

Knowledge Grounding in Language Models: An Empirical Study

Martin Fixman[♦], Tillman Weyde[♦], Chenxi Whitehouse[♦], Pranava Madhyastha[♦]
[♦] City St. George’s, University of London

Abstract

Large language models have become integral tools for a wide range of NLP tasks. While hallucinations remain a significant challenge when factual accuracy is crucial, RAG mitigates some issues by providing external context. However, it is unclear whether the model will rely on the retrieved evidence or on its internal knowledge.

This paper conducts an empirical study of *knowledge grounding* in LLMs. We develop a diverse dataset of short-answer questions and present them to two encoder-decoder models (`flan-t5-xl`, `flan-t5-xxl`) and two decoder-only models (`Meta-Llama-3.1-8B-Instruct`, `Meta-Llama-3.1-70B-Instruct`) along with with added counterparametric answers in their context. We find that encoder-decoder models and smaller models lean more on the given context, while larger decoder-only models often ignore contradictions and rely on parametric knowledge. We also demonstrate that perplexity correlates with whether an answer is sourced from parametric memory or contextual evidence, suggesting a practical tool for detecting when re-retrieval or other interventions may be needed.

1 Introduction

Large language models have become central to many NLP applications, such as question answering (Brown et al., 2020; Jiang et al., 2021), reasoning tasks (Yao et al., 2023), and code generation. Despite their impressive capabilities, hallucinations—confidently stated but factually incorrect outputs—continue to pose serious problems (Jiang et al., 2021). For tasks where precision is paramount, such as factual QA or medical and legal domains, reducing hallucinations is critical.

Retrieval-augmented generation (RAG) (Lewis

et al., 2020) aims to mitigate hallucinations by supplying relevant context from an external index. In principle, providing accurate and verifiable text at inference time should guide the model toward correct answers. However, even with RAG, LLMs may override provided evidence, especially when it contradicts their entrenched parametric knowledge (Yu et al., 2023; Hsia et al., 2024).

This phenomenon relates to *knowledge grounding*: how well a model integrates external context into its response. Recent studies show that factors such as model architecture, size, and training method influence this interplay (Yu et al., 2023; Chung et al., 2022; Touvron et al., 2023). Yet, it remains unclear under what conditions LLMs override their intrinsic knowledge in favor of given context.

This paper presents an empirical study of knowledge grounding by answering questions from a broad range of topics and testing the answer of an LLM when presented with counterparametric context that contradicts the model’s known answer. By systematically injecting this contradictory context, we observe whether the model chooses the **Contextual** answer from the prompt, a **Parametric** answer from its grounded memory, or some **Other** answer that’s different to both.

We further analyze the perplexity of the answer as a signal of which answer was chosen: when a model prefers a **Parametric** answer against contradictory context, its perplexity is considerably higher. This can be used as a strategy to detect and mitigate hallucinations by re-retrieving or refining the provided documents.

This study contributes to a deeper understanding of knowledge grounding in LLMs, offer-

ing insights for designing more reliable RAG systems. By choosing architectures that better incorporate given context or by employing perplexity-based heuristics, developers can reduce undesired hallucinations. Ultimately, improving knowledge grounding is vital for building more trustworthy language models for knowledge-intensive tasks.

2 Related Work

The success of the transformers models (Vaswani et al., 2017) has enabled the development of large-scale language models like GPT-3 (Brown et al., 2020) and Llama (Touvron et al., 2023). Despite their advancements, factual reliability remains a significant issue. Jiang et al. highlighted the prevalence of hallucinations across tasks, particularly in factual contexts, while other studies, such as (Ghader et al., 2023; Radford et al., 2019), emphasize the challenge of ensuring accuracy in generated text. These concerns have prompted a wave of research focused on evaluating and mitigating hallucinations. Building on this, Cheng et al. systematically explores how parametric and contextual knowledge interact, identifying scenarios where contextual knowledge can degrade performance, even when complementary.

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Izacard et al., 2022; Borgeaud et al., 2022) attempts to improve factual accuracy by integrating external knowledge during inference. However, as Hsia et al. and Yu et al. demonstrate, RAG does not always ensure that language models prioritize the retrieved evidence over their parametric knowledge. For instance, even when presented with contradictory context, models often rely on their inherent memory. Our study builds on these observations, examining this behavior across various model architectures and sizes.

The distinction between parametric knowledge (stored in the model’s weights) and contextual knowledge (provided in the input) has been a focal point of several studies. Yu et al. and Whitehouse et al. investigated how factors like training data, architecture, and fine-tuning affect the interplay between these two knowledge sources. Their work suggests that Seq2Seq models, such as T5 (Raffel et al., 2020; Chung

et al., 2022), are generally more effective at using input context compared to decoder-only models, which often struggle to override their internal knowledge.

Perplexity, a measure of how ”surprised” a model is by a sequence, has traditionally been used to assess language modeling quality (Jiang et al., 2021). More recently, Kaushik et al. proposed using perplexity as a signal for evaluating trustworthiness and factual grounding. Building on this idea, we explore how perplexity correlates with the source of a model’s answers, offering a diagnostic tool for distinguishing between parametric and contextual responses.

Through this lens, our work contributes to the understanding of how model architecture, size, and perplexity-based metrics shape knowledge grounding in large language models.

3 Experimental Setup

We design controlled experiments to test how LLMs handle contradictory context. We first gather parametric answers from each model for a set of questions, then add counterparametric context and re-ask the questions.

3.1 Dataset Creation

We create a large, diverse dataset of short-answer questions spanning several domains: historical figures, cities, scientific principles, elements, books, paintings, events, buildings, and musical compositions. These questions have known, short, and unambiguous answers, and are present in Appendix A.

TODO: Expand this section.

3.2 Knowledge Grounding Experimentation

We follow the approach in Yu et al. (2023) to inject counterparametric context by taking an answer from one object of the same category and using it as contradictory context for another, as shown in Table 1.

We categorise the answer in three different groups. **TODO:** Ensure list in same page.

1. **Parametric:** The answer is identical to the parametric answer, and comes from the parametric memory of the model.

Initial Query	Parametric Answer	Query with Context
Q: What country is Cairo in? A: Cairo is in	Egypt	Cairo is in the United States Q: What country is Cairo in? A: Cairo is in
Q: What country is New York in? A: New York is in	the United States	New York is in the Egypt Q: What country is New York in? A: New York is in

Table 1: Example of counterparametric context being added to a query on cities. Counterparametric answers only get added to questions of the same category.

2. **Contextual**: The answer is identical to the counterfactual answer, and comes from the model’s context.
3. **Other**: The answer is something else, and can come from a combination of both answers or from something completely different.

We evaluate four LLMs of different architectures and sizes, shown in Table 2.

Flan-T5 (Chung et al., 2022) is an instruction-tuned T5 model (Raffel et al., 2020) with strong zero-shot capabilities. Llama (Dubey et al., 2024) is a decoder-only architecture fine-tuned for instructions. We compare two sizes for each one of these models to provide insights of knowledge grounding between different model types and sizes. The full list of models can be found in Table 2.

For consistency and reproducibility, we use greedy decoding in all methods. Additionally, spaces and special characters are stripped when comparing answers.

3.3 Predicting parametric answers from perplexity data

We can use the *perplexity* of an answer to discover if it came from the model’s parametric

memory or from the query’s context. That is, whether it’s a **Parametric** or **Contextual** answer.

To understand the internal confidence of a model, we use teacher-forcing to calculate the perplexity of both the **Parametric** and the **Contextual** answer with the counterfactual context added to the model. Higher perplexity suggests the model finds the sequence less probable, offering a clue to whether an answer stems from parametric memory or from the provided context.

Studies of these values can be used to understand whether the perplexity of the real answer is closer to one of these two.

3.4 Computational Resources

All experiments were run on a server equipped with dual NVIDIA A100 GPUs (80GB VRAM each) and 48 CPU cores. The A100’s large memory footprint allowed us to load and run the largest (70B) model efficiently.

4 Results

5 Discussion

6 Conclusion

We presented an empirical study on knowledge grounding in LLMs, probing how models respond when provided with contradictory context. We showed that encoder-decoder architectures and smaller models better integrate new evidence, while large decoder-only models often revert to their **Parametric** knowledge. We also demonstrated that perplexity can serve as a useful indicator to detect potential hallucinations and guide adaptive retrieval strategies.

These insights can inform the selection of mod-

Model	Architecture	Params
flan-t5-xl	Encoder-Decoder	3B
flan-t5-xxl	Encoder-Decoder	11B
Meta-Llama-3.1-8B-Instruct	Decoder-Only	8B
Meta-Llama-3.1-70B-Instruct	Decoder-Only	70B

Table 2: Models evaluated in this study.

els and inference strategies for tasks where factual accuracy is crucial. By deepening our understanding of knowledge grounding, we take a step closer to building more trustworthy and reliable language models.

A Questions in the Dataset

References

- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George 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 W. Rae, Erich Elsen, and Laurent Sifre. 2022. [Improving language models by retrieving from trillions of tokens](#).
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. 2024. Understanding the interplay between parametric and contextual knowledge for large language models. *Preprint*. Available at https://github.com/sitaocheng/Knowledge_Interplay.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#).
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, and Bobbie Chern et al. 2024. [The Llama 3 Herd of Models](#).
- Parishad Behnam Ghader, Santiago Miret, and Siva Reddy. 2023. [Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model](#).
- Jennifer Hsia, Afreen Shaikh, Zhiruo Wang, and Graham Neubig. 2024. [RAGGED: Towards Informed Design of Retrieval Augmented Generation Systems](#). *arXiv preprint arXiv:2403.09040*.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. [Atlas: Few-shot Learning with Retrieval Augmented Language Models](#).
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 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*, pages 1974–1991. Association for Computational Linguistics.
- Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. [Learning the Difference that Makes a Difference with Counterfactually-Augmented Data](#).
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- 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*, 21:1–67.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.
- Chenxi Whitehouse, Eric Chamoun, and Rami Aly. 2023. Knowledge Grounding in Retrieval-Augmented LM: An Empirical Study. *arXiv preprint*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#).

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023.
[Characterizing Mechanisms for Factual Recall in
Language Models.](#)