ReEval: Automatic Hallucination Evaluation for Retrieval-Augmented Large Language Models via Transferable Adversarial Attacks

Xiaodong Yu[♠,*] Hao Cheng[♣] Xiaodong Liu[♣] Dan Roth[♠] Jianfeng Gao[♣]
[♠]University of Pennsylvania [♣]Microsoft Research
https://autodebug-llm.github.io/

Abstract

Despite remarkable advancements in mitigating hallucinations in large language models (LLMs) by retrieval augmentation, it remains challenging to measure the reliability of LLMs using *static* question-answering (QA) data. Specifically, given the potential of data contamination (*e.g.,* leading to memorization), good static benchmark performance does not ensure that model can reliably use the provided evidence for responding, which is essential to avoid hallucination when the required knowledge is new or private. Inspired by adversarial machine learning, we investigate the feasibility of automatically perturbing existing static one for *dynamic* evaluation. Specifically, this paper presents ReEval, an LLM-based framework using prompt chaining to perturb the original evidence for generating new test cases for evaluating the LLMs' reliability in using new evidence for answering.

We implement ReEval using ChatGPT and evaluate the resulting variants of two popular open-domain QA datasets on a collection of LLMs under various prompting settings. Our generated data is human-readable and useful to trigger hallucination in LLM. Accurate models on static data are observed to produce unsupported answers from the perturbed evidence, with pronounced accuracy drops across LLMs including GPT-4. We find that our adversarial examples are transferable across all considered LLMs. The examples generated by a small model can be used to evaluate a much larger model, making our approach cost-effective.

1 Introduction

Due to their superior capability in generating coherent and convincing outputs, large language models (LLMs), such as ChatGPT (OpenAI, 2022), GPT4 (OpenAI, 2023), Claude (Anthropic, 2023) and Palm (Anil et al., 2023), have been extensively used as foundations for language technologies.

Though LLMs excel in memorizing knowledge and understanding natural language, merely depending on parametric knowledge for inquires (closed-book) has inherent limitations. Specifically, these models are unaware of knowledge update and uninformed about new or private information they have not previously encountered. One popular way to mitigate this is to augment LLMs with external relevant evidence (open-book), *e.g.,* retrieval-augmented LLMs (Shi et al., 2023; Peng et al., 2023), outperforming their closed-book counterparts. However, this improvement does not necessarily imply that the model with retrieval augmentation *truly integrates the given evidence for deriving the response*. As most popular datasets used for evaluation are curated using public corpora (*e.g.,* Wikipedia), which are already included in the LLM pretraining, they risk becoming not challenging enough, and models may achieve higher accuracy by mere memorization or by exploiting their familiarity with topics or domains found in static evaluation datasets. Thus, it raises concerns as to whether retrieval-augmented LLMs might resort to fabricating answers that are inconsistent with the presented evidence, resulting in hallucination. Given the wide applications of retrieval-augmented LLMs, it is critical to reliably assess their faithfulness to the context for trustworthy and safe AI, particularly when handling sensitive or recently updated information.

In this work, we propose a new evaluation framework ReEval, which dynamically generate new data to evaluate LLMs. Motivated by using adversarial attacks to trigger undesirable behaviors in machine learning models (Madry et al., 2018; Goodfellow et al., 2014), we focus on perturbing evidence in the prompts to measure the reliability of LLMs' capability of deriving proper responses based on the provided context. Through the perturbation of either the answer span or the rest context in the given evidence, ReEval accordingly provides

[*] Work done during an internship at Microsoft Research.

Findings of the Association for Computational Linguistics: NAACL 2024, pages 1333–1351
June 16-21, 2024 ©2024 Association for Computational Linguistics

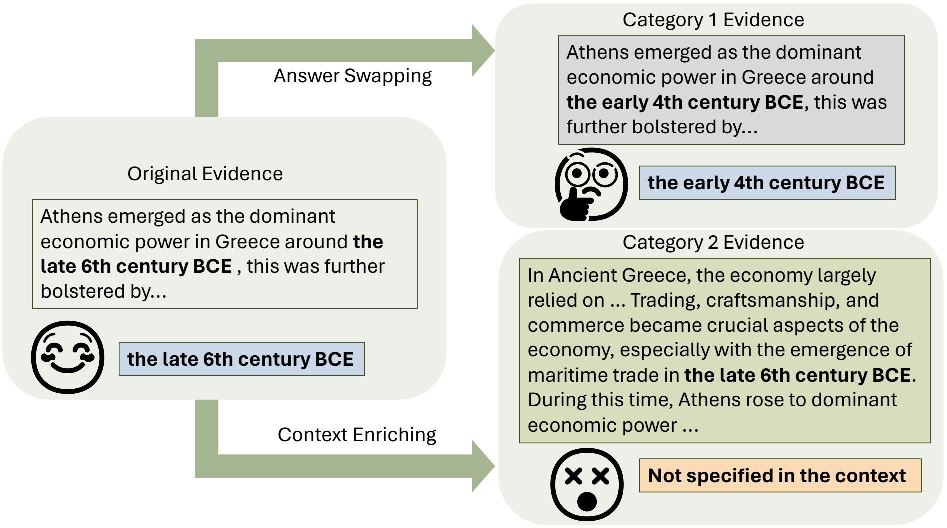
Answer Swapping
Category 1 Evidence
Athens emerged as the dominant economic power in Greece around **the early 4th century BCE**, this was further bolstered by...
the early 4th century BCE
Original Evidence
Athens emerged as the dominant economic power in Greece around **the late 6th century BCE** , this was further bolstered by...
the late 6th century BCE
Category 2 Evidence
In Ancient Greece, the economy largely relied on ... Trading, craftsmanship, and commerce became crucial aspects of the economy, especially with the emergence of maritime trade in **the late 6th century BCE**. During this time, Athens rose to dominant economic power ...
Context Enriching
Not specified in the context

Figure 1: An example of how the original evidence is edited (answer swapping and context enriching) by ReEval. The question is “when did athens emerges as wealthiest greek city state?". “the early 4th century BCE” and “the late 6th century BCE”is the desirable answers for *answer swapping* (Category 1) and *context enriching* (Category 2), respectively. ChatGPT answers are next to the emoji.

two ways of synthesizing evaluation datasets (see examples in Figure 1): 1) *answer swapping* (Category 1), where the original answer is replaced with another valid answer while the remaining context is intact; 2) *context enriching* (Category 2), where more relevant information is added to the provided document while the original supportive information is kept. The former simulates the scenario where only the answer-relevant part of the document is updated while the latter represents the evolving document where more related information is added leading to more complex documentation of specific topics. We then implement ReEval by *prompt chaining* with LLMs, *i.e.,* using LLMs to generate new test cases that are more likely to trigger hallucinations in LLMs.

To verify the effectiveness of the proposed framework, we apply it to two popular open-domain QA dataset, Natural Questions (NQ) (Kwiatkowski et al., 2019) and RealtimeQA (Kasai et al., 2022). Human studies are conducted to verify the naturalness of the generated adversarial attacks, *i.e.,* the updated document is human-readable, supporting the *desirable* answer for the corresponding question. We then evaluate our generated datasets on both open-source (Alpaca (Taori et al., 2023)) and propriety (ChatGPT, Claude, Palm and GPT-4) LLMs under various prompting settings, *e.g.,* zero-shot, few-shot, and more enhanced prompting techniques designed to improve the reliability of prompting with LLMs. Although natural and supportive in the eyes of humans, both probing datasets trigger LLMs to produce inconsistent answers based on the perturbed evidence, regardless of their model sizes and training techniques. We find that the self-attacks are more effective but attacking test examples generated by our method is transferable across all considered LLMs. This enables the possibility of evaluating LLMs using test cases generated by more cost-effective LLMs.

2 Related Work

Faithfulness of Augmented LLM. Recent work shows that, given the correct passages, LLMs could be highly receptive to the provided passage even if the passage is inconsistent with the model memory. For example, (Xie et al., 2023) focus on machine-generated questions from a subject-object-relation triple with machine-generated evidence, and (Zhou et al., 2023) design prompt templates that could force the model to follow the provided context and thus improve the faithfulness of the model. Instead, we use diverse and real-world questions from NQ and focus on editing the passage without compromising the naturalness of the original passages. In addition to including the advanced prompting from (Zhou et al., 2023) in our study, we focus on a more diverse and challenging set of questions rather than a smaller and simpler one with questions that could be answered correctly under the zero-shot closed-

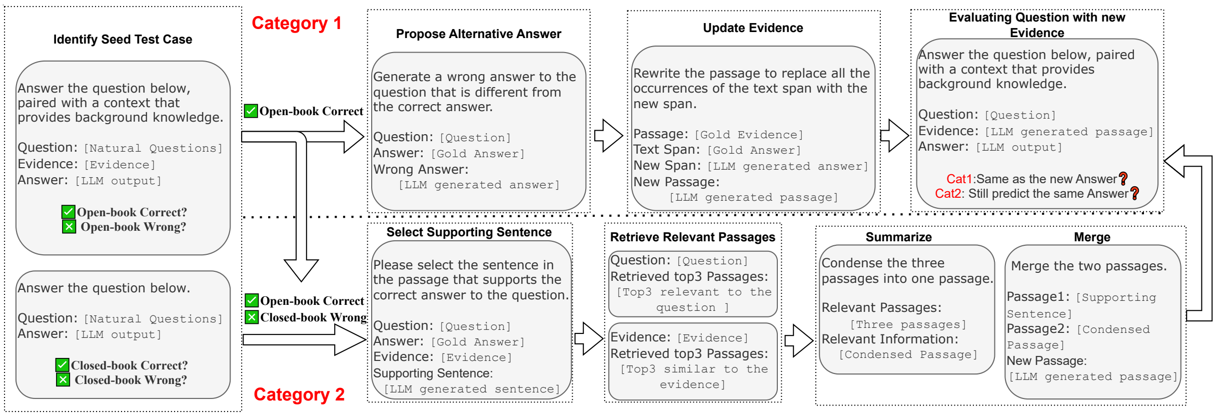
Identify Seed Test Case
Answer the question below, paired with a context that provides background knowledge.
Question: [Natural Questions]
Evidence: [Evidence]
Answer: [LLM output]
☑ Open-book Correct?
☒ Open-book Wrong?
Answer the question below.
Question: [Natural Questions]
Answer: [LLM output]
☑ Closed-book Correct?
☒ Closed-book Wrong?
Category 1
☑ Open-book Correct
☑ Open-book Correct
☒ Closed-book Wrong
Category 2
Propose Alternative Answer
Generate a wrong answer to the question that is different from the correct answer.
Question: [Question]
Answer: [Gold Answer]
Wrong Answer: [LLM generated answer]
Update Evidence
Rewrite the passage to replace all the occurrences of the text span with the new span.
Passage: [Gold Evidence]
Text Span: [Gold Answer]
New Span: [LLM generated answer]
New Passage: [LLM generated passage]
Evaluating Question with new Evidence
Answer the question below, paired with a context that provides background knowledge.
Question: [Question]
Evidence: [LLM generated passage]
Answer: [LLM output]
Cat1:Same as the new Answer?
Cat2: Still predict the same Answer?
Select Supporting Sentence
Please select the sentence in the passage that supports the correct answer to the question.
Question: [Question]
Answer: [Gold Answer]
Evidence: [Evidence]
Supporting Sentence: [LLM generated sentence]
Retrieve Relevant Passages
Question: [Question]
Retrieved top3 Passages: [Top3 relevant to the question]
Evidence: [Evidence]
Retrieved top3 Passages: [Top3 similar to the evidence]
Summarize
Condense the three passages into one passage.
Relevant Passages: [Three passages]
Relevant Information: [Condensed Passage]
Merge
Merge the two passages.
Passage1: [Supporting Sentence]
Passage2: [Condensed Passage]
New Passage: [LLM generated passage]

Figure 2: The pipeline of ReEval, including identifying seed cases, generating new tests, and hallucination evaluation.

book setting. We argue that the difficulty and diversity of the questions as well as the naturalness of evidence passages are crucial for understanding the hallucination of SOTA LLMs for real-world applications. In our framework, we keep the questions natural, and the evidence is from Wikipedia with abundant information. For Category 1 data generation, previous work introduces ideas on altering the entities in the passage (Yan et al., 2021; Longpre et al., 2021; Zhou et al., 2023), while we consider all types of answers (including entities as subcases), and use LLM to automatically substitute the answer properly (making it fit the context). For Category 2 data generation, (Choi et al., 2021) propose to decontextualize the supporting sentence from the passage, and (Jia and Liang, 2017) add distractors to the original passage. In contrast, we want to enrich the original passage by first extracting the supporting sentence with proper decontextualization and then enriching it with other relevant information based on prompting with LLMs.

Adversarial Attacks & Transferability. There is a long line of research in generating adversarial examples to trigger errors or undesirable behaviors from machine learning models (Szegedy et al., 2014; Goodfellow et al., 2014). To improve the robustness of machine learning models, there are also a number of methods proposed to defend against such attacks (Madry et al., 2018; Zhu et al., 2020; Li and Qiu, 2020; Cheng et al., 2021). However, models trained with adversarial learning are found to have at-odds generalization (Tsipras et al., 2019; Zhang et al., 2019), *e.g.,* improving the accuracy on adversarial attacks can compromise the model performance on clean examples. Despite being more challenging due to its discrete nature, different text adversarial attacks with perturbed inputs imperceptible to humans have been proposed for question answering (Jia and Liang, 2017), natural language inference (Nie et al., 2020), and sentiment classification (Iyyer et al., 2018). One surprising phenomenon is that many adversarial examples are *transferable* (Papernot et al., 2016; Wallace et al., 2021). For example, Wallace et al. (2021) show that adversarial prefix optimized for one particular model can also transfer to models of different architectures and sizes. In addition to relying on white-box access to generate effective adversarial examples, recent work even reports that it is difficult to generate reliable examples via automatic search (Carlini et al., 2023). Our work is highly motivated by this long line of work, *i.e.,* making evidence edits while keeping the input legitimate for the targeted task so that the LLMs cannot reliably answer the question. Here, we do not assume any model access except its text outputs, *i.e.,* black-box. We show that our proposed approach of generating adversarial test cases from a pivot LLM can trigger hallucination behaviors across a set of open-source and proprietary LLMs.

3 ReEval Framework

Assessing the hallucination of LLMs is challenging as we often do not know what changes in the prompt would trigger LLMs to hallucinate. In this paper, we present our approach ReEval for automatically constructing a large number of test cases that can surface hallucination issues. Given a pivot LLM, we first prompt it to identify *seed test cases* from a pool of existing data. Then we prompt the pivot LLM again to generate *attacking test cases* based on individual seed test cases. These attacking test cases are used to evaluate the performance of the pivot LLM (**self-attack**) as well as other LLMs (**cross-attack**). While `ReEval` is a general framework, we focus on the QA scenario where the LLMs to be evaluated need to answer open-domain questions based on their supporting evidence. The pipeline is illustrated in Figure 2.

Seed Case Selection. To identify seed test cases, we categorize QA examples based on whether the pivot LLM can answer the question correctly under the open-book and closed-book settings in a zero-shot fashion, similar to typical static evaluation. In the closed-book setting, only the question itself is given and the pivot LLM can only rely on memorization, whereas in the open-book setting, the associated supporting evidence is provided. As we are interested in assessing **whether the LLM can truely comprehend the provided evidence and reliably use that for answering**, only cases that can be answered correctly using open-book prompt are kept as seed. For those cases, `ReEval` generates attacking test cases by perturbing the evidence, potentially updating the answers (*e.g.,* answer swapping). Below is the zero-shot open-book prompt for seed test case selection, and the closed-book version simply drops the evidence part (see more examples in Appendix).

Zero-shot Open-book Prompt

Answer the question below, paired with a context that provides background knowledge. Only output the answer without other context words.
Context: {Evidence}
Question: {Question}
Answer:

Evidence Perturbation. To generate viable attacking test cases, we consider the following two perturbation approaches.

1. **Answer Swapping** (top flow in Figure 2): Update the evidence using a new answer that may lead to a knowledge conflict (§3.1). In the top-right example of Figure 1, we replace *"the late 6th century BCE"* with *"the early 4th century BCE"* in the evidence and test whether the LLM can update its answer accordingly.

2. **Context Enriching** (bottom in Figure 2): Enrich the evidence using extra relevant facts that may dilute the information (§3.2). In the bottom-right example of Figure 1, the evidence becomes much more dense though the answer is unchanged, and we test whether the LLM can still produce the original answer.

For the second approach, we exclude cases where the pivot LLM can answer correctly under the closed-book setting since perturbing the evidence for such cases may not surface the hallucination issue, *i.e.,* the LLM may simply use its internal memory to answer the question correctly and completely ignore the evidence.

Re-evaluation. To assess the hallucination of LLMs, we can simply measure the accuracy of the predicted answers for the attacking test cases. If the LLM faithfully follows the provided context, it should be immune to these perturbations and maintain a high accuracy score. The evaluation considers both zero-shot and few-shot prompting. The zero-shot prompt for evaluation is identical to the one used for seed test selection above. The few-shot version inserts the demonstrations of evidence-question-answer triplets right before the "Context: {Evidence}" line.

Few-shot Open-book Prompt

Answer the question below, paired with a context that provides background knowledge. Only output the answer without other context words.
{Demonstrations of Evidence-Question-Answer tuples}
Context: {Evidence}
Question: {Question}
Answer:

3.1 Category 1: Answer Swapping

Here, we present the first approach to generate test cases by updating the original evidence with alternative answers. Specifically, those alternative answers are proposed by the pivot LLM via prompting. Note that the considered seed test cases are open-book correct with the pivot LLM.

For each question, given the original answer and supportive evidence, we first ask the model to generate an alternative answer that is factually wrong using the following prompt.

Prompt for Generating An Alternative Answer

Generate a wrong answer to the question that is different from the correct answer.
Question: {Question}
Answer: {Gold Answer}
Wrong Answer:

We then instruct the LLM to replace all the occurrences of the original answer with the alternative one.[1]

[1] Although a simple string match can also do the job, it can make the answer occurring sentences inconsistent with the neighboring context, *e.g.,* mismatched pronouns and aliases.

Prompt for Updating Evidence

Rewrite the passage to replace all the occurrences of the text span with the new span.
Passage: {Original Evidence}
Text Span: {Original Answer}
New Span: {LLM generated answer}
New Passage:

Since most context is kept, the newly generated evidence is likely to support the alternative answer for most questions (as verified in §4.3).

3.2 Category 2: Context Enriching

Our second strategy aims to enrich the original evidence with more relevant context, leading to a more complex context for answer reasoning. Unlike Category 1 discussed above, we only keep seed cases that are open-book correct but closed-book wrong to ensure that certain comprehension of the evidence is required to answer the question correctly.

To ensure that the newly generated evidence still provides support for the question, we first extract the supporting sentence from the original evidence.

Prompt for Selecting the Supporting Sentence

Please select the sentence in the passage that supports the correct answer to the question.
Question: {Question}
Answer: {Answer}
Evidence: {Evidence}
Supporting Sentence:

We then gather relevant information from an external database to be used for composing the new evidence. Here, we consider two ways of retrieving passages from Wikipedia for fusion with the supporting sentence above, *i.e.,* evidence-focused expansion and question-focused expansion, where the former uses the original evidence as the query and the question is used for the latter case. As these two expansions bring in different types of relevant information, we create two corresponding copies of new evidence. To make the information more diverse, we select the top-k passages from different Wikipedia pages. To merge these passages into a single passage, we first ask the LLM to summarize the information of the retrieved set, and then merge the supporting sentence into the summary. Here, the pivot LLM needs to extract and summarize key information so that the new evidence is human-readable and still supports the original answer.

Summarize Prompt

Condense the three passages into one passage.
Relevant Passages: {List of Passages}
Relevant Information:

Merge Prompt

Merge the two passages
Passage1: {Supporting Sentence}
Passage2: {Condensed Passage}
New Passage:

4 Experiments

4.1 Experiment Settings

Evaluation Metrics. Three evaluation metrics are reported, *i.e.,* exact match (EM) accuracy, token-level F1, and entailment accuracy. The first two metrics are traditionally used for evaluating QA models. However, they tend to be too strict for evaluating LLM-generated responses, since LLMs often produce long and verbose sequences to explain the answers (partially due to their alignment procedure). The entailment accuracy is a more lenient metric that checks whether "Question + LLM Output" can entail "Question + Answer". In this paper, we use an entailment model `nli-deberta-v3-base`[2] from Sentence-BERT (Reimers and Gurevych, 2019), which is mostly reliable based on our manual inspection. Since we use the pivot model to select the seed cases, the accuracy of other models on the original set is not guaranteed to be 100. To clearly reveal the performance difference, we also report "Normalized Entailment" accuracy, where we normalize the test set to the cases that the corresponding model could answer correctly before perturbation.

Source Data. We use the MRQA version (Fisch et al., 2019) of Natural Questions (Kwiatkowski et al., 2019) and RealTimeQA data (Kasai et al., 2022) from 20220613 to 20231110. and conduct the following filtering steps: 1) remove duplicated Question-Evidence-Answer triplets and only keep one unique instance, 2) remove all evidence passages that are shorter than 10 words, 3) remove all cases with answers longer than 5 words. After this, 7189 instances from NQ and 1380 instances from RealtimeQA are kept. For questions with multiple answers, if the answers are overlapping (*e.g.,* "1871" and "1871 A.D."), we randomly keep one, otherwise, the corresponding examples are removed. Note the same question may still appear in multiple instances because the supporting evidence can be different.

[2] `https://huggingface.co/cross-encoder/nli-deberta-v3-base`

Models	Method	Zero-shot				Few-shot			
		EM	F1	Entail.	Norm Entail.	EM	F1	Entail.	Norm Entail.
Alpaca-7B	Open-book	18.71	36.04	56.65	71.68	21.50	38.46	57.30	67.45
	Faithful Prompt	27.80	43.64	58.75	68.86	33.74	51.10	65.41	74.33
ChatGPT	Open-Book	43.71	59.99	77.31	77.31	40.44	54.58	65.33	65.33
	Faithful Prompt	44.73	40.04	42.98	42.98	40.04	52.75	62.11	62.11
Claude 2	Open-Book	44.62	56.37	59.08	-	20.32	34.09	69.77	-
	Faithful Prompt	52.95	65.05	71.80	-	39.28	50.97	71.83	-
Palm	Open-Book	57.50	65.75	74.71	80.13	65.75	75.74	78.41	83.38
	Faithful Prompt	64.17	68.41	79.20	84.19	68.41	78.61	81.46	86.15
GPT-4	Open-Book	54.11	68.50	81.29	84.73	58.94	72.58	81.01	83.79
	Faithful Prompt	58.49	71.70	82.51	85.52	63.49	75.72	82.25	85.19

Table 1: Zero-shot and few-shot performance of LLMs on Category 1 data of NQ. "Entail." refers to the entailment accuracy. "Norm Entail." refers to the entailment accuracy of the normalized test set that only includes the accurate cases before perturbation.

Models	Method	Zero-shot				Few-shot			
		EM	F1	Entail.	Norm Entail.	EM	F1	Entail.	Norm Entail.
Alpaca-7B	Open-Book	41.64	51.97	74.47	79.74	31.83	43.41	68.67	73.85
	Faithful Prompt	46.05	56.60	76.28	79.74	49.55	61.97	78.08	79.62
ChatGPT	Open-Book	60.06	71.97	84.38	84.38	55.96	68.57	81.08	81.08
	Faithful Prompt	55.16	66.97	80.38	80.38	56.86	68.95	81.18	81.18
Palm	Open-Book	56.26	65.46	73.47	75.03	67.07	74.35	78.78	80.10
	Faithful Prompt	72.17	79.54	82.78	84.46	72.97	79.18	83.18	84.87
GPT-4	Open-Book	66.97	77.81	88.59	90.88	66.17	77.90	88.39	90.04
	Faithful Prompt	66.07	76.57	86.89	89.31	70.77	80.79	88.99	91.19

Table 2: Zero-shot and few-shot performance of LLMs on Category 1 data of RealtimeQA.

Generated Data. Unless otherwise specified, ChatGPT (`gpt-3.5-turbo-0301`) is the pivot LLM for identifying seed test cases and generating attacking test cases. When identifying seed test cases, we treat an answer produced by the pivot LLM as correct if it matches the reference answer exactly or can entail the reference answer in the same way as we compute the entailment accuracy. The retriever used for generating Category 2 cases is based on `all-mpnet-base-v2`[3]. In total, we obtain **3,539** and **2,211** attacking test cases in Category 1 and Category 2 of NQ, and **1,000** and **814** attacking test cases in Category 1 and Category 2 of RealtimeQA respectively.

We evaluate five popular LLMs using the generated attacking test cases: Alpaca-7B (Taori et al., 2023), ChatGPT (`gpt-3.5-turbo-0301`), Claude2, PaLM, and GPT-4 (`gpt-4-0613`), which is considered to be the state-of-the-art (SOTA) LLM. In the few-shot setting, 5 static demonstration examples are used.

4.2 Main Results

We evaluate the five LLMs on the Category 1 and Category 2 data generated by ChatGPT, including both self-attack and cross-attack scenarios. [4] In addition to vanilla zero-shot and few-shot promptings, we consider the recently proposed faithfulness-promoting prompting, *i.e.,* the opinion-based prompt by (Zhou et al., 2023). For each model, we evaluate its closed-book performance, open-book performance, and open-book with faithful prompting performance. The full list of various prompts and error examples is in Appendix.

Category 1. Here, the model is expected to follow the given context, and predict the *altenative answer* proposed by the pivot model. The results are summarized in Table 1 and Table 2. As expected, the model resistance towards our attack is mostly correlated with its model size and capability. Specifically, larger and more capable models are more robust, *e.g.,* GPT-4 is more reliable than

[3] `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

[4] Some numbers of Claude 2 are missing because we lost the access to the model due to Anthropic policy.

Models	Method	Zero-shot				Few-shot			
		EM	F1	Entail.	Norm Entail.	EM	F1	Entail.	Norm Entail.
Alpaca-7B	Open-Book	9.27	39.35	42.79	56.48	14.52	45.56	47.40	58.53
	Faithful Prompt	15.06	43.65	42.65	54.10	20.58	53.40	50.88	60.41
ChatGPT	Open-Book	25.51	57.15	61.78	61.78	27.32	58.94	51.15	51.15
	Faithful Prompt	24.69	53.49	50.38	50.38	24.20	56.26	44.10	44.10
Claude 2	Open-Book	29.99	58.69	43.46	-	12.12	39.83	57.26	-
	Faithful Prompt	35.78	64.89	52.60	-	27.45	54.31	54.68	-
Palm	Open-Book	44.78	71.76	66.76	75.70	50.84	75.23	66.53	75.58
	Faithful Prompt	44.78	70.18	58.75	66.03	47.35	72.03	61.78	69.01
GPT-4	Open-Book	37.68	67.27	68.39	73.55	46.27	74.17	73.04	77.95
	Faithful Prompt	33.60	62.78	58.25	62.36	45.59	72.83	67.57	72.46

Table 3: Zero-shot and few-shot performance of LLMs on Category 2 Data of NQ.

Models	Method	Zero-shot				Few-shot			
		EM	F1	Entail.	Norm Entail.	EM	F1	Entail.	Norm Entail.
Alpaca-7B	Open-Book	31.57	54.69	71.50	77.87	25.92	46.43	49.02	52.69
	Faithful Prompt	38.45	59.31	71.50	75.55	25.80	49.18	60.44	63.03
ChatGPT	Open-Book	42.87	63.77	72.73	72.73	36.00	56.56	64.25	64.25
	Faithful Prompt	31.82	53.06	60.81	60.81	44.10	66.63	75.18	75.18
Palm	Open-Book	62.04	79.14	89.31	91.43	59.46	79.09	85.38	87.47
	Faithful Prompt	67.20	82.51	85.87	87.72	64.86	82.09	84.15	86.06
GPT-4	Open-Book	50.37	71.45	78.38	80.18	58.48	77.26	86.12	87.13
	Faithful Prompt	41.65	61.53	65.36	66.92	57.49	76.28	82.43	84.04

Table 4: Zero-shot and few-shot performance of LLMs on Category 2 Data of RealtimeQA.

Alpaca-7B. Although GPT-4 is the most powerful model, it is still not immune to our attacks, indicating the effectiveness of our approach to trigger hallucination in SOTA LLMs. Though using the human-designed faithful prompt or using in-context examples helps the performance in some cases, there are no consistent improvements compared with zero-shot in general.

Category 2. We require the model to fully understand both the question-focused expansion and evidence-focused expansion cases, and one question is considered correct only when both are answered correctly. We report the merged result in Table 3 and Table 4, and we also report the few-shot performance on each case separately in Table 13 of Appendix. As we can see, there are large performance drops for all models, suggesting they fail to identify the relevant evidence information regardless of prompting techniques. Similar to Category 1, the faithful prompt is observed to have no consistent benefits, which calls for future work to develop more reliable prompting techniques.

4.3 Human Evaluations

To evaluate whether the evidence generated by `ReEval` is supportive and human-readable, we randomly sample 500 cases from Category 1, 1000 cases from Category 2 with 500 examples for question-focused expansion, and 500 for evidence-focused expansion. We use Amazon Mechanical Turk to collect human judgments on this set. Each question is judged by three annotators, who are asked to read the evidence and decide whether it could support them to get the correct answer. To prevent annotators from randomly submitting “Yes” or “No”, 10% of the data is used as validation checks where we know whether the evidence supports the answer. We only accept annotations from the annotators with at least 90% accuracy on the validation check. For each question, if the majority of the annotators think the generated evidence is supportive, it is then counted as human-readable. For all three categories, around 90% of the cases are human readable, supporting the quality of `ReEval`, with 90.8%, 92.4%, and 88.8% human-readable ratios for Category 1, Category 2 question-focused and evidence-focused, respectively.

Models	Method	ChatGPT			GPT-4			Alpaca-7B		
		EM	F1	Entail.	EM	F1	Entail.	EM	F1	Entail.
Alpaca-7B	Open-Book	25.00	40.57	61.20	26.8	43.88	68.2	26.00	43.95	65.80
	Faithful Prompt	37.20	53.46	72.20	39.60	57.49	76.00	36.60	53.93	70.80
ChatGPT	Open-Book	43.00	54.88	66.20	49.60	61.55	71.60	38.40	51.56	61.40
	Faithful Prompt	42.80	53.25	61.80	51.40	61.53	70.40	40.00	52.57	61.20
Palm	Open-Book	70.80	78.51	81.40	75.80	82.58	86.00	67.00	74.55	79.00
	Faithful Prompt	74.20	82.00	84.40	78.80	85.28	89.00	69.20	77.73	82.80
GPT-4	Open-Book	65.20	76.66	84.00	59.20	69.18	76.40	57.00	67.23	73.80
	Faithful Prompt	69.80	79.04	84.80	67.40	75.98	81.80	59.60	70.15	78.40

Table 5: Few-shot case study of backbone LLMs used by `ReEval` (500 examples). The column blocks indicate the Category 1 data generated by ChatGPT, GPT-4, and Alpaca-7B, respectively

Models	Method	ChatGPT			GPT-4		
		EM	F1	Entail.	EM	F1	Entail.
Alpaca-7B	Open-Book	17.80	44.85	52.20	9.00	37.16	42.40
	Faithful Prompt	22.40	53.96	57.00	16.00	46.28	43.80
ChatGPT	Open-Book	29.40	57.12	50.80	23.20	50.76	46.20
	Faithful Prompt	24.40	54.61	41.60	23.20	52.80	43.20
Palm	Open-Book	54.40	76.84	69.60	52.20	73.62	66.40
	Faithful Prompt	53.40	75.93	68.60	48.4	71.91	62.60
GPT-4	Open-Book	49.40	74.38	74.20	24.00	47.18	37.60
	Faithful Prompt	51.80	73.68	71.00	35.00	62.04	52.40

Table 6: Few-shot case study of backbone LLMs used by `ReEval` (500 examples). The column blocks indicate the Category 2 data generated by ChatGPT and GPT-4, respectively

4.4 Case Studies

Is `ReEval` sensitive toward backbone LLMs? To do that, we use alternative LLMs to generate attacking test cases other than ChatGPT. We consider both Alpaca-7b and GPT-4 for Category 1 and only GPT-4 for Category 2 given the task is more demanding. Due to the limitation of budget, we randomly sample 500 examples from NQ for this study. All prompts are similar to those used previously. The few-shot performances of Category 1 and Category 2 are reported in Table 5 and Table 6, respectively. As shown in Table 5, compared with ChatGPT and Alpaca, GPT-4 does not generate stronger attacks. This is probably because the alternative answers from GPT-4 are more receptive to all models. The Category 1 data generated by the smallest model (Alpaca-7B) appears to be very effective for those two larger ones, but we observe that that is because Alpaca sometimes generates invalid answers and also fails to replace all the occurrences of the old answer. On the other hand, compared with ChatGPT, GPT-4 can generate more stronger attacks for Category 2 (Table 6). We find that GPT-4 is better at summarizing multiple pieces of information, leading to more complex evidence. Although all three models are most vulnerable to self-attacks, all `ReEval` attacks are transferable, making it possible to generate attacking test cases using more cost-effective models.

Is `ReEval` sensitive toward the position of the answer? To get the distribution of the answer in the evidence, we only keep the cases where the answer only occurs once in the evidence (2678 in total). There are 55.94% cases where the answer is in the first 1/3 of the evidence, 23.64% cases where the answer is in the middle part of the evidence, and 20.43% cases where the answer is in the last 1/3 of the evidence. We evaluate the accuracy of different models under both few-shot and open-book setting in these 3 cases, and we do not see any significant performance difference except that Alpaca-7B performs worse when the answer is at the end of the evidence. More detailed results are in Table 14 in the Appendix.

5 Conclusion

In this paper, we present `ReEval`, an LLM-based framework that generates transferable adversarial attacks to assess the hallucination of retrieval-augmented LLMs. By swapping the answer in the evidence or adding more relevant information to enrich the context, we successfully trigger hallucination behaviors of existing state-of-the-art LLMs. ReEval is a viable approach in that it can generate transferable attacking examples using more cost-effective LLMs. We believe ReEval could be used to help assess the hallucination of future LLMs, and potentially help mitigate hallucinations. Future directions include further studying ReEval on tasks of different complexities and how to use ReEval for debugging LLM-based applications.

6 Limitations

Although we find our framework effective in evaluating the reliability of retrieval-augmented LLMs, there are some limitations worth discussion here.

First, this study distinctly concentrates on questions with short answers, thereby delineating an intentional boundary from engaging in the exploration of long-form question-answering. For long-form cases, it requires more complex ways of perturbing evidence, *e.g.,* multiple sentences are required to be updated at the same time. The comprehensive investigation into long-form question-answering is deferred to future scholarly endeavors, marking a deliberate scope restriction to refine the focus and depth of the current analysis.

Moreover, the scope of our research rigorously limits its examination to single-hop questions. Consequently, this study does not venture into the evaluation of complex reasoning inaccuracies, often referred to as reasoning hallucinations, which is more likely for multi-hop questions. This delineation underscores a focused approach, yet acknowledges the complexity and necessity of future investigations into multi-hop question-answering, with the need for specialized methodologies to assess and mitigate reasoning errors in such contexts.

In terms of methodology, our study either introduces perturbations within the answer span or modifies the adjacent contextual narrative; however, scenarios that encompass both an altered answer span and a significantly adjusted surrounding context are not within the purview of this investigation. This strategic decision enables the isolation and better understanding of the effects of each type of perturbation independently. Nonetheless, it also marks a critical avenue for further intricate research toward evaluating the compound impacts.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Anthropic. 2023. Claude 2.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, and Ludwig Schmidt. 2023. Are aligned neural networks adversarially aligned?

Hao Cheng, Xiaodong Liu, Lis Pereira, Yaoliang Yu, and Jianfeng Gao. 2021. Posterior differential regularization with f-divergence for improving model robustness. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1078–1089, Online. Association for Computational Linguistics.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of 2nd Machine Reading for Reading Comprehension (MRQA) Workshop at EMNLP*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime qa: What's the answer right now? *arXiv preprint arXiv:2207.13332*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Linyang Li and Xipeng Qiu. 2020. Textat: Adversarial training for natural language understanding with token-level perturbation. *arXiv preprint arXiv:2004.14543*.

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*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

OpenAI. 2022. ChatGPT.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2021. Universal adversarial triggers for attacking and analyzing nlp.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, and Xiang Ren. 2021. On the robustness of reading comprehension models to entity renaming. *arXiv preprint arXiv:2110.08555*.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.

A Appendix

Here, we provide examples of prompt implementations and additional results.

First, we provide selected instances for few-shot prompting in Table 7. In addition, prompts used for Category 1 and Category 2 data generations are listed in Table 8 and Table 9 respectively. Different prompting methods for different language models are detailed in Table 10 (Closed-book), Table 11 (Open-book) and Table 12 (Faithful prompting).

Lastly, more experiment results are in Table 13 (breakdown results of Category 2 experiments in subsection 4.2) and Table 14 (detailed results for section 4.4).

Question: who sings what lovers do with maroon 5

Evidence: “ What Lovers Do ” is a song by American pop rock band Maroon 5 featuring American R&B singer SZA . It was released on August 30 , 2017 , as the lead single from the band ’s sixth studio album Red Pill Blues (2017) . The song contains an interpolation of the 2016 song “ Sexual ” by Neiked featuring Dyo , therefore Victor Rådström , Dyo and Elina Stridh are credited as songwriters .

Answer: American R&B singer SZA

Question: who plays lead guitar on i want you she ’s so heavy

Evidence: John Lennon – lead and harmony vocals , multi-tracked lead guitar , Moog synthesizer Paul McCartney – harmony vocals, bass George Harrison – harmony vocals , multi-tracked lead guitar Ringo Starr – drums , congas , wind machine Billy Preston – Hammond organ

Answer: John Lennon

Question: a long chain of amino acids linked by peptide bonds is a

Evidence: The covalent chemical bonds are formed when the carboxyl group of one amino acid reacts with the amino group of another . The shortest peptides are dipeptides , consisting of 2 amino acids joined by a single peptide bond , followed by tripeptides , tetrapeptides , etc . A polypeptide is a long , continuous , and unbranched peptide chain . Hence , peptides fall under the broad chemical classes of biological oligomers and polymers , alongside nucleic acids , oligosaccharides and polysaccharides , etc .

Answer: polypeptide

Question: when does the school year start in france

Evidence: In Metropolitan France , the school year runs from early September to early July . The school calendar is standardised throughout the country and is the sole domain of the ministry .

Answer: early September

Question: which city is selected under hriday scheme in karnataka

Evidence: With a duration of 4 years (completing in November 2018) and a total outlay of 500 crore (US $78 million) , the Scheme is set to be implemented in 12 identified Cities namely , Ajmer , Amaravati , Amritsar , Badami , Dwarka , Gaya , Kanchipuram , Mathura , Puri , Varanasi , Velankanni and Warangal .

Answer: Ajmer

Table 7: Five Randomly Selected Demo Instances from NQ Training Data for Few-shot Experiments.

Generate Alternative Answer Prompt	A question and its correct answer is below. Generate a wrong answer to the question that is different from the correct answer. Make sure the wrong answer is short, and has the same type as the correct answer. Question: {Question} Answer: {Answer} Wrong Answer:
Replace Old Answer Prompt	A passage and a text span inside the passage is shown below. Rewrite the passage to replace all the occurrences of the text span with the new span. Passage: {Passage} Text Span: {Answer} New Span: {Alternative Answer} New Passage:

Table 8: Prompts for Cat1 Data Generation.

Select Supporting Sentence Prompt	A question, the answer, and a passage are shown below. Please select the sentence in the passage that supports to answer the question correctly. Question: {Question} Answer: {Answer} Passage: {Passage} Sentence:
Summarize Relevant Passages Prompt	Three relevant passages are shown below. Please condense the three passages into one passage. Relevant Passages: [1]: {Passage 1} [2]: {Passage 2} [3]: {Passage 3} Relevant New Information:
Merge Prompt	Two passages and a span are shown below. Please merge the two passages, and make sure to keep the span in the new passage. Passages: [1]: {Supporting Sentence} [2]: {Summarized Passage} Span: {Answer} New Passage:

Table 9: Prompts for Cat2 Data Generation.

Alpaca-7B	Below is an instruction that describes a task. Write a response that appropriately completes the request. Only output the answer without other context words. ### Instruction: {Question} ### Response:
PaLM	You are a helpful and informative bot that answers questions Be sure to respond in a complete sentence, being comprehensive, including all relevant background information. However, you are talking to a non-technical audience, so be sure to break down complicated concepts and strike a friendly and convers- tional tone. Only output the answer without other context words. QUESTION: {Question} ANSWER:
Claude 2	Human: Answer the question below. Only output the answer without other context words. Question: {Question} Assistant:
ChatGPT & GPT-4	system: You are a helpful assistant. user: Answer the question below. Only output the answer without other context words. Question: {Question} Answer:

Table 10: Closed-Book QA prompts for all considered models following their corresponding recommendations.

Alpaca-7B	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. Only output the answer without other context words. ### Instruction: {Question} ### Input: {Evidence} ### Response:
PaLM	You are a helpful and informative bot that answers questions using text from the reference passage included below. Be sure to respond in a complete sentence, being comprehensive, including all relevant background information. However, you are talking to a non-technical audience, so be sure to break down complicated concepts and strike a friendly and conversational tone. If the passage is irrelevant to the answer, you may ignore it. Only output the answer without other context words. QUESTION: {Question} PASSAGE: {Evidence} ANSWER:
Claude 2	Human: Answer the question below, paired with a context that provides background knowledge. Only output the answer without other context words. Context: {Evidence} Question: {Question} Assistant:
ChatGPT & GPT-4	system: You are a helpful assistant. user: Answer the question below, paired with a context that provides background knowledge. Only output the answer without other context words. Context: {Evidence} Question: {Question} Answer:

Table 11: Open-Book Inference Prompts for Different Models Following their Official Instructions.

Alpaca-7B	Instruction: read the given information and answer the corresponding question. Only output the answer without other context words. ### Instruction: Bob said, “{Evidence}” Q: {Question} in Bob’s opinion based on the given text? ### Response:
PaLM	Instruction: read the given information and answer the corresponding question. Only output the answer without other context words. Bob said, “{Evidence}” Q: {Question} in Bob’s opinion based on the given text?
Claude 2	Human: Instruction: read the given information and answer the corresponding question. Only output the answer without other context words. Bob said, “{Evidence}” Q: {Question} in Bob’s opinion based on the given text? Assistant:
ChatGPT & GPT-4	system: You are a helpful assistant. user: Instruction: read the given information and answer the corresponding question. Only output the answer without other context words. Bob said, “{Evidence}” Q: {Question} in Bob’s opinion based on the given text?

Table 12: Opinion-based Inference Prompts for Different Models Following (Zhou et al., 2023)

Models	Method	Few-shot Question Only			Few-shot Evidence Only		
		EM	F1	Entail.	EM	F1	Entail.
Alpaca-7B	Closed-Book	2.67	13.45	13.30	2.40	13.35	12.89
	Open-Book	23.38	44.94	60.65	24.56	46.18	62.87
	Faithful Prompt	30.94	51.88	63.50	33.06	54.93	66.21
ChatGPT	Closed-Book	9.81	25.02	22.03	9.45	24.78	21.66
	Open-Book	40.93	59.10	67.89	40.66	58.78	67.03
	Faithful Prompt	40.89	57.59	64.22	38.22	54.94	60.88
Claude 2	Closed-Book	6.24	19.49	22.75	6.11	19.39	22.70
	Open-Book	22.16	39.63	71.73	22.21	40.03	73.95
	Faithful Prompt	38.13	53.17	68.70	39.35	55.45	70.78
Palm	Closed-Book	11.99	25.23	21.26	11.99	25.23	21.26
	Open-Book	58.44	72.89	73.45	61.96	77.58	78.11
	Faithful Prompt	55.63	70.15	70.28	58.48	73.90	73.32
GPT-4	Closed-Book	20.76	38.04	36.14	20.62	37.98	35.55
	Open-Book	54.23	72.85	80.69	56.54	75.48	83.31
	Faithful Prompt	54.95	71.76	77.25	57.08	73.89	78.79

Table 13: Few-shot result of Question-based Cat2 data and Evidence-based Cat2 data.

Models	Start (< 1/3)	Middle (1/3 - 2/3)	End (> 2/3)
Alpaca-7B	63.68	54.98	50.64
ChatGPT	68.22	68.09	70.02
Claude 2	71.56	72.04	70.93
Palm	79.24	80.09	83.54
GPT-4	82.84	82.46	83.36

Table 14: Few-shot entailment accuracy of Cat1 data. “Start”, “Middle” and “End” indicates the position of the answer span in the evidence.

Cat 1	**Question**: what is the baby elephants name in jungle book **Evidence**: Dumbo - The baby elephant who is the son of Hathi and Winifred and is a good friend of Mowgli. He is voiced by Clint Howard in the first movie and by Jimmy Bennett in The Jungle Book 2 **Answer**: Dumbo **GPT4 Output**: Hathi
Cat 1	**Question**: who brought the idea of castles to england **Evidence**: Castles served a range of purposes , the most important of which were military , administrative , and domestic . As well as defensive structures , castles were also offensive tools which could be used as a base of operations in enemy territory . Castles were established by British rulers of England for both defensive purposes and to pacify the country 's inhabitants . As William the Conqueror advanced through England , he fortified key positions to secure the land he had taken . Between 1066 and 1087 , he established 36 castles such as Warwick Castle , which he used to guard against rebellion in the English Midlands **Answer**: British rulers **GPT4 Output**: William the Conqueror
Cat 1	**Question**: baga beach is in north or south goa **Evidence**: Baga Beach is a popular beach and tourist destination in South Goa. Baga is located at the north end of the contiguous beach stretch that starts from Sinquerim, Candolim, leads to Calangute and then to Baga **Answer**: South Goa **GPT4 Output**: North Goa
Cat 2 Query-based	**Question**: how long prime minister stay in office canada **Evidence**: The Prime Minister of Canada is appointed by the Governor General on the advice of the Prime Minister and serves for an indefinite term, usually around 5 years. The Lieutenant Governors at the provincial level are appointed in a similar manner and also serve for approximately 5 years. The territories have Commissioners who are appointed by the federal cabinet and conventionally serve for about 5 years. The Prime Minister of Canada is the head of government and chooses the ministers that make up the Cabinet. The current Prime Minister, Justin Trudeau, remains in office until he resigns, is dismissed, or dies. **Answer**: until he or she resigns , is dismissed , or dies **GPT4 Output**: Usually around 5 years
Cat 2 Evidence-based	**Question**: what percentage of the earth 's surface is water **Evidence**: Because the oceans that cover roughly 78% of the area of the Earth reflect blue light, the Earth appears blue from space, and is often referred to as the blue planet and the Pale Blue Dot. The Earth's water is distributed across various sources, with oceans holding 97% of surface water, glaciers and polar ice caps holding 2.4%, and other land surface water such as rivers, lakes, and ponds holding 0.6%. Only a small portion of water is contained in aquifers, vapor, clouds, precipitation, biological bodies, and manufactured products. The total volume of water on Earth is estimated to be 1.386 billion km0̆0b3, with 97.5% being saltwater and 2.5% being freshwater. Of the freshwater, only 0.3% is liquid on the surface, while the rest may be present in the lower mantle of the Earth. The United Nations Convention on the Law of the Sea defines all of the ocean as "sea," making Earth the only known planet with liquid water on its surface. Additionally, Earth's water distribution, including oceans, ice caps, and clouds, gives it a distinct blue appearance when viewed from space. Approximately 97.2% of Earth's known water is contained within the seas, which cover more than 70% of its surface. **Answer**: 78% **GPT4 Output**: 70%

Table 15: Error Examples of GPT-4 under the Few-shot Setting.