When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

Alex Mallen*, Akari Asai*,
Victor Zhong, Rajarshi Das,
Daniel Khashabi, Hannaneh Hajishirzi

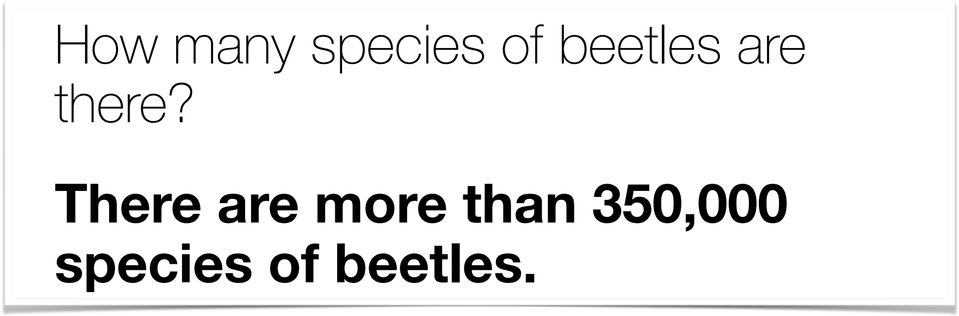
UNIVERSITY of WASHINGTON   Ai2   JOHNS HOPKINS UNIVERSITY

*Equal contribution

Factual knowledge

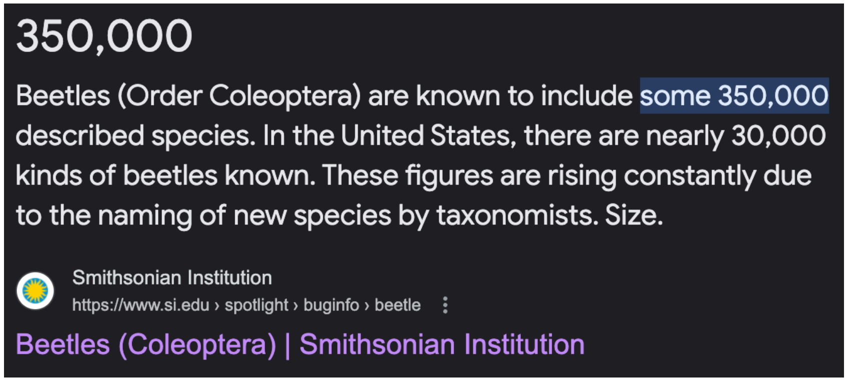
LM parameters encode a wealth of factual information

🤖 text-davinci-002



How many species of beetles are there?

There are more than 350,000 species of beetles.



350,000

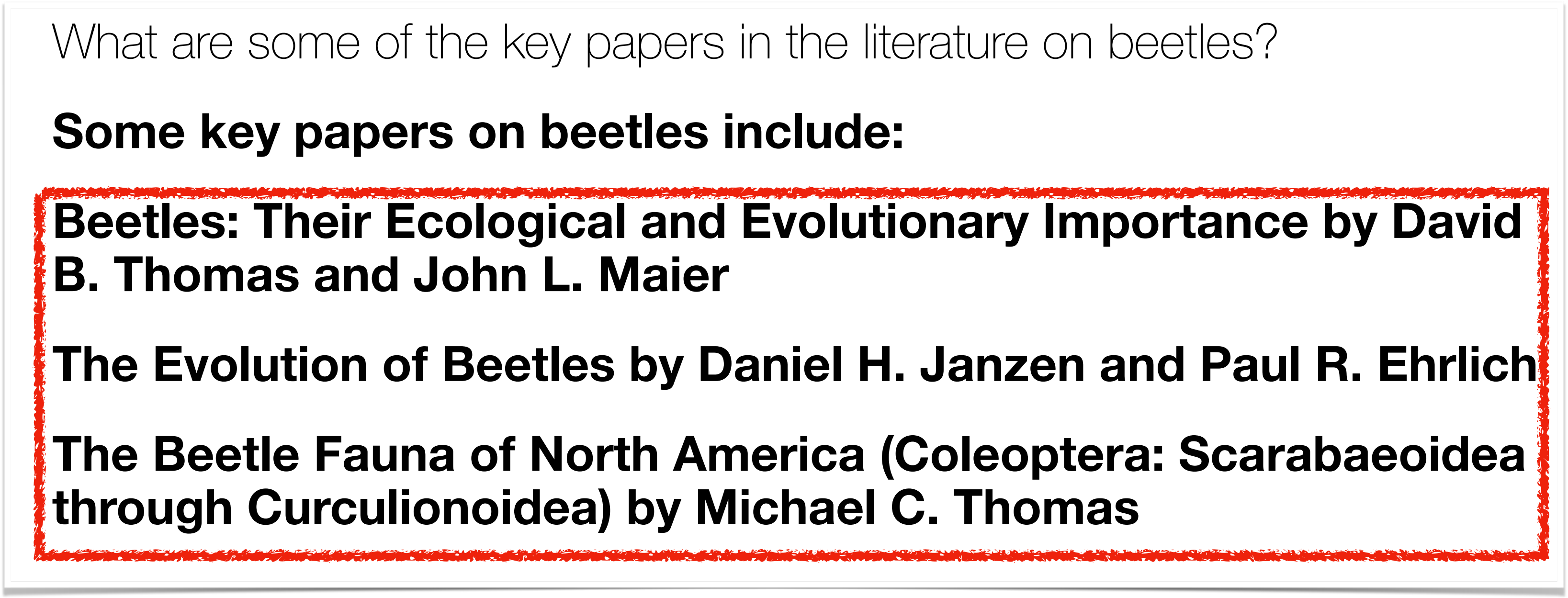
Beetles (Order Coleoptera) are known to include some 350,000 described species. In the United States, there are nearly 30,000 kinds of beetles known. These figures are rising constantly due to the naming of new species by taxonomists. Size.

Smithsonian Institution
https://www.si.edu › spotlight › buginfo › beetle

Beetles (Coleoptera) | Smithsonian Institution

Hallucinations

Their outputs often include **factual errors.**



🤖 text-davinci-002

What are some of the key papers in the literature on beetles?

Some key papers on beetles include:

Beetles: Their Ecological and Evolutionary Importance by David B. Thomas and John L. Maier

The Evolution of Beetles by Daniel H. Janzen and Paul R. Ehrlich

The Beetle Fauna of North America (Coleoptera: Scarabaeoidea through Curculionoidea) by Michael C. Thomas

Parametric and Non-parametric memories

Parametric: Information encoded in NN parameters

Non-Parametric: Usually information stored in a textual corpus (e.g. Wikipedia)

Key questions studied in our analysis

1. Effectiveness and limitations of **parametric** memories
    - When do LMs hallucinate?
    - Does scaling help?
2. Effectiveness and limitations of **non-parametric** memories
3. How do they **complement** each other?

Effectiveness and Limitations of **Parametric** Memories

When do LMs hallucinate?

1. Prior/Concurrent Work

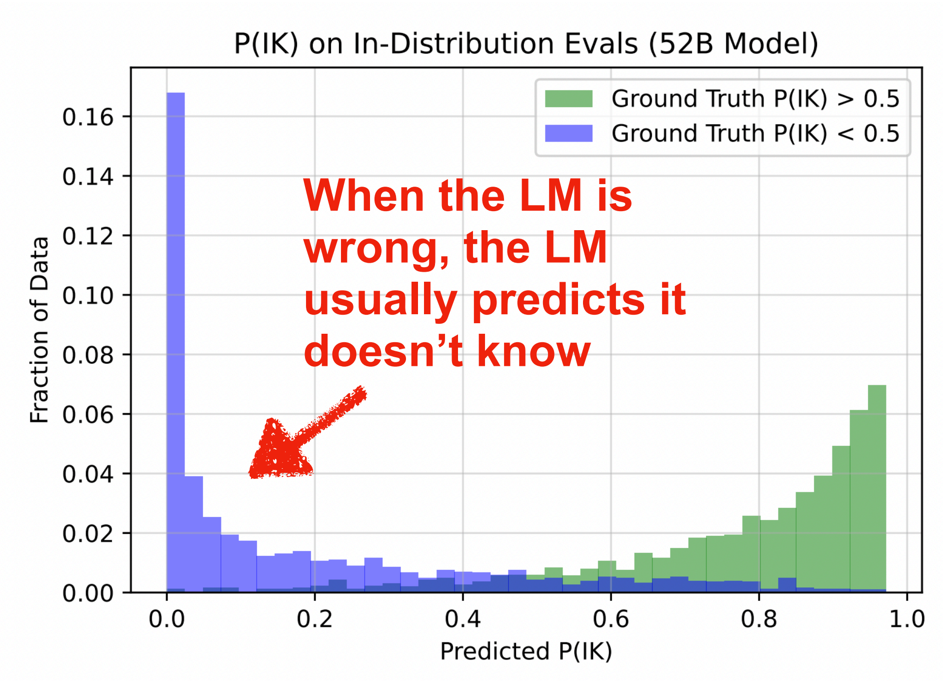
2. Our Approach

When do LMs hallucinate?

1. Prior/Concurrent Work

2. Our Approach

Predicting Failures of Factual Knowledge

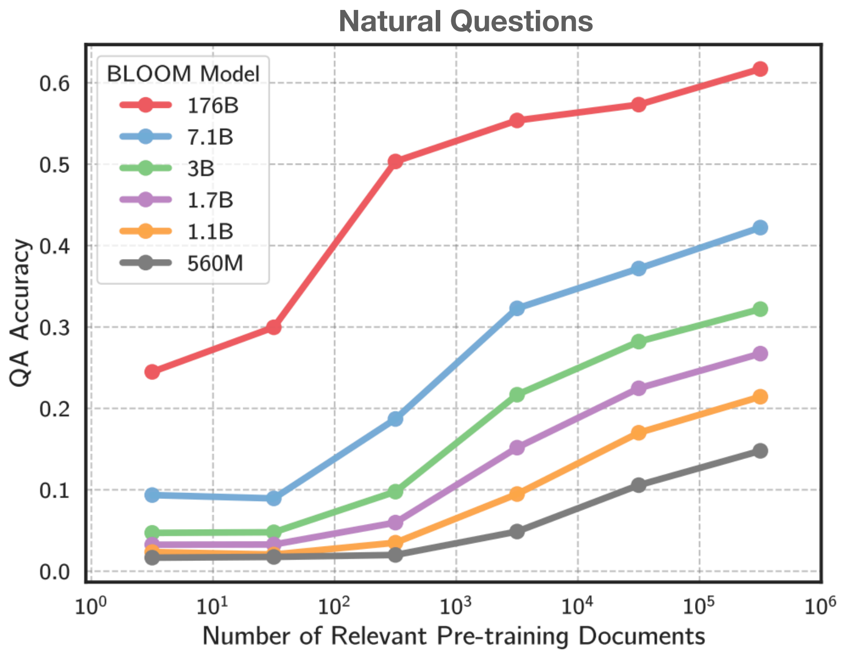


When the LM is wrong, the LM usually predicts it doesn't know

Saurav Kadavath et al., “Language Models (Mostly) Know What They Know” (2022)

Predicting Failures of Factual Knowledge

For each model, QA accuracy increases with the number of relevant pre-training documents



Nikhil Kandpal et al., “Large Language Models Struggle to Learn Long-Tail Knowledge” (2022)

When do LMs hallucinate?

1. Prior/Concurrent Work

2. Our Approach

New dataset: PopQA

Focus: Factual knowledge

WIKIDATA

(Kathy Saltzman, occupation, Politician)

Subject Relationship Object

Task: Open-domain QA

Q: What is the occupation of Kathy Saltzman?
A: politician

Hypothesis: Popularity predicts memorization

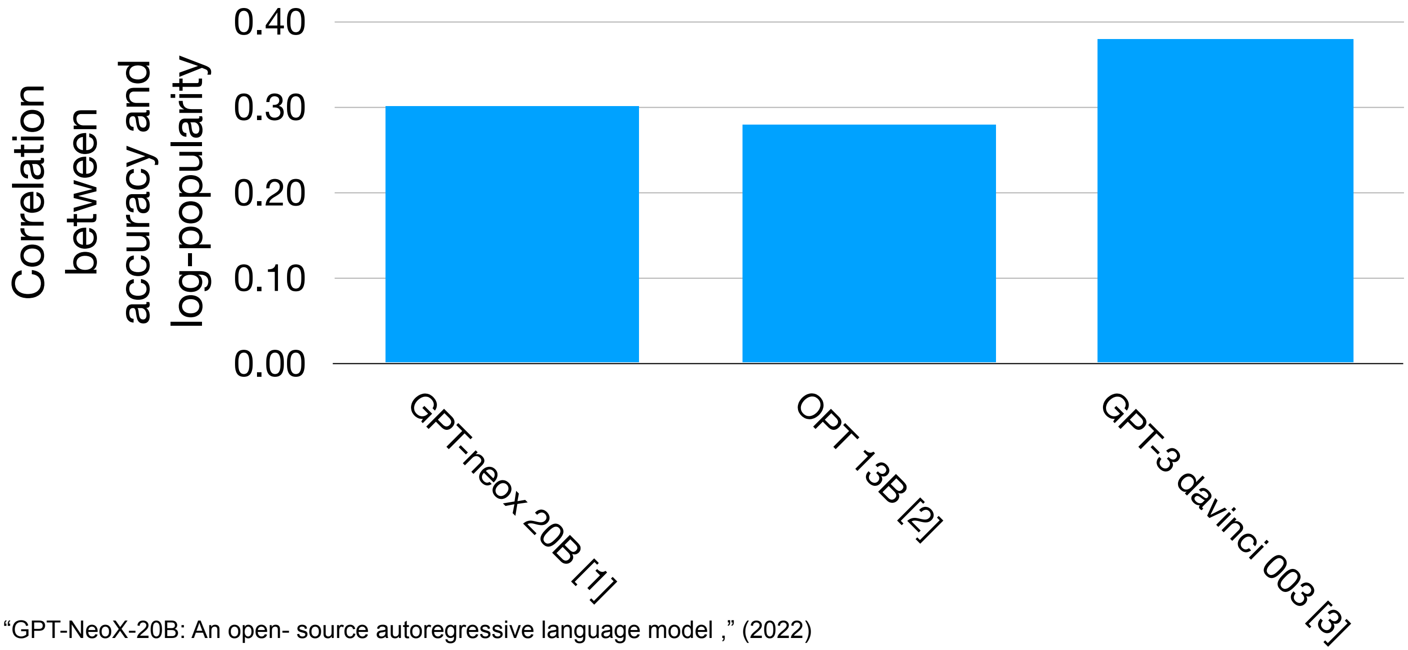
Pop = monthly Wikipedia page views

Pop(Kathy Saltzman) < Pop(Barack Obama)

⇨ Acc_{LM}(**Kathy Saltzman**, occupation, Politician)
< Acc_{LM}(**Barack Obama**, occupation, Politician)

Does popularity predict factual knowledge?

Factual accuracy is positively correlated with popularity across LMs



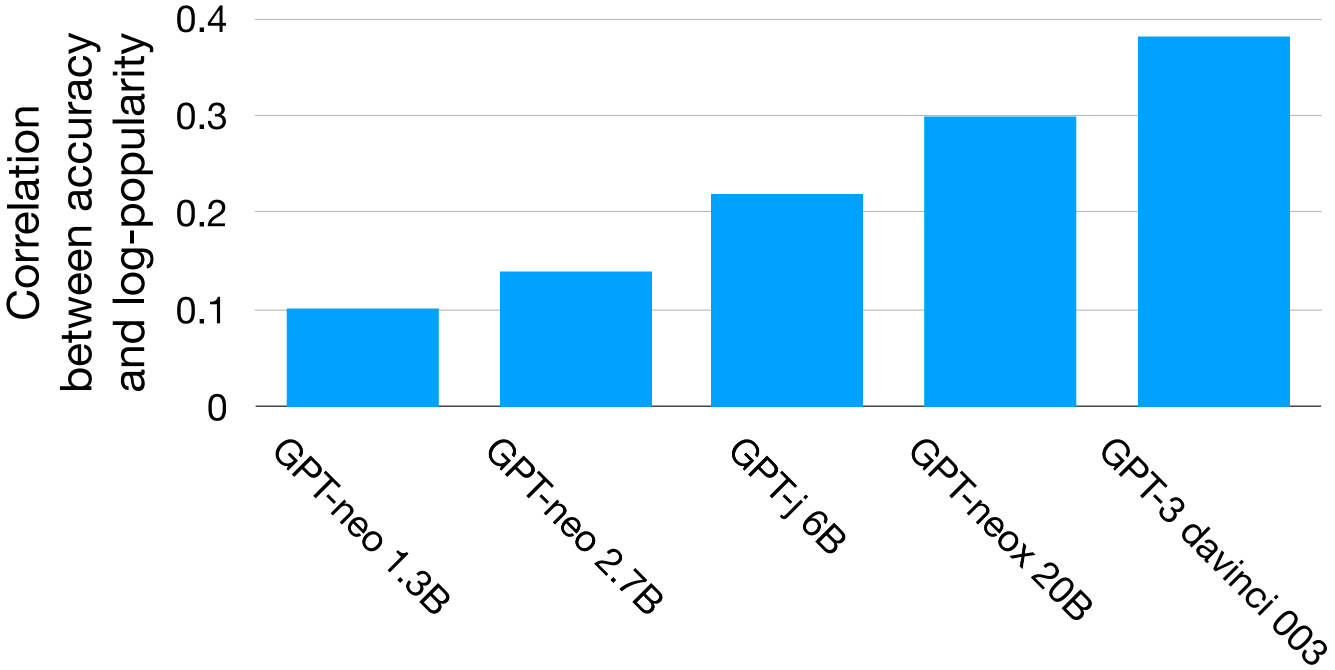
[1] Sidney Black et al., “GPT-NeoX-20B: An open- source autoregressive language model ,” (2022)

[2] Susan Zhang et al., ”Opt: Open pre-trained transformer language models.” (2022)

[3] Tom Brown et al., “Language models are few-shot learners.” (2020)

Does popularity predict factual knowledge?

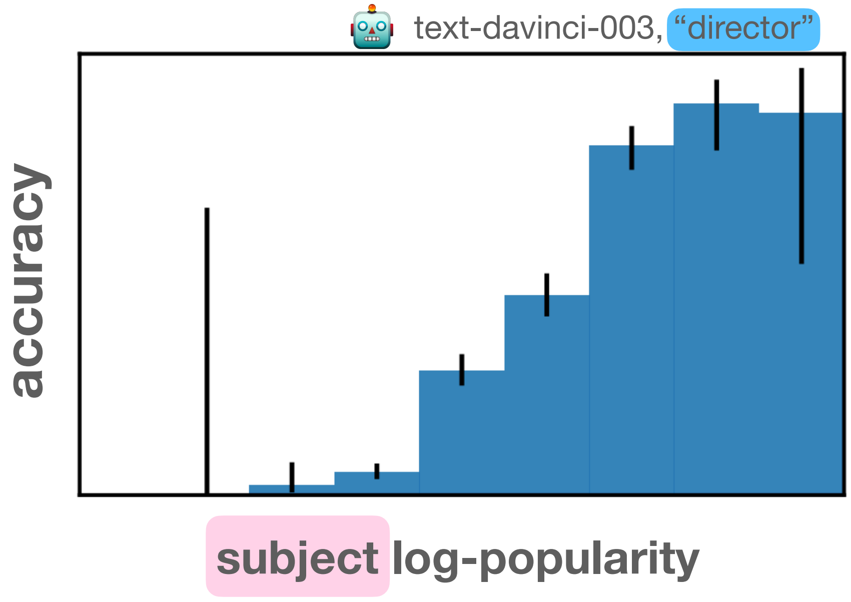
Larger LMs show a greater correlation



Does popularity predict factual knowledge?

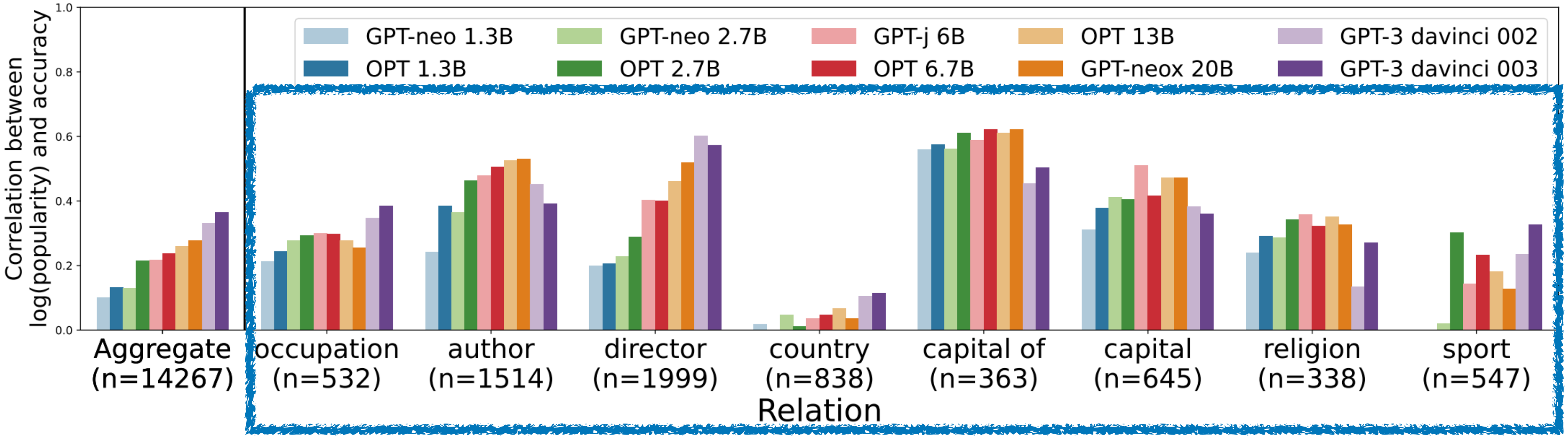
For each relationship type…

Q: Who was the director of The Titanic?
A:



Does popularity predict factual knowledge?

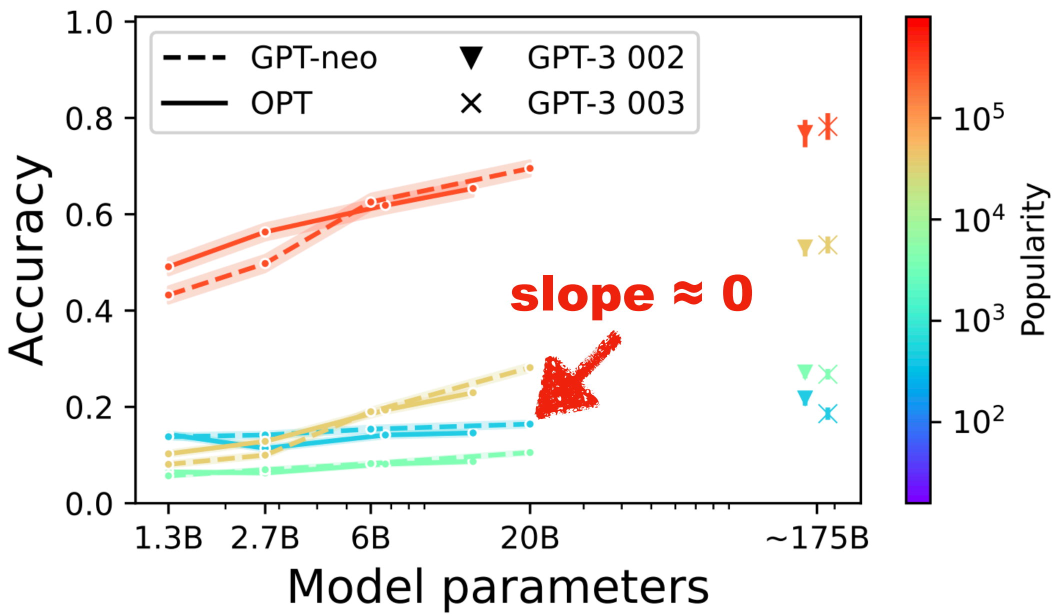
Factual accuracy is positively correlated with popularity across relationship types



Won’t scaling solve LMs’ factual unreliability?

Effects of Scaling

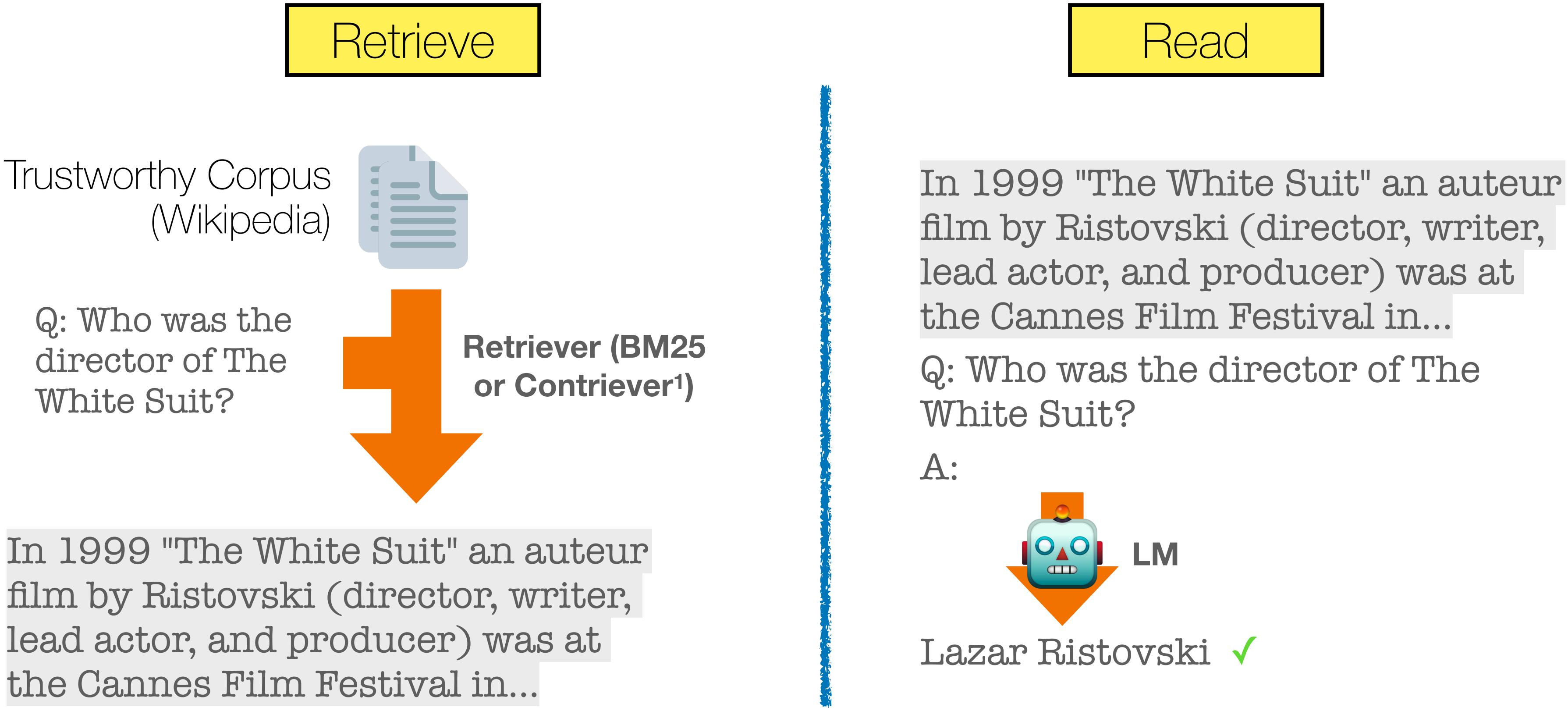
Even the largest LMs barely outperform the smallest LMs for tail questions



Effectiveness and Limitations of ***Non*-Parametric** Memories

How can retrieval help?

Non-parametric memory: Retrieve-and-read



Retrieve

Trustworthy Corpus (Wikipedia)

Q: Who was the director of The White Suit?

Retriever (BM25 or Contriever[1])

In 1999 "The White Suit" an auteur film by Ristovski (director, writer, lead actor, and producer) was at the Cannes Film Festival in…

Read

In 1999 "The White Suit" an auteur film by Ristovski (director, writer, lead actor, and producer) was at the Cannes Film Festival in…

Q: Who was the director of The White Suit?

A:

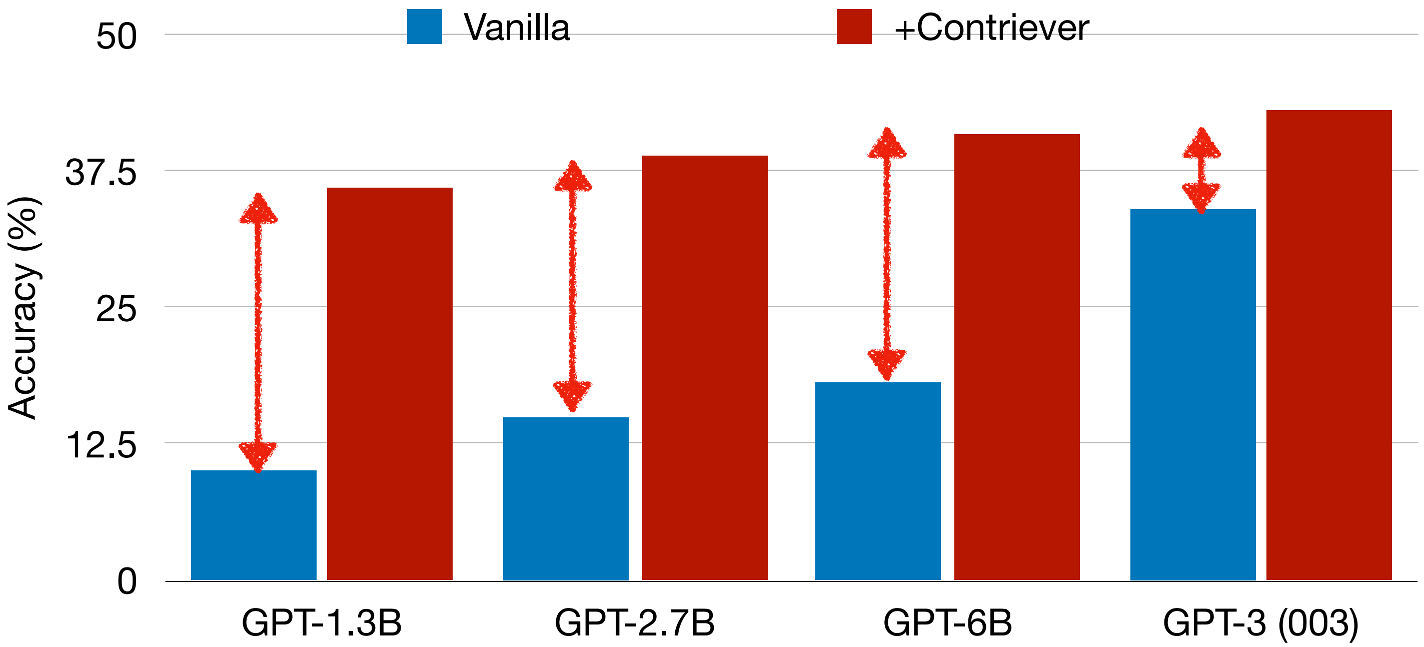
LM

Lazar Ristovski ✓

[1] Gautier Izacard et al., “Unsupervised Dense Information Retrieval with Contrastive Learning” (2022)

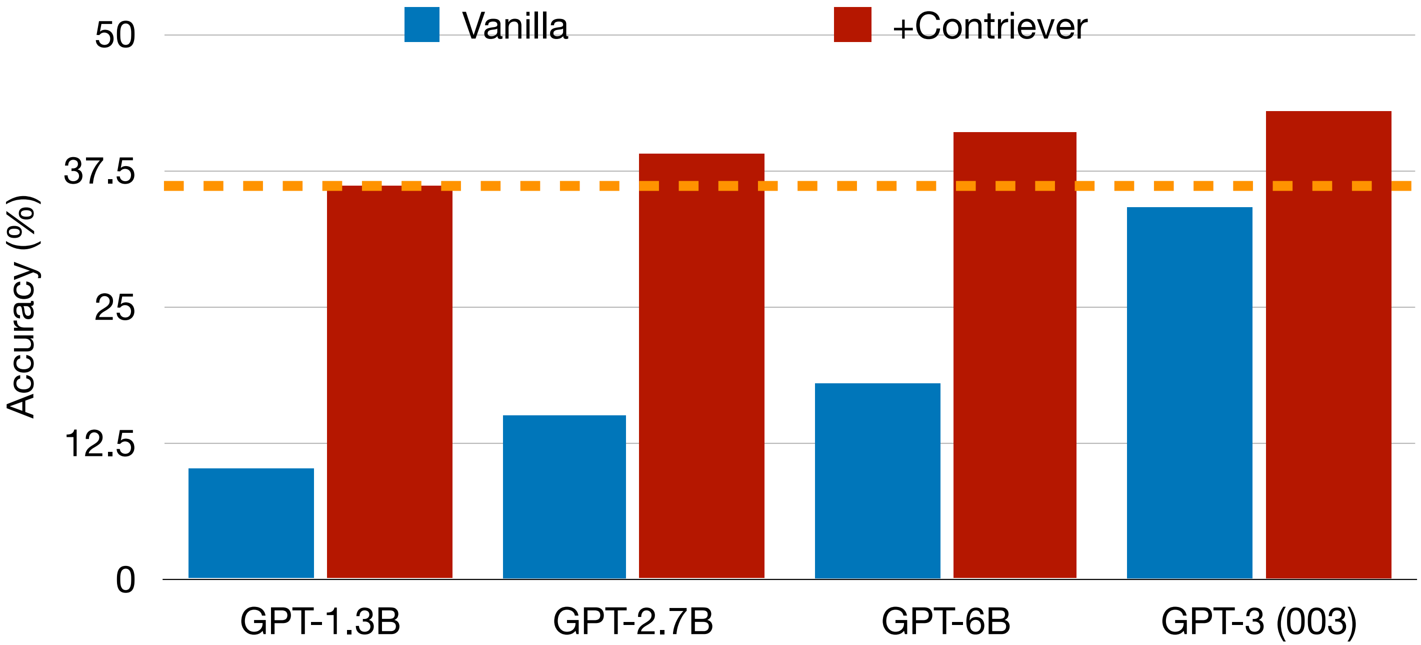
Results of retrieval-augmented LM

Non-parametric retrievers cause large improvements



Results of retrieval-augmented LM

1.3B GPT+retriever is better than GPT-3 (003; 175B?)

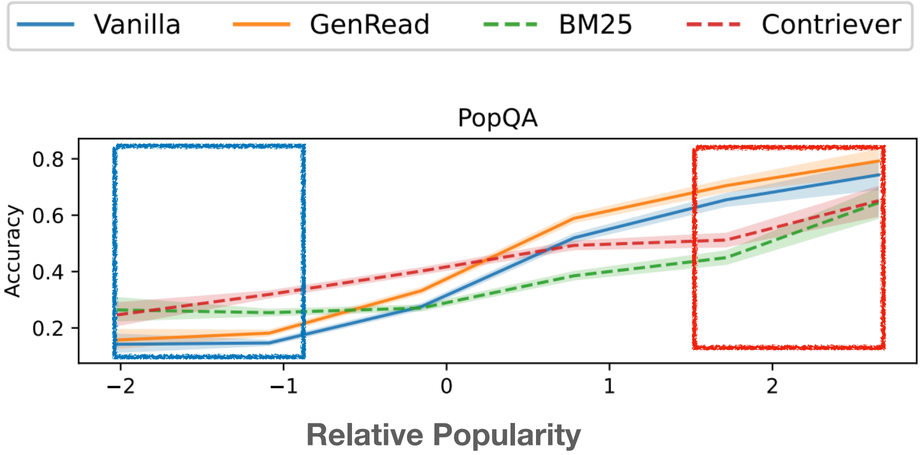


Retrieval in the tail?

When does retrieval help?

Non-parametric knowledge *helps significantly* for **less popular entities**

Non-parametric knowledge is often *often unhelpful* for **more popular entities**



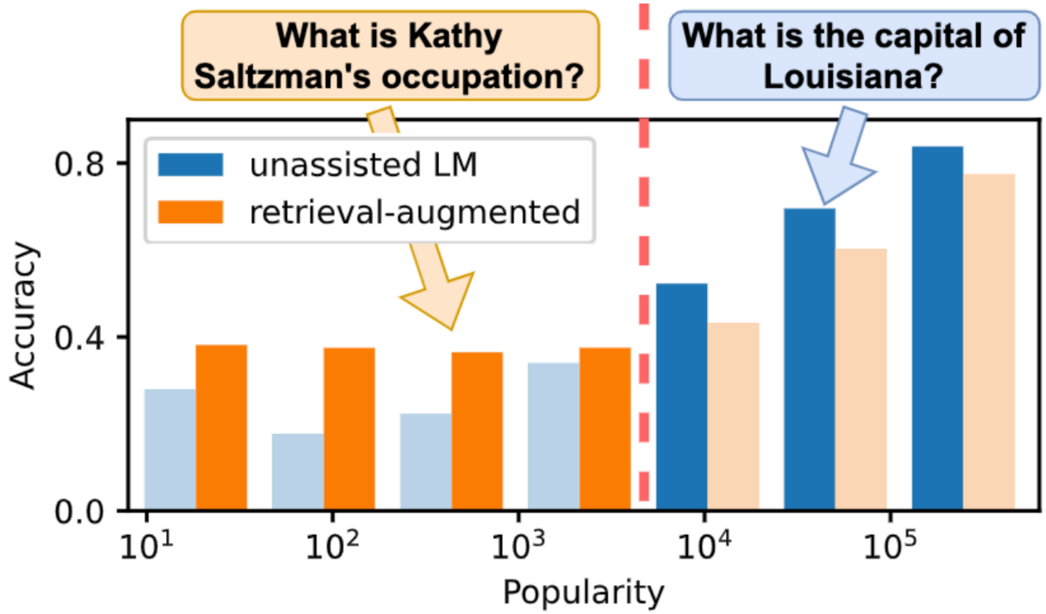
How do they **complement** each other?

Complementarity of parametric and non-parametric memories

Retrieval is…

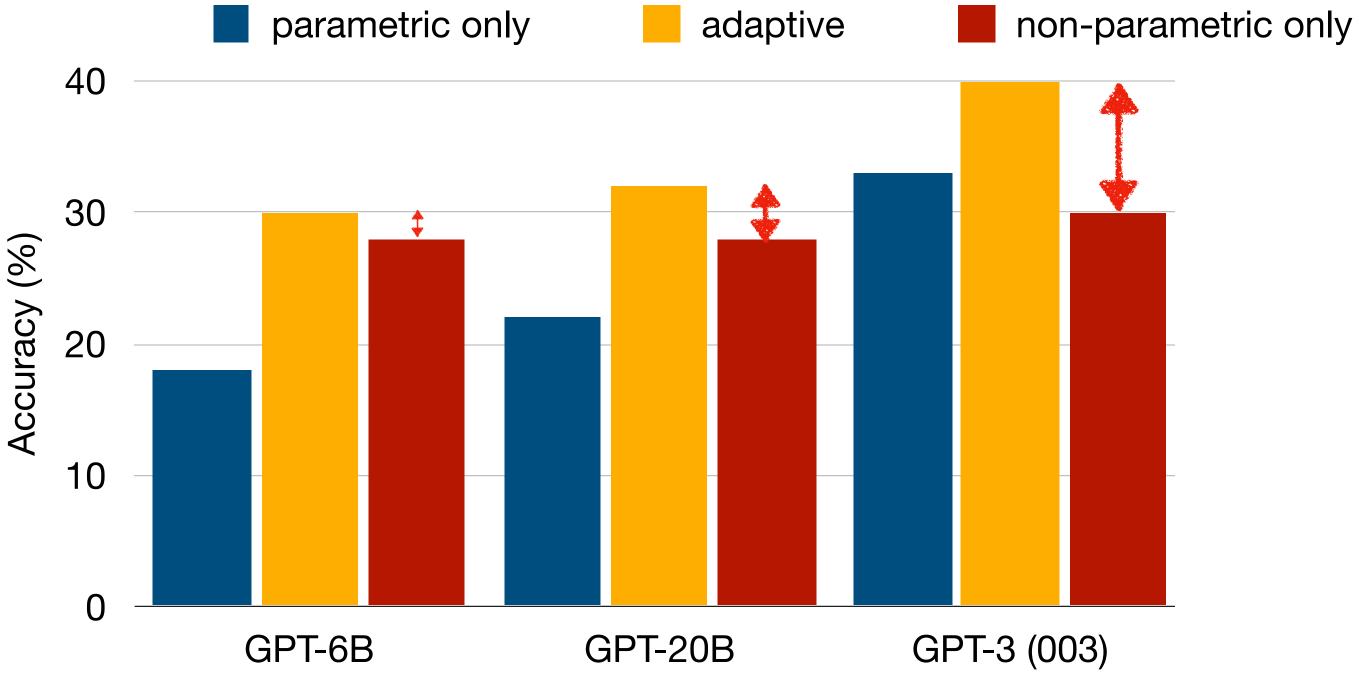
- **especially helpful** in the tail
- **often unhelpful** for popular knowledge

Adaptive Retrieval



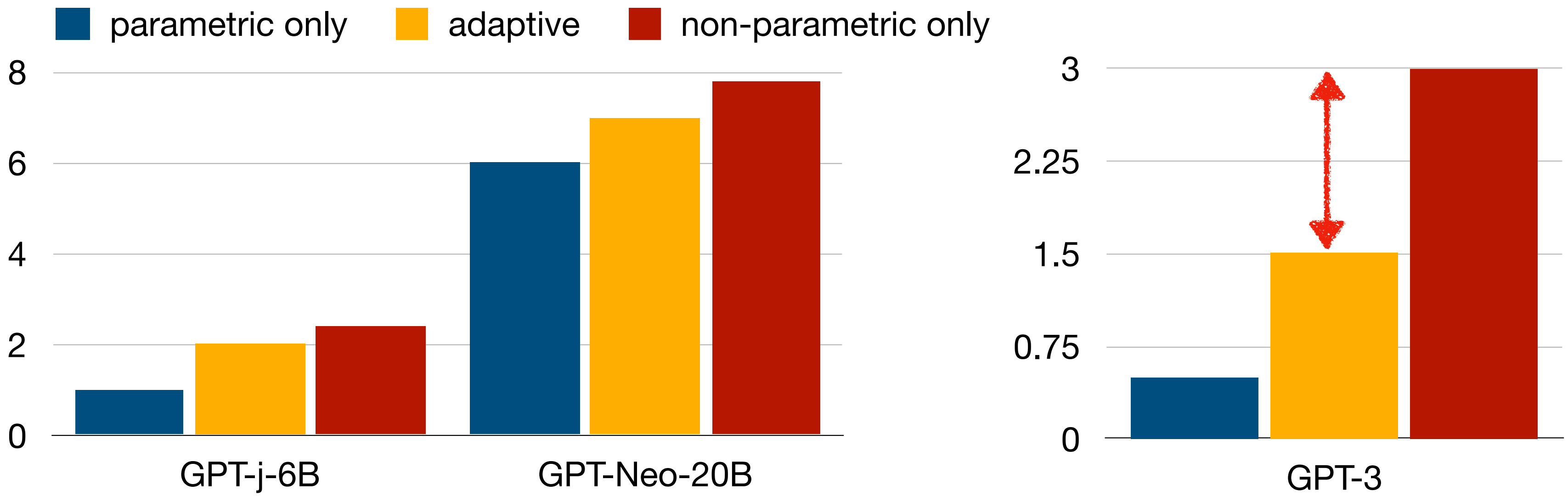
Adaptive retrieval for performance

Adaptive Retrieval improves performance across LMs



Adaptive retrieval for efficiency

Adaptive Retrieval improves efficiency across LMs



Latency (sec) / query

API cost ($) / 1k queries

Summary

- Retrieval complements LM parametric memory:
  - **Retrieval** is especially helpful in the tail
  - **LM parametric memory** is more reliable for popular knowledge
- Scaling is relatively ineffective in the tail
- Adaptive retrieval improves reliability and efficiency

Links ▶

Paper: https://arxiv.org/abs/2212.10511

Code & Data: https://github.com/AlexTMallen/adaptive-retrieval

Contact:
Alex Mallen / atmallen@cs.washington.edu / 🐦 @AlexTMallen
Akari Asai / akari@cs.washington.edu / 🐦 @AkariAsai