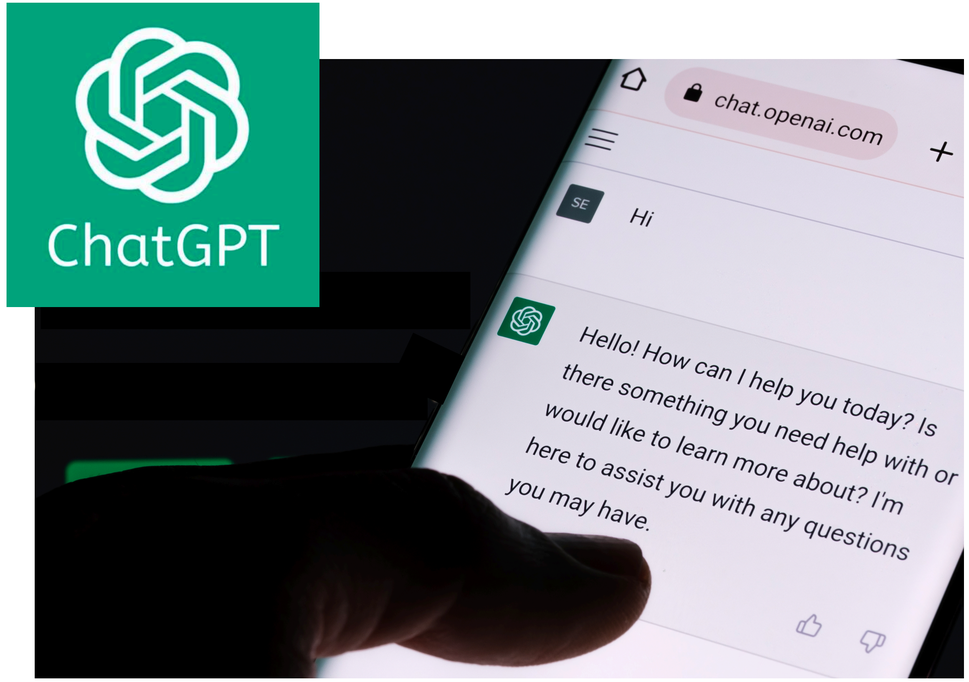
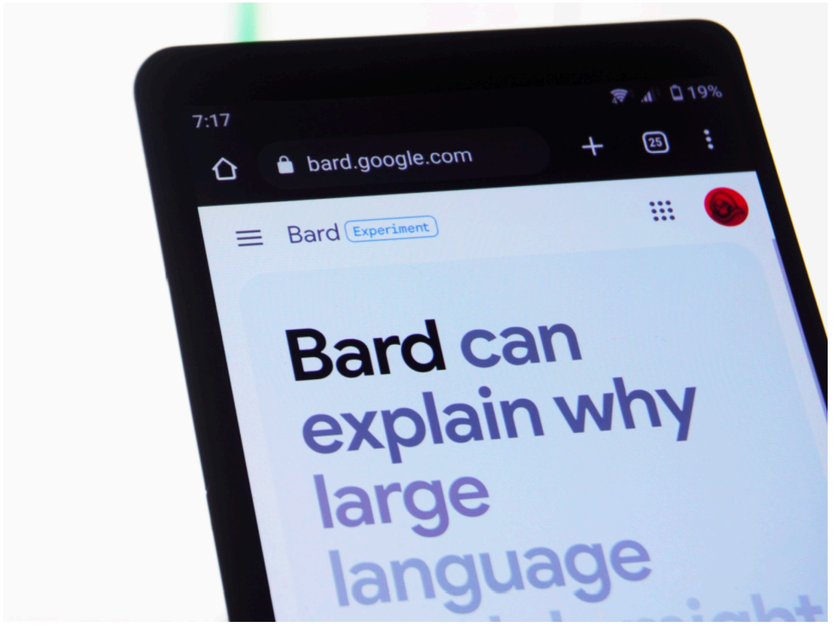
SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore

Sewon Min*, **Suchin Gururangan***, Eric Wallace, Weijia Shi,
Hannaneh Hajishirzi, Noah A. Smith, Luke Zettlemoyer

*co-first authors

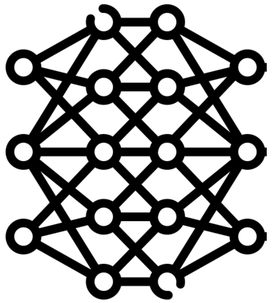
Language Models (LMs)





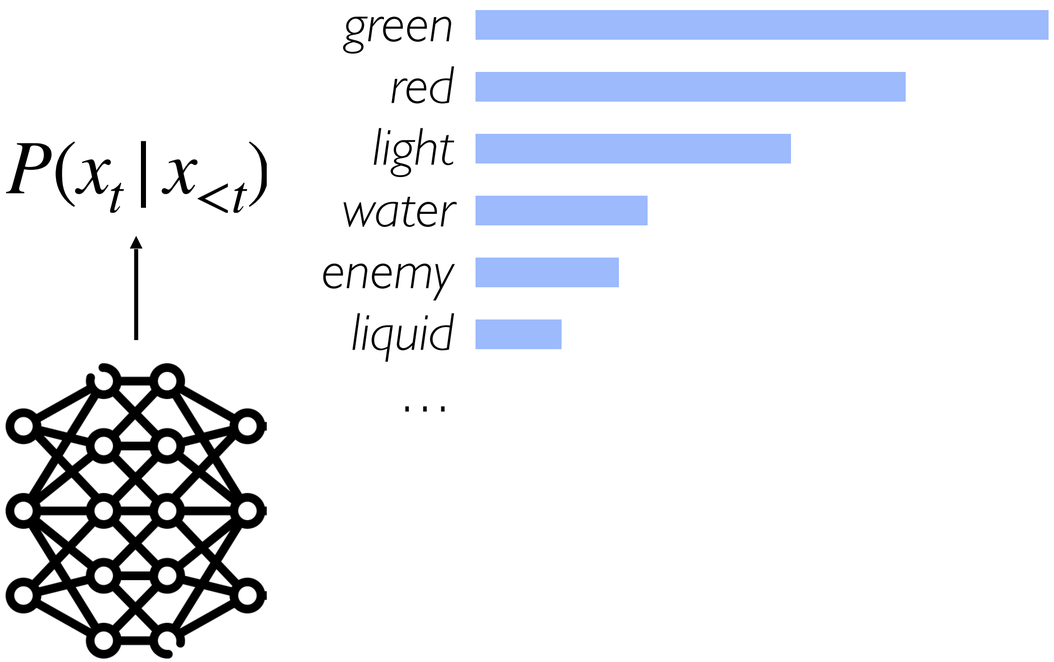
Language Models

$$P(x_t \mid x_{<t})$$



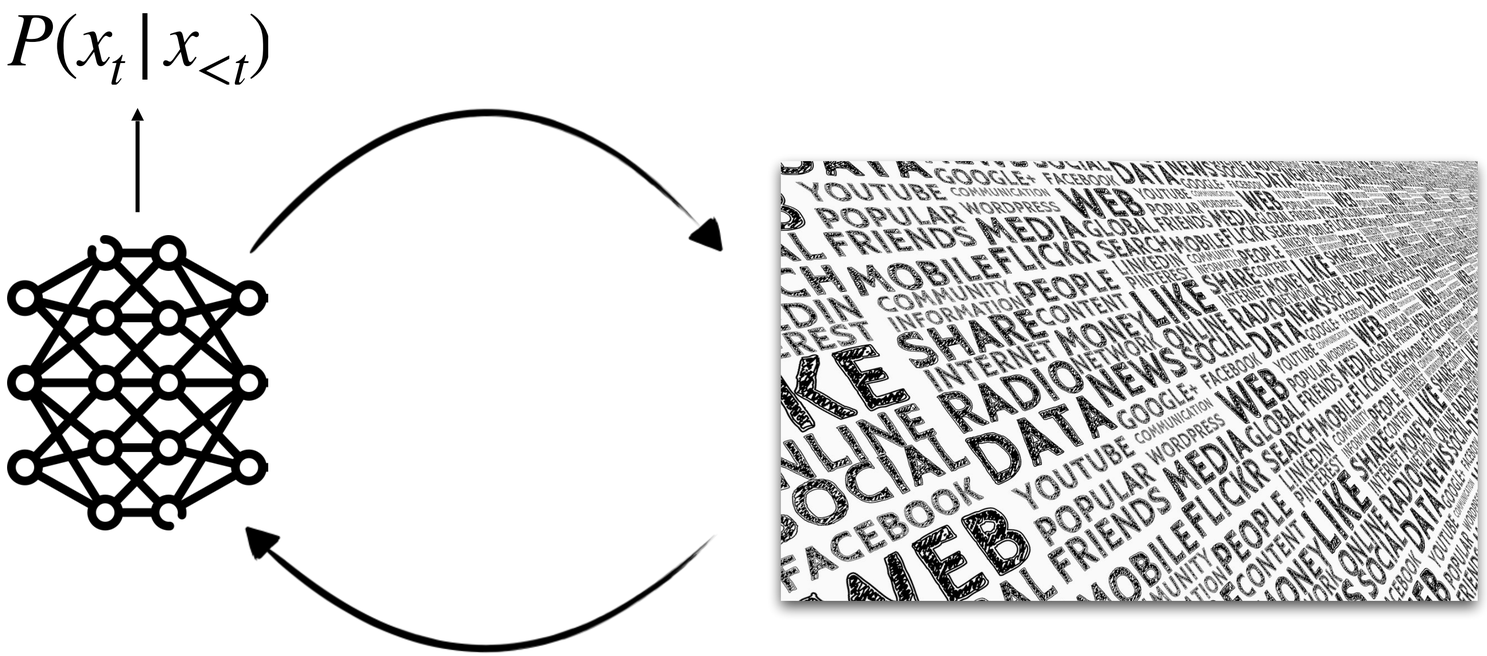
Harry felt Greenback collapse against him … on the floor as a jet of

Language Models



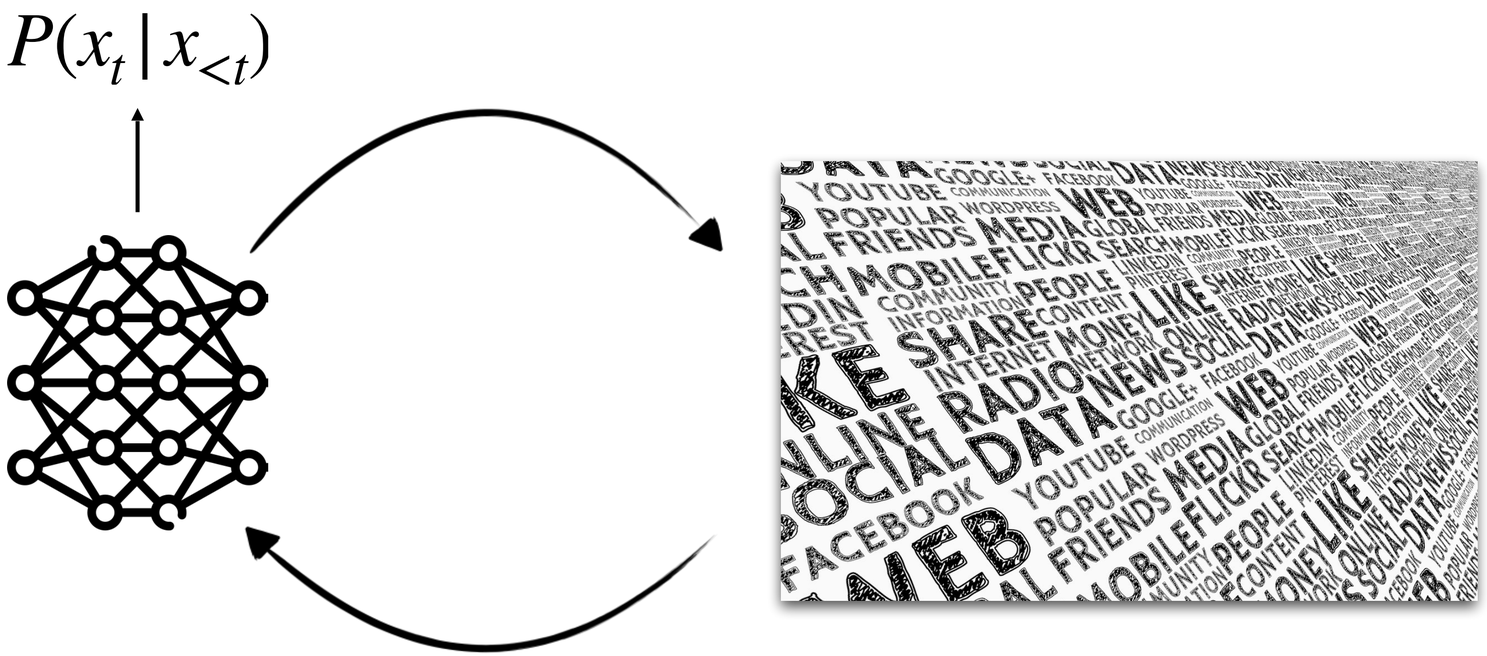
Harry felt Greenback collapse against him … on the floor as a jet of

Language Models



Are trained on all data on the web

Language Models



Are trained on all data on the web

(Mostly copyrighted)

Language Models

The New York Times

THE WALL STREET JOURNAL.

The Washington Post

(Largest data source in Web Crawl according to Dodge et al. 2021)

Language Models

The New York Times

THE WALL STREET JOURNAL.

The Washington Post

(Largest data source in Web Crawl according to Dodge et al. 2021)

CHAPTER 1

The Flame of Tar Valon

The Wheel of Time turns, and Ages come and pass leaving memories that become legend, then fade to myth, and are long forgot when that Age comes again. In one Age, called the Third Age by some, an Age yet to come, an Age long past, a wind rose in the Mountains of Dhoom. The wind was not the beginning. There are neither beginnings nor endings to the turning of the Wheel of Time. But it was a beginning.

Born among black, knife-edged peaks, where death roamed the high passes yet hid from things still more dangerous, the wind blew south across the tangled forest of the Great Blight, a forest tainted and twisted by the touch of the Dark One. The sickly sweet smell of corruption faded by the time the wind crossed that invisible line men called the border of Shienar, where spring flowers

Language Models

The New York Times

THE WALL STREET JOURNAL.

The Washington Post

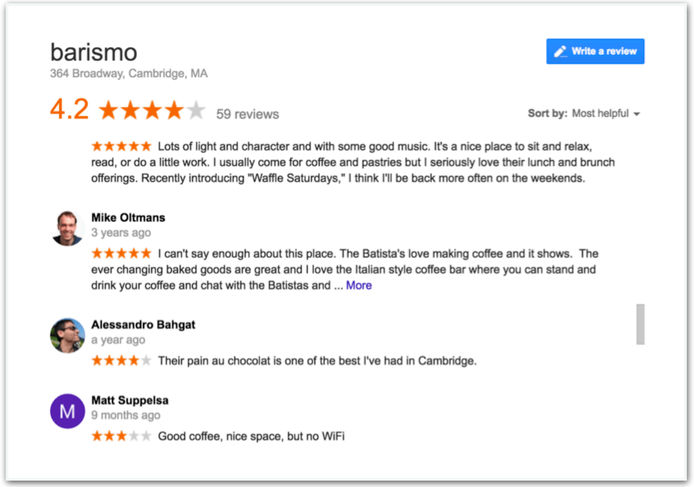
(Largest data source in Web Crawl according to Dodge et al. 2021)

CHAPTER 1

The Flame of Tar Valon

The Wheel of Time turns, and Ages come and pass leaving memories that become legend, then fade to myth, and are long forgot when that Age comes again. In one Age, called the Third Age by some, an Age yet to come, an Age long past, a wind rose in the Mountains of Dhoom. The wind was not the beginning. There are neither beginnings nor endings to the turning of the Wheel of Time. But it was a beginning.

Born among black, knife-edged peaks, where death roamed the high passes yet hid from things still more dangerous, the wind blew south across the tangled forest of the Great Blight, a forest tainted and twisted by the touch of the Dark One. The sickly sweet smell of corruption faded by the time the wind crossed that invisible line men called the border of Shienar, where spring flowers



barismo

364 Broadway, Cambridge, MA

Write a review

4.2 ★★★★☆ 59 reviews

Sort by: Most helpful

★★★★★ Lots of light and character and with some good music. It's a nice place to sit and relax, read, or do a little work. I usually come for coffee and pastries but I seriously love their lunch and brunch offerings. Recently introducing "Waffle Saturdays," I think I'll be back more often on the weekends.

Mike Oltmans
3 years ago
★★★★★ I can't say enough about this place. The Batista's love making coffee and it shows. The ever changing baked goods are great and I love the Italian style coffee bar where you can stand and drink your coffee and chat with the Batistas and ... More

Alessandro Bahgat
a year ago
★★★★☆ Their pain au chocolat is one of the best I've had in Cambridge.

M Matt Suppelsa
9 months ago
★★★☆☆ Good coffee, nice space, but no WiFi

Language Models

The New York Times

THE WALL STREET JOURNAL.

The Washington Post

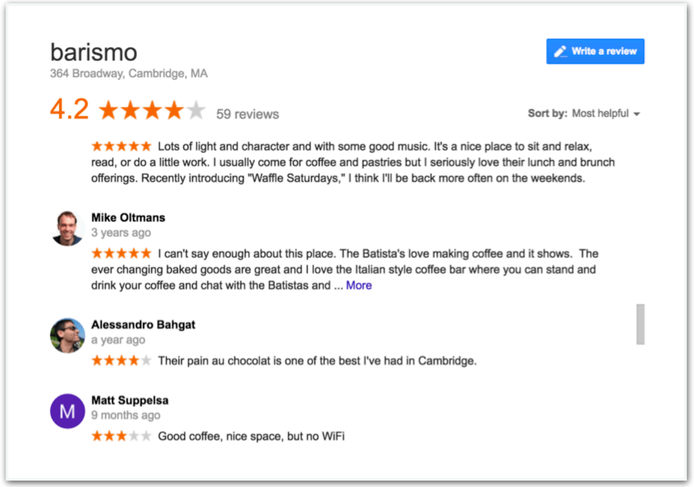
(Largest data source in Web Crawl according to Dodge et al. 2021)

CHAPTER 1

The Flame of Tar Valon

The Wheel of Time turns, and Ages come and pass leaving memories that become legend, then fade to myth, and are long forgot when that Age comes again. In one Age, called the Third Age by some, an Age yet to come, an Age long past, a wind rose in the Mountains of Dhoom. The wind was not the beginning. There are neither beginnings nor endings to the turning of the Wheel of Time. But it was a beginning.

Born among black, knife-edged peaks, where death roamed the high passes yet hid from things still more dangerous, the wind blew south across the tangled forest of the Great Blight, a forest tainted and twisted by the touch of the Dark One. The sickly sweet smell of corruption faded by the time the wind crossed that invisible line men called the border of Shienar, where spring flowers



barismo

364 Broadway, Cambridge, MA

Write a review

4.2 ★★★★☆ 59 reviews

Sort by: Most helpful

★★★★★ Lots of light and character and with some good music. It's a nice place to sit and relax, read, or do a little work. I usually come for coffee and pastries but I seriously love their lunch and brunch offerings. Recently introducing "Waffle Saturdays," I think I'll be back more often on the weekends.

Mike Oltmans
3 years ago
★★★★★ I can't say enough about this place. The Batista's love making coffee and it shows. The ever changing baked goods are great and I love the Italian style coffee bar where you can stand and drink your coffee and chat with the Batistas and ... More

Alessandro Bahgat
a year ago
★★★★☆ Their pain au chocolat is one of the best I've had in Cambridge.

M Matt Suppelsa
9 months ago
★★★☆☆ Good coffee, nice space, but no WiFi

"Since most online content has copyright protections attached at creation, using them for certain purposes could be considered infringement."

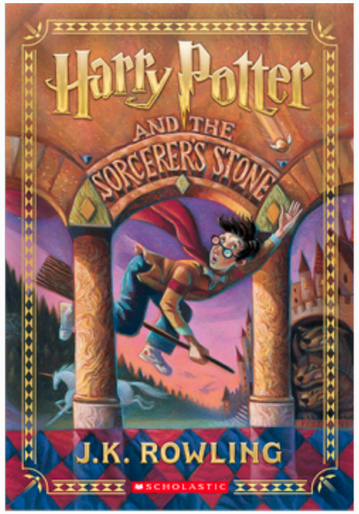
— Henderson et al. 2023. Foundation Models and Fair Use.

Language Models: Risk

Overview Documentation API reference Examples Playground

Playground

Mr and Mrs Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.
Mr Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large moustache. Mrs Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbours. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.
The Dursleys had everything they wanted,



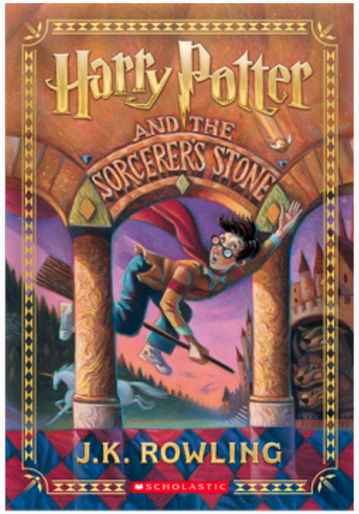
Looking for ChatGPT? Try it now

Submit

Overview Documentation API reference Examples Playground

Playground

Mr and Mrs Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.
Mr Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large moustache. Mrs Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbours. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.
The Dursleys had everything they wanted,



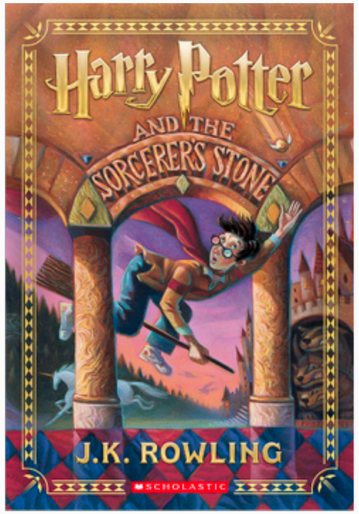
Looking for ChatGPT? Try it now

Submit

Overview Documentation API reference Examples Playground

Playground

Mr and Mrs Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.
Mr Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large moustache. Mrs Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbours. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.
The Dursleys had everything they wanted,



copyright infringement?

Looking for ChatGPT? Try it now

Submit

Language Models: Risk

1) Training on copyrighted material is **high legal risk** to model designers
   - Nowadays, you could get sued!
   - Need to re-train the model if a data owner wants to opt-out

Language Models: Risk

1) Training on copyrighted material is **high legal risk** to model designers
    - Nowadays, you could get sued!
    - Need to re-train the model if a data owner wants to opt-out
2) Training on copyrighted material **disenfranchises data creators**
    - Difficult to attribute or compensate when the model uses your data
    - Difficult to remove your data with strong guarantees

Language Models: Risk

A straightforward solution: filter training data to include permissively-licensed data only

(Kocetkov et al. 2023,
Fried et al. 2023)

peS2o

(Soldaini & Lo, 2023)

Language Models: Risk

A straightforward solution: filter training data to include permissively-licensed data only

(Kocetkov et al. 2023,
Fried et al. 2023)

peS2o

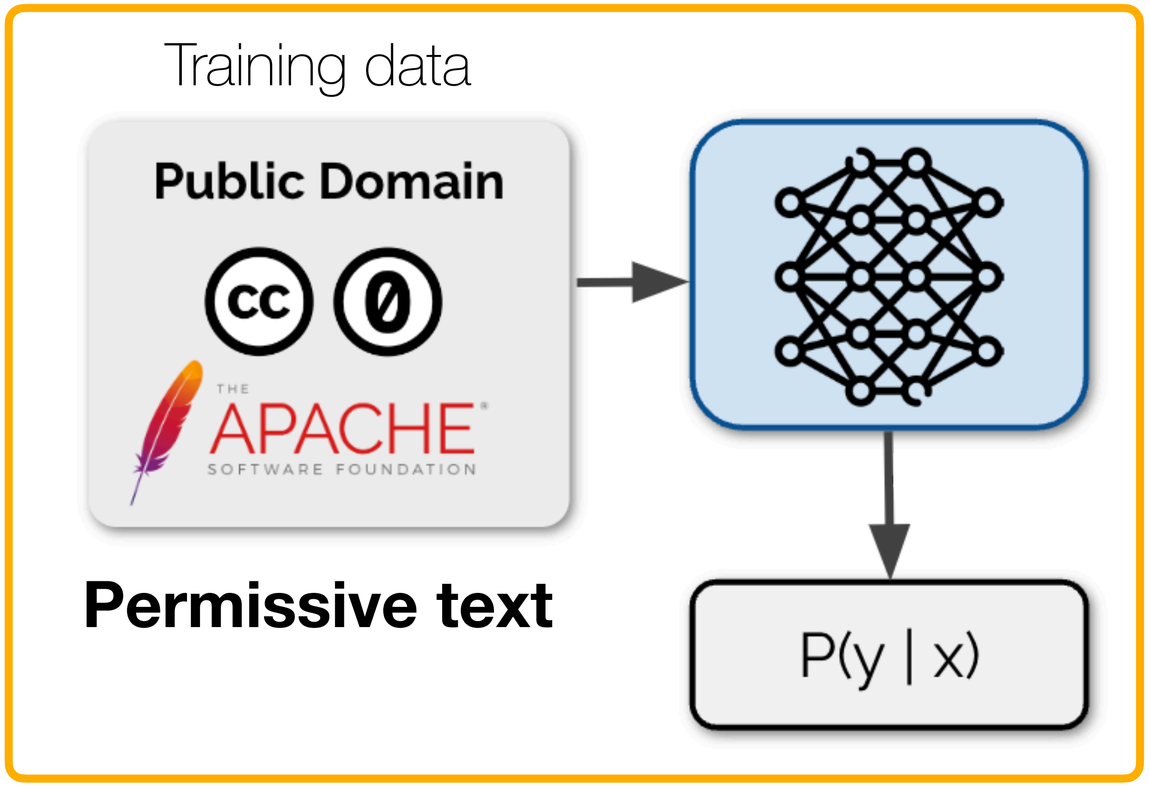
(Soldaini & Lo, 2023)

"[…] the general approach of filtering out any copyrightable content entirely from a dataset […] may not be possible for many practical settings where little to no open domain or permissively-licensed material is available."

— Henderson et al. 2023. Foundation Models and Fair Use.

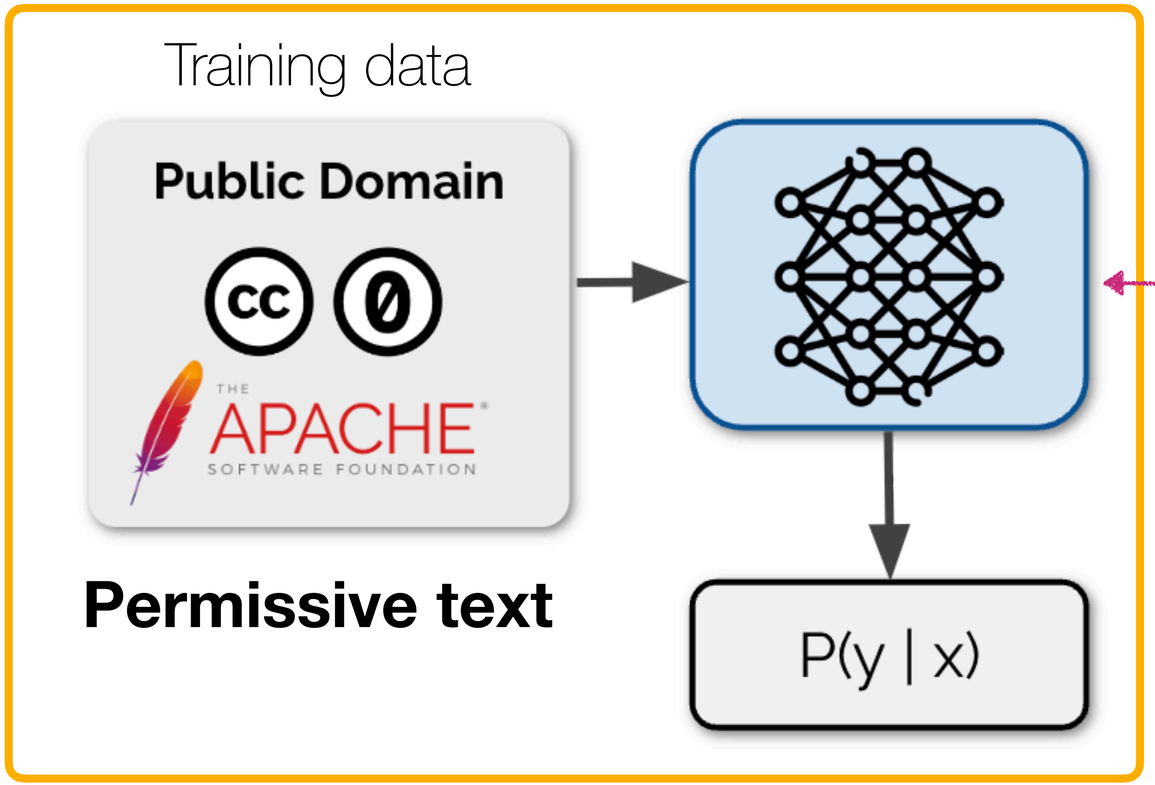
Our Proposal: SILO LM

Training



Our Proposal: SILO LM

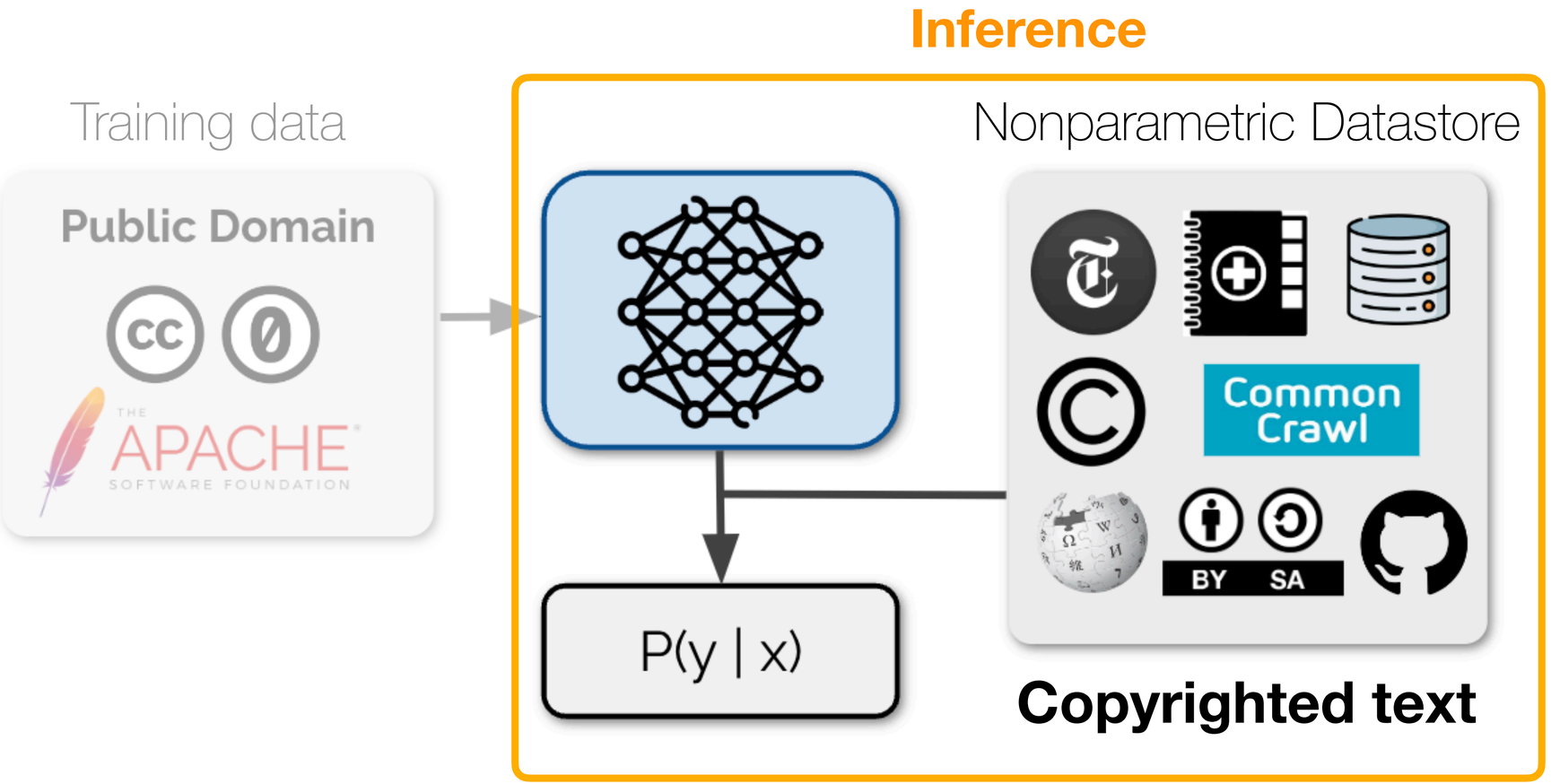
Training



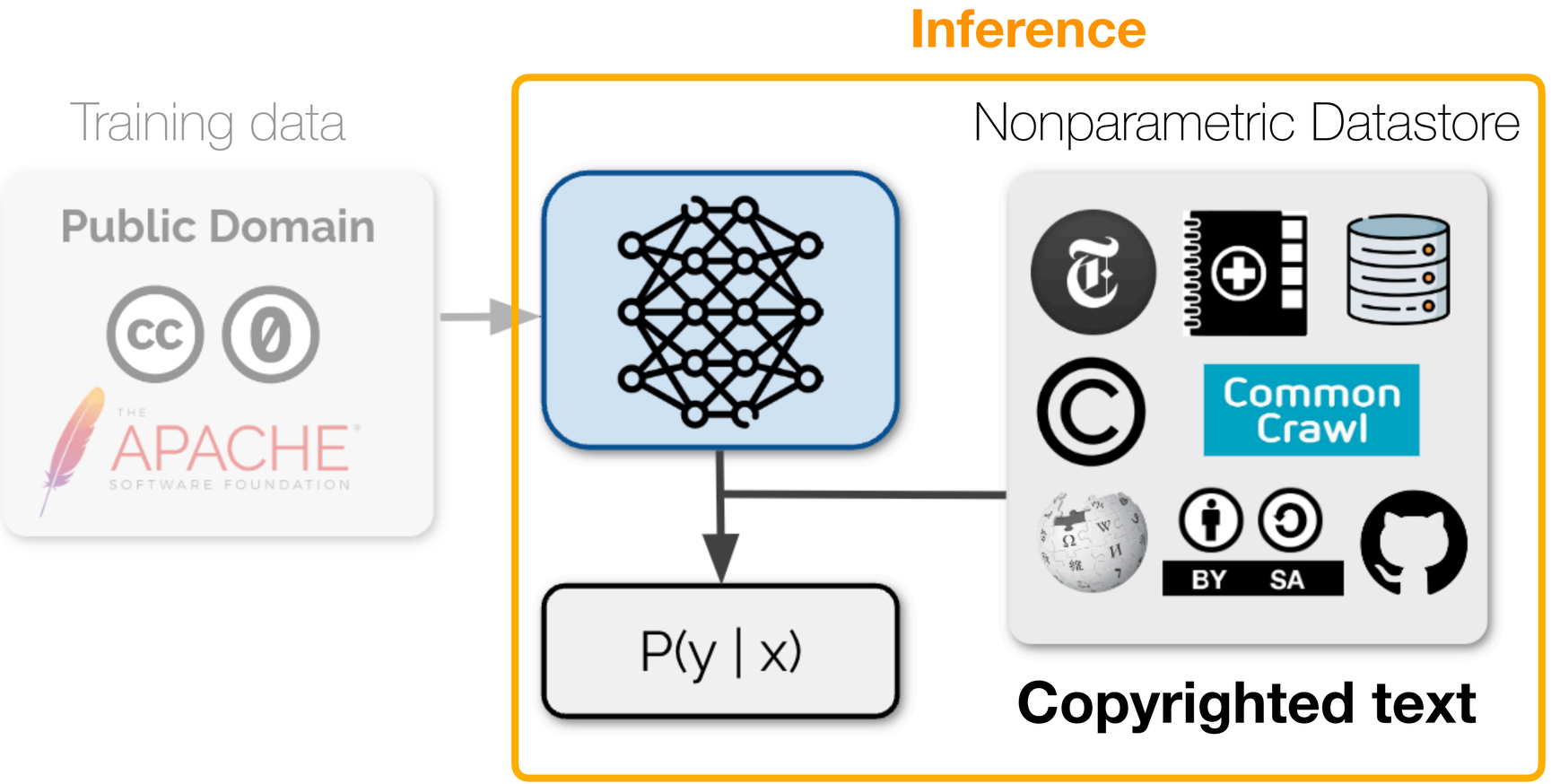
Very low legal risk,
but poor performance
(small-size data, domain shift)

(We'll show you at this talk)

Our Proposal: SILO LM

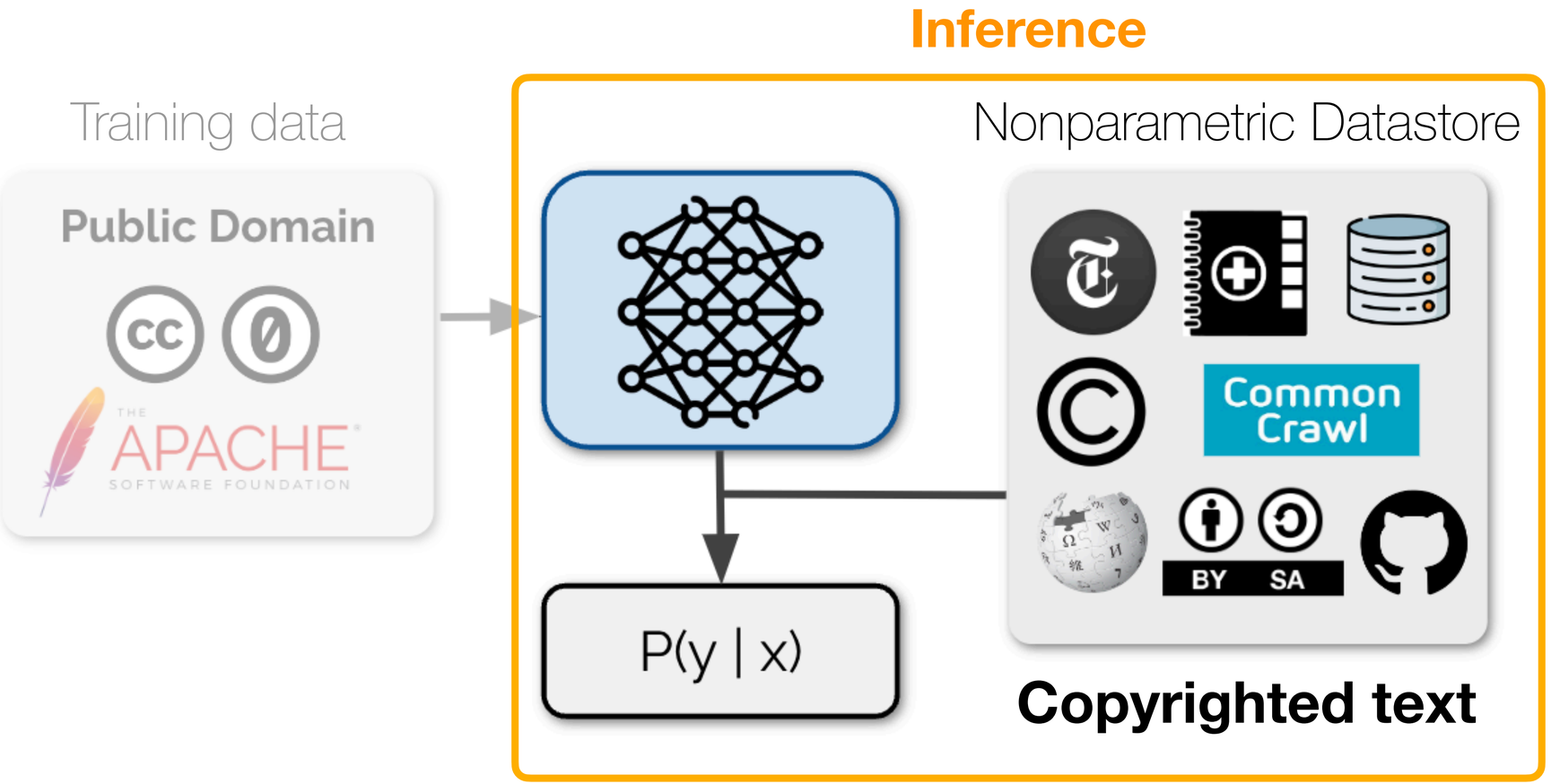


Our Proposal: SILO LM



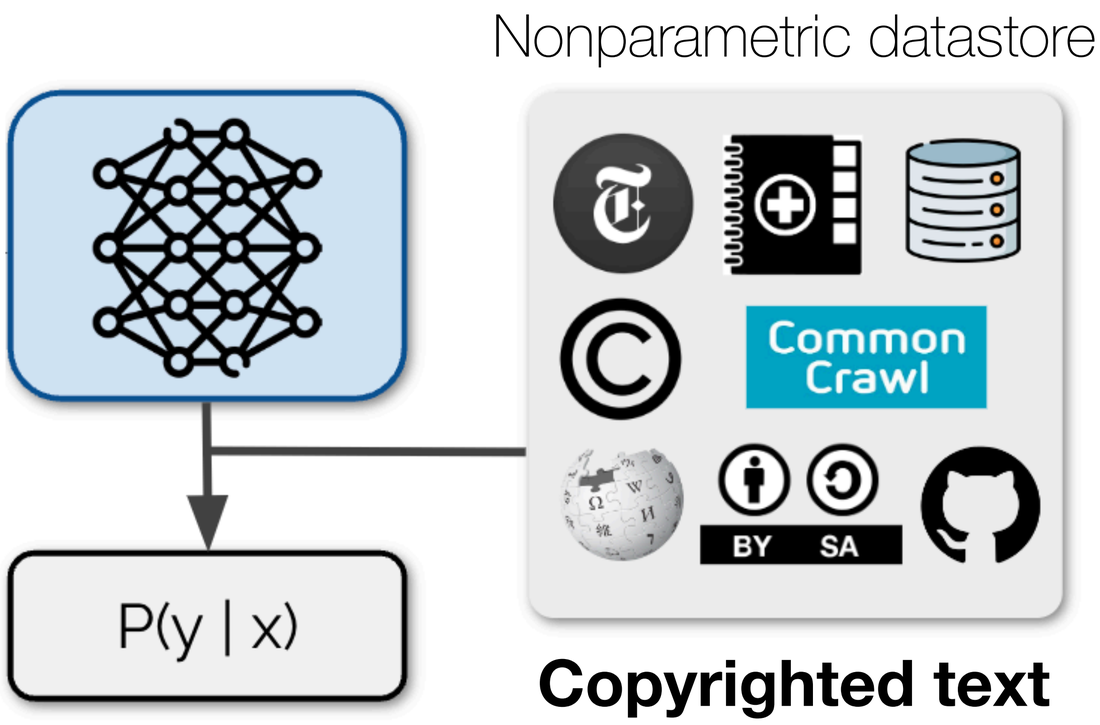
Can trace inherent attribution

Our Proposal: SILO LM



- ☑ Can trace inherent attribution
- ☑ Can modify the datastore at any time

Why SILO LM?

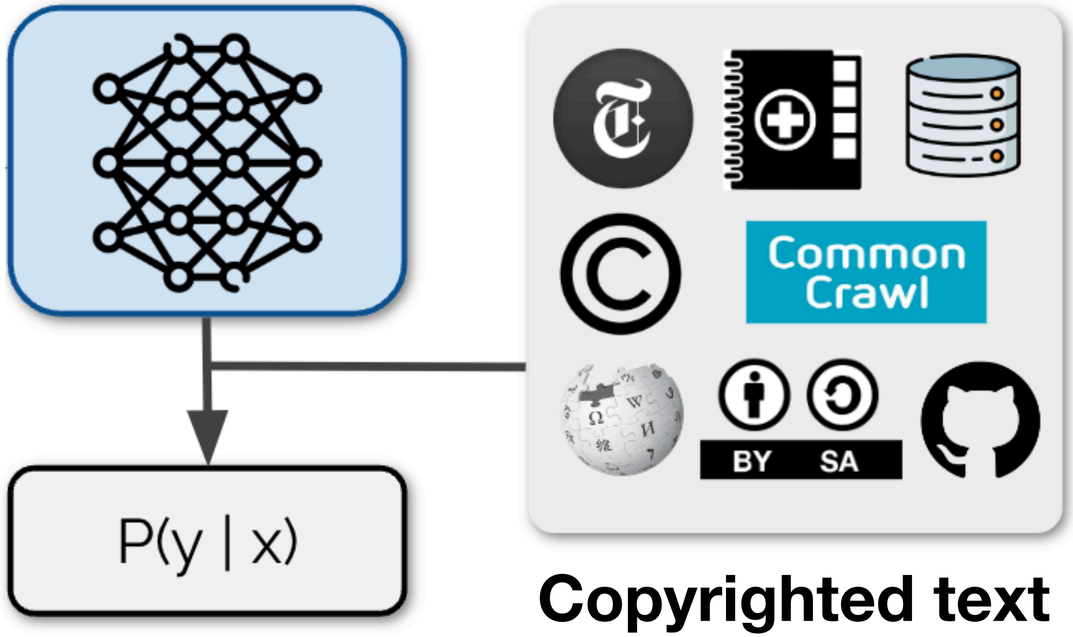


Thanks to Peter Henderson, Mark Lemley and Pamela Samuelson for reviewing our legal claims

Why SILO LM?

Significantly improve generalization → Nonparametric datastore

(We'll show you at this talk)



Thanks to Peter Henderson, Mark Lemley and Pamela Samuelson for reviewing our legal claims

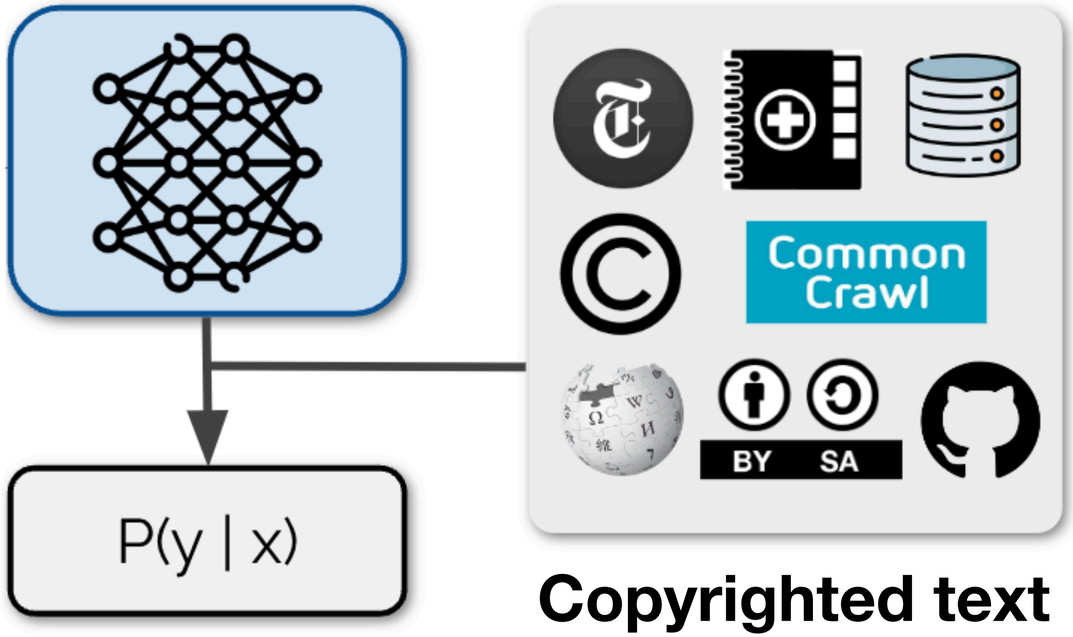
Why SILO LM?

Significantly improve generalization → Nonparametric datastore

(We'll show you at this talk)

☑ Can trace inherent attribution

- Likely defense *fair use*
- Provide copyright notice (alignment to *DMCA*)
- Allow credits (or payment) to data creators

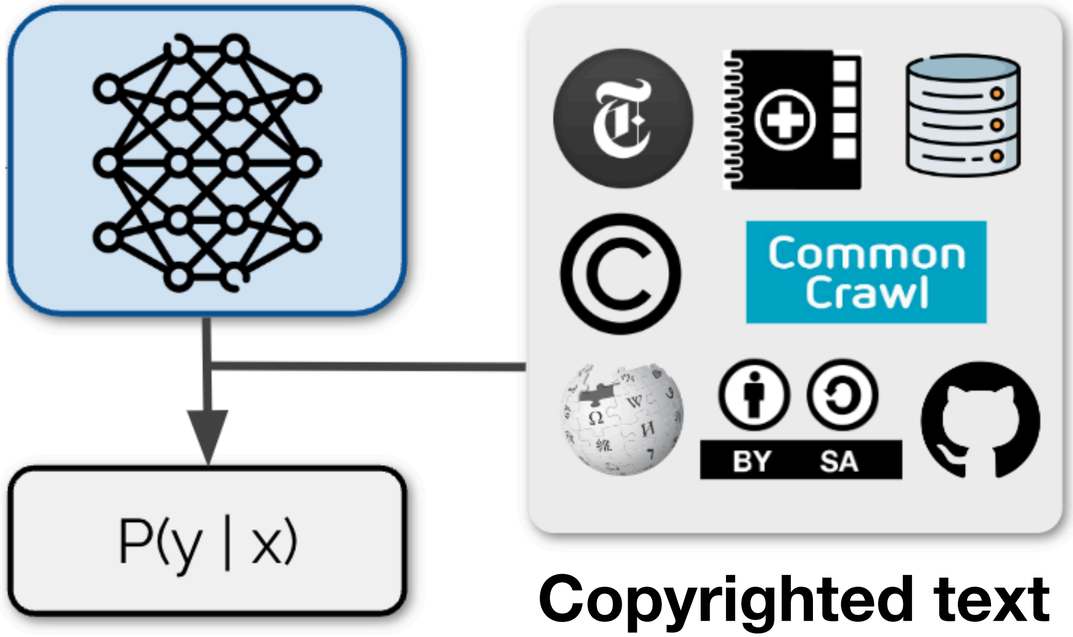


Thanks to Peter Henderson, Mark Lemley and Pamela Samuelson for reviewing our legal claims

Why SILO LM?

Significantly improve generalization → Nonparametric datastore

(We'll show you at this talk)



☑ Can trace inherent attribution

- Likely defense *fair use*
- Provide copyright notice (alignment to *DMCA*)
- Allow credits (or payment) to data creators

☑ Can modify the datastore at any time

- Support removal of data at any time
- Better alignment with *GDPR*

Thanks to Peter Henderson, Mark Lemley and Pamela Samuelson for reviewing our legal claims

Content

Motivation

We are here

A SILO Language Model

Experiments

Conclusion & Future Work

Content

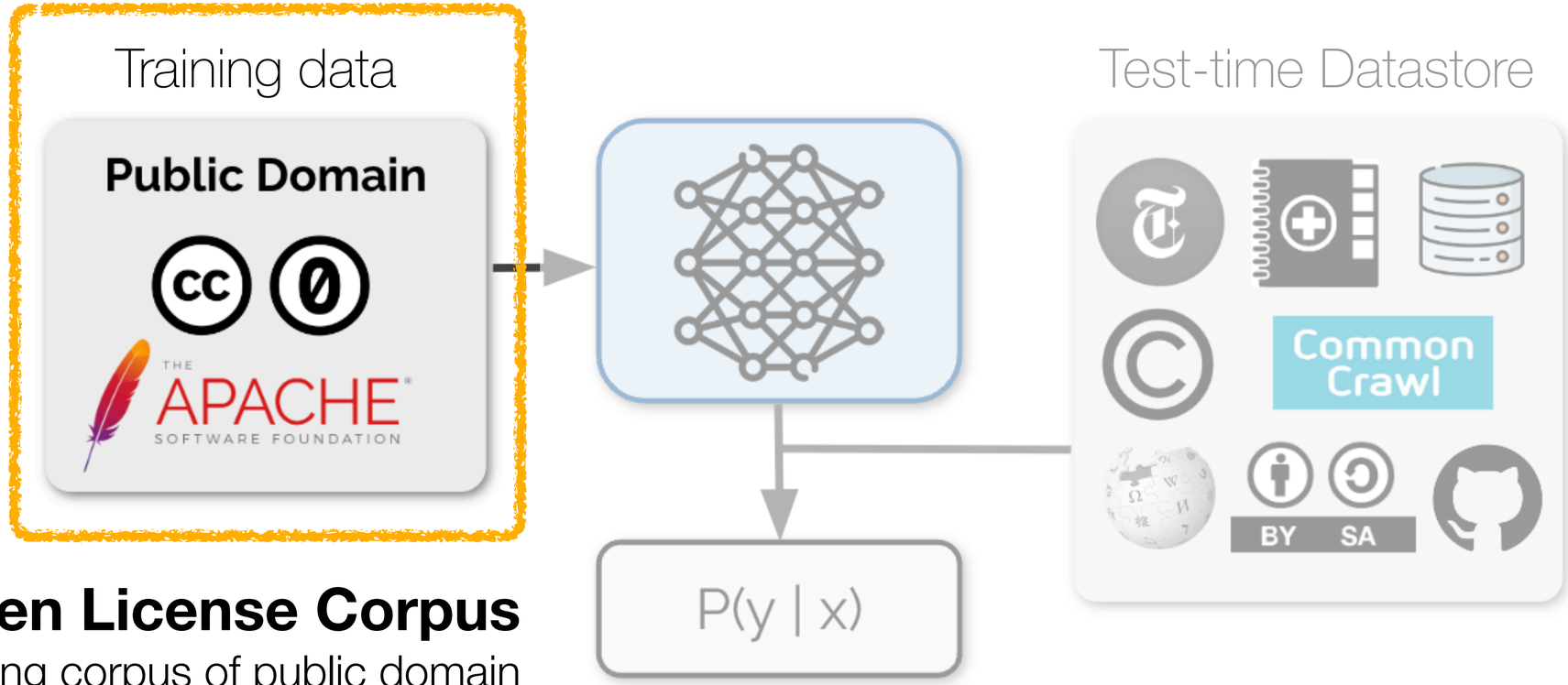
Motivation

A SILO Language Model

Experiments

Conclusion & Future Work

A SILO Language Model

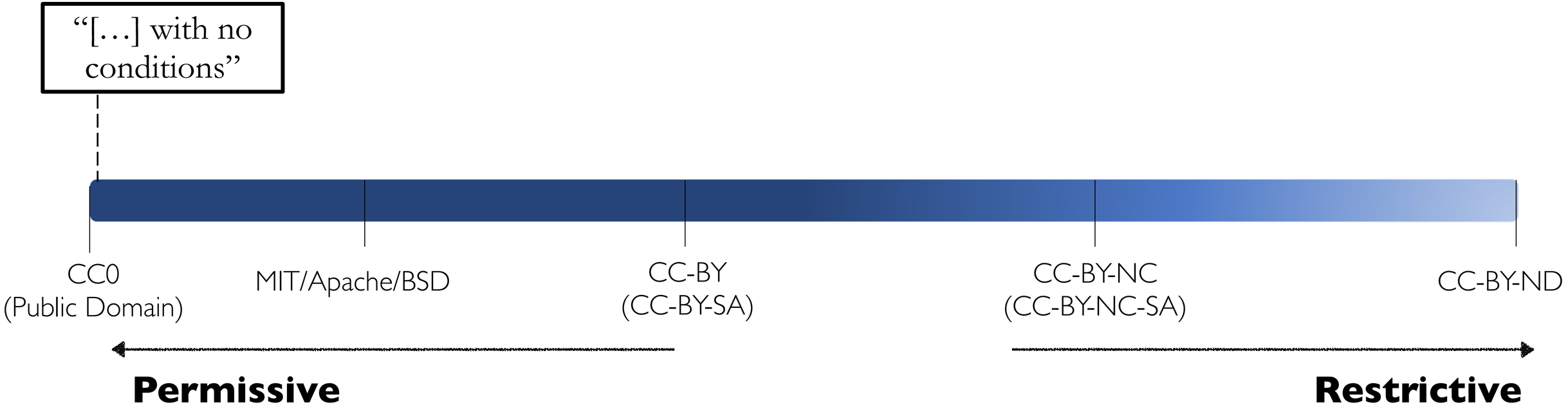


1) Open License Corpus

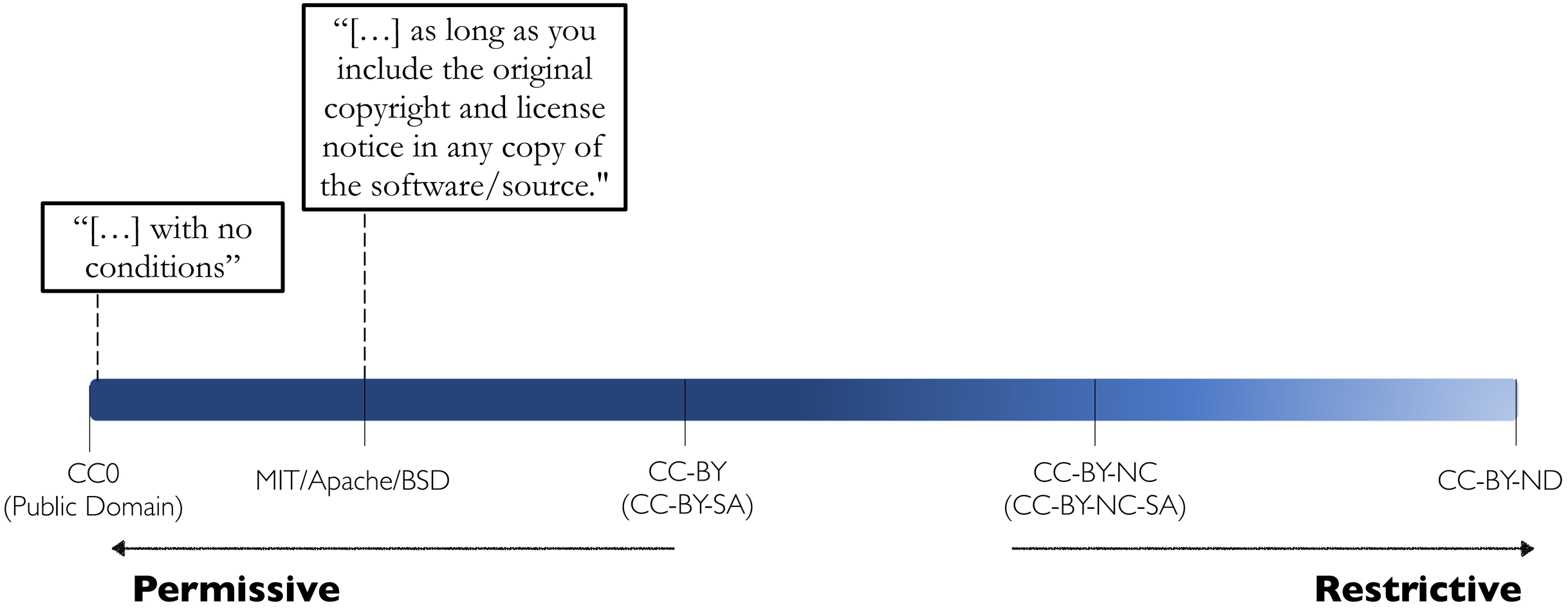
A new training corpus of public domain & permissively-licensed text

Permissive text

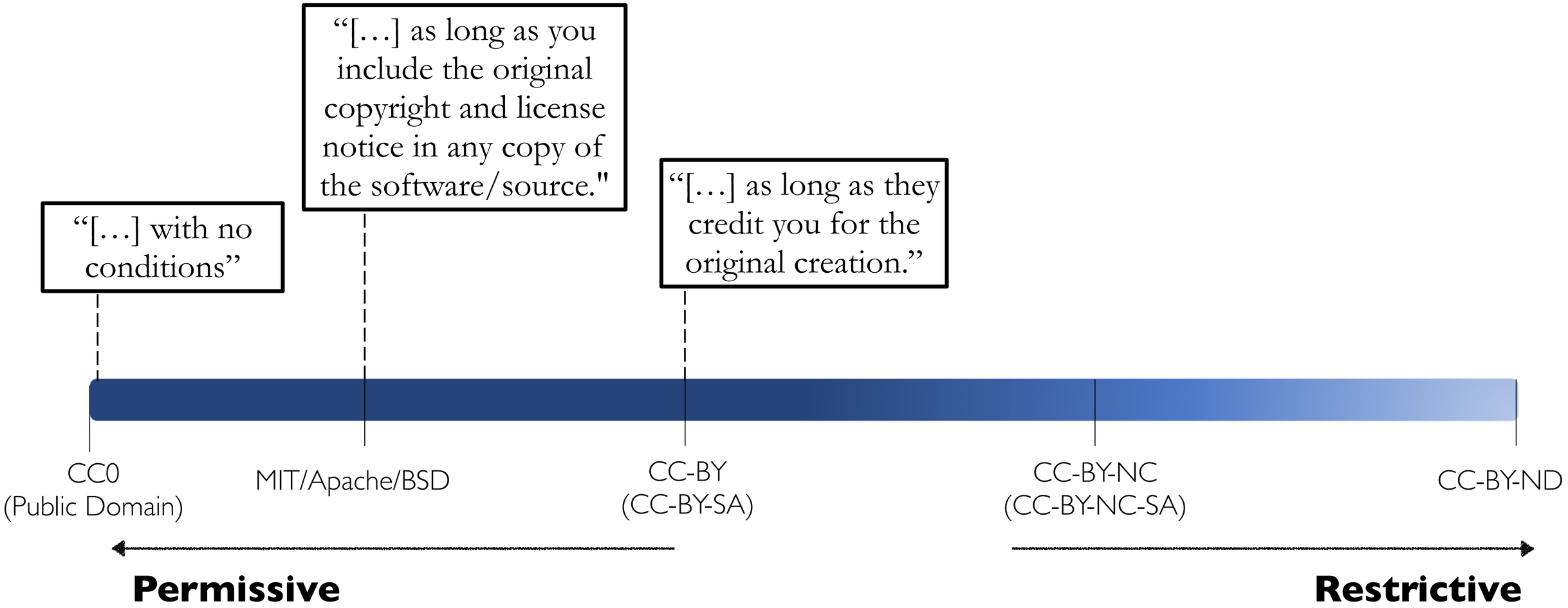
Permissive text



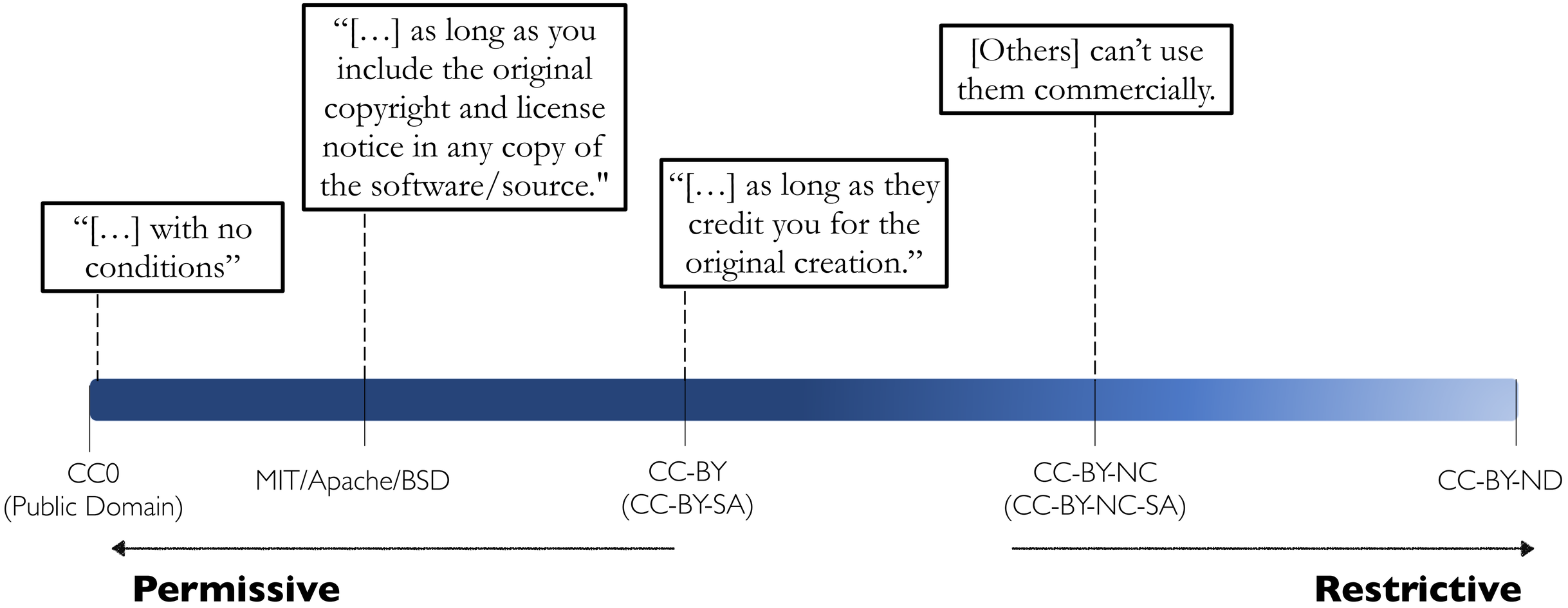
Permissive text



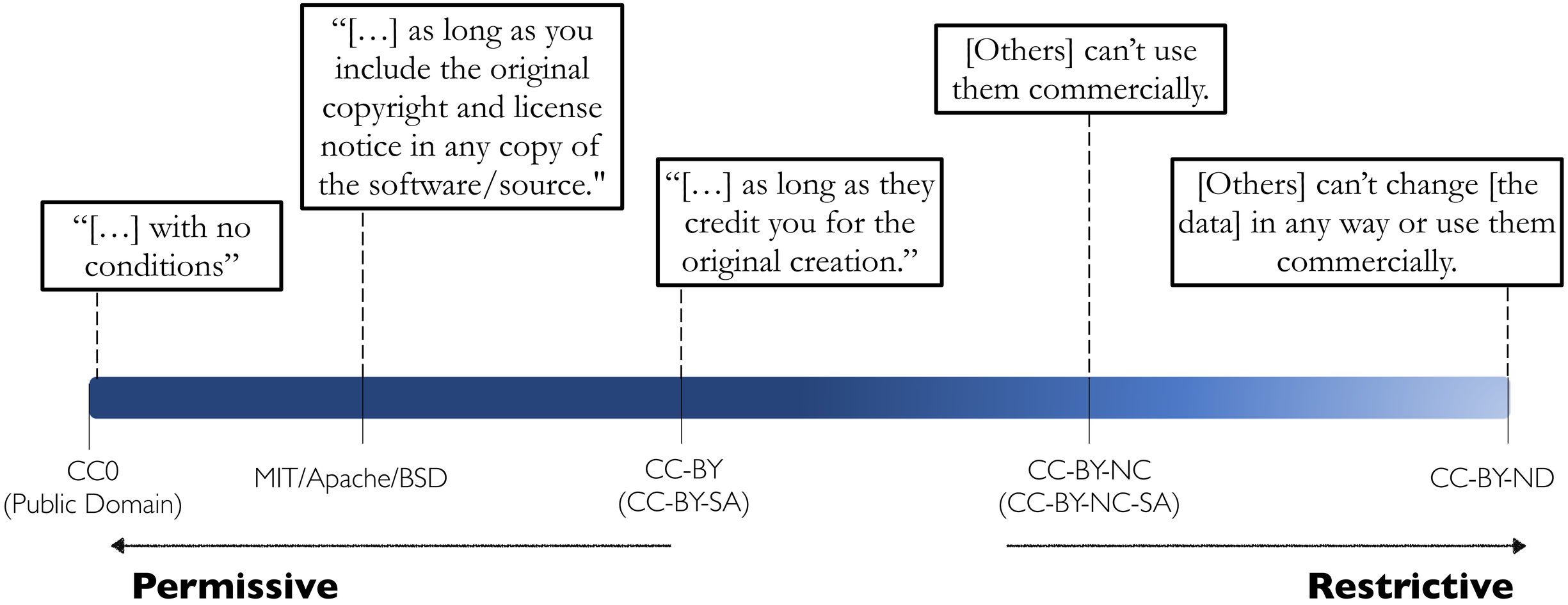
Permissive text



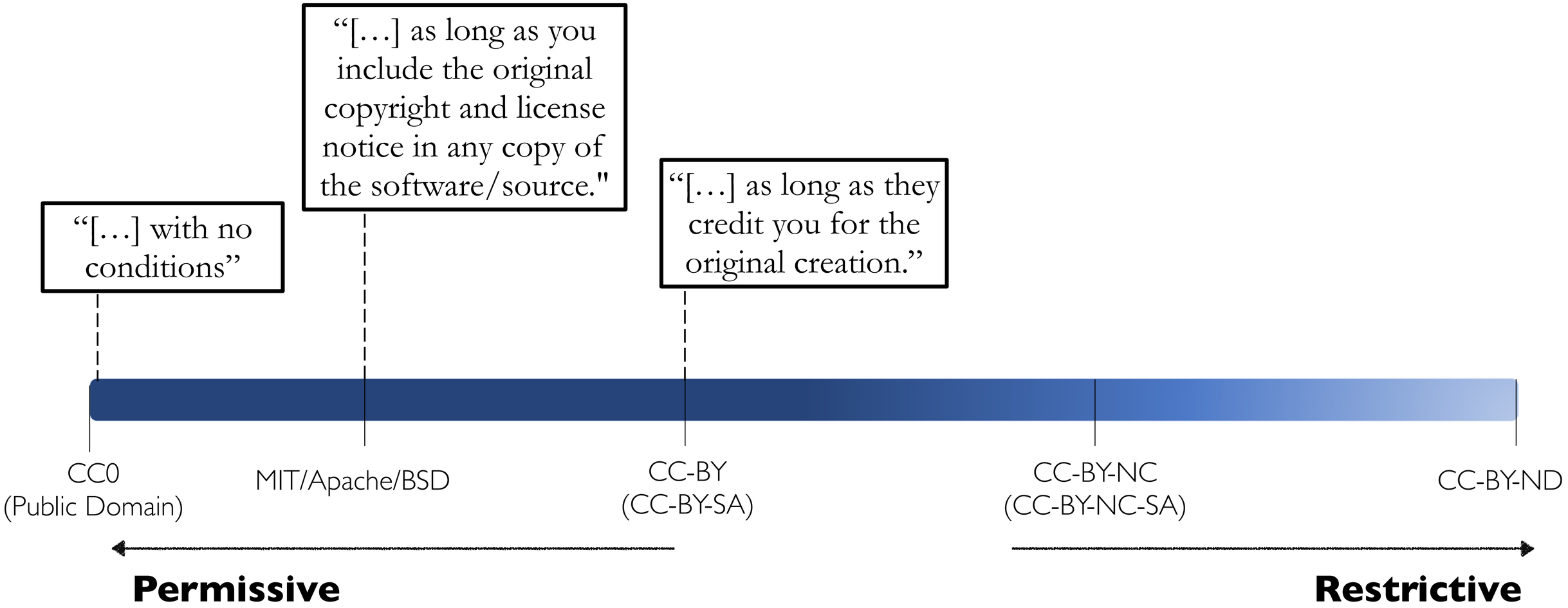
Permissive text



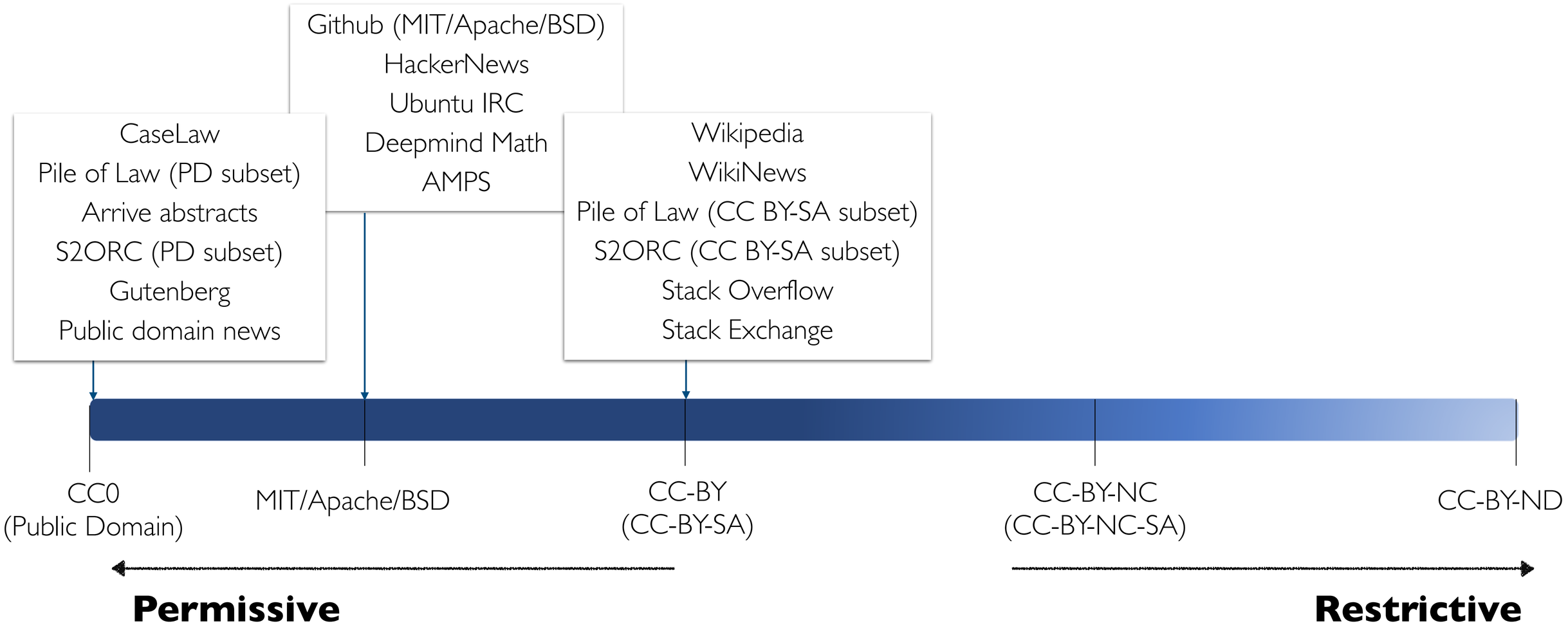
Permissive text



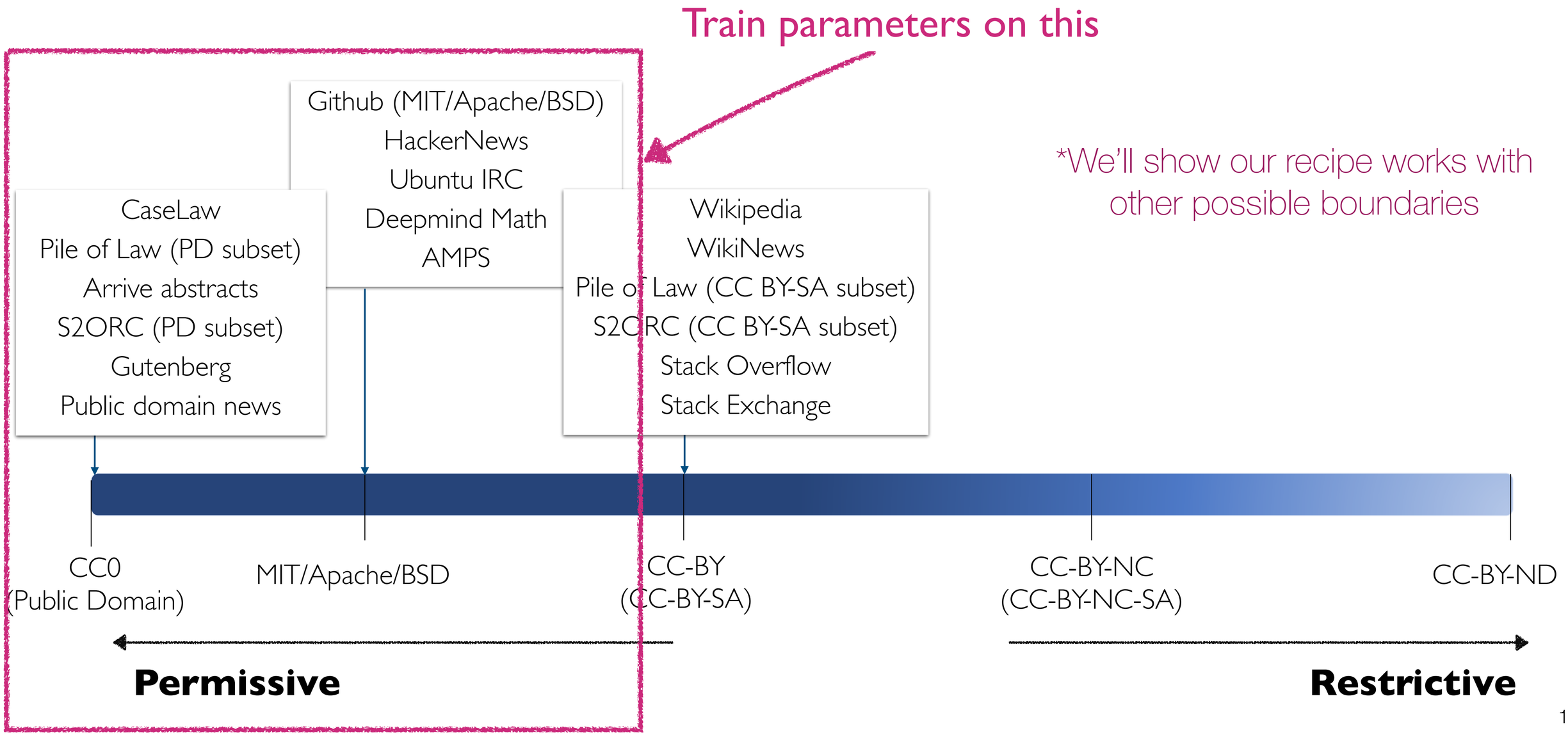
Permissive text



Open License Corpus (OLC)



Open License Corpus (OLC)



OLC Statistics

Our training data

Domain	Tokens (B)	%
Code	58.9	59.1
Legal	27.1	27.2
Conversation	5.9	5.9
Math	3.5	3.5
Books	2.9	2.9
Science	1.2	1.2
News	0.2	0.2
Wikipedia	0.0	0.0
Unverified web	0.0	0.0
Total	99.6	100.0

OLC Statistics

	Our training data		The Pile	
Domain	**Tokens (B)**	**%**	**Tokens (B)**	**%**
Code	58.9	59.1	32.6	9.8
Legal	27.1	27.2	30.8	9.3
Conversation	5.9	5.9	33.1	10.0
Math	3.5	3.5	7.1	2.1
Books	2.9	2.9	47.1	14.2
Science	1.2	1.2	86.0	26.0
News	0.2	0.2	-†	-†
Wikipedia	0.0	0.0	12.1	3.7
Unverified web	0.0	0.0	83.1	25.0
Total	99.6	100.0	331.9	100.0

OLC Statistics

	Our training data		The Pile	
Domain	**Tokens (B)**	**%**	**Tokens (B)**	**%**
Code	58.9	59.1	32.6	9.8
Legal	27.1	27.2	30.8	9.3
Conversation	5.9	5.9	33.1	10.0
Math	3.5	3.5	7.1	2.1
Books	2.9	2.9	47.1	14.2
Science	1.2	1.2	86.0	26.0
News	0.2	0.2	-†	-†
Wikipedia	0.0	0.0	12.1	3.7
Unverified web	0.0	0.0	83.1	25.0
Total	99.6	100.0	331.9	100.0

OLC is smaller than the Pile (but also, larger than previously thought!)

OLC Statistics

	Our training data		The Pile	
Domain	**Tokens (B)**	**%**	**Tokens (B)**	**%**
Code	58.9	59.1	32.6	9.8
Legal	27.1	27.2	30.8	9.3
Conversation	5.9	5.9	33.1	10.0
Math	3.5	3.5	7.1	2.1
Books	2.9	2.9	47.1	14.2
Science	1.2	1.2	86.0	26.0
News	0.2	0.2	-†	-†
Wikipedia	0.0	0.0	12.1	3.7
Unverified web	0.0	0.0	83.1	25.0
Total	99.6	100.0	331.9	100.0

OLC's distribution is different from that of typical pretaining corpora
(likely to cause *extreme domain shift*)

OLC Examples

OLC Examples

Legal and government text

Exhibit 1.1 FIRSTBANK CORPORATION (a Michigan corporation) 33,000 Shares of Fixed Rate Cumulative Perpetual Preferred Stock, Series A Preferred Stock UNDERWRITING AGREEMENT ●, 2012 ● as Representatives of the several Underwriters ● and ● Ladies and Gentlemen: Firstbank Corporation, a Michigan

OLC Examples

Legal and government text

Exhibit 1.1 FIRSTBANK CORPORATION (a Michigan corporation) 33,000 Shares of Fixed Rate Cumulative Perpetual Preferred Stock, Series A Preferred Stock UNDERWRITING AGREEMENT ●, 2012 ● as Representatives of the several Underwriters ● and ● Ladies and Gentlemen: Firstbank Corporation, a Michigan

Old scientific text

1900 M.DR R Dinwiddie EXPERIMENTAL TUBERCULOSIS, HUMAN AND BOVINE, IN THE DOMESTIC ANIMALS. I come under the observation of the veterinarian. The disease in these animals which I have just named presents two features which I wish

OLC Examples

Legal and government text

Exhibit 1.1 FIRSTBANK CORPORATION (a Michigan corporation) 33,000 Shares of Fixed Rate Cumulative Perpetual Preferred Stock, Series A Preferred Stock UNDERWRITING AGREEMENT ●, 2012 ● as Representatives of the several Underwriters ● and ● Ladies and Gentlemen: Firstbank Corporation, a Michigan

Old scientific text

1900 M.DR R Dinwiddie EXPERIMENTAL TUBERCULOSIS, HUMAN AND BOVINE, IN THE DOMESTIC ANIMALS. I come under the observation of the veterinarian. The disease in these animals which I have just named presents two features which I wish

Copyright-expired books

"My First Years As A Frenchwoman 1876-1879 by Mary King Waddington"…

OLC Examples

Legal and government text

Exhibit 1.1 FIRSTBANK CORPORATION (a Michigan corporation) 33,000 Shares of Fixed Rate Cumulative Perpetual Preferred Stock, Series A Preferred Stock UNDERWRITING AGREEMENT ●, 2012 ● as Representatives of the several Underwriters ● and ● Ladies and Gentlemen: Firstbank Corporation, a Michigan

Copyright-expired books

"My First Years As A Frenchwoman 1876-1879 by Mary King Waddington"…

Old scientific text

1900 M.DR R Dinwiddie EXPERIMENTAL TUBERCULOSIS, HUMAN AND BOVINE, IN THE DOMESTIC ANIMALS. I come under the observation of the veterinarian. The disease in these animals which I have just named presents two features which I wish

Public domain news

WASHINGTON - White House economic adviser Larry Kudlow said on Monday the Trump administration wants to end subsidies for electric cars and other items, including renewable energy sources…

OLC is Publicly Available

Datasets: kernelmachine/**open-license-corpus** like 3

Tasks: Text Generation | Languages: English | Size Categories: 100B<n<1T | License: apache-2.0

Dataset card | Files and versions | Community | Settings

Dataset Viewer (First 5GB) Auto-converted to Parquet | API | Go to dataset viewer

Subset: ccby_law (279k rows)

Split: train (279k rows)

text (string)
" JIYA ACQUISITION CORP. 780 Third Avenue November 18, 2020 Jiya Holding Company LLC 638 Middlefield Road Palo Alto, CA 94301 Re: Administrative Services Agreement Ladies and Gentlemen: This letter…
"Name: Commission Regulation (EEC) No 1722/93 of 30 June 1993 laying down detailed rules for the application of Council Regulations (EEC) No 1766/92 and (EEC) No 1418/76 concerning production refunds in the cereals and ric…
"Exhibit 16.1 January 26, 2012 Office of the Chief Accountant Securities and Exchange Commission 100F Street, NE Washington, D.C. 20549 Dear Sir/Madam: We have read the statements included under Item 4.01 in the Form 8-K…
"Exhibit 1.1 FIRSTBANK CORPORATION (a Michigan corporation) 33,000 Shares of Fixed Rate Cumulative Perpetual Preferred Stock, Series A Preferred Stock UNDERWRITING AGREEMENT ●, 2012 ● as Representatives of the several…
"UNITED STATES SECURITIES AND EXCHANGE COMMISSION Washington, D.C.20549 FORM 8-A/A (Amendment No. 1) FOR

Downloads last month 44

Use in dataset library | Edit dataset card

Train in AutoTrain | Evaluate models

HF Leaderboard

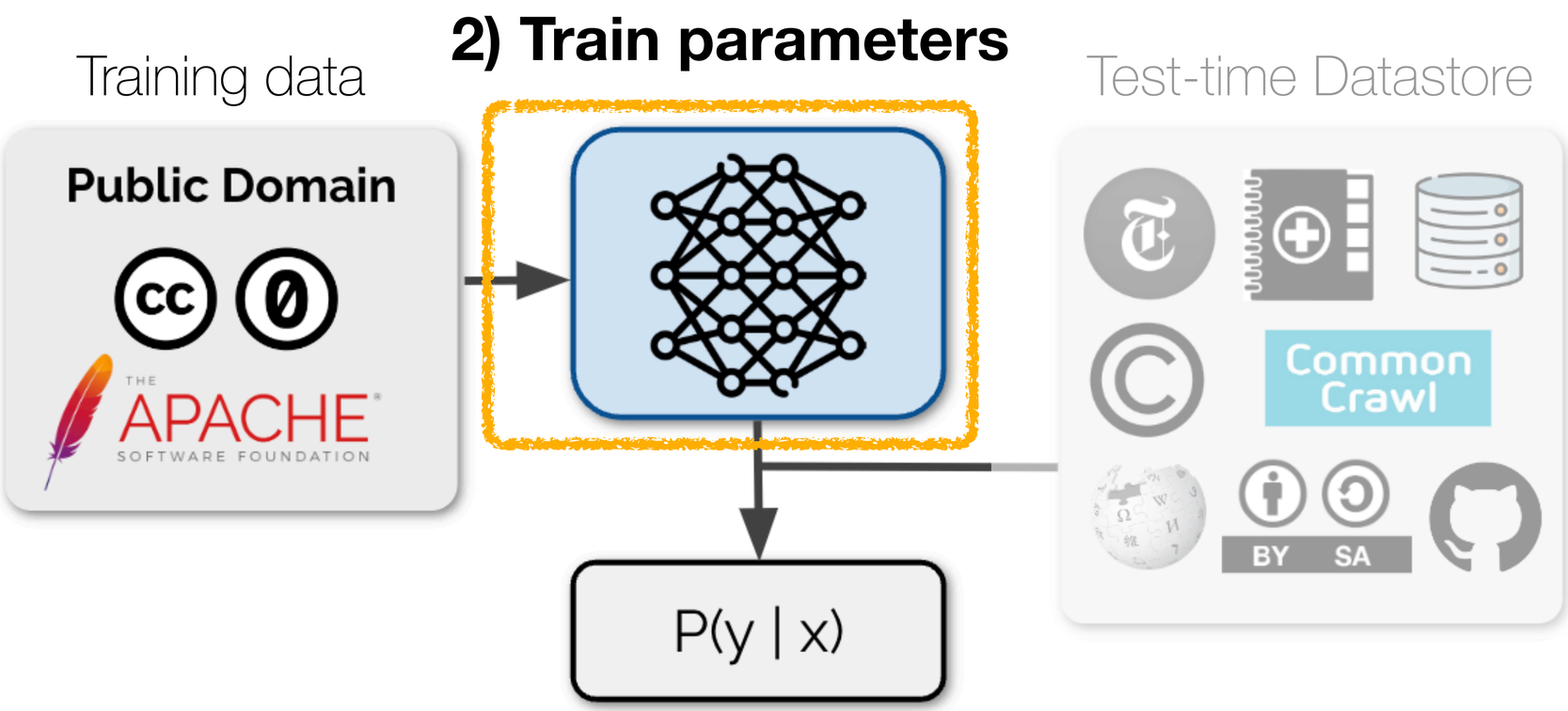
Repository: Silo LM repository

Paper: Silo LM paper

Point of Contact: Suchin Gururangan

Size of the auto-converted Parquet files (First 5GB): 30.2 GB

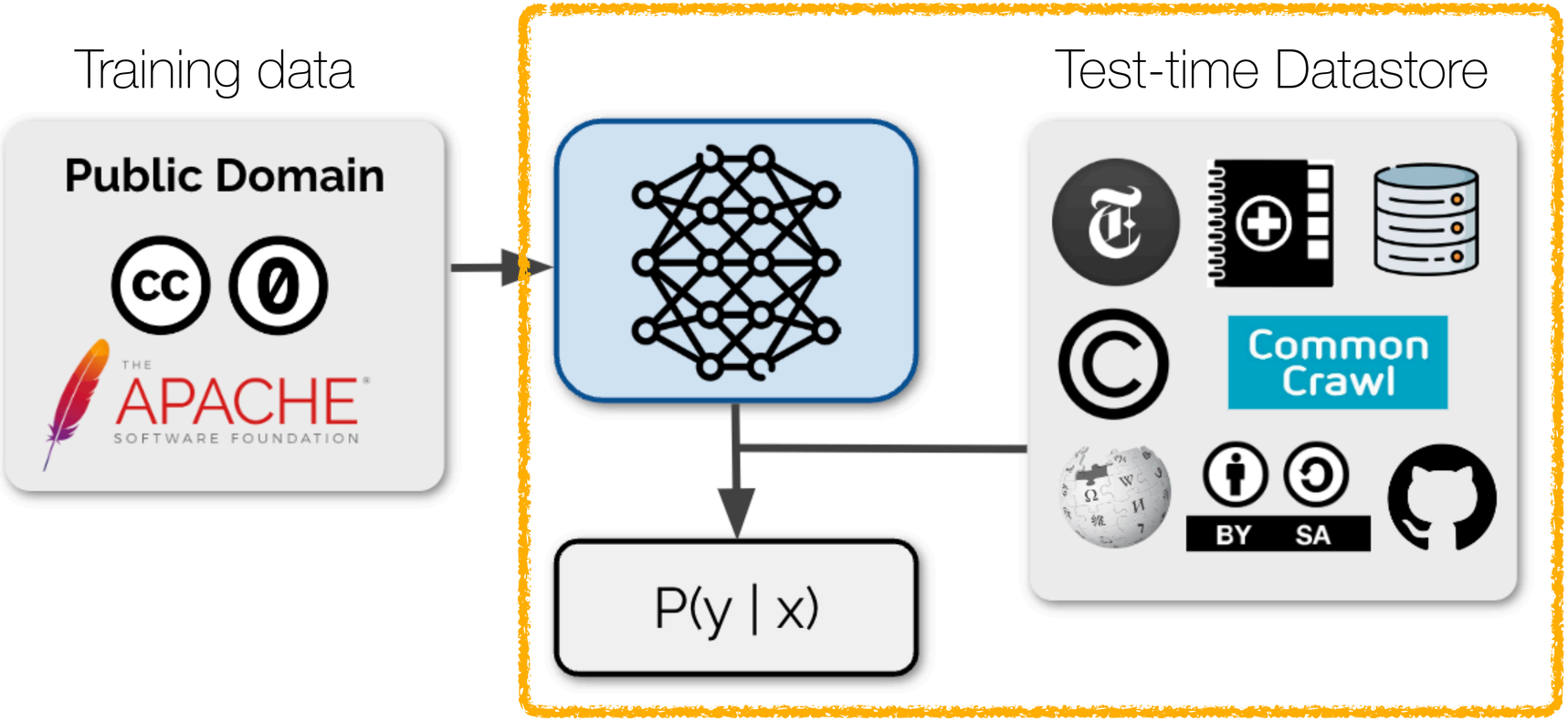
A SILO Language Model



Training a Parametric Component

- Train 1.3B-parameter LMs with the LLaMa architecture
- 128 A100 GPUs across 16 nodes
- Multiple epochs of training (Muennighoff et al 2023) with early stopping
- Updating the tail of the domain distribution by 3x

A SILO Language Model



3) Inference using a nonparametric datastore

How to Use a Nonparametric Datastore?

- Assumption: In-distribution data is available at test time
  - Construct the datastore using the data in-distribution to the eval data

How to Use a Nonparametric Datastore?

- Assumption: In-distribution data is available at test time
  - Construct the datastore using the data in-distribution to the eval data
- Use a retrieval approach to leverage this datastore
  - **Retrieve-in-context LM** (Shi et al 2023, Ram et al 2023)
  - ***k*NN-LM** (Khandelwal et al. 2020)
  - Spoiler: *k*NN-LM outperforms Retrieve-in-context LM but is more expensive

1. Retrieval-in-context LM

Datastore:

©

Ram et al. 2023. “In-Context Retrieval-Augmented Language Models”
Shi et al. 2023. “REPLUG: Retrieval-Augmented Black-Box Language Models”

1. Retrieval-in-context LM

Datastore:

©

Voldemort was ready. As Harry shouted, "Expelliarmus!" Voldemort cried, "Avada Kedavra!" A jet of green light issued from Voldemort's wand just as a jet of red light blasted from Harry's

Ram et al. 2023. "In-Context Retrieval-Augmented Language Models"
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

1. Retrieval-in-context LM

Datastore:

©

Voldemort was ready. As Harry shouted, "Expelliarmus!" Voldemort cried, "Avada Kedavra!" A jet of green light issued from Voldemort's wand just as a jet of red light blasted from Harry's

Test input:

Harry felt Greenback collapse against him … on the floor as a jet of

Ram et al. 2023. "In-Context Retrieval-Augmented Language Models"
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

1. Retrieval-in-context LM

Datastore:

©

Voldemort was ready. As Harry shouted, "Expelliarmus!" Voldemort cried, "Avada Kedavra!" A jet of green light issued from Voldemort's wand just as a jet of red light blasted from Harry's

Test input:

Harry felt Greenback collapse against him … on the floor as a jet of

Ram et al. 2023. "In-Context Retrieval-Augmented Language Models"
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

1. Retrieval-in-context LM

Datastore:

©

Voldemort was ready. As Harry shouted, "Expelliarmus!" Voldemort cried, "Avada Kedavra!" A jet of green light issued from Voldemort's wand just as a jet of red light blasted from Harry's

Voldemort was ready. As Harry shouted, "Expelliarmus!" Voldemort cried, "Avada Kedavra!" A jet of green light issued from Voldemort's wand just as a jet of red light blasted …

Harry felt Greenback collapse against him … a jet of

Test input:

Harry felt Greenback collapse against him … on the floor as a jet of

Ram et al. 2023. "In-Context Retrieval-Augmented Language Models"
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

1. Retrieval-in-context LM

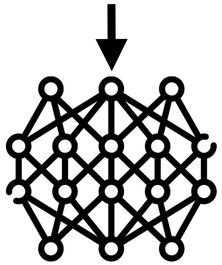
Datastore:

©

Voldemort was ready. As Harry shouted, "Expelliarmus!" Voldemort cried, "Avada Kedavra!" A jet of green light issued from Voldemort's wand just as a jet of red light blasted from Harry's

Voldemort was ready. As Harry shouted, "Expelliarmus!" Voldemort cried, "Avada Kedavra!" A jet of green light issued from Voldemort's wand just as a jet of red light blasted …

Harry felt Greenback collapse against him … a jet of



Test input:

Harry felt Greenback collapse against him … on the floor as a jet of

Ram et al. 2023. "In-Context Retrieval-Augmented Language Models"
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

1. Retrieval-in-context LM

Datastore:

©

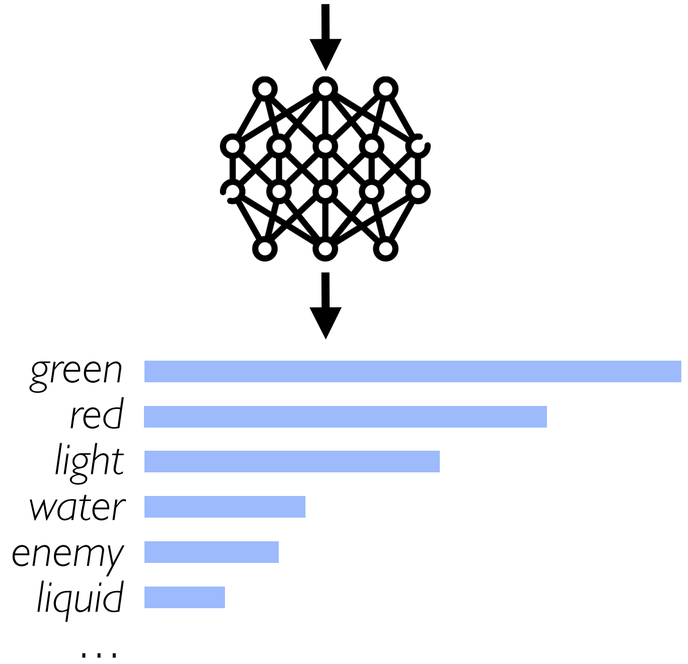
Voldemort was ready. As Harry shouted, "Expelliarmus!" Voldemort cried, "Avada Kedavra!" A jet of green light issued from Voldemort's wand just as a jet of red light blasted from Harry's

Test input:

Harry felt Greenback collapse against him … on the floor as a jet of

Voldemort was ready. As Harry shouted, "Expelliarmus!" Voldemort cried, "Avada Kedavra!" A jet of green light issued from Voldemort's wand just as a jet of red light blasted …

Harry felt Greenback collapse against him … a jet of



Ram et al. 2023. "In-Context Retrieval-Augmented Language Models"
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

2. *k*NN-LM

Datastore:

©

Voldemort's wand just as a jet of **red** light

Voldemort cried, "Avada Kedavra!" A jet of **green**

"The Boy Who Lived." He … **green**

… is operated or driven by a jet of **water**.

Pick up a flat rock, skip it across **Green** River

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

2. *k*NN-LM

Datastore:

©

Voldemort's wand just as a jet of **red** light

Voldemort cried, "Avada Kedavra!" A jet of **green**

"The Boy Who Lived." He … **green**

… is operated or driven by a jet of **water**.

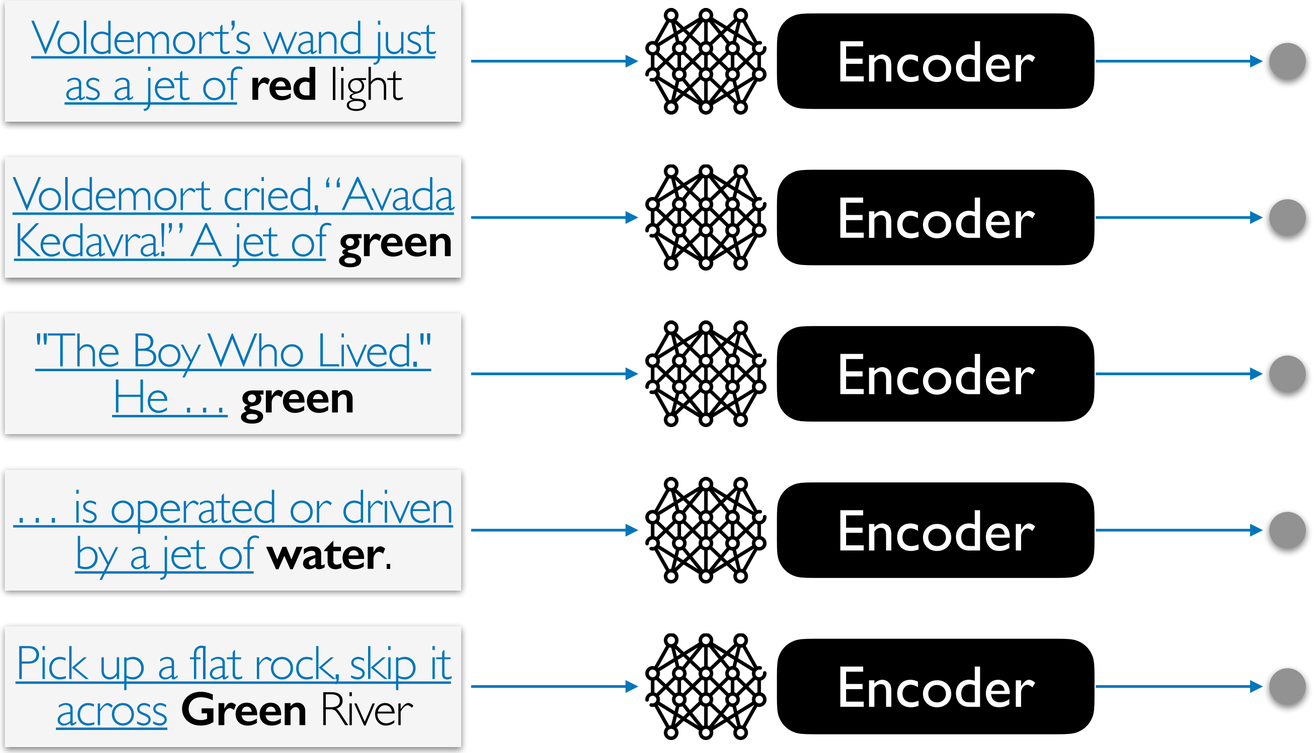
Pick up a flat rock, skip it across **Green** River

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

2. *k*NN-LM

Datastore:

©

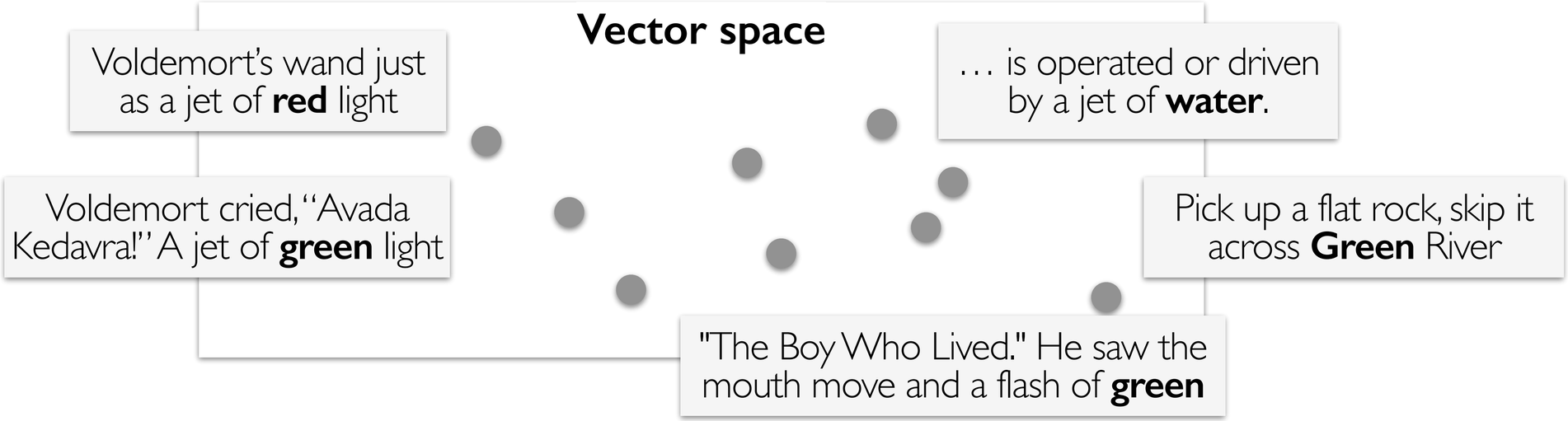


Khandelwal et al. 2020. Generalization through Memorization: Nearest Neighbor Language Models.

2. *k*NN-LM

Datastore:

©



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

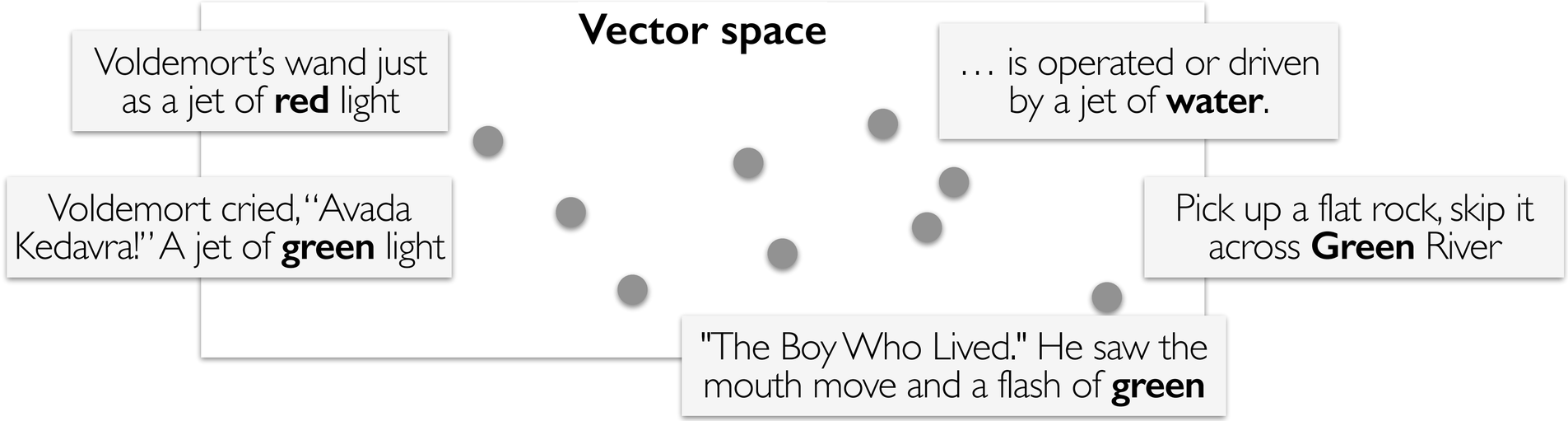
2. *k*NN-LM

Datastore:

©

Test input:

Harry felt Greenback collapse against him … on the floor as a jet of



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

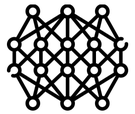
2. *k*NN-LM

Datastore:

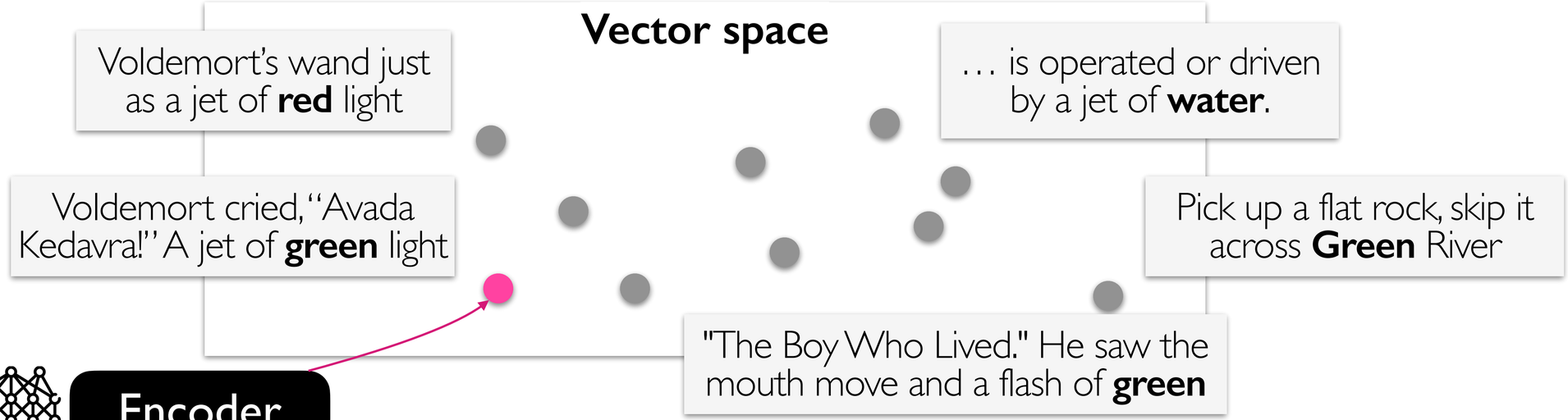
©

Test input:

Harry felt Greenback collapse against him … on the floor as a jet of



Encoder



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

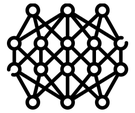
2. *k*NN-LM

Datastore:

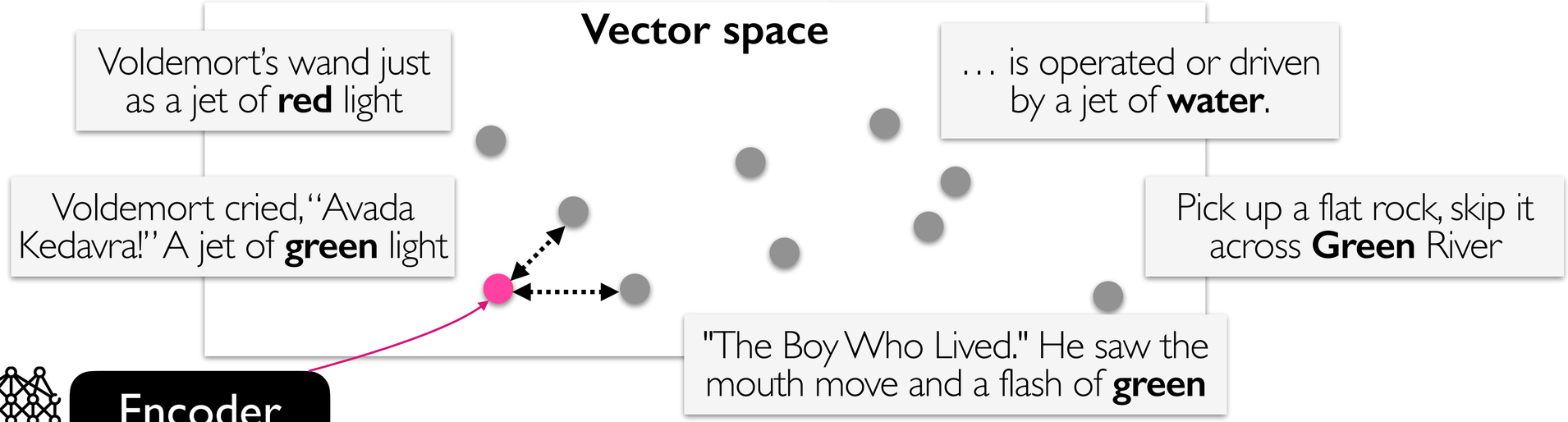
©

Test input:

Harry felt Greenback collapse against him … on the floor as a jet of



Encoder



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

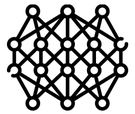
2. *k*NN-LM

Datastore:

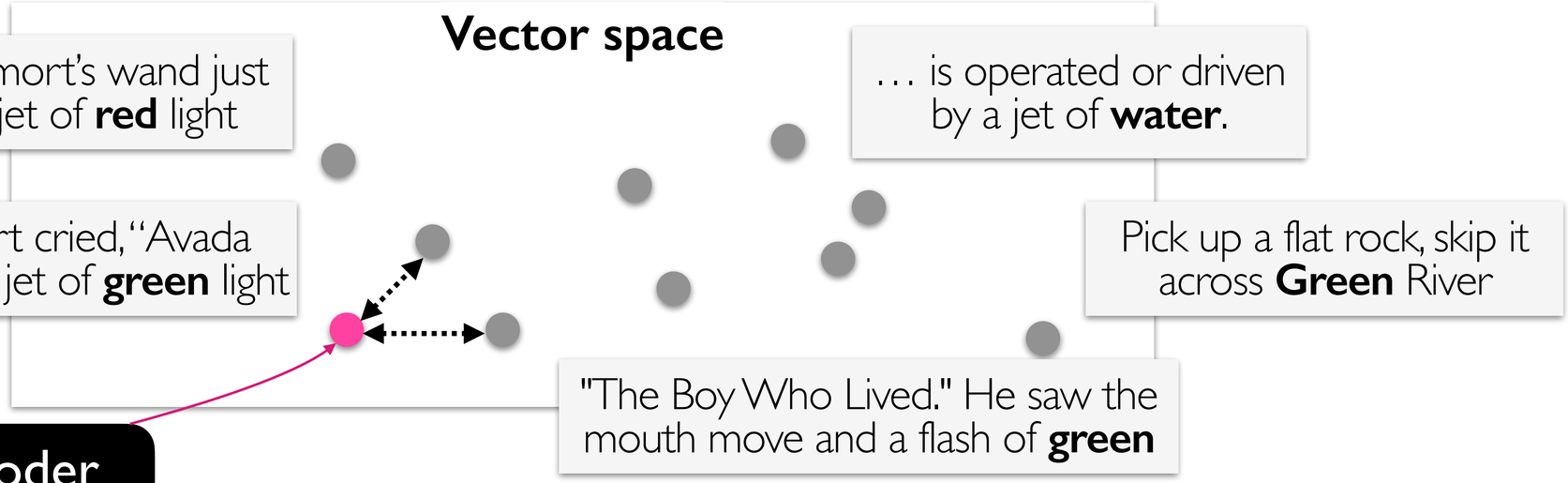
©

Test input:

Harry felt Greenback collapse against him … on the floor as a jet of



Encoder



Voldemort cried, "Avada Kedavra!" A jet of **green**

"The Boy Who Lived." He saw the mouth move and a flash of **green**

Voldemort's wand just as a jet of **red**

… is operated or driven by a jet of **water**.

…

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

2. *k*NN-LM

Datastore:

©

$P_{k\text{NN}}(y \mid x)$

Voldemort cried, ''Avada Kedavra!'' A jet of **green**

"The Boy Who Lived." He saw the mouth move and a flash of **green**

Voldemort's wand just as a jet of **red**

… is operated or driven by a jet of **water**.

…

Test input:

Harry felt Greenback collapse against him … on the floor as a jet of

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

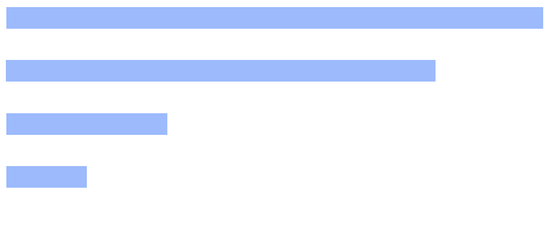
2. *k*NN-LM

Datastore:

©

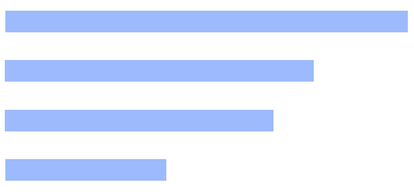
Voldemort cried, ''Avada Kedavra!'' A jet of **green**
"The Boy Who Lived." He saw the mouth move and a flash of **green**
Voldemort's wand just as a jet of **red**
… is operated or driven by a jet of **water**.
…

$P_{k\text{NN}}(y \mid x)$



$P_{\text{LM}}(y \mid x)$

water
light
green
liquid
…



Test input:

Harry felt Greenback collapse against him … on the floor as a jet of

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

2. *k*NN-LM

Datastore:

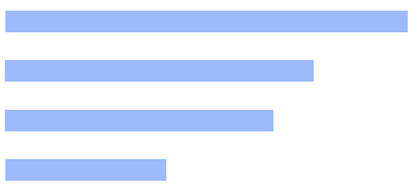
©

Voldemort cried, "Avada Kedavra!" A jet of **green**
"The Boy Who Lived." He saw the mouth move and a flash of **green**
Voldemort's wand just as a jet of **red**
… is operated or driven by a jet of **water**.
…

$P_{k\text{NN}}(y \mid x)$

$P_{\text{LM}}(y \mid x)$

water
light
green
liquid
…



Test input:

Harry felt Greenback collapse against him … on the floor as a jet of

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

λ: hyperparameter

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

RiC-LM vs. *k*NN-LM

RiC-LM	*k*NN-LM

RiC-LM vs. *k*NN-LM

RiC-LM	*k*NN-LM
☑ Only need black box access to LM	

RiC-LM vs. *k*NN-LM

RiC-LM	*k*NN-LM
☑ Only need black box access to LM	
☑ search space = # text blocks, e.g. 1-million search is efficient in speed and disk space	

RiC-LM vs. *k*NN-LM

RiC-LM	*k*NN-LM
☑ Only need black box access to LM	
☑ search space = # text blocks, e.g. 1-million search is efficient in speed and disk space	
✕ Can only affect probability distribution through the context given to the LM	

RiC-LM vs. *k*NN-LM

RiC-LM	*k*NN-LM
✓ Only need black box access to LM	✓ Bypass LM to directly affect probability distribution
✓ search space = # text blocks, e.g. 1-million search is efficient in speed and disk space	
✗ Can only affect probability distribution through the context given to the LM	

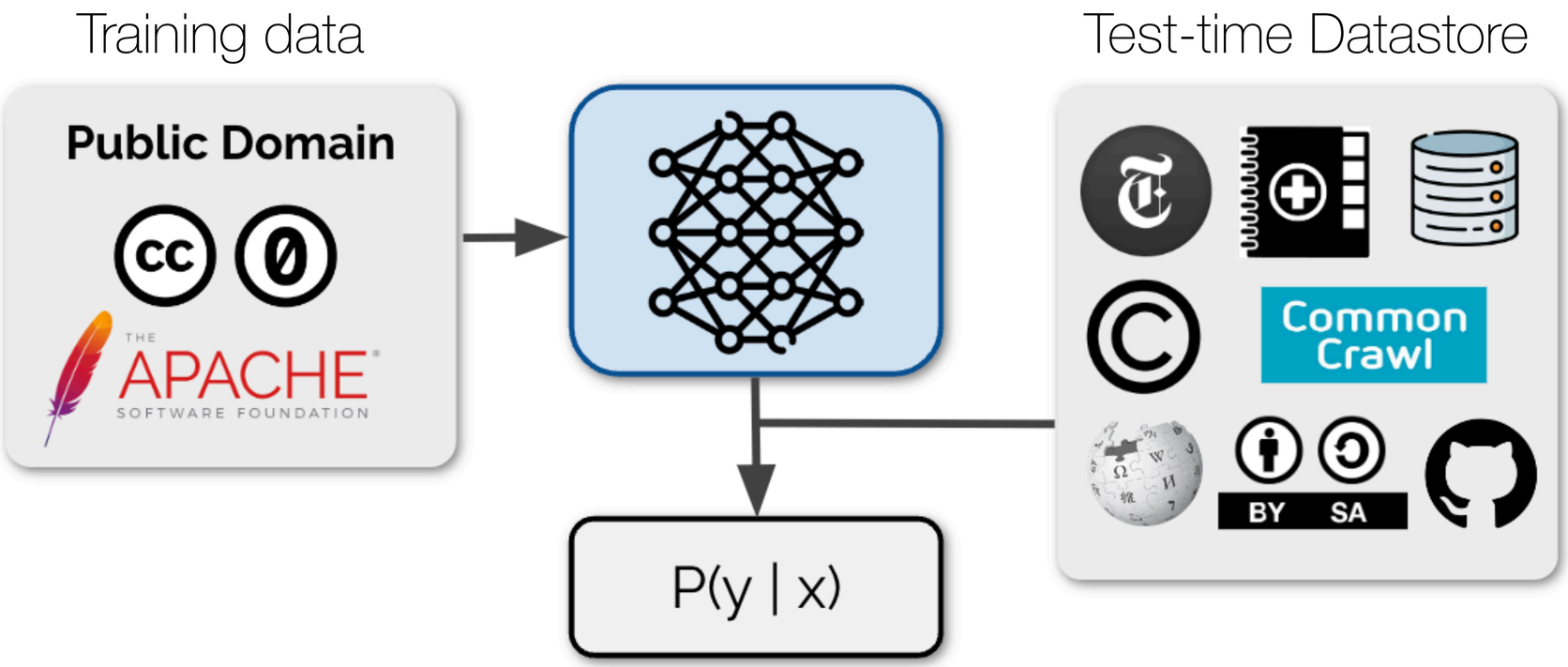
RiC-LM vs. *k*NN-LM

RiC-LM	*k*NN-LM
✓ Only need black box access to LM	✓ Bypass LM to directly affect probability distribution
✓ search space = # text blocks, e.g. 1-million search is efficient in speed and disk space	✗ search space = # tokens, e.g., 1-billion search is expensive in speed and disk space
✗ Can only affect probability distribution through the context given to the LM	

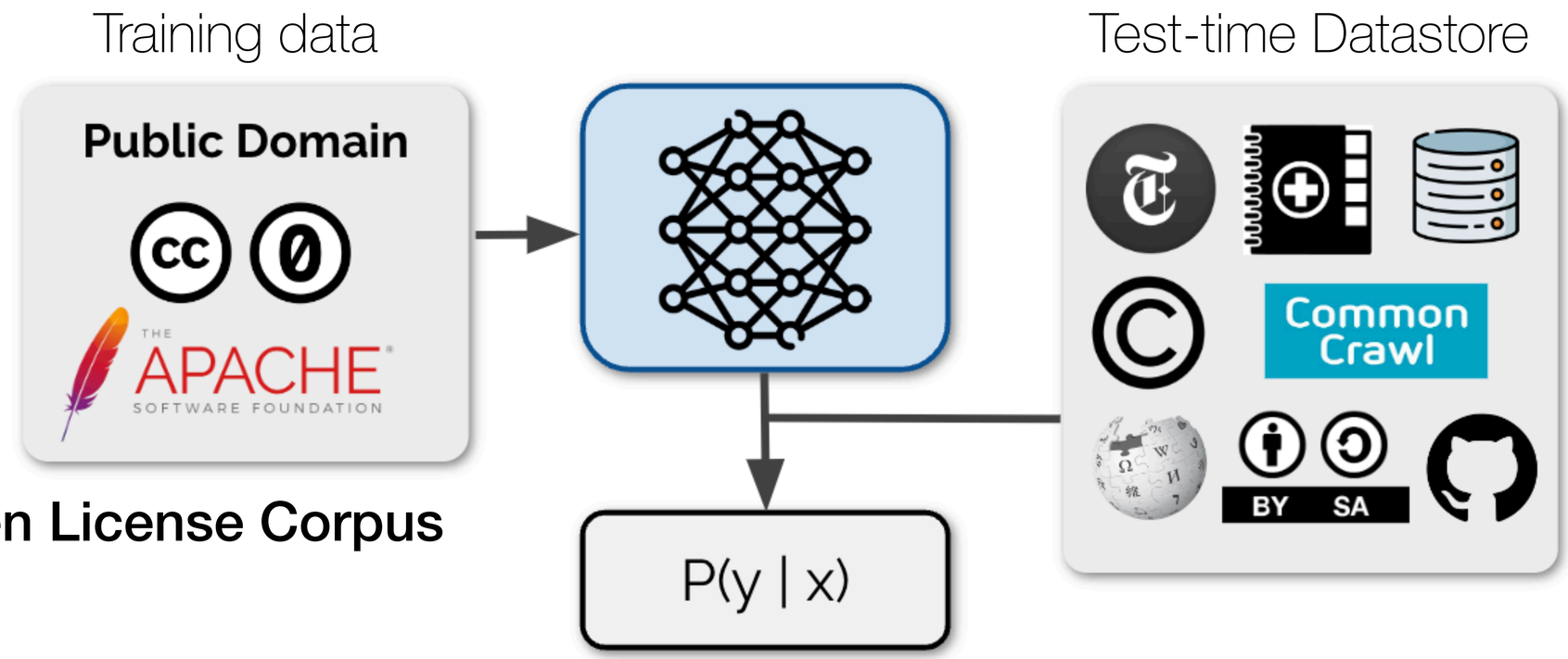
RiC-LM vs. *k*NN-LM

RiC-LM	*k*NN-LM
✓ Only need black box access to LM	✓ Bypass LM to directly affect probability distribution
✓ search space = # text blocks, e.g. 1-million search is efficient in speed and disk space	✗ search space = # tokens, e.g., 1-billion search is expensive in speed and disk space
✗ Can only affect probability distribution through the context given to the LM	✗ Requires similarity search for each new token

A SILO Language Model (Summary)

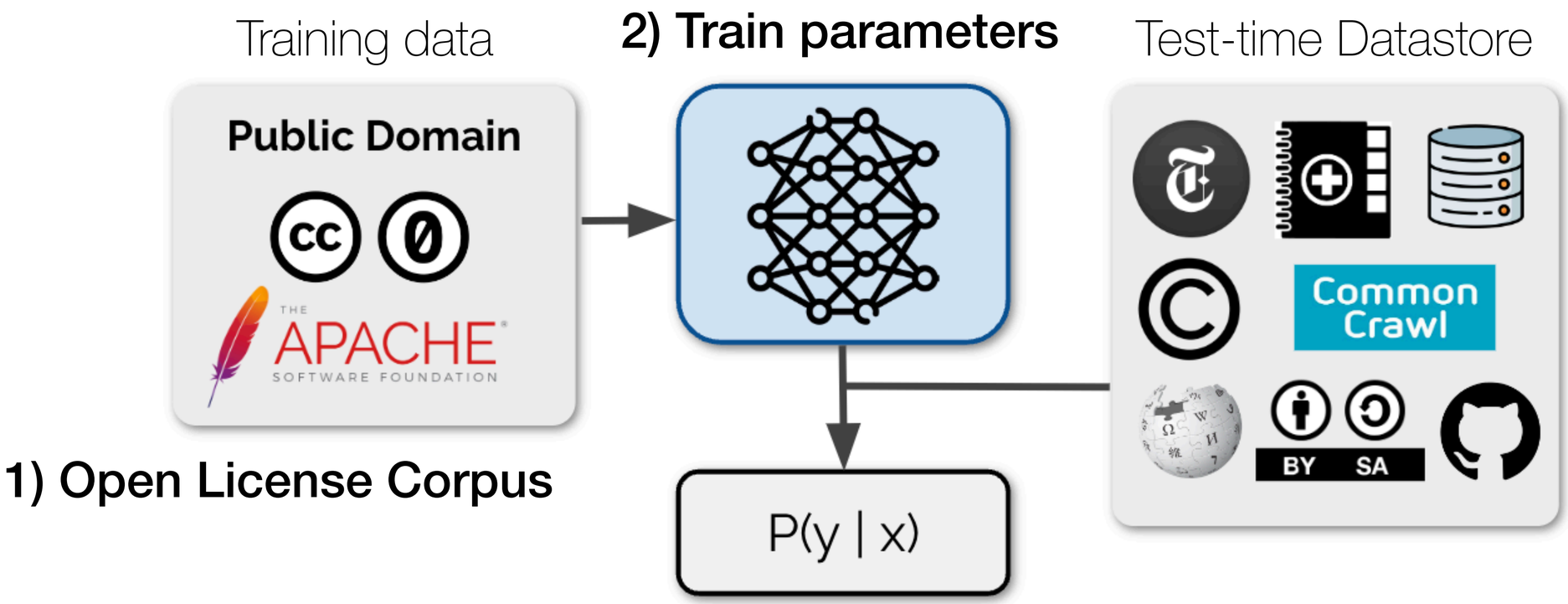


A SILO Language Model (Summary)

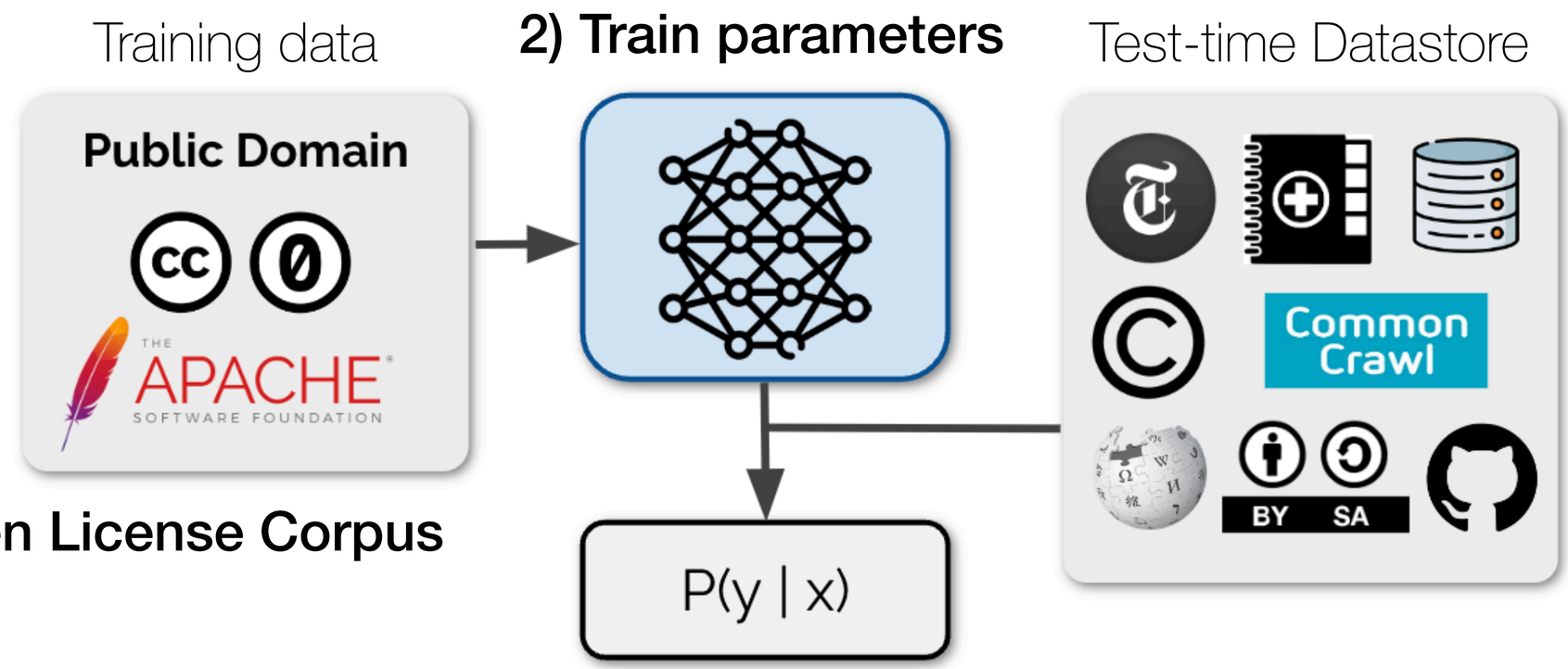


1) Open License Corpus

A SILO Language Model (Summary)



A SILO Language Model (Summary)



1) Open License Corpus

3) Inference using a nonparametric datastore
(RiC-LM or kNN-LM)

Content

Motivation

A SILO Language Model

Experiments

Conclusion & Future Work

Experimental Setup

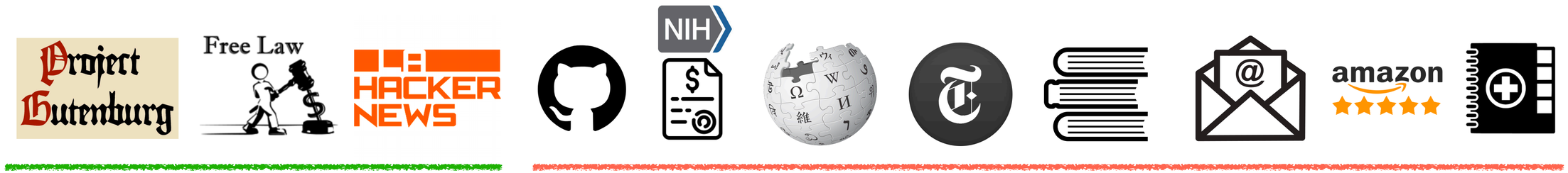
Evaluate Language Modeling Perplexity

Experimental Setup

In-distribution to OLC

Evaluate Language Modeling Perplexity

Experimental Setup

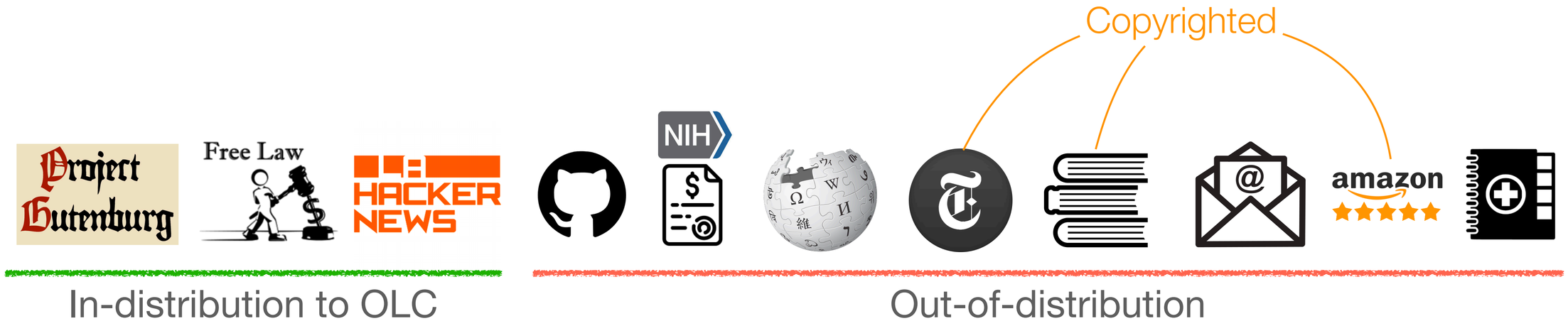


In-distribution to OLC

Out-of-distribution

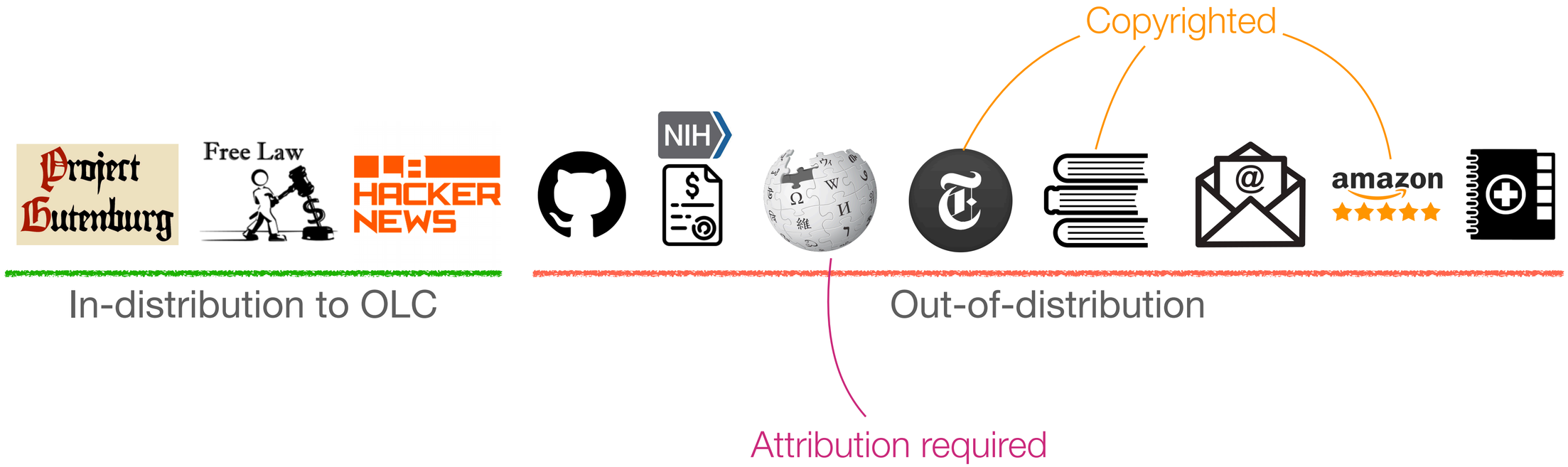
Evaluate Language Modeling Perplexity

Experimental Setup



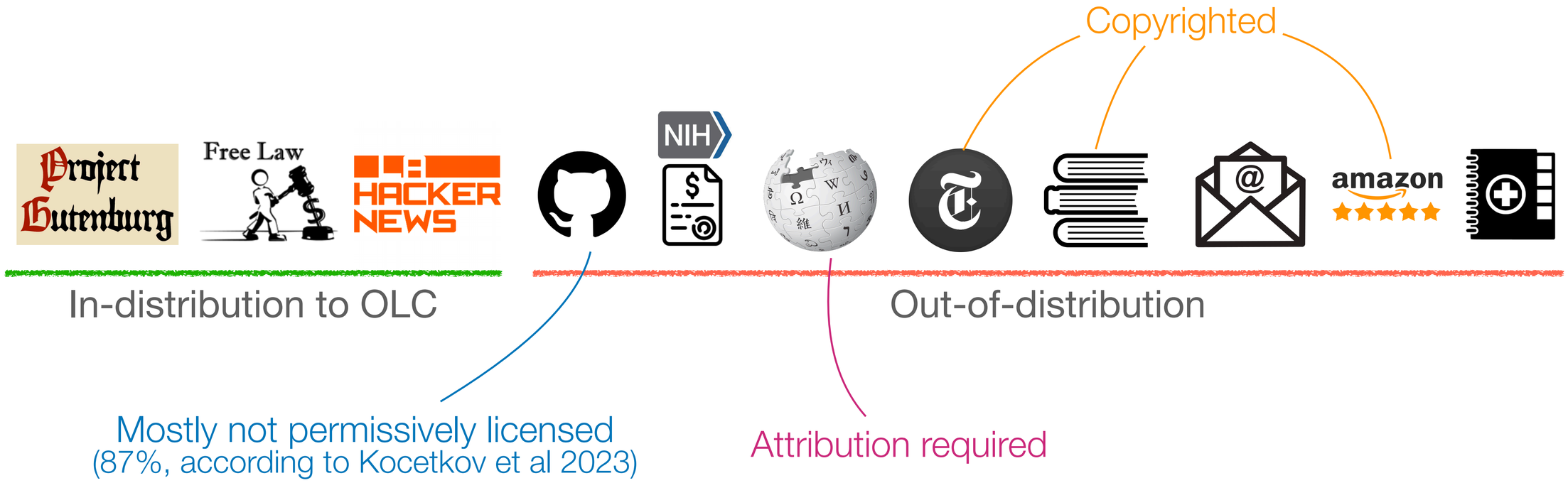
Evaluate Language Modeling Perplexity

Experimental Setup



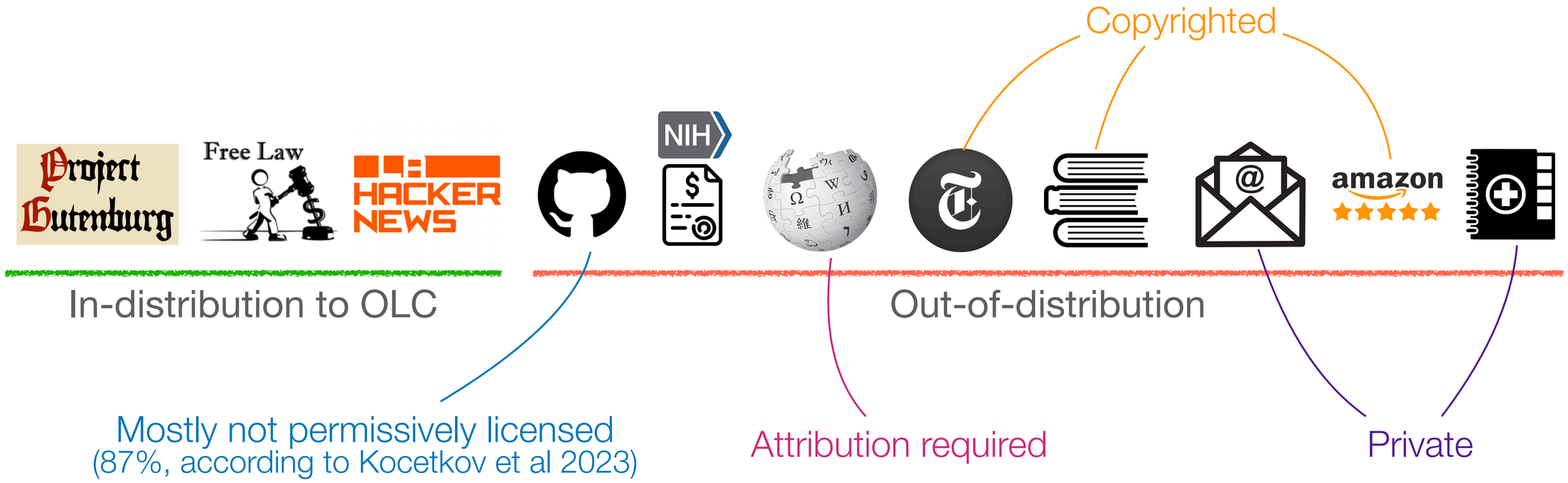
Evaluate Language Modeling Perplexity

Experimental Setup

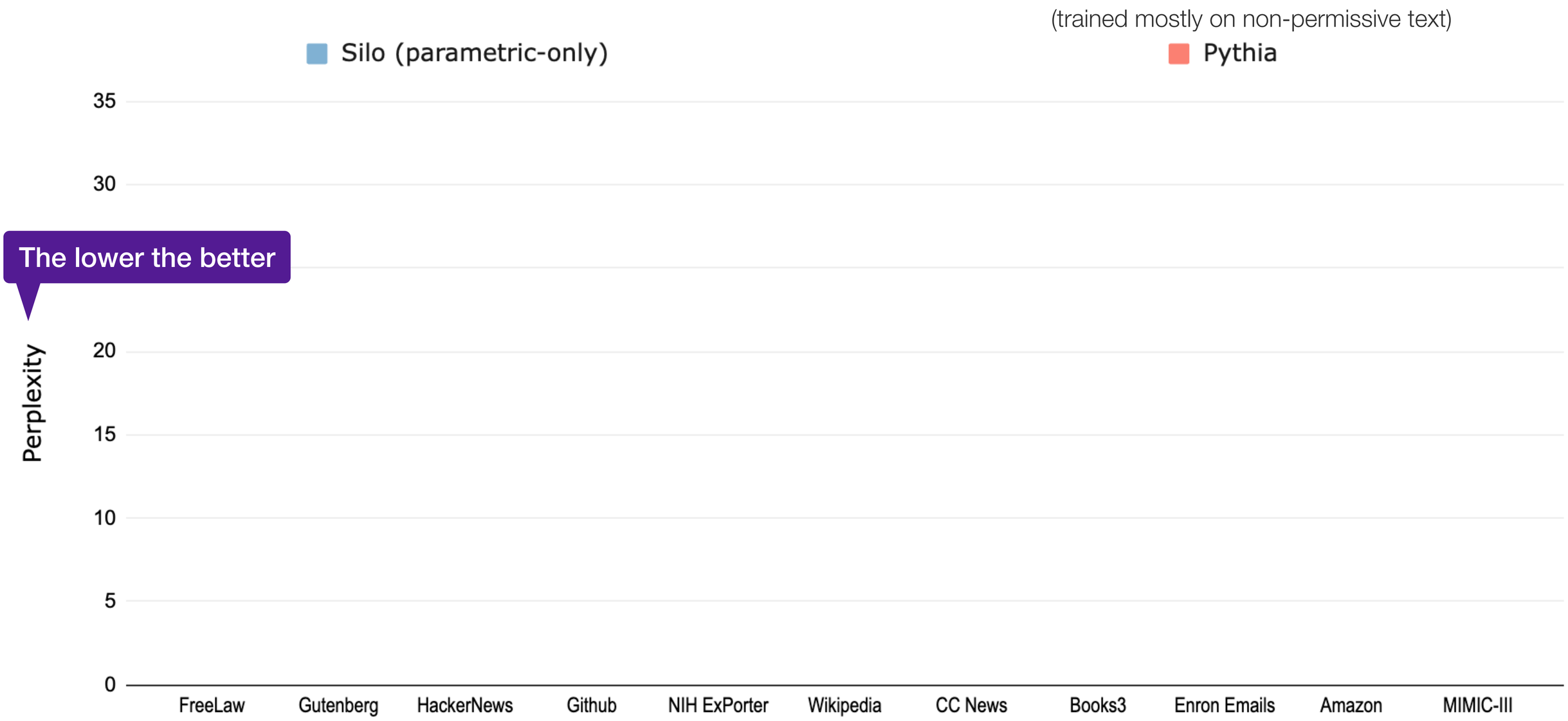


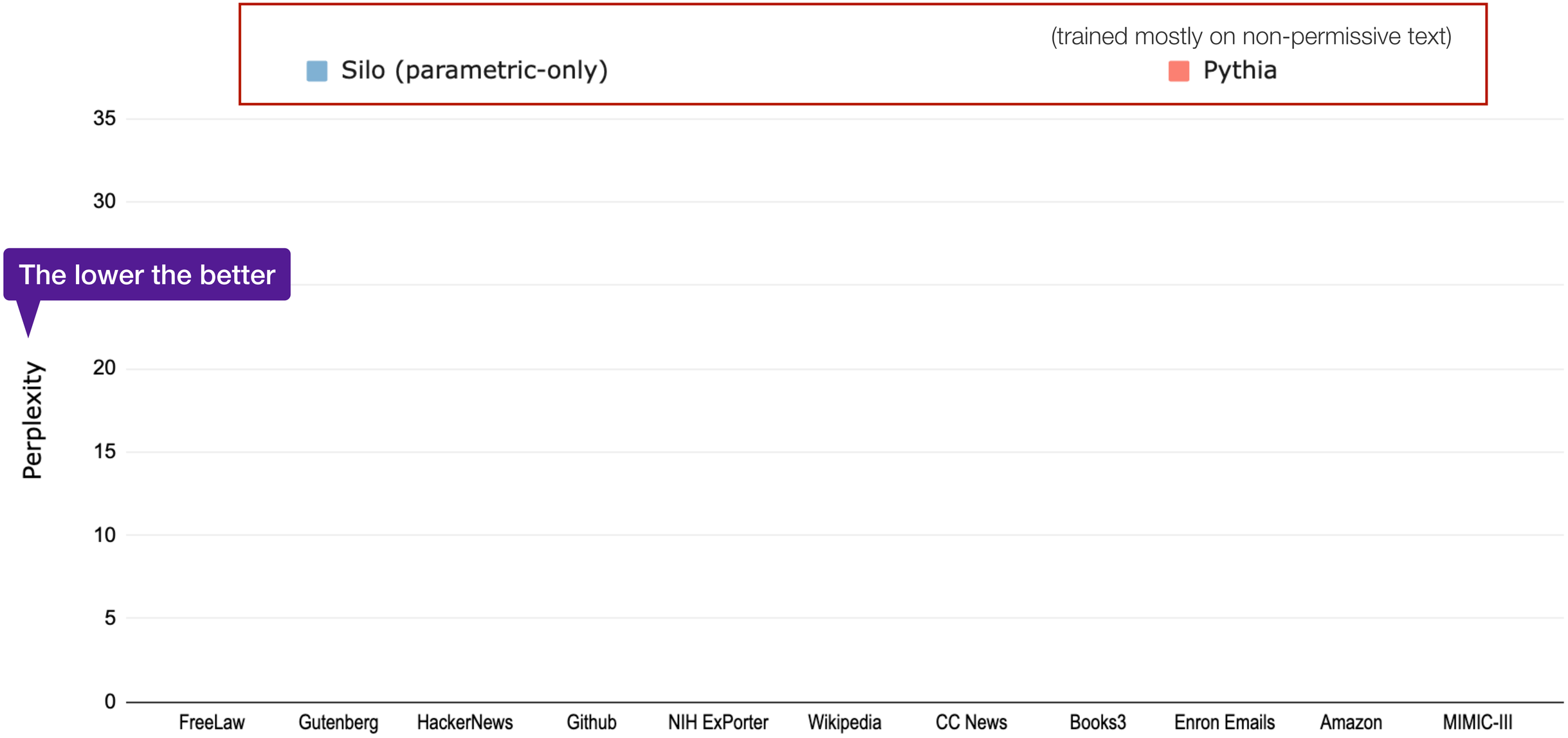
Evaluate Language Modeling Perplexity

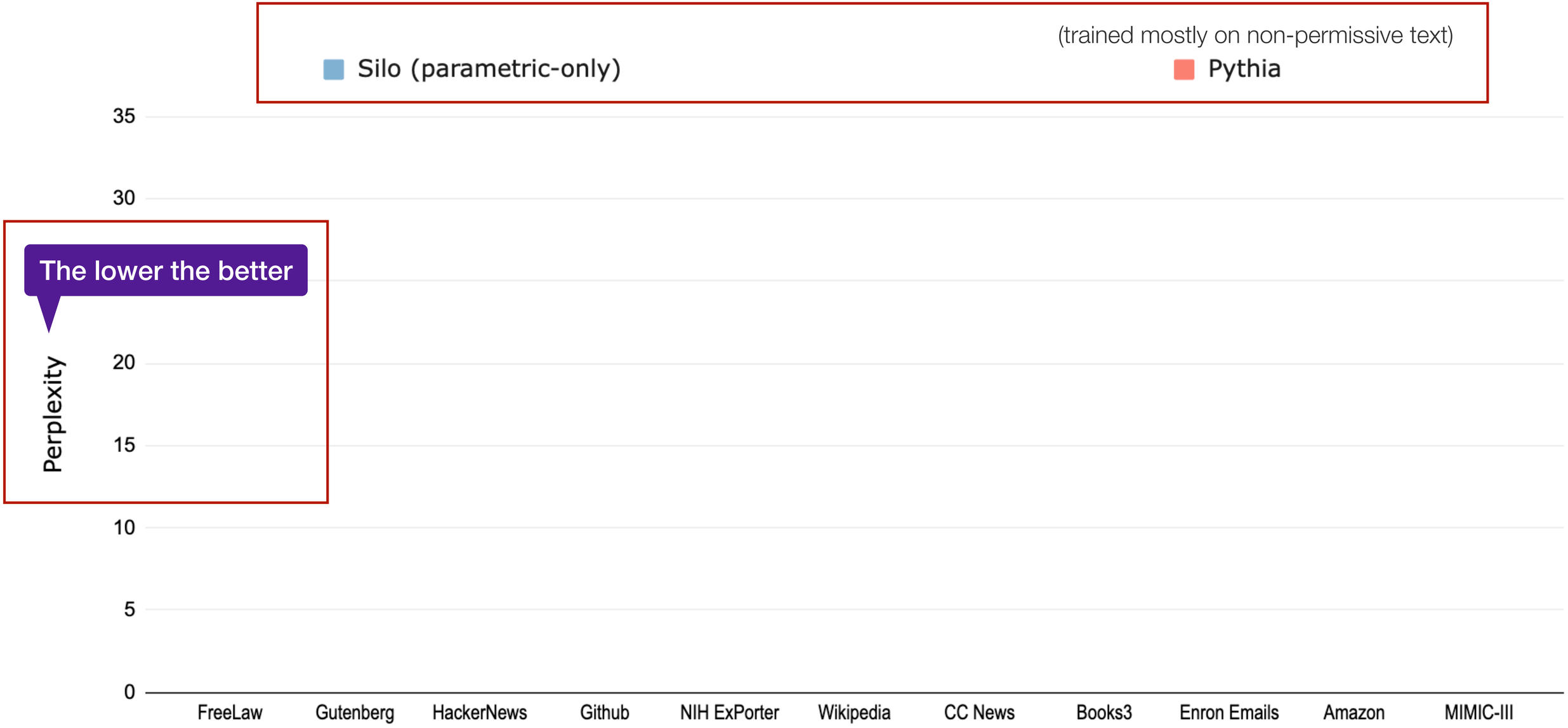
Experimental Setup

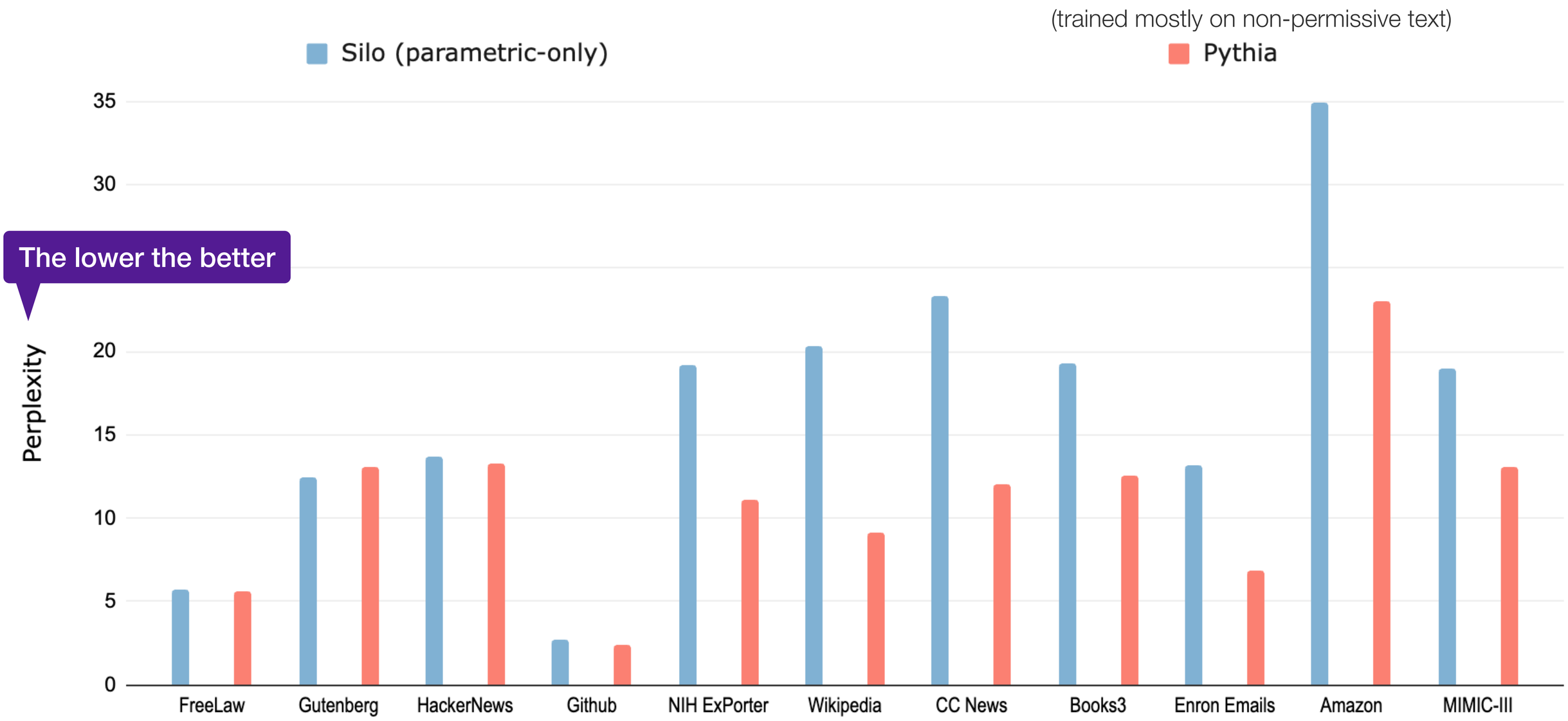


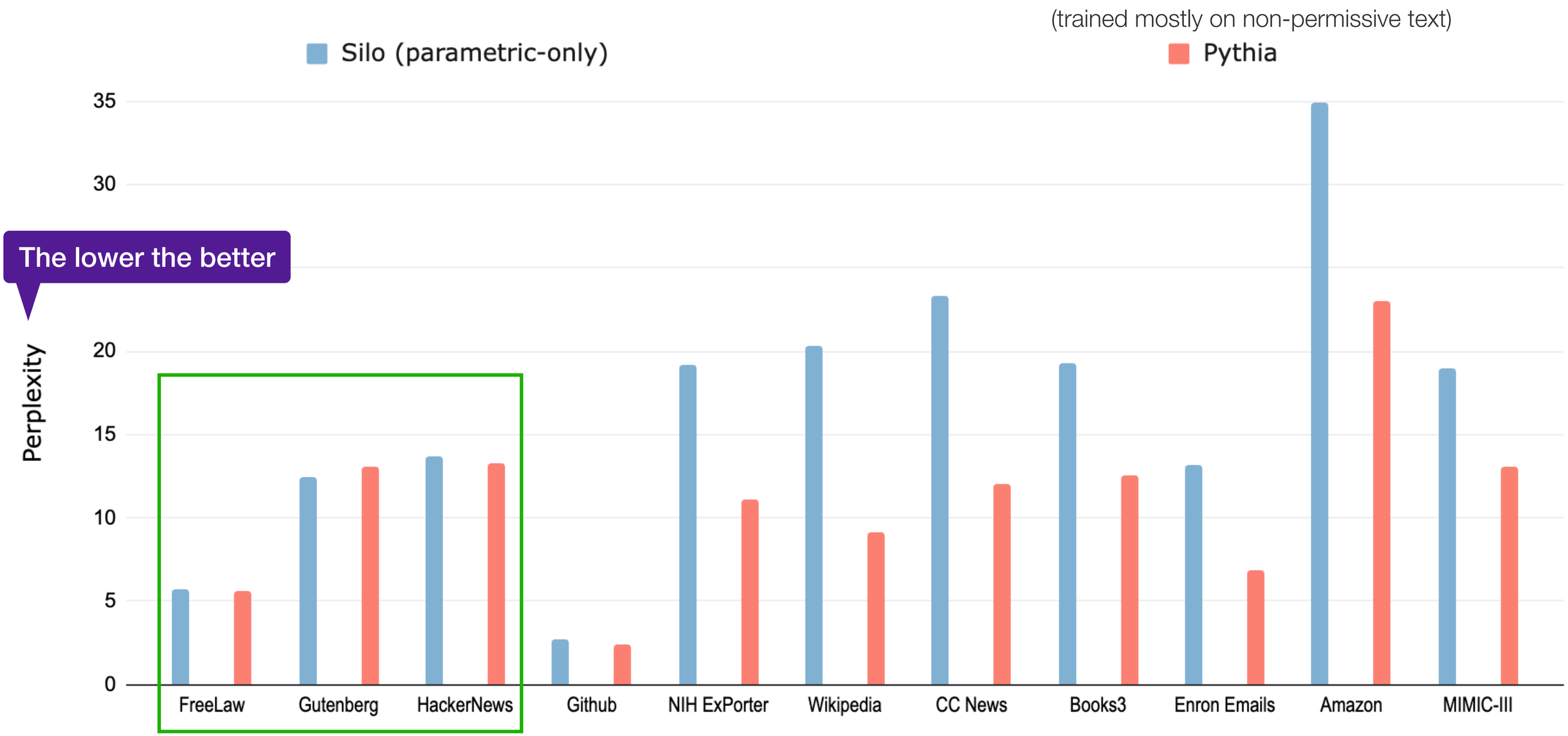
Evaluate Language Modeling Perplexity


(trained mostly on non-permissive text)
Silo (parametric-only)
Pythia
The lower the better
Perplexity
35
30
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III


(trained mostly on non-permissive text)
Silo (parametric-only)
Pythia
The lower the better
Perplexity
35
30
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III

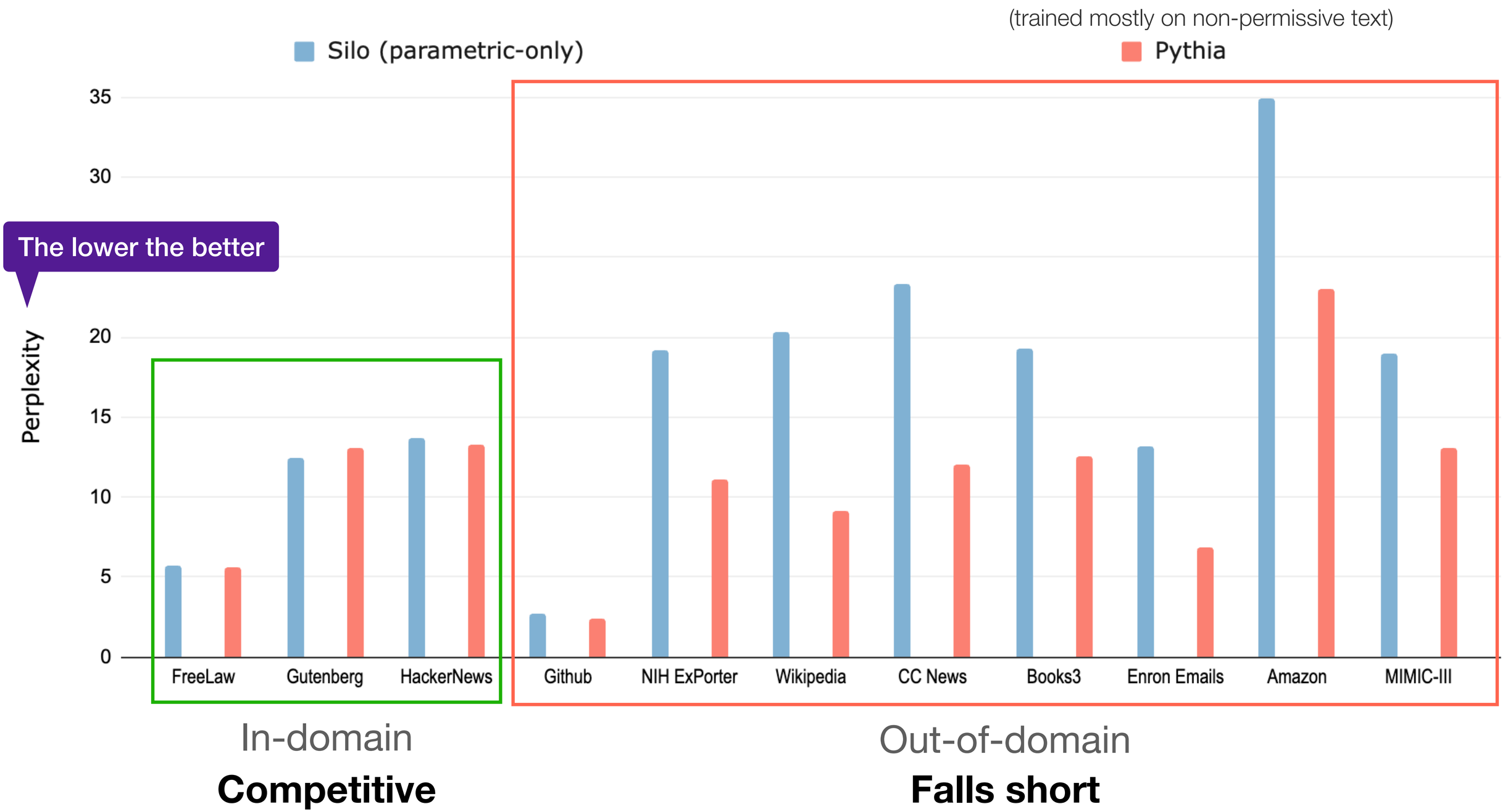

(trained mostly on non-permissive text)
Silo (parametric-only)
Pythia
The lower the better
Perplexity
35
30
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III

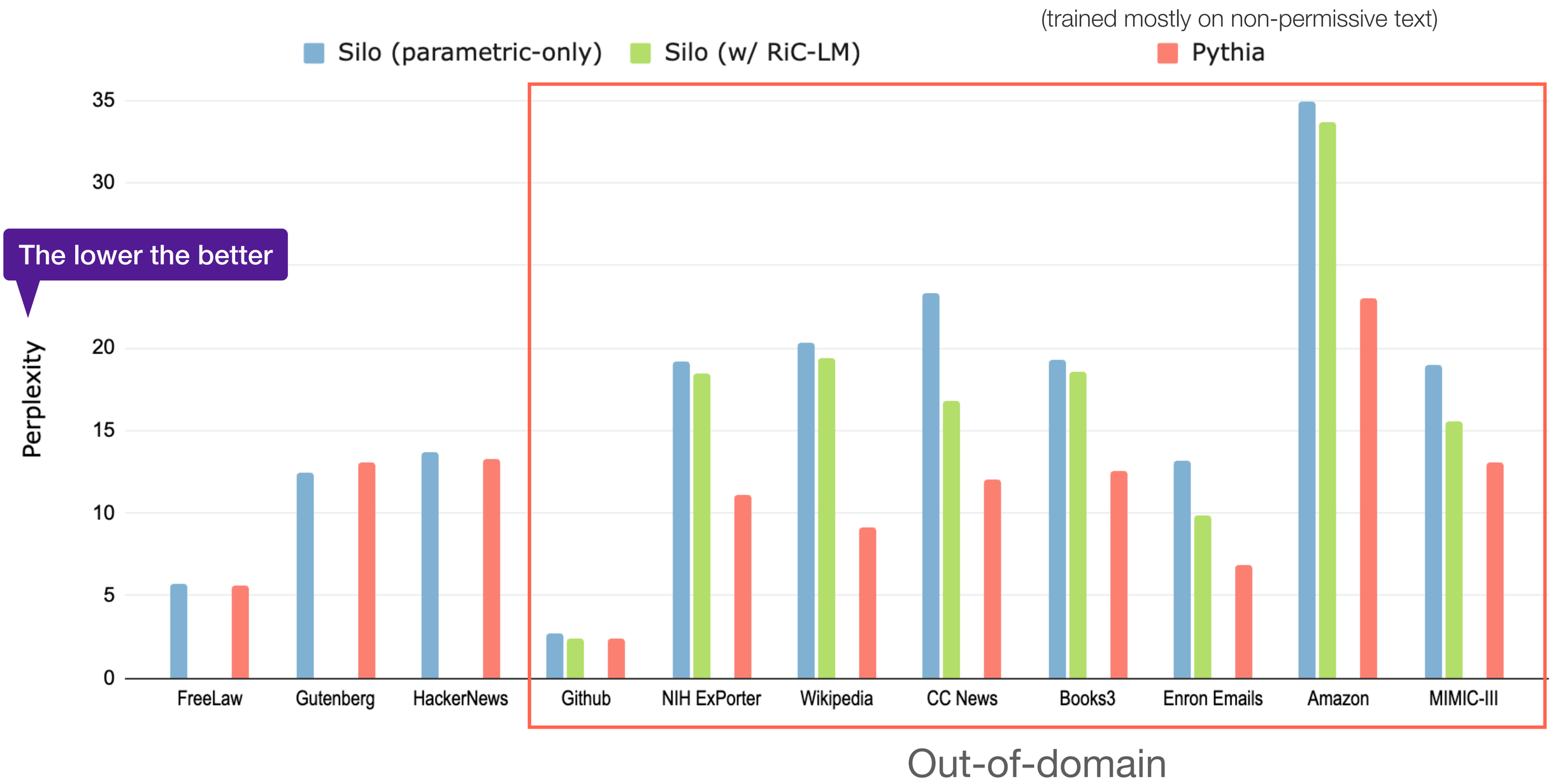

(trained mostly on non-permissive text)
Silo (parametric-only)
Pythia
The lower the better
Perplexity
35
30
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III

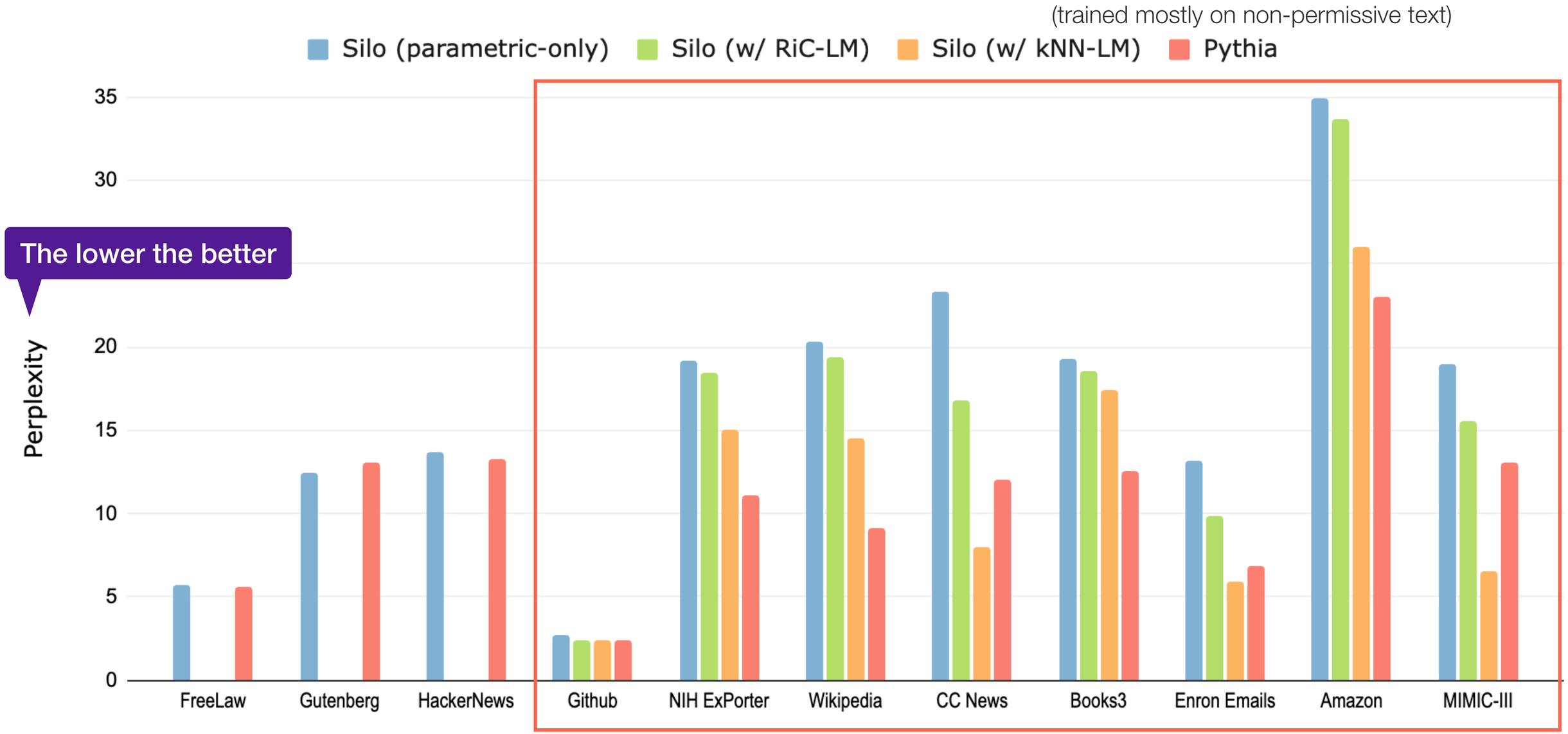


In-domain

Competitive

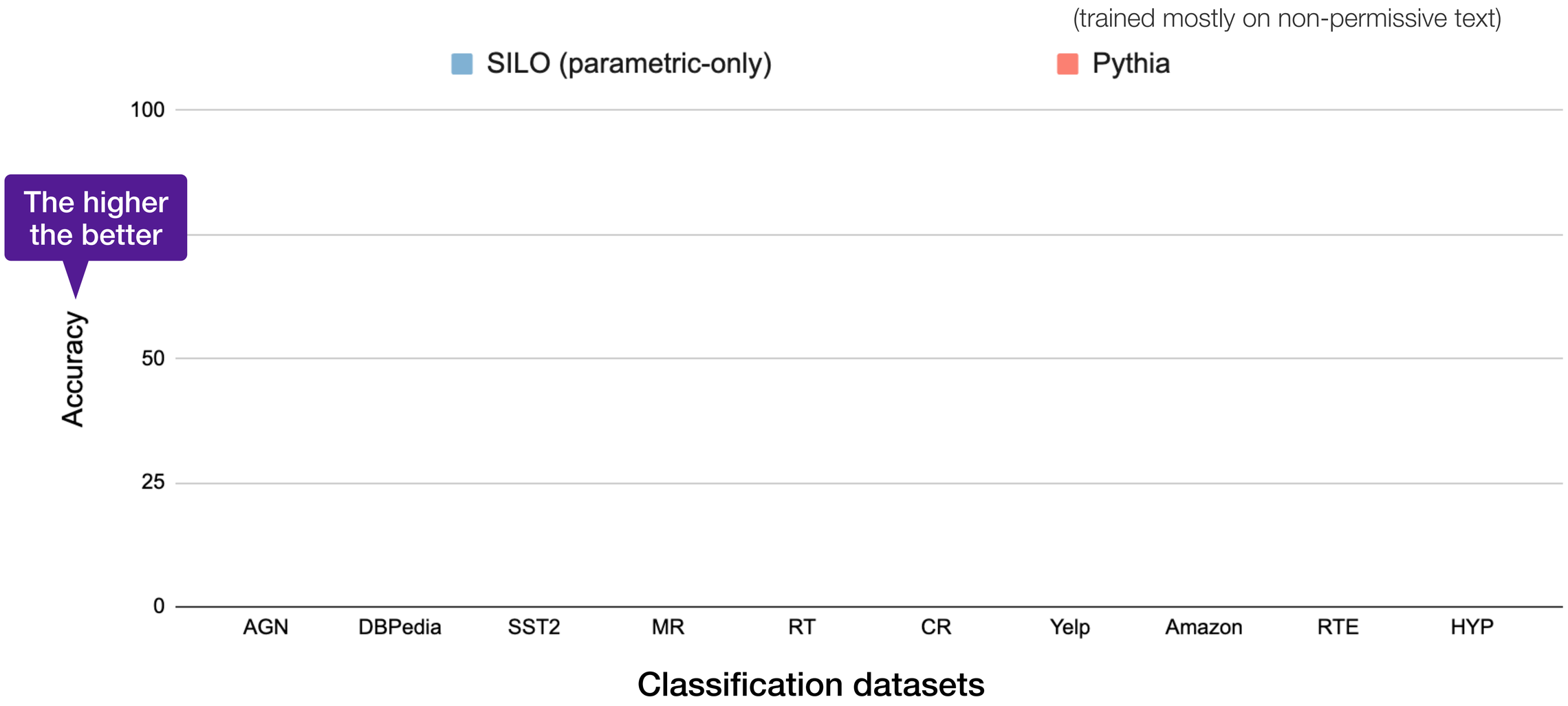

(trained mostly on non-permissive text)
Silo (parametric-only)
Pythia
The lower the better
Perplexity
35
30
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-domain
Competitive
Out-of-domain
Falls short

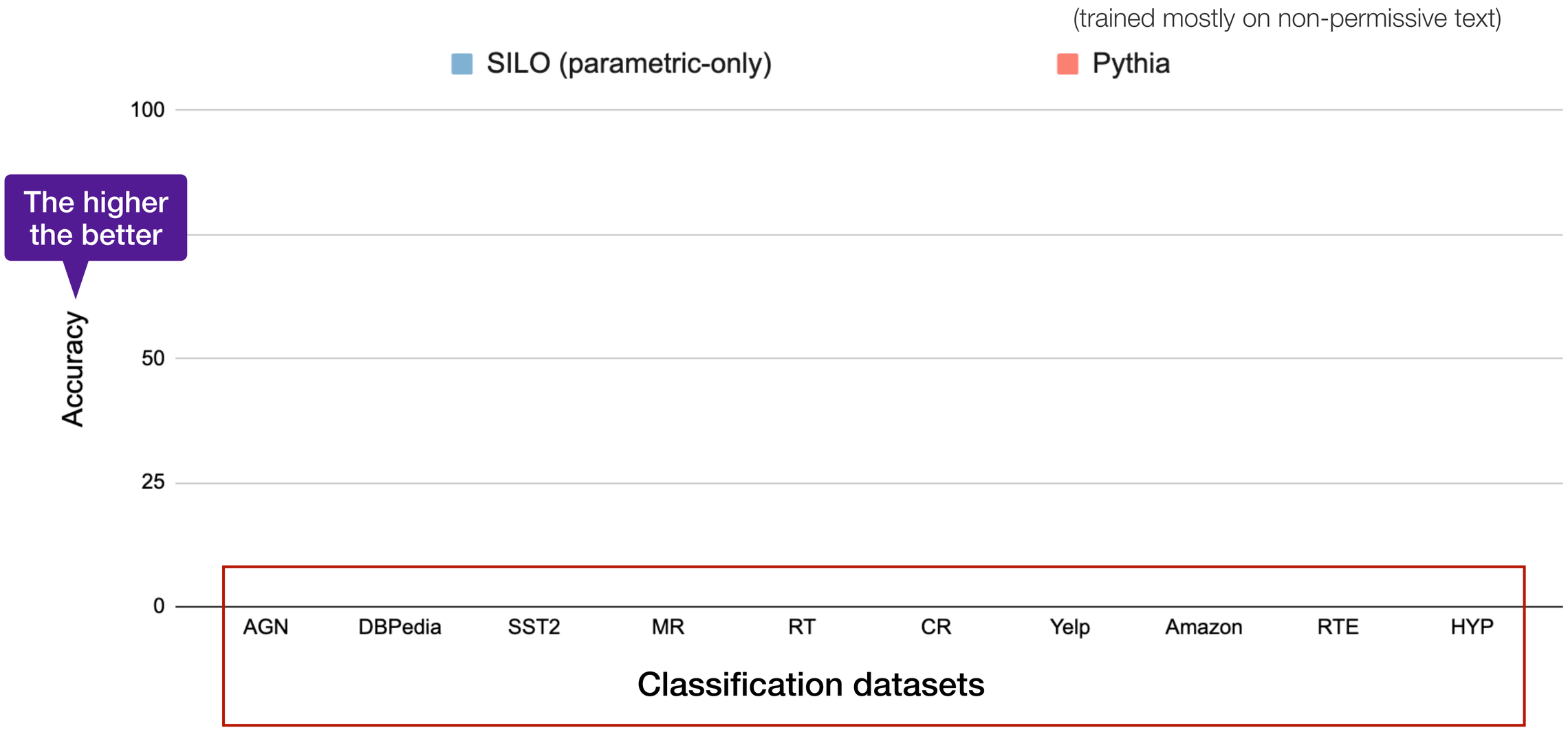

(trained mostly on non-permissive text)
Silo (parametric-only)
Silo (w/ RiC-LM)
Pythia
The lower the better
Perplexity
35
30
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
Out-of-domain

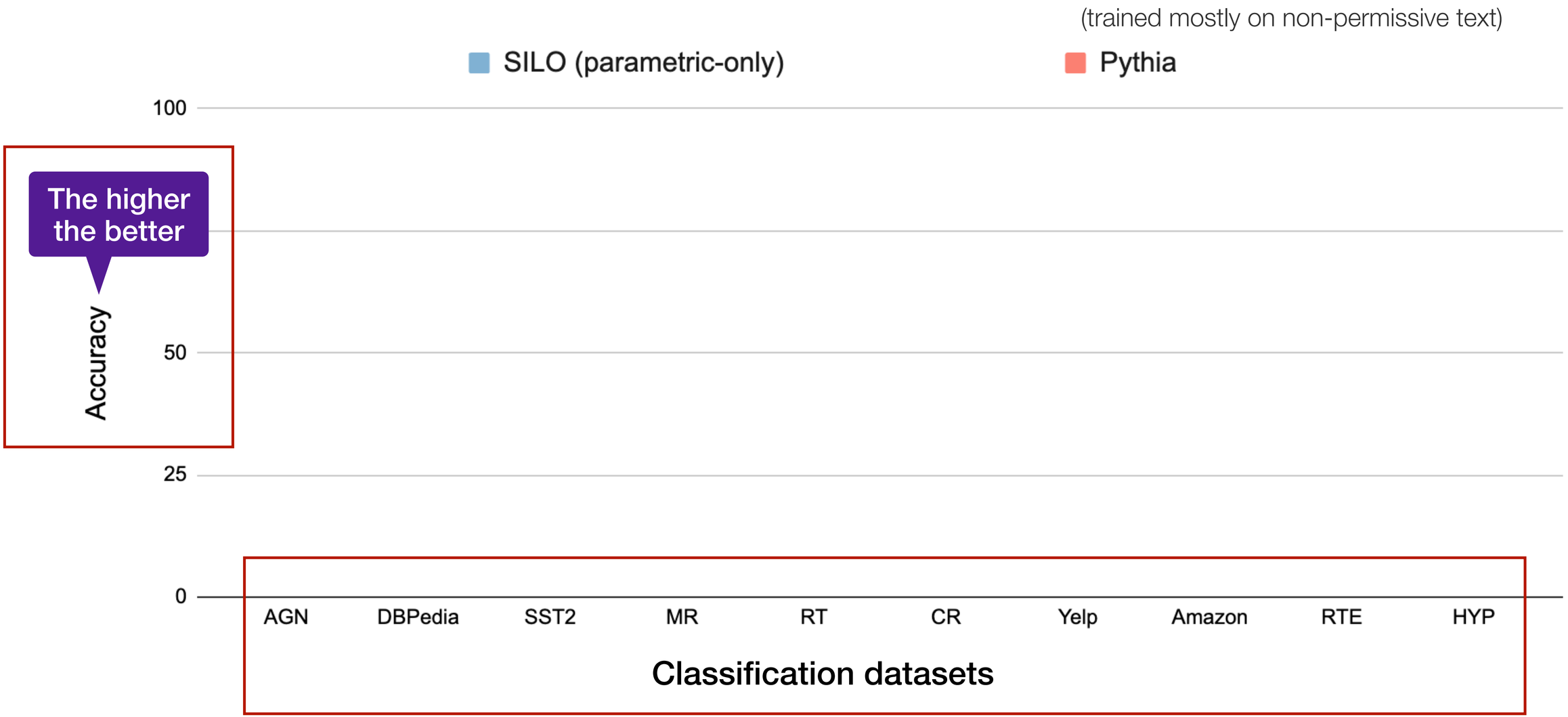


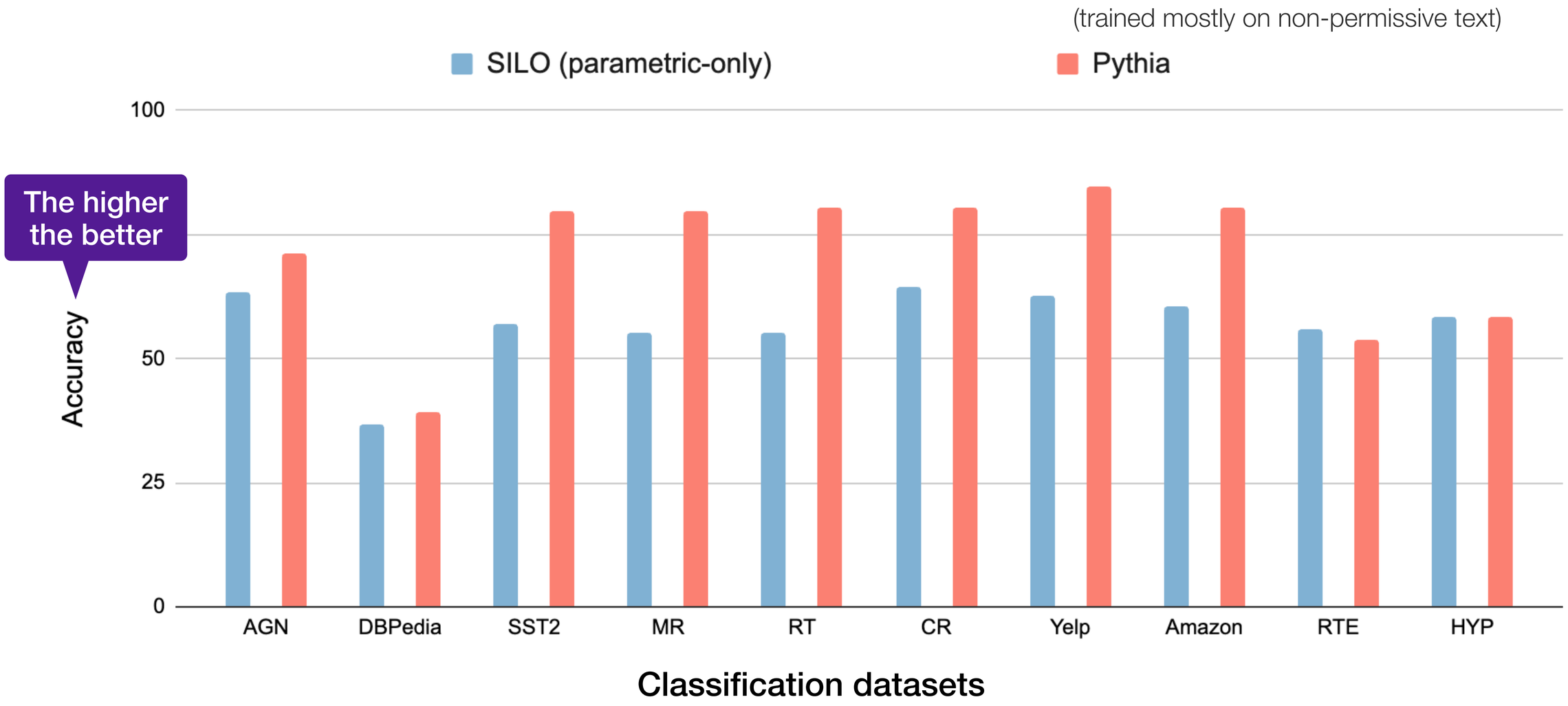
Out-of-domain

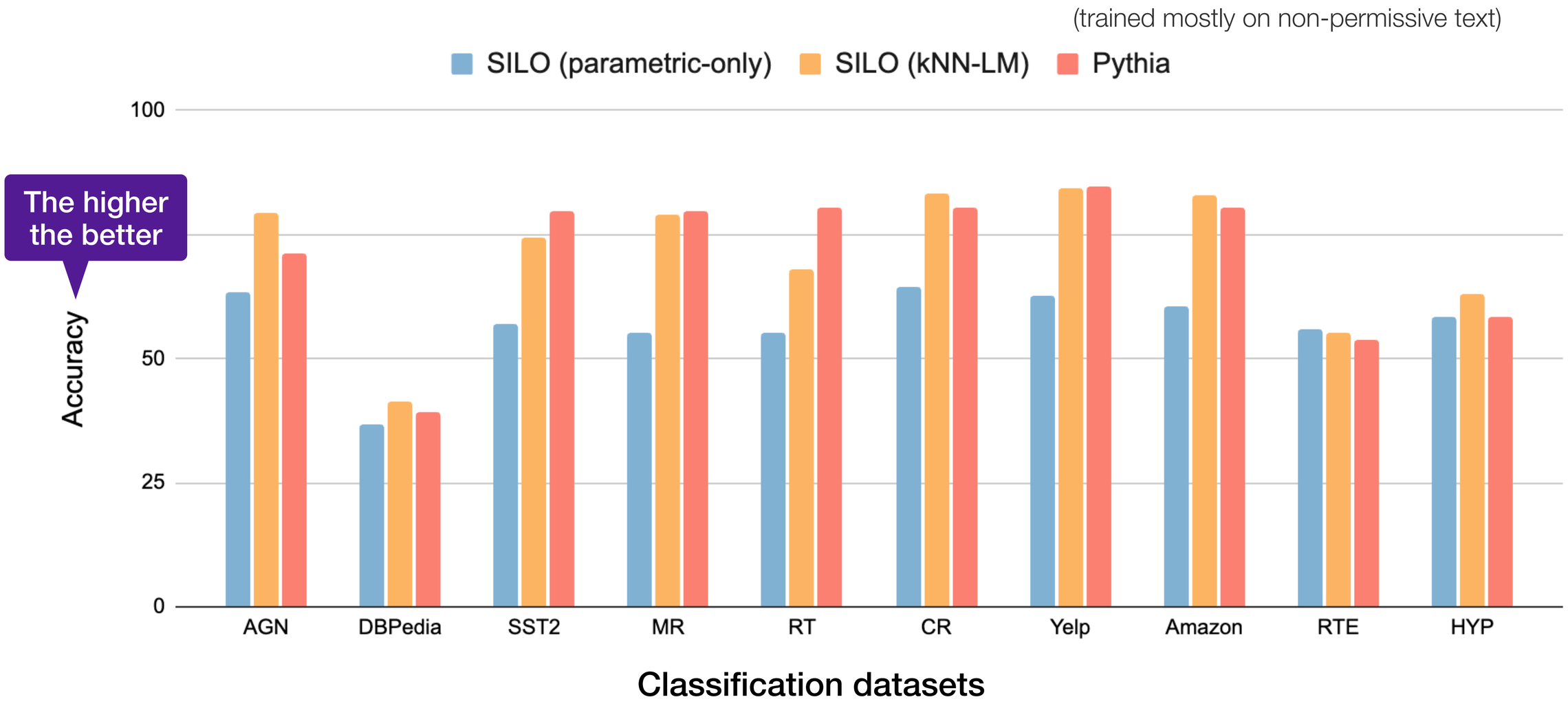
***k*NN-LM reduces the gap by 90% on average**


(trained mostly on non-permissive text)
SILO (parametric-only)
Pythia
The higher the better
Accuracy
100
50
25
0
AGN
DBPedia
SST2
MR
RT
CR
Yelp
Amazon
RTE
HYP
Classification datasets


(trained mostly on non-permissive text)
SILO (parametric-only)
Pythia
100
50
25
0
The higher the better
Accuracy
AGN
DBPedia
SST2
MR
RT
CR
Yelp
Amazon
RTE
HYP
Classification datasets


(trained mostly on non-permissive text)
SILO (parametric-only)
Pythia
100
50
25
0
The higher the better
Accuracy
AGN
DBPedia
SST2
MR
RT
CR
Yelp
Amazon
RTE
HYP
Classification datasets

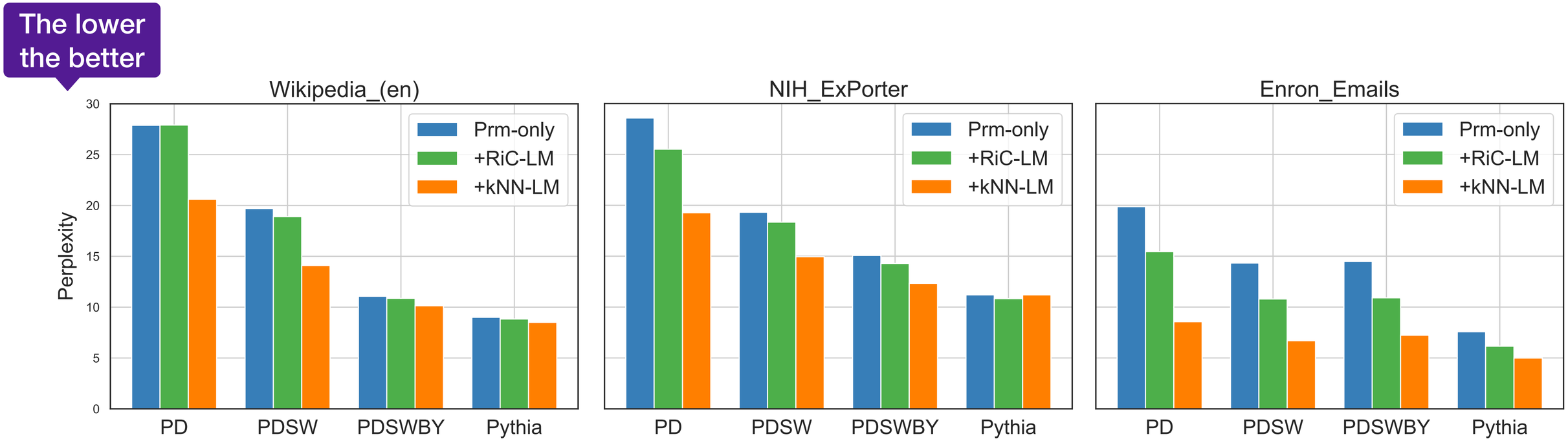

(trained mostly on non-permissive text)
SILO (parametric-only)
Pythia
The higher the better
Accuracy
100
50
25
0
AGN
DBPedia
SST2
MR
RT
CR
Yelp
Amazon
RTE
HYP
Classification datasets



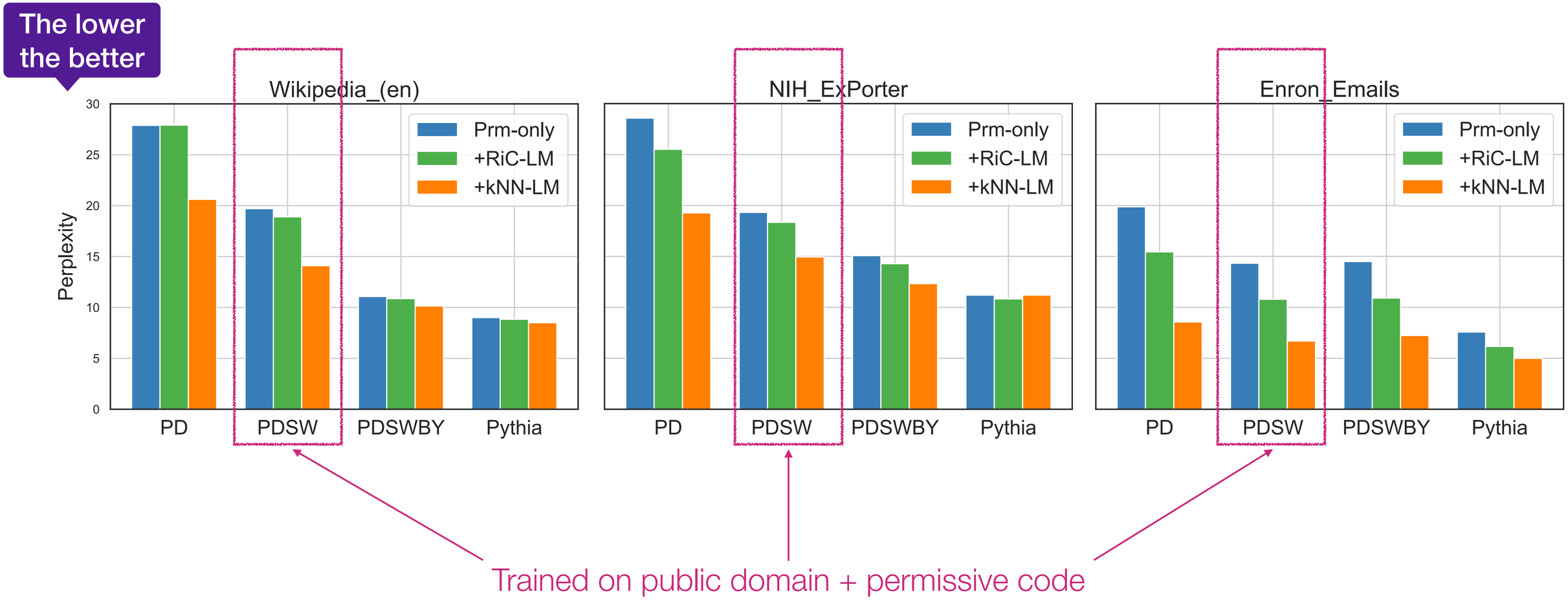
SILO matches or outperforms Pythia on downstream tasks

LMs Trained on Varying Levels of OLC

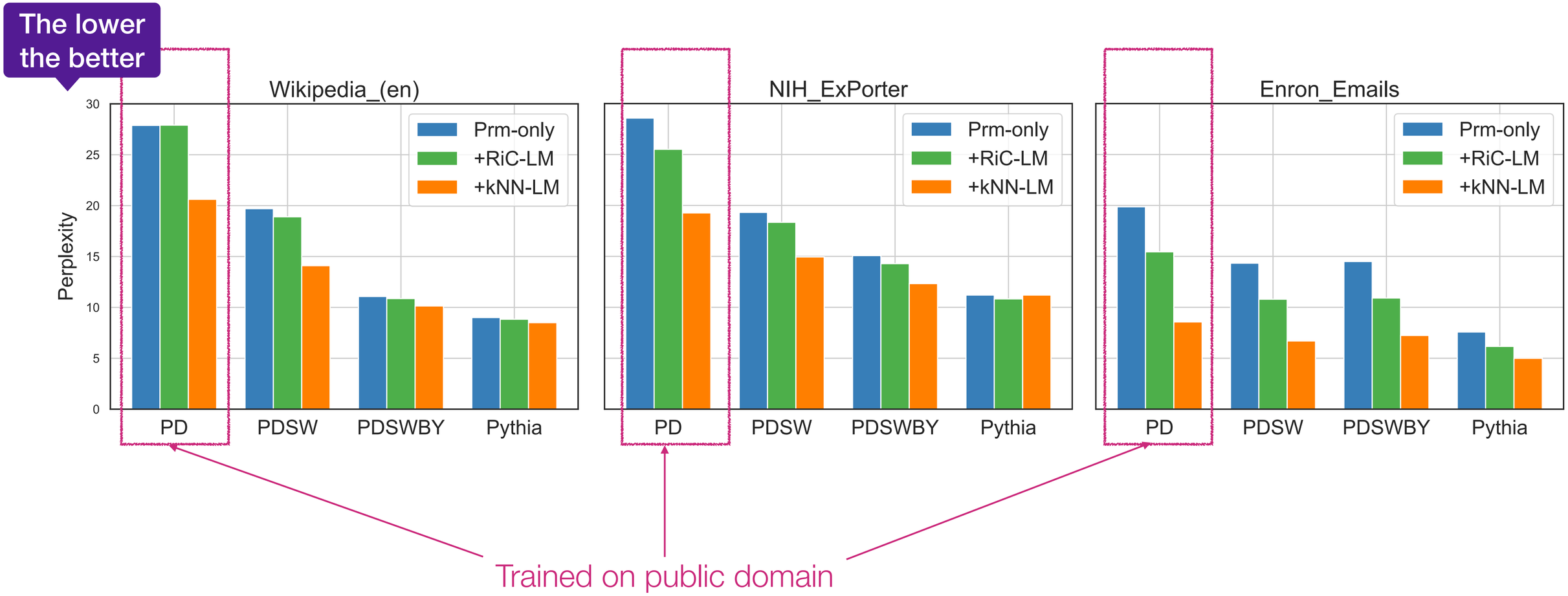
LMs Trained on Varying Levels of OLC



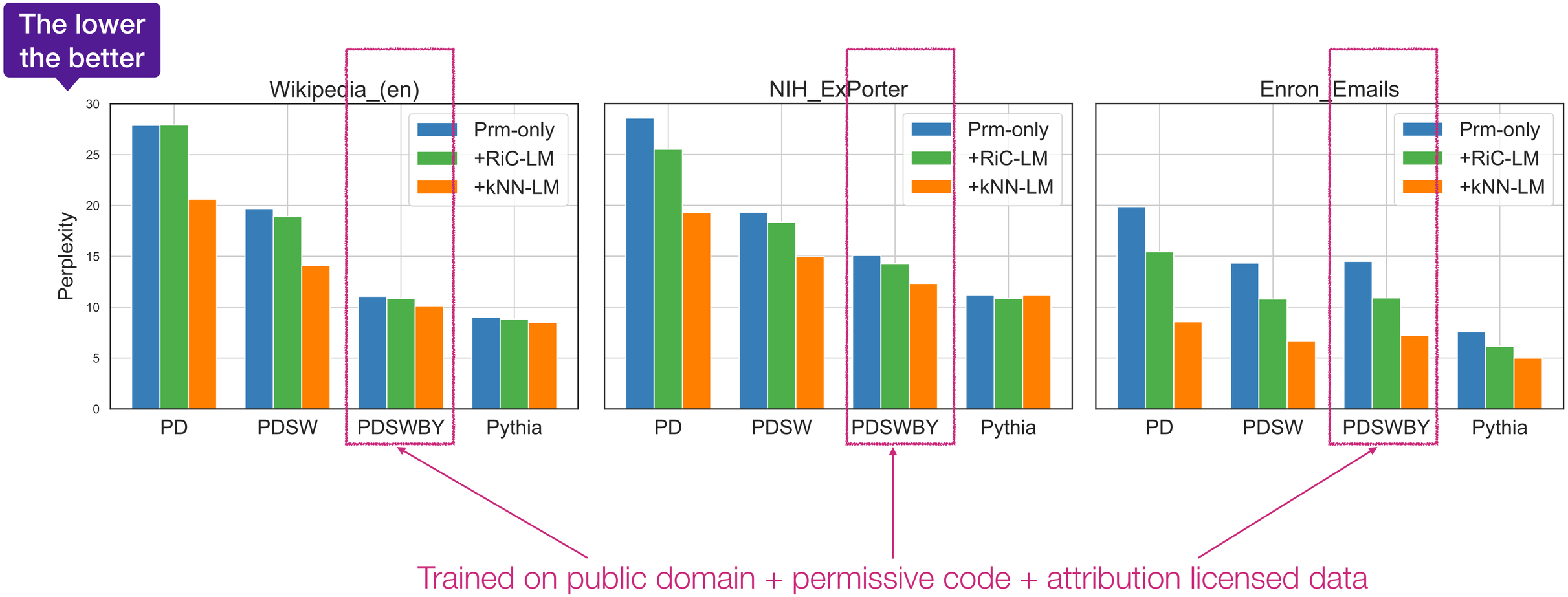
LMs Trained on Varying Levels of OLC



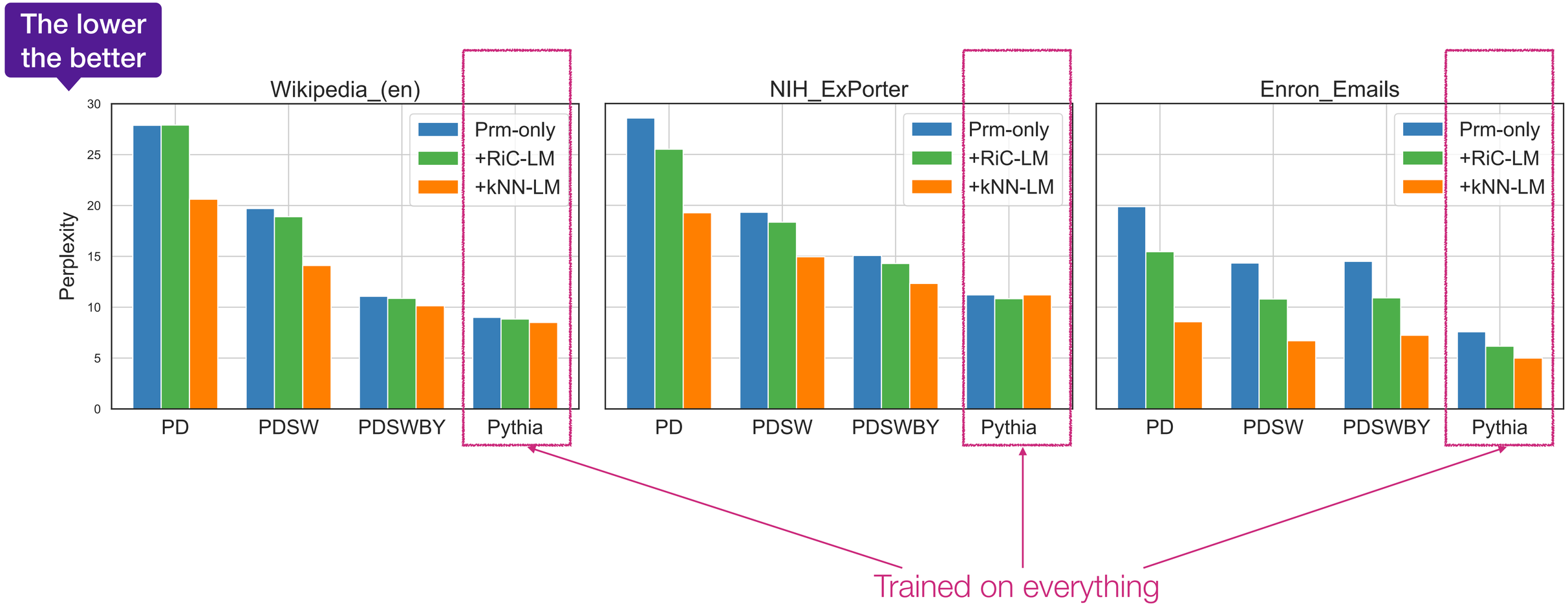
LMs Trained on Varying Levels of OLC



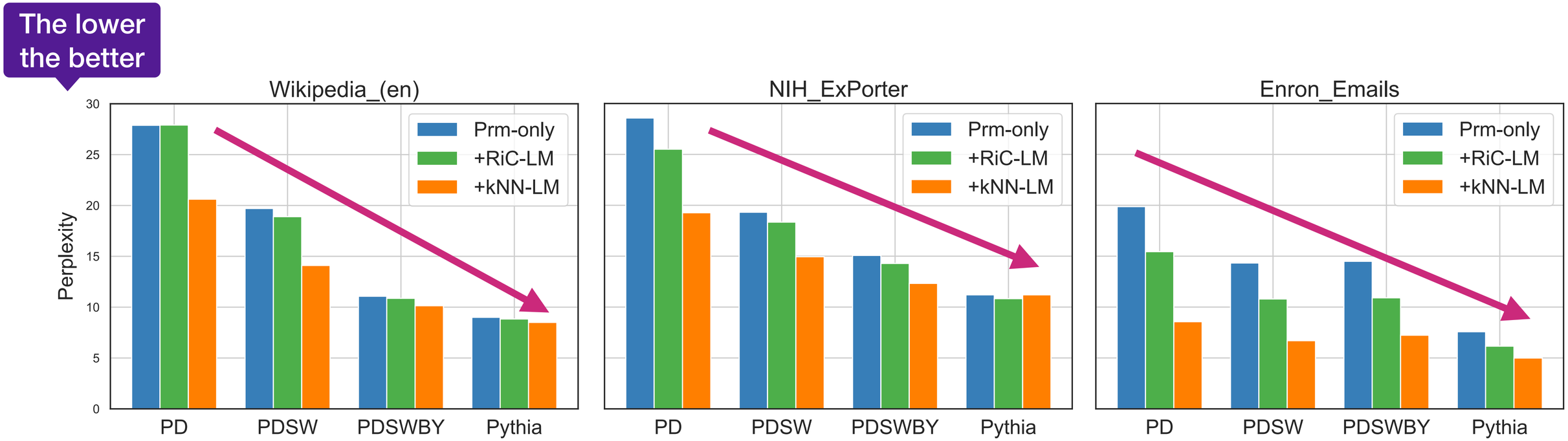
LMs Trained on Varying Levels of OLC

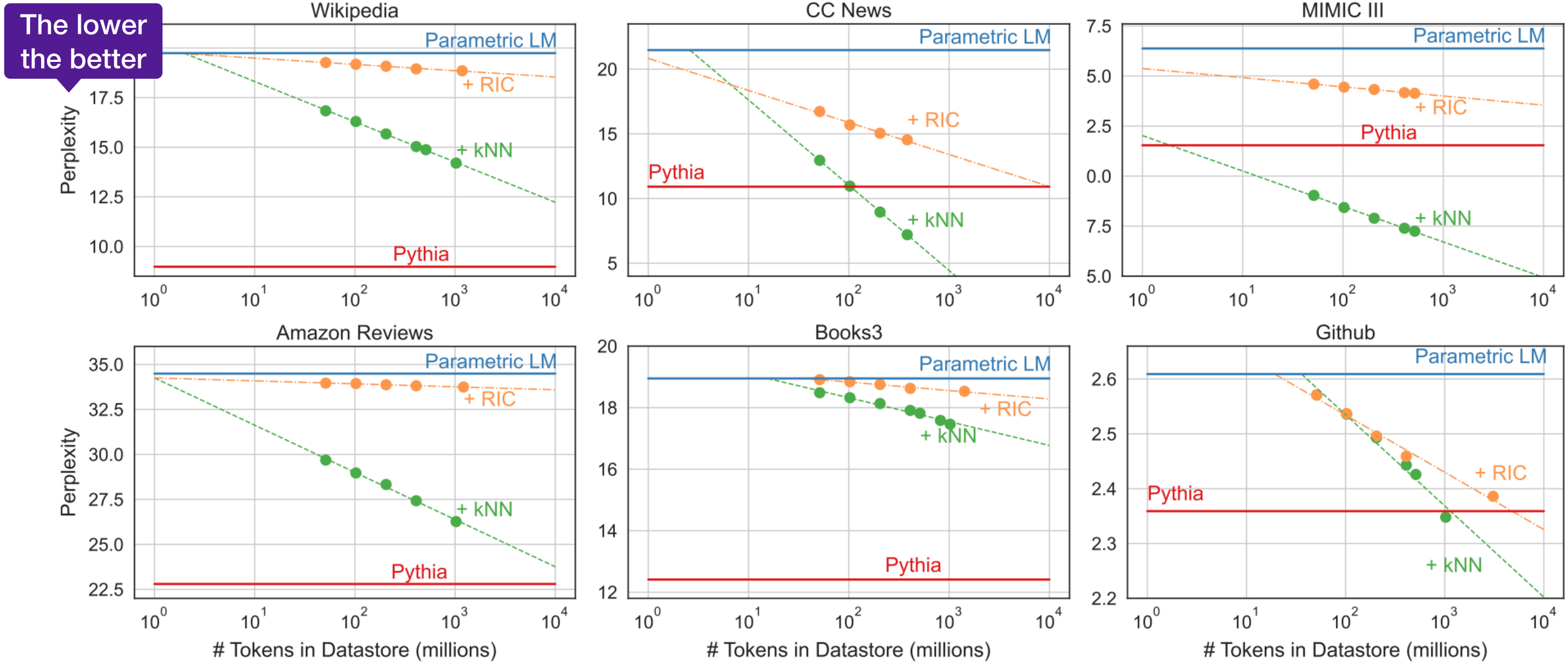


LMs Trained on Varying Levels of OLC

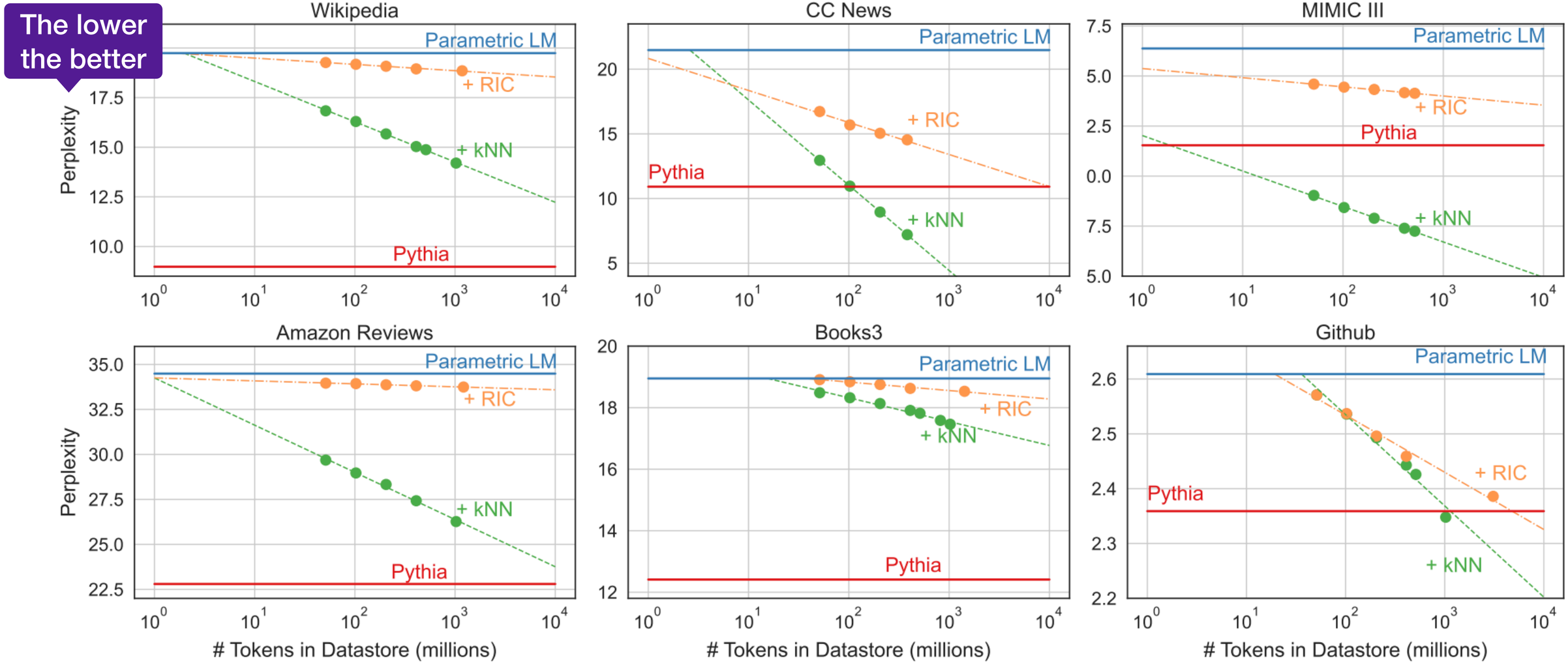


LMs Trained on Varying Levels of OLC



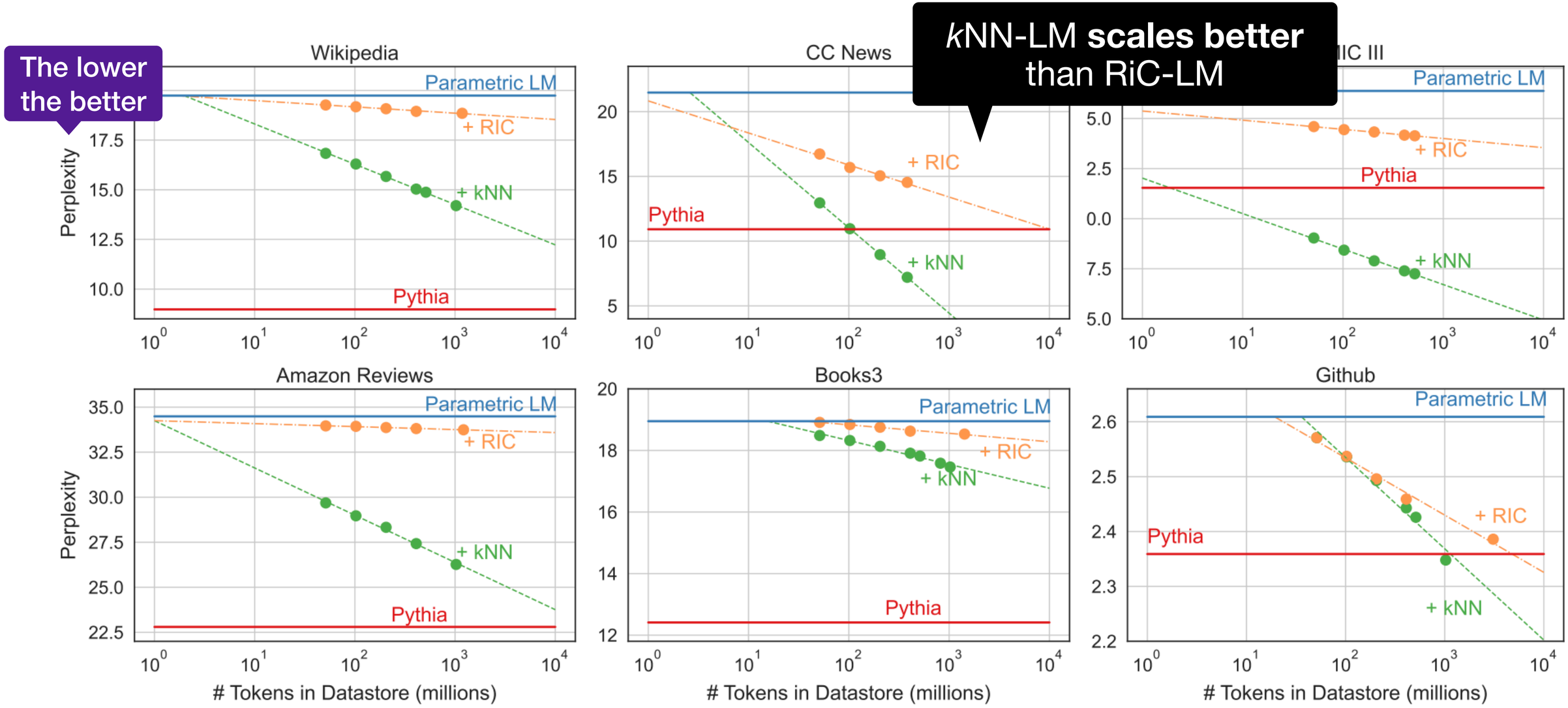


Scaling the test-time datastore improves performance



Scaling the test-time datastore improves performance

(Additional dimension for scaling at inference!)



Scaling the test-time datastore improves performance

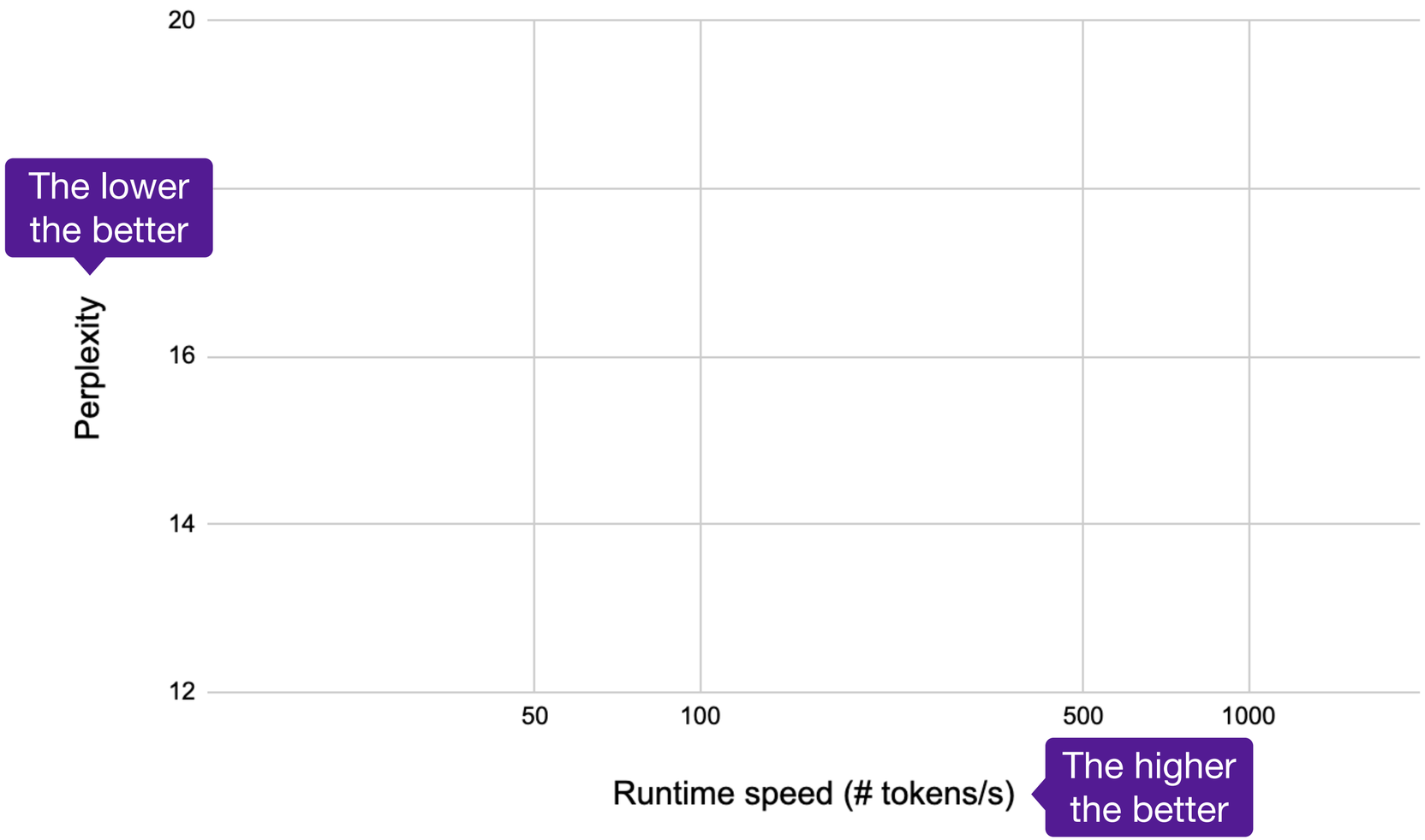
(Additional dimension for scaling at inference!)

Runtime Speed

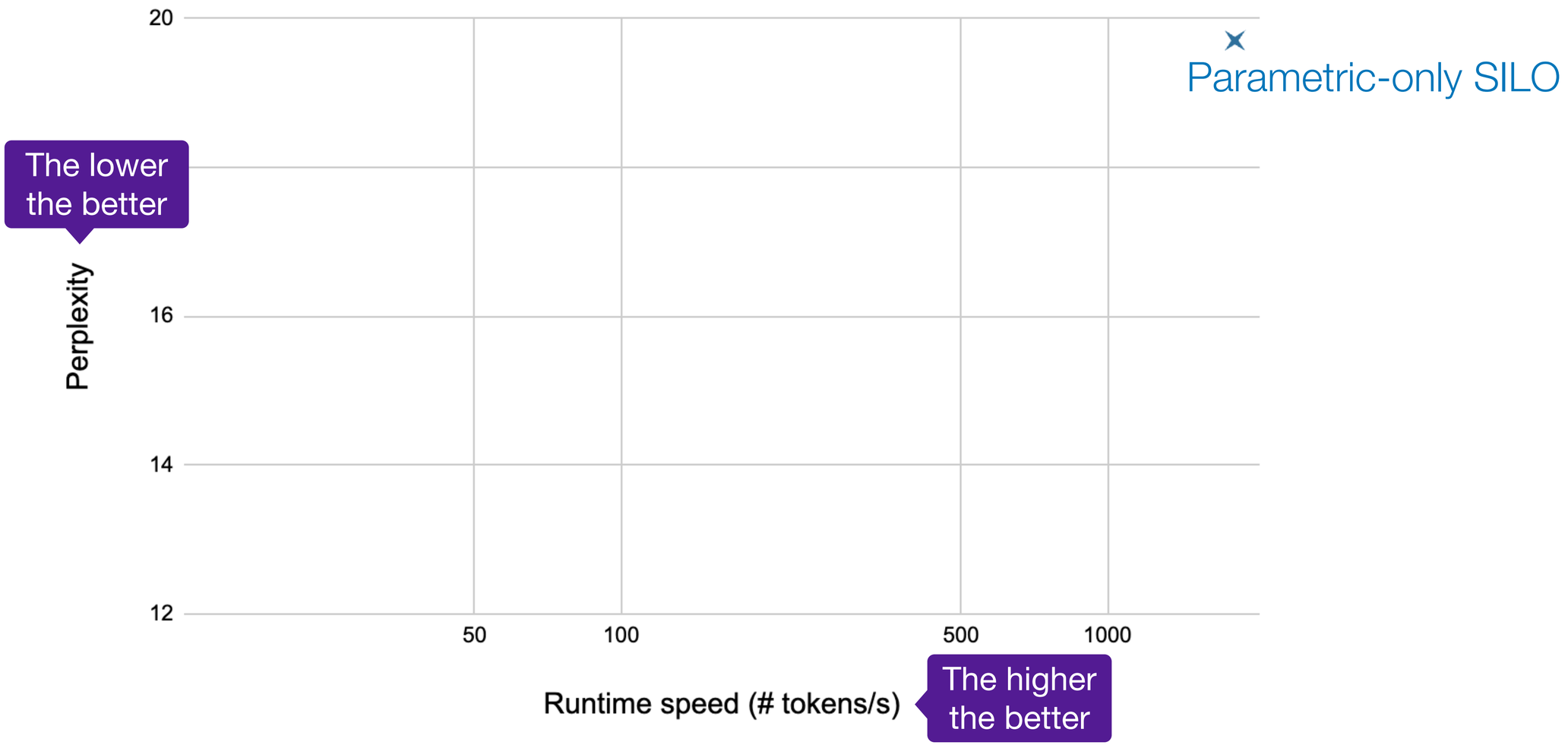
SILO is slower than the same-size parametric-only LM

How much slower?

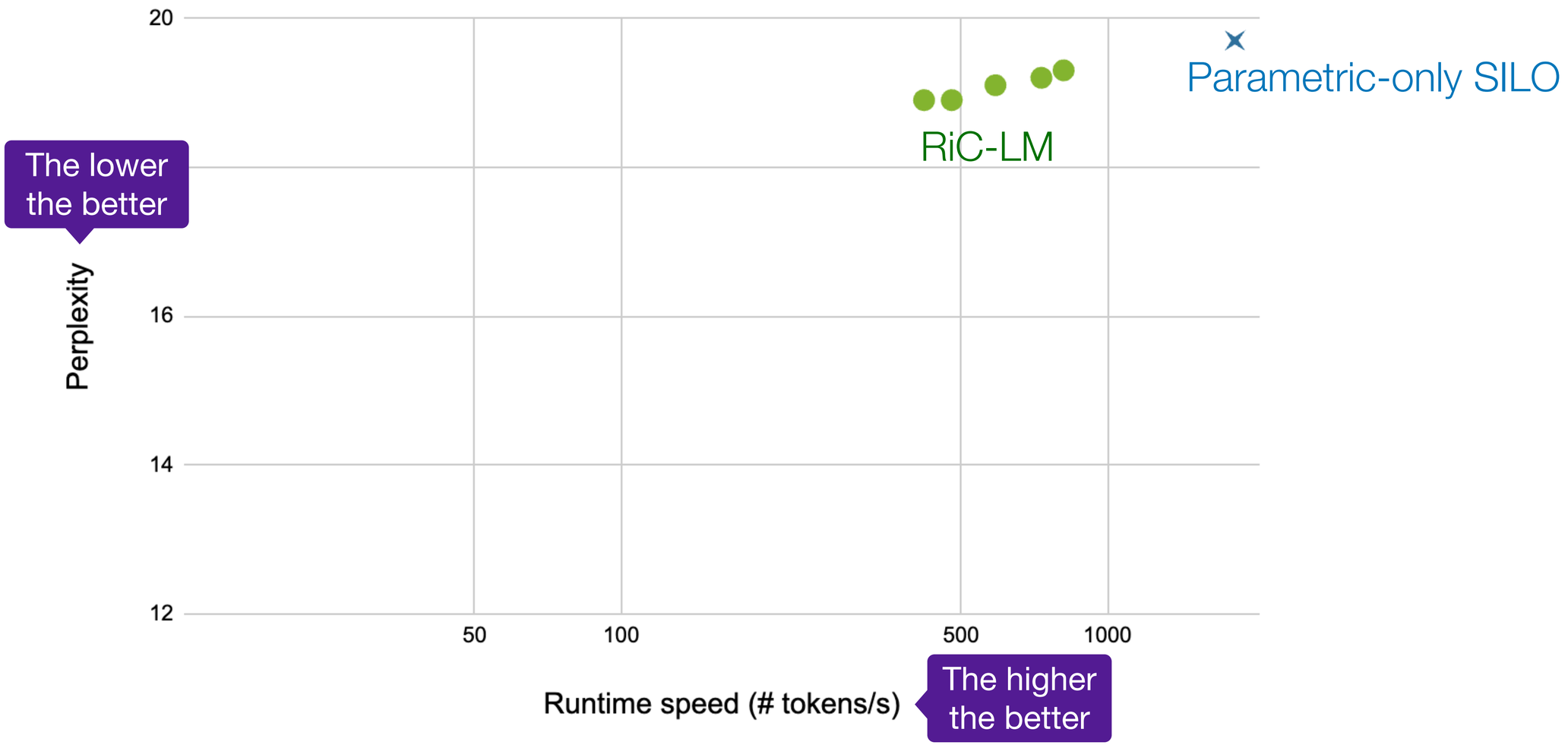
Runtime Speed



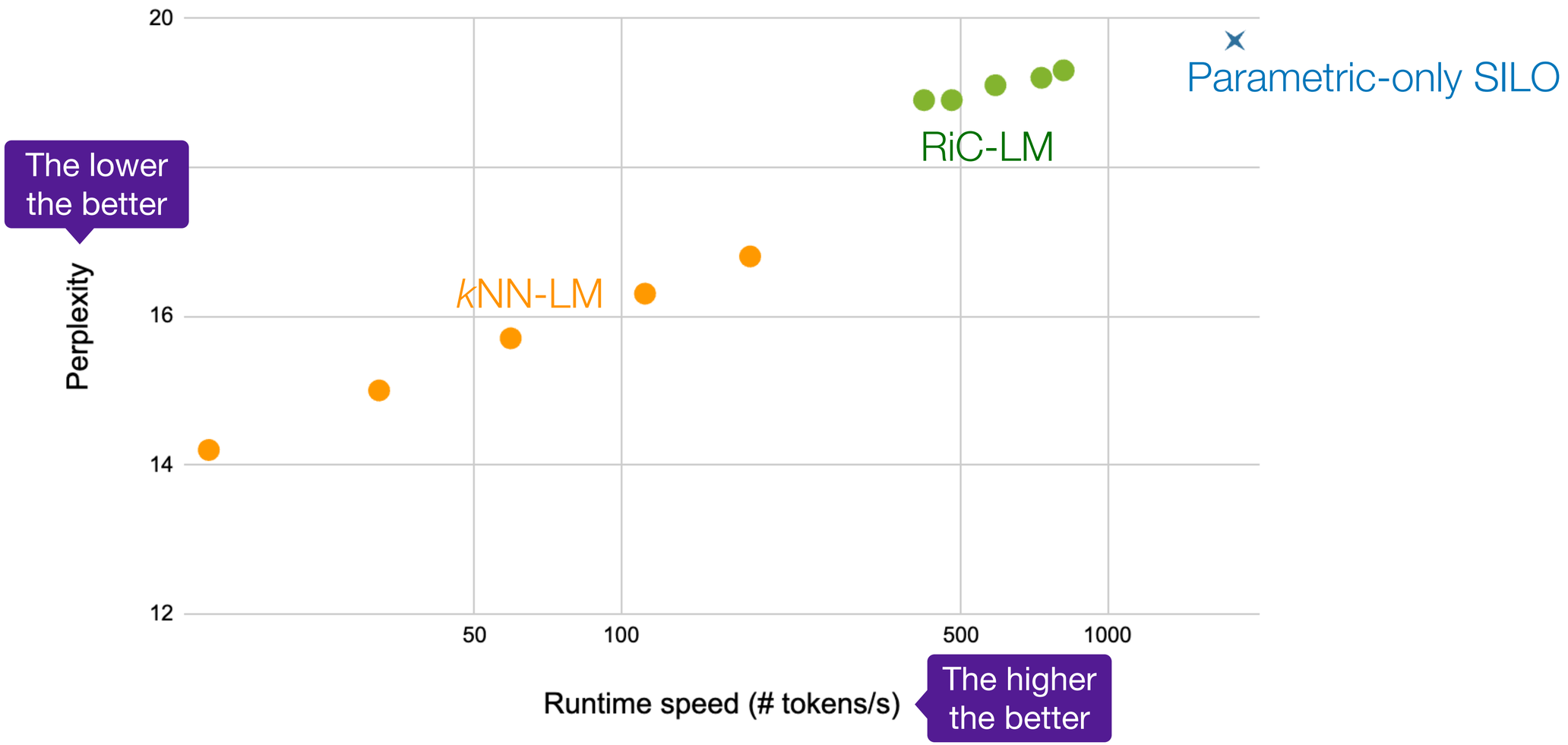
Runtime Speed



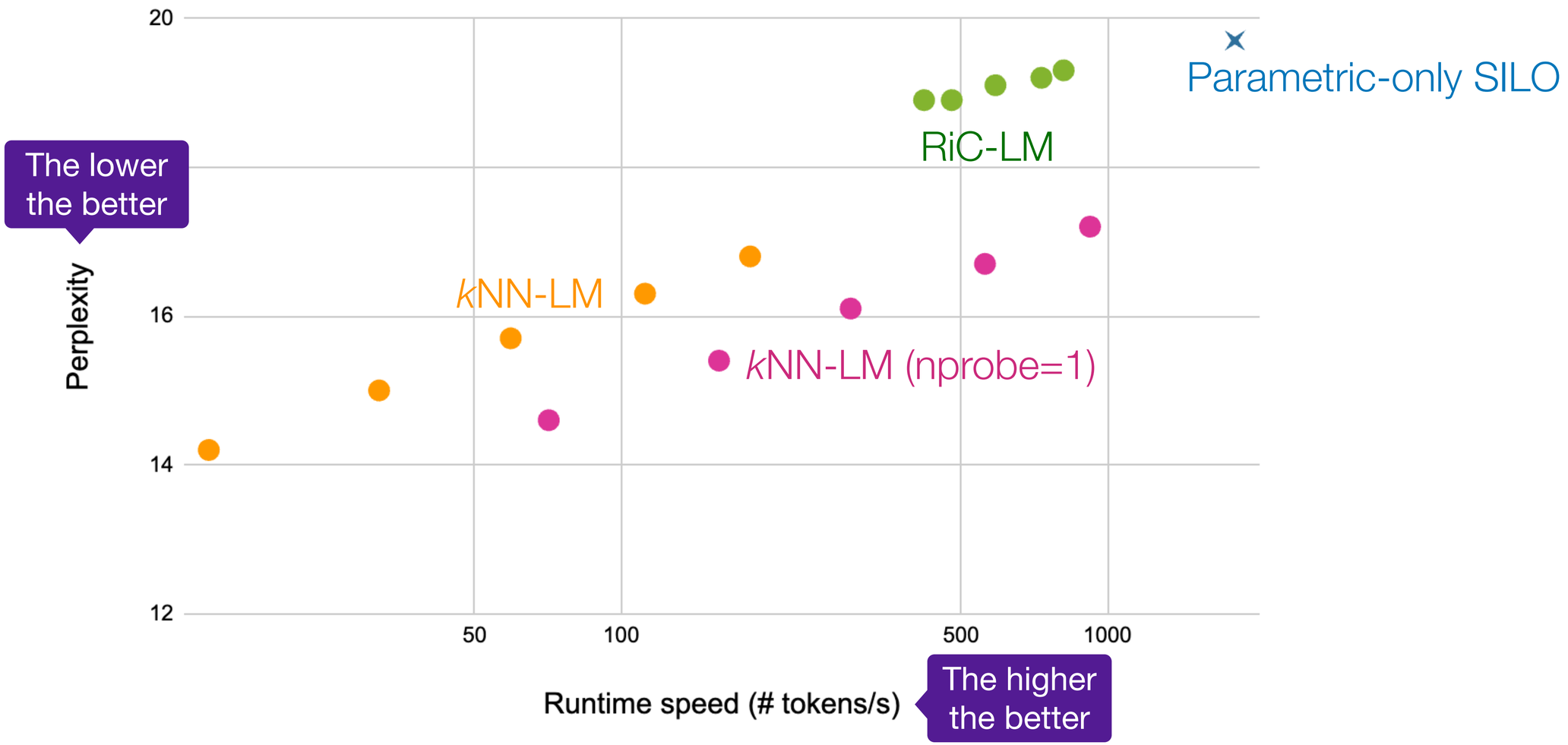
Runtime Speed



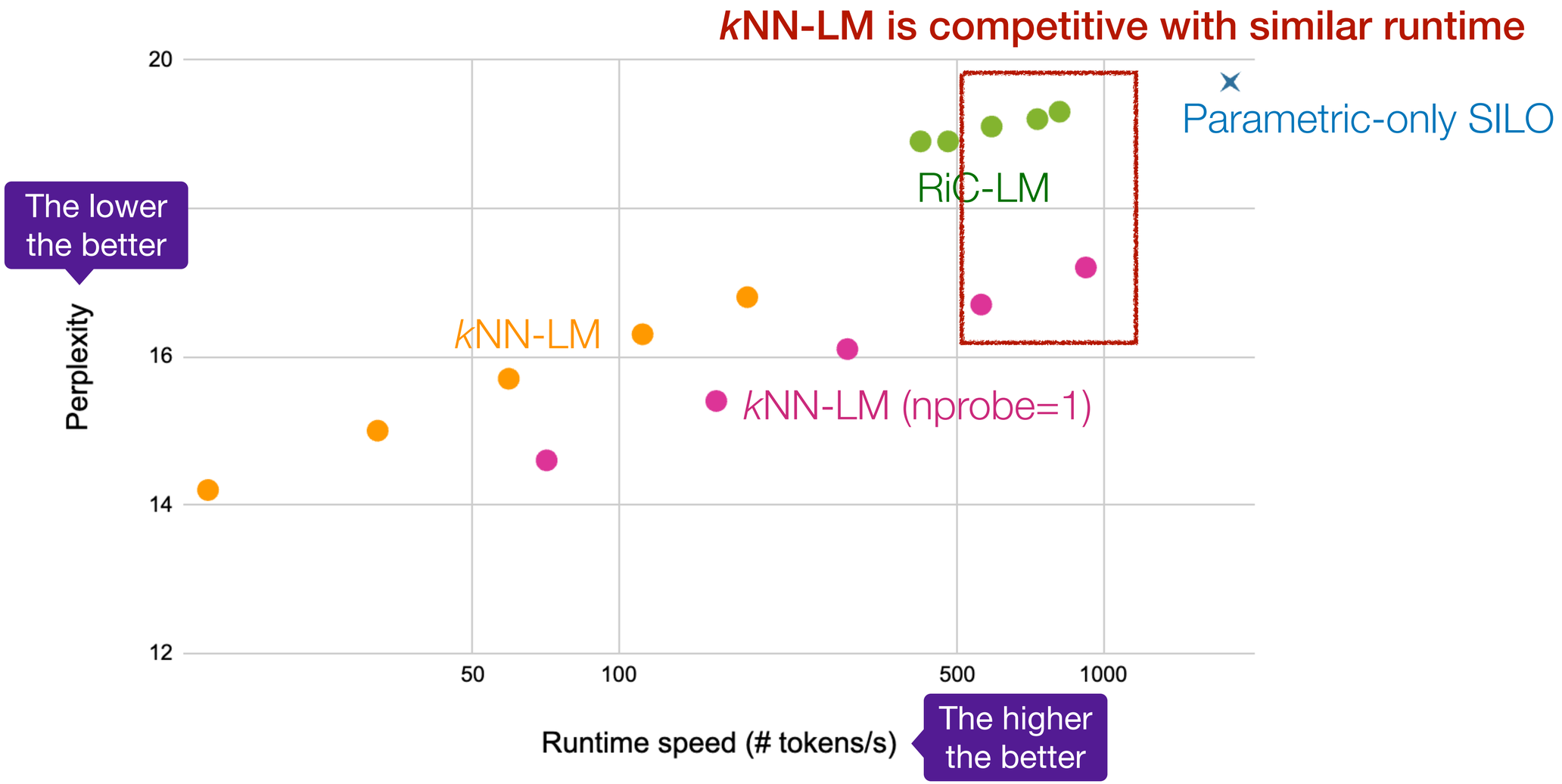
Runtime Speed



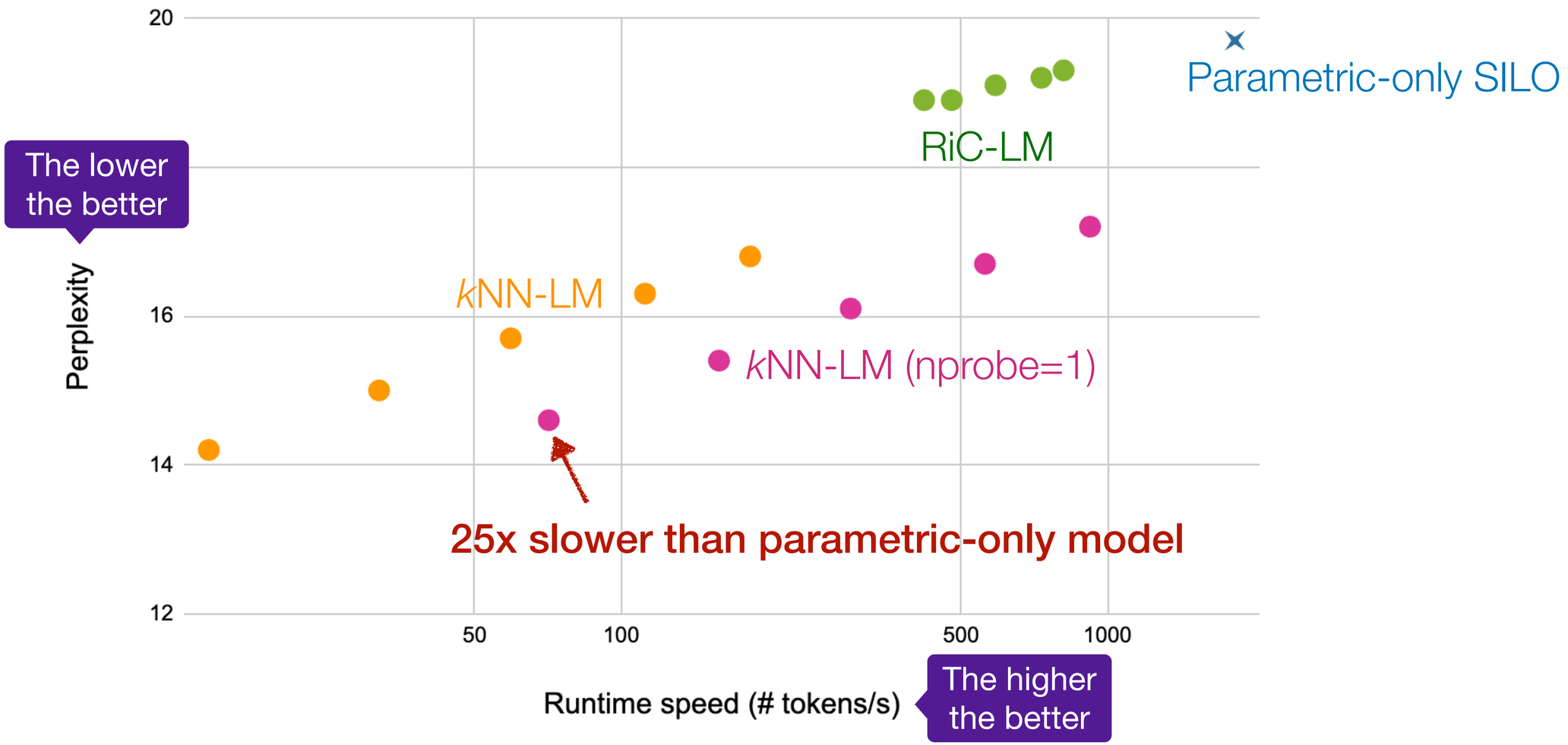
Runtime Speed



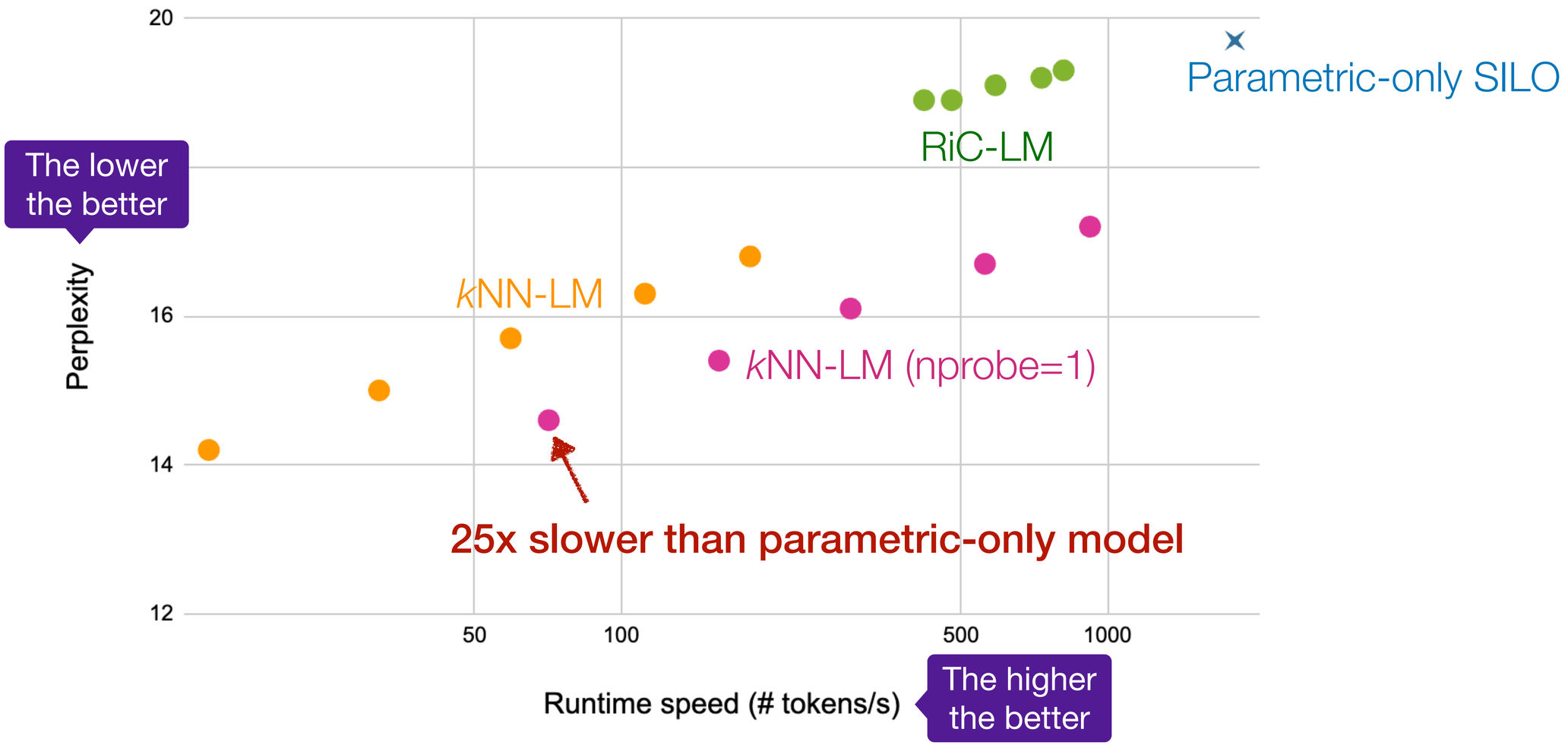
Runtime Speed



Runtime Speed



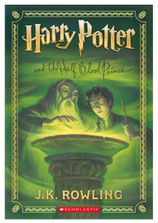
Runtime Speed



Performance-speed tradeoff can be controlled at inference

Attribution Example

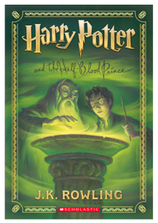
Attribution Example



Test input: Terror tore at Harry's heart … he had to get to Dumbledore and he had to catch Snape… (…) Dumbledore could not have died… (…) Harry felt Greenback collapse against him; with a stupendous effort he pushed the werewolf off and onto the floor as a jet of

Continuation: **[green]** light came flying toward him …

Attribution Example



Test input: Terror tore at Harry's heart … he had to get to Dumbledore and he had to catch Snape… (…) Dumbledore could not have died… (…) Harry felt Greenback collapse against him; with a stupendous effort he pushed the werewolf off and onto the floor as a jet of

Continuation: **[green]** light came flying toward him …

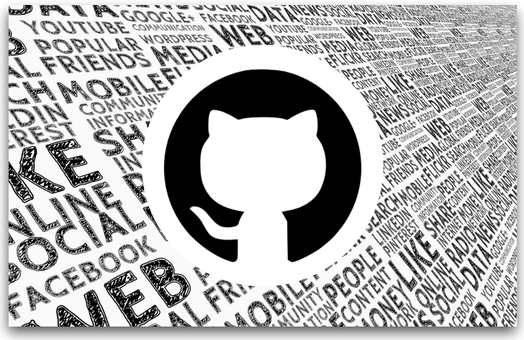
Top-1 retrieved context: Voldemort was ready. As Harry shouted "Expelliarmus!" Voldemort cried "Abada Kedavra!" A jet of **[green]** light issued from Voldemort's want …

Attribution Example

Test input:

```
include '../lib/admin.defines.php';
include '../lib/admin.module.access.php';
include '../lib/admin.smarty.php';
if (! has_right (
```

Continuation: **[AC]**X_BILLING)) { Header …

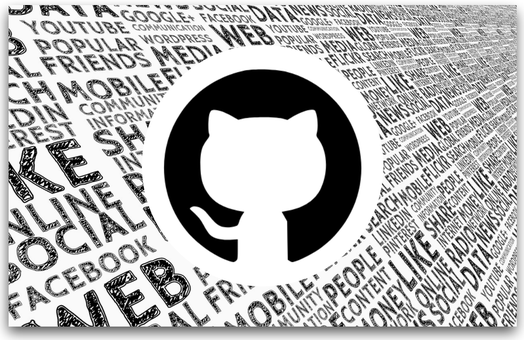


Attribution Example

Test input:

```
include '../lib/admin.defines.php';
include '../lib/admin.module.access.php';
include '../lib/admin.smarty.php';
if (! has_right (
```

Continuation: **[AC]**X_BILLING)) { Header …



Top-1 retrieved context:

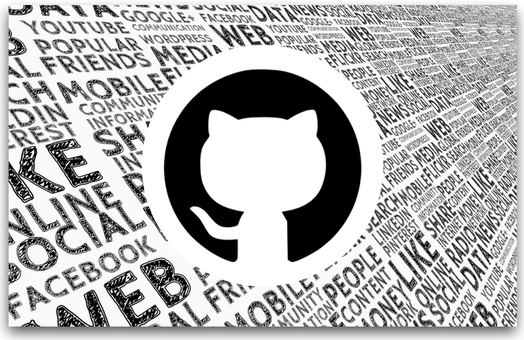
```
* You should have received a copy of the GNU Affero General Public License
* along with this program. If not, see <http://www.gnu.org/licenses/>.
*
*
**/
if (! has_right (
        [AC]X_ACCESS)) { Header …
```

Attribution Example

Test input:

```
include '../lib/admin.defines.php';
include '../lib/admin.module.access.php';
include '../lib/admin.smarty.php';
if (! has_right (
```

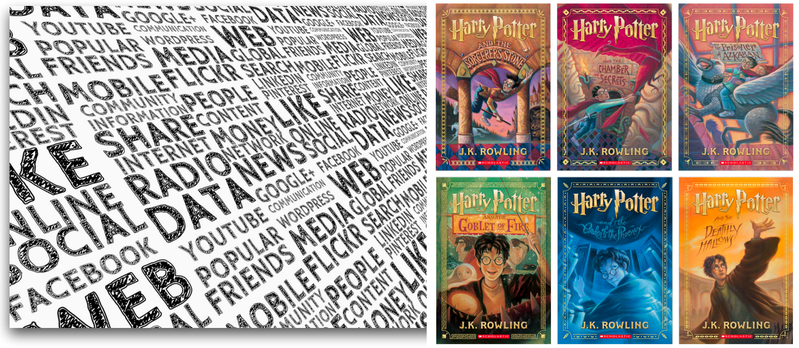
Continuation: **[AC]**X_BILLING)) { Header …



Top-1 retrieved context:

```
*You should have received a copy of the GNU Affero General Public License
* along with this program. If not, see <http://www.gnu.org/licenses/>.
*
*
**/
if (! has_right (
        [AC]X_ACCESS)) { Header …
```

Opt-out Example



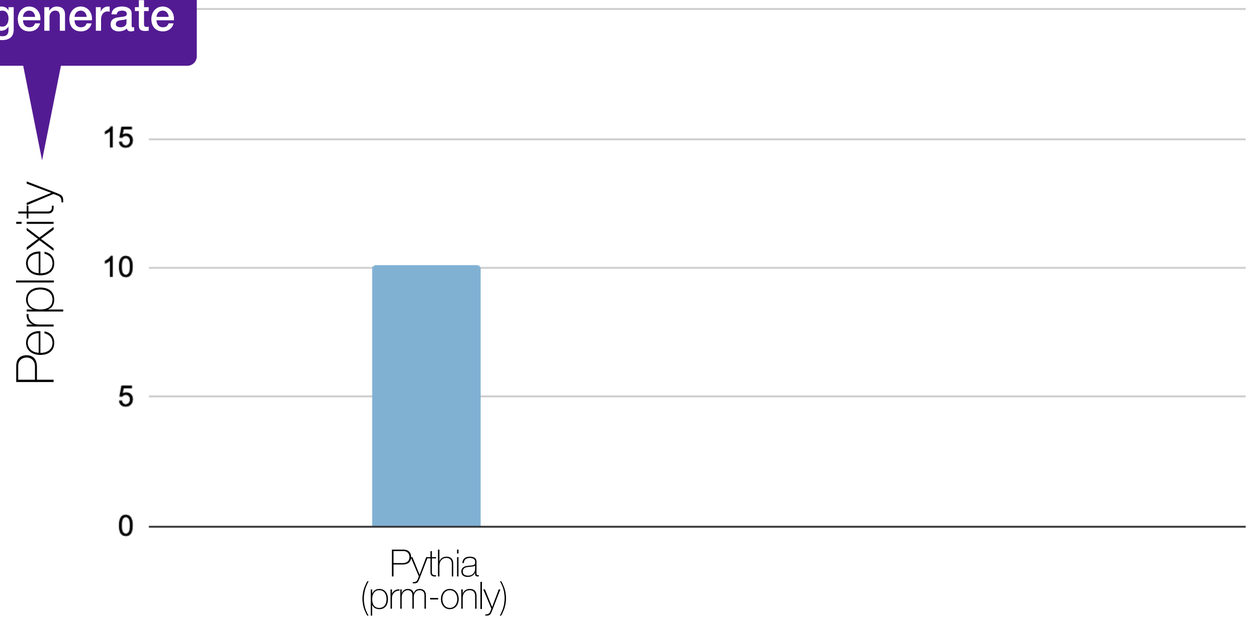
SILO

SILO

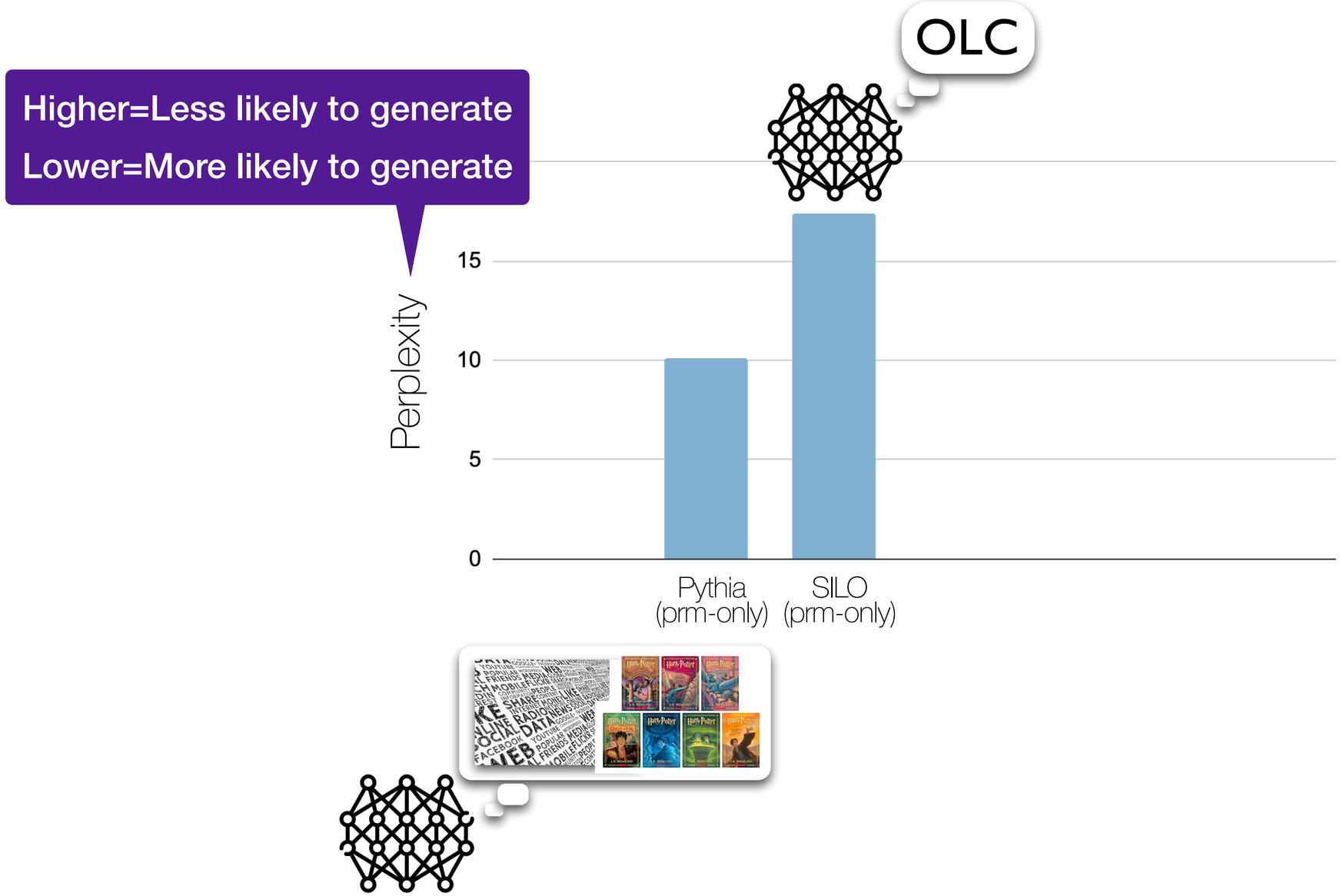
Opt-out Example

Higher=Less likely to generate

Lower=More likely to generate



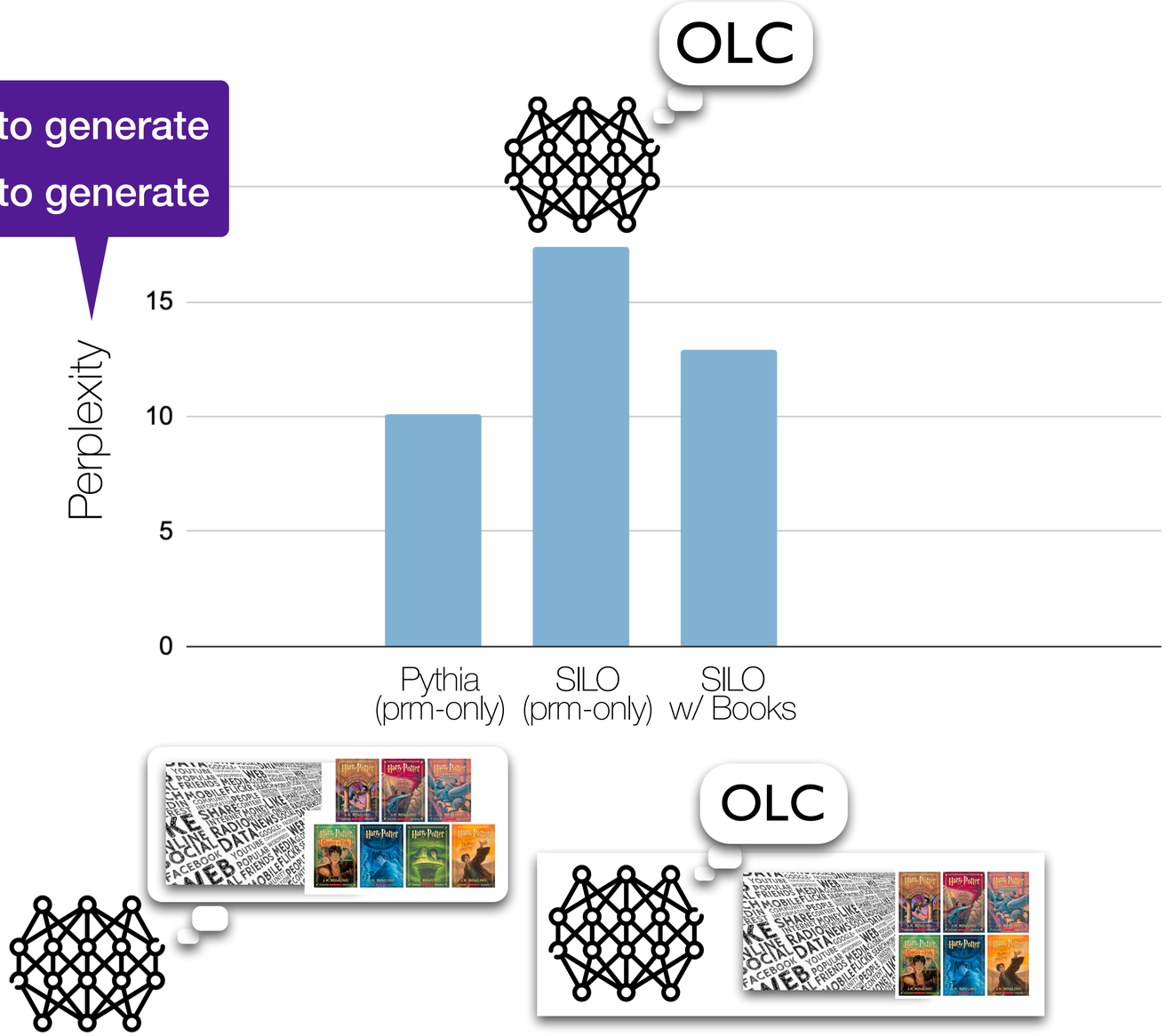
Opt-out Example



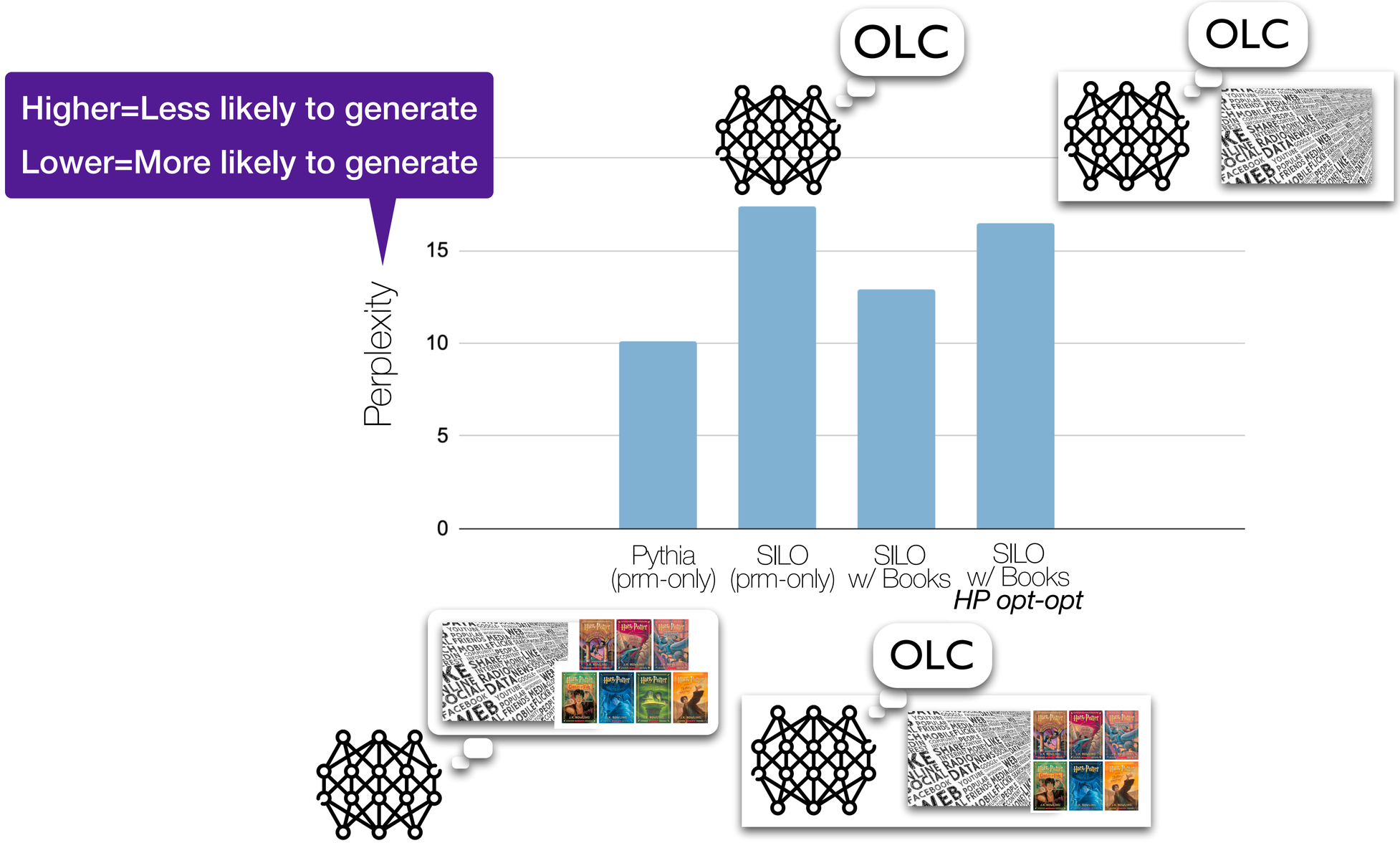
Opt-out Example

Higher=Less likely to generate

Lower=More likely to generate



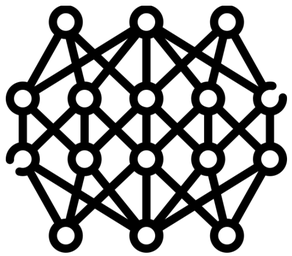
Opt-out Example



Why does *k*NN-LM generalize better than RiC-LM?

Why does *k*NN-LM generalize better than RiC-LM?

As Harry shouted "Expelliarmus!" Voldemort cried "Avada

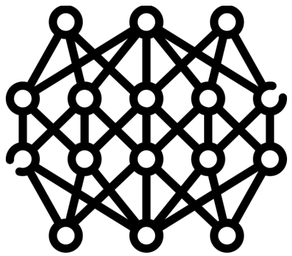


(Ground truth: Kedavra)

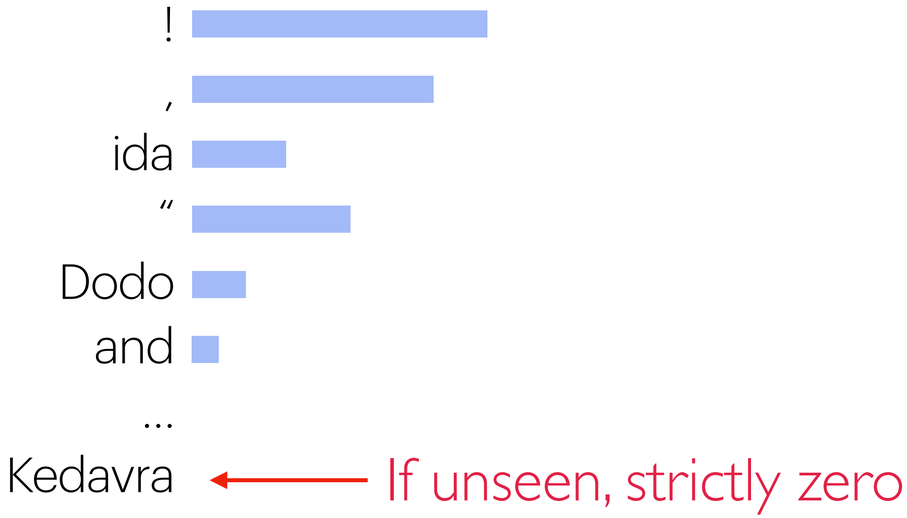
Parametric-only LM

Why does *k*NN-LM generalize better than RiC-LM?

As Harry shouted "Expelliarmus!" Voldemort cried "Avada



(Ground truth: Kedavra)

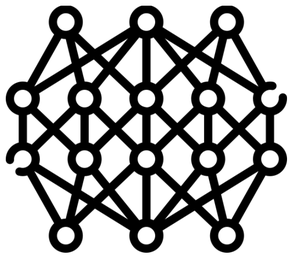


Parametric-only LM

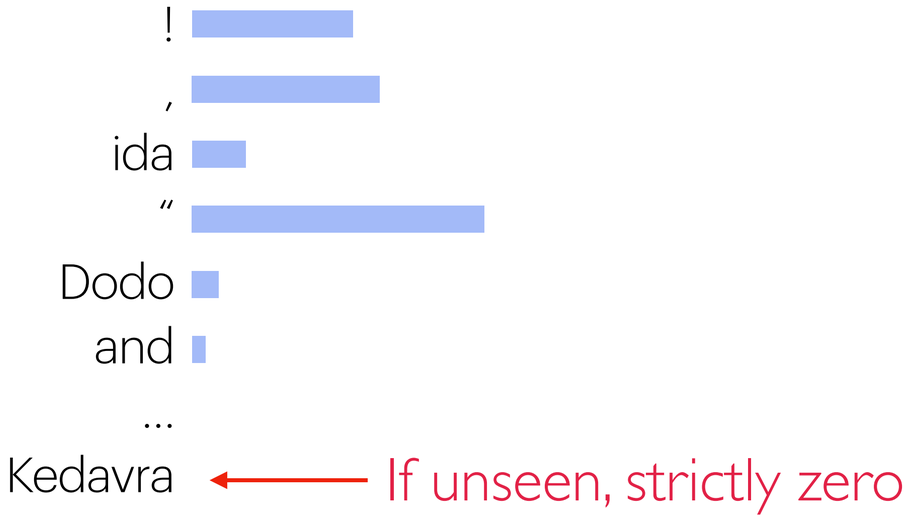
Why does *k*NN-LM generalize better than RiC-LM?

Snape raised his wand and pointed it directly at Dumbledore. "Avada Kedavra!" A jet of green light shot from (...) over the battlements and out of sight."

As Harry shouted "Expelliarmus!" Voldemort cried "Avada



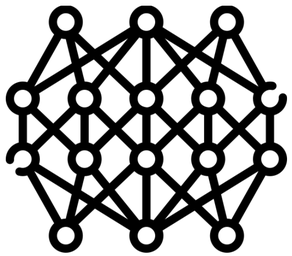
(Ground truth: Kedavra)



RiC-LM

Why does *k*NN-LM generalize better than RiC-LM?

As Harry shouted "Expelliarmus!" Voldemort cried "Avada



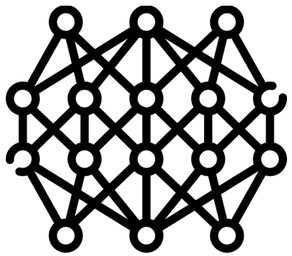
(Ground truth: Kedavra)

...Ron shouted "Harry!
..."Come, Nagini
...ida
...Avada"
...raised his wand and
...
... directly at Dumbledore. "Avada Kedavra
Mad Eye' Moody on the Avada Kedavra
"... come to die. Avada Kedavra

***k*NN-LM**

Why does *k*NN-LM generalize better than RiC-LM?

As Harry shouted "Expelliarmus!" Voldemort cried "Avada



(Ground truth: Kedavra)

...Ron shouted "Harry!
..."Come, Nagini
...ida
...Avada"
...raised his wand and
...
... directly at Dumbledore. "Avada Kedavra
Mad Eye' Moody on the Avada Kedavra
"... come to die. Avada Kedavra

***k*NN-LM**

In this setup with an extreme domain shift, unseen words appear frequently

Content

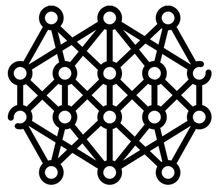
Motivation

A SILO Language Model

Experiments

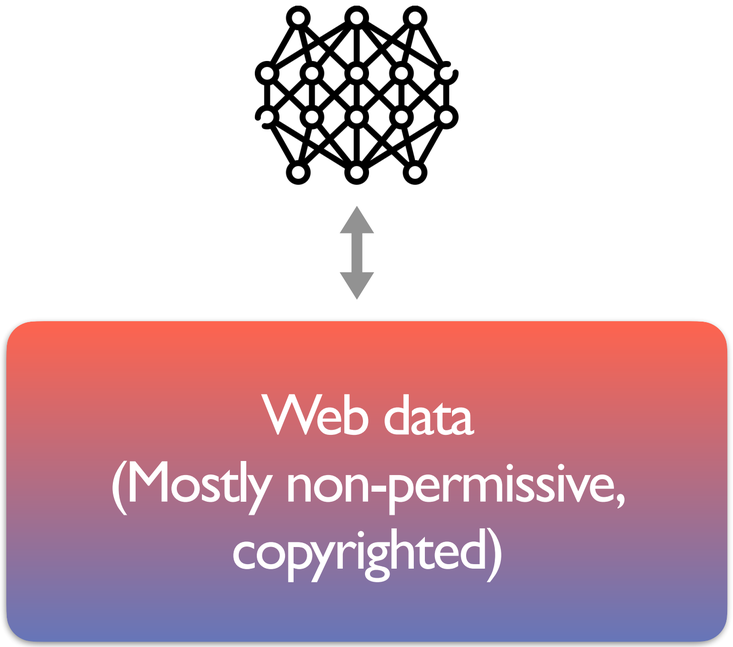
Conclusion & Future Work

Summary



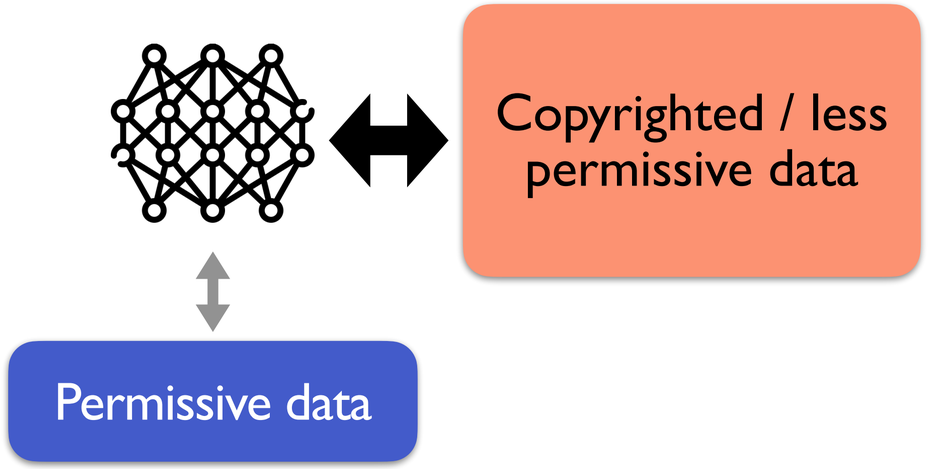
Web data
(Mostly non-permissive, copyrighted)

Summary

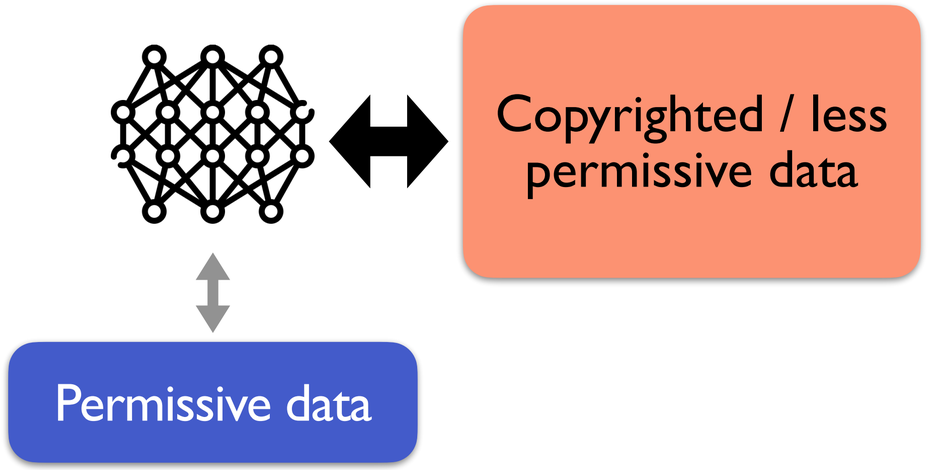


- ✓ Great performance via training on massive text
- ✗ Pose legal risk, e.g., copyright infringement
- ✗ Fail to give credits to data creators
- ✗ Need to re-train for data opt-out

Summary

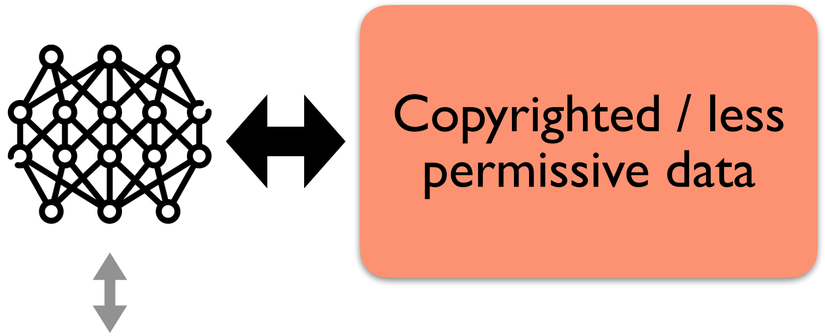


Summary



- ☑ Can opt-out data anytime
- ☑ Data can be attributed/credited
- ☑ Copyright notice can be provided

Summary

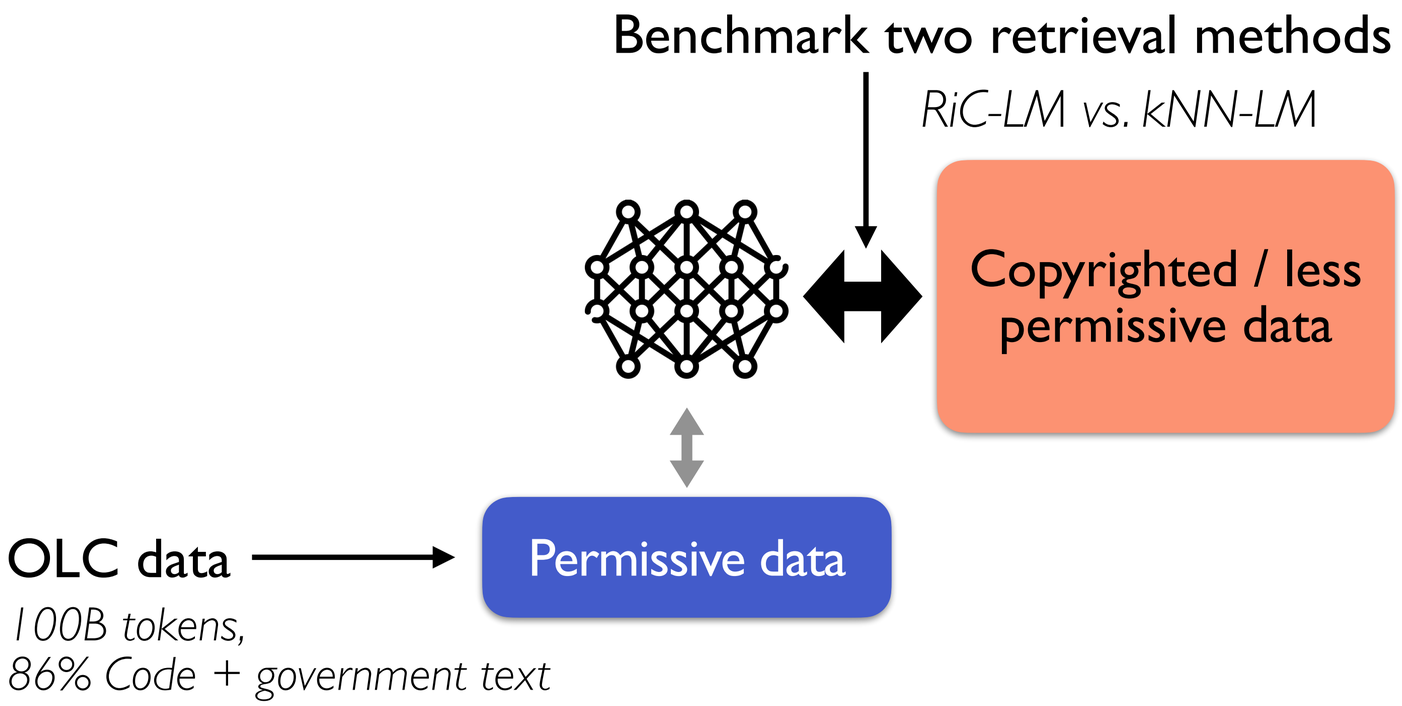


OLC data

100B tokens,
86% Code + government text

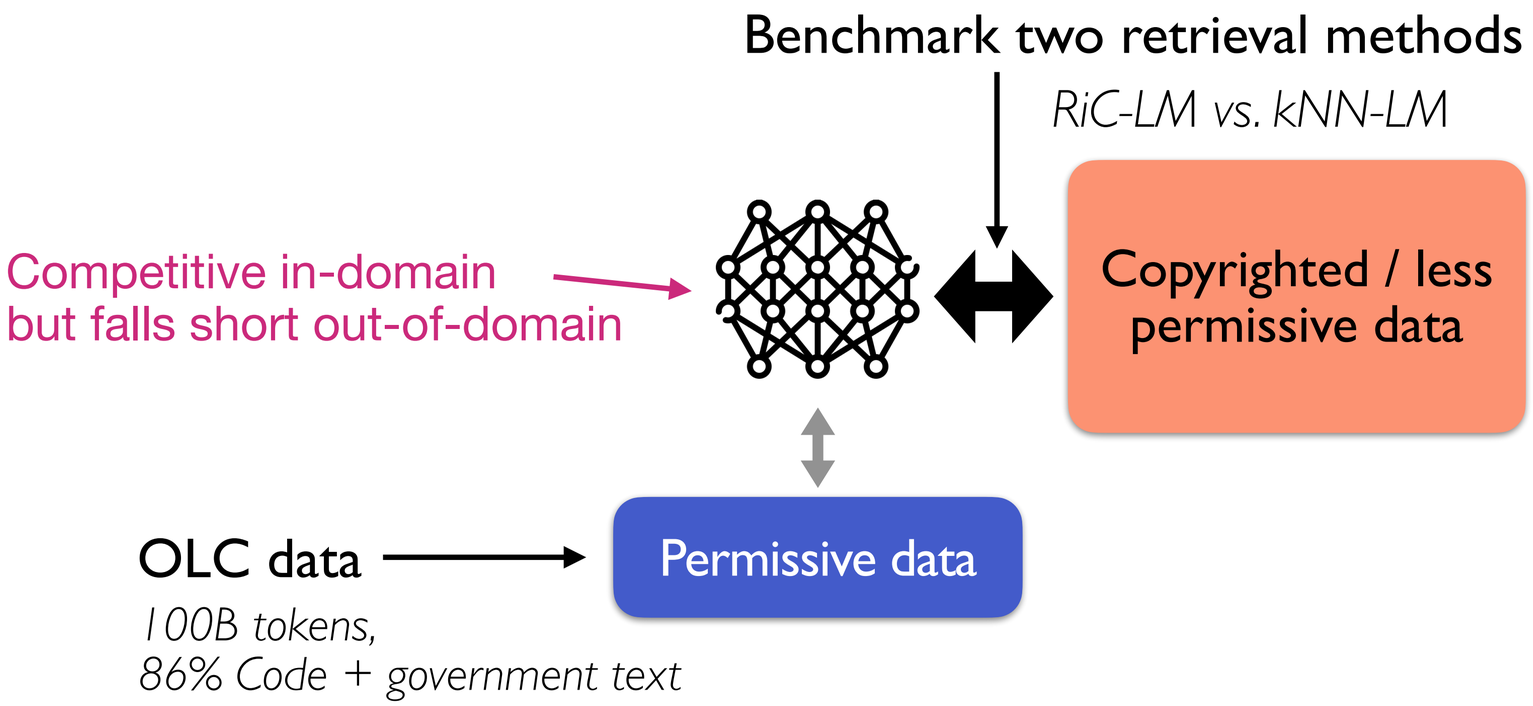
- ☑ Can opt-out data anytime
- ☑ Data can be attributed/credited
- ☑ Copyright notice can be provided

Summary



- ☑ Can opt-out data anytime
- ☑ Data can be attributed/credited
- ☑ Copyright notice can be provided

Summary



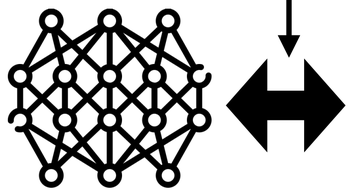
- ☑ Can opt-out data anytime
- ☑ Data can be attributed/credited
- ☑ Copyright notice can be provided

Summary

Benchmark two retrieval methods

RiC-LM vs. kNN-LM

Competitive in-domain
but falls short out-of-domain



Copyrighted / less permissive data

- ☑ Can opt-out data anytime
- ☑ Data can be attributed/credited
- ☑ Copyright notice can be provided

OLC data

100B tokens,
86% Code + government text

Permissive data

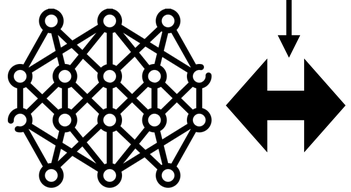
Adding a nonparametric datastore addresses this challenge, closing 90% of the performance gap with Pythia
(kNN-LM is better than RiC-LM, scaling the datastore is important, runtime is slower)

Summary

Benchmark two retrieval methods

RiC-LM vs. kNN-LM

Competitive in-domain
but falls short out-of-domain



Copyrighted / less permissive data

- Can opt-out data anytime
- Data can be attributed/credited
- Copyright notice can be provided

OLC data

100B tokens,
86% Code + government text

Permissive data

Adding a nonparametric datastore addresses this challenge, closing 90% of the performance gap with Pythia
(kNN-LM is better than RiC-LM, scaling the datastore is important, runtime is slower)

- **Doesn't remove legal risk completely**
- **Rather, enables users to have more control →better fair use defense arguments, better alignment with regulations like the GDPR**

Future Work (1/4)

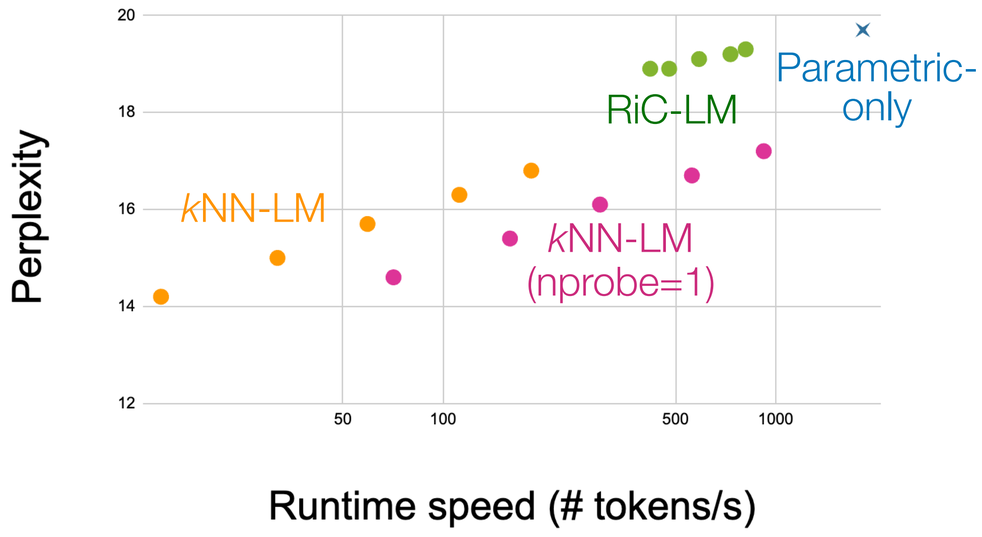
Improving retrieval-based LMs

- Improving the runtime efficiency
- Scaling the datastore
- Training the model to be nonparametric

Future Work (1/4)

Improving retrieval-based LMs

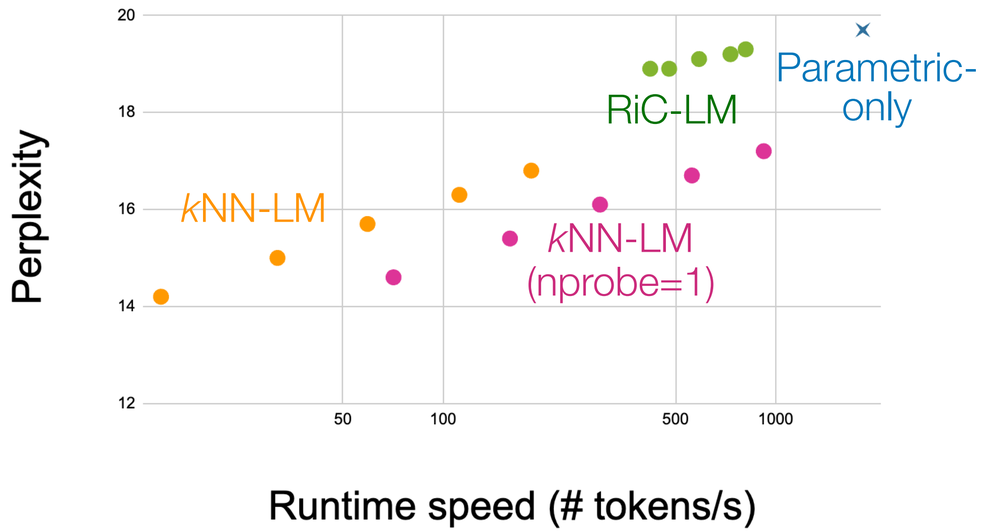
- Improving the runtime efficiency
- Scaling the datastore
- Training the model to be nonparametric

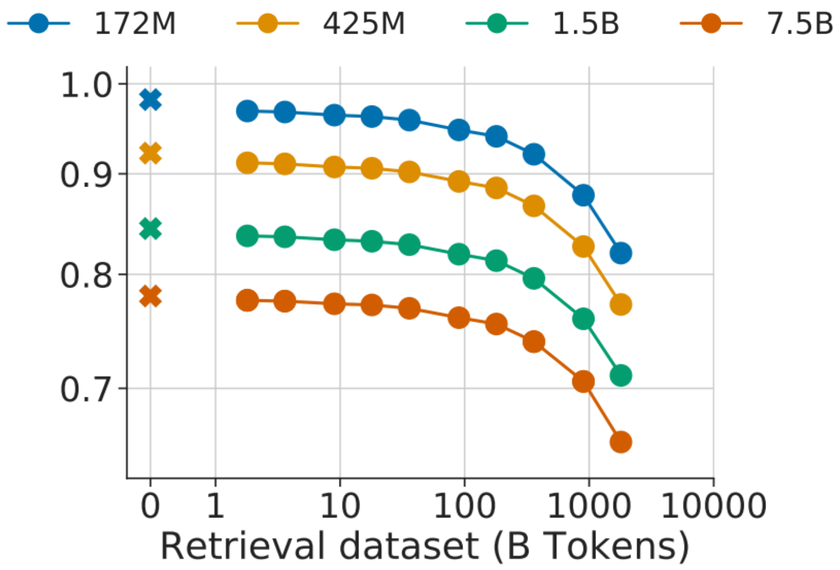


Future Work (1/4)

Improving retrieval-based LMs

- Improving the runtime efficiency
- Scaling the datastore
- Training the model to be nonparametric





Borgeaud et al. 2021

Future Work (2/4)

Improving retrieval-based LMs

Alternative ways to mitigate risk

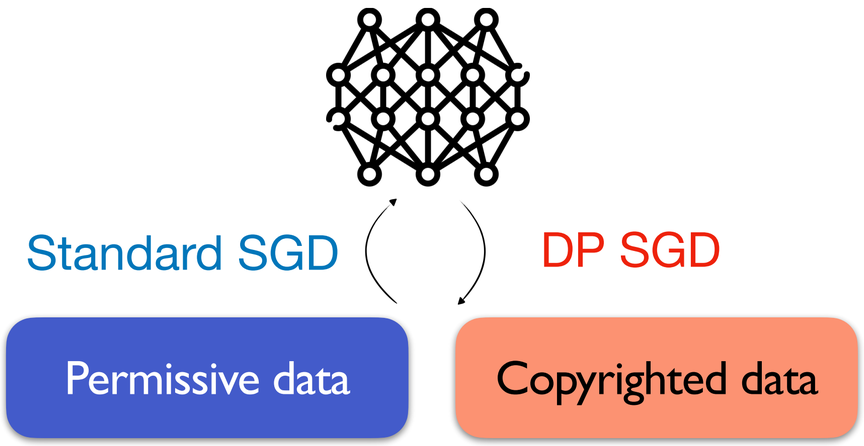
- Near-access free approaches, differential private training / inference
- Expanding OLC, e.g., via inferring data licenses from web crawl
- Modular language modeling

Future Work (2/4)

Improving retrieval-based LMs

Alternative ways to mitigate risk

- Near-access free approaches, differential private training / inference
- Expanding OLC, e.g., via inferring data licenses from web crawl
- Modular language modeling



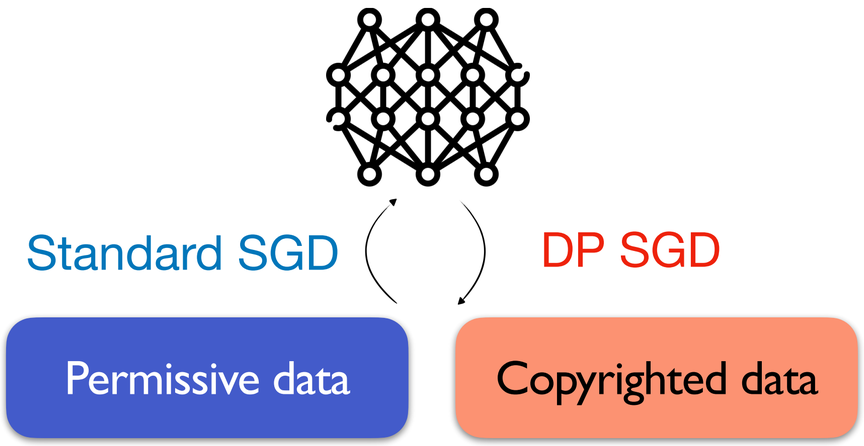
Abadi et al. 2016
Deep Learning with Differential Privacy

Future Work (2/4)

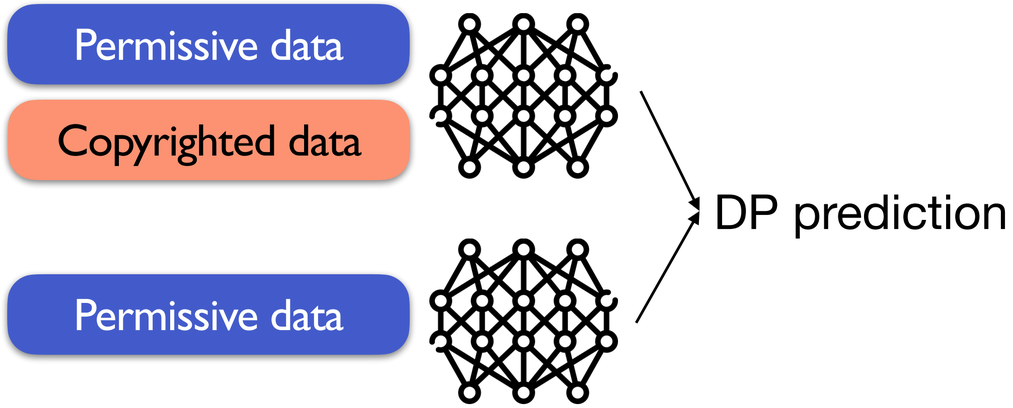
Improving retrieval-based LMs

Alternative ways to mitigate risk

- Near-access free approaches, differential private training / inference
- Expanding OLC, e.g., via inferring data licenses from web crawl
- Modular language modeling



Abadi et al. 2016
Deep Learning with Differential Privacy



Vyas. et al. 2023. On Provable Copyright
Protection for Generative Models

Future Work (3/4)

Improving retrieval-based LMs

Alternative ways to mitigate risk

Other risks & side effect

Project Gutenburg

- SILO may exacerbate other risks (e.g., gender and racial bias, U.S. government positions)

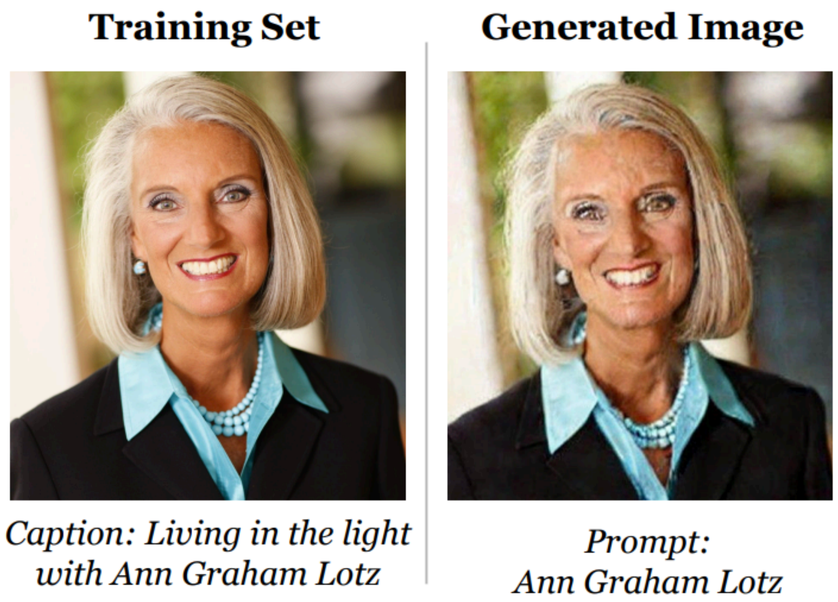
Future Work (4/4)

Improving retrieval-based LMs

Alternative ways to mitigate risk

Other risks & side effect

Beyond text-only



Carlini et al. 2023. Extracting Training Data from Diffusion Models.

Thank you for listening

arXiv LICENSE

arxiv.org/abs/2308.04430

MIT License

github.com/kernelmachine/silo-lm

OLC Data APACHE

huggingface.co/datasets/kernelmachine/open-license-corpus

If you'd like to learn more about Law + Generative AI in general, check out this ICML workshop https://genlaw.github.io
If you'd like to learn more about retrieval-based LMs in general, check out this ACL tutorial: https://acl2023-retrieval-lm.github.io/

Please leave feedback at tinyurl.com/sewon-min-talk