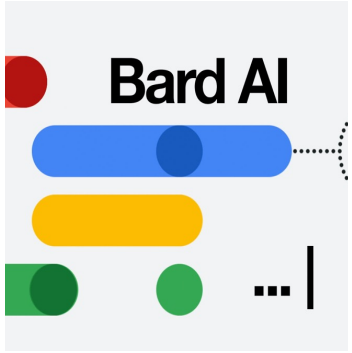
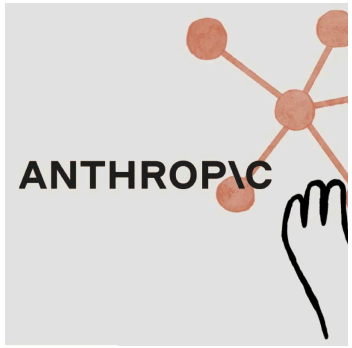
The Uphill Battle to
Making LLMs Reliable

Daniel Khashabi

The success we dreamed of



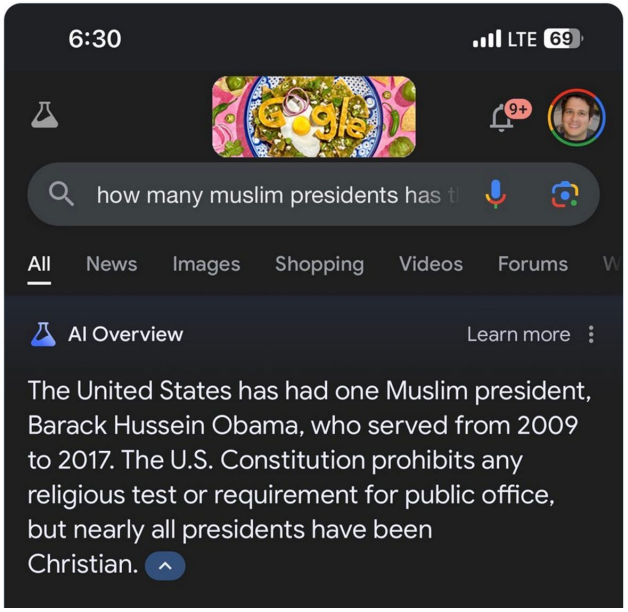


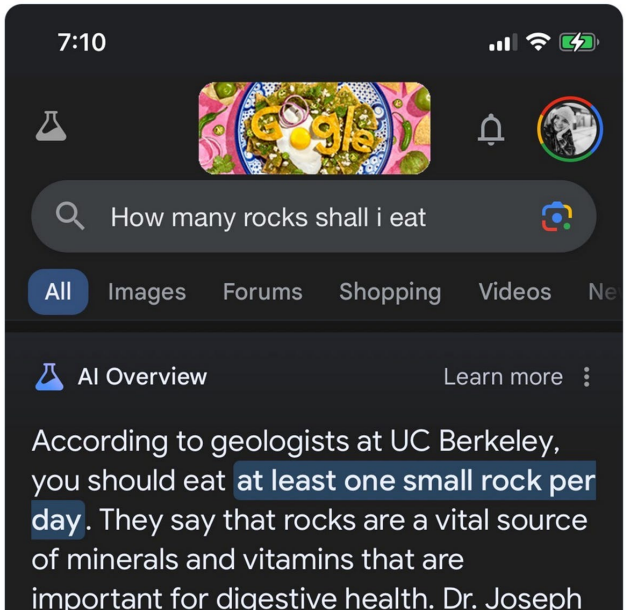
Language models that are remarkably capable at solving many important NLP benchmarks.

Model capabilities—haves vs have-nots

- ✓ Fluent generation
- ✓ Instruction following
- ✓ Several rounds of conversation
- ✗ …

LLMs produce false information





The New York Times

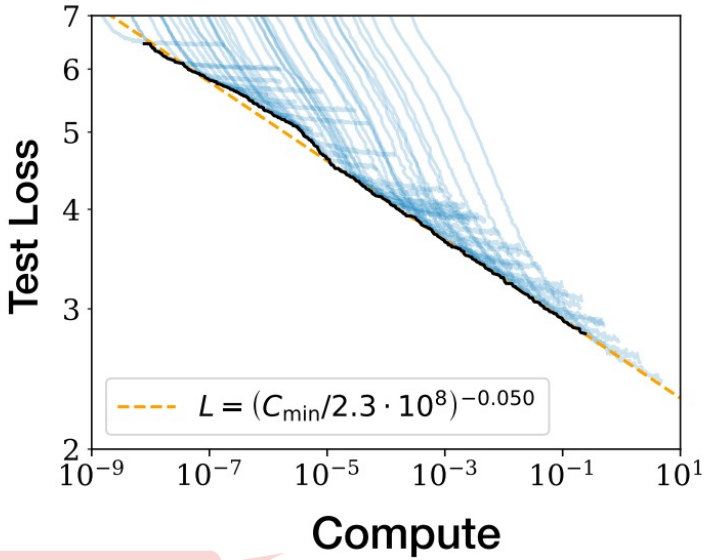
June 1, 2024

Google Rolls Back A.I. Search Feature After Flubs and Flaws

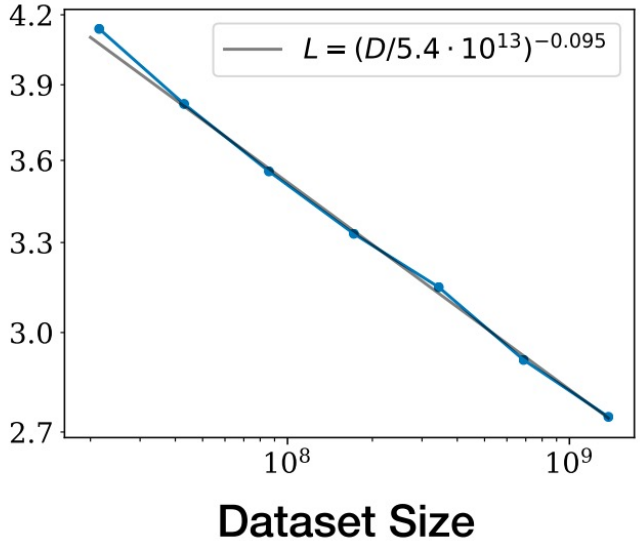
Google appears to have turned off its new A.I. Overviews for a number of searches as it works to minimize errors.

Will "scaling" solve LLM brittleness?

Linear



Exponential



Exponential

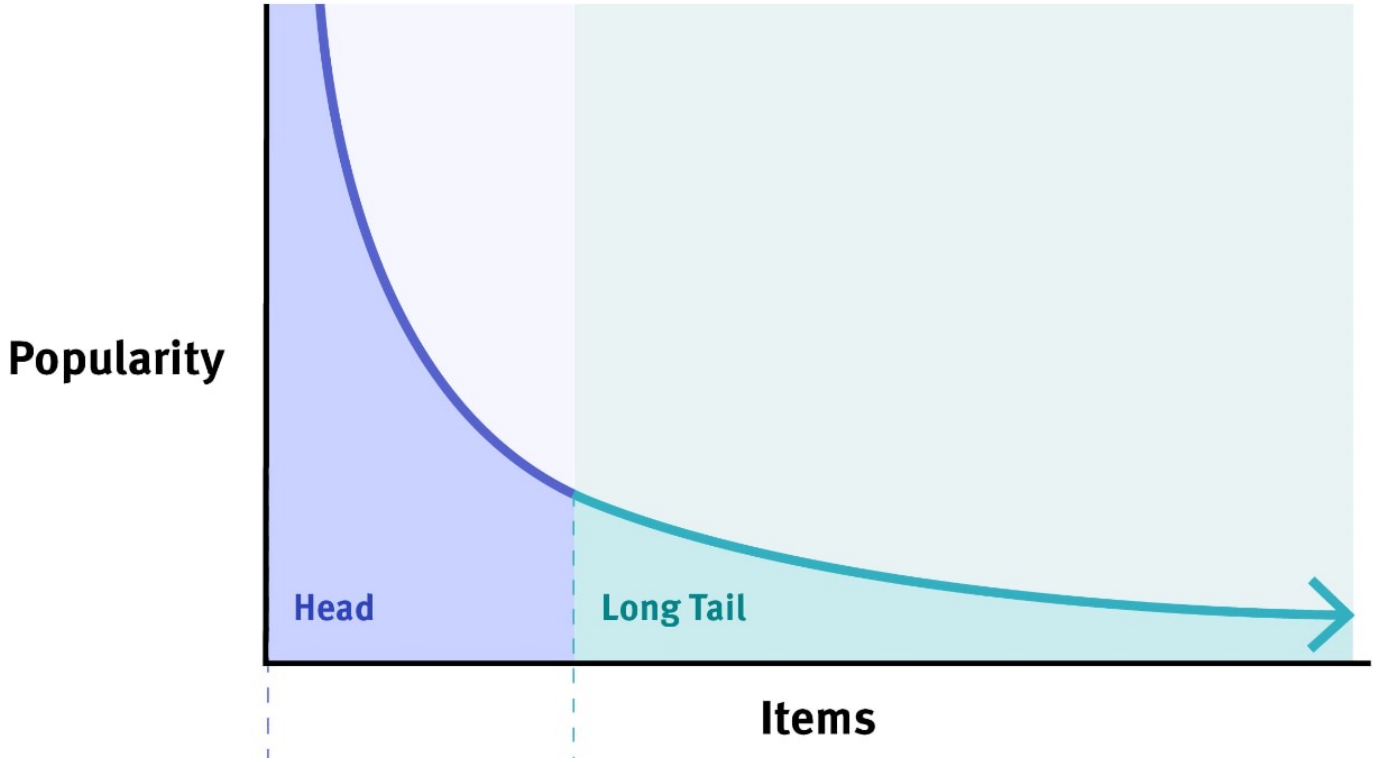
Kaplan et al. 2020; among others

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

Model capabilities—haves vs have-nots

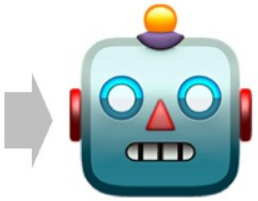
- ✓ Fluent generation
- ✓ Instruction following
- ✓ Several rounds of conversation
- ✗ Cost-inefficient to scale (exponential scale for linear gains)

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

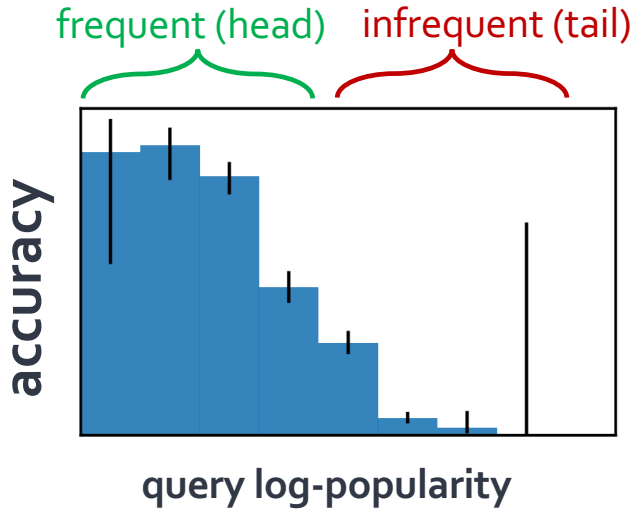


Infrequent things are challenging for LLMs

Who was the director of The Titanic?



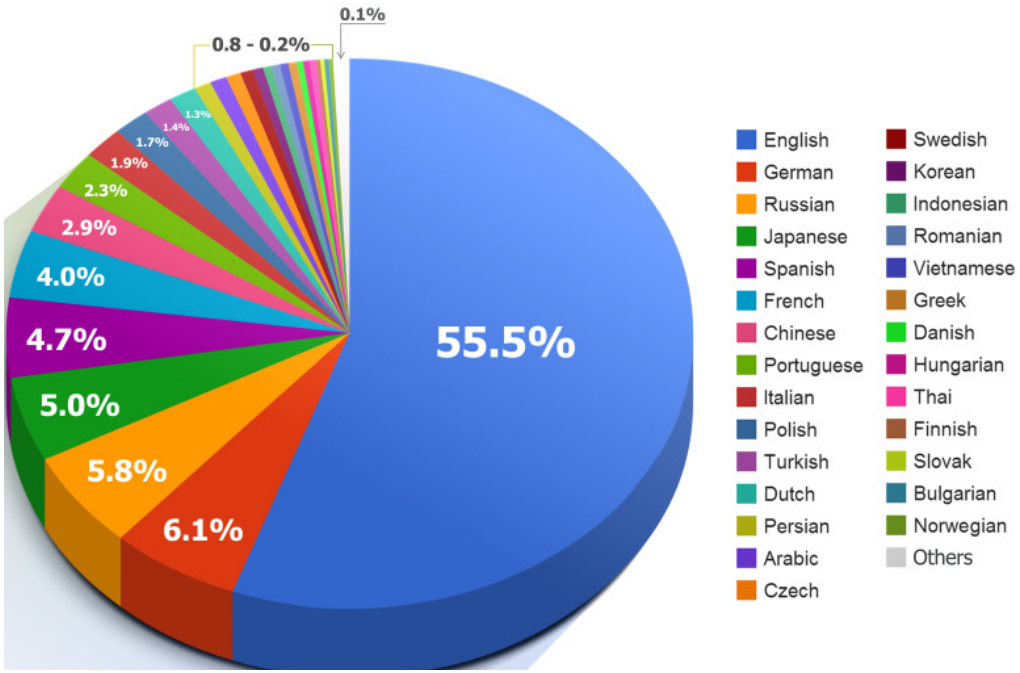
Response



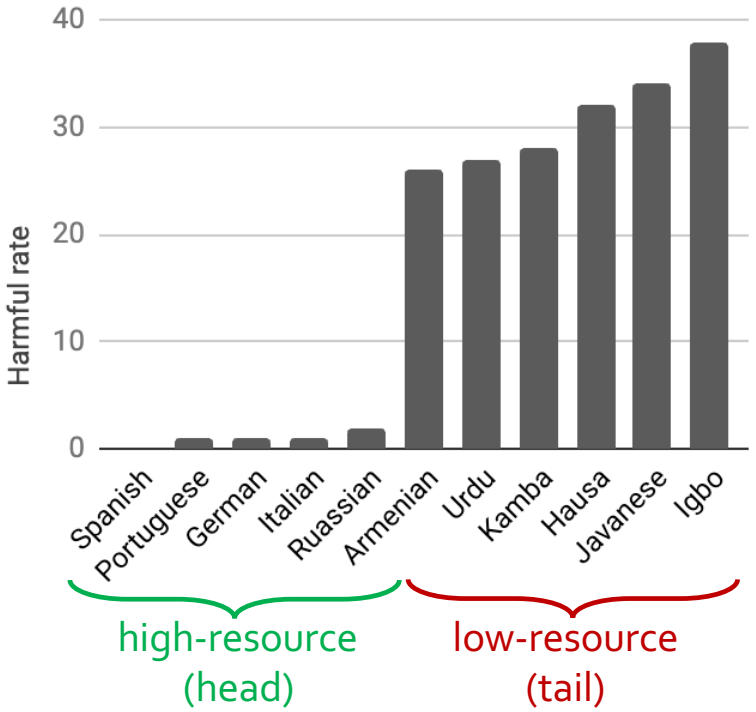
Factual accuracy of LLMs is positively correlated with "popularity" of the input prompts.

Mallen et al. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories, *ACL* 2023

Models are unsafe in low-resource languages



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



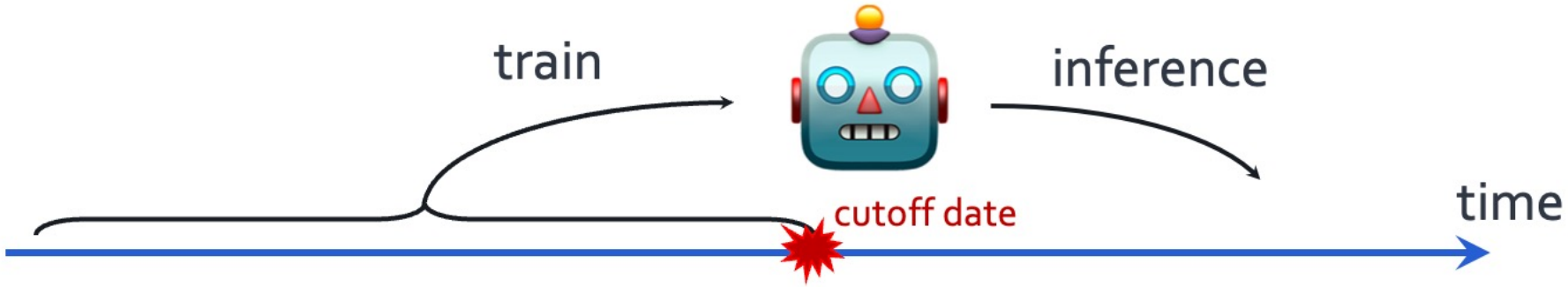
Shen et al. The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Context., *ACL* 2024

Model capabilities—haves vs have-nots

- ✓ Fluent generation
- ✓ Instruction following
- ✓ Several rounds of conversation
- ✗ Cost-inefficient to scale (exponential scale for linear gains)
- ✗ Long tail of problems

Temporal misalignment: LLMs stale over time

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

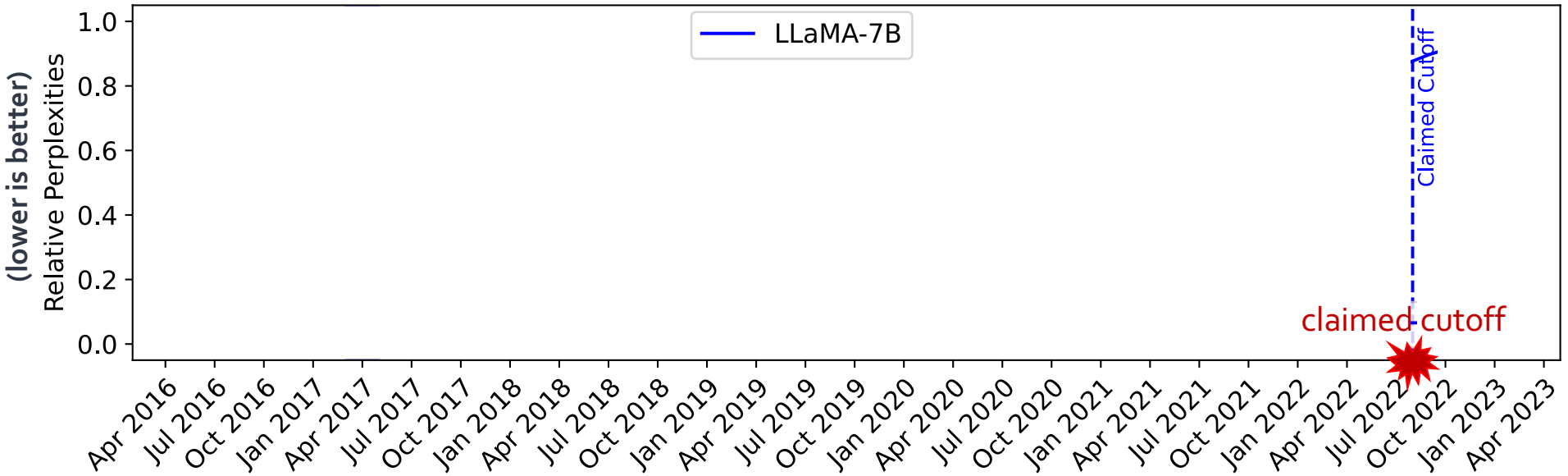


Are LLMs' knowledge before cutoff date consistently good?

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

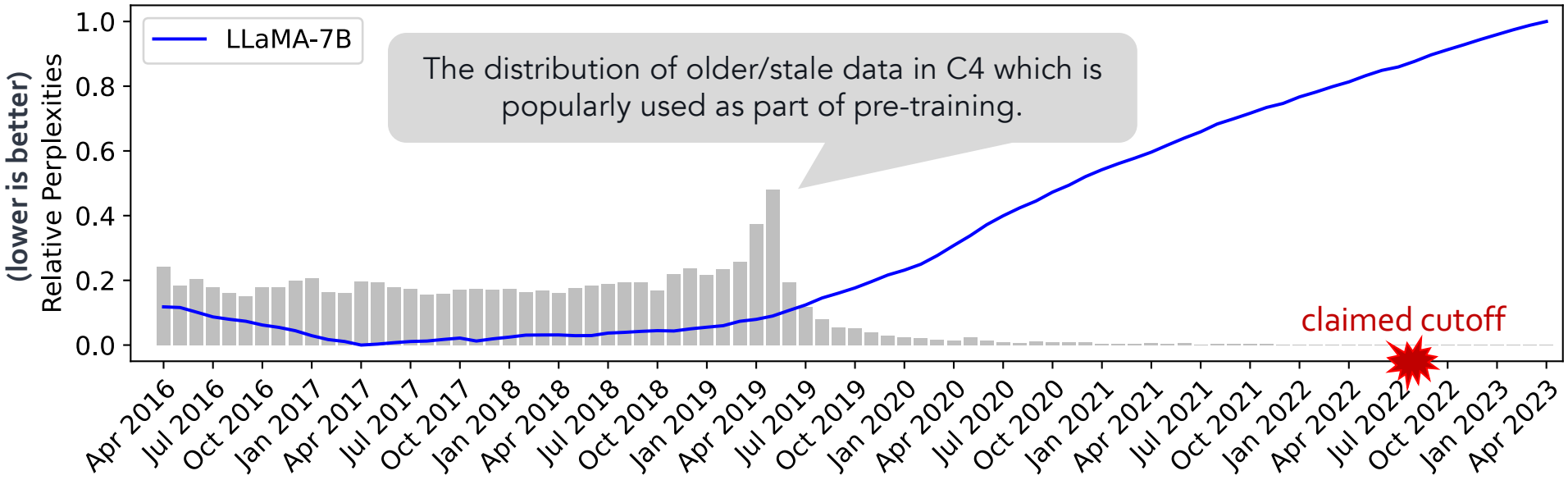
LLM quality in older time-stamped data

- We evaluate LLaMA model on past version of Wikipedia.



Cheng et al. Dated Data: Tracing Knowledge Cutoffs in Large Language Models.., *arXiv* 2024

Pre-training data contain lots of old/stale data



Cheng et al. Dated Data: Tracing Knowledge Cutoffs in Large Language Models.., *arXiv* 2024

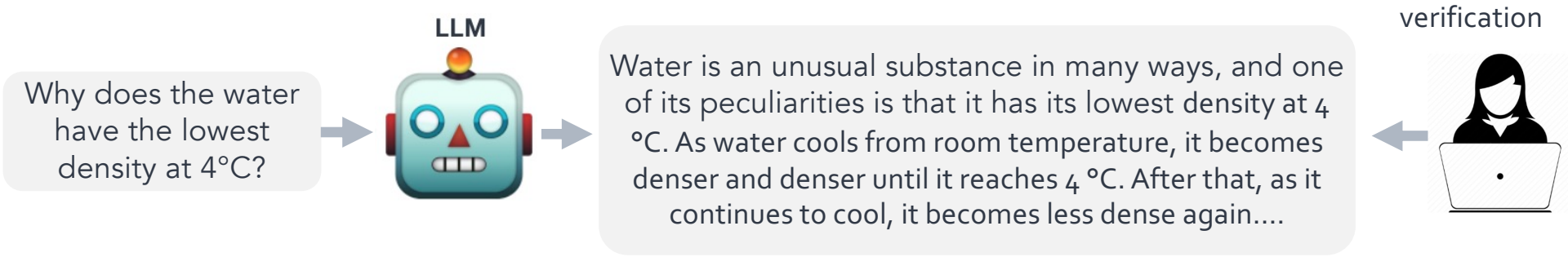
Model capabilities—haves vs have-nots

- ✓ Fluent generation
- ✓ Instruction following
- ✓ Several rounds of conversation
- ✗ Cost-inefficient to scale
- ✗ Long tail of problems
- ✗ Interference of stale knowledge

- General-purpose uses of LMs will remain brittle (at least, in short term)
- What matters is "containing" them.

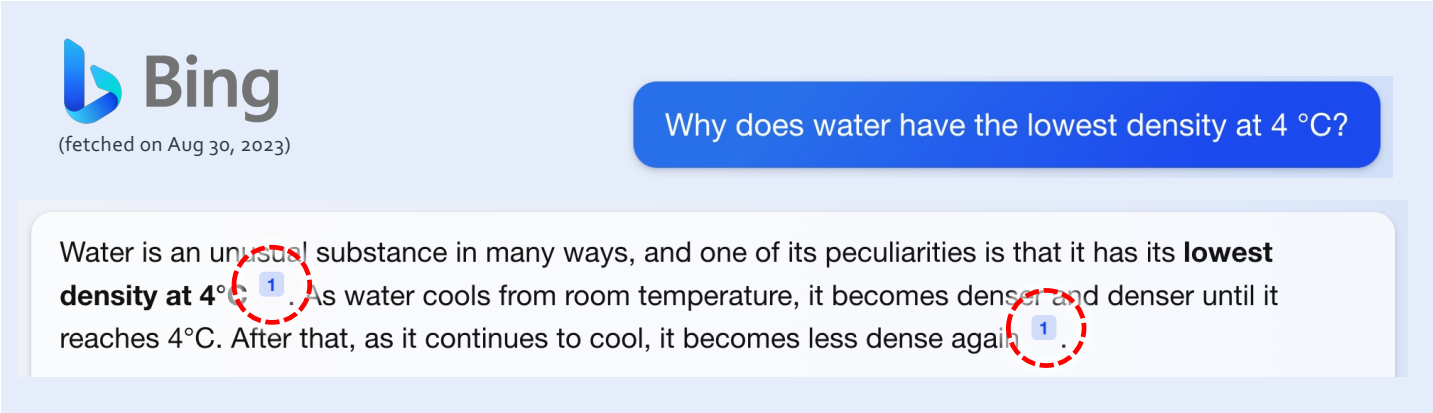
We should make “verifiability” easier

- The burden of LLM mistakes falls on the users.



- A good interface should allow easy “verification” of responses.

Verifying LLM outputs by citing sources?



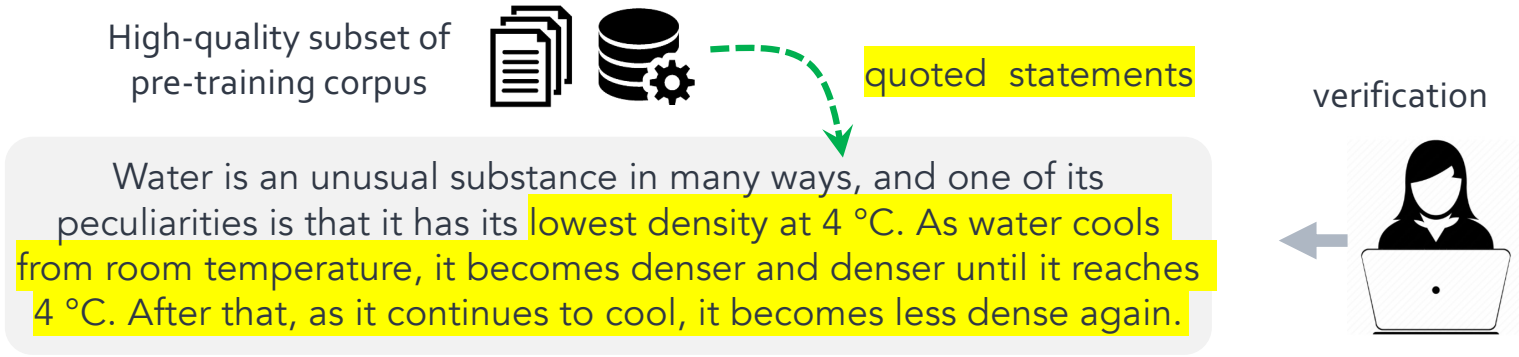
Liu et al. Evaluating Verifiability in Generative Search Engines. In *Findings of EMNLP 2023*

Citation Precision (%; ↑)	Average Over All Queries
Bing Chat	89.5
NeevaAI	72.0
perplexity.ai	72.7
YouChat	63.6
Average	74.5

Retrieval-augmentation helps, but not guaranteed to be correct.

Verifying LLM outputs via "quoting"?

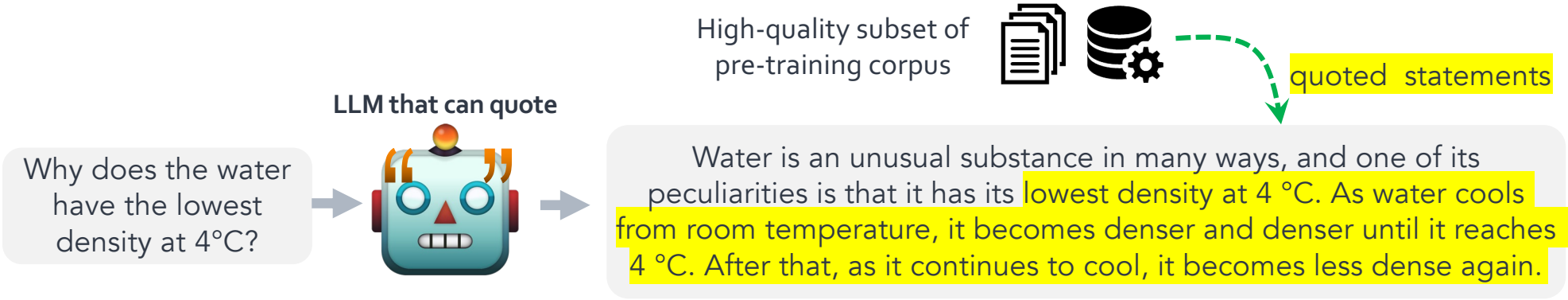
Hypothesis: Verbatim quotes from trusted sources make verifiability trivial.



- Users can focus on verifying the non-quoted portions.

Zhang et al. Verifiable by Design: Aligning Language Models to Quote from Pre-Training Data, 2024

Quote-tuning: LLMs with quoted responses



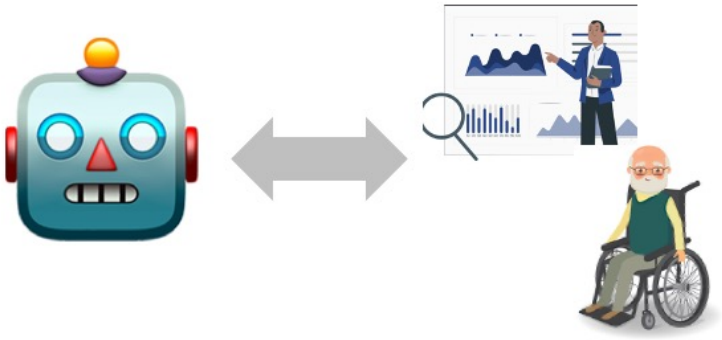
- We have introduced Quote-tuning, a pipeline for training LLMs to produce quoted responses from sources trusted by users.

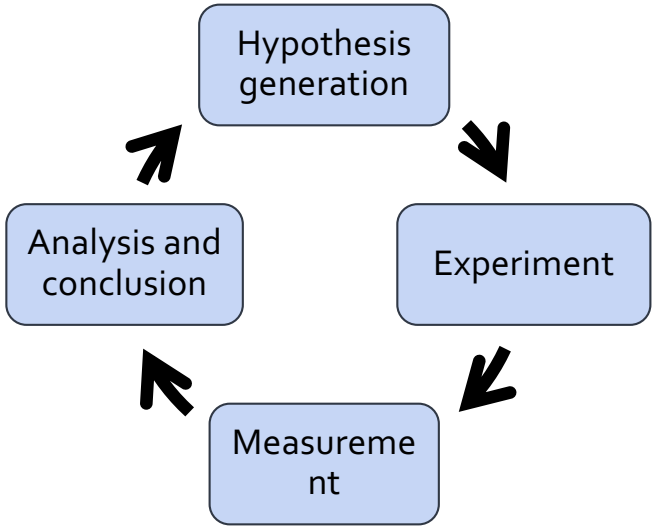
- Not a mature technology, but we are making fast progress on this.

Zhang et al. Verifiable by Design: Aligning Language Models to Quote from Pre-Training Data, 2024

Not AGI: Helpful, specialized applications

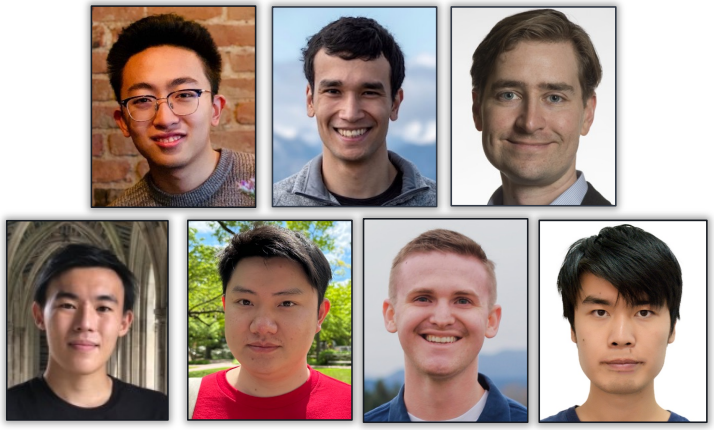
- Specialized models that are robust within well-defined domain, might be better alternatives to generalist brittle models.
- This will allow us to harness specialized feedback.
  - For example, LLMs as part of research cycle.
    - Growing LLMs as part of data ecosystem
    - Requires extensive safety considerations





Thanks for wonderful collaborators on these projects:

Johns Hopkins University



EleutherAI

University of Washington

Funding:

Ai2 Allen Institute for AI

amazon science

Office of Naval Research — Science & Technology