City, University of London MSc in Artificial Intelligence
Project Report
Year 2023/2024

Knowledge Grounding in Language Models: An Empirical Study

Martin Fixman
Supervised By: Tillman Weyde
Collaborators: Chenxi Whitehouse and Pranava Madhyastha

October 2 2024

Declaration

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation.

In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed: *Martin Fixman*

Acknowledgements

Abstract

This is an abstract

Contents

1 Introduction and Objectives

1.1 Problem Background

In recent years, Large Language Models (LLMs) have become ubiquitous in solving general problems across a wide range of tasks, from text generation to question answering and logic problems. However, recent research suggests that using these models alone might not be the most effective way to solve problems that are not directly related to text generation (Yao et al. 2023).

One approach to improving the performance on knowledge problems for LLMs is Retrieval-Augmented Generation (RAG) (Lewis et al. 2020). RAG involves retrieving relevant context related to a query and incorporating it into the model's input, enhancing the model's ability to generate accurate and contextually appropriate responses.

As RAG-enhanced systems become more widespread, studies on the performance of different retrieval systems and their interaction with LLMs have become crucial. Many explore the performance of these downstream tasks depending on both the retriever and the generator (Ghader et al. 2023, Brown et al. 2020), examining whether the knowledge is *grounded* in the context. Retrieval-Augmented models, such as Atlas (Izacard et al. 2022) and Retro (Borgeaud et al. 2022), use this approach to fine-tune a model on both a large body of knowledge and an existing index for context retrieval.

This project aims to understand the performance of various LLMs by measuring their *knowledge grounding* on a dataset consisting of a large variety of questions across a wide range of topics. We follow the approach by Yu et al. of running queries with counterfactual context to understand whether a particular answer originates from the model's inherent knowledge (i.e., its training data) or from the provided context (i.e., the context retrieved by RAG).

This thesis builds on this knowledge and improve our understanding of how different LLMs interact with the given context in the problem of question answering. Specifically, we investigate whether these interactions vary depending on the type of question being answered, contributing to a more nuanced understanding of LLM performance in diverse knowledge domains.

1.2 Thesis Questions & Objectives

This thesis is structured around three different objectives to deepen our understanding knowledge grounding in large language models.

1.2.1 Creating a representative dataset of questions

The research of this thesis requires a large dataset of questions from a variety of categories to test large language models. In order to understand knowledge grounding in these models, we require a dataset with the following properties.

1. The dataset must contain questions that have short, unambiguous answers.
2. The questions must cover a large set of topics.
3. It must allow for the creation of counterfactual answers in the same format as correct ones to test contextual versus inherent knowledge.

The existing literature uses various existing question-and-answer datasets, none of which are useful for this research.*

Natural Questions Dataset Created by Google Research (Kwiatkowski et al. 2019), and commonly used in research related to understanding the answers of LLMs in question-and-answer problems (Hsia et al. 2024, Mallen et al. 2023, Ghader et al. 2023). While the dataset provides an excellent range of questions and existing literature to compare these results to, the lack of categorisation is an obstacle in our objective to generate counterfactual answers.

Human-Augmented Dataset Sometimes used in research related to quality control of large language models (Kaushik et al. 2020). However, the high cost associated with this dataset would limit the size of our questions.

Countries' Capitals Question Dataset Used in "Characterizing Mechanisms for Factual Recall in Language Models" (Yu et al. 2023), this dataset contains a single question about the capital city of certain countries which can be easily transformed to a counterfactual question. This format is ideal for the research done in this thesis, but having a single question pattern will not allow a deep dive into the source of each answer in a general question.

Instead of using an existing dataset, this research takes inspiration from the paper by Yu et al. to create a similar but larger dataset of questions and answers from a wide range of topics, where questions can be grouped by question pattern to ensure that their formats are similar. This way, we can emulate the approach of that paper of using the answer from a certain question as the counterfactual of another.

This dataset will be used to test the remaining questions of this thesis. Since it might be useful for future research, it will also be presented as its own result.

*TODO: Maybe this entire subsubsection should go on Section 2 or Section 3.

1.2.2 When does a model choose the provided context knowledge over its inherent knowledge?

Currently, little is understood about the factors and mechanisms that control whether an LLM will generate text respecting either the context or the memorised information.

Previous research found out that, when the context of a query contradicts the ground knowledge of a model, the answer picked depends on the type and size of the model used (Yu et al. 2023).

This thesis extends this research by testing the representative set of questions and counterfactuals described in the previous section with both Seq2Seq and Decoder-only models of various sizes. We also research the cases when the answer doesn't correspond to either the parametric or contextual knowledge, and why the model chooses a third type of answer when adding counterfactual context.

This thesis also gathers insights from answering this question on different categories and patterns of questions to find out if this depends on what is being asked.

1.2.3 Can we use the perplexity score of an answer to predict whether it came from inherent or contextual knowledge?

Yu et al. showed that there is a correlation between the probability of a large language model choosing a parametric answer over a counterfactual contextual answer and the amount of times this answer appears in the ground truth data of the model. This gives us clues on whether the result of a query came from parametric or contextual knowledge if we have access to this ground truth, as is the case in models like Pythia (Biderman et al. 2023).

Unfortunately, most so-called open-source large language models do not give us access to the source data being used to train it and therefore do not allow this kind of analysis.

The **perplexity** score of answer gives a measure of how "certain" a large language model is of its answer (Jiang et al. 2021). We hypothesise that we can use this metric to serve as a reliable indicator of whether a particular answer was memorised by the LLM or was derived from the provided context.

2 Context

This research is the latest on a long line of academic articles on the topics of retrieval-augmented generation, counterparametric and contextual data, and how to enhance knowledge on large language models.

This section presents a short summary of some of the articles that were useful in researching this topic.

2.1 Foundational Papers on Large Language Models

- "Language models are unsupervised multitask learners"(Radford et al. 2019).
  - The foundational paper for GPT2.
- "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" (Raffel et al. 2020).
  - The foundational paper for T5.
- "Language Models are Few-shot Learners" (Brown et al. 2020).
  - Introduces "in-context learning".
- "Prompt programming for large language models: Beyond the few-shot paradigm" (Reynolds & McDonell 2021).
  - Improves the previous paper.

2.2 Papers working with RAG and contextual data

- "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" (Lewis et al. 2020).
  - Foundational paper for RAG.
- "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection" (Asai et al. 2023).
  - Interesting RAG system.
- "Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model" (Ghader et al. 2023).
  - Nice evaluation of RAG models.

2.3 Retrieval-Augmented Language Models

- "Shall We Pretrain Autoregressive Language Models with Retrieval? A Comprehensive Study" (Wang et al. 2023).

[*]This entire section is in progress — short summaries of the named papers will come soon.

  - Reproduces and pretraines RETRO.
- "Atlas: Few-shot Learning with Retrieval Augmented Language Models" (Izacard et al. 2022).
  - Introduces ATLAS.
- "Improving language models by retrieving from trillions of tokens" (Borgeaud et al. 2022).
- "RAGGED: Towards Informed Design of Retrieval Augmented Generation Systems" (Hsia et al. 2024).
  - Analyses results of these systems; compares Llama to Flan-T5.

2.4 On disentangling parametric and context-augmented counterparametric knowledge

- "DISCO: Distilling Counterfactuals with Large Language Models" (Chen et al. 2023).
  - Does similar analysis with counterfactuals to this thesis

  .

- "DisentQA: Disentangling Parametric and Contextual Knowledge with Counterfactual Question Answering" (Neeman et al. 2022).
  - Also does a similar analysis to this thesis.
- "Characterizing Mechanisms for Factual Recall in Language Models" (Yu et al. 2023).
  - Very simple analysis, but tries to understand WHERE in the model the contextual answers come from.
- "Can We Edit Factual Knowledge by In-Context Learning?" (Zheng et al. 2023).
- "Learning the Difference that Makes a Difference with Counterfactually-Augmented Data" (Kaushik et al. 2020).

3 Methods

3.1 Models and Resources Used

- **Reader Models**
  - Llama-8B.
  - Llama-70B.
  - Flan-T5-XL.
  - Flan-T5-XXL.
  - *Atlas?*
- **Questions**
  - Our own dataset, shown in Appendix A.
  - *Maybe add Natural Questions, HotpotQA, and/or BioASQ as in* RAGGED (Hsia et al. 2024).

3.2 Preprocessing and Inference Methodology

3.2.1 Source Data Preparation

Our source data is prepared by extending the ideas presented by Yu et al.. Instead of using one simple question, our approach consists of separating this data into 7 categories, where each category has a set of base questions and another set of objects that are paired together and presented to our models.

This work contains 7 categories in the configuration shown by Table 1, for a total of 3840 questions. The full list of questions can be found in Appendix A.

Category	**Questions**	**Objects**	**Total**
Person	14	47	658
City	14	60	840
Principle	10	30	300
Element	10	35	350
Book	10	45	450
Painting	14	39	546
Historical Event	6	56	336
Total	68	312	3840

Table 1: The amount of questions for each category. The full list of questions can be found in Appendix A. This is still a work in progress and I expect to add more questions.

We enhance the zero-shot learning prompt used by Brown et al. by using the prompt format example format presented Jiang et al. for calibrating the T5 language model by adding both the question and the first part of the answer.

3.2.2 Prompting

There is plenty of research that suggests that for zero-shot problems (Brown et al. 2020, Reynolds & McDonell 2021), it's convenient to create a minimal prompt (Jiang et al. 2021, Yu et al. 2023). This is helpful when later calculating the perplexity of the answers, as it tends to bias for short answers without any extra information that might change the individual probabilities of each token.

Examples of the prompting format explained in Sections 3.2.1 and 3.2.2 can be found in Table 2. For later queries, this is enhanced with context as in Table 3.

Base Question	Object	Final Question
`What is the date of birth of {person}?` `The date of birth of {person} is`	`Che Guevara`	`Q: What is the date of birth of Che Guevara?` `A: The date of birth of Che Guevara is`
`In what city was {person} born?` `{person} was born in`	`Confucius`	`Q: What is the date of birth of Confucius?` `A: The date of birth of Confucius is`
`What country is {city} in?` `{city} is in`	`Cairo`	`Q: In what city was Che Guevara born?` `A: Che Guevara was born in`
	`Mumbai`	`Q: In what city was Confucius born?` `A: Confucius was born in`
		`Q: What country is Cairo in?` `A: Cairo is in`
		`Q: What country is Mumbai in?` `A: Mumbai is in`

Table 2: Some examples of the base-question and object generation that are fed to the models for finding parametric answers.

3.2.3 Generating and scoring parametric answers

We query each of the models listed in Section 3.1 with the data from the previous subsubsections.

To ensure results are simple to interpret and not affected by randomness, we follow the example of Hsia et. al (Hsia et al. 2024) and use greedy decoding to find the answer. While beam search with a short beam width tends to produce more accurate results for long answers (Sutskever et al. 2014, Wu et al. 2016) and there are many other sampling methods that produce better results (Holtzman et al. 2020), this is likely to not have an effect on experiments that result in shorter answers (Raffel et al. 2020).

The negative log-likelihood of an answer x is calculated in base of the conditional probability of generating each token given the prior tokens. We can use this value to calculate the perplexity, which measures the level of "surprise" of a particular answer.

$$
\begin{aligned}
\mathrm{NLL}\left(x_1, \ldots, x_N | \, Q\right) &= -\frac{1}{N} \sum_{i=1}^{N} \log P\left(x_i \mid Q, x_{i-1}, \ldots, x_1\right) \\
\mathrm{PPL}\left(x_1, \ldots, x_N \mid Q\right) &= e^{\mathrm{NLL}\left(x_1, \ldots, x_N \mid Q\right)}
\end{aligned}
\tag{1}
$$

		Tokens	
		Parametric p	Counterparametric $\overline{p}$
Context	Empty Q	$\mathrm{PPL}\left(p_1, \ldots, p_N \mid Q\right)$	$\mathrm{PPL}\left(\overline{p}_1, \ldots, \overline{p}_{\bar{N}} \mid Q\right)$
	Counterparametric W	$\mathrm{PPL}\left(p_1, \ldots, p_N \mid W\right)$	$\mathrm{PPL}\left(\overline{p}_1, \ldots, \overline{p}_{\bar{N}} \mid W\right)$

Figure 1: Four different perplexity values: one for each set of tokens, and one for each query context..

We can ensure that the probabilities are calculated based on the intended tokens rather than the "most probable" generated ones by using teacher forcing (Lamb et al. 2016).

3.2.4 Shuffling to generate counterparametric answers

Previous work related to finding per token probabilities of answers in large language models focus on either a pre-existing list of questions or on a single question format (Yu et al. 2023). This approach does not work for our use case for three reasons.

1. Having 68 different types of questions, rather than just 1, makes finding counterfactual answers technically challenging.
2. Our focus is not on finding *counterfactual* answers, but *counterparametric* ones. We do not care about correctness; we care about answers not being parametric.
3. Since we are measuring perplexity of these answers, we focus on answers that are generated by the same base question and the same model. This way we ensure that the format of the answer is the same.

We propose a novel way of generating counterparametric answers while focusing on these three points: rather than generating new answers for each question, counterfactual answers are randomly sampled from the parametric answers corresponding to the same base question. An example of this approach can be seen in Table 3.

3.2.5 Counterparametric and contextual perplexity scores

This works extends the approach of analysing answers found in [citation needed] and explained in Section 3.2.3 by also calculating the perplexity of *alternative* answers to each question.

That is, we take the result of applying each model to both the answer with and without counterparametric context, and we calculate the perplexity scores of getting both the parametric and counterparametric answer to each one of these. This produces four different scores which are detailed in Figure 1: one for each answer using either empty and counterparametric context.

*I am finding it hard to explain this subsubsection. Maybe I should add pseudocode here.

Base Question	Parametric Answer	Counterparametric Answer	Question with counter-parametric context
What is the date of birth of Che Guevara?	June 14, 1928	June 21, 1947	Context: [the date of birth of Che Guevara is June 21, 1947]. Q: What is the date of birth of Che Guevara? A: The date of birth of Che Guevara is
What is the date of birth of Ibn al-Haytham?	965 AD	June 14, 1928	Context: [the date of birth of Ibn al-Haytham is June 14, 1928]. Q: What is the date of birth of Ibn al-Haytham? A: The date of birth of Ibn al-Haytham is
What is the date of birth of Boyan Slat?	27 January 1994	February 23, 1868	Context: [the date of birth of Boyan Slat is February 23, 1868]. Q: What is the date of birth of Boyan Slat? A: The date of birth of Boyan Slat is
What is the date of birth of W.E.B Du Bois?	February 23, 1868	June 14, 1928	Context: [the date of birth of W.E.B Du Bois is June 14, 1928]. Q: What is the date of birth of W.E.B Du Bois? A: The date of birth of W.E.B Du Bois is
What is the date of birth of Stephen Hawking?	January 8, 1942	965 AD	Context: [the date of birth of Stephen Hawking is 965 AD]. Q: What is the date of birth of Stephen Hawking? A: The date of birth of Stephen Hawking is
What is the date of birth of Shirin Ebadi?	June 21, 1947	June 14, 1928	Context: [the date of birth of Shirin Ebadi is June 14, 1928]. Q: What is the date of birth of Shirin Ebadi? A: The date of birth of Shirin Ebadi is

Table 3: Example of the sampling done to produce counterparametric answers. Counterparametric answers are generated by sampling a random answer from the parametric answers from the same base questions; to ensure that no parametric and counterparametric pair are identical, we only sample between different parametric answers. Note that the same parametric answer can appear several times as a counterparametric in different questions.

By definition, the tokens of the parametric answer $p_1, \ldots, p_N$ are the ones corresponding to the lowest perplexity answer for the query without any context. This is not the case for the tokens of the counterparametric answer $\overline{p}_1, \ldots, \overline{p}_{\bar{N}}$, which produces the inequality in Equation (2).

$$\mathrm{PPL}\left(p_1, \ldots, p_N \mid Q\right) \leq \mathrm{PPL}\left(\overline{p}_1, \ldots, \overline{p}_{\bar{N}} \mid Q\right) \tag{2}$$

Finding the result of the inequality for the queries with the counterparametric context W is one of the main goals of this research. In fact, we know that if the perplexity of the parametric tokens $p_1, \ldots, p_N$ is greater than the tokens for the counterparametric answer $\overline{p}_1, \ldots, \overline{p}_{\bar{N}}$ then the answer was memorised. Otherwise, the answer was generated in-context.

$$\text{Answer Source } = \begin{cases} \text{Memory} & \text{if } P\left(p_1, \ldots, p_N \mid W\right) < P\left(\overline{p}_1, \ldots, \overline{p}_{\bar{N}} \mid W\right) \\ \text{Context} & \text{otherwise} \end{cases} \tag{3}$$

3.2.6 Comparing the Final Answers

There is a third case that's not present in Equations (2) and (3): the case where the answer comes from neither the model's memory nor the query's context, but that instead the model generates a third answer combining both.

There are several cases where this can happen. The most interesting are explained in **??**, while the full results can be found in Appendix B.

In particular, we categorise the final answers in one of three groups depending on whether the answer with minimal perplexity on the query with the counterfactual context W is equal to the parametric answer, to the counterparametric answer, or to something else.

$$\text{Group} = \begin{cases} \text{Parametric} & \text{if } \left(\nexists\, x_1, \ldots, x_N\right)\ \mathrm{PPL}\left(x_1, \ldots, x_N \mid W\right) < A \\ \text{Counterparametric} & \text{if } \left(\nexists\, x_1, \ldots, x_N\right)\ \mathrm{PPL}\left(x_1, \ldots, x_N \mid W\right) < B \\ \text{Other} & \text{otherwise} \end{cases} \tag{4}$$

where

$$A = \mathrm{PPL}\left(p_1, \ldots, p_N \mid W\right)$$
$$B = \mathrm{PPL}\left(\overline{p}_1, \ldots, \overline{p}_{\bar{N}} \mid W\right)$$

There is a correlation between Equation (4) and Equation (3): an answer in the Parametric group will come from the model's memory, and an answer in the Counterparametric group will come from the query's (counterparametric) context.

4 Results

Some results I want to show.

- Larger models tend to prefer parametric knowledge over contextual knowledge.
  - This is the case in "Characterizing Mechanisms for Factual Recall in Language Models" (Yu et al. 2023), but I'm proving this on a larger set of question.
  - This is using exact match. Maybe attempting Unigram F_1 would produce interesting results (Petroni et al. 2021).
- How this compares between Decoder-only models, Seq2Seq models, and Retrieval-Augmented Language Models.
- How does the perplexity between parametric answers and contextual answers compare within the same model.
  - From the perplexity alone, can we predict whether an answer came from the model's memory or from the context?
  - It might be worth experimenting this with factual answers in the context, to simulate a RAG-difference detector.
- Is there any correlation between the perplexity of the parametric and contextual answer *without any context* and which one will be chosen when adding context?
  - This one is interesting, but I'm not sure we'll get significative results.
- Interesting *"Other"* results.
- **Anything else**?

4.1 Comparing the amounts of each type of answer

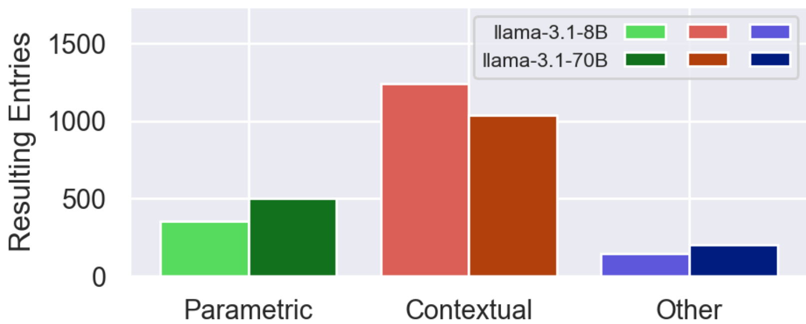

Figure 2: Amount of entries for each result after applying counterfactual context to Llama models. Generally, larger models tend to prefer parametric to contextual knowledge; this is further discussed in Section 5.2.

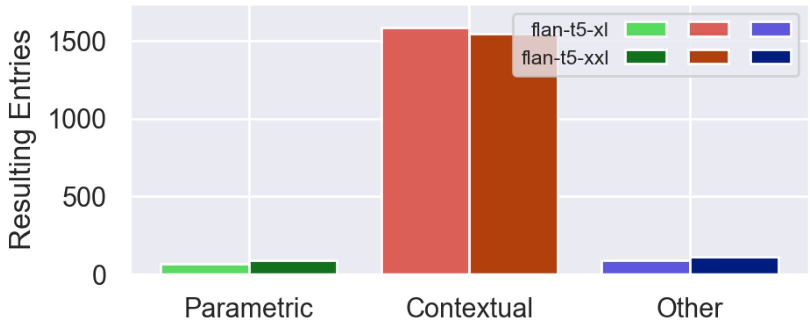

Figure 3: Same results for the Seq2Seq models FLAN-T5. While these models tend to be more biased towards contextual knowledge, as discussed in Section 5.1, larger models still are biased towards parametric knowledge.

4.2 Comparing the perplexity distribution for each type of answer

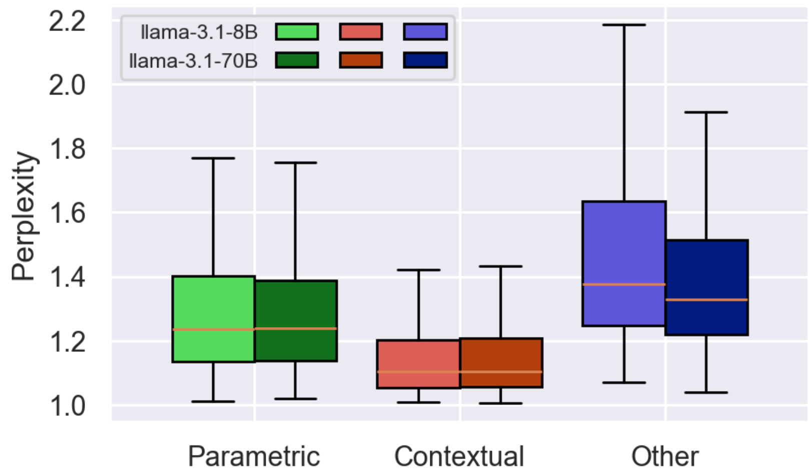

Figure 4: Perplexity box plots for Decoder-only Llama models.

Despite the amount for small and large Llama models being considerably different, the average values and distributions remain roughly the same. This is discussed in Section 5.

Additionally, the perplexity of contextual answers is considerably lower than the one for parametric answers.

Interestingly, the larger models tend to have a much lower perplexity for both paramteric and contextual answers.

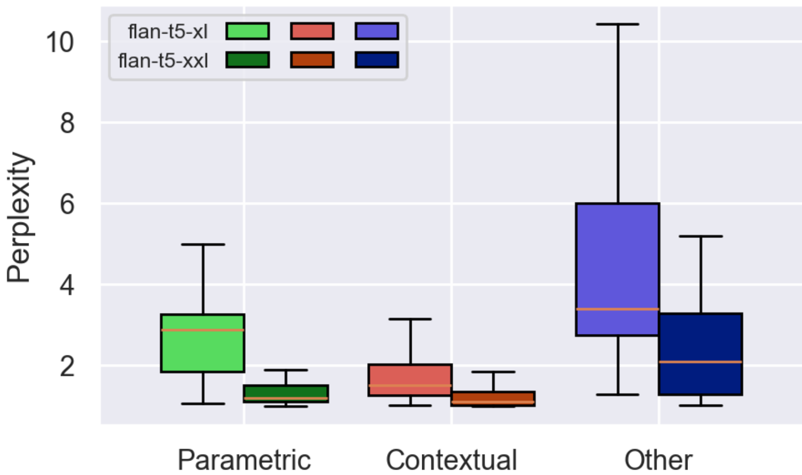

Figure 5: Perplexity box plots for Seq2Seq Flan models.

5 Discussion

5.1 Model type and memorised knowledge

5.2 Model size and memorised knowledge

5.3 Differences in perplexity scores for larger and smaller models

5.3.1 Can we use this to predict from where an answer came from?

5.4 Differences in distributions for different categories and questions.

6 Evaluations, Reflections, and Conclusions

Bibliography

Asai, A., Wu, Z., Wang, Y., Sil, A. & Hajishirzi, H. (2023), Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection, *in* 'International Conference on Learning Representations'.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E. et al. (2023), Pythia: A suite for analyzing large language models across training and scaling, *in* 'International Conference on Machine Learning', PMLR, pp. 2397–2430.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J. W., Elsen, E. & Sifre, L. (2022), 'Improving language models by retrieving from trillions of tokens'.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020), 'Language models are few-shot learners', *arXiv preprint arXiv:2005.14165* .

Chen, Z., Gao, Q., Bosselut, A., Sabharwal, A. & Richardson, K. (2023), 'DISCO: Distilling Counterfactuals with Large Language Models'.
URL: `https://arxiv.org/abs/2212.10534`

Ghader, P. B., Miret, S. & Reddy, S. (2023), 'Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model'.
URL: `https://arxiv.org/abs/2212.09146`

Holtzman, A., Buys, J., Du, L., Forbes, M. & Choi, Y. (2020), 'The curious case of neural text degeneration', *arXiv preprint arXiv:1904.09751* .

Hsia, J., Shaikh, A., Wang, Z. & Neubig, G. (2024), 'RAGGED: Towards Informed Design of Retrieval Augmented Generation Systems', *arXiv preprint arXiv:2403.09040* .

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S. & Grave, E. (2022), 'Atlas: Few-shot Learning with Retrieval Augmented Language Models'.

Jiang, Z., Araki, J., Ding, H. & Neubig, G. (2021), How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering, *in* 'Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, pp. 1974–1991.
URL: `https://aclanthology.org/2021.emnlp-main.150`

Kaushik, D., Hovy, E. & Lipton, Z. C. (2020), 'Learning the Difference that Makes a Difference with Counterfactually-Augmented Data'.
URL: `https://arxiv.org/abs/1909.12434`

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., Lee, K., Toutanova, K. N., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q. & Petrov, S. (2019), 'Natural Questions: a Benchmark for Question Answering Research', *Transactions of the Association of Computational Linguistics* .

Lamb, A., Goyal, A., Zhang, Y., Zhang, S., Courville, A. & Bengio, Y. (2016), Professor Forcing: A New Algorithm for Training Recurrent Networks, *in* 'Advances in Neural Information Processing Systems', Vol. 29, Curran Associates, Inc.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., K"uttler, H., Lewis, M., Yih, W.-t., Rockt"aschel, T., Riedel, S. & Kiela, D. (2020), 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks', *Advances in Neural Information Processing Systems* **33**, 9459–9474.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D. & Hajishirzi, H. (2023), 'When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories'.
URL: `https://arxiv.org/abs/2212.10511`

Neeman, E., Aharoni, R., Honovich, O., Choshen, L., Szpektor, I. & Abend, O. (2022), 'DisentQA: Disentangling Parametric and Contextual Knowledge with Counterfactual Question Answering'.
URL: `https://arxiv.org/abs/2211.05655`

Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T. & Riedel, S. (2021), KILT: a Benchmark for Knowledge Intensive Language Tasks, *in* K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty & Y. Zhou, eds, 'Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', Association for Computational Linguistics, Online, pp. 2523–2544.
URL: `https://aclanthology.org/2021.naacl-main.200`

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019), 'Language models are unsupervised multitask learners', *OpenAI blog* **1**(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J. (2020), 'Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer', *Journal of Machine Learning Research* **21**, 1–67.

Reynolds, L. & McDonell, K. (2021), Prompt programming for large language models: Beyond the few-shot paradigm, *in* 'Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems', pp. 1–7.

Sutskever, I., Vinyals, O. & Le, Q. V. (2014), 'Sequence to Sequence Learning with Neural Networks'.
URL: https://arxiv.org/abs/1409.3215

Wang, B., Ping, W., Xu, P., McAfee, L., Liu, Z., Shoeybi, M., Dong, Y., Kuchaiev, O., Li, B., Xiao, C., Anandkumar, A. & Catanzaro, B. (2023), Shall We Pretrain Autoregressive Language Models with Retrieval? A Comprehensive Study, *in* H. Bouamor, J. Pino & K. Bali, eds, 'Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Singapore, pp. 7763–7786.
URL: https://aclanthology.org/2023.emnlp-main.482

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M. & Dean, J. (2016), 'Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation'.
URL: https://arxiv.org/abs/1609.08144

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y. & Narasimhan, K. (2023), 'Tree of thoughts: Deliberate problem solving with large language models'.
URL: *https://arxiv.org/abs/2305.10601*

Yu, Q., Merullo, J. & Pavlick, E. (2023), 'Characterizing Mechanisms for Factual Recall in Language Models'.
URL: https://arxiv.org/abs/2310.15910

Zheng, C., Li, L., Dong, Q., Fan, Y., Wu, Z., Xu, J. & Chang, B. (2023), 'Can We Edit Factual Knowledge by In-Context Learning?'.
URL: https://arxiv.org/abs/2305.12740

Appendices

A Questions and objects used to form the queries

```
What is the date of birth of {person}? The date of birth of {person} is
In what city was {person} born? {person} was born in
What is the date of death of {person}? The date of death of {person} is
What is the primary profession of {person}? The primary profession of {person} is
What is {person} primarily known for? {person} is primarily known for
What's the main nationality of {person}? {person} is
What educational institution did {person} attend? {person} attended

What country is {city} in? {city} is in
What's the highest administrative subdivision {city} is part of? {city} is part of
In what year was {city} founded? {city} was founded in
What major river is nearest to {city}? The nearest major river to {city} is
What is the time zone of {city}? The time zone of {city} is
What is the current population of {city}? The current population of {city} is
What is the altitude of {city} above sea level? {city} is at an altitude of

Who is credited with the discovery of {principle}? {principle} was discovered by
Which scientific discipline encompasses {principle}? {principle} is encompassed by
What is the primary application of {principle}? The primary application of {principle} is
In which year was {principle} first formulated? {principle} was first formulated in
What is the SI unit most commonly associated with {principle}? The SI unit most commonly associated with
     {principle} is

What's the chemical formula for {element}? The chemical formula for {element} is
When was {element} first isolated? {element} was first isolated in
What's the atomic number of {element}? The atomic number of {element} is
What is the melting point of {element}? The melting point of {element} is
In which group of the periodic table is {element} found? {element} is found in group

What genre does {book} belong to? The genre of {book} is
Who's the author of {book}? {book} was written by
In what year was {book} first published? {book} was first published in
How many pages are in the original publication of {book}? The original publication of {book} has
What is the name of the main protagonist in {book}? The main protagonist in {book} is

Who painted {painting}? {painting} was painted by
When was {painting} completed? {painting} was completed in
What artistic movement does {painting} belong to? {painting} belongs to
What materials were used to create {painting}? {painting} was created with
Where is {painting} primarily housed? {painting} is currently in
What are the dimensions of {painting}? The dimensions of {painting} are
In which museum was {painting} first exhibited? {painting} was first exhibited in

What year did {historical_event} happen? {historical_event} happened in the year
Who was the primary leader associated with {historical_event}? The primary leader associated with
     {historical_event} was
What was the duration of {historical_event}? {historical_event} lasted for
```

Listing 1: All base questions used in this work. Each one of these will get combined with data from Listing 2 as detailed in Section 3.2.1.

```
Ada Lovelace,person
Alan Turing,person
Albert Einstein,person
Alexander Fleming,person
Aristotle,person
Billie Jean King,person
Boyan Slat,person
Catherine the Great,person
Che Guevara,person
Cleopatra,person
Confucius,person
Ernest Rutherford,person
Florence Nightingale,person
Freddie Mercury,person
Frida Kahlo,person
Greta Thunberg,person
Harriet Tubman,person
Ibn al-Haytham,person
Isaac Newton,person
Karl Marx,person
Leonardo da Vinci,person
Mahatma Gandhi,person
```

```
Malala Yousafzai,person
Mansa Musa,person
Marie Curie,person
Martin Luther King Jr.,person
Michelangelo,person
Mohandas Gandhi,person
Mozart,person
Muhammad Ali,person
Neil Armstrong,person
Nelson Mandela,person
Nikola Tesla,person
Pablo Picasso,person
Rosalind Franklin,person
Shirin Ebadi,person
Simon Bolivar,person
Srinivasa Ramanujan,person
Stephen Hawking,person
Sun Yat-sen,person
Virginia Woolf,person
Vladimir Lenin,person
Wangari Maathai,person
W.E.B. Du Bois,person
William Shakespeare,person
Wu Zetian,person
Yuri Gagarin,person
Alexandria,city
Amsterdam,city
Antananarivo,city
Athens,city
Baghdad,city
Berlin,city
Buenos Aires,city
Bukhara,city
Cairo,city
Cape Town,city
Cartagena,city
Chicago,city
Cusco,city
Cuzco,city
Delhi,city
Dubrovnik,city
Fez,city
Havana,city
Istanbul,city
Jerusalem,city
Kyoto,city
La Paz,city
Lhasa,city
Lisbon,city
London,city
Luang Prabang,city
Marrakech,city
Mexico City,city
Montevideo,city
Moscow,city
Mumbai,city
Muscat,city
New York,city
Nur-Sultan,city
Paris,city
Petra,city
Prague,city
Quebec City,city
Reykjavik,city
Rome,city
Sao Paulo,city
Sarajevo,city
Shanghai,city
Singapore,city
St. Petersburg,city
Sydney,city
Tbilisi,city
Tenochtitlan,city
Thimphu,city
Timbuktu,city
Tokyo,city
Ulaanbaatar,city
Varanasi,city
Venice,city
Vienna,city
Wellington,city
Windhoek,city
Xi'an,city
```

```
Yogyakarta,city
Zanzibar City,city
Archimedes' Principle,principle
Bernoulli's Principle,principle
Boyle's Law,principle
Cell Theory,principle
Conservation of Energy,principle
DNA Replication,principle
Electromagnetism,principle
Entropy,principle
Evolution by Natural Selection,principle
Evolution,principle
General Relativity,principle
Germ Theory of Disease,principle
Gravity,principle
Hardy-Weinberg Principle,principle
Heliocentrism,principle
Hubble's Law,principle
Kepler's Laws of Planetary Motion,principle
Le Chatelier's Principle,principle
Mendel's Laws of Inheritance,principle
Newton's Laws of Motion,principle
Pauli Exclusion Principle,principle
Periodic Law,principle
Photosynthesis,principle
Plate Tectonics,principle
Principle of Least Action,principle
Quantum Mechanics,principle
Relativity,principle
Superconductivity,principle
Thermodynamics,principle
Uncertainty Principle,principle
Aluminum,element
Barium,element
Bismuth,element
Bromine,element
Calcium,element
Carbon,element
Chlorine,element
Chromium,element
Copper,element
Gold,element
Helium,element
Hydrogen,element
Iodine,element
Iron,element
Lead,element
Lithium,element
Magnesium,element
Manganese,element
Mercury,element
Neon,element
Nitrogen,element
Oxygen,element
Phosphorus,element
Plutonium,element
Potassium,element
Radon,element
Silicon,element
Silver,element
Sodium,element
Sulfur,element
Thorium,element
Tin,element
Titanium,element
Uranium,element
Zinc,element
1984,book
Anna Karenina,book
Beloved,book
Brave New World,book
Catch-22,book
Crime and Punishment,book
Don Quixote,book
Fahrenheit 451,book
Frankenstein,book
Jane Eyre,book
Midnight's Children,book
Moby-Dick,book
One Flew Over the Cuckoo's Nest,book
One Hundred Years of Solitude,book
Pride and Prejudice,book
Slaughterhouse-Five,book
```

```
The Alchemist,book
The Art of War,book
The Book Thief,book
The Brothers Karamazov,book
The Catcher in the Rye,book
The Chronicles of Narnia,book
The Color Purple,book
The Count of Monte Cristo,book
The Grapes of Wrath,book
The Great Gatsby,book
The Handmaid's Tale,book
The Hitchhiker's Guide to the Galaxy,book
The Hobbit,book
The Hunger Games,book
The Kite Runner,book
The Little Prince,book
The Lord of the Rings,book
The Metamorphosis,book
The Name of the Rose,book
The Odyssey,book
The Picture of Dorian Gray,book
The Pillars of the Earth,book
The Stranger,book
The Sun Also Rises,book
The Wind-Up Bird Chronicle,book
To Kill a Mockingbird,book
Ulysses,book
War and Peace,book
Wuthering Heights,book
American Gothic,painting
Christina's World,painting
Girl with a Pearl Earring,painting
Guernica,painting
Les Demoiselles d'Avignon,painting
Liberty Leading the People,painting
Mona Lisa,painting
School of Athens,painting
Starry Night,painting
The Absinthe Drinker,painting
The Anatomy Lesson of Dr. Nicolaes Tulp,painting
The Arnolfini Portrait,painting
The Astronomer,painting
The Birth of Venus,painting
The Calling of Saint Matthew,painting
The Card Players,painting
The Death of Marat,painting
The Fighting Temeraire,painting
The Garden of Earthly Delights,painting
The Gross Clinic,painting
The Hay Wain,painting
The Kiss,painting
The Last Supper,painting
The Nighthawks,painting
The Night Watch,painting
The Ninth Wave,painting
The Persistence of Memory,painting
The Potato Eaters,painting
The Raft of the Medusa,painting
The Scream,painting
The Sleeping Gypsy,painting
The Son of Man,painting
The Swing,painting
The Third of May 1808,painting
The Tower of Babel,painting
The Treachery of Images,painting
The Triumph of Galatea,painting
The Wanderer above the Sea of Fog,painting
Water Lilies,painting
Decimalisation in the UK,historical_event
Queen Elizabeth II's Platinum Jubilee,historical_event
Queen Victoria's Coronation,historical_event
The Act of Union between England and Scotland,historical_event
The Battle of Adrianople,historical_event
The Battle of Adwa,historical_event
The Battle of Agincourt,historical_event
The Battle of Hastings,historical_event
The Battle of Sekigahara,historical_event
The Battle of Teutoburg Forest,historical_event
The Battle of the Milvian Bridge,historical_event
The Battle of Waterloo,historical_event
The Brexit Referendum,historical_event
The Codification of Roman Law by Justinian,historical_event
The Construction of Hadrian's Wall,historical_event
```

```
The Construction of the Great Pyramid of Giza,historical_event
The Conversion of Constantine,historical_event
The Council of Chalcedon,historical_event
The Crisis of the Third Century,historical_event
The Defeat of the Spanish Armada,historical_event
The Discovery of the Americas by Columbus,historical_event
The Dissolution of the Soviet Union,historical_event
The Division of the Roman Empire,historical_event
The Dunkirk Evacuation,historical_event
The Edict of Caracalla,historical_event
The Fall of Constantinople,historical_event
The Fall of the Aztec Empire,historical_event
The Fall of the Western Roman Empire,historical_event
The First Circumnavigation of the Earth,historical_event
The First Council of Nicaea,historical_event
The First Crusade,historical_event
The Founding of Constantinople,historical_event
The Founding of Rome,historical_event
The Founding of the British Broadcasting Corporation,historical_event
The Founding of the League of Nations,historical_event
The French Revolution,historical_event
The Glorious Revolution,historical_event
The Gothic War in Italy,historical_event
The Great Fire of London,historical_event
The Indian Independence Act,historical_event
The Industrial Revolution,historical_event
The London 7/7 Bombings,historical_event
The Meiji Restoration,historical_event
The Plague of Justinian,historical_event
The Reforms of Diocletian,historical_event
The Reunification of the Empire by Aurelian,historical_event
The Sack of Rome by Alaric,historical_event
The Sack of Rome by the Vandals,historical_event
The Signing of the Good Friday Agreement,historical_event
The Signing of the Magna Carta,historical_event
The Suez Crisis,historical_event
The Treaty of Westphalia,historical_event
The UK Abolition of the Slave Trade Act,historical_event
The Unification of Italy,historical_event
The Wedding of Prince Charles and Lady Diana,historical_event
The Year of the Four Emperors,historical_event
```

Listing 2: All objects which will be conbined with the questions in Listing 1.

B Full Results for Each Question

C Grounder Usage and Documentation

D Grounder Source Code