Knowledge Grounding in Retrieval-Augmented LM: An Empirical Study

Chenxi Whitehouse† **Eric Chamoun**† **Rami Aly**†

Abstract

1 Introduction

Retrieval systems are gaining increasing significance in improving factual and up-to-date generation in LLMs. Various paradigms of retrieval-augmented LLMs have been proposed, which can be categorised based on different architectures such as encoder-decoder models, decoder-only models, or by the integration of the retrieval component, either in pre-training, fine-tuning, or inference (see section 2 for details).

Most work in the literature compares and evaluates retrieval-augmented systems using metrics like perplexity (Guu et al., 2020; Borgeaud et al., 2022; Wang et al., 2023) or focuses on downstream tasks, particularly short-form generation like Natural Questions (Lewis et al., 2020; Izacard and Grave, 2021; Guu et al., 2020). However, the performance of these downstream tasks relies on both the *retriever* (i.e., the relevance of the retrieved context) and the *generator* (i.e., whether the generated content is *grounded* in the context), and very few studies address the conflation between these two aspects.

Addressing this gap, we propose an evaluation framework that specifically focuses on the *grounding* of diverse retrieval-augmented LLMs. A grounded model should demonstrate the capability to adapt its generation based on the provided context, particularly when the context contradicts the model's parametric memorisation.

2 Related Work

(Schwenk et al., 2022)

Retrieval-Augmented LMs Retrieval has been applied to various NLP tasks in recent years, proving to enhance the perplexity of the model (Wang et al., 2023) and the performance of knowledge-intensive downstream tasks (Lewis et al., 2020; Guu et al., 2020; Izacard and Grave, 2021). Various retrieval systems have been proposed, involving retrieval either during pre-training (e.g., REALM (Guu et al., 2020), RETRO (Borgeaud et al., 2022), Altas (Izacard et al., 2023), RETRO++ (Wang et al., 2023)), fine-tuning (e.g., DPR (Karpukhin et al., 2020), RAG (Lewis et al., 2020), FiD (Izacard and Grave, 2021)), or during inference (e.g., KNN-LM (Khandelwal et al., 2020)).

BehnamGhader et al. (2023) evaluate different pre-trained retrievers with LMs, discovering that LMs are imperfect reasoners even when provided with a perfect retriever that retrieves all the essential information. This work extends this assumption of gold retrieved-context and explores the grounding capability of different retrieval systems.

Parametric and Non-parametric Knowledge Many studies have explored parametric knowledge embedded in latent parameters and non-parametric knowledge derived from external context. Neeman et al. (2023) trained a model to distinguish between these two types of knowledge by generating varied responses to the same question in the Natural Questions dataset (Kwiatkowski et al., 2019) when providing factual, counterfactual, and unanswerable context. Yu et al. (2023a) evaluate how GPT models resolve the conflict between parametric knowledge and counterfactual context on the task of capital city prediction. Mallen et al. (2023) propose to use entity popularity to determine when to use non-parametric knowledge over stored memory. Zheng et al. (2023) demonstrate the potential to modify the factual knowledge of GPT models through in-context demonstrations.

This work is distinguished from prior work in that we focus on the grounding specifically in different retrieval systems and long-form generation.

†Core contributors.

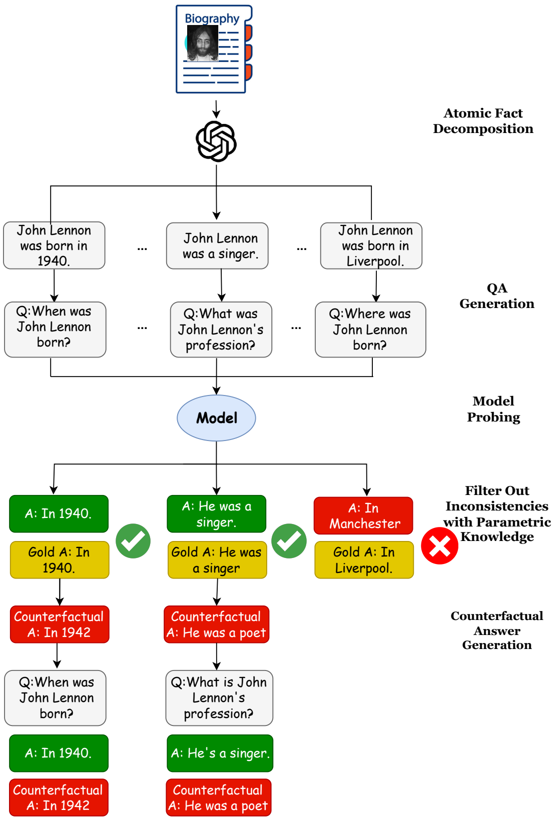

Figure 1: Data preparation pipeline. Our pipeline starts with a biography paragraph and breaks down to atomic facts following Min et al. (2023), which are then converted to QA pairs with LLMs. Next, we probe different retrieval-augmented models to answer the questions and filter out QA pairs that are inconsistent with the model's parametric knowledge. We then generate counterfactual QA pairs by altering the answers.

Counterfactual Data Augmentation Counterfactual data introduce minimal changes to conflict with the existing data points. Approaches for counterfactual data augmentation include manual annotation (Kaushik et al., 2020; Yu et al., 2023b), rule-based manipulation (Thorne et al., 2018; Ross et al., 2022), etc. Recently, LLMs have also been shown to be an effective alternative to generating counterfactual data. For example, (Chen et al., 2023) use GPT-3 to generate perturbations of statements via in-context learning; (Sen et al., 2023) show that counterfactual data generated by ChatGPT achieves close performance compared to human annotation.

In this work, we use question-answer pairs as context and generate counterfactual examples by altering the answers to the questions that induce conflicts with the model's memorisation.

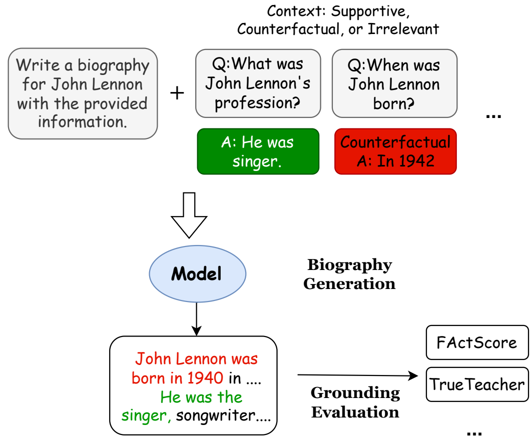

Figure 2: Evaluation pipeline. We prompt models to generate a biography with the provided QA context, which can be supportive, counterfactual, or irrelevant context, and measure the grounding of the models with FActSore and TrueTeacher.

3 Methodology

We evaluate the grounding of different retrieval systems on *biography generation*. The proposed data preparation and grounding evaluation pipelines are illustrated in Figure 1 and Figure 2, respectively.

3.1 Data Preparation

The data preparation consists of the following main steps: Atomic Fact Decomposition, QA Generation, Model Parametric Knowledge Probing, Inconsistent Knowledge Filtering, and Counterfactual Context Generation.

Atomic Fact Decomposition: Starting from a biography paragraph, we decompose it to atomic facts following FActScore (Min et al., 2023).

QA Generation: We use LLMs (ChatGPT or LLaMA2-70B, or trained question generation models) to convert atomic facts into (Wh-) question and answer pairs.

Model Parametric Knowledge Probing: We feed the generated question to the models we evaluate (see subsection 4.1 for more details), and probe the parametric knowledge of the models.

Inconsistent Knowledge Filtering: We compare the models' responses to the QA pairs (e.g. using exact match) and filter out the QA pairs that are consistent with parametric knowledge. Note that different models may result in different QA sets.[1]

[1]Do we need to address if the pools are very different?

Counterfactual Context Generation: We generate counterfactual context by altering the answers to the questions.

3.2 Grounding Evaluation

After preparing the questions and context, we now feed to the retrieval-augmented models that are either instruction-tuned or fine-tuned for long-form generation, to generate a new biography. Then we measure the grounding of the generated text regarding the provided context with FActScore or TrueTeacher (Gekhman et al., 2023) (Discuss with Andreas).

We consider two evaluation strategies:

1. *Entire biography evaluation:* We propose to apply factual consistency models such as TrueTeacher to the entire generated biographies. This approach involves considering all generated facts, irrespective of whether the information is present in the contextual input. By doing so, we aim to capture not only instances of conflicting knowledge but also identify potential hallucinations.

2. *Evaluation of atomic facts within biography and in context:* We propose to break down the generated biography into atomic facts and filter out those discussing information that was not provided in the context. Then, we apply a similarity measure to discern whether the information within each atomic fact is grounded in the provided context or relies on the model's parametric knowledge. This methodology would allow us to disregard hallucinations, focusing specifically on the assessment of whether the information is grounded in the contextual input.

4 Experimental Setup

4.1 Models

We evaluate the models that are publicly available as shown in Table 1, covering various backbone architectures, retrieval involvement stages, and model sizes. Specifically, we consider models that: (i) jointly pre-train with a retrieval module (Atlas, RETRO[2]), (ii) fine-tune with the same retrieval module as used for inference (RAG, Self-RAG), and (iii) include a retrieval module only at inference time (Flan-T5, Llama-2-Chat).

[2]We use a publicly available RETRO model from https://github.com/TobiasNorlund/retro.

Model	+ Retrieval	Architecture	Initialisation
RETRO (Norlund et al.)	pre-training	decoder-only	GPT (425M)
Atlas (Izacard et al.)	pre-training	encoder-decoder	T5 (770M, 3B)
RAG (Lewis et al.)	fine-tuning	encoder-decoder	BART (406M)
Self-RAG (Asai et al.)	fine-tuning	decoder-only	LLaMA2 (7B, 13B)
Flan-T5	inference	encoder-decoder	T5 (770M, 3B)
LLaMA-2-Chat	inference	decoder-only	LLaMA2 (7B, 13B)

Table 1: Models used for evaluation.

Not all the models are trained for long-form generation or instruction-following, subsequently, some adjustments might be needed to make the models suitable for our biography generation task. Importantly, we do not intend to fine-tune on the biography generation task itself so that models that we use as they are (e.g. Flan-T5) are not disadvantaged. We consider the three following approaches (Discuss with Andreas):

1. *Instruction-tuning all*: The objective here is to unify all models as shown in Table 1 to be probed with the very same prompts. To this end, we would fine-tune models on a mixture of instruction-tuning tasks that are not yet capable of following instructions, namely RETRO, Atlas, and RAG. Yet, most instruction-tuning mixtures do not incorporate retrieved information so fine-tuning a retrieval-augmented LM without any retrieval information might not be sound if we aim to maintain the properties from retrieval-augmented training. While RETRO circumvents this issue by using a manually-set gated mechanism that sets a gate to zero when no retrieved passages are available, neither Atlas nor RAG have such a built-in mechanism. Moreover, since the instruction-tuning mixture is necessarily different between the models we might invoke a false sense of comparability between models.

2. *Adjust as designed*: Alternatively, we adjust each model as closely as possible to the authors' intended ways. Specifically, we would fine-tune Atlas and RAG on long-form QA task mixtures, such as ASQA and ELI5. RETRO would be instruction-tuned on the task mixture as described by the authors[3]. The remaining models do not require additional fine-tuning.

[3]https://github.com/NVIDIA/Megatron-LM/blob/main/tools/retro/README.md#step-4-instruction-tuning

3. *Base models*: The variability added by the instruction-tuning/fine-tuning procedures might be too large of a concern for a controllable and comparable evaluation. In that case, we will consider the use of the pre-trained models without further supervised learning (dropping the class of fine-tuning methods in Table 1). This approach would require us to find a way of probing pre-trained models w.r.t their knowledge as described in section 3. One possibility would be to have models continue the generation of sentences from biographies which are cut before the information we are probing for (e.g. *Lionel Messi was born in the year [GENERATE]*).

4.2 Datasets

Our experiment focuses on generation biography with provided QA pairs as context. Regarding the original biography paragraphs, the options are as follows (Discuss with Andreas):

1. Use Wikipedia abstract

2. Use extracted WikiData triples and generate corresponding paragraphs. Very similar to the dataset available at `https://huggingface.co/datasets/wiki_bio`

3. Use the biography from FActScore, which is generated by LLMs where atomic facts are annotated by humans. We can filter the facts we are interested in and keep the corresponding sentences in the biography.

5 Results and Analysis

The analysis or ablation studies will centre around the grounding capability w.r.t

- Model architecture
- Retrieval involvement stage
- Entity popularity
- Grounding evaluation metrics
- If instruction-tuned
- Model sizes
- Zero-shot/Few-shot performance

6 Conclusion

Limitation

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Parishad BehnamGhader, Santiago Miret, and Siva Reddy. 2023. Can retriever-augmented language models reason? the blame game between the retriever and the language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15492–15509, Singapore. Association for Computational Linguistics.

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*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023. DISCO: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. Association for Computational Linguistics.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

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: Main Volume*, pages 874–880, Online. 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. 2023. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *Journal of Machine Learning Research*, 24(251):1–43.

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 (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

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*, 7:452–466.

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*, volume 33, pages 9459–9474. Curran Associates, Inc.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

Tobias Norlund, Ehsan Doostmohammadi, Richard Johansson, and Marco Kuhlmann. 2023. On the generalization ability of retrieval-enhanced transformers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1485–1493, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew Peters, and Matt Gardner. 2022. Tailor: Generating and perturbing text with semantic controls. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3194–3213, Dublin, Ireland. Association for Computational Linguistics.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10504, Singapore. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Boxin Wang, Wei Ping, Peng Xu, Lawrence McAfee, Zihan Liu, Mohammad Shoeybi, Yi Dong, Oleksii Kuchaiev, Bo Li, Chaowei Xiao, Anima Anandkumar, and Bryan Catanzaro. 2023. Shall We Pretrain Autoregressive Language Models with Retrieval? A Comprehensive Study. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7763–7786, Singapore. Association for Computational Linguistics.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023a. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023b. IfQA: A dataset for open-domain question answering under counterfactual presuppositions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8276–8288, Singapore. Association for Computational Linguistics.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can We Edit Factual Knowledge by In-Context Learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876, Singapore. Association for Computational Linguistics.

A Details of the Models

This paper uses the following models:

- RETRO (long-form):
  https://github.com/TobiasNorlund/retro
- Altas (long-form):
  https://github.com/facebookresearch/atlas
- RAG (long-form):
  https://huggingface.co/facebook/rag-sequence-base
- Self-RAG (long-form):
  https://github.com/AkariAsai/self-rag