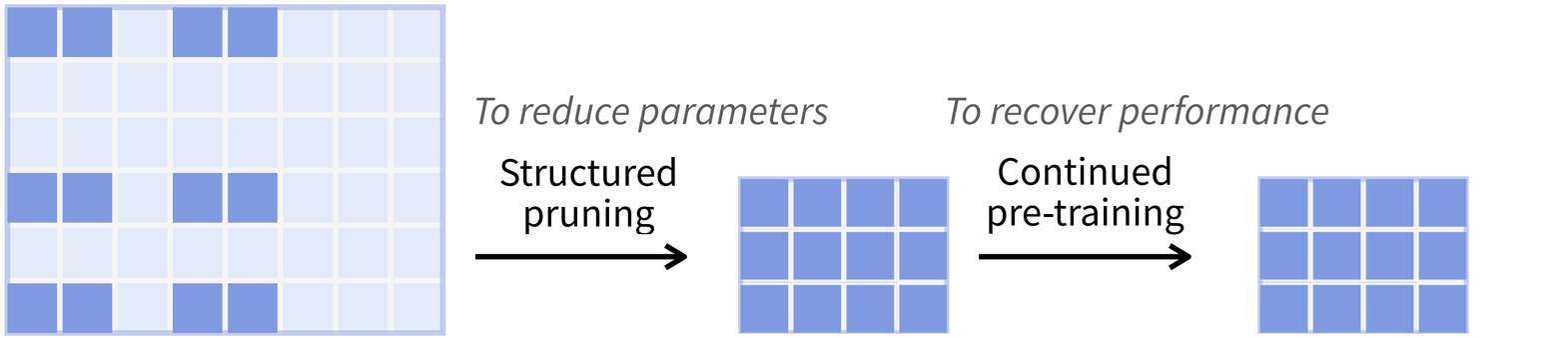
An existing large LM

A small LM

M. Xia, **T. Gao**, Z. Zeng, D. Chen. ICLR 2024. **Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning**.

My work: Cheaply producing small yet capable LMs

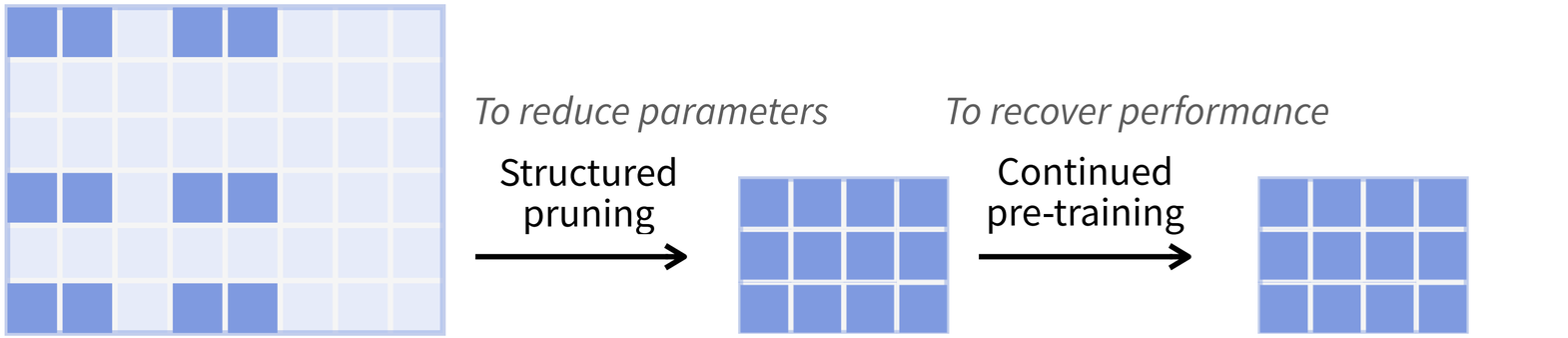
Utilizing existing large models by Sheared Llama



M. Xia, **T. Gao**, Z. Zeng, D. Chen. ICLR 2024. **Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning**.

My work: Cheaply producing small yet capable LMs

Utilizing existing large models by Sheared Llama

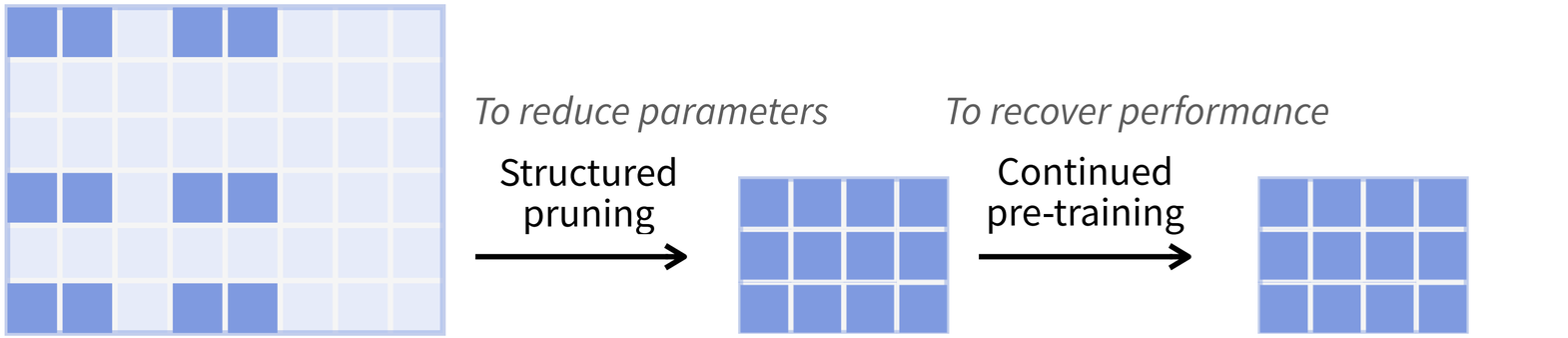


State-of-the-art 1B/3B parameter models at the time (2023)

M. Xia, **T. Gao**, Z. Zeng, D. Chen. ICLR 2024. **Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning**.

My work: Cheaply producing small yet capable LMs

Utilizing existing large models by Sheared Llama



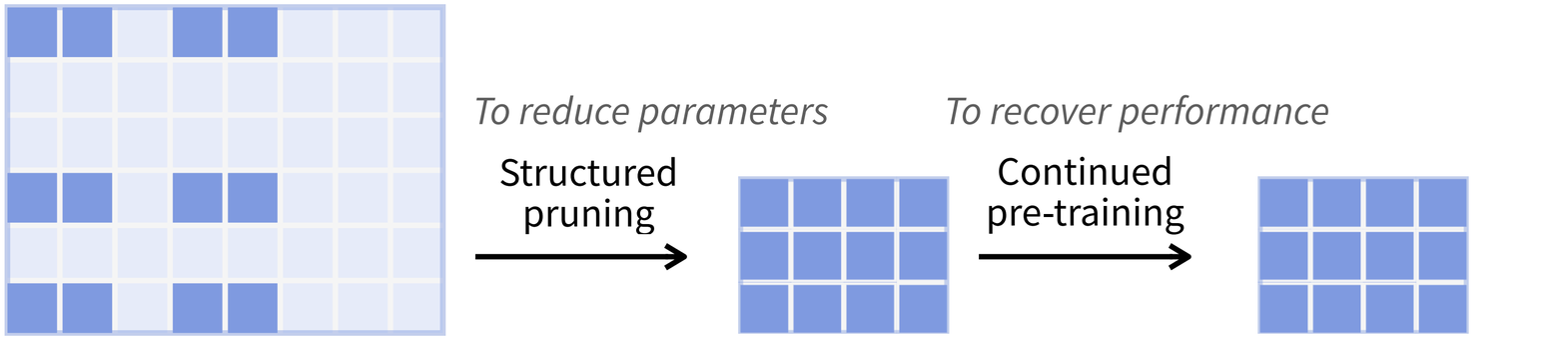
State-of-the-art 1B/3B parameter models at the time (2023)

… with **<3%** of the cost to train them from scratch

M. Xia, **T. Gao**, Z. Zeng, D. Chen. ICLR 2024. **Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning**.

My work: Cheaply producing small yet capable LMs

Utilizing existing large models by Sheared Llama



Small enough to use 1 GPU to fine-tune / to fit on an iPhone!

State-of-the-art 1B/3B parameter models at the time (2023)

… with **<3%** of the cost to train them from scratch

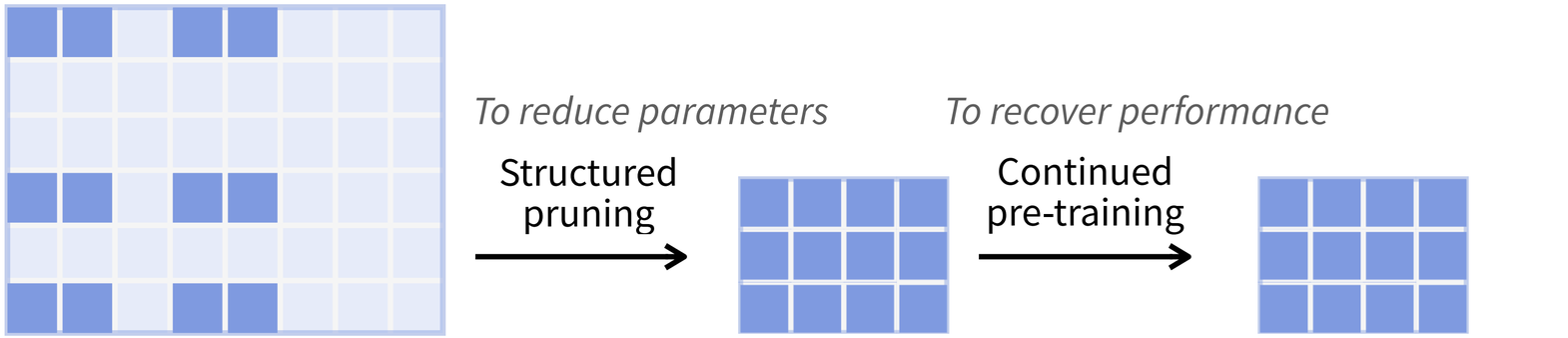
Downloads ⏬

> 800,000

M. Xia, **T. Gao**, Z. Zeng, D. Chen. ICLR 2024. **Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning**.

My work: Cheaply producing small yet capable LMs

Utilizing existing large models by Sheared Llama



Small enough to use 1 GPU to fine-tune / to fit on an iPhone!

State-of-the-art 1B/3B parameter models at the time (2023)

… with **<3%** of the cost to train them from scratch

Downloads ⬇️

> 800,000

M. Xia, **T. Gao**, Z. Zeng, D. Chen. ICLR 2024. **Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning**.

My work: Fine-tuning with a dozen examples and fewer GPUs

My work: Fine-tuning with a dozen examples and fewer GPUs

Using prompts allows fine-tuning with only 32 examples

2 most influential paper

T. Gao*, A. Fisch*, D. Chen. ACL 2021. **Making Pre-trained Language Models Better Few-shot Learners (LM-BFF)**

My work: Fine-tuning with a dozen examples and fewer GPUs

Using prompts allows fine-tuning with only 32 examples

To fine-tune a pre-trained LM for *sentiment classification* and reach >93% accuracy

T. Gao*, A. Fisch*, D. Chen. ACL 2021. **Making Pre-trained Language Models Better Few-shot Learners (LM-BFF)**

My work: Fine-tuning with a dozen examples and fewer GPUs

Using prompts allows fine-tuning with only 32 examples

To fine-tune a pre-trained LM for *sentiment classification* and reach >93% accuracy

Standard fine-tuning: more than 1,000 examples

T. Gao*, A. Fisch*, D. Chen. ACL 2021. **Making Pre-trained Language Models Better Few-shot Learners (LM-BFF)**

My work: Fine-tuning with a dozen examples and fewer GPUs

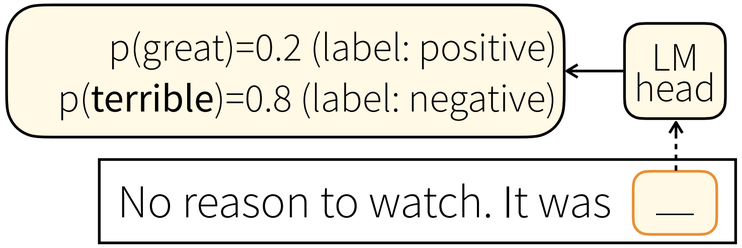
Using prompts allows fine-tuning with only 32 examples

To fine-tune a pre-trained LM for *sentiment classification* and reach >93% accuracy

Standard fine-tuning: more than 1,000 examples

Carefully-engineered prompts (LM-BFF, ours): **32** examples

(Turn any task into a next work prediction format)



2 most influential paper

T. Gao*, A. Fisch*, D. Chen. ACL 2021. **Making Pre-trained Language Models Better Few-shot Learners (LM-BFF)**

My work: Fine-tuning with a dozen examples and fewer GPUs

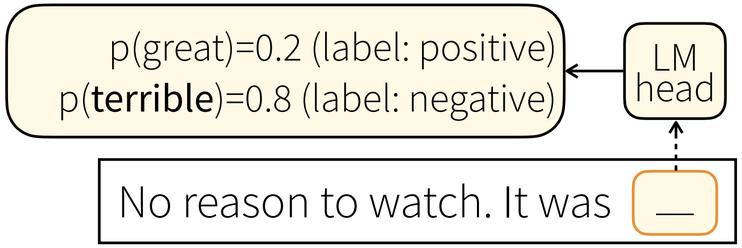
Using prompts allows fine-tuning with only 32 examples

To fine-tune a pre-trained LM for *sentiment classification* and reach >93% accuracy

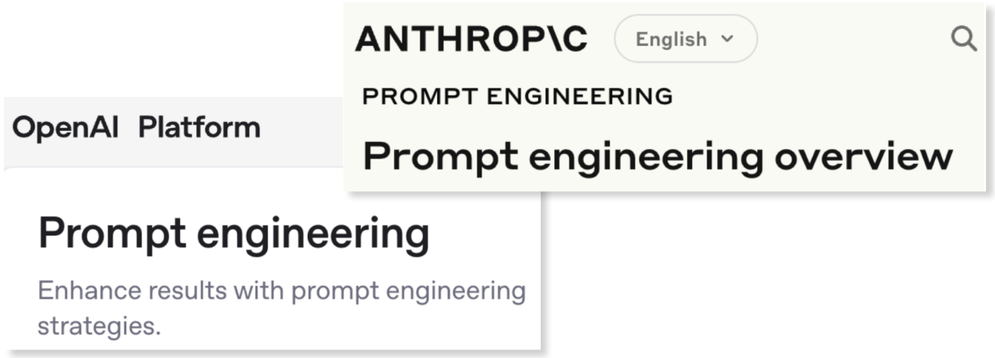
Standard fine-tuning: more than 1,000 examples

Carefully-engineered prompts (LM-BFF, ours): **32** examples

(Turn any task into a next work prediction format)



One of the earliest **prompt engineering** papers



2 most influential paper

T. Gao*, A. Fisch*, D. Chen. ACL 2021. **Making Pre-trained Language Models Better Few-shot Learners (LM-BFF)**

My work: Fine-tuning with a dozen examples and fewer GPUs

My work: Fine-tuning with a dozen examples and fewer GPUs

MeZO saves up to 12x memory for fine-tuning

🥇 Oral presentation (2%)

S. Malladi*, **T. Gao***, et al. NeurIPS 2023. **Fine-Tuning Language Models with Just Forward Passes (MeZO)**

My work: Fine-tuning with a dozen examples and fewer GPUs

MeZO saves up to 12x memory for fine-tuning

MeZO is a **zeroth-order optimizer** that **estimates gradients** by using **two forward passes**

🏅 Oral presentation (2%)

S. Malladi*, **T. Gao***, et al. NeurIPS 2023. **Fine-Tuning Language Models with Just Forward Passes (MeZO)**

My work: Fine-tuning with a dozen examples and fewer GPUs

MeZO saves up to 12x memory for fine-tuning

MeZO is a **zeroth-order optimizer** that **estimates gradients** by using **two forward passes**

→ allows **fine-tuning with the same memory cost as inference**

🏅 Oral presentation (2%)

S. Malladi*, **T. Gao***, et al. NeurIPS 2023. **Fine-Tuning Language Models with Just Forward Passes (MeZO)**

My work: Fine-tuning with a dozen examples and fewer GPUs

MeZO saves up to 12x memory for fine-tuning

MeZO is a **zeroth-order optimizer** that **estimates gradients** by using **two forward passes**

→ allows **fine-tuning with the same memory cost as inference**

	Standard	MeZO
1 x 	3B parameter	**30B** parameter
4 x 	8B parameter	**70B** parameter
8 x 	13B parameter	**175B** parameter

 Oral presentation (2%)

S. Malladi*, **T. Gao***, et al. NeurIPS 2023. **Fine-Tuning Language Models with Just Forward Passes (MeZO)**

My work: Fine-tuning with a dozen examples and fewer GPUs

MeZO saves up to 12x memory for fine-tuning

MeZO is a **zeroth-order optimizer** that **estimates gradients** by using **two forward passes**

→ allows **fine-tuning with the same memory cost as inference**

	Standard	MeZO
1 x 	3B parameter	**30B** parameter
4 x 	8B parameter	**70B** parameter
8 x 	13B parameter	**175B** parameter

If you can run inference, you can fine-tune with MeZO!

 Oral presentation (2%)

S. Malladi*, **T. Gao***, et al. NeurIPS 2023. **Fine-Tuning Language Models with Just Forward Passes (MeZO)**

My work: Fine-tuning with a dozen examples and fewer GPUs

MeZO saves up to 12x memory for fine-tuning

MeZO is a **zeroth-order optimizer** that **estimates gradients** by using **two forward passes**

→ allows **fine-tuning with the same memory cost as inference**

	Standard	MeZO
1 x 	3B parameter	**30B** parameter
4 x 	8B parameter	**70B** parameter
8 x 	13B parameter	**175B** parameter

If you can run inference, you can fine-tune with MeZO!

MeZO has been extended to **distributed training** (Zelikman et al., 2023), **pre-training** (Chen et al., 2024), and **reinforcement learning from human feedback** (Zhang et al., 2024)

🏅 Oral presentation (2%)

S. Malladi*, **T. Gao***, et al. NeurIPS 2023. **Fine-Tuning Language Models with Just Forward Passes (MeZO)**

Summary: Efficient Language Models

- **Improving data efficiency** [W*G*ZC EACL23; GW+ 2025]

G*F*C ACL21; M***G***+ NeurIPS23 oral; W***G***ZC EACL23; X**G**ZC ICLR24; **G**W+ 2025

Summary: Efficient Language Models

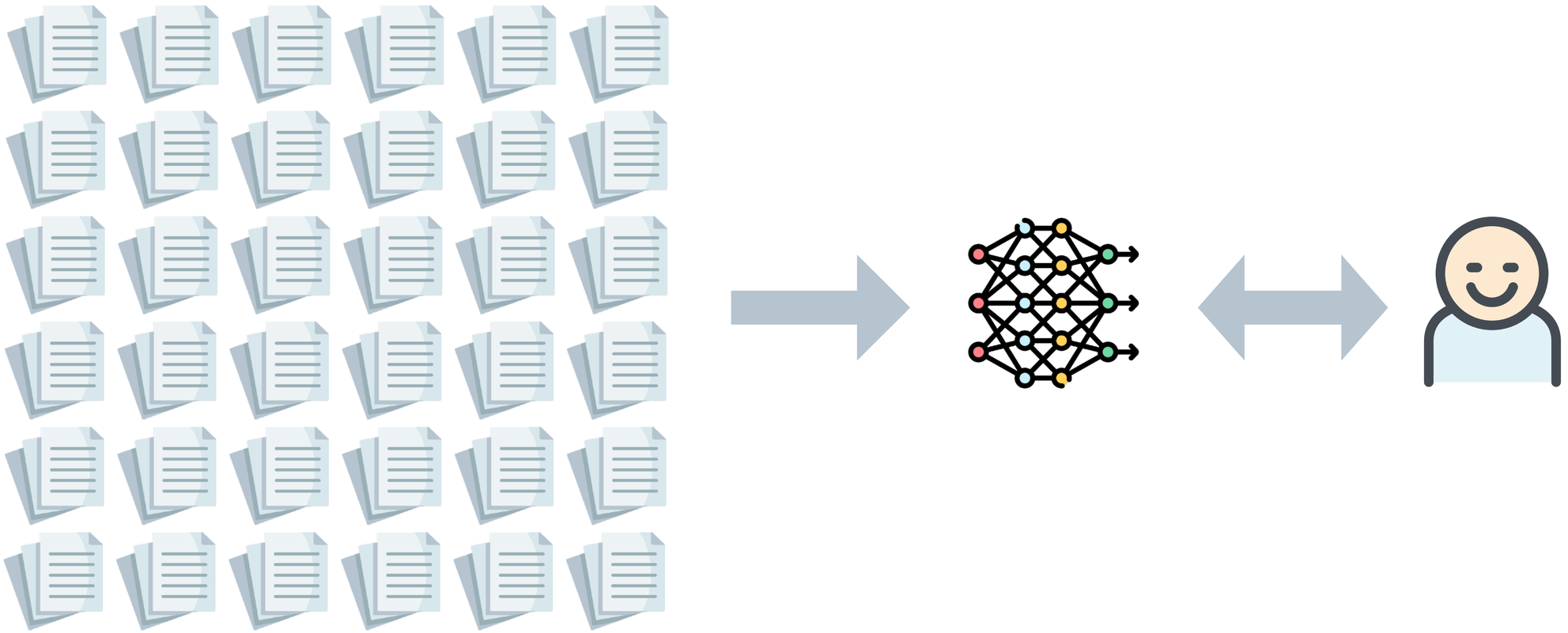
- **Efficient LM pre-training**
  - **Sheared Llama**: existing large LMs → capable small LMs
  - **Improving data efficiency** [W*G*ZC EACL23; GW+ 2025]
- **Efficient fine-tuning**
  - **LM-BFF**: prompt engineering reduces #examples needed
  - **MeZO**: memory-efficient fine-tuning

Democratize LMs

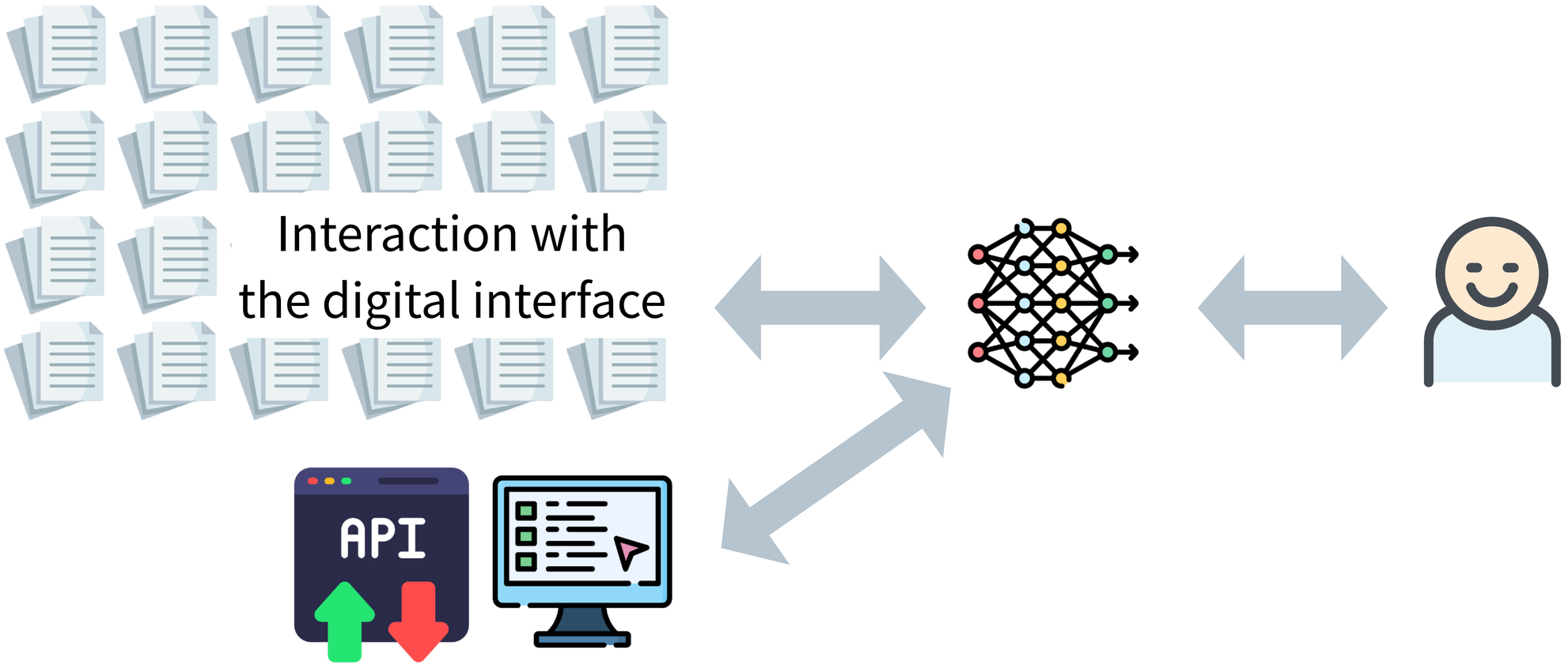
Enable on-device use

G*F*C ACL21; M*G*+ NeurIPS23 oral; W*G*ZC EACL23; XGZC ICLR24; GW+ 2025

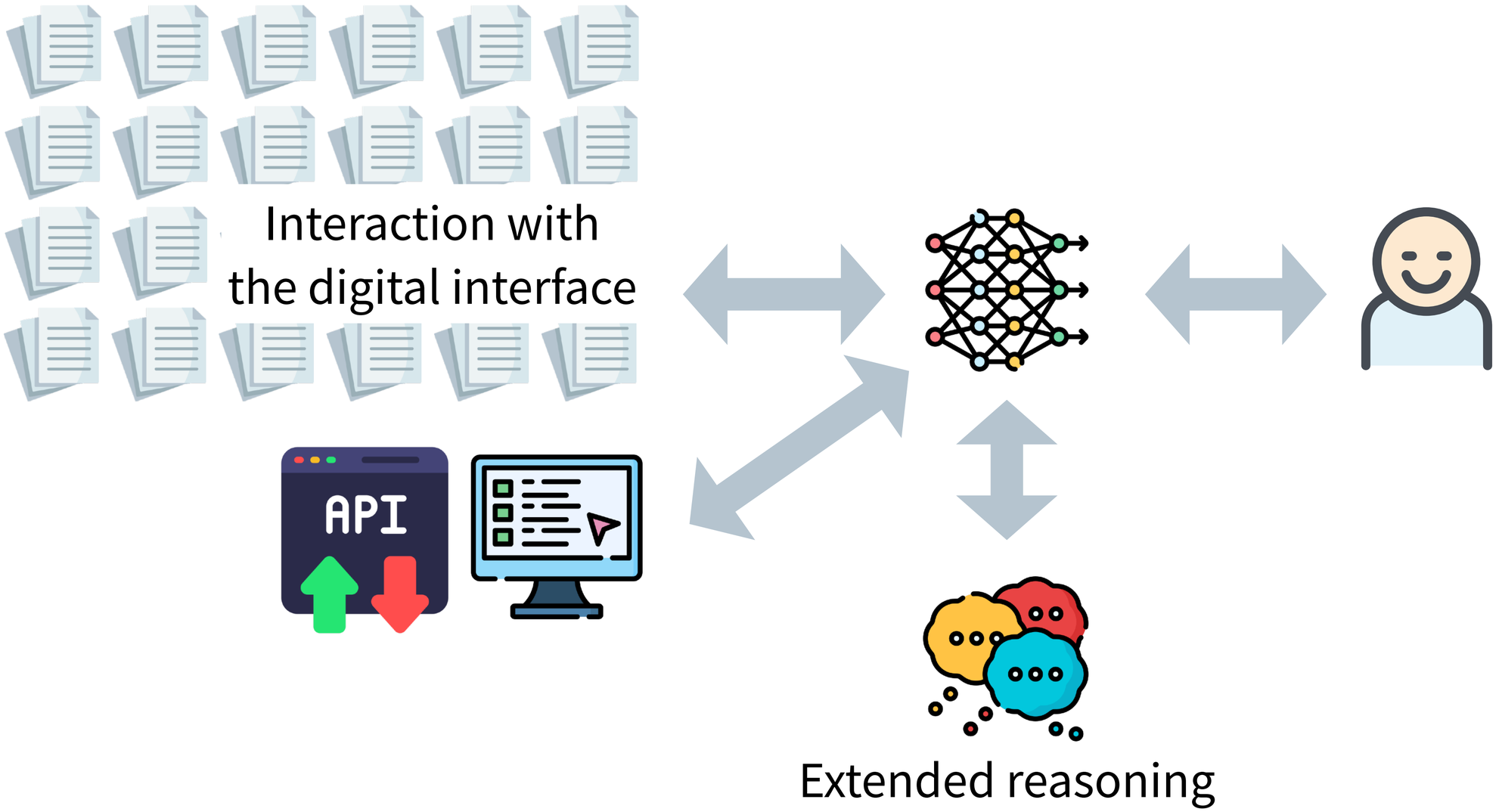
Enabling LMs to process information at scale



Future work: autonomous language models

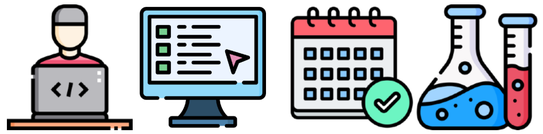


Future work: autonomous language models



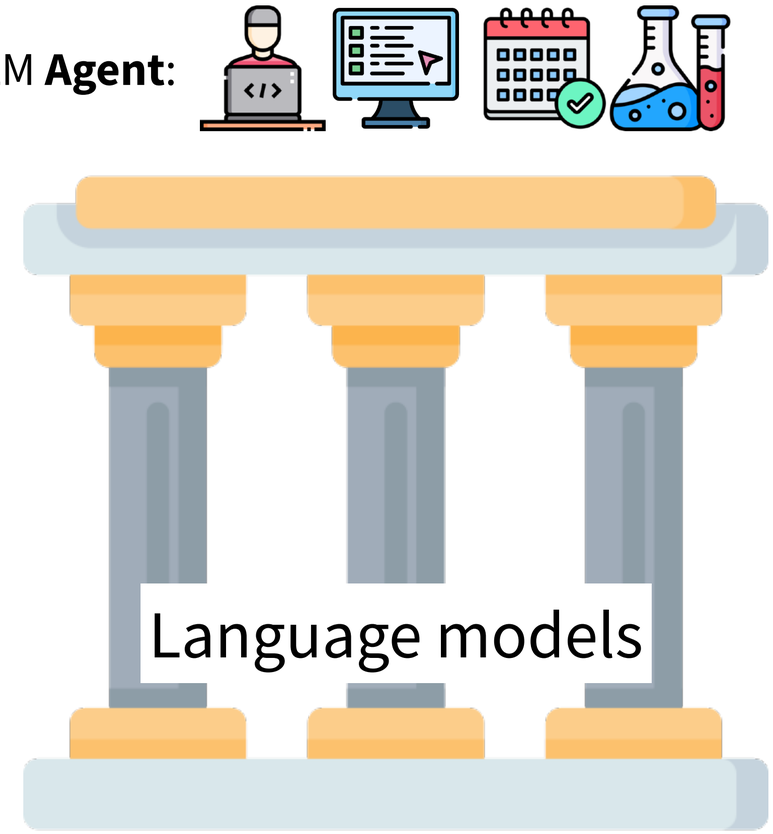
Future work: autonomous language models

Future work: autonomous language models

LM **Agent**: 

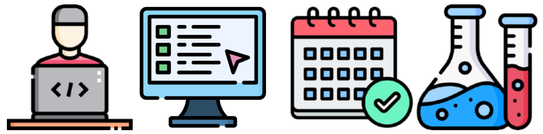
Future work: autonomous language models

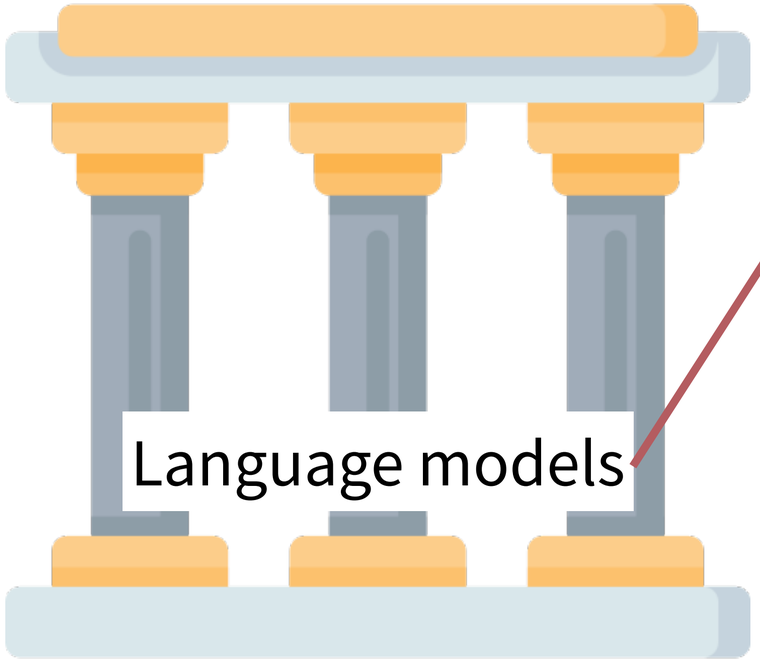
LM **Agent**:



Future work: autonomous language models

LM **Agent**:

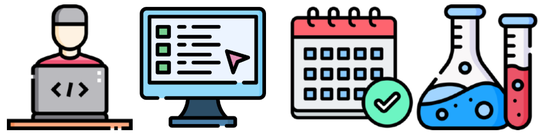


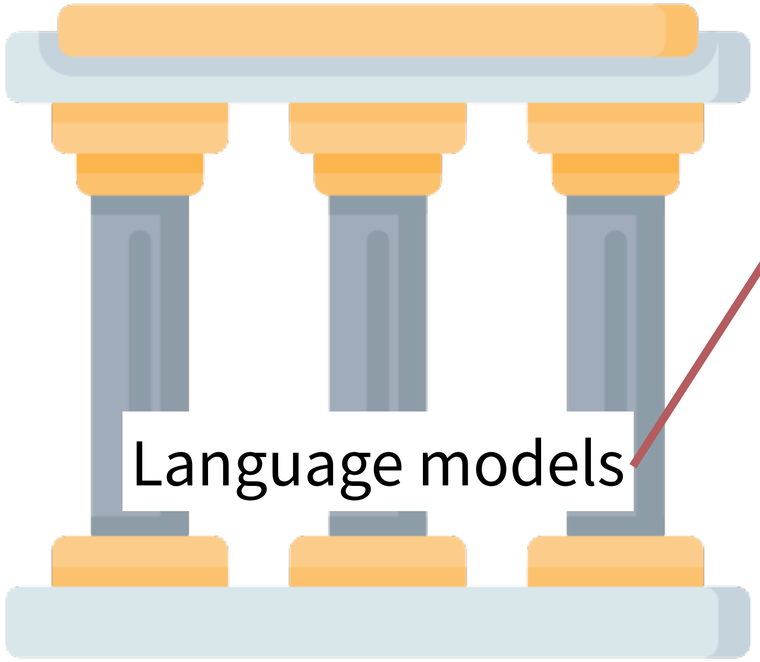


New capabilities required to support agent applications?

Future work: autonomous language models

LM **Agent**:





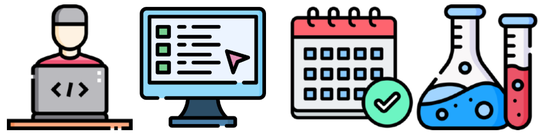
New capabilities required to support agent applications?

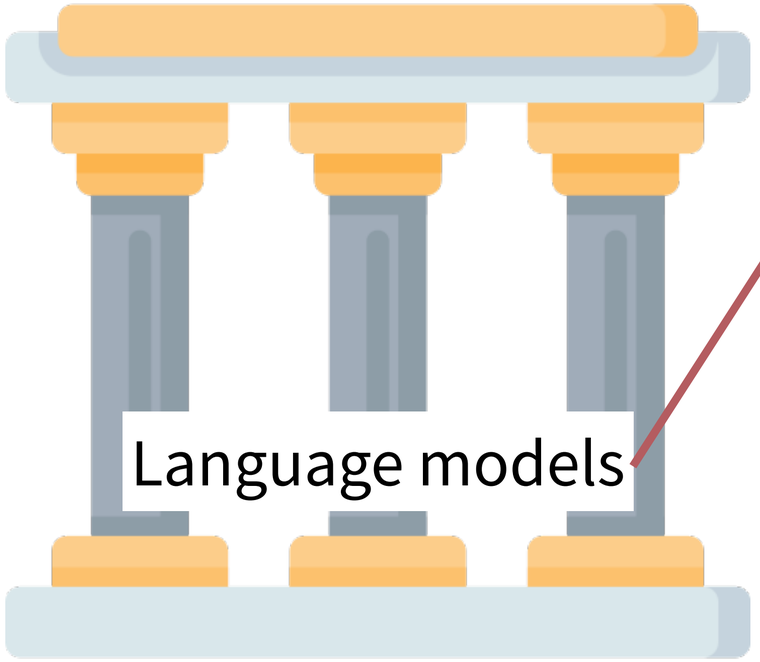
More **affordable**, **accessible** foundation models for agents?

Future work: autonomous language models

foundations for

LM **Agent**:





New capabilities required to support agent applications?

More **affordable**, **accessible** foundation models for agents?

Future work: autonomous language models

foundations for

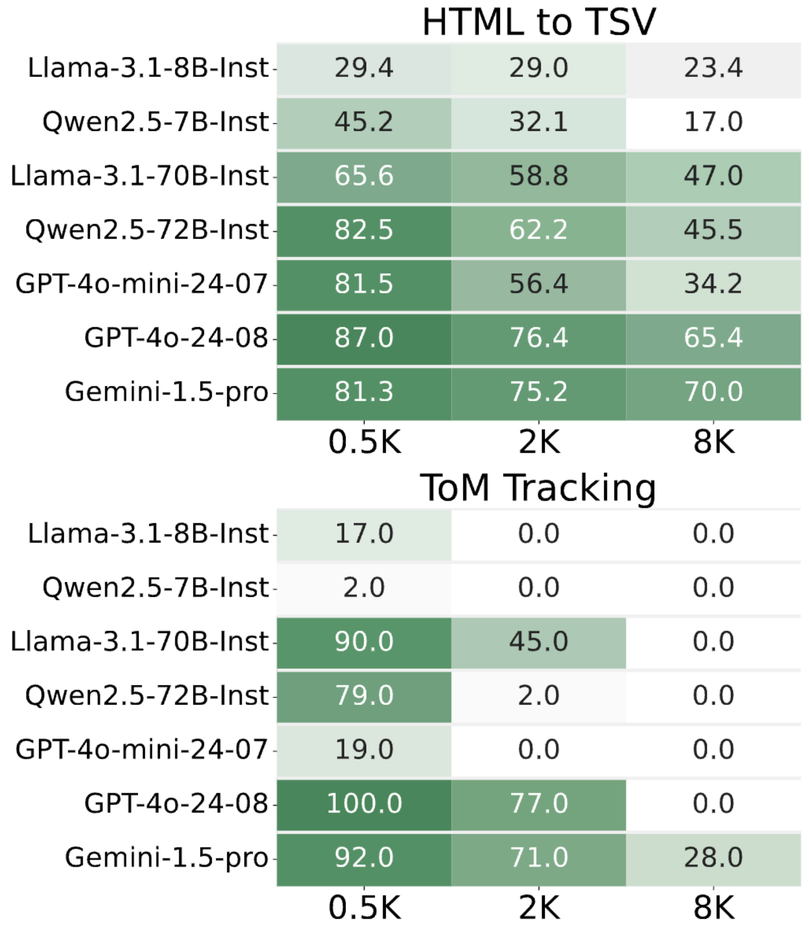
Consistent long-form generations

Future work: autonomous language models

foundations for

Consistent long-form generations

- Existing models struggle to generate long-form content [YYHZY**G**DC, 25]



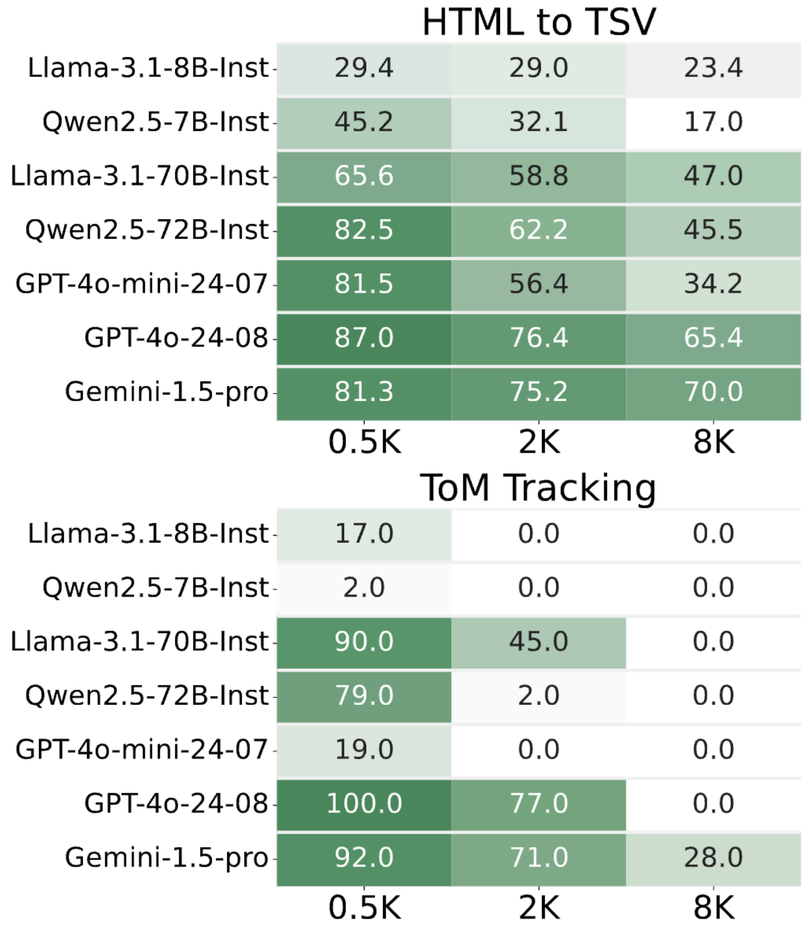
X. Ye, F. Yin, Y. He, J. Zhang, H. Yen, **T. Gao**, G. Durrett, D. Chen.
Preprint 2025. **LongProc: Benchmarking Long-Context Language Models on Long Procedural Generation**

Future work: autonomous language models

foundations for

Consistent long-form generations

- Existing models struggle to generate long-form content [YYHZY**G**DC, 25]
- Data and objective innovations



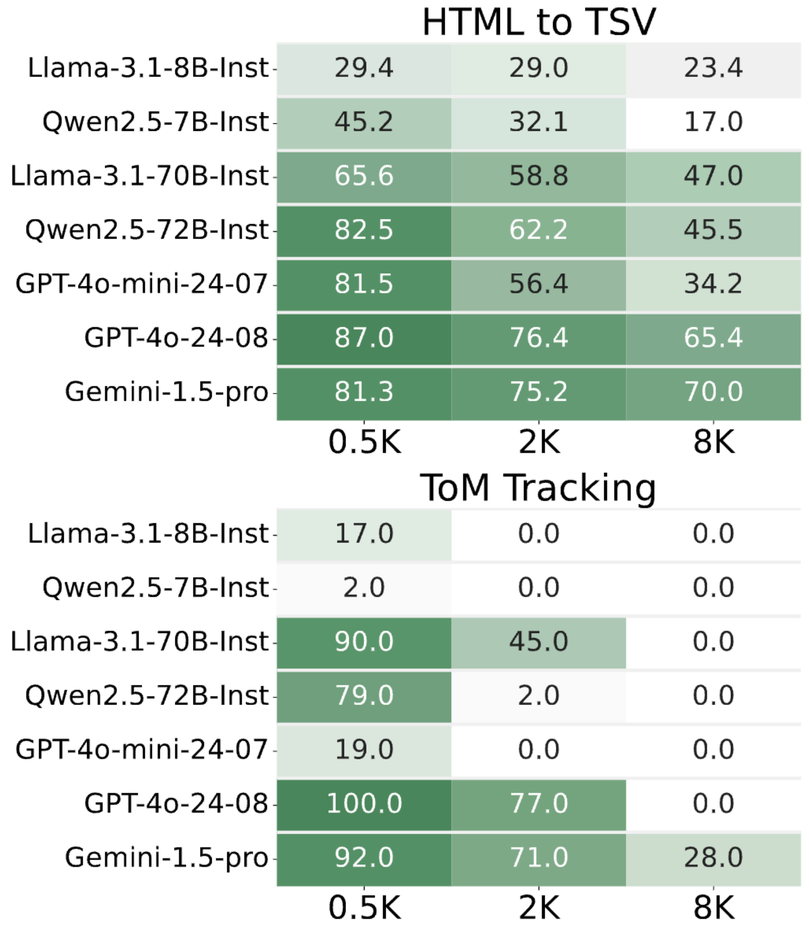
X. Ye, F. Yin, Y. He, J. Zhang, H. Yen, **T. Gao**, G. Durrett, D. Chen.
Preprint 2025. **LongProc: Benchmarking Long-Context Language Models on Long Procedural Generation**

Future work: autonomous language models

foundations for

Consistent long-form generations

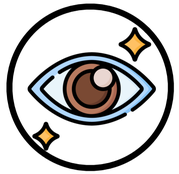
- Existing models struggle to generate long-form content [YYHZY**G**DC, 25]
- Data and objective innovations
- *Synthetic* data



X. Ye, F. Yin, Y. He, J. Zhang, H. Yen, **T. Gao**, G. Durrett, D. Chen.
Preprint 2025. **LongProc: Benchmarking Long-Context Language Models on Long Procedural Generation**

Future work: autonomous language models

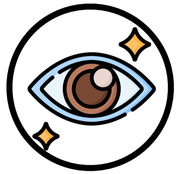
foundations for



Understanding multimodal context

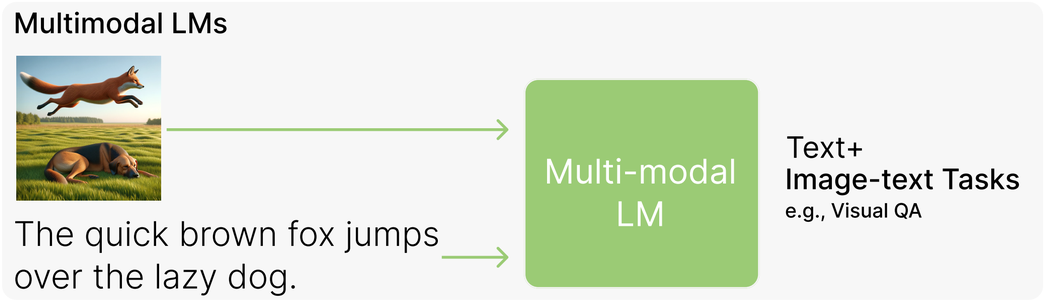
Future work: autonomous language models

foundations for



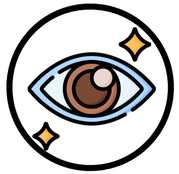
Understanding multimodal context

- Existing models lack the ability to understand complex visually-situated texts (UIs, document scans, diagrams)



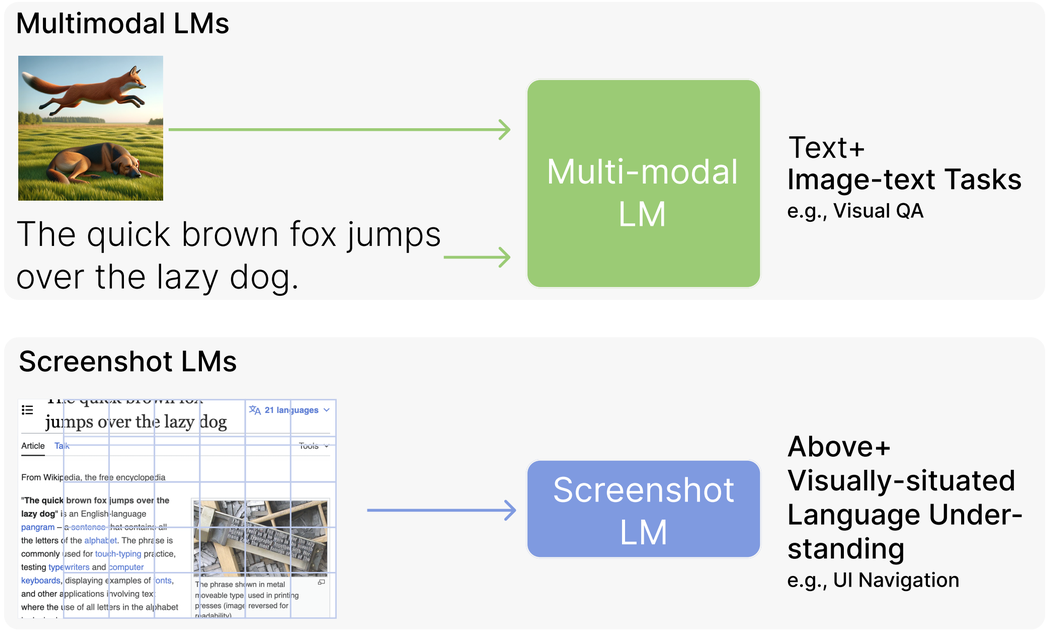
Future work: autonomous language models

foundations for



Understanding multimodal context

- Existing models lack the ability to understand complex visually-situated texts (UIs, document scans, diagrams)
- Unified multimodal interface via *screenshot* language models [**G**WBC 2024]



Future work: autonomous language models

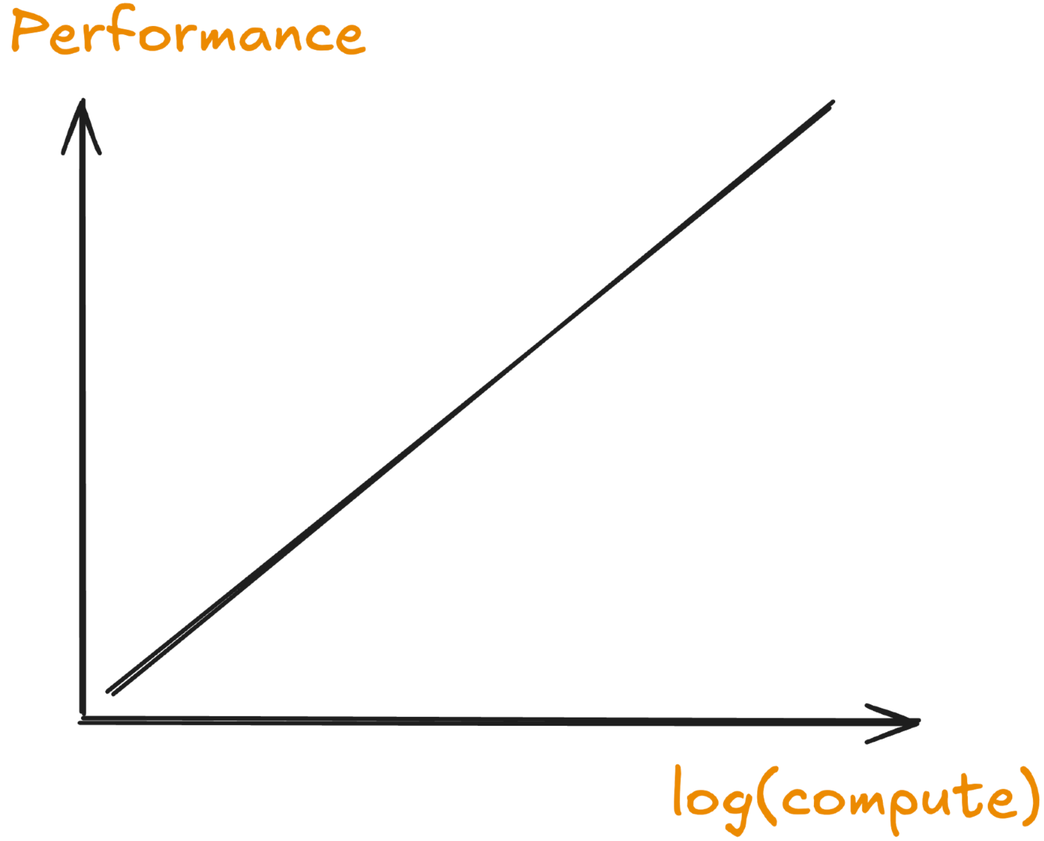
foundations for

Better scaling laws

Future work: autonomous language models

foundations for

Better scaling laws

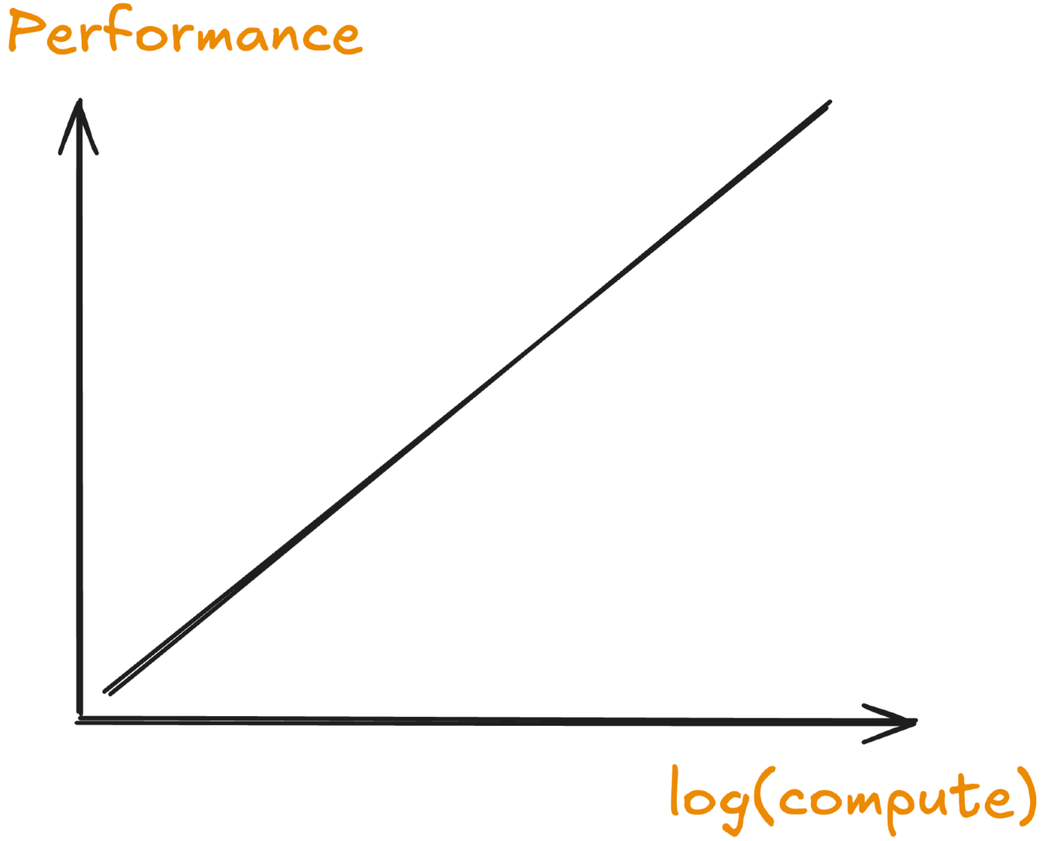


Future work: autonomous language models

foundations for

Better scaling laws

- Advanced capabilities required by agent applications call for more expensive models

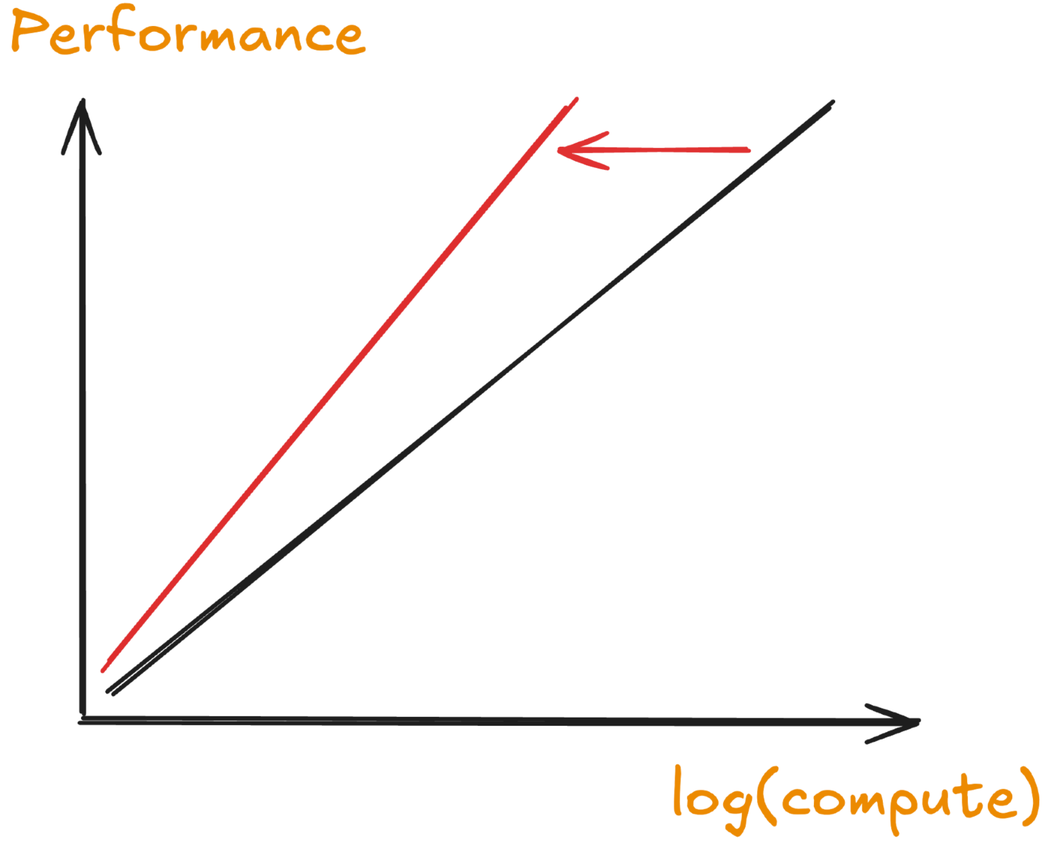


Future work: autonomous language models

foundations for

Better scaling laws

- Advanced capabilities required by agent applications call for more expensive models
- Better scaling laws via architecture innovation

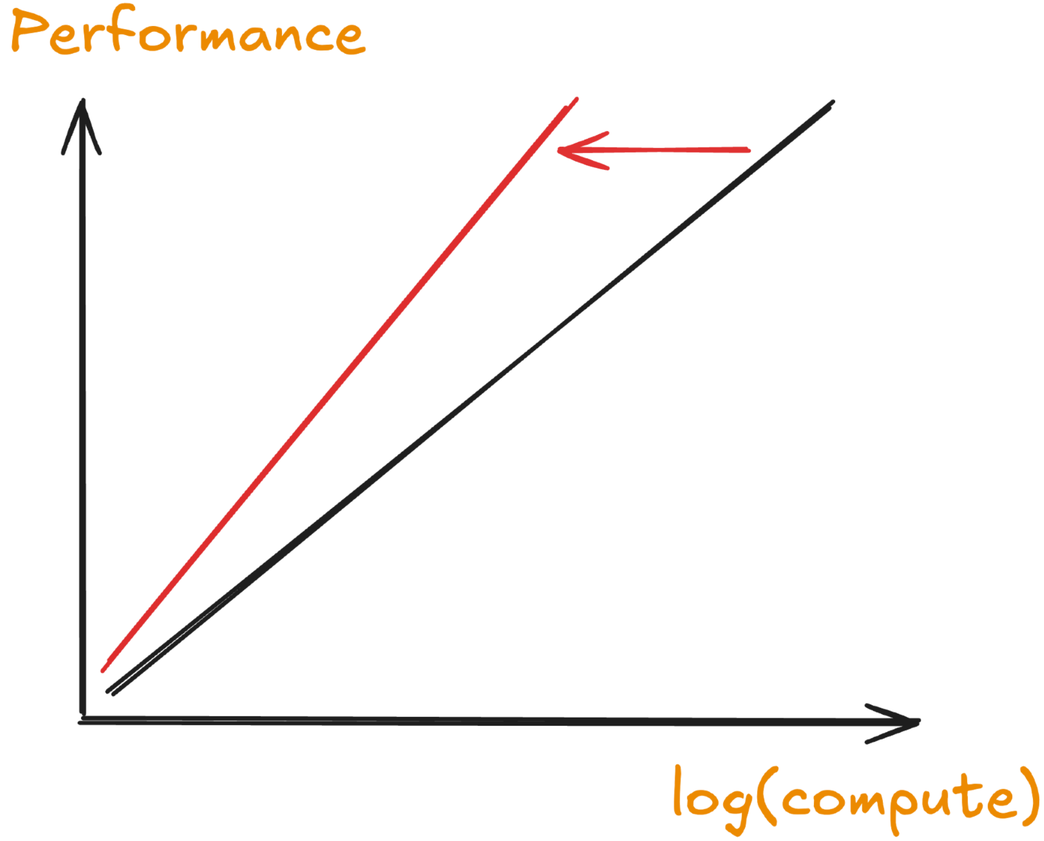


Future work: autonomous language models

foundations for

Better scaling laws

- Advanced capabilities required by agent applications call for more expensive models
- Better scaling laws via architecture innovation
- *Hybrid* models (Transformers+?)



Enabling language models to process information at scale

① Effective long-context processing

Evaluation: ALCE [**G**YYC EMNLP23], HELMET [Y**G**+ ICLR25]
Training: CEPE [Y**G**C ACL24], ProLong [**G***W*YC 2024]

② Accurate search via text embeddings

LM → embedding: SimCSE [**G***Y*C EMNLP21]
Evaluation: LitSearch [AX+**G** EMNLP24]

③ Foundations: Efficient language models

Understanding language models

Conversational evaluation [L***G***GC ACL21 outstanding paper award],
Understanding in-context learning [P**G**CC Findings of ACL23],
Evaluating instruction following: LLMBar [ZY**G**MGC ICLR24]

Accelerating pre-training

MLM masking rates [W***G***ZC EACL23],
Sheared-Llama [X**G**ZC ICLR24], MeCo [**G**W+ 2025]

Efficient fine-tuning

LM-BFF [**G***F*C ACL21], MeZO [M***G***+ NeurIPS23 oral]

Foundations for autonomous language models


Consistent long-form generations


Understanding multimodal contexts


Better scaling laws

PRINCETON UNIVERSITY

清華大學 Tsinghua University

THE UNIVERSITY OF TEXAS AT AUSTIN

Mila