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 [♠]Johns Hopkins University
[♡]Allen Institute for AI
{atmallen,akari,vzhong,rajarshi,hannaneh}@cs.washington.edu
danielk@jhu.edu

Abstract

Despite their impressive performance on diverse tasks, large language models (LMs) still struggle with tasks requiring rich world knowledge, implying the difficulty of encoding a wealth of world knowledge in their parameters. This paper aims to understand LMs' strengths and limitations in memorizing factual knowledge, by conducting large-scale knowledge probing experiments on two open-domain entity-centric QA datasets: PopQA, our new dataset with 14k questions about long-tail entities, and EntityQuestions, a widely used open-domain QA dataset. We find that LMs struggle with less popular factual knowledge, and that retrieval augmentation helps significantly in these cases. Scaling, on the other hand, mainly improves memorization of popular knowledge, and fails to appreciably improve memorization of factual knowledge in the long tail. Based on those findings, we devise a new method for retrieval augmentation that improves performance and reduces inference costs by only retrieving non-parametric memories when necessary.[1]

1 Introduction

Large language models (LMs; Brown et al. 2020; Raffel et al. 2020) have been shown to be competitive on diverse NLP tasks, including knowledge-intensive tasks that require fine-grained memorization of factual knowledge (Chowdhery et al., 2022; Yu et al., 2022). Meanwhile, LMs have also been shown to have limited memorization for less frequent entities (Kandpal et al., 2022), are prone to hallucinations (Shuster et al., 2021), and suffer from temporal degradation (Kasai et al., 2022; Jang et al., 2022). Incorporating *non-parametric knowledge* (i.e., retrieved text chunks) largely helps address those issues stemming from reliance on LMs' *parametric knowledge*—knowledge stored in their parameters (Izacard et al., 2022b)—but it is unclear whether it is strictly superior or complementary to parametric knowledge. Understanding when we should *not* trust LMs' outputs is also crucial to safely deploying them in real-world applications (Kadavath et al., 2022).

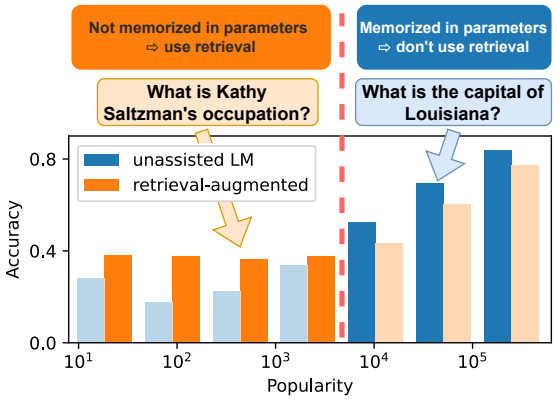

Figure 1: Relationship between subject entity popularity in a question and GPT-3 performance in open-domain QA, with and without retrieved passages. Adaptive Retrieval only retrieves when necessary (orange bars) based on the heuristically-decided threshold (red line).

This work conducts a large-scale knowledge probing of LMs on factual knowledge memorization, to understand when we should and should *not* rely on LMs' parametric knowledge, and how scaling and non-parametric memories (e.g., retrieval-augmented LMs) can help. In particular, we aim to address the following research questions:

(RQ_1) How much factual knowledge is memorized by LMs and what factors affect the memorization? (Section 4)

(RQ_2) To what extent can non-parametric memories alleviate the shortcomings of parametric memories of LMs? (Section 5)

(RQ_3) Can we build a system to adaptively combine non-parametric and parametric memories? (Section 6)

We hypothesize that factual knowledge frequently discussed on the web is easily memorized

[1]Our code and data are available at https://github.com/AlexTMallen/adaptive-retrieval.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 9802–9822
July 9-14, 2023

by LMs, while the knowledge that is less discussed may not be well captured and thus they require retrieving external non-parametric memories. We evaluate ten large LMs of three families (i.e., GPT-Neo, OPT, and GPT-3) with varying scales on the open-domain question answering (QA) task in a zero- or few-shot prompting manner. We construct a new dataset, POPQA, consisting of 14k questions to cover factual information in the long tail that might have been missed in popular QA datasets (Kwiatkowski et al., 2019). We use Wikipedia page views as a measure of popularity and convert knowledge triples from Wikidata, with diverse levels of popularity, into natural language questions, anchored to the original entities and relationship types. We also use EntityQuestions (Sciavolino et al., 2021), an open-domain QA dataset with a long-tail distribution.

On both datasets, LMs' memorization (RQ_1) is often limited to the popular factual knowledge and even GPT-3 `davinci-003` fails to answer the majority of the long-tail questions. Moreover, on such questions, scaling up models does *not* significantly improve the performance (e.g., for the 4,000 least popular questions in POPQA, GPT-j 6B has 16% accuracy and GPT-3 `davinci-003` has 19% accuracy). This also suggests that we can predict if LMs memorize certain knowledge based on the information presented in the input question only.

We next investigate whether a semi-parametric approach that augments LMs with retrieved evidence can mitigate the low performance on questions about less popular entities (RQ_2). Non-parametric memories largely improve performance on long-tail distributions across models. Specifically, we found that retrieval-augmented LMs are particularly competitive when subject entities are not popular: a neural dense retriever (Izacard et al., 2022a)-augmented GPT-neo 2.7B outperforms GPT-3 `davinci-003` on the 4,000 least popular questions. Surprisingly, we also find that retrieval augmentation can hurt the performance of large LMs on questions about popular entities as the retrieved context can be misleading.

As a result, we devise a simple-yet-effective retrieval-augmented LM method, Adaptive Retrieval, which adaptively combines parametric and non-parametric memories based on popularity (RQ_3). This method further improves performance on POPQA by up to 10%, while significantly reducing the inference costs, especially with larger LMs (e.g., reducing GPT-3 API costs by half), indicating the potential for future research in more efficient and powerful retrieval-augmented LMs.

2 Related Work

Parametric and non-parametric knowledge. Petroni et al. (2019) demonstrate that large pre-trained LMs such as BERT (Devlin et al., 2019) memorize the significant amount of world knowledge in their parameters (*parametric knowledge*), and Roberts et al. (2020) show that fine-tuned T5 without any reference documents (closed-book QA) can achieve competitive performance on open-domain QA. More recent and powerful LMs (Brown et al., 2020; Chowdhery et al., 2022) further improve performance on diverse knowledge-intensive tasks, leveraging their strong parametric memories (Kandpal et al., 2022; Yu et al., 2022). However, relying solely on their parameters to encode a wealth of world knowledge requires a prohibitively large number of parameters and the knowledge can become obsolete quickly (Kasai et al., 2022; Jang et al., 2022). Recent work shows that augmenting LMs with non-parametric memories (i.e., retrieved text chunks) enables much smaller models to match the performance of larger models (Izacard et al., 2022b; Khandelwal et al., 2020; Min et al., 2022), although Chen et al. (2022) and Longpre et al. (2021) show that even those models can ignore non-parametric knowledge and rely on parametric knowledge.

Understanding memorization. Several prior work establishes a positive relationship between string frequency in pre-training corpora and memorization (Carlini et al., 2022; Razeghi et al., 2022). Concurrent to our work, Kandpal et al. (2022) show that the co-occurrence of the question and answer entities in pretraining corpora has a positive correlation with models' QA accuracy on popular open-domain QA benchmarks such as Natural Questions (Kwiatkowski et al., 2019). This work, instead, attempts to predict memorization using the variables available in the input question only and uses popularity to obtain a proxy for how frequently an entity is likely to be discussed on the web. Importantly, by constructing a new dataset, we can conduct fine-grained controlled experiments across a wide range of popularities, allowing the investigation of hypotheses that might have been missed in prior analysis using existing open QA datasets. We further analyze the effectiveness and limitations of

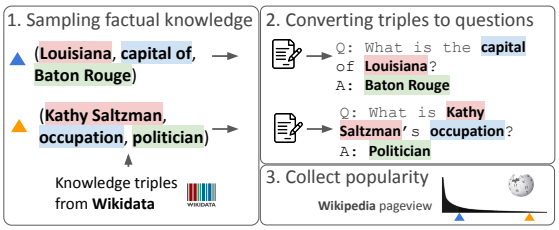

Figure 2: POPQA is created by sampling knowledge triples from Wikidata and converting them to natural language questions, followed by popularity calculation.

retrieval-augmented LMs and introduce Adaptive Retrieval. Prior work investigates the effectiveness of deciding when to use non-parametric memories at the token level in k-nn LM (He et al., 2021). This work is the first work to study the effectiveness of deciding whether to retrieve for each query and show their effectiveness in retrieval-augmented LM prompting.

3 Evaluation Setup

We evaluate LMs' ability to memorize factual knowledge through closed-book QA tasks with few-shot samples. We evaluate LMs on our new dataset, POPQA (Figure 2), and EntityQuestions, both of which have long-tail distributions (Figure 3).

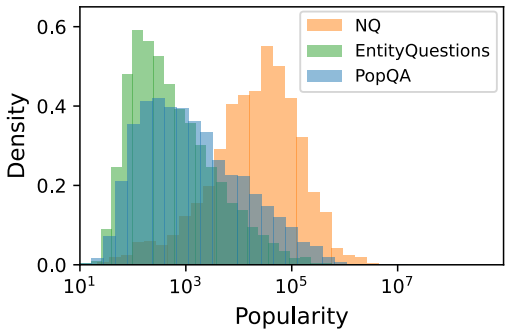

Figure 3: Distribution of subject entity popularity for EntityQuestions, POPQA, and for NQ-open for reference. Details on NQ entities can be found in Appendix A.

3.1 Focus and Task

Focus: factual knowledge. Among diverse types of world knowledge, this work focuses on factual knowledge (Adams, 2015) of entities—knowledge about specific details of the target entities. We define factual knowledge as a triplet of (subject, relationship, object) as in Figure 2 left.

Task format: open-domain QA. We formulate the task as open-domain QA (Roberts et al., 2020): given a question, a model predicts an answer without any pre-given ground-truth paragraph.[2] As in Kandpal et al. (2022), we study few-shot settings and prompt LMs without any parameter updates, instead of fine-tuning them on QA datasets such as in Roberts et al. (2020).

Metrics: accuracy. We mark a prediction as correct if any substring of the prediction is an exact match of any of the gold answers.

3.2 Dimensions of Analysis

We hypothesize that factual knowledge that is less frequently discussed on the web may not be well-memorized by LMs. Previous research often uses the term frequency of object entities in pretraining corpora to understand memorization (Févry et al., 2020; Kandpal et al., 2022; Razeghi et al., 2022). Instead, we investigate whether it's possible to predict memorization based on the input information only, and then apply the findings for modeling improvements, unlike prior analyses. Therefore, our work focuses on the other two variables in a factual knowledge triple: the subject entity and the relationship type.

Subject entity popularity. We use the popularity of the entities measured by Wikipedia monthly page views as a proxy for how frequently the entities are likely to be discussed on the web, instead of using the occurrence of entities or strings in the pretraining corpus (Carlini et al., 2022; Kandpal et al., 2022; Razeghi et al., 2022). Calculating frequencies over large pretraining corpora requires massive computations to link entities over billions of tokens, or can result in noisy estimations.[3] Our initial studies show that this is much cheaper[4] and aligns well with our intuition.

Relationship type. We also consider the relationship types as key factors for factual knowledge memorization. For example, even given the same combinations of the subject and object entities, model performance can depend on the relationship types; relationship types widely discussed can be easier to be memorized, while types that are less discussed may not be memorized much.

[2] Some work conducts knowledge probing of encoder-only models by filling out `[MASK]` tokens (Petroni et al., 2019). We use decoder-only models and thus do not use this `fill-in-the-blank` scheme.

[3] Moreover, several recent models like GPT-3 do not release their pretraining corpora, and it is an open question whether the frequencies in pretraining corpora reflect the frequencies in their private corpora.

[4] We can get page views by calling Wikipedia API.

3.3 Benchmarks

PopQA. In our preliminary studies, we found that existing common open-domain QA datasets such as Natural Questions (NQ; Kwiatkowski et al. 2019) are often dominated by subject entities with high popularity, and it is often hard to identify relationship types due to diverse question surface forms. To enable a fine-grained analysis of memorization based on the aforementioned analysis dimensions, we construct PopQA, a new large-scale entity-centric open-domain QA dataset about entities with a wide variety of popularity, as shown in Figure 3.

To construct PopQA, we randomly sample knowledge triples of 16 diverse relationship types from Wikidata and convert them into natural language questions, using a natural language template (depicted in Figure 2). We verbalize a knowledge triple (S, R, O) into a question that involves substituting the subject S into a template manually written for the relationship type R. The full list of templates is found in Table 2 of the Appendix. The set of acceptable answers to the question is the set of entities E such that (S, R, E) exists in the knowledge graph. We tried various templates and found that the results were fairly robust to the templates. Since PopQA is grounded to a knowledge base, links to Wikidata entities allow for reliable analysis of popularity and relationship types.

EntityQuestions. We test on another popular open-domain QA dataset, EntityQuestions (Sciavolino et al., 2021), which also covers a long-tail entity distribution. They use Wikipedia hyperlink counts as a proxy of the frequency of entities and sample knowledge triples from WikiData, from the frequency distributions. Unlike PopQA, EntityQuestions doesn't provide entity annotations, so we only use 82% of the questions, where the mention of the subject entity has a unique match with a Wikidata entity.

4 Memorization Depends on Popularity and Relationship Type

We evaluate a range of LMs with varying numbers of parameters, to quantify how much factual knowledge they memorize and how different factors affect those memorization behaviors (RQ_1).

4.1 Experimental Setup

Models. We evaluate ten models with a varying scale of model size: OPT (Zhang et al. 2022; 1.3, 2.7, 6.7, and 13 billion), GPT-Neo (Black et al. 2022; 1.3, 2.7, 6, and 20 billion), and GPT-3 (Brown et al. 2020; `davinci-002`, `davinci-003`) on our benchmark without any fine-tuning.[5]

Instructions and demonstrations. We use a simple template "`Q: <question> A:`" to format all of our questions for generative prediction. More sophisticated instructions were attempted in preliminary experiments but they did not improve upon the simple template significantly enough to merit using them, especially given that they may overfit to the model. While we use zero-shot prompting for GPT-3 to reduce API costs,[6] we use 15-shot prompting for all GPT-neo and OPT models.

4.2 Results

Overall model performance. The top left column of Figure 4 illustrates the overall performance on PopQA. As shown, even without using in-context examples, larger LMs exhibit reasonable performance: GPT-3 achieves 35% accuracy, and GPT-Neo 20B achieves 25% accuracy. This indicates that large LMs memorize factual knowledge in their parameters to some extent. This section examines which types of knowledge are better memorized and what factors influence memorization.

Subject entity popularity predicts memorization. Figure 4 (bottom) shows that there is a positive correlation between subject entity popularity and models' accuracy for almost all relationship types. This supports our hypothesis that subject entity popularity can be a reliable indicator of LMs' factual knowledge memorization. In general, the correlations between subject entity popularity and accuracy are stronger for larger LMs; GPT-3 003 shows the highest positive correlation (roughly 0.4) while GPT-Neo-1.3B shows relatively weak positive correlations (approximately 0.1).

Relationship types affects memorization. We find that models have a higher average performance for some relationship types than for others. While this is evidence that factual knowledge of some relationship types are more easily memorized than others, we also observe that questions of certain relationship types can be easily *guessed* without memorizing the knowledge triple. Specifically, certain relationship types (e.g., nationalities) allow models

[5] We did not explore widely-used encoder-decoder models such as T5, as their supervised pretraining consists of QA.

[6] Using 15-shot prompts for GPT-3 would cost upwards of $3000 for the combination of vanilla, Contriever, BM25, and GenRead evaluations on `davinci-002` and `davinci-003`.

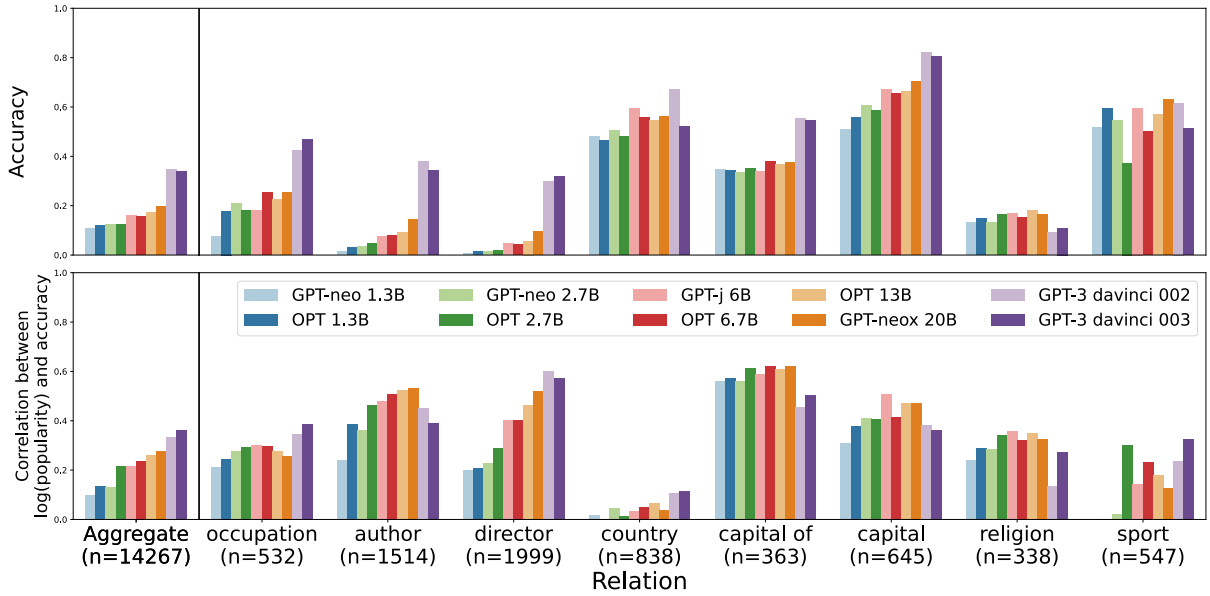

Figure 4: Per relationship type (n = the number of questions) results on PopQA by model, showing overall accuracy and the correlation between accuracy and log popularity. We uniformly bin the questions by $\log(\text{popularity})$, then report the correlation between the bin center and the bin's accuracy. We see that **both subject entity popularity and relationship type are strong predictors of memorization across models.** The correlation with popularity **exists across relationship types and is stronger for larger LMs**. We show a representative subset of relationship types and the complete results are in Figures 16 and 17 in Appendix C.1, including results on EntityQuestions.

to exploit surface-level artifacts in subject entity names (Poerner et al., 2020; Cao et al., 2021). Additionally, models often output the most dominant answer entities for questions about relationship types with fewer answer entities (e.g., red for the color relationship type). In Figure 4, relationships with lower correlation (e.g., country, sport) often shows higher accuracy, indicating that on those relationship types, models may exploit surface-level clues. On the other hand, for relationship types with relatively low accuracy (e.g., occupation, author, director), larger LMs often show a high correlation. Further details are in Appendix C.1.

Scaling may not help with tail knowledge. As seen in the left column of Figure 4, there are clear overall performance improvements with scale on the PopQA dataset. However, Figure 5 shows that on both PopQA and EntityQuestions, most of scaling's positive effect on parametric knowledge comes from questions with high popularity. Specifically, for the questions about the entities whose $\log_{10}(\text{popularity})$ is larger than 4, there is an improvement in accuracy as model size increases (red and yellow lines), while performance on questions with lower popularity remains relatively constant (blue and green lines). For the 4,000 least popular questions, GPT-Neo 6B, 20B, and GPT-3 `davinci-003` have 15%, 16%, and 19% accuracy, respectively.

This somewhat dampens prior works' findings that scaling up models significantly improves their factual knowledge memorization (Roberts et al., 2020; Kandpal et al., 2022). We hypothesize that

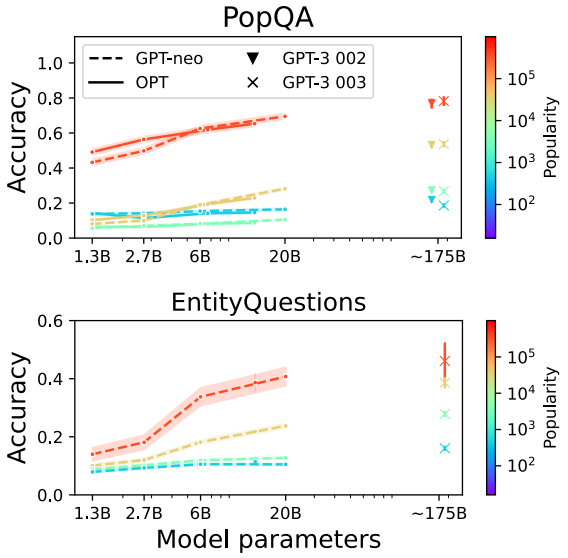

Figure 5: PopQA scaling results, broken down by question popularity level. **Scaling mostly improves memorization of more popular factual knowledge.** Error bars are 95% confidence intervals.

this is because their evaluations are often conducted on QA datasets with popular entities. In sum, scaling lowers the threshold of popularity for knowledge to be reliably memorized, but is not projected to move the threshold far into the long tail for practical model scales.

Relationship type results breakdown. Figure 6 provides a closer look at the relationship between popularity, accuracy, and relationship type; it shows model accuracy over the popularity distributions for director and country. For the first two types, we can see a clear positive trend between popularity and accuracy across models, and as the model size gets larger, the LMs memorize more. On the other hand, in the "country" relationship type, no models show trends, while overall the accuracy is high, indicating the LMs often exploit artifacts to answer less popular questions. We show example models' predictions in Appendix Section C.3.

5 Non-parametric Memory Complements Parametric Memory

Our analysis indicates that even the current state-of-the-art LMs struggle with less popular subjects or certain relationship types, and increasing the model size does not lead to further performance improvements. In light of this, we extend our analysis to non-parametric sources of knowledge, as outlined in Section (RQ_2). Specifically, we investigate the effectiveness of retrieval-augmented LMs (Borgeaud et al., 2022; Lewis et al., 2020), which leverage non-parametric memories (i.e., retrieved text) to improve performance.

5.1 Experimental Setup

Augmenting input. In this work, we try a simple retrieval-augmented LM approach, where we run an off-the-shelf retrieval system off-line to retrieve context from Wikipedia relevant to a question,[7] and then we concatenate the retrieved context with the original question. Although increasing the context size often leads to performance gains (Izacard and Grave, 2021; Asai et al., 2022), we only use the top one retrieved paragraph for simplicity.

Retrieval models. We use two widely-used retrieval systems: **BM25** (Robertson et al., 2009)

[6] 30 PopQA and 26 EntityQuestions questions had popularity less than the smallest popularity bin, and are excluded to avoid showing results for small sample sizes.

[7] We use Wikipedia dump from December 2018.

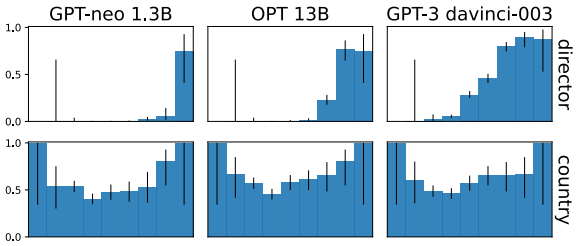

Figure 6: Memorization versus popularity for three models and the relationship types with the largest and smallest correlations. Within a relationship type, generally, there is a **monotonically increasing link between popularity and performance**, except for "country". Error bars show Wilson 95% confidence intervals.

and **Contriever** (Izacard et al., 2022a). BM25 is a static term-based retriever without training, while Contriever is pretrained on large unlabeled corpora, followed by fine-tuning on MS MARCO (Bajaj et al., 2016). We also experiment with a *parametric* augmentation method, **GenRead** (Yu et al., 2022), which prompts LMs to generate rather than retrieve a contextual document to answer a question. We use the ten LMs in Section 4, resulting in 40 LMs and retrieval-augmented LMs.

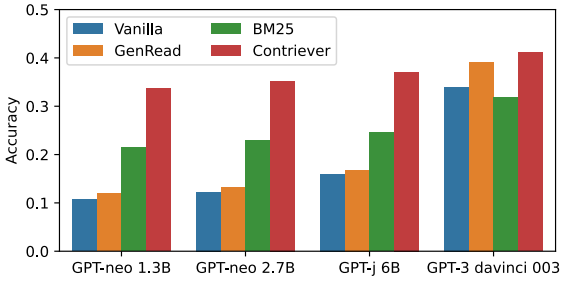

Figure 7: PopQA accuracy of LMs augmented with BM25, Contriever, GenRead, and unassisted (vanilla). **Retrieving non-parametric memories significantly improves the performance of smaller models.** Complete results on PopQA are found in Figure 13. EntityQuestions results are in Figure 14 of the Appendix.

5.2 Results

Retrieval largely improves performance. Figure 7 shows that augmenting LMs with non-parametric memories significantly outperforms unassisted vanilla LMs. A much smaller LM (e.g., GPT-Neo 2.7B) augmented by the Contriever retrieval results outperforms vanilla GPT-3. Large LMs such as GPT-3 also enjoy the benefits of non-parametric memories. Contriever gives 7% accuracy gains on top of GPT-3 `davinci-003`. Gen-

Read shows little-to-no performance improvement over vanilla parametric knowledge for smaller models, while the technique shows sizeable gains for GPT-3, especially `davinci-003`. In addition to its limited effectiveness with smaller LMs, GenRead has potentially prohibitive inference time costs, with GPT-NeoX 20B taking 70 seconds per query.

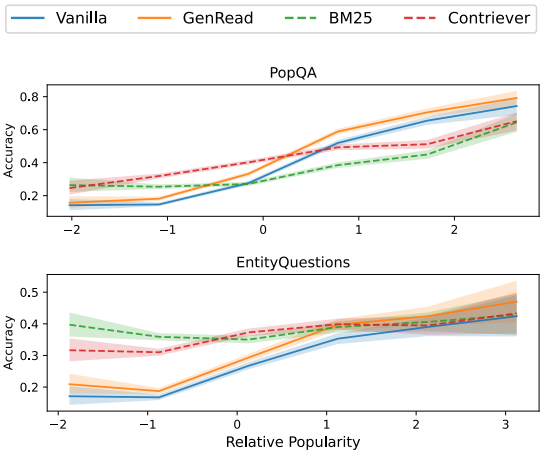

Figure 8: GPT-3 `davinci-003` accuracy versus relative popularity (how popular a question is relative to other questions of its relationship type). **Retrieval-augmented LMs (dashed) outperform LMs' parametric memory (solid) for less popular entities, while parametric memory is competitive for more popular entities.** Relative popularity is defined as the log-popularity of a question, normalized by the mean and standard deviation of log-popularity for the question's relationship type (smaller for less popular entities).[8] Figure 17 shows per-relationship results.

Non-parametric memories are effective for less popular facts. How does retrieval augmentation lead to such significant improvements? Figure 8 shows the relationship between the entity popularity and models' QA performance. It can be seen that retrieval-augmented LMs guided by Contriever or BM25 have a clear advantage over unassisted vanilla LMs, especially on less popular entities, resulting in a significant performance gain. Overall, Contriever-guided LMs outperform BM25-based ones on PopQA, while the BM25-based models perform better on the least popular entities, consistent with the findings from Sciavolino et al. (2021). On the other hand, for more popular entities, parametric knowledge shows equal or higher accuracy, indicating that the state-of-the-art LMs have already memorized the answers, and augmenting input with retrieved-context doesn't help much or even hurts the performance. Interestingly, GenRead generally outperforms vanilla LMs despite relying on LMs' parametric memory. This demonstrates the effectiveness of elicitive prompting (Wei et al., 2022; Sun et al., 2022) as observed in prior work. However, like vanilla LMs, GenRead shows low performance on less popular entities.

[8]Error bars show Wilson 95% confidence intervals. Bins with less than 40 samples have been excluded to avoid showing results with exceedingly wide errorbars.

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

Table 1: The recall@1 of Contriever for questions that GPT-3 `davinci-003` answered correctly and incorrectly with and without retrieval on PopQA. The percent of questions falling in each category is shown in parentheses. **For 10% of questions, retrieval is *harmful* due to low-quality retrieved text (0.14 recall@1).**

Non-parametric memories can mislead LMs. We conduct an in-depth analysis of why retrieval-augmented models suffer in more popular entities. We hypothesize that retrieval results may not always be correct or helpful, and can mislead LMs. To test this hypothesis, we group the questions based on two axes: whether unassisted GPT-3 `davinci-003` predict correctly or not, and whether retrieval-augmented predictions are correct or not. For each of the four categories, we calculate recall@1 (whether a gold answer is included in the top 1 document; Karpukhin et al. 2020).

Table 1 shows recall@1 for each group with percentages of the questions falling into each of the categories. For 10% of questions, retrieval-augmentation causes the LM to incorrectly answer a question it could otherwise answer correctly. We found that on those questions, recall@1 is significantly lower than the overall recall@1 (0.14 vs 0.42 overall), indicating that failed retrieval can result in performance drops. Conversely, for the 17% of questions for which retrieval causes the LM to correctly answer a question it would otherwise have failed to answer, the recall@1 is 0.88. We include examples of both cases in Appendix Section C.3.

6 Adaptive Retrieval: Using Retrieval Only Where It Helps

While incorporating non-parametric memories helps in long-tail distributions, powerful LMs have

already memorized factual knowledge for popular entities, and retrieval augmentation can be harmful. As outlined in (RQ_3), can we achieve the best of both worlds? We propose a simple-yet-effective method, Adaptive Retrieval, which decides when to retrieve passages only based on input query information and augments the input with retrieved non-parametric memories only when necessary. We show that this is not only more powerful than LMs or retrieval-augmented LMs always retrieving context, but also more efficient than the standard retrieval-augmented setup.

6.1 Method

Adaptive Retrieval is based on our findings: as the current best LMs have already memorized more popular knowledge, we can use retrieval only when they do not memorize the factual knowledge and thus need to find external non-parametric knowledge. In particular, we use retrieval for questions whose popularity is lower than a threshold (*popularity threshold*), and for more popular entities, do not use retrieval at all.

Using a development set, the threshold is chosen to maximize the adaptive accuracy, which we define as the accuracy attained by taking the predictions of the retrieval-augmented system for questions below the popularity threshold and the predictions based on parametric knowledge for the rest. We determine the popularity threshold independently for each relationship type.

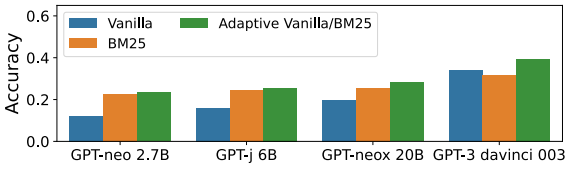

Figure 9: PopQA performance of GPT-neo models and GPT3 `davinci-003`, with different retrieval methods. **Adaptive Retrieval robustly outperforms approaches that always retrieve, especially for larger LMs**.

6.2 Results

Adaptive Retrieval improves performance. Figure 9 shows the results when we adaptively retrieve non-parametric memories based on the per-relationship type thresholds. We can see that adaptively retrieving non-parametric memories is effective for larger models. The best performance on PopQA is using GPT-3 `davinci-003` adaptively with GenRead and Contriever, yielding 46.5% accuracy, 5.3% higher than any non-adaptive method.

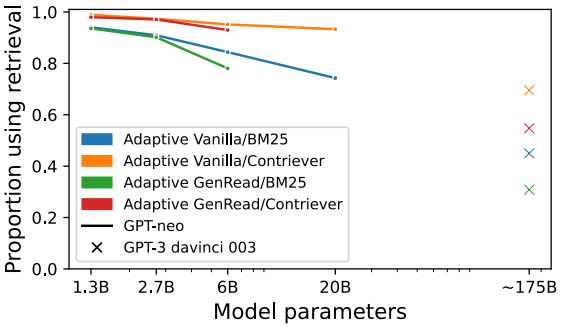

Figure 10: The proportion of questions for which various models use retrieval in the Adaptive Retrieval setup on PopQA. When using Adaptive Retrieval, small models must still rely on non-parametric memory for most questions, while larger models have more reliable parametric memories enabling them to use retrieval less often.

The threshold shifts with LM scale. While Adaptive Retrieval shows performance gains for larger models, smaller models do not realize the same benefits; as shown in Figure 9, the performance gain from Adaptive Retrieval is much smaller when we use models smaller than 10 billion. Why does this happen? Figure 10 shows that smaller LMs almost always retrieve, indicating that there are not many questions for which small LMs' parametric knowledge is more reliable than non-parametric memory. In contrast, large models typically retrieve much less. For example, GPT-3 `davinci-003` only retrieves for 40% of questions when paired with BM25, and even the much smaller GPT-neox 20B does not retrieve documents on more than 20% of the questions. On EntityQuestions (Appendix Figure 15) all of the LMs retrieve much more, as the questions are mostly about less popular entities.

Adaptive Retrieval reduces inference-time costs. We also found that Adaptive Retrieval improves efficiency; if we know we do not need to retrieve documents, we can skip retrieval components and the input length becomes shorter, which improves latency in both retrieval and language model components. Figure 11 shows the inference latency of GPT-J 6B and GPT-neox 20B, and API costs of GPT-3. Especially for larger LMs, concatenating retrieved context results in significantly increased latency (e.g., for GPT-J 6B, the inference time latency almost doubles). Adaptive retrieval enables reducing inference time up to 9% from standard

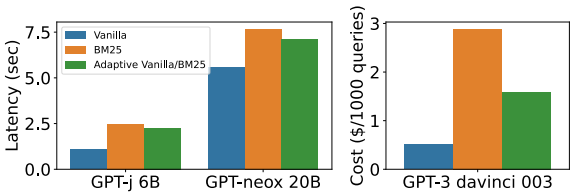

Figure 11: PopQA latency for large GPT-neo models that were run on our machines, and API costs for GPT3. **Adaptive retrieval reduces latency and API costs.**

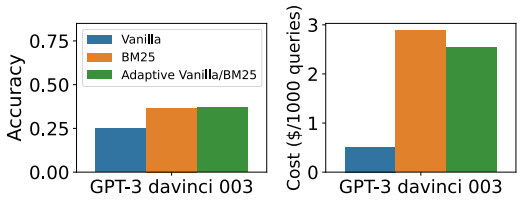

Figure 12: Accuracy and cost savings of Adaptive Retrieval for EntityQuestions. Despite EntityQuestions's lack of popular entities (see Figure 3), Adaptive Retrieval is able to reduce API costs by 15% while maintaining equivalent performance to retrieval only.

retrieval. We also observe cost reduction on EntityQuestions, as shown in Figure 12.

7 Discussion and Conclusions

This work conducts large-scale knowledge probing to examine the effectiveness and limitations of relying on LMs' parameters to memorize factual knowledge and to understand what factors affect factual knowledge memorization. Our results show that memorization has a strong correlation with entity popularity and that scaling up models on long-tail distributions may only provide marginal improvements. We also demonstrate that 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. Based on those findings, we devise simple-yet-effective Adaptive Retrieval, which only retrieves when necessary, using a heuristic based on entity popularity and relationship types. Our experimental results show that this method is not only more powerful than LMs or previous retrieval-augmented LMs but also more efficient.

Limitations

This work focuses on entity-centric factual knowledge and demonstrates that LMs' memorization is heavily affected by the popularity of the entities and the aspect of the entities being asked in the questions. It is important to emphasize that for running controlled experiments, we have relied on two synthetic datasets, and the extent to which our results apply to naturally occurring factual knowledge has not been firmly established. While we can be fairly confident about the relationship between scaling, retrieval, popularity, relationship type, and performance for the kinds of knowledge studied here, the effectiveness of Adaptive Retrieval will depend on many details of the question answering pipeline. Moreover, our work depends on a definition of popularity that is time-dependent and may not perfectly reflect how frequently entities are discussed on the web. Wikipedia page views are one possible definition of popularity for which we observe our results, and we invite others to improve upon it in future work. Further research can expand upon this simple approach, perhaps drawing on insights from Kadavath et al. (2022) to improve the effectiveness of Adaptive Retrieval.

It is an open question if the same findings are applicable to other types of world knowledge such as commonsense. We conjecture that the concept of the subject topic (entity), as well as the aspect (relationship type), can be applied with some minor modifications, which future work can quantify memorization following our scheme.

Ethical Considerations

Recent work (Huang et al., 2022) shows that LMs memorize personal information available on the web, which has significant security issues. Our evaluation focuses on the memorization of general entity-centric knowledge, but our findings can be applicable to those areas. Our findings suggest that LMs are likely to have less reliable knowledge of minority groups. Parrish et al. (2022) established that models often rely on stereotypes to answer in uncertain cases, so our results indicate that LMs are likely to rely on stereotypes disproportionately for minority groups. Future work could investigate whether retrieval augmentation reduces bias in these cases.

Acknowledgements

We thank the UW NLP group members for their helpful discussions, and Joongwon Kim, Wenya Wang, and Sean Welleck for their insightful feedback on this paper. This research was supported by NSF IIS-2044660, ONR N00014-18-1-2826,

ONR MURI N00014- 18-1-2670, and Allen Distinguished Award. AM is funded by a Goldwater Scholarship and AA is funded by the IBM PhD Fellowship.

References

Nancy E Adams. 2015. Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association*.

Akari Asai, Matt Gardner, and Hannaneh Hajishirzi. 2022. Evidentiality-guided generation for knowledge-intensive NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. MS MARCO: A human generated machine reading comprehension dataset.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing systems*.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models.

Hung-Ting Chen, Michael JQ Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Paolo Ferragina and Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Efficient nearest neighbor language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022a. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard

Grave. 2022b. Few-shot learning with retrieval augmented language models.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2022. Large language models struggle to learn long-tail knowledge.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime QA: What's the answer right now?

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wen-tau Yih, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Nonparametric masked language model.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-yet-effective entity embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot reasoning.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models.

Appendix

A Details of PopQA Constructions

List of the relationship types and templates. In this work, we use the following 16 relationship types, and the authors of this paper manually annotated templates to verbalize knowledge triple to natural language questions. We show the final list of the templates used to create PopQA in Table 2.

Figure 3 shows the distribution of subject popularity of PopQAand EntityQuestions versus the popular NQ benchmark. NQ may have multiple entities so the distribution of the least popular entity per question is shown. Subject entities from NQ were extracted using TagMe (Ferragina and Scaiella, 2010) on the NQ-open development set with a score threshold of 0.22. TagMe returns the title of a Wikidata entity which can be directly used to find popularity.

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 2: Full list of the manually annotated templated used for PopQAcreations. `[subj]` denotes a placeholder for subject entities.

Knowledge triples sampling. In the construction of the PopQAdataset, knowledge triples are sampled with higher weight given to more popular entities, otherwise, the distribution would be dominated by the tail and we would not have enough high-popularity entities to complete our analysis. Specifically, when considering whether to sample a particular knowledge triple, we include the knowledge triple if and only if $f > \exp(8R - 6)$, where $R \sim U(0, 1)$ is a unit uniform pseudo-random number and f is the exact match term frequency of the subject entity's aliases in an 800 MB random sample of C4. To increase diversity, once 2000 knowledge triples of a particular relation type have been sampled, they are no longer sampled.

B Experimental Details

Computational resources and API costs. GPT-3 API usage totaled to $275. We ran 14,282 questions through two GPT-3 `davinci` models using four different methods: vanilla experiments cost $13 ($0.46 per 1000 questions), Contriever-augmented experiments cost $88 ($3.08 per 1000 questions), BM25-augmented experiments cost $81 ($2.80 per 1000 questions), and GenRead experiments cost $93 ($3.25 per 1000 questions).

To run experiments using LMs larger than two billion parameters, we use a single V100 Volta GPU with 32GB GPU memories. We use int8bit (Zeng et al., 2022) quantization with OPT 13 billion and GPT-Neo 20 billion models to make them fit our GPUs. In our preliminary experiments using GPT-Neo 6 billion, we did not observe a notable performance drop by using the quantization.

Constructing few-shot contexts. For PopQA, we sample few-shot examples stratified by relationship type to diversify the samples: for each of the 15 relationship types other than the one in the test question, we sample one random question-answer pair to include in the context. For EntityQuestions, we take a simple random sample of 15 question-answer pairs because there are more than 16 relationship types.

Details of deciding thresholds. We 75% of PopQAto determine a popularity threshold for each relation type. Using brute force search, we select the threshold to maximize the adaptive accuracy, which we define as the accuracy attained by taking the predictions of the retrieval-augmented system for questions below the popularity threshold and the predictions based on parametric knowledge for the rest.

We then evaluate adaptive accuracy using the learned thresholds on the remaining 25% of PopQA, and repeat with 100 different random splits and take the mean to obtain the reported adaptive accuracy measurement.

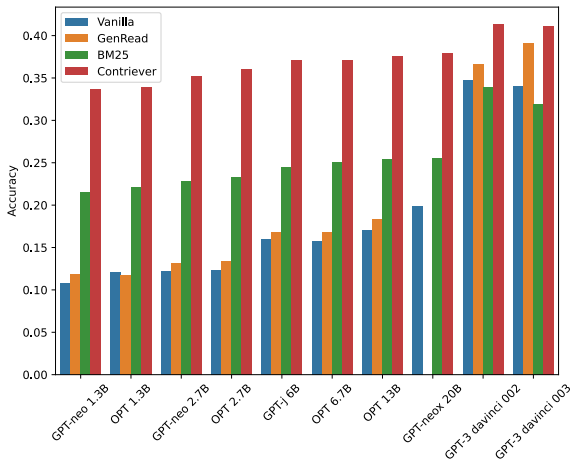

Figure 13: Accuracy by LMs and retrieval-augmented LMs on PopQA. This is an extension of Figure 7

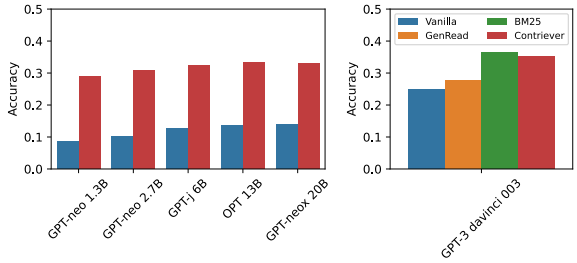

Figure 14: Accuracy by LMs and retrieval-augmented LMs on EntityQuestions.

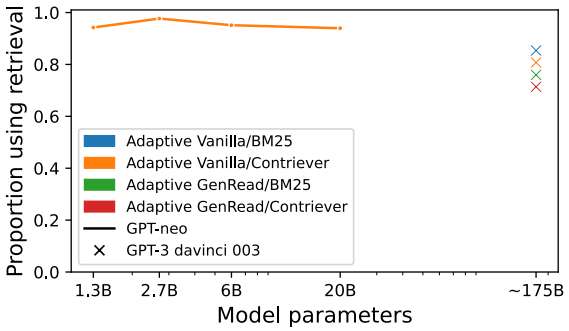

Figure 15: The proportion of questions for which Adaptive Retrieval uses retrieval versus model size for EntityQuestions.

C Detailed Results

C.1 LM results

Full results of per-relationship type accuracy and correlation. Figure 16 shows the full result of per-relationship type accuracy for all relationship types in PopQA. Figure 17 shows the correlations for all relation types. Figures 19 and 18 show the same results for the EntityQuestions dataset.

Negative correlations of capital on EntityQuestions. As shown in Figure 19, the capital relationship types on in EntityQuestions, while on PopQA, this relationship shows relatively high correlations. We found that in EntityQuestions, this capital relationship type has many low-popularity questions whose answers are included in subject entity names (e.g., subject="canton of Marseille-Belsunce", object="Marseille"). This causes performance to have a U-shaped relationship with popularity for the capital relationship type, so if most of the questions sampled come from the top half of popularity, the linear correlation will be positive, and vice versa.

C.2 Retrieval-augmented LM results

Overall performance of retrieval-augmented LMs. Figure 13 shows the overall performance of 40 LMs and retrieval-augmented LMs on PopQA. Retrieval-augmentation largely improves performance across different LMs, and much smaller models (GPT-Neo 1.3B) can perform on per with GPT-3. Figure 14 shows the results on EntityQuestions. Due to computational and time constraints, we were only able to run vanilla and Contriever results for most models.

Adaptive Retrieval for EntityQuestions. Figure 15 shows the proportion of questions above the retrieval threshold for various models using Adaptive Retrieval on EntityQuestions. Because EntityQuestions has a large quantity of low-popularity questions, models (especially smaller ones) must rely heavily on retrieval.

Full results on all relationship types. Figure 20 shows the full results on PopQA of the retrieval-augmented LMs and unassisted LMs on 16 relationship types using three different LMs as backbones. Figure 21 shows these results for GPT-3 `davinci-003` on EntityQuestions.

C.3 Qualitative Results

Table 3 shows several examples on PopQA, where GPT-3 `davinci-003` answers correctly while the Contriever-augmented version fails to answer. Along with the low recall@1 of 0.14 for this group, Table 3 suggests that the most common reason retrieval can be harmful is that it retrieves a document about a mistaken entity, such as a person with the same name as the subject, or an entity that simply is not relevant to the question (as in the case of "Noel Black").

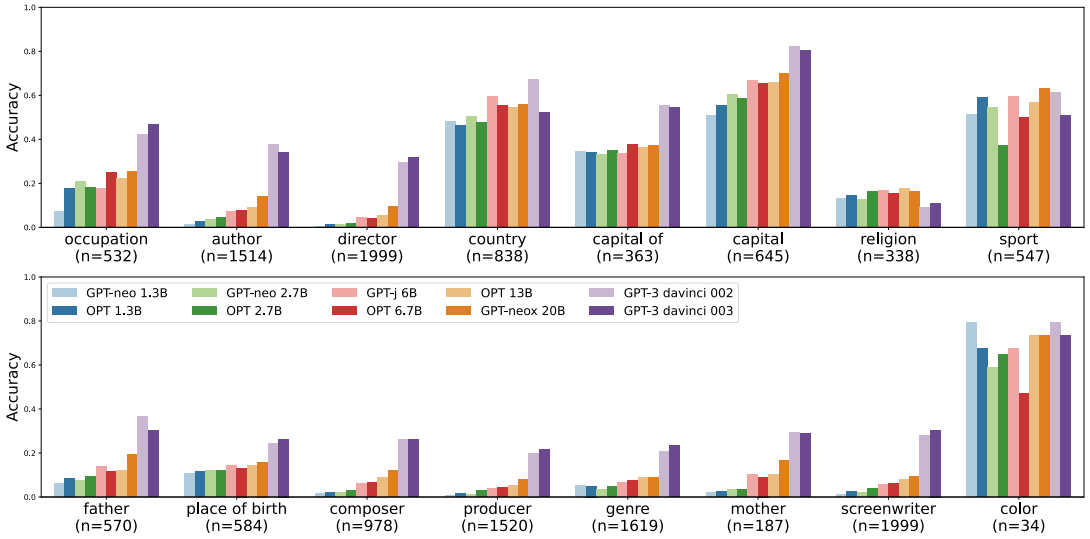

Figure 16: Accuracy on PopQA for all relationship types and models. This is an extension of Figure 4.

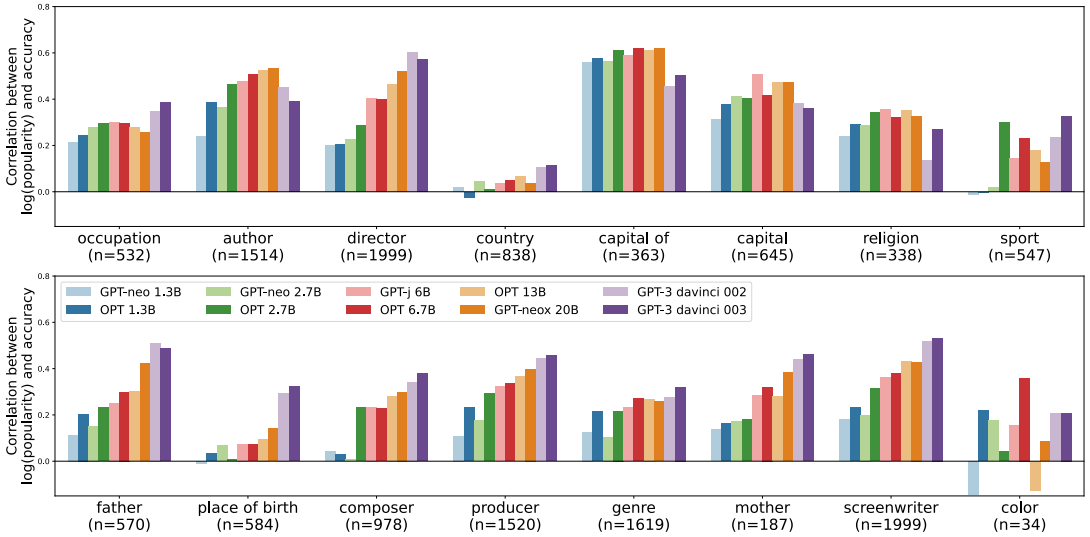

Figure 17: Correlations on PopQA for all relationship types and models. This is an extension of Figure 4.

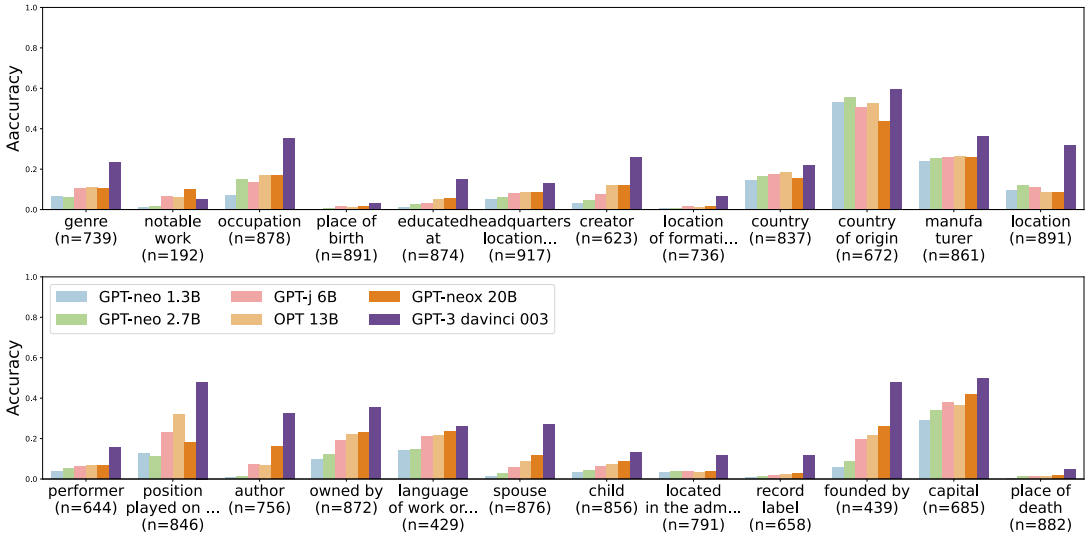

Figure 18: Accuracy on EntityQuestions for all relationship types and models.

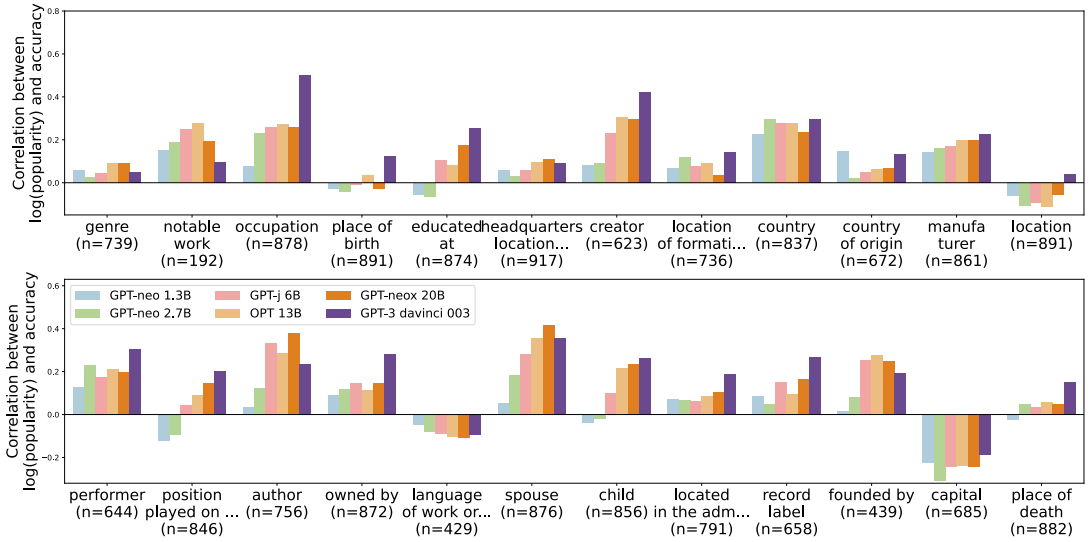

Figure 19: Correlations on EntityQuestions for all relationship types and models.

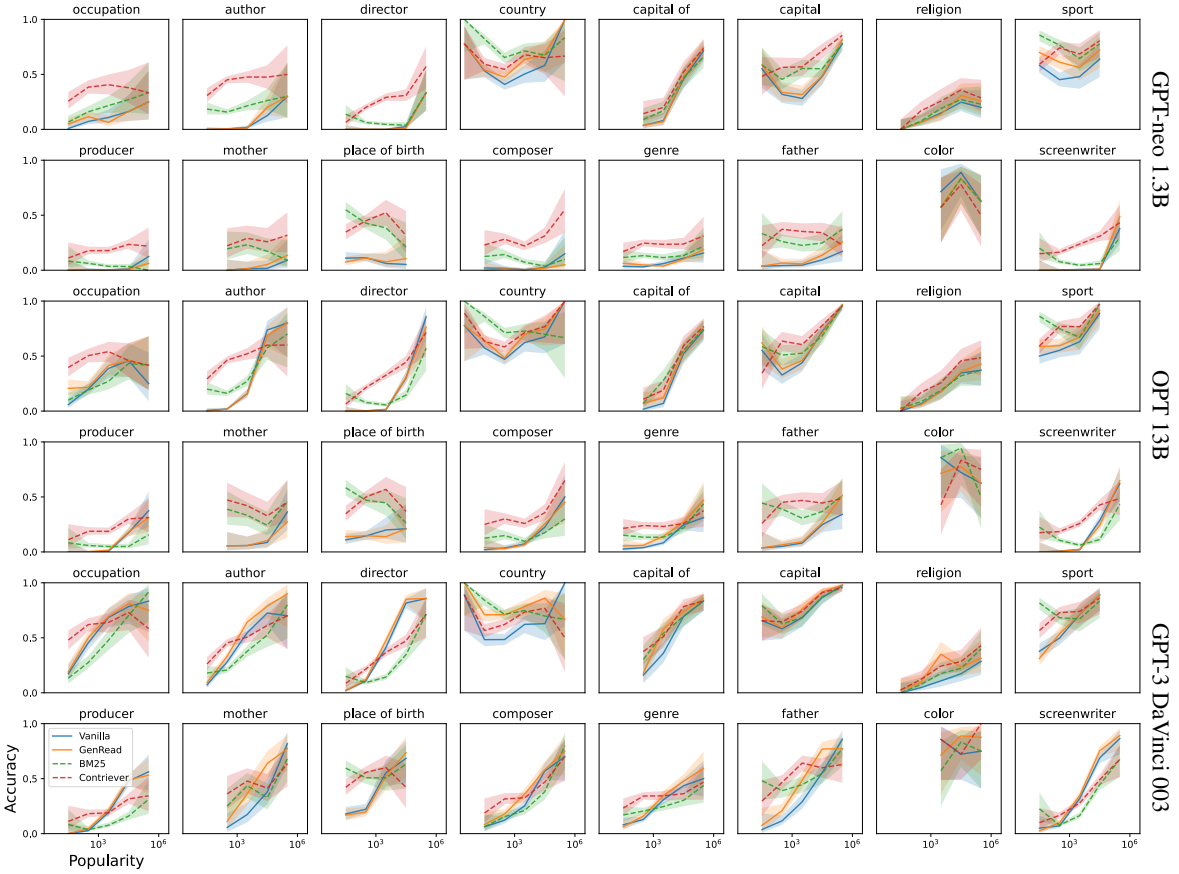

Figure 20: Accuracy for 3 models on POPQA versus popularity as shown in Figure 8 broken down by relationship type. Popularity bins with less than 5 samples are excluded to avoid cluttering the figures with noisy results that have wide error bars.

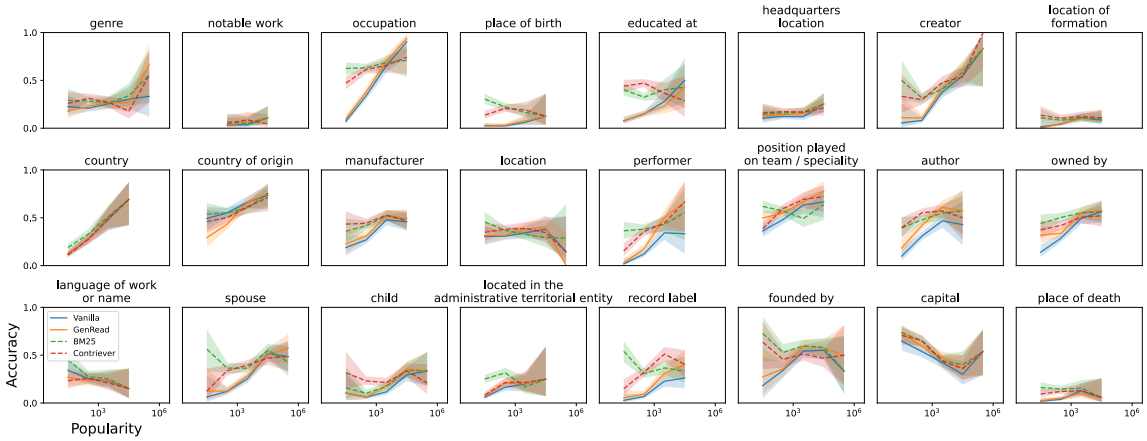

Figure 21: Accuracy versus popularity for GPT-3 `davinci-003` on EntityQuestions broken down by relationship type. Popularity bins with less than 5 samples are excluded.

Table 4 shows several examples on PopQA, where GPT-3 `davinci-003` answers correctly only when augmented with Contriever. The recall@1 for this case is 0.88, which is significantly higher than the overall recall. Note that in the second example, the retrieval caused the LM to answer correctly, but only by coincidence: the subject entity "Pierre" actually refers to the city in South Dakota, not the Basketball player. Otherwise, retrieval appears to be helpful because it provides the relevant information directly.

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.

Table 3: Qualitative examples of the questions where only retrieval-augmented LMs *fail* to answer correctly. The blue underlined text indicates the sub-strings matching the gold answers in the retrieved context.

Question (answer)	LM	Retrieval-LM	Retrieval
What genre is Unknown? (fantasy)	Unknown is not a specific genre of music. It could refer to ✗	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. ✗	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. ✗	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. ✗	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".

Table 4: Qualitative examples of the questions where only retrieval-augmented LMs *successfully* answer correctly. The blue underlined text indicates the sub-strings matching the gold answers in the retrieved context.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7: Limitations

☑ A2. Did you discuss any potential risks of your work?
Section 7: Ethical Considerations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3.3

☑ B1. Did you cite the creators of artifacts you used?
Section 3.3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The license is in our GitHub repository, which will be linked to from the abstract in the non-anonymous version.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Our dataset only contains data from Wikidata, which is widely used for NLP experiments and is already publicly available.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Our dataset only contains data from Wikidata, which is widely used for NLP experiments and is already publicly available.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix A

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Sections 4, 5, and 6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Sections 4.1 and 5.1, Appendix B

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sections 4, 5, and 6

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Sections 4, 5, and 6, Appendix C

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.