ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence

Kevin Wu*
Department of Biomedical Data Science
Stanford University
Stanford, CA 94305
kevinywu@stanford.edu

Eric Wu*
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
wue@stanford.edu

James Zou
Department of Biomedical Data Science
Stanford University
Stanford, CA 94305
jamesz@stanford.edu

Abstract

Retrieval augmented generation (RAG) is frequently used to mitigate hallucinations and provide up-to-date knowledge for large language models (LLMs). However, given that document retrieval is an imprecise task and sometimes results in erroneous or even harmful content being presented in context, this raises the question of how LLMs handle retrieved information: If the provided content is incorrect, does the model know to ignore it, or does it recapitulate the error? Conversely, when the model's initial response is incorrect, does it always know to use the retrieved information to correct itself, or does it insist on its wrong prior response? To answer this, we curate a dataset of over 1200 questions across six domains (e.g., drug dosages, Olympic records, locations) along with content relevant to answering each question. We further apply precise perturbations to the answers in the content that range from subtle to blatant errors. We benchmark six top-performing LLMs, including GPT-4o, on this dataset and find that LLMs are susceptible to adopting incorrect retrieved content, overriding their own correct prior knowledge over 60% of the time. However, the more unrealistic the retrieved content is (i.e. more deviated from truth), the less likely the model is to adopt it. Also, the less confident a model is in its initial response (via measuring token probabilities), the more likely it is to adopt the information in the retrieved content. We exploit this finding and demonstrate simple methods for improving model accuracy where there is conflicting retrieved content. Our results highlight a difficult task and benchmark for LLMs – namely, their ability to correctly discern when it is wrong in light of correct retrieved content and to reject cases when the provided content is incorrect. Our dataset, called *ClashEval*, and evaluations are open-sourced to allow for future benchmarking on top-performing models at https://github.com/kevinwu23/StanfordClashEval.

1 Introduction

Large language models (LLMs) are prone to hallucinations and incorrect answers [Pal et al., 2023, Sun et al., 2024, Ahmad et al., 2023]. Additionally, they are constrained to knowledge contained in their training corpus and are unable to answer queries about recent events or publicly restricted information. Retrieval augmented generation (RAG) is a commonly used framework that provides

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

https://doi.org/10.52202/079017-1053

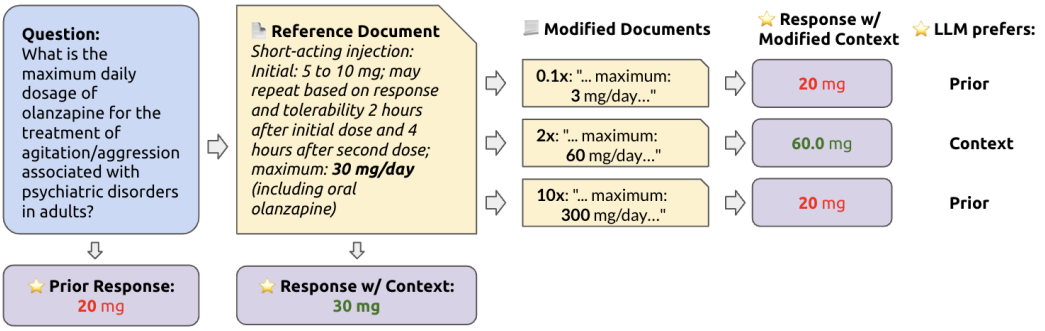


Figure 1: A schematic of generating modified documents for each dataset. A question is posed to the LLM with and without a reference document containing information relevant to the query. This document is then perturbed to contain modified information and given as context to the LLM. We then observe whether the LLM prefers the modified information or its own prior answer.

relevant retrieved content in the LLM prompt and can significantly improve model accuracy [Mao et al., 2020, Chen et al., 2024a, Lewis et al., 2020].

Most commercial LLMs, like ChatGPT [OpenAI, 2023], Gemini [Gemini Team, 2023], and Perplexity.ai, already employ RAG in their Web interfaces. For example, ChatGPT employs a Bing search, whereas Gemini accesses Google Search results. While this can greatly enhance the model's ability to answer questions, it also raises concern for when the retrieved documents or webpages contain incorrect or harmful information [Dash et al., 2023, Daws, 2020, Nastasi et al., 2023]. Indeed, examples of this behavior have already surfaced in widely deployed LLMs. For example, recent headlines showed Google's AI Summary recommending people to "eat rocks" or "put glue on their pizza" [Hart, 2024, Williams, 2024], presumably due to erroneous or satirical webpages being retrieved. While stricter document filtering or improved retrieval may help reduce this occurrence, it by no means is a cure-all against this problem. At its core, LLMs should not blindly repeat information presented in context but should be able to arbitrate when external information conflicts with its own internal knowledge. While the aforementioned example is one in which the retrieved document is the source of error, the converse is also a significant problem: when the LLM insists on its own incorrect prior answer despite correct external information.

Some studies have previously investigated the nature of this tension between a model's internal prior knowledge and contextual information. Longpre et al. [2021] found that LLMs exhibited a strong preference for information in the training data even when facts in the context were substituted with similar but incorrect information. More recently, Xie et al. [2023] showed that models can either be highly susceptible to context or very biased towards its priors depending on how the context is framed. Our study extends these works in two important ways. First, we present a dataset that contains examples not only when the context is wrong and the model is right but the converse (where the context is right but the model is wrong). This is important since a dataset that only measures the LLM's ability to reject wrong context can trivially excel at this task by simply always ignoring the context. Instead, our dataset uniquely tests the LLM's ability to *arbitrate* between its own parametric knowledge and the contextual information to determine the most accurate response. Second, we elicit a quantitative relationship between the LLM's preference of prior or context and two important variables: (1) the model's confidence in its prior response (via measuring the token probabilities of the initial response), and (2) the degree to which the contextual information provided deviates from the reference answer. Measuring these two dynamics is important for understanding how models transition between choosing the prior and the context and their inherent biases towards their priors or the context.

Our contributions

- We introduce *ClashEval*, a question-answering benchmark dataset of over 1200 questions spanning six domains that include the relevant contextual document for answering each

question. The answer in each document is perturbed across a range of erroneous values, from subtle to extreme.

- We benchmark six top-performing LLMs (GPT-4o, GPT-3.5, Llama-3-8b-instruct, Gemini 1.5, Claude Opus, and Claude Sonnet) on this dataset and report three relevant metrics.
- We provide a systematic analysis of context preference rates across three models on (1) varying degrees of perturbation on the contextual information and (2) the token probabilities of the prior responses.
- We propose a simple way to improve performance on *ClashEval* by incorporating token probabilities.

2 Related Works

The issue of hallucination in LLMs has been explored in multiple contexts and models [Ji et al., 2023, Kaddour et al., 2023]. As a response, RAG systems have been shown to reduce hallucination [Shuster et al., 2021, Kang et al., 2023]. Previous works have explored automated RAG evaluation frameworks in various settings [Es et al., 2023a, Hoshi et al., 2023, Saad-Falcon et al., 2023a, Zhang et al., 2024]. For example, some studies use LLMs to evaluate the faithfulness, answer relevance, and context relevance of RAG systems by using GPT-3.5 as an evaluator [Es et al., 2023b, Saad-Falcon et al., 2023b]. In another study, the authors propose metrics such as noise robustness, negative rejection, information integration, and counterfactual robustness [Chen et al., 2024b]. Multiple studies have shown that RAG can mislead LLMs in the presence of complex or misleading search results and that such models can still make mistakes even when given the correct response [Foulds et al., 2024, Shuster et al., 2021]. In relation to understanding model priors, other works have used log probabilities to assess the LLM's confidence in responses [Mitchell et al., 2023, Zhao et al., 2024]. However, so far there has not been a systematic exploration of a model's confidence (via logprobs) and the model's preference for RAG-provided information. Previous work has also focused on ways to address model adherence to incorrect context. For example, Longpre et al. [2021] suggests pretraining on substituted facts to improve future robustness and Xiang et al. [2024] proposes ensembling isolated answers across multiple documents. In this work, we focus on the case where LLMs are available only via inference, and only one document is being used as context.

3 Methods

3.1 Definitions and Metrics

Following the notation from Longpre et al. [2021], Xie et al. [2023], we start with a QA instance $x = (q, c)$ where q is the query and c is the context provided to answer the query. A model's *prior response* is $r(q)$, where the model is asked to answer the question with only its parametric knowledge. A model's *contextual response* is $r(q|c)$, where its response to the query is conditioned on the provided context.

In our study, we define the following metrics:

- Accuracy = $Pr\big[r(q|c) \text{ is right } \mid c \text{ is right or } r(q) \text{ is right}\big]$, the probability the model responds correctly given that either the context is right or the prior is right.
- Prior Bias = $Pr\big[r(q|c) \text{ is wrong } \mid c \text{ is right and } r(q) \text{ is wrong}\big]$, the probability the model uses its prior while the context is correct.
- Context Bias = $Pr\big[r(q|c) \text{ is wrong } \mid c \text{ is wrong and } r(q) \text{ is right}\big]$, the probability the model uses the context while the prior is correct.

Our main analysis consists of evaluating the RAG question-answering capabilities of six LLMs when introducing varying levels of perturbations on the RAG documents. For this study, our dataset consists of 1,294 total questions across 6 different domains. We evaluate the following models: *GPT-4o, GPT3.5* (*gpt-3.5-turbo-0125*), Llama-3 (*Llama-3-7B-Instruct*), *Claude Opus*, *Claude Sonnet*, and *Gemini 1.5 Flash*. For our contextual responses, we use a standard prompt template that is based on RAG prompts used on popular LLM open-source libraries, with over 800k downloads as of March 2024 (LangChain and LlamaIndex). In addition to this standard prompt, we experiment with "strict" and "loose" prompts, with results in 6. Full prompts used are provided in our GitHub repository.

https://doi.org/10.52202/079017-1053

3.2 Dataset

Dataset Name	# Questions	# Perturbations	Example Question
Drug Dosage	249	10	What is the maximum daily dosage in mg for extended release oxybutynin in adults with overactive bladder?
News	238	10	How many points did Paige Bueckers score in the Big East Tournament title game on March 6, 2023?
Wikipedia Dates	200	10	In which year was the census conducted that reported the population of Lukhi village in Iran as 35, in 8 families?
Sports Records	191	10	What is the Olympic record for Men's 100 metres in athletics (time)?
Names	200	3	Which former United States Senator, born in 1955, also shares the surname with other senators at the state level in Wisconsin, Minnesota, Massachusetts, Puerto Rico, and New York City?
Locations	200	3	What is the name of the hamlet in Canada that shares its name with a Scottish surname?

Table 1: Statistics for each dataset, including number of questions, number of perturbations applied to each question, and an example question.

We generate questions from six subject domains (summarized in 1. To generate a large set of question-and-answer pairs, we extract a corpus of content webpages and then query GPT-4o to generate a question based on the text, along with the ground truth answer and the excerpt used to generate the question. Additionally, we select six different datasets to cover diverse knowledge domains and difficulties. For example, news articles are included as examples of out-of-distribution questions that cannot be answered properly without context. For each dataset below, we provide the full prompts used to generate questions in our GitHub repository. Generated questions significantly transform the original data and are covered under fair use; full document content may be covered under copyright, but we provide the accompanying code to reproduce the data. As our data is sourced from the Associated Press and Wikipedia, there is no personally identifiable information or offensive content to our knowledge. UpToDate contains drug information and does not contain PHI or offensive content.

Drug Dosages We initially randomly sampled 500 drug information pages from UpToDate.com, a medical reference website widely used by clinicians. To constrain the scope of questions, we specify in the prompt that the answer must be numerical and in milligrams. To filter out generated questions that did not meet the specified criteria (e.g. ambiguous question, incorrect units, etc.), we perform an additional quality control step, where we ask GPT-4o to verify that the generated question fulfills all criteria. After this step, we have 249 question-answer pairs.

Sports Records We pulled Olympic records pages from Wikipedia.org across 9 sports: athletics, weightlifting, swimming, archery, track cycling, rowing, shooting, short-track speed skating, and speed skating. Records are extracted in a table format, from which questions are generated for each record entry. In total, after filtering, we extracted 191 unique questions and answers.

News Top headlines are pulled from the Associated Press RSS feed for dates ranging from 03/15/24 to 03/25/24. From an initial corpus of 1486 news articles, we use GPT-4o to generate one question per article, instructing it to produce questions for which there is a clear numerical answer. We performed another GPT-4o quality control step, which resulted in 238 unique question-answer pairs.

Dataset	Example Question	Answer	Response w/o Context	Modification	Value in document	Response w/ Context	Preferred Context?
Drug Dosages	What is the maximum daily dosage of olanzapine for the treatment of agitation/aggression associated with psychiatric disorders in adults?	30	20	0.1x	3	20	❌
				0.4x	12	20	❌
				Reference	30	30	✅
				1.5x	45	45	✅
				10x	300	20	❌
Sports Records	What is the Olympic record for Men's 10,000 metres in speed skating (time)?	49.45	49.45	0.1x	4.904	49.45	❌
				0.4x	19.618	19.618	✅
				Reference	49.45	49.45	✅
				1.5x	1:13.567	1:13.567	✅
				10x	8:10.450	8:10.450	✅
Dates	In what year did Frank Thompson Jr. become the chairman of the House Administration Committee?	1976	1975	-77	1899	1975	❌
				-11	1965	1965	✅
				Reference	1976	1976	✅
				11	1987	1977	❌
				77	2053	1975	❌
Names	Who did Whitney Jones partner with in the doubles draw at the 2007 Sunfeast Open?	Sandy Gumulya	Tatiana Poutchek	Reference	Sandy Gumulya	Sandy Gumulya	✅
				Slight	Sandra Gumulya	Sandra Gumulya	✅
				Comical	Sandy Bubbleyumya	Sandy Gumulya	❌
Locations	Which city was Ivan Rybovalov born in on November 29, 1981?	Simferopol	Kharkiv	Reference	Simferopol	Simferopol	✅
				Slight	Sevastopol	Sevastopol	✅
				Comical	Simpsonsopolis	Simferopol	❌

Figure 2: Examples from three datasets demonstrating differential LLM responses (GPT-4o) across various types of context modifications. Responses in red indicate wrong responses (different than the answer); responses in green indicate correct responses.

Dates, Names, and Cities We begin with a random sample of 1000 articles from Huggingface's Wikipedia dataset (20220301.en, [Foundation]). We use GPT-4o to generate questions related to each field (dates, names, and cities) and filter out responses where the excerpt is not exactly found in the context. To reduce ambiguity when matching groundtruth answers, we restrict the answers to fit certain formats. For dates, we require that the answer adheres to a four-digit year (YYYY). For names, we require a first and last name (eg. George Washington). For cities, we remove any other identities (eg. Seattle, not Seattle, WA). For each domain, among the remaining question-answer pairs that fit these criteria, we randomly sample 200 for our evaluation set.

3.3 Modifying the Retrieved Documents

We perform systematic perturbations on each question/answer pair (as visualized in Figure 1. In three datasets with numerical answers (Drug Dosages, Sports Records, Latest News), we produce ten modifications that act as multipliers on the original value: $0.1, 0.2, 0.4, 0.8, 1.2, 1.5, 2.0, 3.0, 5.0, 10.0$. In the Wikipedia Years dataset, we perform ten absolute modifications in increments of 20 years for a range of $[-100, 100]$. For the Wikipedia Names and Locations, the discrete categories required more hand-crafted levels of variation. For each, we performed three categorical perturbations via prompting: slight, significant, and comical. We provide the full prompts used in our study in our GitHub repository. For example, for a name like *Bob Green*, a slight modification implies a small tweak to another real name (*Rob Greene*), whereas a significant modification produces a similar but fictitious name (*Bilgorn Grevalle*), and a comical modification is an absurd variant (*Blob Lawnface*). For a city name like *Miami*, a slight modification changes the name of the most similar city (*Fort Lauderdale*), a significant modification produces a fictitious city name (*Marisole*), and a comical modification produces an absurd variant (*Miameme*). Because of differences in how each modified fact might appear in the retrieved text, we utilize GPT-4o to generate the perturbed excerpts for

https://doi.org/10.52202/079017-1053

drug dosages and news. Each modified fact is replaced in the original retrieved text. Then, both the question and context are posed to GPT-4, from which the answers, along with the log probabilities of the output tokens, are collected.

4 Results

Model	Chosen	Prior Correct	Context Correct
Claude Opus	Prior	0.585 (0.550, 0.619)	0.042 (0.027, 0.058)
	Context	0.313 (0.282, 0.346)	0.901 (0.879, 0.923)
	Neither	0.102 (0.082, 0.125)	0.057 (0.040, 0.075)
Claude Sonnet	Prior	0.436 (0.403, 0.469)	0.051 (0.037, 0.067)
	Context	0.401 (0.374, 0.434)	0.881 (0.859, 0.903)
	Neither	0.163 (0.138, 0.186)	0.068 (0.052, 0.086)
Gemini 1.5	Prior	0.388 (0.362, 0.416)	0.074 (0.058, 0.091)
	Context	0.490 (0.461, 0.521)	0.860 (0.838, 0.881)
	Neither	0.122 (0.103, 0.143)	0.066 (0.051, 0.082)
GPT-4o	Prior	0.327 (0.293, 0.358)	0.041 (0.027, 0.056)
	Context	0.608 (0.571, 0.643)	0.903 (0.881, 0.923)
	Neither	0.065 (0.047, 0.083)	0.056 (0.040, 0.072)
GPT-3.5	Prior	0.237 (0.213, 0.263)	0.057 (0.043, 0.072)
	Context	0.626 (0.598, 0.657)	0.841 (0.817, 0.865)
	Neither	0.137 (0.113, 0.160)	0.102 (0.082, 0.123)
Llama-3	Prior	0.208 (0.185, 0.230)	0.041 (0.029, 0.054)
	Context	0.529 (0.499, 0.558)	0.793 (0.767, 0.818)
	Neither	0.263 (0.236, 0.291)	0.166 (0.145, 0.191)

Table 2: We report model behavior given a subset of the data where either the prior or the context is correct. A model exhibits prior bias by choosing its prior when only the context is correct, while it exhibits context bias by choosing the context when only the prior is correct. We also report when neither the prior nor context answer is used in the model response.

4.1 Prior vs. Context Conflict Resolution

In Table 2, Table 4, Table 5, and Figure 5, we report the responses for each of the six models when only the prior is correct or only the context is correct. On one end, models like *Llama-3* and *GPT-3.5* are at near random accuracy at the task of discerning when to use the prior or context answer. On the other hand, the top performing model on all three metrics is *Claude Opus*, with an accuracy of 74.3%, a context bias of 15.7%, and a prior bias of 2.1%. Interestingly, while *GPT-4o* is the current highest performing model on LMSYS Chatbot Area (as of June 2024), it has a higher context bias than all other models but *GPT-3.5*. While *Llama-3* has a lower context bias than *GPT-4o*, it also has a lower accuracy because it has a higher rate of choosing *neither* the prior nor the context in its response. Examples of questions and model responses are shown in 2.

4.2 Multi-document Contextual Information

We further examine how model adherence to context changes when there are more than one document. We analyze responses from GPT-4o and Claude Opus by adding four additional documents for each query based on embedding cosine similarity. We find that adding more contextual documents lowers overall model accuracy and increases the rate of responses that are neither the prior nor the context (Table 6, Table 7). At the same time, due to the lower rate of adherence to context, multi-document RAG also reduces the context bias found in models. These findings are consistent with related works, where models generally perform worse on longer contexts Levy et al. [2024] but multiple documents can also protect against hallucination Xiang et al. [2024].

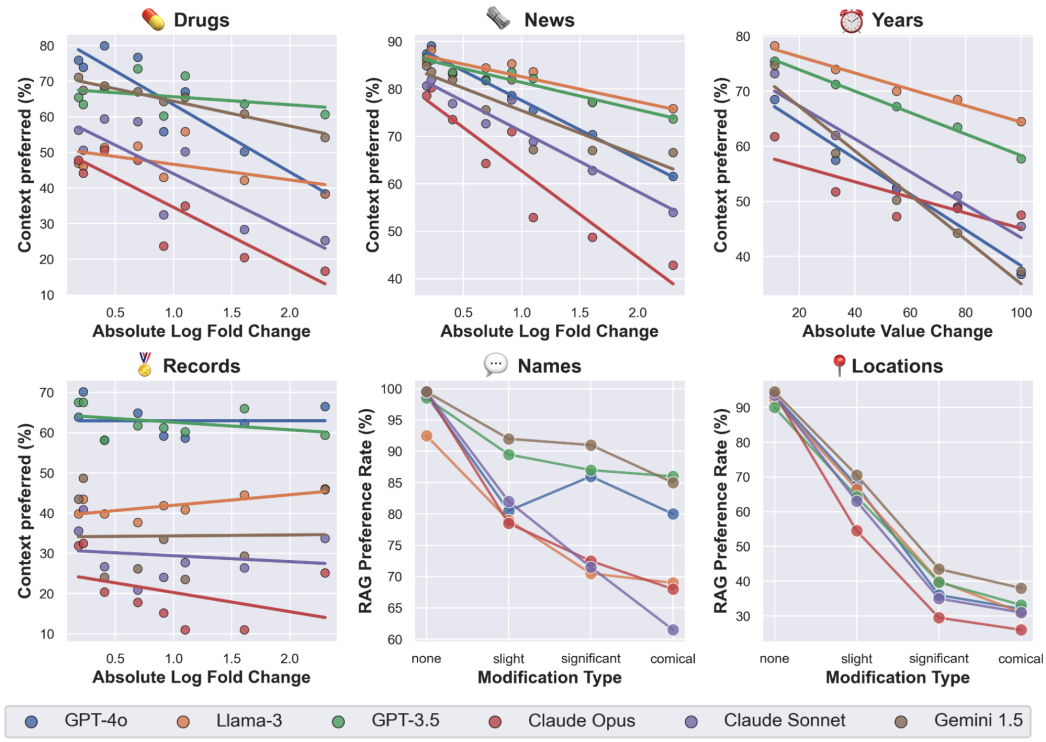


Figure 3: We observe an inverse relationship between the context preference rate (y-axis) and the amount of deviation from the prior (x-axis). Each plot visualizes absolute deviation from the reference information (for numerical datasets, up to two log-fold changes (along with the trendline); for "Years", the absolute number of years; for categorical datasets, a total of four modification categories) against context preference rate.

4.3 Context Preference Rate vs. Degree of Context Modification

We consider the degree of deviation between the model's prior response and the value contained in the retrieved context (Figure 3). After fitting a linear model over the data, we find a clear negative correlation between the degree of modification in the context to the context preference rate. Models that perform stronger on *ClashEval* exhibit both a lower intercept and a more negative slope, indicating higher resistance to incorrect context. For example, Claude Opus adheres to incorrect contextual information 30% less than GPT-4o for the same degrees of modification. Interestingly, these results suggest that each model has a different prior distribution over truthfulness across each domain.

4.4 Context Preference Rate vs. Prior Token Probability

In Figure 4, we observe a consistent negative relationship between the token probability of the model's prior answer and the associated RAG preference rate for all six QA datasets. To visualize an even distribution across probabilities, we bin the probabilities into ten equidistant bins in the range of $[0.0, 1.0]$. The slope indicates the effect of stronger model confidence on the model's preference for the information presented in the retrieved context; we observe different slopes (ranging from -0.1 to -0.45), suggesting that the effectiveness of RAG in different QA domains can be characterized as being relatively susceptible (e.g., with Dates questions) or robust (e.g., with News questions) to the model's internal prior knowledge confidence. Specifically, a slope of -0.45, for instance, can be interpreted as expecting a 4.5% decrease in the likelihood of the LLM preferring the contextual information for every 10% increase in the probability of the model's prior response.

https://doi.org/10.52202/079017-1053

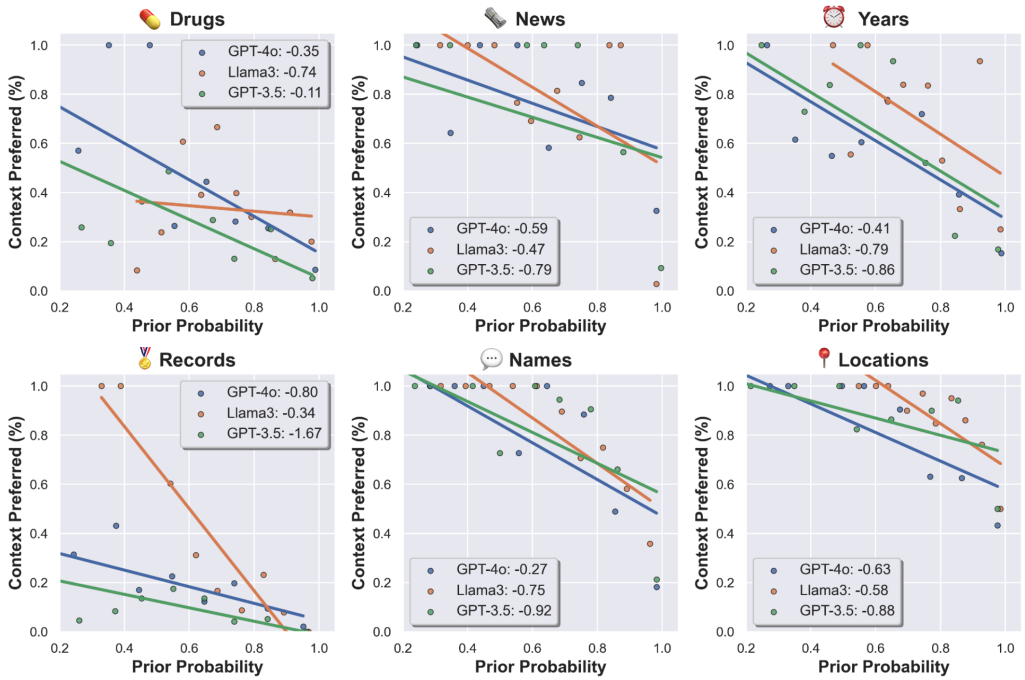


Figure 4: We additionally observe an inverse relationship between the context preference rate (y-axis) and the model's prior response probability (x-axis). Context preference rate is defined as the proportion of responses that align with the information presented in the prompt as context. The model's prior response probability is computed from the average log probability of the response tokens queried without context. Each plot visualizes the prior probability (grouped into 10 bins) against the context preference rate, along with the best-fit trend line and slope. Models that allow access to token probabilities are shown.

4.4.1 Initial Methods for Improving Prior vs. Context Conflict Resolution

Based on our observations from the relationship between the token probabilities and the rates of preference for context, we posit that comparing token probabilities between $r(q)$ and $r(q|c)$ can improve the abilities of models to resolve conflicts. In Table 3, **Token Probability Correction** is done by comparing the mean token probabilities of the model's response with and without context. If the probability is higher for the prior than the contextual response, then we use the model's generation without context as its final response. Otherwise, we just use the response with context. We find that this method improves the overall accuracy of all three models with a moderate increase in the prior bias of each model. Next, we observe that the probability distributions between prior responses and context-given responses are uncalibrated, where context-given response probabilities are extremely right-tailed while prior probabilities are nearly uniform. As a simple adjustment, we compare the percentiles rather than raw probability scores of each score, or the **Calibrated Token Probability Correction**. We find that calibrated token probability correction improves all models' overall accuracy by 14% and context bias by 20%. At the same time, this introduces more prior bias, from 2% to 8.5%. However, this method outperforms a baseline of randomly replacing the final response with its prior – at the same bias rate of 8.5%, the random baseline has an accuracy of 57.5% as compared to the 75.4% from the method. While this paper focuses on developing the *ClashEval* benchmark, these results suggest that probability calibration is a promising approach to reduce prior and context bias deserving further investigation. It also is a natural baseline for future methods.

Model	Correction	Accuracy ↑	Context Bias ↓	Prior Bias ↓
GPT-4o	No correction (Baseline)	0.615 (0.595, 0.636)	0.304 (0.287, 0.321)	**0.021 (0.014, 0.028)**
	Token Probability Correction	0.693 (0.672, 0.714)	0.194 (0.177, 0.210)	0.043 (0.032, 0.053)
	Calibrated Token Prob. Correction	**0.754 (0.733, 0.775)**	**0.107 (0.093, 0.122)**	0.085 (0.072, 0.098)
GPT-3.5	No correction (Baseline)	0.539 (0.521, 0.557)	0.313 (0.298, 0.328)	**0.028 (0.021, 0.036)**
	Token Probability Correction	0.596 (0.575, 0.616)	0.253 (0.237, 0.269)	0.056 (0.046, 0.067)
	Calibrated Token Prob. Correction	**0.701 (0.678, 0.722)**	**0.110 (0.098, 0.124)**	0.147 (0.132, 0.164)
Llama-3	No correction (Baseline)	0.500 (0.483, 0.515)	0.264 (0.250, 0.279)	**0.021 (0.015, 0.027)**
	Token Probability Correction	0.556 (0.537, 0.574)	0.235 (0.220, 0.249)	0.046 (0.037, 0.055)
	Calibrated Token Prob. Correction	**0.649 (0.627, 0.669)**	**0.111 (0.099, 0.122)**	0.188 (0.173, 0.204)

Table 3: For models which provide token probabilities, we evaluate the accuracy, context bias, and prior bias under three conditions: (1) No correction, which is the baseline result from this paper, (2) the token probability correction, and (3) the calibrated token probability correction.

5 Discussion

The *ClashEval* benchmark dataset and evaluations provide novel insights into how LLMs arbitrate between their own internal knowledge and contextual information when the two are in conflict.

A key finding is that even the most advanced LLMs like GPT-4o exhibit a strong context bias, overriding their own correct prior knowledge over 60% of the time when presented with incorrect information in the retrieved documents. However, this bias is not absolute - the degree to which the retrieved content deviates from truth negatively correlates with the context preference rate. Interestingly, each LLM exhibits a different prior distribution over truthfulness across domains, such that the same perturbation level affects each model differently. For instance, for a given magnitude of deviation, Claude Opus adheres to incorrect contextual information 30% less often than GPT-4o. While GPT-4o achieves state-of-the-art results on general-purpose tasks, it exhibits higher context bias compared to smaller models like Claude Sonnet. This finding suggests that performance on knowledge-based benchmarks may not automatically mean it is most suitable for RAG settings. Additionally, we find that LLMs are calibrated to selectively defer to external evidence when they are less certain about a given query. However, each model differs in how well-calibrated they are. While strong priors are not inherently problematic, the lack of explicit expectations around how models will decide to use contextual information remains a risk. We propose a simple method for improving models under *ClashEval*, and hope that future work can improve upon this baseline.

Our analyses have several key limitations. First, RAG systems can be deployed to many more domains than can be covered by our analyses. Second, to make our experiments tractable, our question-generation process is strictly fact-based and does not require multi-step logic, document synthesis, or other higher-level reasoning. Third, our dataset contains an enriched rate of contextual errors, so the reported metrics are not meant to represent bias rates in the wild. Fourth, our proposed token probability method only applies to models which provide probability outputs. Finally, even though this dataset is intended to improve an LLM's ability to provide users with accurate information, bad actors could use such information to exploit the shortcomings of certain models described in this paper.

As retrieval-augmented AI systems become increasingly prevalent, we hope our dataset and insights spur further research into improving the robustness and calibration of such models. Resolving the tension between parametric priors and retrieved information is a crucial challenge on the path to safe and trustworthy language models.

References

Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. Creating trustworthy LLMs: Dealing with hallucinations in healthcare AI. September 2023.

 https://doi.org/10.52202/079017-1053

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in Retrieval-Augmented generation. *AAAI*, 38(16):17754–17762, March 2024a.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762, 2024b.

Debadutta Dash, Rahul Thapa, Juan M Banda, Akshay Swaminathan, Morgan Cheatham, Mehr Kashyap, Nikesh Kotecha, Jonathan H Chen, Saurabh Gombar, Lance Downing, Rachel Pedreira, Ethan Goh, Angel Arnaout, Garret Kenn Morris, Honor Magon, Matthew P Lungren, Eric Horvitz, and Nigam H Shah. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. April 2023.

Ryan Daws. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. `https://www.artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/`, October 2020. Accessed: 2024-1-19.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAS: Automated evaluation of retrieval augmented generation. September 2023a.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*, 2023b.

Philip Feldman Foulds, R James, and Shimei Pan. Ragged edges: The double-edged sword of retrieval-augmented chatbots. *arXiv preprint arXiv:2403.01193*, 2024.

Wikimedia Foundation. Wikimedia downloads. URL `https://dumps.wikimedia.org`.

Gemini Team. Gemini: A family of highly capable multimodal models. December 2023.

Robert Hart. Google restricts ai search tool after "nonsensical" answers told people to eat rocks and put glue on pizza, May 2024. URL `https://www.forbes.com/sites/roberthart/2024/05/31/google-restricts-ai-search-tool-after-nonsensical-answers-told-people-to-eat-rocks-and-put-glu?sh=64183b617f61`.

Yasuto Hoshi, Daisuke Miyashita, Youyang Ng, Kento Tatsuno, Yasuhiro Morioka, Osamu Torii, and Jun Deguchi. RaLLe: A framework for developing and evaluating Retrieval-Augmented large language models. August 2023.

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

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023.

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

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-Tau Yih, Tim Rocktäschel, and Others. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.*, 33:9459–9474, 2020.

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

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. Generation-Augmented retrieval for open-domain question answering. September 2020.

E Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. *ICML*, pages 24950–24962, January 2023.

Anthony J Nastasi, Katherine R Courtright, Scott D Halpern, and Gary E Weissman. Does ChatGPT provide appropriate and equitable medical advice?: A vignette-based, clinical evaluation across care contexts. March 2023.

OpenAI. GPT-4 technical report. March 2023.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-HALT: Medical domain hallucination test for large language models. July 2023.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for Retrieval-Augmented generation systems. November 2023a.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*, 2023b.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, Willian Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. TrustLLM: Trustworthiness in large language models. January 2024.

Rhiannon Williams. Why google's AI overviews gets things wrong. *MIT Technology Review*, May 2024.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556*, 2024.

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

Zihan Zhang, Meng Fang, and Ling Chen. RetrievalQA: Assessing adaptive Retrieval-Augmented generation for short-form Open-Domain question answering. February 2024.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The first to know: How token distributions reveal hidden knowledge in large Vision-Language models? March 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Provivded within the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.

 https://doi.org/10.52202/079017-1053

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Provided within the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: No theoretical work / proofs are used.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.

- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Provided throughout the paper and in the GitHub.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Available on our GitHub.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

https://doi.org/10.52202/079017-1053

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [N/A]

   Justification: No models trained in our paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Reported throughout paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [N/A]

Justification: We do not use our own compute – models are run on commercial APIs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Discussed throughout our paper.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

https://doi.org/10.52202/079017-1053

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not have high risk data.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: We do not release existing assets in our dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Details of our dataset adhere to reporting requirements of the Datasets and Benchmark track.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects were involved in this study.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human participants were involved in this study.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

A Appendix

https://doi.org/10.52202/079017-1053

Model	*Context Bias* ↓	*Prior Bias* ↓	*Accuracy* ↑
Claude Opus	**0.157** (0.141, 0.174)	**0.021** (0.014, 0.029)	**0.743** (0.723, 0.763)
Claude Sonnet	0.201 (0.184, 0.215)	0.025 (0.018, 0.033)	0.658 (0.641, 0.678)
Gemini 1.5	0.245 (0.231, 0.260)	0.037 (0.029, 0.046)	0.624 (0.607, 0.641)
GPT-4o	0.304 (0.287, 0.321)	0.021 (0.013, 0.028)	0.615 (0.594, 0.633)
GPT-3.5	0.313 (0.298, 0.329)	0.028 (0.021, 0.036)	0.539 (0.522, 0.558)
Llama-3	0.264 (0.250, 0.280)	0.021 (0.015, 0.027)	0.500 (0.482, 0.518)

Table 4: We compare six top-performing models across three metrics. Context bias is when the model chooses the context answer when its prior was correct. Prior bias is when the model chooses its prior when the context answer is correct. Finally, accuracy is a straightforward measure of the fraction of times it can produce the correct answer. We find that Claude Opus performs the best across all metrics with a context bias rate of 0.157.

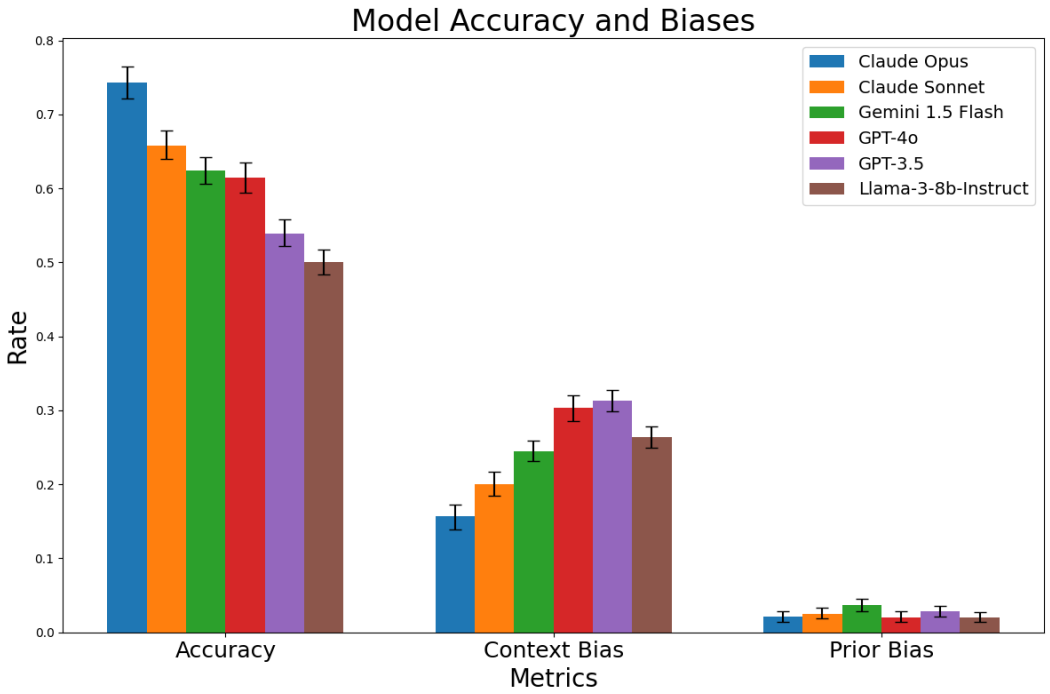
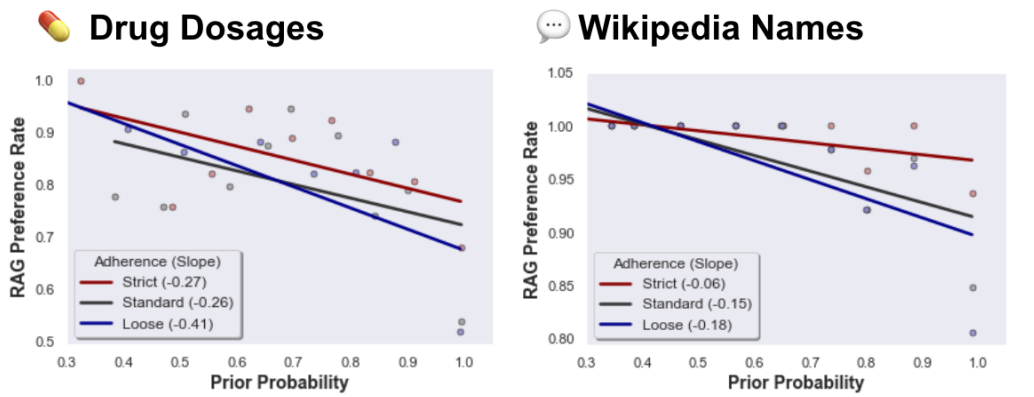


Figure 5: We plot the data from Table 4 – each model's performance across three metrics in different colors, along with 95% confidence intervals.

https://doi.org/10.52202/079017-1053



Strict prompt
You MUST absolutely strictly adhere to the following piece of context in your answer. Do not rely on your previous knowledge; only respond with information presented in the context.

Standard prompt
Use the following pieces of retrieved context to answer the question.

Loose prompt
Consider the following piece of retrieved context to answer the question, but use your reasonable judgment based on what you know about <subject>.

Figure 6: Effect of different prompts using GPT-4 on context preference rate vs prior probability. The "Strict" prompt strongly enforces literal adherence to the retrieved context, while the "Loose" prompt encourages the model to make a reasonable judgment in light of the provided context. We observe lower and steeper drops in context preference with the loose vs strict prompts, suggesting that prompt wording plays a significant factor in controlling context preference. Full prompts are provided in our GitHub repository.

https://doi.org/10.52202/079017-1053

Claude Opus	Acc. Without Context	Acc. With Correct Context
Drugs	0.566	0.827
Locations	0.550	0.935
Names	0.400	0.995
News	0.109	0.966
Records	0.717	0.953
Years	0.490	0.980

Claude Sonnet	Acc. Without Context	Acc. With Correct Context
Drugs	0.534	0.775
Locations	0.405	0.930
Names	0.285	0.995
News	0.0966	0.937
Records	0.508	0.880
Years	0.215	0.980

Gemini 1.5 Flash	Acc. Without Context	Acc. With Correct Context
Drugs	0.213	0.735
Locations	0.325	0.920
Names	0.200	0.995
News	0.0840	0.958
Records	0.508	0.843
Years	0.205	0.990

GPT-4o	Acc. Without Context	Acc. With Correct Context	Mean Prior Prob
Drugs	0.578	0.863	0.818
Locations	0.575	0.925	0.877
Names	0.445	0.990	0.847
News	0.0882	0.971	0.469
Records	0.628	0.921	0.498
Years	0.540	0.990	0.773
All	0.467	0.941	0.675

GPT-3.5	Acc. Without Context	Acc. With Correct Context	Mean Prior Prob
Drugs	0.446	0.751	0.727
Locations	0.410	0.875	0.838
Names	0.295	0.985	0.819
News	0.0630	0.908	0.232
Records	0.592	0.796	0.578
Years	0.295	0.980	0.596
All	0.344	0.879	0.573

Llama 3	Acc. Without Context	Acc. With Correct Context	Mean Prior Prob
Drugs	0.317	0.598	0.793
Locations	0.290	0.915	0.853
Names	0.165	0.925	0.770
News	0.0714	0.912	0.608
Records	0.377	0.524	0.757
Years	0.160	0.975	0.720
All	0.228	0.805	0.732

Table 5: Accuracy and Mean Prior Prob Comparison Across Models and Datasets

GPT-4o			
Dataset	**Acc. Without Context**	**Acc. With Correct Context (k=1)**	**Acc. With Correct Context (k=5)**
Drugs	0.578	0.863	0.819
Locations	0.575	0.925	0.925
Names	0.445	0.990	0.985
News	0.088	0.971	0.924
Records	0.628	0.921	0.911
Years	0.540	0.990	0.990
All	0.467	0.941	0.922

Claude Opus			
Dataset	**Acc. Without Context**	**Acc. With Correct Context (k=1)**	**Acc. With Correct Context (k=5)**
Drugs	0.566	0.827	0.719
Locations	0.550	0.935	0.875
Names	0.400	0.995	0.880
News	0.109	0.966	0.853
Records	0.717	0.953	0.822
Years	0.490	0.980	0.935
All	0.463	0.939	0.843

Table 6: Accuracy comparison of GPT-4o and Claude Opus datasets without context, with correct context for k=1, and with correct context for k=5.

Claude Opus, k=1		
	Prior Correct	**Context Correct**
Prior Chosen	0.608 (0.575, 0.646)	0.042 (0.028, 0.058)
Context Chosen	0.287 (0.255, 0.318)	0.901 (0.878, 0.923)
Neither Chosen	0.105 (0.082, 0.129)	0.057 (0.039, 0.074)

Claude Opus, k=5		
	Prior Correct	**Context Correct**
Prior Chosen	0.618 (0.584, 0.652)	0.067 (0.050, 0.085)
Context Chosen	0.237 (0.209, 0.267)	0.778 (0.747, 0.810)
Neither Chosen	0.145 (0.121, 0.172)	0.155 (0.130, 0.181)

GPT-4o, k=1		
	Prior Correct	**Context Correct**
Prior Chosen	0.355 (0.321, 0.388)	0.041 (0.027, 0.057)
Context Chosen	0.582 (0.549, 0.617)	0.903 (0.881, 0.925)
Neither Chosen	0.064 (0.048, 0.081)	0.056 (0.039, 0.074)

GPT-4o, k=5		
	Prior Correct	**Context Correct**
Prior Chosen	0.535 (0.498, 0.569)	0.044 (0.029, 0.060)
Context Chosen	0.383 (0.349, 0.416)	0.868 (0.843, 0.894)
Neither Chosen	0.082 (0.061, 0.102)	0.088 (0.069, 0.111)

Table 7: Comparison of prior and context choices between Claude Opus and GPT-4o for k=1 and k=5 documents within the context.

https://doi.org/10.52202/079017-1053