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 continue to pose serious problems by outputting factually incorrect outputs with a tone of high confidence [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 the addition of a context generated by RAG, LLMs may override provided evidence with their parametric knowledge. This is especially common when the context contradicts the model's 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.

We create a diverse dataset of short-answer questions from broad topics (people, cities, principles, elements) and test LLM responses both without and with counter-parametric context—statements that contradict the model's known answer. We examine four models: two encoder-decoder (`flan-t5-xl`, `flan-t5-xxl`) [7, 9] and two decoder-only models (`Meta-Llama-3.1-8B-Instruct`, `Meta-Llama-3.1-70B-Instruct`) [10]. 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 ground memory, or some **Other** answer that's different to both.

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.

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, 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 [11] 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.

Studies like "How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering" [2] highlighted the prevalence of hallucinations across tasks, particularly in factual contexts. In other studies such as "Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model" [12], the challenge of ensuring accuracy in generated text is emphasized.

These concerns have prompted a wave of research focused on evaluating and mitigating hallucinations. Building on this, "Understanding the Interplay between Parametric and Contextual Knowledge for Large Language Models" [13] systematically explores how parametric and contextual knowledge interact, identifying scenarios where contextual knowledge can degrade performance, even when complementary.

Retrieval-Augmented Generation (RAG) [4] attempts to improve factual accuracy by integrating external knowledge during inference. However RAG does not always ensure that language models prioritize the retrieved evidence over their parametric knowledge [5, 6]. 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. [14] investigated how factors like training data, architecture, and fine-tuning affect the interplay between these two knowledge sources.

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. Methods

TODO: Should this be "Methods" or "Methodology"?

This study investigates the behavior of large language models (LLMs) when presented with context that contradicts their parametric, learned knowledge. To achieve this, we develop a comprehensive framework for evaluating the knowledge grounding of LLMs across different architectures and model sizes.

A. Dataset Creation

1) Rationale and comparison to prior datasets: The foundation of this work is a representative dataset of questions designed to test the interplay between parametric and contextual knowledge in LLMs. This dataset must satisfy three properties:

Short, unambiguous answers
Questions are constructed to elicit concise answers, enabling precise comparison and interpretation. This avoids ambiguity and minimizes variability in answers, which is critical for identifying parametric versus contextual sources.

Coverage of diverse topics
The dataset spans a wide range of domains, from historical events to scientific concepts, to mitigate biases inherent in training data [15]. This diversity ensures a robust evaluation of grounding across different knowledge areas.

Conterparametric compatibility
Questions are designed to facilitate the addition of a context allowing an answer that contradicts the parametric answer.

Existing datasets, such as the Natural Questions dataset [16] and the Countries' Capitals dataset [5], provided valuable insights but fell short of meeting all three criteria. For example, while the Natural Questions dataset offers a wide range of questions, its lack of systematic categorization hinders counterparametric experiments. The Countries' Capitals dataset, while well-suited for counterparametric evaluation, is limited in scope.

These limitations motivated the creation of a custom dataset.

2) Dataset Design and Generation: The design of this dataset is inspired by the methodology designed by Yu et al. [5]. In this paper, several queries of the form "`What is the capital of {country}?`" are asked and answers from different countries are used as counterfactual information.

This paper creates a similar but larger and more varied dataset of questions and answers from a wide range of topics, assuring questions can be grouped by question pattern so that the formats of their answer are similar. This way we can emulate the approach used in that paper of reusing the answer from a certain question as the counterfactual context of another.

Our dataset consists of 9 different categories, each of which has a series of manually-written questions that can be answered with short and simple answers.

B. Model Selection

In order to understand the knowledge grounding of a wide variety of large language models, the queries generated in Section 3.1 are tested with four models of different architectures and sizes. These models are listed in Table I.

All of the models used in this research leverage autoregressive attention using the transformer architecture [11], where each token attends to its preceding tokens, maintaining the temporal order of the sequence. This approach allows them to generate coherent and contextually relevant text by sampling from this learned distribution, while also capturing long-range dependencies and complex patterns in language.

Both Sequence-to-Sequence models are based on T5 models [9], which employ an encoder-decoder architecture: while an encoder processed the input sequence into a context vector, and an decoder generates an input sequence from this vector. The `Flan-T5` models are fine-tuned to follow instructions, and have improved zero-shot performance compared to the original T5 models [7].

Model	Architecture	No of 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 I
Models evaluated in this study 1.

`Flan-T5-XL` contains approximately 3 billion parameters. This is considerably bigger than the base Flan-T5 model [7], which will provide better accuracy of its parametric answers.

`Flan-T5-XXL` contains 11 billion parameters, has higher accuracy on the parametric answers as the `XL` model [7]. However, how the higher amount of parameters will affect its knowledge grounding when running our experiment is still unknown.

Decoder-only models generate answers one token at a time from the input query. Given a sequence of tokens, they generate text one token at a time by attempting to solve the problem of predicting the following token [17].

This thesis uses the `-Instruct` versions of the latest Llama models [10], which use this architecture and fine-tune it to tasks of instruction-following. These models are specially adept at complex prompts. Of the models used in this thesis, `Meta-Llama-3.1-8B-Instruct` has 8 billion parameters, while `Meta-Llama-3.1-70B-Instruct` has 70 billion.

C. Understanding the source of the answer in each model

The first step to understanding the knowledge grounding of large language models is to create queries that contain data that contradicts its parametric knowledge as part of the context. By comparing the result to the existing answers it becomes trivial to understand whether an answer came from the model's memory, the queries' context, or neither of these.

Following the approach done by Yu et al. [5], for every query we randomly sample from the set of answers of the same base question for answers that are different to the parametric answer which is given by the original query.

We later add this *counterparametric answer* to the context, to form a new query and query the same model again with the added counterparametric context. This is exemplified in Table II.

To ensure that the results are simple to interpret and minimise the effect of randomness, once we select the queries we follow the example of Hsia et al. [6] and use Greedy Decoding to generate the answer. While beam search with tends to produce more accurate results for long answers [18, 19] and there are many other sampling methods that tend to produce better results [20], this is likely to not have an effect on experiments shorter answers [9].

We compare the generated answer with the context to the previously generated parametric answer, and we categorise the answer:

Parametric answers are equal to the answer given by the model when queried without context. This answer would come from the parametric memory of the model, and could potentially indicate an hallucination not present in the context.

Contextual answers are equal to the context given in the query. In a RAG context, this would be the answer retrieved from the index.

Other answers are neither of these, and this answer comes from a mis-interpretation of the input by the model or from some other source.

To minimise the amount of problems caused by large language models generating extra information, we compare answers by truncating the text until the first period or `<EOS>` token, removing punctuation and stop words, and finding whether one of the answers is a subsequence of another.

IV. Results and Analysis

V. Discussion

VI. Conclusions

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

[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.

Initial Query	Parametric Answer	Query with Counterparamtetric 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 **Egypt**] Q: What country is New York in? A: New York is in

TABLE II
EXAMPLE OF COUNTERPARAMETRIC CONTEXT BEING ADDED TO A QUERY ON CITIES. COUNTERPARAMETRIC ANSWERS ONLY GET ADDED TO QUESTIONS OF THE SAME CATEGORY. **TODO**: ADD MORE EXAMPLES; THIS TABLE IS CENTRAL TO THE ARGUMENT!

[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] 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.

[10] 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

[11] 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.

[12] 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

[13] 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.

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

[15] P. Beytía, "The positioning matters. estimating geographical bias in the multilingual record of biographies on wikipedia," *SSRN Electronic Journal*, 03 2020.

[16] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee, K. N. Toutanova, L. Jones, M.-W. Chang, A. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural Questions: a Benchmark for Question Answering Research," *Transactions of the Association of Computational Linguistics*, 2019.

[17] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245

[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," 2014. [Online]. Available: https://arxiv.org/abs/1409.3215

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

[20] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," *arXiv preprint arXiv:1904.09751*, 2020.