

Science of AI and AI for Science

Daniel Khashabi



JOHNS HOPKINS
UNIVERSITY

Go Quakers!!



REASONING-DRIVEN QUESTION-ANSWERING
FOR NATURAL LANGUAGE UNDERSTANDING

Daniel Khashabi

A DISSERTATION

in

Computer and Information Sciences

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

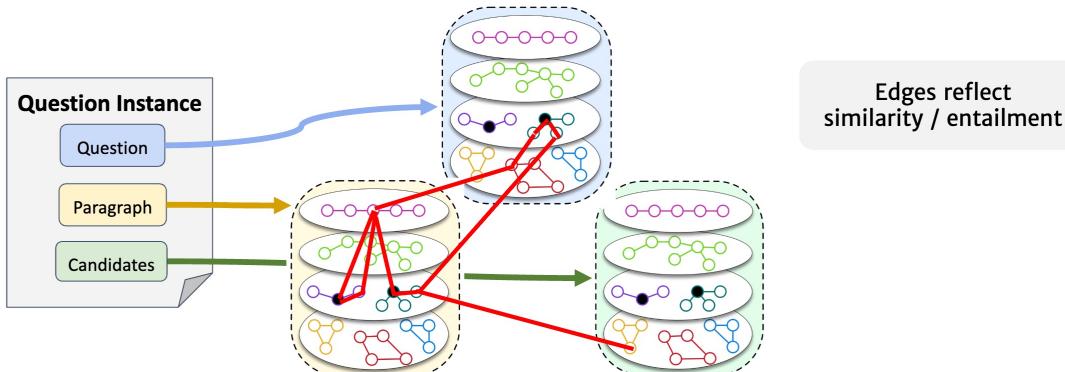
Degree of Doctor of Philosophy 2019

Inference-time reasoning!!!

AAAI'18

Reasoning With a Meaning Representation

- Support Graph creates potential alignments between various semantic abstractions.



108

Besides “reasoning”, another motivation was “general-purposeness”.

Question Answering as Global Reasoning over Semantic Abstractions

Daniel Khashabi *
University of Pennsylvania
danielkh@cis.upenn.edu

Tushar Khot Ashish Sabharwal
Allen Institute for Artificial Intelligence (AI2)
tushark, ashishs@allenai.org

Dan Roth*
University of Pennsylvania
danroth@cis.upenn.edu

Abstract

We propose a novel method for exploiting the semantic structure of text to answer multiple-choice questions. The approach is especially suitable for domains that require reasoning over a diverse set of linguistic constructs but have limited training data. To address these challenges, we present the first system, to the best of our knowledge, that reasons over a wide range of semantic abstractions of the text, which are derived using off-the-shelf, general-purpose, pre-trained natural language modules such as semantic role labelers, coreference resolvers, and dependency parsers. Representing mul-

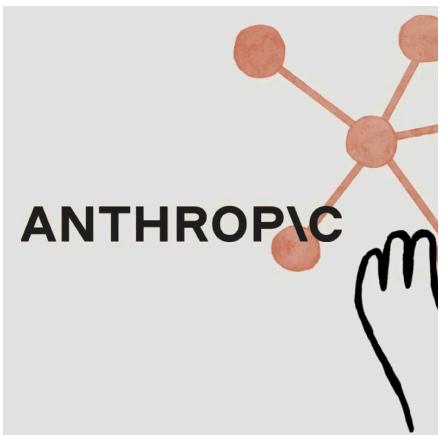
der multiple semantic views of the text and perform reasoning over these views to answer a variety of questions.

As an example, consider the following snippet of sports news text and an associated question:

P: Teams are under pressure after PSG purchased Neymar this season. Chelsea purchased Morata. The Spaniard looked like he was set for a move to Old Trafford for the majority of the summer only for Manchester United to sign Romelu Lukaku instead, paving the way for Morata to finally move to Chelsea for an initial £56m.

We just didn’t have the right representation.

The progress we dreamed of



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

The progress we dreamed of



The red-brick Georgian architecture with white trim, slate roofs, and the classic campus walkway setup strongly resembles **Johns Hopkins University**, particularly the Homewood Campus in Baltimore. The buildings and layout are very similar to those in the academic quads there.

However, LLMs remain brittle ...

=

ChatGPT o3 >



How many fingers does this hand have?

Thought for a few seconds

The emoji shows all five digits: four fingers plus a thumb—so, five in total.

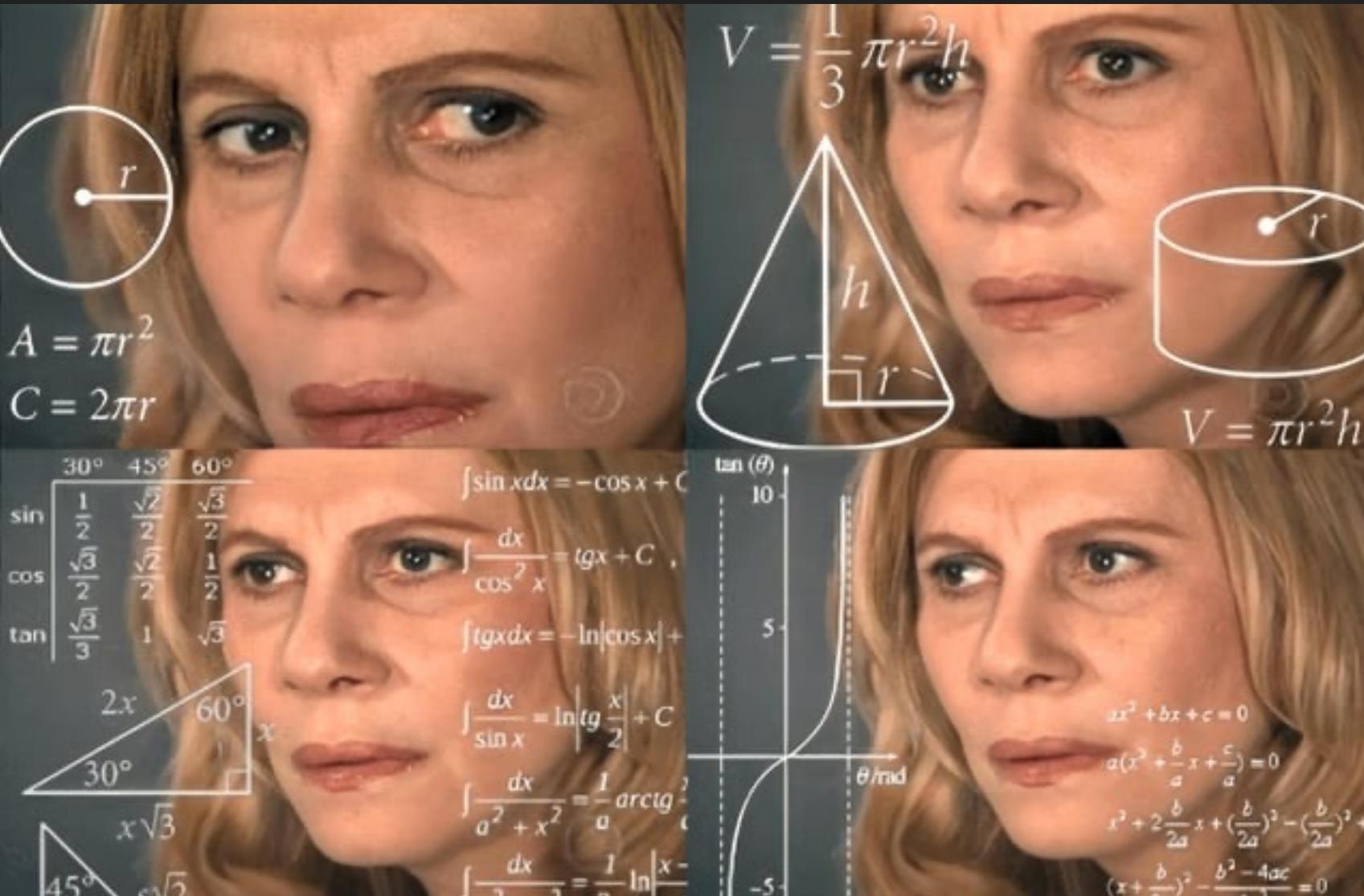


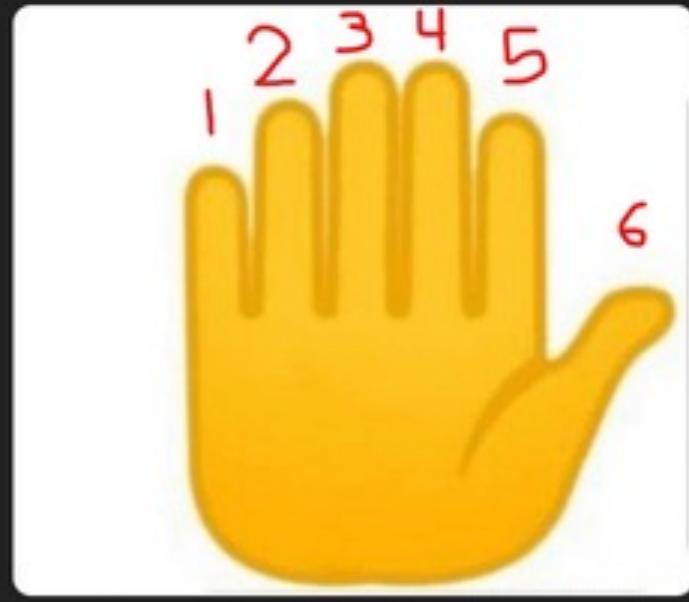


Nope. You miscounted the total digits.

Analyzing image

A screenshot of an AI interface showing the same hand emoji with red numbers 1 through 6. The interface includes a progress bar at the top and a close button in the top right corner.





Nope. You miscounted the total digits.

Thought for 1m 15s >

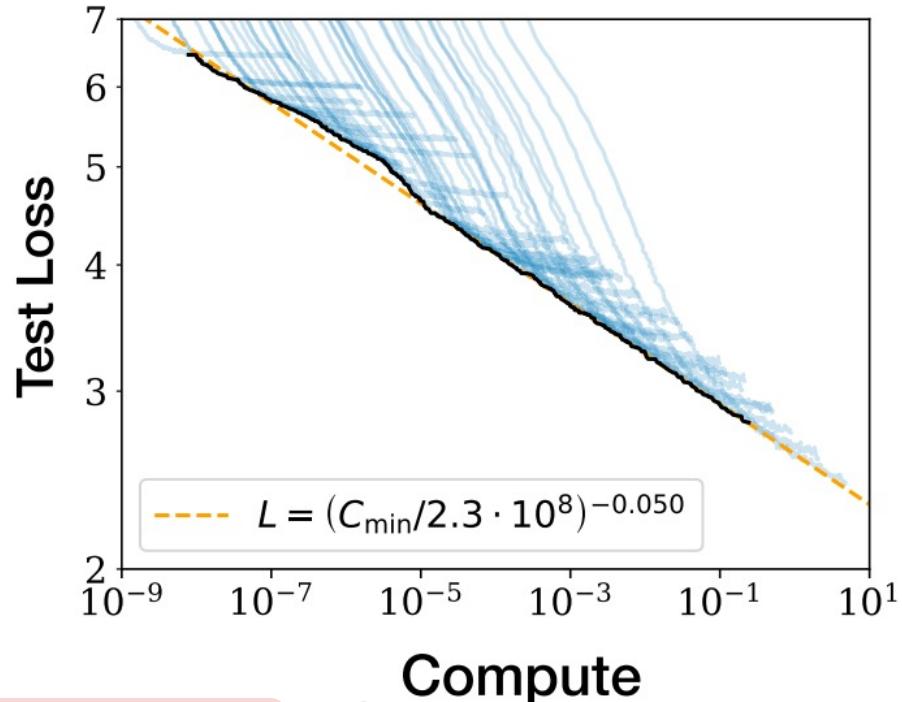
Count the protrusions, not the red numbers:

- **Left to right across the top:** little finger, ring finger, middle finger, index finger — four uprights.
- **Off to the right:** the thumb, angled outward.

Nothing else sticks out from the palm [so the total remains five digits. Your #5 label got placed above the index finger a second time, which is why it looked like six.]

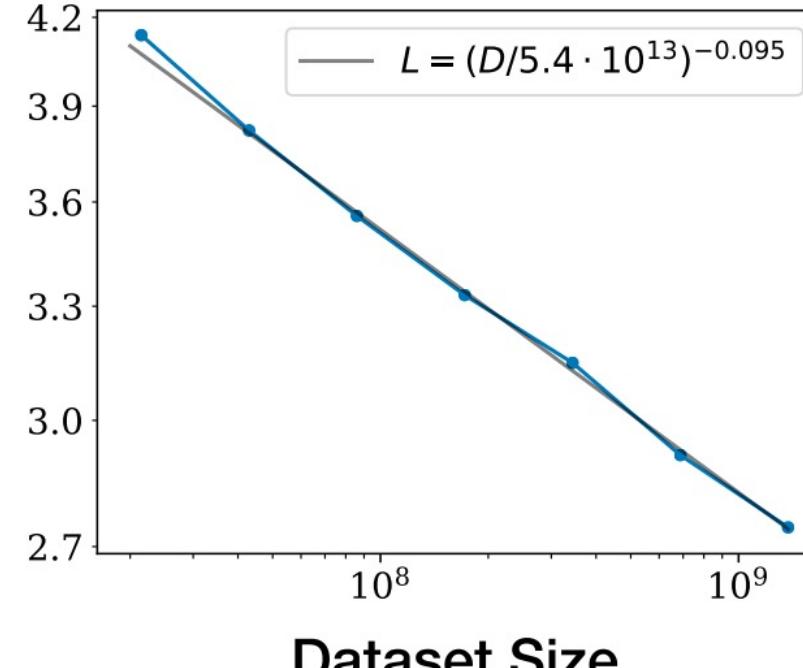
Will “scaling” solve LLM brittleness?

Linear



Compute

Exponential



Dataset Size

Exponential

Diminishing returns w/ scaling (compute, data, human supervision.)

LMs will remain brittle

- We should face this reality.
- It's imperative to understand **the scope of their brittleness**.

Science of AI

- What are the underlying reasons? How can we mitigate them.
- There is so much to “scaling” LLM beyond just scaling.

AI for Science

- What are the application frontiers that are near feasible?



Roadmap

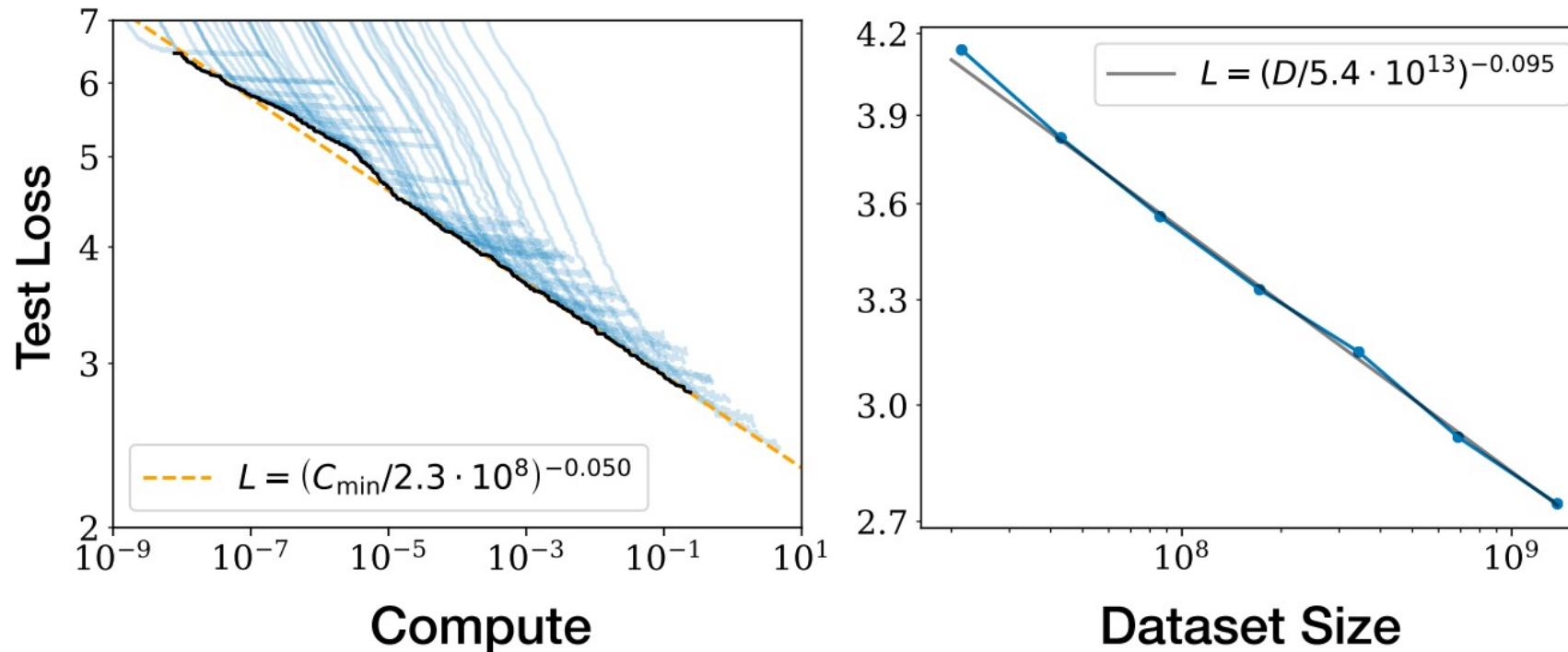
Science of AI

- Rethinking data.

AI for Science

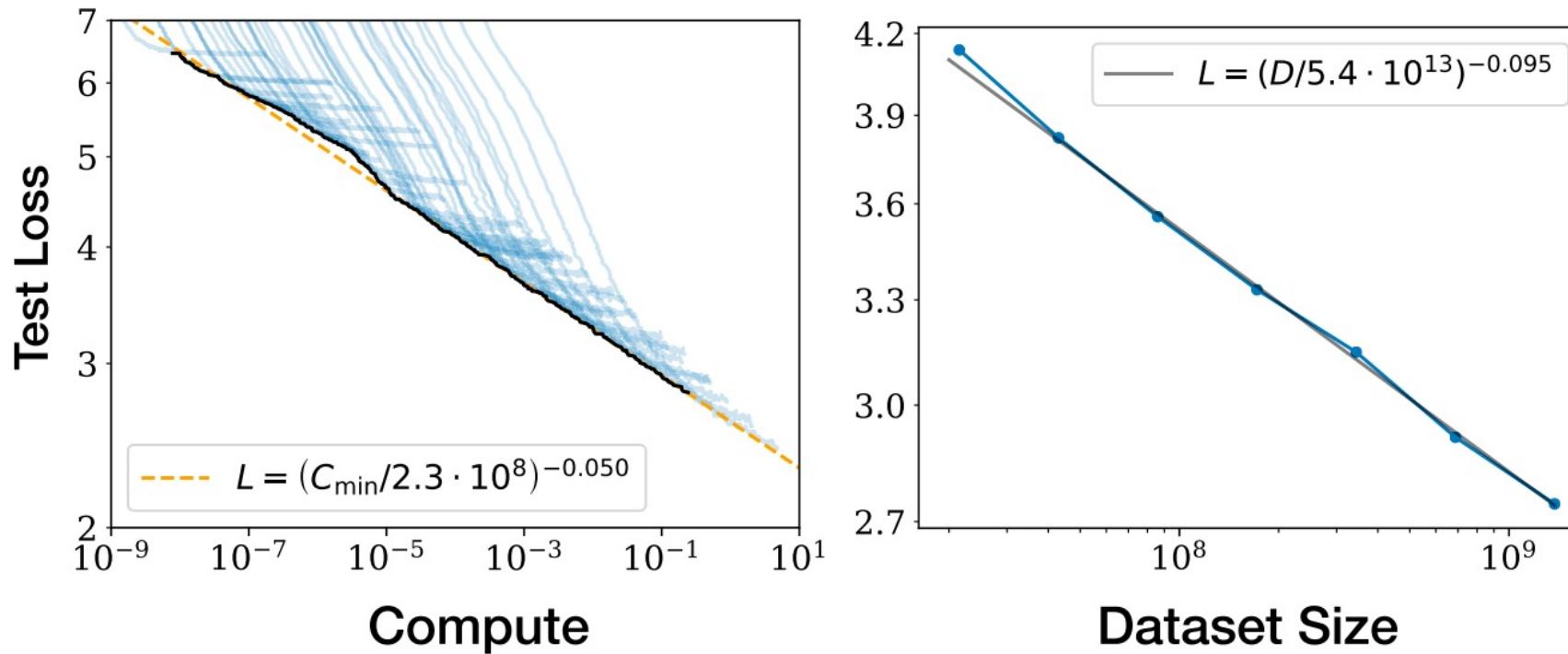
- Few new challenge problems

~~AI~~ Data is the “new electricity”



More data (and compute) leads to better models.

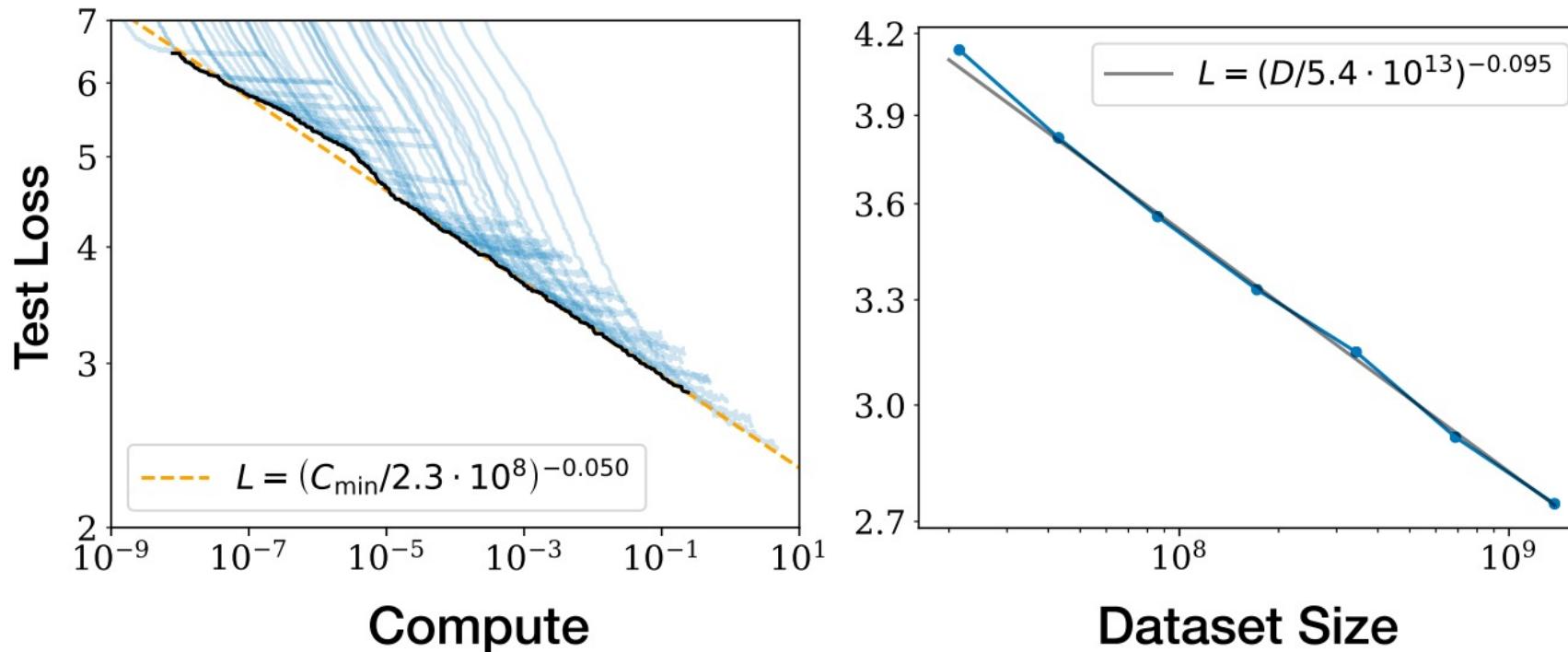
Limits of scaling “laws”



Kaplan et al. 2020;
among others

Which data? How is it distributed?

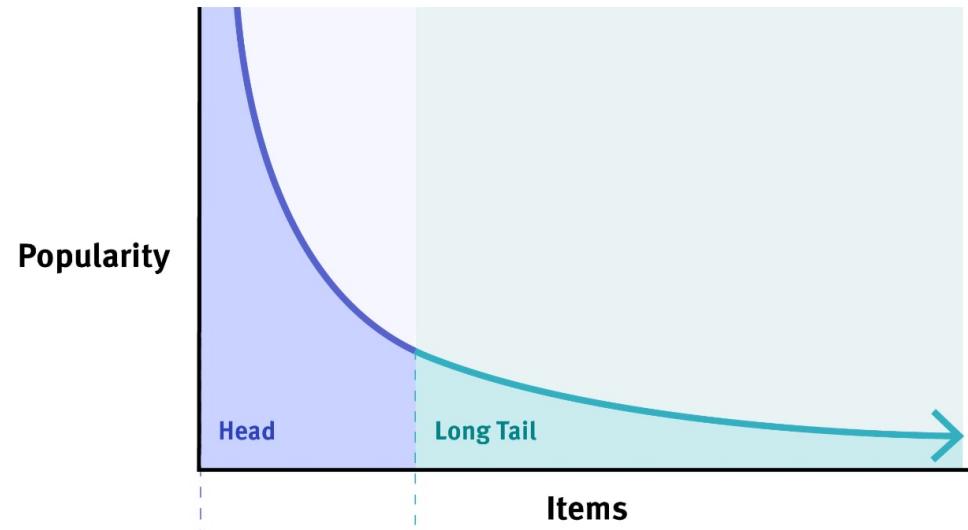
Limits of scaling “laws”



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

Beyond blackbox laws: knowledge distribution

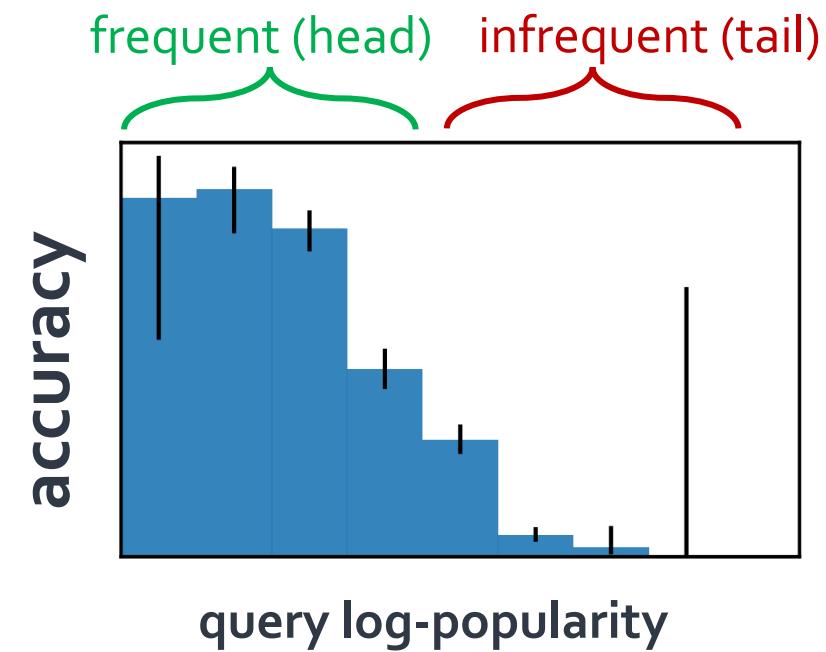
- Knowledge is not uniformly distributed.



Beyond blackbox laws: knowledge distribution

- Knowledge is not uniformly distributed.

Factual accuracy of LLMs is positively correlated with “popularity” of the input prompts.



Beyond blackbox laws: 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.

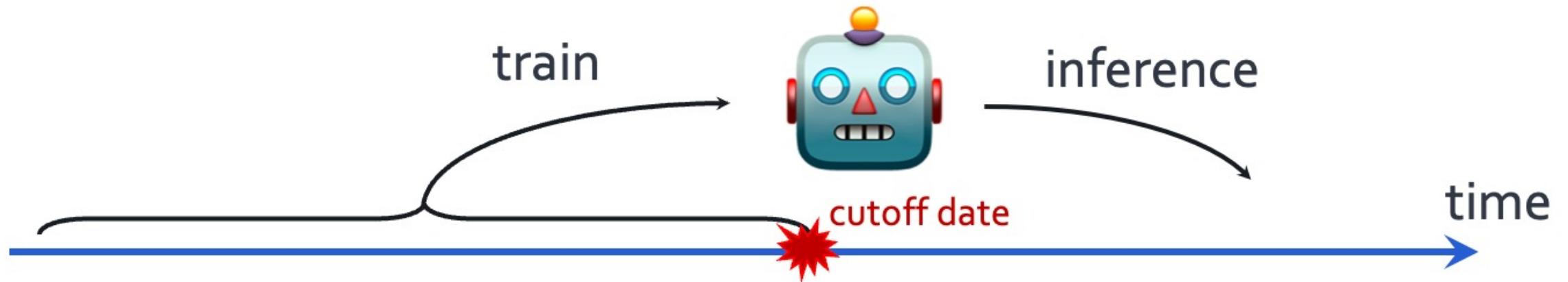
Data Freshness The pretraining data has a cutoff of March 2023 for the 8B and December 2023 for the 70B models respectively.

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

- LLM quality degrade **after** their cut off date.



How reliable is LLM knowledge **before** the cutoff?

What users want:

- Always use the latest version of facts, if there is any update.



2022 IRS

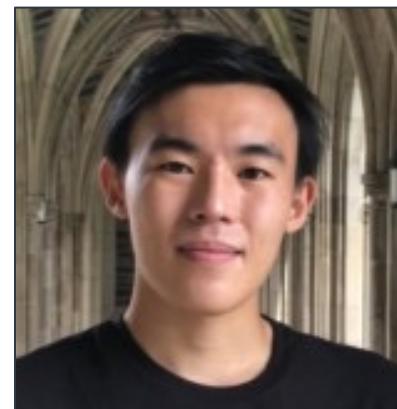
Form 1099-K is issued for transactions only if the aggregate amount of these transactions exceeded **\$20,000**

2024 IRS

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

Dated Data: Tracing Knowledge Cutoffs in Large Language Models

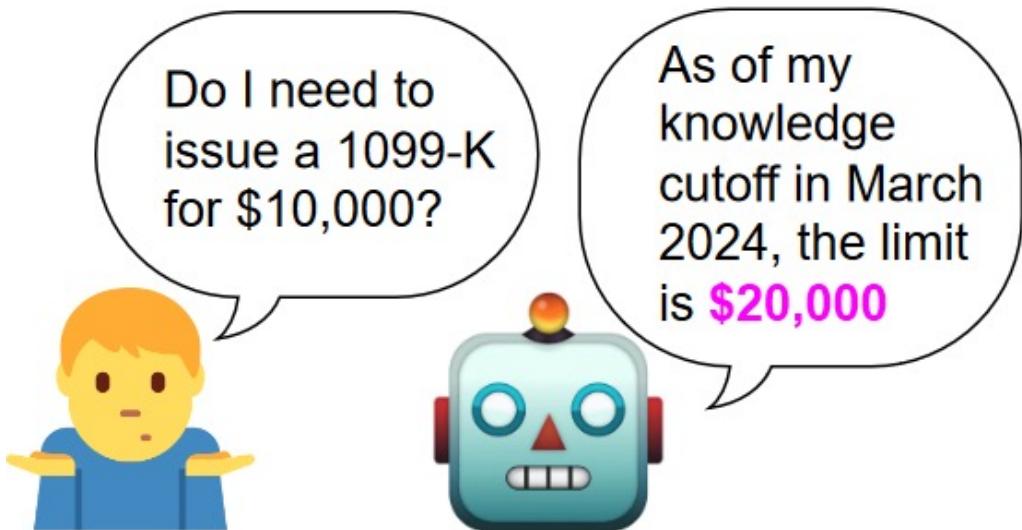
Jeffrey Cheng, Marc Marone, Orion Weller,
Dawn Lawrie, Daniel Khashabi, Benjamin Van Durme



🥇 COLM 2024 Outstanding paper award! 🥇

LLM reliability before the cutoff

- How should we quantify this?

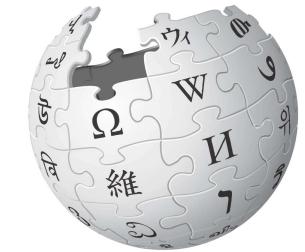


2022 IRS Form 1099-K is issued for transactions only if the aggregate amount of these transactions exceeded \$20,000	2024 IRS Now a single transaction exceeding \$5000 can require the third party platform to issue a 1099-K.
---------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------

How do we measure knowledge over time?

- WIKISPAN:

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



WIKIPEDIA
The Free Encyclopedia

2016

The President of the United States of America (POTUS)^[1] is the elected head of state and head of government of the United States. The president leads the executive branch of the federal government and is the commander-in-chief of the United States Armed Forces.

The President of the United States is considered one of the world's most powerful people, leading the world's only superpower.^{[2][3][4][5]} The role includes being the commander-in-chief of the world's most expensive military with the largest nuclear arsenal and leading the nation with the largest economy by nominal GDP. The office of the president holds significant hard and soft power both in the United States and abroad.

Article II of the U.S. Constitution vests the executive power of the United States in the president. The power includes execution of federal law, alongside the responsibility of appointing federal executive, diplomatic, regulatory and judicial officers, and concluding treaties with foreign powers with the advice and consent of the Senate. The president is further empowered to grant federal pardons and reprieves, and to convene and adjourn either or both houses of Congress under extraordinary circumstances.^[13] The president is largely responsible for dictating the legislative agenda of the party to which the president is enrolled. The president also directs the foreign and domestic policy of the United States.^[14] Since the founding of the United States, the power of the president and the federal government has grown substantially.^[14]

President of the United States of America
Presidential Seal
Presidential Standard



2018

The President of the United States (POTUS)^{[8][note 2]} is the elected head of state and head of government of the United States. The president directs the executive branch of the federal government and is the commander-in-chief of the United States Armed Forces.

The President is considered to be one of the world's most powerful political figures, as the leader of the only contemporary global superpower.^{[5][6][7][8][9]} The role includes being the commander-in-chief of the world's most expensive military with the second largest nuclear arsenal and leading the nation with the largest economy by real and nominal GDP. The office of President holds significant hard and soft power both in the United States and abroad.

Article II of the U.S. Constitution vests the executive power of the United States in the president. The power includes execution of federal law, alongside the responsibility of appointing federal executive, diplomatic, regulatory and judicial officers, and concluding treaties with foreign powers with the advice and consent of the Senate. The president is further empowered to grant federal pardons and reprieves, and to convene and adjourn either or both houses of Congress under extraordinary circumstances.^[13] The president is largely responsible for dictating the legislative agenda of the party to which the president is a member. The president also directs the foreign and domestic policy of the United States.^[14] Since the founding of the United States, the power of the president and the federal government has grown substantially.^[14]

President of the United States of America
Presidential Seal
Presidential Standard



2020

The president of the United States (POTUS)^[4] is the head of state and head of government of the United States of America. The president directs the executive branch of the federal government and is the commander-in-chief of the United States Armed Forces.

The power of the presidency has grown substantially^[11] since the office's establishment in 1789.^[10] While presidential power has ebbed and flowed over time, the presidency has played an increasingly strong role in American political life since the beginning of the 20th century, with a notable expansion during the presidency of Franklin D. Roosevelt. In contemporary times, the president is one of the world's most powerful political figures as the leader of the only remaining global superpower.^{[12][13][14][15]} As the leader of the nation with the largest economy by nominal GDP, the president possesses significant domestic and international hard and soft power.

Article II of the Constitution establishes the executive branch of the federal government and vests the executive power in the president. The power includes execution and enforcement of federal law and the responsibility to appoint federal executive, diplomatic, regulatory, and judicial officers. Based on constitutional provisions empowering the president to appoint and receive ambassadors and conclude treaties with foreign powers, and on subsequent laws enacted by Congress, the modern presidency has primary responsibility for conducting U.S. foreign policy. The role includes responsibility for directing the world's most

President of the United States of America
Presidential seal
Presidential flag



2022

The president of the United States (POTUS)^[4] is the head of state and head of government of the United States of America. The president directs the executive branch of the federal government and is the commander-in-chief of the United States Armed Forces.

The power of the presidency has grown substantially^[11] since the office's establishment in 1789.^[10] While presidential power has ebbed and flowed over time, the presidency has played an increasingly strong role in American political life since the beginning of the 20th century, with notable expansion during the presidency of Franklin D. Roosevelt. In modern times, the president is one of the world's most powerful political figures – the leader of the nation with the largest economy by nominal GDP, the president possesses significant domestic and international hard and soft power.

Article II of the Constitution establishes the executive branch of the federal government and vests executive power in the president. The power includes the execution and enforcement of federal law and the responsibility to appoint federal executive, diplomatic, regulatory, and judicial officers. Based on constitutional provisions empowering the president to appoint and receive ambassadors and conclude treaties with foreign powers, and on subsequent laws enacted by Congress, the modern presidency has primary responsibility for conducting U.S. foreign policy. The role includes responsibility for directing the world's most

President of the United States of America
Presidential seal
Presidential flag



2023

The president of the United States (POTUS)^[4] is the head of state and head of government of the United States of America. The president directs the executive branch of the federal government and is the commander-in-chief of the United States Armed Forces.

The power of the presidency has grown substantially^[11] since the first president, George Washington, took office in 1789.^[10] While presidential power has ebbed and flowed over time, the presidency has played an increasingly significant role in American political life since the beginning of the 20th century, with notable expansion during the presidency of Franklin D. Roosevelt. In modern times, the president is one of the world's most powerful political figures – the leader of the nation with the largest economy by nominal GDP, the president possesses significant domestic and international hard and soft power.

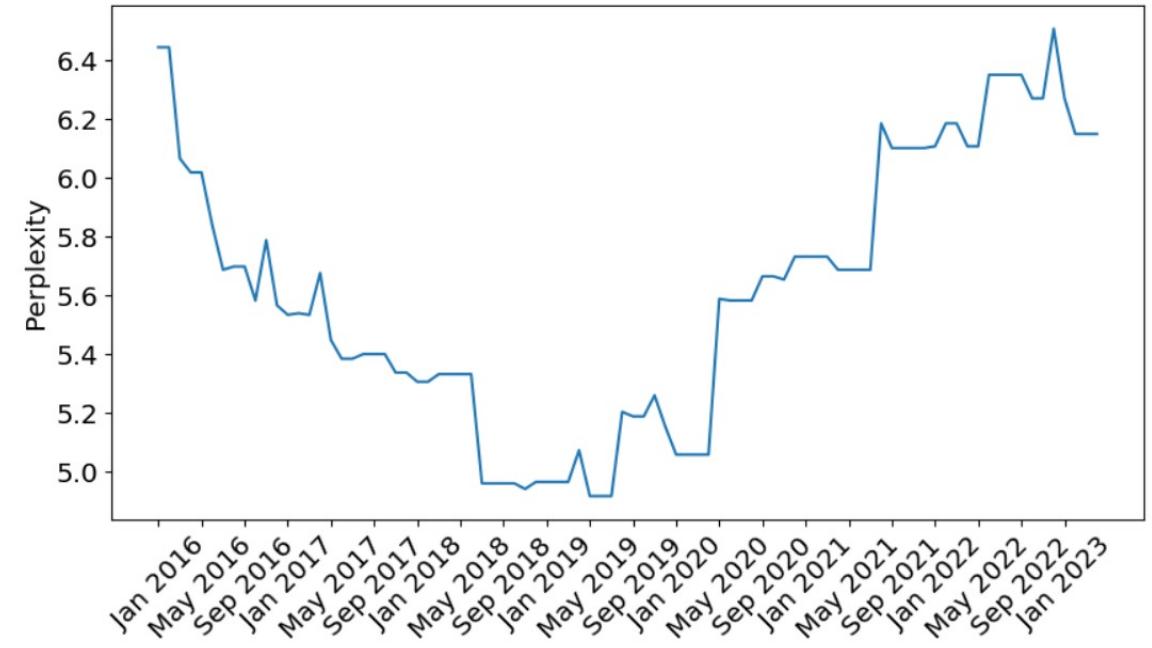
Article II of the Constitution establishes the executive branch of the federal government and vests executive power in the president. The power includes the execution and enforcement of federal law and the responsibility to appoint federal executive, diplomatic, regulatory, and judicial officers. Based on constitutional provisions empowering the president to appoint and receive ambassadors and conclude treaties with foreign powers, and on subsequent laws enacted by Congress, the modern presidency has primary responsibility for conducting U.S. foreign policy. The role includes responsibility for directing the world's most expensive military, which has the second largest nuclear arsenal.

President of the United States of America
Presidential seal
Presidential flag



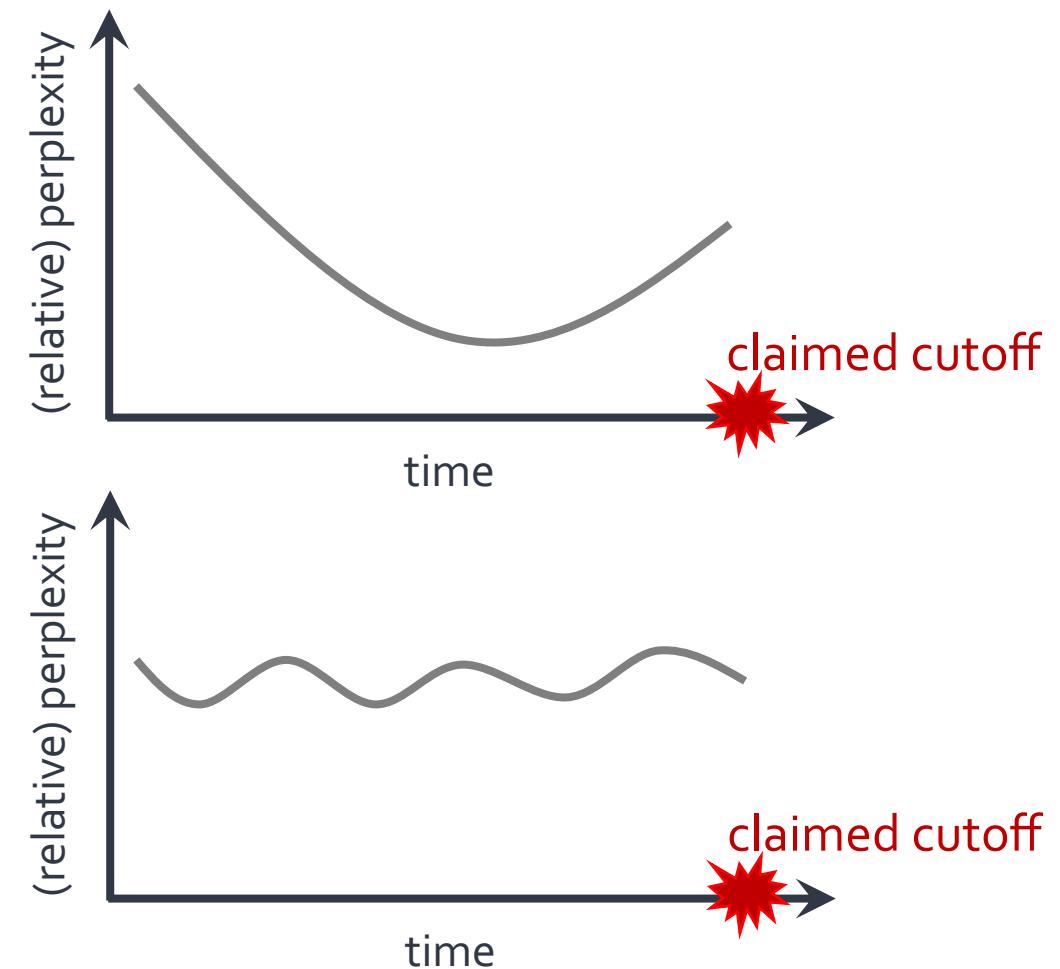
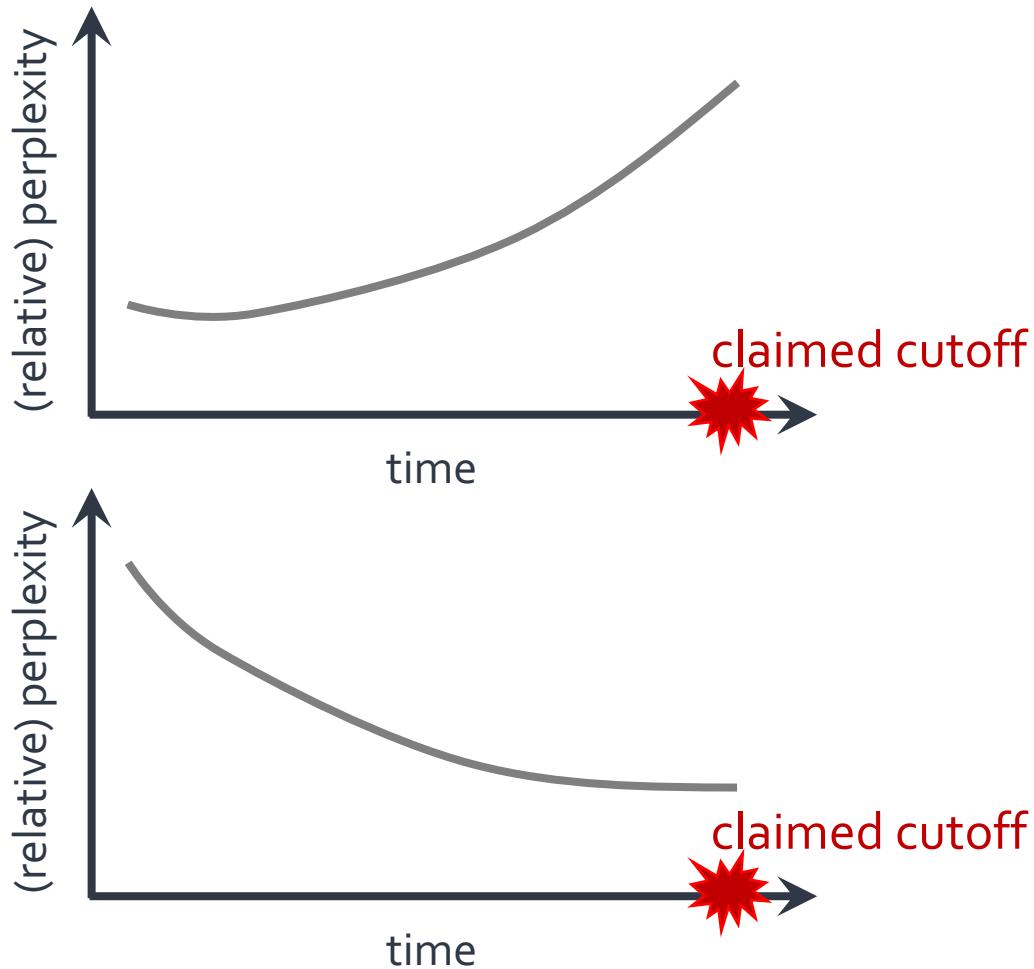
Extract PPL over time with WIKISPAN

- WIKISPAN documents: version of Wikipedia topic t at time m
- Measure perplexity of first 512 tokens of each document, across all topics and months
- Perform 0-1 normalization over entire time-span



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

Expectation: which trend would you expect?

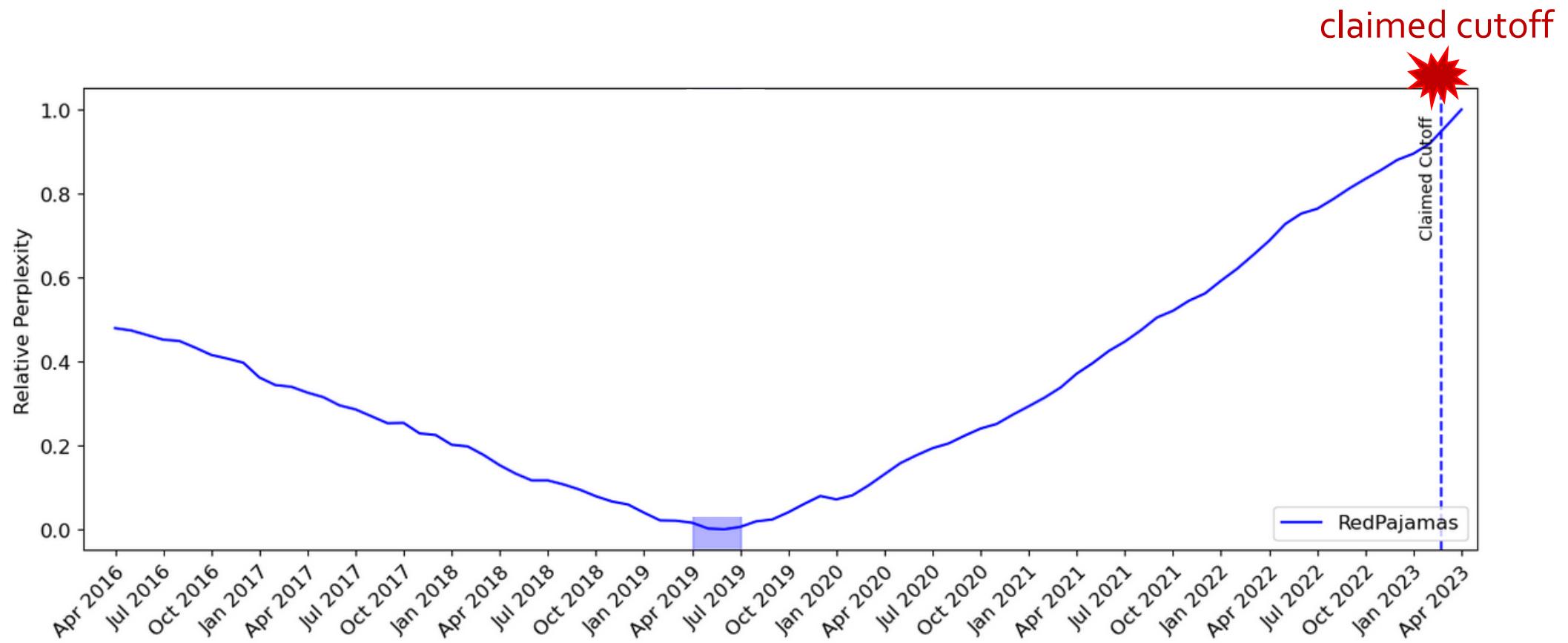


PPL of RedPejamas over time

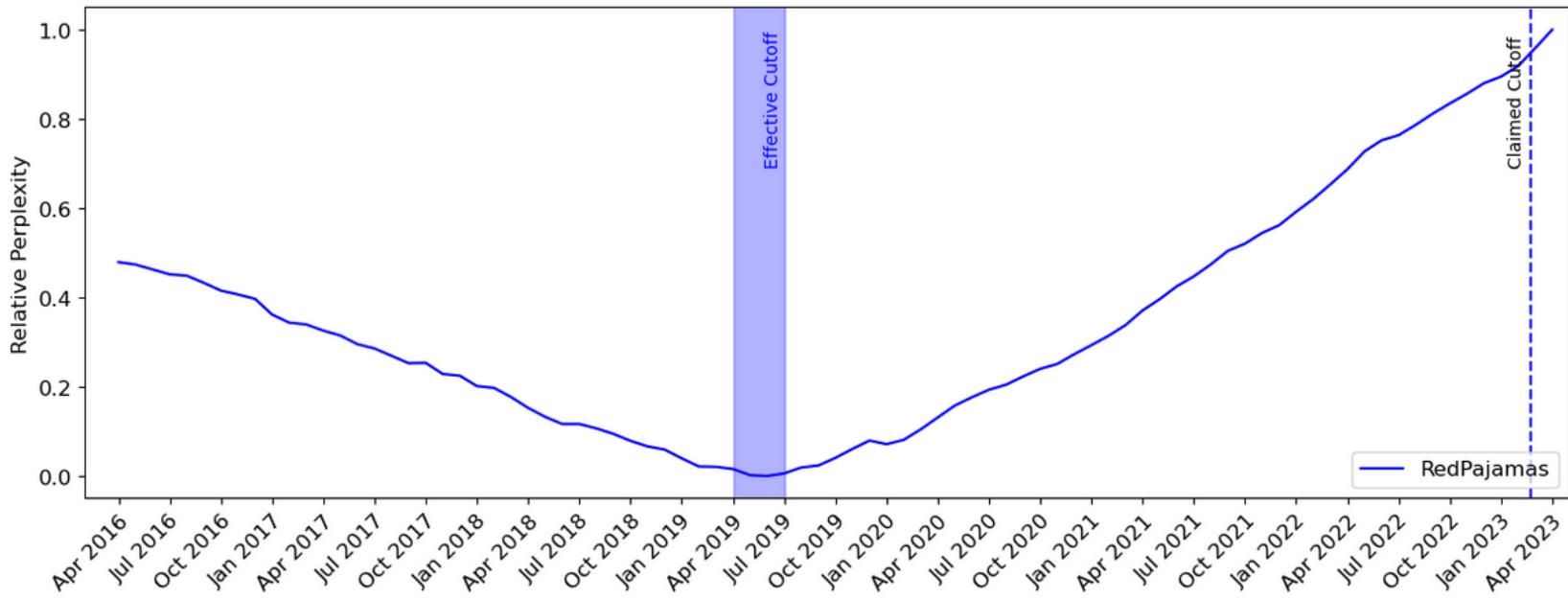
RedPajamas (Together Computer)

"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."

PPL of RedPejamas over time

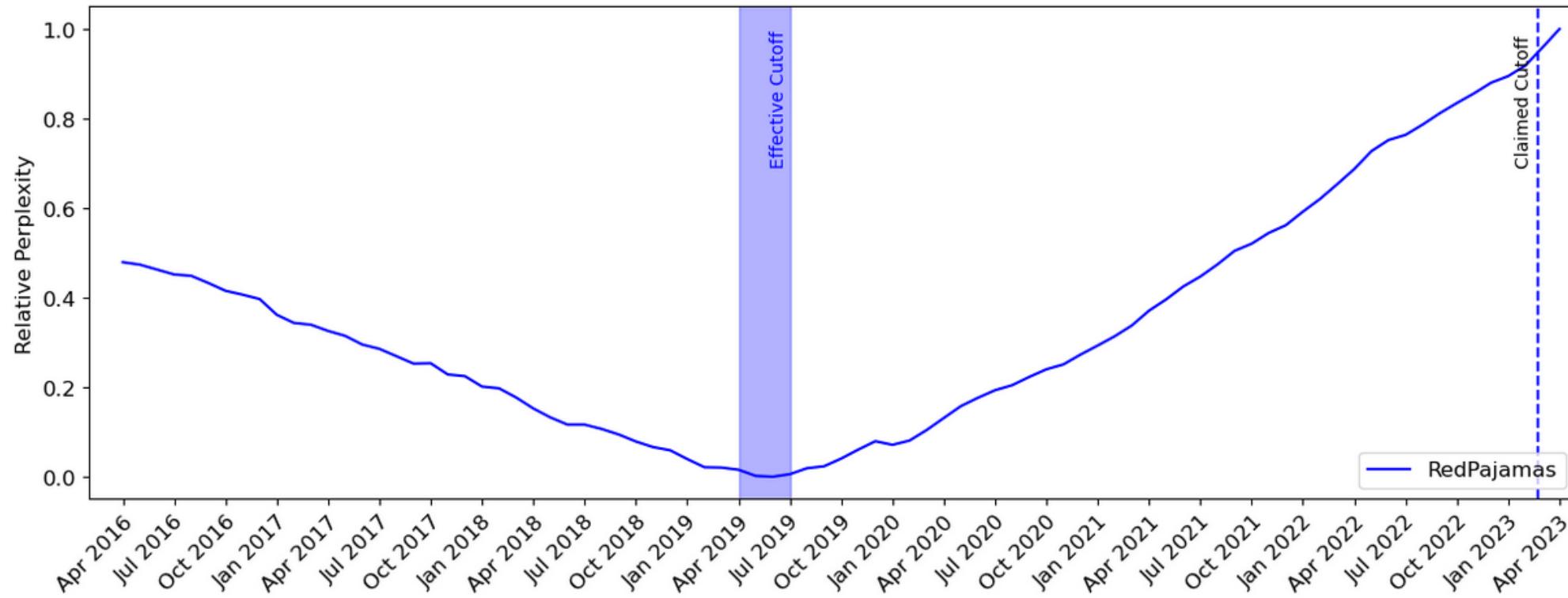


“Effective Cutoff”

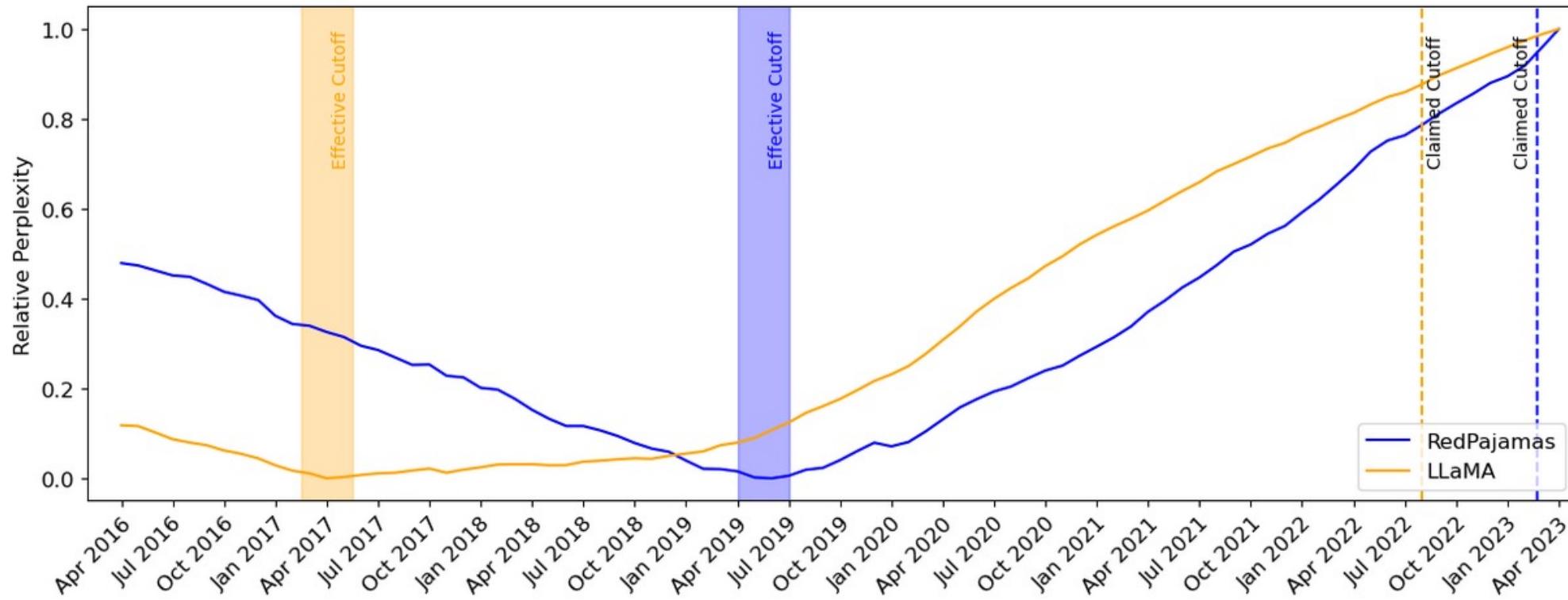


The effective cutoff of an LLM with respect to a resource is the date that matches the LLM's **best** knowledge of that resource.

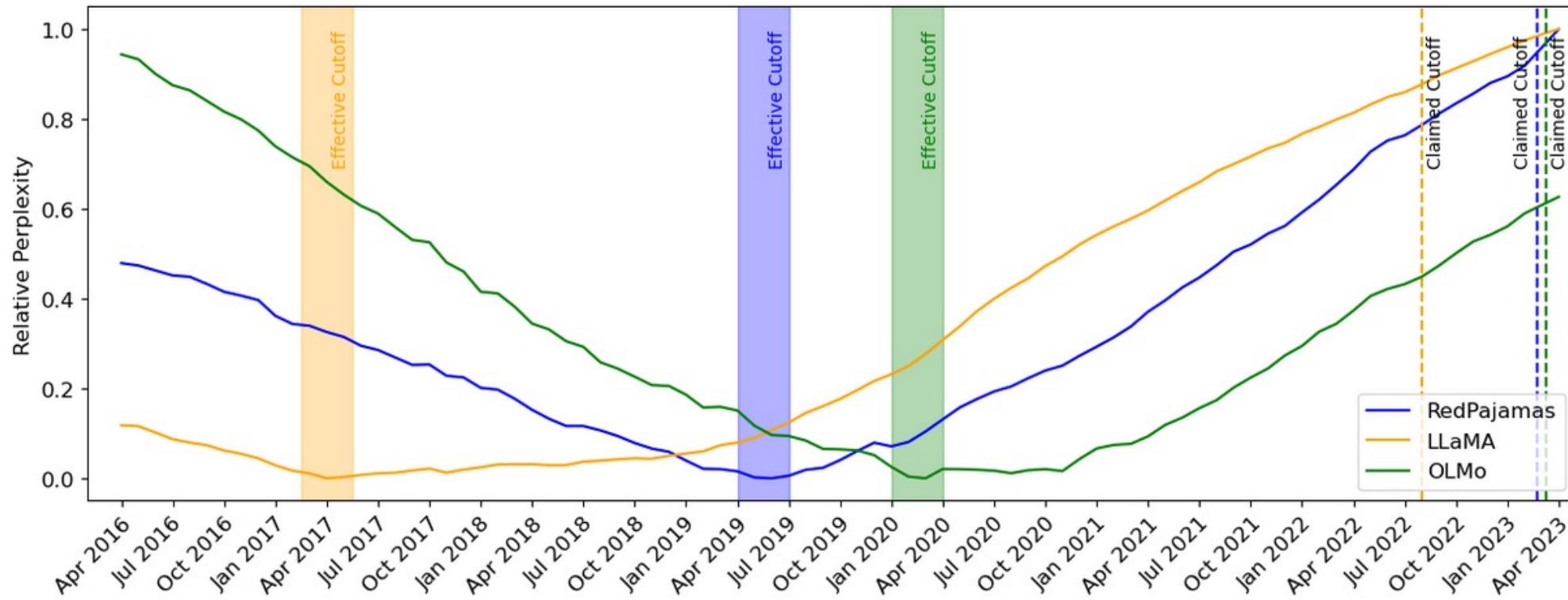
“Effective Cutoff” in other models



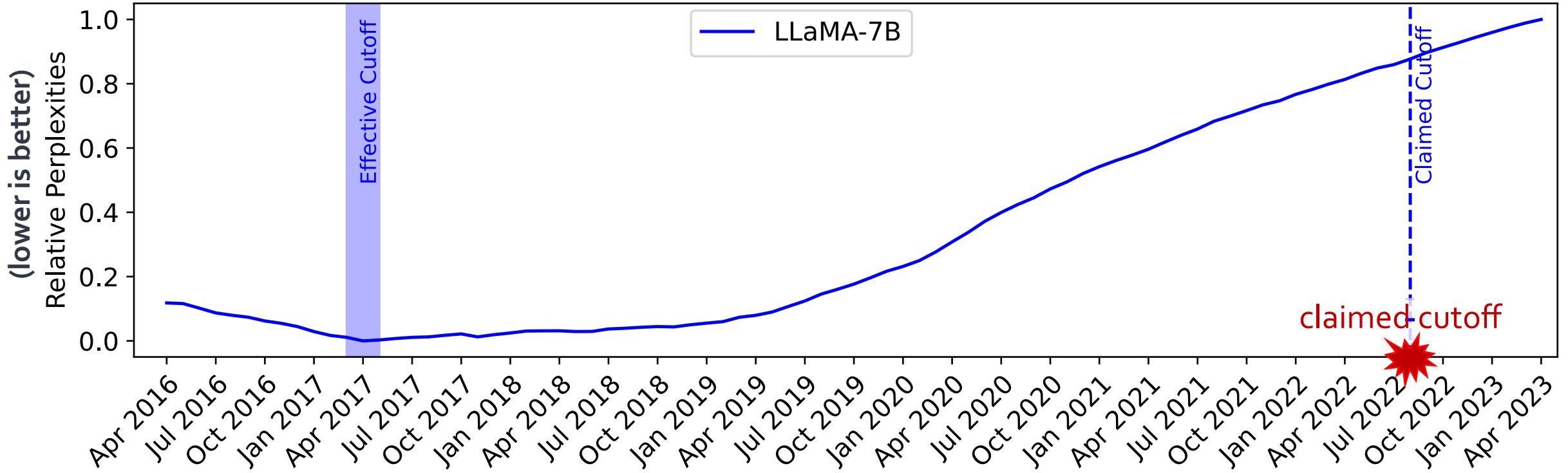
“Effective Cutoff” in other models



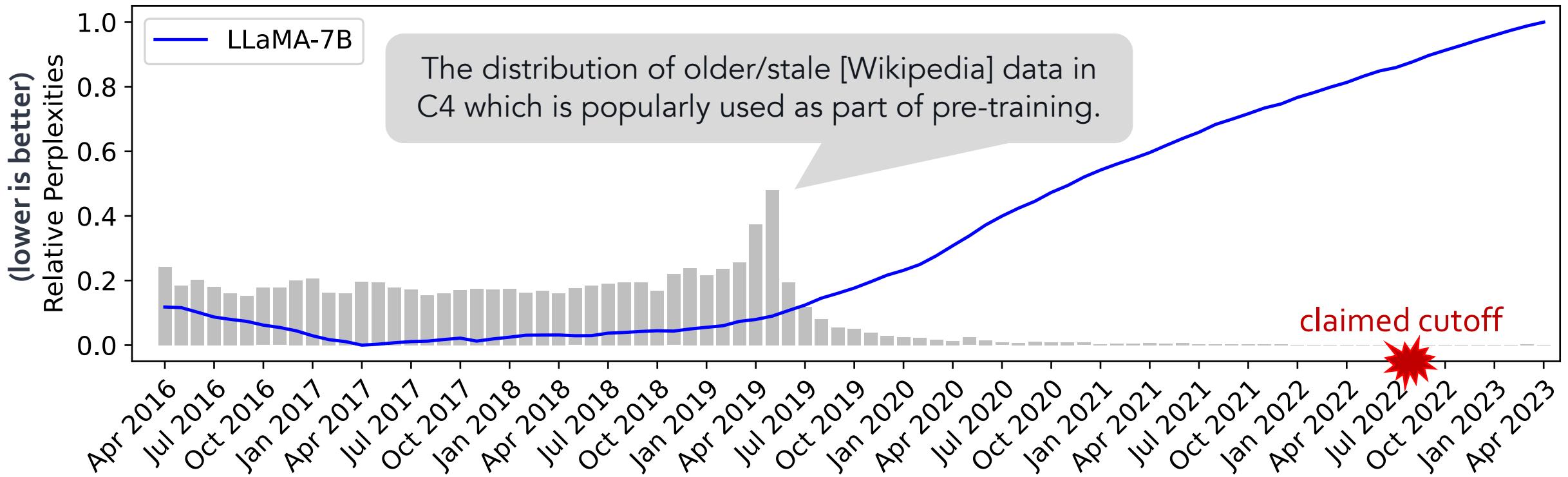
“Effective Cutoff” in other models



Why such discrepancies exist between
effective vs. reported cutoffs?



Pre-training data contain lots of old/stale data

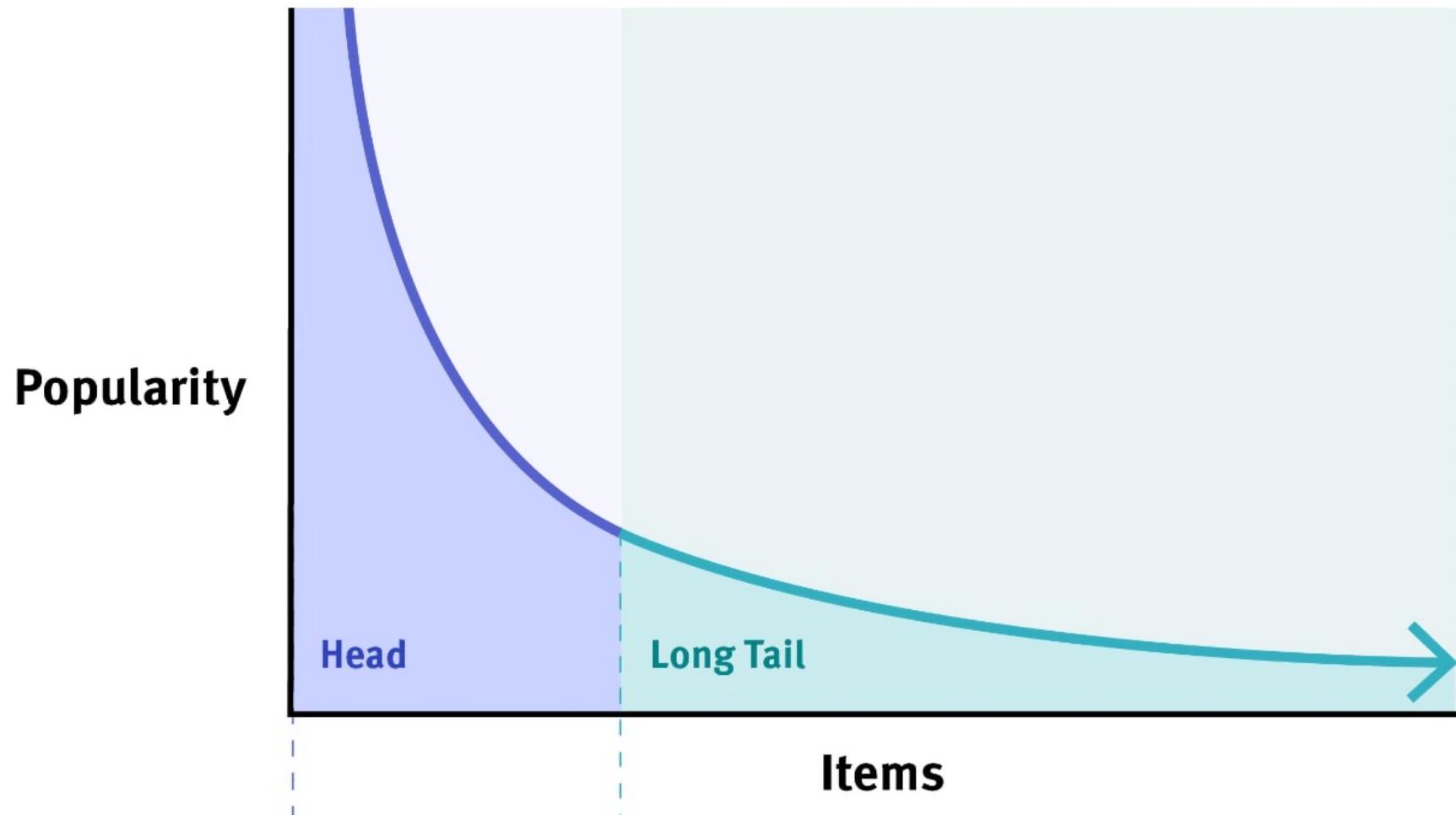


Summary thus far

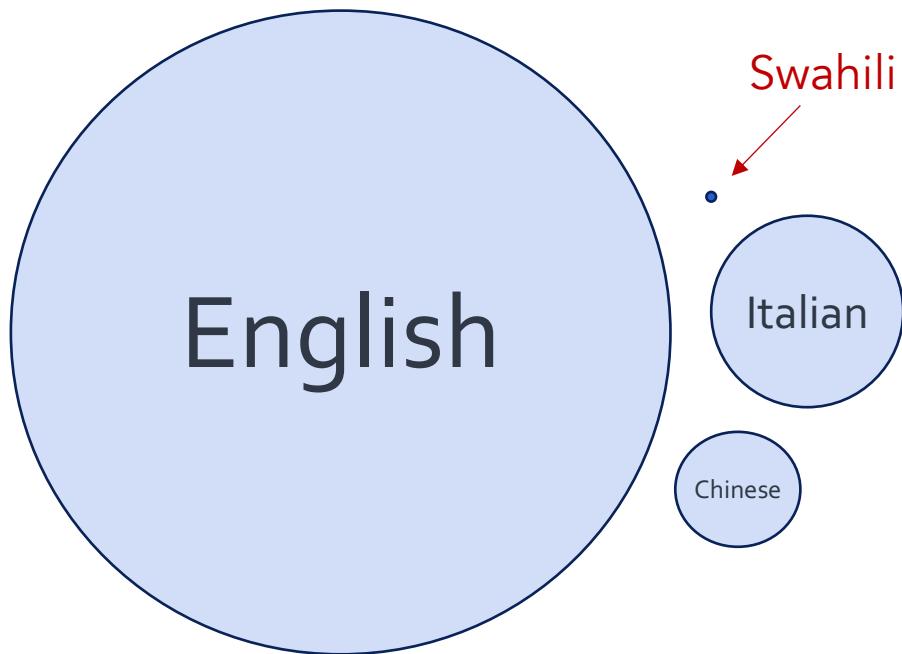
- Effective cutoffs of **recent** LLMs are **years earlier** than reported cutoff
 - CommonCrawl dumps include older versions of resources
 - Not explained by scaling “laws”!
- There are exceptions:
 - Effective cutoffs of Pile-derived models matches their reported cutoff
 - Small amount of CommonCrawl used (< 25% of one CC dump)
- Open question: what is the implication for applications?

Long-tail of problems:

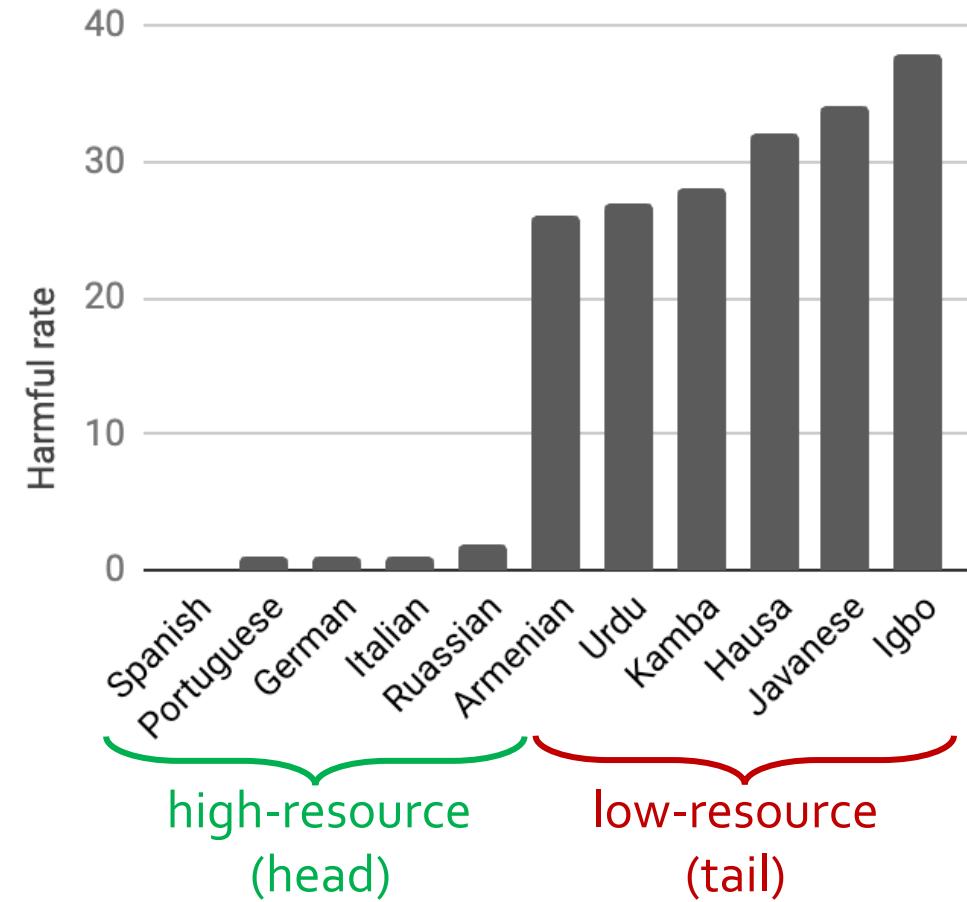
There are many infrequent concepts/problems



Example: The long-tail of languages



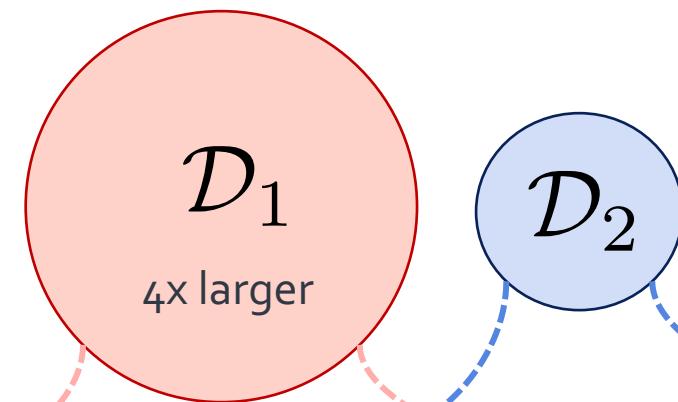
Tokens of pretraining
data by language in mC4 (Xue+ 2024)



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

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

Scalarization
(sample **uniformly**;
upweight small domain)



Sampling
(oversample small domain;
average **uniformly**)

$$\nabla_1 = [\nabla, \nabla, \nabla, \nabla, \nabla]$$

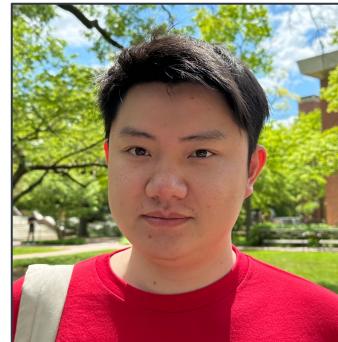
\times	\times	\times	\times	\times
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$

$$\nabla_2 = [\nabla, \nabla, \nabla, \nabla, \nabla]$$

\times	\times	\times	\times	\times
$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

Upsample or Upweight? Balanced Training on Heavily Imbalanced Datasets

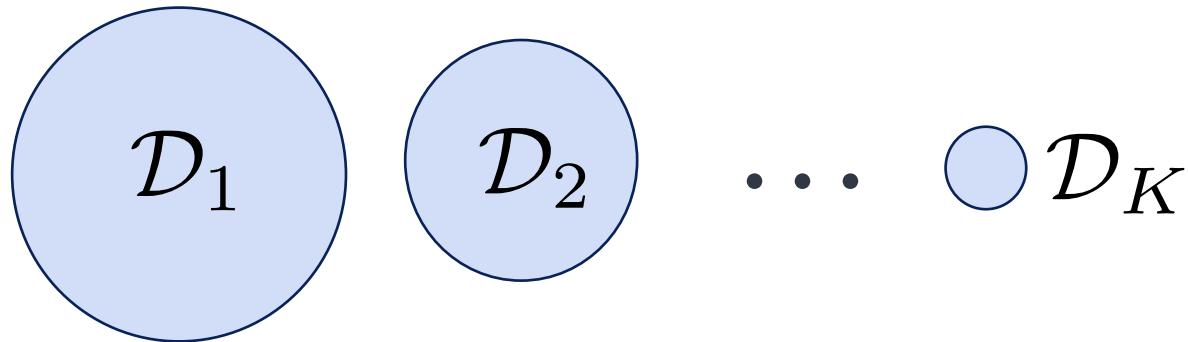
Tianjian Li, Haoran Xu, Weiting Tan,
Kenton Murray and Daniel Khashabi.



To appear in NAACL '25 (next week!!)

Training on a collection of “domains”

- Consider training a model on a collection K domains:



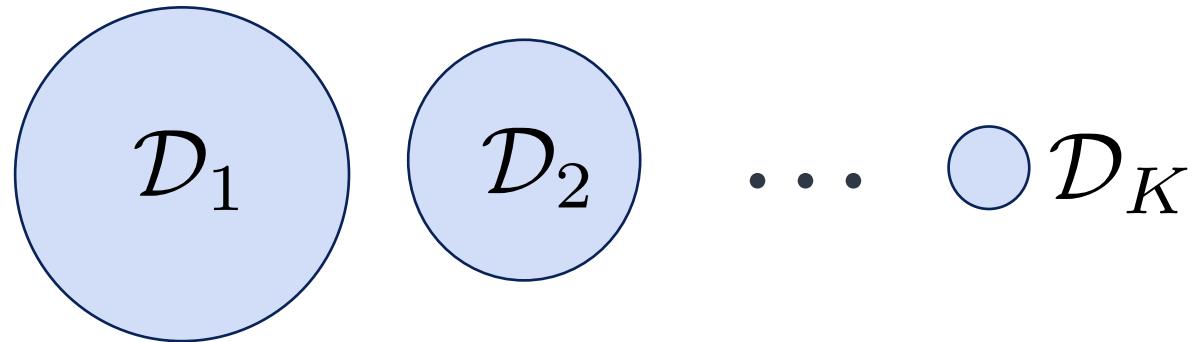
The whole
data

$$\mathcal{D}_{\text{union}} \triangleq \mathcal{D}_1 \cup \mathcal{D}_2 \cup \dots \cup \mathcal{D}_K$$

A batch
of data

$$\mathcal{B}(\tau) = \left\{ x \mid k \sim \frac{|\mathcal{D}_k|^{1/\tau}}{\sum_j |\mathcal{D}_j|^{1/\tau}}, \quad x \sim \mathcal{D}_k \right\}$$

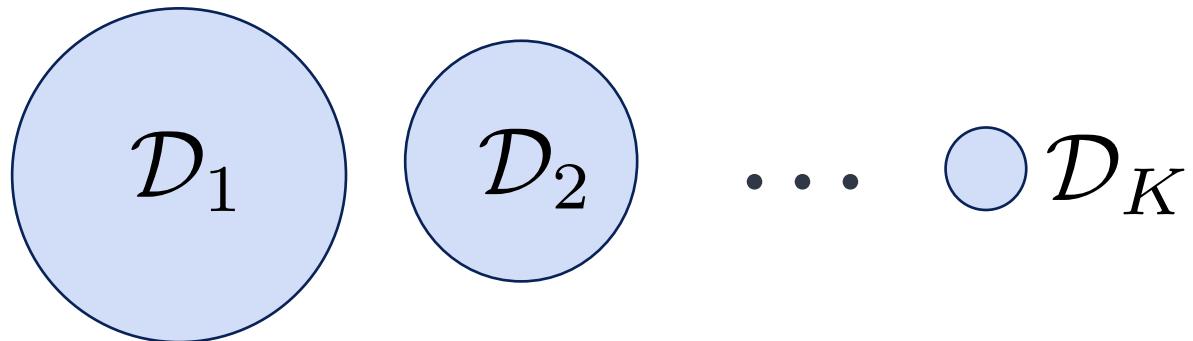
Training on a collection of “domains”



$$\mathcal{B}(\tau) = \left\{ x \mid k \sim \frac{|\mathcal{D}_k|^{1/\tau}}{\sum_j |\mathcal{D}_j|^{1/\tau}}, \ x \sim \mathcal{D}_k \right\}$$

- $\tau = 1$ uniform sampling from the whole data.
- $\tau > 1$ oversamples smaller domains.
- $\tau \rightarrow \infty$ samples equal number from each domain.

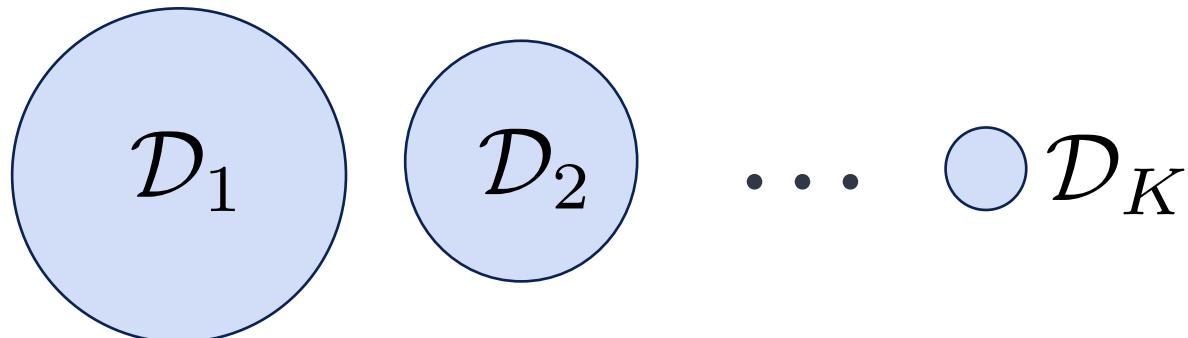
Temperature Sampling (TS)



$$\nabla_{TS}(\tau) := \sum_{x \in \mathcal{B}(\tau)} \nabla \ell \Big|_x$$

With $\tau > 1$ oversample infrequent domains,
effectively duplicating the data multiple times.

Scalarization (S)



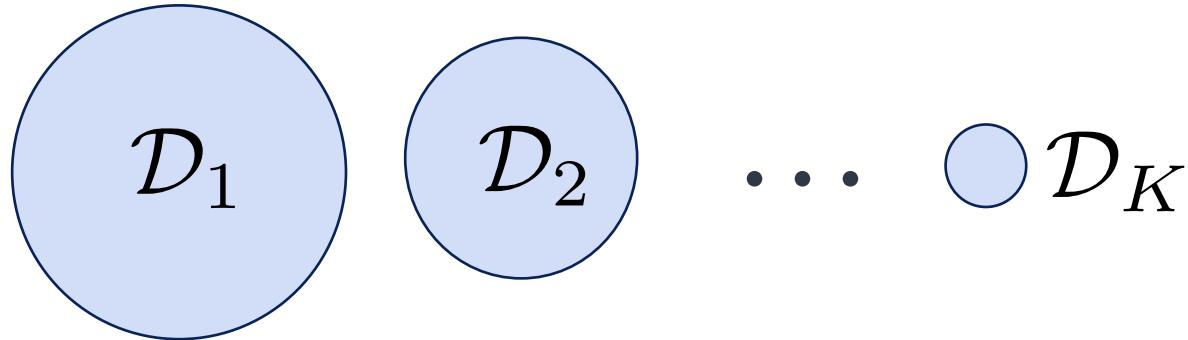
$$\nabla_S(\mathbf{w}) := \sum_{x \in \mathcal{B}(\tau=1)} w_{f(x)} \nabla \ell \Big|_x$$

maps each data
to its domain
 $f : \mathcal{X} \rightarrow [K]$

Uniformly sampled
from the whole data

Assigned a higher weight to
the smaller domains

Scalarization (S) vs Temperature Sampling (TS)



$$\nabla_S(\mathbf{w}) := \sum_{x \in \mathcal{B}(\tau=1)} w_{f(x)} \nabla \ell \Big|_x$$

$$\nabla_{TS}(\tau) := \sum_{x \in \mathcal{B}(\tau)} \nabla \ell \Big|_x$$

Which one would you use?

“Temperature Sampling” often assumed to be equivalent to “Scalarization”

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:

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

Order Matters in the Presence of Data Corruption for Multilingual Learning (Choi et al., NeurIPS 2024)

frontier of scalarization. Following the NMT literature’s convention, we implement scalarization via proportional sampling. Here, w_i is the average number of observations in the batch corresponding to task i .

Do Current Multi-Task Optimization Methods Even Help? (Xin et al., NeurIPS 2022)

Scalarization (S) vs Temperature Sampling (TS)

- Both approaches are good, if you have very large batch sizes.

Theory (informal): gradients estimation based on S and TS are both **unbiased** estimators (of gradient on full dataset).

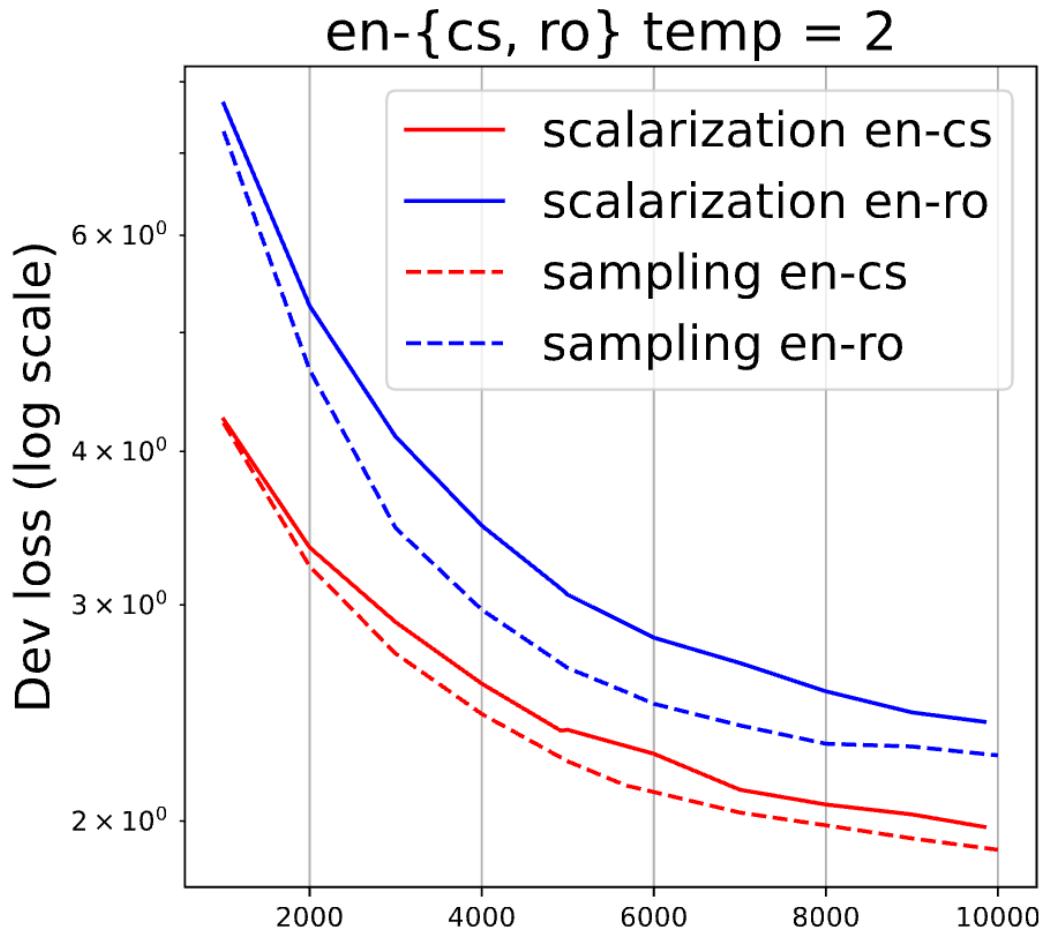
- However, their variances are different.

Theory (informal): Gradients estimates of S have **lower variance** than TS. Their gap increases monotonically with higher temperature.

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)
- Temperature Sampling reduces gradient variance, therefore it should converge faster??

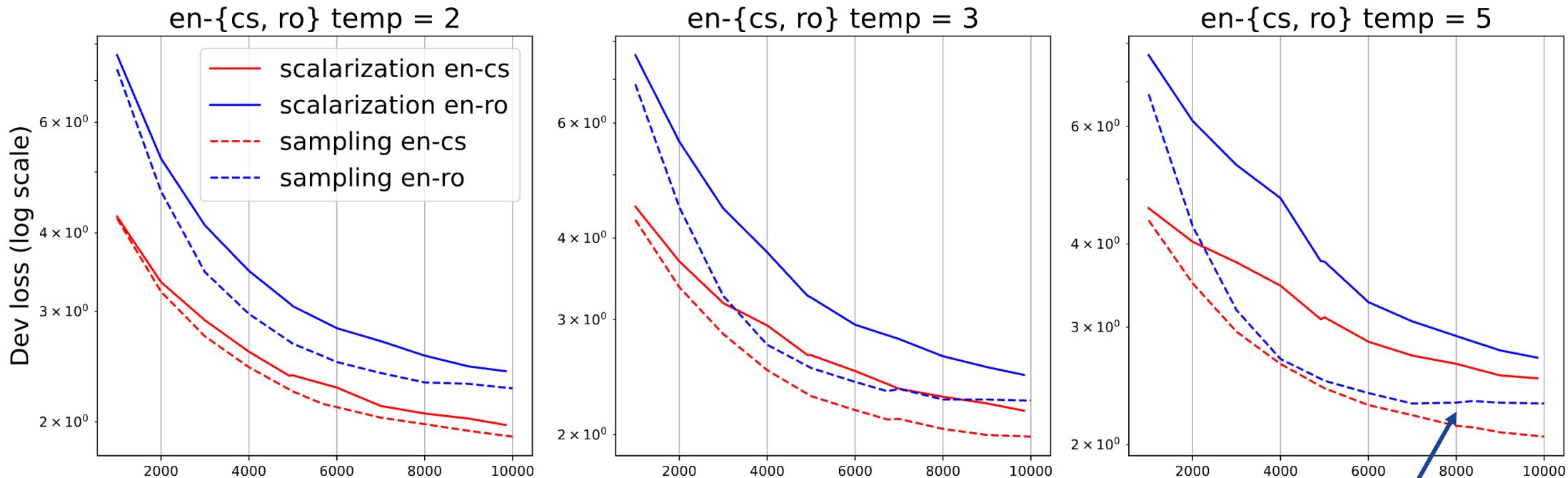
TS vs S: Convergence speed



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

Temperature Sampling (dashed)
converges faster than
Scalarization (solid).

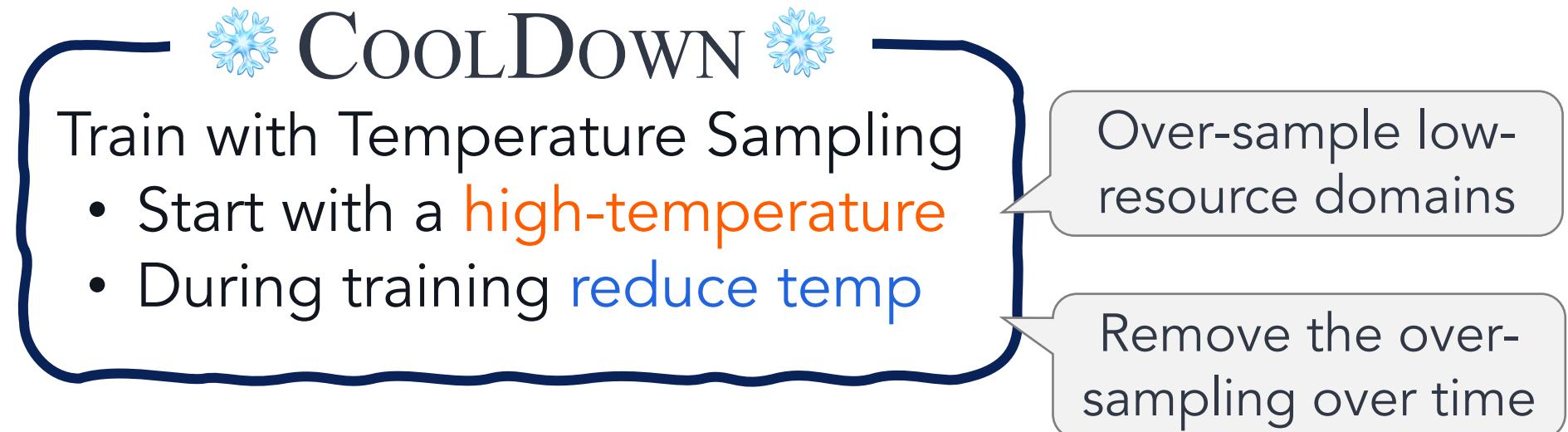
TS vs S: Convergence speed



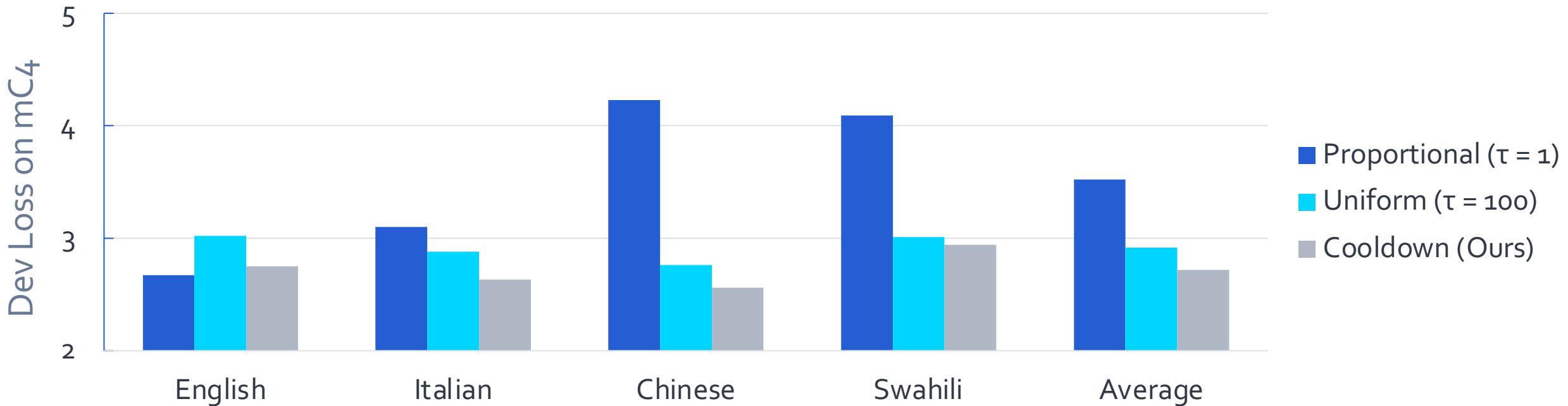
Increasing temperature (2 to 5) makes the convergence even faster, but easy to overfit (to small domains).

❄️ CoolDown ❄️ : Adapting sampling temperature

- When using **temperature sampling**, we need balance between:
 - High-temperature (faster convergence)
 - Low-temperature (avoid overfitting)

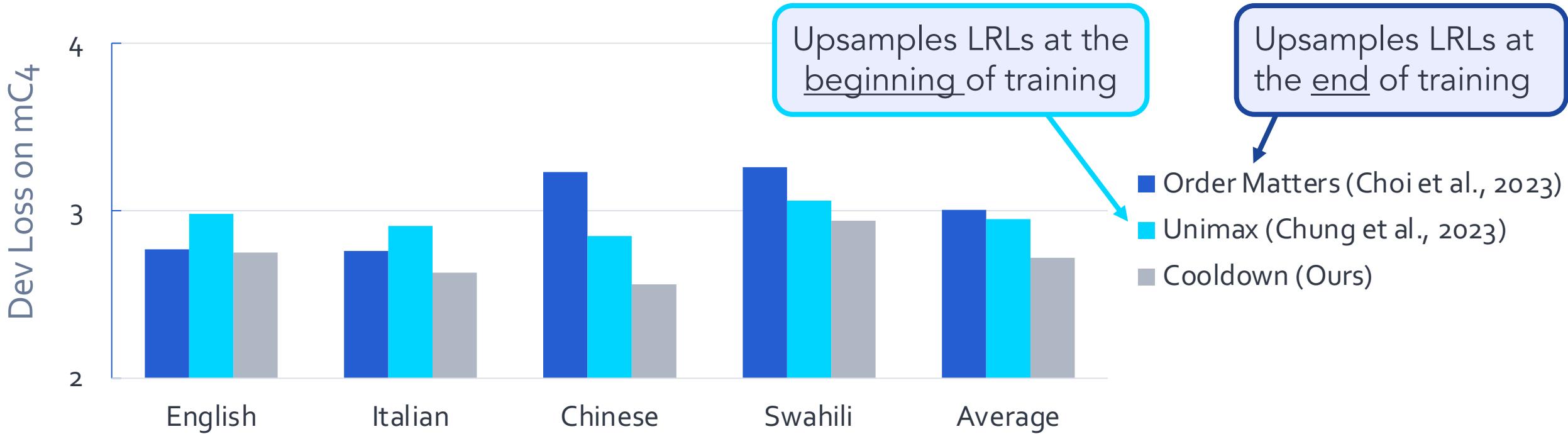


Multilingual Language Modeling



 CoolDown  outperforms
fixed-temperature sampling.

Multilingual Language Modeling



COOLDOWN  outperforms existing work that dynamically adjusts the sampling temperature!

Summary

- Common approaches for training on imbalanced data:
 - Sampling vs Scalarization
- Despite common perception, these two are not equivalent.
- ❄️ CoolDown ❄️
 - A suggested recipe for imbalanced [pre-]training
- Open question: Do these actually scale?

Data \leftrightarrow LLM behavior

- These (data staleness, imbalance, etc.) are not captured by scaling laws.
- Our understanding of data and its impact on behavior remains quite primitive.
- The heavy focus on “scaling laws” may be have been counterproductive, as it disincentivizing a deeper understanding of data.



Today

Science of AI

- Rethinking data.

AI for Science

- Few new challenge problems



Today

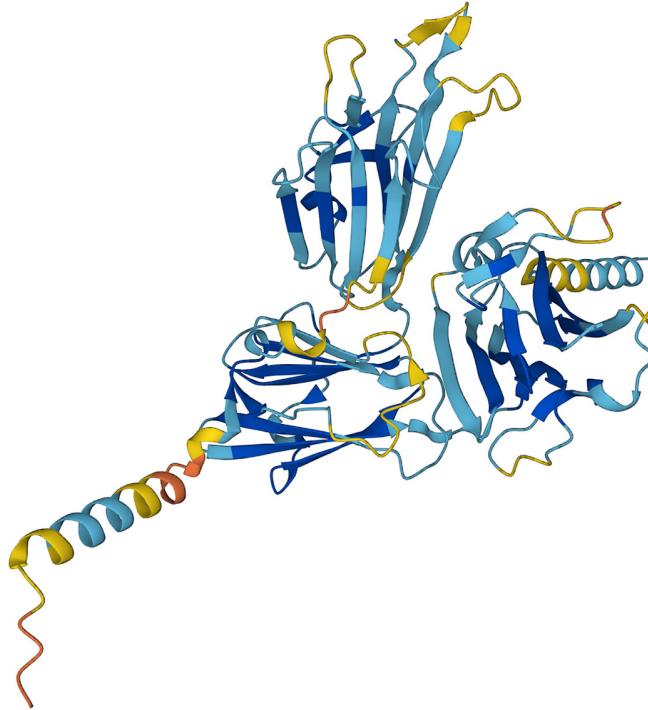
Science of AI

- Rethinking data.

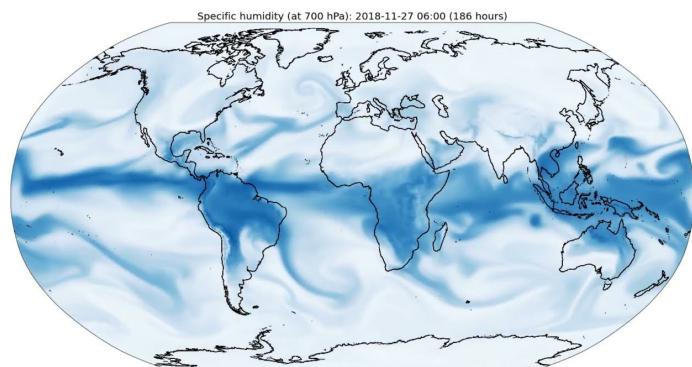
AI for Science

- Few new challenge problems

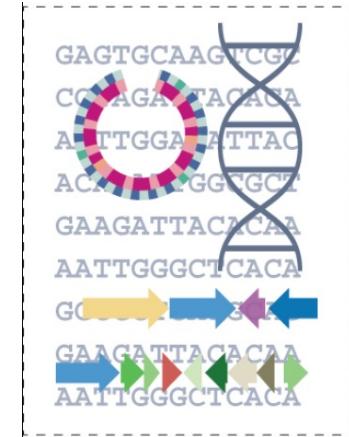
AI for Science



Protein Folding
(e.g., AlphaFold, 2022)

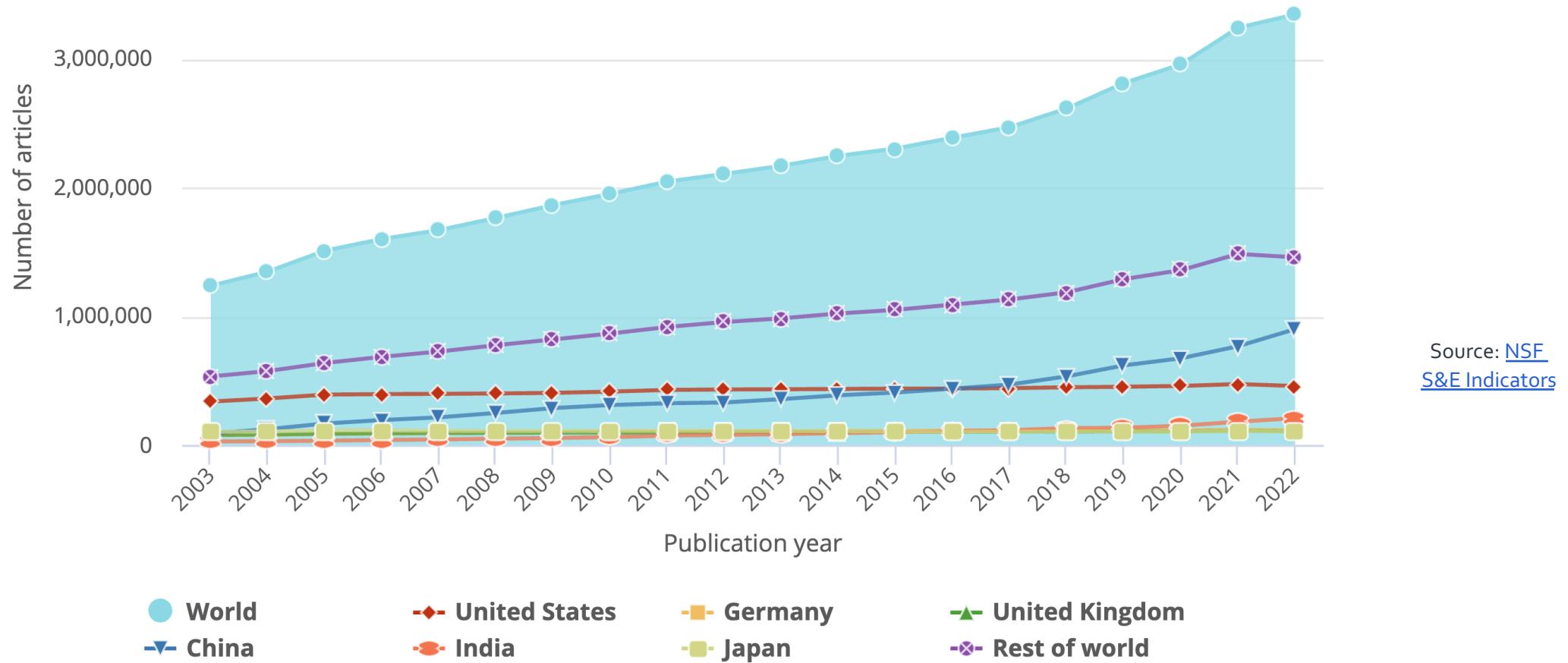


Climate Modeling
(e.g., AlphaFold, 2022)



Drug Discovery
(e.g., Insilico Medicine)

“Digesting” science is getting difficult



Source: [NSF](#)
[S&E Indicators](#)

It's near impossible to keep track of science, even within its sub-fields.

Data ↔ LLM behavior ↔ Science discovery

- Cutoff dates and data staleness:
 - Science keeps evolving. How do you know that your model retains the latest version of findings?
- Data imbalance:
 - Few findings/papers are overly repeated. A lot of science is less repeated (if any).

Science queries: Information seeking vs Aggregative

- Information seeking queries:



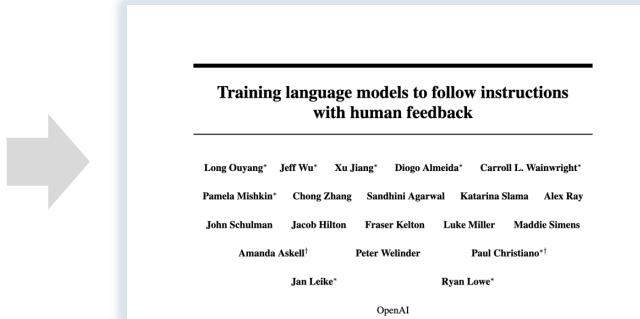
Query: What was GPT3's approach to alignment?



Repositories of science: arXiv, Google Scholar, etc.



Potentially relevant papers



Summary

Science queries: Information seeking vs Aggregative

- Aggregative queries:



Query: Compare alignment approaches used in notable recent LLMs.



Information retrieval systems:
Google Scholar etc.



Relevant papers, sorted

Precision issues



Chatbots
(parametric or retrieval-augmented)



Response

Recall issues

Not able to identify a meaningful set of column values (table schema)

Here's a table summarizing recent language models and their alignment processes:

ChatGPT

Create a table where the rows are recent language model and the columns are different details about their alignment process.

A lot of details are missing here!
(many open-source models)

Language Model	Alignment Process	Description
InstructGPT	Reinforcement Learning from Human Feedback (RLHF)	Trained to follow instructions by learning from human-provided feedback, enhancing its ability to align with user intentions.
ChatGPT	Reinforcement Learning from Human Feedback (RLHF)	Utilizes human feedback to fine-tune responses, aiming to improve relevance and appropriateness in conversational contexts.
Claude	Reinforcement Learning from AI Feedback (RLAIF)	Employs AI-generated feedback based on constitutional principles to guide its behavior, reducing reliance on human input.
Gemini	Reinforcement Learning from Human Feedback (RLHF)	Incorporates human feedback to align its outputs with user preferences, aiming for safer and more accurate responses.
Sparrow	Reinforcement Learning from Human Feedback (RLHF)	Focuses on aligning dialogue agents through targeted human judgments to improve safety and usefulness.
DeepSeek R1	Reinforcement Learning with Novel Training Methods	Introduces innovative training techniques that prioritize correct answers over human-legible reasoning, leading to unexpected behaviors.

What I manually created for my course:

Create a table where the rows are recent language model and the columns are different details about their alignment process.

Language Model	Release	Base	Alignment Algorithm(s) Used	Alignment Data Sources for alignment
GPT-3-instruct	2020	GPT-3	SFT --> RLHF/PPO	Curated datasets with human-labeled prompts and responses
GPT-4	2023	GPT-4 pre-trained?	SFT --> RLHF/PPO	Curated datasets with human-labeled prompts and responses
Gemini	2023	Gemini pre-trained?	SFT --> RLHF/PPO	Curated datasets with human-labeled prompts and responses
LLaMA2	2023	LLaMA2 pre-trained	SFT --> RLHF/PPO	Curated datasets with human-labeled prompts and responses
LLaMA3	2024	LLaMA3 pre-trained	Iterate: Rejection sampling --> SFT --> DPO	10 million human-annotated examples. The alignment process was conducted over multiple rounds, with each round involving the collection of new preference annotations and SFT data.
Alpaca	2023	LLAMA 1	SFT	Self-Instruct, 52,000 input-output pairs
Qwen2.5	2024	Qwen2.5 pre-trained	SFT -> DPO -> GRPO	1 million samples
Tulu 3	2024	Llama 3.1	SFT -> DPO -> RLVR	near 1 million samples
DeepSeek (V3)	2024	DeepSeek pre-trained	SFT -> GRPO	1.5 million samples (reasoning + non-reasoning tasks). Reasoning data was generated by specialized models. Non-reasoning data was produced by DeepSeek-V2.5 and validated by human reviewers.

arXiv2Table: A benchmark for aggregative questions

- 2.1K user demand prompts + tables (inherited from prior work).
- Evaluation framework based on utilization.

<https://github.com/JHU-CLSP/arXiv2Table>



Prompt: Generate a table comparing video datasets



Potentially relevant papers

	Dataset size	Annotation method	Intended Application	Evaluation Metric
Paper 1	1,200 video sequences	Subjectively annotated	Objective VQA method development	Subjective Mean Opinion Score
Paper 2	585 videos	Subjective video quality scores via crowdsourcing	NR video quality prediction advancement	Subjective video quality scores
Paper 3	153,841 videos	Coarsely annotated set with five quality ratings each	Deep-learning VQA model training	Spearman rank-order correlation coefficient
Paper 4	1 million YouTube videos	N/A	Large-scale video classification and action recognition	Performance improvements over baselines

Tabular summary

Another type of aggregative questions: Distribution of science contributions

- Aggregative queries:
 - *Query: How many papers are in AI?*
 - *Query: How many papers in AI are about extending RL?*
 - *Query: How many papers in AI are about extending DPO?*
 - *Query: How many papers use RL in surgical applications?*
 - ...

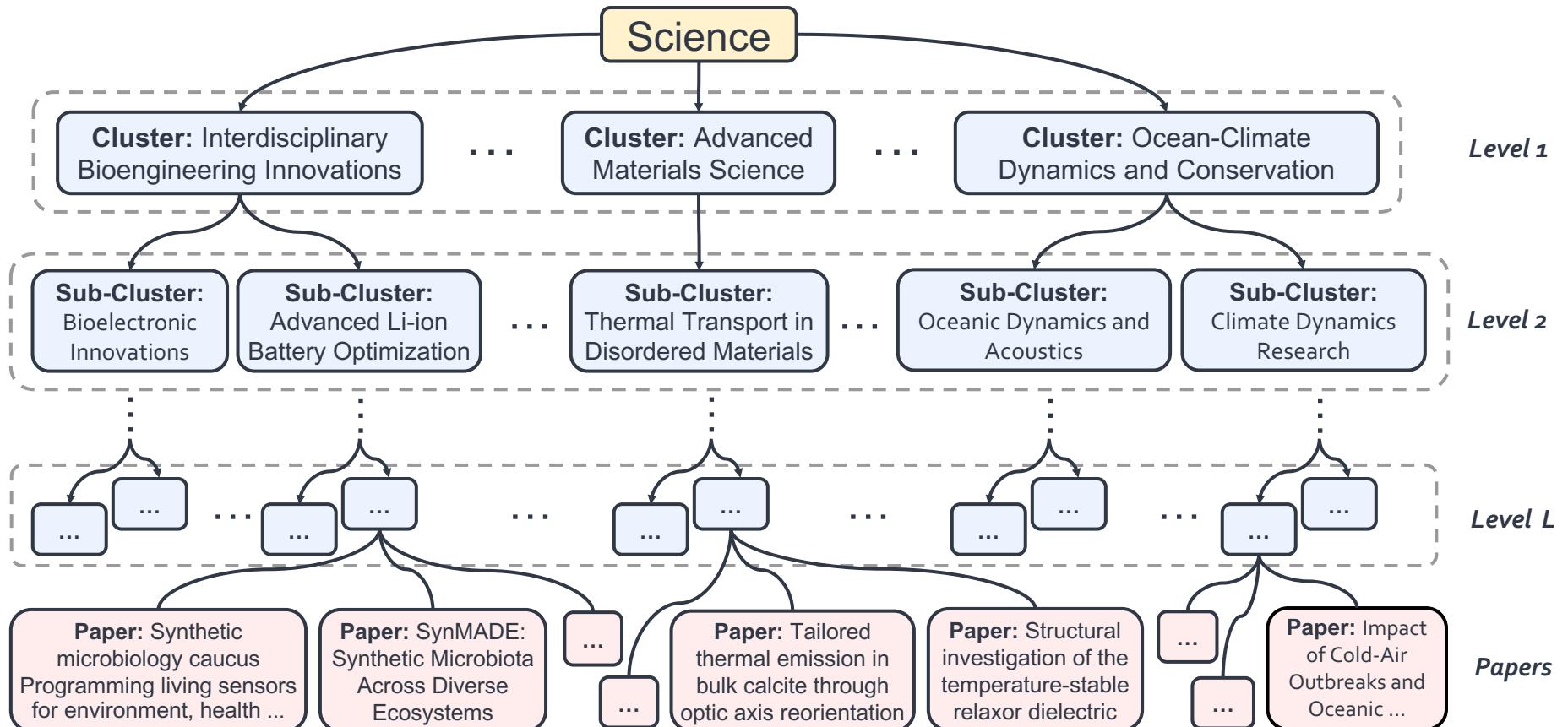
Problem statement: How do we compute the distribution of scientific efforts at varying levels of granularity?

Distribution of science contributions: why?

- Why we should care?
 - Allows us (e.g., policy makers) identify areas that are **over/under-explored**.
 - Necessary for ensuring effective investment in science.
 - Not just science! Many institutions (e.g., venture capital firms) need to understand the distribution of their investments.

Problem statement: How do we compute the distribution of scientific efforts at varying levels of granularity?

An idealistic hierarchy of science



Problem statement: How do we compute the distribution of scientific efforts at varying levels of granularity?

Existing hierarchies of science



- Manually curated with human curation
- They also have papers and their summaries within each category.

[Home](#) » [Engineering](#) » [Mechanical Engineering](#)



https://orkg.org/home/R261/Mechanical_Engineering

Existing hierarchies of science



- Manually curated with human curation
- They also have papers and their summaries within each category.

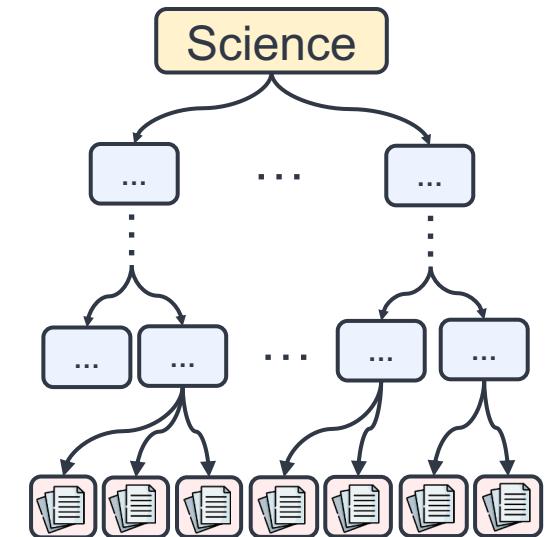
They have an ad for
hiring people!!

ORKG CURATION GRANTS
RECEIVE 400 EURO MONTHLY
APPLICATION DEADLINE:
MAY 11TH, 2025

<https://orkg.org/page/call-curation-grants-2025>

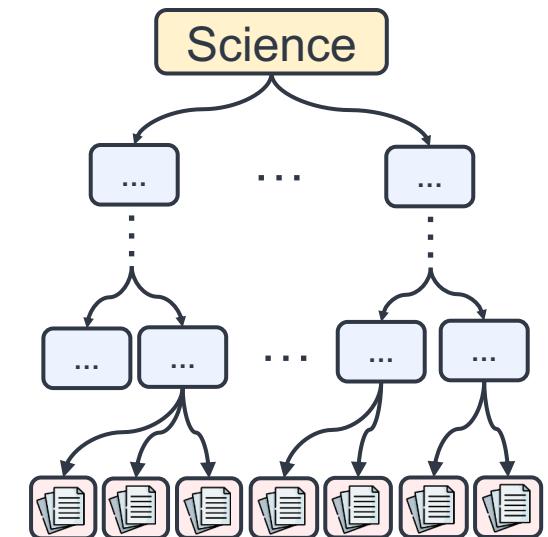
Hierarchies of science paper

1. Representation of the node (*what is a “paper”?*)
2. Evaluating hierarchies
3. Inferring them



Hierarchies of science paper

1. Representation of the node (*what is a “paper”?*)
2. Evaluating hierarchies
3. Inferring them



How do you represent a “paper”?

- We extract different contribution types from each paper:

The Tin Pest Problem as a Test of Density Functionals Using High-Throughput Calculations

Michael J. Mehl,^{1,*} Mateo Ronquillo,² David Hicks,³ Marco Esters,³ Corey Osse,³ Rico Friedrich,³ Andriy Smolyanyuk,³ Eric Gossett,³ Daniel Finkenstadt,⁴ and Stefano Curtarolo^{3,†}

¹*Center for Autonomous Materials Design, Duke University, Durham NC 27708*
²*U. S. Nuclear Power School, Goose Creek, South Carolina 29445*
³*Center of Autonomous Materials Design, Duke University, Durham NC 27708*
⁴*Physics Department, U.S. Naval Academy, Annapolis, Maryland 21402*

(Dated: October 15, 2020)

At ambient pressure tin transforms from its ground-state semi-metal α -Sn (diamond structure) phase to the compact metallic β -Sn phase at 13°C (286K). There may be a further transition to the simple hexagonal γ -Sn above 450K. These relatively low transition temperatures are due to the small energy differences between the structures, $\approx 20 \text{ meV/atom}$ between α - and β -Sn. This makes tin an exceptionally sensitive test of the accuracy of density functionals and computational methods. Here we use the high-throughput Automatic-FLOW (AFLOW) method to study the energetics of tin in multiple structures using a variety of density functionals. We look at the successes and deficiencies of each functional. As no functional is completely satisfactory, we look Hubbard U corrections and show that the Coulomb interaction can be chosen to predict the correct phase transition temperature. We also discuss the necessity of testing high-throughput calculations for convergence for systems with small energy differences.

<https://arxiv.org/pdf/2010.07168>

GPT4
→

Problem statement:

- Problem domain
- Challenges
- Research goal

Solution:

- Approach
- Novelty

Results:

- Findings
- Impact

How do you represent a “paper”?

- We extract different **contribution** types from each paper:



Problem statement:

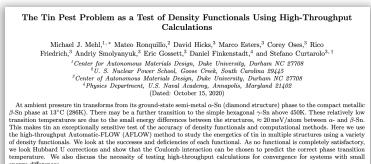
- **Problem domain:** Condensed matter physics
- **Challenges:** Lack of full rotational symmetry in solids leading to new excitations beyond Dirac and Weyl fermions
- **Research goal:** Investigate sixfold excitations in electrides

Solution:

- **Approach:** Propose that a single linear dispersive sixfold excitation can be found in the electride $\text{Li}_{12}\text{Mg}_3\text{Si}_4$ and its derivatives
- **Novelty:** Unique topological bulk-surface-edge correspondence for the spinless sixfold excitation

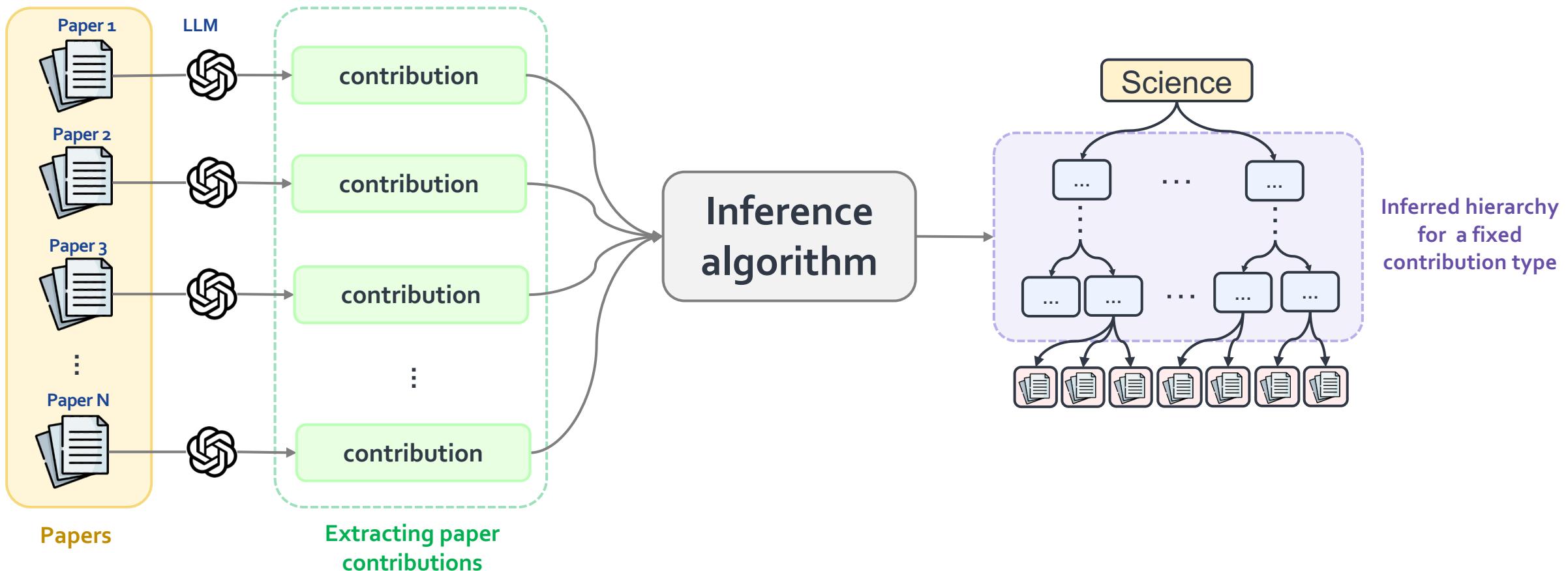
Results:

- **Findings:** The sixfold excitation is formed by floating bands of elementary band representation A@12a. All gapped kz slices belong to a two-dimensional higher-order topological insulating phase, characterized by a quantized fractional corner charge $Q_{\text{corner}} = 3|e|/4$. Hinge arcs are obtained in the hinge spectra of the S_4z -symmetric rod structure.
- **Impact:** Electrides are promising platforms for systematic studies of different topological phases.



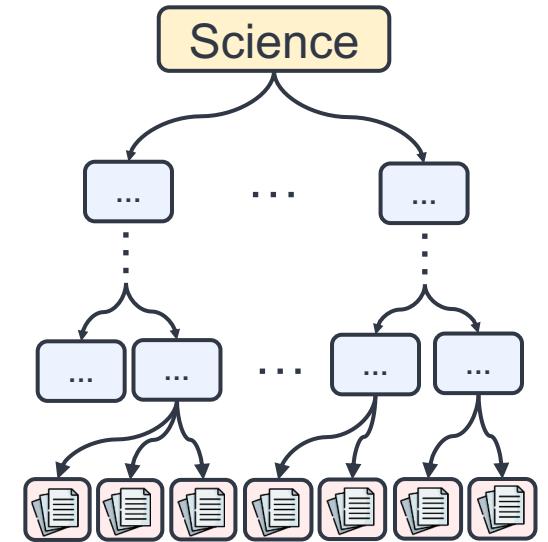
One hierarchy per contribution type

- We extract different contribution types from each paper:



Hierarchies of science paper

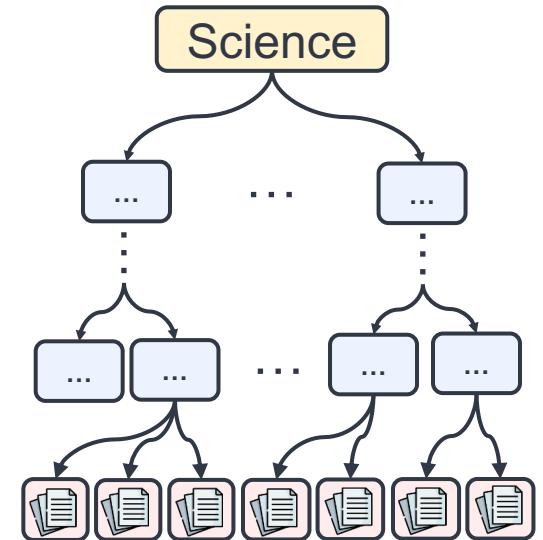
1. Representation of the node (*what is a “paper”?*)
2. **Evaluating hierarchies**
3. Inferring them



How should we evaluate a science hierarchy?

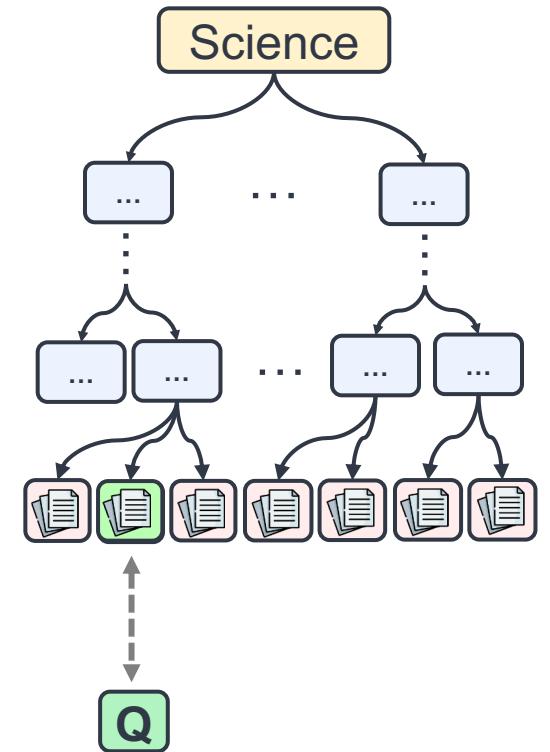
- There is **no reference data** for this task.
- Even if it did, it would need to be updated every few weeks!

In absence of any gold standard,
how can we evaluate a *given* hierarchy?



Evaluation as Utilization

- Idea: A good hierarchy allows one to identify things.
- Setup:
 - Input: A query **Q** that is answered at a leaf node.
 - Starting from root node, let a judge (an LLM) traverse the tree to identify the related leaf node.
- Metric: The accuracy of finding the relevant leaf.



Evaluation as Utilization: Example

Q

Title: Controlling nonlinear interaction in a many-mode laser by tuning disorder

Abstract: Many-mode lasers with nonlinear modal interaction can model many-body physics...

We introduce random phase fluctuation to tune lasing modes...

Level 1 Clusters:

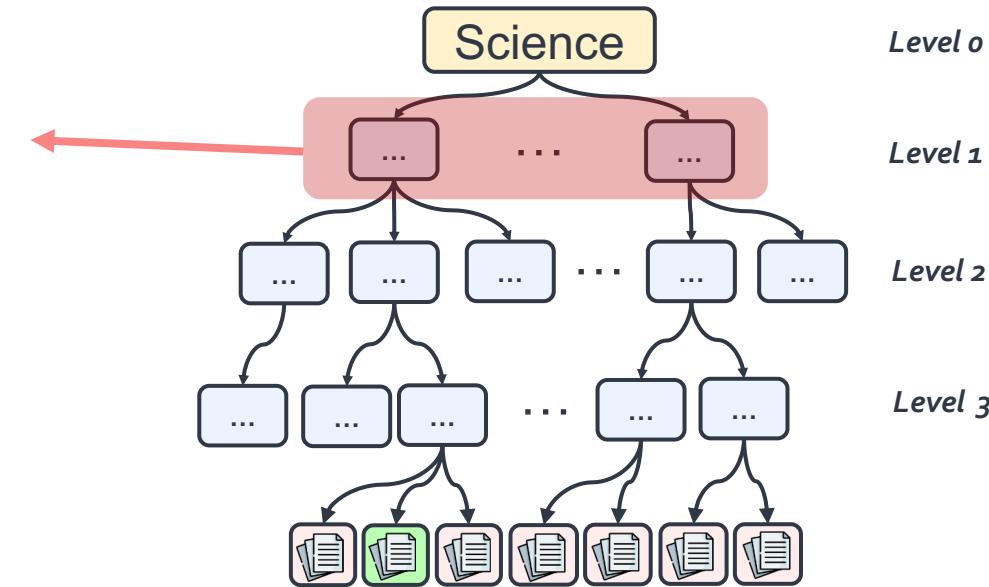
1. Materials for energy/env/health
2. AI for complex systems
3. Ocean & atmosphere dynamics
4. Topological & quantum materials
5. Photonic & quantum systems
6. Sustainable energy materials



The most appropriate cluster is:
5. Photonic & quantum systems

Evaluator:

Qwen2.5 (32B)



Evaluation as Utilization: Example

Q

Title: Controlling nonlinear interaction in a many-mode laser by tuning disorder

Abstract: Many-mode lasers with nonlinear modal interaction can model many-body physics...

We introduce random phase fluctuation to tune lasing modes...

Level 2 Clusters:

1. Advancements in Nano-Optical Resonance, Photonic Devices, and Nonlinear Optical Materials
2. Advances in Photonic Systems and Quantum Technologies for Enhanced Optical Control and Manipulation
3. Advancements in Quantum Optomechanics and Optical Physics for Enhanced Precision and Control
4. Advanced Photonic Technologies for Turbulence-Resilient Optical Communications and Wave Manipulation
5. Experimental Investigations of Quantum Gravity and Macroscopic Quantum Systems

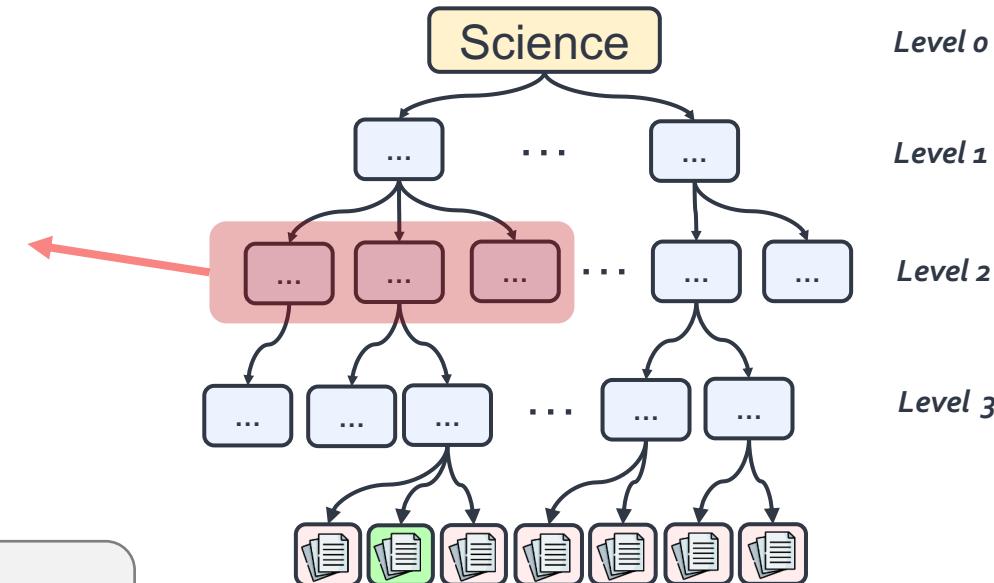


Evaluator:

Qwen2.5 (32B)

The most appropriate cluster is:

1. **Advancements in Nano-Optical Resonance, Photonic Devices, and Nonlinear Optical Materials**



Evaluation as Utilization: Example

Q

Title: Controlling nonlinear interaction in a many-mode laser by tuning disorder

Abstract: Many-mode lasers with nonlinear modal interaction can model many-body physics...
We introduce random phase fluctuation to tune lasing modes...

Level 3 Clusters:

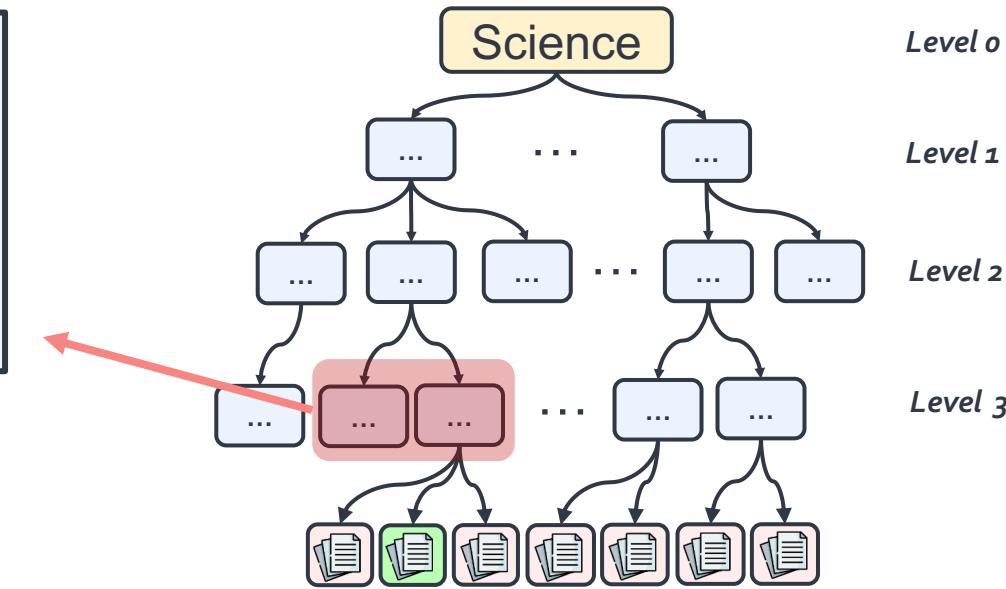
1. Advancements in Nano-Optical Resonance and Photonic Devices
2. Advanced Nanophotonic Materials and Metamaterials for Enhanced Optical Interactions
3. Advancements in Nonlinear Photonics and Quantum Optics
4. Advanced Optical Materials and Phenomena for Enhanced Metrology and Control
5. Advanced Nonlinear Optical Materials and Photonics



Evaluator:

Qwen2.5 (32B)

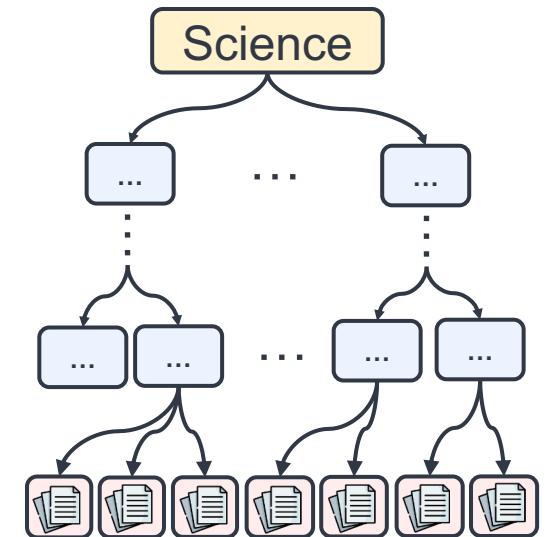
The most appropriate cluster is:
3 - Advancements in Nonlinear Photonics and Quantum Optics



A good hierarchy allows one to identify things!

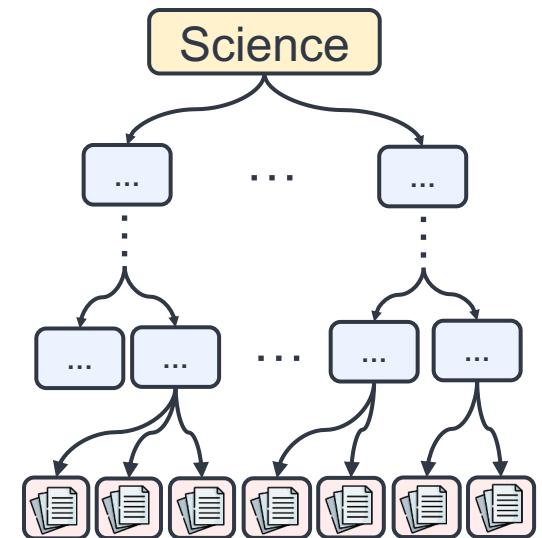
Hierarchies of science paper

1. Representation of the node (*what is a “paper”?*)
2. Evaluating hierarchies
3. Inferring them



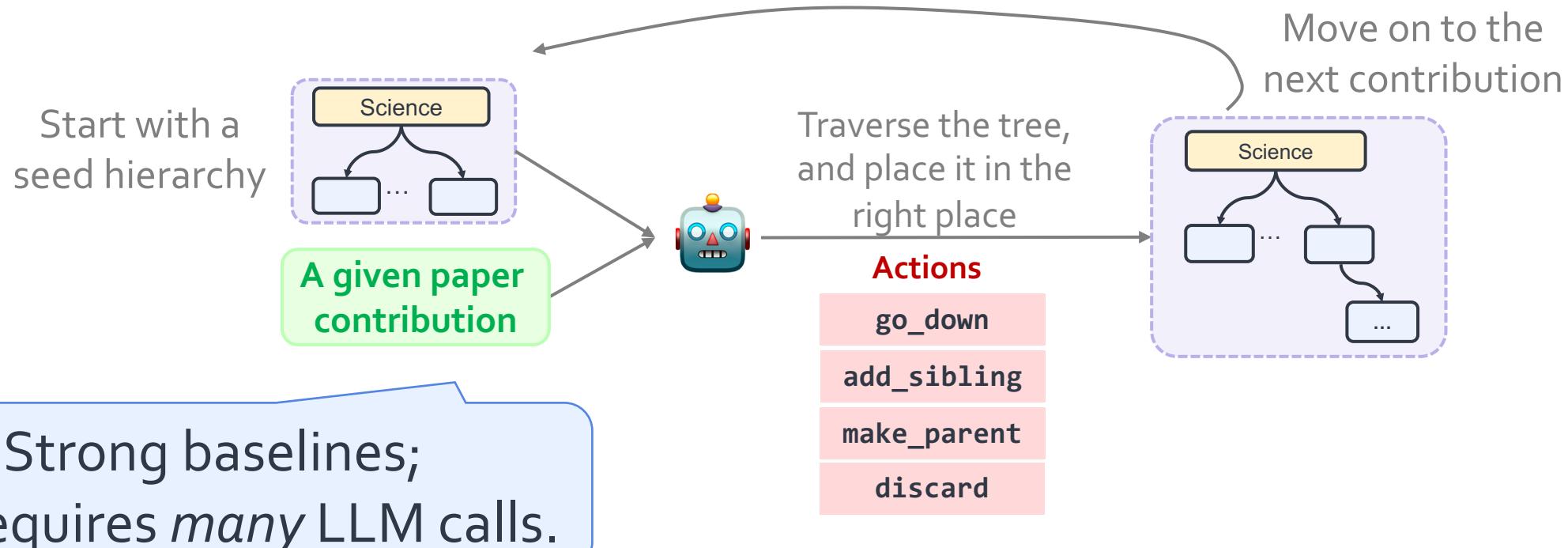
Can LLMs solve this?

- LLMs are good at local decisions (if they're common).
 - Example: RL is a variant of AI.
- Challenges :
 - The task here requires a global coherence, but you cannot fit the whole hierarchy/papers in the context.
 - A lot of science is part of the long-tail where LLMs are less reliable.



Pure-LLM based approaches

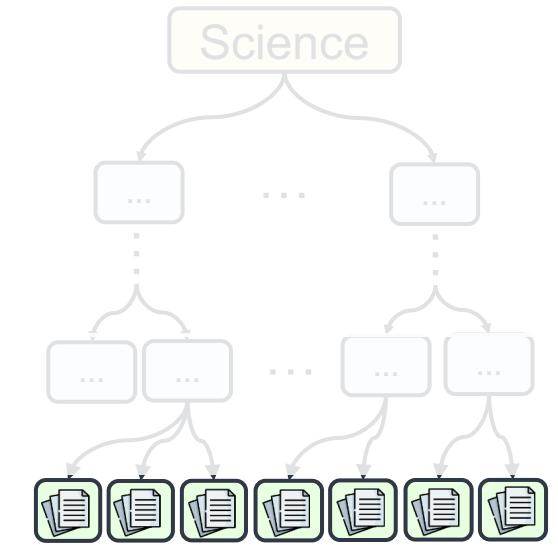
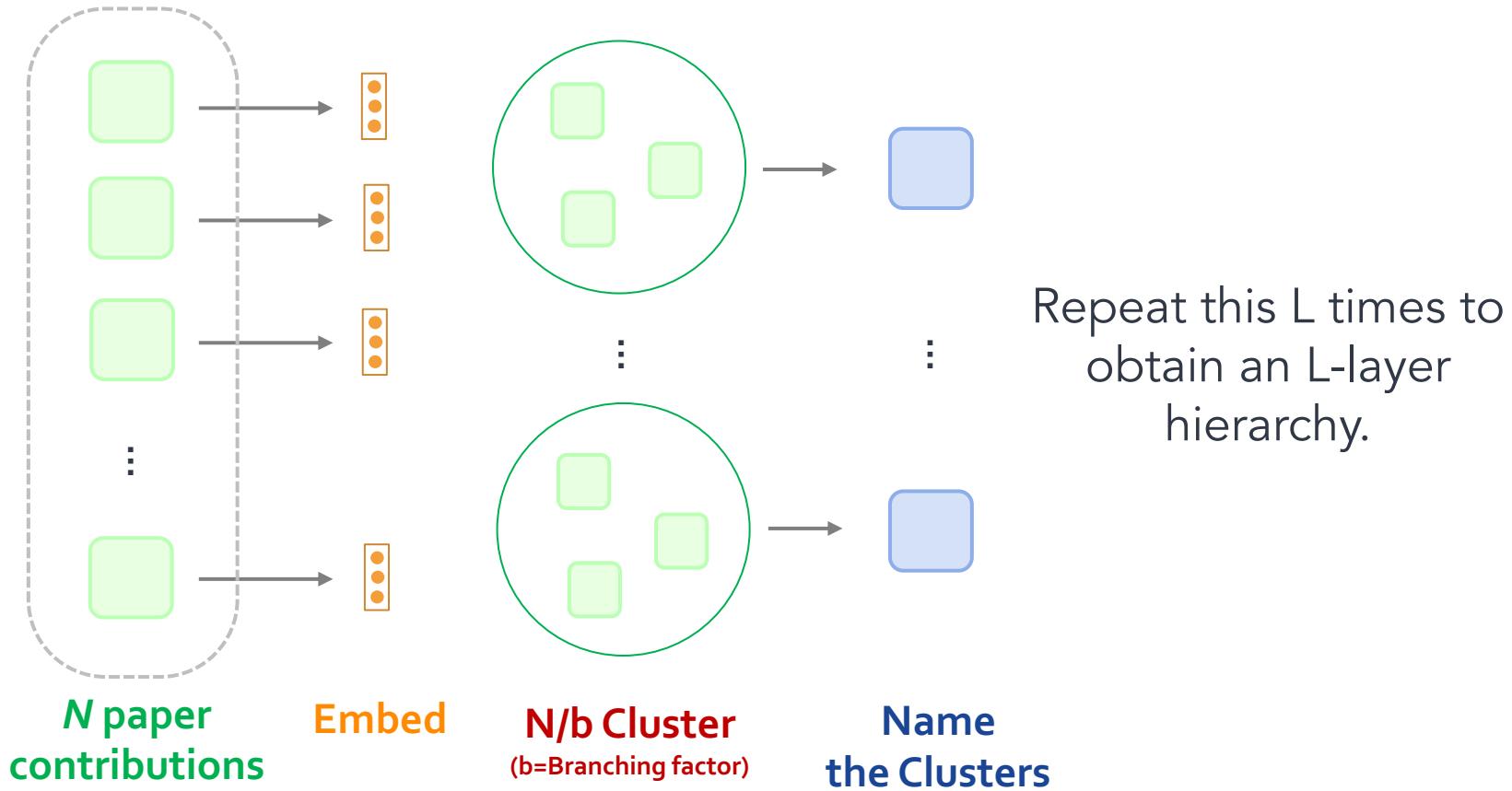
- Approach: Place one item at a time in a seed hierarchy.



$O(C \cdot \log C)$
(C : number of contributions)

Scychic: Combining LLM and Clustering

- Cluster papers and then name with LLMs (bottom-up)



Number of LLM calls: $O(C/b)$
(C : number of contributions;
 b : branching factor)

Evaluation

- A semi-random set of 2K papers
 - (results on 10K papers in the paper)

Scychic has high accuracy while requiring much less LLM calls.

The task is difficult.
Help us solve it!

- Examples here: <https://ia1.wse.jhu.edu/port4000/>

Method	Accuracy ↑	# of LLM Calls ↓
<i>Contributions type: Topic</i>		
SCYCHIC	14.9	
↳ Top-down	14.9	322
↳ Bottom-up	13.9	
↳ Pure LLM baseline	18.0	61K
<i>Contributions type: Problem Statement</i>		
SCYCHIC	51.1	
↳ Top-down	49.0	322
↳ Bottom-up	45.9	
<i>Contributions type: Solution Statement</i>		
SCYCHIC	48.8	
↳ Top-down	45.9	322
↳ Bottom-up	14.8	
<i>Contributions type: Results Statement</i>		
SCYCHIC	46.4	
↳ Top-down	45.2	322
↳ Bottom-up	40.0	

Putting things together

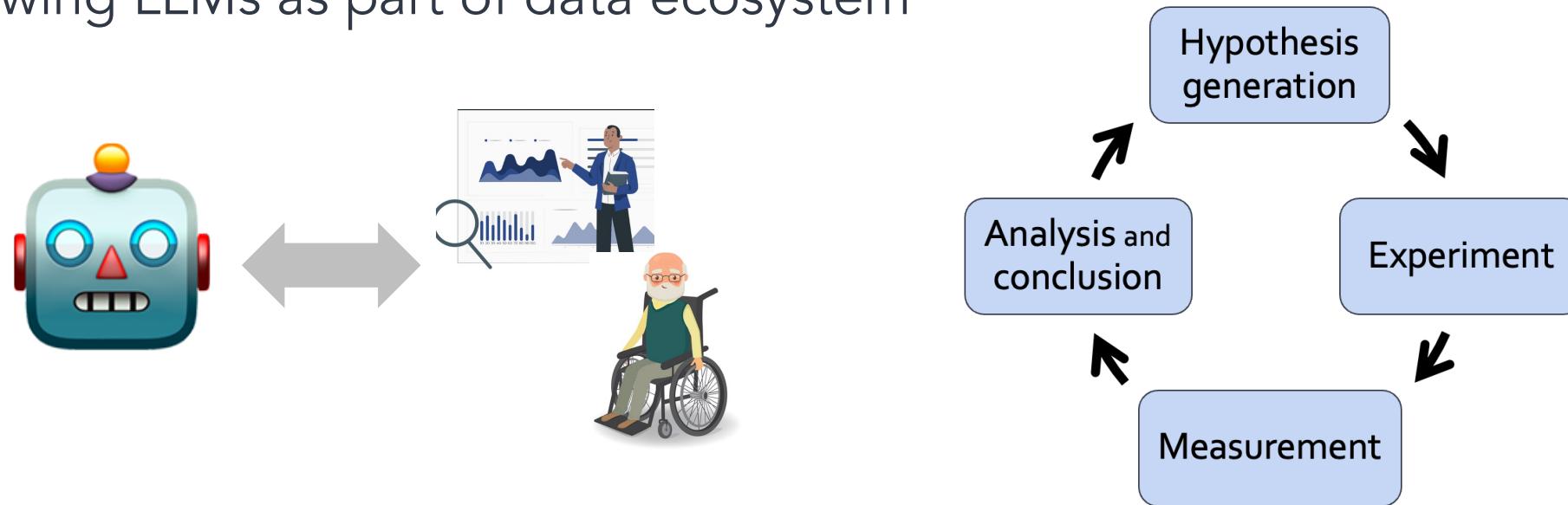
- Motivation: Aggregative queries that require understanding the distribution of scientific efforts across different sub-areas.
- Science Hierarchography: Large-scale hierarchical summarization of scientific literature.
- We proposed framework to
 1. Evaluate: measuring hierarchy quality, in absence of gold data.
 2. Infer hierarchies: balancing quality and scalability. Far from perfect!

AI for facilitating science: how far are we?

- LLM have brought up **massive opportunities** to speed up science.
 - Accessing and digesting, reviewing, ...
- Not just LLMs— we should think about “systems” that are built upon them. There are many other challenges remain here.
- At the same time, LMs **are brittle** and likely to remain brittle.
 - We need innovative ways to **scope them** and **contain their brittleness**.

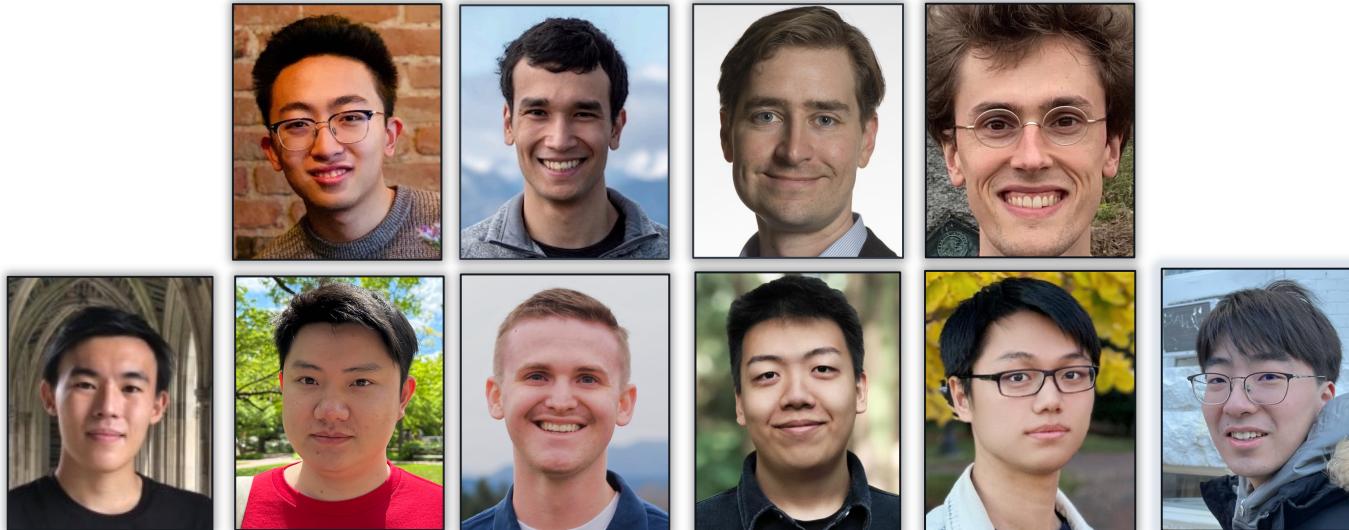
Not AGI: Helpful agents w/ bounded autonomy

- Systems that are **robust within well-defined domain**, might be better alternatives to generalist brittle models.
 - Growing LLMs as part of data ecosystem



- It's more helpful to pursue **scoped agents** that **augment us**.
 - Not to replace us.

Thanks for wonderful collaborators on these projects:



Funding:



