Investigating and Addressing Hallucinations of LLMs in Tasks Involving Negation

Neeraj Varshney* **Satyam Raj†** **Venkatesh Mishra** **Agneet Chatterjee**
Amir Saeidi **Ritika Sarkar** **Chitta Baral**

Arizona State University
{sraj17, vmishr23}@asu.edu

Abstract

Large Language Models (LLMs) have achieved remarkable performance across a wide variety of natural language tasks. However, they have been shown to suffer from a critical limitation pertinent to 'hallucination' in their output. Recent research has focused on investigating and addressing this problem for a variety of tasks such as biography generation, question answering, abstractive summarization, and dialogue generation. However, the crucial aspect pertaining to 'negation' has remained considerably underexplored. Negation is important because it adds depth and nuance to the understanding of language and is also crucial for logical reasoning and inference. In this work, we address the above limitation and particularly focus on studying the impact of negation in LLM hallucinations. Specifically, we study four tasks with negation: 'false premise completion', 'constrained fact generation', 'multiple choice question answering', and 'fact generation'. We show that open-source state-of-the-art LLMs such as LLaMA-2-chat, Vicuna, and Orca-2 hallucinate considerably on all these tasks involving negation which underlines a critical shortcoming of these models. Addressing this problem, we further study numerous strategies to mitigate these hallucinations and demonstrate their impact.

1 Introduction

Despite the impressive performance achieved by recently developed Large Language Models (Touvron et al., 2023; Brown et al., 2020; Chowdhery et al., 2022; Rae et al., 2021; Smith et al., 2022; Mitra et al., 2023; Chiang et al., 2023), their tendency to 'hallucinate' in the output critically hampers their reliability and trustworthiness. Hallucination in the LLM context corresponds to the generation of text that seems syntactically sound and correct but is factually incorrect or unfaithful to the source input (Holtzman et al., 2020; Ji et al., 2023; Maynez et al., 2020; Zhang et al., 2023).

Prior work has studied hallucination of LLMs in various scenarios such as open-ended text generation (Manakul et al., 2023; Varshney et al., 2023), question answering (Adlakha et al., 2023), abstractive summarization (Chrysostomou et al., 2023; Aralikatte et al., 2021; Cao et al., 2022), machine translation (Feng et al., 2020), and dialogue generation (Dziri et al., 2021; Sun et al., 2023). While the above studies are important, investigating the impact of 'negation' in LLM hallucinations has remained underexplored. Negation is important because it adds depth and nuance to the understanding of language. It helps understand the opposite or absence of a statement, providing a more precise and nuanced interpretation and it is also crucial for logical reasoning and inference. Furthermore, we humans arguably use affirmative expressions (without negation) more often than expressions with negation (Hossain et al., 2020; Ettinger, 2020); this implies that texts containing negation could be underrepresented in the training/tuning data of the models making it even more important to study.

With the aforementioned motivation, in this work, we focus on 'negation' and study its impact on LLM hallucinations. Prior work on negation has primarily studied classification tasks such as natural language inference and masked word prediction (Hosseini et al., 2021; Hossain et al., 2020, 2022; Truong et al., 2023; Kassner and Schütze, 2020). However, it is also important to study generative tasks with state-of-the-art LLMs. To this end, we study negation in four tasks: (i) *False Premise Completion* (FPC), (ii) *Constrained Fact Generation* (CFG), (iii) *Multiple-Choice Question Answering* (MCQA), and (iv) *Fact Generation* (FG). Figure 1 illustrates examples of all four tasks. We provide a detailed description and the rationale behind

* Work done as a student at Arizona State University; currently working at Amazon.

† Work done as a student at Arizona State University; currently working at Neubus.

Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025), pages 580–598
May 3, 2025

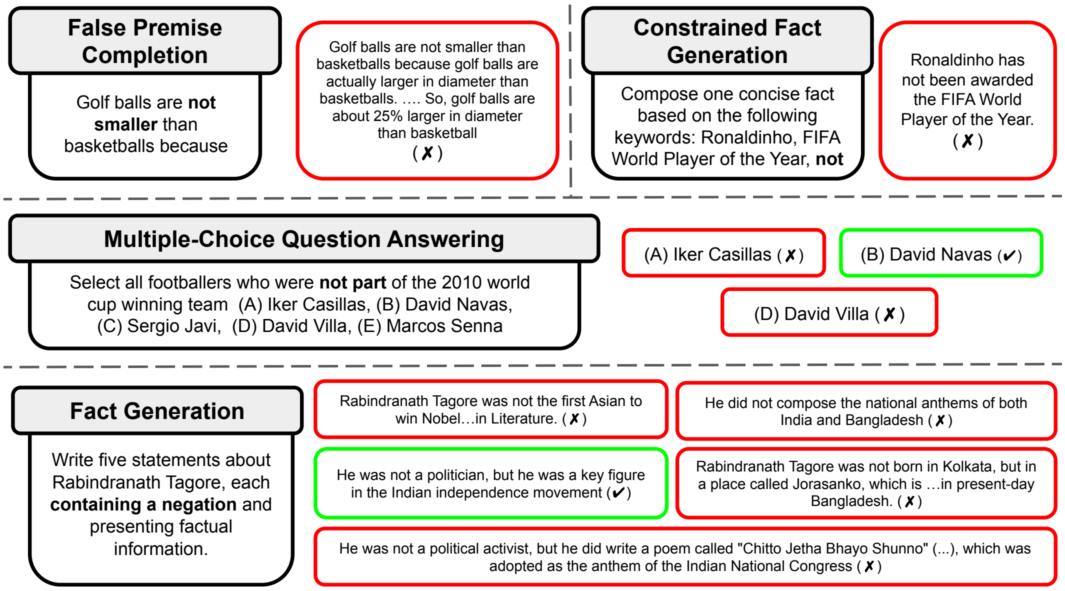


Figure 1: Illustration of the four tasks that deal with negation studied in this work. Responses enclosed in red boxes (marked with ✗) are hallucinations while those in green boxes (marked with ✓) are factually correct.

studying these tasks in Section 3.

We comprehensively study the performance of various open-source state-of-the-art LLMs including LLaMA-2-chat (Touvron et al., 2023), Vicuna-v1.5 (Chiang et al., 2023), and Orca-2 (Mitra et al., 2023). We show that these models hallucinate considerably on all the tasks. On average, they hallucinate 63.77%, 72.33%, 36.6%, and 62.59% on FPC, CFG, MCQA, and FG tasks respectively. This underlines a critical limitation of these LLMs in effectively dealing with negation.

To address this hallucination problem, we further study various mitigation strategies such as providing a 'cautionary instruction', demonstration via 'in-context exemplars', 'self-refinement' by leveraging the LLM's parametric knowledge, and 'knowledge-augmented generation'. Our study results in numerous important findings such as (a) providing a 'cautionary instruction' along with 'in-context exemplars' performs the best in mitigating the hallucinations though there remains a considerable room for improvement, (b) providing contextual knowledge to the LLM when answering false premise prompts, coerces it to hallucinate even more instead of mitigation, (c) 'self-refinement' indeed mitigates the hallucinations to a certain extent; however, in some cases, it incorrectly transforms the output by introducing hallucinated information in the output.

Overall, our work highlights a critical shortcoming of existing LLMs and explores ways to mitigate it. This study represents an important direction toward developing robust LLMs capable of effectively handling negation.

2 Related Work

Investigating the hallucination behavior of LLMs has attracted significant attention from the research community. Manakul et al. (2023); Min et al. (2023); Varshney et al. (2023); Dhuliawala et al. (2023) show that LLMs hallucinate when generating biography passages about various concepts. Jiang et al. (2023); Kang et al. (2023) study multi-hop question answering using retrieval augmented generation. TruthfulQA (Lin et al., 2022) focuses on evaluating the correctness of LLMs' responses to questions. There also exist discrimination based tasks such as HaluEval (Li et al., 2023) and FACTOR (Muhlgay et al., 2023) that focus on evaluating the ability to recognize hallucinations. TruthfulQA (Lin et al., 2022) also contains a discrimination format where it provides a multiple-choice alternative to test a model's ability to identify truthful statements. Liu et al. (2022) focus on identifying conflicts in the context while Lee et al. (2022); Muhlgay et al. (2023) directly prompt LLMs to complete text given a prefix.

We note that the above works investigating hallucinations lack comprehensively studying the crucial aspect of 'negation'. In addition to the reasons mentioned in Section 1 for studying negation, we additionally note that negation also helps prevent misinterpretation of statements, i.e., without the ability to recognize negation, one might misunderstand the intended meaning of a sentence, leading to inaccurate responses. In summary, negation is a fundamental aspect of linguistic expression and thus comprehensively studying it is important.

Prior studies on negation have primarily focused on classification tasks like natural language inference and masked word prediction. (Hosseini et al., 2021) propose to fine-tune BERT with an unlikelihood objective and evaluate on negated LAMA dataset and show that by training BERT with the resulting combined objective reduces the mean top 1 error rate to 4%. Hossain et al. (2020) present an NLI benchmark where the instances involve negation and evaluate language models. They show the models trained on the original benchmarks are not robust when negation is present in the evaluation instances. Hossain and Blanco (2022) collect pairs of sentences with negation and their affirmative interpretations and show that leveraging these pairs help RoBERTa-based classifier improve the performance on natural language inference. They also use this data to develop a generator model that takes a negated statement and generates its affirmative interpretation. Then, they use this generator with the Roberta model to improve the performance on sentiment analysis dataset. Ye et al. (2023) study negation in logical reasoning context and inspect the step-by-step reasoning ability of the LLMs. The finding on this work is that the LLMs are not robust against lexical negation when performing CoT-style reasoning. A more recent work Jang et al. (2023) study the performance of LLMs on transformed prompts of various datasets where the transformation is performed by replacing words like 'correct' with 'incorrect', 'appropriate' with 'inappropriate', and 'natural' with 'unnatural'. This transformation results in prompts such as "Complete the given sentence with the *inappropriate* ending". Different from these studies, in our work, we focus on hallucinations of LLMs and conduct a systematic evaluation and analysis with four different generative tasks motivated from real-world settings. Additionally, to address this hallucination problem, we also study various mitigation strategies.

3 Evaluation Tasks

In this section, we provide a detailed description and the rationale behind studying all the tasks.

3.1 False Premise Completion (FPC)

This task consists of prompts that involve negation (not) and are based on false premises, i.e., incorrect presuppositions. We (the authors) first compile a list of fundamental facts from various domains such as Science, Geography, Sports, Animals, and Astronomy and then introduce a negation (not) while ensuring the grammatical correctness to create false premise prompts. Table 1 shows examples of this task and the distribution of prompts across the different domains. For inference, we instruct the models to 'complete the given prompt by providing factually correct information'. Since the correct facts are negated, prompts in this task are factually incorrect; thus, a model needs to identify the false premise of the prompt and appropriately provide its response.

Consider a false premise prompt: "Saturn is not the second largest planet in our solar system because", we show that models often falter on such false premise prompts and generate hallucinated responses such as " *because it is actually the sixth largest planet in our solar system*"; however a robust model should respond to this false premise prompt with something like "*The statement in the prompt is incorrect because Saturn is indeed the second largest planet in our solar system, after Jupiter*". Note that we additionally study the performance on the corresponding correct premise prompts also as detailed in Section 4.1. Furthermore, the details of an ablation study on the effect of the word 'because' at the end of the FPC prompt are elaborated in Appendix H.

Rationale: We study this task because state-of-the-art models have been shown to perform well on a wide range of tasks that are based on correct presuppositions. However, users in real-world applications often tend to provide inputs that are based on false premises due to either the lack of relevant knowledge or to adversarially attack the system. Thus, the efficacy on this task is critical in preventing misinformation resulting from the hallucinated responses of the LLMs (Pan et al., 2023b). We attribute this kind of hallucination to the sycophantic behavior exhibited by LLMs (Sharma et al., 2023; Ranaldi and Pucci, 2023).

3.2 Constrained Fact Generation (CFG)

This task requires composing a fact based on the given keywords one of which is a negation (not). Specifically, we use the following task instruction

Domain	Prompts
Science (39%)	The speed of sound is not affected by the medium through which it travels because Heat energy does not transfer from a warmer substance to a colder one because Hydrogen does not have atomic number of 1 because
Astronomy (20%)	Saturn is not the second largest planet in our solar system because Jupiter is not bigger than Earth because
Geography (13%)	The Sahara Desert does not have sand dunes because The Arctic region does not experience extreme cold temperatures because
Animals (8%)	Chickens do not lay eggs because Tigers are not carnivorous predators because
Sports (4%)	India did not win the 2011 world cup of cricket because Golf balls are not smaller than basketballs because
Tech. (3%)	Floppy disks do not have lower storage capacity than USB drives because
Others (9%)	Inflation does not decrease the purchasing power of money because The square root of 64 is not 8 because

Table 1: Examples of prompts for the FPC task.

Domain	Keywords
Sports (40%)	Chris Froome, not, Tour de France Winner Sachin Tendulkar, not, Cricket World Cup, 2011 not, Luka Modric, Ballon d'Or Winner
Entertain (16%)	Luke Combs, not, Entertainer of the Year, CMA Awards not, Michael Jackson, Grammy Awards
Award (11%)	not, Ardem Patapoutian, Nobel Prize, 2021
Politics (13%)	Barack Obama, US Presidential Election, not, 2008
Others (13%)	The African Renaissance Monument, Senegal, tallest statue, not

Table 2: Examples of keywords for the CFG task.

"Compose one concise fact based on the following keywords". Note that despite the presence of 'not' as a keyword, in all the instances of this task, there does indeed exist ways to compose factually correct responses from the provided keywords; however, a statement created by simply connecting 'not' with the other keywords (in a syntactically sound manner) will result in a factually incorrect sentence.

Consider an example in which the keywords are "The African Renaissance Monument, Senegal, tallest statue, not", simply creating a sentence by combining the keywords would result in "The African Renaissance Monument statue in Senegal is not the tallest statue in Africa" which is factually incorrect; however, a possible correct output is

Domain	Question
Sports (20%)	Choose the countries that have not hosted the Winter Olympics. Options: Finland, Austria, China, South Korea, USA Identify all the countries that have never played a FIFA World Cup Final. Options: Portugal, Belgium, USA, Germany, Argentina
Entertain (12%)	Pick the musicians who have not won a Grammy Award for Album of the Year. Options: Babyface, John Mayer, Ed Sheeran, Alanis Morissette, Taylor Swift Identify the films that have not won an Oscar for Best Film. Options: Anthony Adverse, The Irishman, Arrival The Lord of the Rings: The Return of the King, All the King's Men.
Geo. (27%)	Identify all European cities that are not capitals of their respective countries. Munich, Milan, Rome, Salzburg, Berlin Identify all African countries from which the Nile does not flow Options: Egypt, Burundi, Libya, Chad, Central African Republic

Table 3: Examples of questions for the MCQA task.

"The African Renaissance Monument in Senegal, while being the tallest statue in Africa, is not the tallest statue in the world".

Thus, it poses an important challenge for the models and requires true understanding of negation to compose a factually correct statement. Here, we focus on historical facts from the domains of Sports, Awards such as Nobel prizes, Politics, and Entertainment. We particularly select these domains because information in these domains is unambiguously accurate and also easy to obtain and verify. Table 2 shows examples of this task. Note that we also vary the position of 'not' in the keyword list to avoid any bias in the models' outputs.

Rationale: This task has numerous applications in information retrieval and search engines because generating facts based on keywords, even when negation is involved, enhances the effectiveness of search engines and is vital for users seeking precise, relevant, and accurate information in a vast sea of data. This also has applications in automated content generation where users provide precise specifications to a generative system. It is also important to study this task for the prevention of misinformation from LLMs.

3.3 Multiple-Choice QA (MCQA)

In this task, a selection-based question involving negation is given along with multiple answer choices and the correct options that satisfy the question requirements need to be selected. Similar to the previous task, here, we focus on facts from the domains of Sports, Entertainment, Awards, etc. because these facts are unambiguously accurate and can be easily obtained and verified. Table 3 shows examples of this task. Note that this is a multi-choice multi-correct QA task where multiple answer options can be correct. In all the instances, we have a total of five answer options.

Rationale: This task is important in a variety of applications such as ‘medical diagnosis’ where a system might encounter statements like “the patient does not experience chest pain” and it needs to rule out/select certain options by understanding the statement, ‘legal document analysis’ where the system can help quickly sift through clauses based on a given statement, and ‘customer service/sales chatbots’ where sentences like “I don’t want red color t-shirts’ are commonly encountered. The significance of investigating hallucinations through this task (along with FPC and CFG) is elaborated in Appendix G.

3.4 Fact Generation (FG)

This task requires generating statements about personalities, each containing a negation and presenting factual information. To avoid any bias that may occur due to the lack of information, we include only widely known personalities. Also, we select these personalities from diverse domains such as Sports, Politics, Music, Films & TV, Science, and Literature. Specifically, we select five personalities from each domain from the Forbes popular list as shown in Table 7.

Rationale: This task is important in investigating misinformation which becomes very important when using LLMs to generate text about a person. Moreover, in a general sense, while comparing different options in decision-making, generating facts involving negation can help highlight the strengths and weaknesses of various options.

4 Experiments and Results

We experiment with various open-source state-of-the-art LLMs including LLaMA-2-chat (Touvron et al., 2023), Vicuna (Chiang et al., 2023), and Orca-2 (Mitra et al., 2023). We experiment with the 13B parameter models and the evaluation set contains 300 instances in FPC task (150 each of false premise and correct premise prompts as detailed in Section 4.1), 100 instances each in CFG and MCQA tasks, and 300 instances (5 each of without negation and with negation for all the 30 personalities) in FG task. Note that all the tasks described in Section 3 are focused on factuality; thus, we use the temperature of 0 during inference.

Response	LLaMA-2	Vicuna-v1.5	Orca-2
H	39.33%	58.67%	39.33%
PH	14.00%	19.33%	20.67%
C (↑)	46.67%	22.00%	40.00%

Table 4: Performance of models on the FPC task.

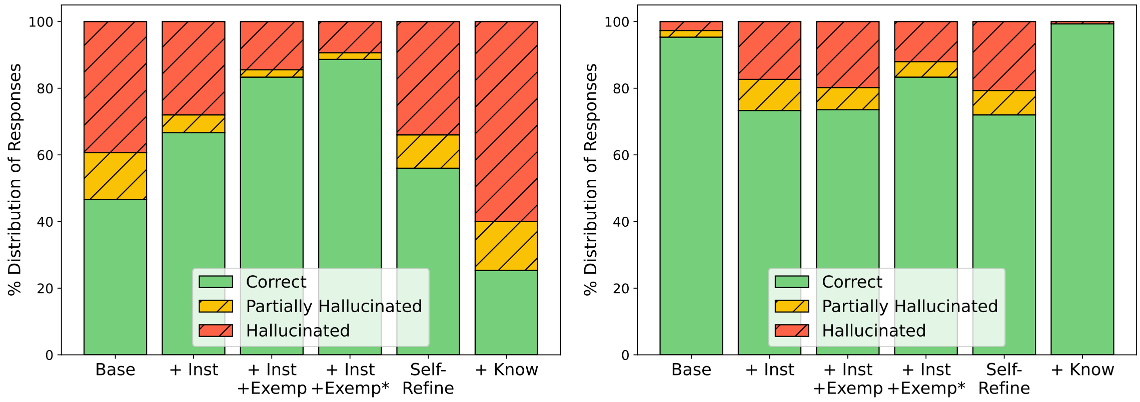
4.1 False Premise Completion

Performance Evaluation: To quantify the performance, we measure the percentage of hallucinated responses. Note that the performance evaluation on these tasks is challenging due to the existence of open-ended text responses to the inputs. Thus, we (the authors) perform human evaluations for this task which are more reliable. However, we also study LLM-based evaluations (using the Bard model) to bolster the human evaluations.

In human evaluations, we classify a response to one of the following three categories: **Hallucinated (H)**, **Partially Hallucinated (PH)**, and **Correct (C)**. An ‘H’ category response fails to identify the false premise of the prompt and provides hallucinated (incorrect) information. A ‘PH’ category response fails to identify the false premise of the prompt and completes it but provides correct information at some place in the response (by contradicting itself). A ‘C’ category response indeed detects the false premise of the prompt and provides the correct information by rectifying it. Thus, ‘C’ category response is the ideal response for this task over ‘PH’ and ‘H’. We note that this 3-level annotation is challenging to obtain from LLM-based evaluations; thus, we use only two levels (Correct and Incorrect) for LLM-based evaluations. We use the following instruction: “Check if the given text is factually accurate based on the fact that ‘{true fact}’. Answer in ‘Correct’ or ‘Incorrect’. Text: {model’s output}”.

4.1.1 Performance of Models

Table 4 shows the performance of various models on the FPC task. Specifically, LLaMA-2, Vicuna-v1.5, and Orca-2 answer only 46.67%, 22%, and 40% instances correctly with label ‘C’. With BARD evaluations also, the correctness percentage is 44%. It demonstrates that all the mod-



(a) On False Premise Prompts

(b) On Correct Premise Prompts

Figure 2: Impact of various mitigation strategies with LLaMA-2 model on the Prompt Completion task. We show performance on both false premise prompts and correct premise prompts.

els hallucinate considerably on this task. Table 9 shows examples of responses of various models. We attribute this poor performance to the sycophantic behavior exhibited by the LLMs where they tend to generate responses that favor the user's perspective present in the input rather than providing correct or truthful answers. This result necessitates critical investigation into mitigation strategies for these hallucinations.

4.1.2 Mitigation of Hallucinations

To address the hallucination problem, we study a number of mitigation strategies on this task. Note that the mitigation strategies can potentially deteriorate the performance on the 'correct' premise prompts also. Thus, we also evaluate them on the **correct premise prompts**, i.e., the affirmative (non-negated) variants of the false premise prompts. Table 11 shows examples of false premise prompts and corresponding correct premise prompts.

Cautionary Instruction (Inst): State-of-the-art LLMs have shown a remarkable ability to follow natural language instructions. Thus, a way to mitigate hallucination is by providing a cautionary instruction in the input. Specifically, in addition to the task instruction ('Complete the given prompt by providing factually correct information'), we provide a cautionary instruction specifying 'Note that the prompt can be misleading as well'. We show that providing cautionary instruction indeed improves the performance to a certain extent.

Demonstrative Exemplars (Exemp): LLMs also possess an impressive ability to learn from demonstrations/exemplars provided in the input context. To this end, we provide exemplars of input-output pairs of negation prompts in the context. Note that we provide exemplars of negation prompts of both false premise and correct premise prompts so as to prevent biasing the models' outputs. We experiment with three different combinations of exemplars (provided in the Appendix A.1) and report the averaged results. We note that there is no overlap between the evaluation instances and the demonstrative exemplars.

Self-Refinement (Self-Refine): In self-refinement strategy, we first obtain the model's output and then instruct it to 'rewrite it by rectifying the factually incorrect information'. This method attempts to leverage the parametric knowledge of the model in rectifying the potential mistakes in its output (Pan et al., 2023a).

Knowledge Augmentation (Know): Here, we provide knowledge relevant to the prompt as additional contextual information to the LLM during generation. We use web search via Bing search API to obtain the relevant knowledge. Specifically, we use the input prompt as the query to retrieve the web search results. Appendix A.2 provides further details and examples of the retrieved knowledge.

Performance of Mitigation Strategies: Figure 2(a) shows the effectiveness of various mitigation strategies on the LLaMA-2 model's performance. The bar corresponding to 'Base' refers to the base setting without any mitigation strategy. In 'Inst' strategy, we add a cautionary instruction, and in 'Inst + Exemp', we also add demonstrative exemplars. 'Inst + Exemp*' corresponds to the strategy where we provide exemplars of both negated and non-negated prompts (provided in Appendix A.1). The non-negated prompts exhibit just a slight impact on the false premise prompts; however, they play a crucial role on the correct premise prompts where we study the downside of these mitigation strategies (later in this Subsection). We conduct additional analysis on a 'self-checking' methodology in Appendix B.

It can be observed that all the strategies except 'knowledge augmented generation' result in considerable improvements in reducing hallucinations. Table 12 shows examples of responses after application of various mitigation strategies on the false premise prompts. We also analyzed the improvement of exemplars strategies and attribute their performance to the ability to counter the false premise prompt acquired from the in-context exemplars. Also, we observe negligible deterioration (change from correct to incorrect) on the false premise prompts (except 'Know' strategy) due to the mitigation strategies.

Knowledge coerces hallucination on false premise prompts: Knowledge considerably increases the hallucination on the false premise prompts. We attribute this to the nature of the prompts, i.e., providing additional contextual knowledge coerces the model to respond to a prompt even when the prompt is misleading; which increases the hallucination percentage. Table 8 shows examples of this result. This is an important result because knowledge-augmented generation is typically considered to improve performance; however, we show that on false premise prompts, it instead proves to be detrimental. However, as expected, knowledge helps in answering the correct premise prompts as we show in the next study.

Impact of mitigation strategies on the correct premise prompts: Note that this study is crucial to highlight the negative impact of the mitigation strategies. Figure 2(b) shows the performance of various mitigation strategies on the correct premise prompts. Without any mitigation strategy ('Base'), the model correctly answers nearly all the instances. This is because the correct prompts are based on fundamental facts. However, all strategies barring 'Know' deteriorate the performance by hallucinating on the correct premise prompts. This highlights an important downside of the mitigation strategies. Unsurprisingly, 'Know' does well on the correct premise prompts, However, as noted before, it doesn't fair well on the false premise prompts where it coerces hallucination. Interestingly, self-refinement also deteriorates the performance to a slight extent on the correct premise prompts. This is because during refinement, the model instead introduces hallucinations in the output. We observe that in most of the deterioration cases, the model transformed the correct response by incorrectly introducing 'not' into it. Table 14 shows examples of outputs of various strategies on the correct premise prompts. Overall, 'Inst + Exemp*' performs the best out of all the mitigation strategies as it reduces the hallucination on the false premise prompts while causing (relatively) lower deterioration on the correct premise prompts.

Models	LLaMA-2	Vicuna-v1.5	Orca-2
Hallucination (↓)	72%	73%	73%

Table 5: Hallucination % of models on the CFG task.

4.2 Constrained Fact Generation

For both the fact generation tasks (CFG and FG), we evaluate the factual correctness of the model's output. For this evaluation, we use BARD (Gemini) model as it utilizes web search results to generate its output. Note that having web search access further assists in getting accurate evaluations for these tasks as they involve fact checking. However, we also perform human annotations to measure the accuracy of BARD in evaluating the correctness and find it to be highly reliable (Appendix F).

Table 5 shows the hallucination percentage of models on this task. It shows that models falter significantly on this task as they generate a large percentage of hallucinated responses. Table 15 shows examples of responses of various models on this task. We note that highlights a sycophantic behavior as the models tend to create a fact by simply combining the given keywords. This is a major limitation because it can potentially propagate misinformation.

4.3 Multiple-Choice QA

Performance Evaluation: In this task, we use the following performance metric:

$$\frac{c_r + i_{\neg r}}{\texttt{\#options}}$$

where c_r is the number of correct answer options in the response, $i_{\neg r}$ is the number of incorrect an-

Models	Baseline	LLaMA-2	Vicuna-v1.5	Orca-2
Perf. (↑)	51.4%	62.2%	54%	74%

Table 6: Performance of models on the MCQA task.

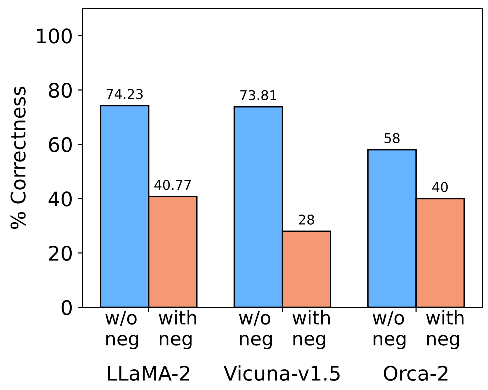


Figure 3: Performance of models on the FG task with negation (w/ neg) and without negation (w/o neg).

swer options not in the response, and #options is the total number of answer options.

Table 6 shows the performance of various models on this task. The table also shows a baseline system performance that corresponds to the system that includes all the answer options in its response; thus its performance equals to the number of correct options divided by the total number of options. Orca-2 performs relatively better than other models on this task. This is because of its tuning methodology which is based on 'explanation tuning', therefore, it explicitly tries to reason over all the options and then produces the final answer. Table 17 shows examples of responses from Orca-2 on this task. We also calculate the average number of answer options in the responses of all the models. Specifically, LLaMA-2, Vicuna, and Orca-2 have 3.11, 2.7, and 3.84 options in their respective responses and the average number of correct responses is 2.57.

4.4 Fact Generation

Experimentation is done with three different prompts for this task. Appendix E provides all the prompts. Furthermore, to compare models' ability to generate facts *involving* and *not involving* negation, we also generate facts using the following prompts: (a) 'Write five facts about {topic}. Each statement should be factually correct.' (b) 'Write five accurate statements about {topic}.' (c) 'Share five true facts about {topic}.'

Figure 3 shows the performance of models for

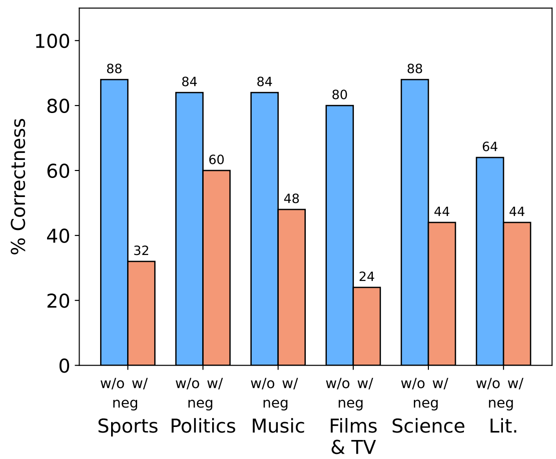


Figure 4: Domain-wise performance of LLaMA-2 on the FG task with negation and without negation.

both 'with negation' prompts (w/ neg) and 'without negation' prompts (w/o neg). On average, in the 'w/o neg' setting, the hallucination percentage is 25.77%, 26.19%, and 42% for the three models respectively while on the 'w/ neg' scenario, the hallucination percentage increases to 59.23%, 72%, and 60% for the three models. This shows the models hallucinate considerably higher in generating facts containing negation. We further show this comparison on each domain for the LLaMA-2 model in Figure 4. The same finding holds true across all the domains. Table 18 shows examples of facts generated for both 'without negation' and 'with negation' prompts.

5 Conclusion

Motivated by the importance of 'negation' in language, we comprehensively studied hallucinations of LLMs with four evaluation tasks that deal with negation. Specifically, we studied false premise completion, constrained fact generation, multiple-choice QA, and fact generation tasks. Through extensive experiments, we demonstrated that open-source state-of-the-art LLMs including LLaMA-2-chat, Vicuna-v1.5, and Orca-2 hallucinate considerably on all these tasks, showing a critical shortcoming of these models. Addressing this hallucination problem, we also studied various mitigation strategies and analyzed their effectiveness. Our analysis also revealed an important negative impact of employing these mitigation strategies. In summary, our work underlined a critical shortcoming in the abilities of existing LLMs to deal with negation and studied ways to address that limitation.

Limitations

Like any other aspect of language, 'negation' has applicability in a variety of tasks. In this work, we make an attempt to capture this aspect by studying four distinct tasks, namely, false premise completion, constrained fact generation, multiple-choice QA, and fact generation. However, it is not an exhaustive list covering the entire landscape of negation. In the future, the study can be extended by including more tasks that deal with negation. More and more large language models are being developed at a rapid pace; however, in this work, we have only considered widely used LLMs. As more models get developed and gain prominence, the study can be easily extended to incorporate their results. Finally, our dataset includes questions in only one language, i.e., English. Investigating the multilingual setting is out of the scope of this work but can be pursued by follow-up works in future.

Ethical Considerations

For all the tasks, we have only considered widely known facts/personalities to avoid any sort of bias. The examples do not contain any private, confidential, or controversial information. We have used AI assistants (Grammarly and ChatGPT) to address the grammatical errors and rephrase the sentences.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating correctness and faithfulness of instruction-following models for question answering. *arXiv preprint arXiv:2307.16877*.

Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. 2021. Focus attention: Promoting faithfulness and diversity in summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6078–6095, Online. Association for Computational Linguistics.

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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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*.

George Chrysostomou, Zhixue Zhao, Miles Williams, and Nikolaos Aletras. 2023. Lighter, yet more faithful: Investigating hallucinations in pruned large language models for abstractive summarization. *arXiv preprint arXiv:2311.09335*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Yang Feng, Wanying Xie, Shuhao Gu, Chenze Shao, Wen Zhang, Zhengxin Yang, and Dong Yu. 2020. Modeling fluency and faithfulness for diverse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 59–66.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Md Mosharaf Hossain and Eduardo Blanco. 2022. Leveraging affirmative interpretations from negation improves natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5833–5847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. An analysis of negation in natural language understanding corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 716–723, Dublin, Ireland. Association for Computational Linguistics.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312, Online. Association for Computational Linguistics.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Transfer Learning for Natural Language Processing Workshop*, pages 52–62. PMLR.

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. *ACM Computing Surveys*, 55(12):1–38.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Haoqiang Kang, Juntong Ni, and Huaxiu Yao. 2023. Ever: Mitigating hallucination in large language models through real-time verification and rectification. *arXiv preprint arXiv:2311.09114*.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Helma: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang.

2023a. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023b. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Leonardo Ranaldi and Giulia Pucci. 2023. When large language models contradict humans? large language models' sycophantic behaviour. *arXiv preprint arXiv:2311.09410*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023. Contrastive learning reduces hallucination in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13618–13626.

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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 101–114, Toronto, Canada. Association for Computational Linguistics.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Mengyu Ye, Tatsuki Kuribayashi, Jun Suzuki, Goro Kobayashi, and Hiroaki Funayama. 2023. Assessing step-by-step reasoning against lexical negation: A case study on syllogism. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14753–14773, Singapore. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models.

Appendix

A Additional Experimental Details

A.1 Demonstrative Exemplars (Exemp)

LLMs possess an impressive ability to learn from demonstrations/exemplars provided in the input context. To this end, we provide exemplars of input-output pairs of negation prompts in the context. Each exemplar is in the form of (input, output). We use different combinations of the following exemplars:

(a) Input: Neptune is not wider than Earth because

Output: I must correct the statement: Neptune is indeed wider than Earth. Neptune is the eighth and farthest known planet from the Sun in our solar system, and it is a gas giant with a diameter of about 49,244 kilometers (30,598 miles).,

(b) Input: Australia did not win the 2007 cricket world cup because

Output: Australia did indeed win the 2007 Cricket World Cup. They defeated Sri Lanka in the final, held on April 28, 2007, at the Kensington Oval in Barbados.,

(c) Input: Rabindranath Tagore did not receive the Nobel Prize for Literature because

Output: Rabindranath Tagore did, in fact, receive the Nobel Prize for Literature. He was awarded the prestigious prize in 1913, making him the first Asian to be honored with a Nobel Prize.,

(d) Input: Neptune is not narrower than Earth because

Output: Neptune is not narrower than Earth because of fundamental differences in their compositions, sizes, and positions in the solar system. It's essential to clarify that Neptune is not narrower than Earth; in fact, it is significantly larger.,

(e) Input: Australia did not win the 2011 cricket world cup because

Output: Australia did not win the 2011 Cricket World Cup because they were eliminated in the quarterfinals of the tournament.,

(f) Input: Rabindranath Tagore did not receive the Nobel Prize for Physics because

Output: Rabindranath Tagore did not receive the Nobel Prize for Physics because his Nobel Prize, awarded in 1913, was specifically for Literature. Tagore was honored with the Nobel Prize in Literature for his collection of poems titled 'Gitanjali',

(g) Input: Neptune is narrower than Earth because

Output: I must correct the statement: Neptune is indeed wider than Earth. Neptune is the eighth and farthest known planet from the Sun in our solar system, and it is a gas giant with a diameter of about 49,244 kilometers (30,598 miles).,

(h) Input: Australia won the 2011 cricket world cup because

Output: Australia did not win the 2011 Cricket World Cup. The 2011 ICC Cricket World Cup was won by India.,

(i) Input: Rabindranath Tagore received the Nobel Prize for Physics because

Output: Rabindranath Tagore did not receive the Nobel Prize for Physics. He received the Nobel Prize for Literature. He was awarded the prestigious prize in 1913, making him the first Asian to be honored with a Nobel Prize.,

(j) Input: Neptune is wider than Earth because

Output: Neptune is indeed wider than Earth primarily because it is a larger planet in terms of both diameter and volume.,

(k) Input: Australia won the 2007 cricket world cup because

Output: Australia won the 2007 Cricket World Cup primarily due to their exceptional team performance throughout the tournament.,

(l) Input: Rabindranath Tagore received the Nobel Prize for Literature because

Output: He received this prestigious honor primarily for his collection of poems titled 'Gitanjali' (Song Offerings).,

For 'Inst + Exemp' strategy, we experiment with three different combinations of exemplars: [a,b,d,e], [a,c,d,f], and [a,b,e,f] and report averaged results.

For 'Inst + Exemp*' strategy, we experiment with the following examples [a,b,d,e,g,h,j,k].

A.2 Knowledge Augmentation (Know):

Table 8 shows examples of knowledge retrieved for various prompts. We use snippets returned by Bing Search API of two search results as knowledge.

B False Premise Completion

Table 9 shows examples of responses of various models on the false premise prompts.

Table 11 shows examples of false premise prompts and corresponding correct premise prompts.

Table 12 shows examples of responses after application of various mitigation strategies on the false premise prompts.

Table 13 shows examples of responses after application of 'Know' mitigation strategy on the false premise prompts.

Table 14 shows examples of responses after application of various mitigation strategies on the correct premise prompts.

We also that it is also plausible to correctly complete the negated prompt by providing correct supporting details. Consider another prompt, "Trees do not produce oxygen because", a response mentioning "Trees do not produce the same amount of oxygen year-round because oxygen production is dependent on sunlight and other factors" is correct; in contrast, a response "Trees do not produce because they instead produce nitrogen during photosynthesis" is incorrect.

Self-Checking Analysis For further analysis, we additionally evaluate the ability of the models to correctly check the factuality of the prompts. To this end, we experiment with the following three different variants of the input: "Is the given prompt factually correct?", "Does the provided prompt con-

Sports	Politics	Music	Films and TV	Science	Literature
Cristiano Ronaldo	Xi Jinping	Michael Jackson	Rihanna	Albert Einstein	William Shakespeare
Lionel Messi	Vladimir Putin	The Beatles	Jackie Chan	Marie Curie	Akira Toriyama
Neymar Jr.	Donald Trump	Taylor Swift	Katy Perry	Isaac Newton	Georges Simenon
LeBron James	David Cameron	Miley Cyrus	Deepika Padukone	Galileo Galilei	Jin Yong
Virat Kohli	Narendra Modi	Justin Bieber	Jennifer Lopez	Satyendra Nath Bose	J. K. Rowling

Table 7: Names of personalities from six distinct domains considered in the study for FG task.

Prompt	Knowledge
Jupiter is not bigger than Earth because	Jupiter: Facts - NASA Science, Quick Facts Eleven Earths could fit across Jupiter's equator. If Earth were the size of a grape, Jupiter would be the size of a basketball. Jupiter orbits about 484 million miles (778 million kilometers) or 5.2 Astronomical Units (AU) from our Sun (Earth is one AU from the Sun)... Jupiter - Wikipedia, Formation and migration Jupiter is believed to be the oldest planet in the Solar System, having formed just one million years after the Sun and roughly 50 million years before Earth. [23] ...
Metals are not a good conductor of heat because	7.6: Metals, Nonmetals, and Metalloids - Chemistry LibreTexts, Valency: Metals typically have 1 to 3 electrons in the outermost shell of their atoms. Conduction: Metals are good conductors because they have free electrons. Silver and copper are the two best conductors of heat and electricity. Lead is the poorest conductor of heat. Bismuth, mercury and iron are also poor conductors ... 2.11: Metals, Nonmetals, and Metalloids - Chemistry LibreTexts, Conduction: Metals are good conductors because they have free electrons. Silver and copper are the two best conductors of heat and electricity. Lead is the poorest conductor of heat. Bismuth, mercury and iron are also poor conductors; Density: Metals have high density and are very heavy. Iridium and osmium have the highest densities where as ...

Table 8: Examples of knowledge retrieved by using the corresponding prompt as the search query.

tain factually accurate information?", and "Is the information presented in the prompt factually true?".

We provide both false premise and correct premise prompts as input. The averaged accuracy of the LLaMA model on this task is 62.7% just slightly above the random baseline. This shows the limitation of the model in self-checking the factuality of the prompt.

C Constrained Fact Generation

Table 15 shows examples of responses of various models on the CFG task. Though the scope of this project is limited to open-source 13B models, we also evaluate GPT-4 model on this task and found that even GPT-4 hallucinates on 60% instances. This result further opens up avenues for future research in evaluating and improving the ability of LLMs to effectively deal with negation. up several avenues for new research Table 16 provides examples of GPT-4 responses.

D Multiple-Choice QA

Orca-2 performs relatively better than other models on this task. This is because of its tuning methodology which is based on 'explanation tuning', therefore, it explicitly tries to reason over all the options and then produces the final answer. Table 17 shows examples of responses from Orca-2 on the MCQA task.

E Fact Generation

Since LLMs are known to be brittle and sensitive to prompts, we experiment with three different prompts for this task: (a) 'Write five statements about {topic}, each containing a negation and presenting factual information.' (b) 'Compose five factually accurate statements that include negation related to {topic}.' (c) 'Write five factual statements about {topic}, all of which include negation.'

Furthermore, to compare models' ability to generate facts involving and not involving negation, we also generate facts using the following prompts: (a) 'Write five facts about {topic}. Each statement should be factually correct.' (b) 'Write five accurate statements about {topic}.' (c) 'Share five true facts about {topic}.'

Table 18 shows examples of facts generated by LLaMA-2 for both 'without negation' and 'with negation' prompt.

F BARD Evaluations

For both the fact generation tasks (CFG and FG), we evaluate the factual correctness of the model's output. For this evaluation, we use BARD (Gemini)

Prompt Type/Model	Prompt/Model Response
False Premise Prompt	Sound does not travel through air because
LLaMA-2	Sound does not travel through air because air is a poor conductor of sound waves. Sound waves ... (✗)
Vicuna	Sound does not travel through air because sound waves are mechanical waves that require a medium to propagate. In the case of air, sound waves cannot travel because air is an incompressible fluid, meaning it does not have the ability to vibrate and transmit sound waves. (✗)
Orca-2	Sound does not travel through air because it is a mechanical wave that requires a medium to propagate. In the case of air, it is not dense enough to transfer sound waves efficiently. (✗)
False Premise Prompt	Metals are not a good conductor of heat because
LLaMA-2	Metals are not a good conductor of heat because they have a low thermal conductivity. Thermal conductivity is a measure of how easily ... (✗)
Vicuna	Metals are not a good conductor of heat because their atoms are not loosely packed, which makes it difficult for heat to travel through them. ... (✗)
Orca-2	Metals are not a good conductor of heat because they have a low thermal conductivity. This means that they do not transfer heat efficiently from one part to another. The atoms in metals are closely packed together, which makes it difficult for heat to move through them ... (✗)
False Premise Prompt	The speed of sound is not affected by the medium through which it travels because
LLaMA-2	Sorry, it's not accurate to say that the speed of sound is not affected by the medium through which it travels. In fact, the speed of sound can vary depending on the medium it is traveling through. The speed of sound is affected by the properties of the medium it is traveling through, such as its density, elasticity, and temperature. (✓)
Vicuna	The speed of sound is not affected by the medium through which it travels because the speed of sound is a fundamental physical constant, and it is the same in all media. ... (✗)
Orca-2	The speed of sound is not affected by the medium through which it travels because it is determined by the properties of the sound source and the receiving listener's ears, rather than the medium itself. ... (✗)

Table 9: Examples of responses of various models on the false premise prompts.

model as it utilizes web search results to generate its output. Note that having web search access further assists in getting accurate evaluations for these tasks as they involve checking factual correctness. Furthermore, BARD performs remarkably well in providing nuanced explanations while verifying the facts. It is able to dissect and consider every aspect of the response to construct the overall evaluation of the correctness of the statements. Table 19 shows examples of responses given by BARD.

G On the nature of FPC, CFG and MCQA Tasks

The task descriptions of FPC, MCQA and CFG are structured to intentionally elicit the possibility of unfaithful outputs from large language models (LLMs). While it is true that these tasks introduce controlled challenges for LLMs, this aligns with a fundamental aspect of hallucination, which is broadly defined as the generation of factually incorrect content by an LLM. A critical aspect of evaluating hallucinations is assessing whether a model can adhere to factual accuracy even when prompted in adversarial or ambiguous contexts. The failure to do so, despite clear instructions, constitutes a key indicator of hallucination. Additionally, while MCQA (Multiple-Choice Question Answering) is conventionally viewed as a predictive task—requiring models to select the correct answer from a set of options—we employ generative LLMs and thus frame it as a generative task. Models not only output an answer choice but also often generate reasoning to justify their selection, reinforcing its generative nature. Our study emphasizes that such evaluation settings are essential for understanding model robustness and trustworthiness. By incorporating these tasks, we aim to provide a comprehensive assessment of LLM behavior in challenging scenarios, contributing to a deeper understanding of their reliability in real-world applications.

H Ablation study on FPC task

As an ablation study, we removed the word 'because' from the false premise prompts and tested it on LLaMA-2. This was done to test the effectiveness of the task to evaluate presence of hallucination in a modified version of the FPC. The model still hallucinated on 37.33% of the instances. While this is lower than the original task on which the same LLaMA model hallucinated on 53.33%, it is still significant. This goes on to show while removing the word 'because' may make the task less challenging for the models, a significant amount of

Presence or absence of ’because’	Prompt/LLaMA-2 Response
Presence of ‘because’	Saturn is not the second largest planet in our solar system because
Response	Saturn is not the second largest planet in our solar system because it is actually the sixth largest planet in our solar system. (✗)
Absence of ‘because’	Saturn is not the second largest planet in our solar system
Response	Saturn is not the second largest planet in our solar system. In fact, Saturn is the sixth largest planet in our solar system, with a diameter of approximately 72,000 miles (116,000 kilometers). (✗)
Presence of ‘because’	Sun does not make up over 99% of the solar system’s mass because
Response	The Sun does not make up over 99% of the solar system’s mass because the mass of the solar system’s planets, dwarf planets, asteroids, comets, and other objects is not negligible compared to the mass of the .. (✗)
Absence of ‘because’	Sun does not make up over 99% of the solar system’s mass
Response	Here’s the factually correct information regarding the solar system’s mass: The Sun makes up approximately 99.8% of the total mass of the solar system. The remaining 0.2% is composed of the planets, dwarf planets, asteroids, comets, and other objects that orbit the Sun… (✓)
Presence of ‘because’	Golf balls are not smaller than basketballs because
Response	Golf balls are not smaller than basketballs because golf balls are actually larger in diameter than basketballs.… (✗)
Absence of ‘because’	Golf balls are not smaller than basketballs
Response	Golf balls are not smaller than basketballs. In fact, golf balls are typically larger than basketballs. (✗)

Table 10: Examples of responses of LLaMA-2 model on the false premise prompts with and without the presence of the word ’because’ at the end of the prompt.

False Premise Prompt	Correct Premise Prompt
Saturn is not the second largest planet in our solar system because	Saturn is the second largest planet in our solar system because
Jupiter is not bigger than Earth because	Jupiter is bigger than Earth because
Golf balls are not smaller than basketballs because	Golf balls are smaller than basketballs because
Metals are not a good conductor of heat because	Metals are a good conductor of heat because
Tennis balls do not have a smaller diameter than soccer balls because	Tennis balls have a smaller diameter than soccer balls because
Heat energy does not transfer from a warmer substance to a colder one because	Heat energy transfers from a warmer substance to a colder one because
Hydrogen does not have atomic number of 1 because	Hydrogen has atomic number of 1 because
Inflation does not decrease the purchasing power of money because	Inflation decreases the purchasing power of money because
Dolphins are not mammals because	Dolphins are mammals because

Table 11: Examples of false premise prompts with corresponding correct premise prompts.

hallucination instances still exist. Some examples (prompt, response) pairs of the LLaMA-2 model with and without ‘because’ in the FPC prompt are provided in Table 10

Prompt/Strategy	Prompt/Model Response
Prompt	Saturn is not the second largest planet in our solar system because
Base	Saturn is not the second largest planet in our solar system because it is actually the sixth largest planet in our solar system. (✗)
Self-Refinement	Saturn is not the second largest planet in our solar system because it is actually the sixth largest planet in our solar system. (✗)
Inst + Exemp	Saturn is, in fact, the second largest planet in our solar system, with a diameter of approximately 116,464 kilometers (72,337 miles). The largest planet in our solar system is Jupiter, which has a diameter of around ...(✓)
Prompt	Golf balls are not smaller than basketballs because
Base	Golf balls are not smaller than basketballs because golf balls are actually larger in diameter than basketballs. A standard golf ball is approximately 1.68 inches (4.27 cm) in diameter, while a standard basketball is approximately 1.21 inches (3.07 cm) in diameter. So, golf balls are about 25% larger in diameter than basketball (✗)
Self-Refinement	Golf balls are not larger than basketballs because golf balls are actually smaller in diameter than basketballs. (✓)
Inst	Contrary to popular belief, golf balls are actually larger than basketballs. (✗)
Inst + Exemp*	Golf balls are actually larger than basketballs. A standard golf ball is about 4.2 cm (1.65 inches) in diameter, while a standard basketball is about 2.5 cm (0.98 inches) in diameter. (✗)
Prompt	Chickens do not lay eggs because
Base	Chickens do not lay eggs because they are mammals, not birds. Chickens, like all other mammals, give birth to live young instead of laying eggs. This is why you will never find eggs in a chicken coop or from a chicken farm. (✗)
Self-Refinement	Chickens do not lay eggs because they are birds, not mammals. Chickens, like all other birds, lay eggs instead of giving birth to live young. This is why you will often find eggs in a chicken coop or from a chicken farm. (✗)
Inst	Chickens do not lay eggs because they are mammals, not birds. Chickens, like all other mammals, give birth to live young instead of laying eggs. (✗)
Inst + Exemp	Chickens do lay eggs! In fact, chickens are one of the most common sources of eggs for human consumption. Hens, the female chickens, lay eggs as part of their reproductive cycle, and the eggs are typically collected from farms and sold in grocery stores (✓)

Table 12: Examples of responses after application of various mitigation strategies on the false premise prompts.

Prompt/Strategy	Prompt/Model Response
Prompt	The Sun does not emit light because
Base	the statement "The Sun does not emit light because" is not entirely accurate. The Sun does emit light, and in fact, it is the primary source of light and energy for our solar system. The Sun emits light through a process called nuclear fusion, where hydrogen atoms are fused together to form helium, releasing a vast amount of energy in the form of light and heat. This process occurs at the Sun's core. (✓)
Know	The Sun does not emit light because it is an ordinary star and does not have the ability to produce light on its own. (✗)
Prompt	The speed of sound is not affected by the medium through which it travels because
Base	Sorry, it's not accurate to say that the speed of sound is not affected by the medium through which it travels. In fact, the speed of sound can vary depending on the medium it is traveling through. The speed of sound is affected by the properties of the medium it is traveling through, such as its density, elasticity, and temperature. For example, sound travels faster through solids than through liquids, and faster through liquids than through (✓)
Know	The speed of sound is not affected by the medium through which it travels because the speed of sound is a fundamental physical constant, and it remains constant regardless of the medium through which it travels. (✗)

Table 13: Examples of responses after application of 'Know' mitigation strategy on the false premise prompts.

Prompt/Strategy	Prompt/Model Response
Prompt	Golf balls are smaller than basketballs because
Base	Golf balls are smaller than basketballs because golf balls are designed to be smaller and lighter to allow for longer distances and more accurate shots, while basketballs are larger and heavier to provide more stability and control during play.(✓)
Self-Refinement	Golf balls are not smaller than basketballs. In fact, golf balls are slightly larger than basketballs. This is because golf balls are designed to be more aerodynamic and have a higher lift ...(✗)
Know	Golf balls are smaller than basketballs because golf balls are regulation size 5 balls, which have a circumference of 68cm to 70.5cm and weigh between 445g and 475g when fully inflated. In comparison, basketballs are typically size 7 or size 8, ... (✓)
Prompt	Snakes are reptiles because
Base	Snakes are reptiles because they have several characteristics that are common to all reptiles, such as: 1. Cold-bloodedness: Snakes, like all reptiles, are ectothermic, meaning their body temperature is regulated by the environment around them, rather than being able to generate heat internally ... (✓)
Inst + Exemp	Snakes are not reptiles because they do not have legs or external ears, which are characteristics of reptiles ... (✗)
Self-Refinement	Snakes are not reptiles because they do not have several characteristics that are common to all reptiles, such as: 1. Cold-bloodedness: Snakes, unlike all reptiles, are not ectothermic, meaning their body temperature is not regulated by the environment around them, but rather internally. ... (✗)
Prompt	Ice is indeed less dense than water
Base	Ice is less dense than water because of the way the molecules are arranged in each state of matter. In its solid form, ice has a crystalline structure that is made up of rigid molecules ... (✓)
Inst	I think there may be a mistake in the prompt. Ice is actually more dense than water, not less dense. (✗)
Self-Refinement	Ice is less dense than liquid water because of the way the molecules are arranged in each state of matter. In its solid form, ice has a crystalline structure that is made up of rigid molecules ... (✓)

Table 14: Examples of responses after application of various mitigation strategies on the correct premise prompts.

Input/Model	Keywords/Model Response
Keywords	Michael Clarke, Cricket World Cup, not, 2015
LLaMA-2	Michael Clarke did not play in the 2015 Cricket World Cup. (✗)
Vicuna	Michael Clarke did not participate in the 2015 Cricket World Cup. (✗)
Orca-2	Michael Clarke did not play in the 2015 Cricket World Cup due to a hamstring injury. (✗)
Possible Correct Generation	Michael Clarke did not score a century in the 2015 Cricket World Cup. (✓)
Keywords	not, Finland, Men's Ice Hockey World Championships
LLaMA-2	Finland has never won the Men's Ice Hockey World Championships, despite being one of the top teams in the tournament's history. (✗)
Vicuna	Finland did not win the Men's Ice Hockey World Championships in 2021. (✓)
Orca-2	Finland did not qualify for the Men's Ice Hockey World Championships in 2021. (✗)

Table 15: Examples of responses of various models on the CFG task.

Input/Model	Keywords/Model Response
Keywords	8, not, Oxygen, Atomic Number
GPT-4	The atomic number of Oxygen is not 8. (✗)
Keywords	Roger Federer, not, Wimbledon
GPT-4	Roger Federer has not won Wimbledon since 2017.(✓)
Keywords	2016, Denver Broncos, Super Bowl, Not
GPT-4	In 2016, the Denver Broncos did not lose the Super Bowl; they won it.(✓)
Keywords	Claire Foy, not, emmy awards, 2018
GPT-4	Claire Foy did not win an Emmy Award in 2018.(✗)

Table 16: Examples of responses of GPT-4 on the CFG task.

Question **Options**	Identify the Formula 1 drivers who have not won the Monaco Grand Prix. (A) Max Verstappen, (B) Ayrton Senna, **(C) Carlos Sainz**, (D) Lewis Hamilton, **(E) Lando Norris**
Orca-2 Response	The question is asking for the Formula 1 drivers who have not won the Monaco Grand Prix. The options provided are Max Verstappen, Ayrton Senna, Carlos Sainz, Lewis Hamilton, and Lando Norris. First, let's eliminate the drivers who we know have won the Monaco Grand Prix. Ayrton Senna and Lewis Hamilton are both well-known for their success in Formula 1, including multiple Monaco Grand Prix wins. Therefore, we can eliminate options (B) and (D). Now, let's consider the remaining options: Max Verstappen, Carlos Sainz, and Lando Norris. All three of these drivers are currently active in Formula 1, and it is possible that they have not yet won the Monaco Grand Prix. However, Max Verstappen has won the Monaco Grand Prix in the past, so we can eliminate option (A). This leaves us with two options: Carlos Sainz and Lando Norris. Both of these drivers are still active in Formula 1, and as of the time of this writing, neither of them has won the Monaco Grand Prix. Therefore, both options (C) and (E) are correct. ### Final Answer: (C) Carlos Sainz, (E) Lando Norris
Question **Options**	Select all cricketers who were not part of the 2007 world cup winning team. **(A) Ryan Harris**, (B) Shaun Tait, **(C) Brett Lee**, **(D) Shane Warne**, (E) Brad Hogg
Orca-2 Response	Key concepts and information: - Cricketers - 2007 world cup winning team Breakdown of the question: - Identify which cricketers were not part of the 2007 world cup winning team Relevant information: - The 2007 world cup winning team was Australia Answer choices: (A) Ryan Harris (B) Shaun Tait (C) Brett Lee (D) Shane Warne (E) Brad Hogg Elimination of answer choices: - Shane Warne was part of the 2007 world cup winning team, so eliminate (D) Narrowing down choices: - We are left with (A) Ryan Harris, (B) Shaun Tait, (C) Brett Lee, and (E) Brad Hogg Comparison of remaining choices: - All four remaining choices were not part of the 2007 world cup winning team Final Answer: (A) Ryan Harris, (B) Shaun Tait, (C) Brett Lee, (E) Brad Hogg

Table 17: Examples of Orca-2 responses on the MCQA task.

Name	**Without Negation**	**With Negation**
Virat Kohli	Kohli has won numerous awards for his cricketing achievements, including the ICC Cricketer of the Year award in 2017 and 2018. (✓) He is a right-handed batsman and has been ranked as the number one Test batsman in the world by the International Cricket Council (ICC) on several occasions. (✓) He has scored over 11,000 runs in One Day Internationals (ODIs) and has a batting average of over 50 in this format. (✓)	Kohli has never scored a double century in Test cricket, despite having a career average of over 50 in the format. (✗) 4. Virat Kohli has never won the ICC Cricketer of the Year award, although he has been nominated several times. (Negation: has never won) (✗) Virat Kohli does not have a perfect record in Test cricket, having been dismissed for a duck (zero) on several occasions.(✓)

Table 18: Examples of facts generated by LLaMA-2 for both 'without negation' and 'with negation' prompt. (✗) and (✓) correspond to hallucinated and factually correct statements respectively.

Text	BARD Response
Cristiano Ronaldo was born on February 5, 1985, in Funchal, Madeira, Portugal.	This statement is correct. Here's why: Multiple credible sources report Cristiano Ronaldo's birth date and place as February 5, 1985, in Funchal, Madeira, Portugal. These sources include: Wikipedia: https://en.wikipedia.org/wiki/Cristiano_Ronaldo Britannica: https://www.britannica.com/biography/Cristiano-Ronaldo Biography: https://www.biography.com/athletes/cristiano-ronaldo Ronaldo himself has confirmed his birth date and place in interviews and on social media. Therefore, based on the available evidence, the claim that Cristiano Ronaldo was born on February 5, 1985, in Funchal, Madeira, Portugal, is correct.
Cristiano Ronaldo has never represented Portugal in a World Cup tournament.	The statement that Cristiano Ronaldo has never represented Portugal in a World Cup tournament is incorrect. He has actually represented Portugal in five World Cups: 2006, 2010, 2014, 2018, and 2022. In fact, he is the all-time leading scorer for Portugal in the World Cup with eight goals.
Despite being a prolific scorer, Ronaldo has never won the Golden Boot award at the UEFA Champions League, although he has come close several times.	Incorrect. Explanation: Cristiano Ronaldo has actually won the UEFA Champions League Golden Boot award a record seven times, not never. He holds the distinction of being the top scorer in the competition the most times, surpassing his long-time rival Lionel Messi who has won it six times. Here's a breakdown of Ronaldo's Champions League Golden Boot wins: 2007-08: Manchester United 2013-14: Real Madrid (17 goals, record for most goals in a single season) 2014-15: Real Madrid (tied with Neymar and Messi, all at 10 goals) 2015-16: Real Madrid 2016-17: Real Madrid 2017-18: Real Madrid 2018-19: Juventus Therefore, the claim that Ronaldo has never won the Champions League Golden Boot is factually incorrect.

Table 19: Examples of BARD evaluations.