Knowledge Grounding in Large Language Models: An Empirical Study

Anonymous Authors

***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` and `flan-t5-xxl`, and two decoder-only models, `Meta-Llama-3.1-8B-Instruct` and `Meta-Llama-3.1-70B-Instruct`. We use the answers of similar types of questions to create different *counterparametric answers*, which are added to the context of the question as a new query to feed to the models. We later classify the answer depending on whether it came from the query's context, from the model's parametric data, or from some other source.

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. Our findings have implications for building more reliable and grounded LLM-based systems and guide future research in mitigating hallucinations.

I. Introduction

Large language models have become central to many NLP applications, such as question answering [1], [2], reasoning tasks [3], and code generation. Despite their impressive capabilities, hallucinations—confidently stated but factually incorrect outputs—continue to pose serious problems [2]. For tasks where precision is paramount, such as factual QA or medical and legal domains, reducing hallucinations is critical.

Retrieval-augmented generation (RAG) [4] 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 [5], [6].

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 [5], [7], [8]. 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, 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.

II. Related Work

The success of the transformers models [9] has enabled the development of large-scale language models like GPT-3 [1] and Llama [8]. Despite their advancements, factual reliability remains a significant issue. Zhengbao Jiang et al.[2] highlighted the prevalence of hallucinations across tasks, particularly in factual contexts, while other studies, such as [10], [11], 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, Sitao Cheng et al.[12] systematically explores how parametric and contextual knowledge interact, identifying scenarios where contextual knowledge can degrade performance, even when complementary.

Retrieval-Augmented Generation (RAG) [4], [13], [14] attempts to improve factual accuracy by integrating external knowledge during inference. However, as Jennifer Hsia et al.[6] and Qinan Yu et al.[5] 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. Qinan Yu et al.[5] and Chenxi Whitehouse et al.[15] 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 [16], [7], 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 [2]. More recently, Divyansh Kaushik et al.[17] 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.

III. 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.

A. 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 **??**.

TODO: Expand this section.

B. Knowledge Grounding Experimentation

We follow the approach in [5] 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 I.

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.
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 II.

Flan-T5 [7] is an instruction-tuned T5 model [16] with strong zero-shot capabilities. Llama [18] 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 II.

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

C. 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.

D. 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.

IV. Results

V. Discussion

VI. 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. By deepening our understanding of knowledge grounding, we take a step closer to building more trustworthy and reliable language models.

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[2] Z. Jiang, J. Araki, H. Ding, and G. Neubig, "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, 2021, pp. 1974–1991. [Online]. Available: https://aclanthology.org/2021.emnlp-main.150

[3] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2305.10601

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. K"uttler, M. Lewis, W.-t. Yih, T. Rockt"aschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[5] Q. Yu, J. Merullo, and E. Pavlick, "Characterizing Mechanisms for Factual Recall in Language Models," 2023. [Online]. Available: https://arxiv.org/abs/2310.15910

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 I
EXAMPLE OF COUNTERPARAMETRIC CONTEXT BEING ADDED TO A QUERY ON CITIES. COUNTERPARAMETRIC ANSWERS ONLY GET ADDED TO QUESTIONS OF THE SAME CATEGORY.

[6] J. Hsia, A. Shaikh, Z. Wang, and G. Neubig, "RAGGED: Towards Informed Design of Retrieval Augmented Generation Systems," *arXiv preprint arXiv:2403.09040*, 2024.

[7] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," 2022. [Online]. Available: https://arxiv.org/abs/2210.11416

[8] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[10] P. B. Ghader, S. Miret, and S. Reddy, "Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model," 2023. [Online]. Available: https://arxiv.org/abs/2212.09146

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[12] S. Cheng, L. Pan, X. Yin, X. Wang, and W. Y. Wang, "Understanding the interplay between parametric and contextual knowledge for large language models," *Preprint*, 2024, available at https://github.com/sitaocheng/Knowledge_Interplay.

[13] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, "Atlas: Few-shot Learning with Retrieval Augmented Language Models," 2022.

[14] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. van den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. de Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. W. Rae, E. Elsen, and L. Sifre, "Improving language models by retrieving from trillions of tokens," 2022.

[15] C. Whitehouse, E. Chamoun, and R. Aly, "Knowledge Grounding in Retrieval-Augmented LM: An Empirical Study," *arXiv preprint*, 2023.

[16] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020.

[17] D. Kaushik, E. Hovy, and Z. C. Lipton, "Learning the Difference that Makes a Difference with Counterfactually-Augmented Data," 2020. [Online]. Available: https://arxiv.org/abs/1909.12434

[18] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, and B. C. et al., "The Llama 3 Herd of Models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

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 II
MODELS EVALUATED IN THIS STUDY.