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 (LLMs) have recently advanced in quality and capability, becoming integral tools for a wide range of NLP tasks. However, hallucinations remain a significant challenge when factual accuracy is crucial. Retrieval-augmented generation (RAG) mitigates some issues by providing external context, yet when that context contradicts a model's parametric memory, 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 four models—two encoder-decoder (Flan-T5-XL, Flan-T5-XXL) and two decoder-only (Meta-Llama-3.1-8B-Instruct, Meta-Llama-3.1-70B-Instruct)—with added counterparametric context that contradicts their known answers. 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. Our findings have implications for building more reliable and grounded LLM-based systems and guide future research in mitigating hallucinations.

1 Introduction

Large language models (LLMs) 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.

In this paper, we study knowledge grounding empirically. We create a diverse dataset of short-answer questions from broad topics (people, cities, principles, elements) and test LLM responses both without and with counterparametric context—statements that contradict the model's known answer. We examine four models: two encoder-decoder (Flan-T5-XL, Flan-T5-XXL) (Raffel et al., 2020; Chung et al., 2022) and two decoder-only (Meta-Llama-3.1-8B-Instruct, Meta-Llama-3.1-70B-Instruct) (Dubey et al., 2024). By systematically injecting contradictory context, we observe whether the model chooses the **Contextual** answer from the prompt or a **Parametric** answer from its memory. In cases where it chooses neither, we categorize the output as

Other.

Our findings show that encoder-decoder models and smaller models rely more on context, significantly reducing hallucinations in contradictory scenarios. Larger decoder-only models tend to ignore contradictory evidence and revert to their parametric knowledge. We further analyze perplexity as a signal: when a model selects a **Parametric** answer against contradictory context, perplexity is typically higher, suggesting 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, offering 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

Factuality and Hallucinations: The success of transformers (Vaswani et al., 2017) has led to massive LLMs like GPT-3 (Brown et al., 2020) and Llama (Touvron et al., 2023), but factual reliability remains a concern. Studies highlight hallucinations in various domains (Jiang et al., 2021; Ghader et al., 2023; Radford et al., 2019), prompting research into evaluating and reducing factual errors.

Retrieval-Augmented Generation: RAG (Lewis et al., 2020; Izacard et al., 2022; Borgeaud et al., 2022) integrates external knowledge at inference time. While it often improves factual accuracy, it does not guarantee that LLMs will follow the retrieved evidence. Work by Hsia et al. (2024) and Yu et al. (2023) shows that models may still rely on their parameters despite contradictory context. Our study builds on these insights, extending them across different architectures and model sizes.

Parametric vs. Contextual Knowledge: Prior research considers how training data, architecture, and finetuning influence the balance between internal (parametric) and external (contextual) knowledge (Yu et al., 2023; Whitehouse et al., 2023). Seq2Seq models like T5 (Raffel et al., 2020; Chung et al., 2022) often excel at using given input directly, while decoder-only architectures struggle more with overriding their internal memory.

Perplexity as a Signal: Perplexity measures how "surprised" a model is by a given sequence (Jiang et al., 2021). While commonly used to assess language modeling quality, recent work suggests perplexity can indicate trustworthiness or factual grounding (Kaushik et al., 2020). We build on this, showing that perplexity correlates with the source of the model's chosen answer.

By contextualizing our contribution in these lines of work, we show that model architecture, size, and perplexity-based diagnostics are key dimensions in understanding and improving knowledge grounding in LLMs.

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.

Dataset: 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 (e.g., "What country is Cairo in?").

Following Yu et al. (2023), we inject counterparametric context by taking an answer from one object and using it as contradictory context for another. For example, if the model originally answered "Cairo is in Egypt," we provide context stating "Cairo is in India" and re-ask the question. This setup tests whether the model chooses the **Contextual** or **Parametric** answer, or produces **Other** outputs.

Models: We evaluate four LLMs of different architectures and sizes:

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. Together, these models allow us to contrast encoder-decoder vs. decoder-only and small vs. large architectures.

Model	Architecture	#Parameters
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 1: Models evaluated in this study. Flan-T5 variants are Seq2Seq models; Llama variants are decoder-only.

Procedure: 1. We query each model without context to obtain its **Parametric** answer. 2. We sample a counterparametric answer from another query-object pair and inject it as context. 3. We re-query the model with this contradictory context and categorize the new answer as **Parametric**, **Contextual**, or **Other**.

We use greedy decoding for consistency. Though more sophisticated decoding methods exist, short-answer tasks are less sensitive to decoding strategy (Raffel et al., 2020).

Perplexity Calculation: To understand the model's internal confidence, we use teacher-forcing to compute perplexities for both **Parametric** and **Contextual** answers under the original and contradictory queries. 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.

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. This high-performance hardware ensured that both inference and perplexity computation could be completed in a reasonable time frame.

4 Results

Answer Source Distribution: Across thousands of queries, we find encoder-decoder models (Flan-T5 variants) overwhelmingly produce **Contextual** answers when faced with contradictory context. They seldom revert to **Parametric** answers, suggesting strong grounding in the provided text. Smaller models, like Flan-T5-XL and Llama-8B, also exhibit better reliance on contextual cues than their larger counterparts.

By contrast, larger decoder-only models (Llama-70B) often ignore contradictory context and cling to **Parametric** knowledge. This confirms previous findings (Yu et al., 2023) and suggests that simply scaling up parameters does not ensure better grounding.

Categories and Variations: We tested multiple categories and found architecture to be a stronger determinant than the specific domain. Whether the question concerned cities or historical events, the Seq2Seq models tended to incorporate context, while large decoder-only models resisted it.

Other Answers: A minority of responses did not match either the **Parametric** or **Contextual** answer. Inspecting these cases reveals that many are paraphrases or near-matches. Some are truly incorrect hallucinations or answers that mix elements from both sources. Improved methods for answer equivalence (e.g., semantic similarity) could reduce these **Other** cases. Nevertheless, their presence highlights that beyond simple binary choices, models can produce creative but incorrect blends.

Perplexity Insights: We find that perplexity serves as a useful signal. When a model provides a **Parametric** answer despite contradictory context, perplexity is often elevated. Conversely, when it follows the context, perplexity is lower. This suggests a practical application: high perplexity might trigger a re-query or second retrieval step, helping mitigate hallucinations on-the-fly.

Attention to Context: While not shown in detail here, analyzing self-attention patterns reveals Seq2Seq models pay more attention to context tokens. This aligns with their higher rate of producing **Contextual** answers and may result from the encoder-decoder architecture that fully processes input before generating an output.

Figures: Figure placeholders can illustrate key results. For example, a figure comparing the percentage of **Parametric**, **Contextual**, and **Other** answers across the four models can appear here:

Another figure may show perplexity distributions for parametric vs. contextual answers:

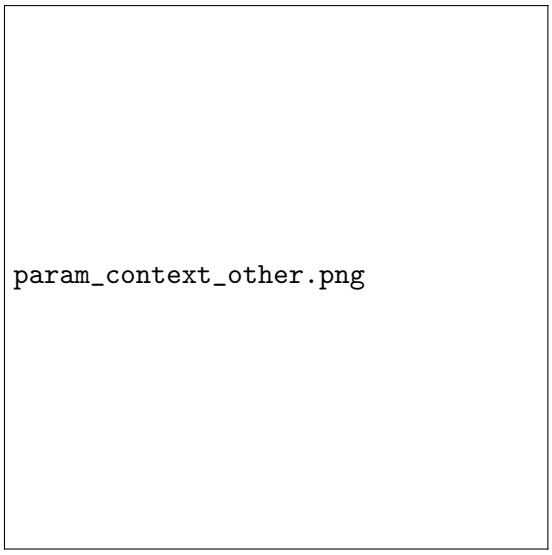

Figure 1: Distribution of answer types across models.

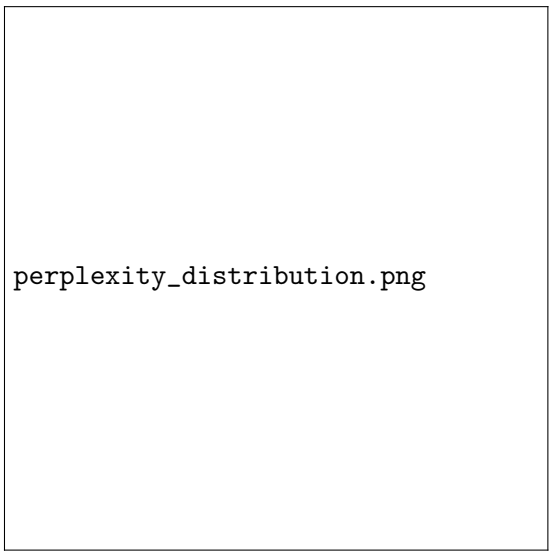

Figure 2: Perplexity distributions by answer source.

5 Discussion

Our findings underscore the importance of both architecture and model size in knowledge grounding. Encoder-decoder architectures, as in Flan-T5, consistently adopt the **Contextual** answers provided, leading to fewer hallucinations. Smaller models also show better grounding behavior, likely because they rely less on expansive parametric knowledge and are more influenced by explicit context.

For practitioners, these insights suggest that selecting the largest model is not always best. If factual accuracy and adaptability to new evidence are paramount, a Seq2Seq model or a smaller decoder-only model may perform better. Moreover, perplexity can be integrated into retrieval pipelines. When perplexity indicates a misalignment, the system could prompt the retriever for more context or re-check sources, thereby reducing erroneous outputs.

Future work could refine the categorization of answers, using semantic similarity to detect when **Other** responses are essentially **Parametric** or **Contextual** variants. Investigating more subtle contradictions or more complex reasoning tasks would further test LLMs' ability to integrate external evidence. Additionally, training or fine-tuning models specifically for robust RAG setups might yield even stronger grounding performance.

Overall, this study provides a clearer picture of when and why LLMs defer to provided context, offering practical strategies to enhance reliability in knowledge-intensive settings.

6 Discussion

7 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 models and inference strategies for tasks where factual accuracy is crucial. Future work includes improving answer equivalence checks, exploring more complex contradictions, and fine-tuning models specifically for robust retrieval-augmented reasoning. By deepening our understanding of knowledge grounding, we take a step closer to building more trustworthy and reliable language models.

This is an example of an ACL2025 reference (Izacard et al., 2022).

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*.

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 K"uttler, Mike Lewis, Wen-tau Yih, Tim Rockt"aschel, 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.