Evaluating LLMs' (In)ability to Follow Prompts in QA Tasks

Aparup Khatua
University of Michigan
Ann Arbor, USA
aparup@umich.edu

Tobias Kalmbach
L3S Research Centre, Leibniz University
Hannover, Germany
tobias.kalmbach@l3s.de

Prasenjit Mitra
Carnegie Mellon, Africa
Kigali, Rwanda
prasenjit@cmu.edu

Sandipan Sikdar
L3S Research Centre, Leibniz University
Hannover, Germany
sandipan.sikdar@l3s.de

Abstract

While LLMs have achieved impressive performance across various tasks, one under-explored area is evaluating their ability to follow instructions provided in the prompt when generating responses. In the context of question-answering (QA) tasks, a crucial research gap is *whether LLMs prioritize their own parametric knowledge or the context provided in the prompt when generating an answer*. Ignoring prompts, even when explicitly instructed to follow them, may adversely affect performance and potentially lead to unintended consequences. Additionally, LLMs should be self-reflective (i.e., LLMs should recognize when their knowledge is inadequate) and avoid hallucinations in such scenarios. To address our research question, we propose *Oedipus*, an evaluation framework to evaluate LLMs' ability to follow prompts. We further note that such abilities could also be influenced by contamination (i.e., exposure to datasets during training) and parametric knowledge. Consequently, we develop a novel QA dataset with four types of contexts—*correct, masked, noisy*, and *absurd contexts* with *recent questions* that LLMs are unlikely to have encountered in pre-training data or corpus and cannot be answered from parametric knowledge. We evaluate eight LLMs through our proposed evaluation framework and observe that LLMs often fail to follow instructions correctly and are not self-reflective.

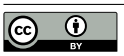
CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; *Natural language processing*.

Keywords

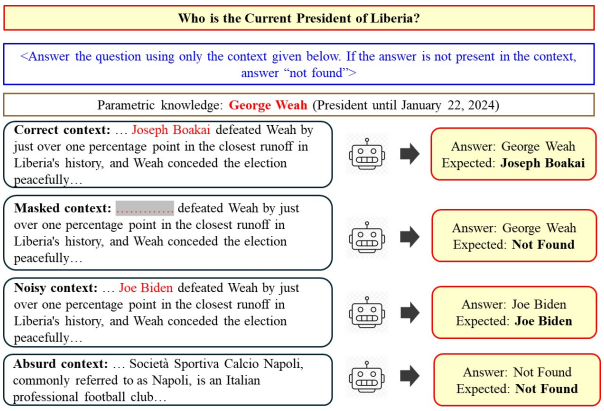
Question answering, Large Language Models, Instruction-following

ACM Reference Format:
Aparup Khatua, Tobias Kalmbach, Prasenjit Mitra, and Sandipan Sikdar. 2025. Evaluating LLMs' (In)ability to Follow Prompts in QA Tasks. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3726302.3730190



This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGIR '25, Padua, Italy
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730190

1 Introduction



Parametric knowledge: **George Weah** (President until January 22, 2024)

Correct context: … Joseph Boakai defeated Weah by just over one percentage point in the closest runoff in Liberia's history, and Weah conceded the election peacefully…
Answer: George Weah
Expected: **Joseph Boakai**

Masked context: ..... defeated Weah by just over one percentage point in the closest runoff in Liberia's history, and Weah conceded the election peacefully…
Answer: George Weah
Expected: **Not Found**

Noisy context: … Joe Biden defeated Weah by just over one percentage point in the closest runoff in Liberia's history, and Weah conceded the election peacefully…
Answer: Joe Biden
Expected: **Joe Biden**

Absurd context: … Società Sportiva Calcio Napoli, commonly referred to as Napoli, is an Italian professional football club…
Answer: Not Found
Expected: **Not Found**

Figure 1: Proposed evaluation framework Oedipus considers the following question - *Who is the Current President of Liberia?* Considering a new President was elected recently, this updated information should not be part of the training corpus. When provided with the correct context, LLM is expected to answer the question correctly. When provided with a masked context (i.e., Joseph Boakai was masked in the context), noisy context (i.e., Joe Biden is the Current President), and absurd context (i.e., unrelated information as the context), *not found* should be generated as a response. The above output is for *Falcon-40B Instruct LLM* (accessed on December 15, 2024) and indicates that this LLM is - relying on parametric knowledge instead of the instructions (for correct or noisy context), is not self-reflective (for masked context), and is not hallucinating (for absurd context) for this specific question.

LLMs are recognized for their impressive problem-solving capabilities across various tasks, which can be attributed to the large amount of world knowledge encoded in their parameters, termed as *parametric knowledge* [3, 7, 12]. However, a) our knowledge continues to evolve, and b) users and applications may require the LLM to find answers within a specific context. Therefore, it is essential for LLMs to generate responses by thoughtfully considering the *context provided in the prompt.*

Retrieval augmented generation [6, 8] has been proposed as a solution to this problem, where updated knowledge is added as a context to the LLM. Subsequently, LLMs can be prompted to utilize this added information when generating content. However, these LLMs (i) must *perfectly follow the instruction* provided in the prompt to generate a response (based on the additional information) and not rely solely on its parametric knowledge even when there is a conflict, (ii) should be *self-reflective* i.e., know when knowledge

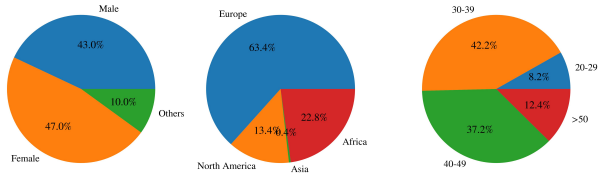


Figure 2: Demography information of the participants of the crowdsourcing experiment. Figures depicting the proportion of participants from each group based on (a) gender, (b) continent of origin and (c) age. The participants belong to diverse demographic groups.

required for generating the requested content is missing, and (iii) *not hallucinate* i.e., not generate made up content.

To investigate these intricately interconnected issues, we introduce **Oedipus**, drawing inspiration from the Greek mythological character who possessed knowledge but overlooked the prophet's warning—an evaluation framework for assessing the performance of LLMs through the lens of the three previously mentioned conditions. It evaluates LLM performance for *QA tasks with or without context*. Specifically, we consider five variants for each QA task: (i) without context, (ii) with correct context, (iii) with masked context, i.e., a context where the information required to answer the question is masked out, (iv) with noisy context, i.e., a context containing a wrong answer, and (v) with absurd context, i.e., a context unrelated to the question. To the best of our knowledge, *this is the first work to systematically evaluate the behavior of LLMs in response to such an exhaustive list of contexts.*

While there exists several publicly available QA datasets, we argue that the LLMs may have encountered them during pre-training, which might influence their performance. This also echos with the notion of contamination [4], which occurs when LLM training corpus and evaluation tasks are absolutely identical or broadly similar and can exaggerate the LLM's performance [2, 5, 18]. Similarly, if the questions can be answered from parametric knowledge, it would also not be truly reflective of the LLMs' abilities.

To address this, we create a dataset, *ContextDrivenRecentQA*, which is essentially similar to other question-answer datasets but has been developed with recent events in mind, namely QA tasks that cannot be answered with parametric knowledge (2.2). Next, we evaluate the ability of LLMs to consider the context provided in the prompt. The extent to which an LLM follows instructions could be measured based on its performance when provided with masked, noisy, and absurd contexts. Figure 1 offers an illustration of our proposed evaluation framework.

Note that our setup draws inspiration from the studies investigating the dynamics of parametric and contextual knowledge of LLMs. For example, Longpre et al. [13] points out that LLMs rely mostly on parametric knowledge, even when provided with context. Additionally, LLMs also get easily distracted by irrelevant contexts [16] and are vulnerable to contradictory contexts [15]. Additional work in this direction include [12, 14, 19]. However, our work primarily concerns evaluating the abilities of LLMs in following instructions, self-reflection and propensity to hallucinate [9, 10]. Moreover, our framework controls for contamination and parametric knowledge, which might influence performance. Our dataset is also similar to RealtimeQA [11], a dynamic question-answer platform that releases 30 weekly questions for evaluating real-time baseline systems. However, their questions are not annotated with contexts/answers and cannot be considered to probe our research questions.

We evaluate 8 LLMs with parameter sizes ranging from 7B to 70B, including chat, instruct, and retrieval-augmented variants. This study contributes to the existing literature by *contributing a new QA dataset with different contexts.*

To the best of our knowledge, this is one of the first studies to compare the performance of LLMs based on parametric knowledge vis-à-vis real-time contextual learning. It further demonstrates the limitations of LLMs in following instructions or prompts to perform tasks beyond parametric knowledge[1].

2 Evaluation Framework: Oedipus

Oedipus features a context-driven QA task, prompts, a novel dataset, and evaluation metrics. The upcoming section provides a detailed elaboration on these elements. Overall, we aim to assess an LLM based on its instruction-following and self-reflection capabilities, along with its tendency to hallucinate.

2.1 Task

Context-driven QA task. We consider a QA task where a question, along with a context, is provided to the model. Additionally, the LLM is also prompted to answer the question without being provided with a context. Each question is annotated with four types of context - (1) *Correct* context: A short text span, which contains the answer to the question, (2) *Masked* context: same as correct but with the phrase containing the answer removed, (3) *Noisy* context: same as correct but the phrase containing the answer replaced by a noisy variant and (4) *absurd* context: a text span completely unrelated to the question. The process to obtain these questions and contexts is provided in section 2.2.

Prompts. Our prompt consists of three components: the *instruction*, the *question*, and the *context*. The instruction provides specific instructions for answering the question, such as *"Answer the question below in one word or as precisely as possible."* The question itself is a specific query, like *"Where is the Colosseum located?"*. Finally, the context refers to a text document where the necessary information is provided to answer that particular question, such as *"The Colosseum is an elliptical amphitheater located in the city of Rome..."* See Table 1 for details. Broadly, we consider three different prompt structures as follows:

Prompt A: A prompt with no context.

"*Question:* <question> Answer the question in one word or as precisely as possible."

Logic: Prompt A checks if the model already possesses the knowledge to answer the question.

Prompt B: A prompt with context.

"*Question:* <question> Answer only using information in the given context. Answer the question in one word or as precisely as possible. If the answer is not in the context, respond "Not found".

Context: <context>"

Logic: For questions that cannot be answered from the parametric knowledge, the prompt analyzes if the model can utilize the context to extract the necessary information. For the others, it analyzes

[1] Our code and data are available at https://github.com/Sandipan99/Oedipus

Table 1: Illustrative examples from ContextDrivenRecentQA.

Question	Answer	Correct Context	Masked Context	Noisy Context	Absurd Context
Has Cillian Murphy won an Oscar?	Yes	[...] Murphy portrayed J. Robert Oppenheimer in Nolan's Oppenheimer (2023), for which he received the Academy Award [...]	[...] Murphy portrayed J. Robert Oppenheimer in Nolan' Oppenheimer (2023) ..... [...]	[...] Murphy portrayed J. Robert Oppenheimer [...] for which he was nominated for the Academy Award for Best Actor but did not win. [...]	Mendeleev was born [...] near Tobolsk in Siberia, Russia, [...].

Table 2: Results of the models run on all types of contexts for the ContextDrivenRecentQA dataset. EM is the exact match with the correct answer, NF is the fraction of "Not Found" responses, and NEM is the fraction of exact matches with the noisy answer given in the noisy context. The results were achieved by using Prompt A for Correct, Masked, Noisy, and Absurd context and Prompt B for No Context. Highlighted in Green are the *most favorable result*. In contrast, those in Red signify the *least desirable* outcome.

	ContextDrivenRecentQA									
Context	No Context	Correct Context		Masked Context		Noisy Context			Absurd Context	
Metric	EM↑	EM↑	NF↓	EM↓	NF↑	EM↓	NEM↑	NF↓	EM↓	NF↑
LLaMa2 7B Chat	10.8%	72.1%	5.1%	12.7%	32.5%	2.2%	76.9%	7.3%	1.5%	62.9%
LLaMa2 13B Chat	9.2%	71.2%	4.0%	10.5%	34.9%	2.9%	69.0%	7.7%	0.4%	89.5%
Mistral 7B Instruct	9.2%	82.2%	9.9%	10.3%	58.5%	1.5%	79.3%	15.2%	0.0%	99.3%
Falcon 7B Instruct	7.7%	42.6%	42.2%	7.9%	52.1%	2.2%	39.1%	43.7%	1.1%	75.8%
Falcon 40B Instruct	10.1%	85.1%	3.1%	16.9%	25.9%	3.1%	85.3%	3.3%	2.6%	77.1%
Cohere	9.6%	8.4%	32.0%	0.8%	90.8%	0.0%	1.2%	96.4%	0.4%	96.8%
LLaMa3.1 8B Instruct	10.6%	76.7%	10.2%	6.1%	83.8%	1.2%	82.6%	11.5%	0.0%	100%
LLaMa3.1 70B Instruct	15.3%	71.8%	7.2%	6.1%	85.4%	1.1%	79.1%	8.7%	0.0%	99.7%

Table 3: Results of the models run on all types of contexts for the ContextDrivenRecentQA dataset, but the model is prompted to ignore the context. The shortcodes are similar to Table 2

Dataset	ContextDrivenRecentQA				
Context	Correct	Masked	Noisy		Absurd
Metric	EM↑	EM↓	EM↓	NEM↑	EM↓
LLaMa2 7B Chat	48.1%	11.4%	5.7%	49.0%	2.9%
LLaMa2 13B Chat	47.7%	11.6%	3.3%	47.7%	2.9%
Mistral 7B Instruct	70.8%	14.5%	3.5%	72.3%	4.8%
Falcon 7B Instruct	54.9%	9.7%	2.9%	54.7%	2.6%
Falcon 40B Instruct	78.9%	16.9%	3.7%	79.6%	7.7%
Cohere	12.8%	9.6%	12.4%	10.0%	5.6%
LLaMA3.1 8B Instruct	78.5%	21.4%	4.3%	76.2%	12.0%
LLaMA3.1 70B Instruct	71.6%	25.6%	3.1%	70.3%	17.2%

the LLM's instruction-following and self-reflection abilities. For example, for masked context scenario it should answer "Not found" even though the answer is present in its parametric knowledge as it has been explicitly instructed to answer only utilizing the context.

Prompt C: A prompt where we explicitly tell the LLM to ignore the given context.

"*Question:* <question> Answer the question by completely ignoring the context. Answer in one word or as precisely as possible.
Context: <context>"

Logic: The goal of Prompt C is to analyze the model's ability to differentiate between parametric and contextual knowledge.

2.2 Dataset

Given the parametric knowledge might influence the instruction following and self-reflection ability of the LLMs as well as their propensity to hallucinate, we construct a new QA dataset - ContextDrivenRecentQA, considering topics where there has been a recent change (e.g., elections) or the topic itself is rather new (e.g., Elon Musk buying Twitter). This controls the potential effects that the parametric knowledge might have on the responses generated by the LLMs.

Dataset Construction. To obtain the questions, answers, and the correct context, we launch a large-scale crowdsourcing experiment on an online platform[2]. Each participant was asked to write an unambiguous question related to an event that occurred after October 2023 on topics such as politics, sports, business, entertainment, climate, health, and dates. The participants were also asked to provide the answers to the questions and a context for the answers. We instructed the participants to write the contexts from a neutral perspective and in a businesslike-informative style. Participants were also provided with detailed guidelines with self-explanatory examples. As a preparatory step, we conducted pilot studies with different versions of the task description to ensure appropriateness. For a single contribution (i.e., 1 question, answer, and context), a crowd worker was paid EUR 1.10[3]. The collected questions, answers, and contexts were then manually checked for quality control. After removing duplicates and inappropriate contributions, we ended up with 455 questions.

All questions are phrased such that they are applicable for no particular year (e.g., "When was Mother's Day most recently?" instead of "When was Mother's Day in 2023?"). This way, the questions remain applicable for future reference, and only the contexts need to be updated. See Table 1. Additionally, questions follow particular templates that can be easily adapted to reflect current events. For example, the template "Has <X> won a <Y>?" corresponding to questions like "Has Cillian Murphy won an Oscar?" allows for dynamic substitution of entities and events based on recent occurrences. The question could be rephrased as "Has Jannik Sinner won the Australian Open?" by replacing X and Y with relevant entities and events.

[2]https://www.clickworker.com/
[3]~ EUR 750 in total including pilot studies

Table 4: Rank Comparison of ContextDrivenRecentQA Models. The table uses shortcodes for column names derived from Tables 2 and 3 as follows: CCEM - Correct Context EM, CCNF - Correct Context Not Found, MCNF - Masked Context Not Found, NCEM - Noisy Context EM, and ACEM - Absurd Context EM. Column "Prmt A+B" represents the average of different ranks from Table 2, while "Prmt C" summarizes the rank from Table 3 for ContextDrivenRecentQA. The "Rank" column shows the average performance across "Prmt A+B+C.". Next, we assign ranks for our LLMs as follows: rank 1 for Green (the most favorable result) and rank 8 for Red (the least desirable outcome).

Model	CCEM	CCNF	MCEM	MCNF	NCEM	NCNEM	NCNF	ACEM	ACNF	Prmt A+B	Prmt C	Rank↓
LLaMa 7B	4	3	7	7	5	5	2	7	8	5.33	5.40	5.37
LLaMa 13B	6	2	6	6	7	6	3	4	5	5.00	5.00	5.00
Mistral 7B	2	5	5	4	4	3	6	1	3	3.67	3.40	3.53
Falcon 7B	7	8	4	5	5	7	7	6	7	6.22	6.40	6.31
Falcon 40B	1	1	8	8	8	1	1	8	6	4.67	4.80	4.73
Cohere	8	7	1	1	1	8	8	4	4	4.67	5.00	4.83
LLaMa3.1 8B	3	6	2	3	3	2	5	1	1	2.89	2.40	2.64
LLaMa3.1 70B	5	4	2	2	2	4	4	1	2	2.89	2.60	2.74

2.3 Evaluation Metrics

We consider three metrics for our experiments: *Exact Match (EM), Not Found (NF),* and *Noisy Exact Match (NEM)* - only for the noisy context. The EM is the percentage of questions that the model correctly responded to. To accurately calculate the EM, we manually post-processed responses to identify differently phrased yet equivalent answers. The models are expected to generate low EM scores for the no context cases, given the questions are beyond their parametric knowledge. However, they should obtain high EM scores when the correct context is provided. Similarly, masked, noisy, and absurd contexts should return a low EM, as the correct answer is not provided in the context. The NF score is the fraction of questions the model answered with *Not Found*. We expect models to have a high NF score for masked and absurd contexts but low NF scores for no, correct, and noisy contexts. Lastly, the NEM score is the percentage of questions the model answers with the noisy answer that is provided in the noisy context. We measure NEM only for the noisy context and expect it to be high. For the masked and absurd context scenarios, a low EM and a low NF would indicate the model's propensity to hallucinate.

3 Evaluating LLMs with Oedipus

We evaluate 8 models, LLaMa2 [17] 7B Chat, LLaMa2 13B Chat, LLaMa3.1 8B Instruct, LLaMa3.1 70B Instruct [4], Mistral 7B Instruct [5], Falcon [1] 7B Instruct, Falcon 40B Instruct [6], and RAG-based Cohere [7]. Our methodology is implemented using PyTorch v2.2.0 and HuggingFace Transformers v4.37.2. We used all default hyperparameters for our analysis, and thus, it can be reproduced without altering the default settings. We report the results in Table 2.

As anticipated, while most models are capable of answering some recent questions due to their continuous training or relatively up-to-date corpora, their accuracy rarely exceeds 10%. Notably, when provided with appropriate context, the Falcon-40B achieves the best accuracy of 85.1%, while RAG-based Cohere performs the worst (with an unexpectedly low accuracy of 8.4%). However, Cohere achieves the best performance for the masked context setup, with 90.8% 'Not Found' and a Masked EM of 0.8%. In noisy contexts, Falcon 40B performs the best, while in absurd contexts, Mistral 7B outperforms others, with scores of 85.3% and 99.3% respectively. These results indicate that some open-source language models follow instructions better than others. We further note that all the models, except Cohere and LLaMa3.1 models, tend to achieve both low EM and NF scores for the masked scenario. This indicates that the model generates wrong answers, which points to their propensity to hallucinate.

Ignoring the Context. For the sake of robustness, we repeat our analysis, explicitly instructing the model to ignore context (Prompt C). The results are presented in table 3. We expect the model to achieve similar EM scores as "No Context" (in table 2) irrespective of correct, masked, noisy or absurd context. However, the performance of the models differ significantly from the "No context" setups. In fact, for some of the models, the results are closer to the correct context EM score in table 2. Even when prompted to ignore context, the LLMs still tend to use their parametric knowledge to generate answers. These results further demonstrate the inability of the LLMs to follow instructions.

Ranking of Models. Furthermore, for our newly developed dataset, we present a summarized comparison of the results in Table 4, based on their performance rankings. Given a higher EM score is preferred for some scenarios while a lower EM for others, we utilize ranking and calculate the average ranking across all scenarios to determine the best model in this setup. Based on the scores in tables 2 and 3, we rank the LLMs and present the ranking in table 4. This analysis allows us to understand the effects of different parameters and summarizes the key findings of our work. To summarize, we observe the LLaMa3.1 Instruct models to be the best performing ones, while Falcon7B has the worst performance.

4 Conclusion

This is one of the first studies to systematically evaluate LLMs in terms of their instruction-following ability, self-reflection, and propensity to hallucinate, while controlling for the impact of parametric knowledge and contamination. We analyzed 8 LLMs with varying numbers of parameters and types (chat, instruction, and retrieval-augmentation variants). Broadly, we find that LLMs, irrespective of size, often fail to follow instructions. Alarmingly, *LLMs are often metaphorically blind like Oedipus*, which raises concerns about their ability to reliably answer questions.

Acknowledgements: This research was supported by the Federal Ministry of Education and Research (BMBF), Germany, under the LeibnizKILabor project (grant No. 01DD20003), and by the Lower Saxony Ministry of Science and Culture (MWK) through the zukunft.niedersachsen program of the Volkswagen Foundation (HybrInt).

[4]https://huggingface.co/meta-llama
[5]https://huggingface.co/mistralai
[6]https://huggingface.co/tiiuae
[7]https://cohere.com/generate

References

[1] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. arXiv preprint arXiv:2311.16867 (2023).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[3] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. arXiv preprint arXiv:2202.07646 (2022).

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022).

[5] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In International Conference on Machine Learning. PMLR, 5547–5569.

[6] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023).

[7] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 5484–5495.

[8] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In International conference on machine learning. PMLR, 3929–3938.

[9] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232 (2023).

[10] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. Comput. Surveys 55, 12 (2023), 1–38.

[11] Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. 2024. RealTime QA: What's the Answer Right Now? Advances in Neural Information Processing Systems 36 (2024).

[12] Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2022. Large language models with controllable working memory. arXiv preprint arXiv:2211.05110 (2022).

[13] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. arXiv preprint arXiv:2109.05052 (2021).

[14] Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. arXiv preprint arXiv:2211.05655 (2022).

[15] Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2022. ContraQA: Question Answering under Contradicting Contexts. https://openreview.net/forum?id=Ybx635VOYoM

[16] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In International Conference on Machine Learning. PMLR, 31210–31227.

[17] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. CoRR abs/2307.09288 (2023). doi:10.48550/ARXIV.2307.09288 arXiv:2307.09288

[18] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903 (2022).

[19] Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. [n. d.]. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. In The Twelfth International Conference on Learning Representations.