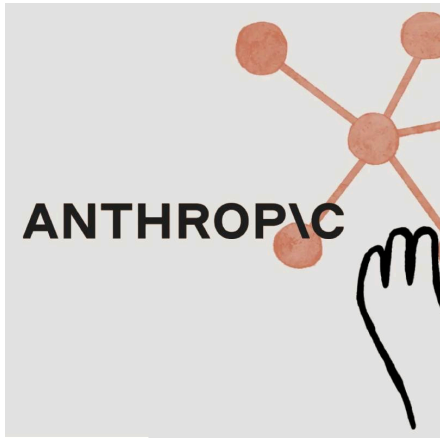
Subtleties about [Pre-]Training Data:
Imbalance and Staleness

Daniel Khashabi

The success we dreamed of

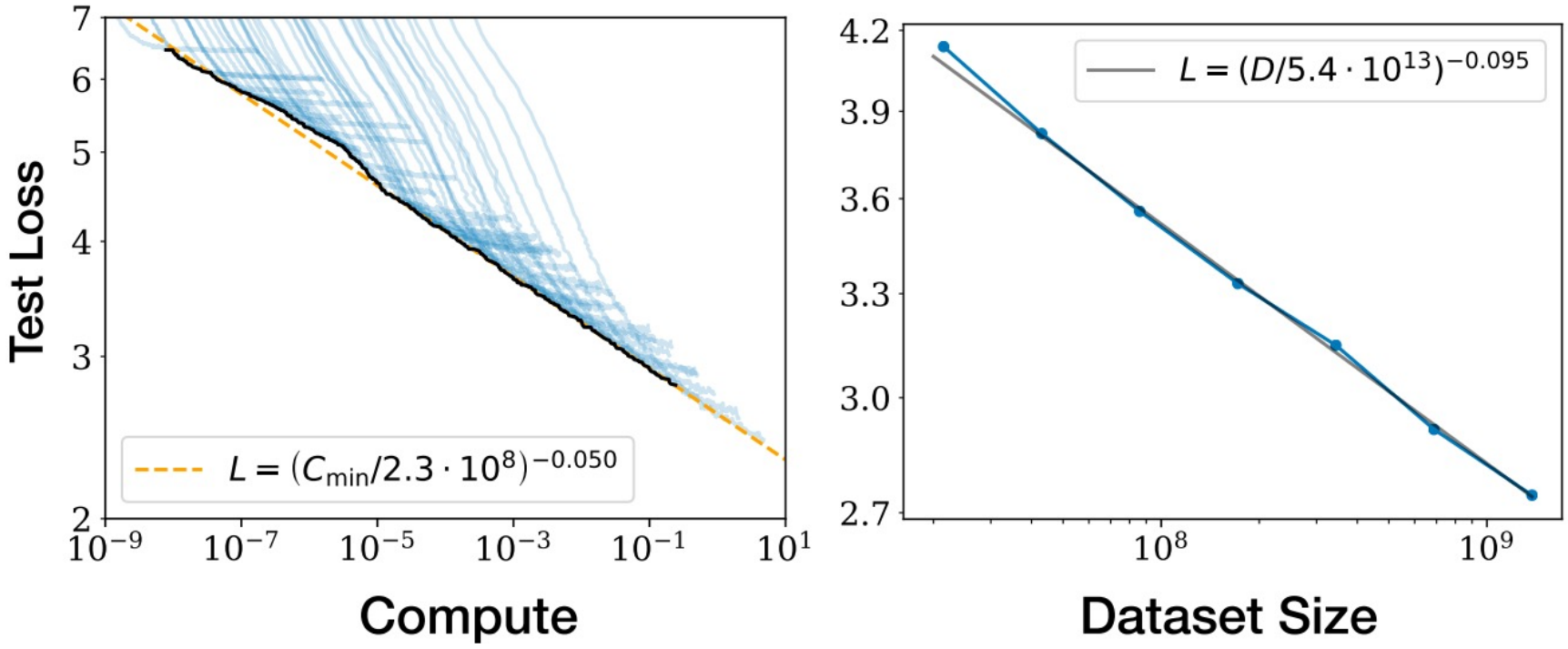
Gemini
Google



OLMoE
Ai2

Language models that are remarkably capable at solving many important NLP benchmarks.

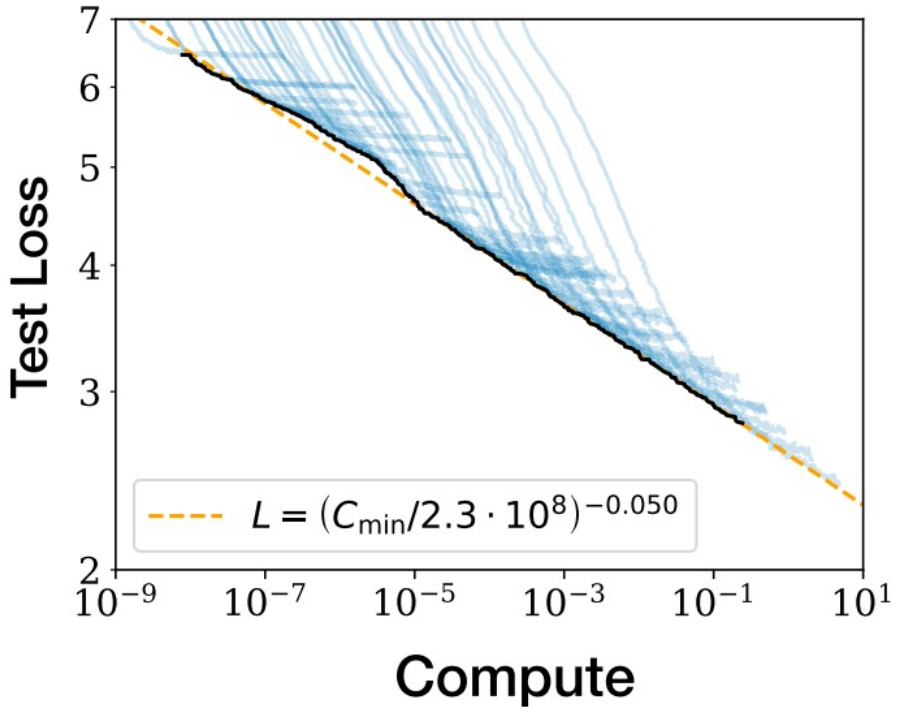
~~AI~~ Data is the "new electricity"

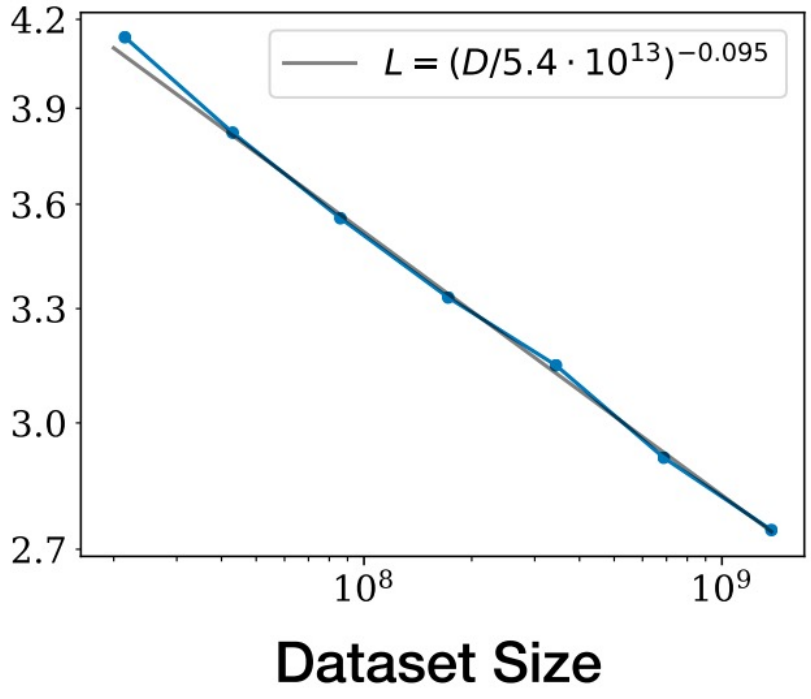


Kaplan et al. 2020; among others

More data (and compute) leads to better models.

Limits of scaling "laws"





Kaplan et al. 2020; among others

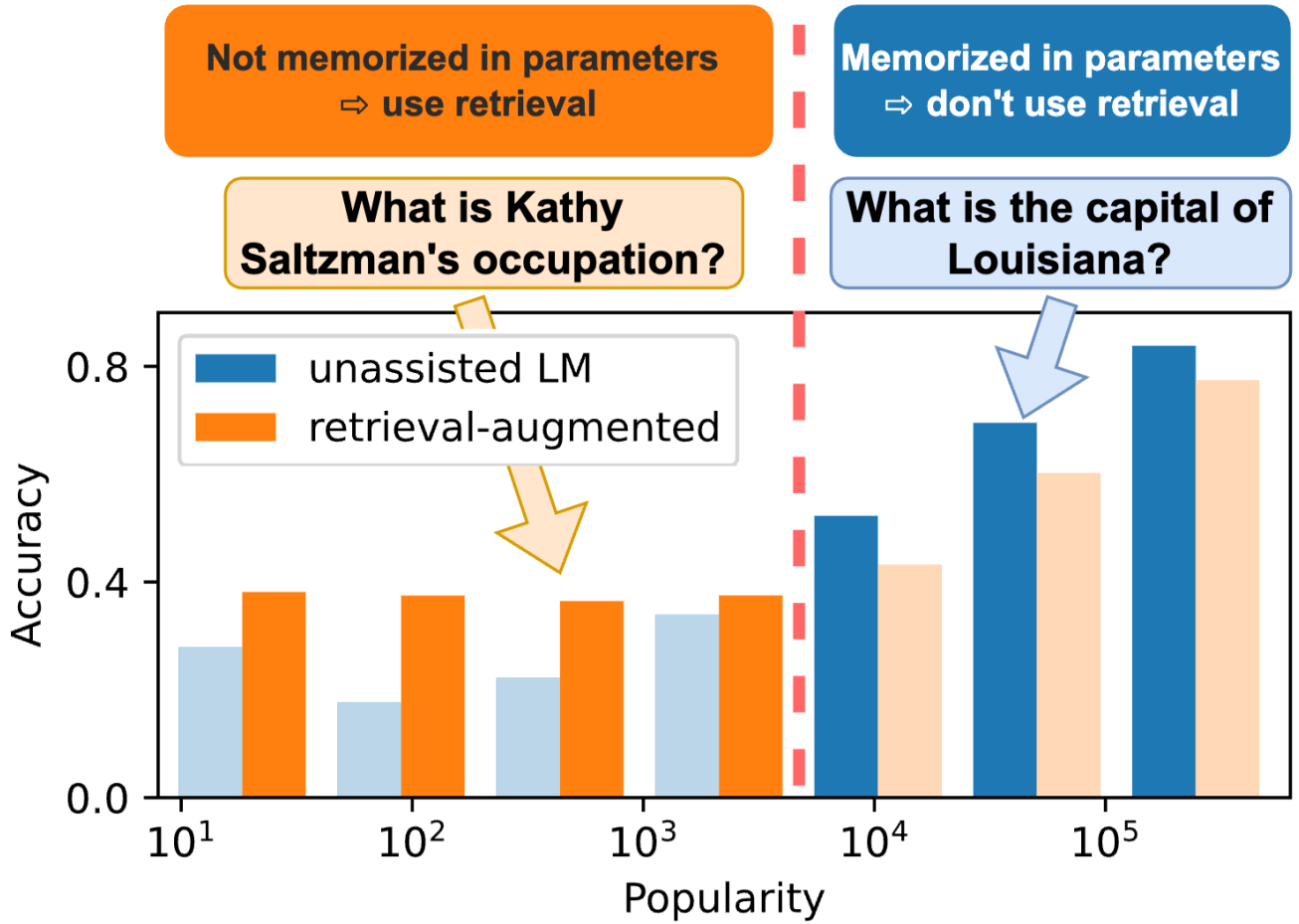
Which data? How is it distributed?

Our understanding of data remains limited

- Not all knowledge is uniformly distributed.

- Factual accuracy of LLMs is positively correlated with "popularity" of the inputs.

(Mallen et al. 2023; among others)



Our understanding of data remains limited

- The recent progress (e.g., Llama2 to Llama3) seems to be mostly due to better data engineering (better data filtering).

There are various nuances about data that are not explained any "laws" and remain poorly understood.

Today

Stale data

- A ton world data is old.
  How does that impact our models?

Imbalanced data

- Most data follow a long-tail distribution.
  How should we deal with it?

Today

Stale data

- A ton world data is old.
  How does that impact our models?

Imbalanced data

- Most data follow a long-tail distribution.
  How should we deal with it?

Knowledge Cutoffs

My knowledge cutoff date is September 2023. Anything that has occurred or been released after that date may not be included in my responses unless I use real-time browsing. Let me know if you'd like me to look up more recent information.

My knowledge was last updated in April 2024. When discussing events or developments, I approach it as a well-informed individual from April 2024 would if speaking to someone in October 2024. If you have questions about more recent events, I'll do my best to help based on trends and information available up to my knowledge cutoff, but I may not be aware of specific developments that occurred after April 2024.

Training Data

Overview Llama 3 was pretrained on over 15 trillion tokens of data from publicly available sources. The fine-tuning data includes publicly available instruction datasets, as well as over 10M human-annotated examples. Neither the pretraining nor the fine-tuning datasets include Meta user data.

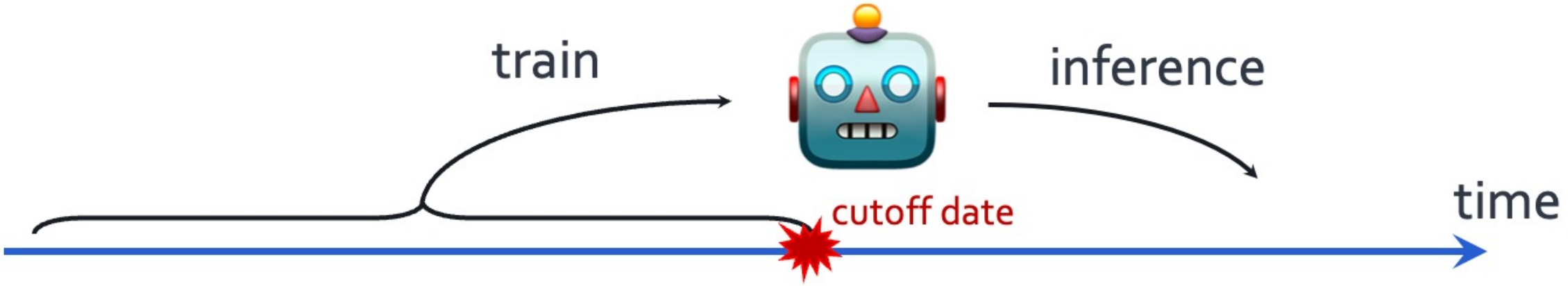
Data Freshness The pretraining data has a cutoff of March 2023 for the 8B and December 2023 for the 70B models respectively.

Overview: Llama 3.1 was pretrained on ~15 trillion tokens of data from publicly available sources. The fine-tuning data includes publicly available instruction datasets, as well as over 25M synthetically generated examples.

Data Freshness: The pretraining data has a cutoff of December 2023.

Temporal misalignment: LLMs stale over time

- Known: Their quality degrade **after** their cut off date.

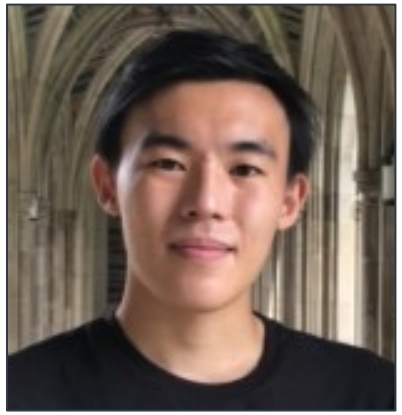


Are LLMs' knowledge before cutoff date consistently good?

Luu et al. Time Waits for No One! Analysis and Challenges of Temporal Misalignment., *NAACL* 2022
Lazaridou et al. Mind the Gap: Assessing Temporal Generalization in Neural Language Models, *NeurIPS* 2021

Dated Data:
Tracing Knowledge Cutoffs in Large Language Models

Jeffrey Cheng, Marc Marone, Orion Weller,
Dawn Lawrie, Daniel Khashabi, Benjamin Van Durme

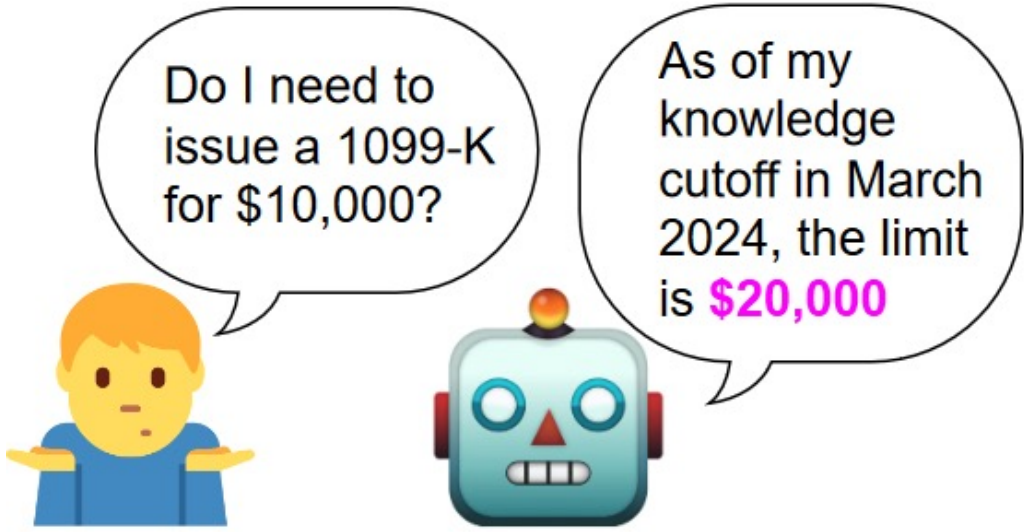


🥇 COLM 2024 Outstanding paper award! 🥇

https://arxiv.org/abs/2403.12958

Knowledge before the [claimed] cutoff date

- Do all resources in the training data share the same reported knowledge cutoff?



2022 IRS

Form 1099-K is issued for transactions only if the aggregate amount of these transactions exceeded **$20,000**

2024 IRS

Now a single transaction exceeding **$5000** can require the third party platform to issue a 1099-K.

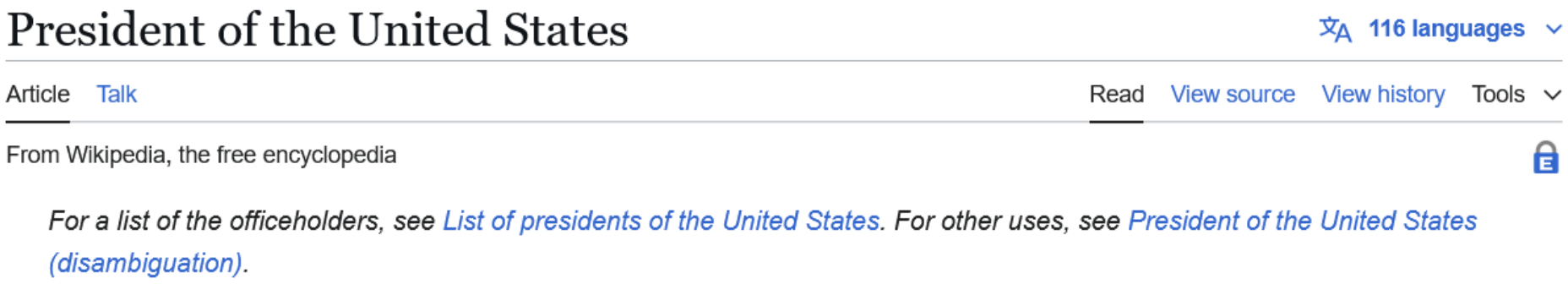
Knowledge before the [claimed] cutoff date

- Do all resources in the training data share the same reported knowledge cutoff?

The *effective cutoff* of an LLM (with respect to a resource) is the date that matches the LLM's best knowledge of that resource.

Effective cutoff == reported cutoffs?

- It's possible that they're not the same.
  - Claimed cutoff shows the latest version of data.
  - Effective cutoff depends on how much old/stale information are there.



President of the United States

116 languages

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

For a list of the officeholders, see List of presidents of the United States. For other uses, see President of the United States (disambiguation).

The president is elected indirectly through the Electoral College to a four-year term, along with the vice president. Under the Twenty-second Amendment, ratified in 1951, no person who has been elected to two presidential terms may be elected to a third. In addition, nine vice presidents have become president by virtue of a president's intra-term death or resignation.[C] In all, 45 individuals have served 46 presidencies spanning 58 four-year terms.[D] Joe Biden is the 46th and current president, having assumed office on January 20, 2021.

Through the Electoral College, registered voters indirectly elect the president and vice president to a four-year term. This is the only federal election in the United States which is not decided by popular vote.[19] Nine vice presidents became president by virtue of a president's intra-term death or resignation.[C]

Donald Trump of New York is the 45th and current president of the United States. He assumed office on January 20, 2017.

How do we measure knowledge over time?

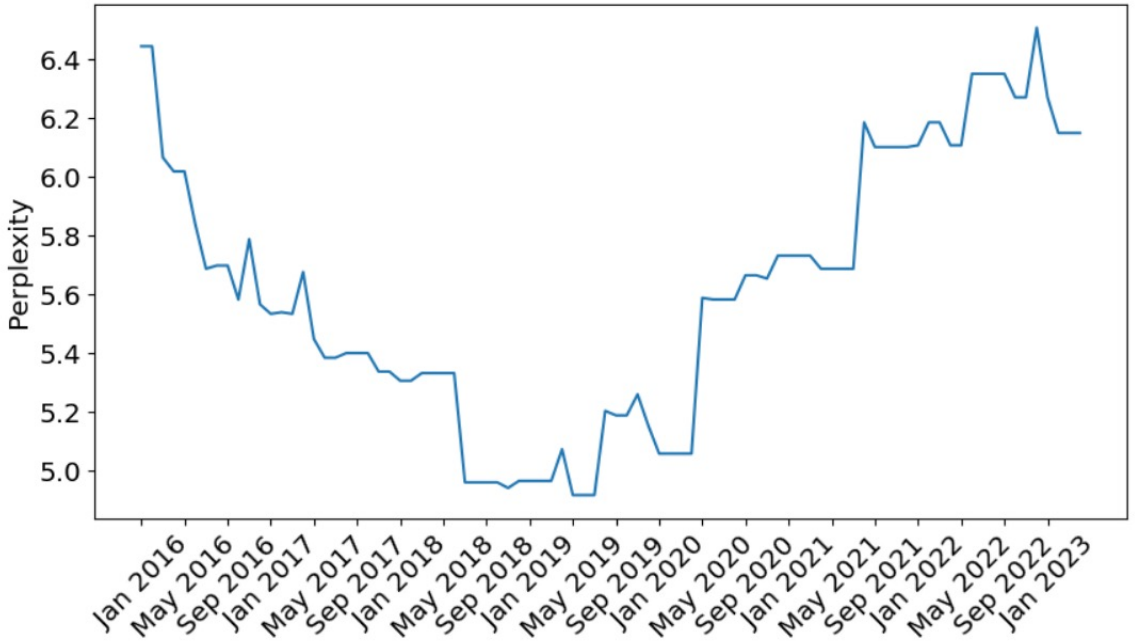
- WIKISPAN:
  - Collect 5000 most edited topics
  - Scrape monthly versions from April 2016 to April 2023

- We also built NEWSPAN based on New York Times articles. Feel free to checkout the details in the paper.

Extracting PPL over time with WIKISPAN

- WIKISPAN documents: version of Wikipedia topic *t* at time *m*
- Measure perplexity of first 512 tokens of each document, across all topics and months

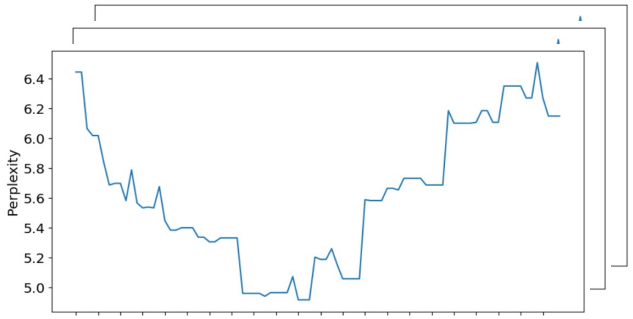
$$\mathrm{PPL}(X) = \exp\left(-\frac{1}{t}\sum_{i=1}^{t}\log p_\theta(x_i|x_{<i})\right)$$



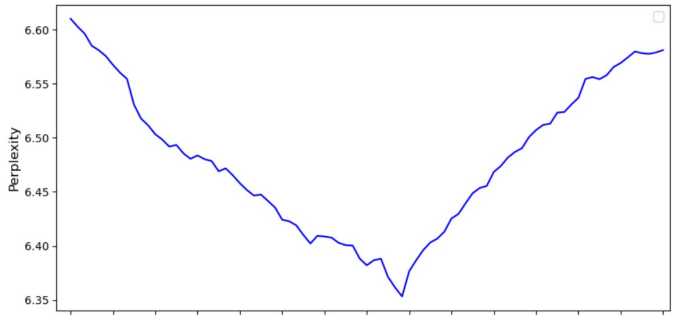
Perplexity of the Wikipedia document "Liverpool" under Pythia-7b. Each point is the perplexity of the document at that time.

Extracting PPL over time with WIKISPAN

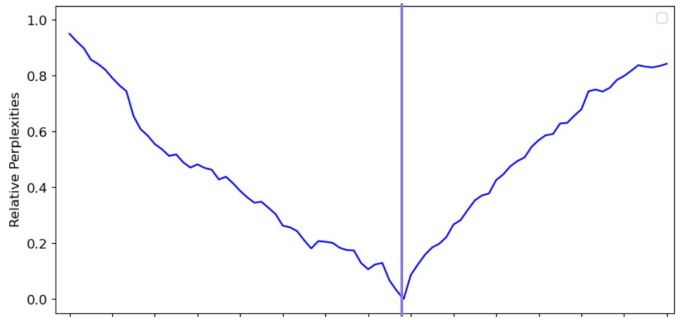
- Normalization
  - Aggregate perplexities with 95% truncated mean within each month
  - Perform 0-1 normalization over entire time-span
- Effective knowledge cutoffs are the argmin of relative perplexity curves



Perplexity measurements of documents



Aggregate along month axis with truncated mean



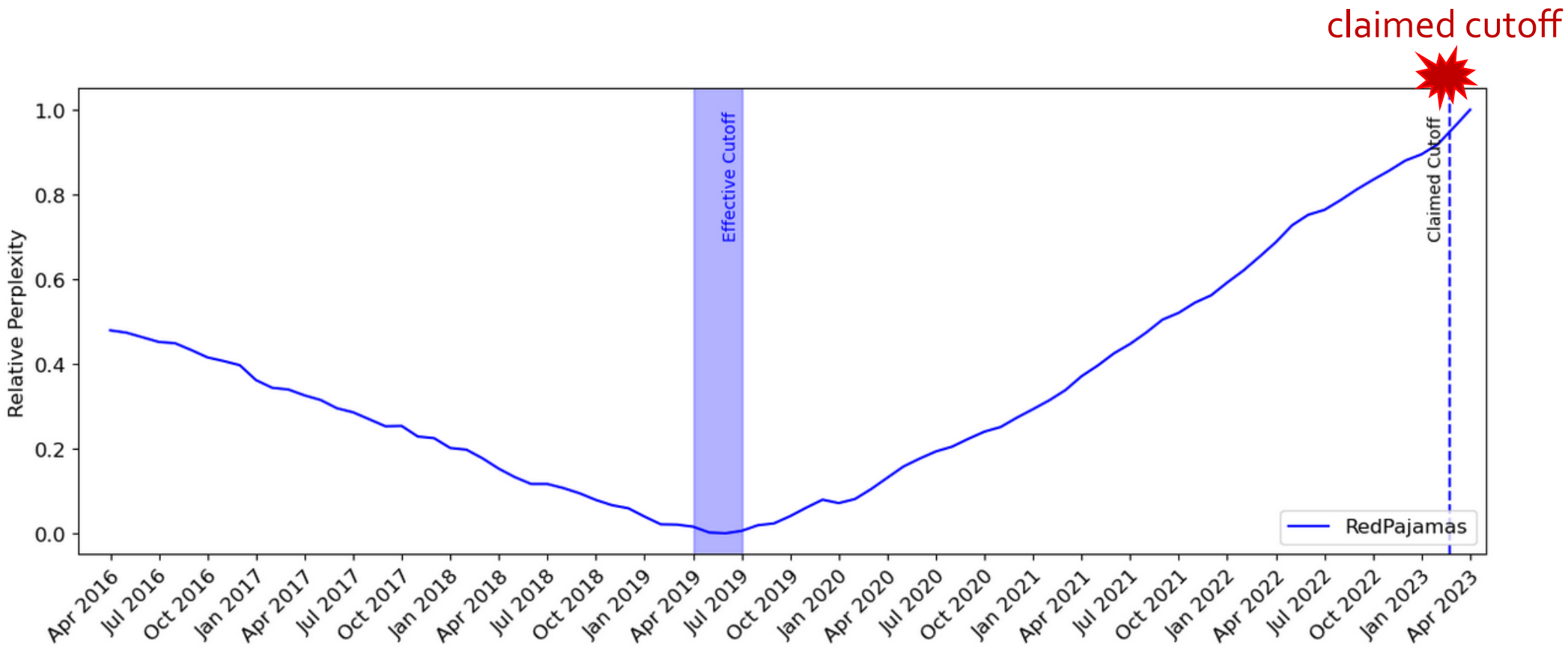
Convert to relative perplexities by 0-1 scaling

C4-derived models on WIKISPAN

RedPajamas (Together Computer)

"We use the Wikipedia dataset available on Huggingface, which is based on the Wikipedia dump from 2023-03-20 and contains text in 20 different languages. The dataset comes in preprocessed format, so that hyperlinks, comments and other formatting boilerplate has been removed."

C4-derived models on WIKISPAN

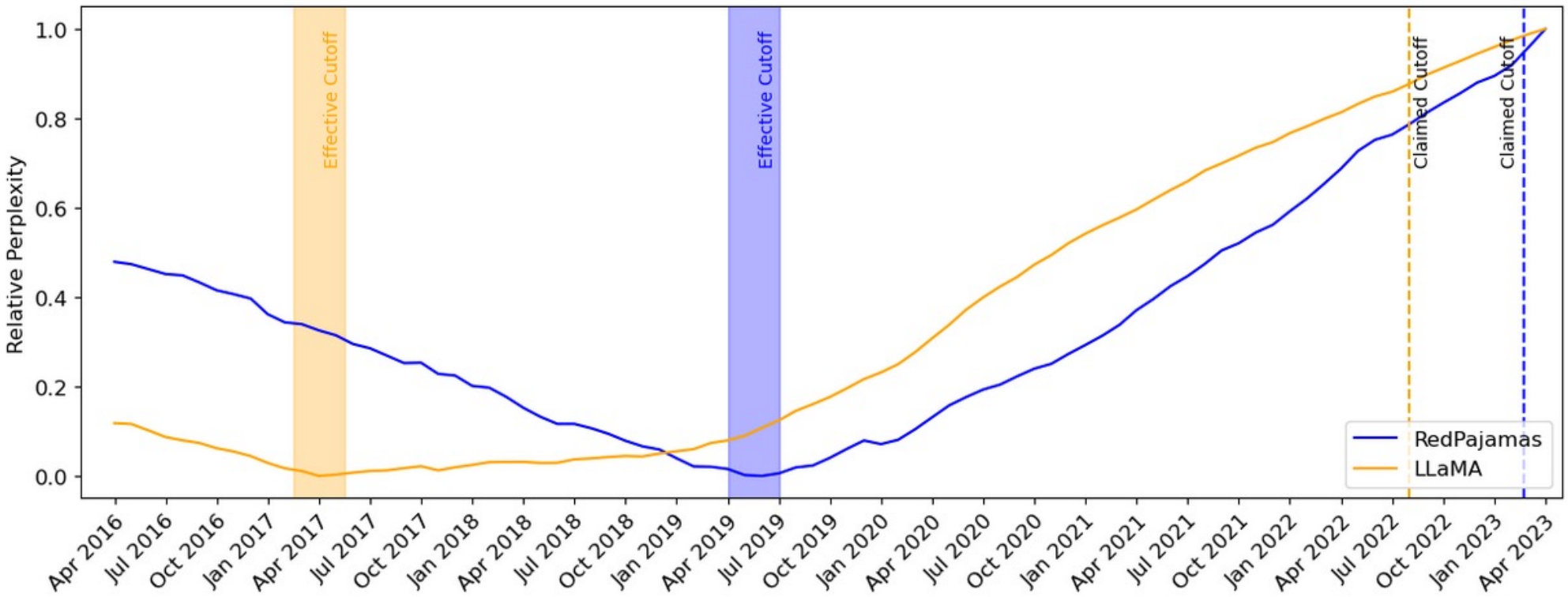


C4-derived models on WIKISPAN

LLaMA (Meta)

"We add Wikipedia dumps from the June-August 2022 period, covering 20 languages, which use either the Latin or Cyrillic scripts: bg, ca, cs, da, de, en, es, fr, hr, hu, it, nl, pl, pt, ro, ru, sl, sr, sv, uk. We process the data to remove hyperlinks, comments and other formatting boilerplate."

C4-derived models on WIKISPAN

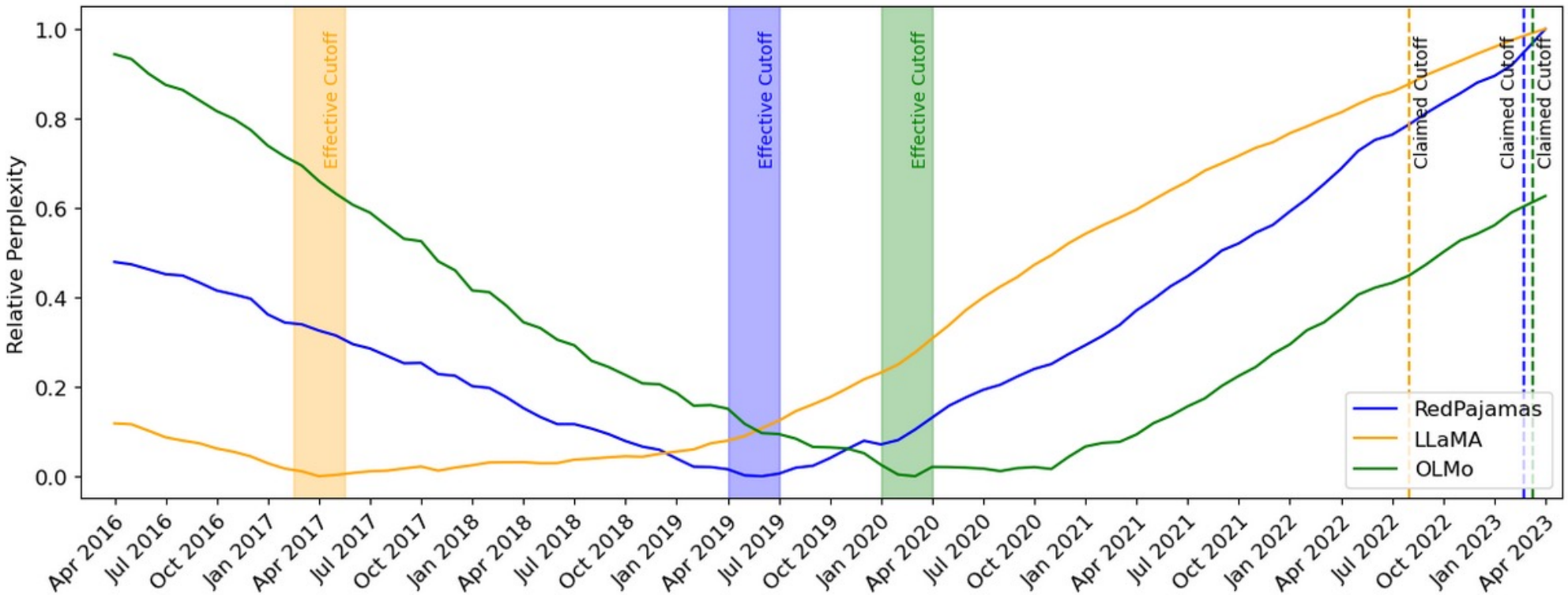


C4-derived models on WIKISPAN

OLMo (Al2)

"Dumps were downloaded from Wikimedia's website. We use the dump from March 20th, 2023."

C4-derived models on WIKISPAN



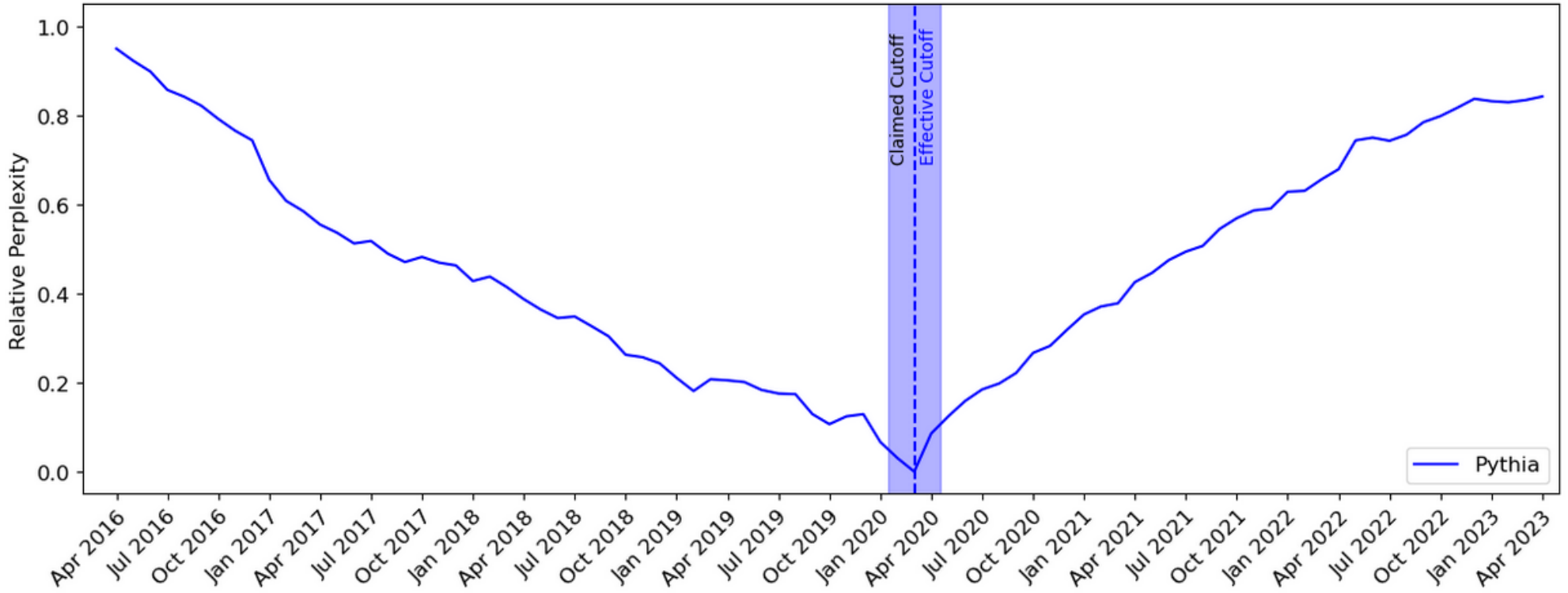
Effective cutoff date can be much earlier than the claimed cutoff date.

Pile-derived models on WIKISPAN

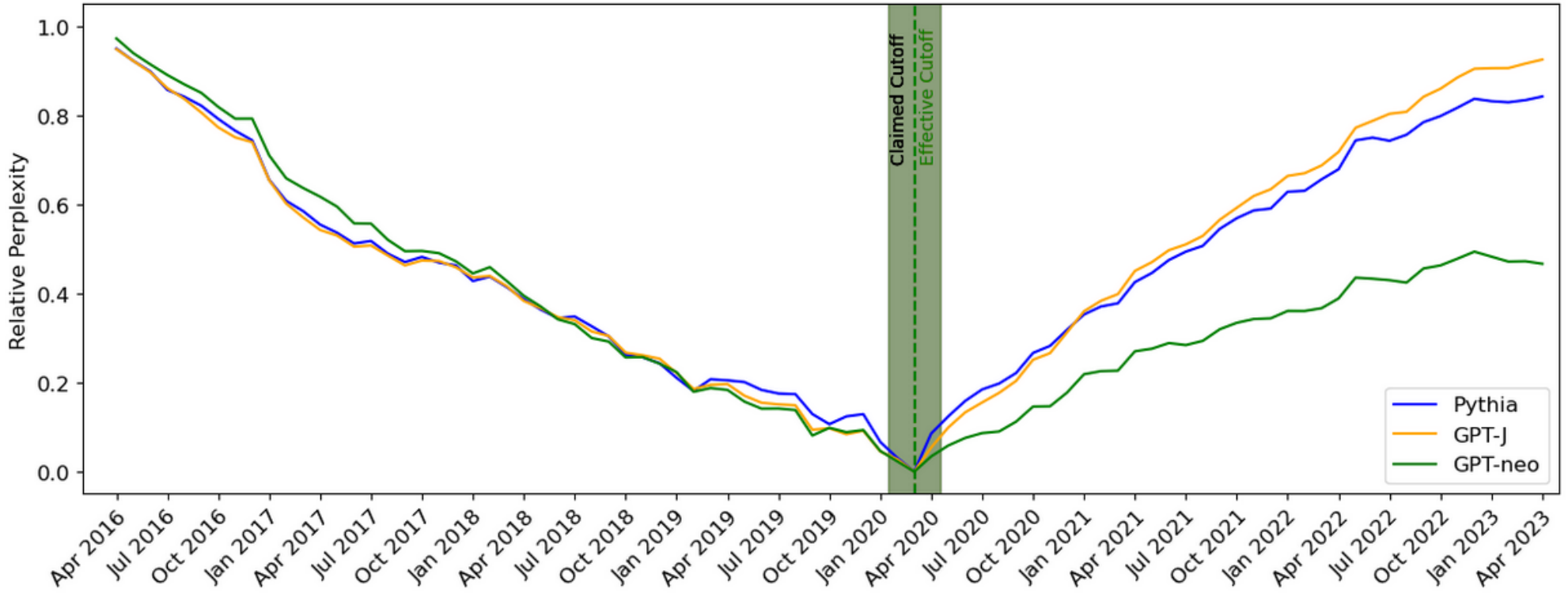
Pile (EleutherAI)

"We use the wikipedia/20200301.en dataset from TensorFlow Datasets. We prepend the title to the body of each article, separated by two newlines."

Pile-derived models on WIKISPAN



Pile-derived models on WIKISPAN



Effective cutoff and the claimed cutoff date can be the same!

Why do there exist discrepancies between effective and reported cutoffs?

Factors contributing to the misalignment between effective and reported cutoff:

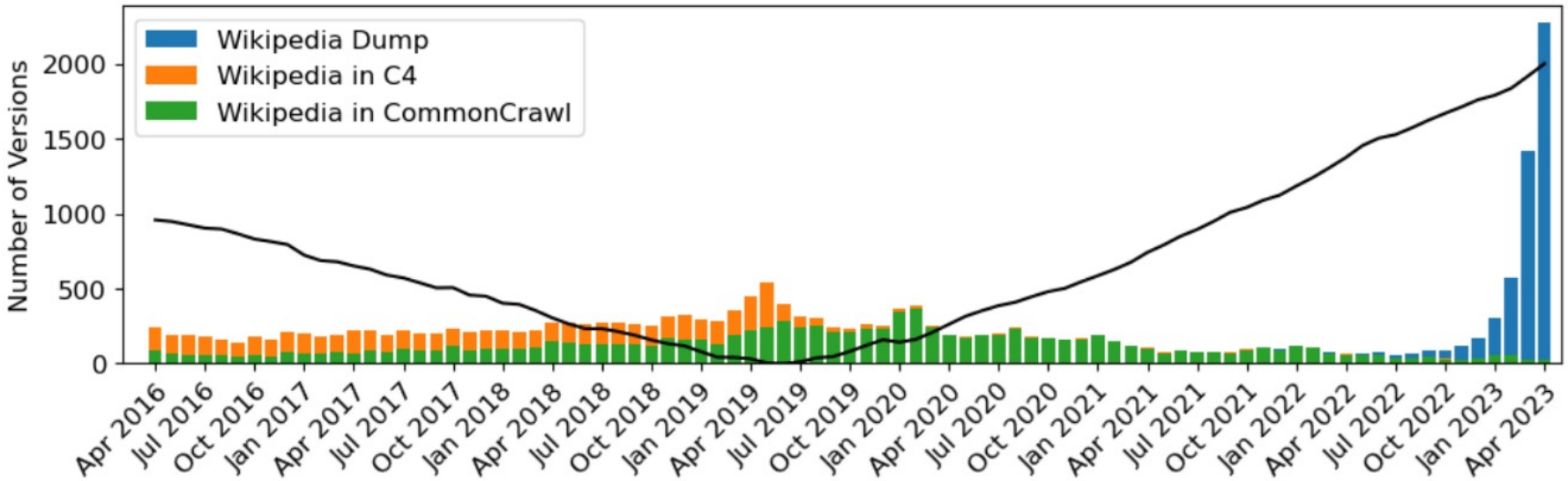
- Prevalence of stale data
- Complications in deduplication pipelines

Factors contributing to the misalignment between effective and reported cutoff:

- Prevalence of stale data
- Complications in deduplication pipelines

Prevalence of stale data

- RedPajamas is trained on: C4, CommonCrawl dumps and Wikipedia
- Breakdown of RedPajamas training corpus:



CommonCrawl dumps contain old data that bias effective cutoffs.

Factors contributing to the misalignment between effective and reported cutoff:

- Prevalence of stale data
- Complications in deduplication pipelines

Deduplication Issues

- It is common practice to deduplicate pre-training datasets
  - Fuzzy deduplication should remove different versions of Wikipedia documents
  - Exact deduplication should remove exact copies of Wikipedia document
- We empirically find many duplicates in pretraining datasets!

By the end of the 17th century, the Chinese economy had recovered from the devastation caused by the wars in which the Ming dynasty were overthrown, and the resulting breakdown of order.[147] In the following century, markets continued to expand as in the late Ming period, but with more trade between regions, a greater dependence on overseas markets and a greatly increased population.[148][149] The government broadened land ownership by returning land that had been sold to large landowners in the late Ming period by families unable to pay the land tax.[150] To give people more incentives to participate in the market, they reduced the tax burden in comparison with the late Ming, and replaced the corvée system with a head tax used to hire laborers.[151] The administration of the Grand Canal was made more efficient, and transport opened to private merchants.[152] A system of monitoring grain prices eliminated severe shortages, and enabled the price of rice to rise slowly and smoothly through the 18th century.[153] Wary of the power of wealthy merchants, Qing rulers limited their trading licenses and usually refused them permission to open new mines, except in poor areas ...

By the end of the 17th century, the Chinese economy had recovered from the devastation caused by the wars in which the Ming dynasty were overthrown, and the resulting breakdown of order.[148] In the following century, markets continued to expand as in the late Ming period, but with more trade between regions, a greater dependence on overseas markets and a greatly increased population.[149][150] The government broadened land ownership by returning land that had been sold to large landowners in the late Ming period by families unable to pay the land tax.[151] To give people more incentives to participate in the market, they reduced the tax burden in comparison with the late Ming, and replaced the corvée system with a head tax used to hire laborers.[152] The administration of the Grand Canal was made more efficient, and transport opened to private merchants.[153] A system of monitoring grain prices eliminated severe shortages, and enabled the price of rice to rise slowly and smoothly through the 18th century.[154] Wary of the power of wealthy merchants, Qing rulers limited their trading licenses and usually refused them permission to open new mines, except in poor areas ...

Summary thus far

- There **do** exist discrepancies between effective and reported knowledge cutoffs in modern LLMs

- Effective cutoffs of modern LLMs are years earlier than reported cutoff
  - CommonCrawl dumps include older versions of resources
  - Old versions and their duplicates are not removed by deduplication pipelines

- Open problem: how should we strike a balance between data coverage and recency?

Today

Stale data

- A ton world data is old.
  How does that impact our models?

Imbalanced data

- Most data follow a long-tail distribution.
  How should we deal with it?

Today

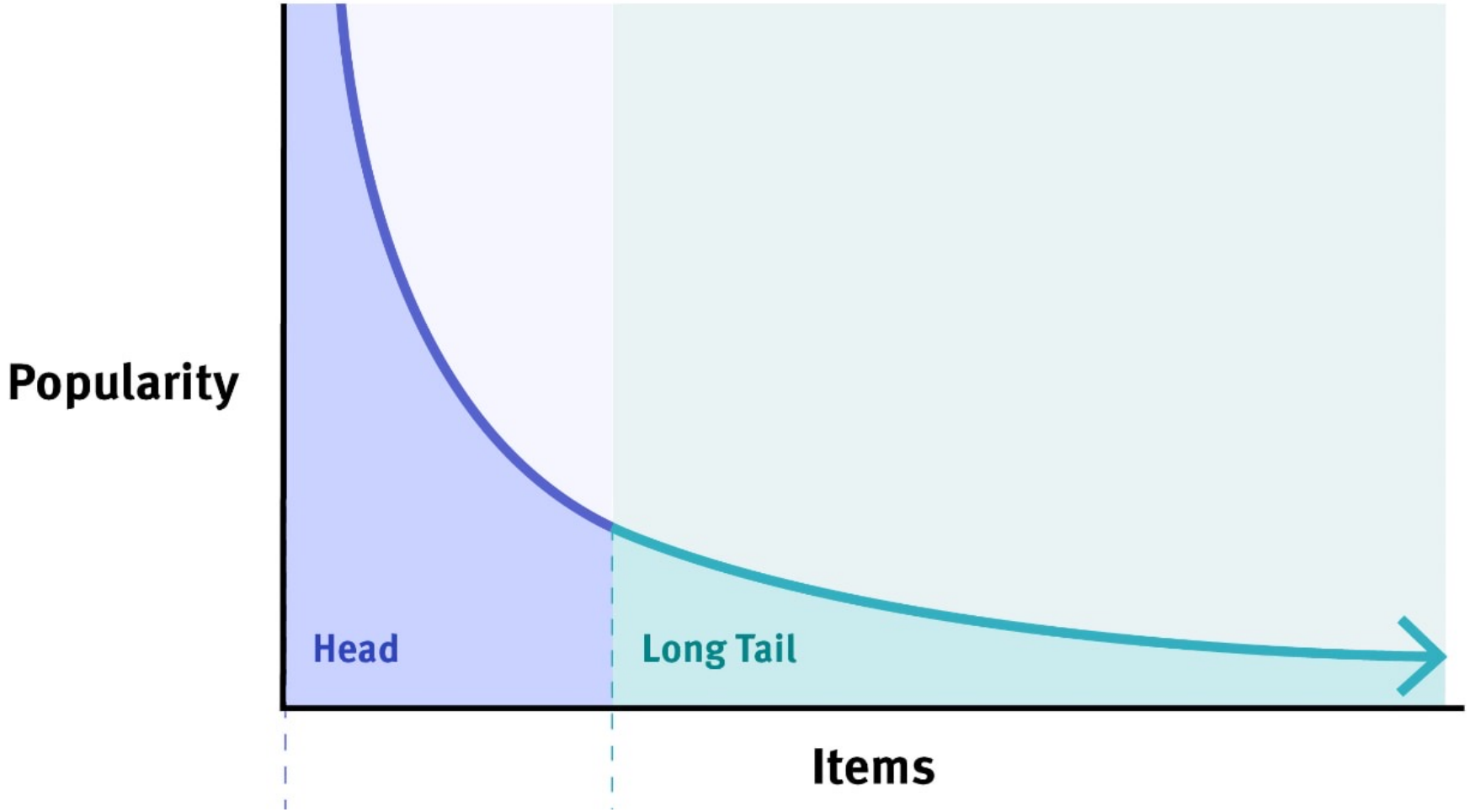
Stale data

- A ton world data is old.
  How does that impact our models?

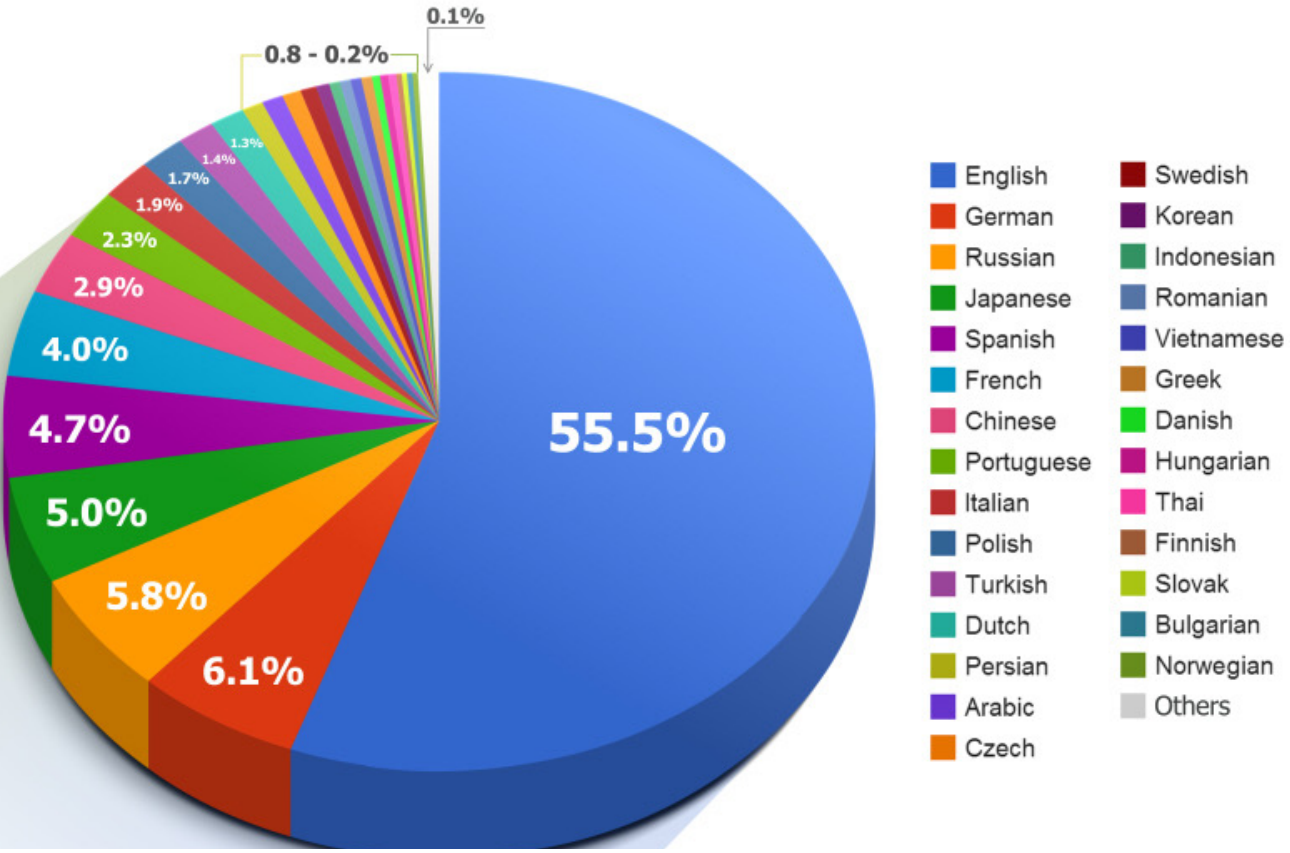
Imbalanced data

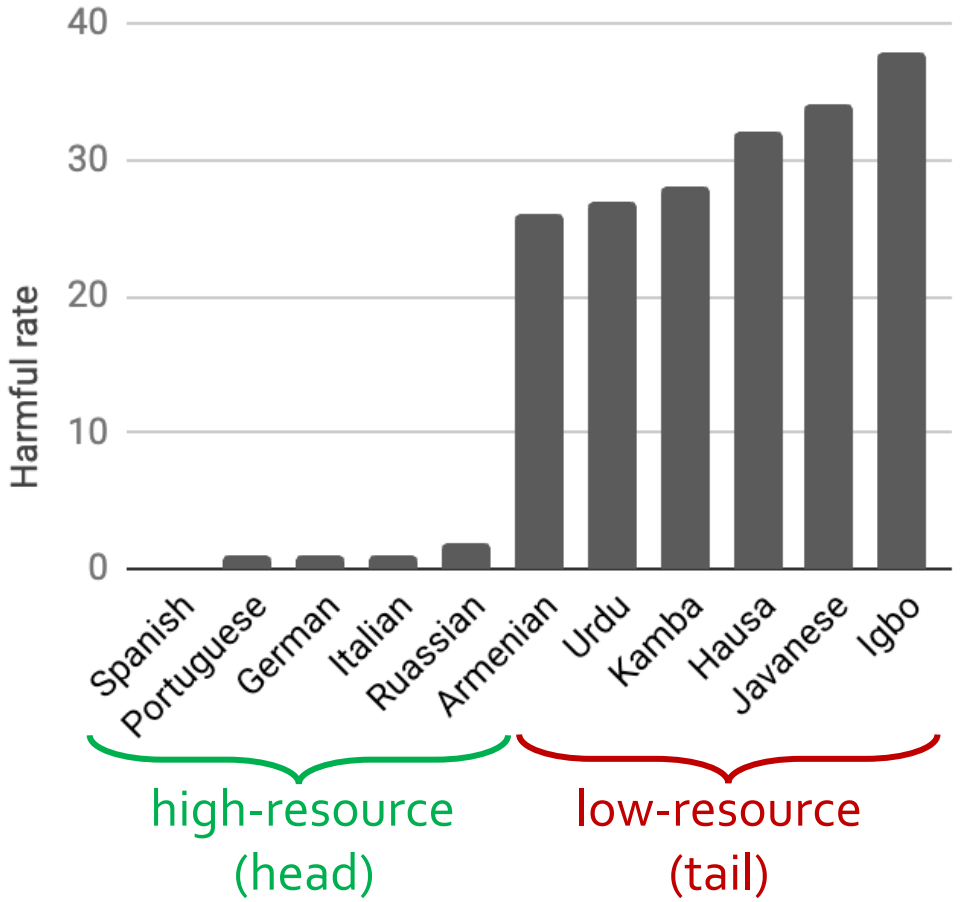
- Most data follow a long-tail distribution.
  How should we deal with it?

Long-tail of problems:
There are many infrequent concepts/problems



The long tail of languages



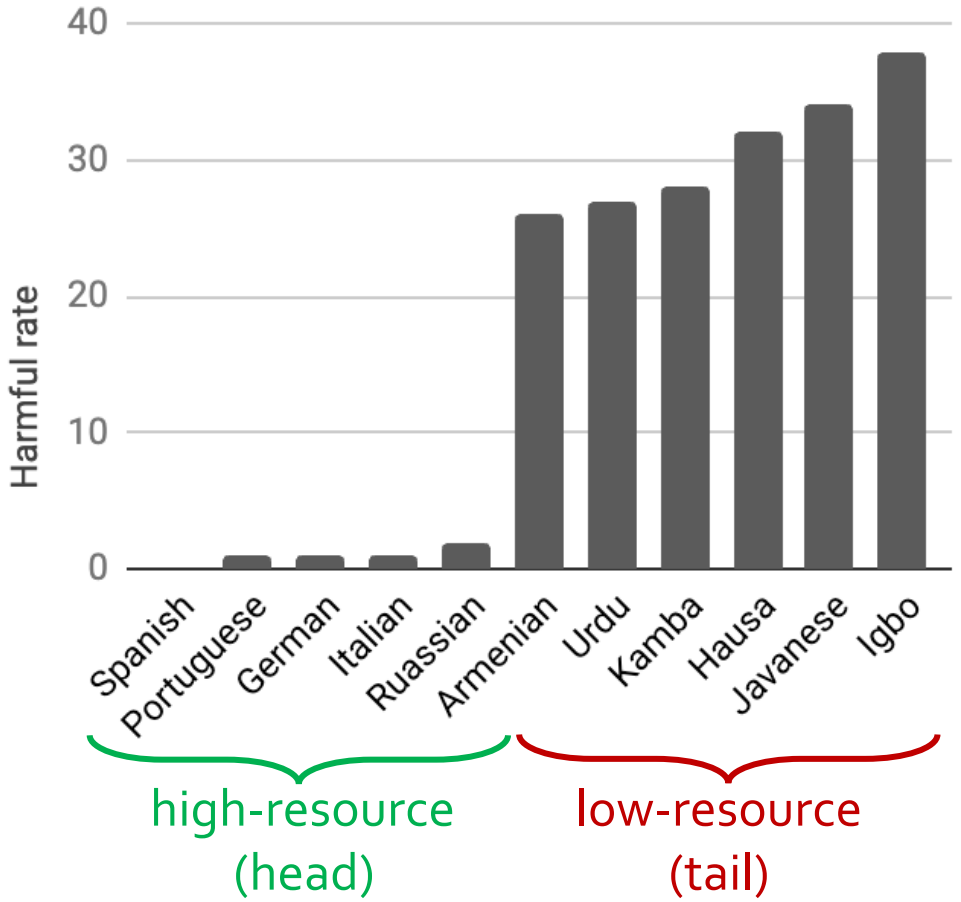


https://commons.wikimedia.org/wiki/File:2014_Distribution_of_Languages_on_Internet_Websites.jpg

Shen et al. The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Context., *ACL* 2024

The long tail of languages

- What is the effective way to train models on such imbalanced data?



Shen et al. The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Context., *ACL* 2024

Upsample or Upweight?
Balanced Training on Heavily Imbalanced Datasets

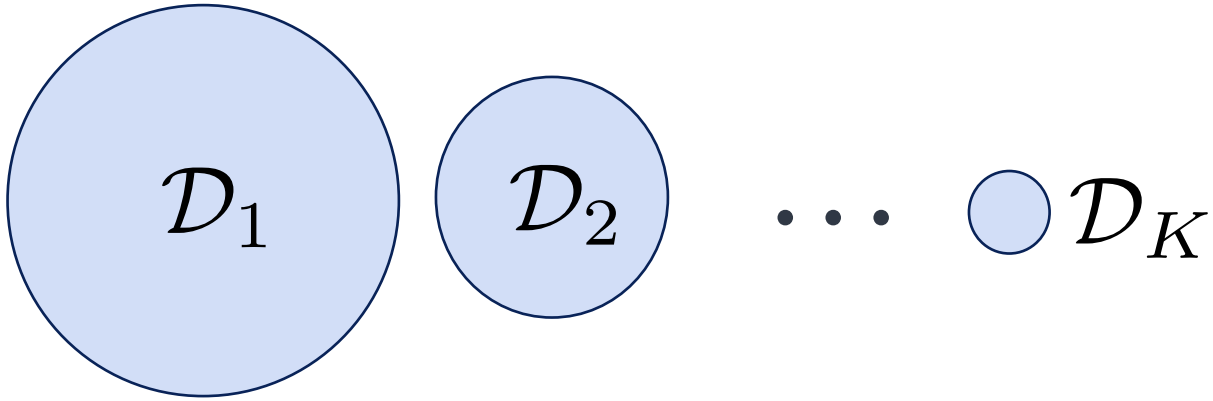
Tianjian Li, Haoran Xu, Weiting Tan,
Kenton Murray and Daniel Khashabi.

https://arxiv.org/abs/2410.04579
(under review)

Training on a collection of "domains"

- Consider training a model on a collection K domains:

$$\mathcal{D}_{\text{total}} \triangleq \mathcal{D}_1 \cup \mathcal{D}_2 \cup \ldots \cup \mathcal{D}_K$$



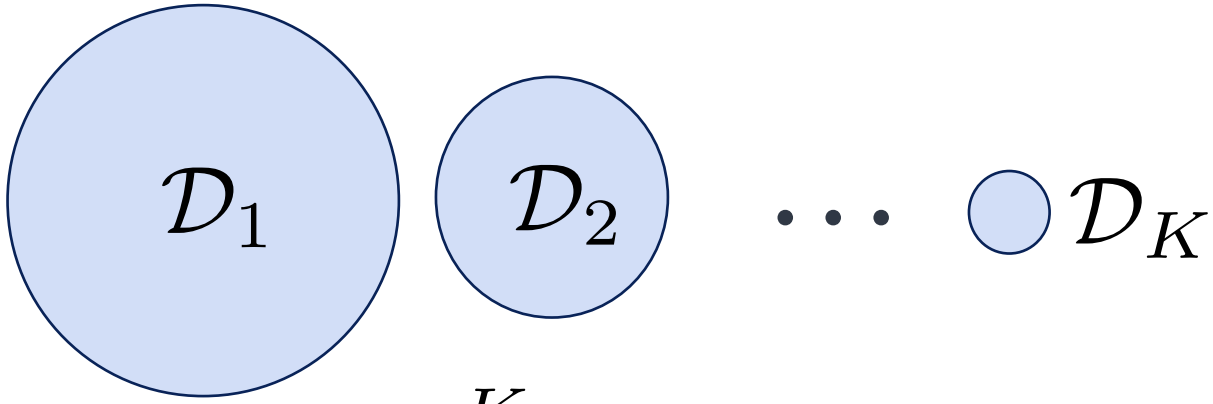
- Loss, treating all domains uniformly:

$$\sum_{x \in \mathcal{D}_{\text{total}}} \ell(x)$$

Scalarization (S)

- Define a per-domain weights (scalars) to adjust their weights:

$$w_1 \quad w_2 \quad \dots \quad w_K$$



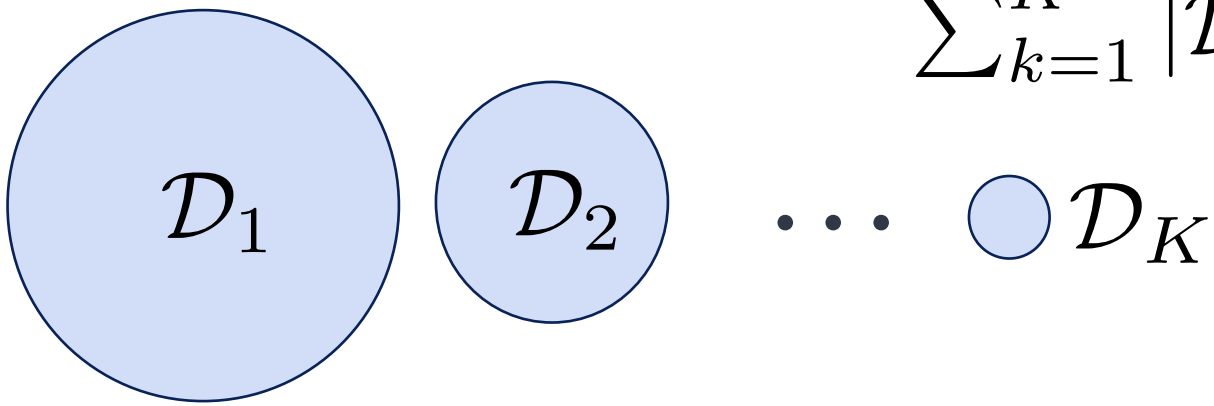
$$L_{\mathrm{S}} = \sum_{k=1}^{K} w_k \sum_{x \in \mathcal{D}_k} \ell(x)$$

Assigned a higher weight to the smaller/harder domains

Temperature Sampling (TS)

- Define a probability distribution for sampling instances:

$$\forall i \in \{1, 2, ..., K\} : \; p(i; \tau) = \frac{|\mathcal{D}_i|^{\frac{1}{\tau}}}{\sum_{k=1}^{K} |\mathcal{D}_k|^{\frac{1}{\tau}}},$$

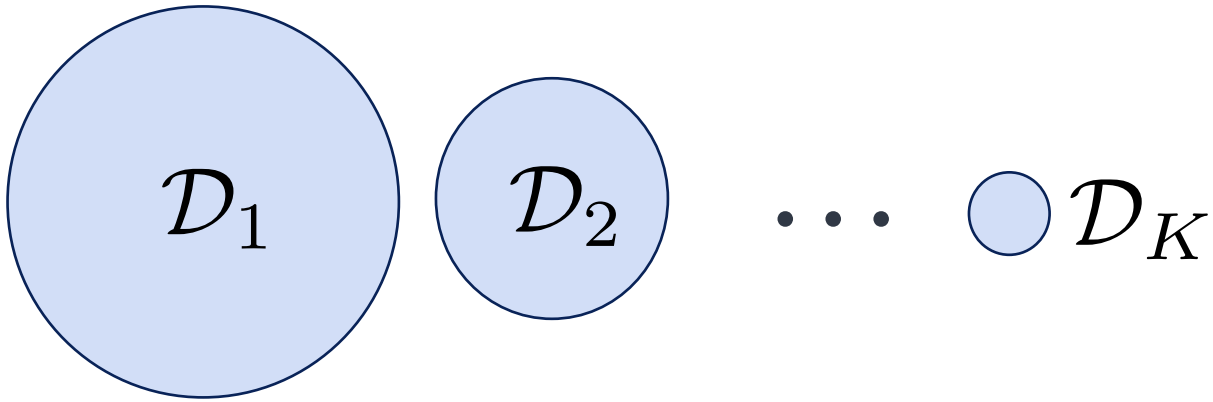


$$L_{\text{TS}} = \mathop{\mathbb{E}}_{\substack{k \sim p \\ x \sim \mathcal{D}_k}} \Big[\mathcal{L}(x) \Big]$$

Temperature Sampling vs. Scalarization

$$L_{\mathrm{S}} = \sum_{k=1}^{K} w_k \sum_{x \in \mathcal{D}_k} \ell(x)$$

$$L_{\mathrm{TS}} = \mathop{\mathbb{E}}_{\substack{k \sim p \\ x \sim \mathcal{D}_k}} \Big[\mathcal{L}(x) \Big]$$



Is Temperature Sampling == Scalarization?

- The common understanding is that these two are equivalent.

In our work, we follow convention and implement scalarization via proportional sampling, where data from task i is sampled with probability equal to w_i. In this case, the expected loss is equal to the loss from scalarization:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}}\left[\ell(\boldsymbol{x};\boldsymbol{\theta})\right] = \sum_{i=1}^{K} \mathbb{P}(\text{task } i)\mathbb{E}[\ell(\boldsymbol{x};\boldsymbol{\theta})|\boldsymbol{x} \in \text{task } i] = \sum_{i=1}^{K} w_i \mathcal{L}_i(\boldsymbol{\theta}). \quad (2)$$

Choi et al. Order Matters in the Presence of Dataset Imbalance for Multilingual Learning, *NeurIPS* 2023

Spoiler: they're not the same!

frontier of scalarization. Following the literature's convention, we implement scalarization via proportional sampling. Here, the number of observations in the batch corresponding to task i is proportional to w_i. In this case, the expected training loss is equal to

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}}[\ell(\boldsymbol{x};\boldsymbol{\theta})] = \sum_{i=1}^{K} \mathbb{P}(\boldsymbol{x} \in \text{task } i)\mathbb{E}_{\boldsymbol{x}}[\ell(\boldsymbol{x};\boldsymbol{\theta})|\boldsymbol{x} \in \text{task } i] = \sum_{i=1}^{K} w_i \mathcal{L}_i(\boldsymbol{\theta}).$$

Xin et al. Do Current Multi-Task Optimization Methods in Deep Learning Even Help?, *NeurIPS* 2022

Xie et al. (NeurIPS, 2023) and Fan et al. (ICML, 2024) implement TS via S.

TS-loss = S-loss

Theorem: For any sampling temperature τ, there exists a set of weights $\{w_1, w_2, \ldots, w_K\}$ for which S-loss is equivalent to TS-loss (on the whole data).

$$L_{\mathrm{TS}} = \mathop{\mathbb{E}}_{\substack{k \sim p \\ x \sim \mathcal{D}_k}} \Big[\mathcal{L}(x) \Big] \qquad L_{\mathrm{S}} = \sum_{k=1}^{K} w_k \sum_{x \in \mathcal{D}_k} \ell(x)$$

TS gradients have lower variance than S

Theorem: Gradient estimates of S-loss have higher variance than those by TS-loss:

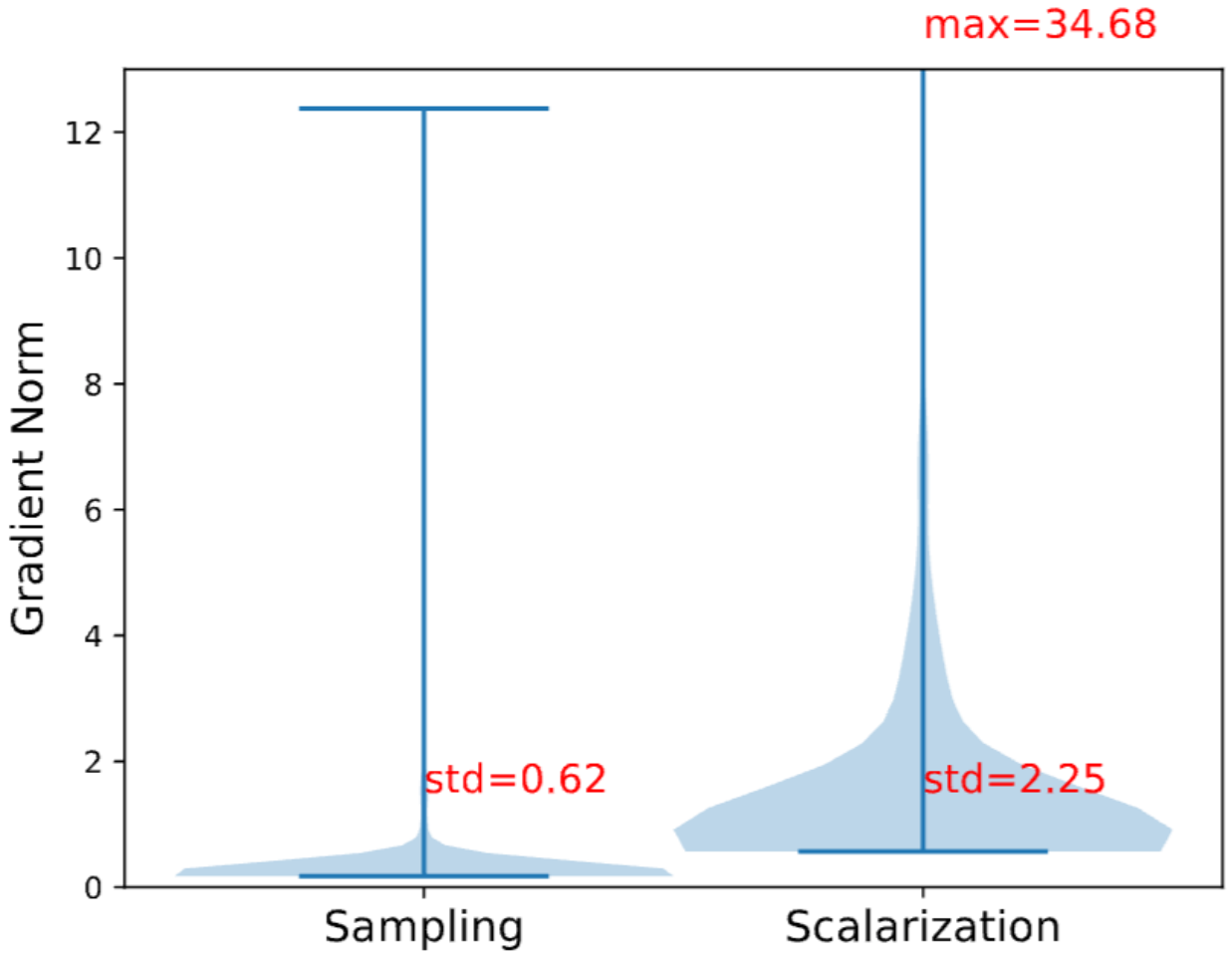
$$\mathrm{Var}(\nabla L_{\mathrm{S}}(x)) \geq \mathrm{Var}(\nabla L_{\mathrm{TS}}(x))$$

Theorem: The variance gap

$$\delta = \mathrm{Var}(\nabla L_{\mathrm{S}}(x)) - \mathrm{Var}(\nabla L_{\mathrm{TS}}(x))$$

increases monotonically for $\tau \geq 1$.

TS vs S gradients: empirical evidence



Summary thus far

- Training on imbalanced data:
  - Scalarization: weighting domains
  - Temperature Sampling: resampling from domains

- Despite the common perception, these two are not equivalent.

- TS offers gradients estimates with lower variance.

- Next: how is this useful for model training?

TS vs S: How do they affect my optimization?

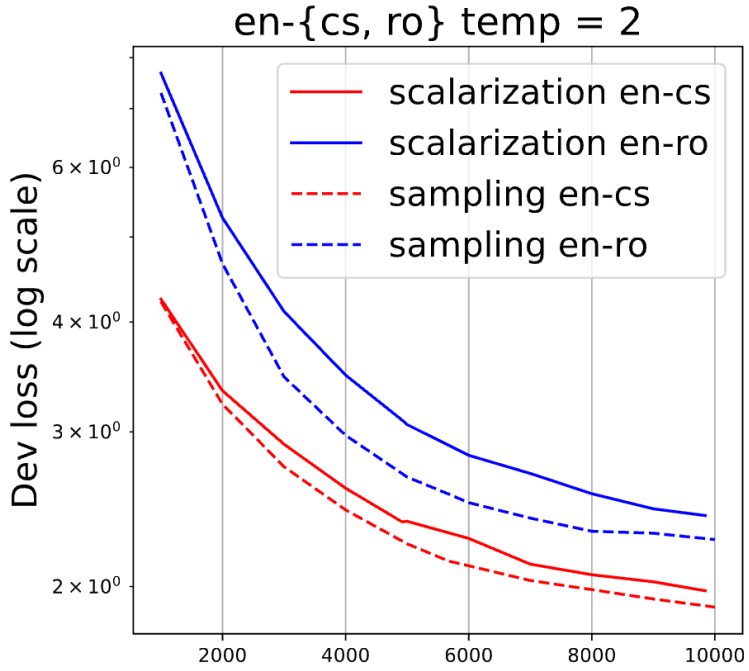
- It is well-known that variance-reduction accelerates the convergences of SGD (Sutskever et al., 2013; Kingma and Ba, 2015)

- TS provides lower variance gradient estimates than S.

Hypothesis. Temperature Sampling (at higher temperature) converges [much] faster than Scalarization on heavily imbalanced domain distributions.

Temperature Sampling's faster convergence

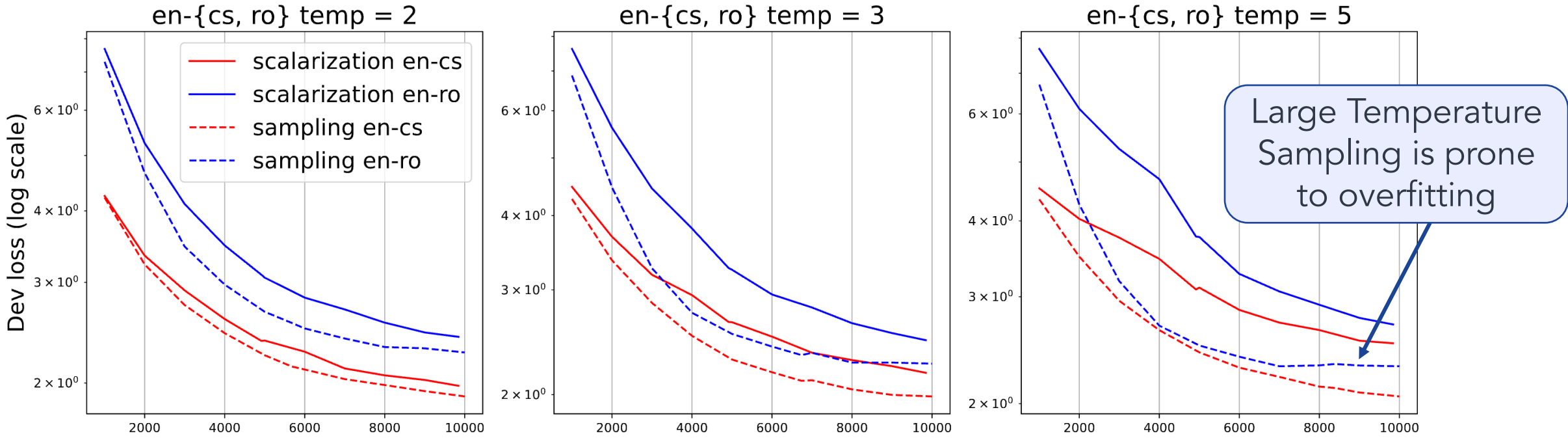
- We train an MT model for pair of high and low-resource languages.



Temperature Sampling converges faster than Scalarization.

Temperature Sampling's faster convergence

- We train an MT model for pair of high and low-resource languages.



Larger temperature --> faster convergence for Temperature Sampling (dashed)

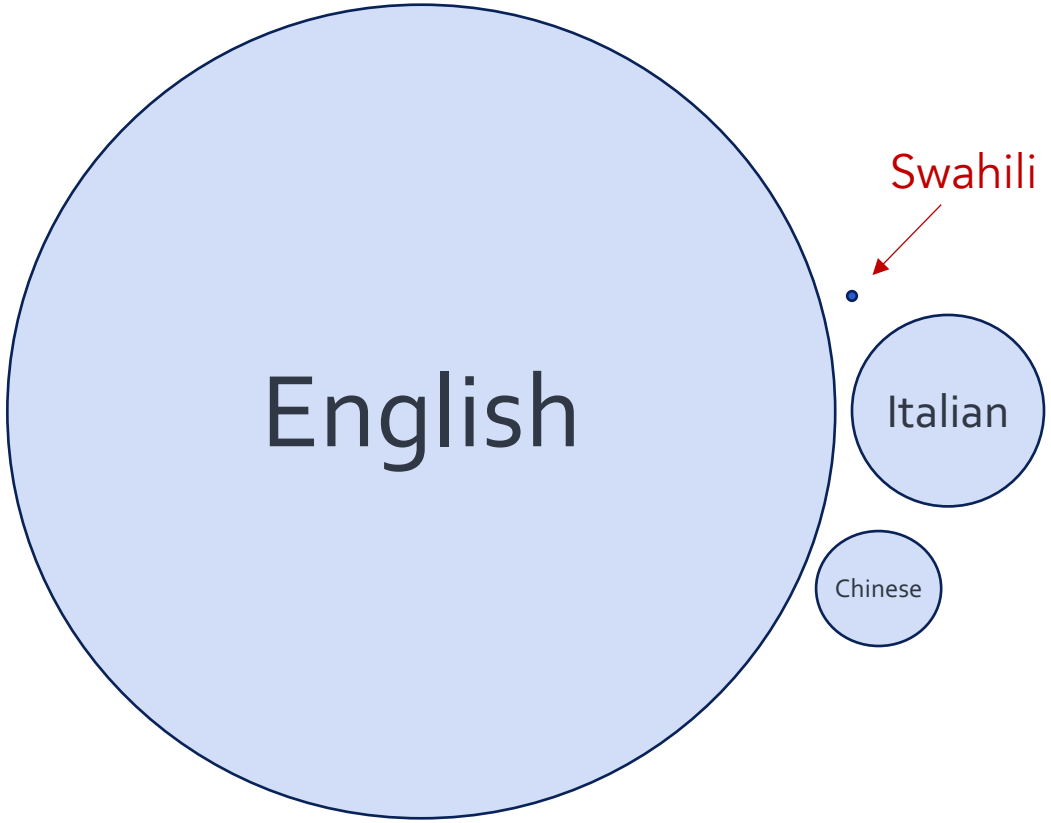
Summary thus far

- Temperature Sampling converges faster than Scalarization.
- If temperature is set too high, this faster convergence might lead to overfitting.

- Next: how is this useful for training on imbalanced data?

Training on a collection of “domains”

Tokens of pretraining data by language in mC4 (Xue+ 2024)



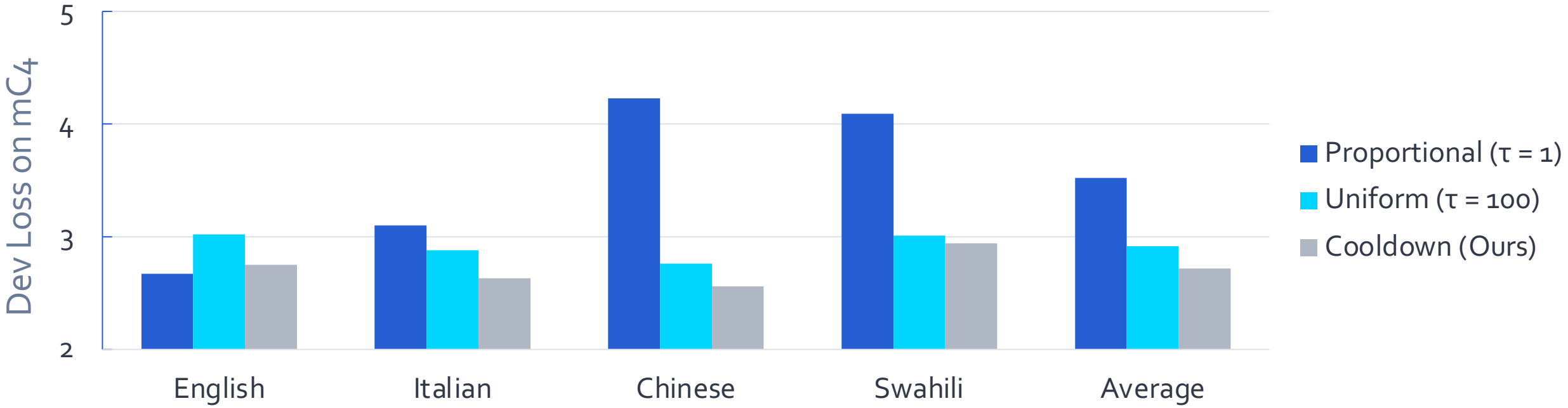
How should we [pre-]train on imbalanced data?

- We need to strike a good balance between
  - High-temperature (faster convergence)
  - Low-temperature (avoid overfitting)

❄ CoolDown ❄

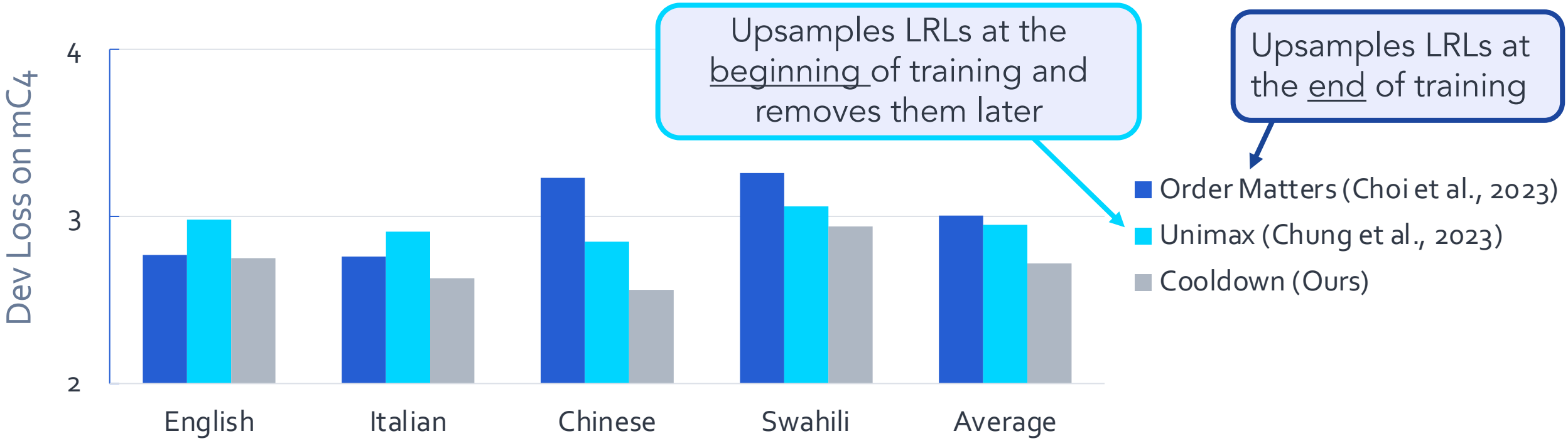
Train with Temperature Sampling
- Start with a high-temperature
- Over the course of training

Multilingual Language Modeling



CoolDown performs better than fixed-temperature (high or low) sampling.

Multilingual Language Modeling



CoolDown performs better than fixed-temperature (high or low) sampling.

Summary

- Two common approaches for dealing with imbalanced data:
  - Temperature sampling
  - Scalarization

- Despite common perception, these two are not equivalent.
  - Temperature sampling leads to lower-variance grad estimates
  - ... and faster convergence.

- ❄ CoolDown ❄
  - A suggested recipe for imbalanced [pre-]training

Overall conclusions

Stale data

- Effective cutoffs of LLMs are years earlier than reported cutoff.
- Old data are not removed by deduplication pipelines.
- Open problem: how should we strike a balance between data coverage and recency?

Imbalanced data

- Resampling is a competitive strategy—better than weighting.
- An effective resampling (in quality and speed) requires adaptively adjusting the sampling distribution (temperature).

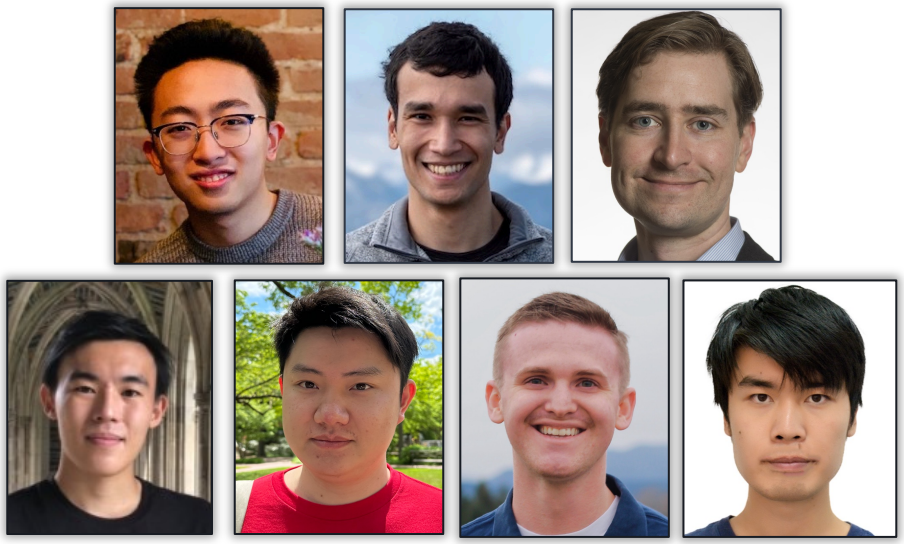
Data! Data! Data! … I can’t make bricks without clay!

---Sherlock Holmes

- Understanding data (and how to use it effectively) is difficult but necessary for our progress.

Thanks for wonderful collaborators on these projects:

Johns Hopkins University



EleutherAI

University of Washington

Funding:

Ai2

amazon science

Office of Naval Research — Science & Technology