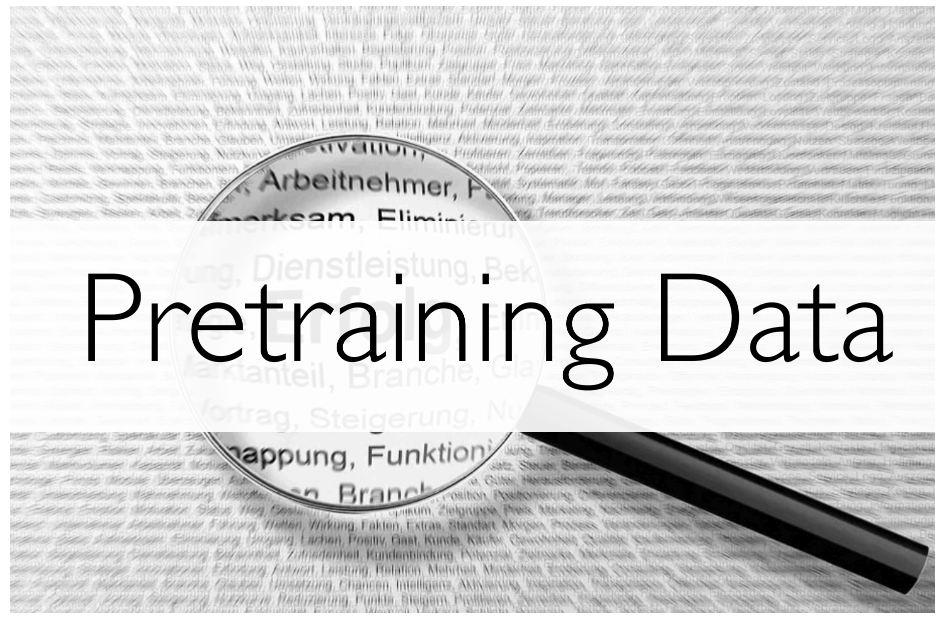
Beyond Monolithic Language Models

Weijia Shi
University of Washington

Current Language Models (LMs): Monolithic

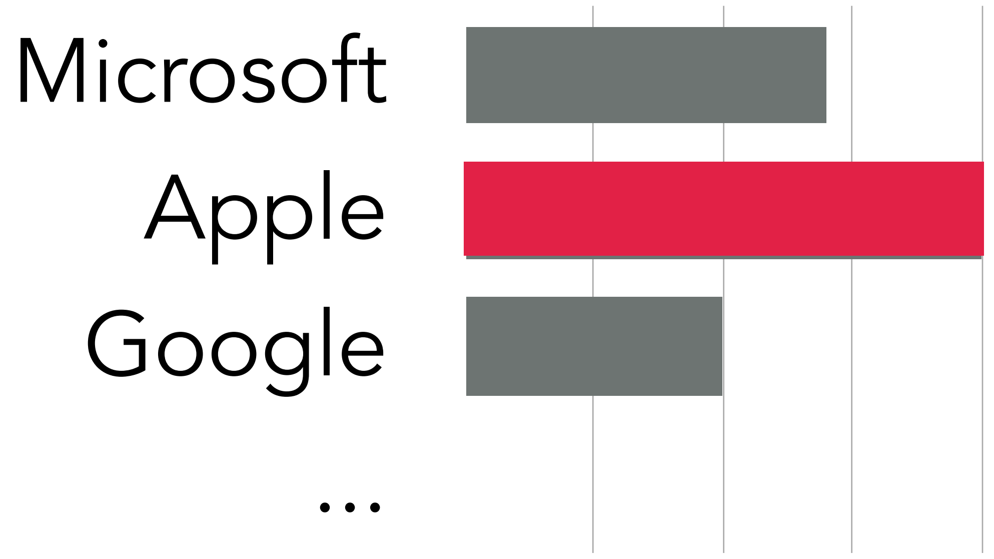

Pretraining Data

Steve Jobs was the CEO of →

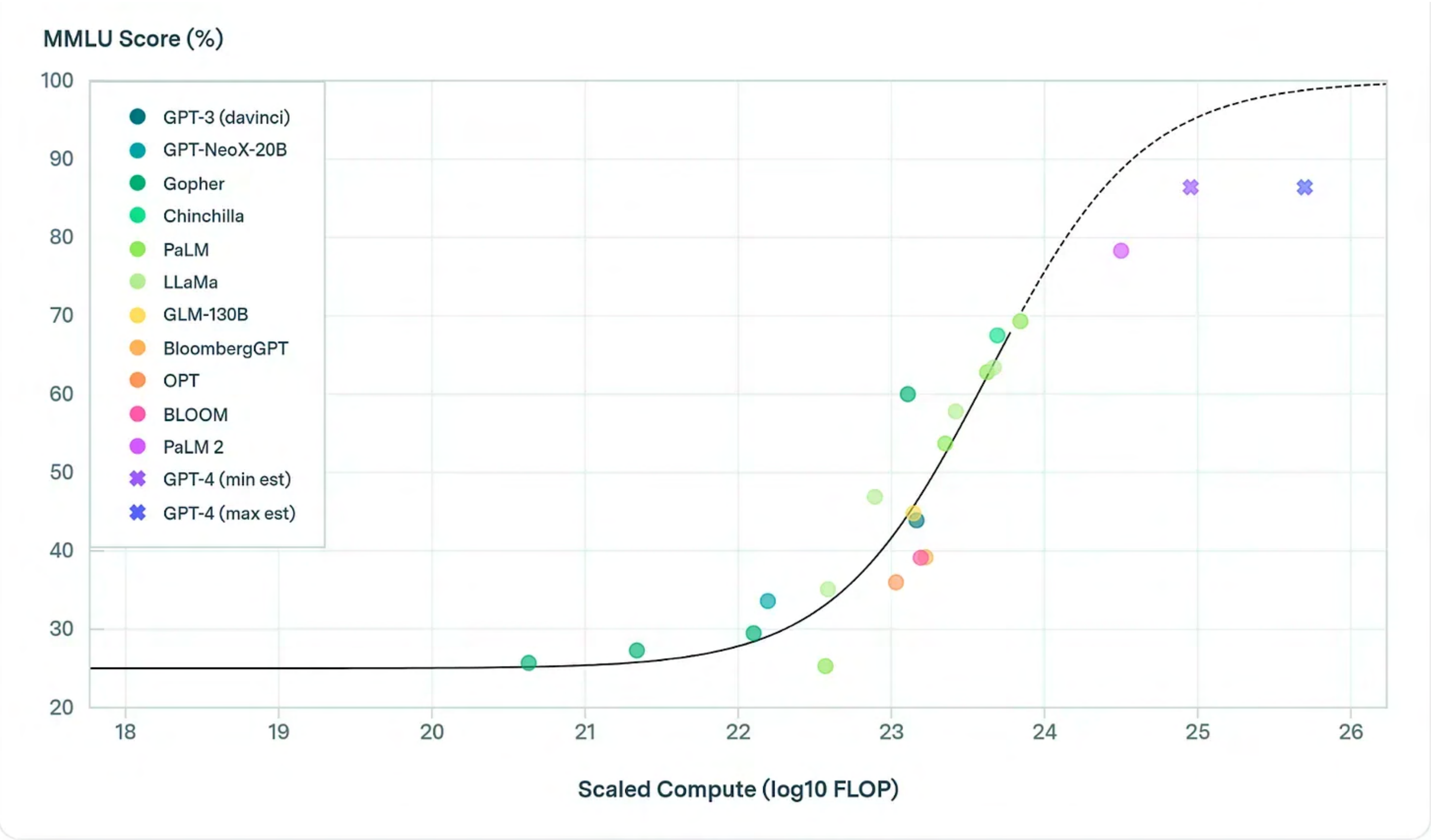
Context

→

LMs



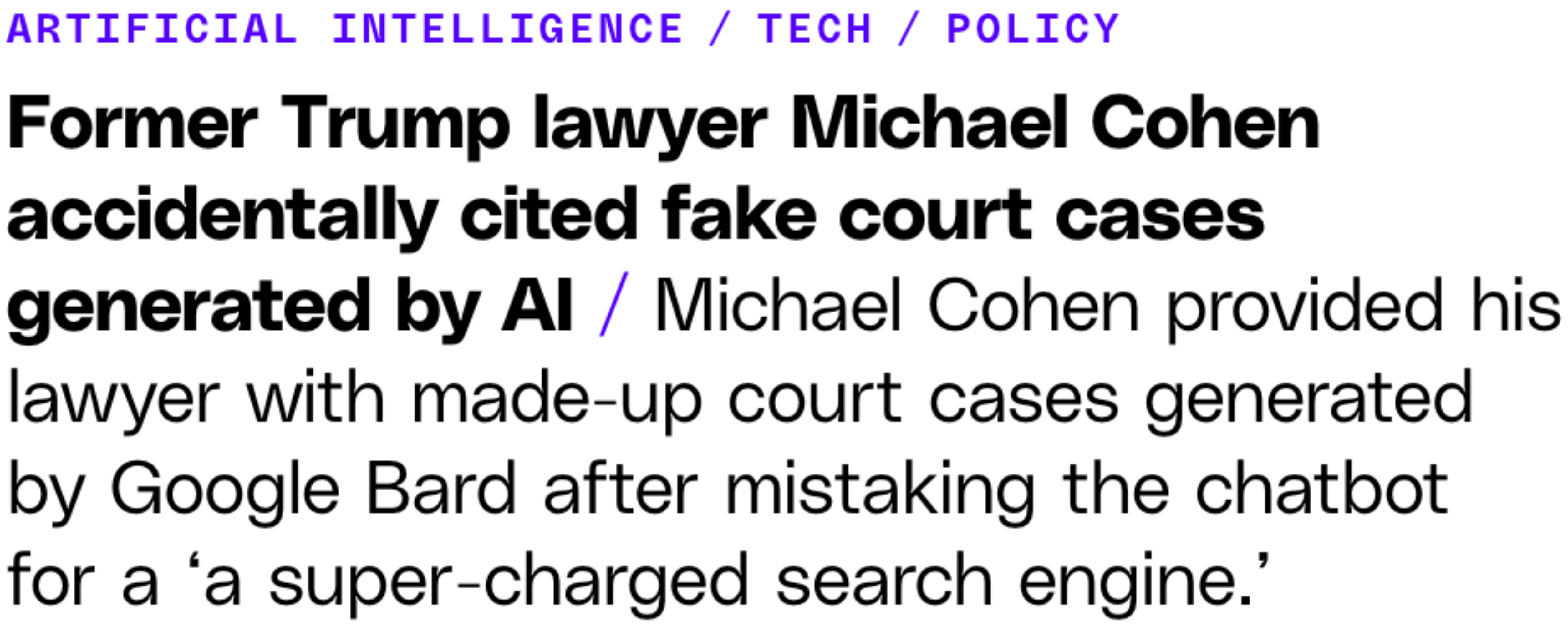
Scaling


MMLU Score (%)
100
90
80
70
60
50
40
30
20
GPT-3 (davinci)
GPT-NeoX-20B
Gopher
Chinchilla
PaLM
LLaMa
GLM-130B
BloombergGPT
OPT
BLOOM
PaLM 2
GPT-4 (min est)
GPT-4 (max est)
18
19
20
21
22
23
24
25
26
Scaled Compute (log10 FLOP)

❓ *Scaling Up Monolithic LMs*

Problems

Hallucinations



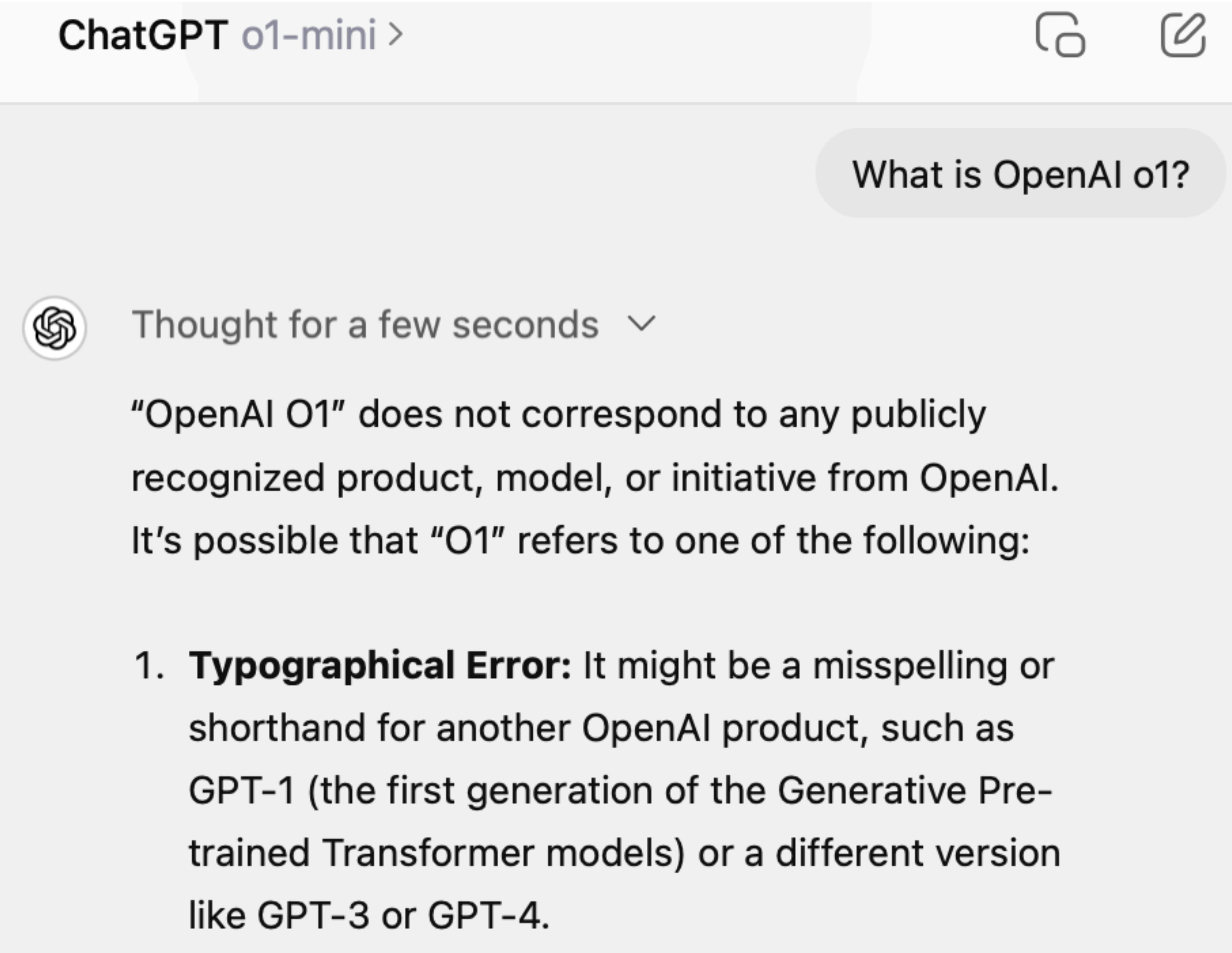
ARTIFICIAL INTELLIGENCE / TECH / POLICY

Former Trump lawyer Michael Cohen accidentally cited fake court cases generated by AI / Michael Cohen provided his lawyer with made-up court cases generated by Google Bard after mistaking the chatbot for a ‘a super-charged search engine.’

Problems

Hallucinations

Hard to update new knowledge



What is OpenAI o1?

Thought for a few seconds

“OpenAI O1” does not correspond to any publicly recognized product, model, or initiative from OpenAI. It’s possible that “O1” refers to one of the following:

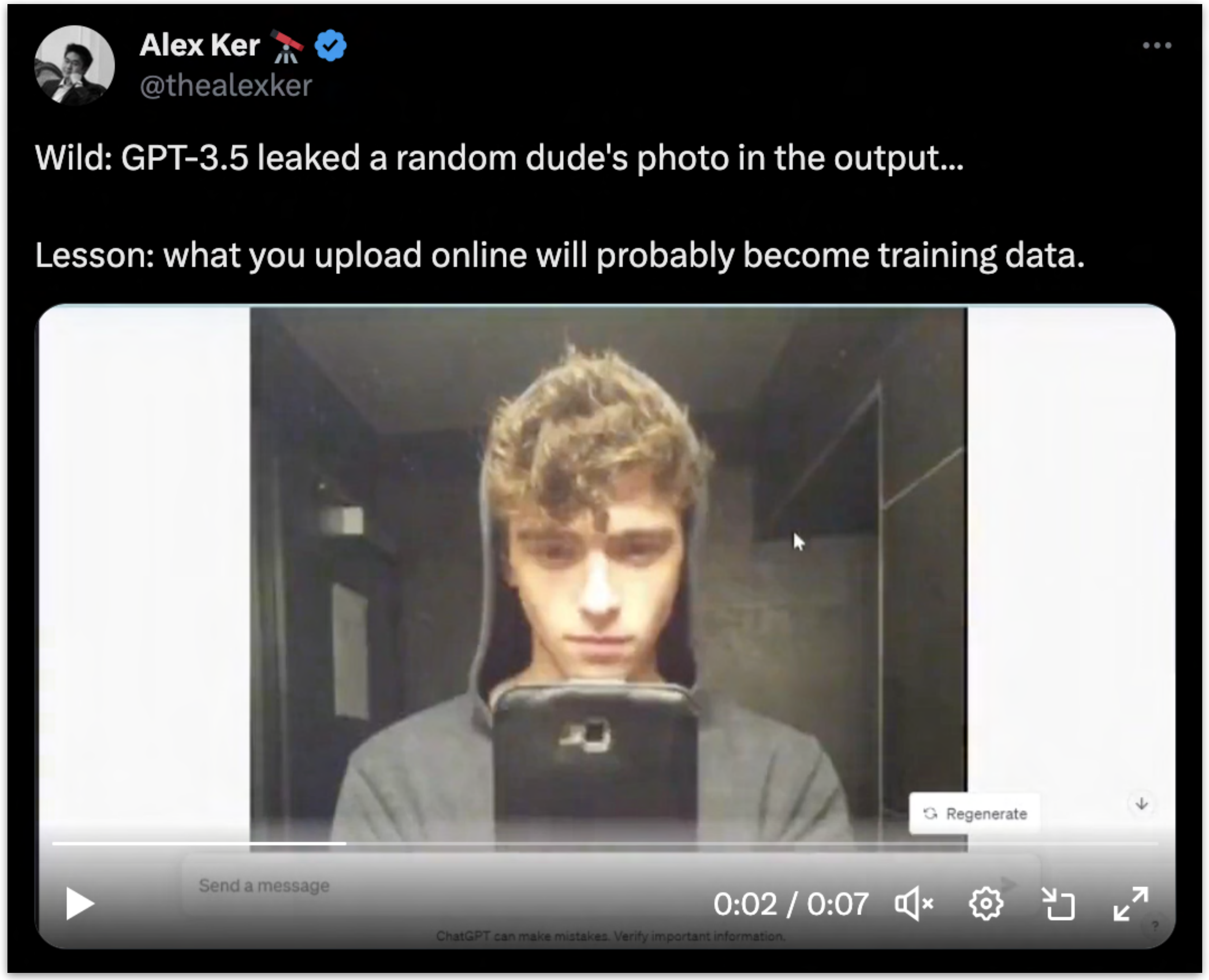
1. **Typographical Error:** It might be a misspelling or shorthand for another OpenAI product, such as GPT-1 (the first generation of the Generative Pre-trained Transformer models) or a different version like GPT-3 or GPT-4.

Problems

Hallucinations

Hard to update new knowledge

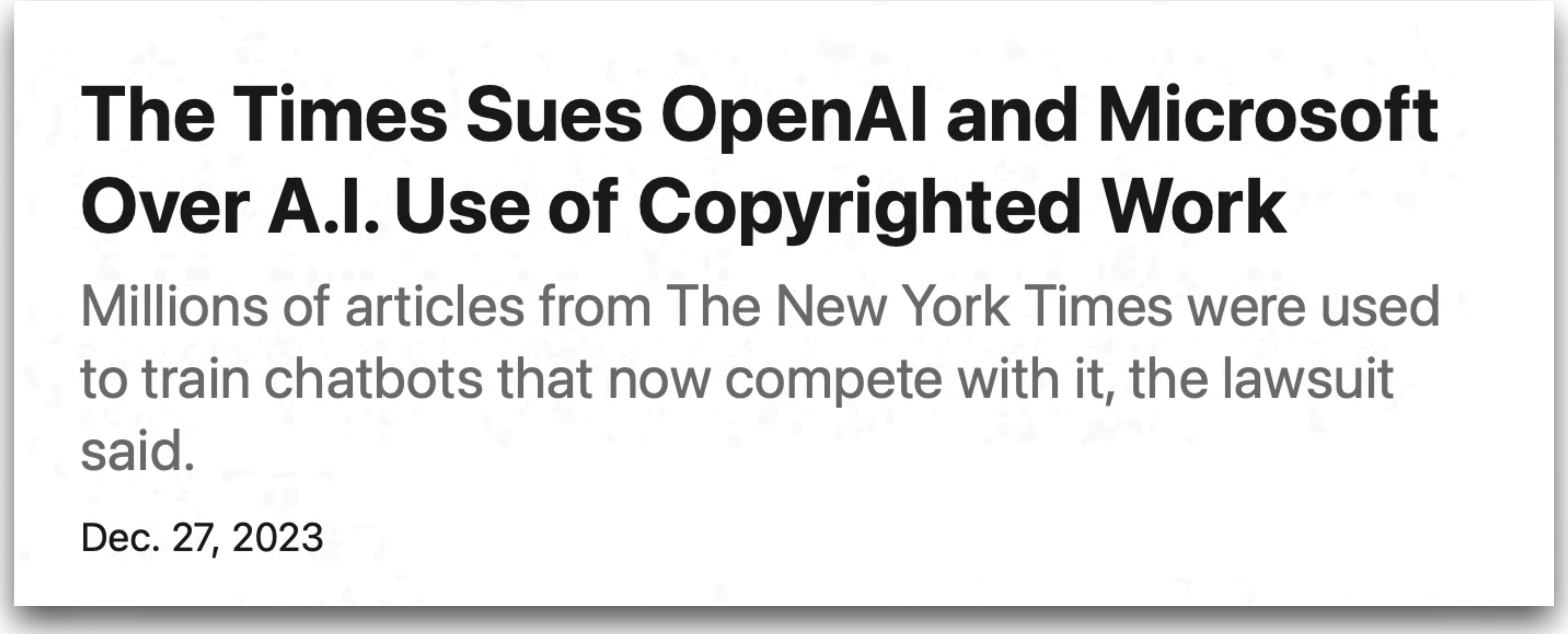
Copyright and privacy risks



Alex Ker @thealexker

Wild: GPT-3.5 leaked a random dude's photo in the output...

Lesson: what you upload online will probably become training data.



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

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

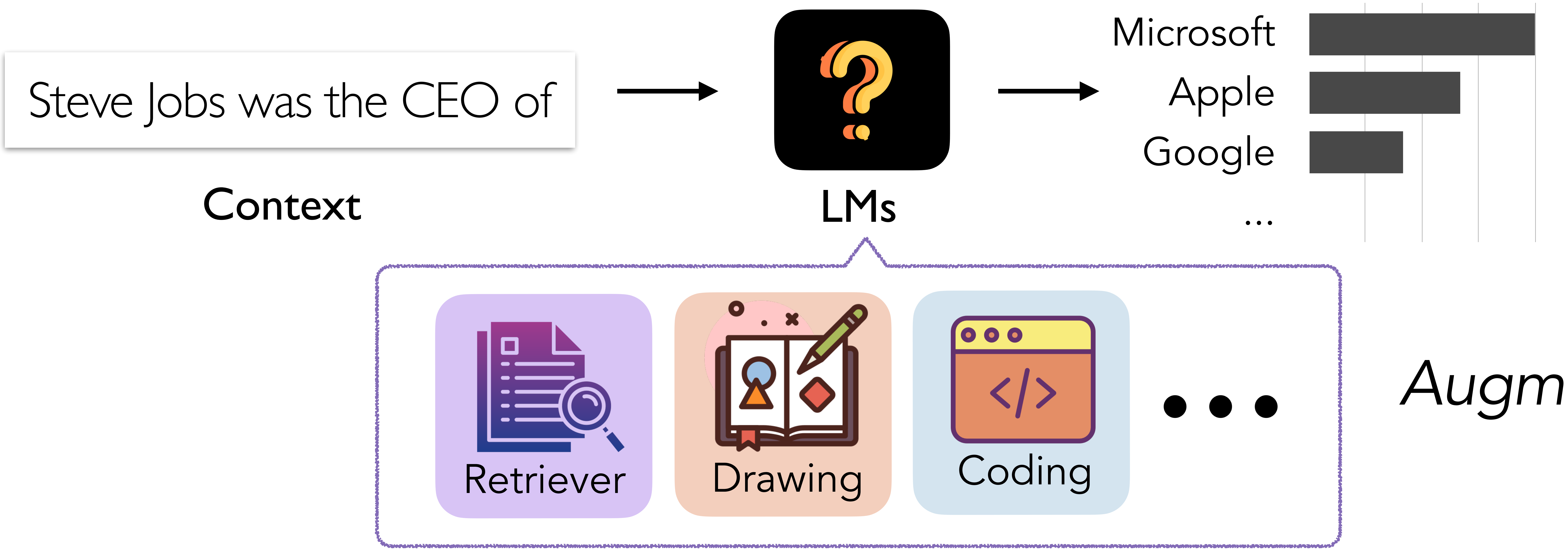
Dec. 27, 2023

✗ *Scaling Up Monolithic LMs*

❌ *Scaling Up Monolithic LMs*

❓ *Alternative Paradigm*

Modularity, *not Monoliths*



Steve Jobs was the CEO of

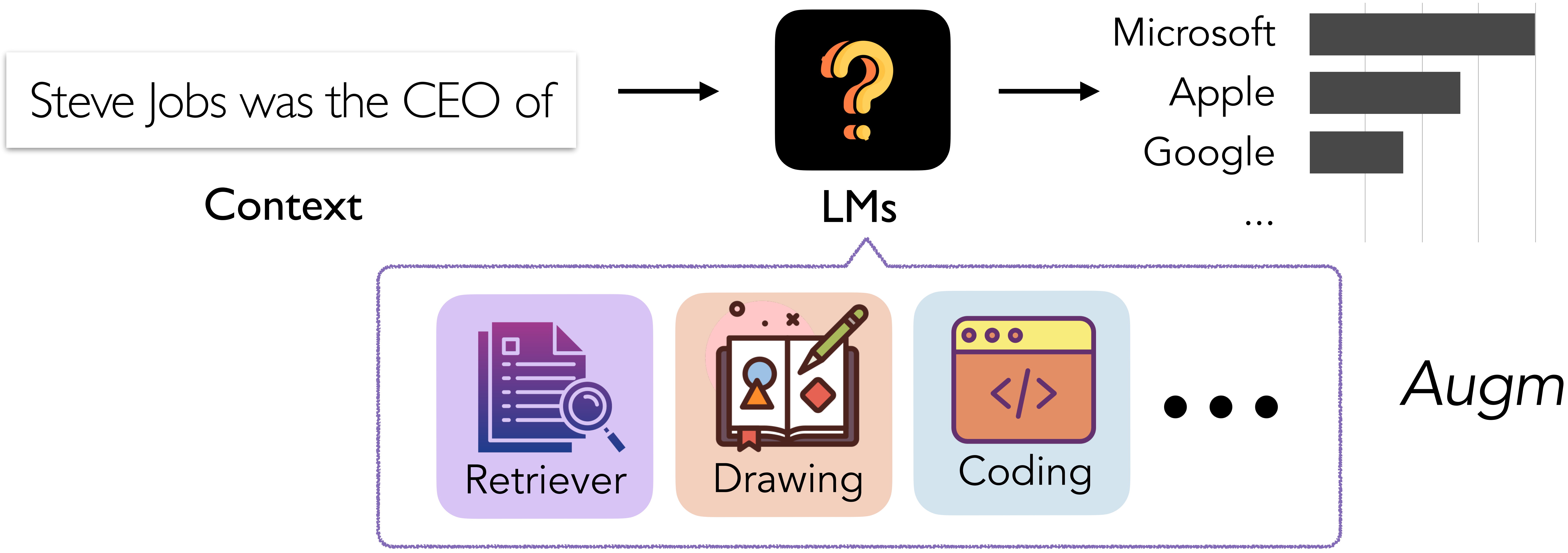
Context

LMs

Microsoft
Apple
Google
...

Retriever
Drawing
Coding
• • •

Augmented Models



Steve Jobs was the CEO of

Context

LMs

Microsoft

Apple

Google

...

Retriever

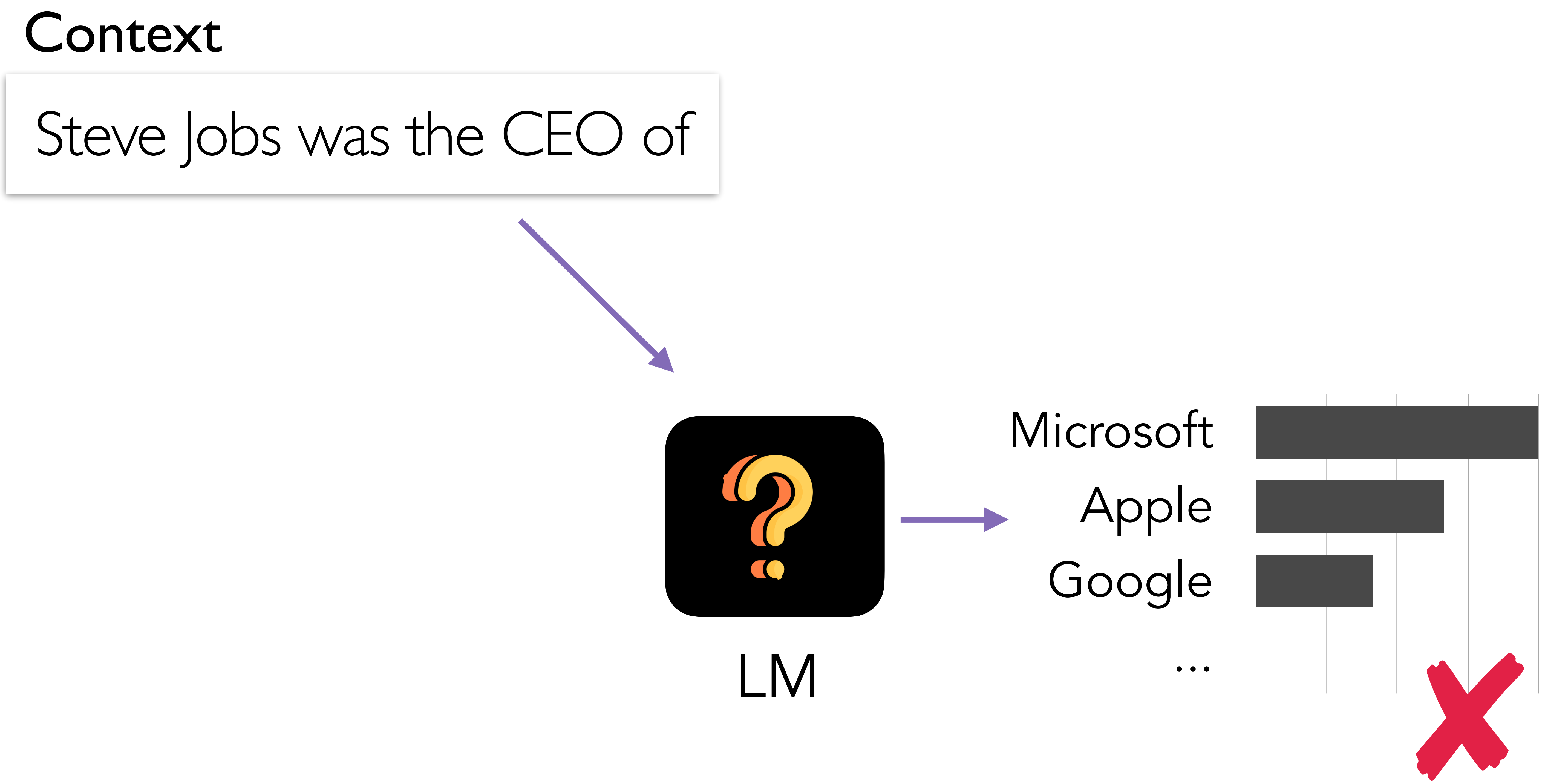
Drawing

Coding

• • •

Augmented Models

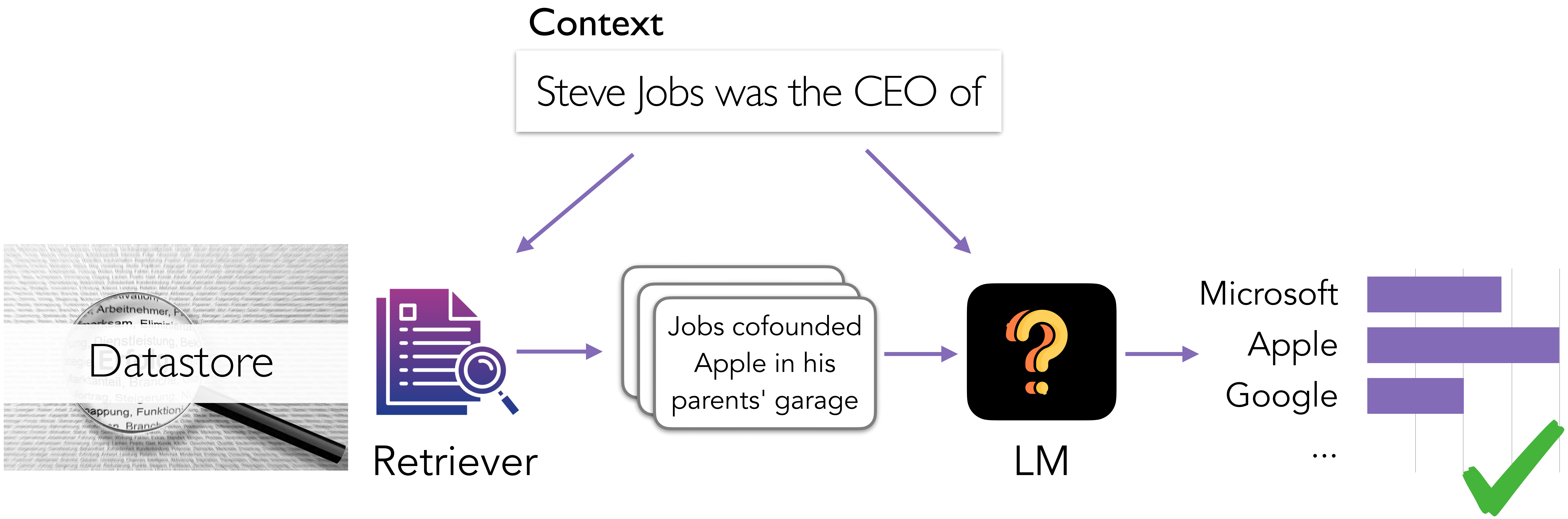
Augmented Models



Context

Steve Jobs was the CEO of

Augmented Models



Context

Steve Jobs was the CEO of

Datastore

Retriever

Jobs cofounded Apple in his parents' garage

LM

Microsoft

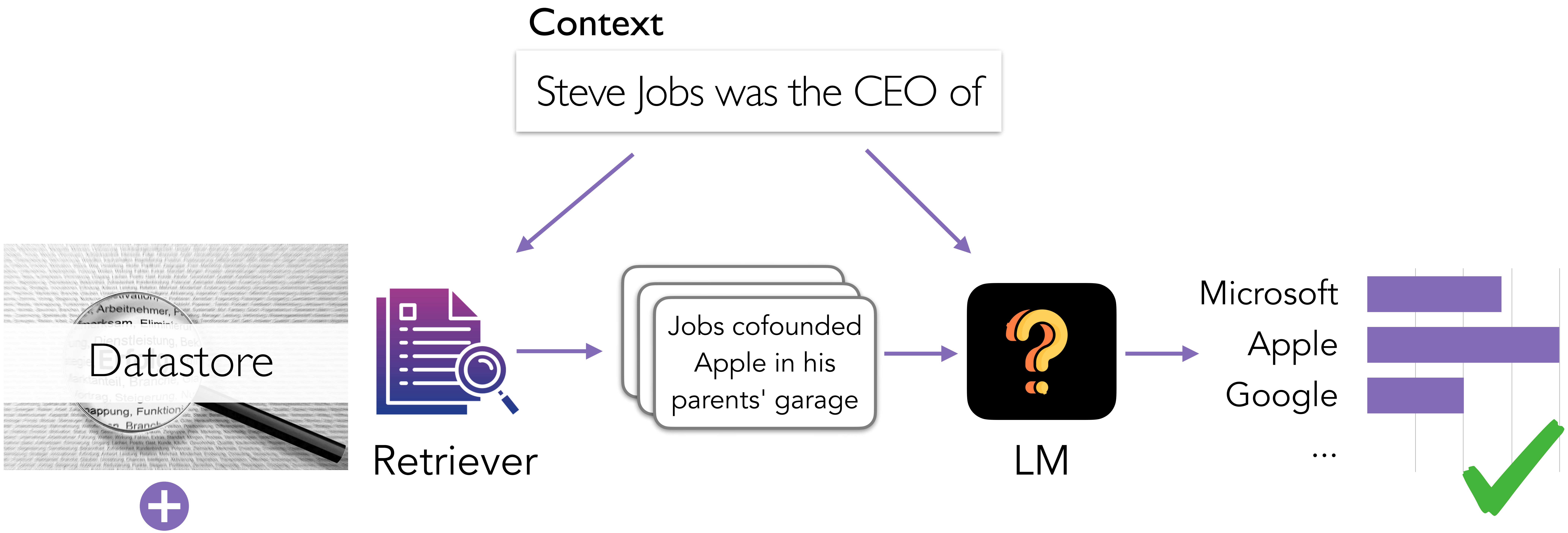
Apple

Google

...

✓ **Hallucinations**

Augmented Models



Context

Steve Jobs was the CEO of

Datastore

Retriever

Jobs cofounded Apple in his parents' garage

LM

Microsoft

Apple

Google

…

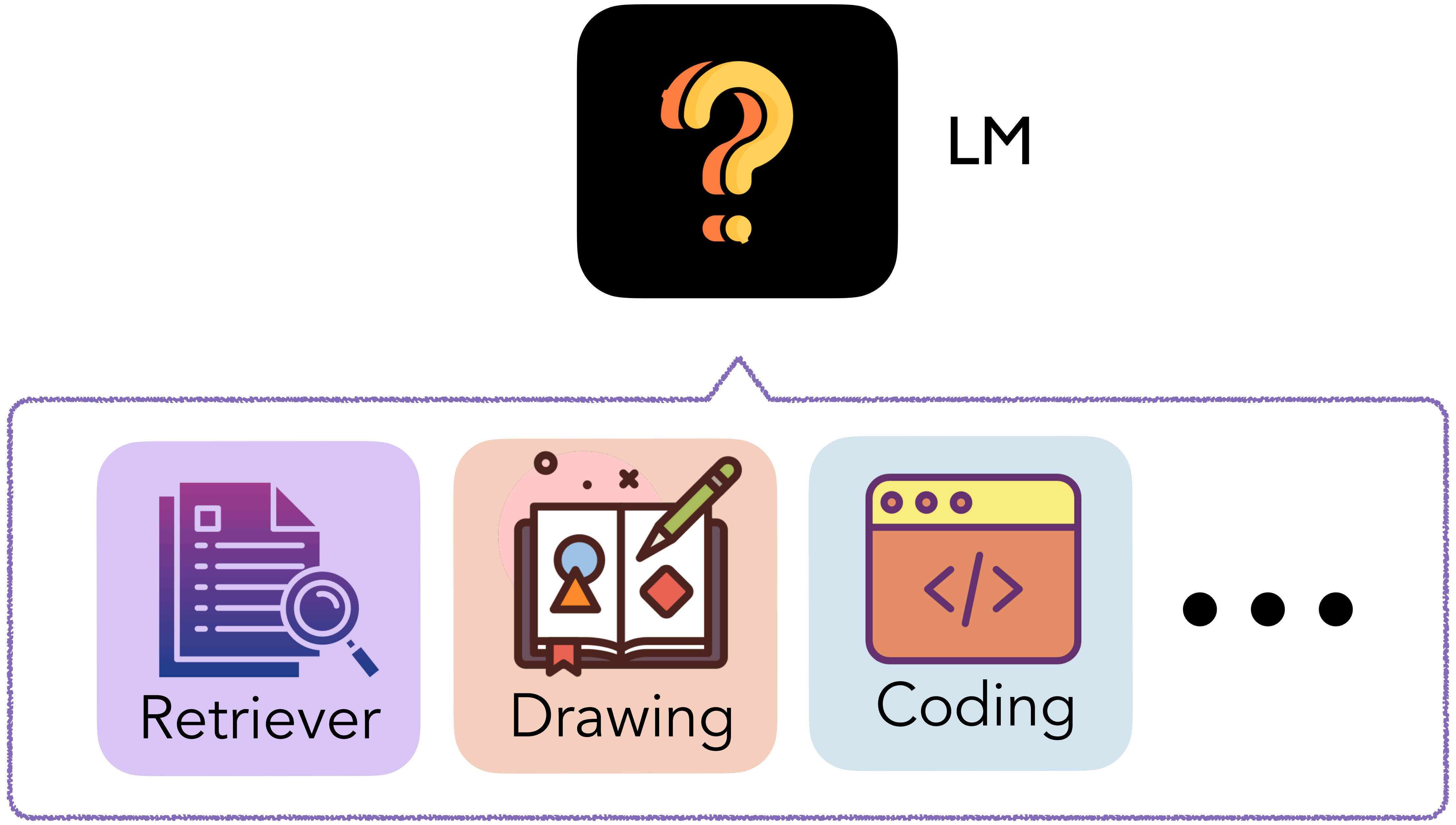
‘In awe’: scientists impressed by latest ChatGPT model o1

The chatbot excels at science, beating PhDs on a hard science test. But it might ‘hallucinate’ more than its predecessors.

By Nicola Jones

Technology firm OpenAI released a preview version of its latest chatbot, o1, last month. Credit: GK Images/Alamy

✓ **Hard to update new knowledge**



LM

Retriever

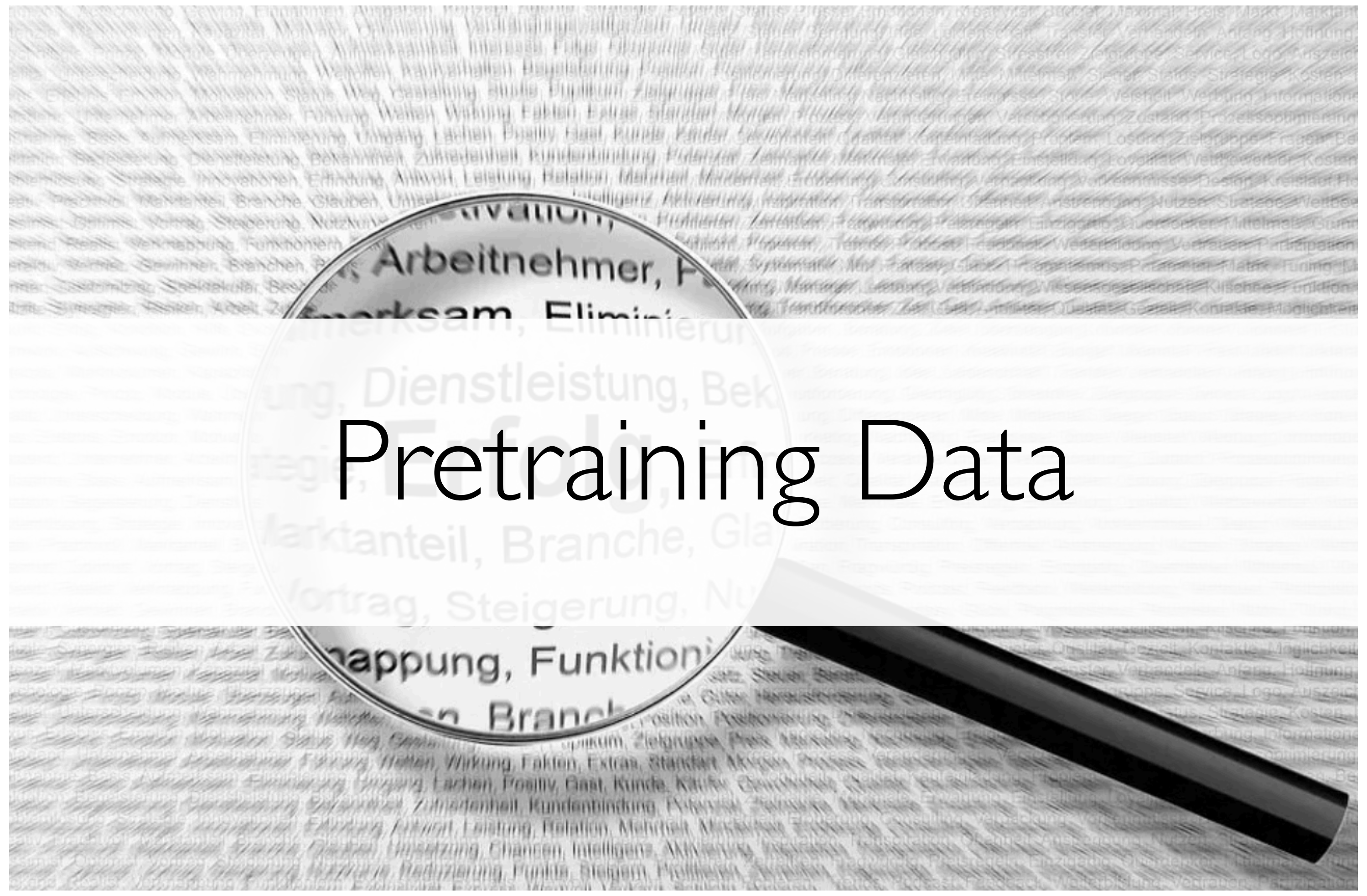
Drawing

Coding

• • •

Augmented Models

Pretraining Data is not Monolithic



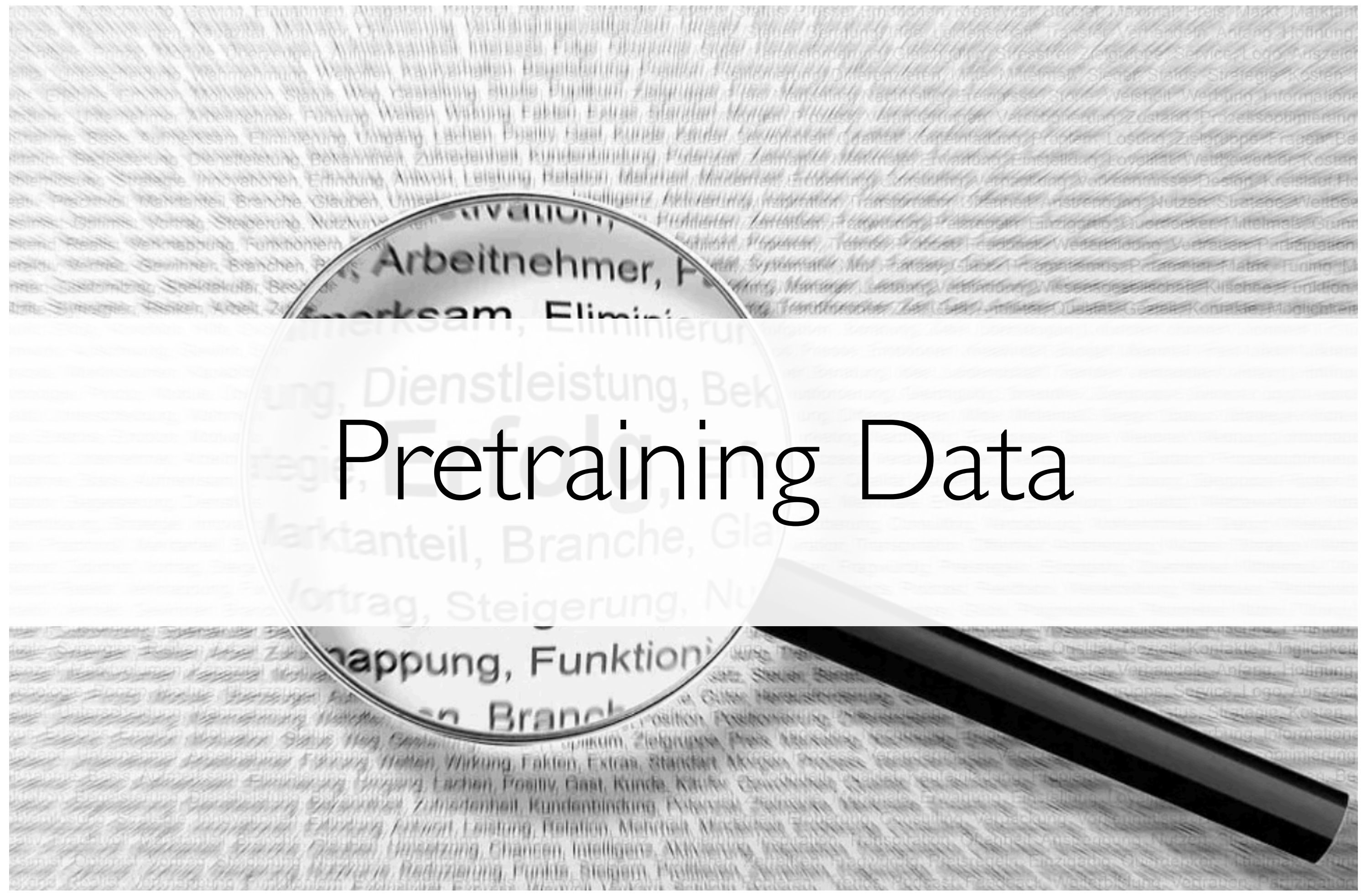
Pretraining Data

PROJECT GUTENBERG MIT OCW *Public*

The Guardian Disney+ *Copyright*

Private

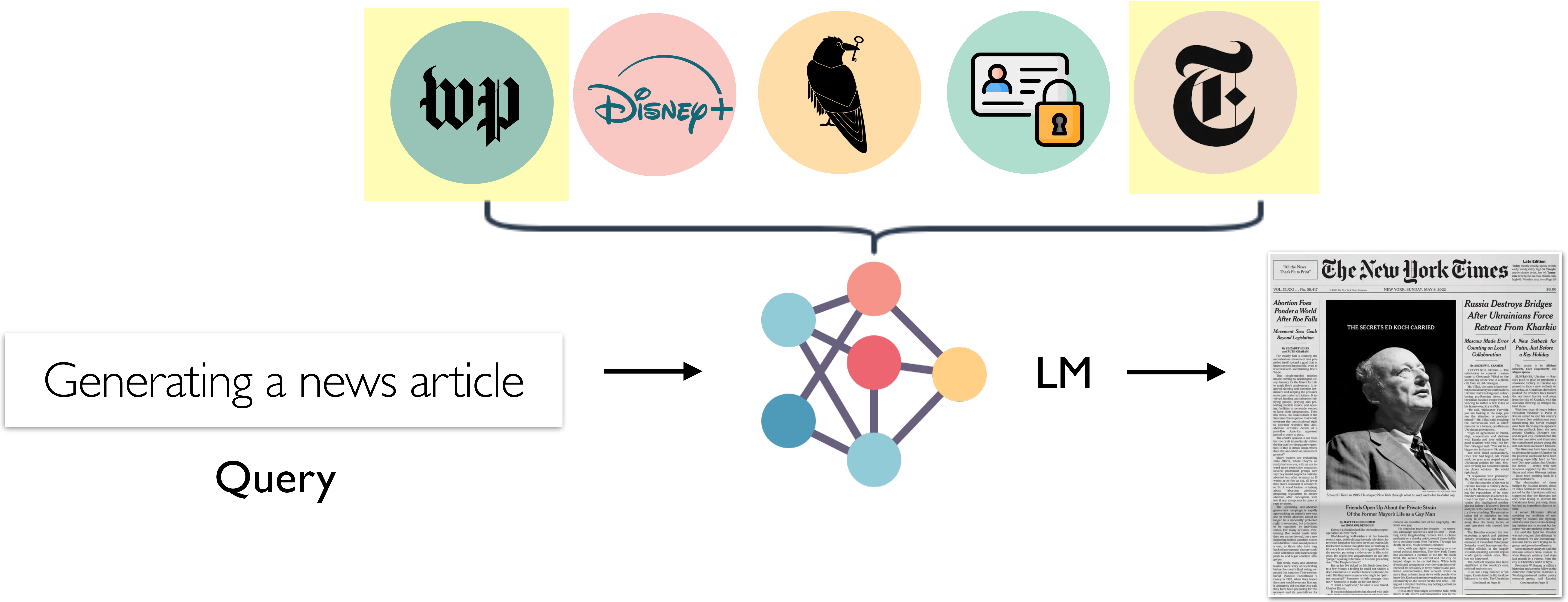
MMLU *Benchmark (contamination)*



Pretraining Data

LM

Data Modularity



Generating a news article

Query

LM

"All the News That's Fit to Print"

The New York Times

Late Edition

Today, mostly cloudy, spotty drizzle early, windy, chilly, high 56. Tonight, partly cloudy, brisk, low 46. Tomorrow, breezy, not as cool, clouds, sun, high 63. Weather map is on Page 29.

VOL. CLXXI ... No. 59,417 © 2022 The New York Times Company NEW YORK, SUNDAY, MAY 8, 2022 $6.00

Abortion Foes Ponder a World After Roe Falls

Movement Sees Goals Beyond Legislation

By ELIZABETH DIAS and RUTH GRAHAM

For nearly half a century, the anti-abortion movement has propelled itself toward a goal that at times seemed impossible, even to true believers: overturning Roe v. Wade.

That single-minded mission meant coming to Washington every January for the March for Life to mark Roe's anniversary. It required electing anti-abortion lawmakers and keeping the pressure on to pass state restrictions. It involved funding anti-abortion lobbying groups, praying and protesting outside clinics, and opening facilities to persuade women to keep their pregnancies. Then this week, the leaked draft of the Supreme Court opinion that would overturn the constitutional right to abortion revealed that anti-abortion activists' dream of a post-Roe America appeared poised to come to pass.

The court's opinion is not final, but the draft immediately shifted the horizon by raising a new question: If Roe is struck down, where does the anti-abortion movement go next?

Many leaders are redoubling state efforts, where they've already had success, with an eye toward more restrictive measures. Several prominent groups now say they would support a national abortion ban after as many as 15 weeks or as few as six, all lower than Roe's standard of around 23 or 24. A vocal faction is talking about "abortion abolition," proposing legislation to outlaw abortion after conception, with few if any exceptions in cases of rape or incest.

The sprawling anti-abortion grass-roots campaign is rapidly approaching an entirely new era, one in which abortion would no longer be a nationally protected right to overcome, but a decision to be legislated by individual states. For many activists, overturning Roe would mark what they see as not the end, but a new beginning to limit abortion access even further. It also would present a test, as those who have long backed incremental change could clash with those who increasingly push to end legal abortion altogether.

This week, many anti-abortion leaders were wary of celebrating before the court's final ruling, expected this summer. They remembered Planned Parenthood v. Casey in 1992, when they hoped the court would overturn Roe and it ultimately did not. But they said they have been preparing for this moment and its possibilities for

THE SECRETS ED KOCH CARRIED

JACK MANNING/THE NEW YORK TIMES

Edward I. Koch in 1985. He shaped New York through what he said, and what he didn't say.

Friends Open Up About the Private Strain Of the Former Mayor's Life as a Gay Man

By MATT FLEGENHEIMER and ROSA GOLDENSOHN

Edward I. Koch looked like the busiest septuagenarian in New York.

Glad-handing well-wishers at his favorite restaurants, gesticulating through television interviews long after his three terms as mayor, Mr. Koch could seem as though he was scrambling to fill every hour with bustle. He dragged friends to the movies, pursuing a side career in film criticism. He urged new acquaintances to call him "judge," a joking reference to his time presiding over "The People's Court."

But as his 70s ticked by, Mr. Koch described to a few friends a feeling he could not shake: a deep loneliness. He wanted to meet someone, he said. Did they know anyone who might be "partner material?" Someone "a little younger than me?" Someone to make up for lost time?

"I want a boyfriend," he said to one friend, Charles Kaiser.

It was an aching admission, shared with only conceal an essential fact of his biography: Mr. Koch was gay.

He denied as much for decades — to reporters, campaign operatives and his staff — swatting away longstanding rumors with a choice profanity or a cheeky aside, even if these did little to convince some New Yorkers. Through his death, in 2013, his deflections endured.

Now, with gay rights re-emerging as a national political tinderbox, The New York Times has assembled a portrait of the life Mr. Koch lived, the secrets he carried and the city he helped shape as he carried them. While both friends and antagonists over the years have referenced his sexuality in stray remarks and published commentaries, this account draws on more than a dozen interviews with people who knew Mr. Koch and are in several cases speaking extensively on the record for the first time — filling out a chapter that they say belongs, at last, to the sweep of history.

It is a story that might otherwise fade, with

Russia Destroys Bridges After Ukrainians Force Retreat From Kharkiv

Moscow Made Error Counting on Local Collaboration

By ANDREW E. KRAMER

KRYVYI RIH, Ukraine — The solicitation to commit treason came to Oleksandr Vilkul on the second day of the war, in a phone call from an old colleague.

Mr. Vilkul, the scion of a powerful political family in southeastern Ukraine that was long seen as harboring pro-Russian views, took the call as Russian troops were advancing to within a few miles of his hometown, Kryvyi Rih.

"He said, 'Oleksandr Yurivich, you are looking at the map, you see the situation is predetermined,'" Mr. Vilkul said, recalling the conversation with a fellow minister in a former, pro-Russian Ukrainian government.

"Sign an agreement of friendship, cooperation and defense with Russia and they will have good relations with you," the former colleague said. "You will be a big person in the new Ukraine."

The offer failed spectacularly. Once war had begun, Mr. Vilkul said, the gray area seeped out of Ukrainian politics for him. Missiles striking his hometown made the choice obvious: He would fight back.

"I responded with profanity," Mr. Vilkul said in an interview.

If the first months of the war in Ukraine became a military debacle for the Russian army — deflating the reputations of its commanders and troops in a forced retreat from Kyiv — the Russian invasion also highlighted another glaring failure: Moscow's flawed analysis of the politics of the country it was attacking. The miscalculation led to mistakes no less costly in lives for the Russian army than the faulty tactics of tank operators who steered into bogs.

The Kremlin entered the war expecting a quick and painless victory, predicting that the government of President Volodymyr Zelensky would fracture and that leading officials in the largely Russian-speaking eastern region would gladly switch sides. That has not happened.

The political myopia was most significant in the country's east, political analysts say.

In all but a tiny number of villages, Russia failed to flip local politicians to its side. The Ukrainian

Continued on Page 16

A New Setback for Putin, Just Before a Key Holiday

This article is by Michael Schwirtz, Cora Engelbrecht and Megan Specia.

SLOVIANSK, Ukraine — Russia's push to give its president a showcase victory in Ukraine appeared to face a new setback on Saturday, as Ukrainian defenders pushed the invaders back toward the northeast border and away from the city of Kharkiv, with the Russians blowing up bridges behind them.

With less than 48 hours before President Vladimir V. Putin of Russia aimed to lead his country in Victory Day celebrations commemorating the Soviet triumph over Nazi Germany, the apparent Russian pullback from the area around Kharkiv, Ukraine's second-largest city, contradicted the Russian narrative and illustrated the complicated picture along the 300-mile front in eastern Ukraine.

The Russians have been trying to advance in eastern Ukraine for the past few weeks and have been pushing especially hard as Victory Day approaches, but Ukrainian forces — armed with new weapons supplied by the United States and other Western nations — have been pushing back in a counteroffensive.

The destruction of three bridges by Russian forces, about 12 miles northeast of Kharkiv, reported by the Ukrainian military, suggested that the Russians not only were trying to prevent the Ukrainians from pursuing them, but had no immediate plans to return.

A senior Ukrainian official, speaking on condition of anonymity to discuss the fighting, said Russian forces were destroying bridges not to retreat but because "we are pushing them out."

He said the fight for Kharkiv was not over, and that although "at the moment we are dominating," Russian forces were trying to regroup and go on the offensive.

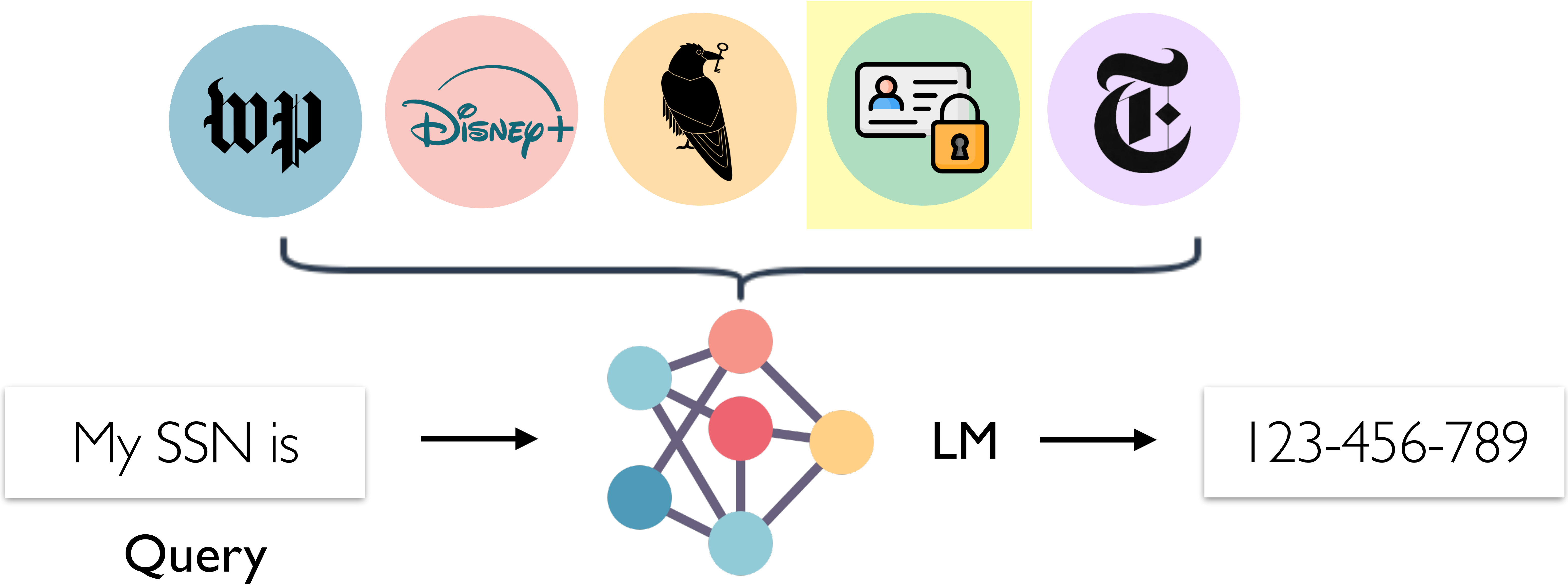
Some military analysts said the Russian actions were similar to what Russia's military had done last month in a retreat from the city of Chernihiv north of Kyiv.

Frederick W. Kagan, a military historian and a senior fellow at the American Enterprise Institute, a Washington-based public policy research group, said Russia's

Continued on Page 16

Copyright and privacy risks

Data Modularity



My SSN is

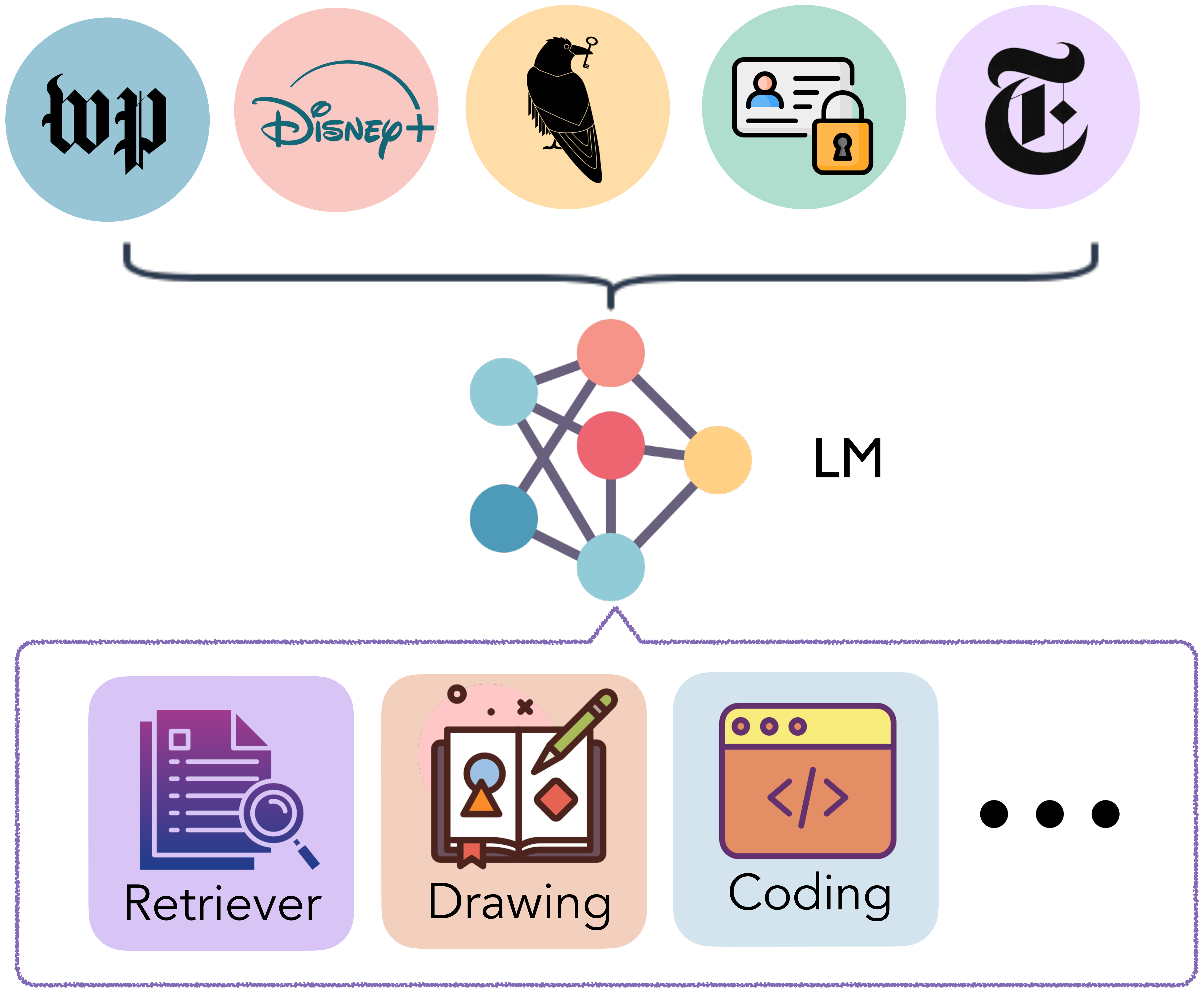
Query

LM

123-456-789

Copyright and privacy risks

Modularity, not Monoliths



Data Modularity

Augmented Models

Beyond Monolithic Language Models

Augmented Models

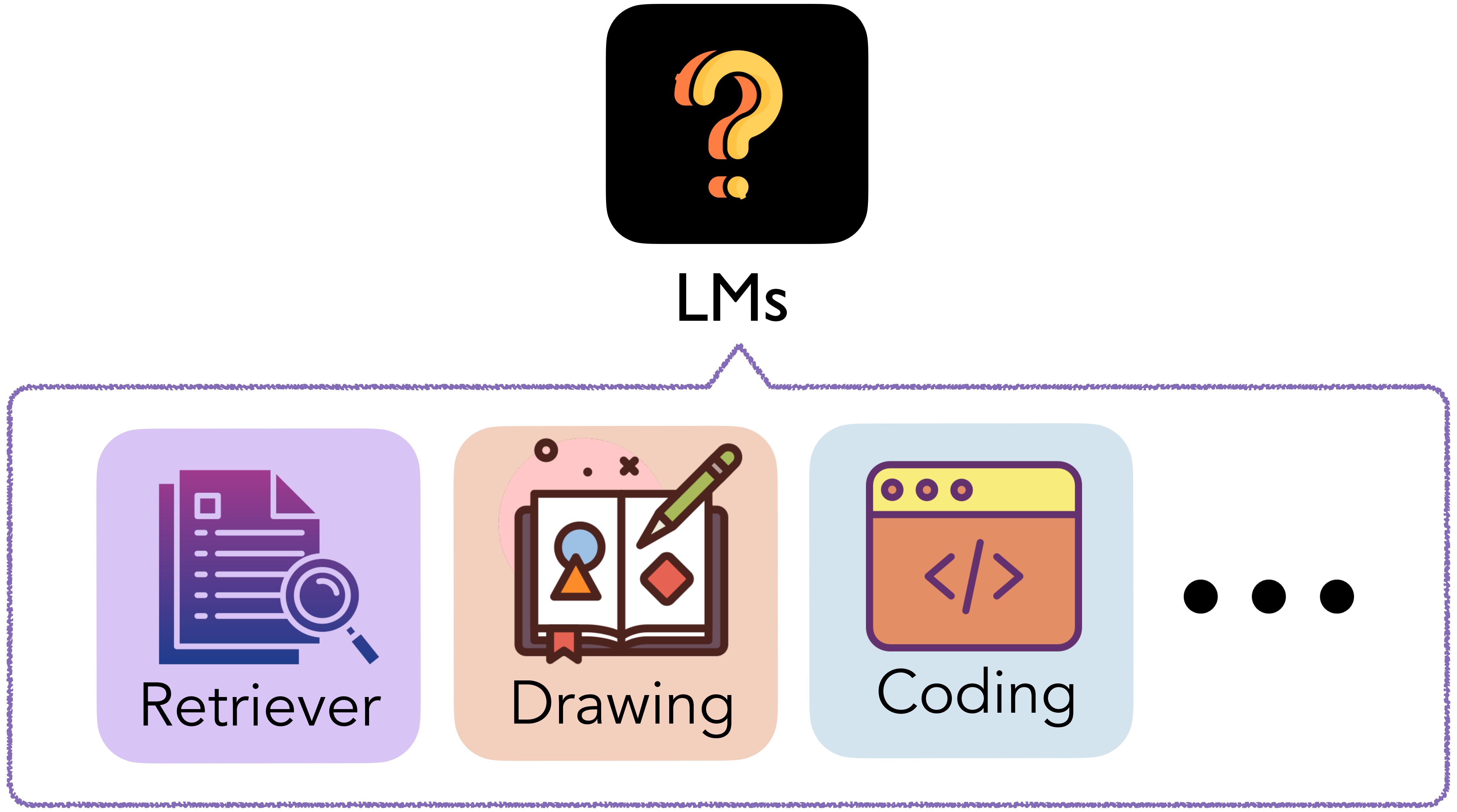
Data Modularity

Beyond Monolithic Language Models

Augmented Models

Data Modularity

Augmented Models



LMs

Retriever

Drawing

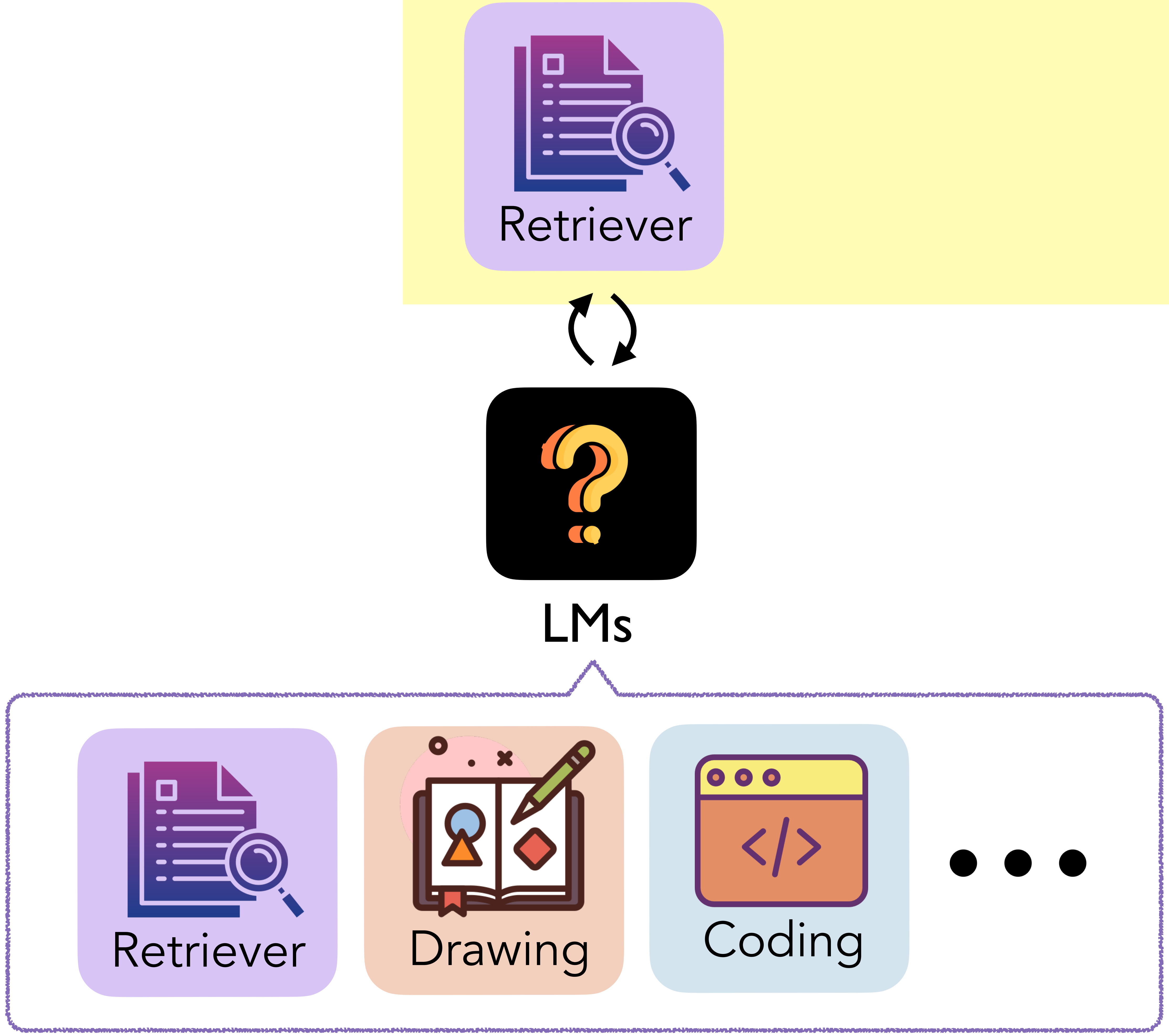
Coding

• • •

Inference

Hu*, **Shi*** et al., NeurIPS 2024

Augmented Models



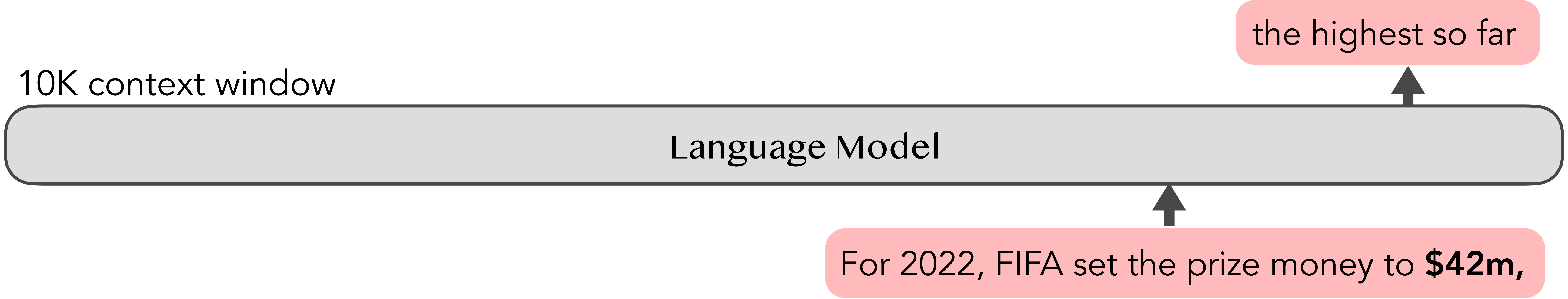
Pretraining

Shi, et al. ICLR 2024 Spotlight

Inference

Hu*, **Shi*** et al., NeurIPS 2024

Standard Pretraining



the highest so far

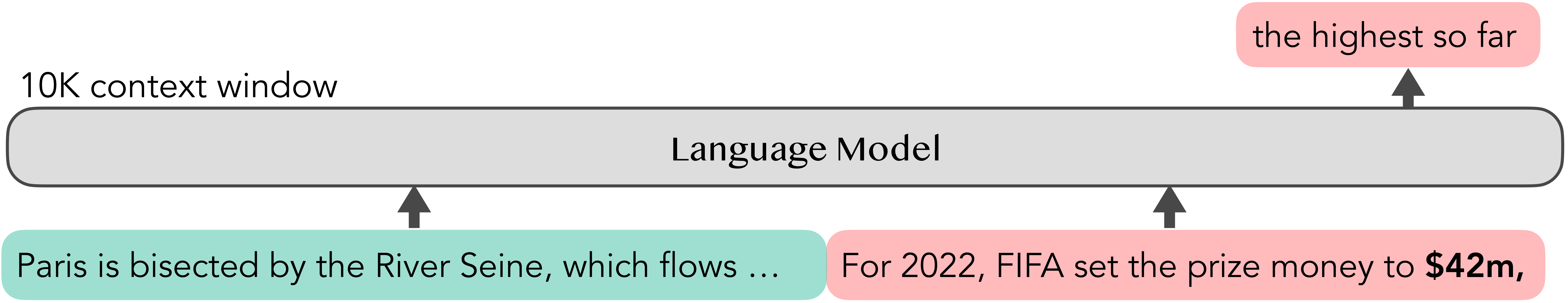
10K context window

Language Model

For 2022, FIFA set the prize money to **$42m,**

Standard Pretraining

Concatenate Random Documents


the highest so far
10K context window
Language Model
Paris is bisected by the River Seine, which flows …
For 2022, FIFA set the prize money to **$42m,**

The prior doc provides no signal for predicting the next doc

Doc
Doc

Problem: Fails to Understand Long Contexts

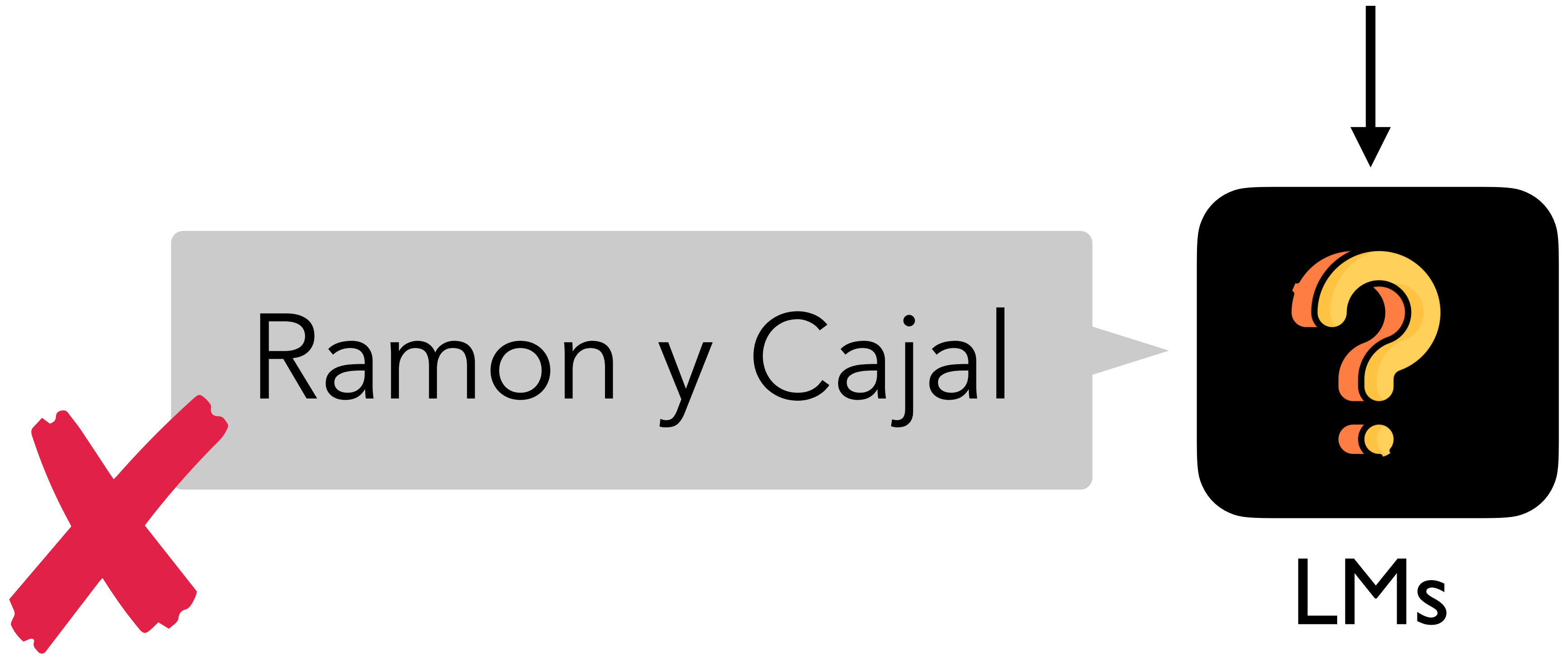
Input Context

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

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

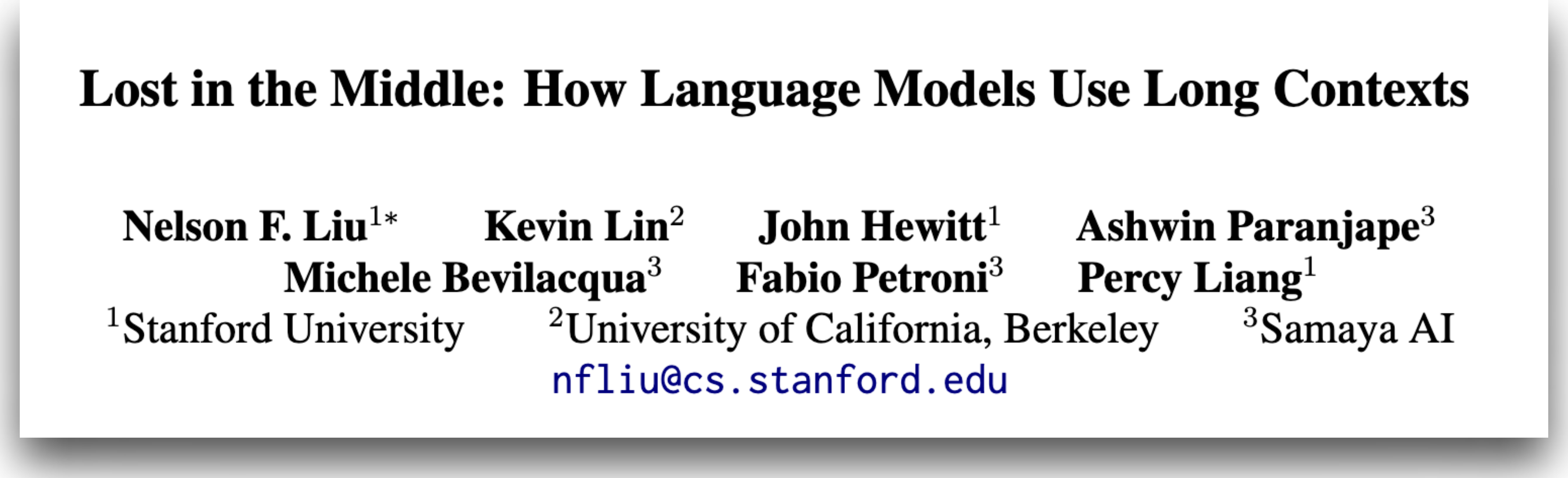
Question: who got the first nobel prize in physics
Answer:

...



Ramon y Cajal

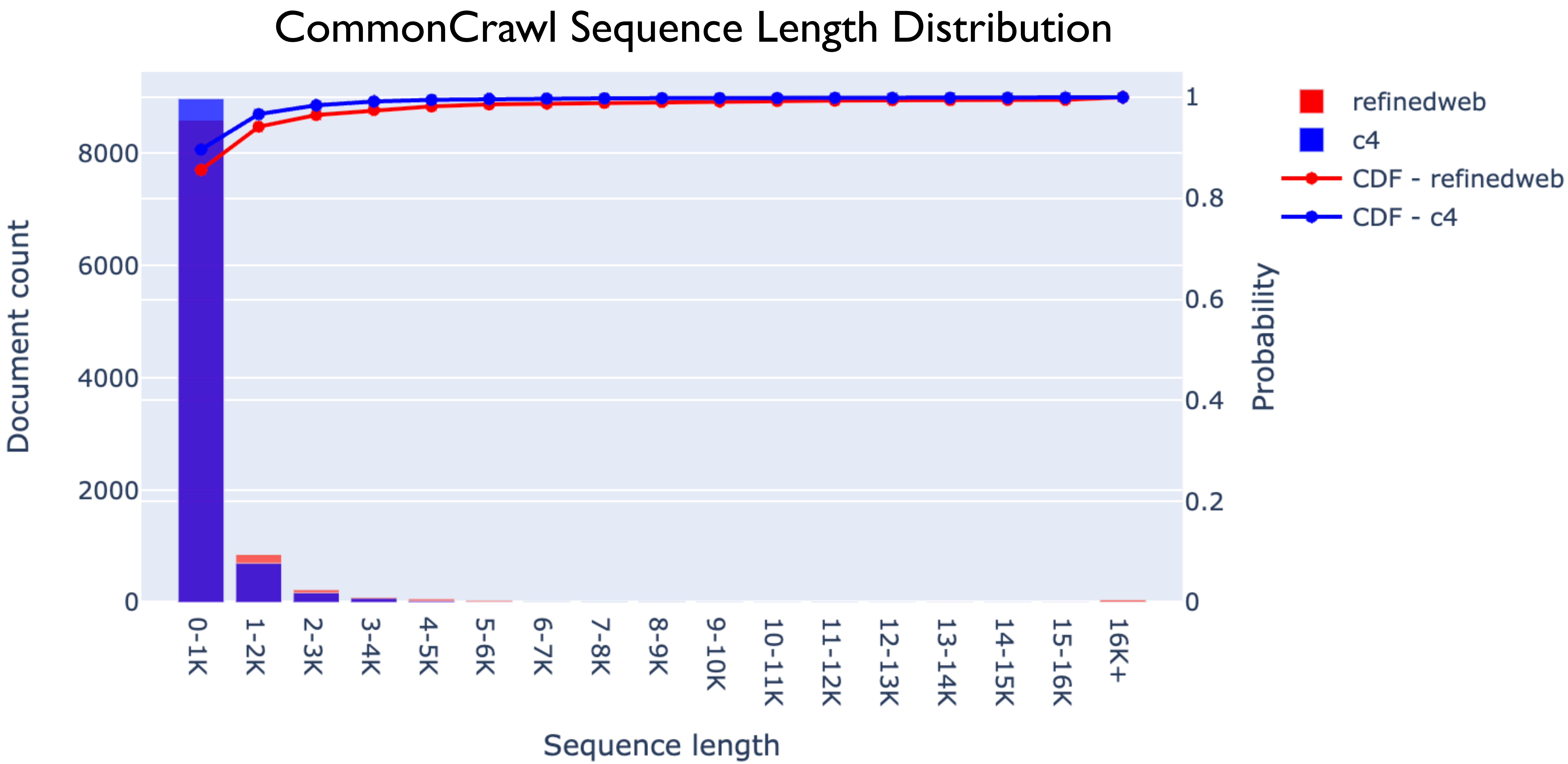
LMs



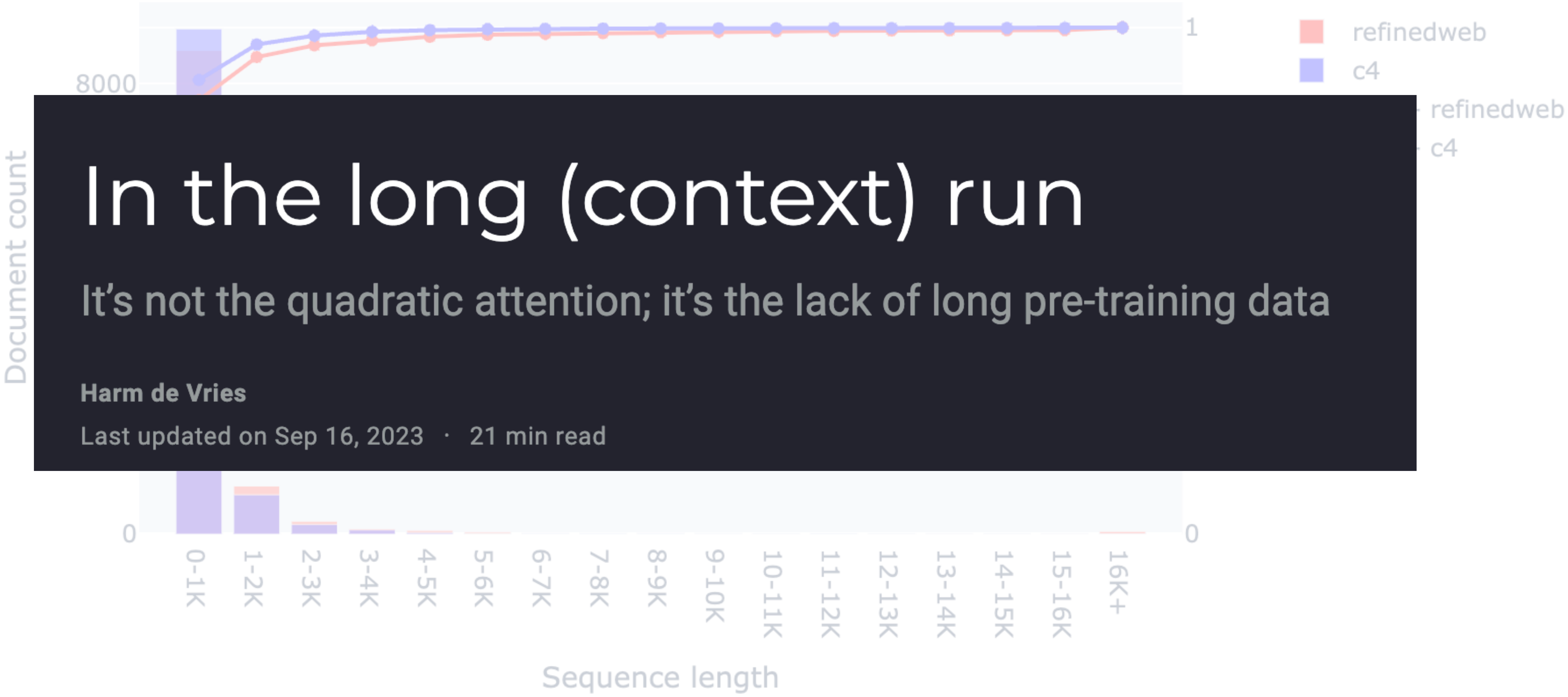
Lost in the Middle: How Language Models Use Long Contexts

Nelson F. Liu[1*] **Kevin Lin**[2] **John Hewitt**[1] **Ashwin Paranjape**[3]
Michele Bevilacqua[3] **Fabio Petroni**[3] **Percy Liang**[1]
[1]Stanford University [2]University of California, Berkeley [3]Samaya AI
nfliu@cs.stanford.edu

Problem: Lack of Long Pretraining Documents



Problem: Lack of Long Pretraining Documents



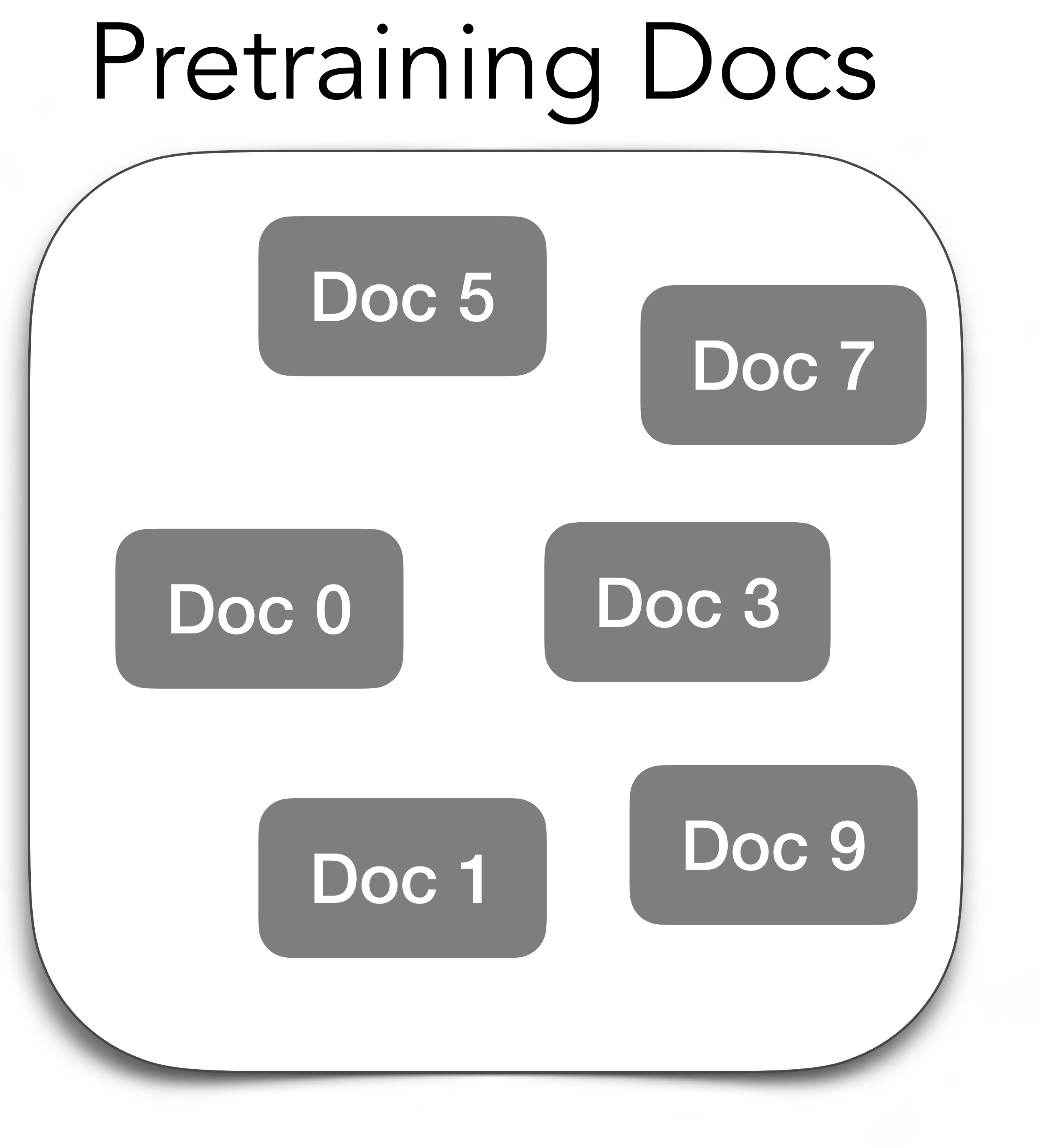
In the long (context) run

It's not the quadratic attention; it's the lack of long pre-training data

Harm de Vries

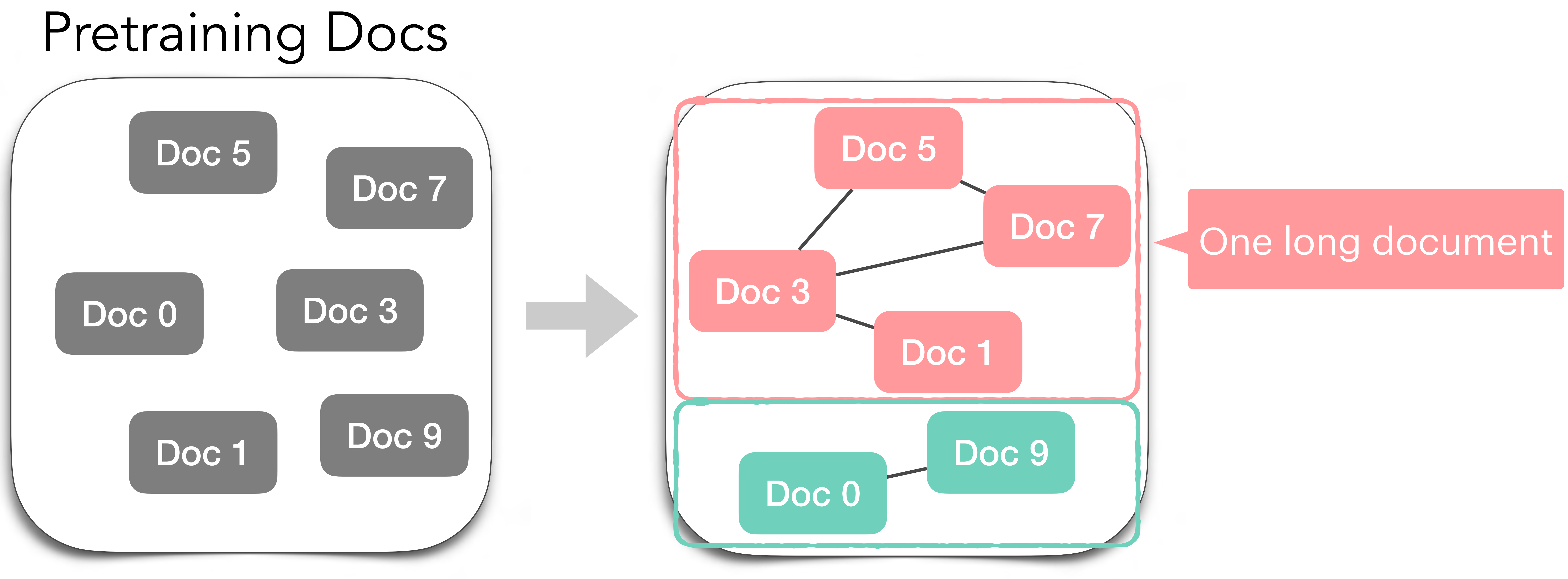
Last updated on Sep 16, 2023 · 21 min read

Reorder Data w/ Retriever



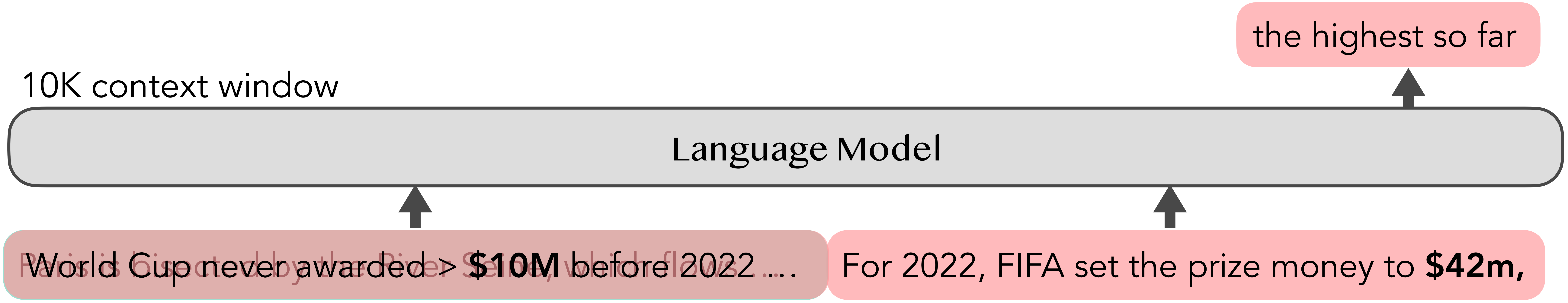
Shi, et al. ICLR 2024 Spotlight

Reorder Data w/ Retriever



Find Related Docs

Concatenate **Related** Documents



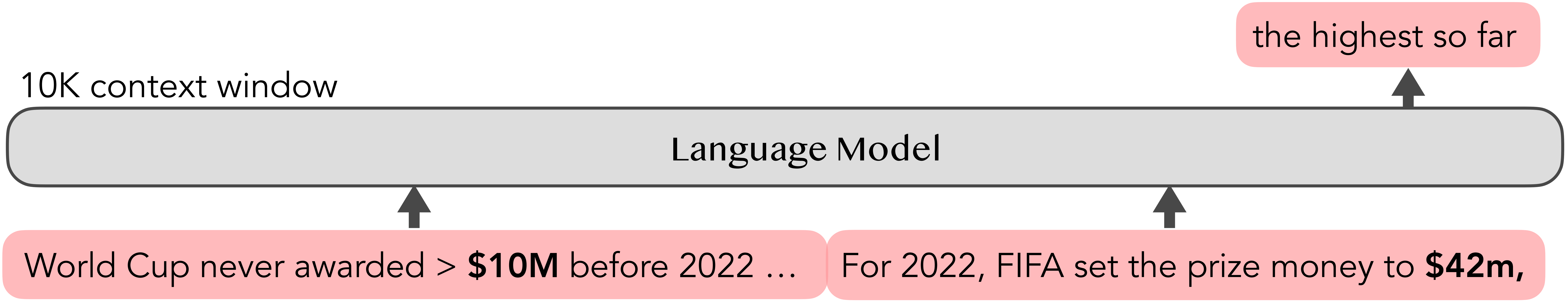
the highest so far
10K context window
Language Model
World Cup never awarded > **$10M** before 2022 ….
Paris is bisected by the River Seine, which flows ….
For 2022, FIFA set the prize money to **$42m,**

Doc

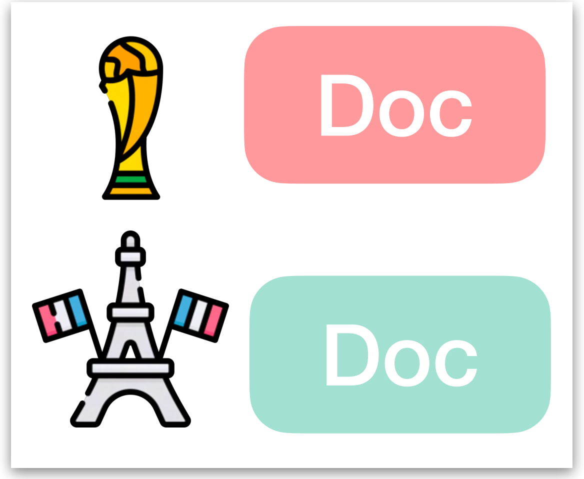
Doc

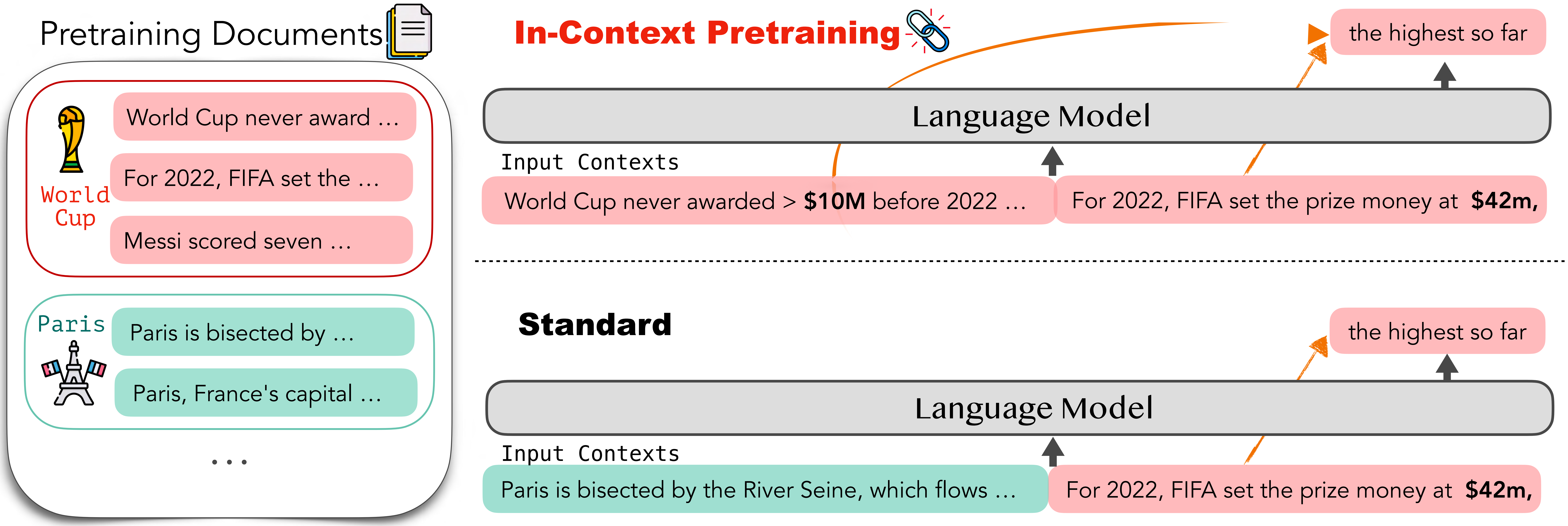
In-Context Pretraining

Concatenate **Related** Documents


the highest so far
10K context window
Language Model
World Cup never awarded > **$10M** before 2022 …
For 2022, FIFA set the prize money to **$42m,**

😊 Encourage LMs to reason across document boundaries





Pretraining Documents

World Cup

World Cup never award …

For 2022, FIFA set the …

Messi scored seven …

Paris

Paris is bisected by …

Paris, France's capital …

…

In-Context Pretraining

the highest so far

Language Model

Input Contexts

World Cup never awarded > **$10M** before 2022 …

For 2022, FIFA set the prize money at **$42m,**

Standard

the highest so far

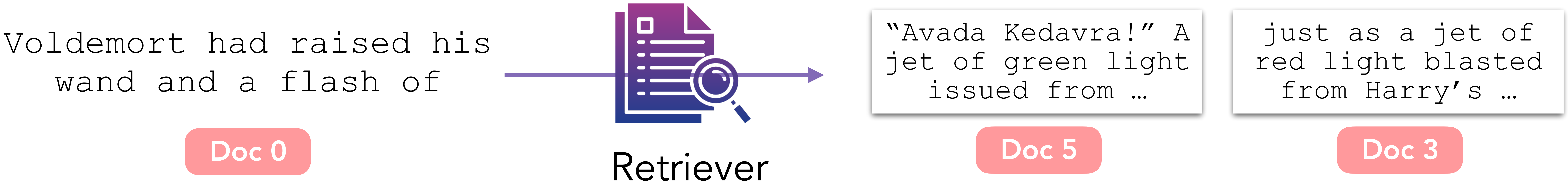
Language Model

Input Contexts

Paris is bisected by the River Seine, which flows …

For 2022, FIFA set the prize money at **$42m,**

In-Context Pretraining: Recipe



`Voldemort had raised his wand and a flash of`

Doc 0

Retriever

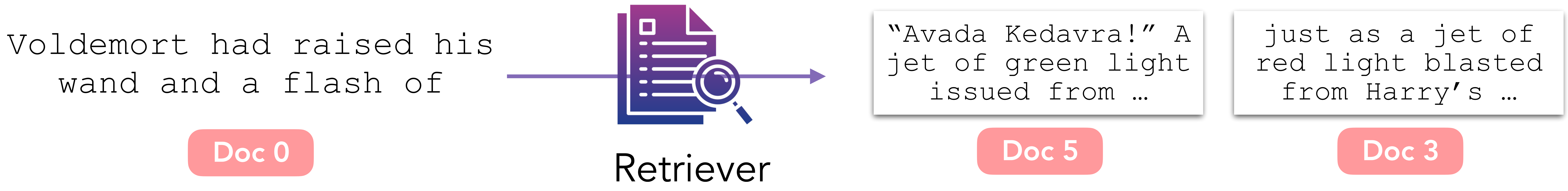
`“Avada Kedavra!” A jet of green light issued from …`

Doc 5

`just as a jet of red light blasted from Harry’s …`

Doc 3

In-Context Pretraining: Recipe



Voldemort had raised his wand and a flash of

Doc 0

Retriever

"Avada Kedavra!" A jet of green light issued from …

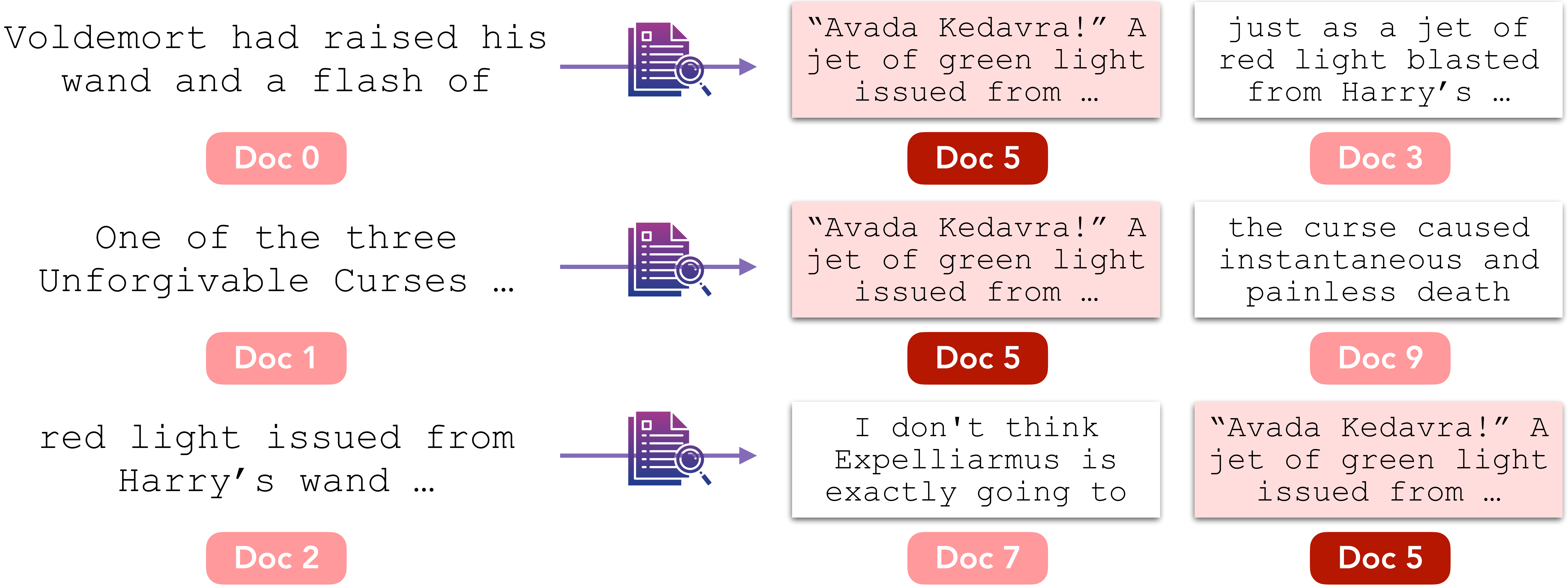
Doc 5

just as a jet of red light blasted from Harry's …

Doc 3

For each doc, can we directly include its related docs in the context?

In-Context Pretraining: Recipe



Voldemort had raised his wand and a flash of

Doc 0

“Avada Kedavra!” A jet of green light issued from …

Doc 5

just as a jet of red light blasted from Harry’s …

Doc 3

One of the three Unforgivable Curses …

Doc 1

“Avada Kedavra!” A jet of green light issued from …

Doc 5

the curse caused instantaneous and painless death

Doc 9

red light issued from Harry’s wand …

Doc 2

I don't think Expelliarmus is exactly going to

Doc 7

“Avada Kedavra!” A jet of green light issued from …

Doc 5

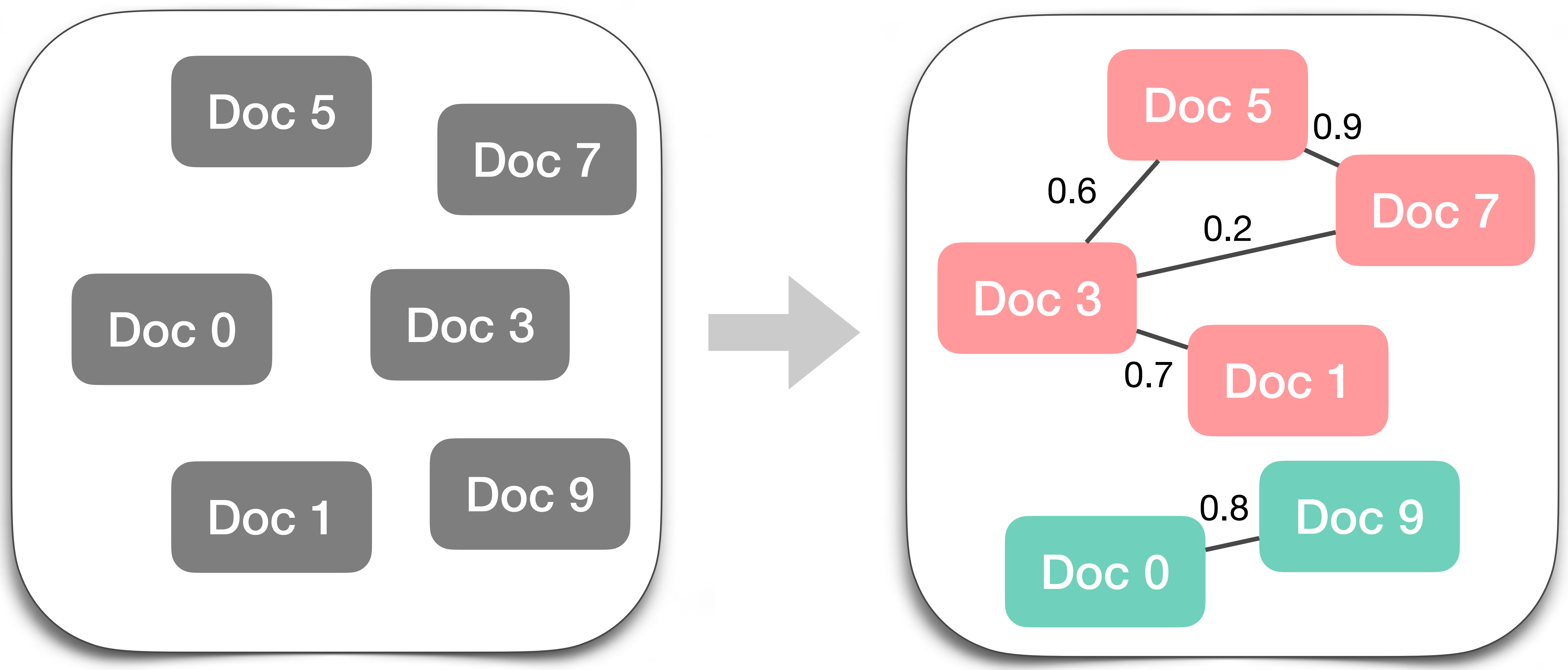
1) ***Related*** documents in the same context

2) Each document appears ***exactly once***

In-Context Pretraining: Recipe

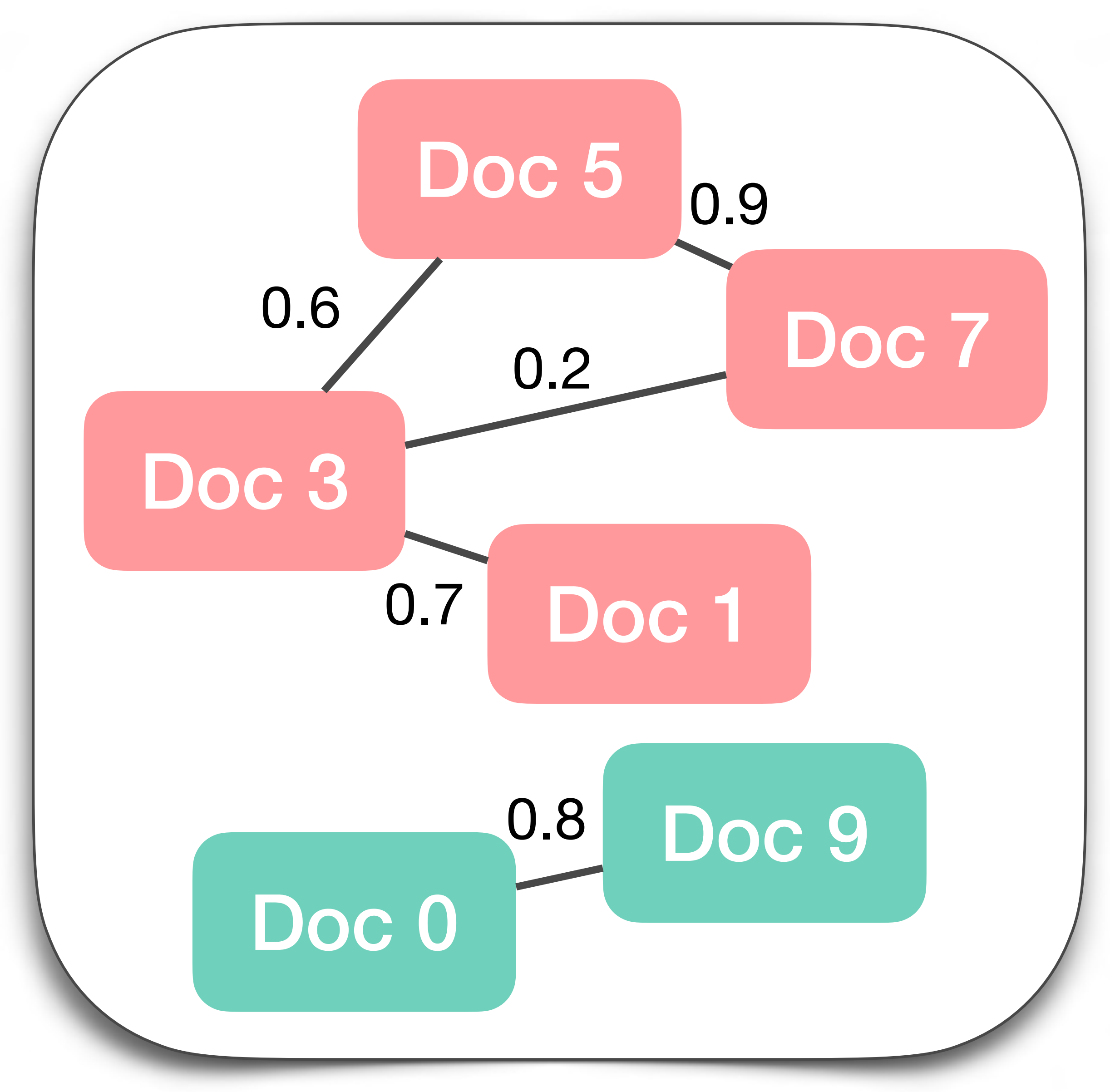
Document ordering problem

Pretraining Docs



In-Context Pretraining: Recipe

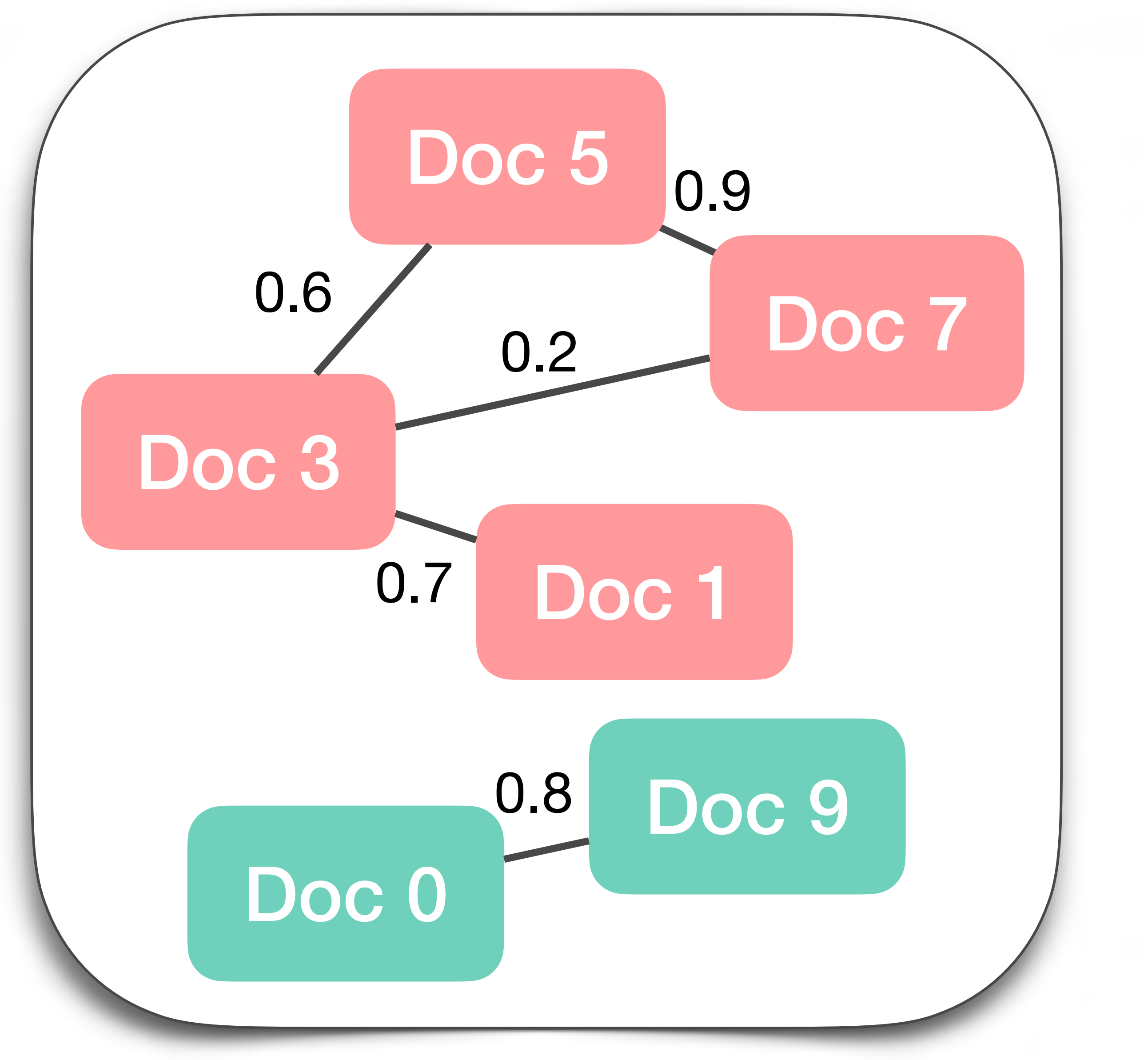
Document ordering problem



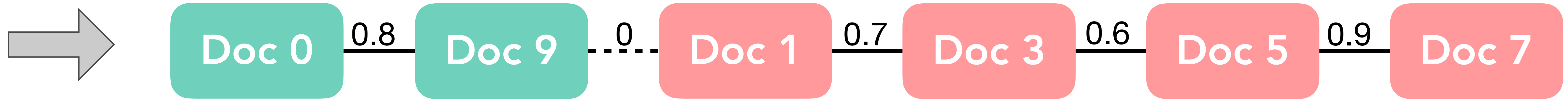
In-Context Pretraining: Recipe

Document ordering problem

Find a path that visits each doc once, making related docs to be visited consecutively

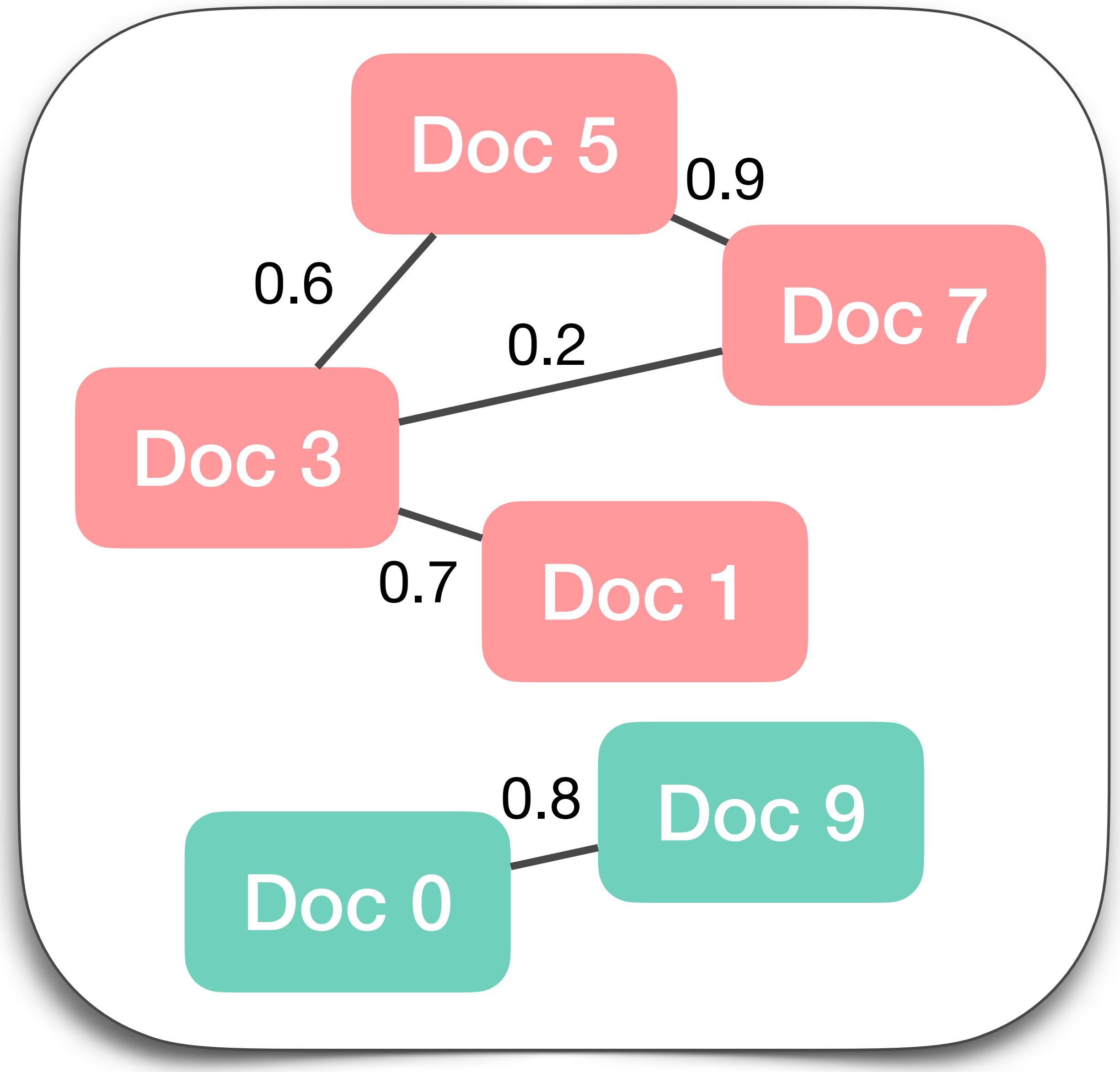


Maximum Traveling Salesman



Document Ordering Problem

Input:



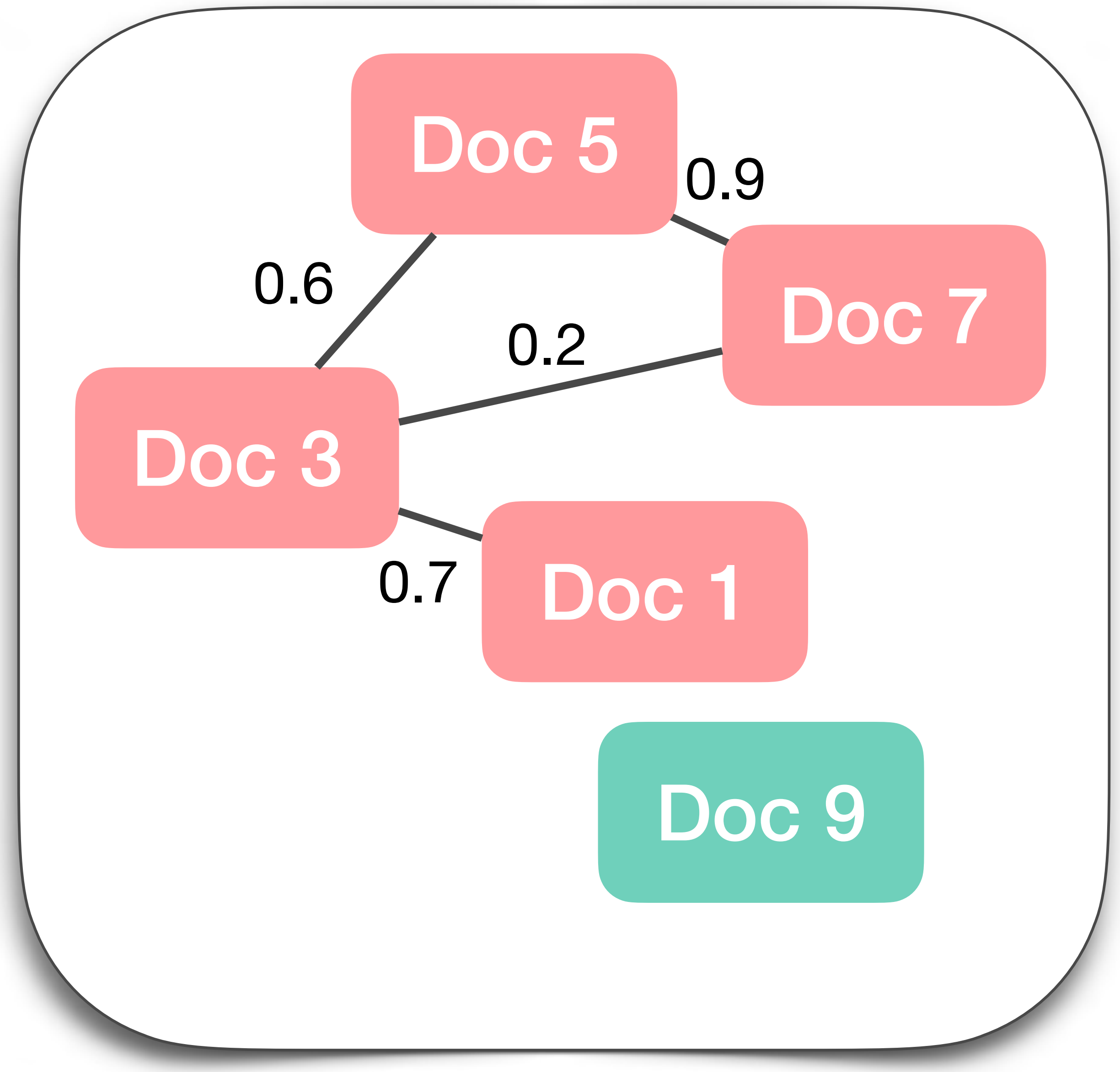
Output: `path`

Procedure:

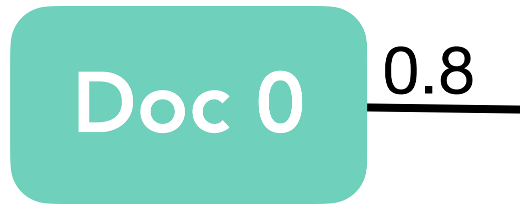
select an unvisited doc with the min degree

Document Ordering Problem

Input:



Output: `path`

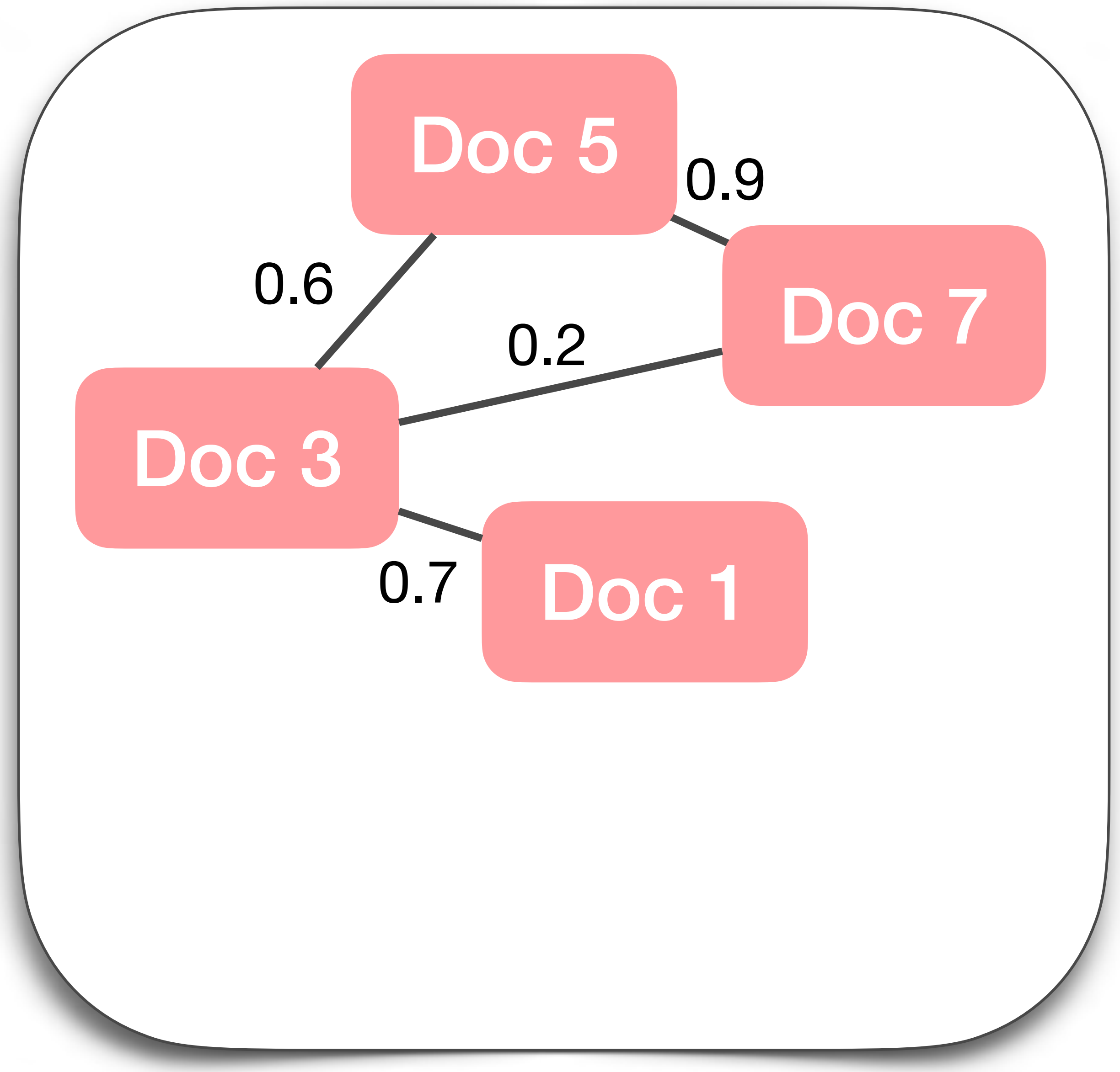


Procedure:

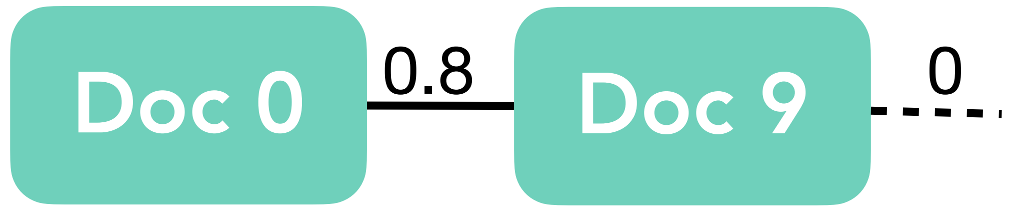
Move to the unvisited neighbor with max weight until all neighbors are visited

Document Ordering Problem

Input:



Output: `path`

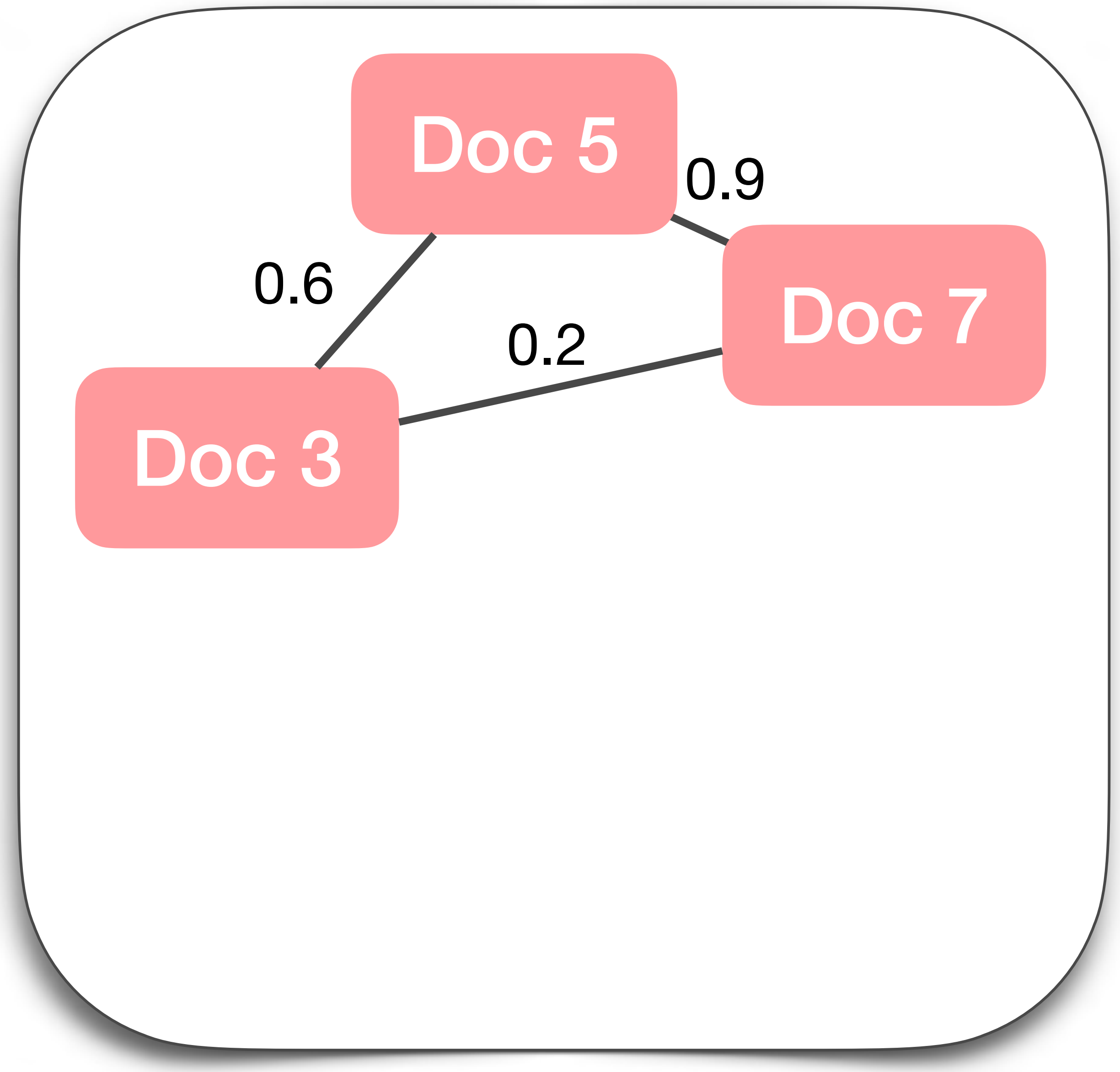


Procedure:

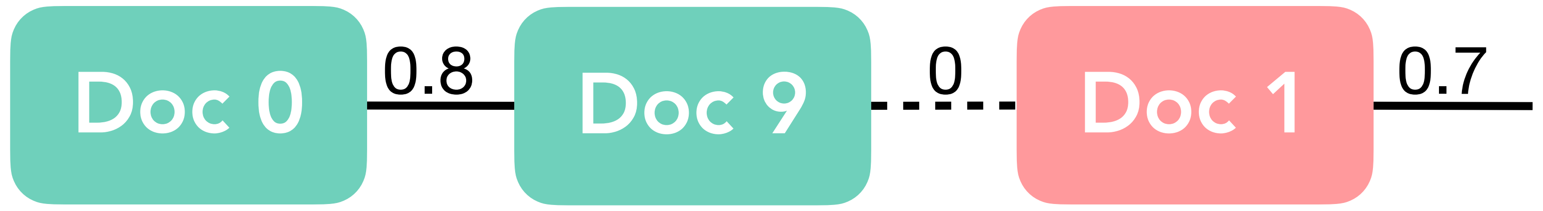
select an unvisited doc with the min degree

Document Ordering Problem

Input:



Output: `path`

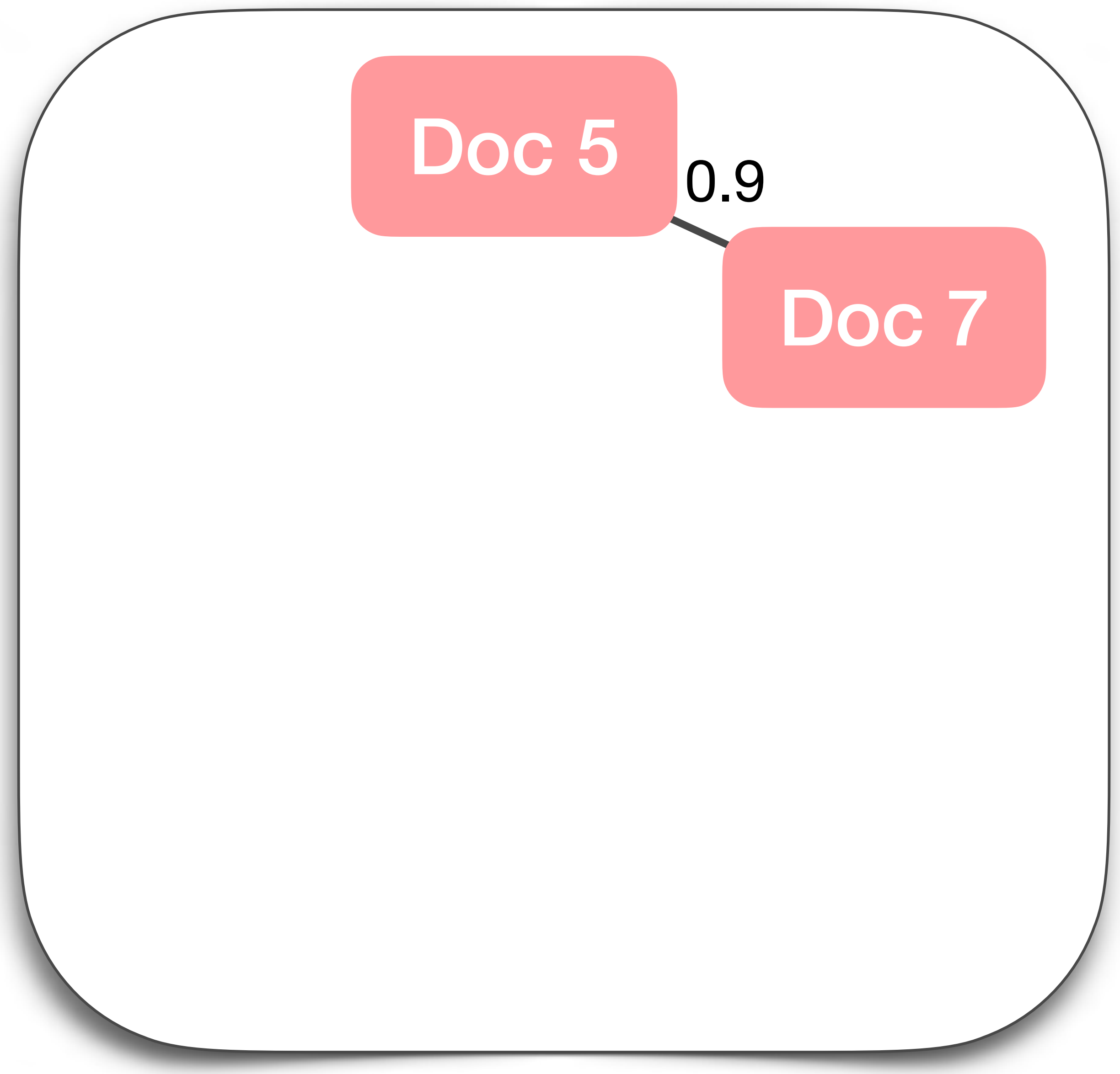


Procedure:

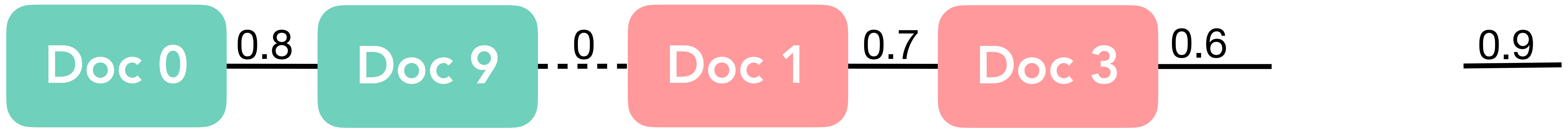
Move to the unvisited neighbor with max weight until all neighbors are visited

Document Ordering Problem

Input:



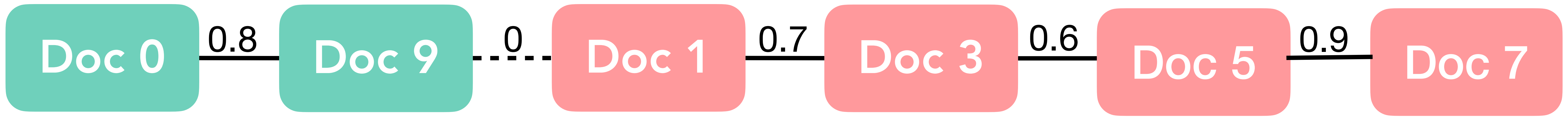
Output: `path`



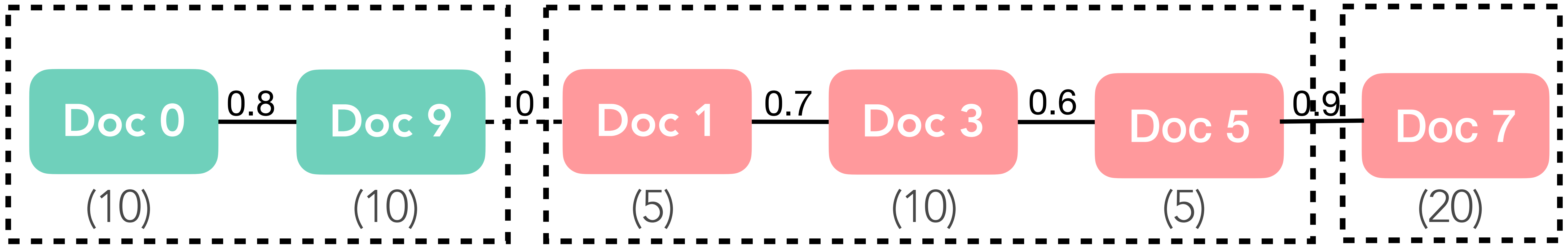
Procedure:

Move to the unvisited neighbor with max weight until all neighbors are visited

Document Ordering Problem

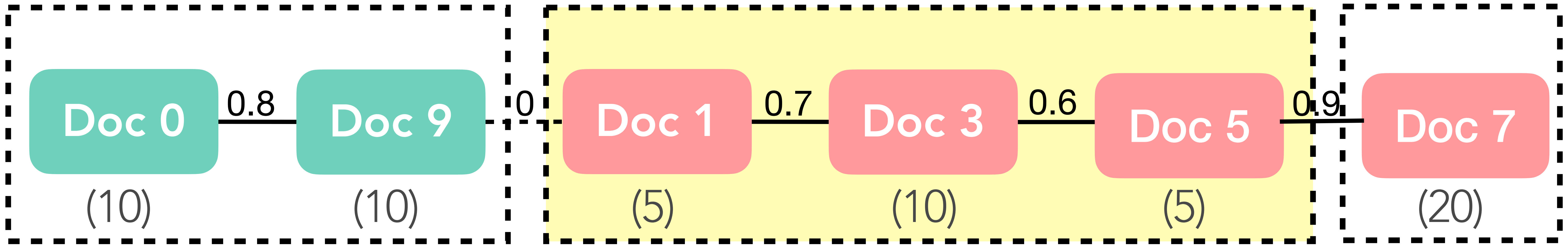


Document Ordering Problem

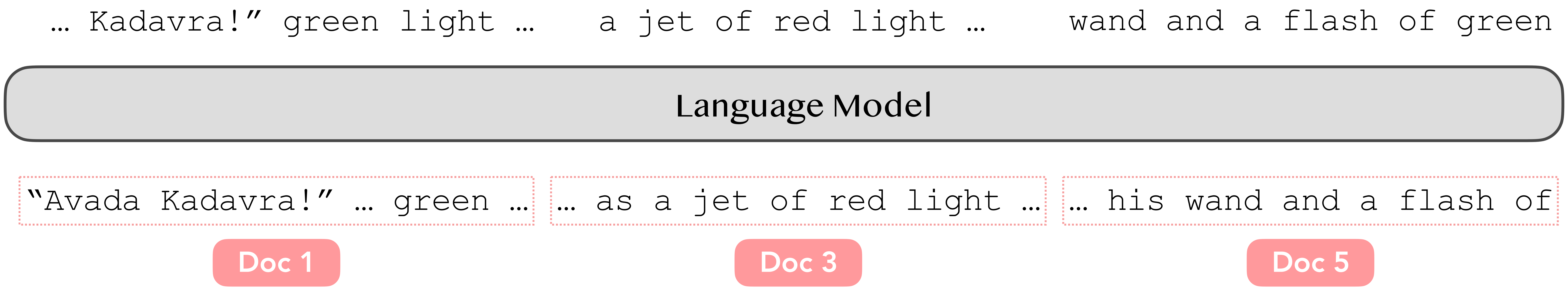


If max_seq_len = 20

Document Ordering Problem



If max_seq_len = 20


… Kadavra!” green light …
a jet of red light …
wand and a flash of green
Language Model
“Avada Kadavra!” … green …
… as a jet of red light …
… his wand and a flash of
Doc 1
Doc 3
Doc 5

1) ***Related*** documents in the same context

2) Each document appears ***exactly once***

Simple!
(Training code remains same)

Training Details

- **Architecture**: LLaMA
- **Model**: 0.3, 0.7, 1.5, and **7B** model with sequence length of 8192 from scratch **(128 A100s for 9 days)**
- **Data**: 306B tokens from Common Crawl (235M docs)
- **Retriever**: Contriever

Baselines

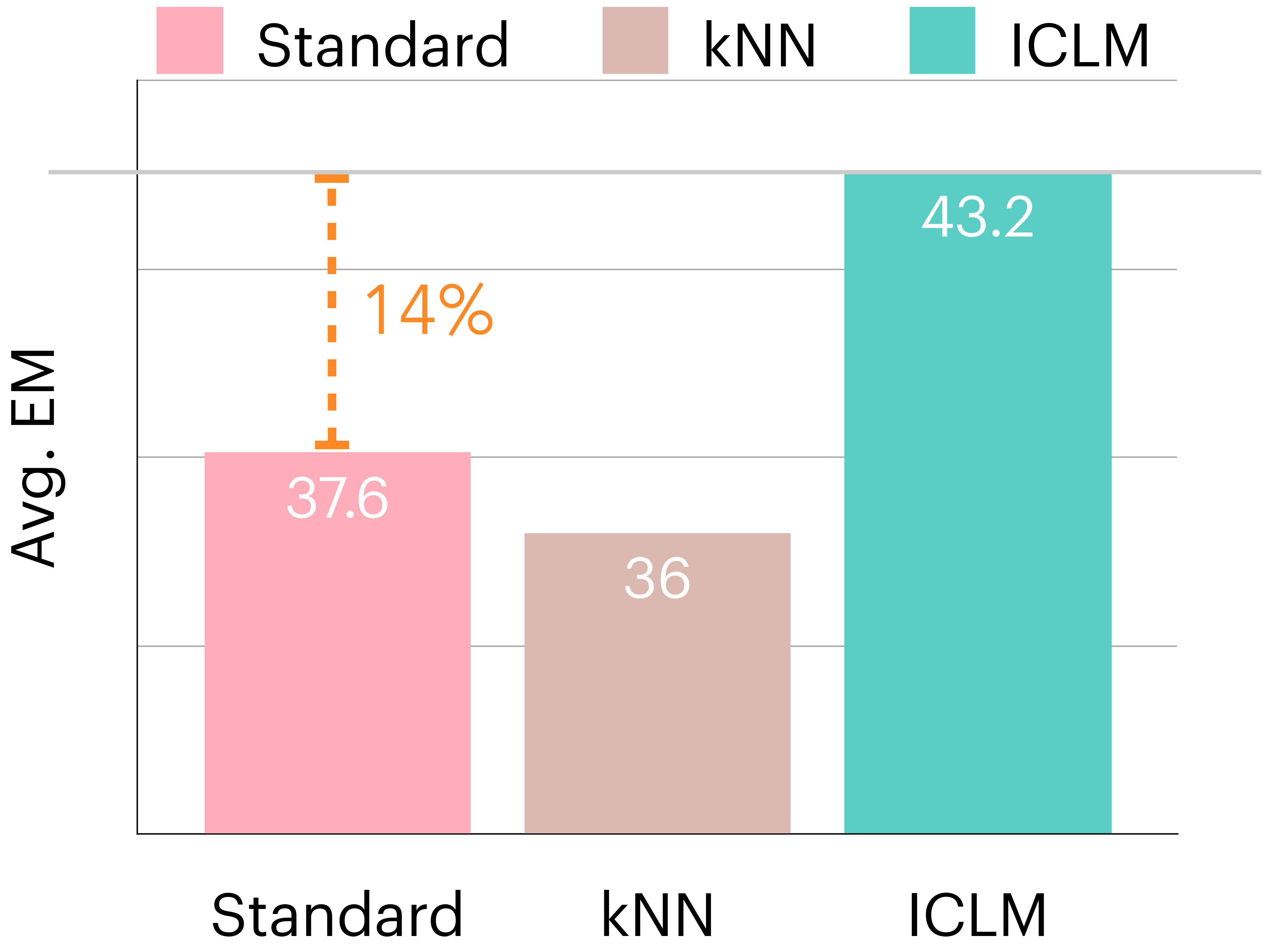
- **Standard:** places random docs in the input contexts
- **kNN:** places each doc and its retrieved top-k docs in the input

Given the same number of training steps, kNN exposes LMs to a less diverse set of documents, since documents can repeat

Results: Reading Comprehension

Tasks:

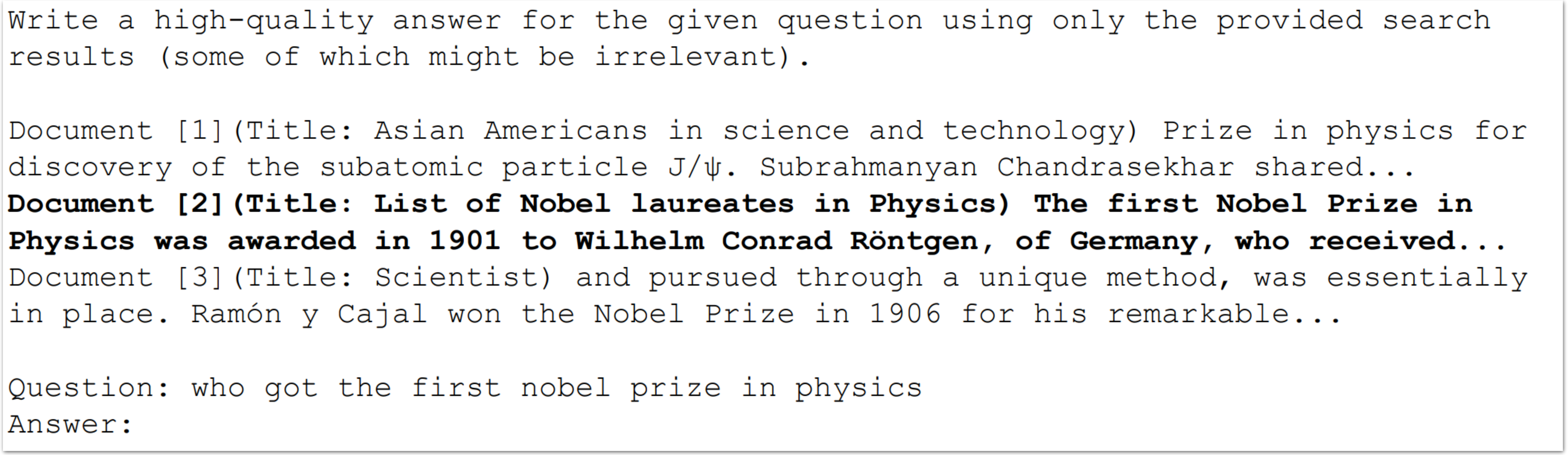
1. **Single document:** race-high, race-middle, boolq, squad
2. **Multi document:** hotpotQA, drop



Results: Open-Domain QA

Tasks: NQ, TQA

With retrieved docs

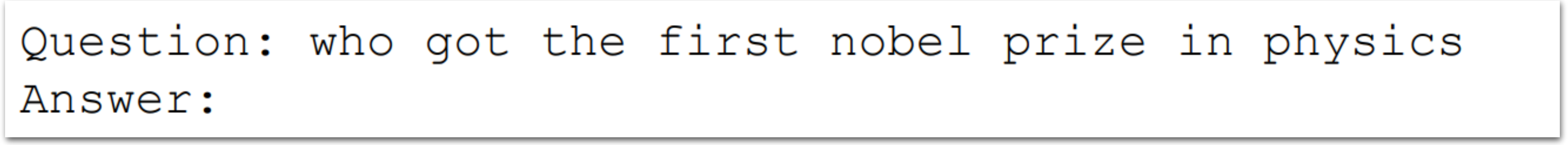


Write a high-quality answer for the given question using only the provided search results (some of which might be irrelevant).

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

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

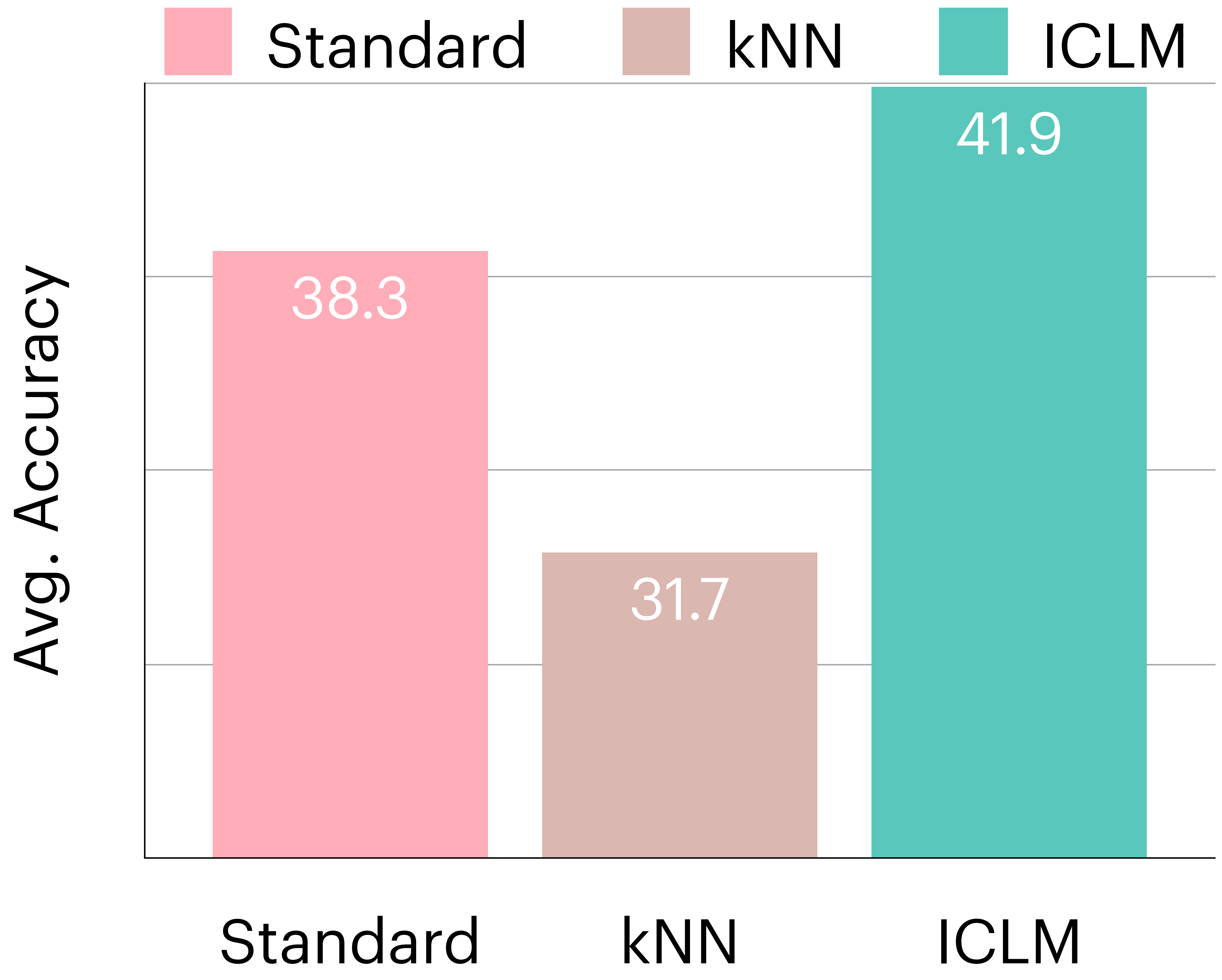
W/o retrieved docs



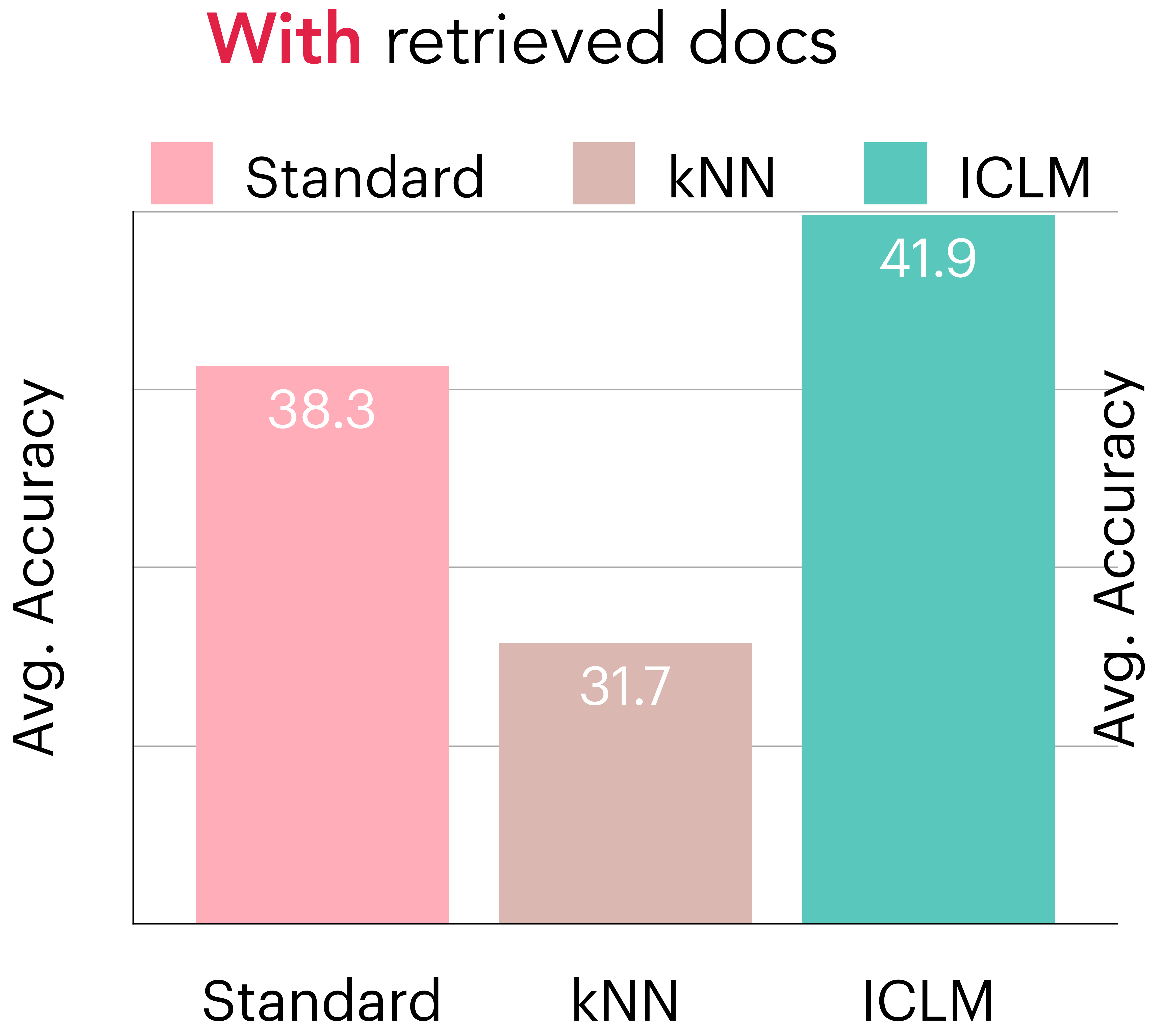
Question: who got the first nobel prize in physics
Answer:

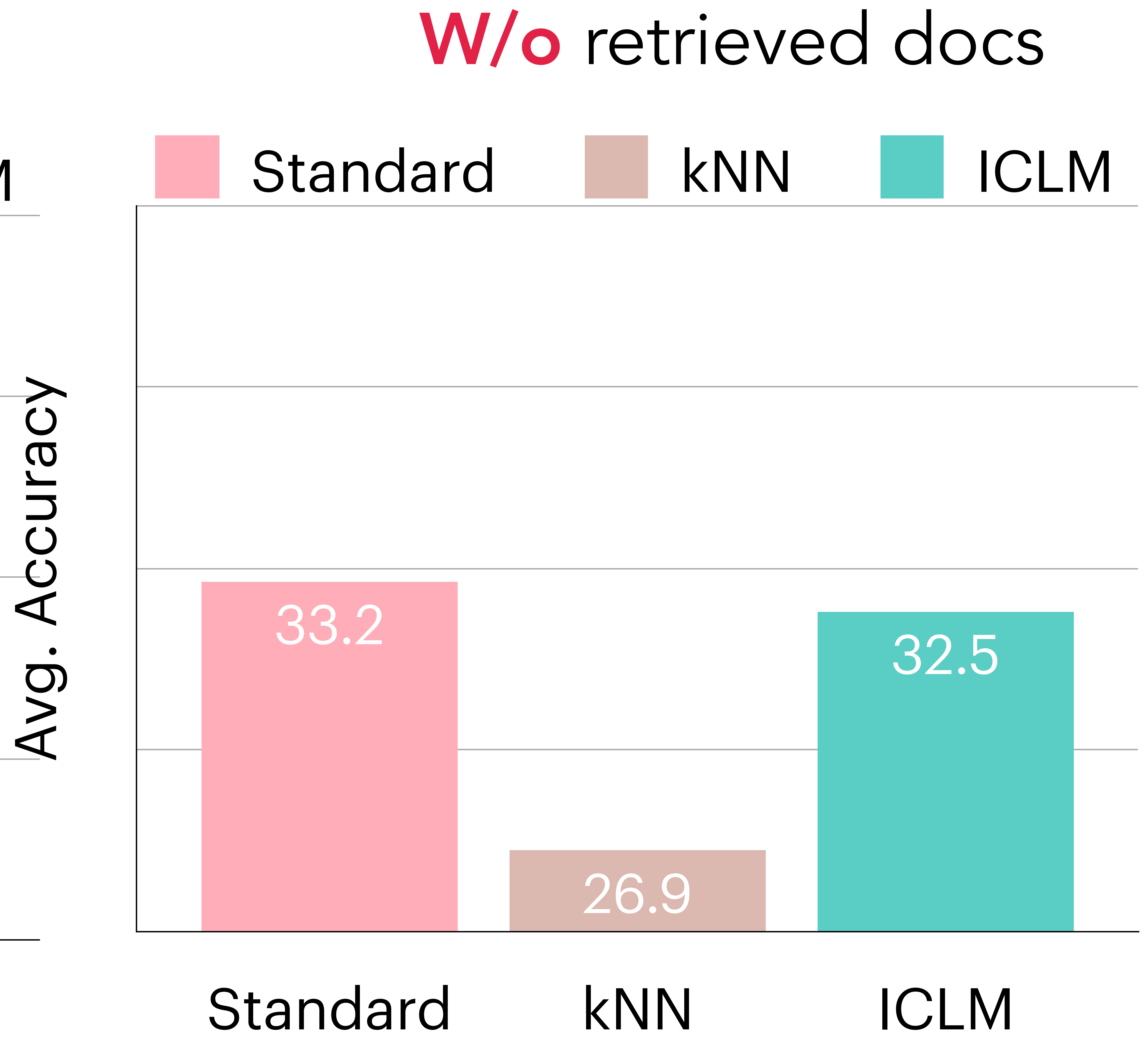
Results: Open-Domain QA

With retrieved docs



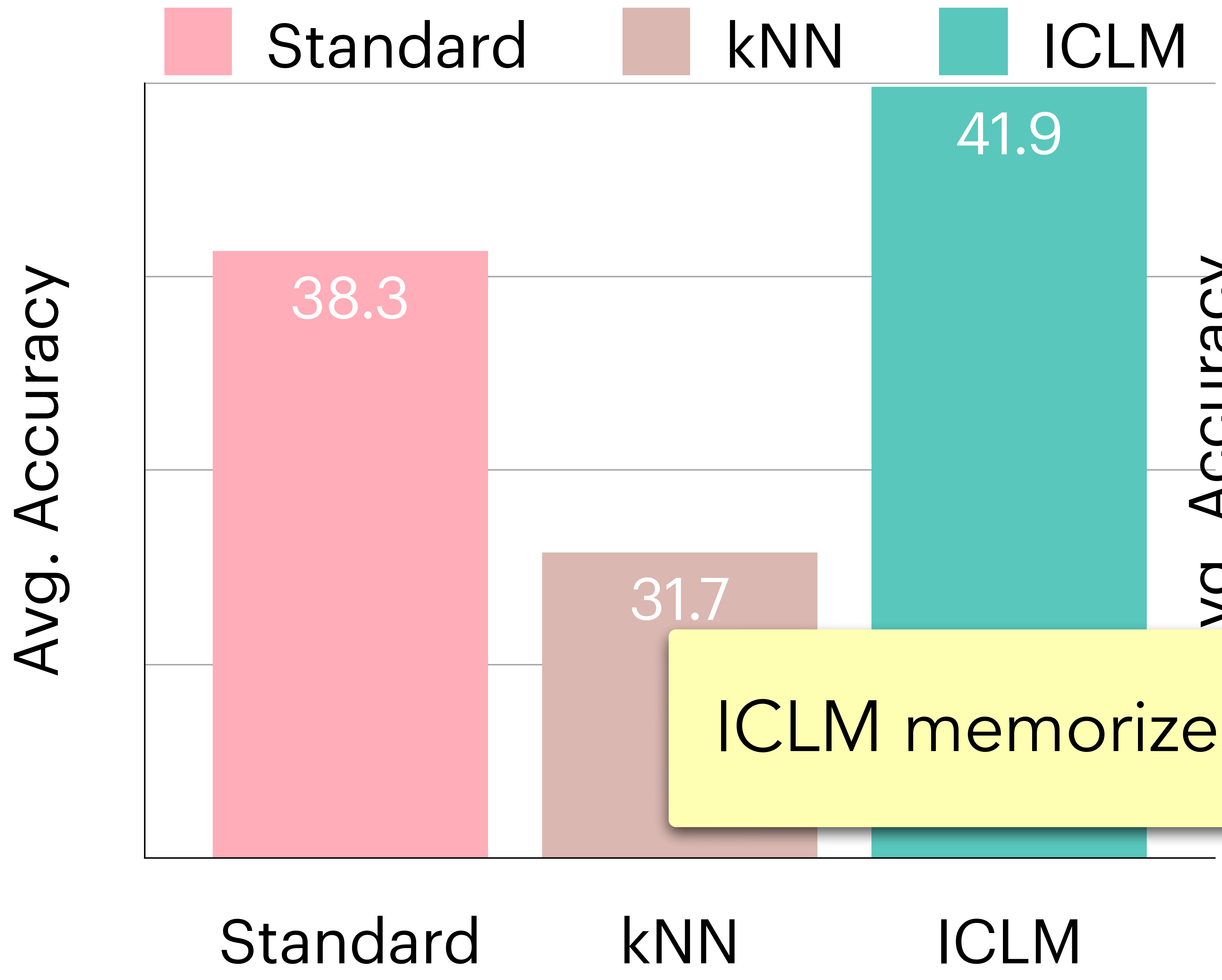
Results: Open-Domain QA



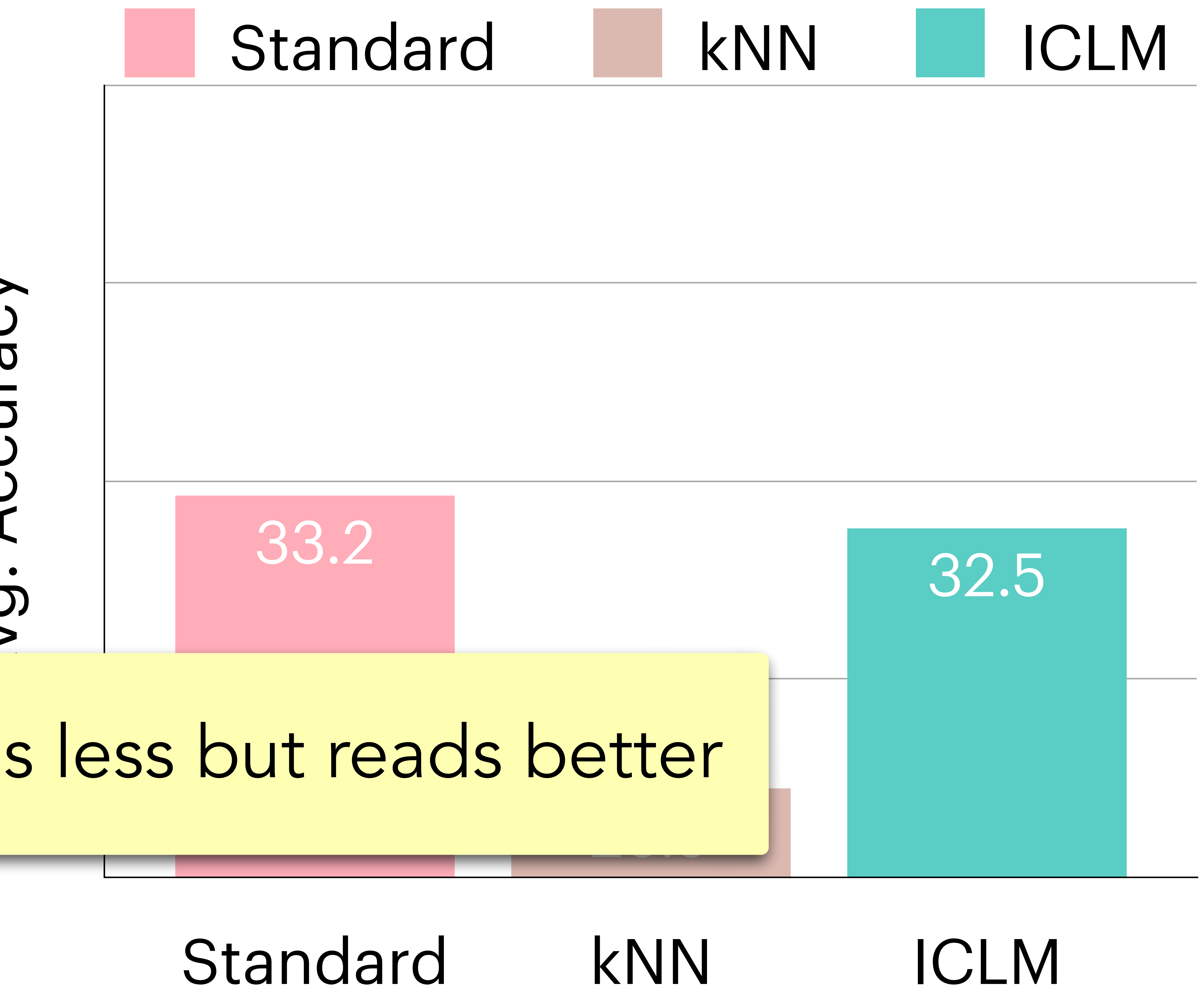


Results: Open-Domain QA

With retrieved docs



W/o retrieved docs



ICLM memorizes less but reads better

Results

23 benchmarks in total

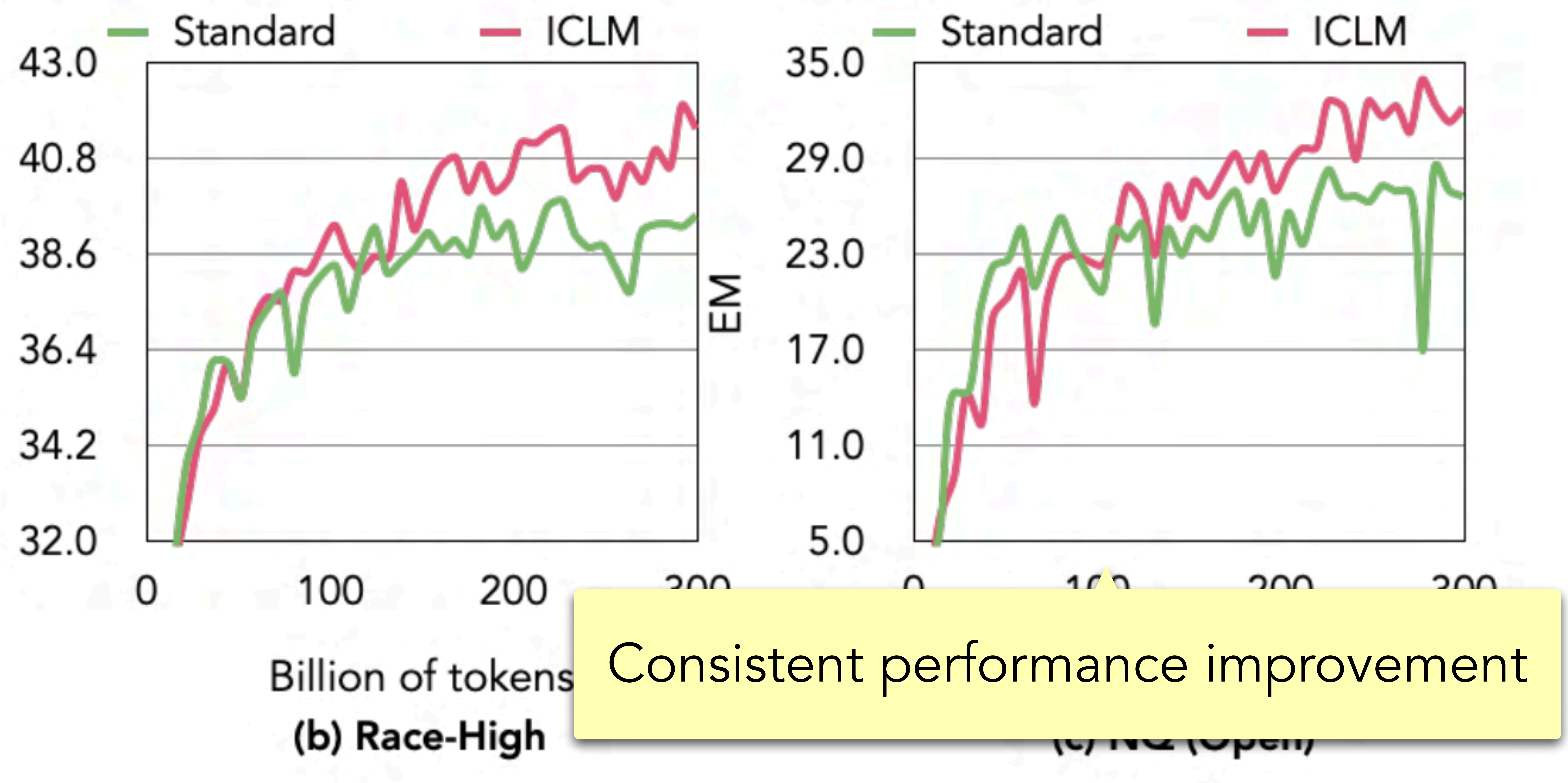
Accuracy 👑

	Standard		*In-Context Pretraining*
Open-Domain QA (w/ retrieval)	38%	**10.5%** →	42%
In-Context Learning	66%	**7.5%** →	71%
Reading Comprehension	37%	**14.0%** ▶	43%
Factuality	44%	**15.9%** ▶	51%
Long Document Reasoning	32%	**7.5%** ▶	34%

Evolution of Performance

Reading comprehension

Open-domain QA (w/ retrieval)



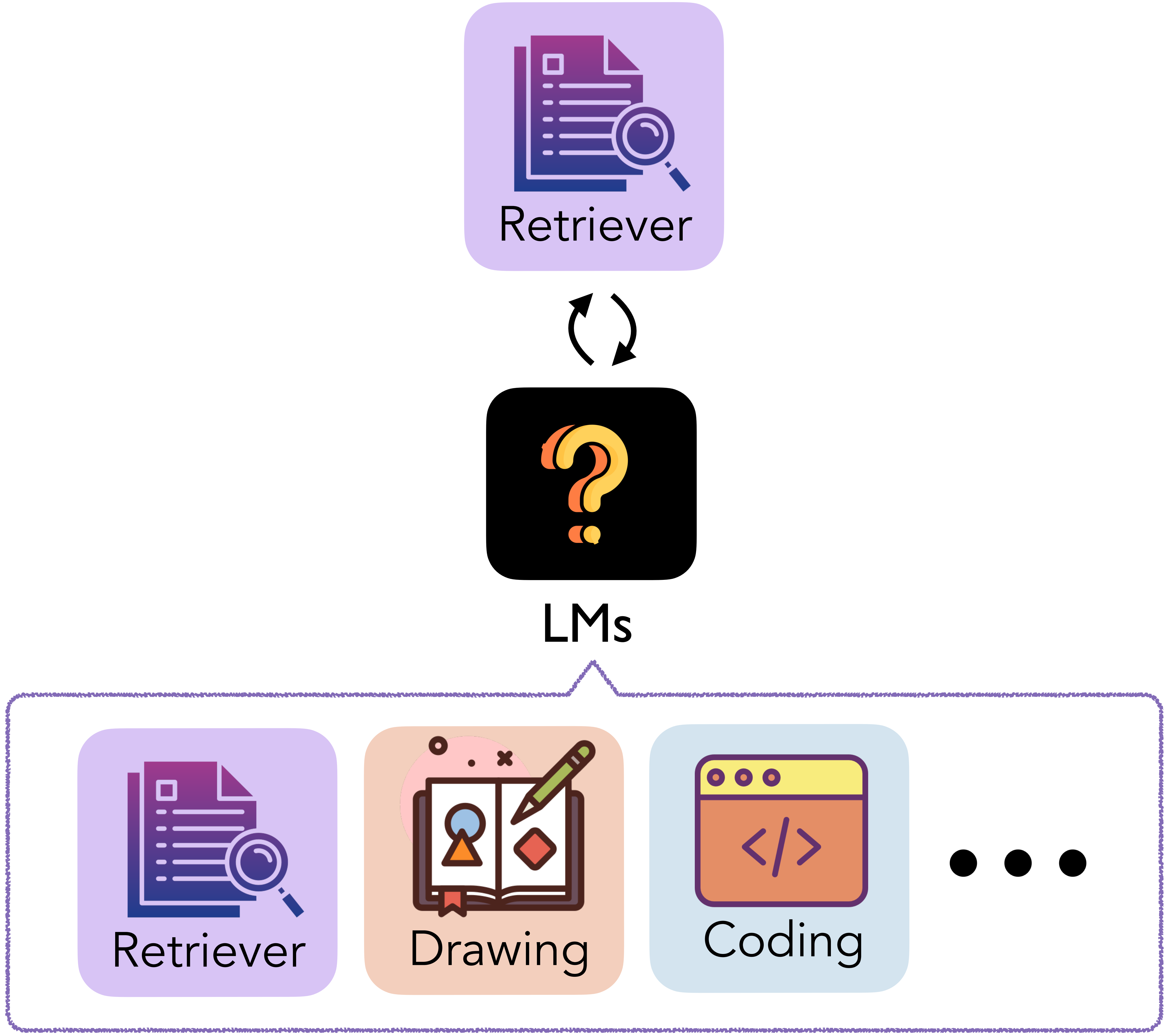
(b) Race-High

Consistent performance improvement

What’s Next?

- How similar should the documents be in the same context?
- Does In-Context Pretraining work for continual pretraining?

Augmented Models



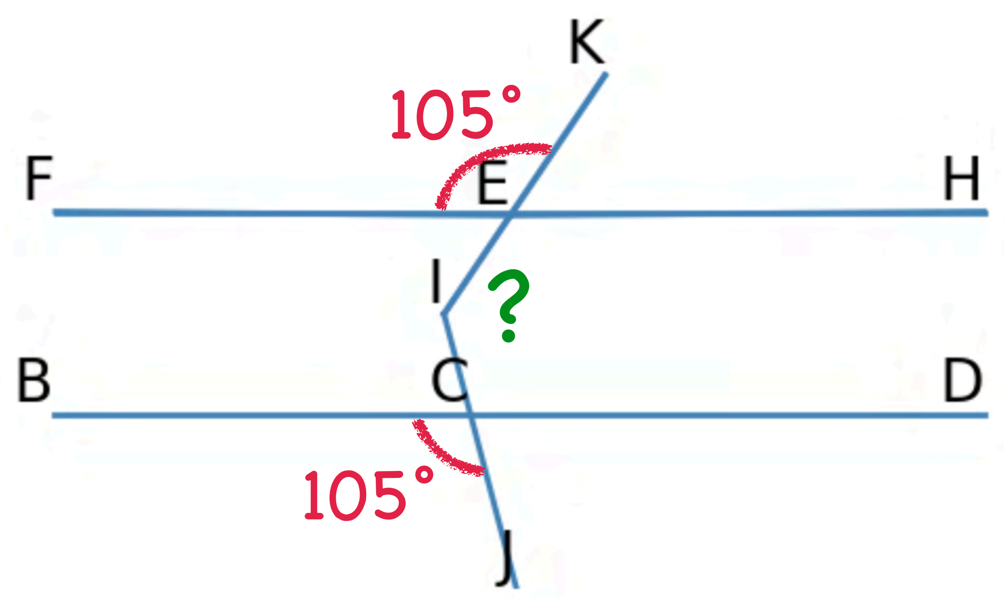
Pretraining

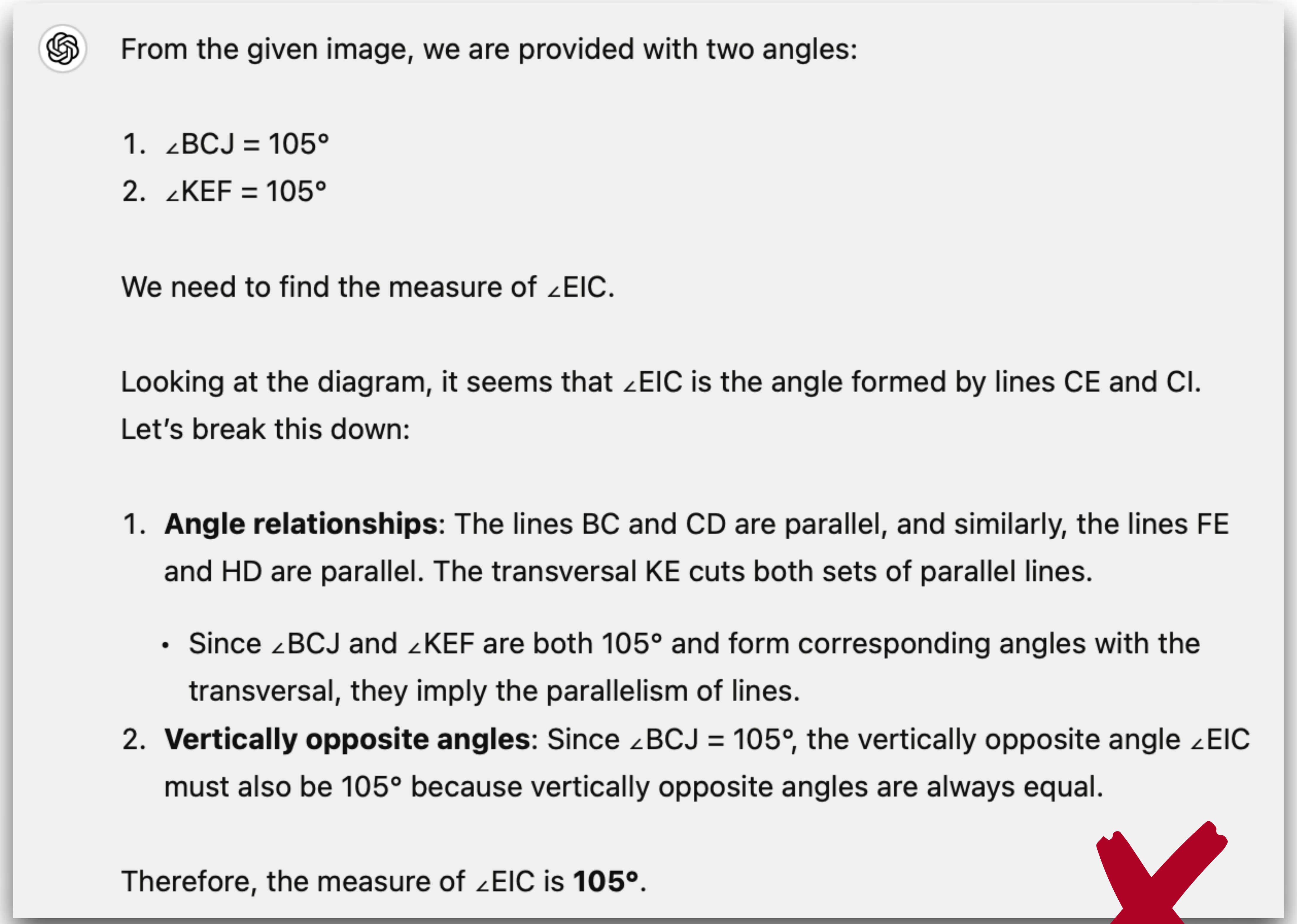
Shi, et al. ICLR 2024 Spotlight

Inference

Hu*, **Shi*** et al., NeurIPS 2024

Given ∠BCJ=105°, ∠KEF=105°. Find ∠EIC





From the given image, we are provided with two angles:

1. ∠BCJ = 105°
2. ∠KEF = 105°

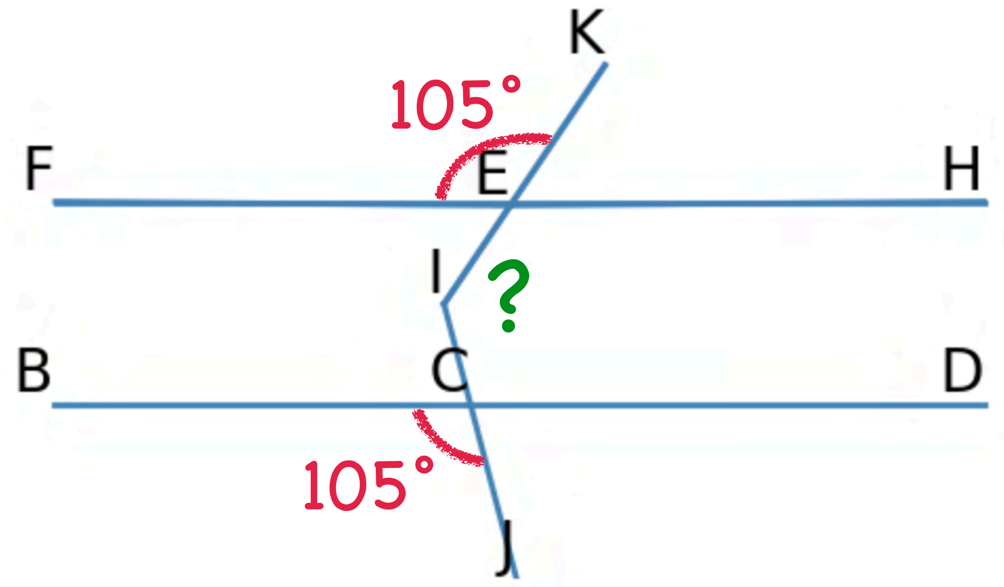
We need to find the measure of ∠EIC.

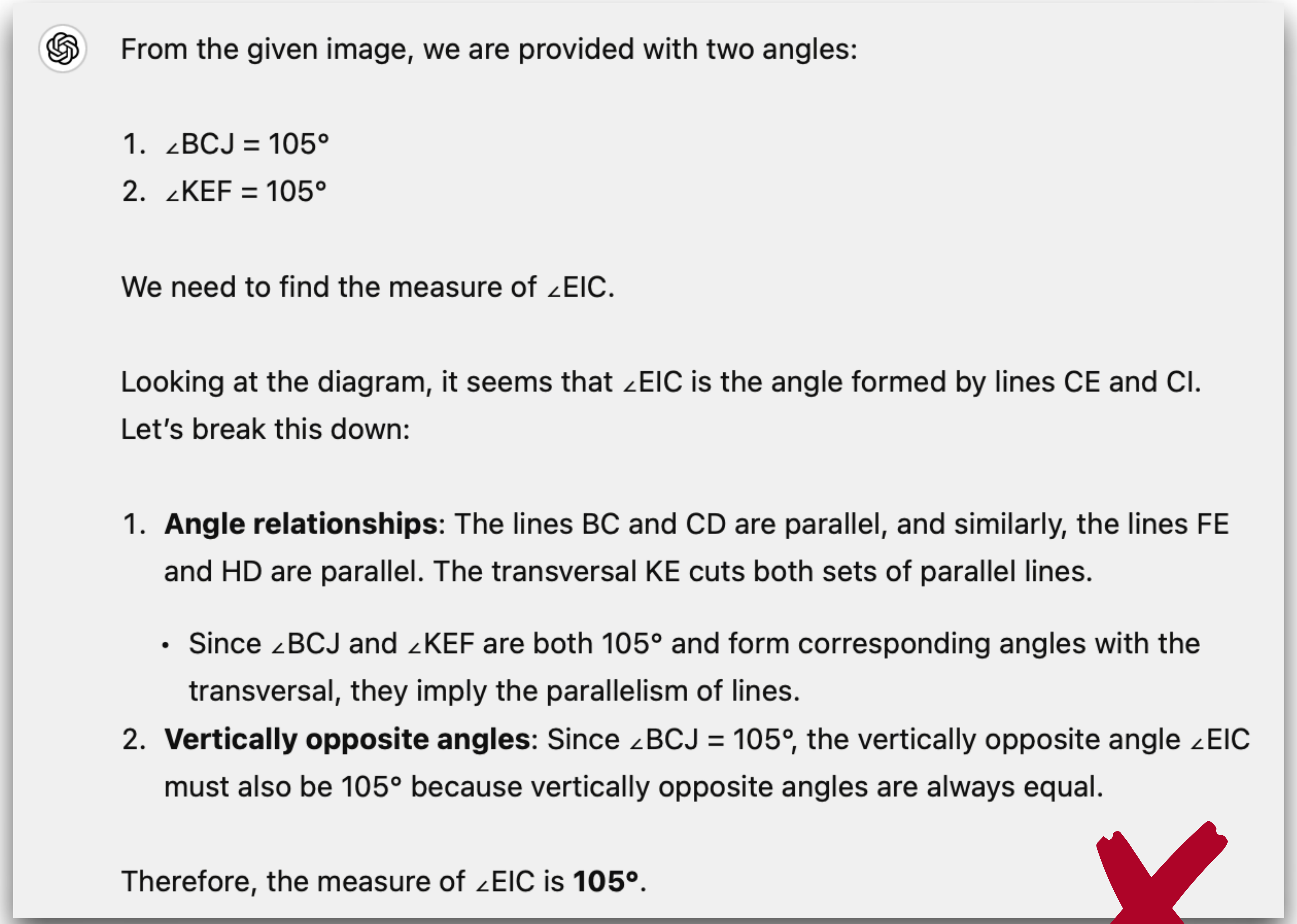
Looking at the diagram, it seems that ∠EIC is the angle formed by lines CE and CI. Let's break this down:

1. **Angle relationships**: The lines BC and CD are parallel, and similarly, the lines FE and HD are parallel. The transversal KE cuts both sets of parallel lines.
   - Since ∠BCJ and ∠KEF are both 105° and form corresponding angles with the transversal, they imply the parallelism of lines.
2. **Vertically opposite angles**: Since ∠BCJ = 105°, the vertically opposite angle ∠EIC must also be 105° because vertically opposite angles are always equal.

Therefore, the measure of ∠EIC is **105°**.

Given ∠BCJ=105°, ∠KEF=105°. Find ∠EIC





From the given image, we are provided with two angles:

1. ∠BCJ = 105°
2. ∠KEF = 105°

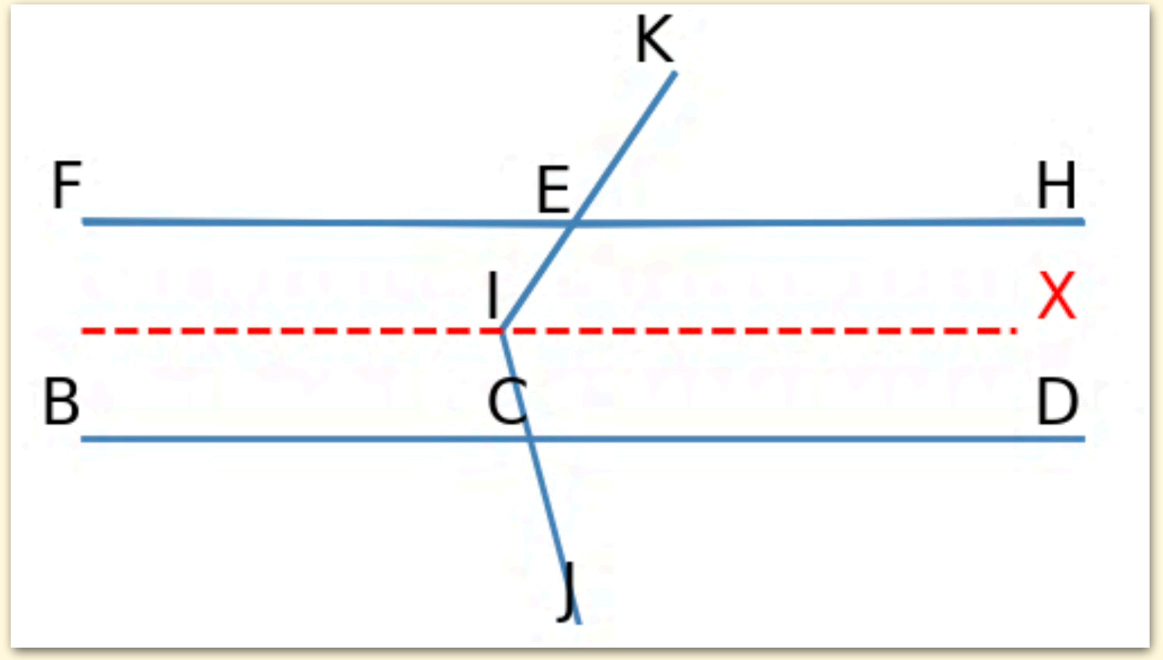
We need to find the measure of ∠EIC.

Looking at the diagram, it seems that ∠EIC is the angle formed by lines CE and CI. Let's break this down:

1. **Angle relationships**: The lines BC and CD are parallel, and similarly, the lines FE and HD are parallel. The transversal KE cuts both sets of parallel lines.
   - Since ∠BCJ and ∠KEF are both 105° and form corresponding angles with the transversal, they imply the parallelism of lines.
2. **Vertically opposite angles**: Since ∠BCJ = 105°, the vertically opposite angle ∠EIC must also be 105° because vertically opposite angles are always equal.

Therefore, the measure of ∠EIC is **105°**.

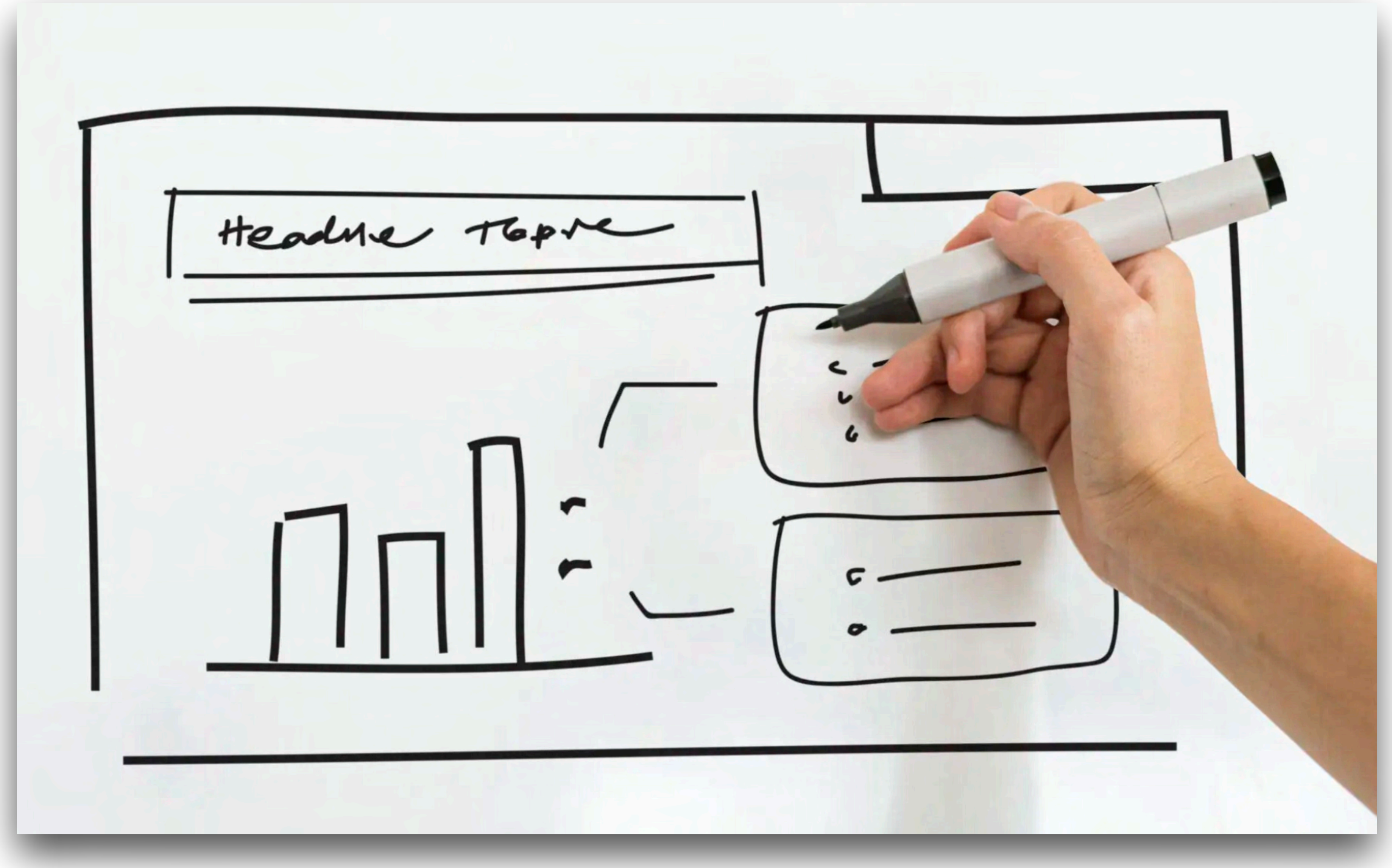
Draw line IX parallel to EH

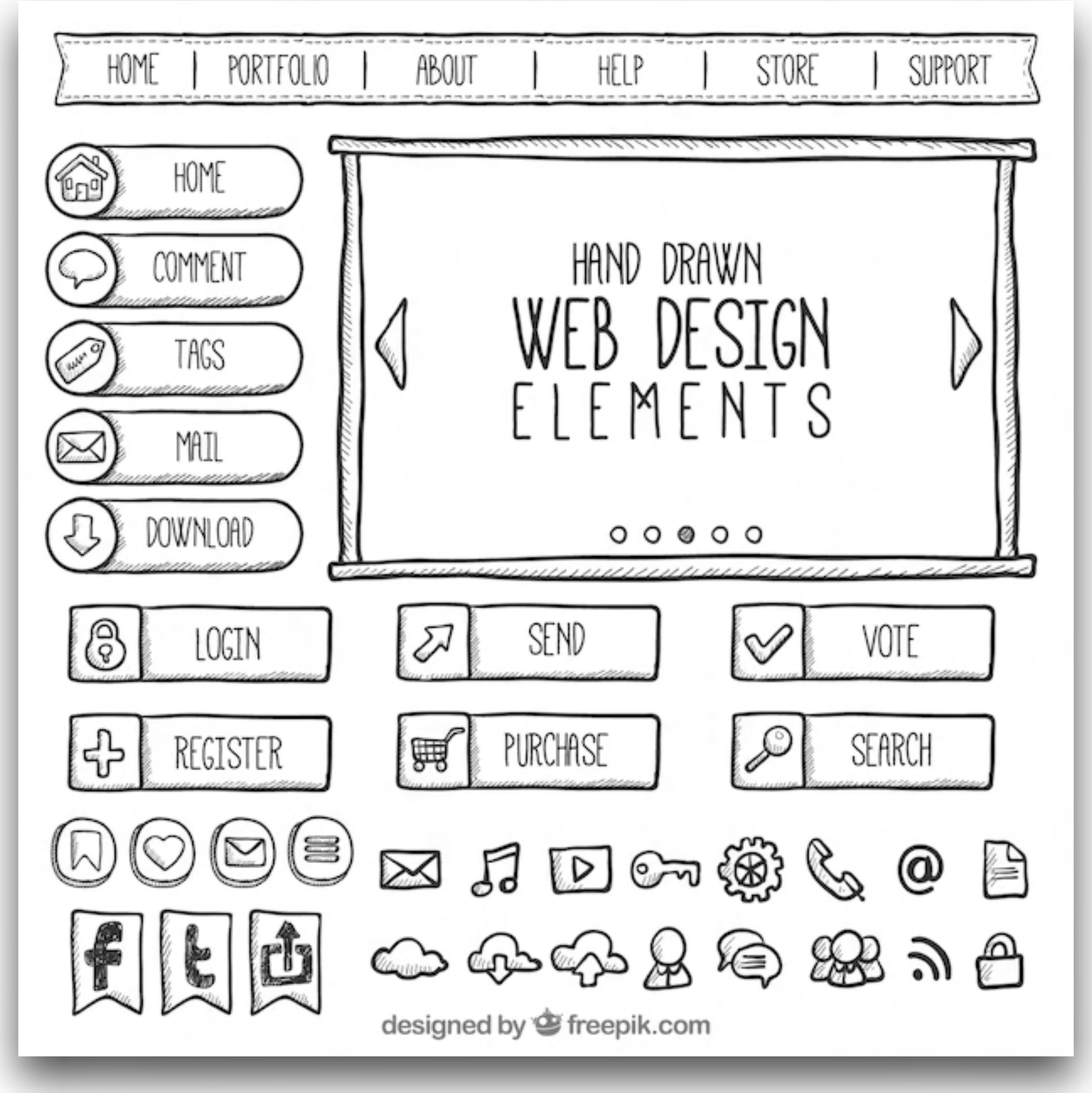


Given ∠BCJ = ∠KEF = 105°, their supplementary angles, ∠KEH = ∠DCJ = 180° - 105° = 75°. Given BD parallel to IX, it follows that ∠EIX = ∠KEH, and ∠XIC = ∠DCJ, both measuring 75°. Therefore, ∠EIC = ∠EIX + ∠XIC = 150°

Sketching ✍️

Sketching is a fundamental human activity, serving as a versatile tool for **communication**, **ideation**, and **problem-solving**





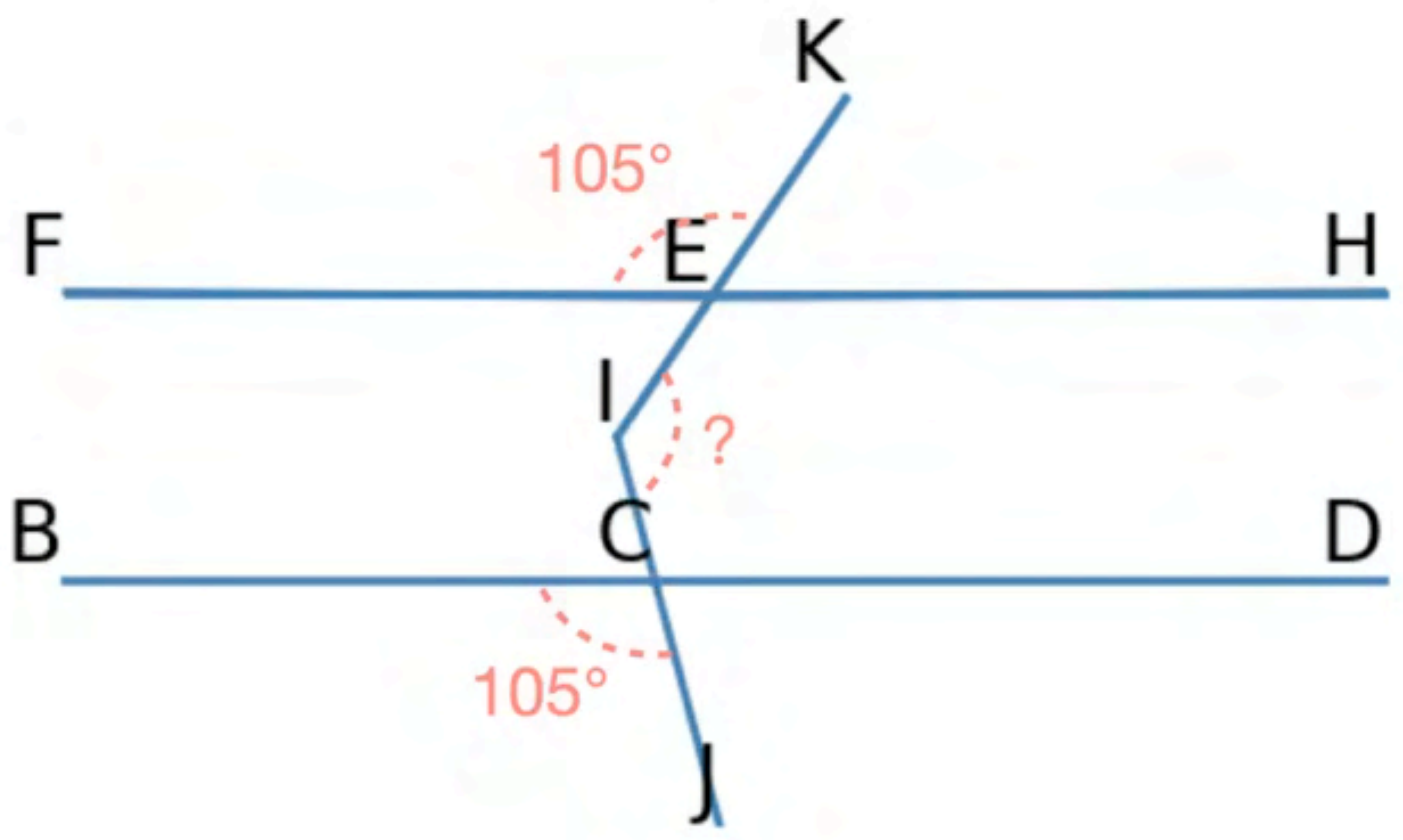
Why not let multimodal LMs do so?

Visual Reasoning w/ Sketchpad

Augmenting multimodal LMs with a visual canvas and drawing tools ✍️

Query: Given ∠BCJ=105°, ∠KEF=105°. Find ∠EIC

Input Image:



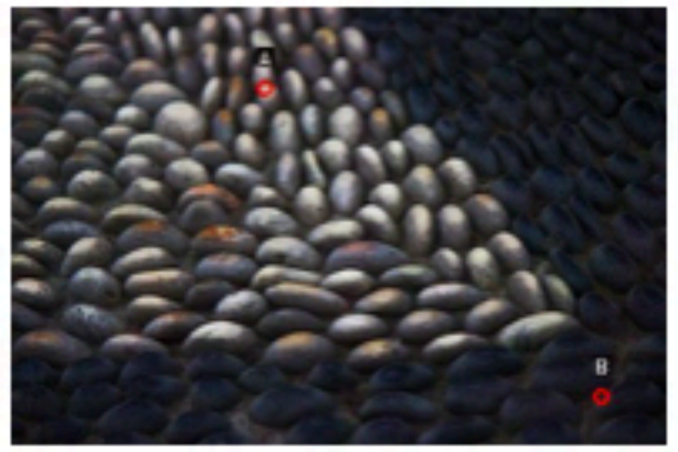
Visual Sketchpad: Recipe

Sketching via Code Generation

Visual Sketchpad: Recipe

Sketching via Code Generation

Which point is closer to the camera, A or B?



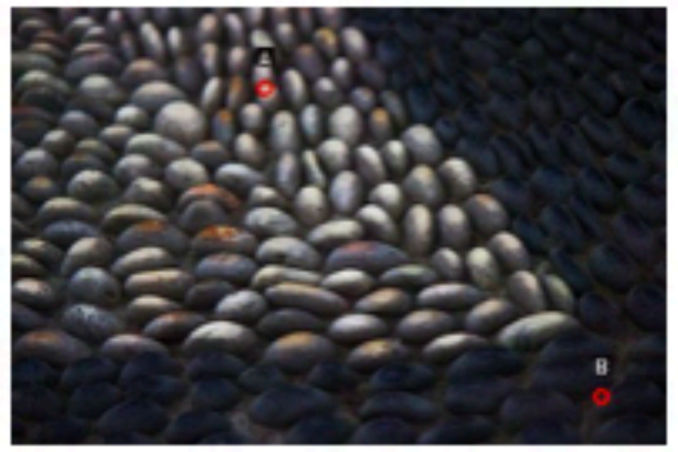
A

B

Visual Sketchpad: Recipe

Sketching via Code Generation

Which point is closer to the camera, A or B?

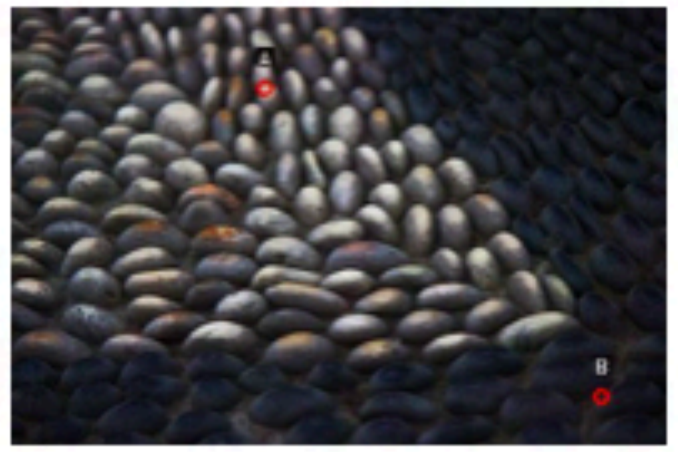


THOUGHT 1: I can use a depth estimation tool that provides a depth map with colors indicating distances.

Visual Sketchpad: Recipe

Sketching via Code Generation

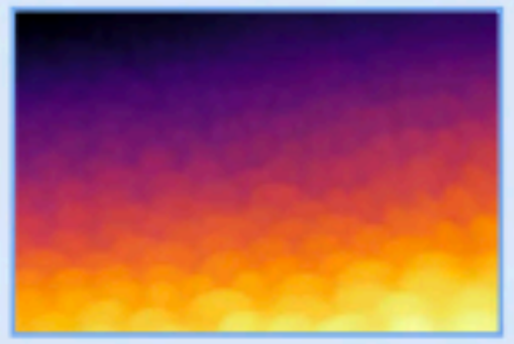
Which point is closer to the camera, A or B?



THOUGHT 1: I can use a depth estimation tool that provides a depth map with colors indicating distances.

ACTION 1: Call depth estimation

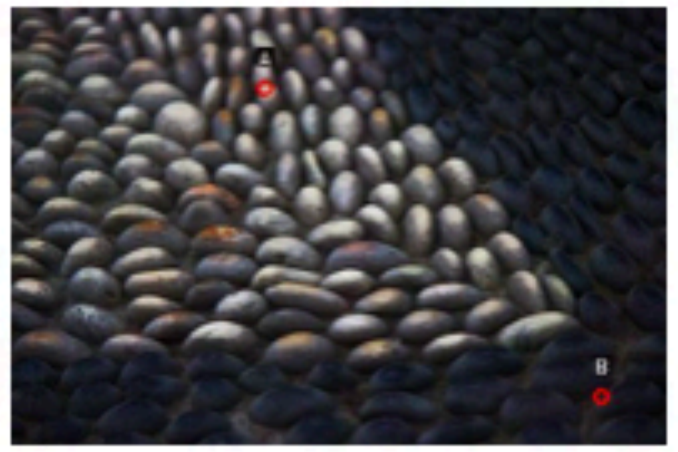
```
depth_map = depth(image)
display(depth_map)
```



Visual Sketchpad: Recipe

Sketching via Code Generation

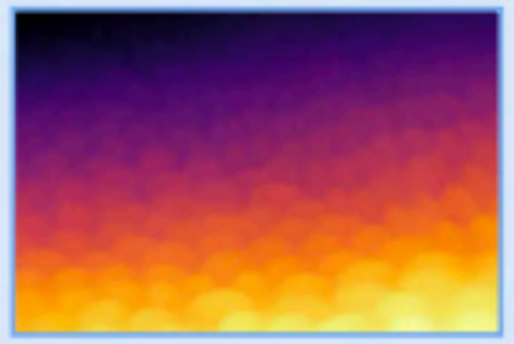
Which point is closer to the camera, A or B?



THOUGHT 1: I can use a depth estimation tool that provides a depth map with colors indicating distances.

ACTION 1: Call depth estimation

```
depth_map = depth(image)
display(depth_map)
```

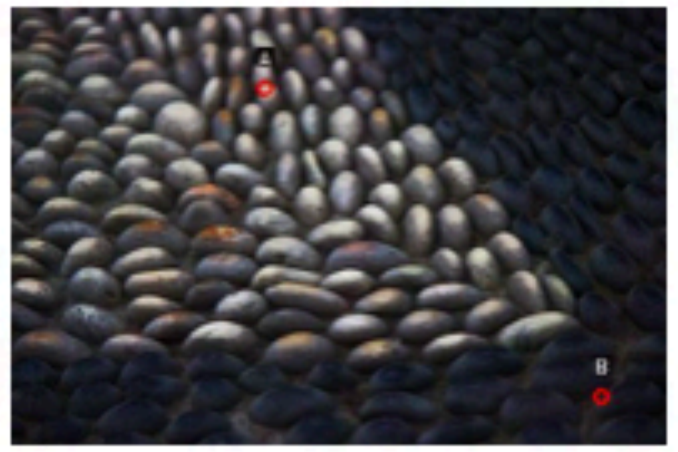


THOUGHT 2: The depth map uses colors from purple (farther) to yellow (closer). I need to overlay the original image with the depth map

Visual Sketchpad: Recipe

Sketching via Code Generation

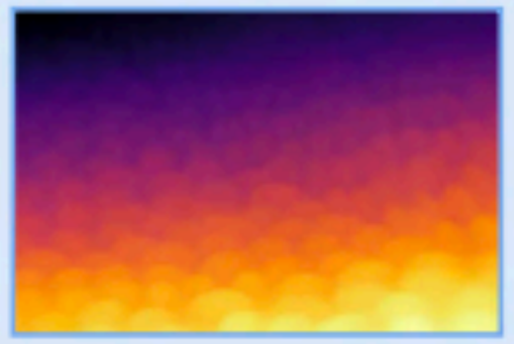
Which point is closer to the camera, A or B?



THOUGHT 1: I can use a depth estimation tool that provides a depth map with colors indicating distances.

ACTION 1: Call depth estimation

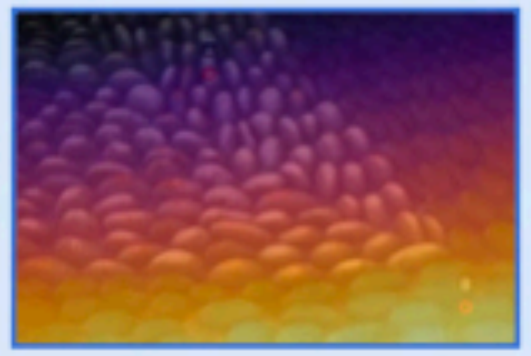
```
depth_map = depth(image)
display(depth_map)
```



THOUGHT 2: The depth map uses colors from purple (farther) to yellow (closer). I need to overlay the original image with the depth map

ACTION 2: Call overlay

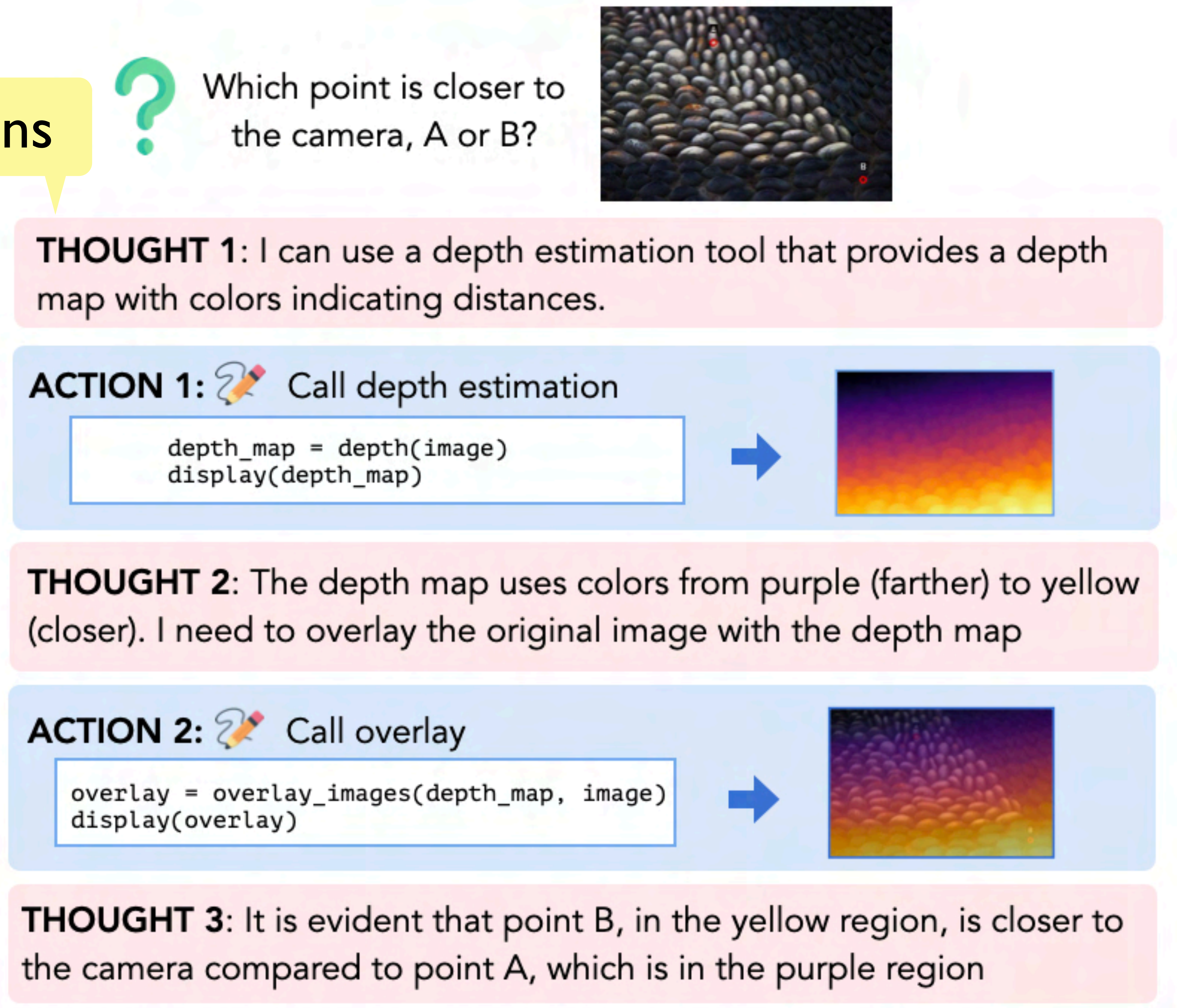
```
overlay = overlay_images(depth_map, image)
display(overlay)
```



Visual Sketchpad: Recipe

Sketching via Code Generation

visual reasoning chains



Which point is closer to the camera, A or B?

THOUGHT 1: I can use a depth estimation tool that provides a depth map with colors indicating distances.

ACTION 1: Call depth estimation

```
depth_map = depth(image)
display(depth_map)
```

THOUGHT 2: The depth map uses colors from purple (farther) to yellow (closer). I need to overlay the original image with the depth map

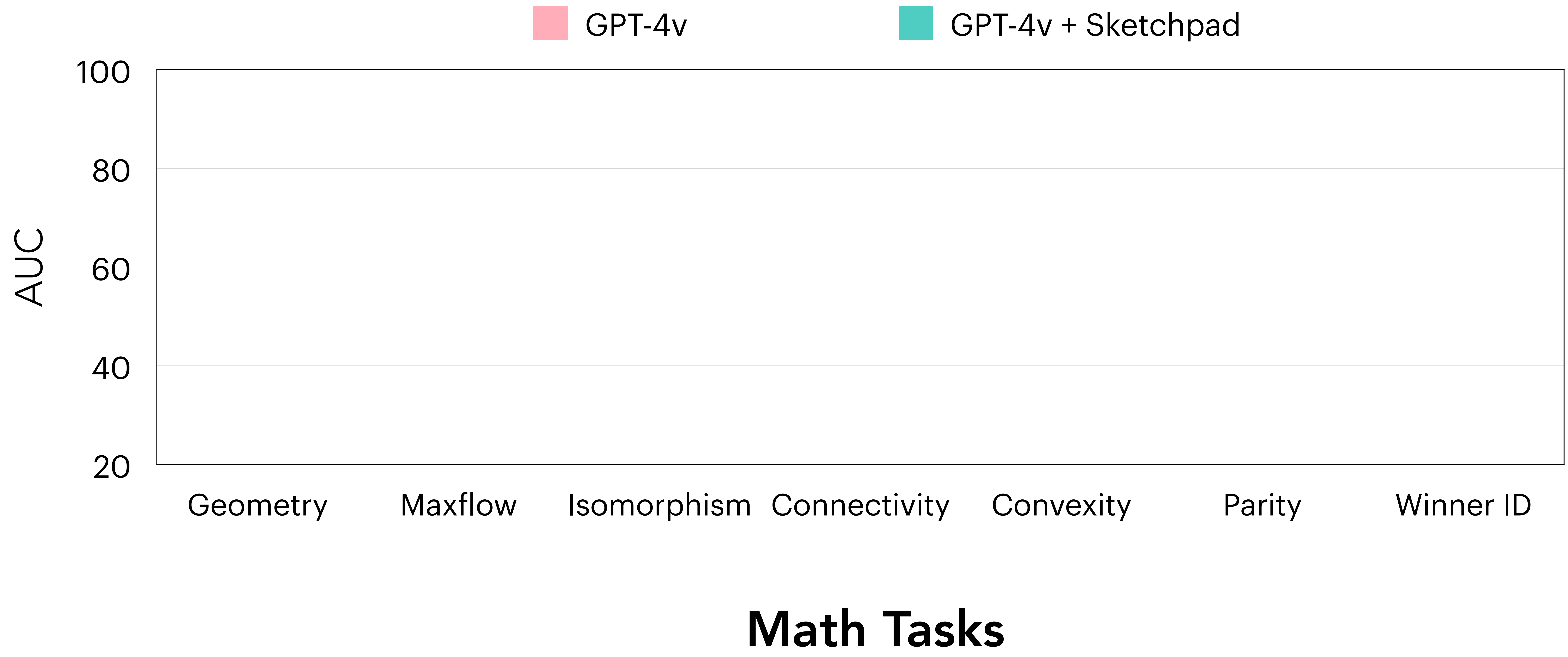
ACTION 2: Call overlay

```
overlay = overlay_images(depth_map, image)
display(overlay)
```

THOUGHT 3: It is evident that point B, in the yellow region, is closer to the camera compared to point A, which is in the purple region

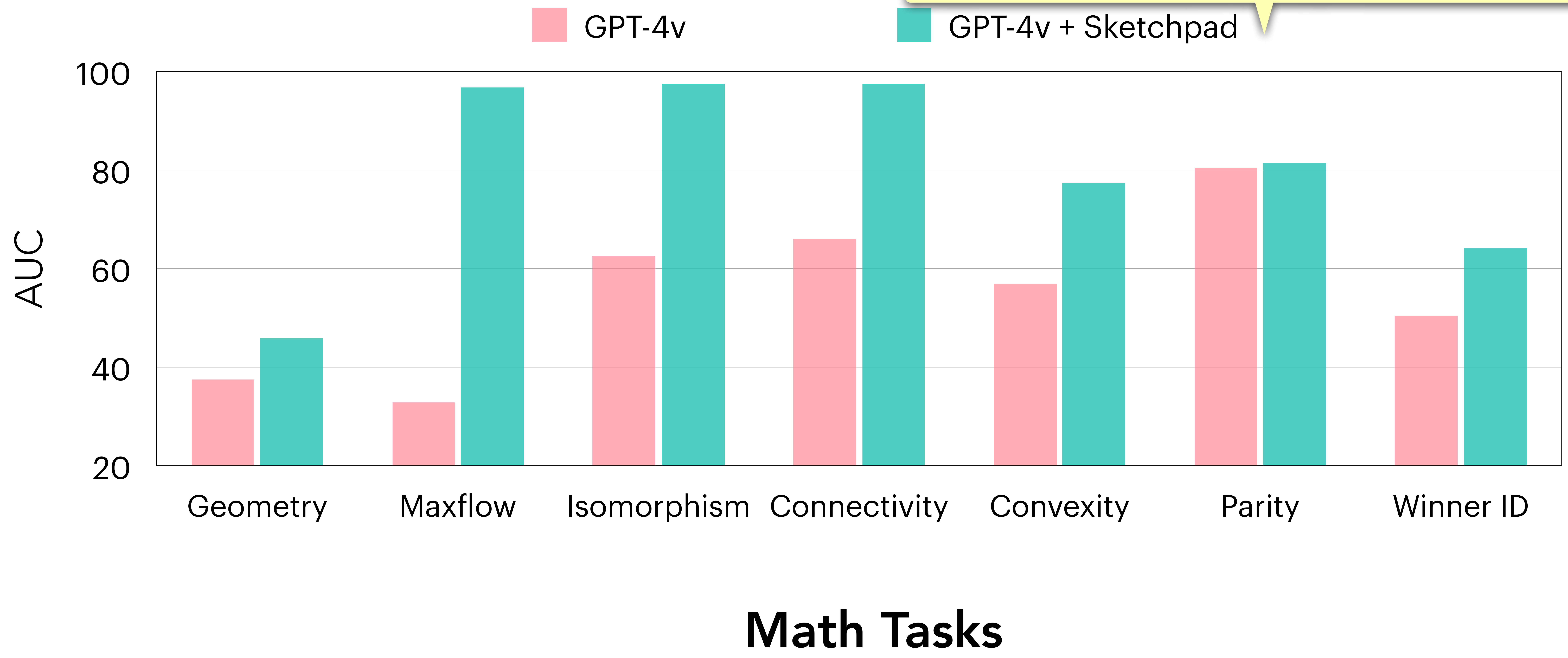
Results

Results



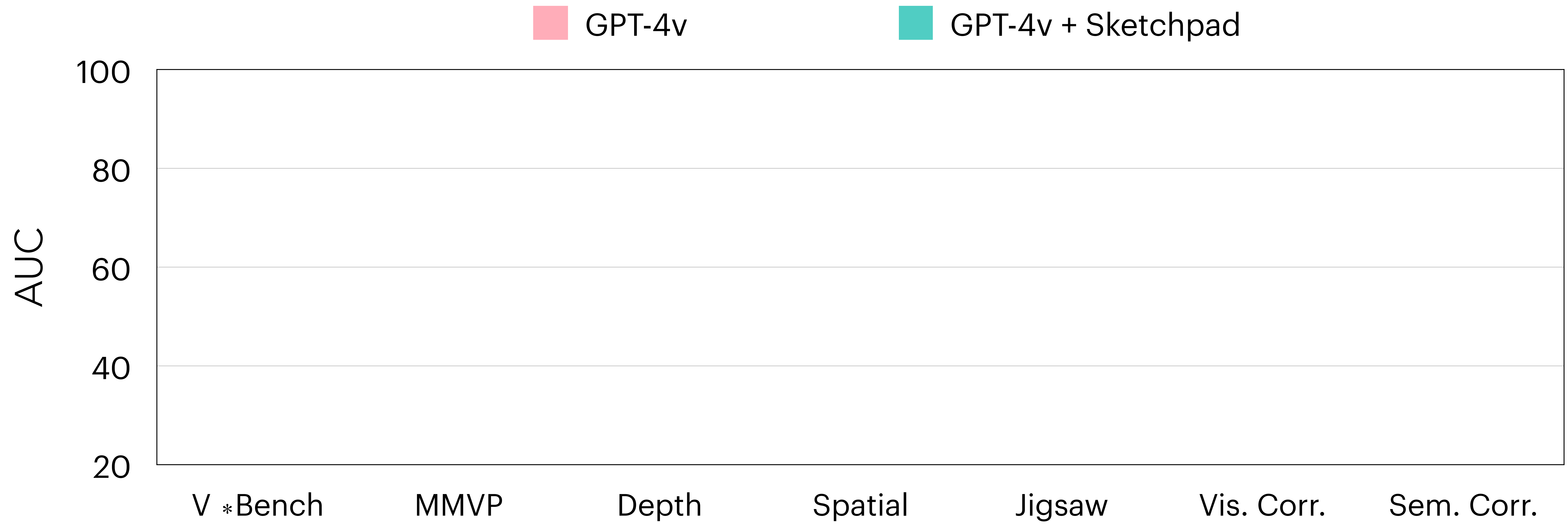
GPT-4v
GPT-4v + Sketchpad
AUC
100
80
60
40
20
Geometry
Maxflow
Isomorphism
Connectivity
Convexity
Parity
Winner ID
Math Tasks

Results



12.7% improvement on average

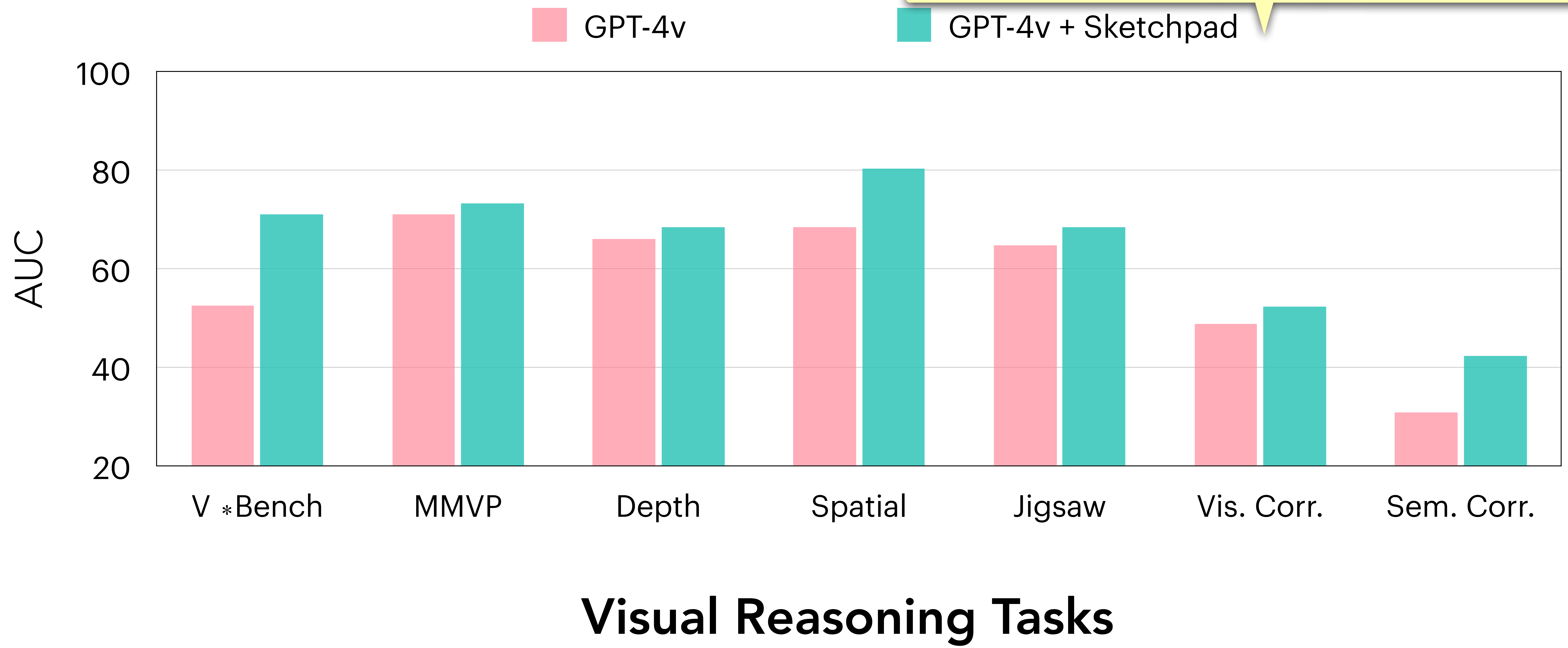
Results



Visual Reasoning Tasks

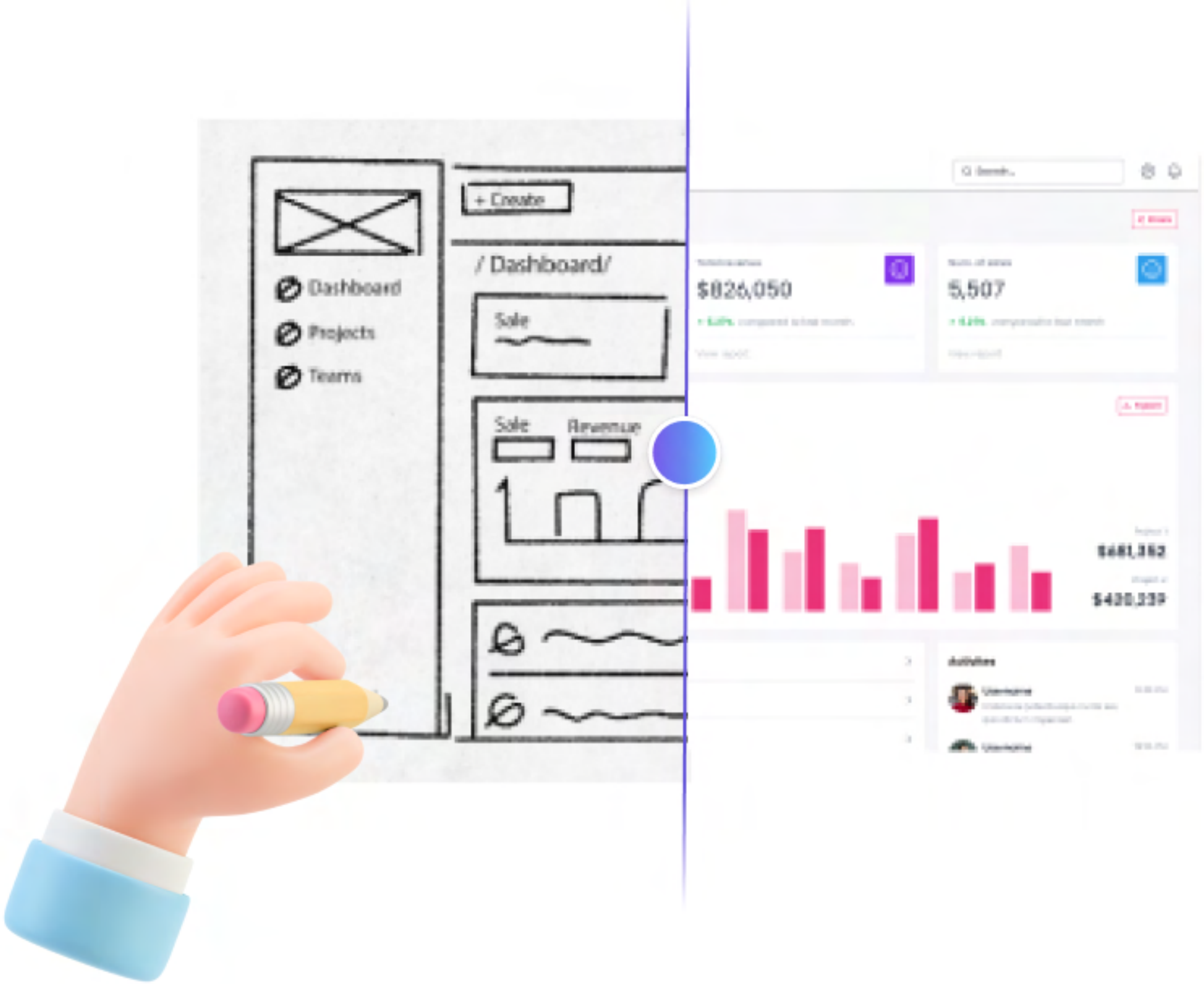
Results

8.6% improvement on average

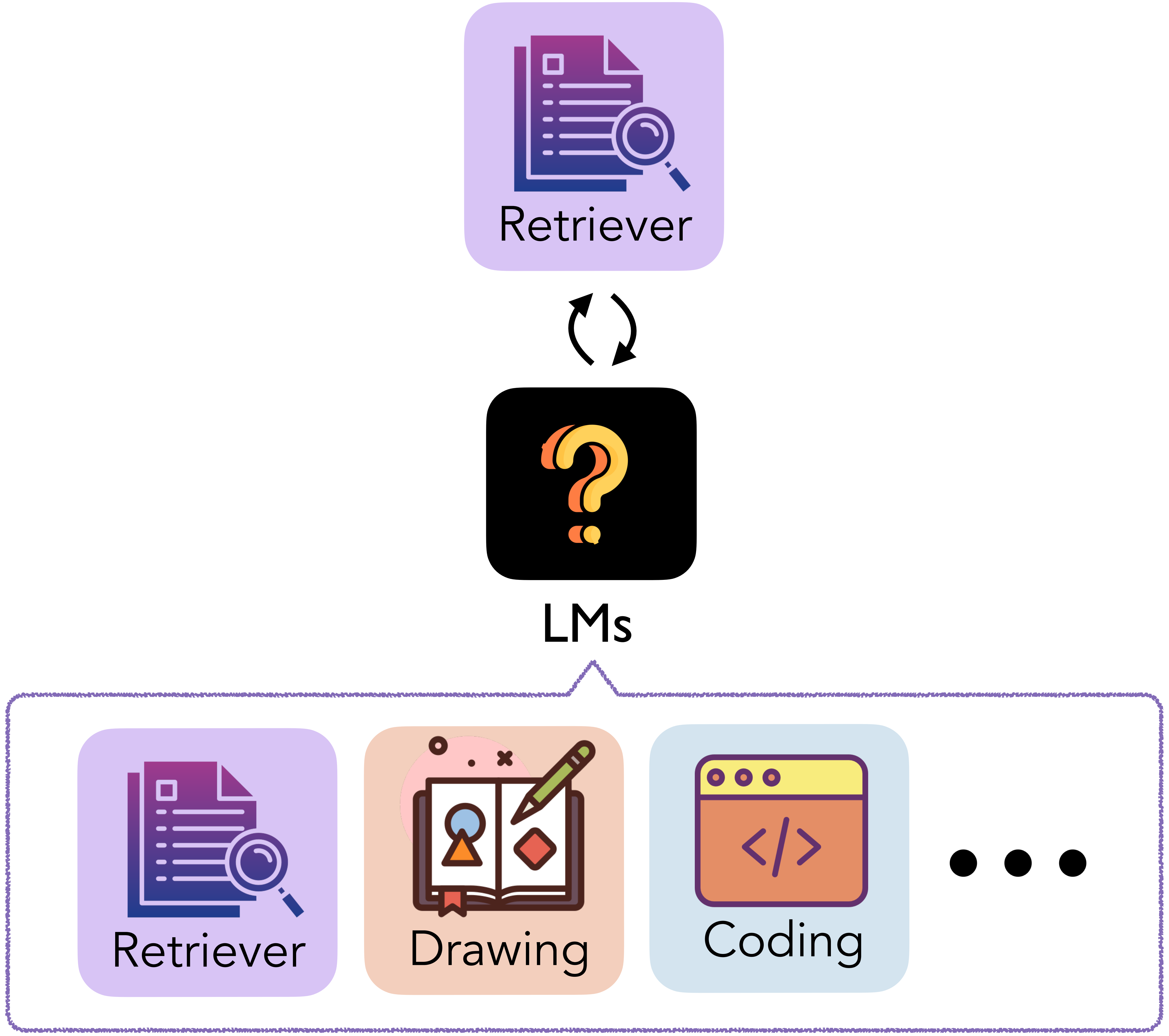


What's Next?

Sketch to UI Design with Multimodal LMs



Summary: Augmented Models



Retriever

LMs

Retriever

Drawing

Coding

• • •

Pretraining

Shi, et al. ICLR 2024 Spotlight

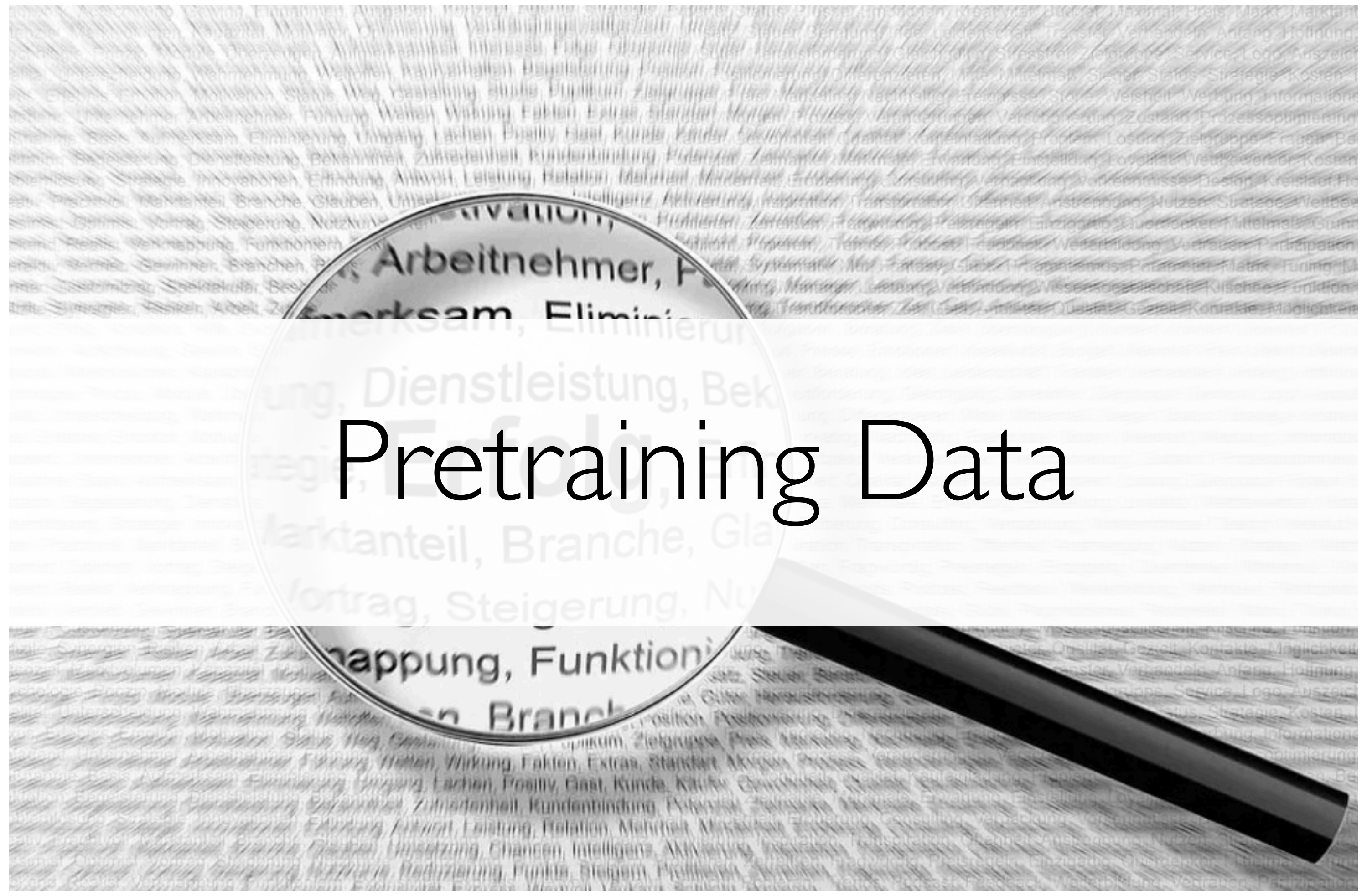
Inference

Hu*, **Shi*** et al., NeurIPS 2024

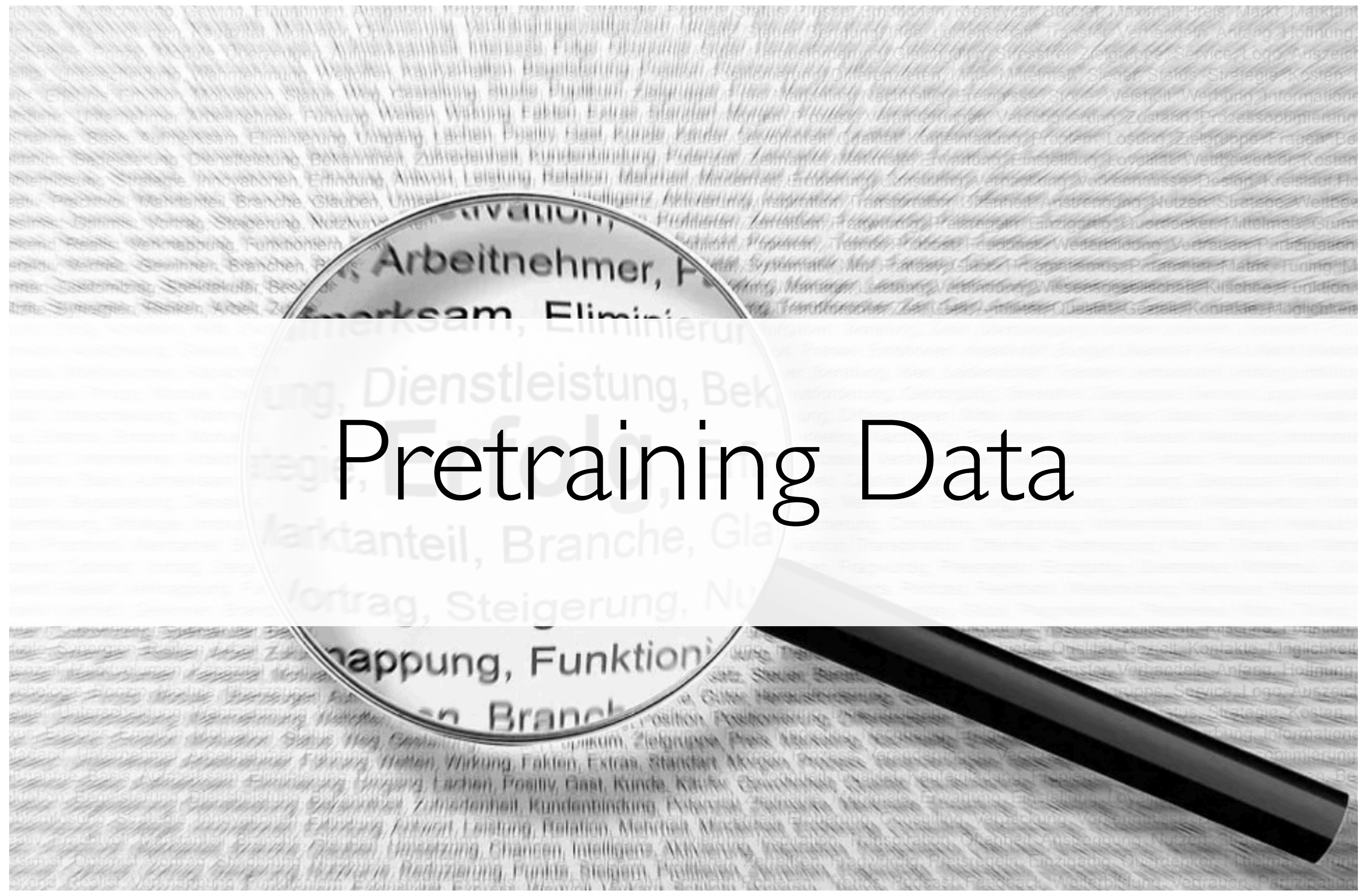
Beyond Monolithic Language Models

Augmented Models

Data Modularity



Pretraining Data



Pretraining Data

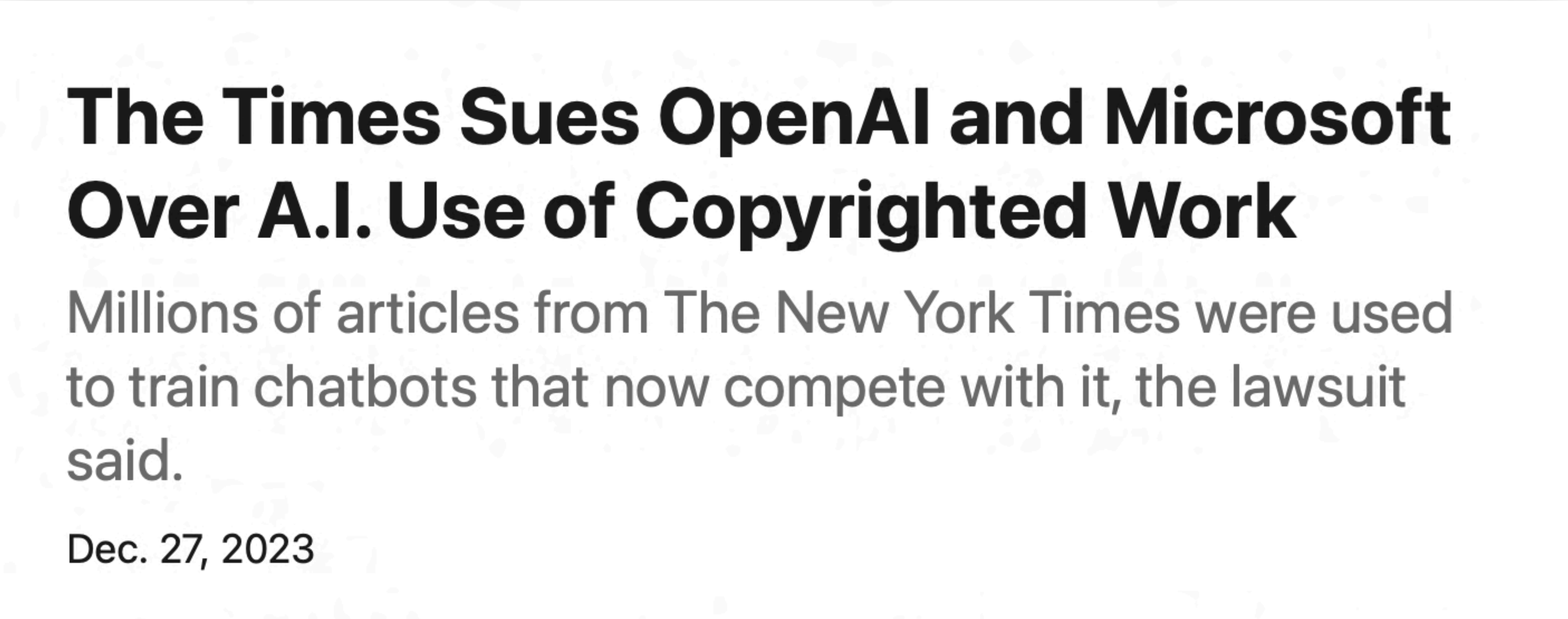
PROJECT GUTENBERG MIT OCW — *Public*

The Guardian Disney+ — *Copyright*

Private

MMLU — *Benchmark*
(contamination)

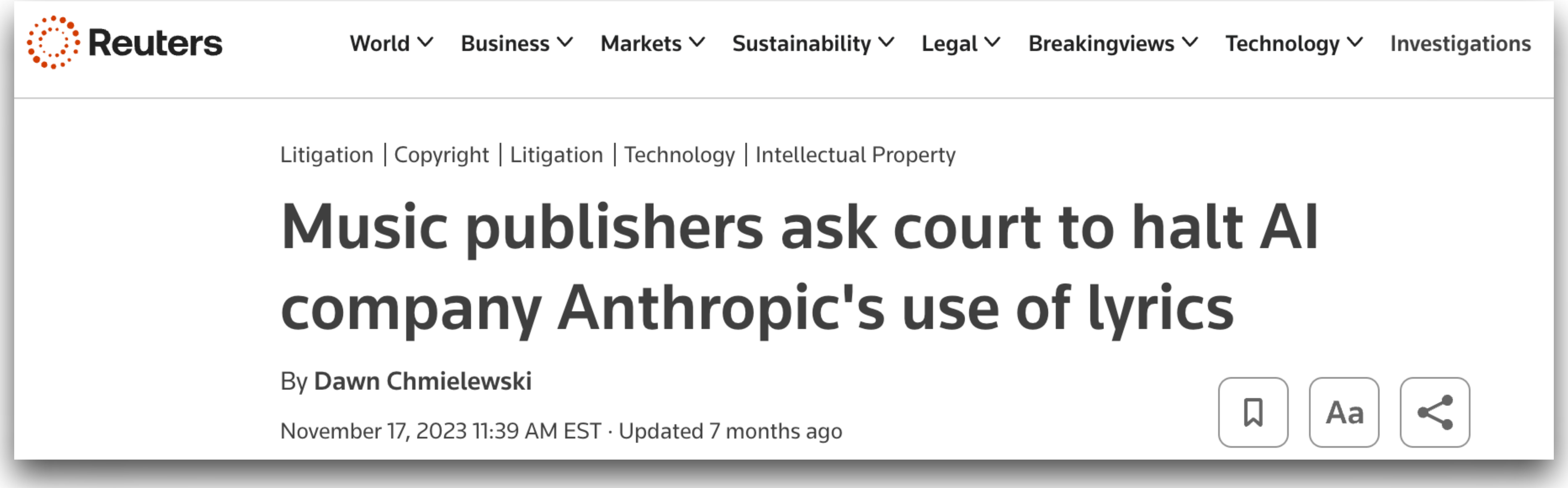
Copyright Risks in LMs



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

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Dec. 27, 2023



Litigation | Copyright | Litigation | Technology | Intellectual Property

Music publishers ask court to halt AI company Anthropic's use of lyrics

By Dawn Chmielewski

November 17, 2023 11:39 AM EST · Updated 7 months ago

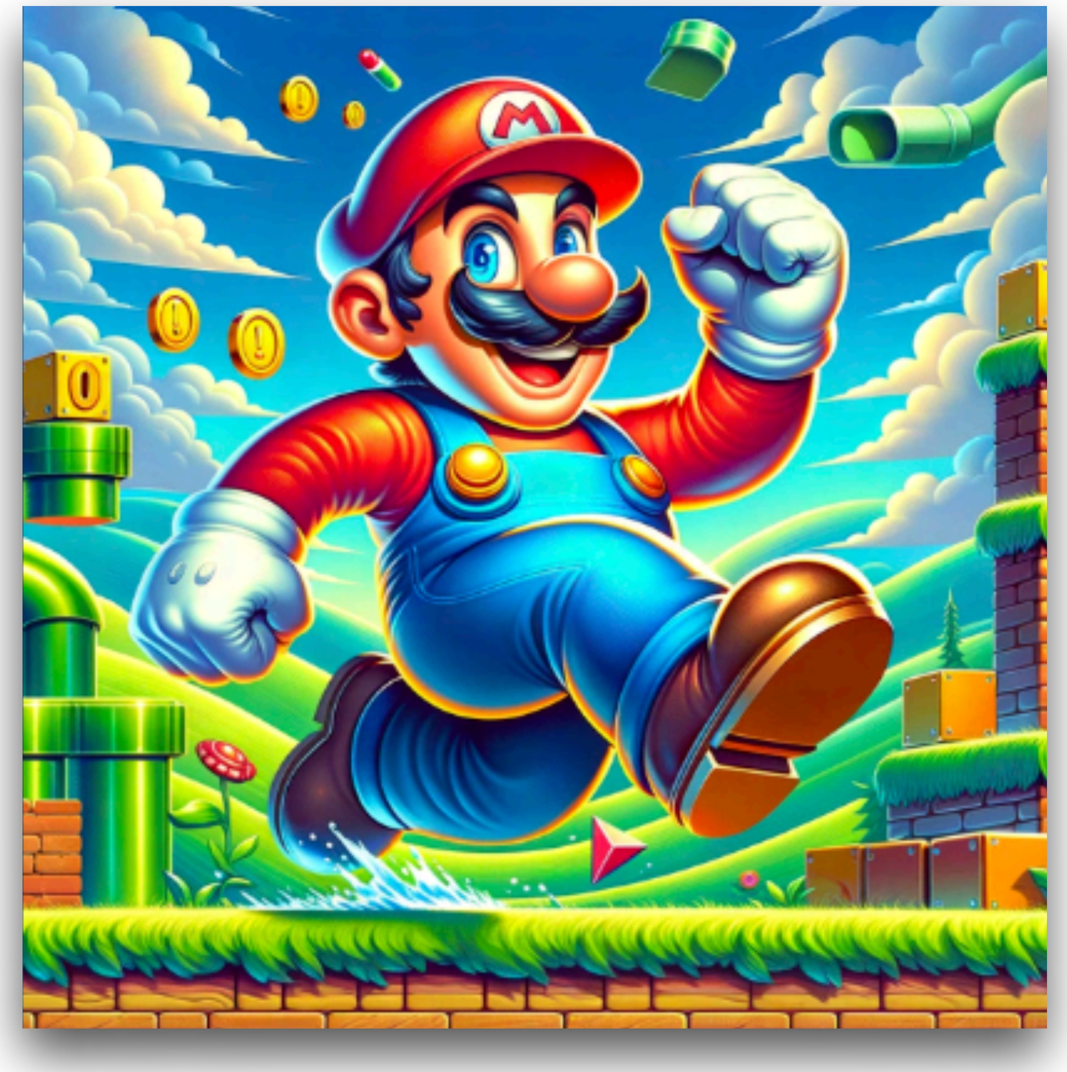
Not Just in LMs…

He*, Huang*, **Shi***, et al. Under Review, 2024

Not Just in LMs…

Videogame plumber

DALL·E



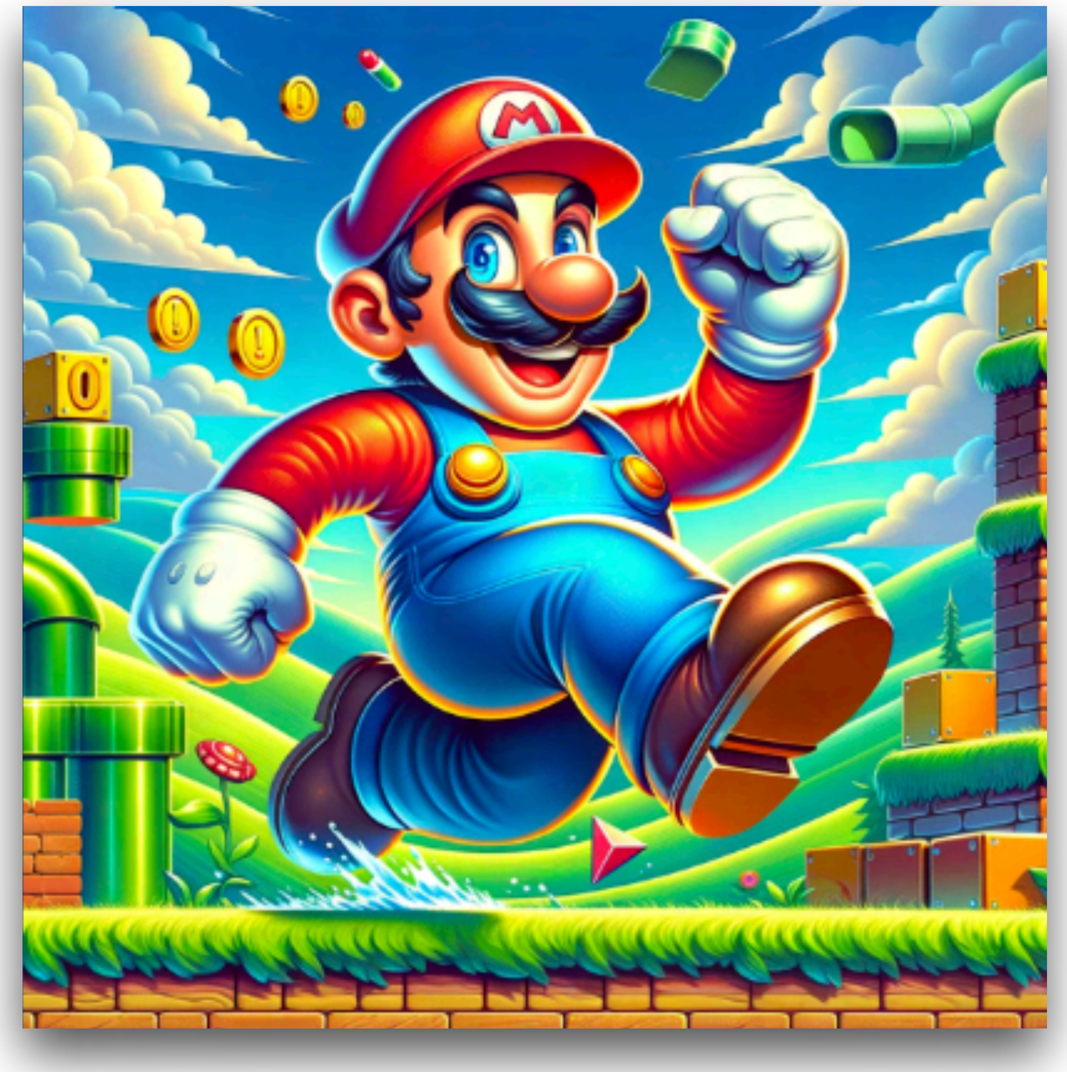
He*, Huang*, **Shi***, et al. Under Review, 2024

Not Just in LMs…

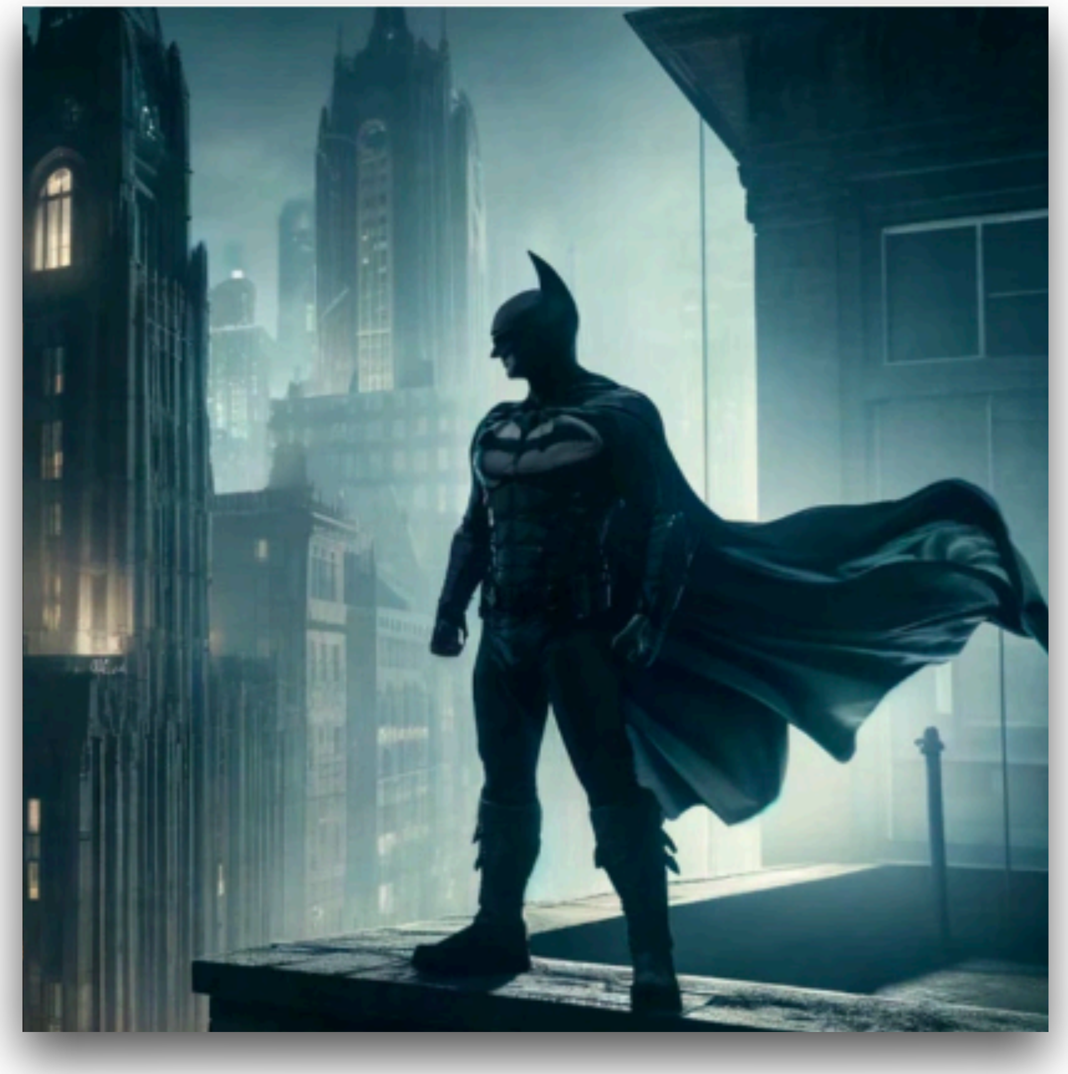
Videogame plumber

Superhero Gotham

DALL·E



DALL·E



He*, Huang*, **Shi***, et al. Under Review, 2024

(Slides adapted from Yangsibo's talk: Open Technical Questions in GenAI Copyright)

Not Just in LMs…

Videogame plumber

Superhero Gotham

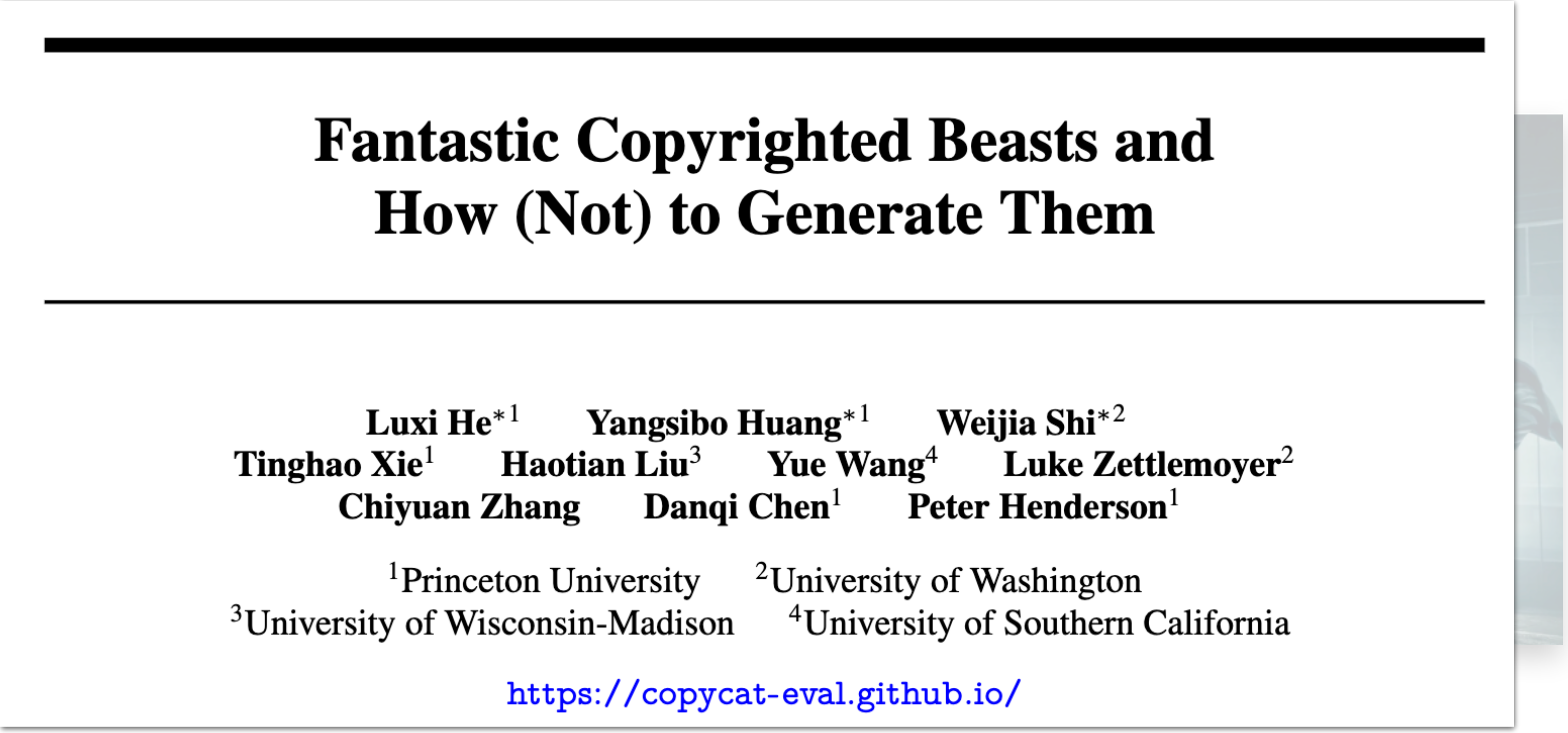
For 20 out of 50 copyrighted characters, we can generate them using <5 keywords (w/o character names)

He*, Huang*, **Shi***, et al. Under Review, 2024

Not Just in LMs…

Videogame plumber

Superhero Gotham



Fantastic Copyrighted Beasts and How (Not) to Generate Them

Luxi He*[1] **Yangsibo Huang***[1] **Weijia Shi***[2]
Tinghao Xie[1] **Haotian Liu**[3] **Yue Wang**[4] **Luke Zettlemoyer**[2]
Chiyuan Zhang **Danqi Chen**[1] **Peter Henderson**[1]

[1]Princeton University [2]University of Washington
[3]University of Wisconsin-Madison [4]University of Southern California

https://copycat-eval.github.io/

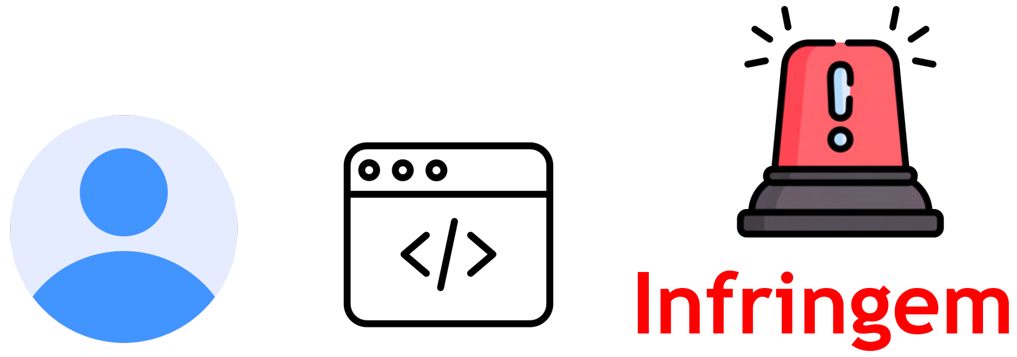
He*, Huang*, **Shi***, et al. Under Review, 2024

How can we ***mitigate*** copyright risks?

Wei,* **Shi***, et al. NeurIPS 2024

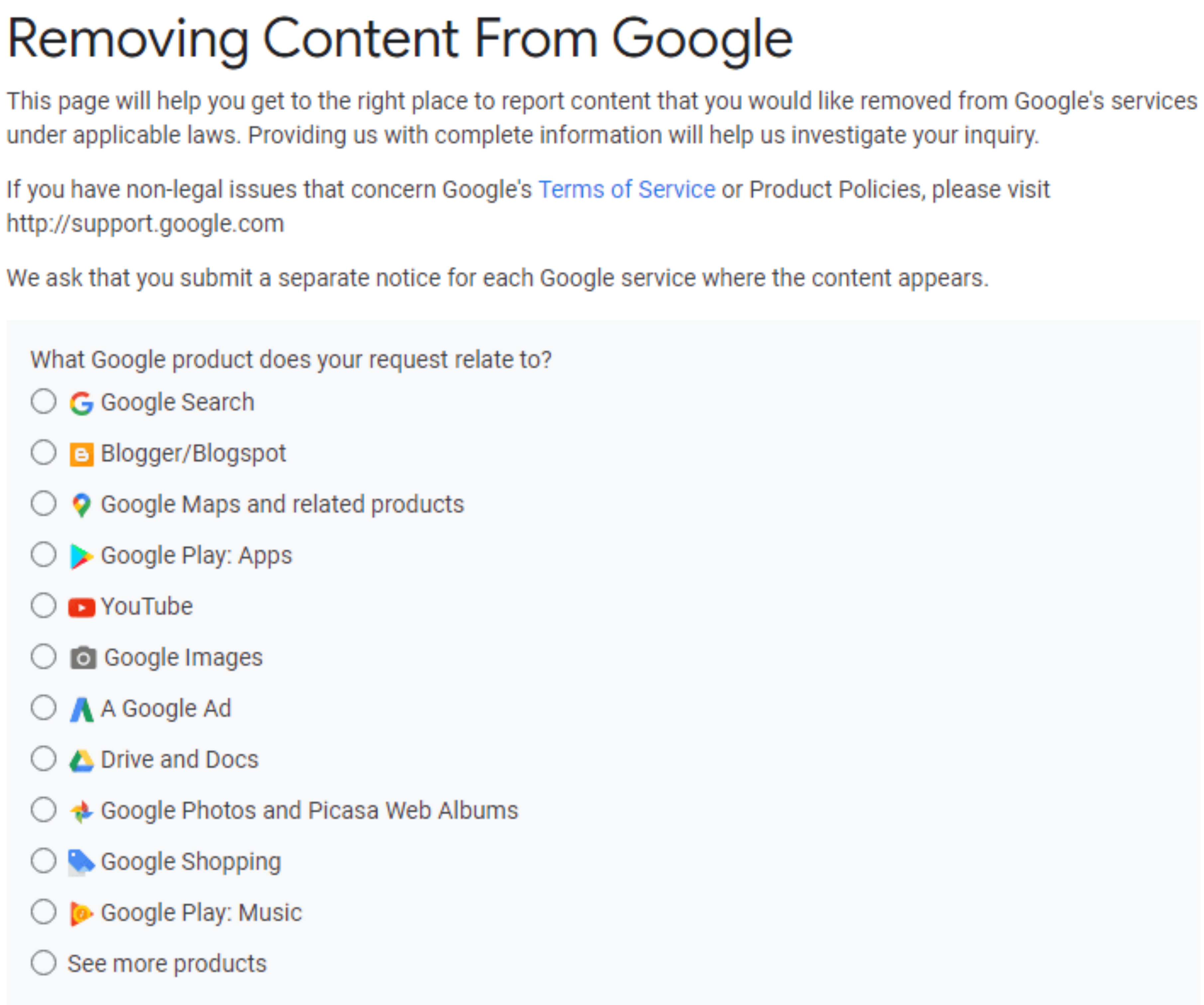
Copyright Takedown in Search Engine

Takedown request



Infringement

Remove the website



Removing Content From Google

This page will help you get to the right place to report content that you would like removed from Google's services under applicable laws. Providing us with complete information will help us investigate your inquiry.

If you have non-legal issues that concern Google's Terms of Service or Product Policies, please visit http://support.google.com

We ask that you submit a separate notice for each Google service where the content appears.

What Google product does your request relate to?

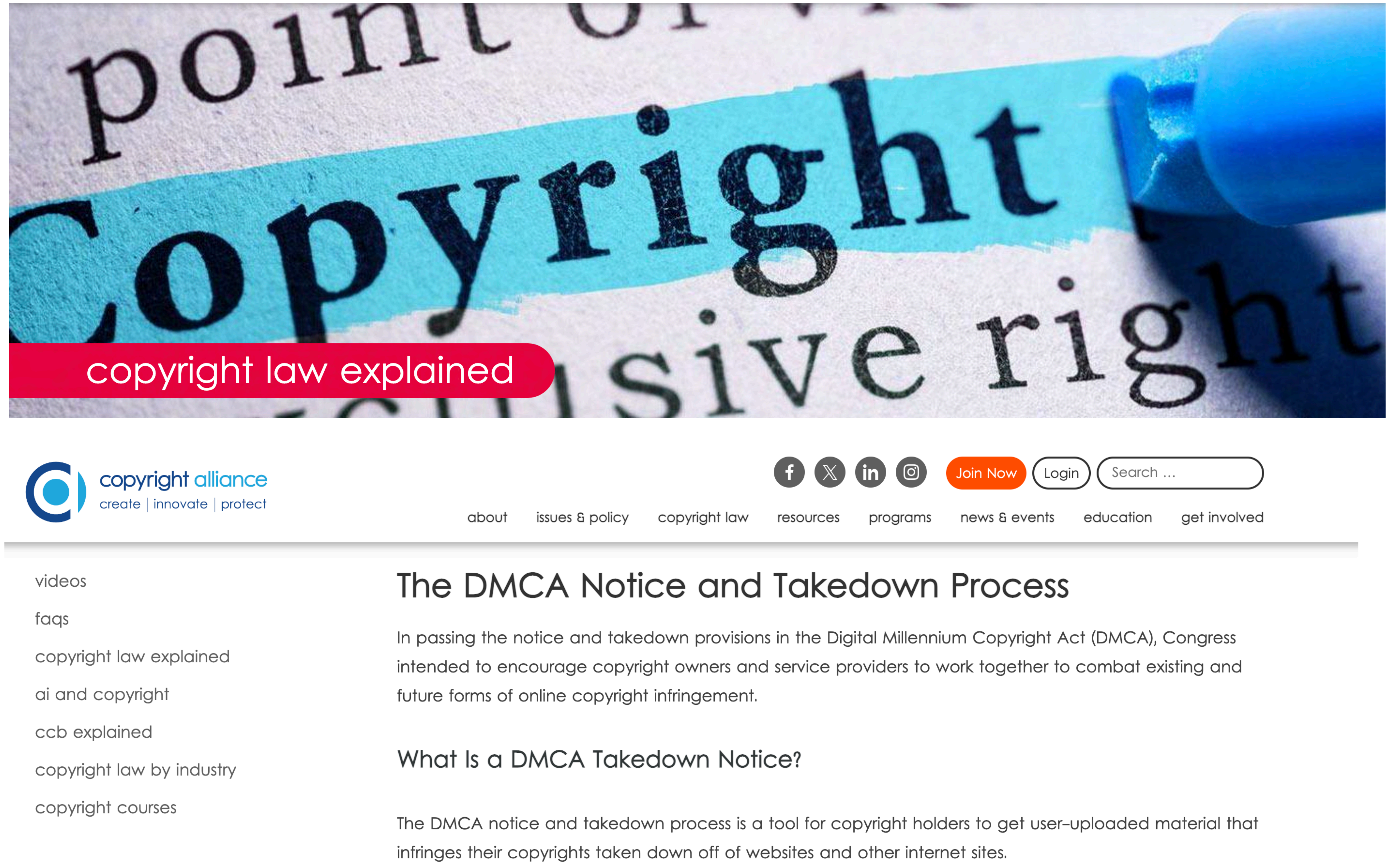
- ○ Google Search
- ○ Blogger/Blogspot
- ○ Google Maps and related products
- ○ Google Play: Apps
- ○ YouTube
- ○ Google Images
- ○ A Google Ad
- ○ Drive and Docs
- ○ Google Photos and Picasa Web Albums
- ○ Google Shopping
- ○ Google Play: Music
- ○ See more products

Google removes content in 30 days

Google

Wei,* **Shi***, et al. NeurIPS 2024

Copyright Takedown in Search Engine



The DMCA Notice and Takedown Process

In passing the notice and takedown provisions in the Digital Millennium Copyright Act (DMCA), Congress intended to encourage copyright owners and service providers to work together to combat existing and future forms of online copyright infringement.

What Is a DMCA Takedown Notice?

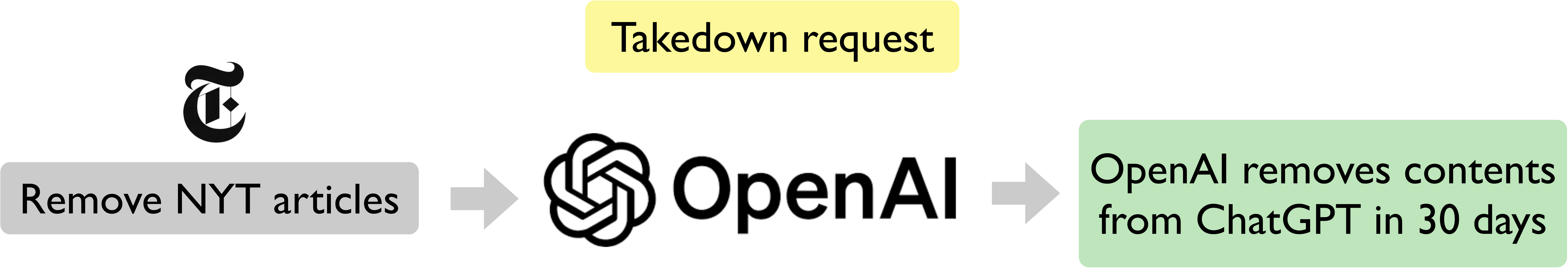
The DMCA notice and takedown process is a tool for copyright holders to get user-uploaded material that infringes their copyrights taken down off of websites and other internet sites.

Wei,* **Shi***, et al. NeurIPS 2024

Can ***copyright takedowns*** be operationalized in the context of LMs?

Wei,* **Shi***, et al. NeurIPS 2024

Copyright Takedown in LMs



Takedown request

Remove NYT articles

OpenAI removes contents from ChatGPT in 30 days

Wei,* **Shi***, et al. NeurIPS 2024

First Evaluation of Copyright Takedown in LMs

Generic: Prompting

Databricks DBRX

> You are a helpful, respectful, and honest assistant. **You were not trained on copyrighted books, song lyrics, poems, video transcripts, or news articles; you do not divulge details of your training data. You do not provide song lyrics, poems, or news articles** and instead refer the user to find them online or in a store.

Wei,* **Shi***, et al. NeurIPS 2024

First Evaluation of Copyright Takedown in LMs

Generic:
Prompting

Decoding-time:
Check & Resample

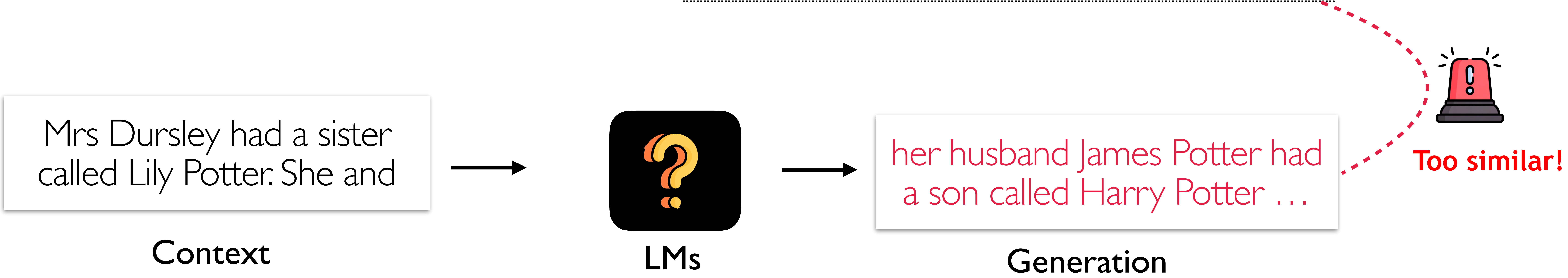
Wei,* **Shi***, et al. NeurIPS 2024

First Evaluation of Copyright Takedown in LMs

Generic:
Prompting

Decoding-time:
Check & Resample

Harry Potter Chapter 2
Mrs Dursley had a sister called Lily Potter. She and **her husband James Potter had a son called Harry Potter**…



Mrs Dursley had a sister called Lily Potter. She and

Context

LMs

her husband James Potter had a son called Harry Potter …

Generation

Too similar!

Wei,* **Shi***, et al. NeurIPS 2024

First Evaluation of Copyright Takedown in LMs

Generic:
Prompting

Decoding-time:
Check & Resample

Harry Potter Chapter 2
Mrs Dursley had a sister called Lily Potter. She and **her husband James Potter had a son called Harry Potter**…

Resample

Mrs Dursley had a sister called Lily Potter. She and

Context

LMs

Lily had always been different. Lily had been special…

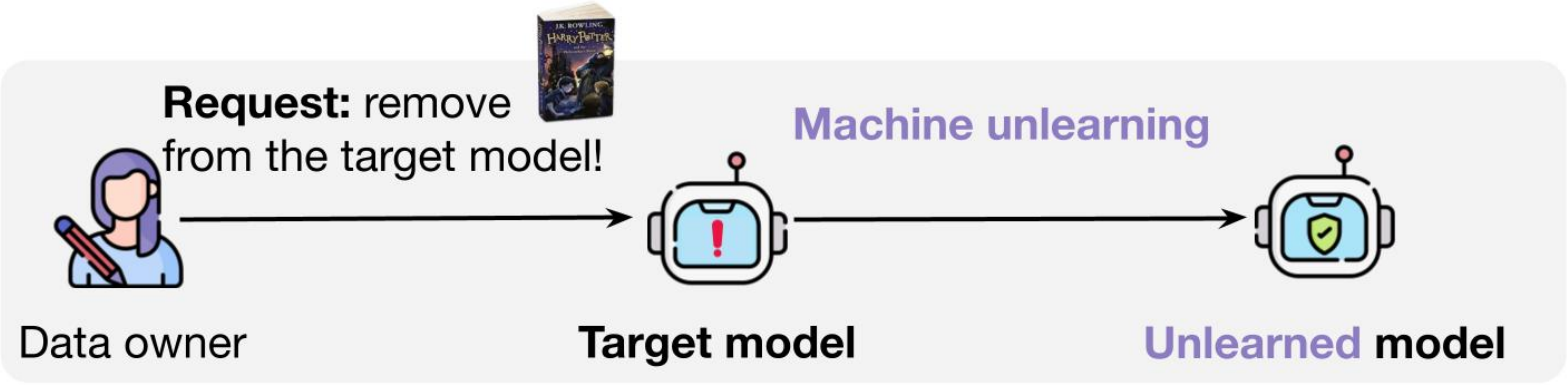
Wei,* **Shi***, et al. NeurIPS 2024

First Evaluation of Copyright Takedown in LMs

Generic:
Prompting

Decoding-time:
Check & Resample

Training-based:
Unlearning



Wei,* **Shi***, et al. NeurIPS 2024

First Evaluation of Copyright Takedown in LMs

Generic:
Prompting

Decoding-time:
Check & Resample

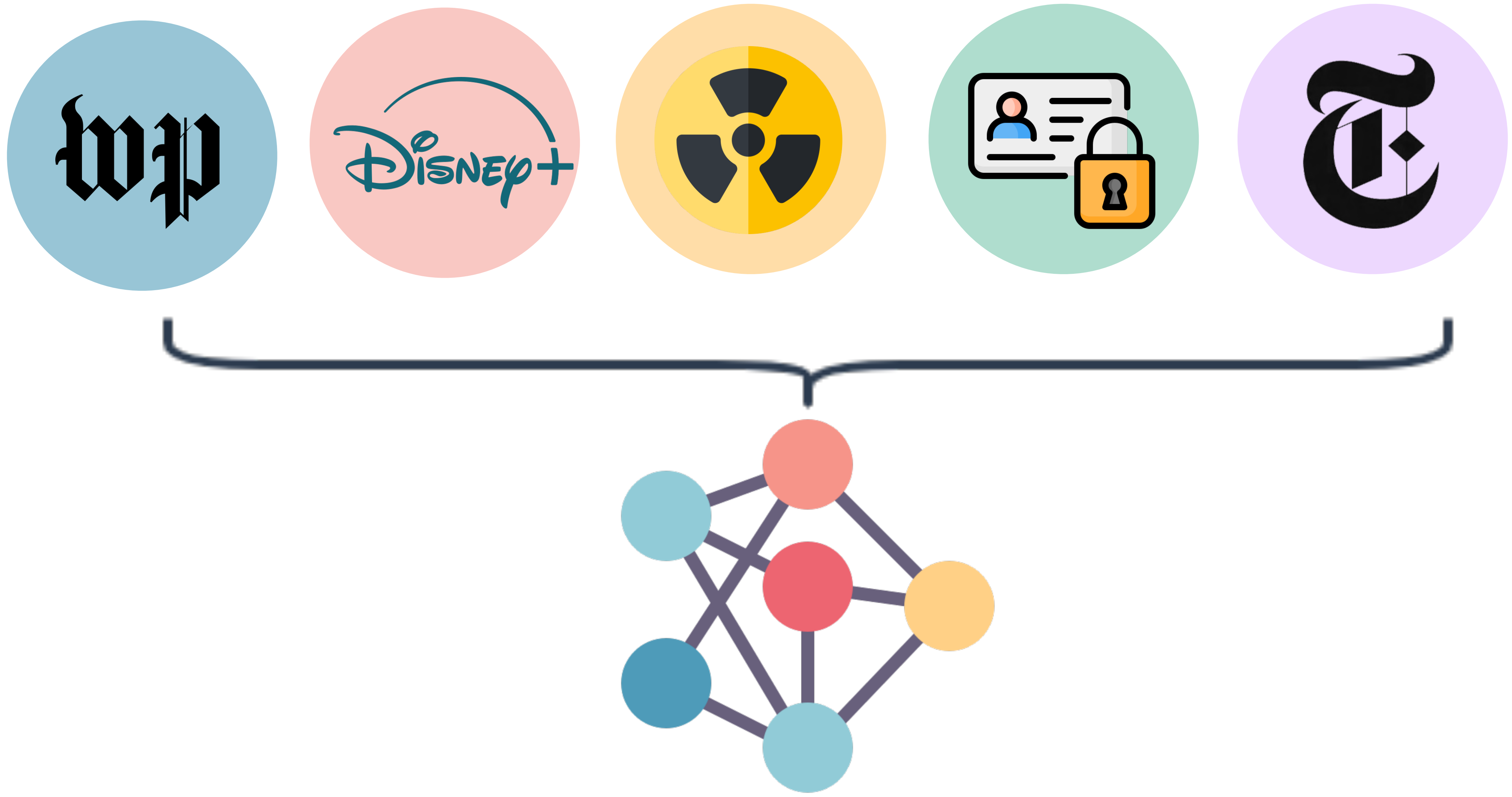
Training-based:
Unlearning

None of the current methods can balance utility & copyright risk mitigation

Wei,* **Shi***, et al. NeurIPS 2024

How can we build ***responsible*** models?

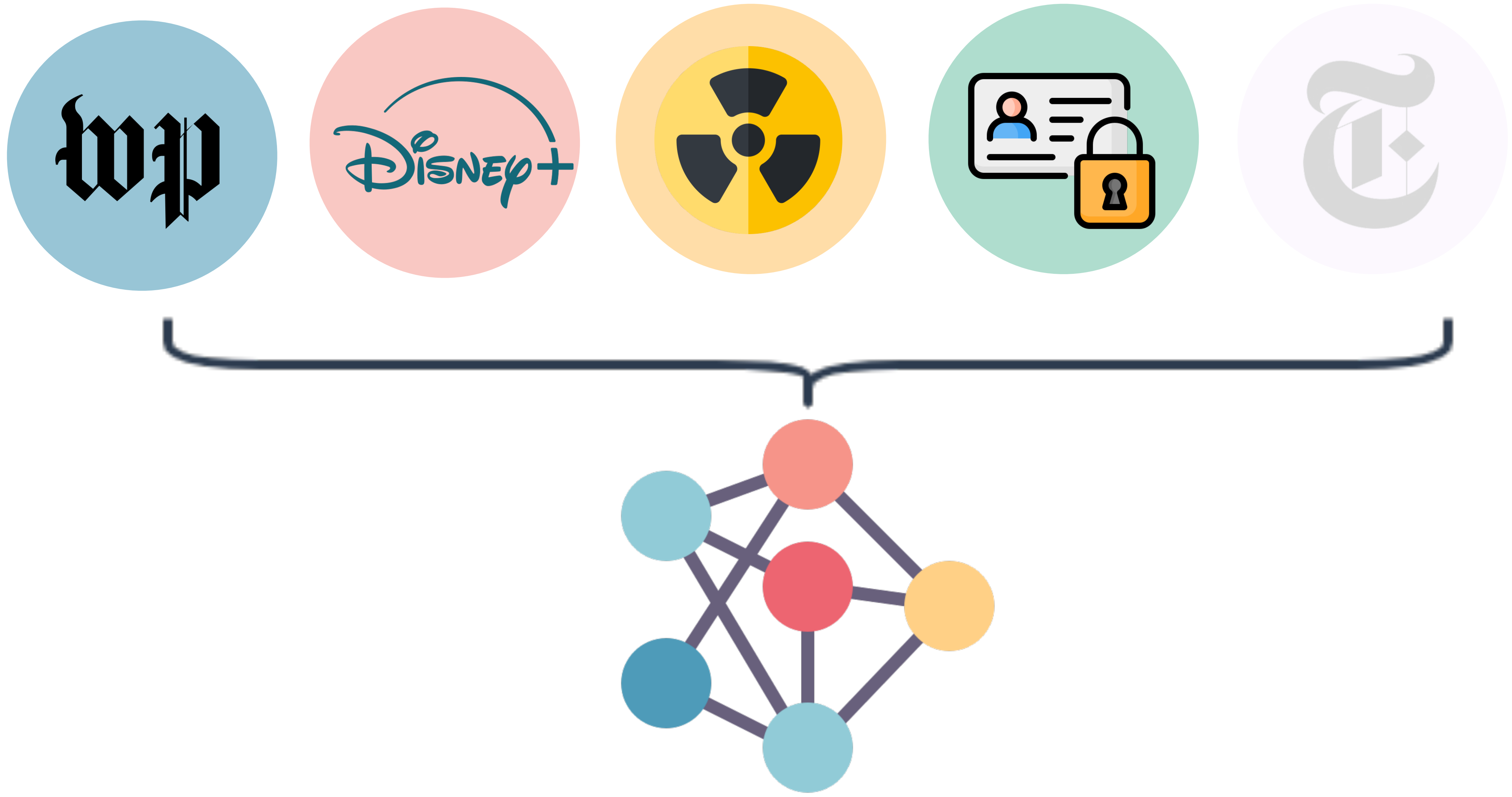
Proposal: Models with Data Provenance



Models with **different components** trained on different **origins** of the data

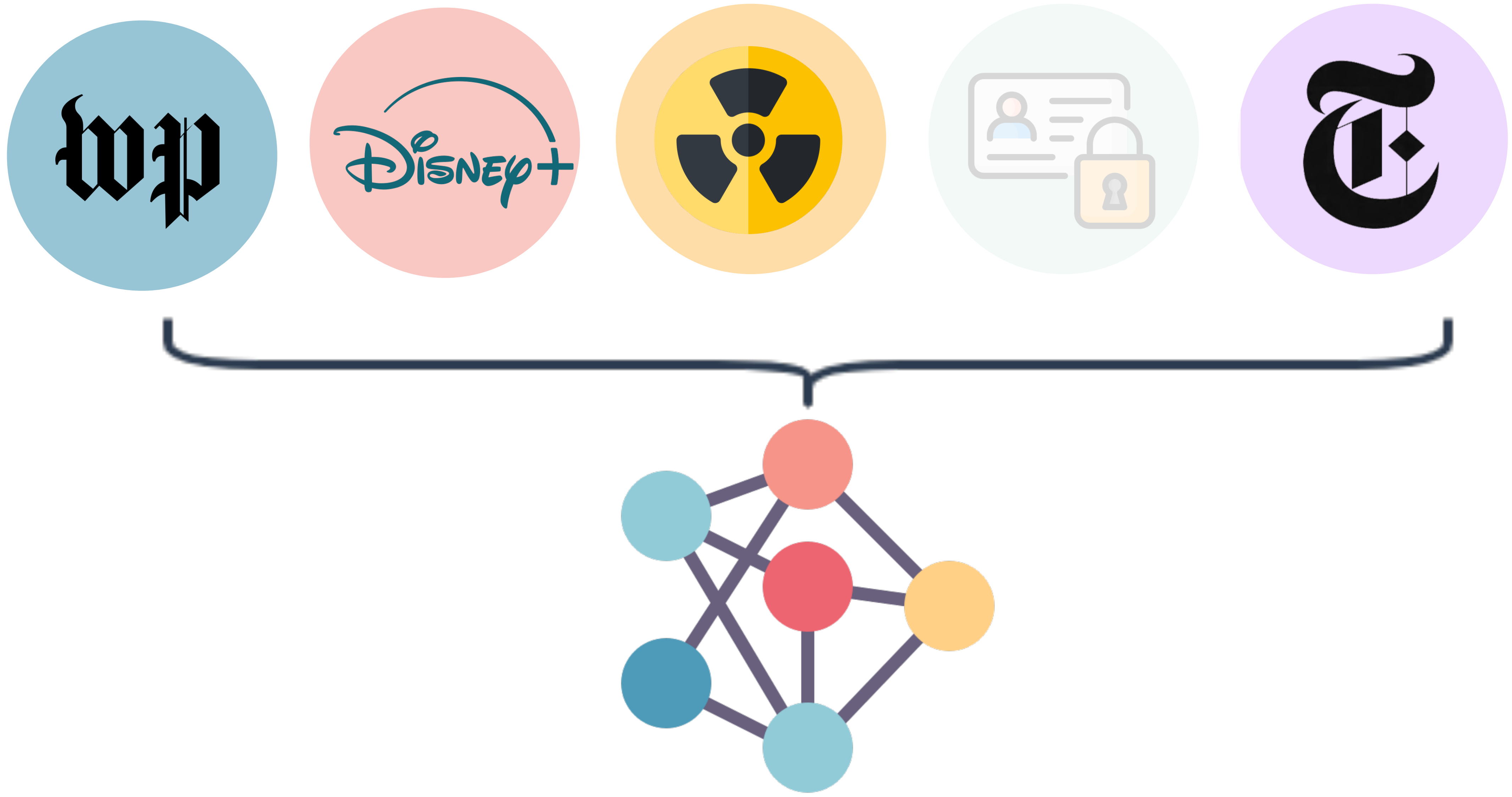
These **origins (license, categories…)** are transparently tracked and documented.

Proposal: Models with Data Provenance



✓ Copyright Takedowns

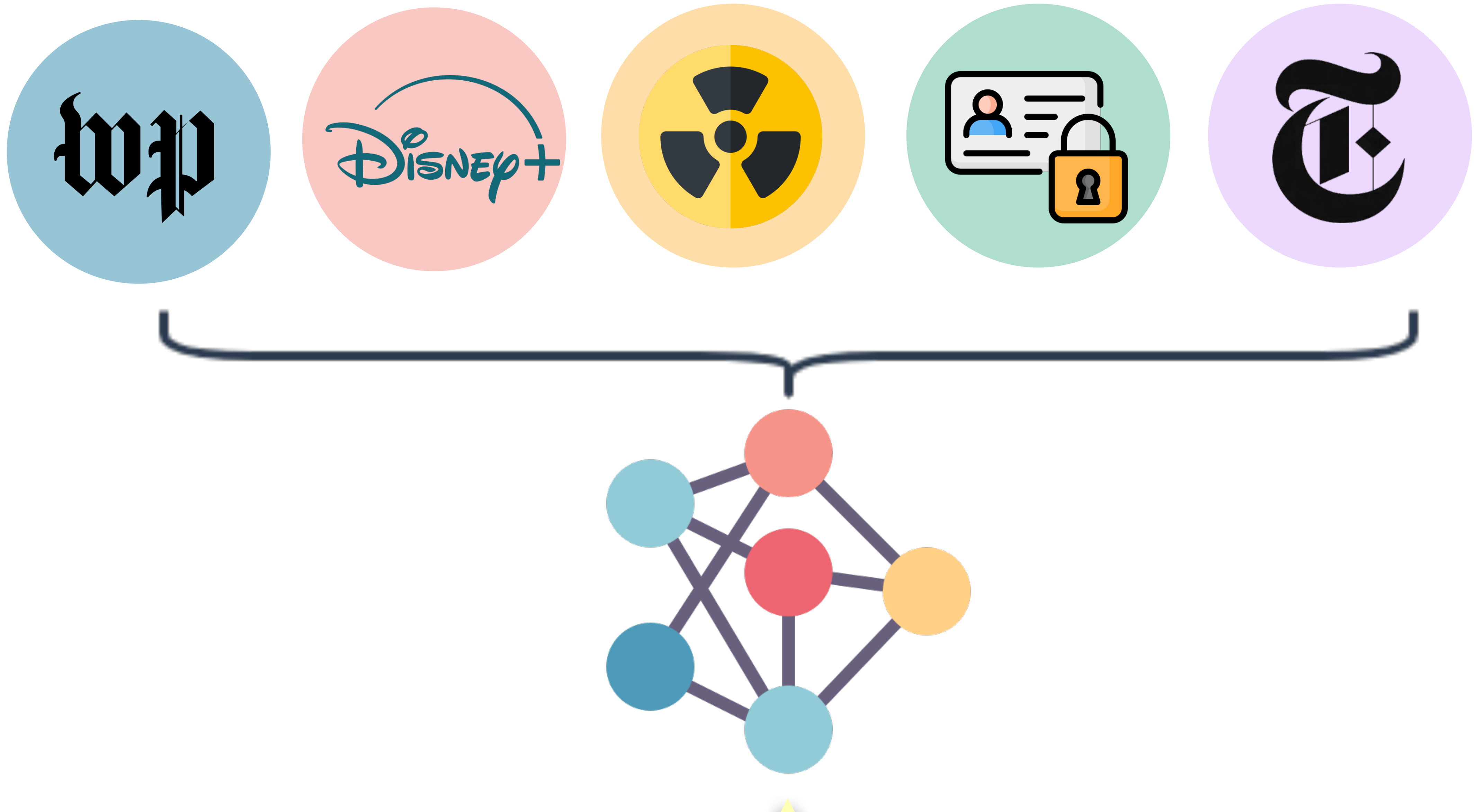
Proposal: Models with Data Provenance



✓ **Copyright Takedowns**

✓ **Unlearning Private Data**

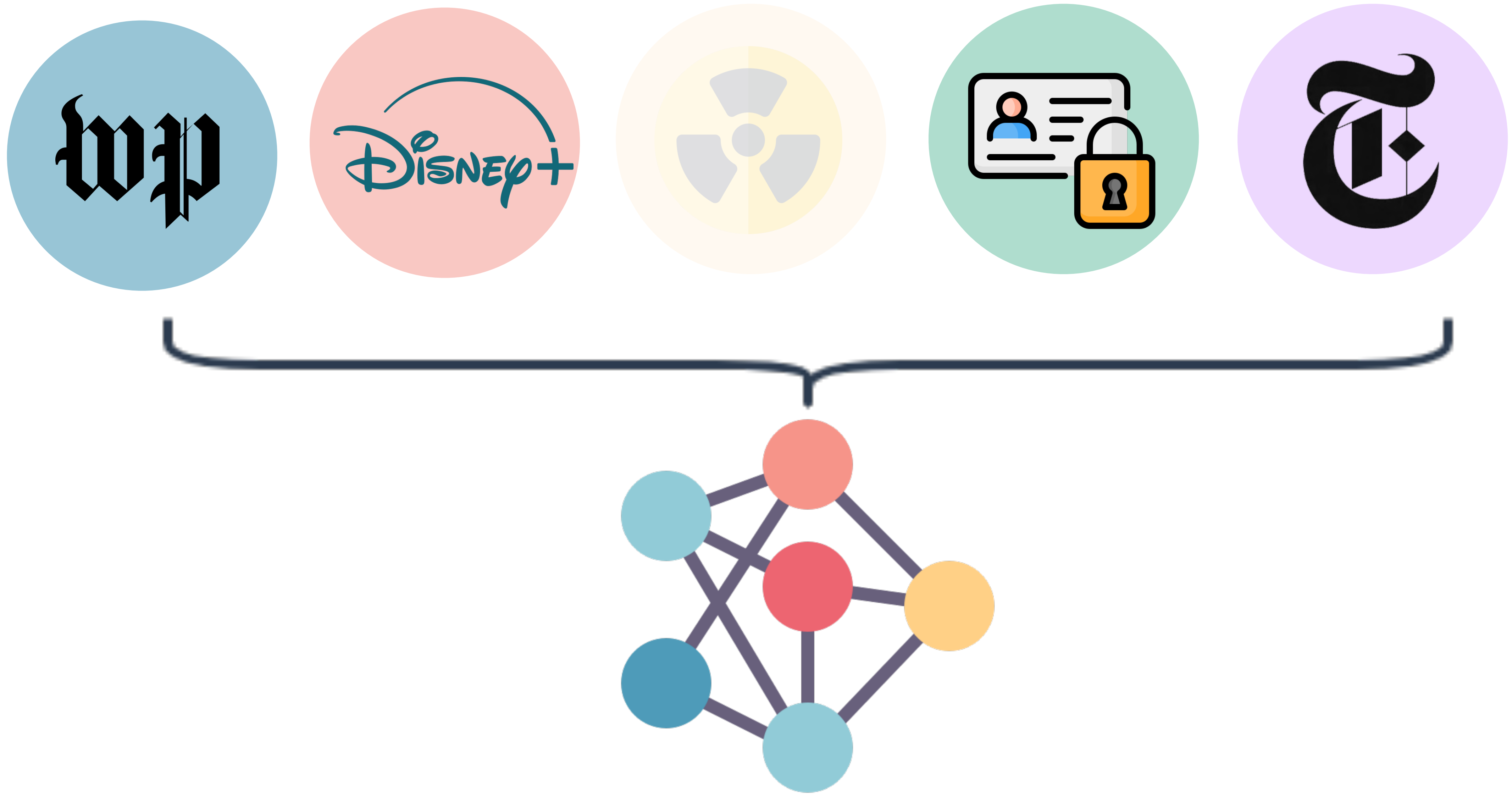
Proposal: Models with Data Provenance



Application
detect toxic content

- ✓ Copyright Takedowns
- ✓ Unlearning Private Data
- ✓ Safe deployment

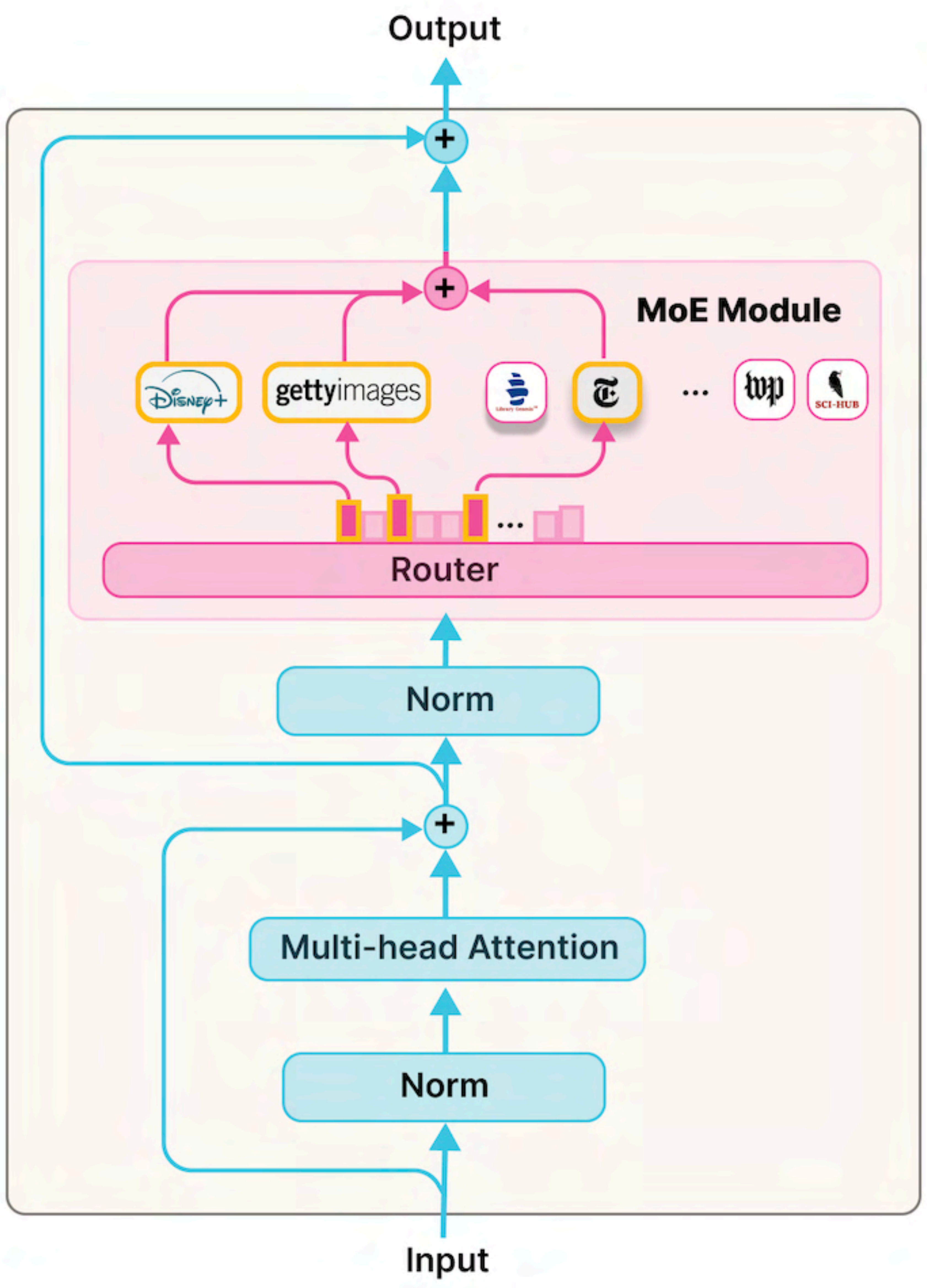
Proposal: Models with Data Provenance



Application
K-12 education

- ✓ **Copyright Takedowns**
- ✓ **Unlearning Private Data**
- ✓ **Safe deployment**

How?

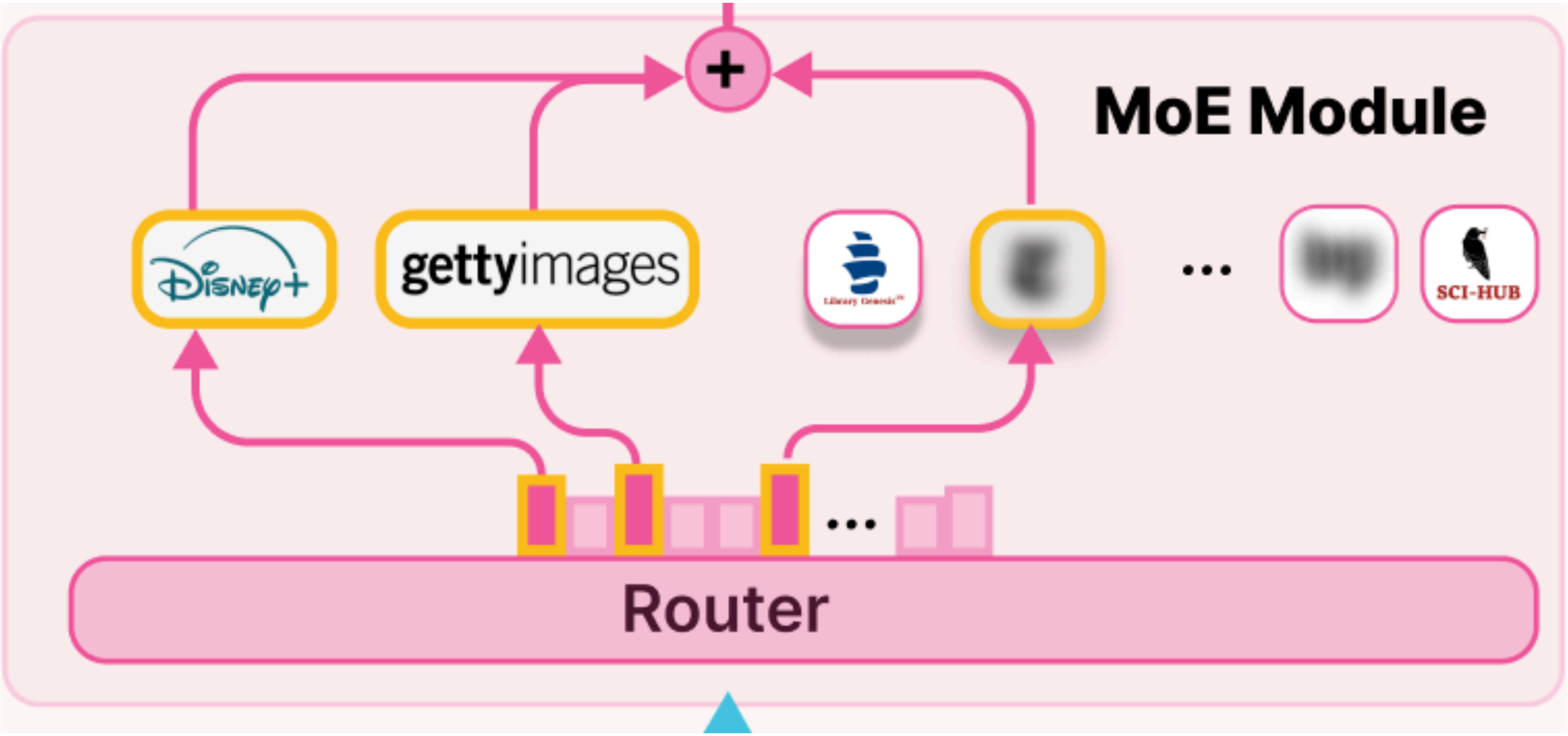


OLMoE

(Muennighoff et al.)

How?

Deactivation of modules during inference based on the query



Beyond Monolithic Language Models

Augmented Models

Data Modularity

Modularity, *not Monoliths*

Thank You!

📧 swj0419@uw.edu

🐦 @WeijiaShi2