UNIVERSITY of WASHINGTON

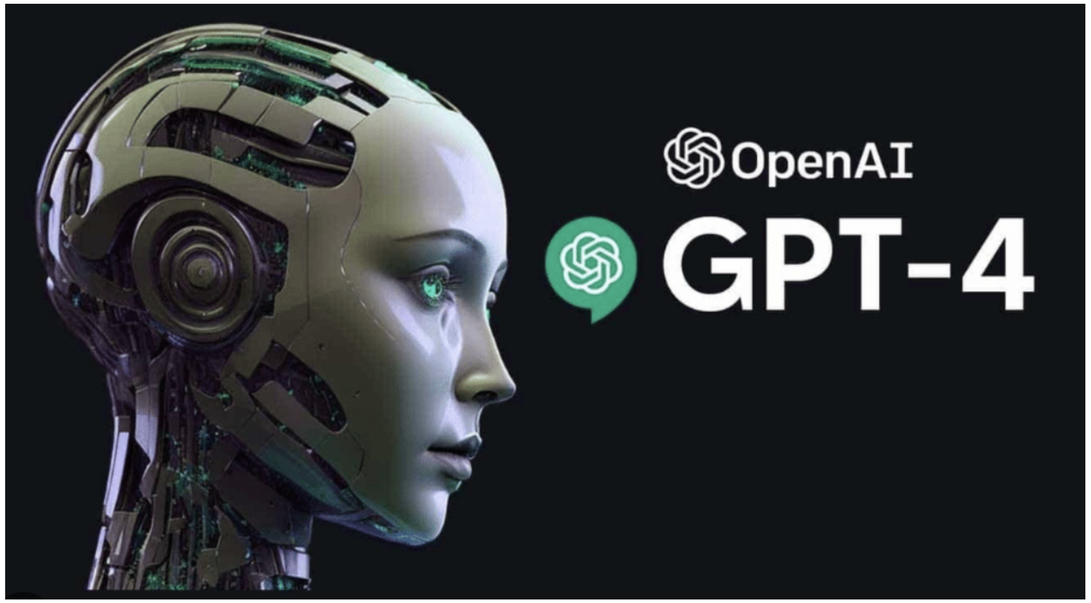
Memorization Dilemma of Language Models

Knowing What They Shouldn't, Yet Missing What They Should

Weijia Shi

swj0419.github.io/

Slides adapted from Akari Asai's tutorial on Retrieval-augmented Language Models (ACL 2023)



Benchmark	GPT-4 Evaluated few-shot
MMLU Multiple-choice questions in 57 subjects (professional & academic)	86.4% 5-shot
HellaSwag Commonsense reasoning around everyday events	95.3% 10-shot
AI2 Reasoning Challenge (ARC) Grade-school multiple choice science questions. Challenge-set.	96.3% 25-shot
WinoGrande Commonsense reasoning around pronoun resolution	87.5% 5-shot

Human-level intelligence?

Dataset Contamination

(Train on test subset unintentionally)

Horace He
@cHHillee

I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

This strongly points to contamination.

Which codeforces is AquaMoon and Two Arrays from?

AquaMoon and Two Arrays is a problem from Codeforces Round #731 (Div. 3) that took place on July 6, 2021. The problem can be found on the Codeforces platform by visiting the following link: https://codeforces.com/problemset/problem/1546/B

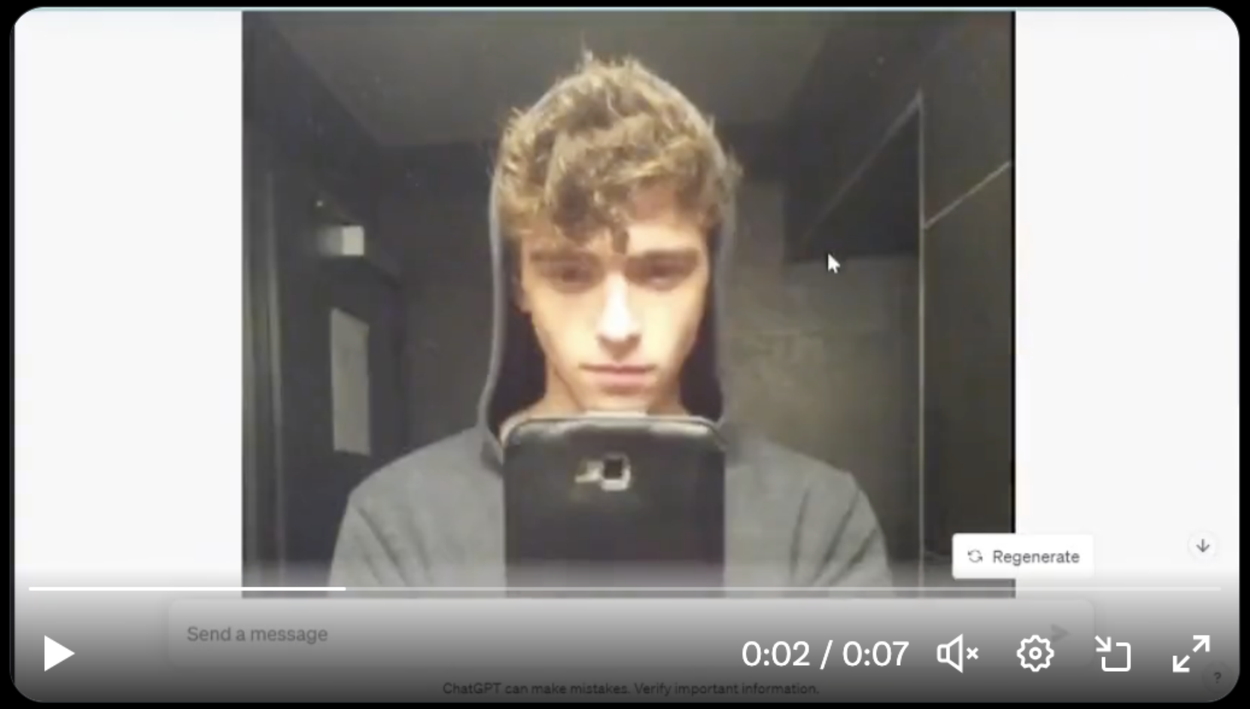
GPT-4 memorizes Codeforces problems before its training cutoff date.

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

A lawsuit by The New York Times could test the emerging legal contours of generative A.I. technologies.Sasha Maslov for The New York Times

Leer en español

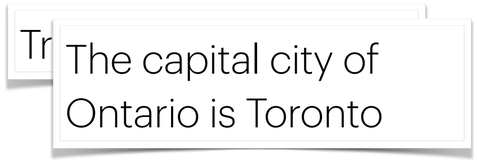
This Talk

Memorization Dilemma of Language Models

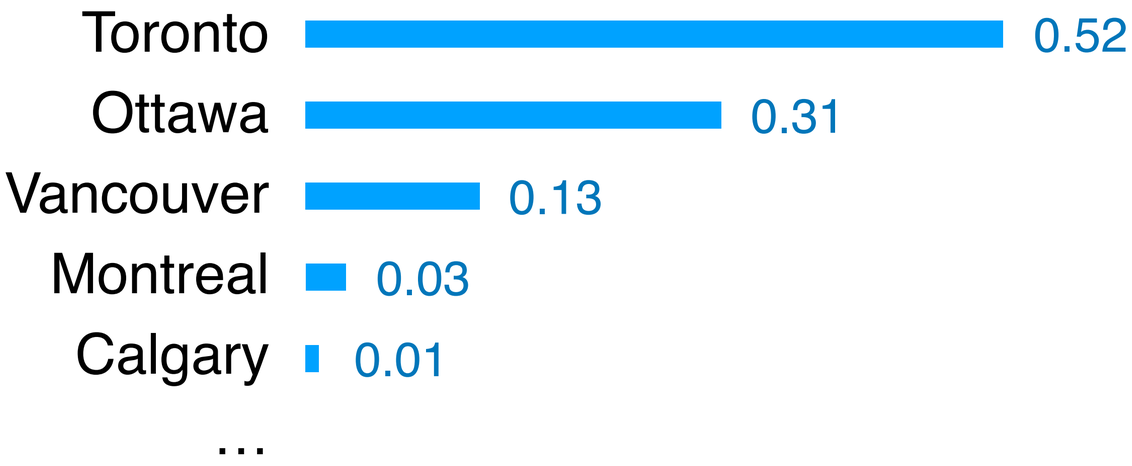
Knowing what they shouldn’t

How do such parametric LMs work?

$$P(x_n \mid x_1, x_2, \cdots, x_{n-1})$$



Large-scale pre-training corpus
(e.g., 1T tokens)



Language model (Transformers)

The capital city of Ontario is

x_1 x_2 … x_{n-1}

How do such parametric LMs work?

The capital city of Ontario is **Toronto**

LM

Training time

The capital city of Ontario is _____

LM

Test time

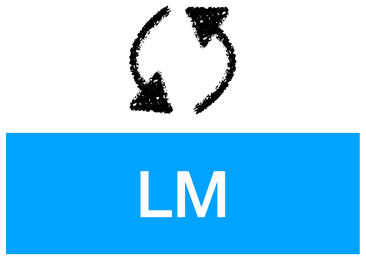
LMs know what they shouldn't?

Pretraining Corpus

Public data

Copyrighted, Private

Benchmark data



Training time

Dataset contamination

Copyright infringement

Privacy Risk

This Talk

Memorization Dilemma of Language Models

Knowing what they shouldn’t

This Talk

Memorization Dilemma of Language Models

Knowing what they shouldn't

Yet missing what they should

Hallucinations in LM outputs

List 5 important papers authored by Geoffrey Hinton

ChatGPT

Geoffrey Hinton is a renowned computer scientist …
Here are five important papers authored by him:

1. "**Learning Internal Representations by Error Propagation**" (with **D. E. Rumelhart** and **R. J. Williams**) - This paper, published in **1986**, ..

2. "**Deep Boltzmann Machines**" (with **R. Salakhutdinov**) - Published in **2009**, ..

…

4. "**Deep Learning**" (with Y. Bengio and A. Courville) - Published as a book in **2016**,…

5. "**Attention Is All You Need**" (with V. Vaswani, N. Shazeer, et al.) - Published in **2017**, this paper introduced the Transformer model,…

Catastrophic Errors as Results of LM Hallucinations

TECH · LAW

Humiliated lawyers fined $5,000 for submitting ChatGPT hallucinations in court: ‘I heard about this new site, which I falsely assumed was, like, a super search engine’

BY RACHEL SHIN
June 23, 2023 at 9:41 AM PDT

Lawyers who filed legal documents with false citations generated by ChatGPT have been fined.
ERIK MCGREGOR—LIGHTROCKET/GETTY IMAGES

Air Canada must honor refund policy invented by airline’s chatbot

Air Canada appears to have quietly killed its costly chatbot support.

ASHLEY BELANGER - 2/16/2024, 12:12 PM

This Talk

Memorization Dilemma of Language Models

Knowing what they shouldn't

Yet missing what they should

This Talk

Memorization Dilemma of Language Models

Knowing what they shouldn't

Yet missing what they should

↓

Detecting when it happens

↓

Solving the dilemma

This Talk

Memorization Dilemma of Language Models

Knowing what they shouldn't

Yet missing what they should

Detecting when it happens

Solving the dilemma

Detecting when LMs know what they should not know

Detecting Pretraining Data from Large Language Models
Shi *et al., ICLR 2024*

UNIVERSITY of WASHINGTON

PRINCETON UNIVERSITY

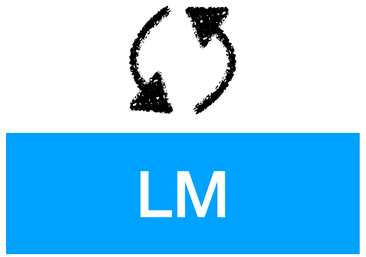
LMs know what they shouldn't?

Pretraining Corpus

Others

Copyrighted, Private

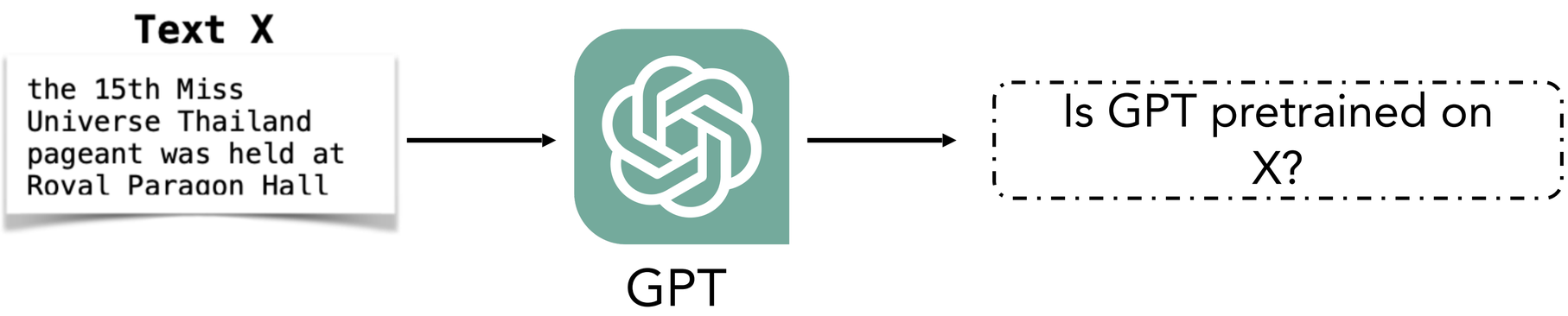
Benchmark data



Training time

Detect Pretraining Data from LLMs

Given a piece of text and black-box access to an LLM (only output logits), can we determine if the model was pretrained on the provided text



🕵️ Detecting ***copyrighted*** or ***private*** data from black-box LMs

Min-K% Prob


Text X
the 15th Miss Universe Thailand pageant was held at Royal Paragon Hall
Min-K% Prob
GPT-3.5
is pretrained on X

The 15th Miss Universe Thailand pageant was held at Roval Paragon Hall.

Member

pageant = 99.91%
Page = 0.09%
page = 0.00%
Page = 0.00%
= 0.00%

Total: -0.00 logprob on 1 tokens
(100.00% probability covered in top 5 logits)

Gemini is an AI model developed by Google. It is trained on a variety of data including video, images,

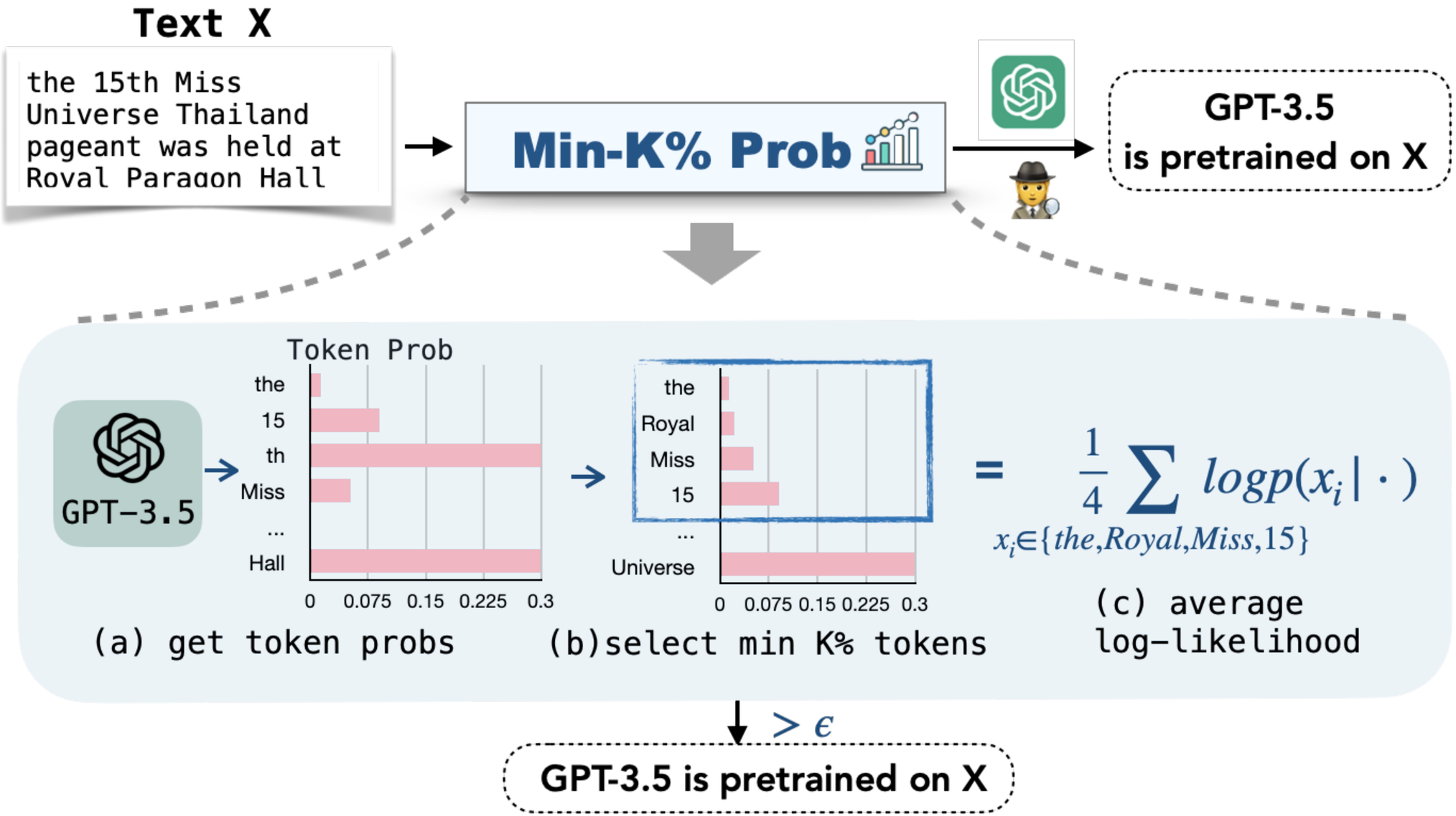
Non-member

designed = 36.98%
the = 17.95%
trained = 13.01%
a = 8.70%
capable = 6.39%

Total: -2.04 logprob on 1 tokens
(83.02% probability covered in top 5 logits)

A non-member tends to include outlier words with low prob

Our Method: Min-K% Prob



Detecting Copyrighted Books in Pretraining Data

Contamination %	Book Title	Author	Year
100	The Violin of Auschwitz	Maria Àngels Anglada	2010
100	North American Stadiums	Grady Chambers	2018
100	White Chappell Scarlet Tracings	Iain Sinclair	1987
100	Lost and Found	Alan Dean	2001
100	A Different City	Tanith Lee	2015
100	Our Lady of the Forest	David Guterson	2003
100	The Expelled	Mois Benarroch	2013
99	Blood Cursed	Archer Alex	2013
99	Genesis Code: A Thriller of the Near Future	Jamie Metzl	2014
99	The Sleepwalker's Guide to Dancing	Mira Jacob	2014
99	The Harlan Ellison Hornbook	Harlan Ellison	1990
99	The Book of Freedom	Paul Selig	2018
99	Three Strong Women	Marie NDiaye	2009
99	The Leadership Mind Switch: Rethinking How We Lead in the New World of Work	D. A. Benton, Kylie Wright-Ford	2017
99	Gold	Chris Cleave	2012
99	The Tower	Simon Clark	2005
98	Amazon	Bruce Parry	2009
98	Ain't It Time We Said Goodbye: The Rolling Stones on the Road to Exile	Robert Greenfield	2014
98	Page One	David Folkenflik	2011
98	Road of Bones: The Siege of Kohima 1944	Fergal Keane	2010

This Talk

Memorization Dilemma of Language Models

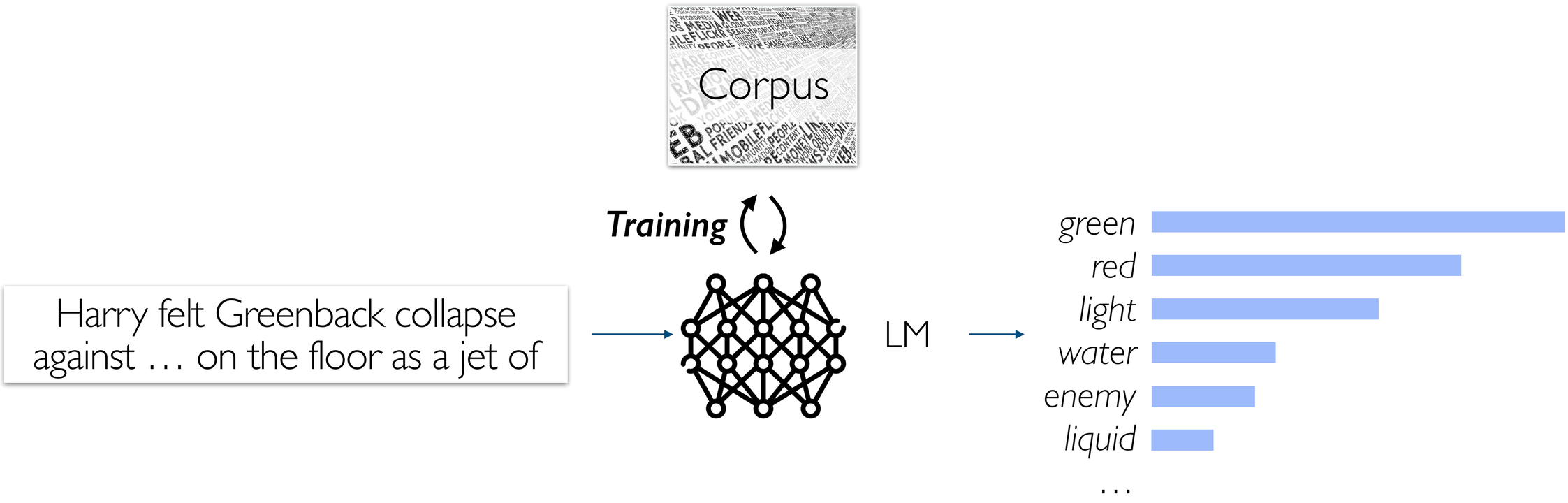
Knowing what they shouldn't

Yet missing what they should

Detecting when it happens

Solving the dilemma

Memorize knowledge in its parameters



Hallucinations in LM outputs

List 5 important papers authored by Geoffrey Hinton

ChatGPT

Geoffrey Hinton is a renowned computer scientist …
Here are five important papers authored by him:

1. "**Learning Internal Representations by Error Propagation**" (with **D. E. Rumelhart** and **R. J. Williams**) - This paper, published in **1986**, ..

2. "**Deep Boltzmann Machines**" (with **R. Salakhutdinov**) - Published in **2009**, ..

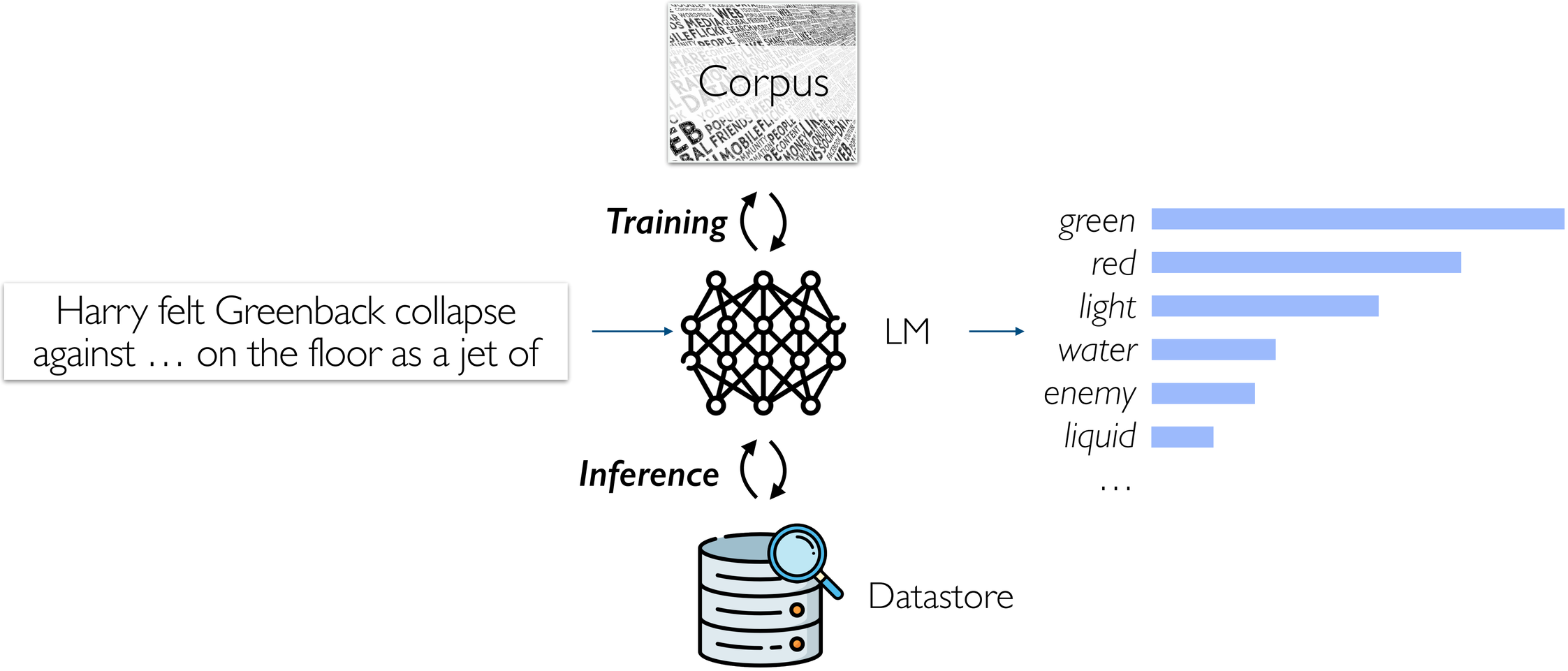
…

4. "**Deep Learning**" (with Y. Bengio and A. Courville) - Published as a book in **2016**,…

5. "**Attention Is All You Need**" (with V. Vaswani, N. Shazeer, et al.) - Published in **2017**, this paper introduced the Transformer model,…

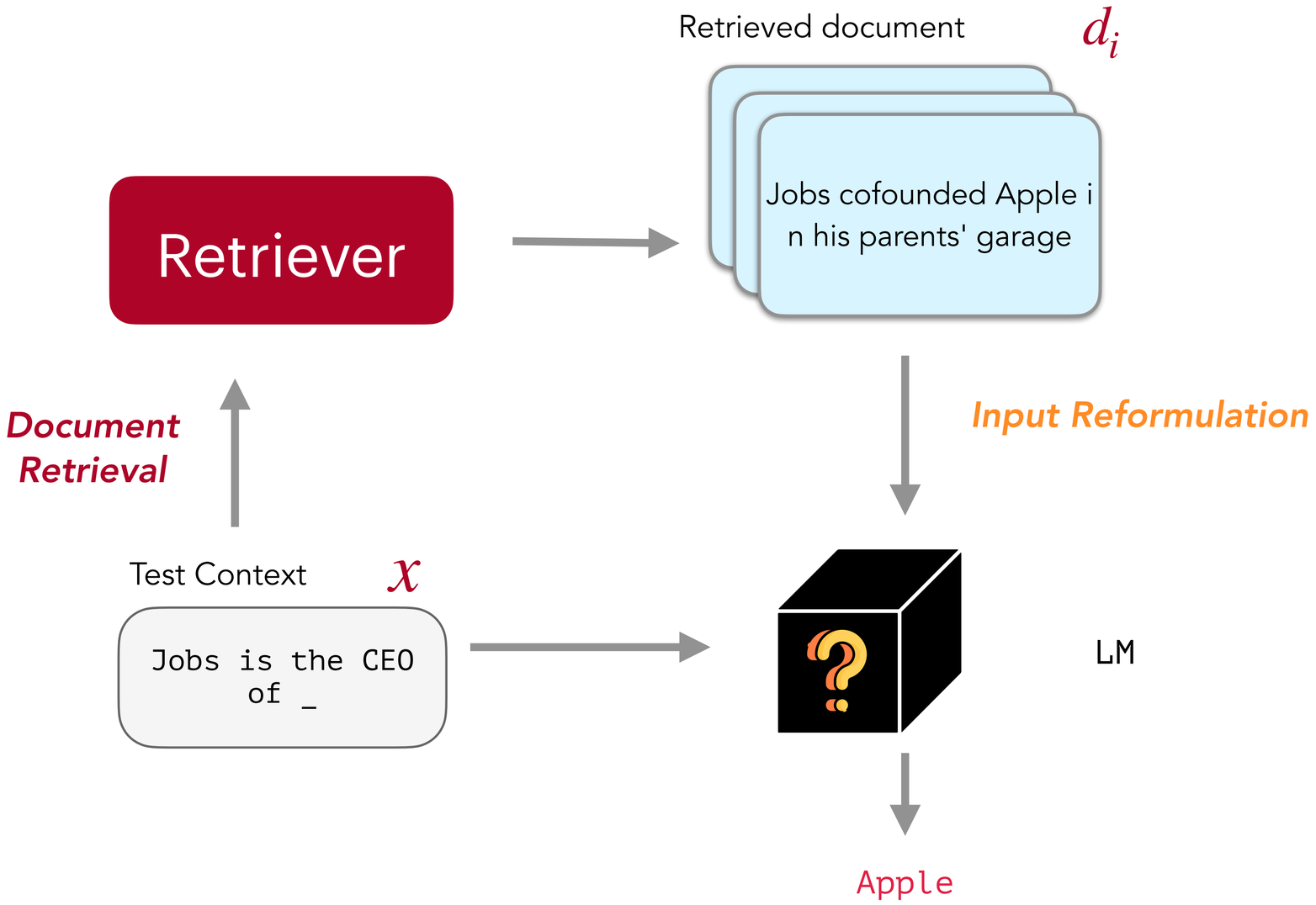
Memorize knowledge in its parameters

+ external knowledge during inference



Retrieval-augmented LM (RAGs)

Retrieval-augmented LM overview

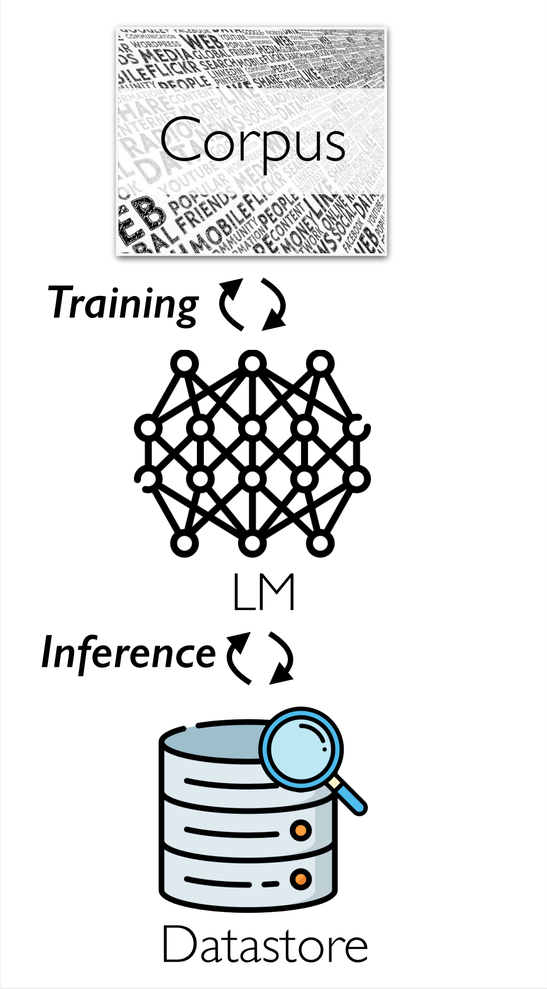


How RAGs solve the memorization dilemma?

Hallucinations

Privacy and copyright risks

Retrieval-based LM



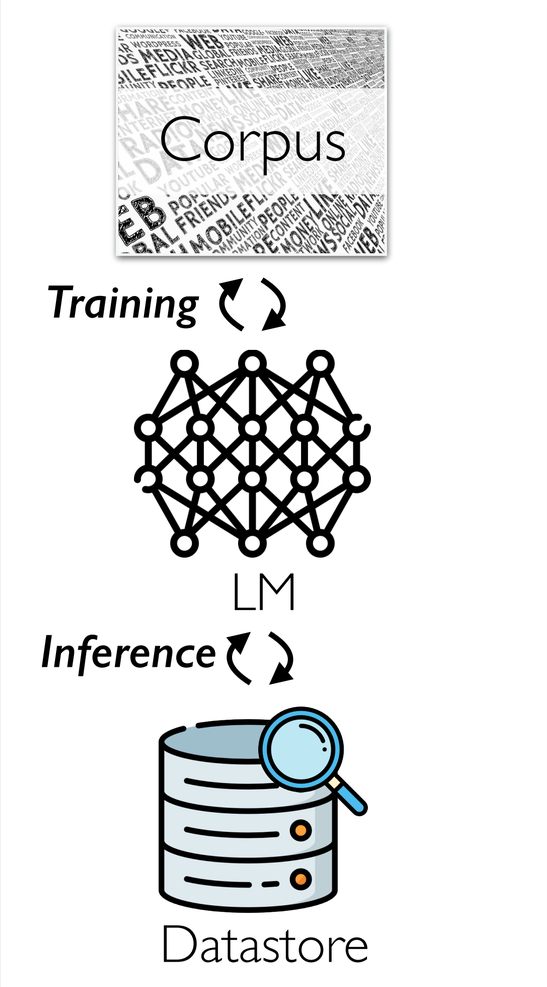
How RAGs solve the memorization dilemma?

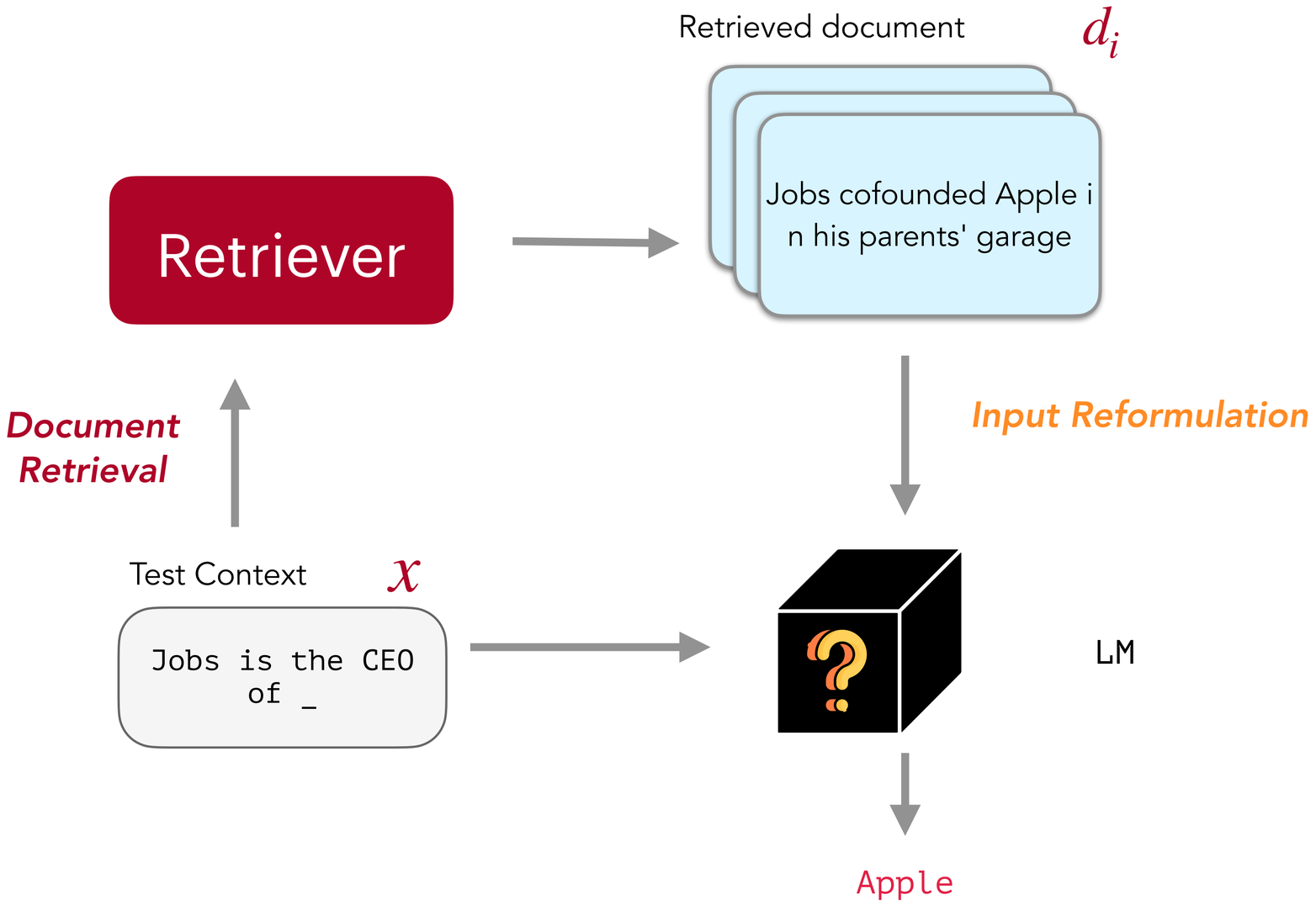
~~Hallucinations~~

Look up the datastore

Privacy and copyright risks

Retrieval-based LM




Retrieved document
d_i
Jobs cofounded Apple i
n his parents' garage
Retriever
Document
Retrieval
Input Reformulation
Test Context
x
Jobs is the CEO
of _
LM
Apple

How RAGs solve the memorization dilemma?

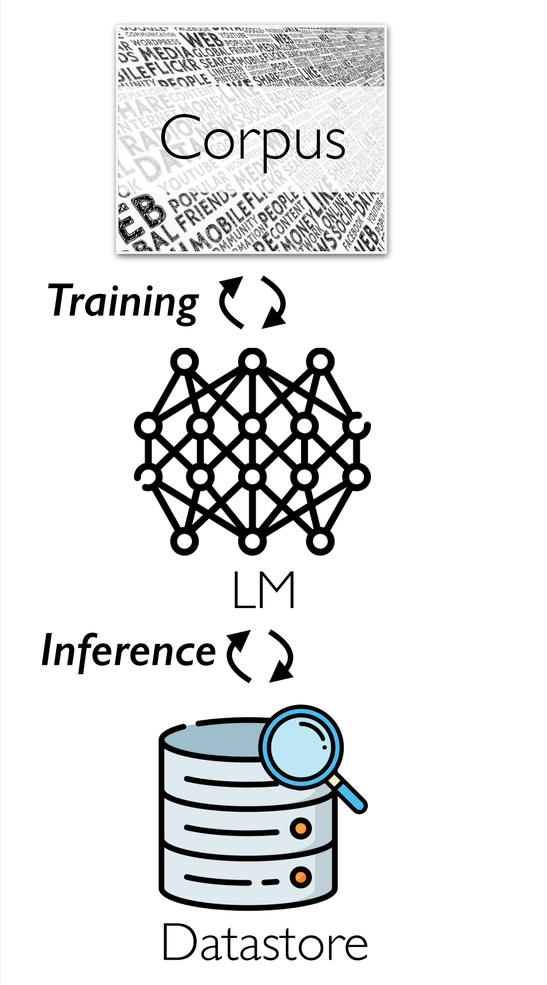
~~Hallucinations~~

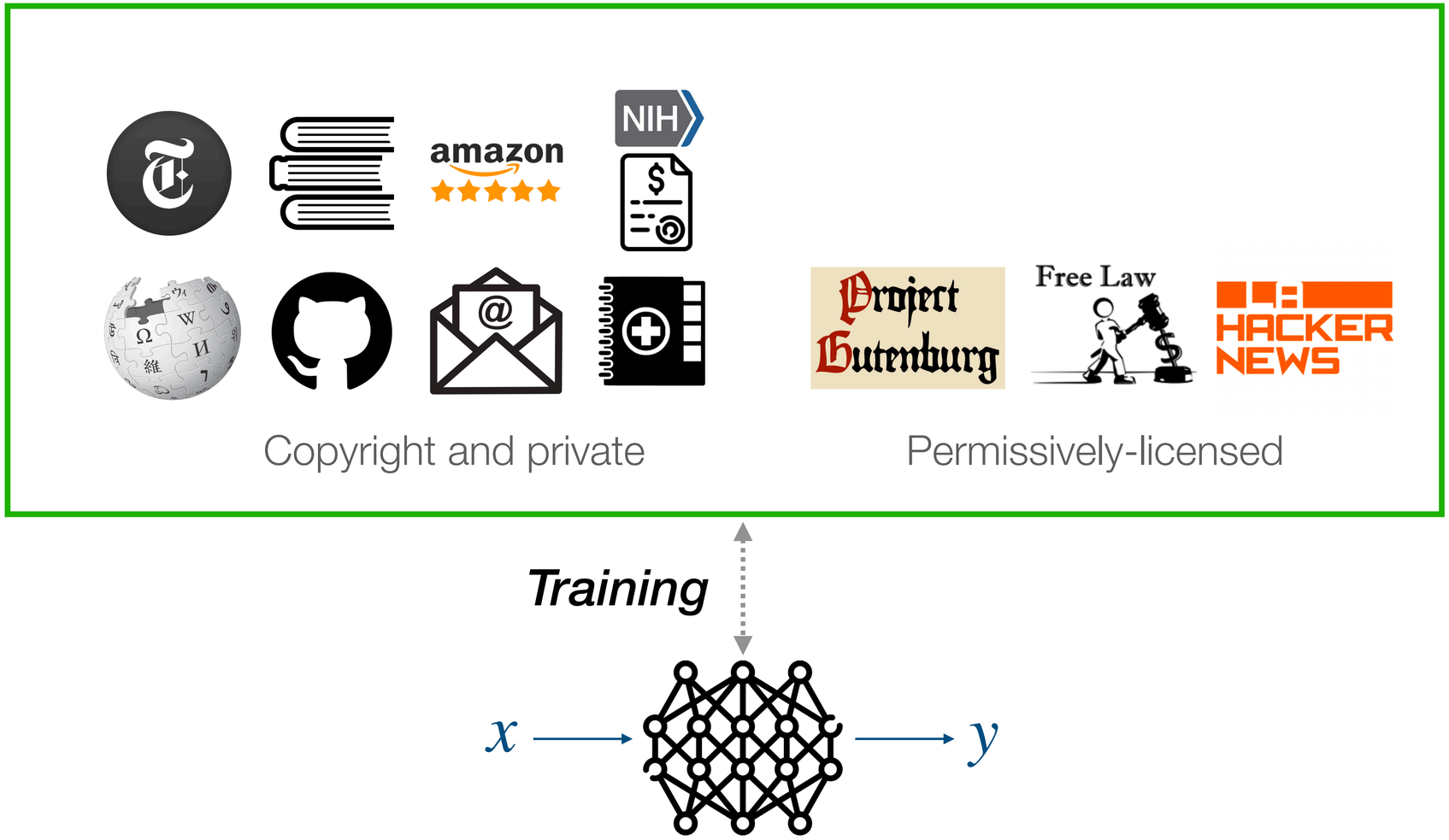
Look up the datastore

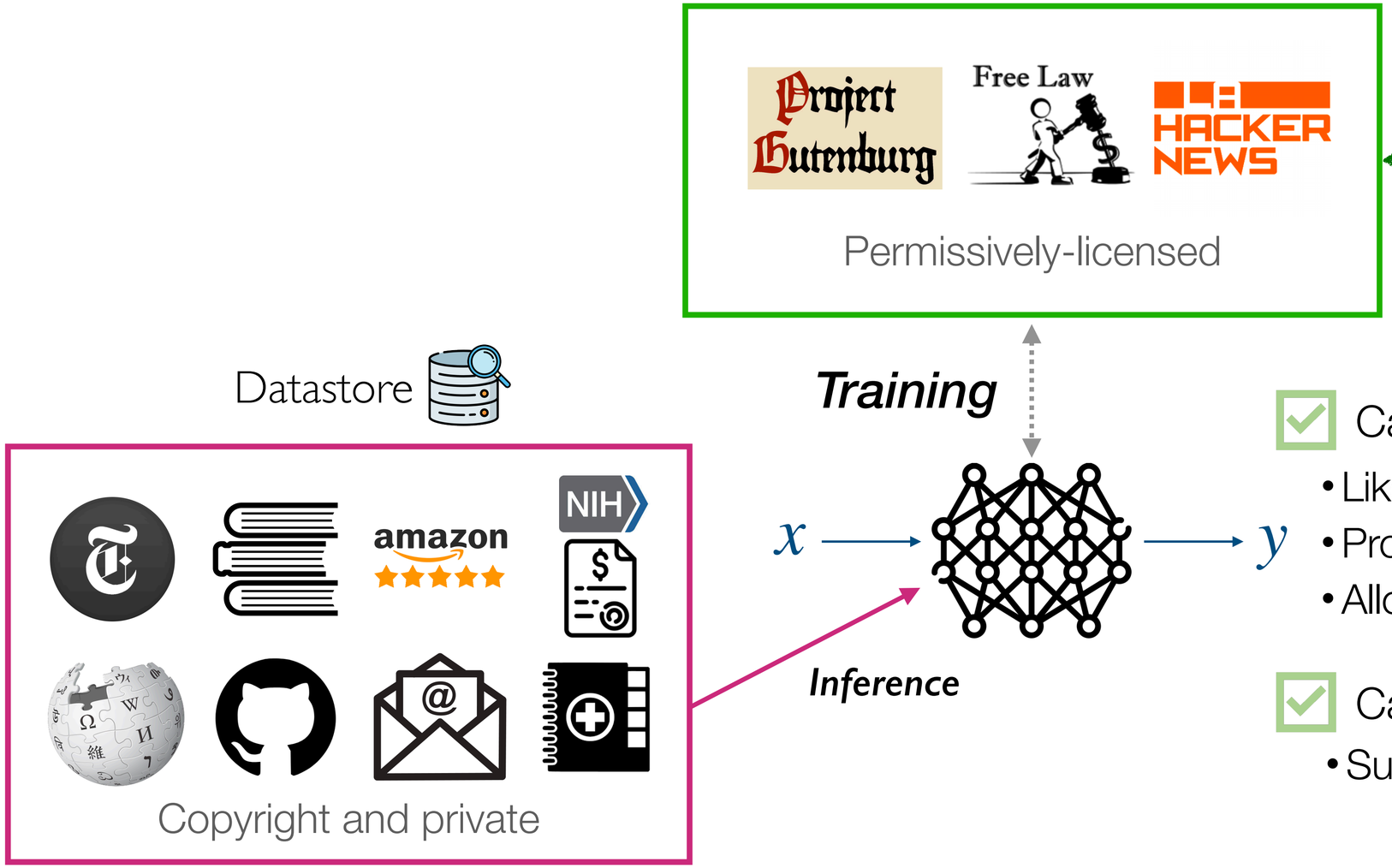
~~Privacy and copyright risks~~

Store the sensitive data in the datastore

Retrieval-based LM




amazon
NIH
Project Gutenburg
Free Law
HACKER NEWS
Copyright and private
Permissively-licensed
Training
x
y



Very low legal risk,
but poor performance
(small-size data, domain shift)

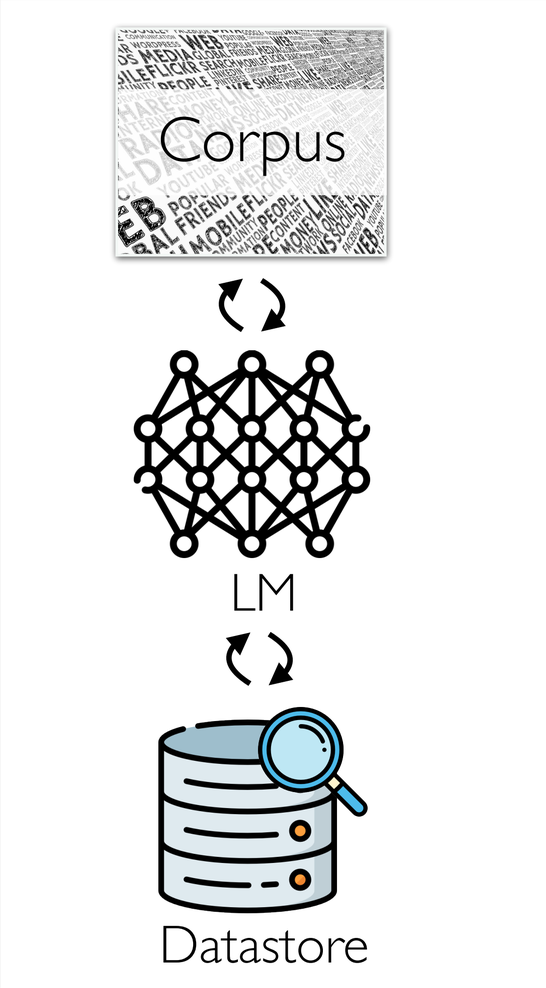
- ☑ Can trace inherent attribution
  - Likely defense *fair use*
  - Provide copyright notice
  - Allow credits (or payment) to data creators
- ☑ Can modify the datastore at any time
  - Support removal of data at any time

Min, Gururangan, Wallace, **Shi** et al. ICLR 2024, Spotlight. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

Retrieval-augmented LMs (RAGs)

1. Why do need RAGs?
2. Architectures of RAGs
3. Training of the retriever
4. Training of the LMs

Retrieval-based LM



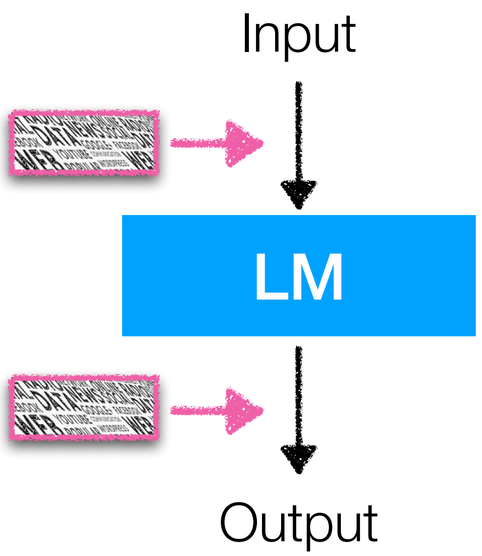
Categorization of retrieval-augmented LMs

What to retrieve?

Query

Text chunks (passages)?
Tokens?

How to use retrieval?



Two representative architectures

What: Text chunks
How: Input

Input augmentation

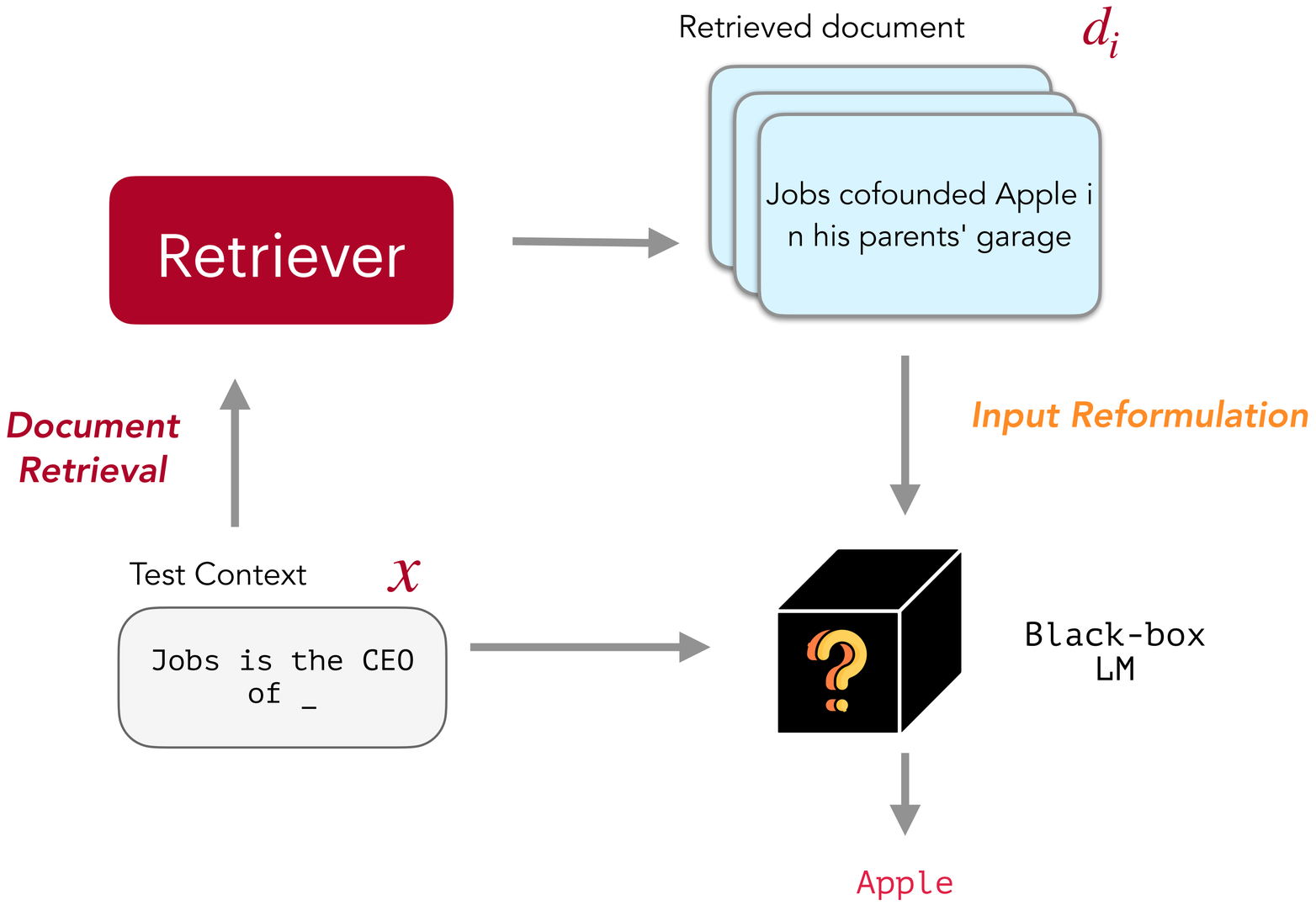
REALM (Guu et al., 2020)
REPLUG (Shi et al., 2023)

What: Tokens
How: Output

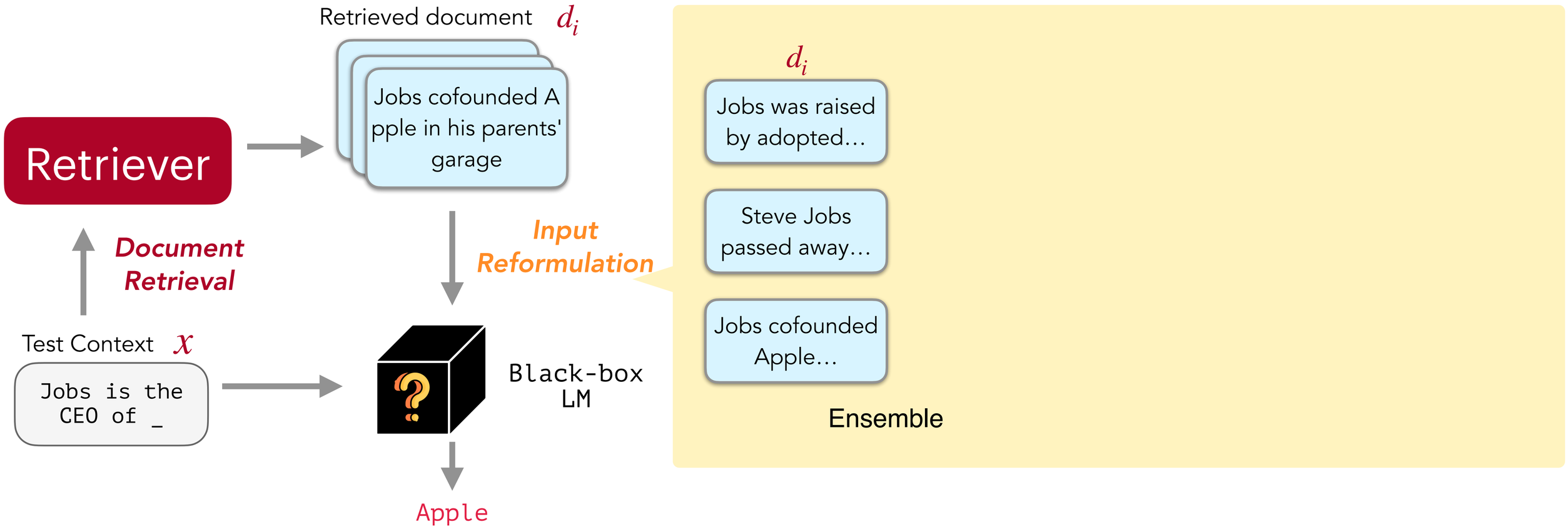
Output interpolations

kNN-LM, kNN-Prompt
(Khandelwal et al., 2020, Shi et al, 2023)

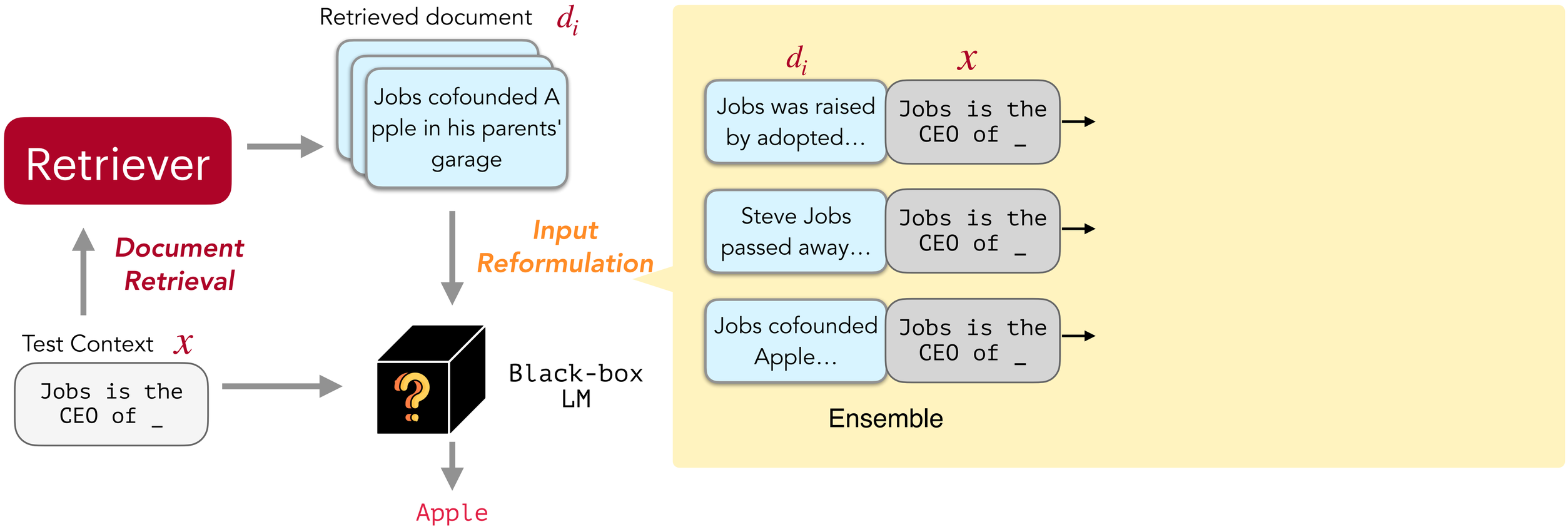
REPLUG (Shi et al 2023)



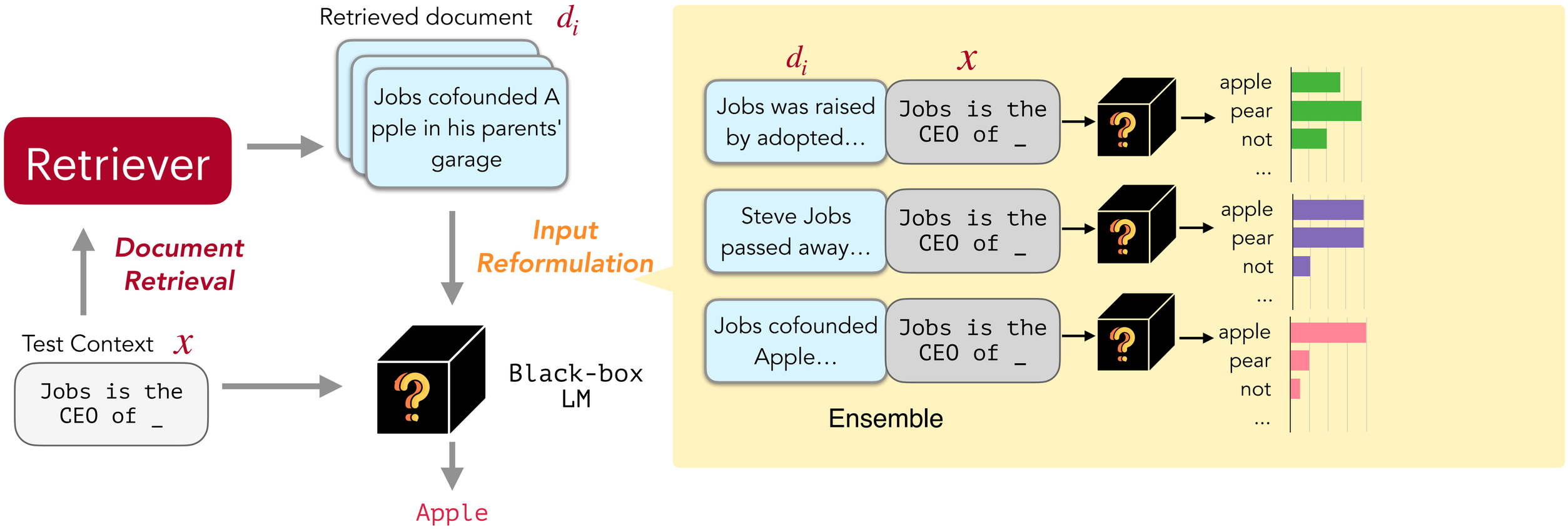
REPLUG inference: input reformulation



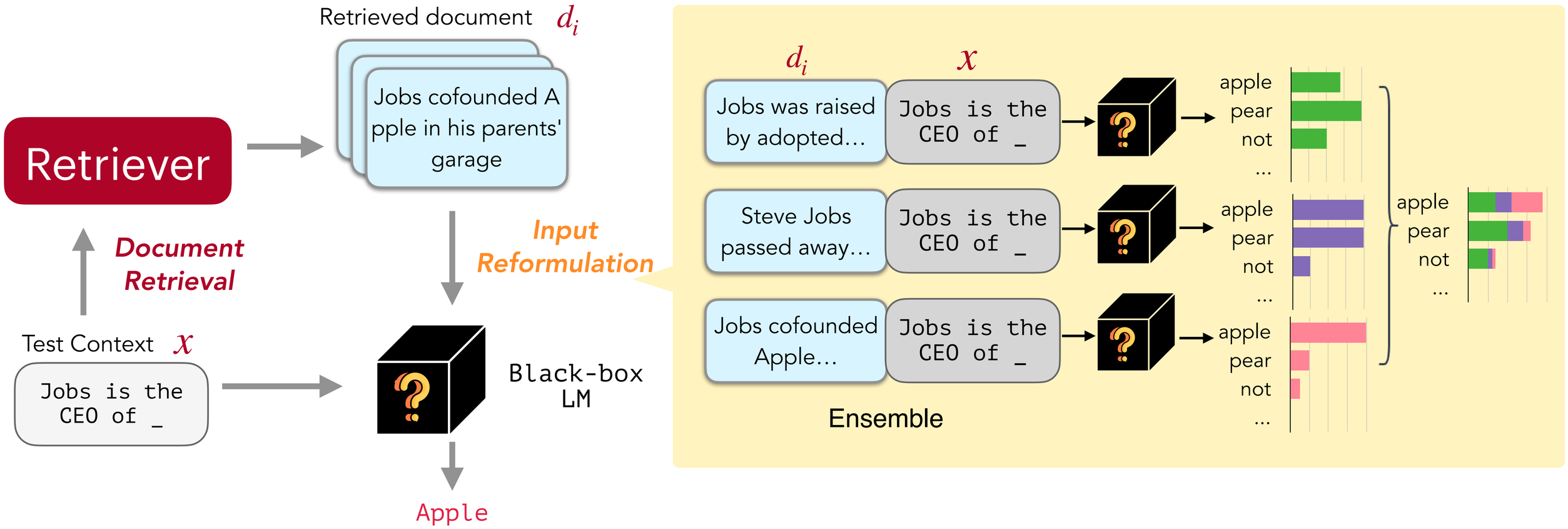
REPLUG inference: input reformulation



REPLUG inference: input reformulation



REPLUG inference: input reformulation



Two representative architectures

What: Text chunks
How: Input

Input augmentation

REALM (Guu et al., 2020)
REPLUG (Shi et al., 2023)

What: Tokens
How: Output

Output interpolations

kNN-LM, kNN-Prompt
(Khandelwal et al., 2020, Shi et al, 2022)

kNN-LM (Khandelwal et al. 2020)

- ✓ A different way of using retrieval, where the LM outputs a nonparametric distribution over every token in the data.
- ✓ Can be seen as an incorporation in the “output” layer

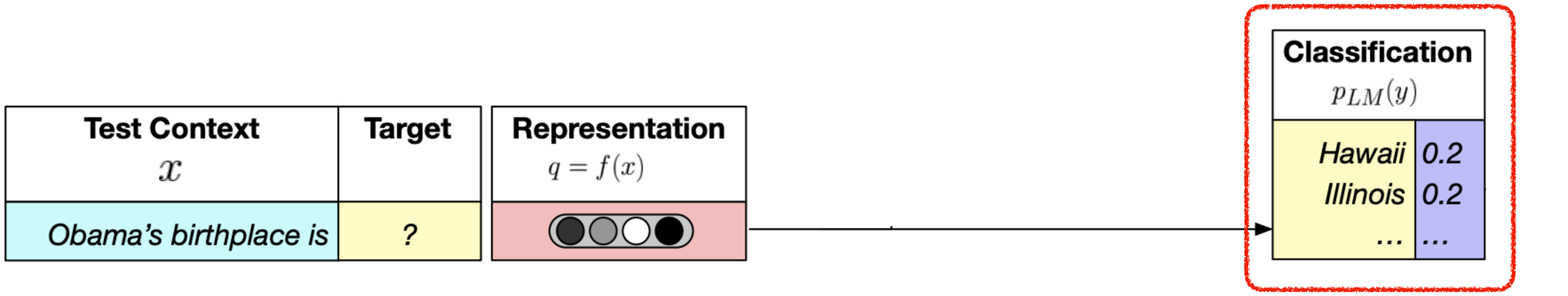
Khandelwal et al. Generalization through Memorization: Nearest Neighbor Language Models. ICLR 2020.

kNN-LM (Khandelwal et al. 2020)

Test Context x	Target
Obama's birthplace is	*?*

Khandelwal et al. Generalization through Memorization: Nearest Neighbor Language Models. ICLR 2020.

kNN-LM (Khandelwal et al. 2020)



Parametric distribution

Khandelwal et al. Generalization through Memorization: Nearest Neighbor Language Models. ICLR 2020.

kNN-LM (Khandelwal et al. 2020)

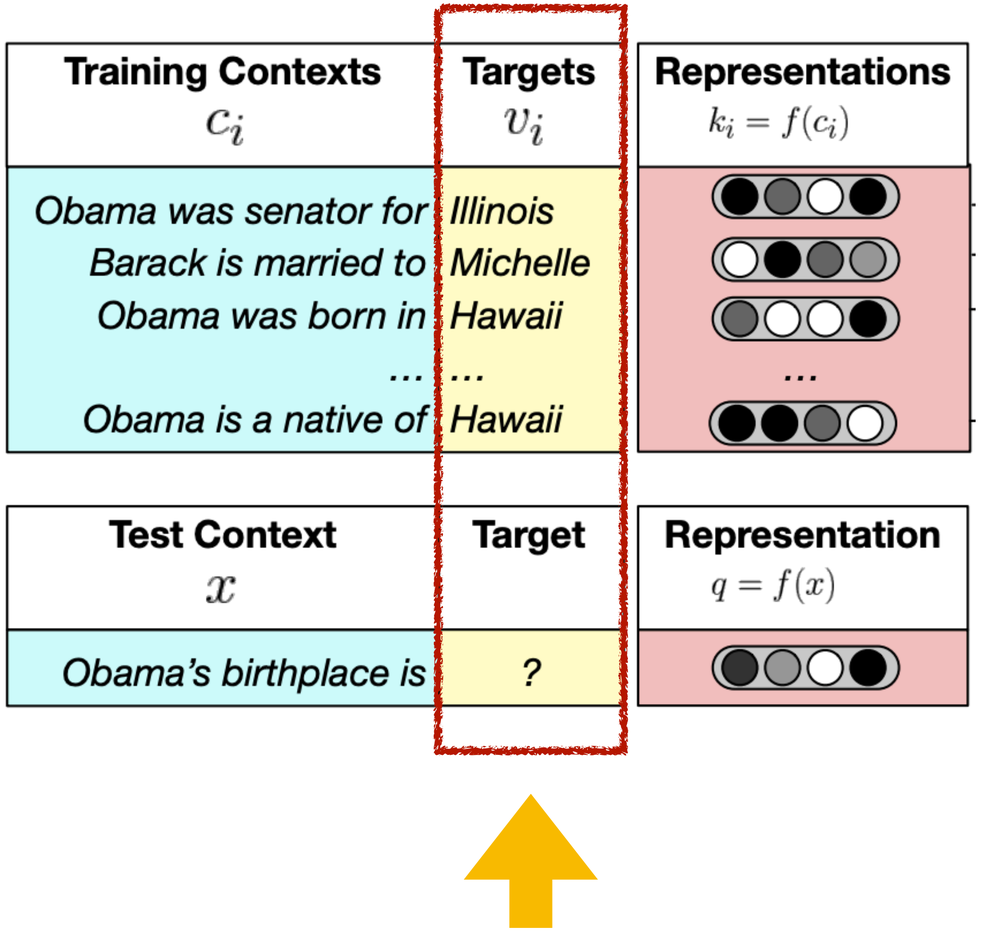
Training Contexts c_i	Targets v_i
Obama was senator for	*Illinois*
Barack is married to	*Michelle*
Obama was born in	*Hawaii*
...	*...*
Obama is a native of	*Hawaii*

... Obama was senator for Illinois from 1997 to 2005, ... Barack is Married to Michelle and their first daughter, ... Obama was born in Hawaii, and graduated from Columbia University. ... Obama is a native of Hawaii, ...

Test Context x	Target	Representation $q = f(x)$
Obama's birthplace is	*?*	

Khandelwal et al. Generalization through Memorization: Nearest Neighbor Language Models. ICLR 2020.

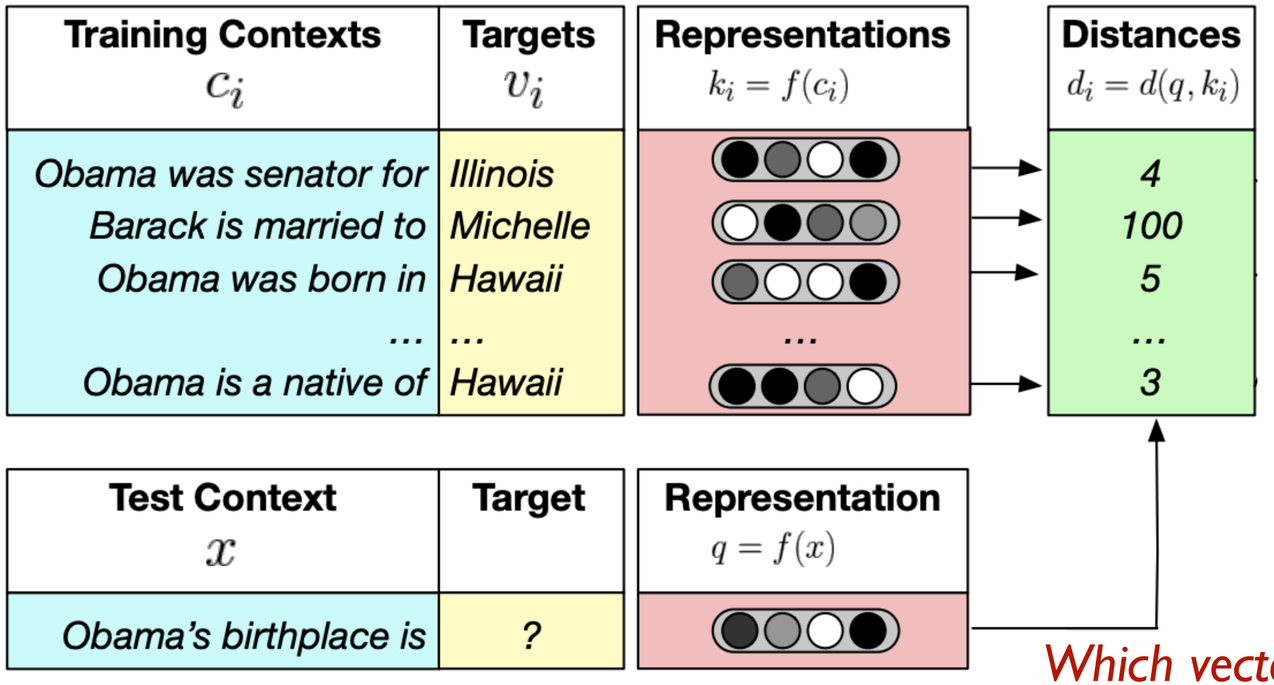
kNN-LM (Khandelwal et al. 2020)



Which tokens <u>in a datastore</u> are close to the next token?

Khandelwal et al. Generalization through Memorization: Nearest Neighbor Language Models. ICLR 2020.

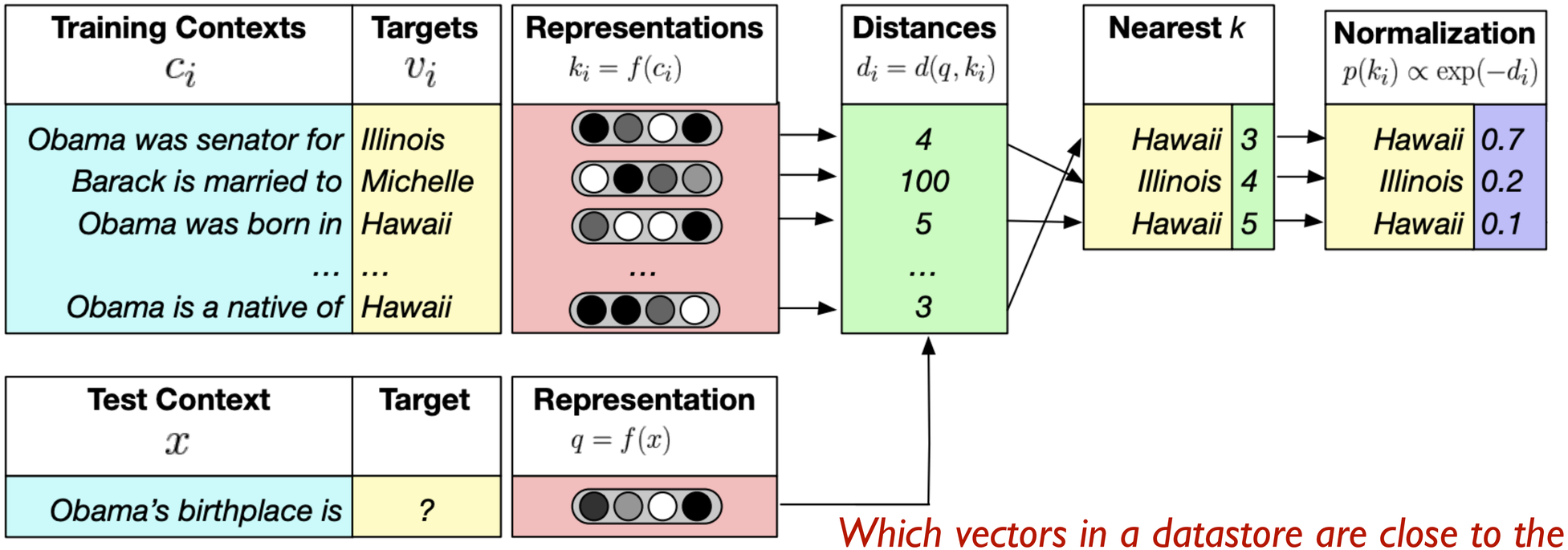
kNN-LM (Khandelwal et al. 2020)



Which vectors in a datastore are close to the vector we have?

Khandelwal et al. Generalization through Memorization: Nearest Neighbor Language Models. ICLR 2020.

kNN-LM (Khandelwal et al. 2020)

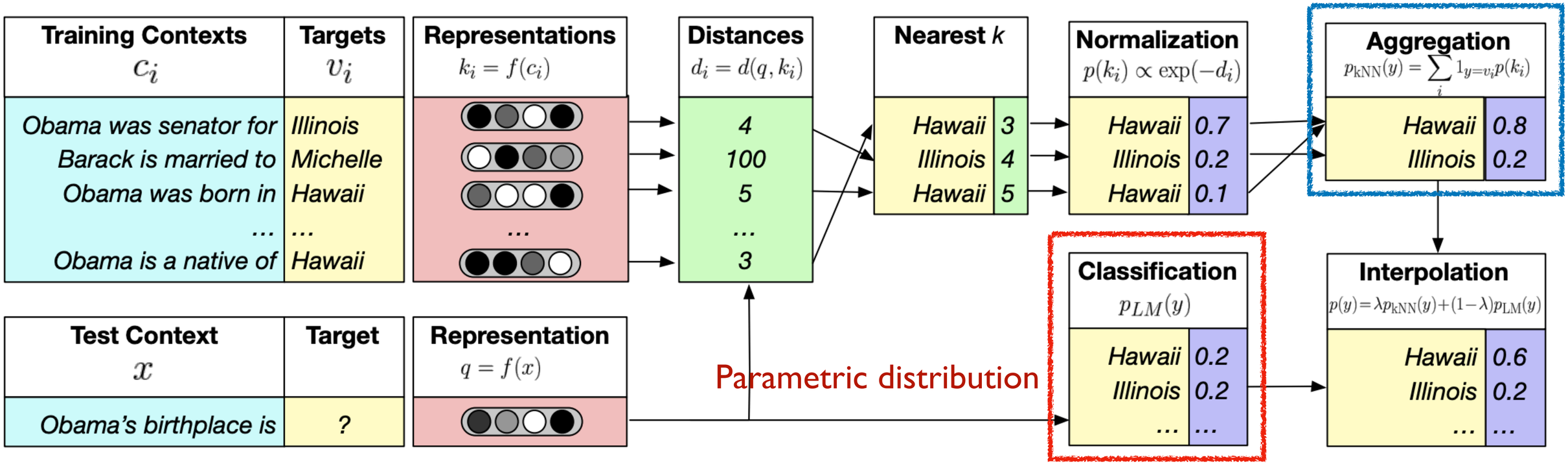


Which vectors in a datastore are close to the vector we have?

Khandelwal et al. Generalization through Memorization: Nearest Neighbor Language Models. ICLR 2020.

kNN-LM (Khandelwal et al. 2020)

Nonparametric distribution



λ: hyperparameter

$$P_{k\text{NN-LM}}(y \mid x) = (1 - \lambda) P_{\text{LM}}(y \mid x) + \lambda P_{k\text{NN}}(y \mid x)$$

Khandelwal et al. Generalization through Memorization: Nearest Neighbor Language Models. ICLR 2020.

Two representative architectures

What: Text chunks
How: Input

Input augmentation

REALM (Guu et al., 2020)
REPLUG (Shi et al., 2023)

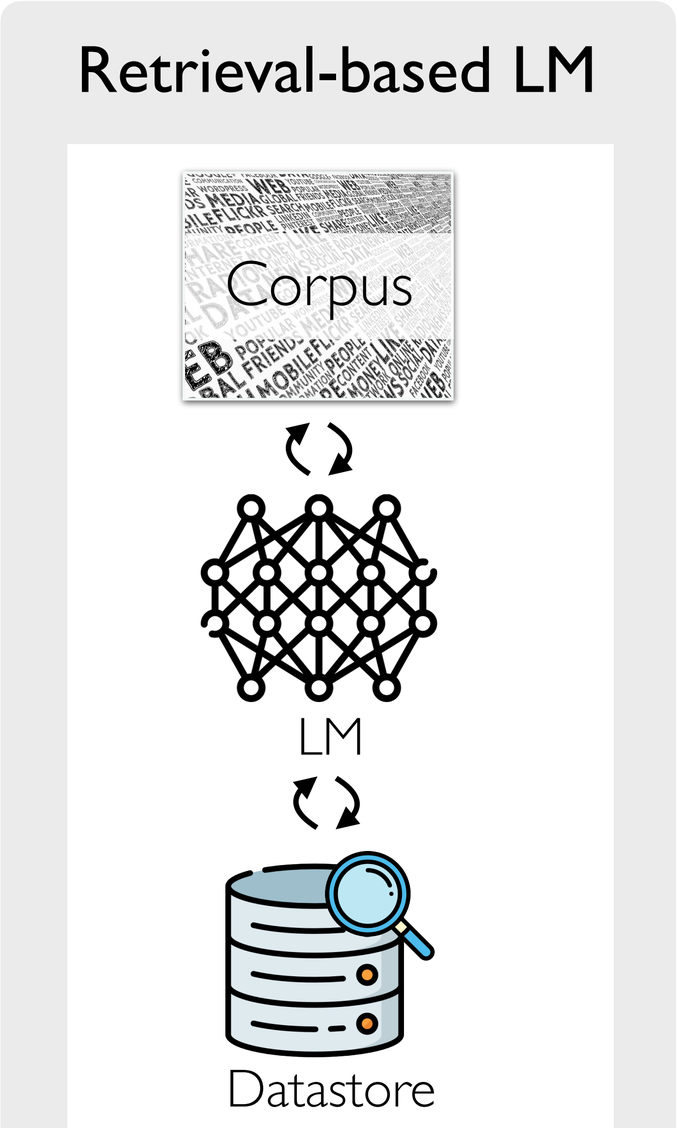
What: Tokens
How: Output

Output interpolations

kNN-LM, kNN-Prompt
(Khandelwal et al., 2020, Shi et al, 2022)

Retrieval-augmented LMs (RAGs)

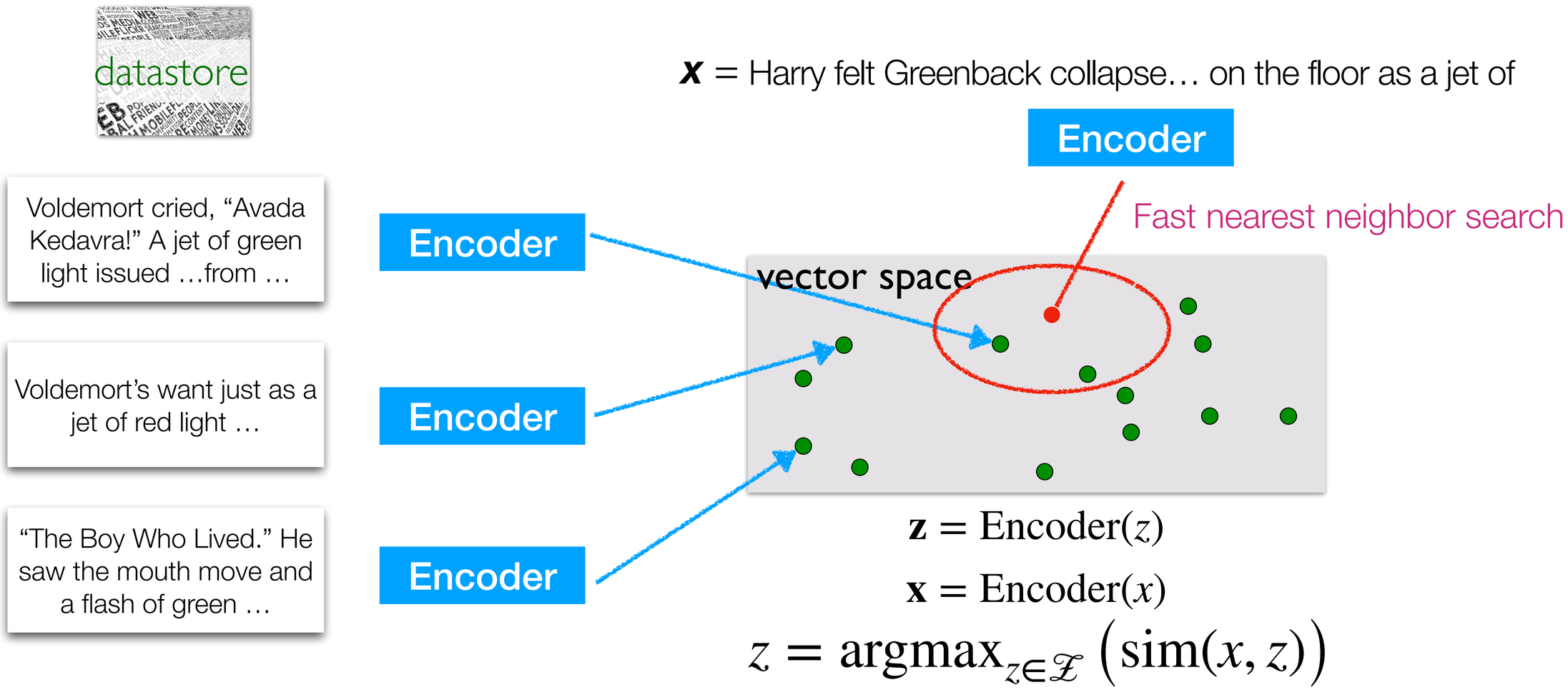
1. Why do need RAGs?
2. Architectures of RAGs
3. Training of the retriever
4. Training of the LMs



Training of the retriever

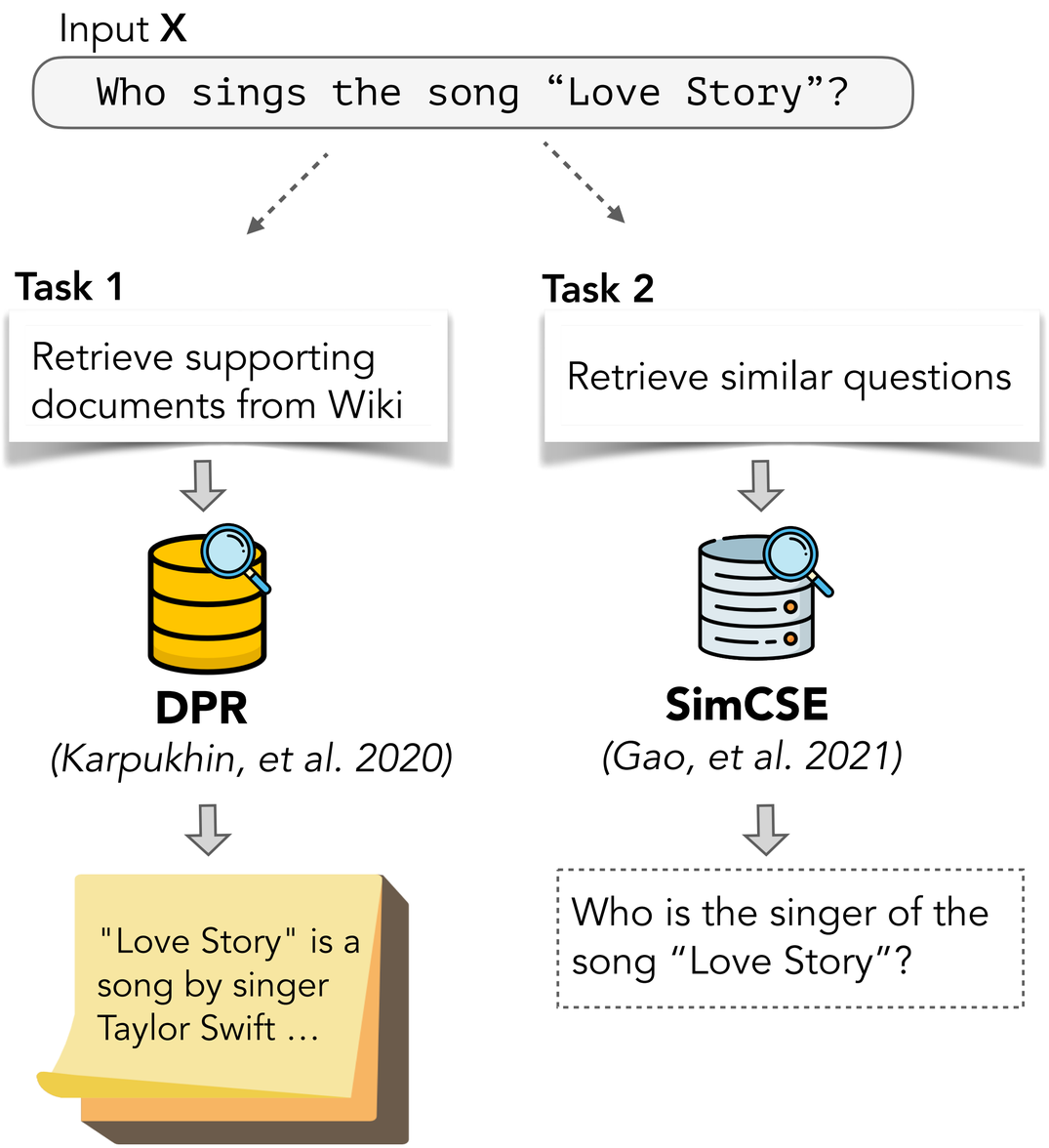
One Embedder, Any Task 👨‍🏫: Instruction-Finetuned Text Embeddings
Su, **Shi*** et al., 2023*

Dense retriever overview



Source: Retrieval-based Language Models Tutorial

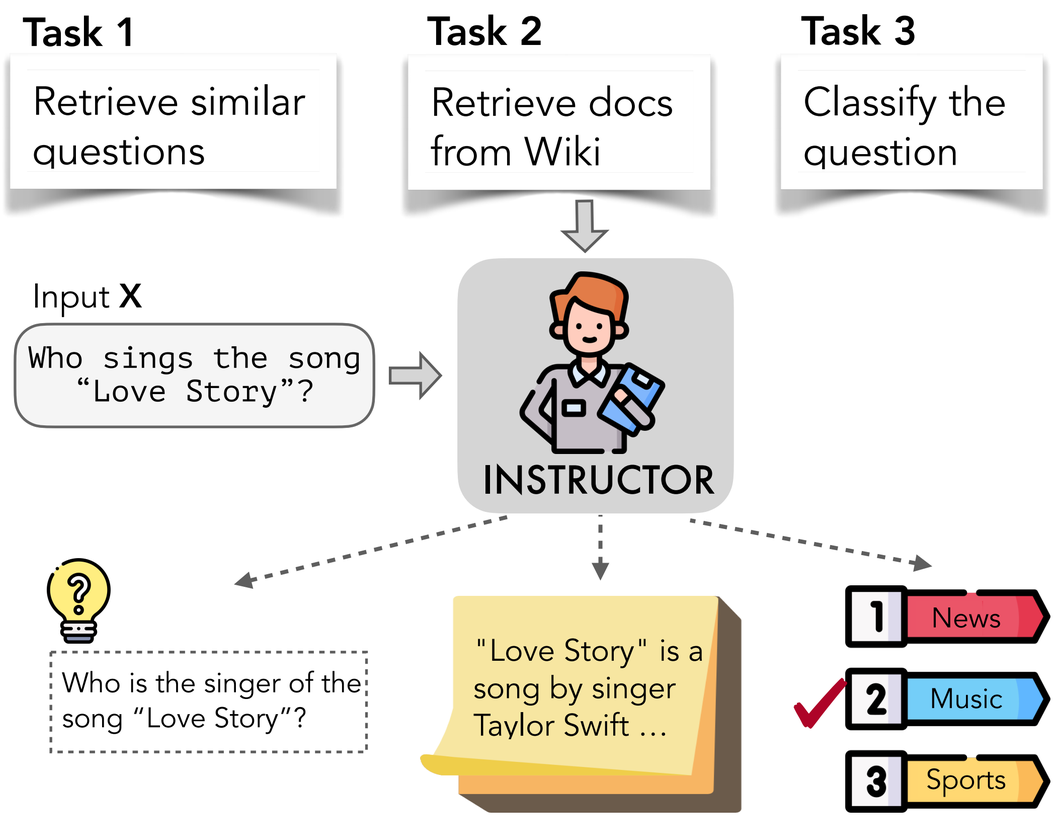
Previous task-specific retriever



Problem

Existing ***task-specific*** retrievers struggle to generalize to new tasks

Our customizable retriever: Instructor 👨‍🏫



Problem

Existing ***task-specific*** retrievers struggle to generalize to new tasks

Our approach

***A customizable* retriever** tailored to any task without further training

Instructor inference

Simply write an instruction

Instructor benefits

Efficient and simple: task-aware embeddings without any further training by simply providing the task instruction

Training

MEDI 330 datasets

Text Similarity

Measuring the similarity between sentences:

How can I be a good geologist?
What should I do to be a great geologist?

Question Answering

Retrieve documents that can help answer the question:

Why do rockets look white?

...

Fact Checking

Find documents that can help verify the fact:

The Ten Commandments is an epic film.

Sentiment Analysis

Classify the sentiment of the sentence:

You should see their decadent dessert menu

Train →

Trained on 330 datasets

Instruction format

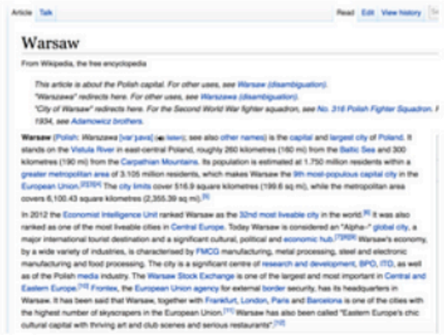
Our instruction format contains three elements:

text type, task objective, domain

Datasets	Instruction
Natural Question	Encode the Wikipedia question to retrieve supporting documents

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Document Retriever



Instruction format

Our instruction format contains three elements:

text type, task objective, domain

Datasets	Instruction
Natural Question	Encode the Wikipedia question to retrieve supporting documents
SummEval	Encode the Biomedical summary to retrieve duplicate summaries
IMDB Classification	Encode the movie review to classify emotions as positive or negative

MEDI: large-scale Instruction finetuning datasets

330 datasets

MEDI

Text Similarity

Measuring the similarity between sentences:

How can I be a good geologist?
What should I do to be a great geologist?

Question Answering

Retrieve documents that can help answer the question:

Why do rockets look white?

...

Fact Checking

Find documents that can help verify the fact:

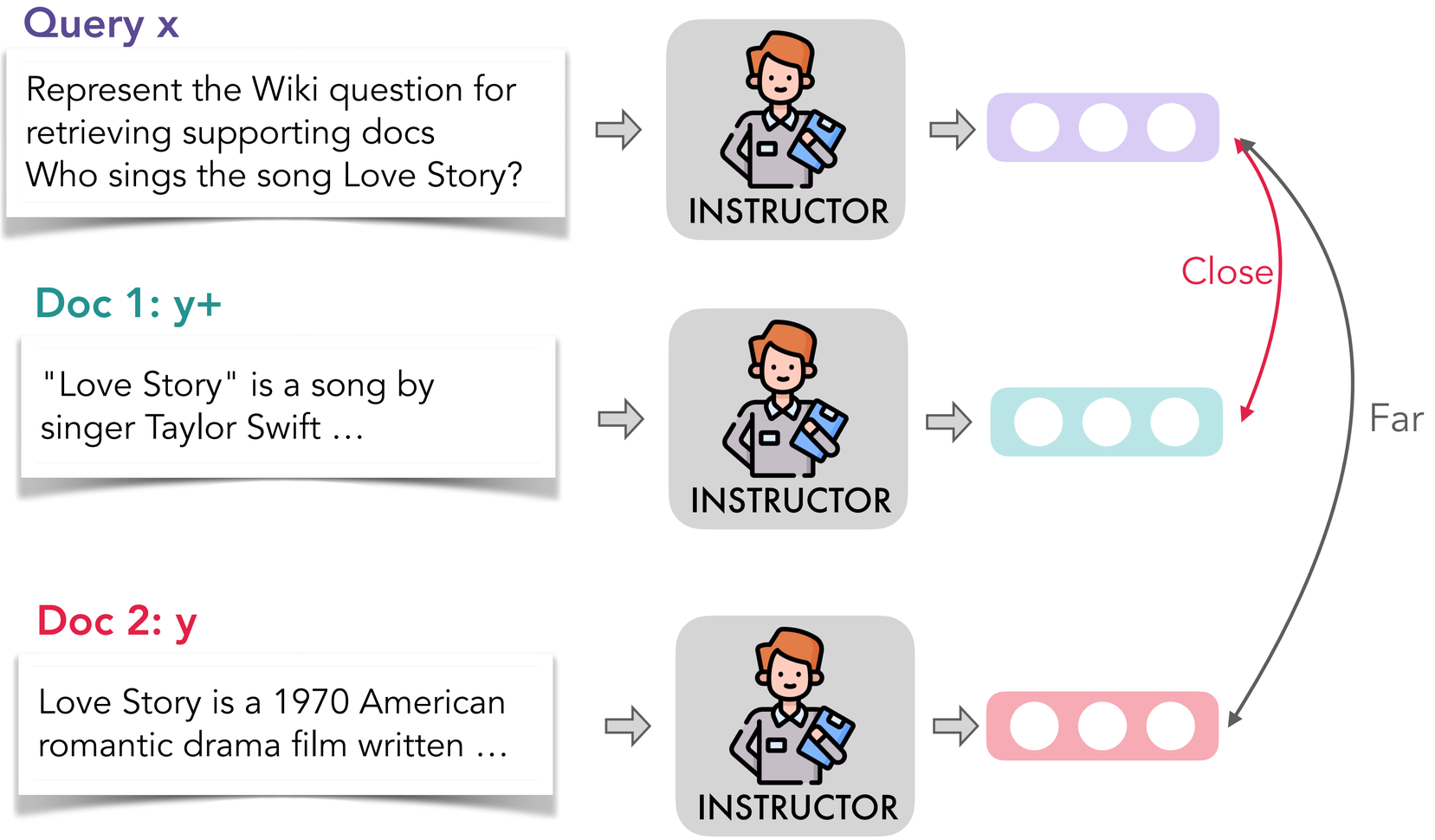
The Ten Commandments is an epic film.

Sentiment Analysis

Classify the sentiment of the sentence:

You should see their decadent dessert menu

Training the Retriever with MEDI



$$L = \frac{e^{s(x,y^+)/\gamma}}{\sum_{y \in \{y^+, y, \ldots\}} e^{s(x,y)/\gamma}}$$

Evaluation

MEDI 330 datasets

Text Similarity

Measuring the similarity between sentences:

How can I be a good geologist?
What should I do to be a great geologist?

Question Answering

Retrieve documents that can help answer the question:

Why do rockets look white?

...

Fact Checking

Find documents that can help verify the fact:

The Ten Commandments is an epic film.

Sentiment Analysis

Classify the sentiment of the sentence:

You should see their decadent dessert menu

Trained on 330 datasets

Train → INSTRUCTOR → Eval

70 datasets

Text Evaluation 4 datasets

Prompt Retrieval 11 datasets

Classification 12 datasets

Retrieval 15 datasets

Semantic Similarity 10 datasets

Reranking 4 datasets

Pair Classification 3 datasets

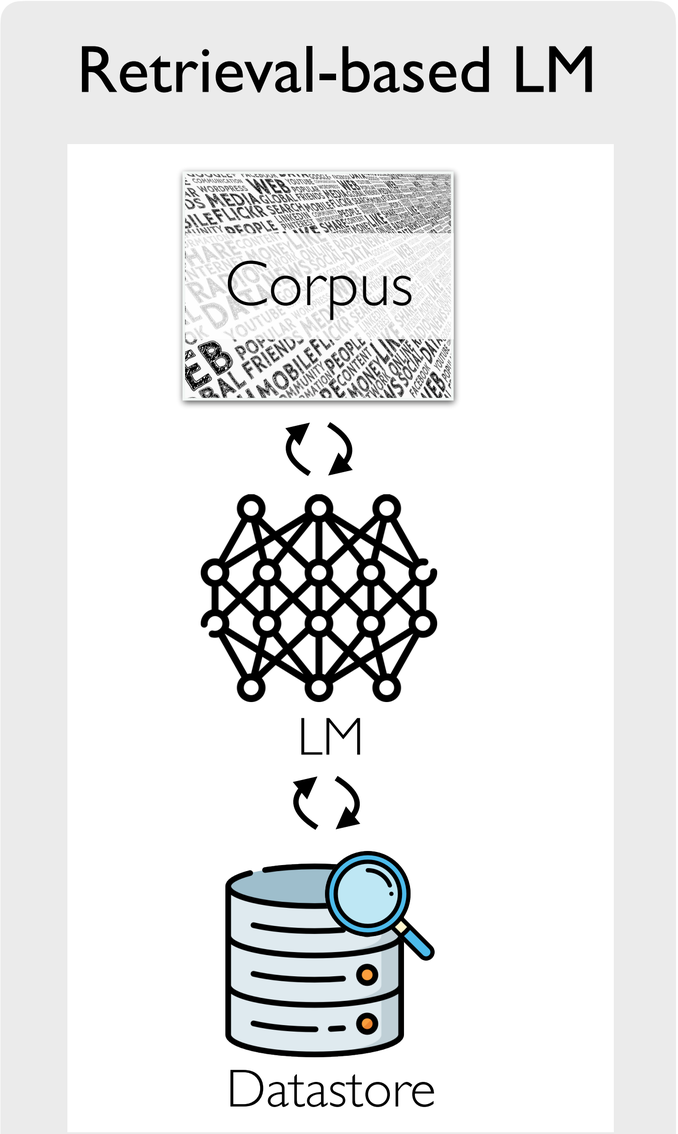
Clustering 11 datasets

⬆️ 7% compared with the best baseline

Evaluation

Retrieval-augmented LMs (RAGs)

1. Why do need RAGs?
2. Architectures of RAGs
3. Training of the retriever
4. **Training of the LMs**



Training of LMs for RAG

In-Context Pretraining: Language Modeling Beyond Document Boundaries
Shi *et al., ICLR 2024 Spotlight*

UNIVERSITY of WASHINGTON Meta AI

Input Context

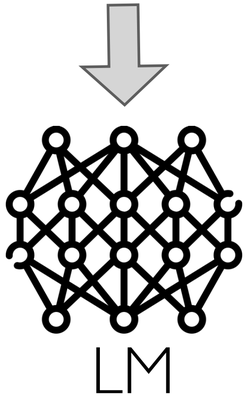
```
Write a high-quality answer for the given question using only the provided search
results (some of which might be irrelevant).

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:
```

Problem

LMs fail to use information in the context



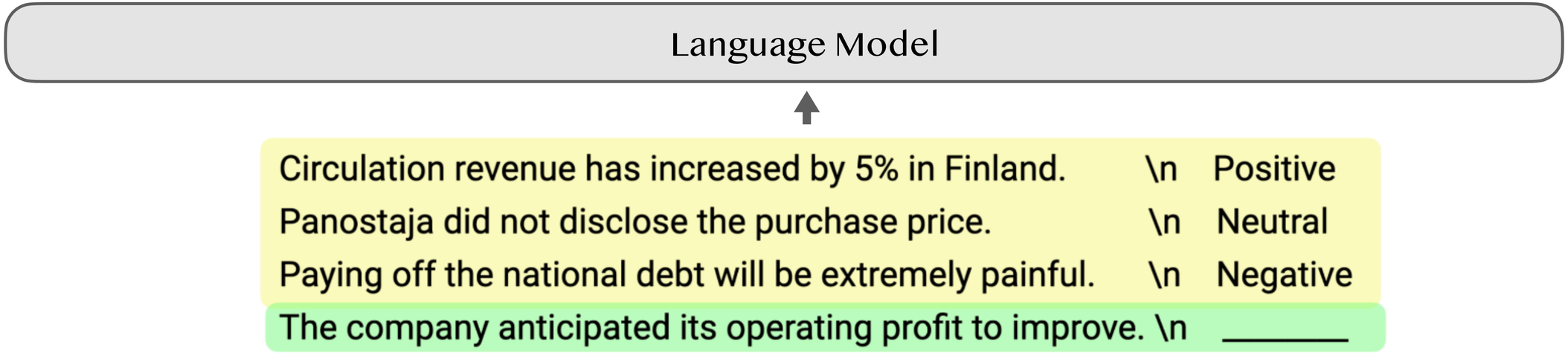
Ramon y Cajal ✗

(Liu et al., 2023)

Lack of context understanding

It impacts:

1. Retrieval augmentation
2. In-context learning

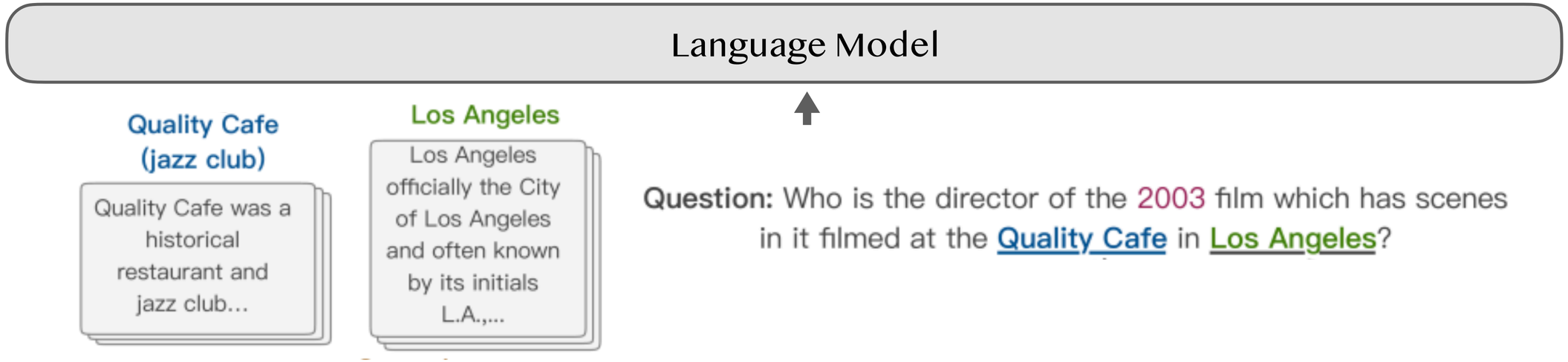


In-context learning

Lack of context understanding

It impacts:

1. Retrieval augmentation
2. In-context learning
3. Multidocument reasoning



Multidocument reasoning

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:

Problem

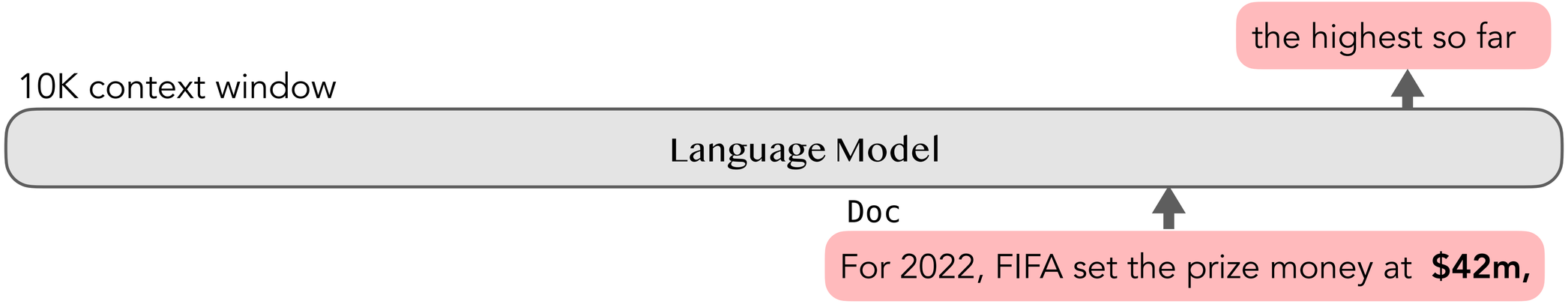
LMs fail to understand information in the context

Why does it happen?

(Liu et al., 2023)

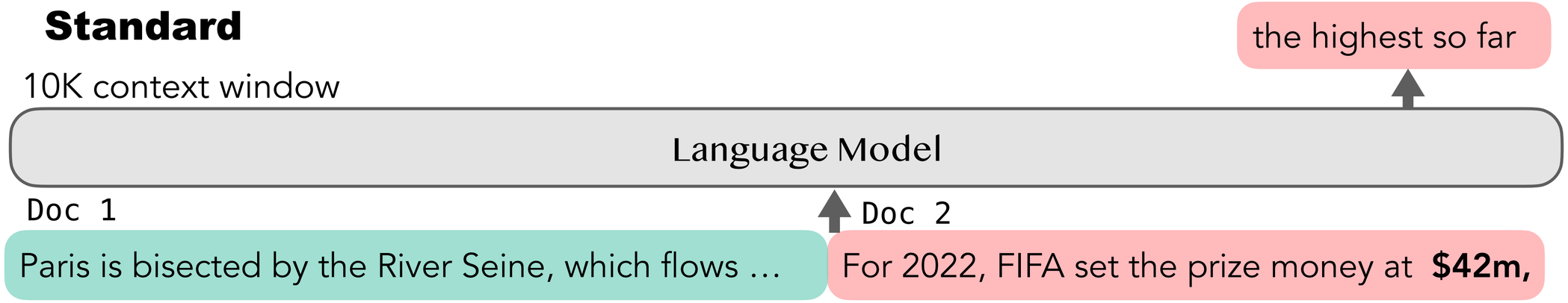
How are LMs pretrained?

- **Objective**: Predict the next token based on the prior input context

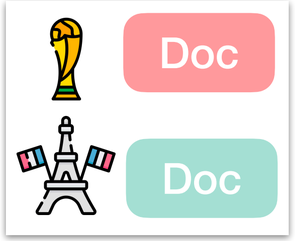


How are LMs pretrained?

- **Objective**: Predict the next token based on the prior input context
- **Input contexts**: Concatenate random documents in the same context window



the prior docs provide no signal for predicting the next doc

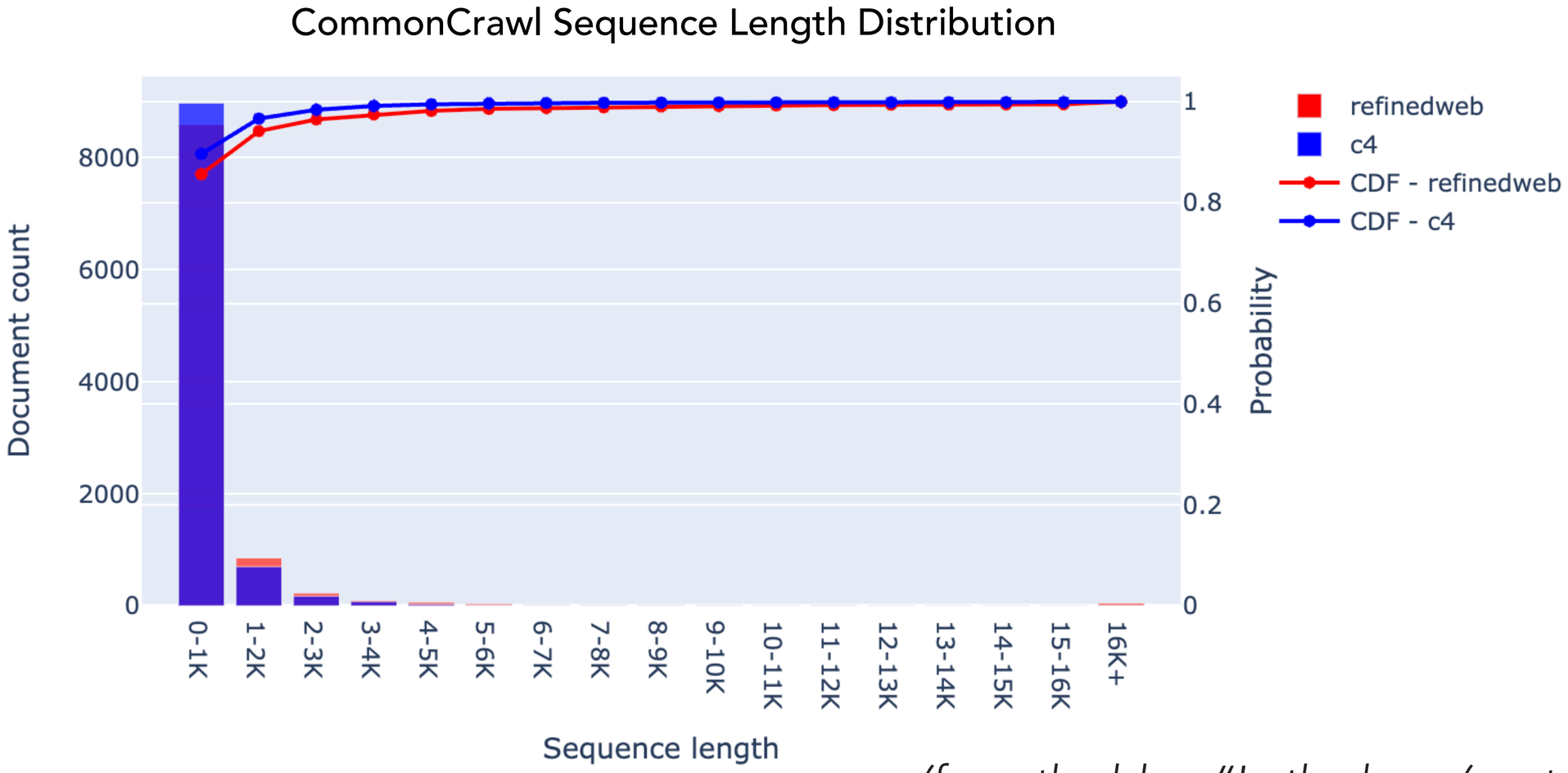


Different color indicates different doc. topic

No training signals from prior documents

- **Lack of long documents during pretraining**

  ~13% of CommonCrawl documents contain > 1K tokens



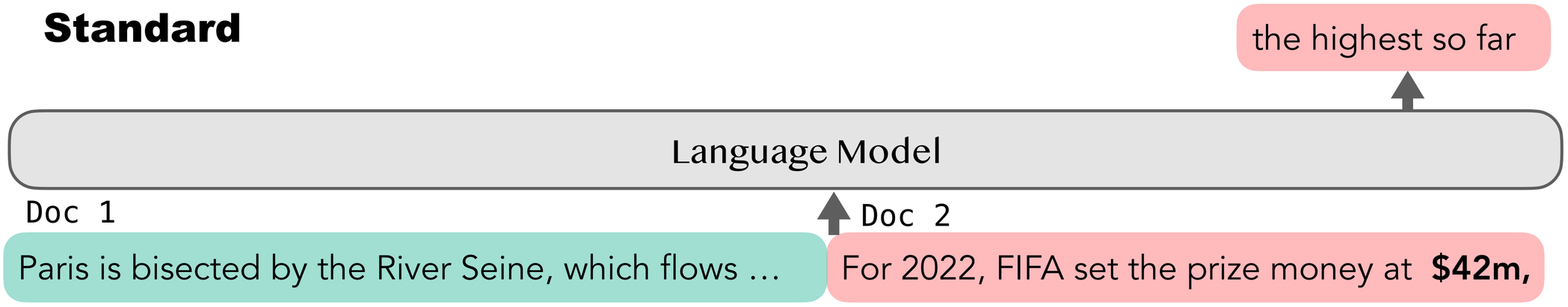
(from the blog "In the long (context) run")

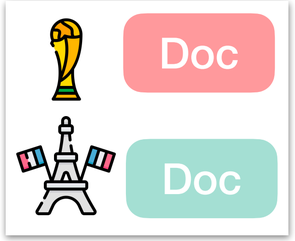
Proposed: In-Context Pretraining

Place related docs in the same context

Proposed: In-Context Pretraining

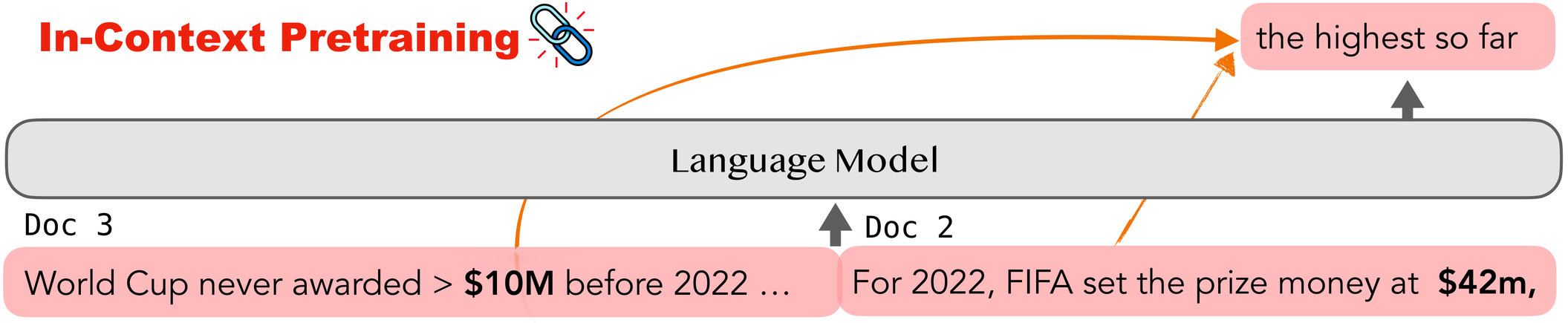
Place related docs in the same context



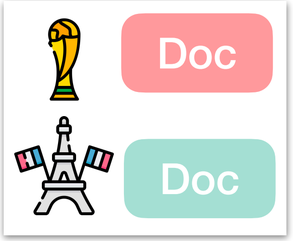


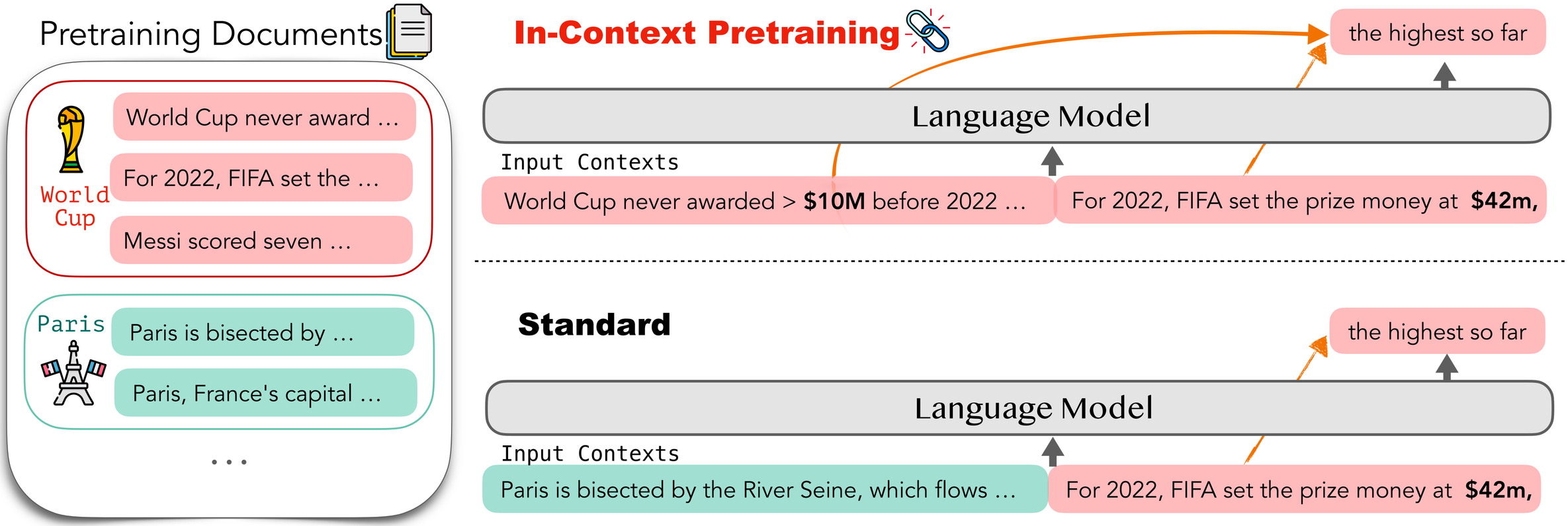
Proposed: In-Context Pretraining

Place related docs in the same context

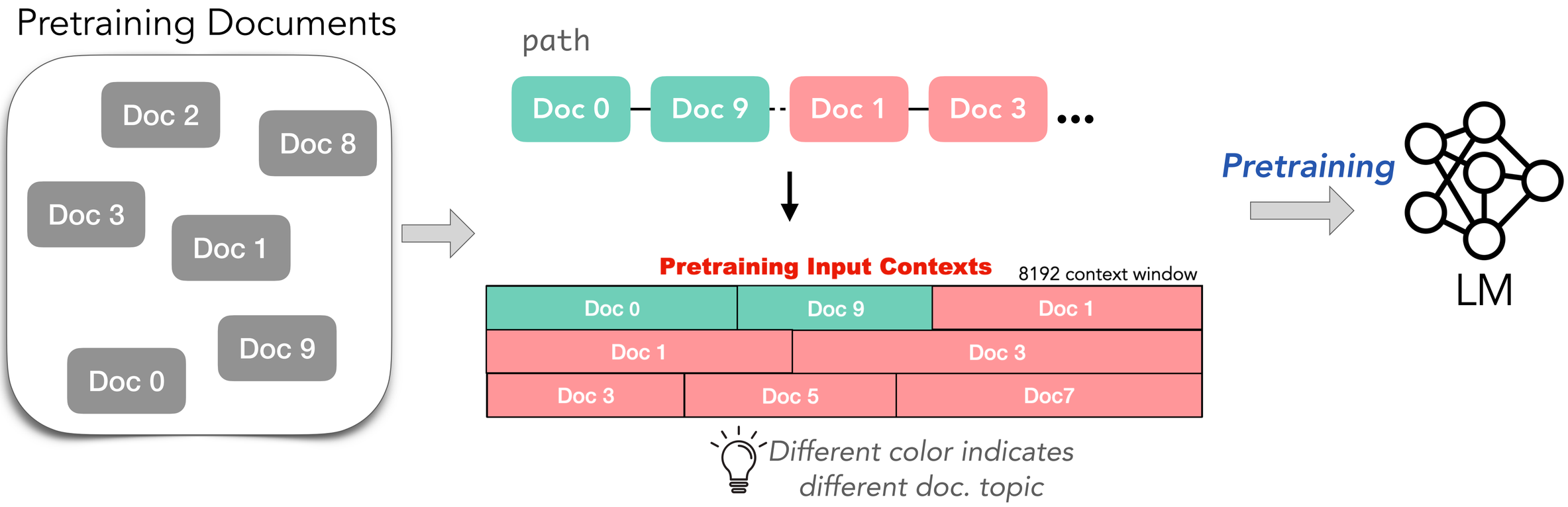


Encourage LMs to read and reason across document boundaries




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,

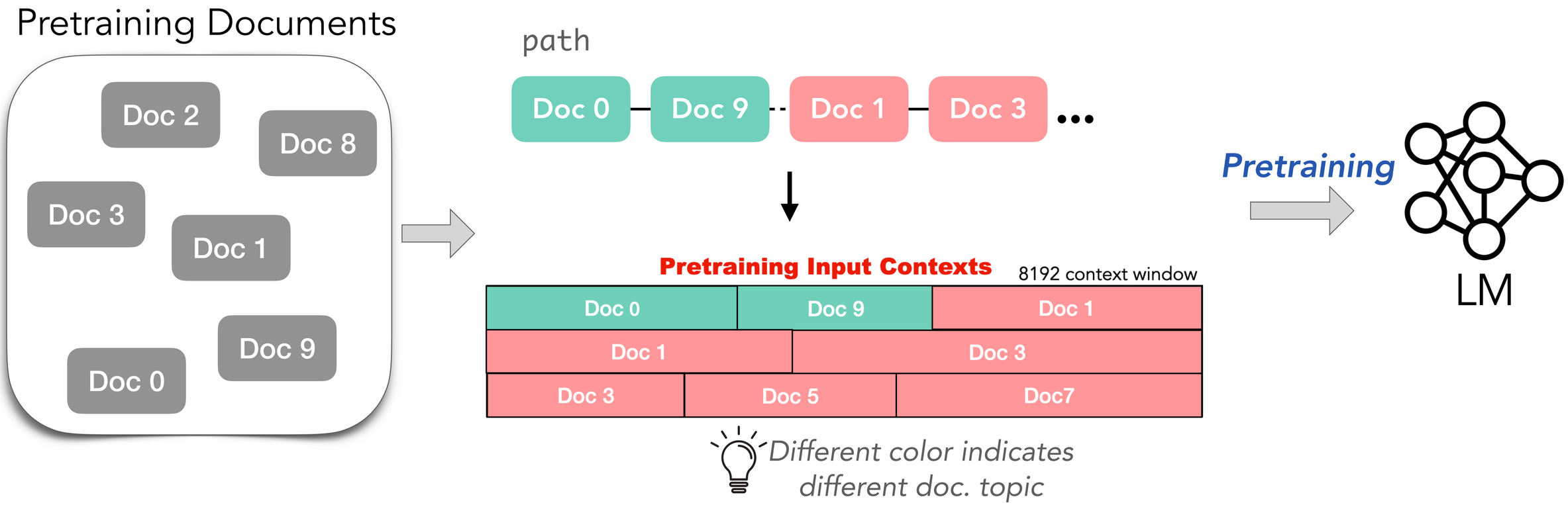
In-Context Pretraining overview



Step 1: Find Related Docs

Step 2: Create Input Contexts

In-Context Pretraining overview

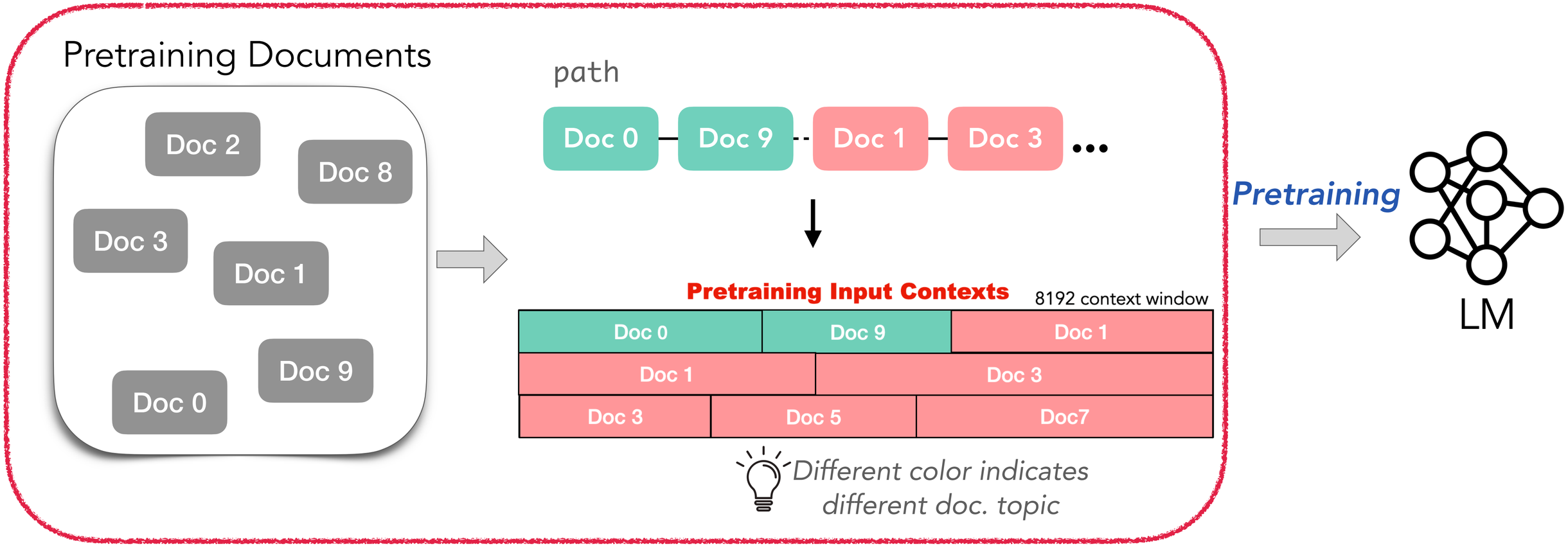


Step 1: Find Related Docs

Step 2: Create Input Contexts

In-Context Pretraining overview

It only changes the document ordering during pretraining



Step 1: Find Related Docs ***Step 2: Create Input Contexts***

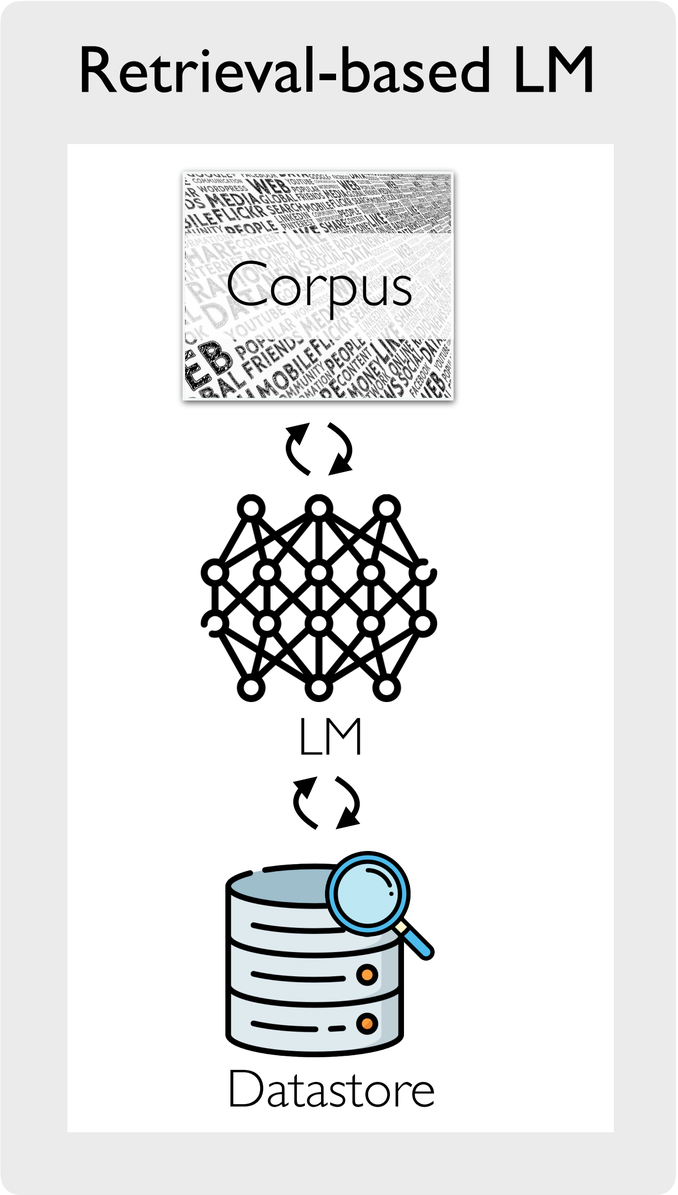
Summary

	Standard		*In-Context Pretraining*
Retrieval Augmentation	38%	**10.5%** →	42%
In-Context Learning	66%	**7.5%** →	71%
Reading Comprehension	37%	**14.0%** →	43%
Knowledge conflicts	44%	**15.9%** →	51%
Long Document Reasoning	32%	**7.5%** →	34%

23 datasets in total

Retrieval-augmented LMs (RAGs)

1. Why do need RAGs?
2. Architectures of RAGs
3. Training of the retriever
4. Training of the LMs



Q & A

Thank you for listening!

Slides adapted from Akari Asai's tutorial on Retrieval-augmented Language Models (ACL 2023)