Distributed Language Models: Isolating Data Risks

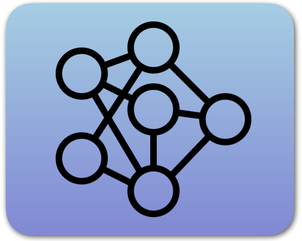
Sewon Min

sewonmin.com

Language models (LMs)

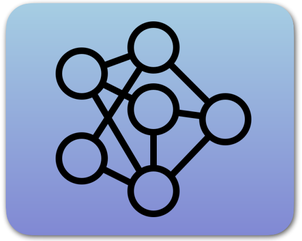
Language models (LMs)

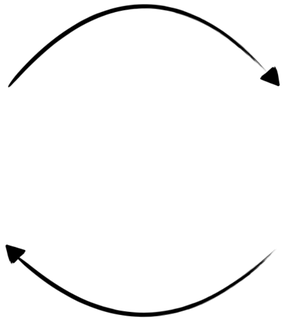
$P(x_t \mid x_{<t})$

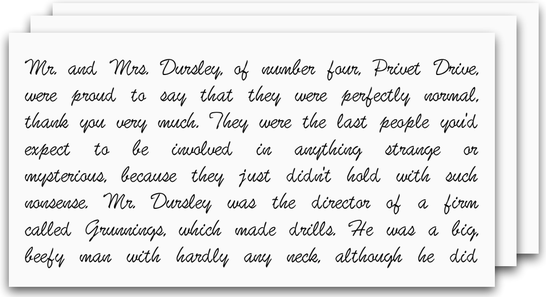


Language models (LMs)

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

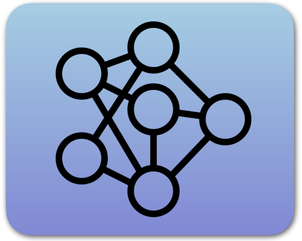


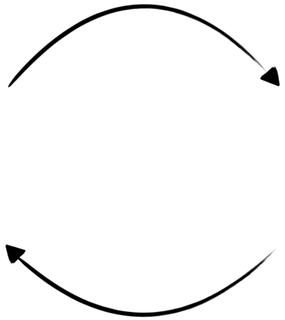


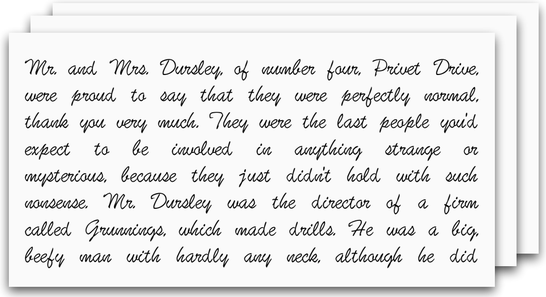


Language models (LMs)

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



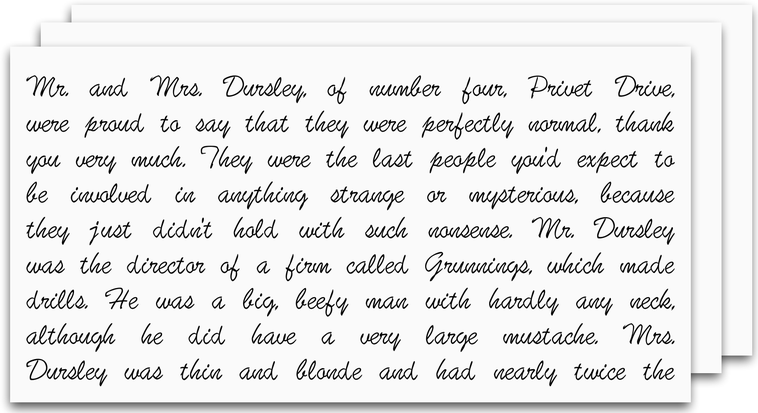




Language models treat all data **homogeneously**

Data is *not* homogenous

Project Gutenberg



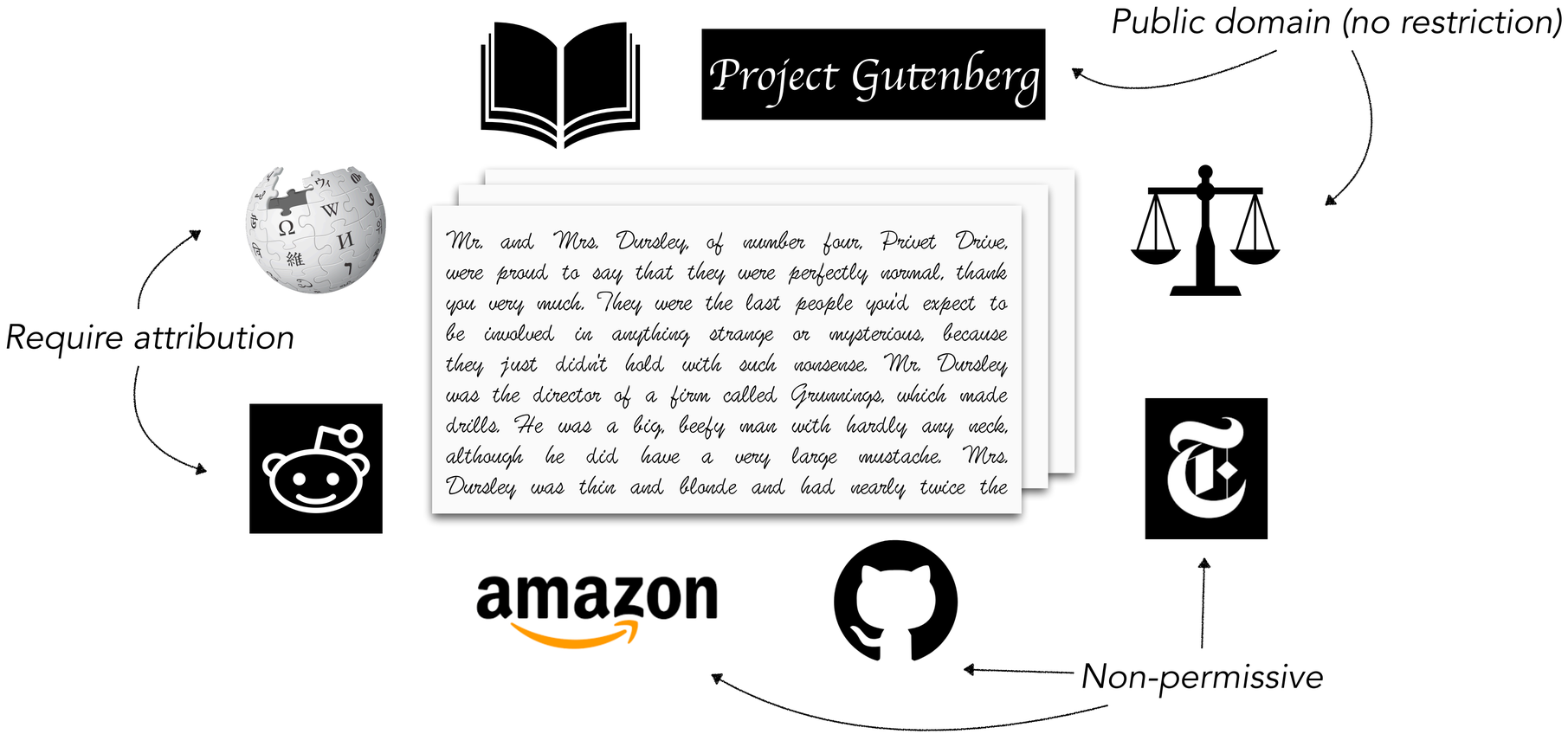
Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the

amazon

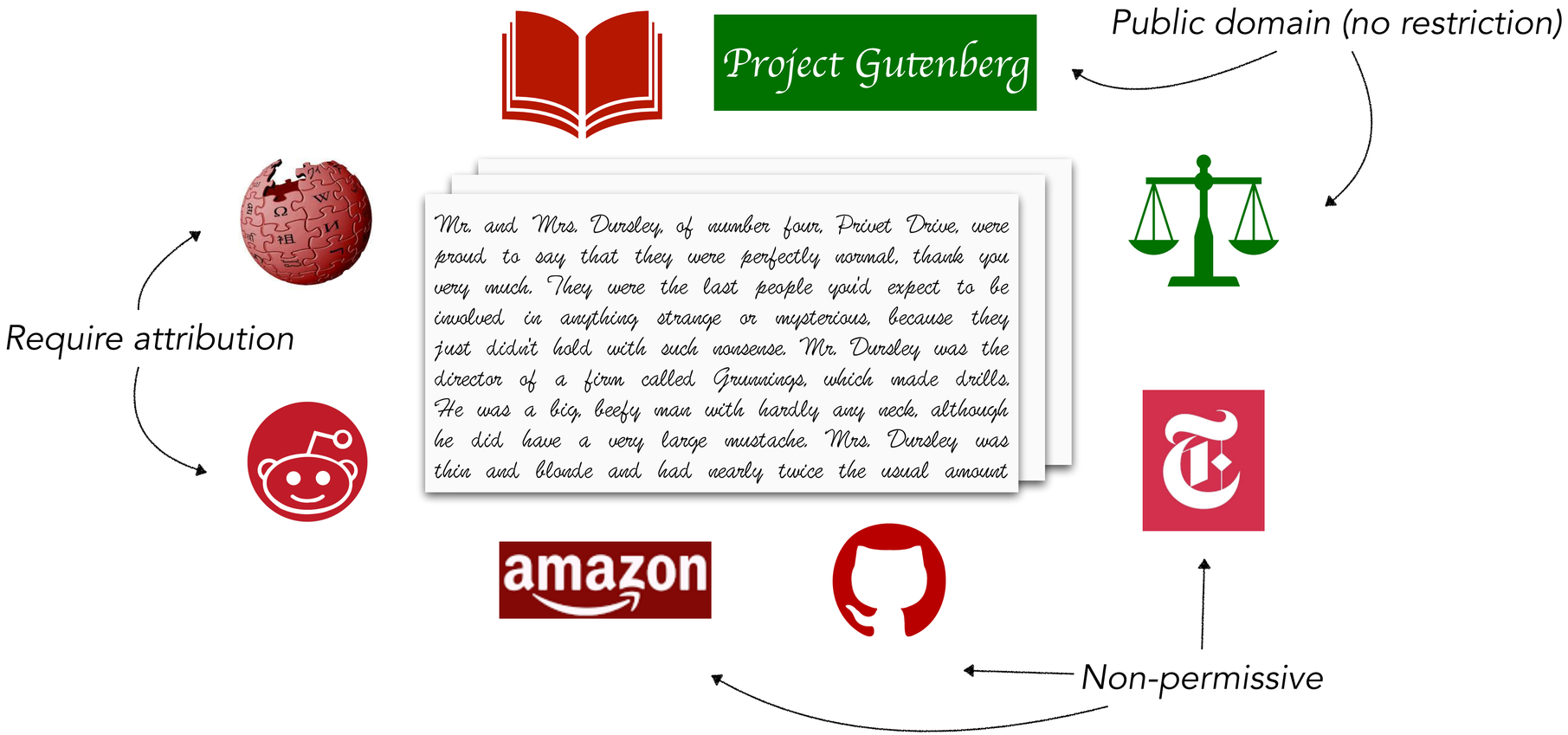
Data is *not* homogenous

Data is *not* homogenous

Data is *not* homogenous



Data is *not* homogenous



Example: Licenses

Example: Licenses

Public domain

"[…] with no conditions"

Example: Licenses

Public domain

"[…] with no conditions"

CC-BY

(CC-BY-SA)

"[…] as long as they credit you for the original creation."

Example: Licenses

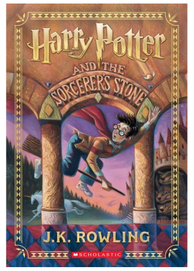
Public domain	CC-BY (CC-BY-SA)	CC-BY-NC (CC-BY-NC-SA)
“[…] with no conditions”	“[…] as long as they credit you for the original creation.”	[Others] can’t use them commercially.

Example: Licenses

Public domain	CC-BY (CC-BY-SA)	CC-BY-NC (CC-BY-NC-SA)
"[…] with no conditions"	"[…] as long as they credit you for the original creation."	[Others] can't use them commercially.

Close access journals/textbooks, most books, most news articles, …

Library Genesis



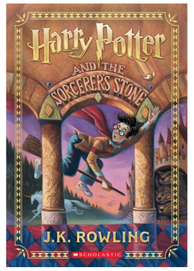
The New York Times
THE WALL STREET JOURNAL.
The Washington Post

Example: Licenses

Public domain	CC-BY (CC-BY-SA)	CC-BY-NC (CC-BY-NC-SA)
"[…] with no conditions"	"[…] as long as they credit you for the original creation."	[Others] can't use them commercially.

Close access journals/textbooks, most books, most news articles, …

Library Genesis



The New York Times
THE WALL STREET JOURNAL.
The Washington Post

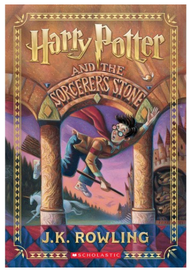
Current LM training does not consider these restrictions

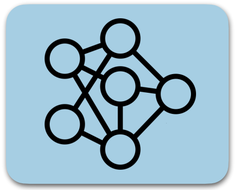
Risks in LM training

Difficulty in training data attribution

Risks in LM training

Difficulty in training data attribution

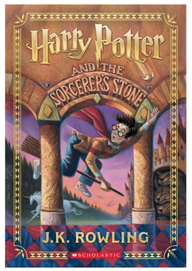


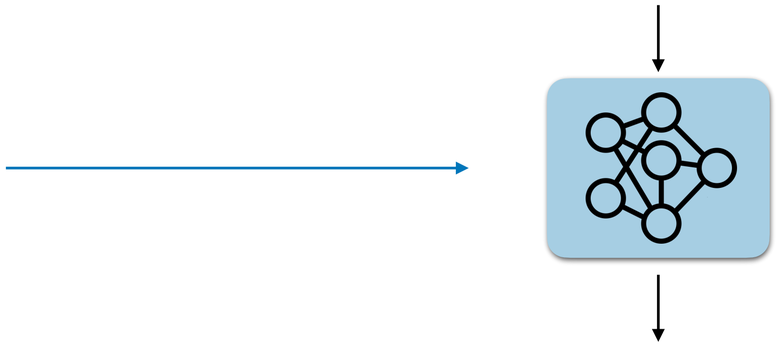


Risks in LM training

Difficulty in training data attribution

Write an original story about a boy who discovers he is a famous wizard on his 11th birthday.



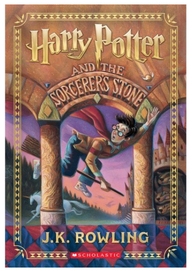


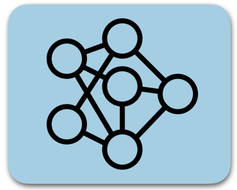
[…] He goes to Hogwarts School of Witchcraft and Wizardry and makes friends with Ron Weasley and Hermione Granger. […]

Risks in LM training

Difficulty in training data attribution

Write an original story about a boy who discovers he is a famous wizard on his 11th birthday.





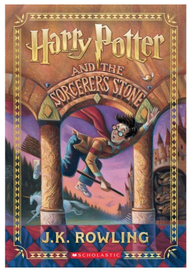
[…] He goes to Hogwarts School of Witchcraft and Wizardry and makes friends with Ron Weasley and Hermione Granger. […]

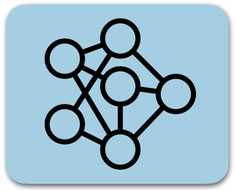
Risks in LM training

Difficulty in training data attribution

Reproduction of original work

Write an original story about a boy who discovers he is a famous wizard on his 11th birthday.





[…] He goes to Hogwarts School of Witchcraft and Wizardry and makes friends with Ron Weasley and Hermione Granger. […]

Chen, …., **Min** et al. Submitted to NeurIPS 2024

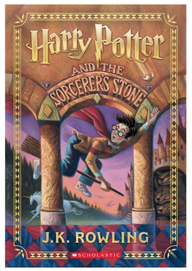
Risks in LM training

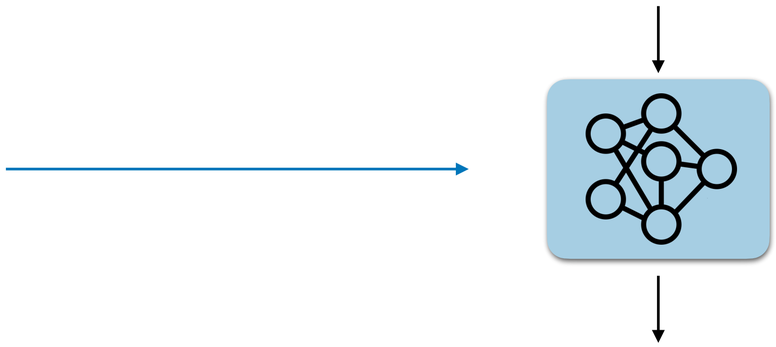
Difficulty in training data attribution

Reproduction of original work

Competition with original data's market

Write an original story about a boy who discovers he is a famous wizard on his 11th birthday.





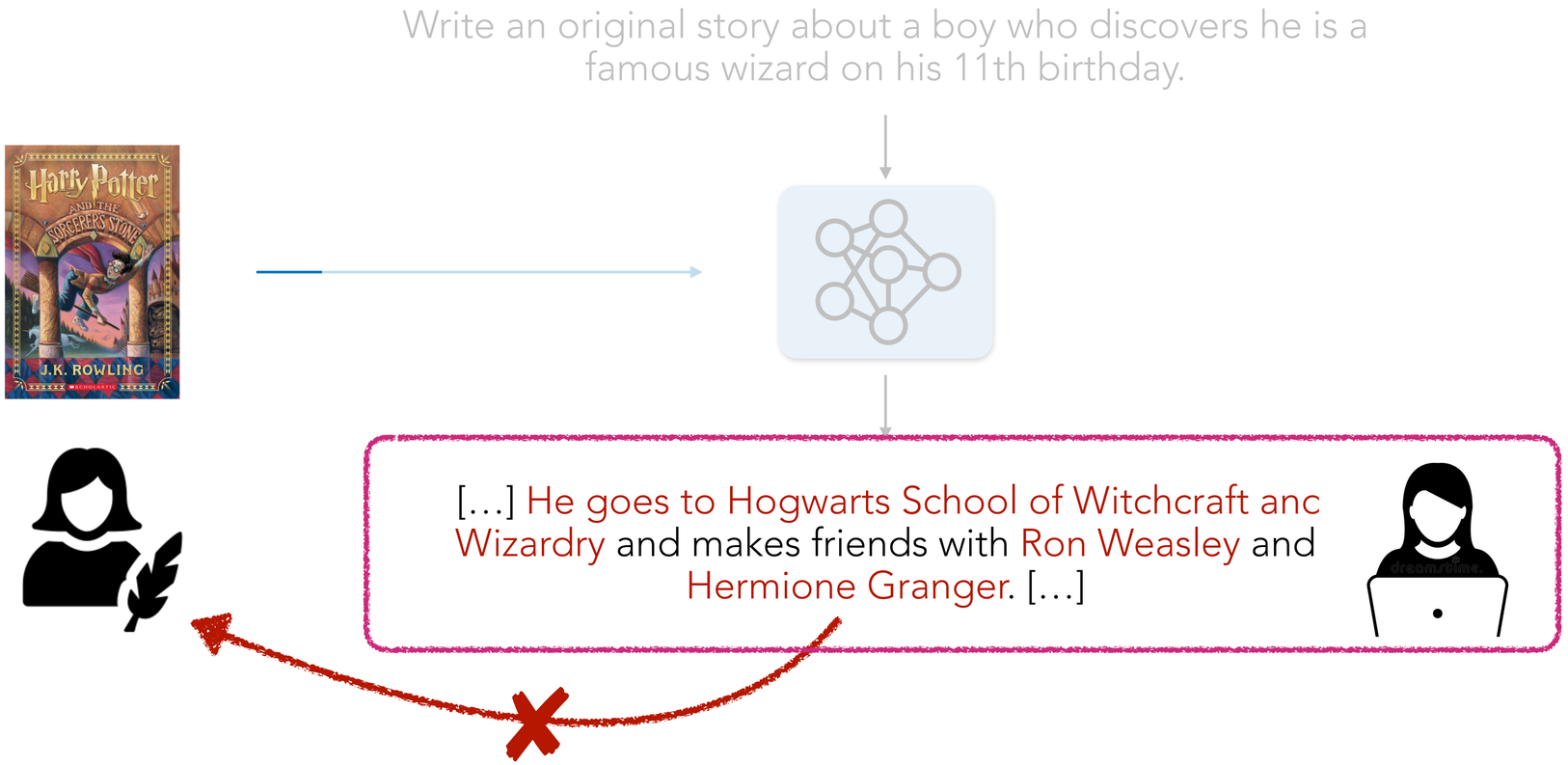
[…] He goes to Hogwarts School of Witchcraft and Wizardry and makes friends with Ron Weasley and Hermione Granger. […]

Risks in LM training

Difficulty in training data attribution

Reproduction of original work

Competition with original data's market

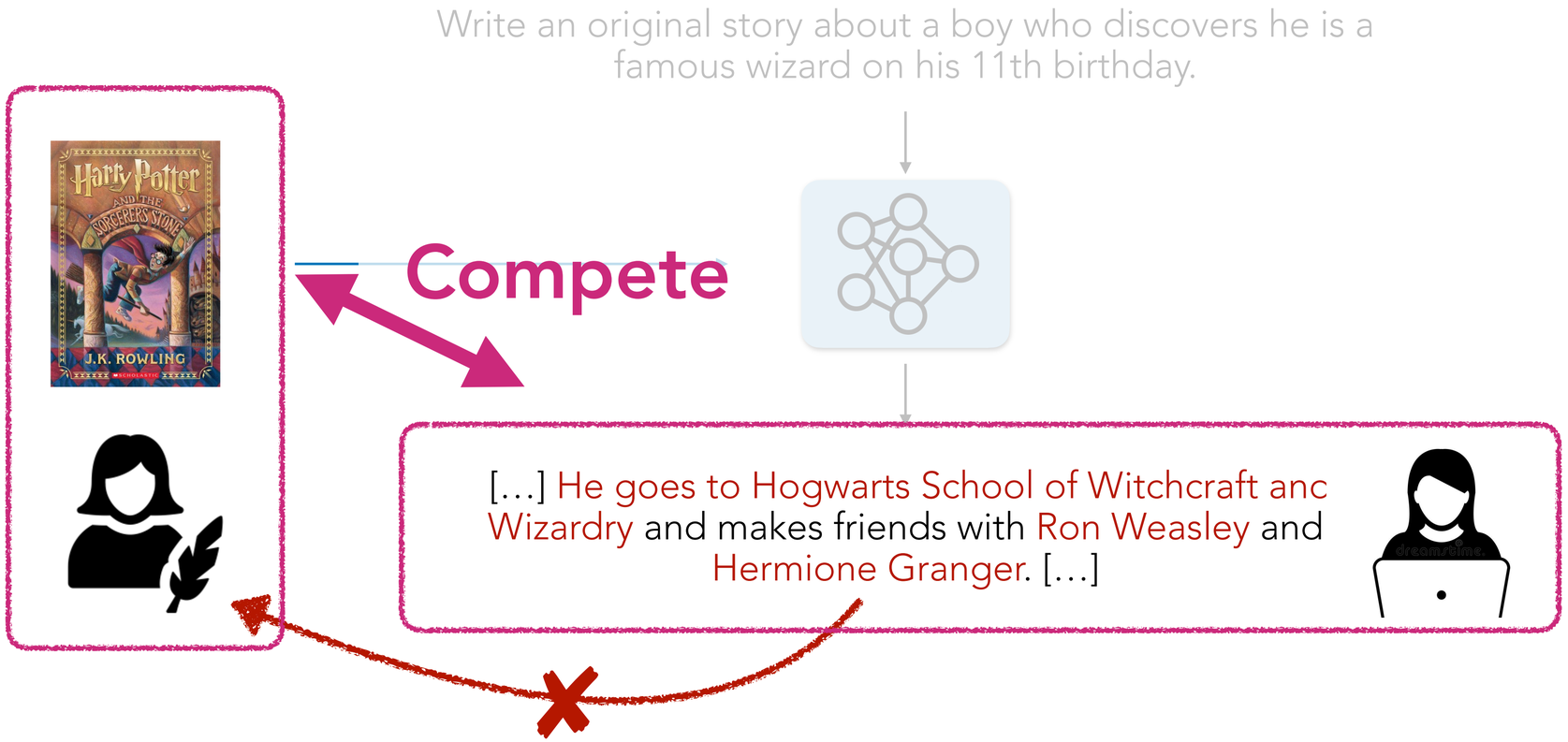


Risks in LM training

Difficulty in training data attribution

Reproduction of original work

Competition with original data's market



Risks in LM training

I want my books to be excluded.

Risks in LM training

I want my books to be excluded.

I want to get credited whenever the model uses my articles.

Risks in LM training

I want my books to be excluded.

I want to get credited whenever the model uses my articles.

I want to delete my private information.

Risks in LM training

I want my books to be excluded.

I want to get credited whenever the model uses my articles.

I want to delete my private information.

Risks in LM training

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I want to delete my private information.

Risks in LM training

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

I want to delete my private information.

Risks in LM training

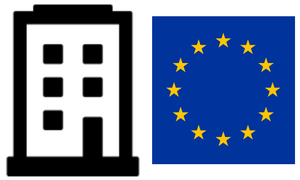
I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

I want to delete my private information.



Risks in LM training

I want my books to be excluded.

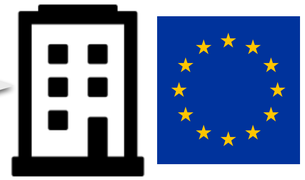
I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

I want to delete my private information.

I need to delete user data to comply with GDPR.



Risks in LM training

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

I want to delete my private information.

I need to delete user data to comply with GDPR.

Risks in LM training

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

I want to delete my private information.

I need to delete user data to comply with GDPR.

- Train on permissive data only?

Risks in LM training

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

I want to delete my private information.

I need to delete user data to comply with GDPR.

- Train on permissive data only? → **Poor performance, too conservative**

Risks in LM training

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

I want to delete my private information.

I need to delete user data to comply with GDPR.

- Train on permissive data only? → **Poor performance, too conservative**
- Remove problematic data ad-hoc?

Risks in LM training

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

I want to delete my private information.

I need to delete user data to comply with GDPR.

- Train on permissive data only? → **Poor performance, too conservative**
- Remove problematic data ad-hoc? → **Re-training is very expensive**

My research

LMs for responsible data use

My research

LMs for responsible data use

Building LMs for responsible data use

- Nonparametric LMs with a datastore (**Min***, Gururangan* et al. ICLR 2024 Spotlight)
- Distributed LMs (Proposal)

My research

LMs for responsible data use

Building LMs for responsible data use

- Nonparametric LMs with a datastore (**Min***, Gururangan* et al. ICLR 2024 Spotlight)
- Distributed LMs (Proposal)

Auditing LMs for responsible data use

- Identify whether the model is trained on the data (Duan, …., **Min** et al. Submitted to COLM 2024)
- Identify whether the model reproduces the data (Chen, …., **Min** et al. Submitted to NeurIPS 2024)

My research

LMs for responsible data use

Building LMs for responsible data use

- Nonparametric LMs with a datastore (**Min***, Gururangan* et al. ICLR 2024 Spotlight)
- Distributed LMs (Proposal)

Auditing LMs for responsible data use

- Identify whether the model is trained on the data (Duan, …., **Min** et al. Submitted to COLM 2024)
- Identify whether the model reproduces the data (Chen, …., **Min** et al. Submitted to NeurIPS 2024)

Talk outline

Talk outline

Data: Open License Corpus

100B token text corpus with public domain & permissively-licensed text

Talk outline

Data: Open License Corpus

100B token text corpus with public domain & permissively-licensed text

Our Model: SILO

A nonparametric LM that isolates risks in a datastore

Talk outline

Data: Open License Corpus

100B token text corpus with public domain & permissively-licensed text

Our Model: SILO

A nonparametric LM that isolates risks in a datastore

Proposal: Distributed LMs

LMs with a set of components, allowing flexible activation at test time

Talk outline

Data: Open License Corpus

100B token text corpus with public domain & permissively-licensed text

Our Model: SILO

A nonparametric LM that isolates risks in a datastore

Proposal: Distributed LMs

LMs with a set of components, allowing flexible activation at test time

Permissive text

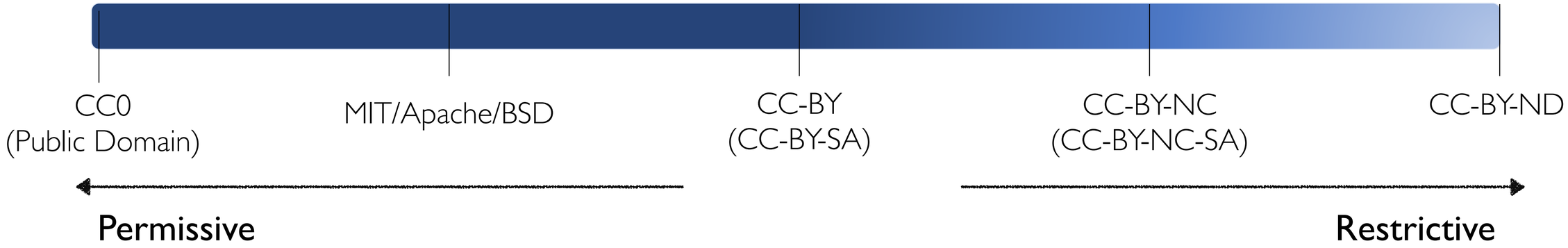
Existing open access data: Common Crawl, Dolma, FineWeb, …

→ Still mostly copyrighted!

Permissive text

Existing open access data: Common Crawl, Dolma, FineWeb, …

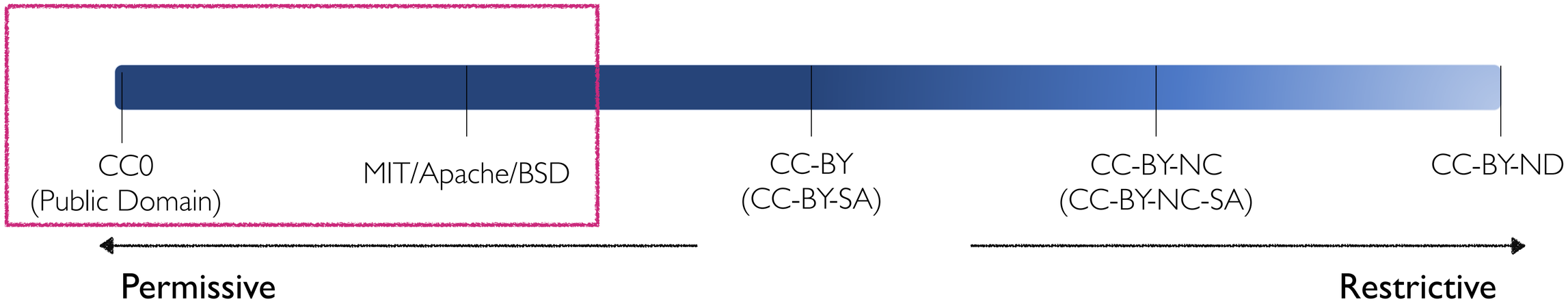
→ Still mostly copyrighted!



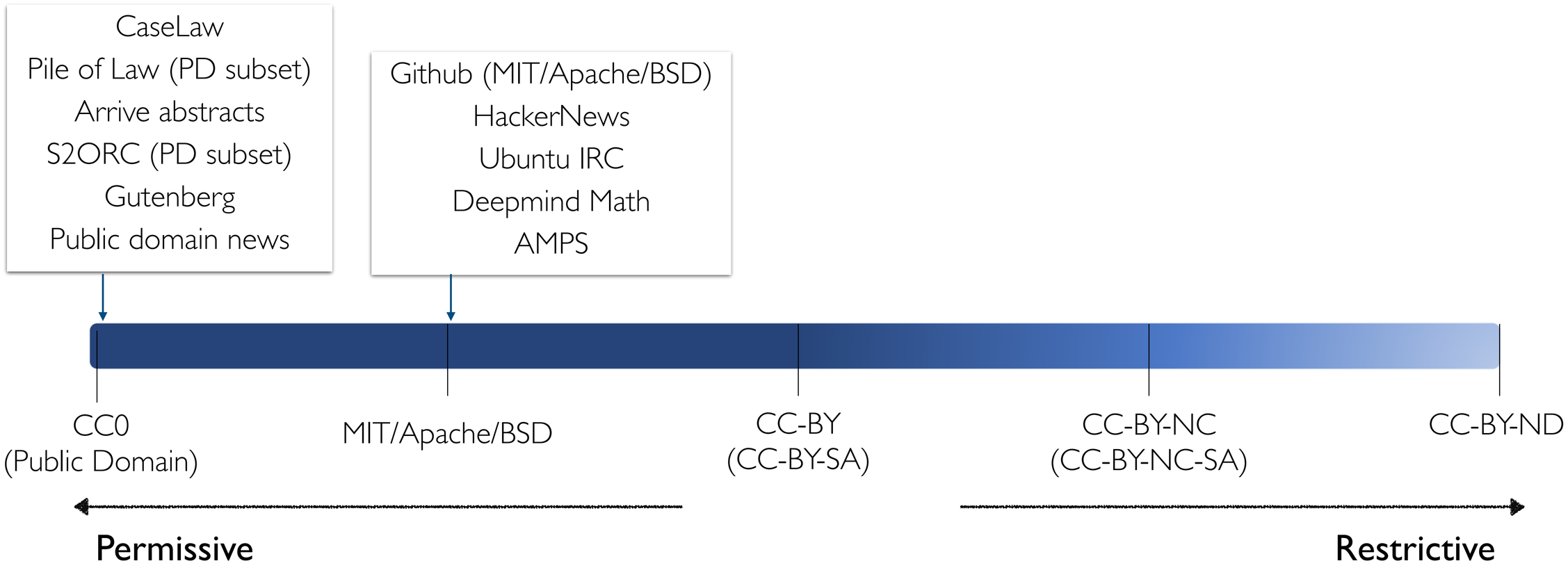
Permissive text

Existing open access data: Common Crawl, Dolma, FineWeb, …

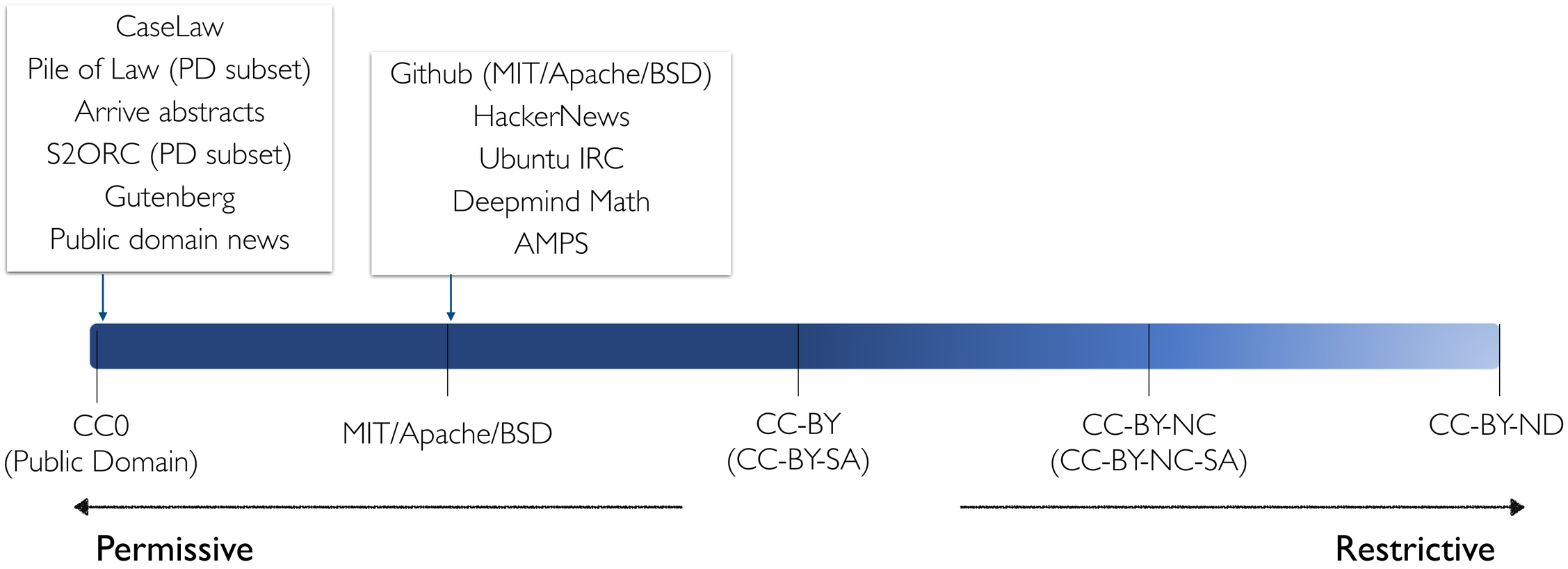
→ Still mostly copyrighted!



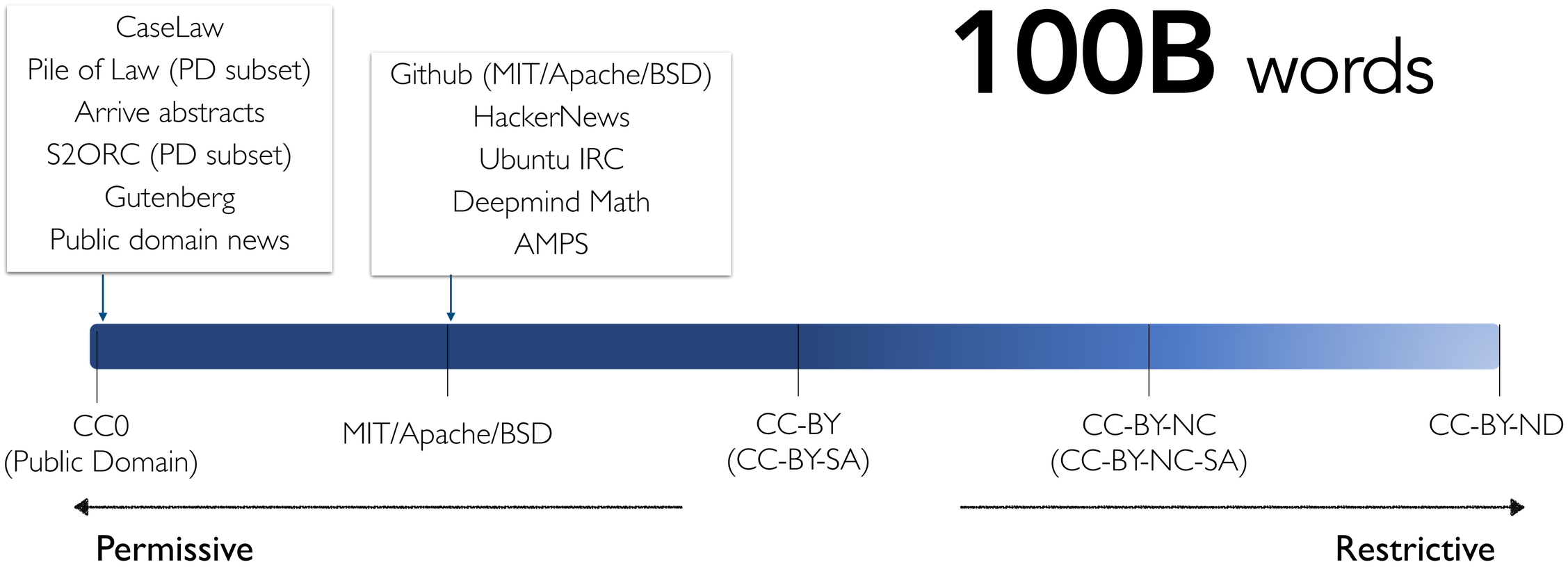
Permissive text



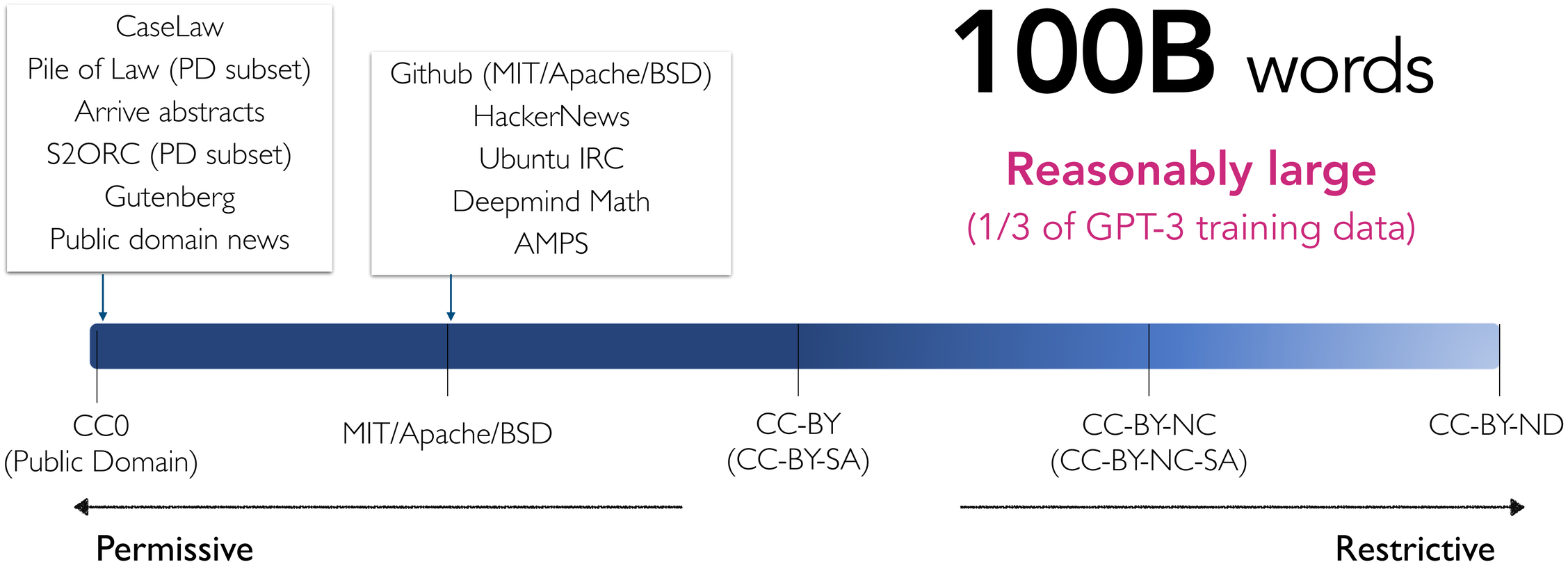
Open License Corpus (OLC)



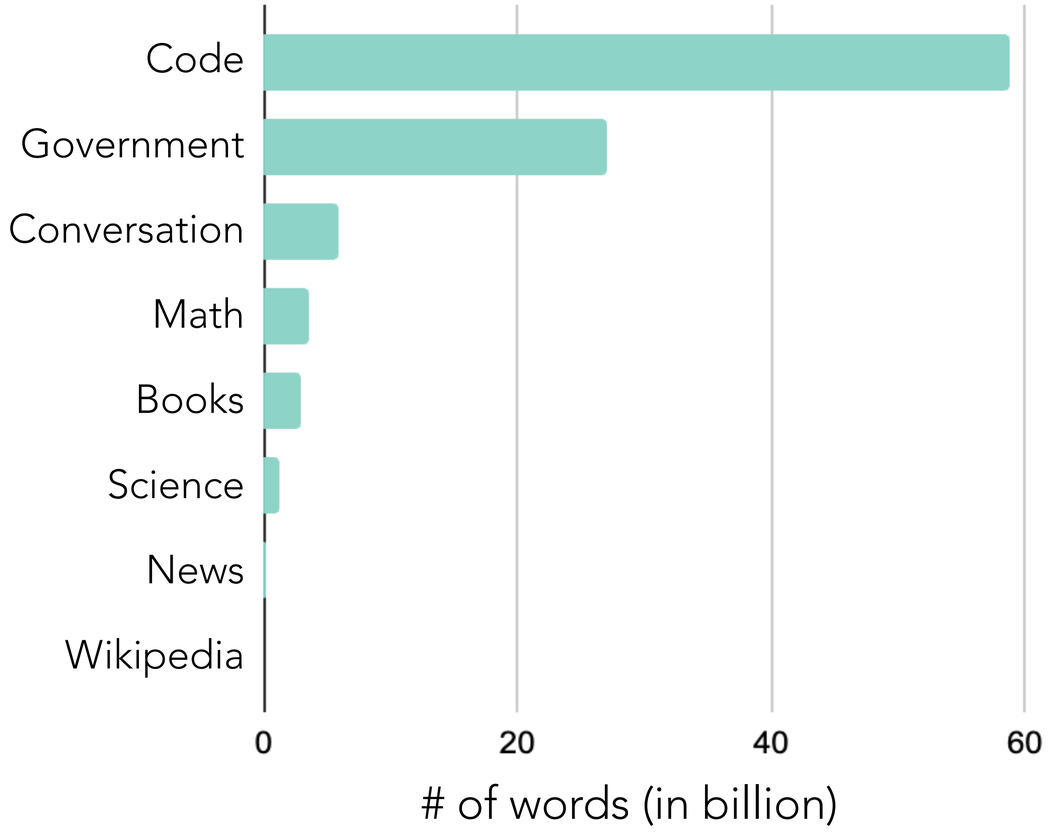
Open License Corpus (OLC)



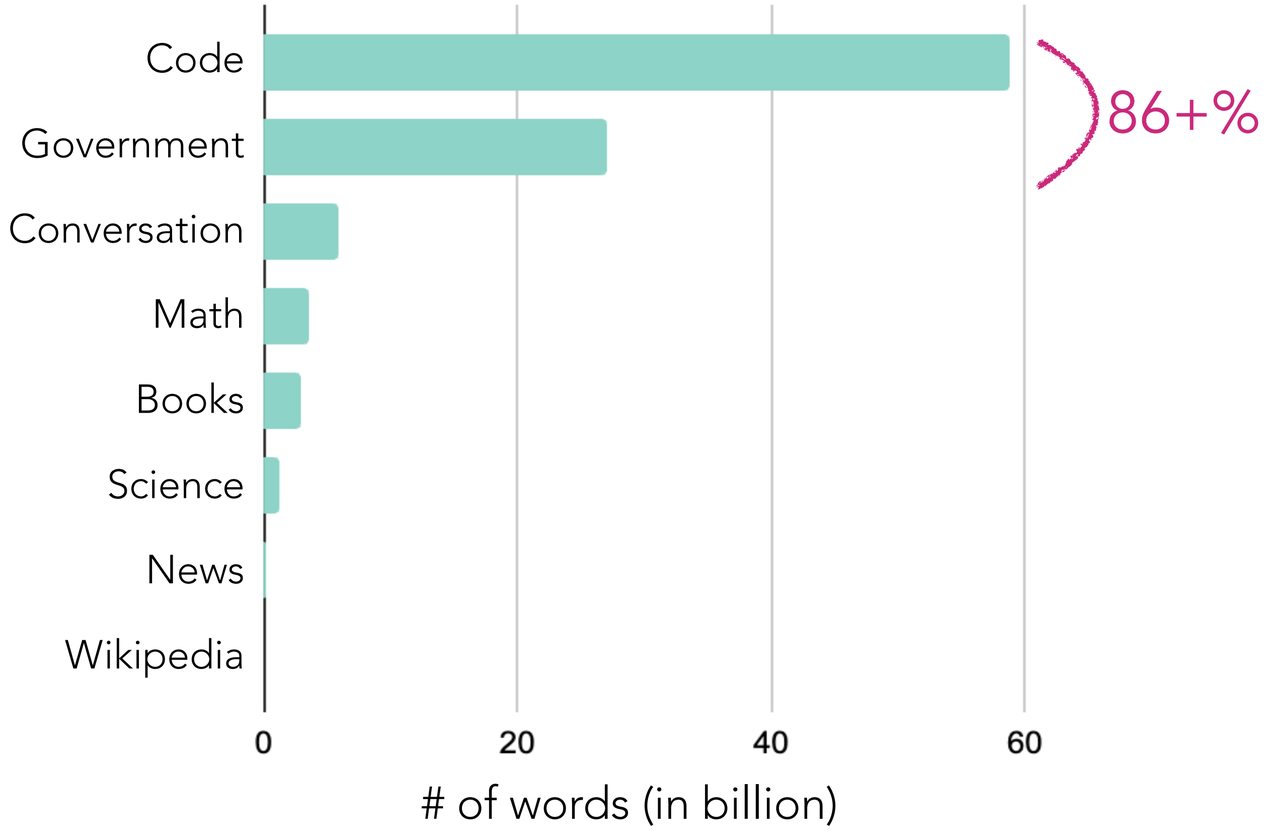
Open License Corpus (OLC)



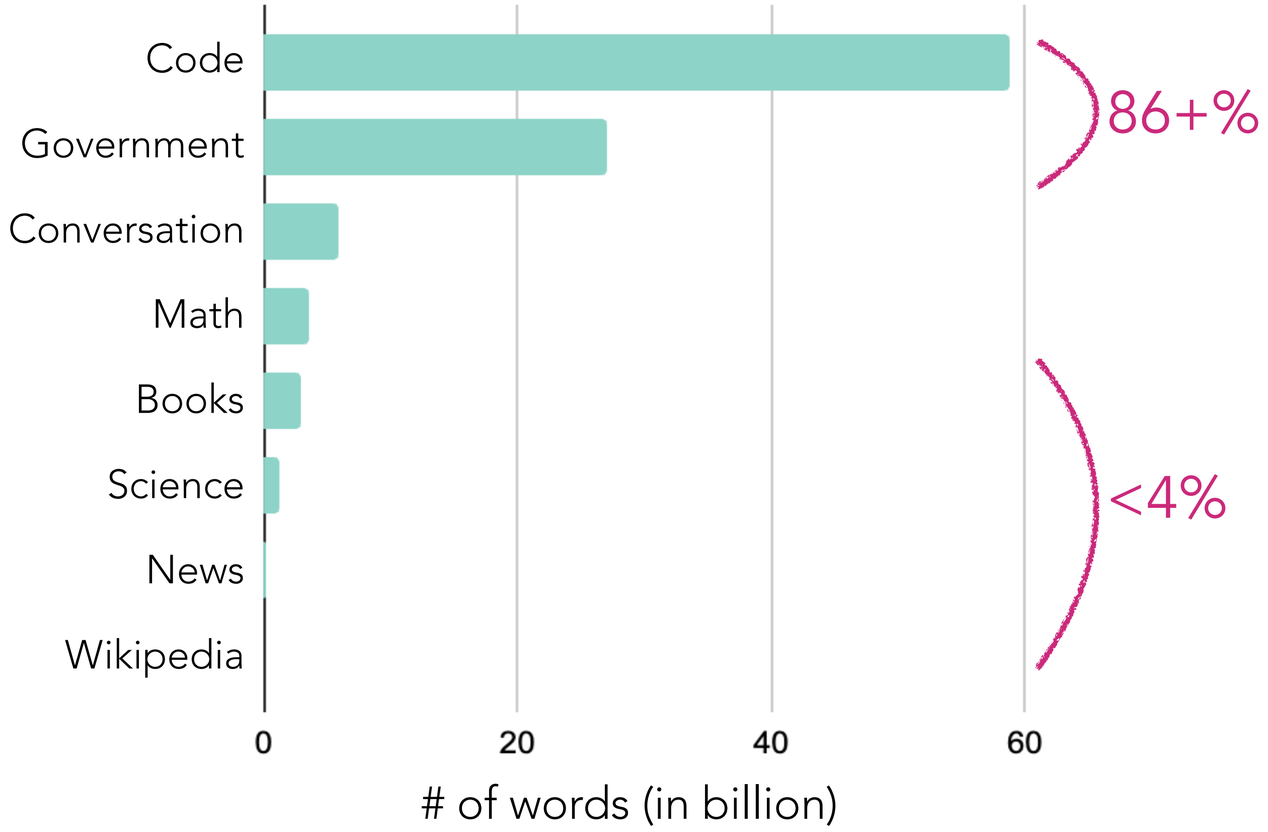
Open License Corpus (OLC)



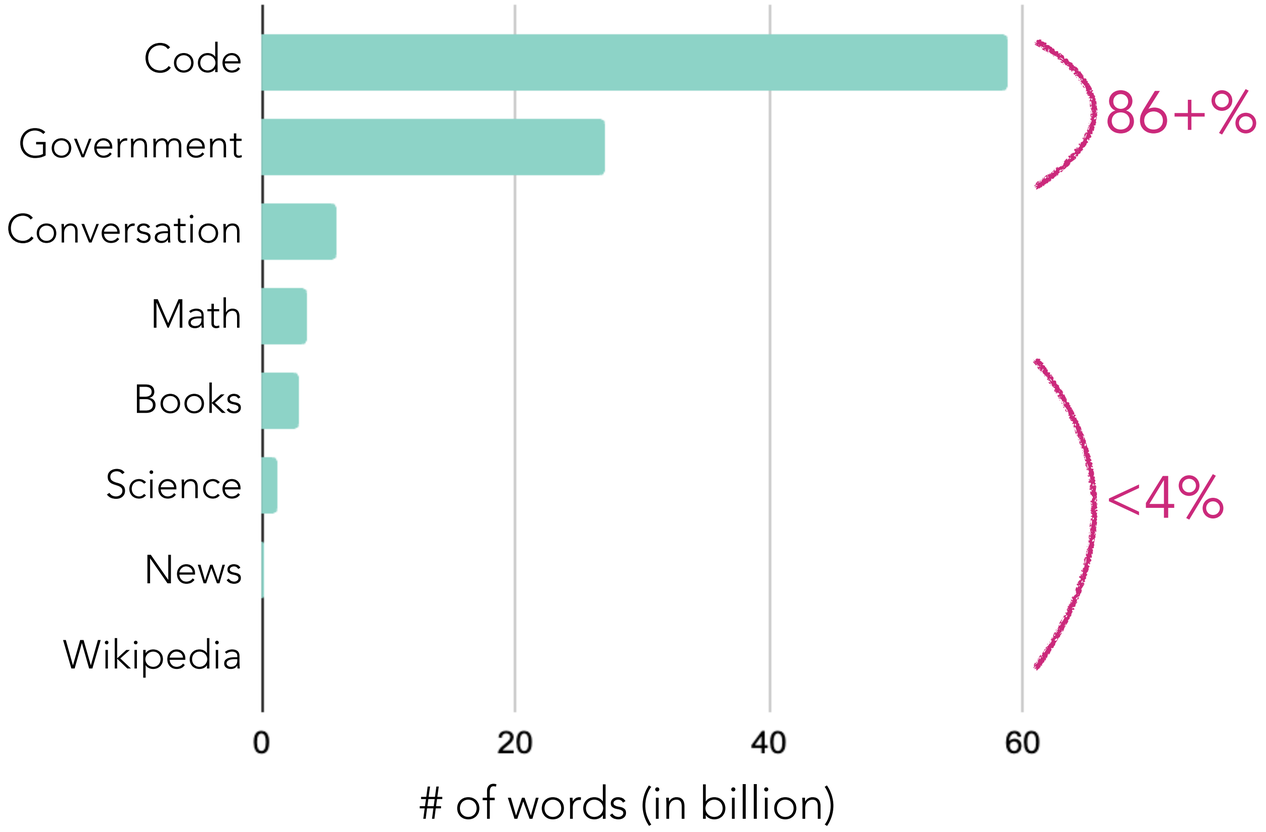
Open License Corpus (OLC)



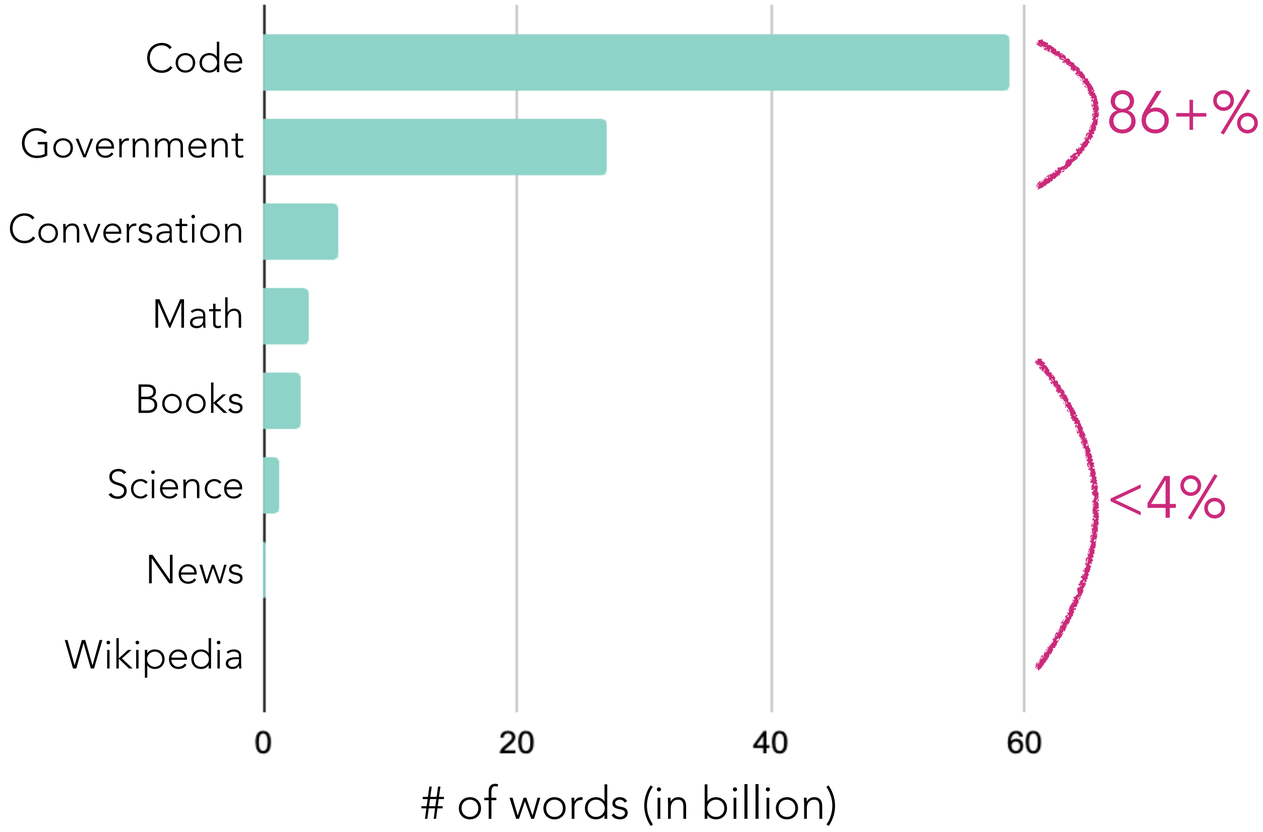
Open License Corpus (OLC)



Open License Corpus (OLC)

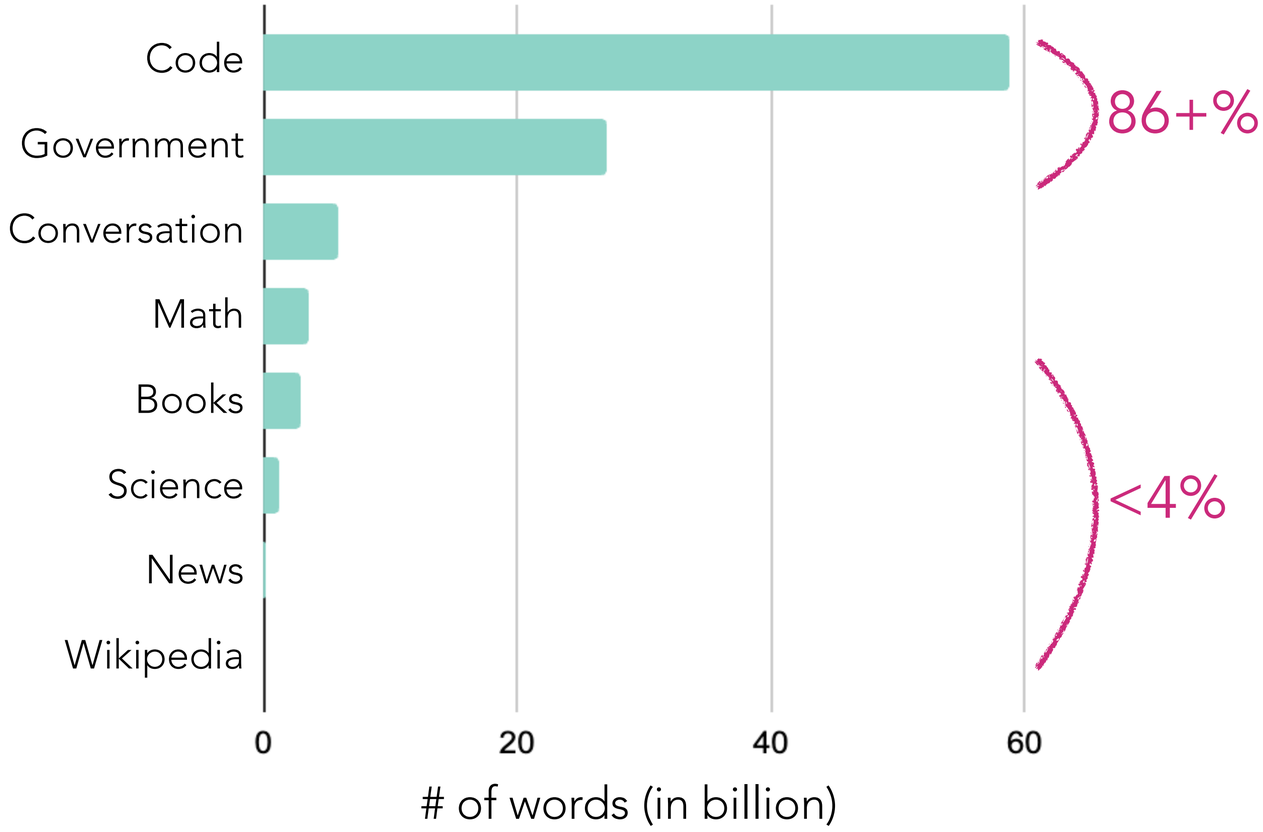


Open License Corpus (OLC)



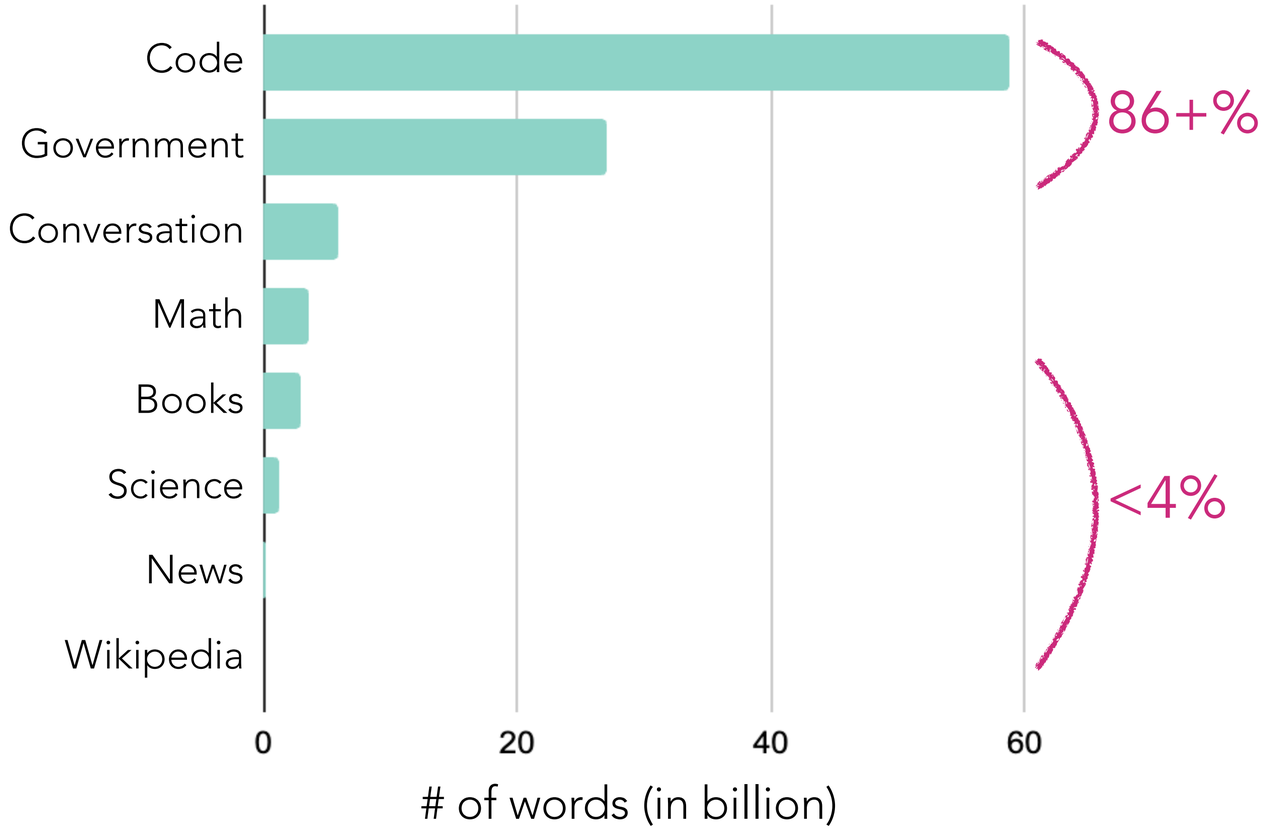
- Small size
  (although larger than we thought)

Open License Corpus (OLC)



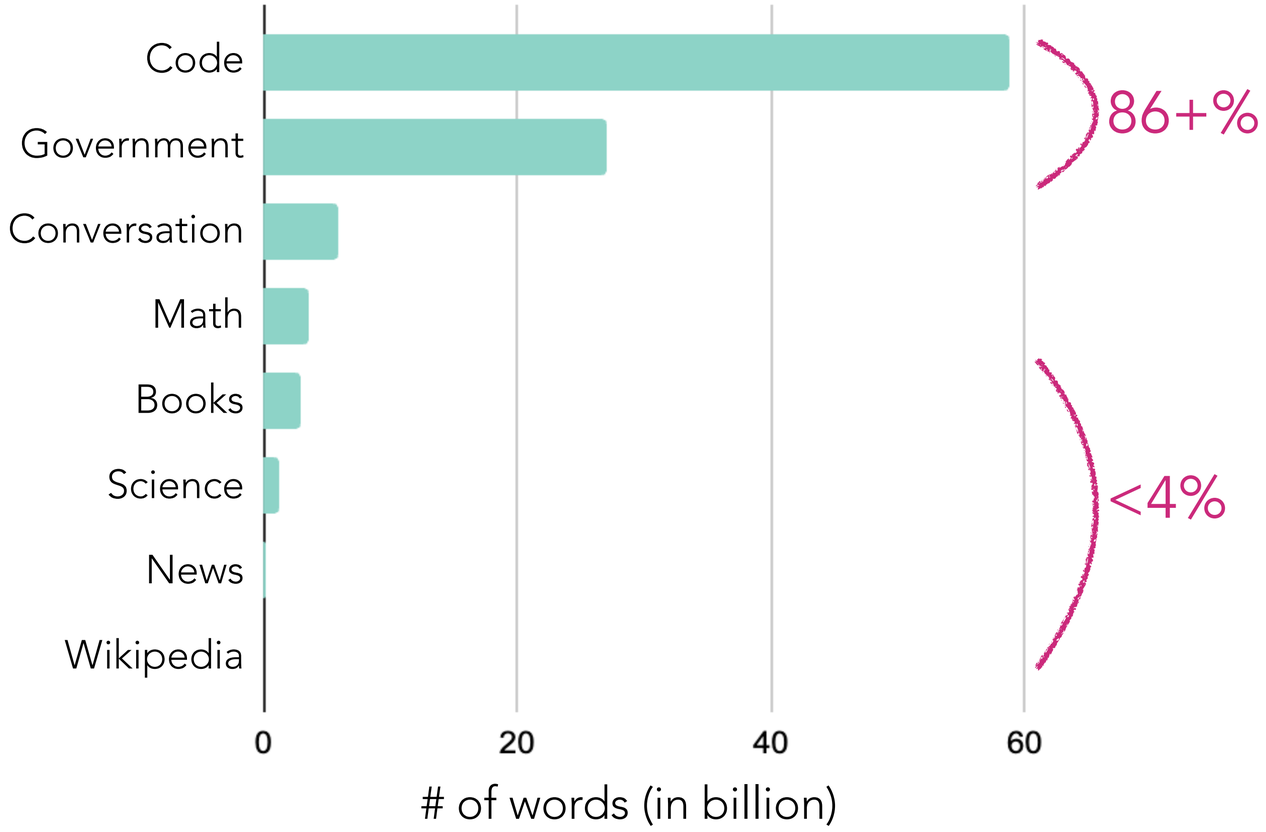
- Small size
  (although larger than we thought)
- Extreme domain distribution

Open License Corpus (OLC)



- Small size
  (although larger than we thought)
- Extreme domain distribution
- Lack of high-quality data
  (e.g., books, news)

Open License Corpus (OLC)



- Small size
  (although larger than we thought)
- Extreme domain distribution
- Lack of high-quality data
  (e.g., books, news)

Models trained on OLC must address these issues

Talk outline

Data: Open License Corpus

100B token text corpus with public domain & permissively-licensed text

Our Model: SILO

A nonparametric LM that isolates risks in a datastore

Proposal: Distributed LMs

LMs with a set of components, allowing flexible activation at test time

Recap: Data risks

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

Copilot

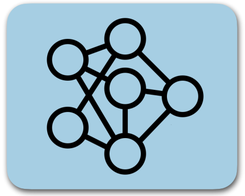
I want to delete my private information.

I need to delete user data to comply with GDPR.

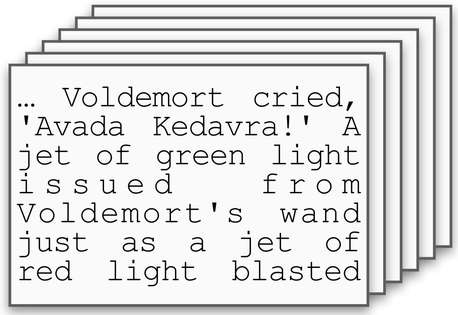
- Train on permissive data only? → **Poor performance, too conservative**
- Remove problematic data ad-hoc? → **Re-training is very expensive**

Our proposal: A nonparametric LM

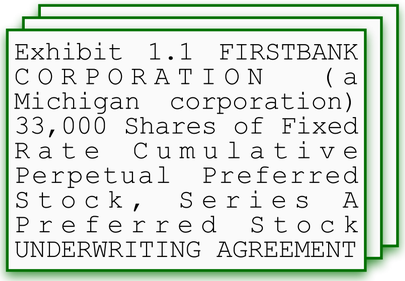
Parameters



Datastore

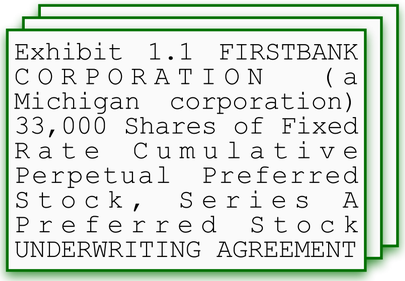


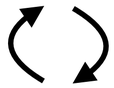
Our proposal: A nonparametric LM

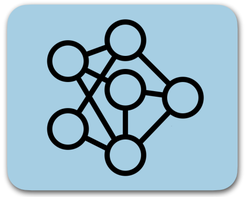


Our proposal: A nonparametric LM

Training data







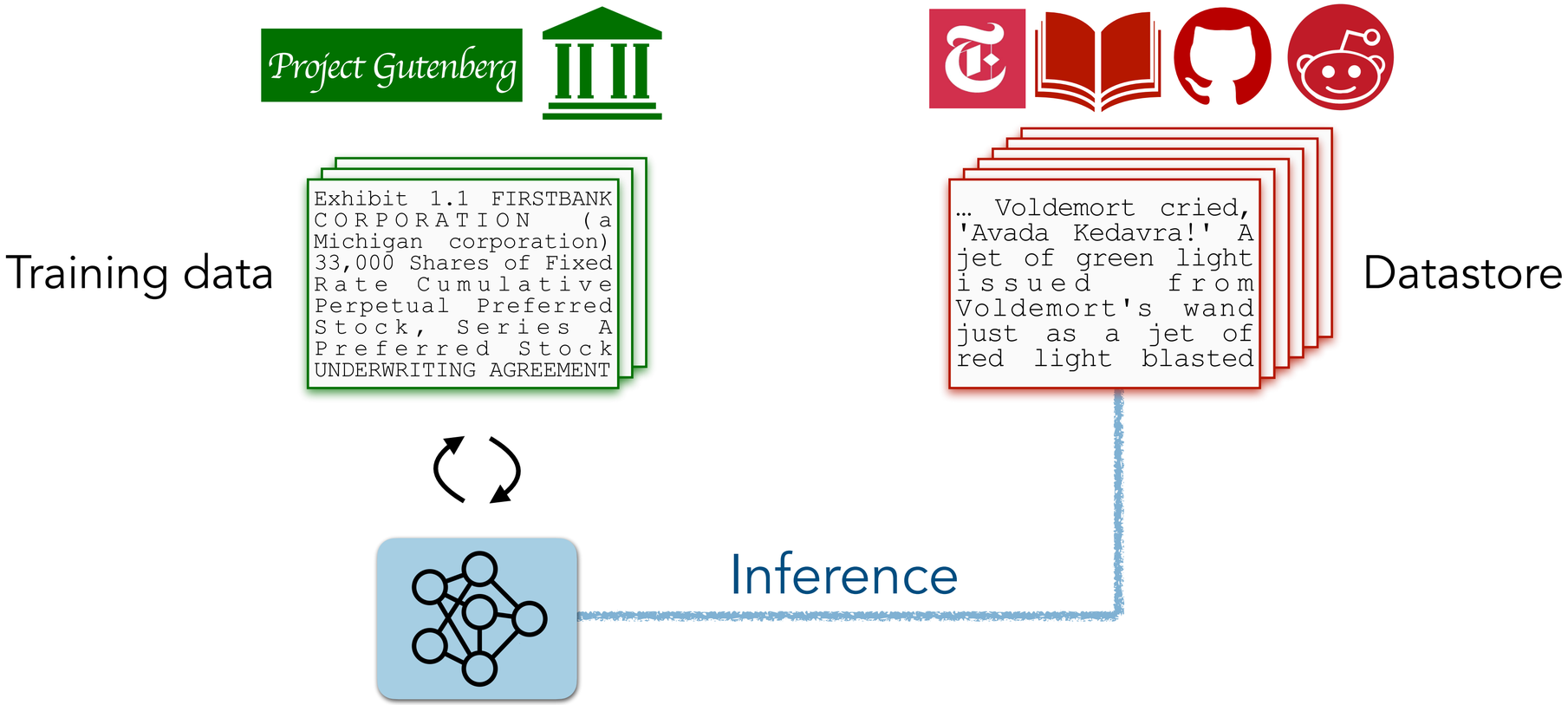
Our proposal: A nonparametric LM

Training data

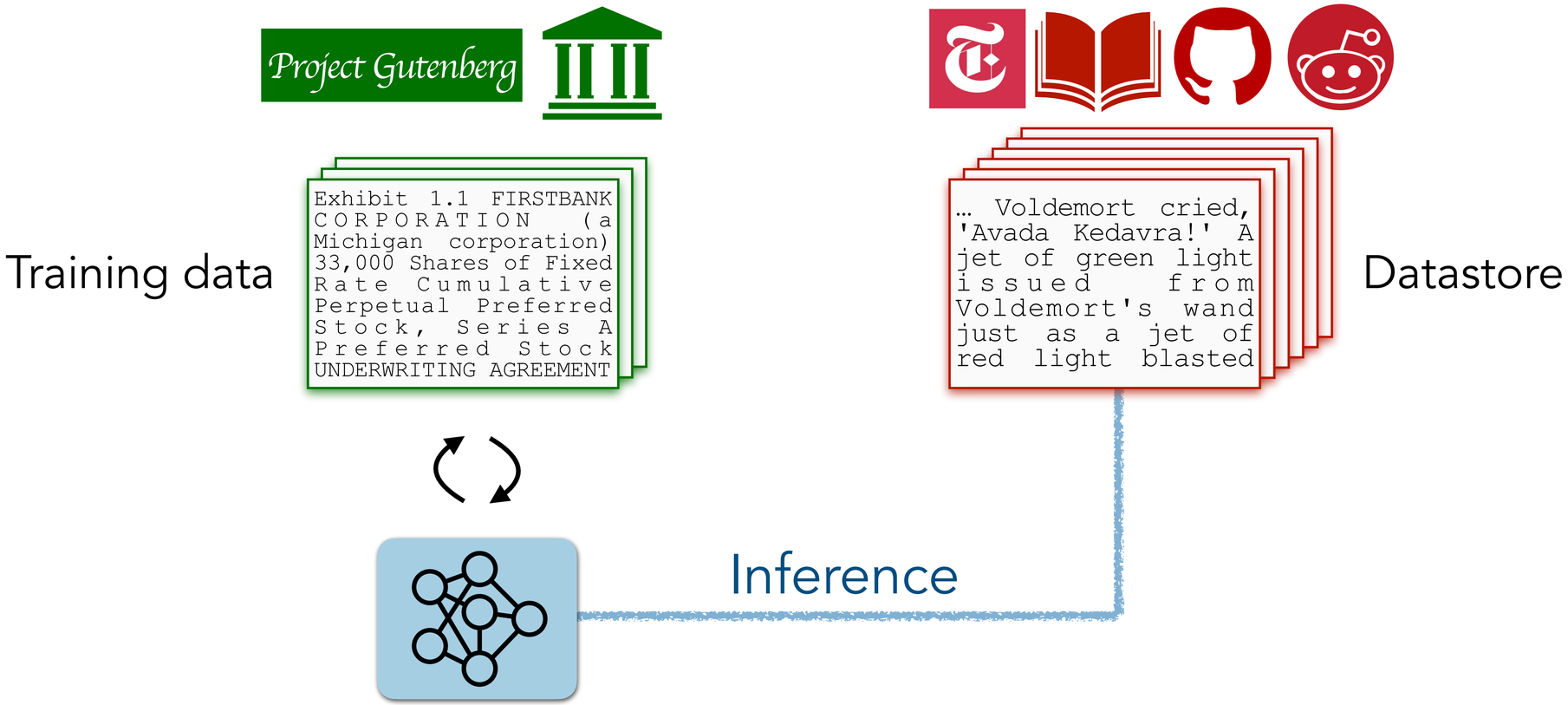
Very low legal risk, but poor performance

(We'll show you at this talk)

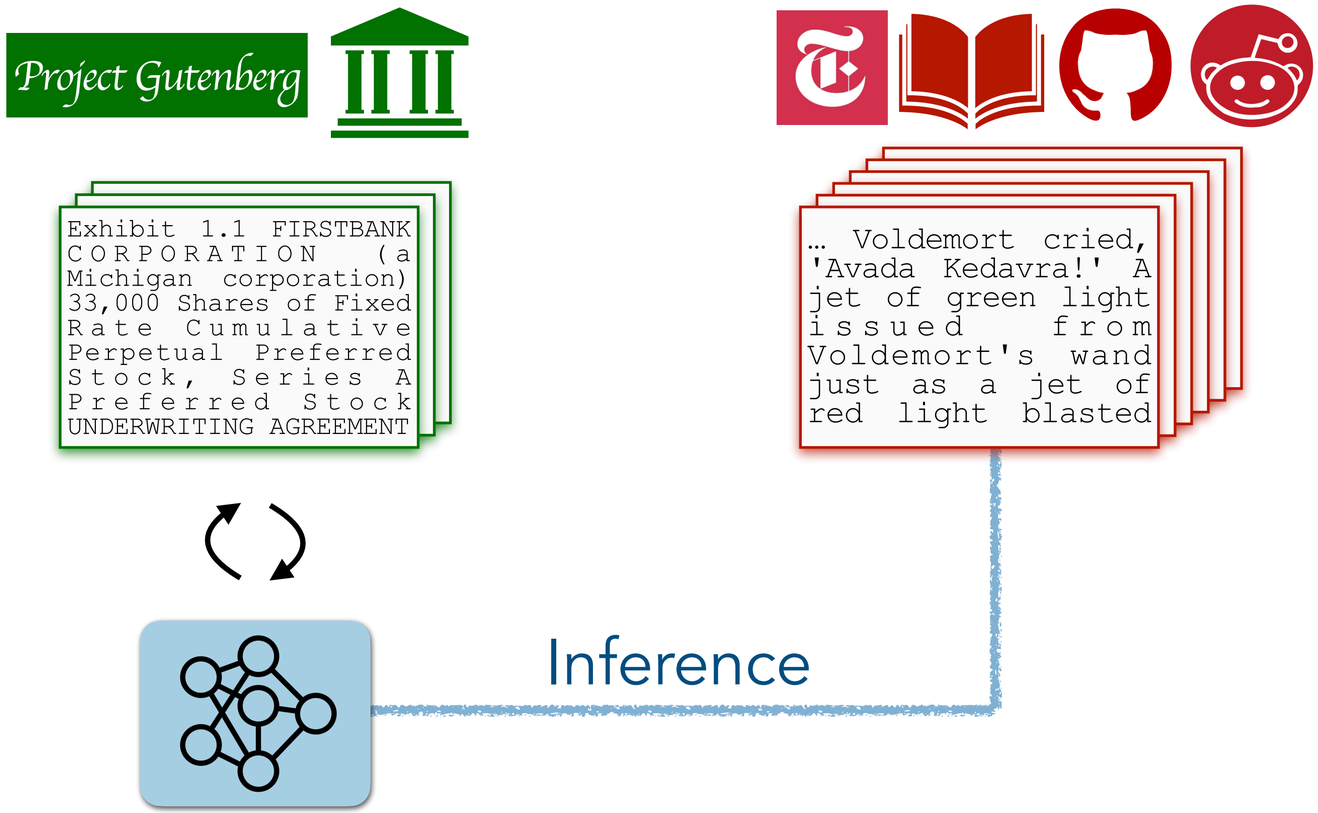
Our proposal: A nonparametric LM



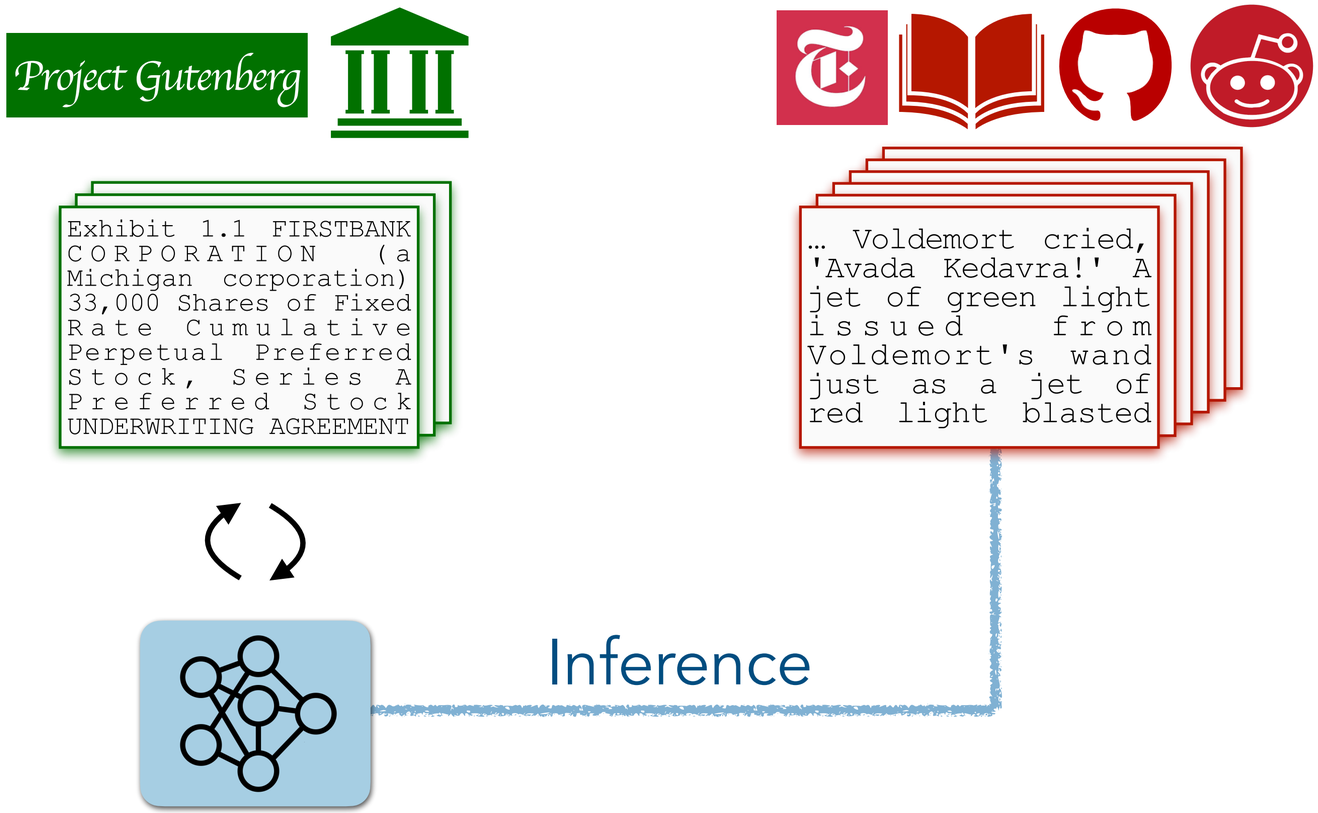
Our proposal: A nonparametric LM



Our proposal: A nonparametric LM

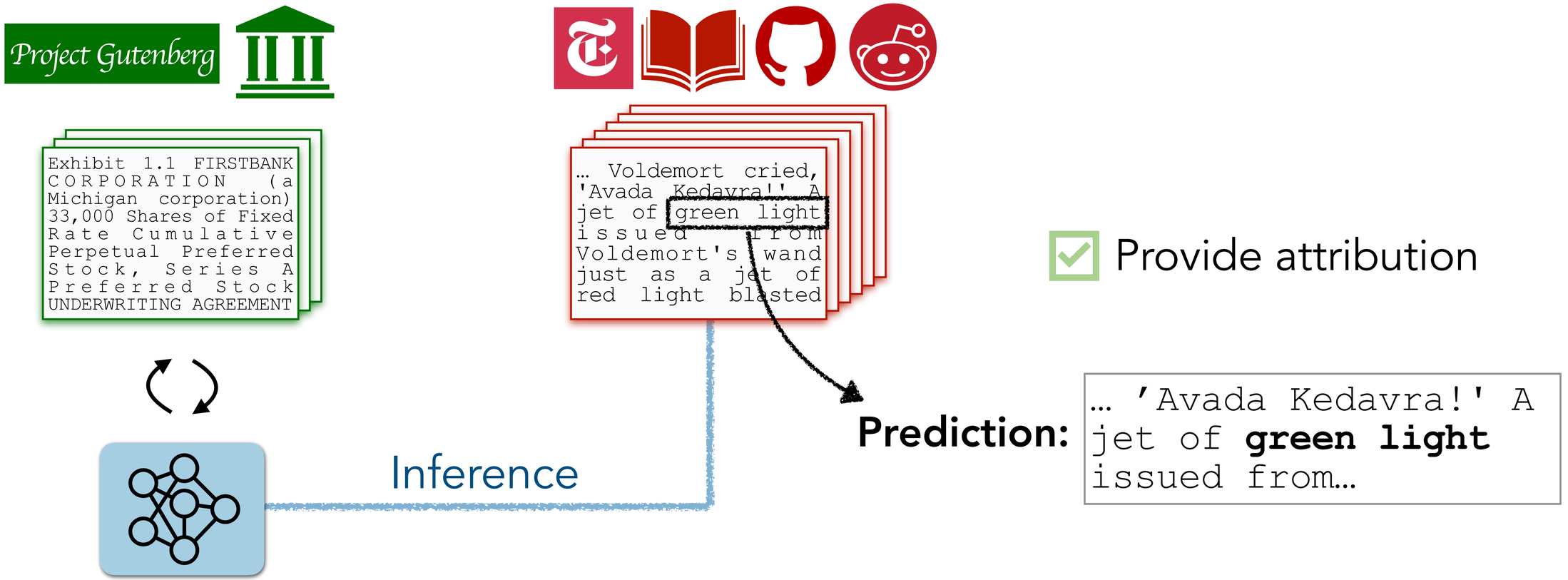


Our proposal: A nonparametric LM

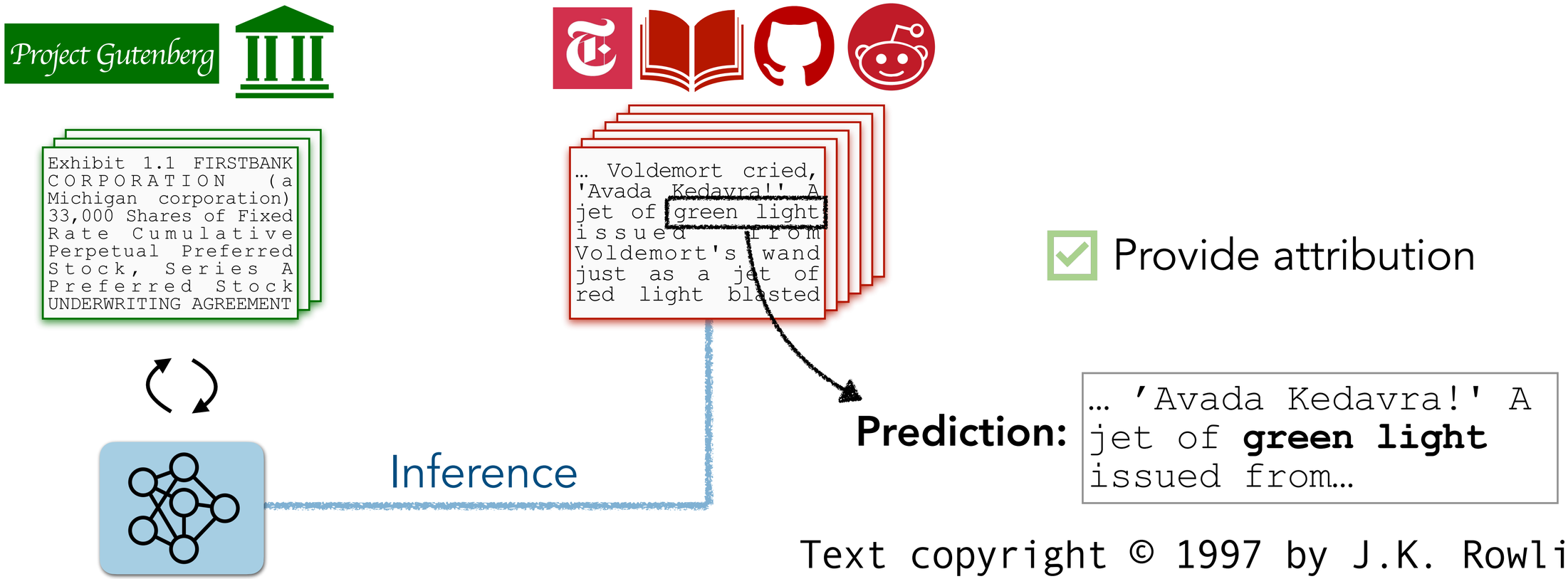


☑ Provide attribution

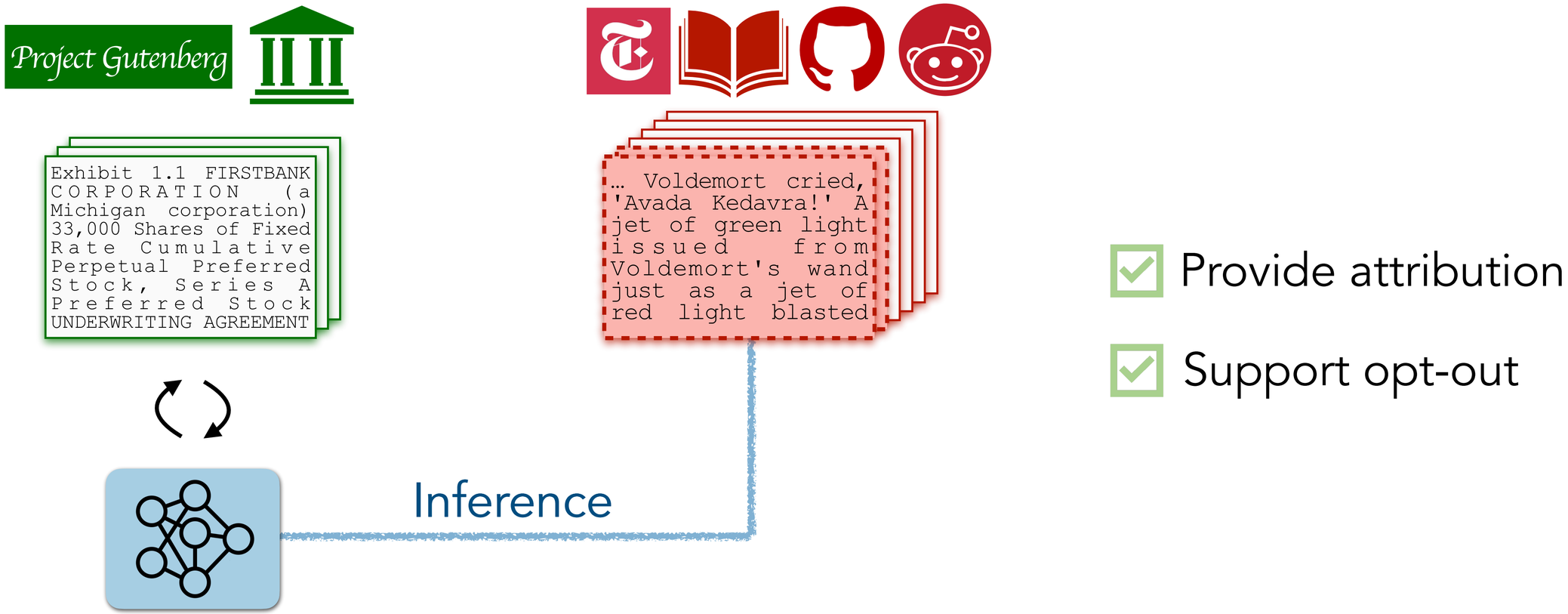
Our proposal: A nonparametric LM



Our proposal: A nonparametric LM

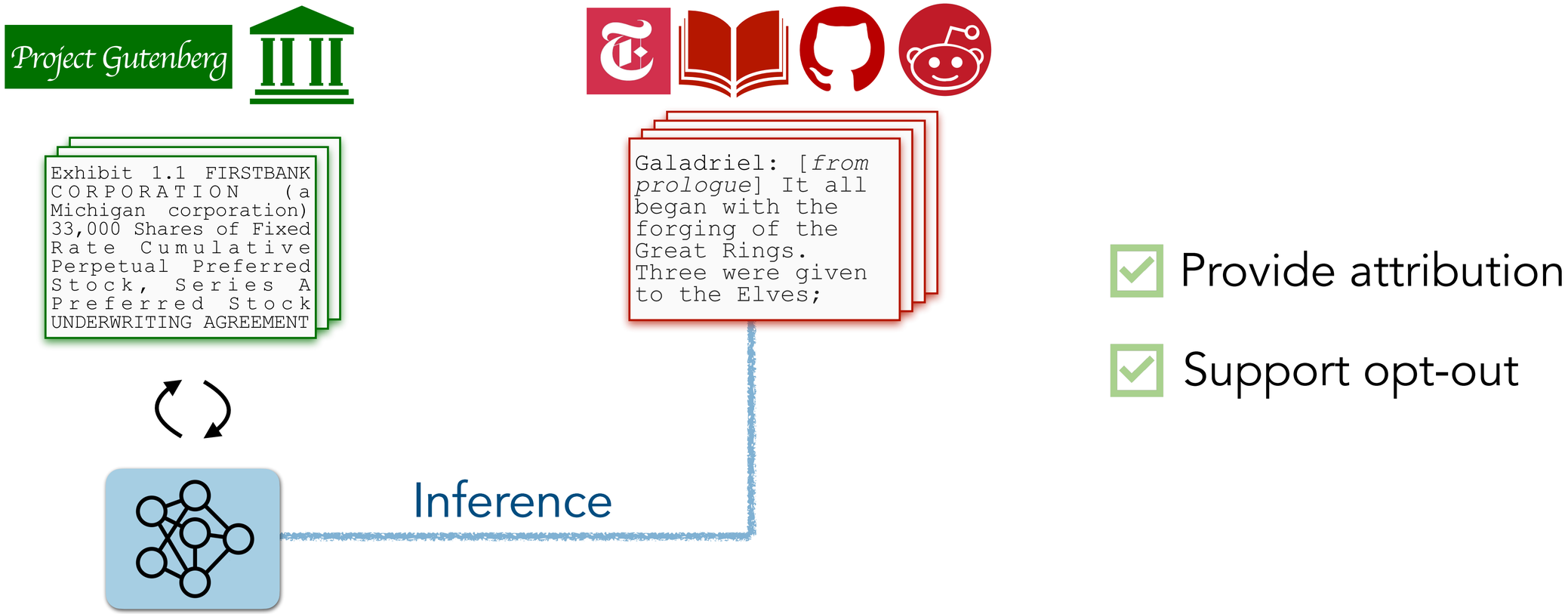


Our proposal: A nonparametric LM



- ☑ Provide attribution
- ☑ Support opt-out

Our proposal: A nonparametric LM



- ☑ Provide attribution
- ☑ Support opt-out

Our proposal: A nonparametric LM

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

I want to delete my private information.

I need to delete user data to comply with GDPR.

Our proposal: A nonparametric LM

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

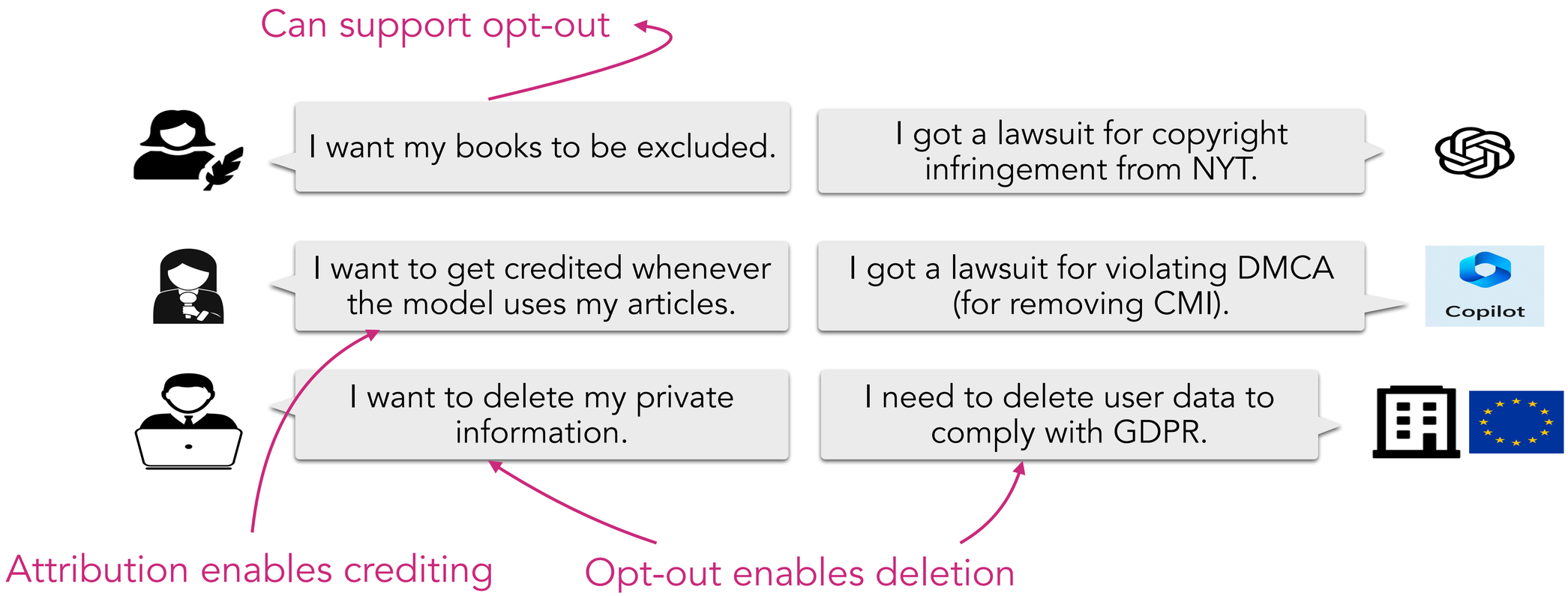
I got a lawsuit for violating DMCA (for removing CMI).

I want to delete my private information.

I need to delete user data to comply with GDPR.

Attribution enables crediting

Our proposal: A nonparametric LM



Our proposal: A nonparametric LM

Can support opt-out

Attribution strengthens fair use defense,
Opt-out can also provide defense

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

Copilot

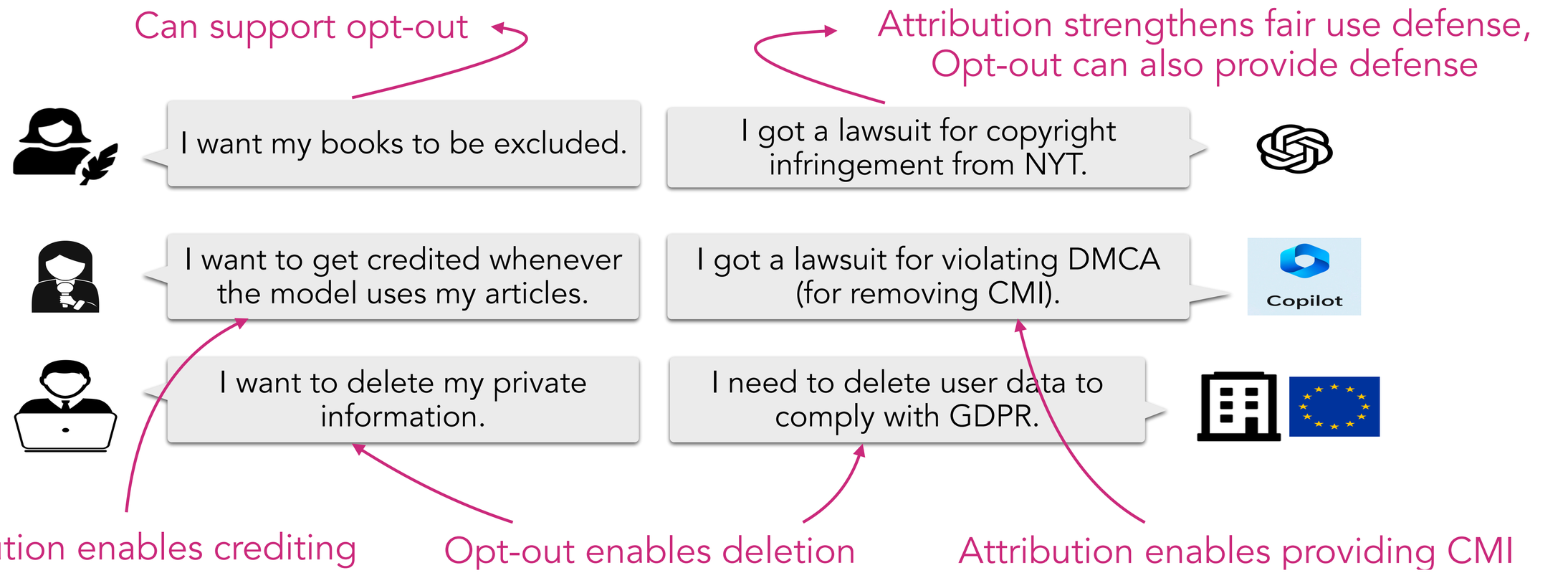
I want to delete my private information.

I need to delete user data to comply with GDPR.

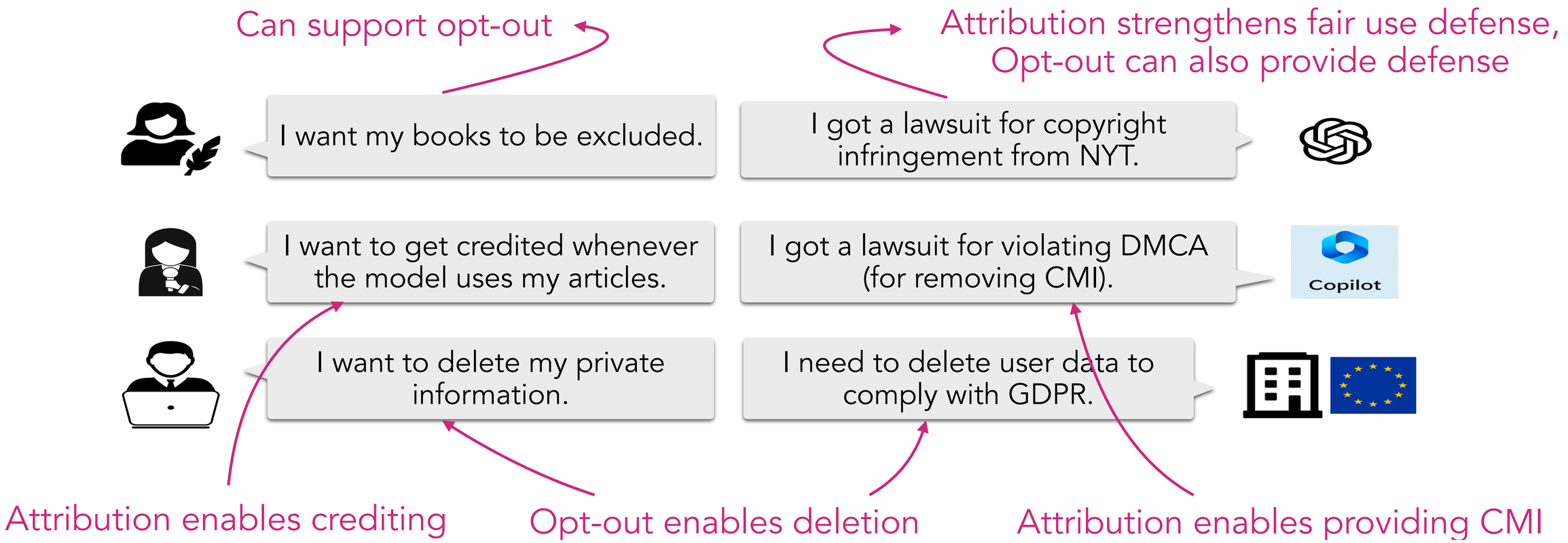
Attribution enables crediting

Opt-out enables deletion

Our proposal: A nonparametric LM

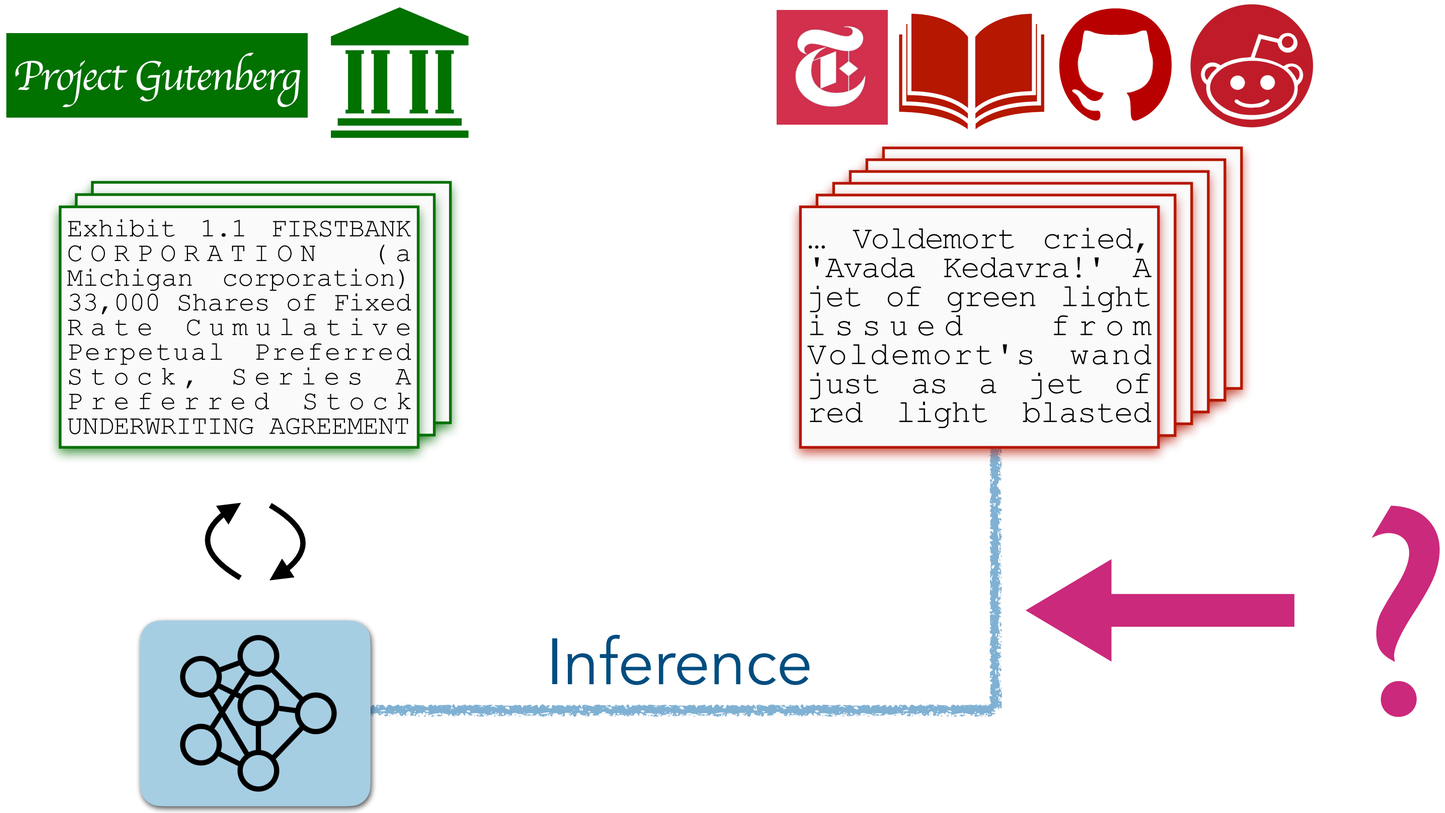


Our proposal: A nonparametric LM

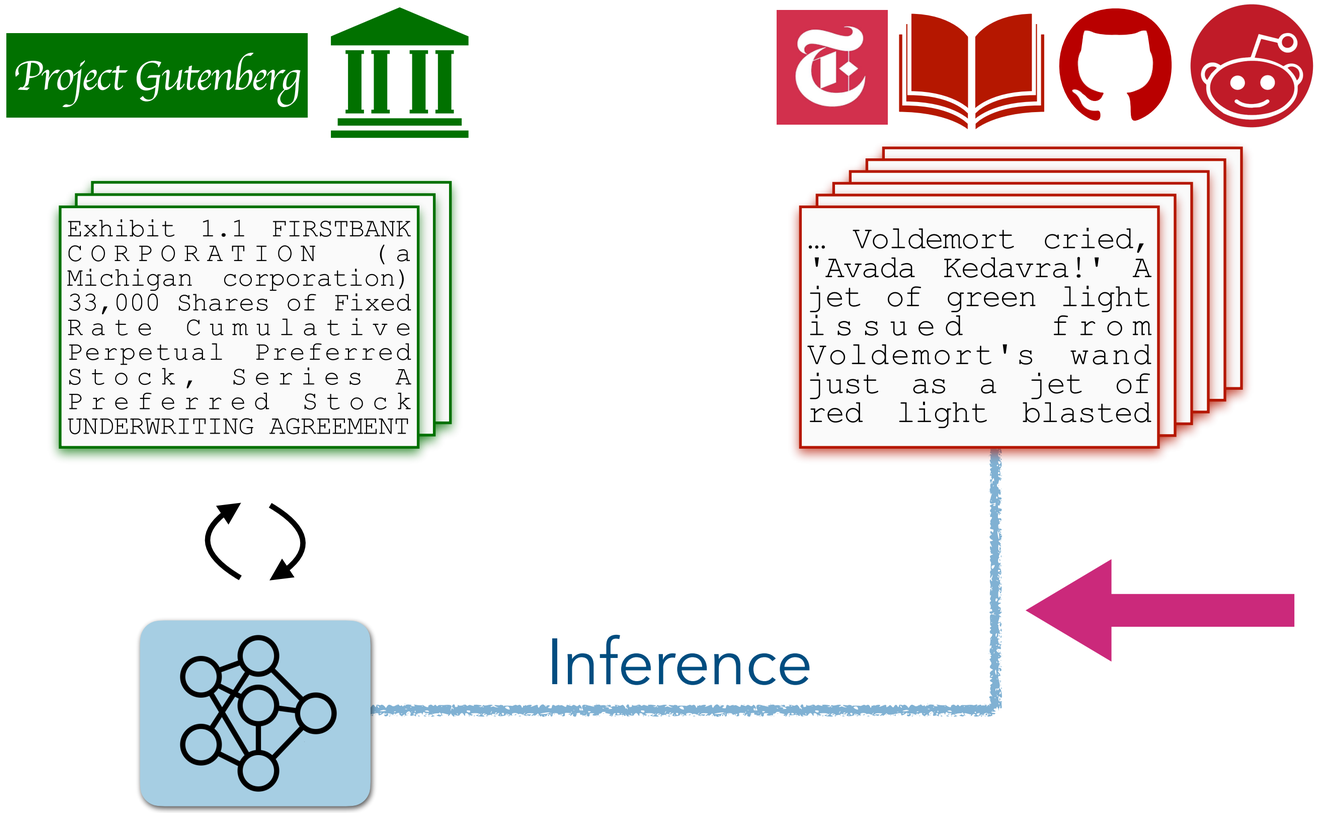


Significantly improve generalization (we'll show you soon!)

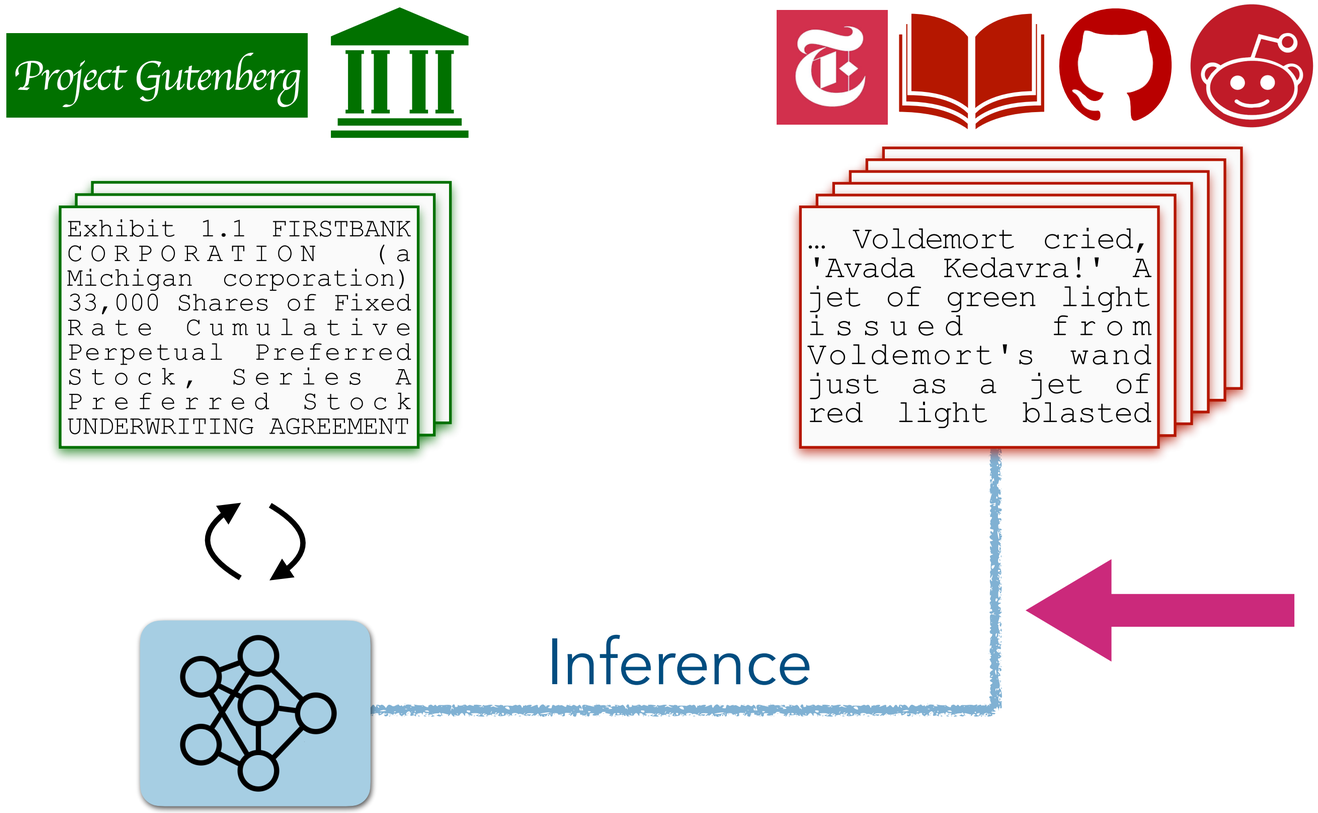
How does LM use a datastore?



Use a *retrieval* approach



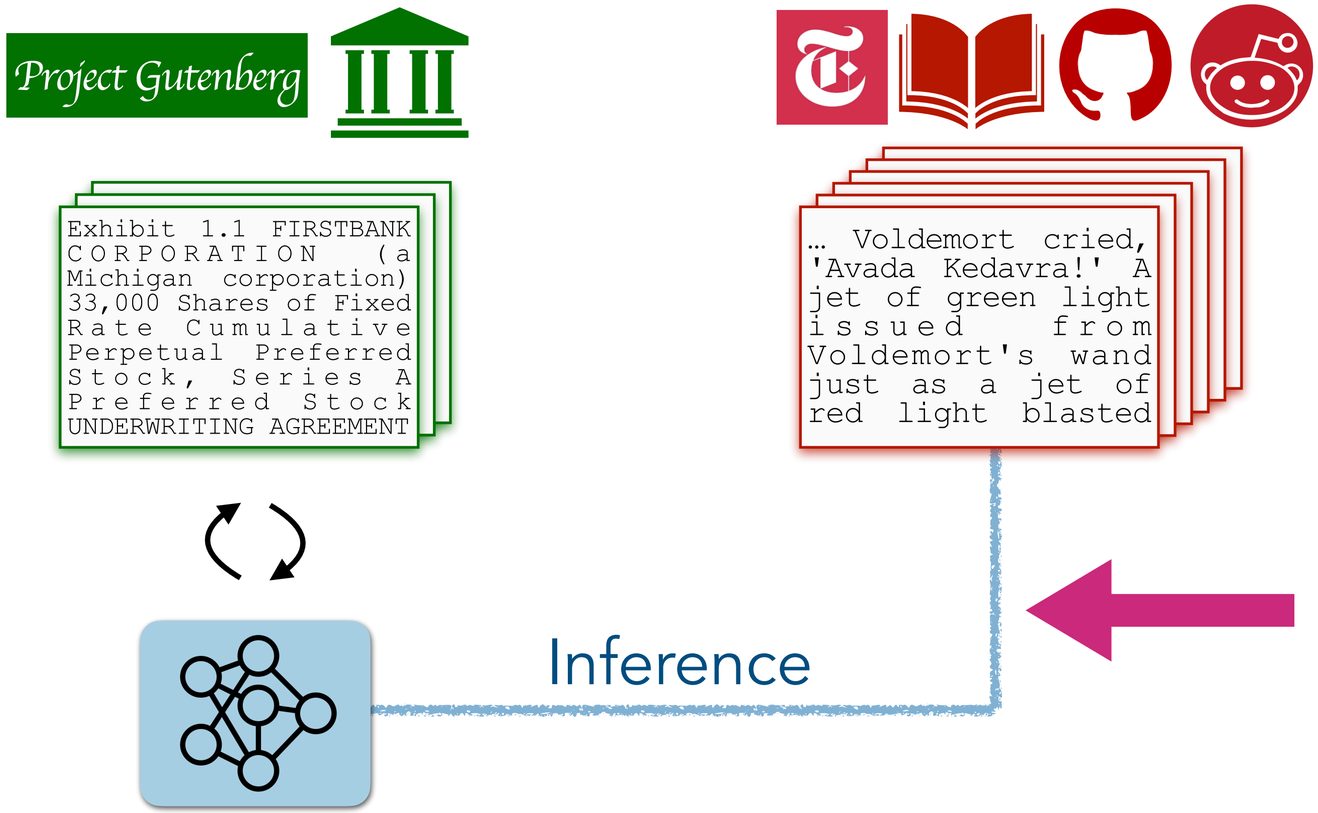
Use a *retrieval* approach



1. **Retrieve-in-context LM**

(Shi et al 2023, Ram et al 2023)

Use a *retrieval* approach



1. **Retrieve-in-context LM**
   (Shi et al 2023, Ram et al 2023)
2. ***k*NN-LM** (Khandelwal et al. 2020)

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

Retrieval model

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

Retrieval model

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"

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

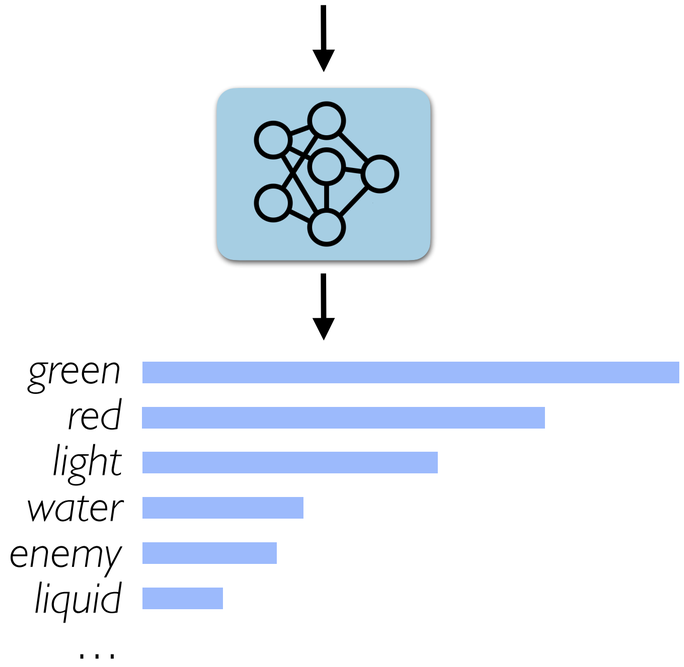
Retrieval model

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

Encoder

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 → Encoder →

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

"The Boy Who Lived." He … **green** → Encoder →

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

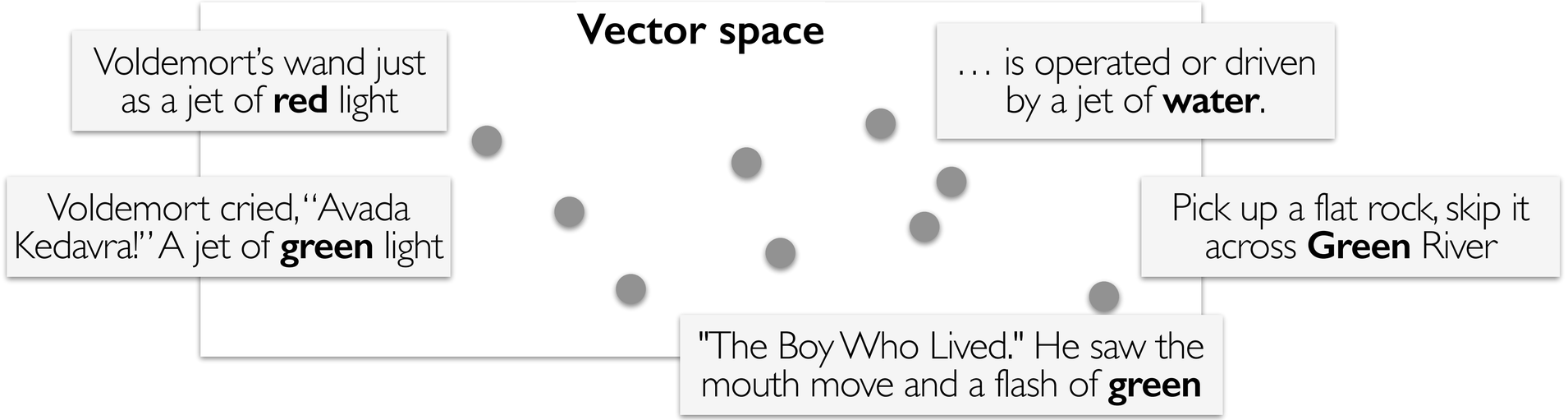
Pick up a flat rock, skip it across **Green** River → Encoder →

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.

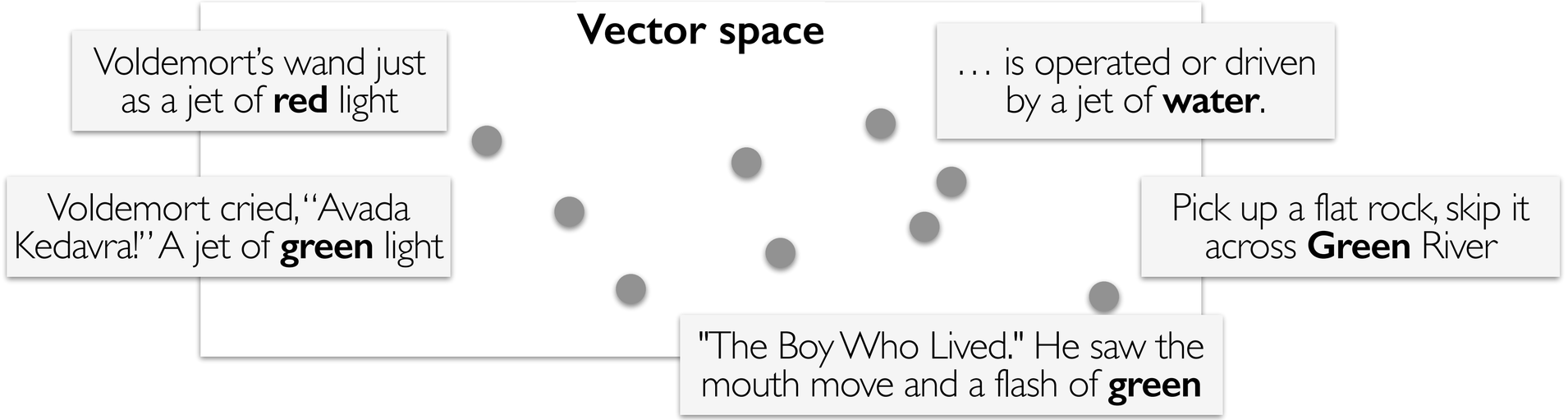
2. *k*NN-LM

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.

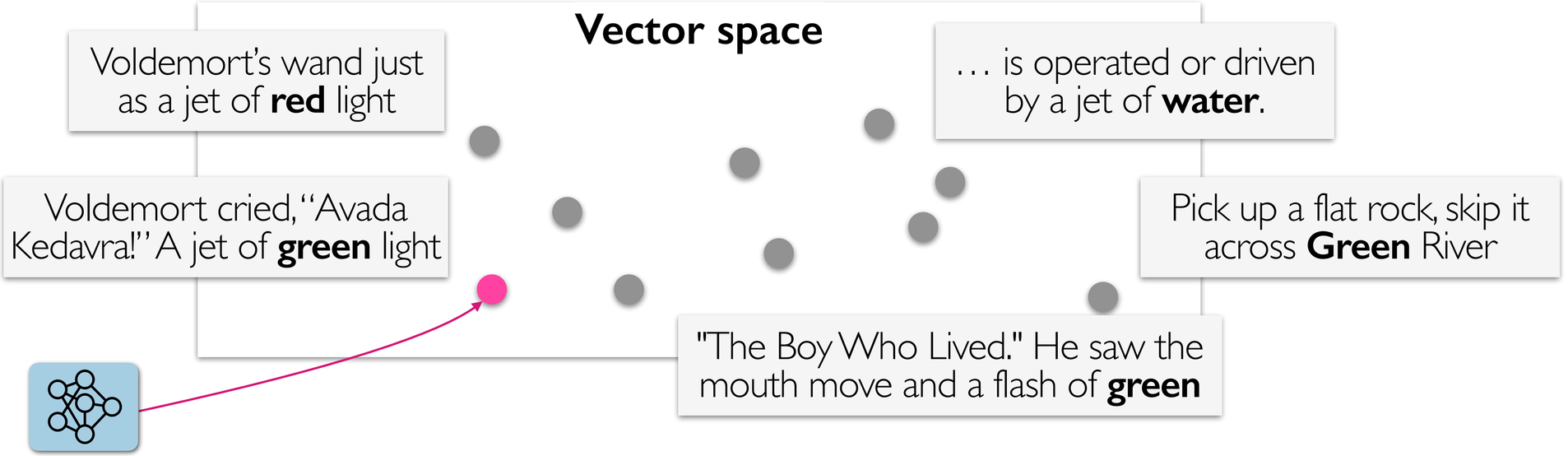
2. *k*NN-LM

Datastore:

©

Test input:

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



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

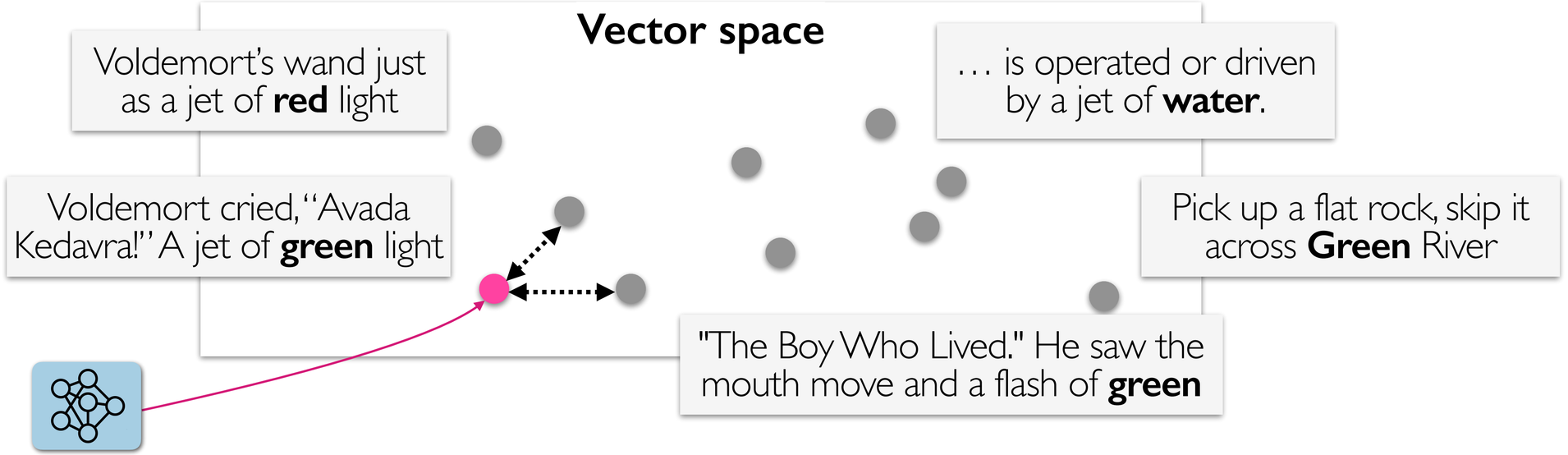
2. *k*NN-LM

Datastore:

©

Test input:

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



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.

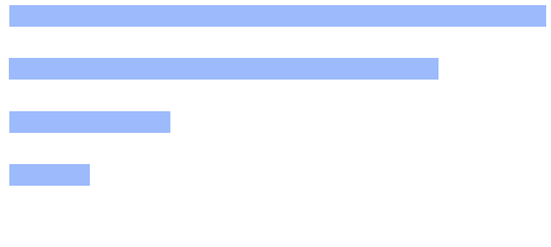
2. *k*NN-LM

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**.
…

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

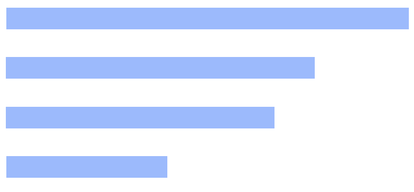


Test input:

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

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

water
light
green
liquid
…



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

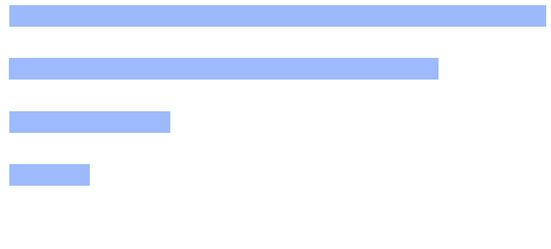
2. *k*NN-LM

Datastore:

©

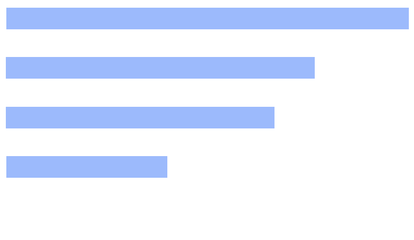
Voldemort cried, "Avada Kedavra!" A jet of **green**
"The Boy Who Lived." He saw the mouth move and a flash of **green**
Voldemort's wand just as a jet of **red**
… is operated or driven by a jet of **water**.
…

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



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

water
light
green
liquid
…



Test input:

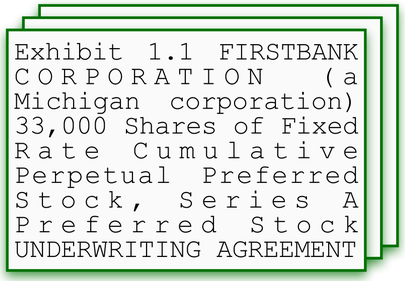
Harry felt Greenback collapse against him … on the floor as a jet of

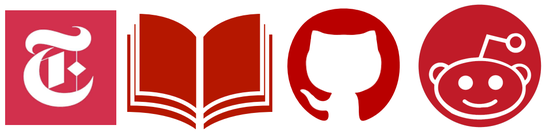
$$P_{k\text{NN-LM}}(y \mid x) = (1 - \lambda)P_{\text{LM}}(y \mid x) + \lambda P_{k\text{NN}}(y \mid x)$$

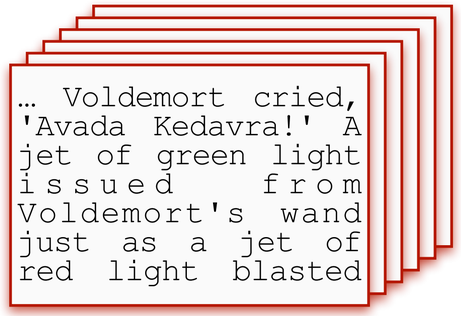
λ: hyperparameter

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

SILO: Summary

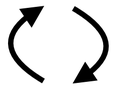


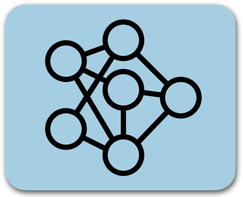


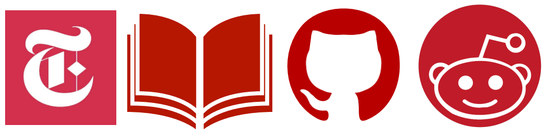


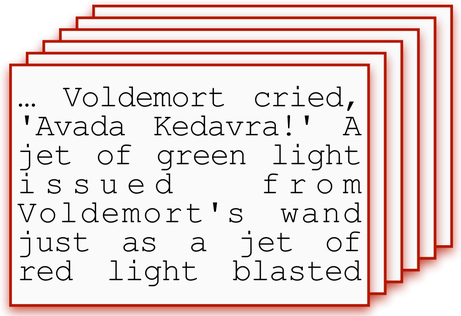
SILO: Summary

Training data

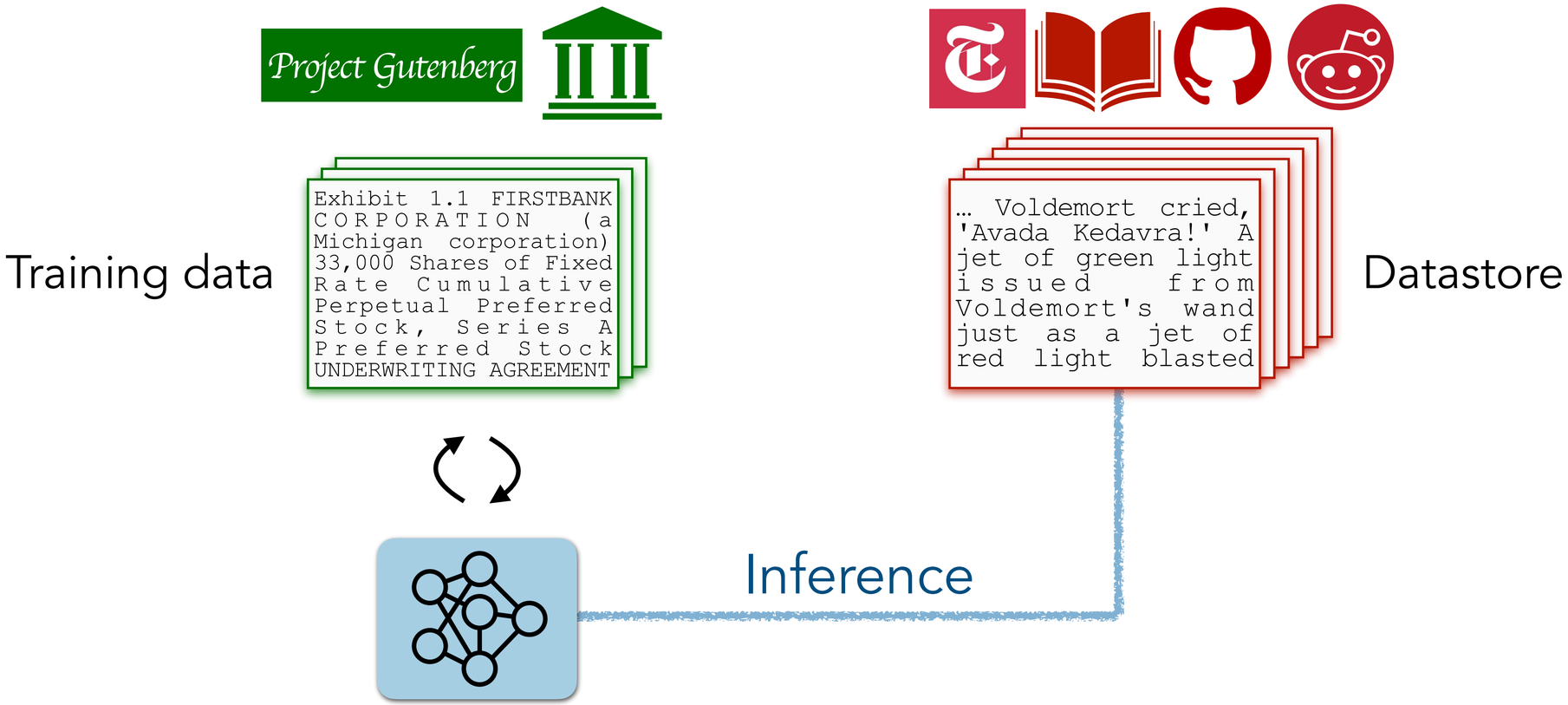




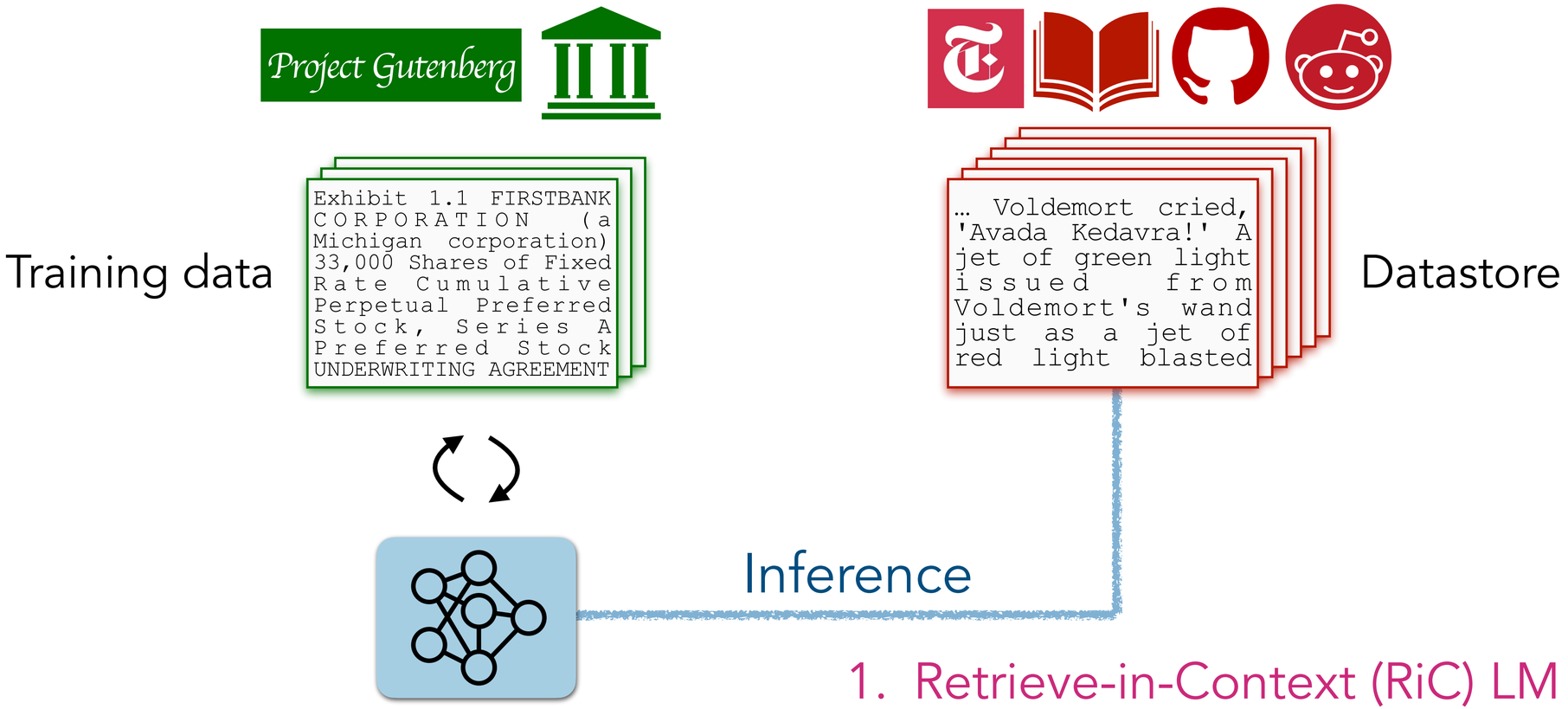




SILO: Summary



SILO: Summary



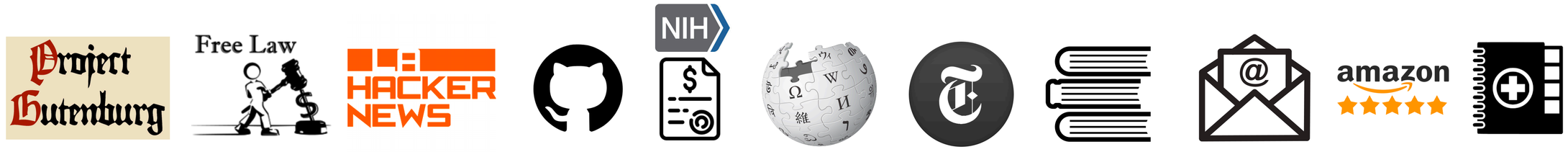
1. Retrieve-in-Context (RiC) LM
2. *k*NN-LM

Experiments

Evaluate Language Modeling Perplexity

Experiments

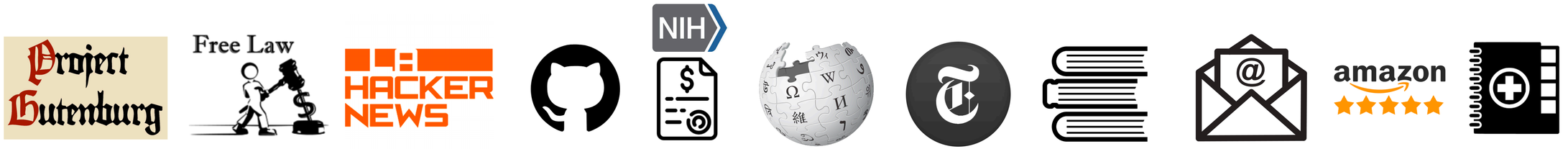
Evaluate Language Modeling Perplexity



In-distribution to OLC

Experiments

Evaluate Language Modeling Perplexity



In-distribution to OLC

Out-of-distribution

Experiments

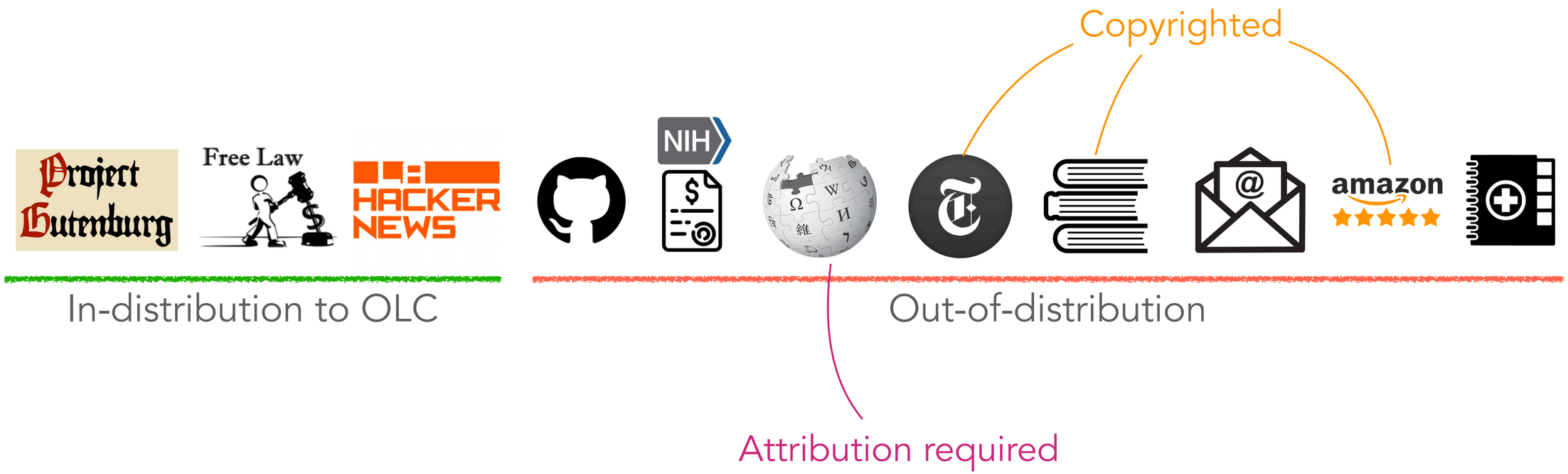
Evaluate Language Modeling Perplexity

In-distribution to OLC

Out-of-distribution

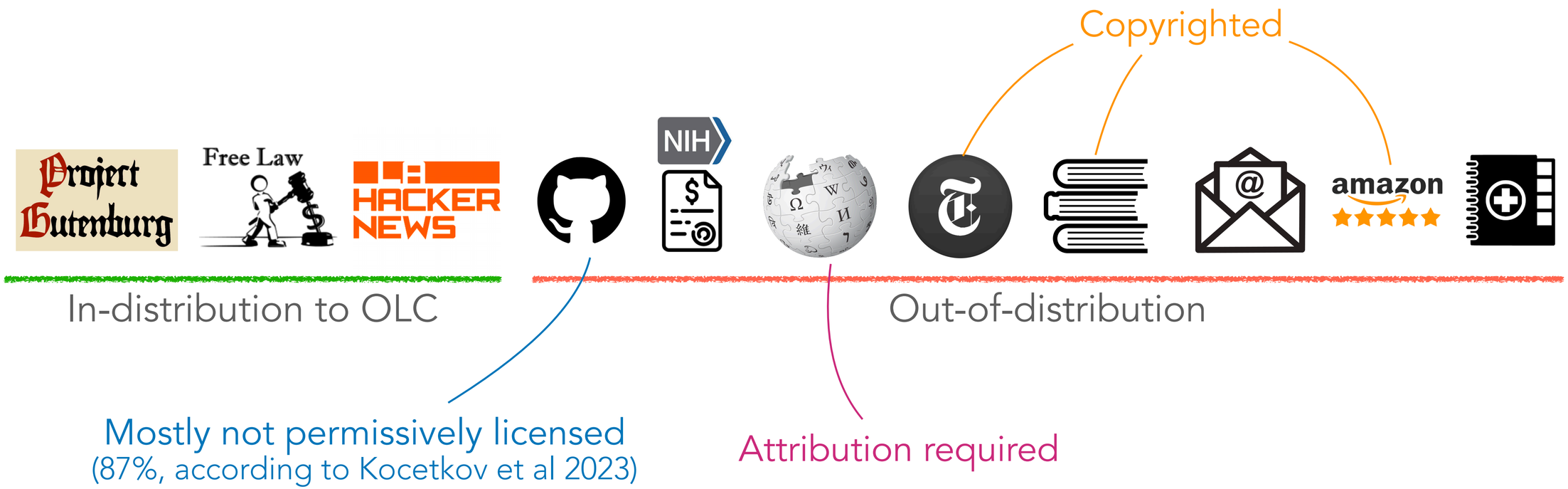
Experiments

Evaluate Language Modeling Perplexity



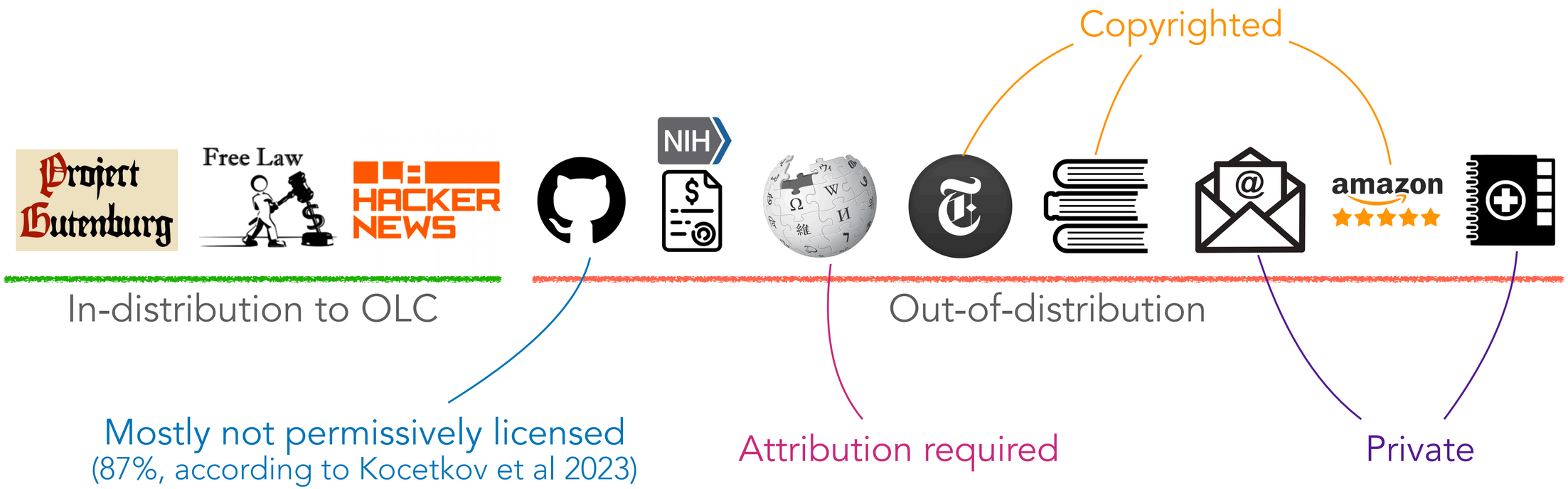
Experiments

Evaluate Language Modeling Perplexity

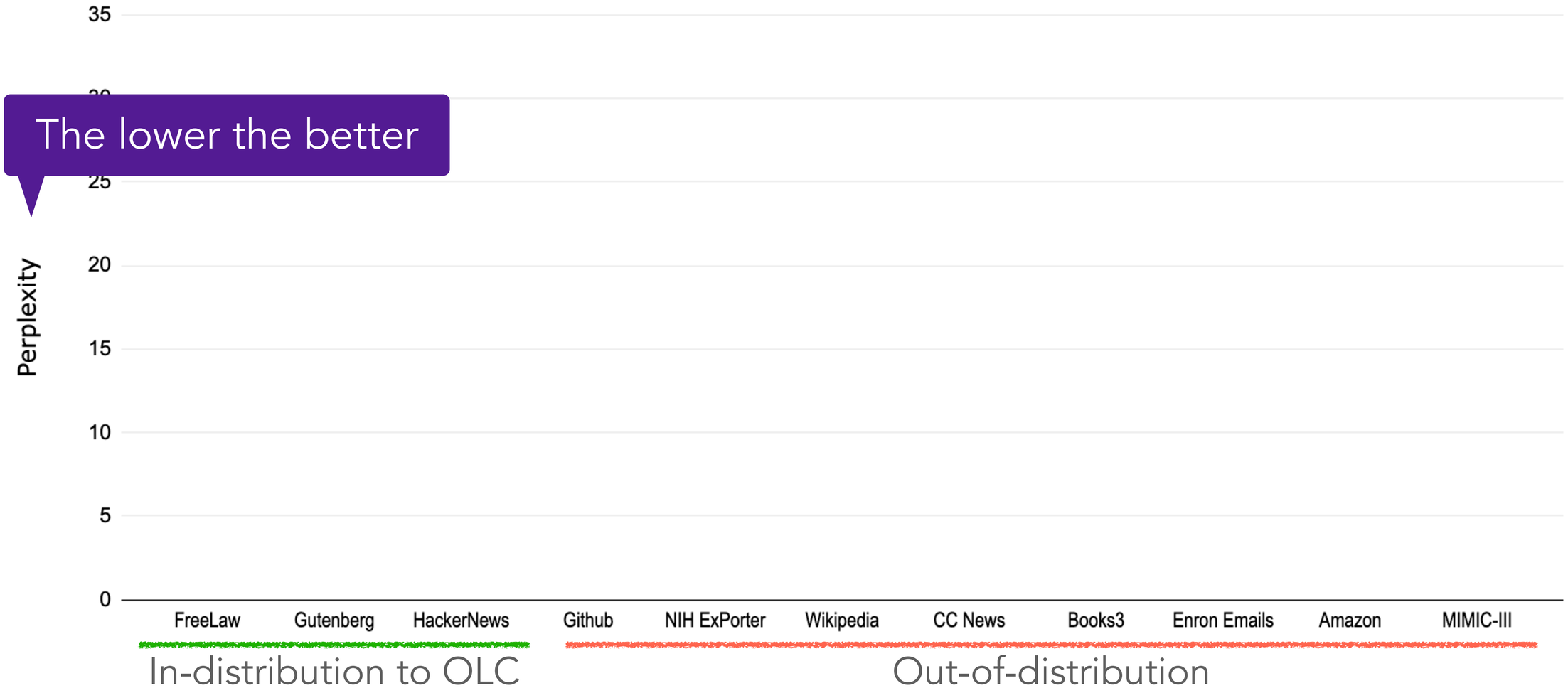


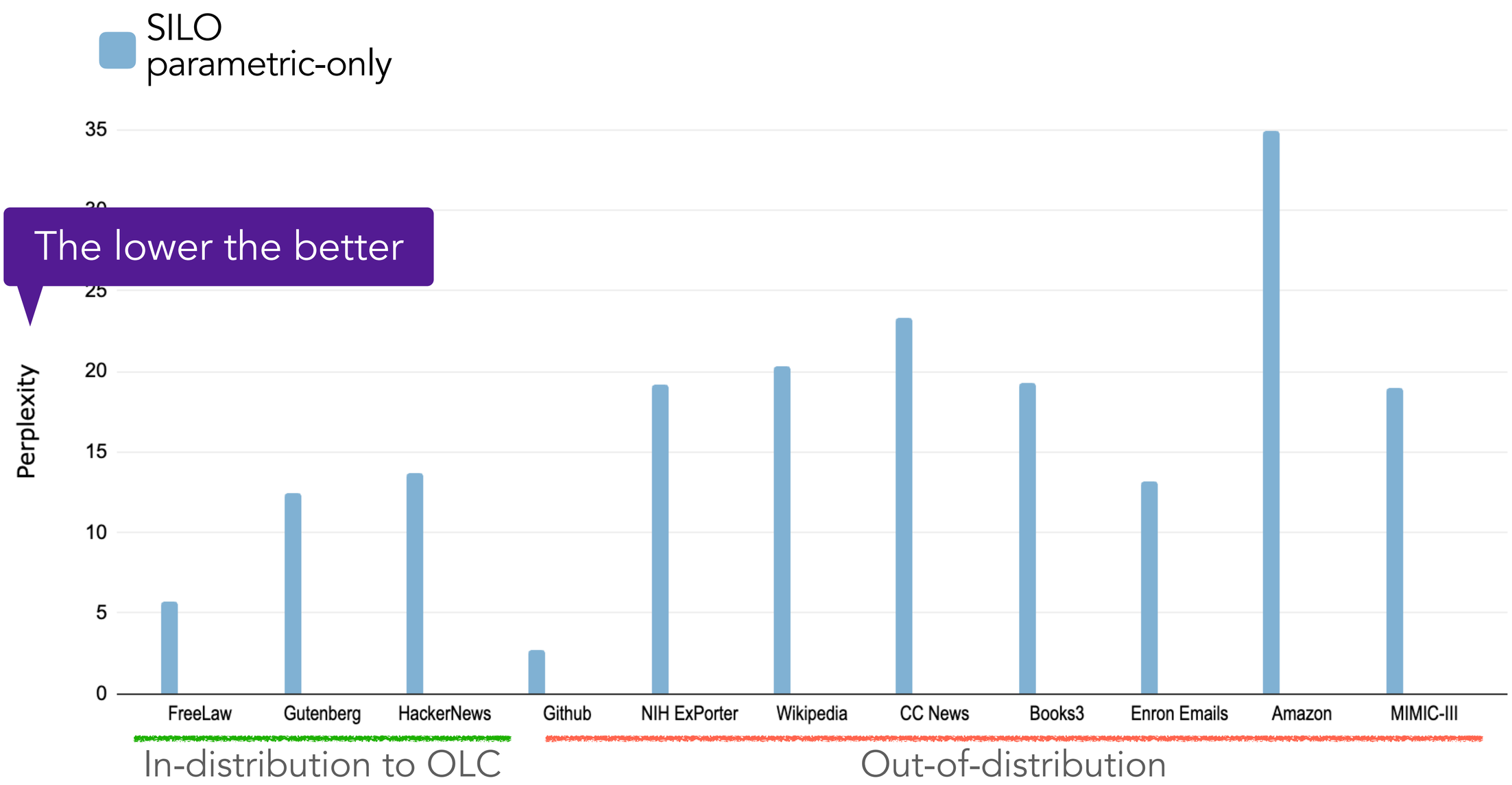
Experiments

Evaluate Language Modeling Perplexity



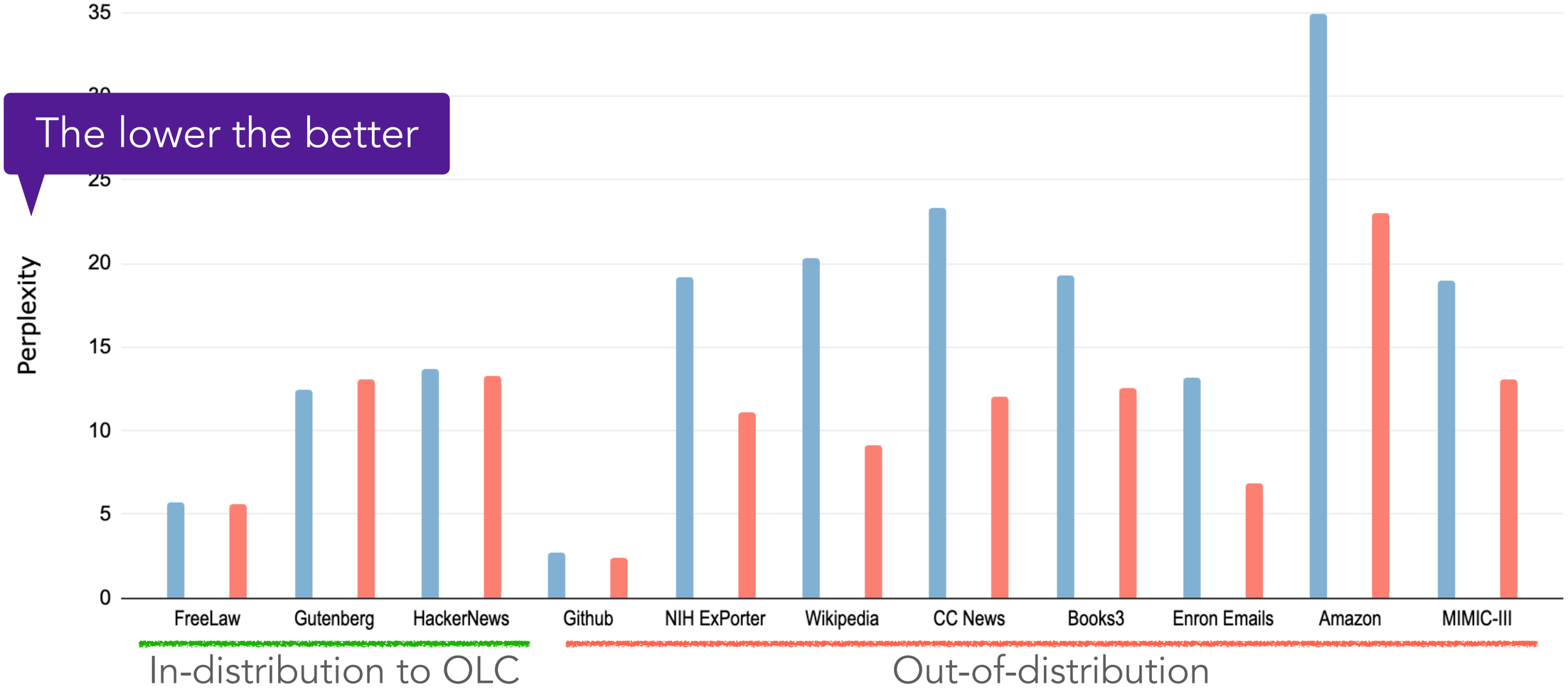
Experiments

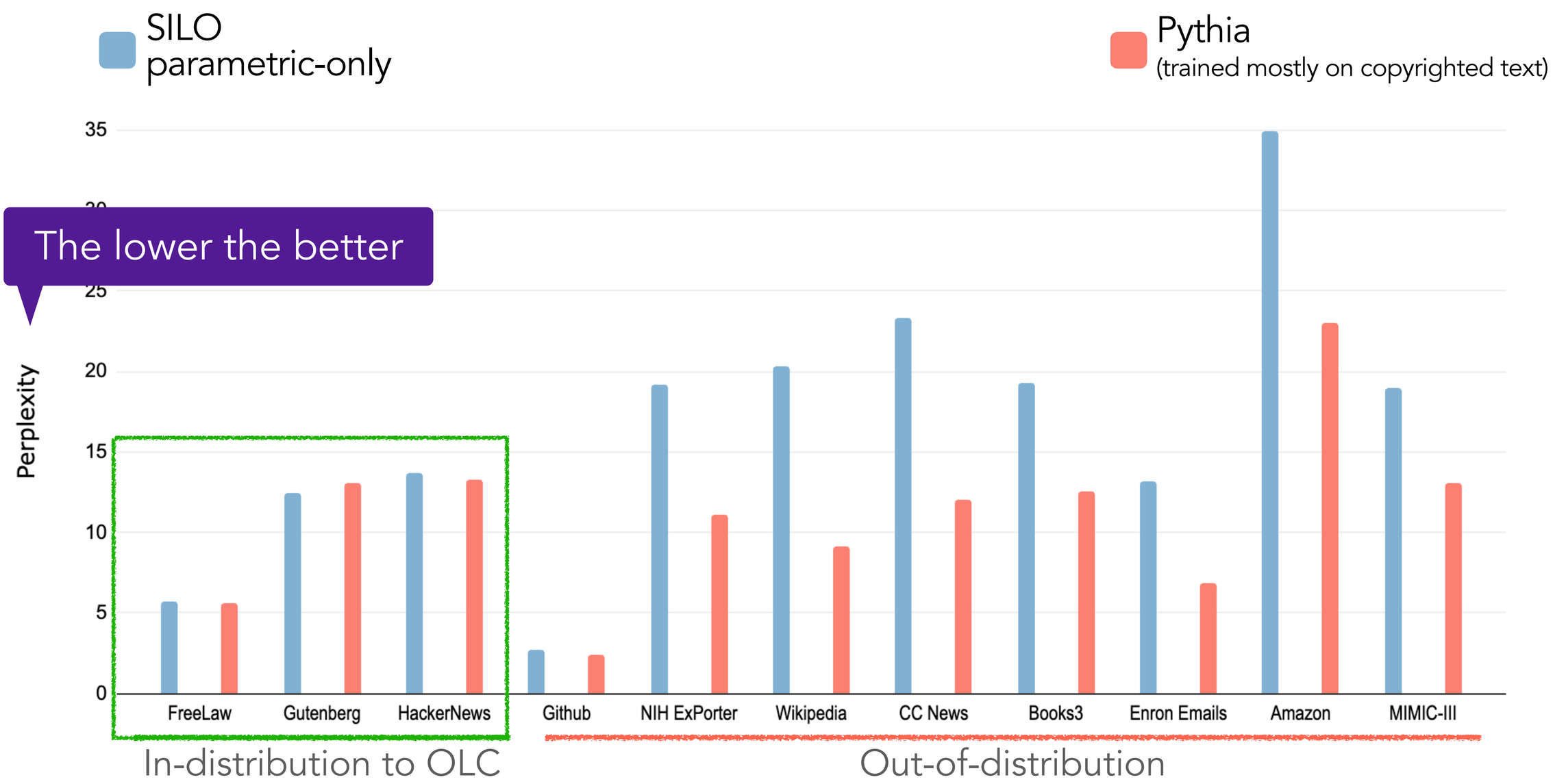


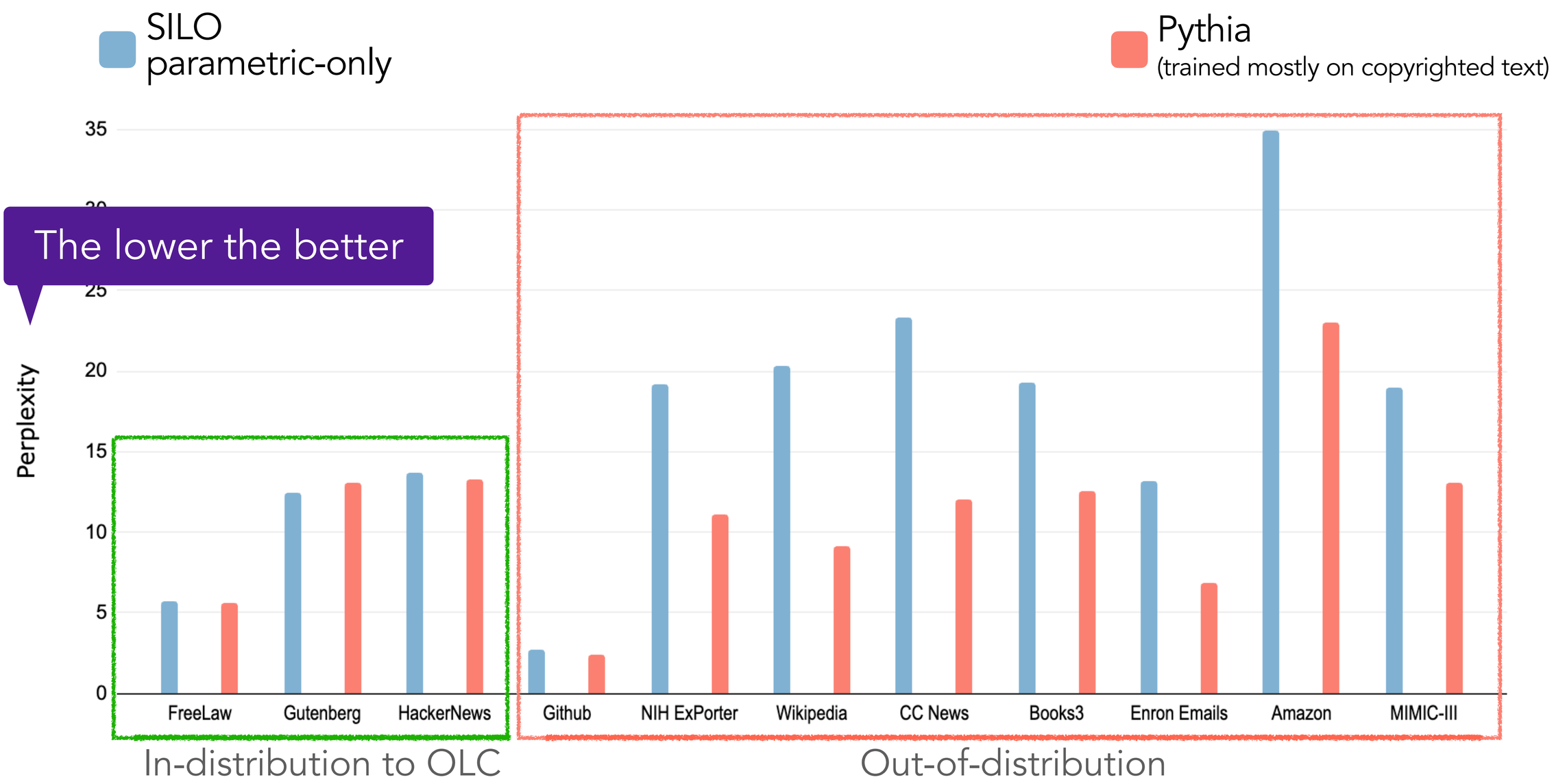

SILO
parametric-only
The lower the better
Perplexity
35
25
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution

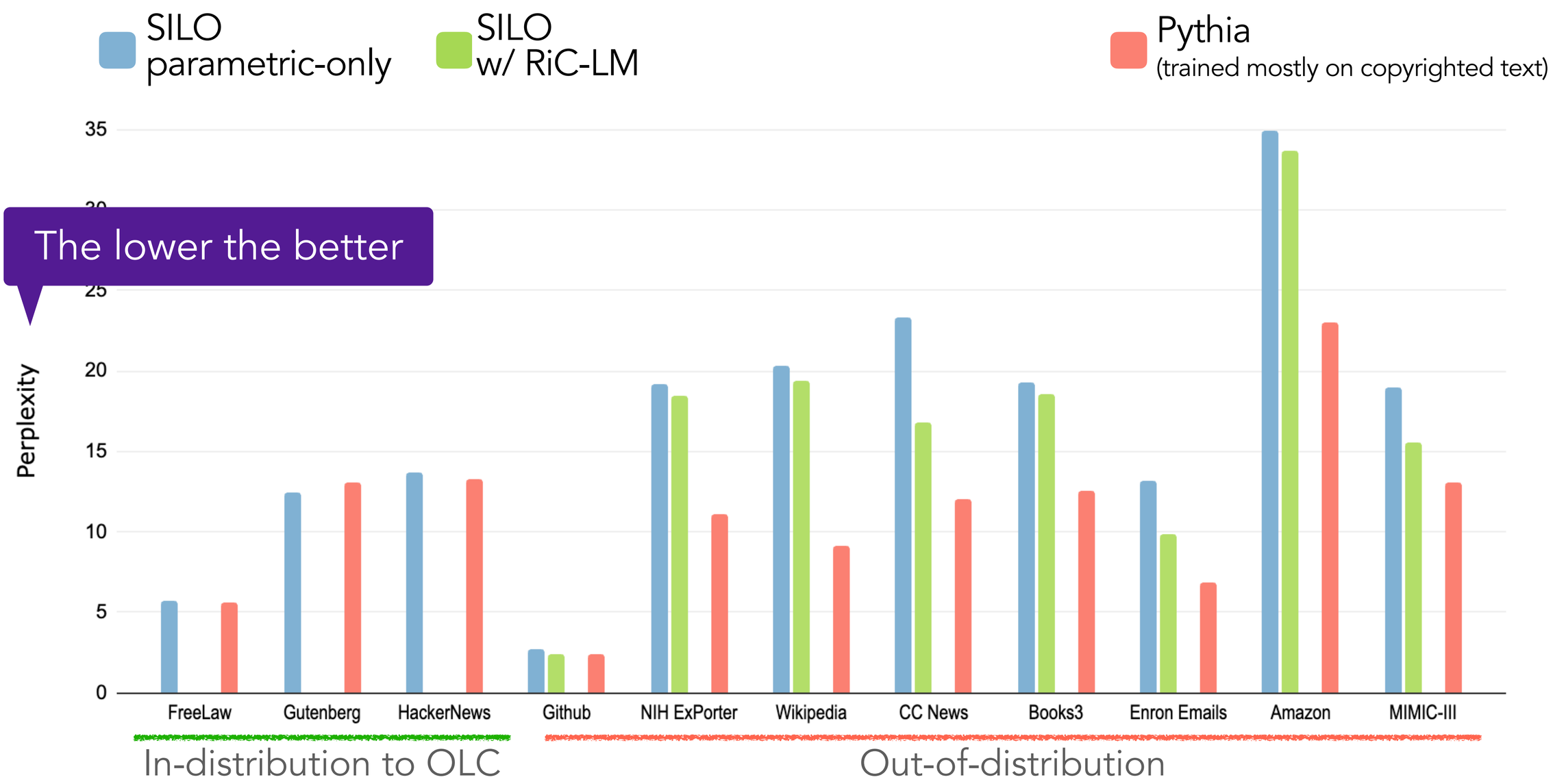
SILO
parametric-only

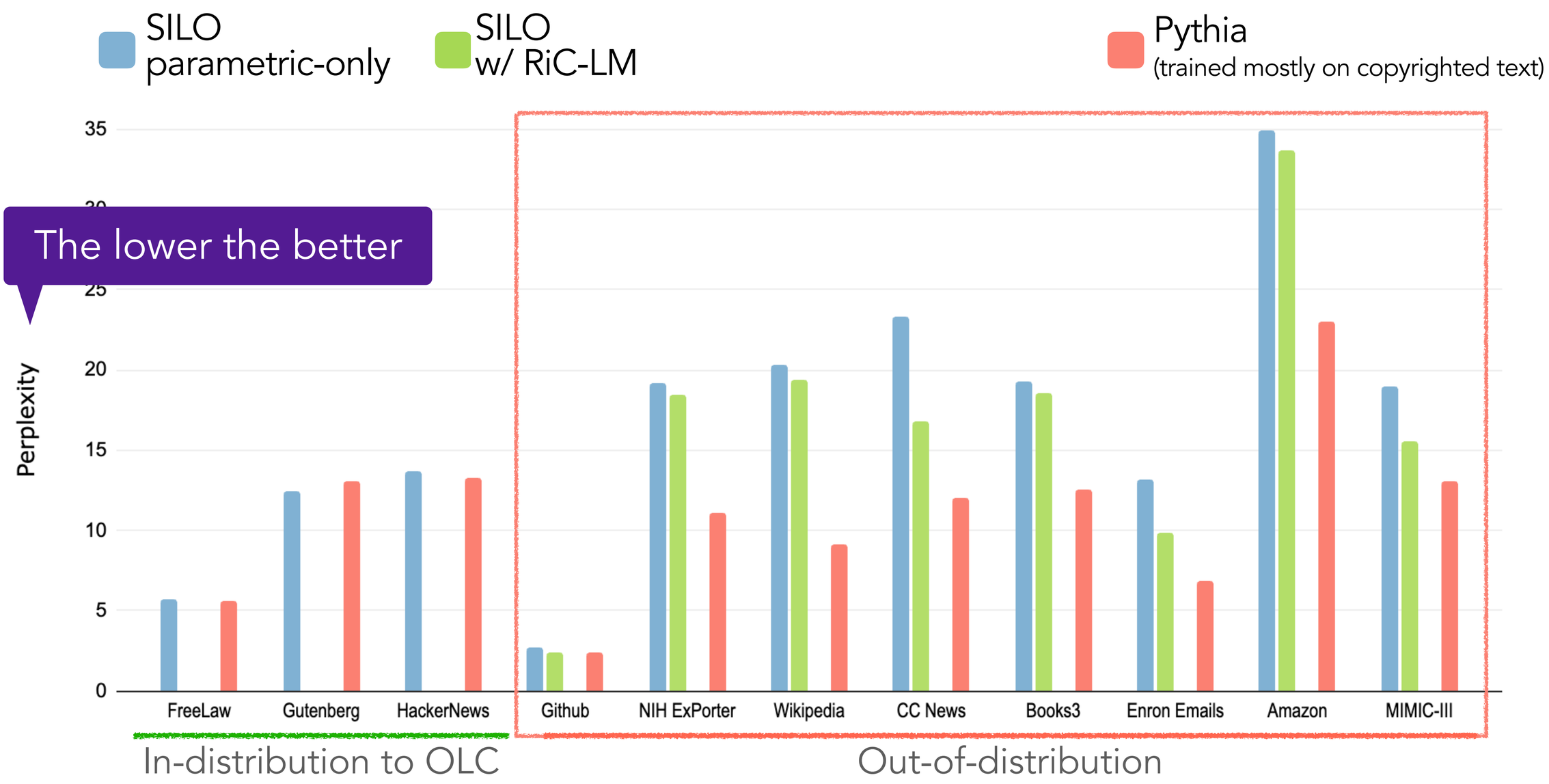
Pythia
(trained mostly on copyrighted text)

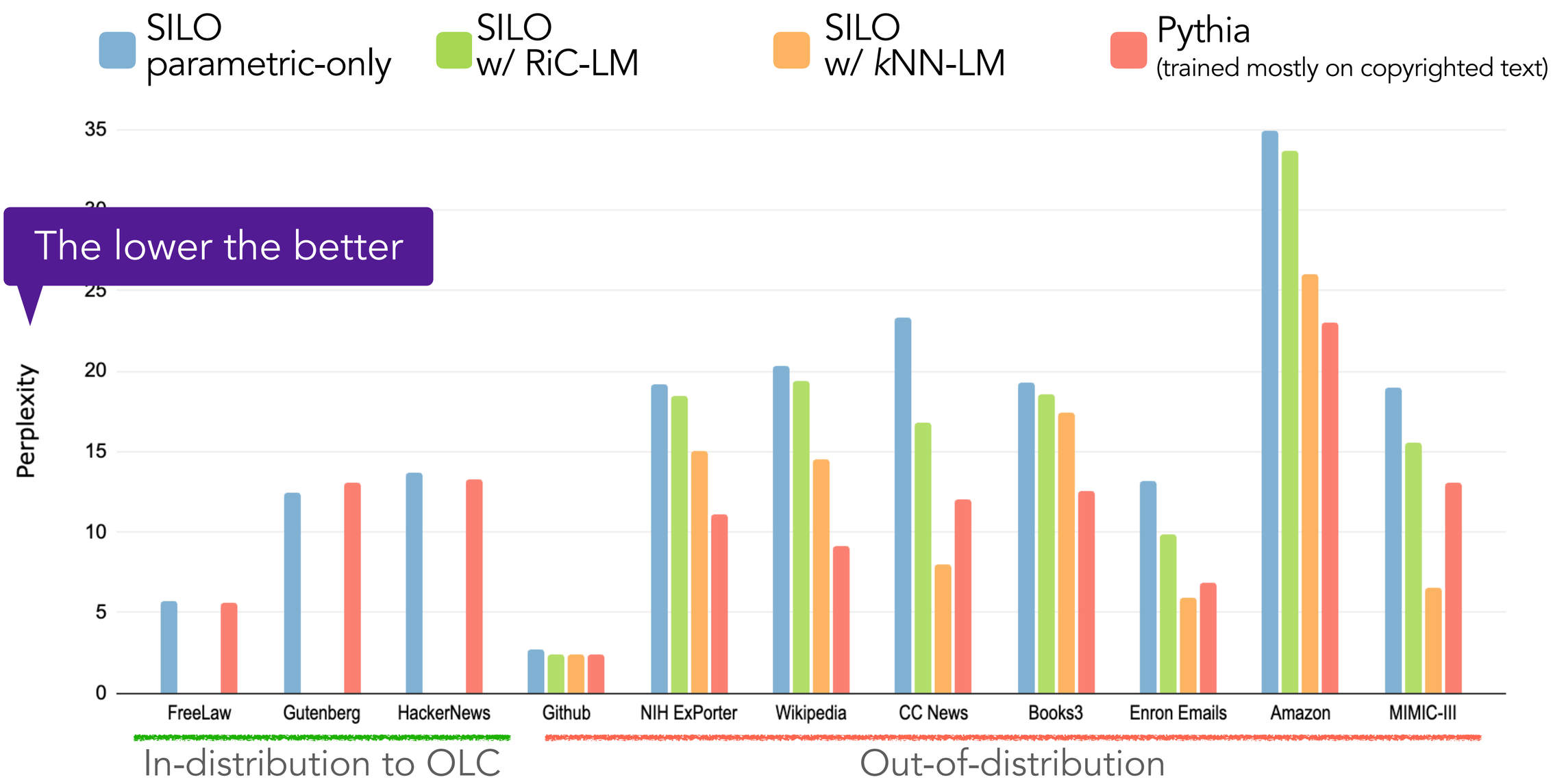


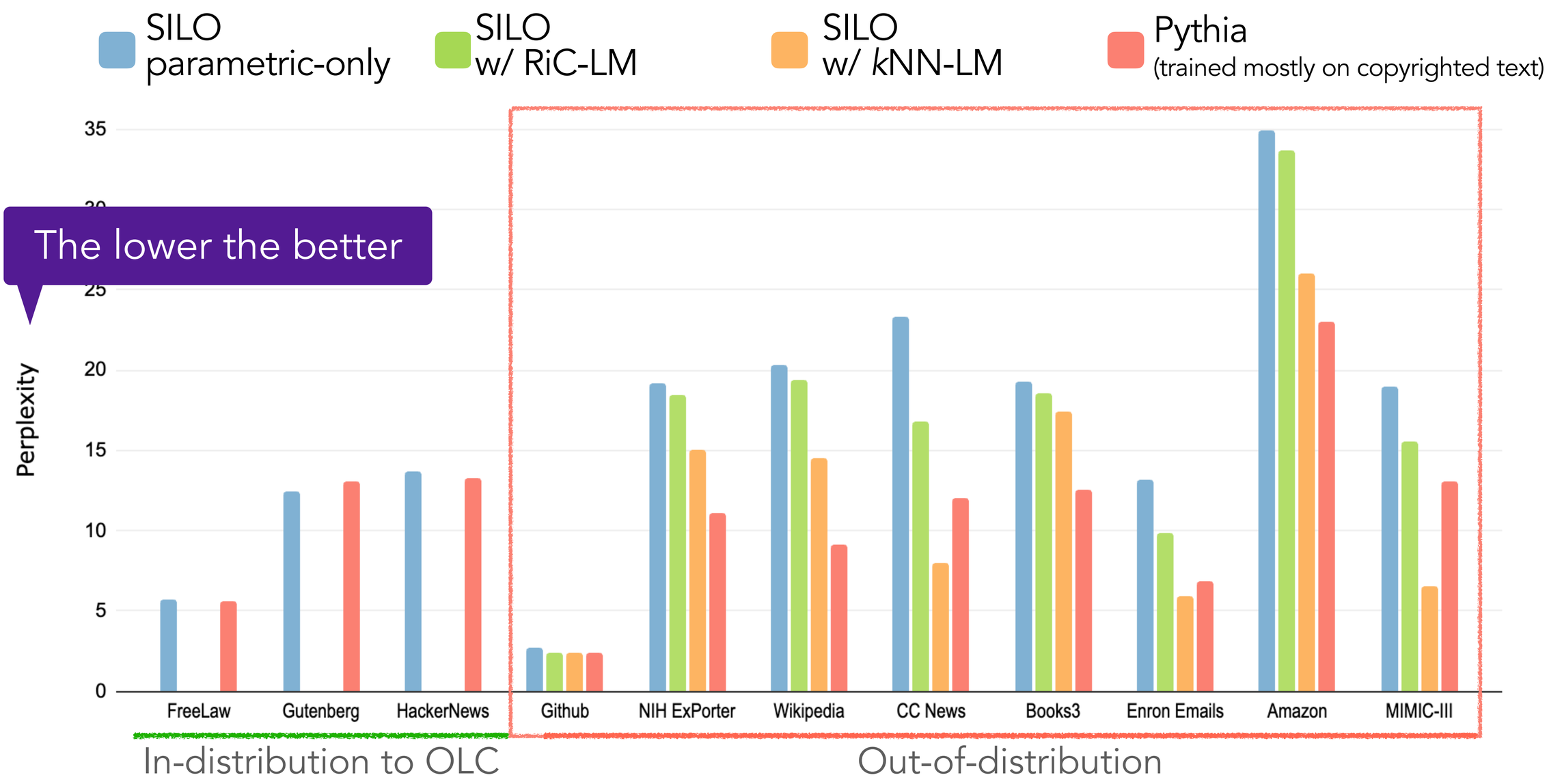

SILO
parametric-only
Pythia
(trained mostly on copyrighted text)
The lower the better
Perplexity
35
25
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution

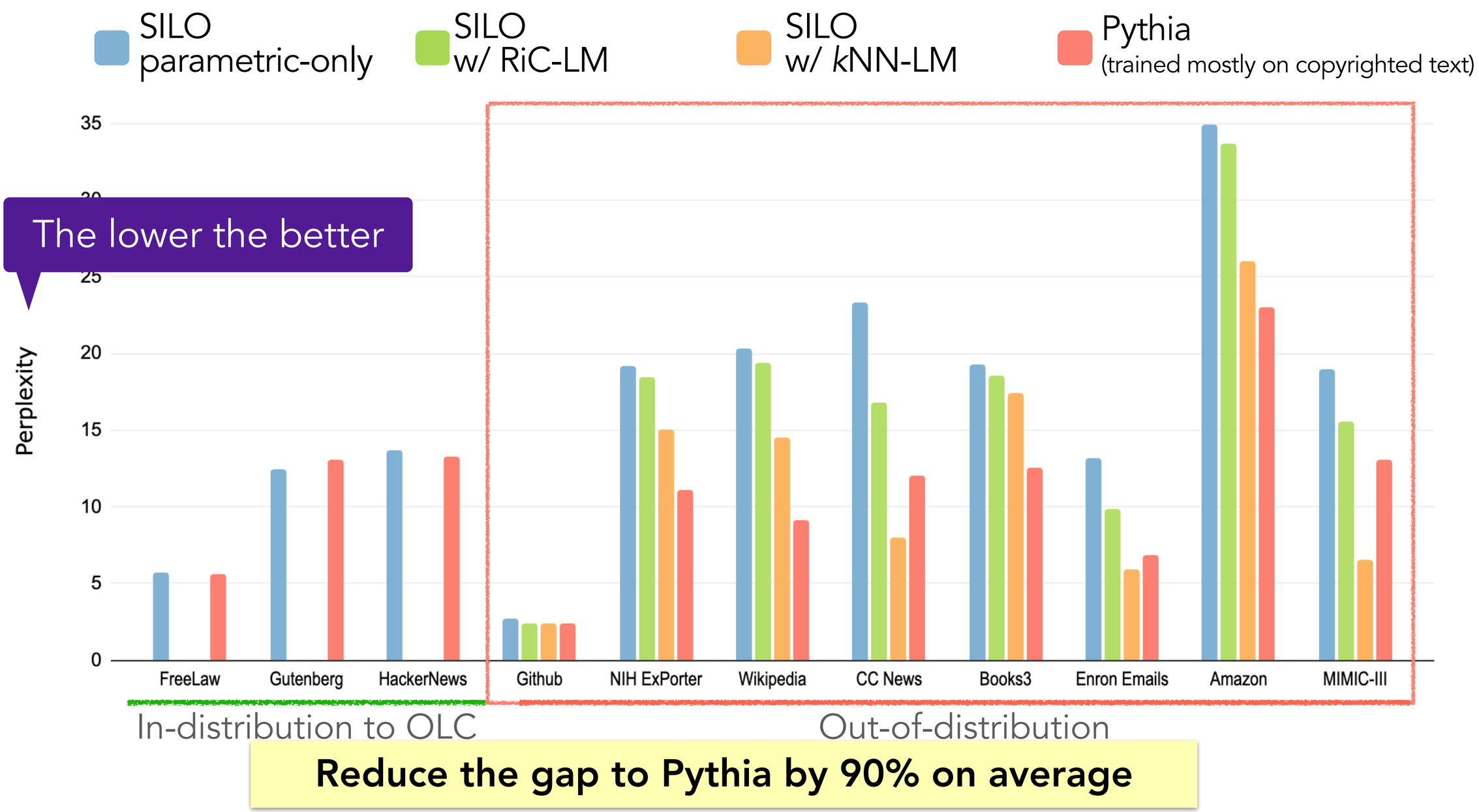

SILO
parametric-only
Pythia
(trained mostly on copyrighted text)
The lower the better
Perplexity
35
25
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution


SILO
parametric-only
SILO
w/ RiC-LM
Pythia
(trained mostly on copyrighted text)
The lower the better
Perplexity
35
25
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution


SILO
parametric-only
SILO
w/ RiC-LM
Pythia
(trained mostly on copyrighted text)
The lower the better
Perplexity
35
25
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution


SILO parametric-only
SILO w/ RiC-LM
SILO w/ kNN-LM
Pythia (trained mostly on copyrighted text)
The lower the better
Perplexity
35
25
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution

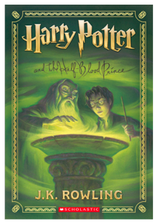

SILO parametric-only
SILO w/ RiC-LM
SILO w/ kNN-LM
Pythia (trained mostly on copyrighted text)
The lower the better
Perplexity
35
25
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution



Reduce the gap to Pythia by 90% on average

Attribution Example

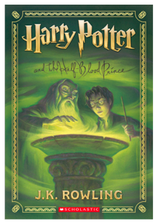
Attribution Example



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 …

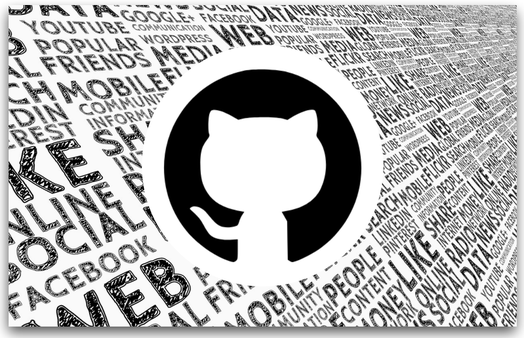
Top-1 retrieved context: Voldemort was ready. As Harry shouted "Expelliarmus!" Voldemort cried "Abada Kedavra!" A jet of **[green]** light issued from Voldemort's want …

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 …

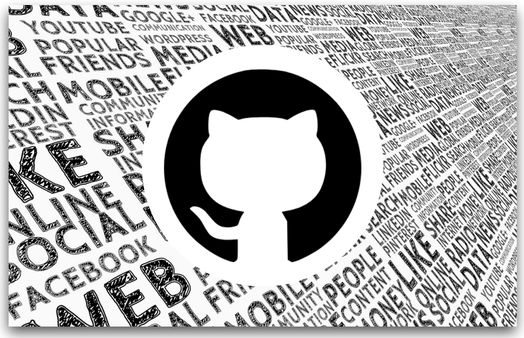


Attribution Example

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:

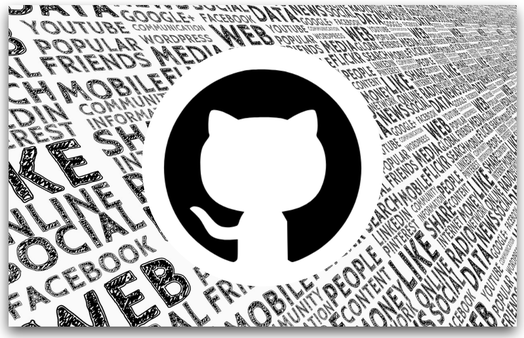
*You should have received a copy of the GNU Affero General Public License
* along with this program. If not, see <http://www.gnu.org/licenses/>.
*
*
**/
if (! has_right (
[AC]X_ACCESS)) { Header …

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 …

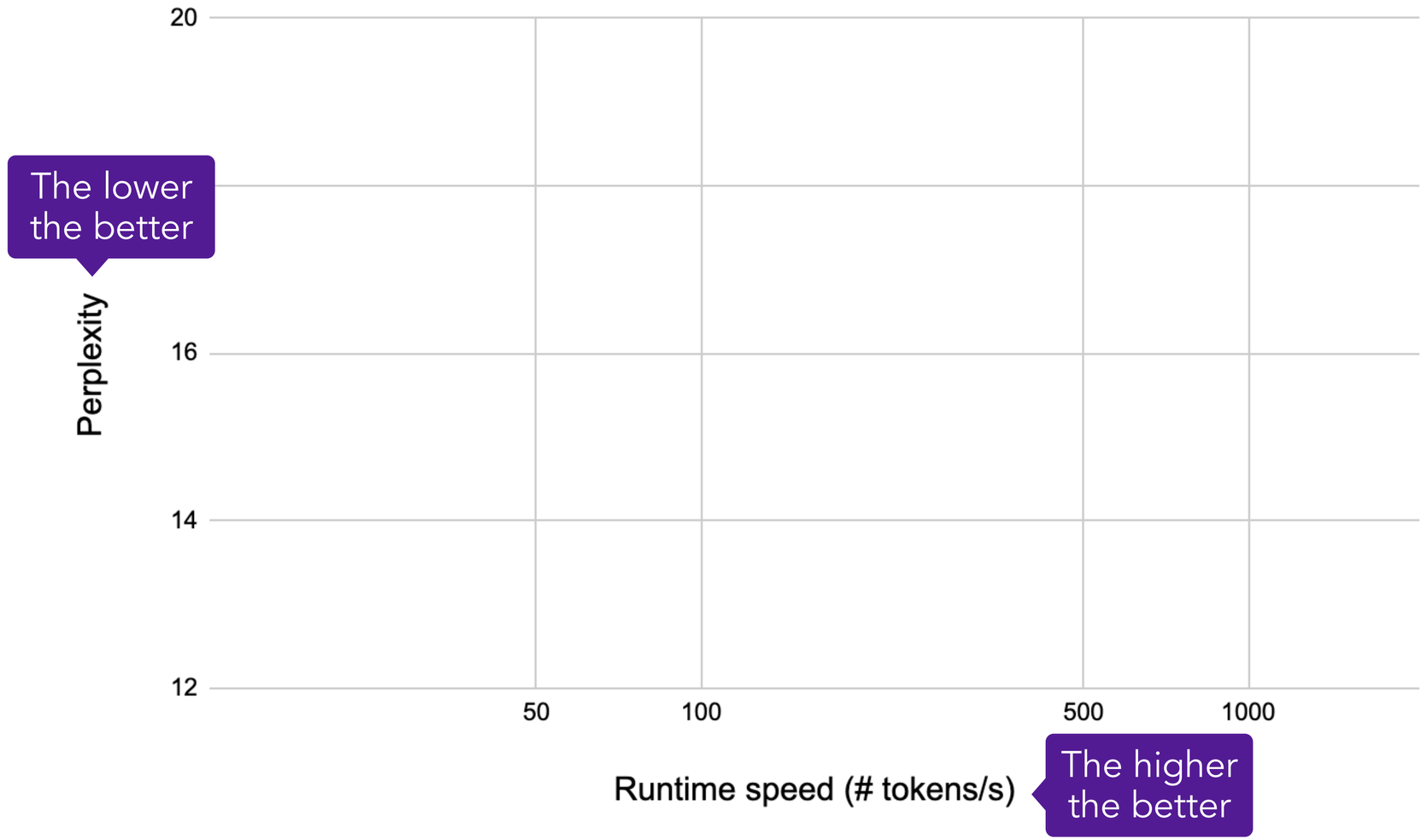


Top-1 retrieved context:

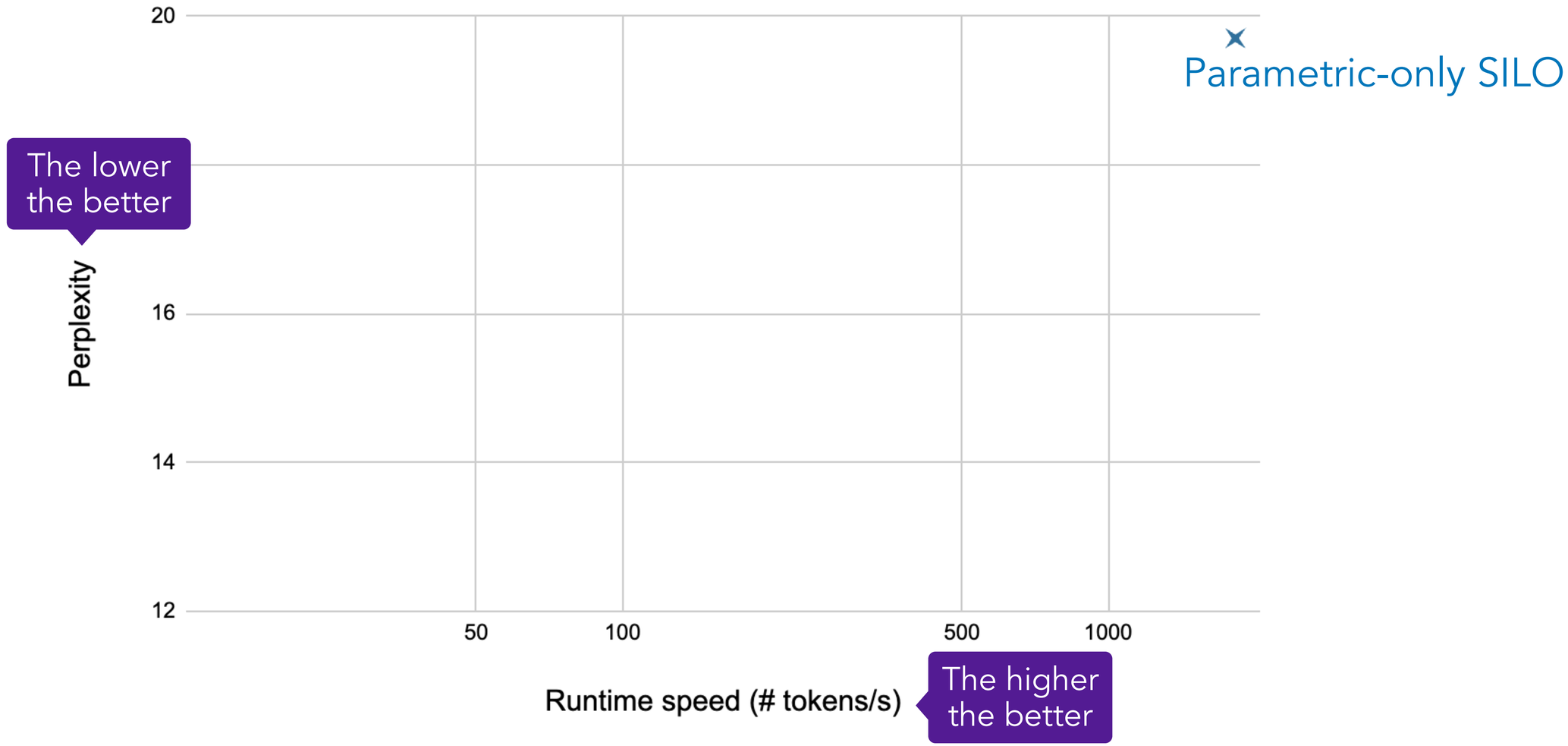
*You should have received a copy of the GNU Affero General Public License
* along with this program. If not, see <http://www.gnu.org/licenses/>.
*
*
**/
if (! has_right (
[AC]X_ACCESS)) { Header …

Runtime Speed

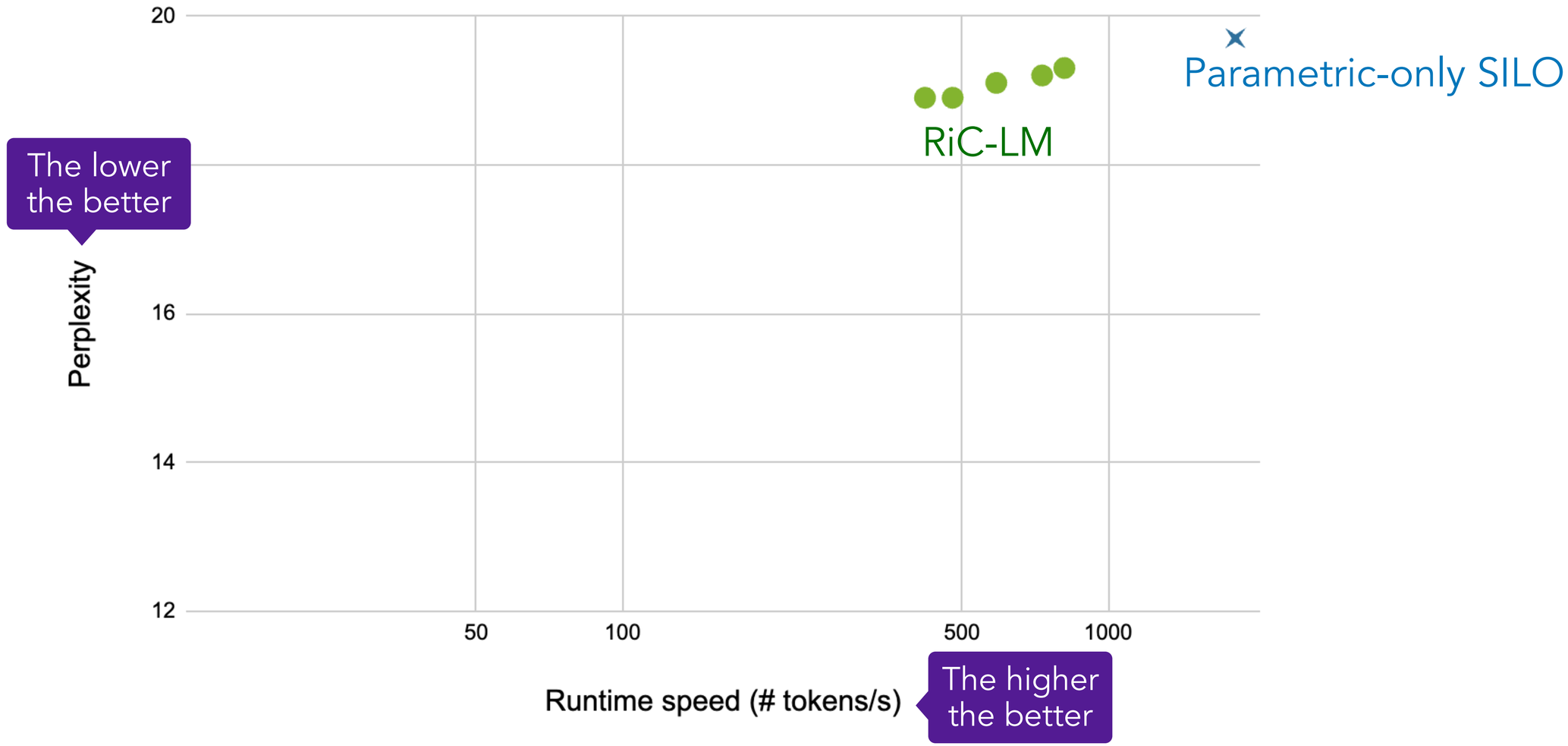
Runtime Speed



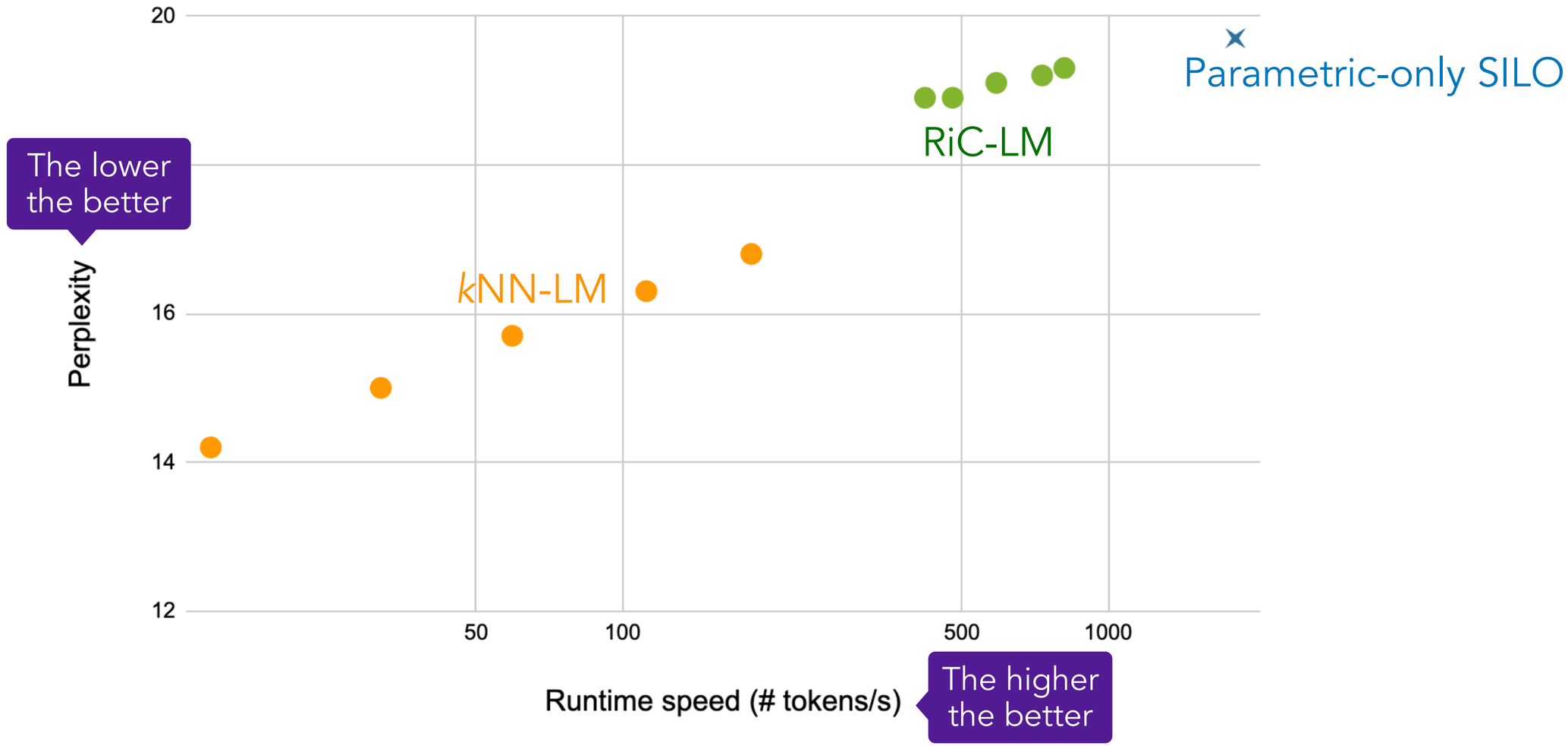
Runtime Speed



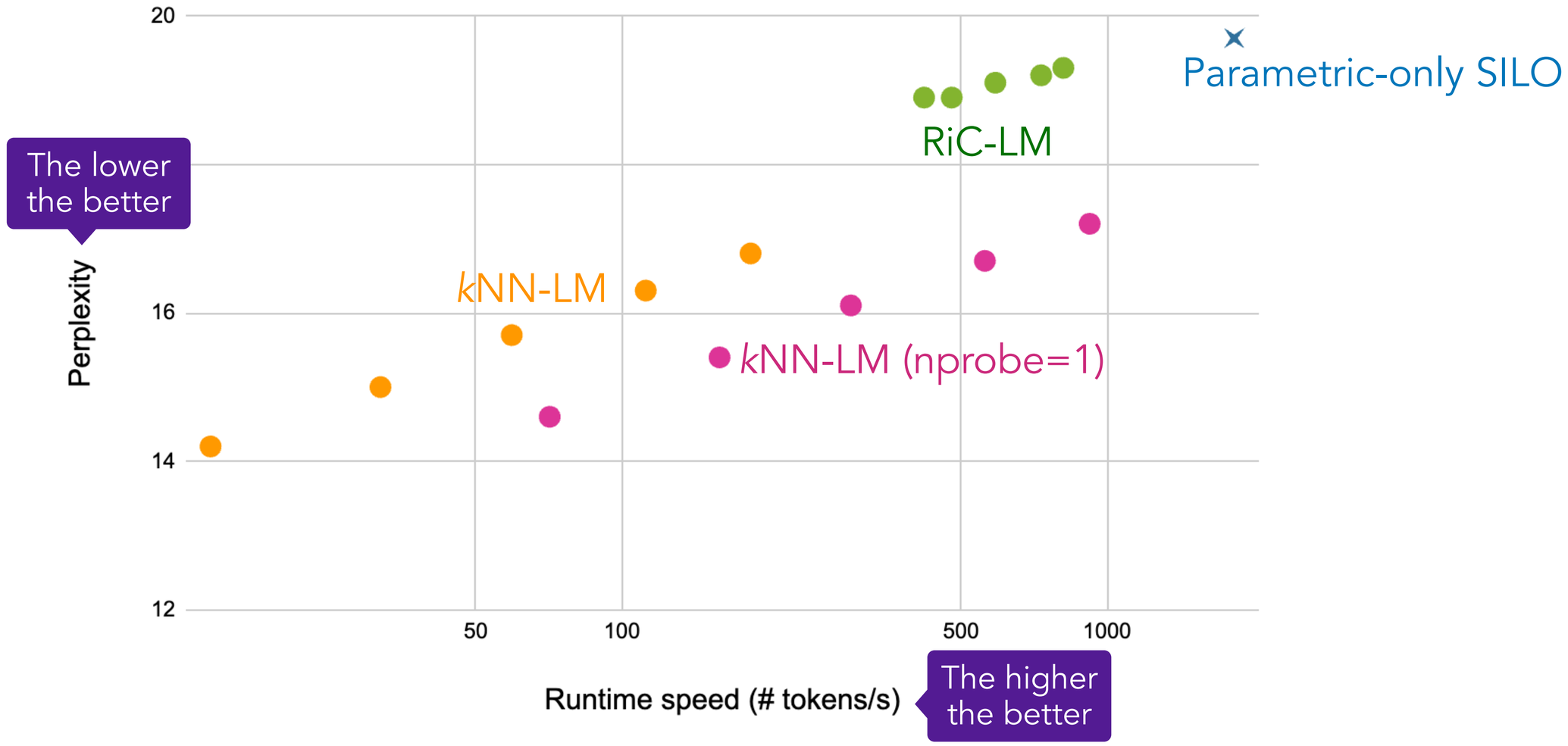
Runtime Speed



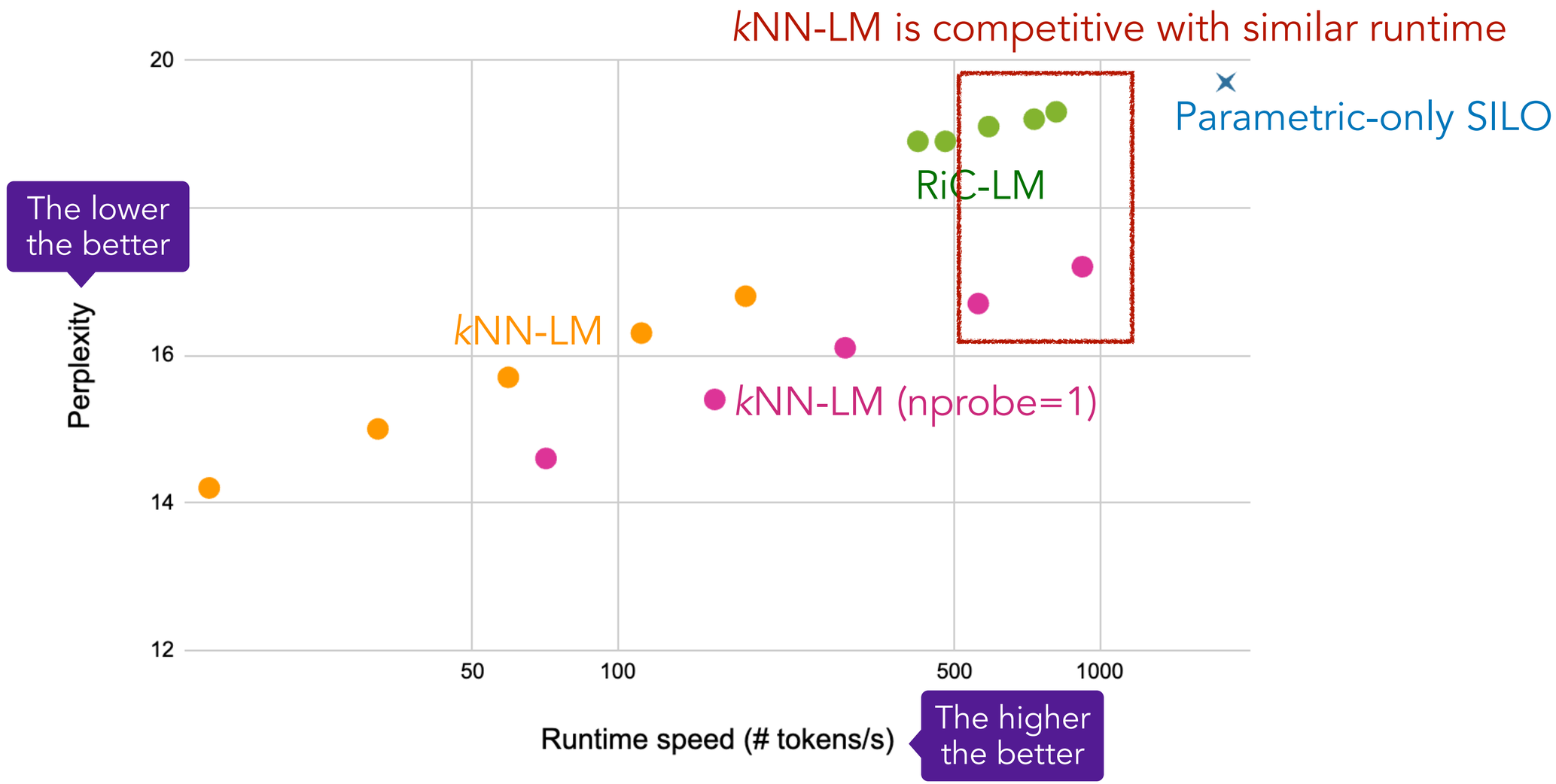
Runtime Speed



Runtime Speed



Runtime Speed



Talk outline

Data: Open License Corpus

100B token text corpus with public domain & permissively-licensed text

Our Model: SILO

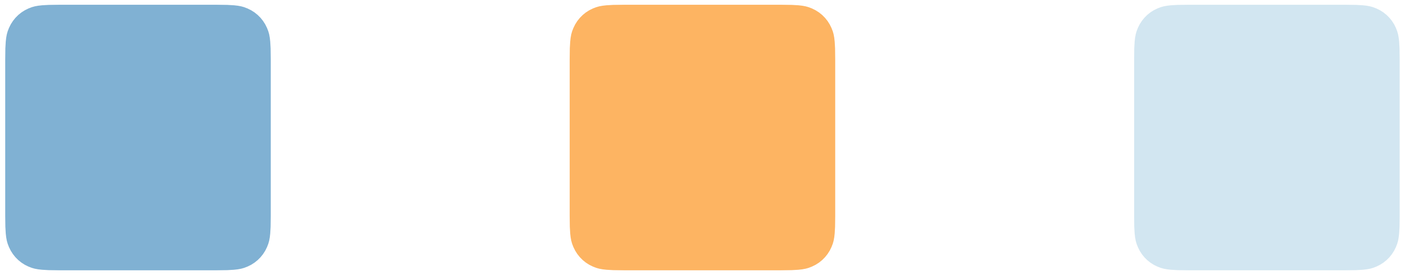
A nonparametric LM that isolates risks in a datastore

Proposal: Distributed LMs

LMs with a set of components, allowing flexible activation at test time

Distributed LMs

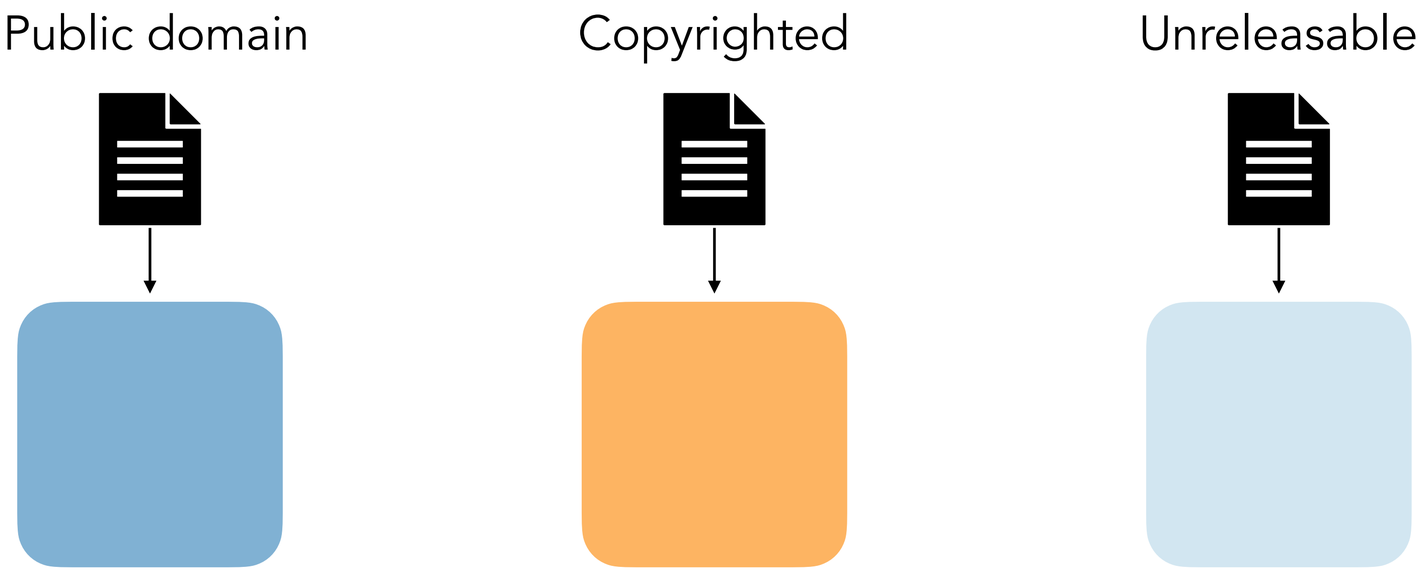
A set of model components:



Distributed LMs

A set of model components:

- Use different datasets

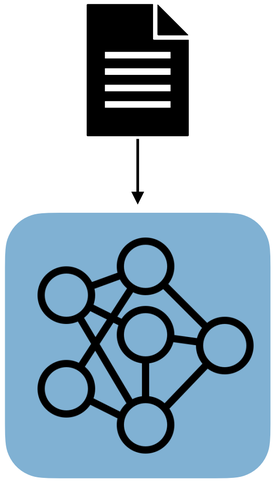


Distributed LMs

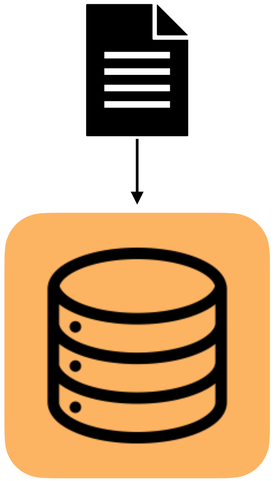
A set of model components:

- Use different datasets
- Use data differently

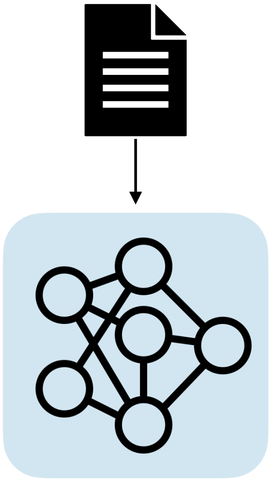
Public domain



Copyrighted



Unreleasable

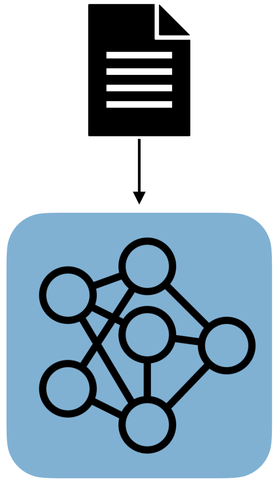


Distributed LMs

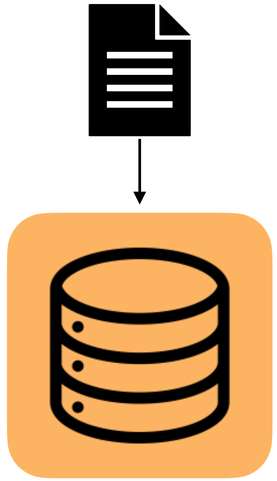
A set of model components:

- Use different datasets
- Use data differently
- Hosted differently

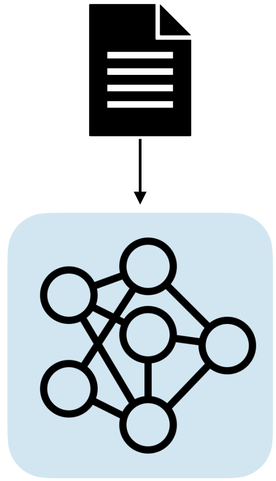
Public domain



Copyrighted



Unreleasable



Can be local

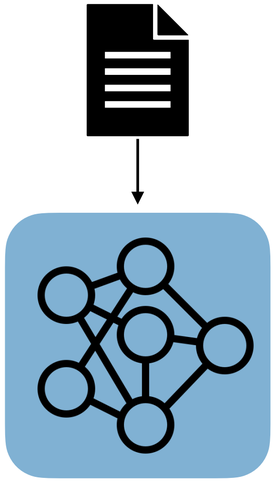
Hosted by data owners

Distributed LMs

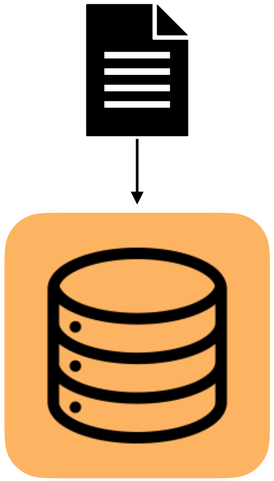
A set of model components:

- Use different datasets
- Use data differently
- Hosted differently
- **Which components to activate is flexible at test time**

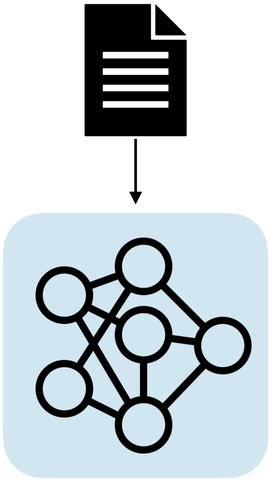
Public domain



Copyrighted



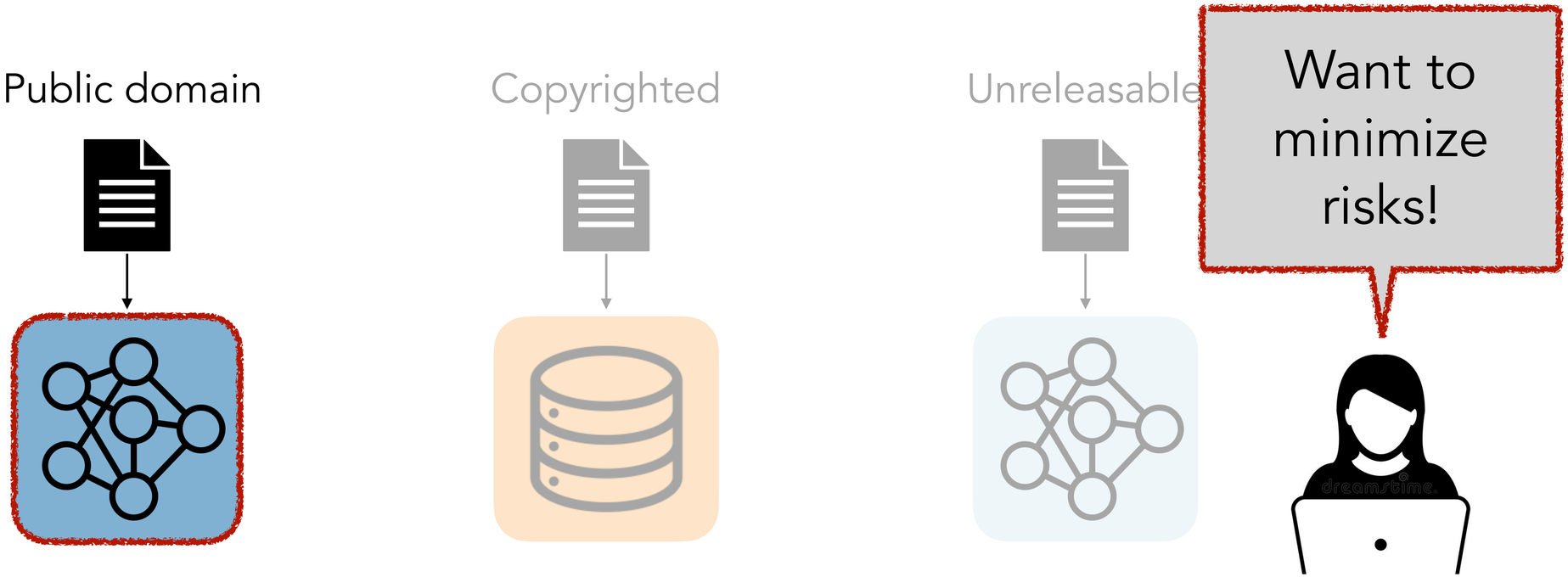
Unreleasable



Distributed LMs

A set of model components:

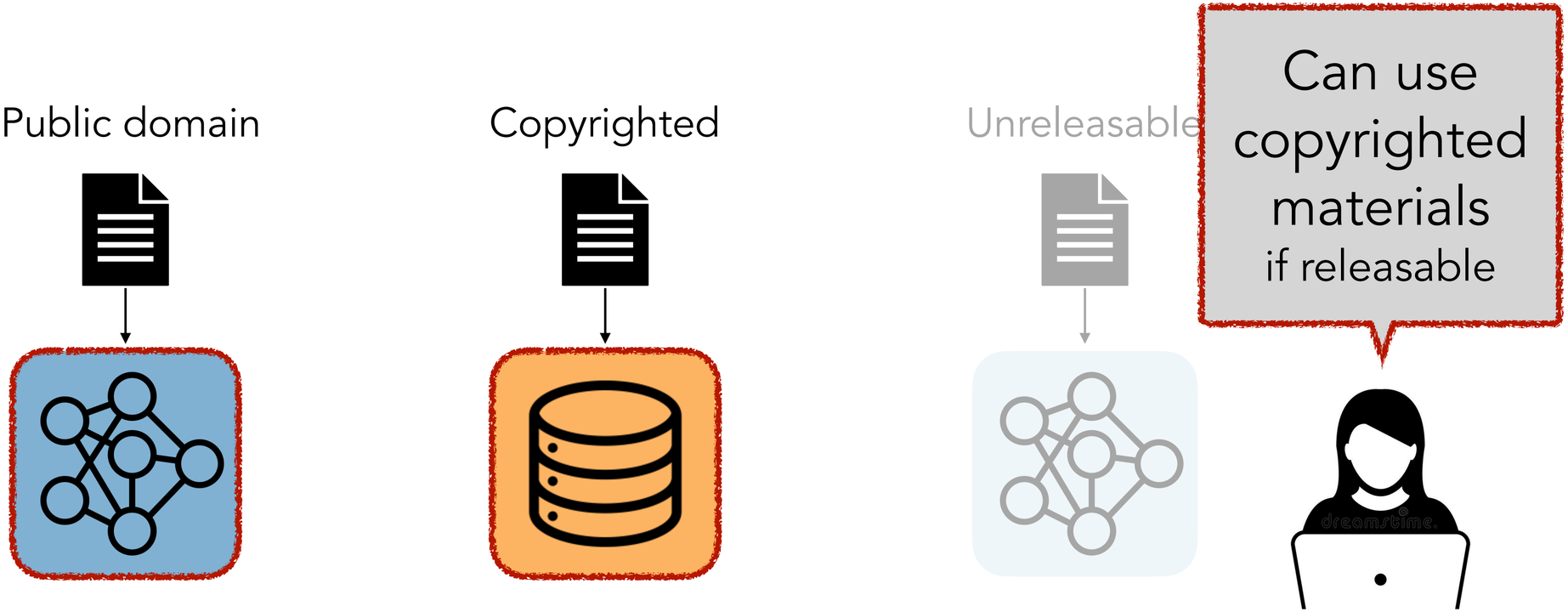
- Use different datasets
- Use data differently
- Hosted differently
- **Which components to activate is flexible at test time**



Distributed LMs

A set of model components:

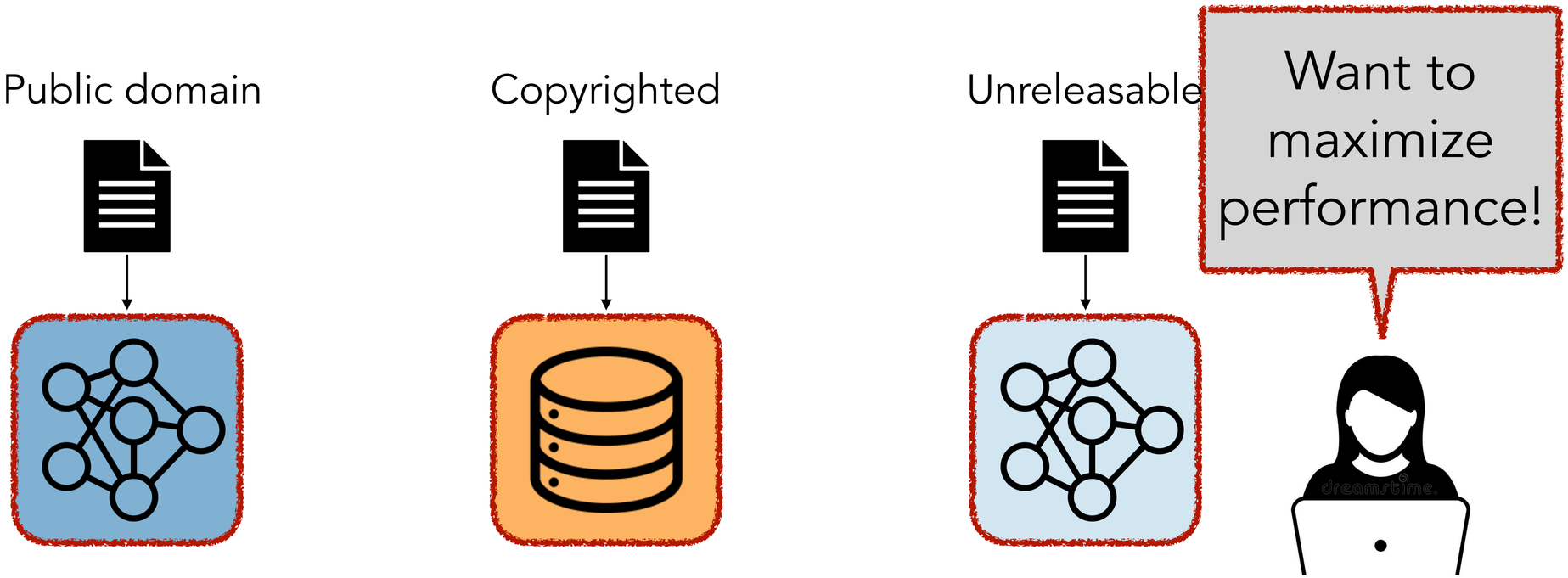
- Use different datasets
- Use data differently
- Hosted differently
- **Which components to activate is flexible at test time**



Distributed LMs

A set of model components:

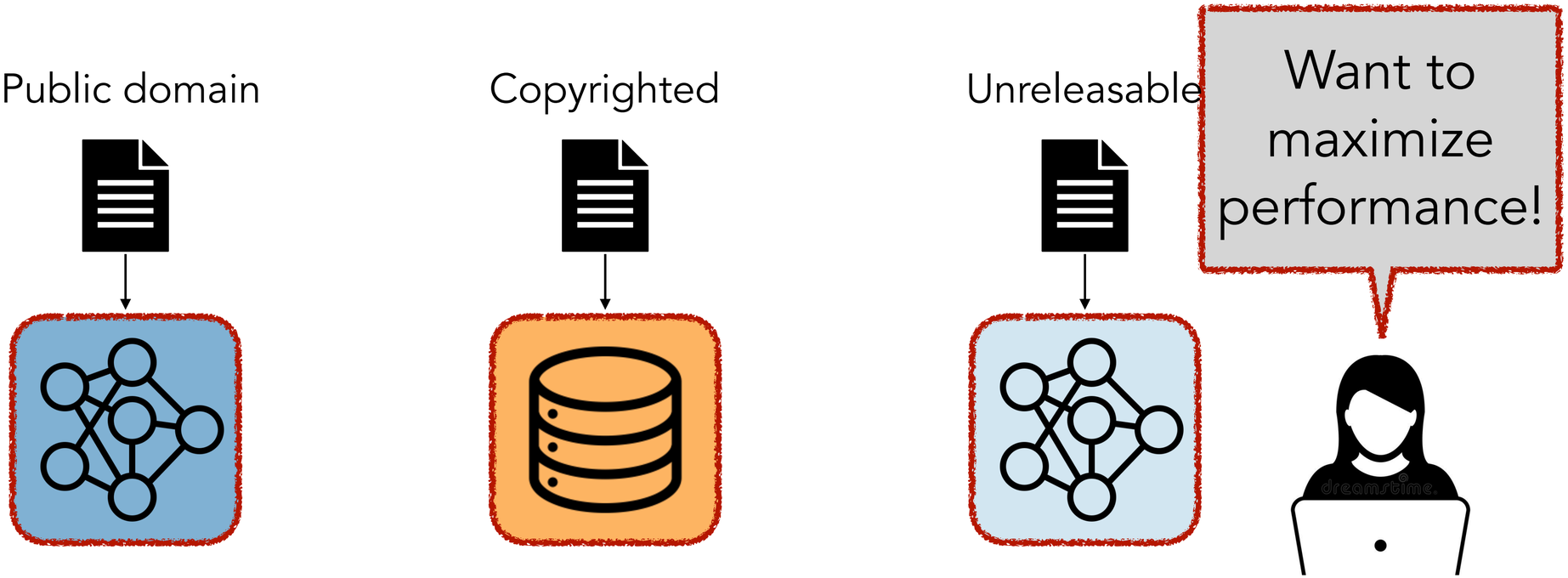
- Use different datasets
- Use data differently
- Hosted differently
- **Which components to activate is flexible at test time**



Distributed LMs

A set of model components:

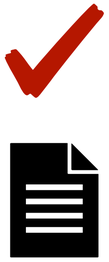
- Use different datasets
- Use data differently
- Hosted differently
- **Which components to activate is flexible at test time**



Distributed LMs (1): Nonparametric LMs

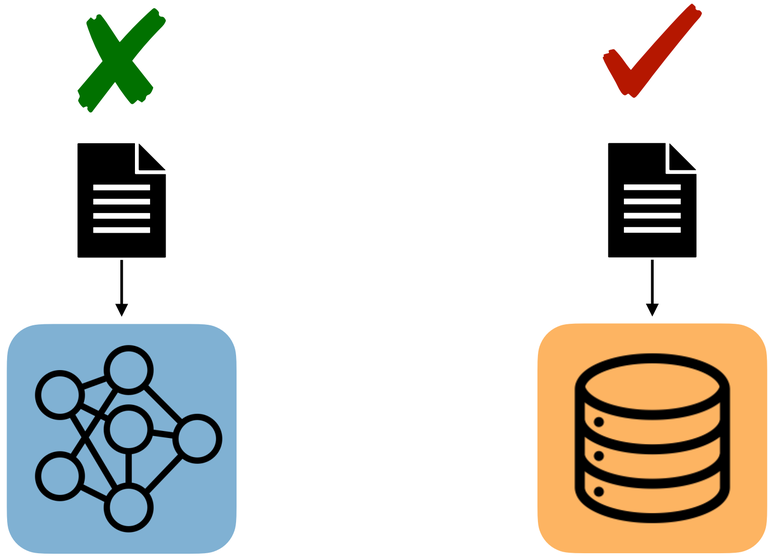
Distributed LMs (1): Nonparametric LMs

Attribution requirement?

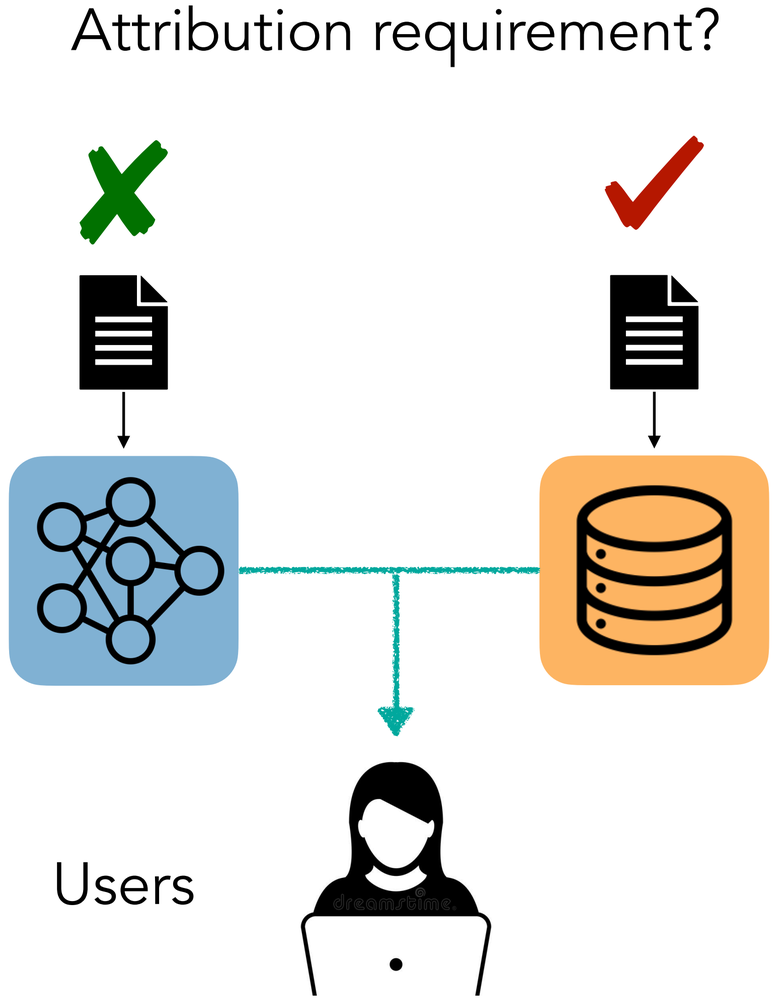


Distributed LMs (1): Nonparametric LMs

Attribution requirement?



Distributed LMs (1): Nonparametric LMs



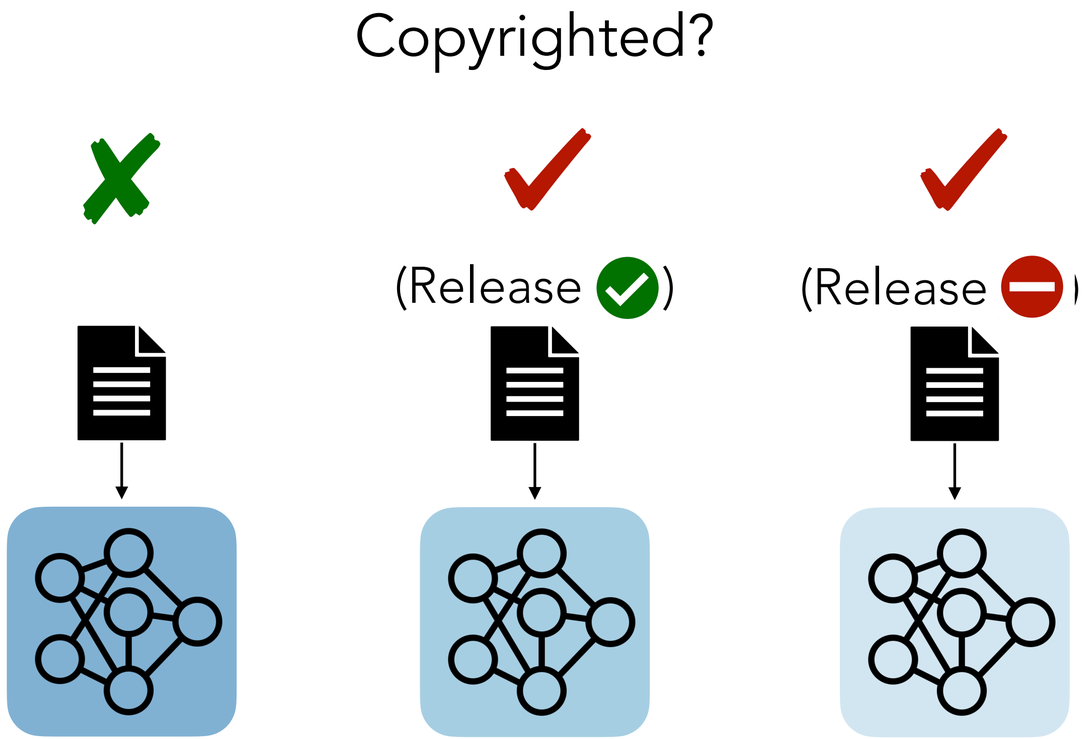
Distributed LMs (2): Mixture of Experts

Li et al. 2022. Branch-Train-Merge: Embarrassingly Parallel Training of Expert Language Models

Distributed LMs (2): Mixture of Experts

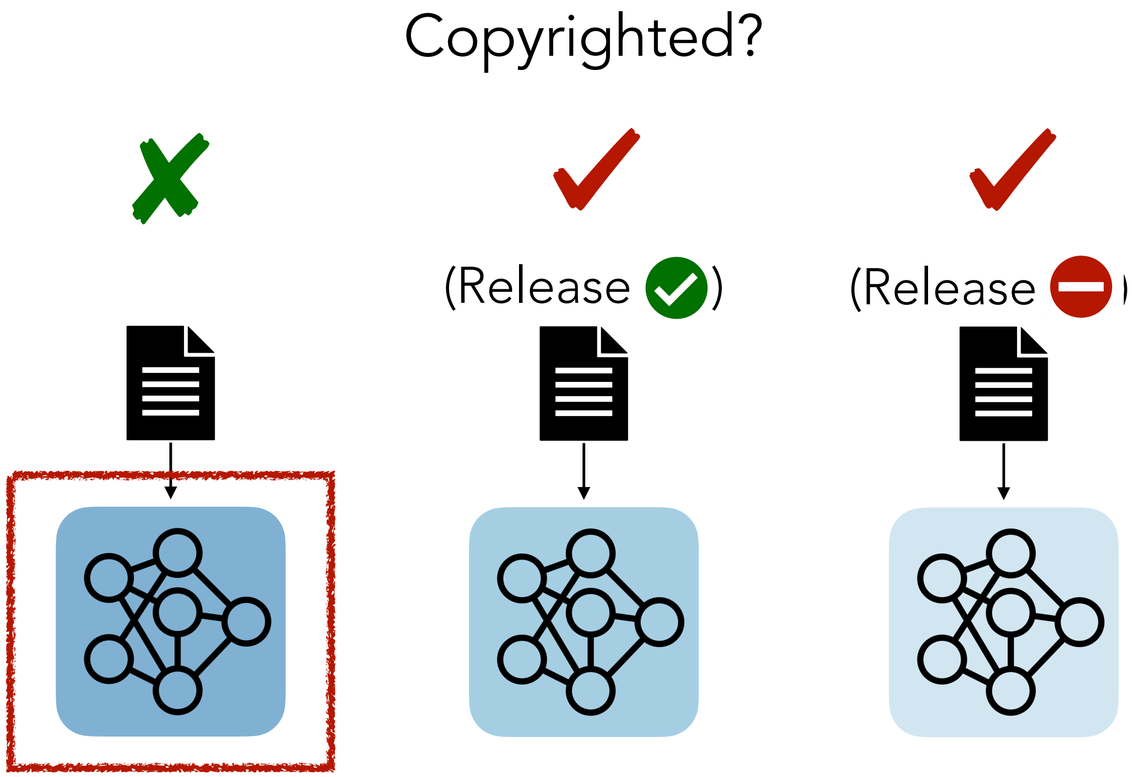
Li et al. 2022. Branch-Train-Merge: Embarrassingly Parallel Training of Expert Language Models

Distributed LMs (2): Mixture of Experts



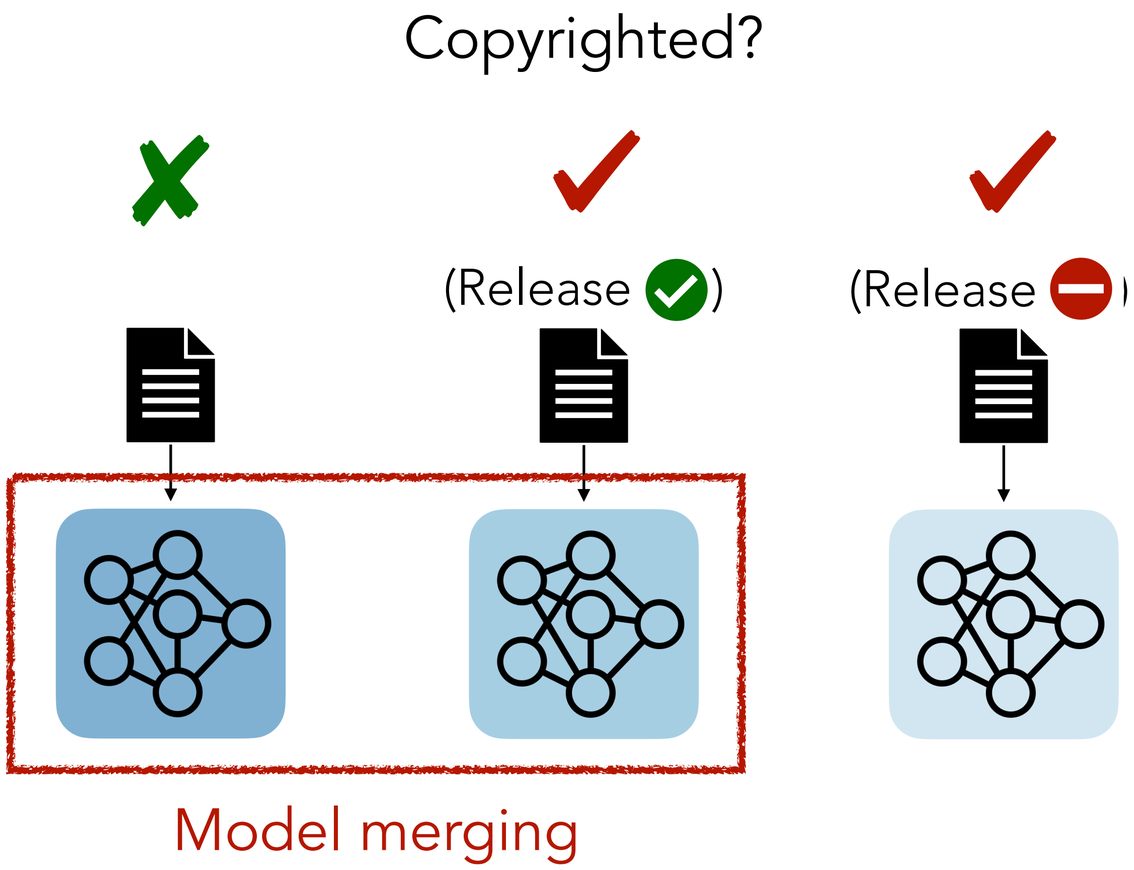
Li et al. 2022. Branch-Train-Merge: Embarrassingly Parallel Training of Expert Language Models

Distributed LMs (2): Mixture of Experts



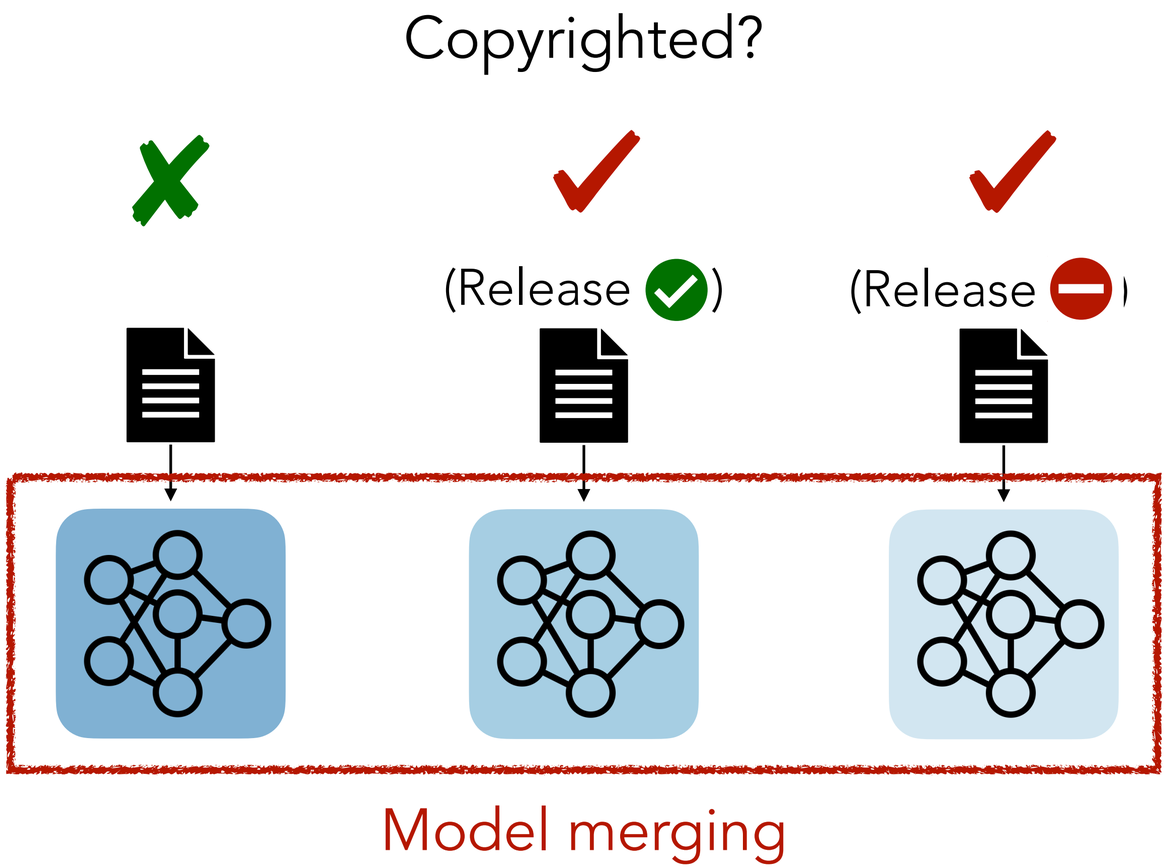
Li et al. 2022. Branch-Train-Merge: Embarrassingly Parallel Training of Expert Language Models

Distributed LMs (2): Mixture of Experts



Li et al. 2022. Branch-Train-Merge: Embarrassingly Parallel Training of Expert Language Models

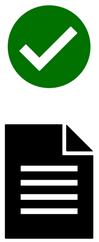
Distributed LMs (2): Mixture of Experts



Li et al. 2022. Branch-Train-Merge: Embarrassingly Parallel Training of Expert Language Models

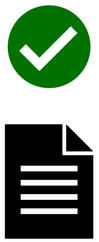
Distributed LMs (3): Collaborative LMs

Reproduction allowed?



Distributed LMs (3): Collaborative LMs

Reproduction allowed?

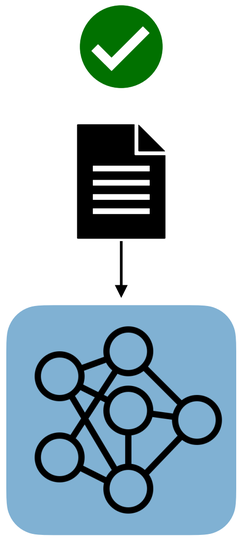


*"[…] is likely considered fair use in circumstances where the final model **does not directly generate content**. When it comes to […] generative use cases, the analysis becomes more complex."*

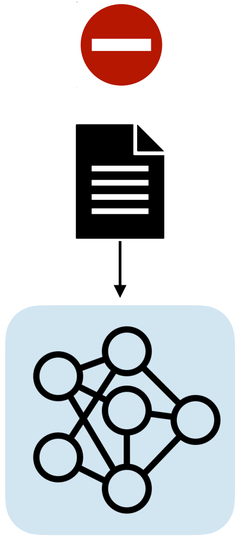
— Henderson et al. 2023. Foundation Models and Fair Use.

Distributed LMs (3): Collaborative LMs

Reproduction allowed?



A generative LM

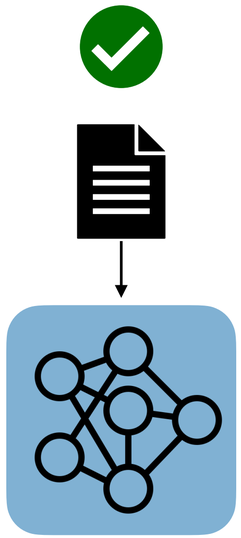


A non-generative LM

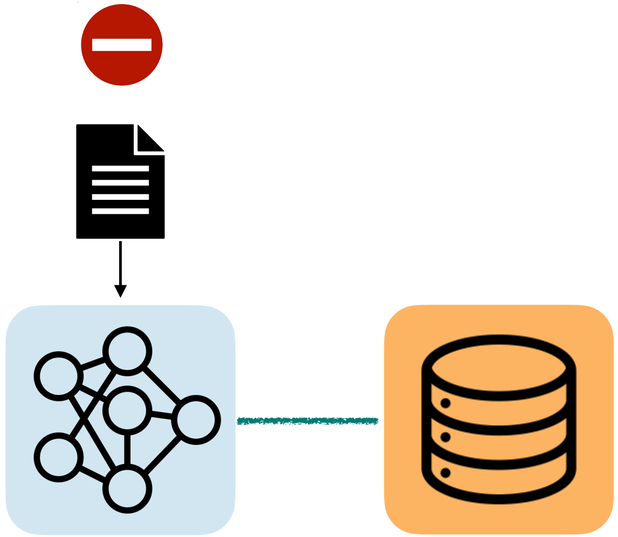
Colin Raffel. 2023. Collaborative, Communal, & Continual Machine Learning

Distributed LMs (3): Collaborative LMs

Reproduction allowed?



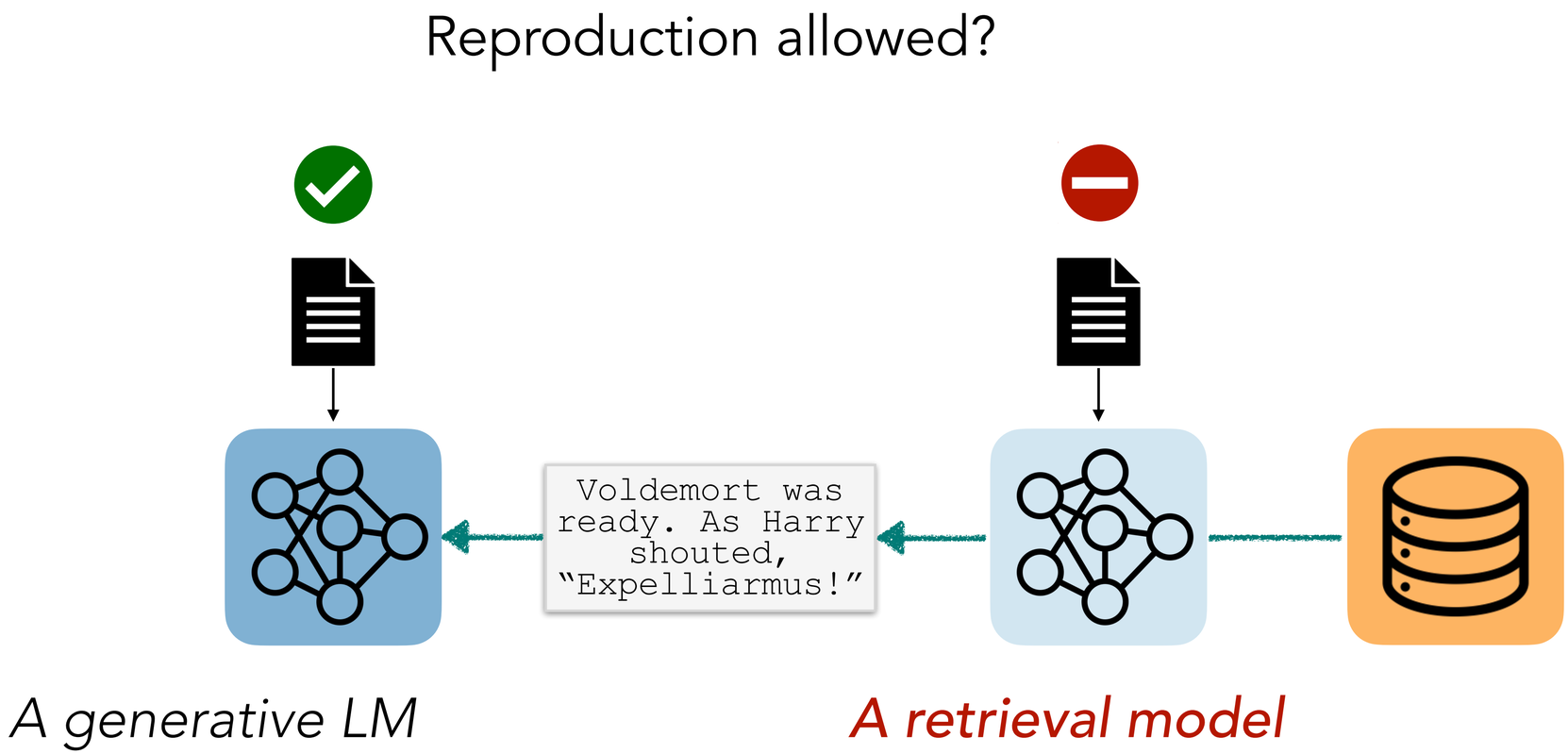
A generative LM



A retrieval model

Colin Raffel. 2023. Collaborative, Communal, & Continual Machine Learning

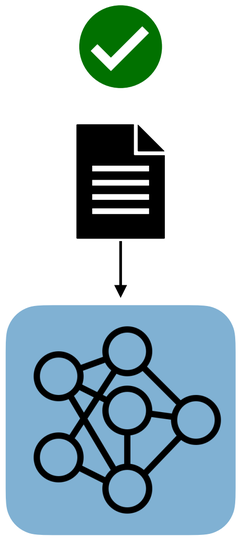
Distributed LMs (3): Collaborative LMs



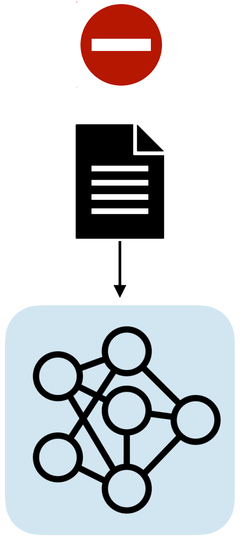
Colin Raffel. 2023. Collaborative, Communal, & Continual Machine Learning

Distributed LMs (3): Collaborative LMs

Reproduction allowed?



A generative LM

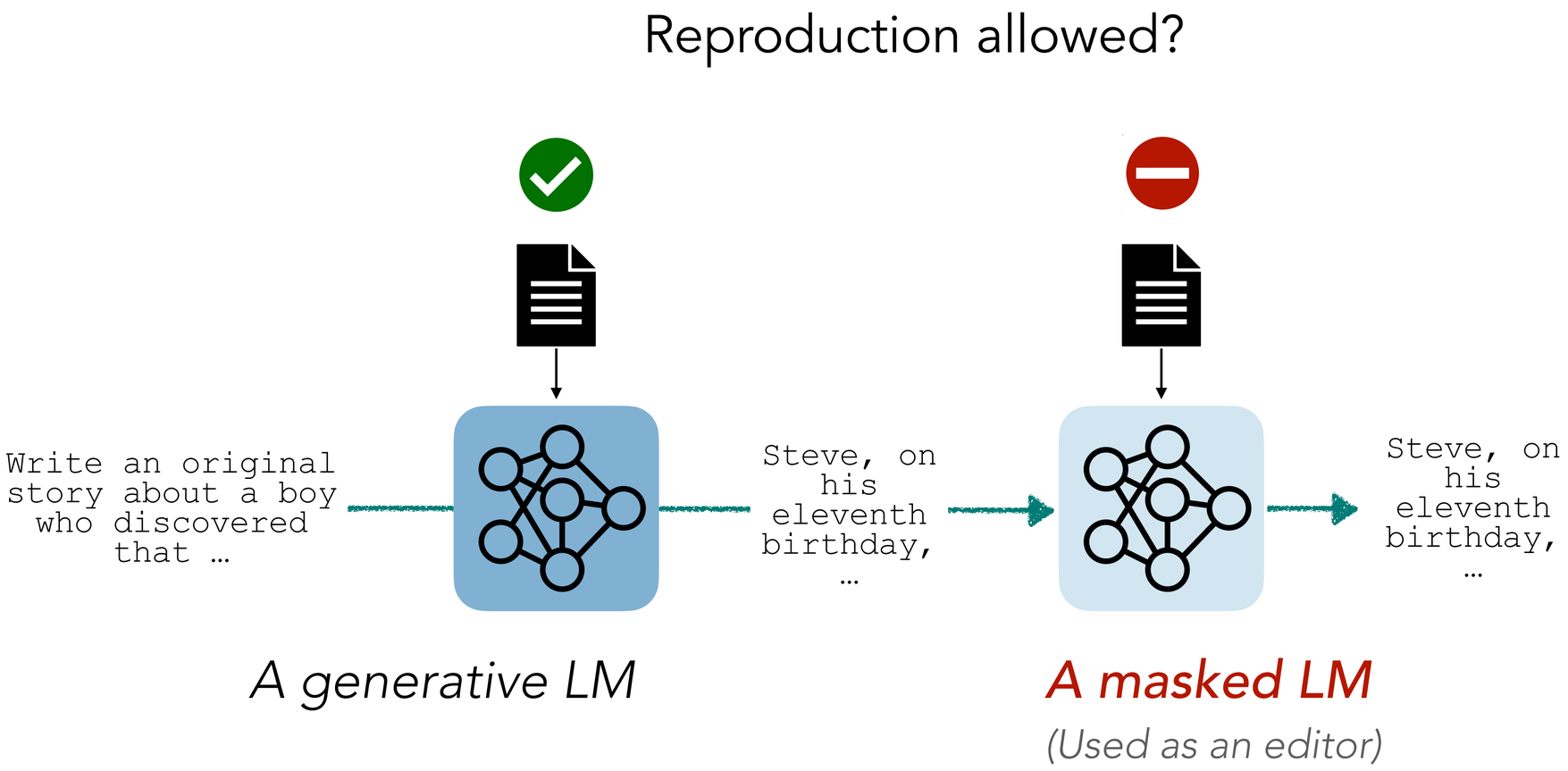


A masked LM

(Used as an editor)

Colin Raffel. 2023. Collaborative, Communal, & Continual Machine Learning

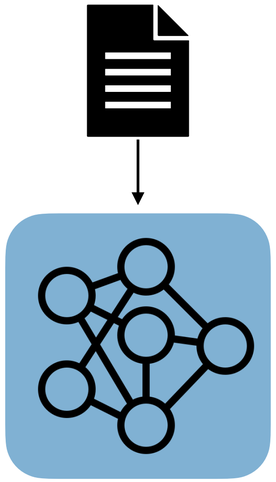
Distributed LMs (3): Collaborative LMs



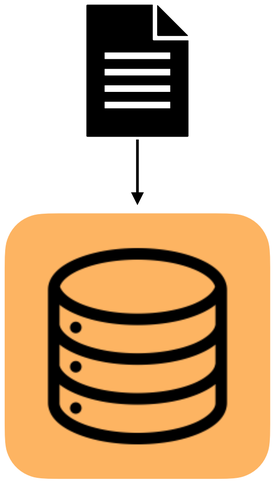
Colin Raffel. 2023. Collaborative, Communal, & Continual Machine Learning

Distributed LMs: Summary

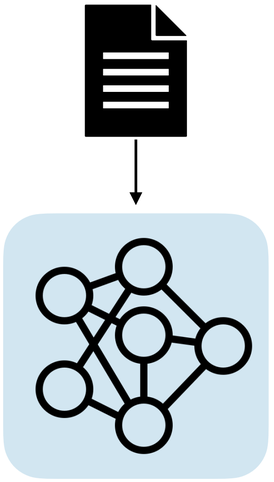
Public domain



Copyrighted



Unreleasable



Distributed LMs: Summary

Challenges

- Domain generalization

Distributed LMs: Summary

Challenges

- Domain generalization
- Methods for better communication
  (model merging, methods for exchanging outputs, etc)

Talk outline

Data: Open License Corpus

100B token text corpus with public domain & permissively-licensed text

Our Model: SILO

A nonparametric LM that isolates risks in a datastore

Proposal: Distributed LMs

LMs with a set of components, allowing flexible activation at test time

Talk outline

Every date source has their own restrictions

Data: Open License Corpus

100B token text corpus with public domain & permissively-licensed text

Our Model: SILO

A nonparametric LM that isolates risks in a datastore

Proposal: Distributed LMs

LMs with a set of components, allowing flexible activation at test time

Talk outline

Every date source has their own restrictions

Data: Open License Corpus

100B token text corpus with public domain & permissively-licensed text

Our Model: SILO

A nonparametric LM that isolates risks in a datastore

Proposal: Distributed LMs

LMs with a set of components, allowing flexible activation at test time

Talk outline

Every date source has their own restrictions

Data: Open License Corpus

100B token text corpus with public domain & permissively-licensed text

Our Model: SILO

A nonparametric LM that isolates risks in a datastore

Proposal: Distributed LMs

LMs with a set of components, allowing flexible activation at test time

Talk outline

Every date source has their own restrictions

Data: Open License Corpus

100B token text corpus with public domain & permissively-licensed text

Our Model: SILO

A nonparametric LM that isolates risks in a datastore

Proposal: Distributed LMs

LMs with a set of components, allowing flexible activation at test time

Talk outline

Every date source has their own restrictions

Data: Open License Corpus

100B token text corpus with public domain & permissively-licensed text

Our Model: SILO

A nonparametric LM that isolates risks in a datastore

Proposal: Distributed LMs

LMs with a set of components, allowing flexible activation at test time

Exciting follow up work! (Next slide)

Follow up (1)

Expanding OLC

common-pile Public

Repo to hold code and track issues for the collection of permissively licensed data

blester125 Changes made while scraping the whole Project Gutenberg ...	796b6cc · 4 hours ago	86 Commits
.github/workflows	add ci for linting	last month
arxiv	Create README.md	4 months ago
bhl	run linters over all files	last month
courtlistener	Data Provenance data (#61)	5 days ago
data_provenance	Data Provenance data (#61)	5 days ago
food	Updates needed to fully scrape foodista data. (#77)	6 hours ago
gutenberg	Changes made while scraping the whole Project Gut...	4 hours ago
licensed_pile	Data Provenance data (#61)	5 days ago
news	Lintangsutawika news (#68)	yesterday
public_domain_review	Update DPR scraping	last month
pubmedcentral	update readme with created	last month
stackexchange	updates needed when processing the full dump (#79)	4 hours ago
stackv2	Stackv2 (#72)	6 hours ago
ubuntu	filter empty chats (#80)	4 hours ago

github.com/r-three/common-pile

Common Corpus updated Mar 20

The largest public domain dataset for training LLMs.

- PleIAs/US-PD-Newspapers — Viewer • Updated Mar 22 • 9 • 35
- PleIAs/French-PD-Books — Viewer • Updated Mar 19 • 7.56k • 39
- PleIAs/French-PD-Newspapers — Viewer • Updated Mar 19 • 5 • 60
- PleIAs/German-PD — Viewer • Updated Mar 21 • 1 • 9
- PleIAs/Spanish-PD-Books — Viewer • Updated Mar 21 • 4 • 3

PleIAs huggingface.co/collections/PleIAs/common-corpus-65d46e3ea3980fdcd66a5613

Follow up (1)

Expanding OLC

common-pile Public

Repo to hold code and track issues for the collection of permissively licensed data

blester125 Changes made while scraping the whole Project Gutenberg ...	796b6cc · 4 hours ago	86 Commits
.github/workflows	add ci for linting	last month
arxiv	Create README.md	4 months ago
bhl	run linters over all files	last month
courtlistener	Data Provenance data (#61)	5 days ago
data_provenance	Data Provenance data (#61)	5 days ago
food	Updates needed to fully scrape foodista data. (#77)	6 hours ago
gutenberg	Changes made while scraping the whole Project Gut...	4 hours ago
licensed_pile	Data Provenance data (#61)	5 days ago
news	Lintangsutawika news (#68)	yesterday
public_domain_review	Update DPR scraping	last month
pubmedcentral	update readme with created	last month
stackexchange	updates needed when processing the full dump (#79)	4 hours ago
stackv2	Stackv2 (#72)	6 hours ago
ubuntu	filter empty chats (#80)	4 hours ago

github.com/r-three/common-pile

Common Corpus updated Mar 20

The largest public domain dataset for training LLMs.

- PleIAs/US-PD-Newspapers — Viewer • Updated Mar 22 • 9 • 35
- PleIAs/French-PD-Books — Viewer • Updated Mar 19 • 7.56k • 39
- PleIAs/French-PD-Newspapers — Viewer • Updated Mar 19 • 5 • 60
- PleIAs/German-PD — Viewer • Updated Mar 21 • 1 • 9
- PleIAs/Spanish-PD-Books — Viewer • Updated Mar 21 • 4 • 3

100 billion → 500 billion tokens

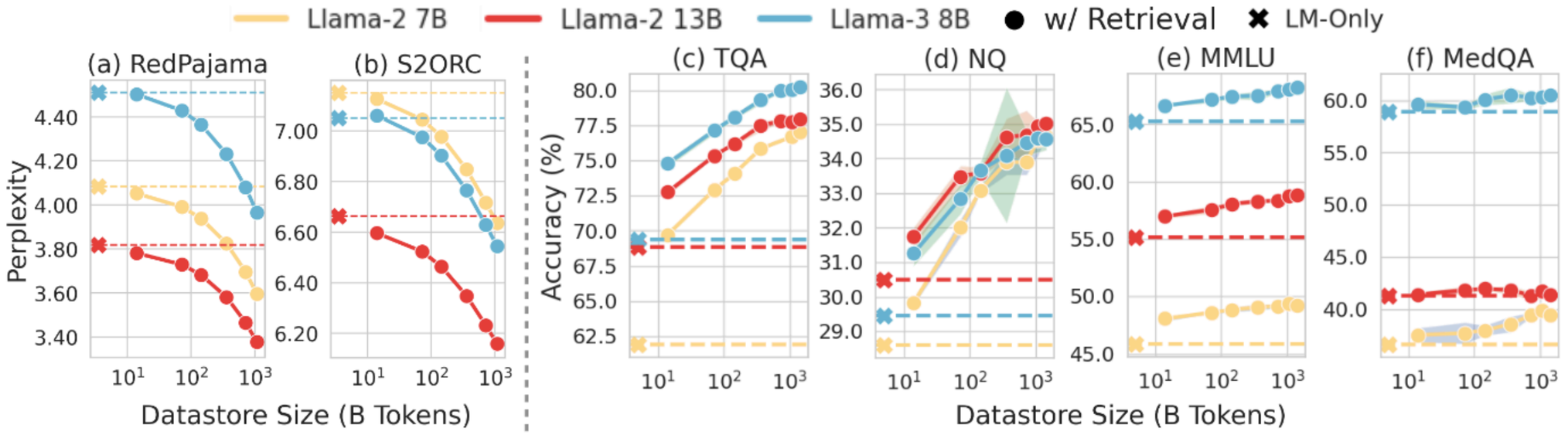
PleIAs

huggingface.co/collections/PleIAs/common-corpus-65d46e3ea3980fdcd66a5613

Follow up (2)

Expanding OLC

Expanding nonparametric LMs

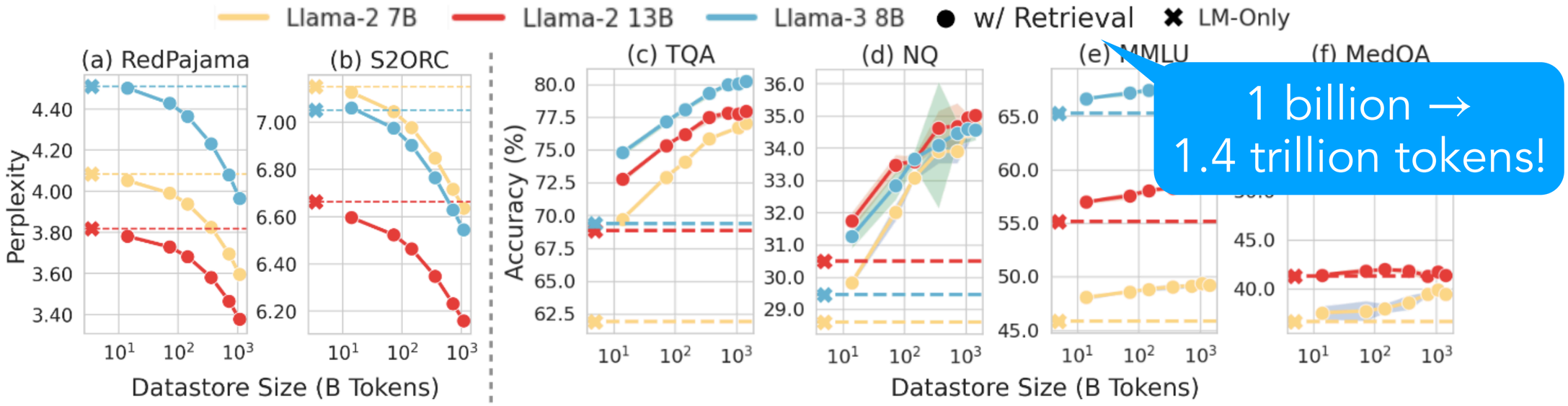


Shao, …., **Min** et al. Submitted to NeurIPS 2024

Follow up (2)

Expanding OLC

Expanding nonparametric LMs



Shao, …., **Min** et al. Submitted to NeurIPS 2024

Thanks to collaborators

UWNLP AI2 Berkeley UNIVERSITY OF CALIFORNIA

And stability.ai for providing compute

sewonmin.com

sewon@cs.washington.edu

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