

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**

**CNeRG Reading Group Presentation
(18th January, 2024)**

**Paramita Das
Kiran Purohit**

Factual Knowledge

LM Parameters encode a wealth of factual information

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.

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?

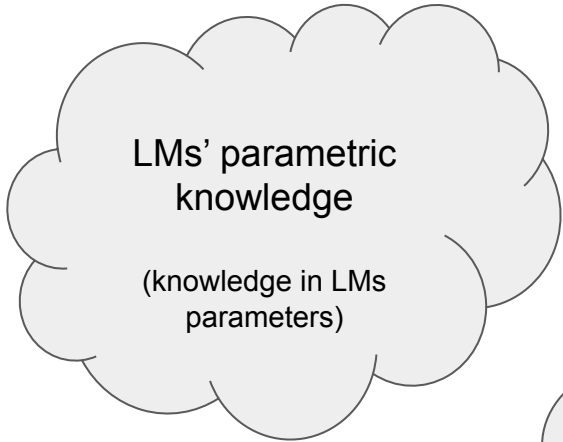
Some key papers on beetles include:

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

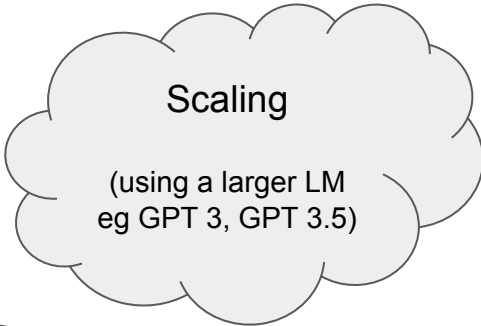
The Beetle Fauna of North America (Coleoptera: Scarabaeoidea through Curculionoidea) by Michael C. Thomas

When we should and should not rely on:



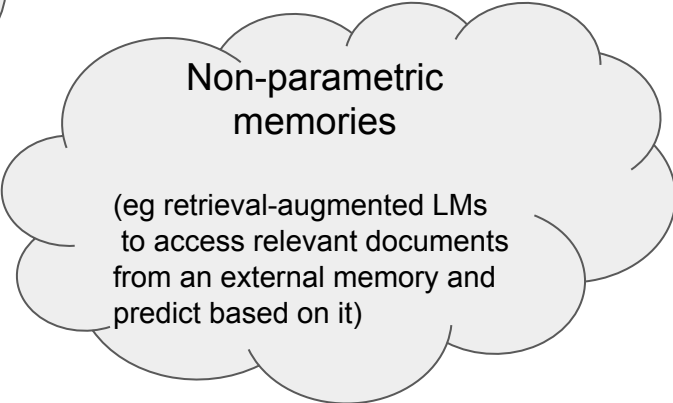
LMs' parametric
knowledge

(knowledge in LMs
parameters)



Scaling

(using a larger LM
eg GPT 3, GPT 3.5)



Non-parametric
memories

(eg retrieval-augmented LMs
to access relevant documents
from an external memory and
predict based on it)

Research Questions



(RQ₁) How much factual knowledge is memorized by LMs and what factors affect the memorization ?

Research Questions



(RQ₁) How much factual knowledge is memorized by LMs and what factors affect the memorization ?

(RQ₂) To what extent can non-parametric memories alleviate the shortcomings of parametric memories of LMs?

Research Questions



(RQ₁) How much factual knowledge is memorized by LMs and what factors affect the memorization ?

(RQ₂) To what extent can non-parametric memories alleviate the shortcomings of parametric memories of LMs?

(RQ₃) Can we build a system to adaptively combine non-parametric and parametric memories?

Task: Open Domain QA

(model predict the answer without any pre-given ground truth paragraph)

Metrics: Accuracy

(Prediction is correct if any substring of the prediction is an exact match of any of the gold answers.)

Existing Dataset's Limitations

- Natural Questions (NQ)
- EntityQuestions
- PopQA (*proposed*)

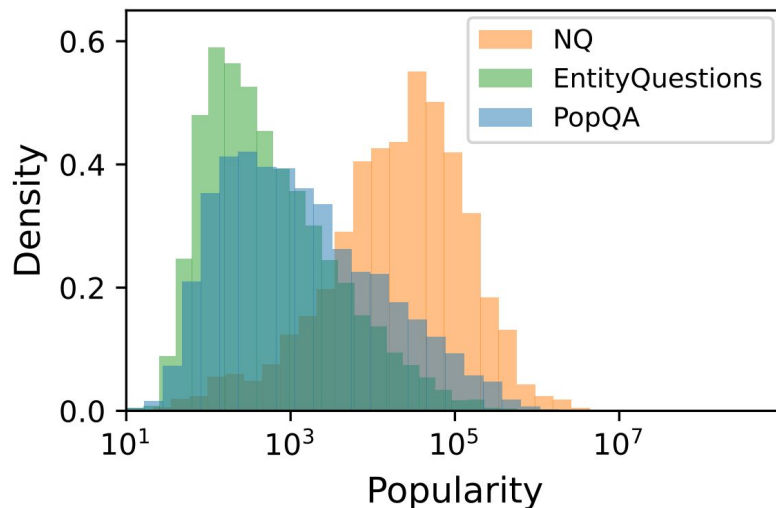


Figure : Distribution of subject entity popularity for EntityQuestions, POPQA, and for NQ-open for reference.

PopQA

Sampling factual
Knowledge from
wikipedia

(Kathy Saltzman, occupation, Politician)

Subject Relationship Object

Converting triples
to questions using
a template

“Q: <question> A:”

Q: What is the occupation of
Kathy Saltzman?

A: politician

Collect popularity

Wikipedia pageview



PopQA

- They randomly sample **16 diverse relationship types** from wikipedia

Relationship	Template
occupation	What is [subj] 's occupation?
place of birth	In what city was [subj] born?
genre	What genre is [subj]?
father	Who is the father of [subj] ?
country	In what country is [subj] ?
producer	Who was the producer of [subj] ?
director	Who was the director of [subj] ?
capital of	What is [subj] the capital of?
screenwriter	Who was the screenwriter for [subj] ?
composer	Who was the composer of [subj] ?
color	What color is [subj] ?
religion	What is the religion of [subj] ?
sport	What sport does [subj] play?
author	Who is the author of [subj] ?
mother	Who is the mother of [subj] ?
capital	What is the capital of [subj] ?

Table . Full list of the manually annotated templated used for POPQAcreations. [subj] denotes a placeholder for subject entities.

Hypothesis: Popularity predicts memorization

Pop = monthly Wikipedia page views

$$\text{Pop}(\text{Kathy Saltzman}) < \text{Pop}(\text{Barack Obama})$$

Hypothesis: Popularity predicts memorization

Pop = monthly Wikipedia page views

Pop(Kathy Saltzman) < Pop(Barack Obama)

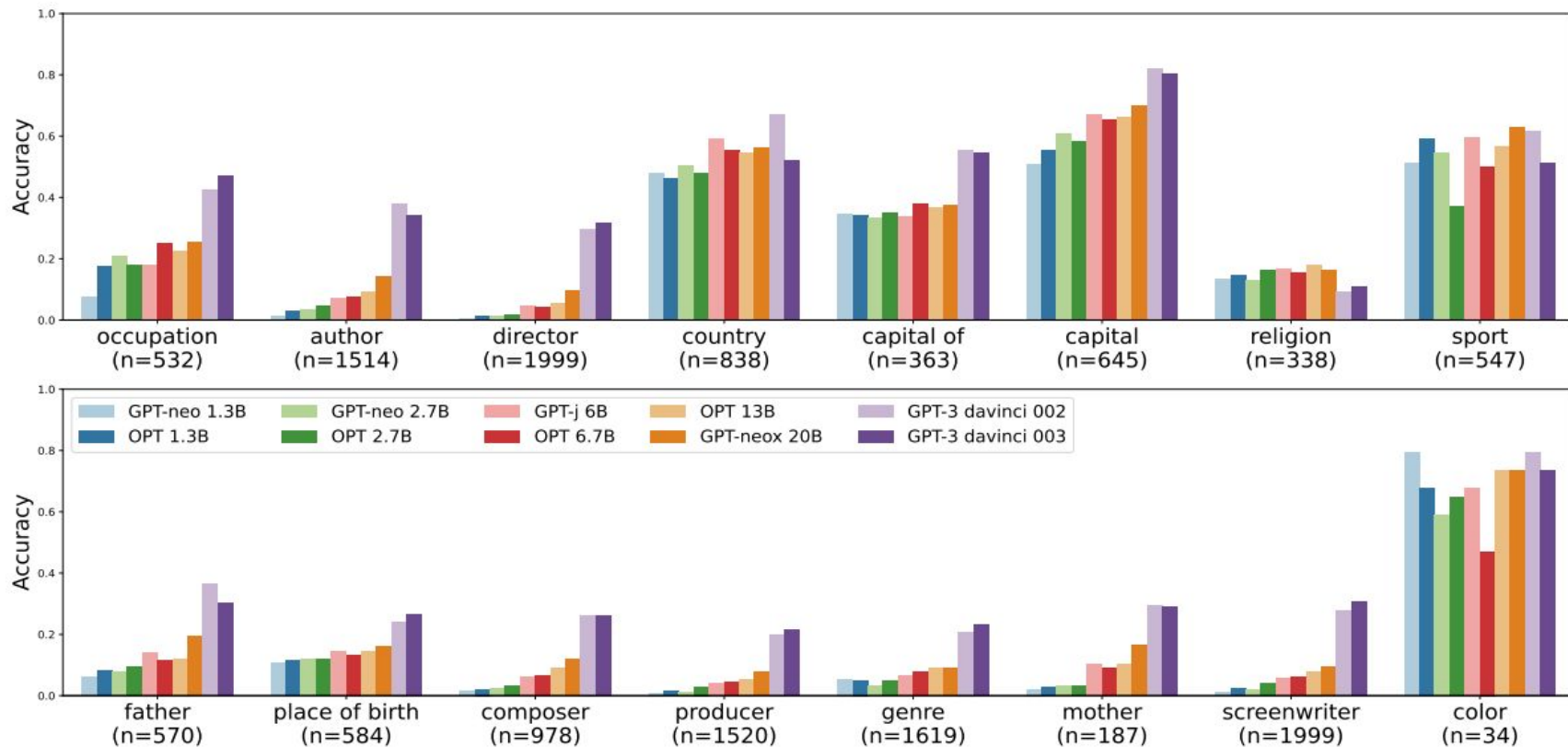
⇒ Acc_{LM}(Kathy Saltzman, occupation, Politician)
< Acc_{LM}(Barack Obama, occupation, Politician)

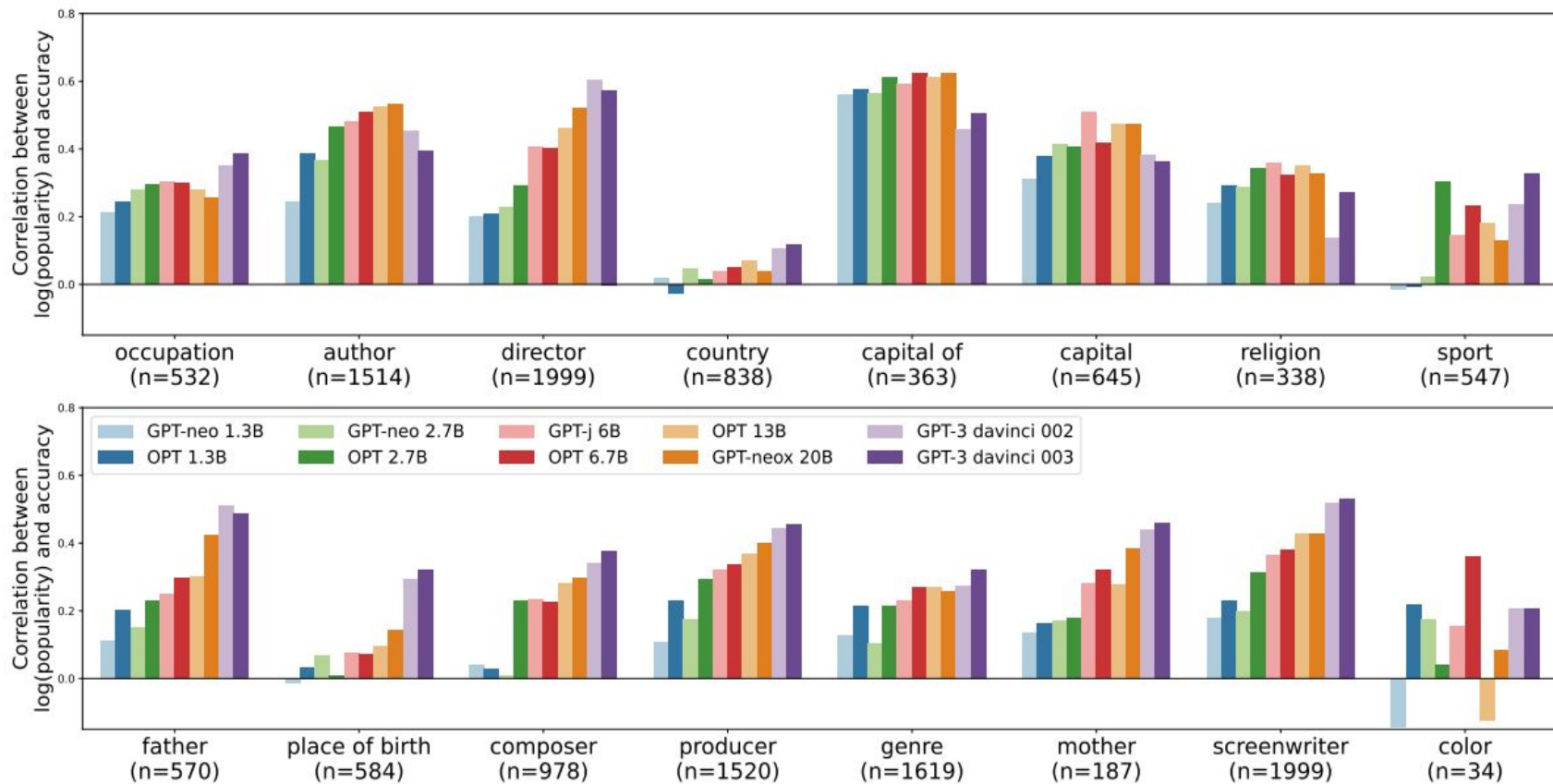
RQ1: Memorization depends on Popularity

Experimental Setup:

- Models used: OPT (1.3, 2.7, 6.7, and 13 billion), GPT-Neo (1.3, 2.7, 6, and 20 billion), GPT-3 (davinci-002, davinci-003).
- Template used: Q: <Question> A:
- Prompts used: zero-shot prompting for GPT-3, 15-shot prompting for all GPT-neo and OPT models.
- GPT3 → 35% accuracy, GPT-Neo 20B → 25% accuracy.
- Large LMs memorize factual knowledge in their parameters to some extent.
 - which types of knowledge are better memorized?
 - what factors influence memorization?

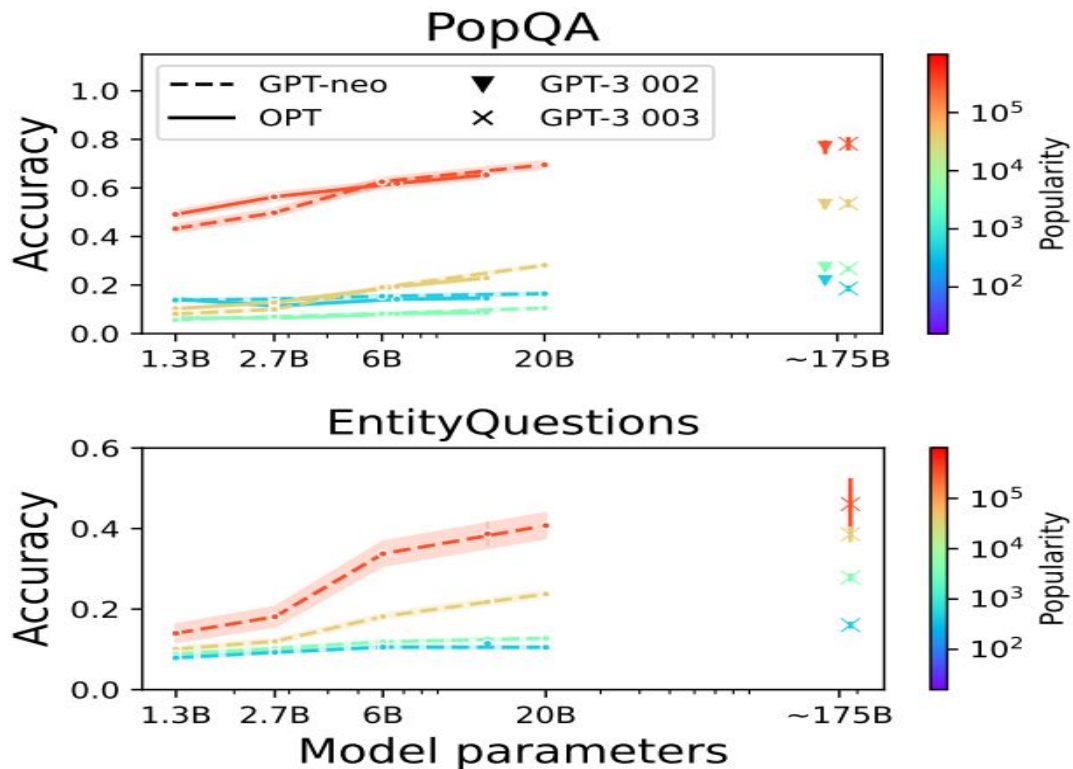
Subject entity popularity predicts memorization





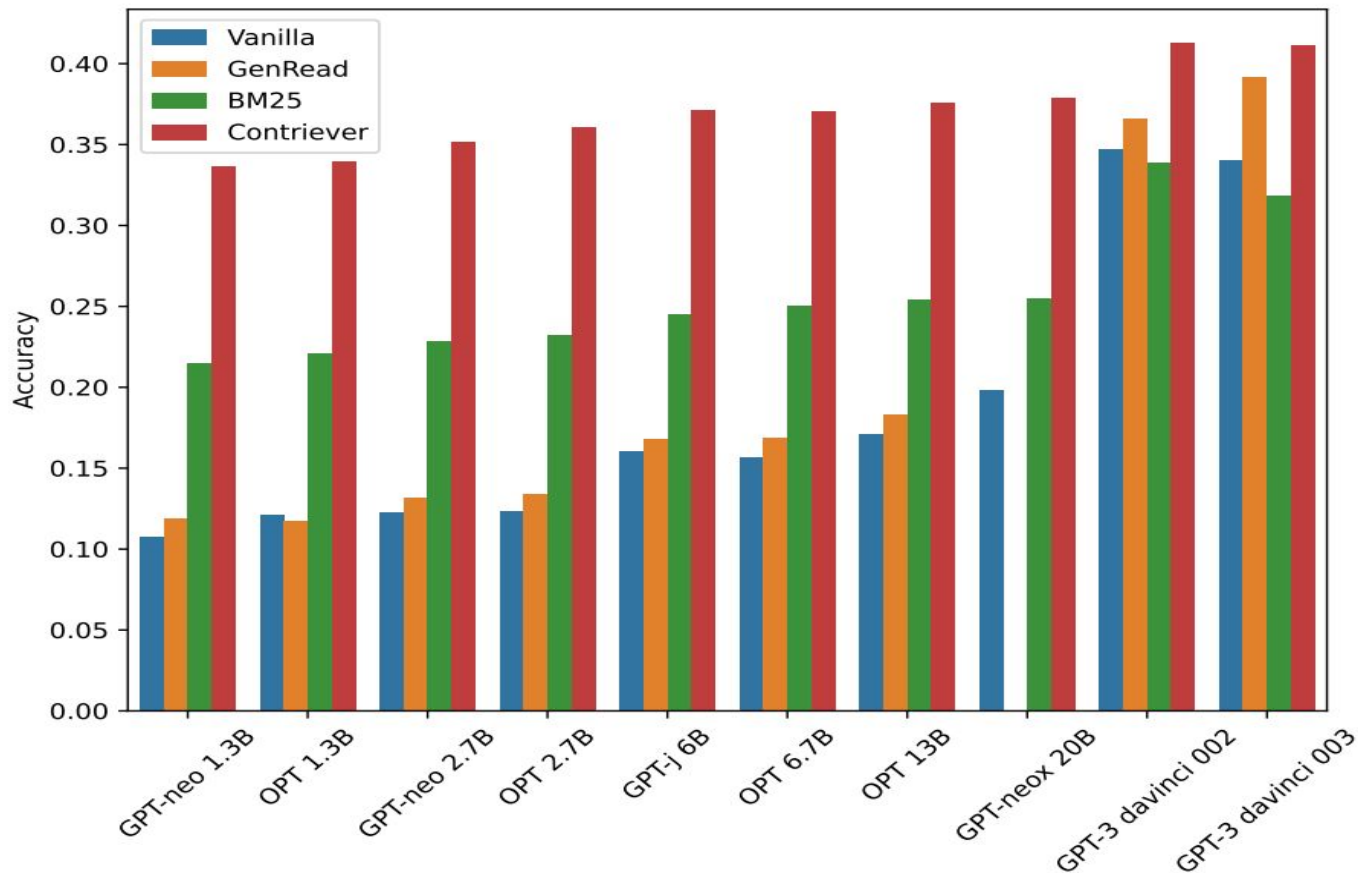
Correlations on PopQA for all relationship types and models.

Scaling may not help with tail knowledge

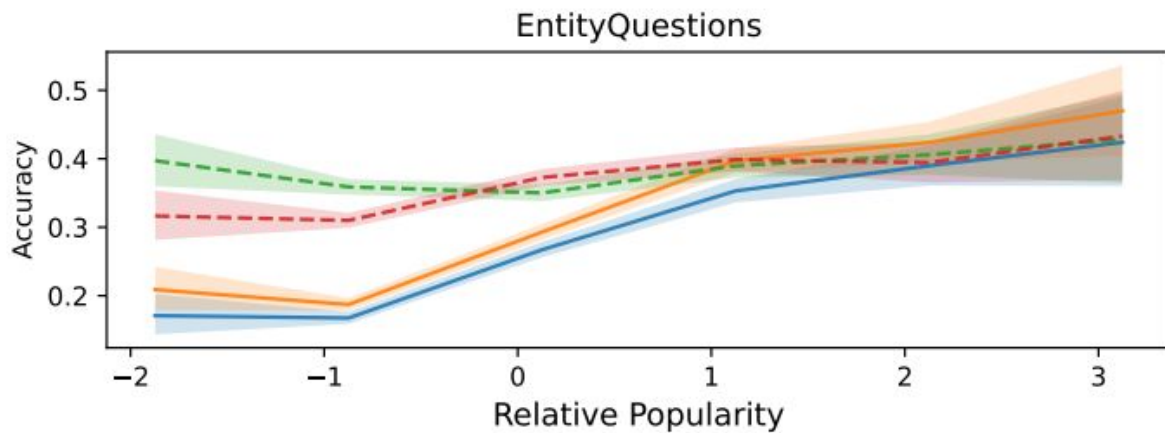
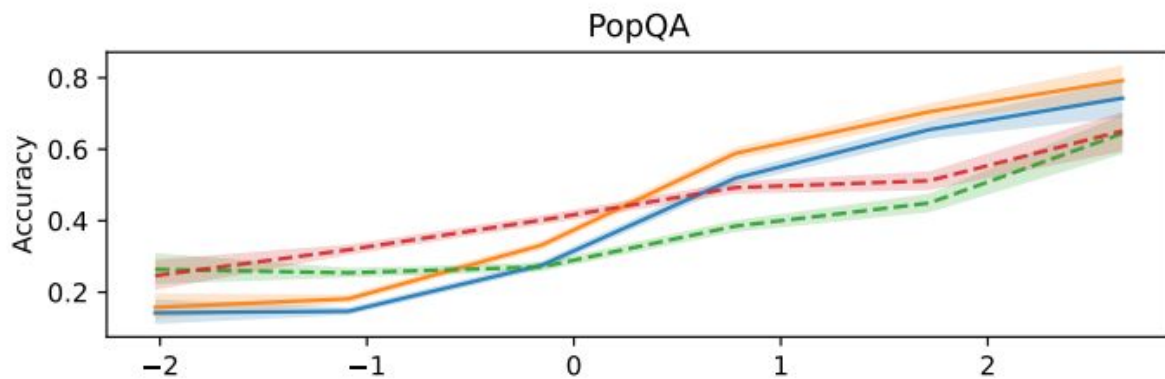


RQ2: Non-parametric memory complements parametric memory

- SOTA LMs exhibit low accuracy with --
 - Less popular subjects.
 - Certain relation types, e.g.-occupation, author, director etc.
 - Increasing model size does not help.
- Hypothesis:: non-parametric memories with the help of retrieval-augmented models can improve performance.
- Augmenting input- Additional context retrieved from Wikipedia (off-line) relevant to a question and then concatenated with the original question.
- Retrieval models- BM25, Contriever, GenRead.



PopQA accuracy of LMs augmented with BM25, Contriever, GenRead, and unassisted (vanilla). **Retrieving non-parametric memories significantly improves the performance of smaller models.**



GPT-3 davinci-003 accuracy versus relative popularity (how popular a question is relative to other questions of its relationship type).

Retrieval-augmented LMs (dashed) outperform LMs' parametric memory (solid) for less popular entities, while parametric memory is competitive for more popular entities.

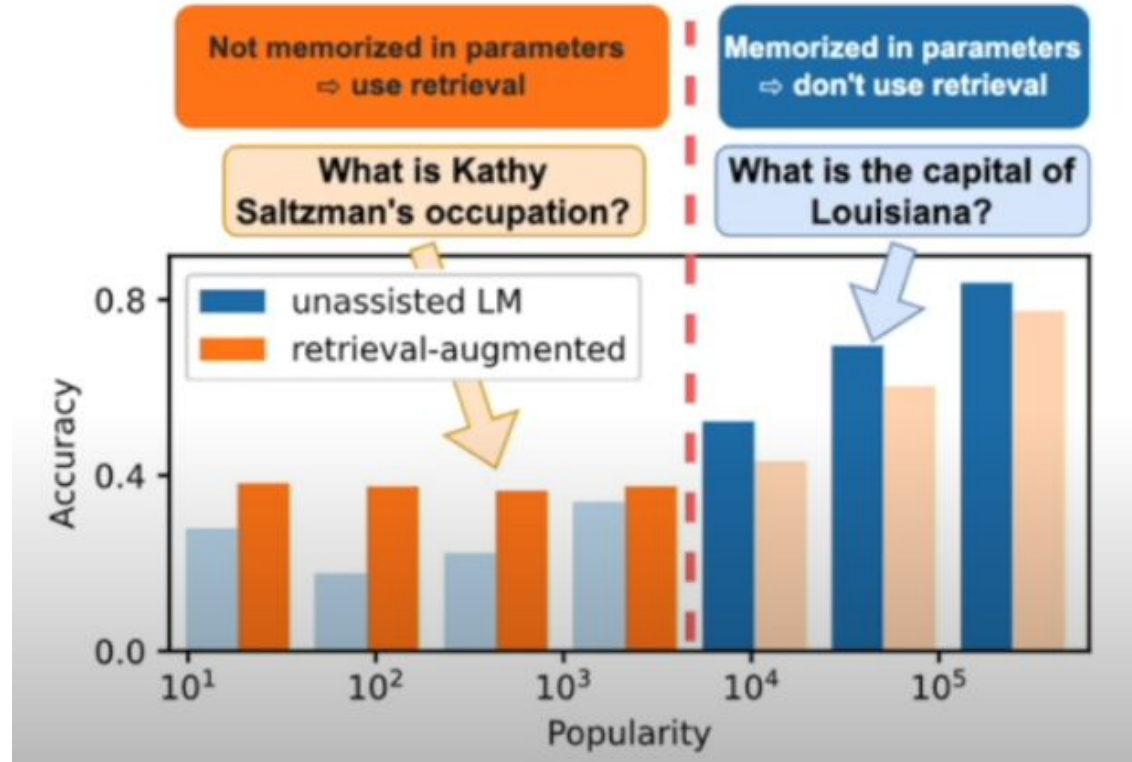
Non-parametric memories can mislead LMs.

	Contriever-augmented LM	
	succeeded	failed
LM succeeded	0.83 (24%)	0.14 (10%)
LM failed	0.88 (17%)	0.11 (49%)

The recall@1 of Contriever for questions that GPT-3 davinci-003 answered correctly and incorrectly with and without retrieval on POPQA. **For 10% of questions, retrieval is harmful due to low-quality retrieved text (0.14 recall@1).**

Question (answer)	LM	Retrieval-LM	Retrieval
Who was the director of Black? (Sanjay Leela Bhansali)	The director of Black is Sanjay Leela Bhansali. ✓	Noel Black ✗	Noel Black (June 30, 1937 – July 5, 2014) was an American film and television director, screenwriter, and producer. Black was born in Chicago, Illinois. He won awards at the 1966 ... and subsequently concentrated on directing for television, occasionally directing films such as "Private School".
Who was the producer of The Faculty? (Robert Rodriguez)	The Faculty was produced by Elizabeth Avellan and Robert Rodriguez. ✓	The producer of The Faculty was Elizabeth Avellan. ✗	The Faculty is a 1998 American science fiction teen horror film written by Kevin Williamson, directed by Robert Rodriguez , and starring Elijah Wood,... The film was released on Christmas Day, 1998, by Miramax Films through Dimension Films. ... The film has developed a cult following in the years since its release.
What is Michael Shelley's occupation? (singer-songwriter)	Michael Shelley is a singer-songwriter and musician. ✓	Michael Shelley is an American applied mathematician. ✗	Michael Shelley (mathematician) Michael J. Shelley (born August 17, 1959) is an American applied mathematician who works on the modeling and simulation of complex systems arising in physics and biology. This has included free-boundary problems in fluids and materials science... He is also the co-founder and co-director of the Courant Institute's Applied Mathematics Lab.
In what city was Zijah Sokolović born? (Sarajevo)	Zijah Sokolović was born in Sarajevo ✓	Zijah Sokolović was born in Orahovac, Kingdom ✗	Ali Sokol ... (born 8 May 1921 in Orahovac, Kingdom of Serbs, Croats and Slovenes, died 23 September 1974) was a Yugoslav pulmonologist . Ali Sokol was born into an agricultural family. He was the third of four children of father and mother Hatixhes Solomon. It is unknown the exact date of birth but the most reliable date is May 8 year in 1921.

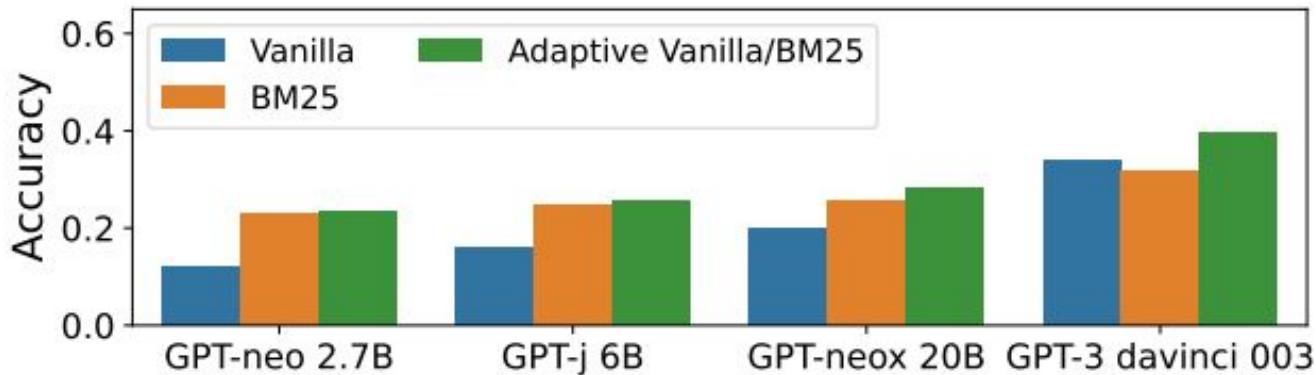
Question (answer)	LM	Retrieval-LM	Retrieval
What genre is Unknown? (fantasy)	Unknown is not a specific genre of music. It could refer to X	Unknown is a pulp fantasy fiction magazine. ✓	Unknown (magazine) Unknown (also known as Unknown Worlds) was an American pulp fantasy fiction magazine, published from 1939 to 1943 by Street & Smith, and edited by John W. Campbell. ... The leading fantasy magazine in the 1930s was "Weird Tales", which focused on shock and horror.
In what country is Pierre? (United States)	Pierre is in France. X	Pierre is from the United States. ✓	Jonas Farley Pierre (born May 30, 1980) is an American professional basketball player. A center, he has played with three teams in the National Basketball League of Canada (NBL). Pierre last played with the Saint John Mill Rats in Canada. ... He then joined Lincoln University in Missouri for his third year. In the summer of 2003, he signed a letter of intent to play with Nicholls State in the NCAA Division I.
Who was the producer of The Cocoanuts? (Walter Wanger)	The Cocoanuts was produced by Florenz Ziegfeld. X	The Cocoanuts was produced for Paramount Pictures by Walter Wanger, who ✓	The Cocoanuts is a 1929 musical comedy film starring the Marx Brothers. Produced for Paramount Pictures by Walter Wanger , who is not credited, the film stars the four Marx Brothers, Oscar Shaw, Mary Eaton, and Margaret Dumont. It was the first sound film to credit more than one director (Robert Florey and Joseph Santley), and was adapted to the screen by Morrie Ryskind from the George S. Kaufman Broadway musical play. ...
Who was the director of The White Suit? (Lazar Ristovski)	The White Suit was directed by Sachin Kundalkar. X	Lazar Ristovski ✓	In 1999 "The White Suit" an auteur film by Ristovski (director, writer, lead actor, and producer) was at the Cannes Film Festival in the Critics Week program. "The White Suit" was the Serbian entry for the 1999 Academy Awards. Lazar Ristovski is the sole owner of Zillion Film Company In 2006, he made a small appearance in the James Bond film "Casino Royale". He played Caruso in the 2004 movie "King of Thieves". He starred as Đorđe in the award-winning 2009 film "St. George Shoots the Dragon".



Retrieval is --

- Helpful in long-tail distributions.
- Often harmful for popular knowledge.

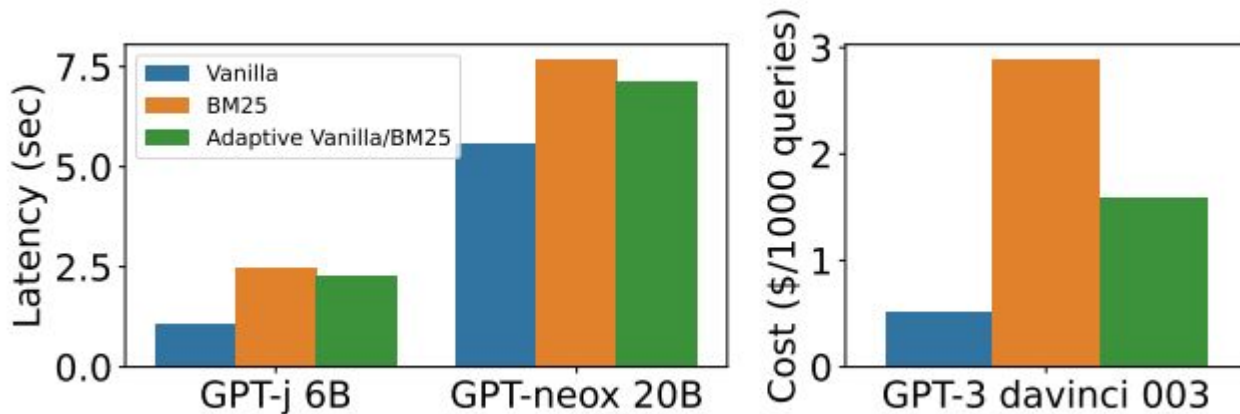
RQ3: Adaptive Retrieval



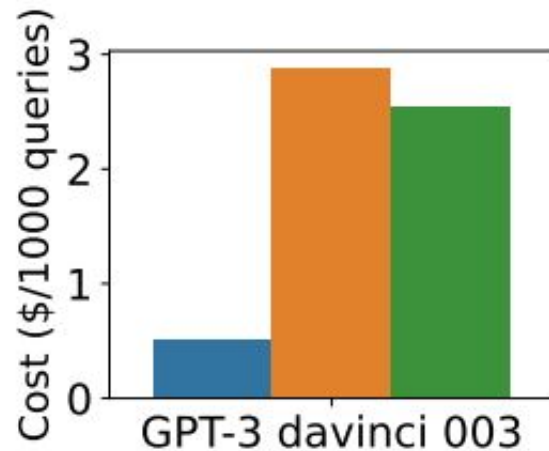
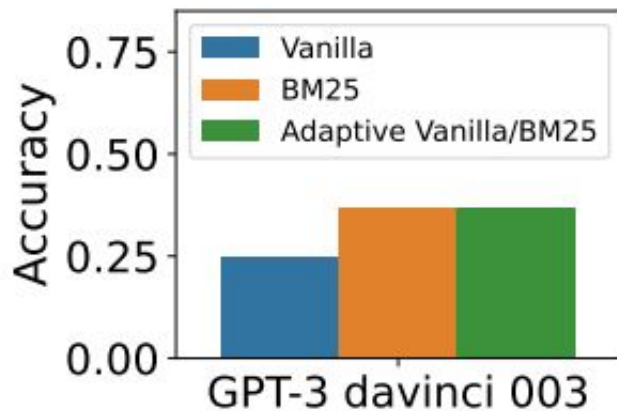
POPQA performance of GPT-neo models and GPT3 davinci-003, with different retrieval methods. **Adaptive Retrieval robustly outperforms approaches that always retrieve, especially for larger LMs.** Use retrieval for questions whose popularity is lower than a threshold.

Adaptive Retrieval reduces inference-time and costs

PopQA:



EntityQuestions:



Key Points

1. Memorization has a strong correlation with entity popularity and that scaling up models on long-tail distributions may only provide marginal improvements.
2. Non-parametric memories can greatly aid LMs on these long-tail distributions, but can also mislead LMs on questions about well-known entities, as powerful LMs have already memorized them in their parameters.
3. Adaptive Retrieval, which only retrieves when necessary, using a heuristic based on entity popularity and relationship types.

Thank You