

---

# Evaluating the Sensitivity of LLMs to Prior Context

---

**Robert Hankache<sup>1</sup>**   **Kingsley Nketia Acheampong<sup>1</sup>**   **Liang Song<sup>1</sup>**  
**Marek Brynda<sup>1</sup>**   **Raad Khraishi<sup>1,2</sup>**   **Greig A. Cowan<sup>1</sup>**

<sup>1</sup>NatWest AI Research  
<sup>2</sup>University College London

## Abstract

As large language models (LLMs) are increasingly deployed in multi-turn dialogue and other sustained interactive scenarios, it is essential to understand how extended context affects their performance. Popular benchmarks, focusing primarily on single-turn question answering (QA) tasks, fail to capture the effects of multi-turn exchanges. To address this gap, we introduce a novel set of benchmarks that systematically vary the volume and nature of prior context. We evaluate multiple conventional LLMs, including GPT, Claude, and Gemini, across these benchmarks to measure their sensitivity to contextual variations. Our findings reveal that LLM performance on multiple-choice questions can degrade dramatically in multi-turn interactions, with performance drops as large as 73% for certain models. Even highly capable models such as GPT-4o exhibit up to a 32% decrease in accuracy. Notably, the relative performance of larger versus smaller models is not always predictable. Moreover, the strategic placement of the task description within the context can substantially mitigate performance drops, improving the accuracy by as much as a factor of 3.5. These findings underscore the need for robust strategies to design, evaluate, and mitigate context-related sensitivity in LLMs.

## 1 Introduction

The rapid evolution and deployment of large language models (LLMs) in various natural language processing applications has sparked substantial interest in understanding the capacity of models when handling complex contextual settings and interactions in natural language, particularly, settings that involve conversation sequences. With LLMs increasingly underpinning applications across domains such as conversational agents [1, 2], content generation [3], document editing [4], and educational tutoring systems [5], there is a pressing need to evaluate their robustness beyond single-turn tasks.

A critical aspect of LLM usage in real-world systems is their ability to manage dynamic contextual information across varied multi-turn user interaction settings [6, 7]. In such settings, the sequence of interactions can vary greatly in terms of domain specificity, task type, and semantic coherence. While LLMs have been successful in tasks such as few-shot learning [8], where minimal context is provided to infer or adapt to new tasks, and in addressing needle-in-haystack problems [9], where the goal is to extract or identify highly specific information from extensive datasets, these tasks fundamentally differ from maintaining contextual coherence across extended multi-turn interactions. Recent insights from research have shown that maintaining coherence and accuracy over multiple interactions is a challenging task for LLMs, often leading to degradation in response quality [10, 11]. Few-shot learning often relies on pre-curated prompts or demonstrations to guide the model, while needle-in-haystack challenges prioritize locating relevant information in a static context. In contrast, multi-turn interactions demand robust memory mechanisms to interpret and effectively integrate prior conversational turns [12]. It is also important in long-running agentic AI systems that require robustness and consistency [13].

Evaluating model performance in multi-turn conversations remains a critical challenge. While existing single question-answer (Single QA) benchmarks such as MMLU and GPQA have provided valuable insights into various aspects of LLM capabilities, they fall short in adequately assessing multi-turn interactions and the subtle role of context within these settings [7].

In this study, we address the gap in understanding how LLMs cope with extended multi-turn interactions by demonstrating that performance consistency is strongly influenced by the nature of preceding interactions. Also, we assess the performance of smaller and larger LLMs under varying lengths of context. Specifically, we assess whether models maintain domain consistency if introduced to different context and lengths. We introduce new benchmarks derived from GPQA Diamond [14] to evaluate the sensitivity of LLMs to prior context to investigate these challenges. We then conduct a series of controlled experiments on popular LLMs, including GPT, Claude and Gemini models. Our experiments systematically vary both the knowledge domain of preceding interactions and the depth of conversation history, analysing LLM performance degradation in the presence of prior context. Our findings suggest that performance may degrade drastically with long prior context, as high as 73% drop compared to performance when no prior context is added. This demonstrates the invalidity of using a single-turn benchmark such as GPQA as a proxy for performance over multi-turn interactions. As such, our new benchmarks enable the assessment of LLMs' multi-turn capabilities and contextual awareness, which are crucial for enhancing LLM performance in practical scenarios. Moreover, this study further reinforces the necessity for design improvements in LLMs, such as mechanisms for better contextual memory management and strategies for handling variable coherence in sequential user interactions to enhance the reliability of these models in dynamic real-world applications.

## 2 Related Work

The ability of LLMs to incorporate and leverage prior context has been a focal point of recent work. Floridi and Chiriaci [15] introduced GPT-3, demonstrating the model's remarkable zero-shot and few-shot learning capabilities while also highlighting its sensitivity to prompt phrasing and context design. Subsequent iterations, including OpenAI's GPT-4 and o3, and other open-source models like Deepseek-R1, have sought to improve the robustness and adaptability of the models to understanding context [2]. However, prior studies have shown that LLMs struggle to maintain coherence across long multi-turn conversations, especially when prior context contains minimal information to aid the LLMs in reasoning [16, 17].

To better evaluate model performance, previous studies have presented novel benchmarks for context understanding using multiple datasets or through contrastive decoding [18, 19]. These LLM benchmarks, which include HellaSwag, BigBench, TruthfulQA, and Chatbot Arena [20, 21, 22], have emerged as standards for assessing LLM performance. However, LLM benchmarks have certain limitations [23]. Issues such as data contamination [24], narrow focus [23], and an emphasis on single-turn interactions have been previously reported. New benchmarks, such as MT-Eval [25], are being introduced to handle the evaluation of multi-turn and multilingual interactions; for example, Multi-IF utilizes a hybrid framework that combines LLMs and human annotators [26].

Moreover, early models, such as GPT-1, were limited to 512 tokens, which constrained their ability to generate coherent responses over extended inputs [27]. Recent advances have dramatically increased these limits, with many models supporting hundreds or even millions of input tokens. The ability to handle longer context lengths has enabled LLMs to process larger volumes of information, thereby improving performance in tasks such as summarisation and dialogue systems [28]. However, increasing context length introduces challenges such as the "lost in the middle" phenomenon [29], and performance degradation, particularly when domains shift mid-conversation [30, 31]. Several efforts have aimed to address the limitations of large language models (LLMs) in handling context [32, 33, 34, 35, 36, 37, 38], and consequently, necessitating development of novel evaluation benchmarks.

## 3 Methodology

In this section, we describe how we create our benchmark datasets for evaluating the QA performance of LLMs in situations with varying past context types, specify the LLMs we evaluated, and the evaluation criteria for assessing models' sensitivity to prior context.

### 3.1 Datasets

We design experiments to mimic a user’s textual interactions with an LLM assistant or a chatbot. Each experiment is formed from three inputs:

1. An optional prior context or conversational history.
2. A target query ( $q_t$ ).
3. Four answer choices for each  $q_t$ , where only one choice is the correct response.

To evaluate model performance, we measure the model’s accuracy in selecting the correct response.

#### 3.1.1 Target Queries

All target queries,  $q_t$ , on which the models are evaluated, are sourced from the STEM categories: Biology, Physics, and Chemistry of the Graduate-level Professional Question Answering (GPQA) benchmark dataset [14] used by top AI institutions to evaluate their LLMs’ expert-level reasoning in STEM fields. In this study, the GPQA questions are presented to LLMs following the prior contexts to evaluate their ability to maintain accuracy and domain fidelity. We use the diamond variant of GPQA to align with benchmark consistency in evaluating scientific reasoning and knowledge [2, 39, 40, 41] which consists of 198 graduate-level multiple-choice questions selected to challenge highly capable and motivated PhD-level non-experts. Only questions where experts answer correctly, and the majority of non-experts answer incorrectly are in this variant. Each target query from GPQA,  $q_t$ , is accompanied by four of answer choices. For each query instance presented to the experimental models, we shuffle the order of answer choices randomly to mitigate positional bias.

#### 3.1.2 Prior Context

Variety of context is crucial to evaluating LLMs’ sensitivity in different context scenarios. As such, prior context was sourced from two main data sources: LMSYS-Chat-1M dataset [42], and MMLU dataset [43]. Context lengths for the created dataset range from 4k to 64k tokens.<sup>1</sup>

**Free-chat.** The *free-chat* context scenario is used to evaluate how authentic, open-ended prior human interactions with LLMs affect their performance. In real-world use, users frequently engage in highly varied and non-linear conversations, often shifting rapidly between lines of thought as seen in LMSYS-Chat-1M dataset.

The LMSYS-Chat-1M dataset is a large-scale corpus of approximately one million real-world user chats with LLMs. It comprises both single- and multi-turn dialogues originally collected from public chatbot conversations. For our study, we selected chat samples generated by three of the top-performing models in the dataset: vicuna-33b, llama-2-13b-chat, and mpt-30b-chat. The selection contains enough volume of conversations for a varied experimentation, more than 45k single- and multi-turn chats. Additionally, as a preprocessing step to maintain data quality, we removed any empty messages from the sampled conversations, eliminating minor artifacts likely introduced by the original LMSYS-Chat-1M cleaning process.

**Multi-turn QA.** The multi-turn QA context scenarios make use of multiple STEM and Non-STEM examples large enough to cover thousands of tokens in context length. Given that the GPQA dataset is restricted to STEM subjects and has a limited size, we chose to utilize an alternative dataset for the prior context to enhance the scope of our experiments. We selected the Measuring Massive Multitask Language Understanding (MMLU) dataset, due to its volume, broad domain coverage, and high-quality multiple-choice format, ensuring that prior context is both structured and relevant for testing. The data, used to test models multitask accuracy, covers 57 domains including STEM and Non-STEM subjects. It was designed to be more challenging than then-existing benchmarks such as General Language Understanding Evaluation (GLUE) on which new language models were achieving better-than-human accuracy. Similar to GPQA, the benchmark is used as an industry standard for analysing performance of LLM models. We employed the MMLU dataset for generating prior context in two experimental settings:

---

<sup>1</sup>The context length (tokens count) is computed using the OpenAI tokeniser for GPT-4o and GPT-4o-mini.

- **Same-domain Context.** For this setting, we extract STEM-related question-response  $qr$  from the MMLU dataset. These include subjects such as Biology, Physics, Chemistry, Clinical Knowledge, and Mathematics.
- **Cross-domain Context.** For this setting, Non-STEM  $qr$  pairs from the MMLU dataset are selected. These include History, Law, Social Sciences, and Philosophy subjects.

See Appendix F for a complete list of selected subjects.

### 3.2 Task Position

The task description<sup>2</sup> for the target query, which includes the expected answer formatting, is placed in the last message after the prior context and before the target query for the *free-chat* context, since the tasks are not related. For the multi-turn QA experiments, we evaluate the following two settings. First is *task-at-top*, where the task description is provided only once at the very top of the prior context, i.e. the first message of the conversation. This priming technique is motivated in Brown et al. [29] and Liu et al. [44]. Given the initial low performances (see section 4.2) and as a result of the detailed investigations (see Section 4.2.1), we also evaluate a second setting, *task-repeated*, where the task description is reiterated both at the very top of the prior context, and right before the target query.

### 3.3 Models

We evaluate several state-of-the-art large language models, namely: GPT-4o-mini, GPT-4o [39], Claude Haiku, Claude Sonnet<sup>3</sup>, Gemini Flash, and Gemini Pro [40] (refer to Appendices E and G for models versions and experimental cost). These LLMs were selected purposefully for their demonstrated ability in multi-turn conversations and wide use in the AI community, also accounting for the distinct subtleties inherent in the architectures and training paradigms of various providers, thereby ensuring the generalisability of our findings across various LLM implementations.

**Temperature Settings.** Various temperature settings, varying from 0 to 1, have been used when evaluating LLM models on Question Answering, Maths and reasoning tasks. For our experiments, a temperature setting of 0.5 is used for all models to balance creativity, randomness, and determinism in responses [2, 39, 41]. Accordingly, the models' seeds are varied across evaluation runs to assess the models' response variability.

## 4 Results

Using the context scenarios described in the previous section, we run several experiments to evaluate the sensitivity of the selected large language models to prior context. We set up two primary experimental conditions. In each, the target query is a STEM-related query from the GPQA dataset (Section 3.1.1). In the first experiment, prior context consists of multi-turn *free-chat* conversations (Section 3.1.2). The task description and target query ( $q_t$ ) are added as the last message in the sequence. In the second experiment, we present each model with prior context consisting of *same-* or *cross-domain* multi-turn QA conversation-styled interactions (Section 3.1.2). In addition, the location of the task description in the prior context is varied between *task-at-top* and *task-repeated* scenarios, to better ascertain task location influence on the overall model performance. Hence in this second experiment, the total unique conditions are four. In all experiments, the various LLMs were given the same list of target queries and additional context. For every target query,  $q_t$ , in the experiments, we present models with four varying lengths of prior context: 4k, 16k, 32k and 64k. The variation in context lengths enables us to not only assess sensitivity in the presence of context, but also the effect of increasing prior context lengths.

Only complete  $qr$  pairs are presented as input context; a soft truncation is applied on the  $qr$  level. We add a *no-context* scenario which serves as a base experiment during the context sensitivity experiments.<sup>4</sup> Three evaluation runs across three different seeds are used for each combination of

---

<sup>2</sup>Our prompts follow OpenAI "simple-evals" framework.

<sup>3</sup>Claude models only accept a maximum of one thousand conversational messages. In cases where we exceed this limit, we combine several rounds of conversation into one message, respecting the conversation formatting and keeping the same content.

<sup>4</sup>For the base *no-context* experiments, the average context length of the target queries is 270 tokens.

Table 1: LLM performance results, in terms of accuracy, across different context scenarios (various sources and lengths). Accuracies are reported as mean values of three runs, with standard deviations. Context length refers to token counts of the entire text sent to the LLM, which includes the target query and the prior context if added. "No context" refers to the scenario in which no additional context, apart from the target query, is provided to the models.

Model	Context Length	Free-chat	Same Domain		Cross Domain	
			Task-at-top	Task-repeated	Task-at-top	Task-repeated
Claude Haiku	no context	0.365±0.020	0.365±0.020	0.365±0.020	0.365±0.020	0.365±0.020
	4k	0.365±0.021	0.357±0.018	0.362±0.029	0.343±0.013	0.386±0.023
	16k	0.332±0.036	0.360±0.015	0.318±0.035	0.323±0.013	0.367±0.023
	32k	0.354±0.036	0.359±0.005	0.372±0.018	0.350±0.008	0.367±0.034
	64k	0.352±0.006	0.347±0.031	0.359±0.005	0.315±0.006	0.345±0.042
	no context	0.412±0.008	0.412±0.008	0.412±0.008	0.412±0.008	0.412±0.008
Claude Sonnet	4k	0.375±0.034	0.375±0.043	0.389±0.009	0.249±0.034	0.386±0.006
	16k	0.367±0.019	0.375±0.008	0.411±0.028	0.172±0.017	0.407±0.028
	32k	0.382±0.011	0.350±0.030	0.389±0.010	0.180±0.037	0.412±0.023
	64k	0.384±0.031	0.380±0.016	0.375±0.025	0.190±0.051	0.362±0.044
	no context	0.476±0.006	0.476±0.006	0.476±0.006	0.476±0.006	0.476±0.006
	4k	0.412±0.008	0.313±0.040	0.470±0.025	0.305±0.023	0.446±0.006
Gemini Flash	16k	0.421±0.011	0.340±0.008	0.480±0.018	0.180±0.015	0.443±0.029
	32k	0.439±0.020	0.354±0.005	0.461±0.020	0.162±0.027	0.443±0.008
	64k	0.421±0.034	0.340±0.028	0.475±0.015	0.128±0.026	0.438±0.021
	no context	0.545±0.013	0.545±0.013	0.545±0.013	0.545±0.013	0.545±0.013
	4k	0.545±0.018	0.483±0.018	0.537±0.016	0.502±0.008	0.517±0.011
	16k	0.488±0.015	0.480±0.033	0.539±0.011	0.471±0.008	0.551±0.033
Gemini Pro	32k	0.502±0.030	0.500±0.033	0.552±0.018	0.451±0.008	0.502±0.013
	64k	0.481±0.044	0.473±0.013	0.552±0.041	0.460±0.028	0.534±0.013
	no context	0.524±0.019	0.524±0.019	0.524±0.019	0.524±0.019	0.524±0.019
	4k	0.453±0.030	0.449±0.023	0.500±0.015	0.461±0.046	0.510±0.018
	16k	0.461±0.031	0.439±0.010	0.510±0.022	0.389±0.018	0.502±0.012
	32k	0.451±0.013	0.423±0.016	0.475±0.013	0.372±0.015	0.525±0.015
GPT-4o	64k	0.458±0.013	0.441±0.008	0.483±0.016	0.355±0.024	0.510±0.028
	no context	0.451±0.044	0.451±0.044	0.451±0.044	0.451±0.044	0.451±0.044
	4k	0.412±0.011	0.348±0.036	0.394±0.013	0.320±0.016	0.423±0.019
	16k	0.406±0.026	0.276±0.013	0.409±0.026	0.226±0.024	0.423±0.018
	32k	0.396±0.026	0.278±0.028	0.382±0.023	0.190±0.006	0.418±0.011
	64k	0.392±0.048	0.293±0.022	0.404±0.045	0.205±0.028	0.416±0.025

queries, context types, and models. The following randomisations are done for each run: i) shuffle the target query's choices, ii) randomise the additional prior context, and iii) change the model's seed. The mean and standard deviation of the accuracy is computed from the three runs and are shown in Table 1. In addition to absolute values, relative scores with respect to the base *no-context* experiments are recorded in the appendix (see Table 4). We now discuss the results of the *free-chat* experiments and multi-turn QA experiments.

#### 4.1 Free-chat Context Experiments

**Prior context leads to performance degradation.** It is clear from Figure 1 that prior contexts cause models' performances to degrade. This is true for all the six models we tested from the three different providers. It is reasonable to assume that the inclusion of unrelated free chat would not influence the model's performance, particularly since the target question is independent of information from the prior context. Nonetheless, the performance drops observed are non-negligible. The average accuracy drops in the highest two bins ranges from 3.5% to 13% relative to *no-context* scenarios, revealing a varied effect depending on the model, and especially between model providers. Although the degradation is bigger for some models, we do not observe any flip in accuracy rankings. Instead, we see a reduction of the accuracy spread within the same prior context experiments. With *no-context*, the top performing model (Gemini Pro) accuracy is 49% higher than the worst performing model (Claude Haiku). At 64k-tokens prior context, this difference is reduces to 37%.

**Model performance degradation is affected by the length of interactions.** The length of multi-turn interactions also plays a critical role in performance stability, as expected. The longer the conversational interaction, the higher the drop in performance that is observed for all the models, up to a certain length. After 4k to 16k context length, models' performances drop very slightly or even stabilise. For instance, for the top performing model Gemini Pro, the accuracy drops 5.7% from

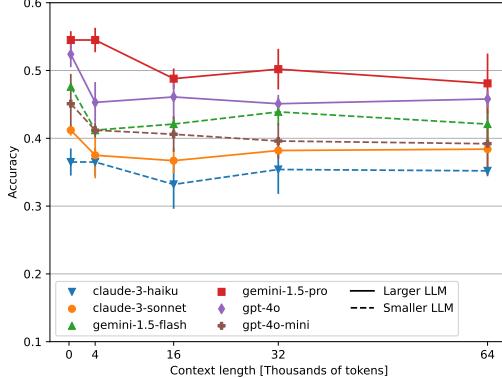


Figure 1: Experimental LLM performance over variable *free-chat* context. Scores are mean accuracies across three runs with the standard deviations shown as error bars. Dashed and continuous lines represent smaller and larger LLMs respectively.

no- to 16k-token context; then the remaining drop until 64k-token context is just 0.7%. At 4k-token context, GPT-4o model had the largest relative drop of 12%, with a negligible remaining drop until 64k-token context. This early stabilisation might be due to the following: queries  $q_t$ , that were correctly answered in the *no-context* scenario but with lower confidence, might have their answers easily flipped even with short prior context; whereas the ones answered with high confidence will have the same answer even at 64k context lengths. More studies are needed to understand and explain this behaviour.

**Larger models do not necessarily improve robustness to context sensitivity.** Contrary to our expectations, we found large models to be just as susceptible to the effect of prior context as their smaller counterparts. Comparing models within each family, smaller and larger GPT (Gemini) models have similar accuracy drops in the highest two bins, averaging 13% (12%) (see Figure 1). Claude models exhibit a different effect; in those same two bins, the larger Sonnet model’s performance drops 7%, compared to 3.5% only for the smaller Haiku model. The additional intriguing observation is that Claude Haiku, which is the least affected by the prior context, is at the same time the worst performing model between the six. This shows empirically that it is not guaranteed that a more powerful and high-performing model in basic benchmarks will maintain its advantage in other scenarios with varying context.

## 4.2 Multi-turn QA Context Experiments

**Prior interactions have large impact on models’ performances.** In Figures 2a and 2b, we see the effect of multi-turn QA context on model performance. Notably, we find that the effect of context is more significant for the *task-at-top* scenario, where most models suffer from significant performance degradation, with some experiencing as high as a 73% relative drop (Gemini Flash with *cross-domain* context). Although locating the task at the top of the context is intuitive and well-motivated as the task does not change with our final query similar to few-shot prompting scenarios, we see that having any type of context preceding a target query, whether it’s QA from the *same-* or *cross-domain*, significantly influences the models’ performances. The only exception observed among the six models is Claude Haiku, which remains almost flat for the *same-domain* experiments. Furthermore, it is interesting to see the models’ ranking change between no prior context and the longest context scenario. In *same-domain* experiments, both Claude models climbed two ranks up from the bottom. In *cross-domain* experiments, Claude Haiku went up from last to third place, in contrast with Gemini Flash that went down from third to last place. The accuracy spread of models increases with the addition of more prior context. With no prior context, the highest-accuracy over lowest-accuracy ratio is 1.5, but at 64k-tokens it went up to 1.6 for *same-domain* and a staggering 3.6 for *cross-domain* scenarios.

In contrast, the *task-repeated* scenario seems to be much more robust, as can be seen in Figures 2c and 2d, with a maximum drop for models being 15% (GPT-4o-mini) and 12% (Claude Sonnet) for the *same-domain* and *cross-domain* experiments, respectively. The performance drops for the

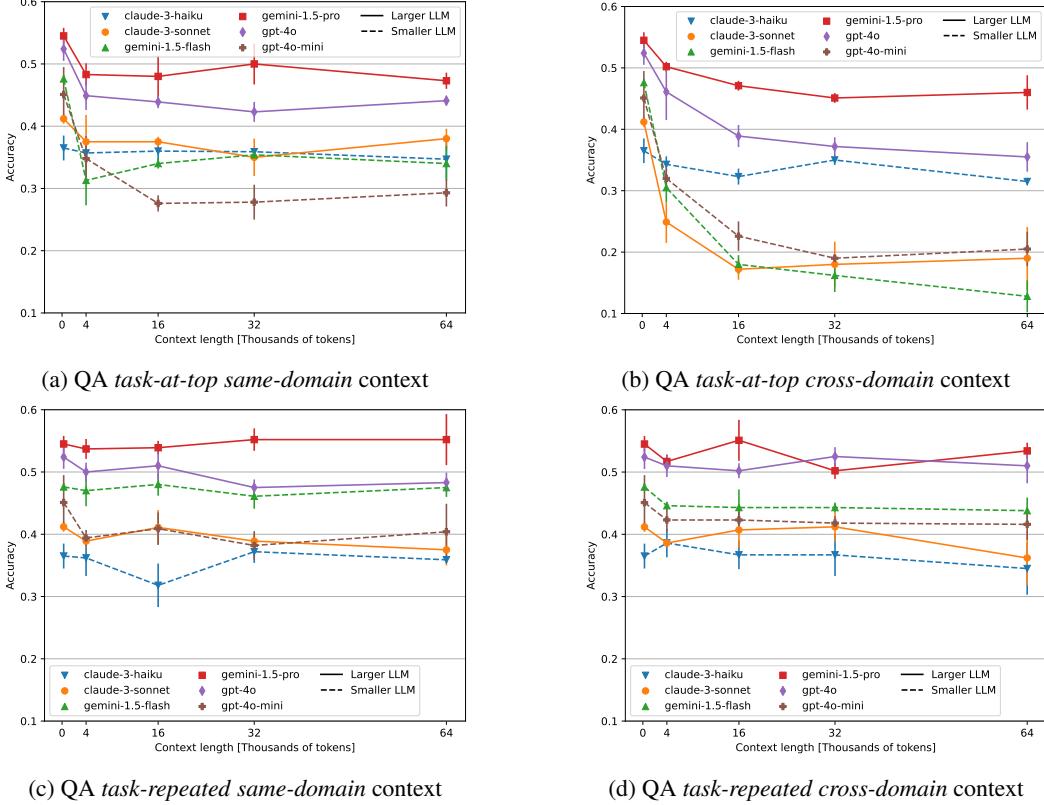


Figure 2: Experimental LLM performance over variable multi-turn *task-at-top* (top row), *task-repeated* (bottom row), *same-domain* (left column) and *cross-domain* (right column) contexts. Scores are mean relative accuracies across three runs with the standard deviations shown as error bars. Dashed and continuous lines represent smaller and larger LLMs respectively.

*task-repeated* scenario are more comparable to the *free-chat* scenario; the only noticeable difference is that Gemini models performed slightly worse in the latter. Same as with the *free-chat* results, no models ranking changes were observed for the *task-repeated* scenario. As for the highest-accuracy over lowest-accuracy ratio, it stays almost unchanged for both *same-* and *cross-domain* scenarios between no and longest prior context. This finding motivates repeated task placement as a mitigation strategy for degradation of model performance due to prior context.

**Task location influences significantly the model performance.** As highlighted earlier, we find repeated task placement to be an effective mitigation for the performance degrading effects of the prior context. The accuracies for the *task-repeated* scenario were significantly higher in some experiments compared to *task-at-top* ones. To better quantify the improvement after repeating the task before the query question, we plot in Figure 3a the accuracy ratio of *task-repeated* over *task-at-top* scenarios for both *same-domain* and *cross-domain* context. The biggest improvements can be seen in *cross-domain* experiments, where Gemini Flash accuracy improved by a factor of 3.4. For Claude Sonnet, the larger model within the Claude family, the accuracy increased from 19% to 36.2% when the task description was repeated in the *cross-domain* context of 64k length, representing nearly a doubling of performance. Another way to assess this improvement is to evaluate the amount of performance drop that was recovered after repeating the task description, as illustrated in Figure 3b. For GPT and Gemini models in a *cross-domain* scenario, 85% or more of the drop is recovered. For example, for GPT-4o-mini, the average accuracy drop in the highest two bins is 56% relative to the base *no-context* experiment for the *cross-domain task-at-top* context. In the *task-repeated* experiment, this drop is only 8%, indicating that 86% of the drop is recovered. Claude Haiku, which was the most stable in *task-at-top* experiments, has seen the lowest average improvements across the two domains, where both are compatible with no change in accuracy. This difference in performance regarding the two task-location scenarios is unexpected but has very important implications: shedding light on how

LLMs use information in different locations in the context and, as a consequence, on how to structure the prompt for best performance. These results suggest that reiterating the task description is an effective mitigation technique where long context separates the description and the query. A detailed investigation of this observation is discussed in Section 4.2.1.

**Larger models performance can become equal to or worse than smaller models.** Claude Sonnet model, which is larger than Claude Haiku model, has higher performance in the base *no-context* experiment (+13% relatively). For the *same-domain* context in *task-at-top* experiments (Figure 2a), Claude Sonnet's accuracy decreases with more added prior context, until it becomes similar to Claude Haiku's accuracy. However, for the *cross-domain* context (Figure 2b), Claude Sonnet's accuracy drops below that of Claude Haiku after just a few thousand tokens of additional context, reaching as low as half of Haiku's accuracy. Same as seen in the *free-chat* scenario, this observation shows the importance of benchmarks assessing the effect of prior context.

**Model performance worsens when the QA prior context and the target query are from different knowledge domains.** It can be argued that LLMs may focus on important and relevant information to answer the target query while disregarding non-relevant information. However, we found a surprising performance disparity between the *same-domain* and *cross-domain* experiments in the *task-at-top* scenario, as demonstrated empirically through the experiments. All models performed worse in the *task-at-top* experiments where prior context and target queries came from cross domains (Figure 2b), as opposed to those from the *same-domain* context (Figure 2a). For some results at 4k tokens length, the difference is negligible and can even go in the opposite direction due to fluctuations. In comparison, the models most affected by the *same-domain* context were GPT-4o-mini and Gemini Flash, with average accuracy drops in the highest two bins of 36% and 27% relative to the baseline *no-context* experiment. For the *cross-domain* context, however, the average drops in those same bins were 56% and 70%.

For the more robust *task-repeated* experiments, the performance differences are significantly smaller between *same-* and *cross-domain* contexts. The trend of lower performances for the latter is usually observed, except for GPT models, where the accuracy fluctuations are greater than the differences, making it difficult to draw a conclusive observation.

Overall, the observed difference in performance between *same-* and *cross-domain* contexts suggests that the model's internal representations favour context continuity. Furthermore, the performance of the models in both *same-domain* and *cross-domain* experiments becomes comparable when the task is repeated, indicating that the context, and the task location within it, significantly influences the contextual behaviour of models more than previously thought.

**Model performance declines with lengthier interactions.** For *same-domain* experiments in the *task-at-top* scenario, the performance drop flattens after 4k tokens, except for the GPT-4o-mini model, which continues to drop until 16k. In *cross-domain* experiments, the general trend is that performances keep dropping but at slower rates (except for a few statistically insignificant fluctuations). The two most performant models across the various context lengths are GPT-4o and Gemini Pro. Their accuracies drop by 16% (32%) and 14% (16%) relatively between *no-context* and 64k-tokens *same-domain* (*cross-domain*) context experiments.

In *task-repeated* scenarios, the model performance variations are small and comparable in magnitude to fluctuations, which makes it harder to detect the general trend in performance. For some models, most of the degradation occurs in the experiments with a 4k-token context length, while for others, small degradation continues until the 64k-token experiments. Ultimately, it is important to note that repeating task descriptions within the context proves effective in reducing the repercussions of long context.

#### 4.2.1 Understanding the Effect of Task Location

Given the large performance differences between *task-at-top* and *task-repeated* scenarios, it is important to understand why the task location has this significant influence on models. The hypothesis is that the task information is being diluted with longer context, and this was the initial motivation for having the two task location scenarios.

Qualitative analysis of the models responses reveals three primary sources of errors contributing to low performance scores: i) reasoning errors, where the model selects the wrong choice; ii) answer format deviations, where the correct choice was selected but the wrong formatting leads to a failed

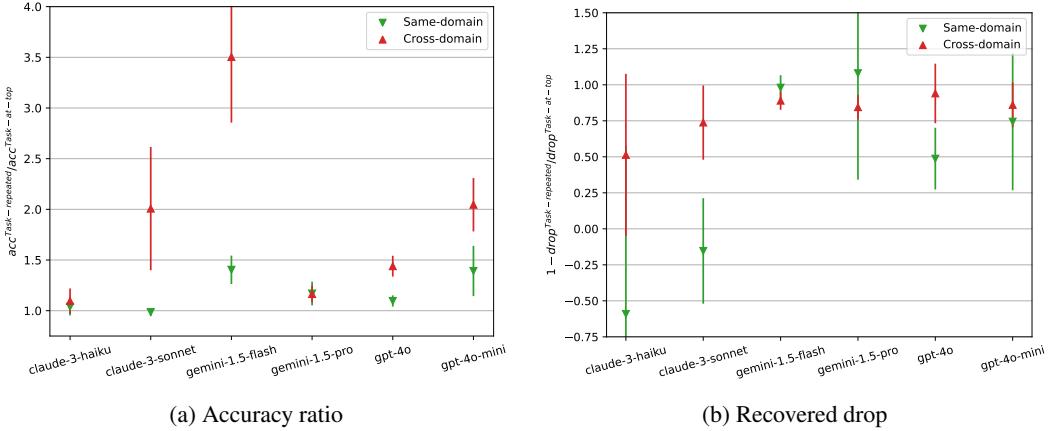


Figure 3: Performance improvements when repeating the task description before the query question. (a) Accuracy ratio of *task-repeated* over *task-at-top* scenarios for both *same-domain* and *cross-domain* multi-turn QA context and 64k-token length. (b) Amount of recovered performance drop after repeating the task (zero means no performance recovered, one mean full performance recovered, values smaller than zero or larger than one are due to statistical fluctuations).

answer extraction; and iii) models not answering the given task, where the output is not related to answering the question (e.g. becoming stuck in repetitive token sampling).

The most significant factor identified for the performance degradation under the *task-at-top* condition was formatting issues, where models frequently diverged from the expected response format and occasionally produced output in unintended formats, such as LaTeX, which contradicted the task description. This observation is non-intuitive, since having the task at the beginning primes the models, and all QA examples within the context adhere to the expected answer format, effectively functioning as few-shot learning instances for the models. While we expected that the adherence to the correct formatting will be reinforced more in such conditions, the results show the opposite effect.

Low model performance scores in *task-at-top* scenarios stems from multiple factors, especially the dilution of task-relevant information in longer contexts. Repeating the task towards the end appears to refocus the model’s attention on the key instructions, thereby ensuring the output complies with the specified format. Ultimately, these experimental results emphasize the importance of strategic task placement within the context, particularly for tasks that involve extensive inputs. By reintroducing task-specific details at key locations, we can greatly improve the performance of language models. These findings are important for any LLM application where descriptions and guidelines are presented at the beginning of the inputs. For example, it is common to define the persona, tone of voice, personalisation, and guidelines of chat models in the first conversation message as a system prompt. Also, in agentic application where multi-round of (self-)interaction are present, the agent guidelines and tools are also defined at the beginning. In those cases, it is important to evaluate the model on lengthy interactions and anticipate any performance degradation.

## 5 Conclusion and Future Work

In this study, we introduce novel benchmarks comprising various multi-turn conversations, and explore the sensitivity of several LLMs to prior context, evaluating their accuracies and robustness in various contexts and prompting scenarios. Our findings reveal the substantial influence that prior context as well as priming techniques have on LLM performance, which is more significant than previously thought. Relative accuracy drops as high as 73% are observed on GPQA questions. This finding was unanticipated, especially in some of our scenarios that are similar to few-shots prompting where one would expect more aligned responses. Our results show the inadequacy of relying on a single-turn benchmark to accurately reflect performance across multi-turn interactions. The impact of prior context varies depending on the content, even when it is unrelated to the current task. We also find that multi-turn *cross-domain* QA context harms models’ performances even more compared to *same-domain* context. Moreover, larger models are not necessarily more robust to the effects

of prior context, where, in some scenarios, their accuracies dropped below that of smaller models. Task information dilution contributes significantly to the degradation of accuracy. Repeating the task before the last query proves to be an effective mitigation strategy. This approach helps recover a substantial portion of the drop, with many models gaining twice as much or more in accuracy.

This variability highlights the need for a deeper understanding of LLM behaviour in contexts where sustained engagement and contextual awareness are required. It also emphasises the importance of benchmarks that evaluate the effect of prior context, such as the ones we introduced in this paper. This will guide the selection of LLMs that are not only proficient in focused reasoning, but also capable of maintaining consistent performance in real-world applications where context is nearly inevitable. Moreover, when engineering model prompts, one should account for the risk of information dilution when task description is introduced early in the conversation.

Future research may extend this study to include more varied context, such as different languages and code context, to capture a wider range of real-life scenarios; as well as apply this study to alternative benchmarks other than GPQA. Next, we aim to develop methods that assess the relevance of prior interactions in context and assess LLMs ability to differentiate between contextually dependent and independent queries, leading to more accurate and coherent responses. We will also evaluate the effect of chain-of-thought, prior context summarisation and selective memory retention techniques on the sensitivity to prior context.

## 6 Limitations

Our analysis is limited by the set of questions in LMSYS-1M-Chat and QA benchmark datasets (i.e., MMLU and GPQA), which may not capture the full diversity prior context LLM users may experience in real-world applications. In the prior context, the same given answers are used for all models to maintain experimental consistency. Alternatively, one could use the actual model answers to test for additional error propagation. Moreover, the models were evaluated in controlled settings, which, while necessary for consistency, may not fully represent the variability found in everyday natural language interactions, such as cultural and linguistic diversities, and more varied conversation history.

## Acknowledgements

We wish to express appreciation Graham Smith and Zachery Anderson of NatWest Group for the time and support needed to develop this research paper.

## References

- [1] Bryan Wang, Gang Li, and Yang Li. Enabling conversational interaction with mobile ui using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [3] Steven Moore, Richard Tong, Anjali Singh, Zitao Liu, Xiangen Hu, Yu Lu, Joleen Liang, Chen Cao, Hassan Khosravi, Paul Denny, et al. Empowering education with llms—the next-gen interface and content generation. In *International Conference on Artificial Intelligence in Education*, pages 32–37. Springer, 2023.
- [4] Philippe Laban, Jesse Vig, Marti Hearst, Caiming Xiong, and Chien-Sheng Wu. Beyond the chat: Executable and verifiable text-editing with llms. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–23, 2024.
- [5] Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. Large language models in education: Vision and opportunities. In *2023 IEEE international conference on big data (BigData)*, pages 4776–4785. IEEE, 2023.

- [6] Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19368–19376, 2024.
- [7] Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. A survey on recent advances in llm-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*, 2024.
- [8] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070, 2021.
- [9] Elliot Nelson, Georgios Kollias, Payel Das, Subhajit Chaudhury, and Soham Dan. Needle in the haystack for memory based large language models. *arXiv preprint arXiv:2407.01437*, 2024.
- [10] Erxin Yu, Jing Li, Ming Liao, Siqi Wang, Zuchen Gao, Fei Mi, and Lanqing Hong. Cosafe: Evaluating large language model safety in multi-turn dialogue coreference. *arXiv preprint arXiv:2406.17626*, 2024.
- [11] Zhiting Fan, Ruizhe Chen, Tianxiang Hu, and Zuozhu Liu. Fairmt-bench: Benchmarking fairness for multi-turn dialogue in conversational llms. *arXiv preprint arXiv:2410.19317*, 2024.
- [12] Chen Zhang, Xinyi Dai, Yaxiong Wu, Qu Yang, Yasheng Wang, Ruiming Tang, and Yong Liu. A survey on multi-turn interaction capabilities of large language models. *arXiv preprint arXiv:2501.09959*, 2025.
- [13] Izunna Okpala, Ashkan Golgoon, and Arjun Ravi Kannan. Agentic ai systems applied to tasks in financial services: Modeling and model risk management crews, 2025. URL <https://arxiv.org/abs/2502.05439>.
- [14] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.
- [15] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.
- [16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [17] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.
- [18] Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. Can large language models understand context? In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2004–2018, St. Julian’s, Malta, March 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.findings-eacl.135/>.
- [19] Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. Enhancing contextual understanding in large language models through contrastive decoding. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4225–4237, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.237. URL <https://aclanthology.org/2024.naacl-long.237/>.
- [20] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL <https://aclanthology.org/P19-1472/>.

- [21] Hanjun Luo, Haoyu Huang, Ziye Deng, Xuecheng Liu, Ruizhe Chen, and Zuozhu Liu. Bigbench: A unified benchmark for social bias in text-to-image generative models based on multi-modal llm. *arXiv preprint arXiv:2407.15240*, 2024.
- [22] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL <https://aclanthology.org/2022.acl-long.229/>.
- [23] Timothy R McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N Halgamuge. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *arXiv preprint arXiv:2402.09880*, 2024.
- [24] Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models, 2024. URL <https://arxiv.org/abs/2404.18824>.
- [25] Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. MT-eval: A multi-turn capabilities evaluation benchmark for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20153–20177, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1124. URL <https://aclanthology.org/2024.emnlp-main.1124/>.
- [26] Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024.
- [27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [28] Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, pages 1154–1156, 2021.
- [29] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [30] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhui Chen. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024.
- [31] Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. Bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277, 2024.
- [32] Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.
- [33] Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. Beyond the limits: A survey of techniques to extend the context length in large language models. *arXiv preprint arXiv:2402.02244*, 2024.
- [34] Yiran Ding, Li Lyra Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.
- [35] Qingyang Wu and Zhou Yu. Stateful memory-augmented transformers for dialogue modeling. *arXiv preprint arXiv:2209.07634*, 2022.

- [36] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36, 2024.
- [37] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL <https://aclanthology.org/2021.emnlp-main.243/>.
- [38] Junhua Liu, Yong Keat Tan, Bin Fu, and Kwan Hui Lim. Balancing accuracy and efficiency in multi-turn intent classification for llm-powered dialog systems in production. *arXiv preprint arXiv:2411.12307*, 2024.
- [39] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [40] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, and others. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv e-prints*, art. arXiv:2403.05530, March 2024. doi: 10.48550/arXiv.2403.05530.
- [41] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [42] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. Lmsys-chat-1m: A large-scale real-world llm conversation dataset, 2024. URL <https://arxiv.org/abs/2309.11998>.
- [43] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- [44] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35, 2023.
- [45] Herman Melville. *Moby-Dick; or, The Whale*. Project Gutenberg, 2001. URL <https://www.gutenberg.org/ebooks/2701>. Accessed: 2025-02-11.
- [46] Mary Shelley. *Frankenstein; or, The Modern Prometheus*. Project Gutenberg, 1993. URL <https://www.gutenberg.org/ebooks/84>. Accessed: 2025-02-11.

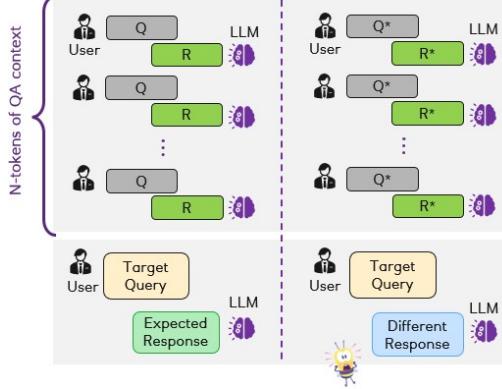


Figure 4: An illustration of varying context scenario in interactive settings where LLMs respond differently to the same target query when preceded with multiple turns of interactions.

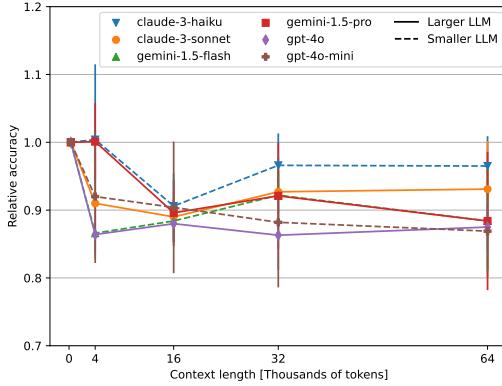


Figure 5: Experimental LLM performance, relative to the base *no-context* case, over variable *free-chat* context. Scores are mean accuracies across three runs with the standard deviations shown as error bars. Dashed and continuous lines represent smaller and larger LLMs respectively.

## A Illustration

Figure 4 illustrates the prior context experimentation idea.

## B Relative Accuracy Plots

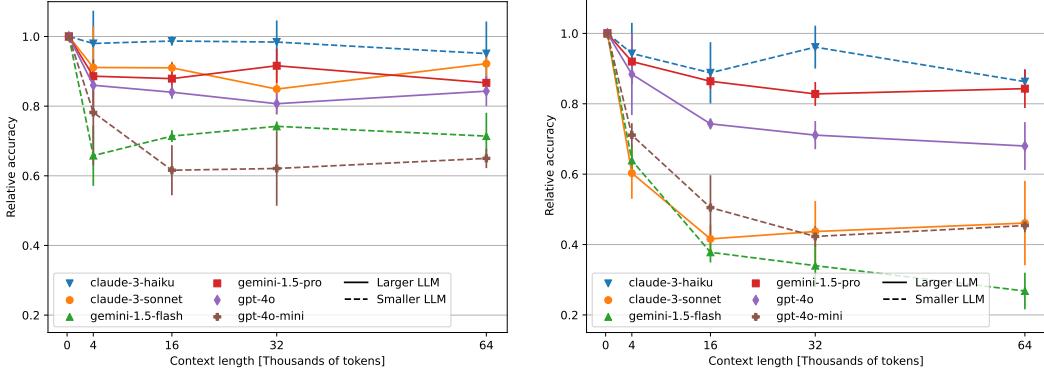
Figures 5, 6 and 7 show relative accuracies for the main experiments with respect to the base *no-context* case.

## C Task Location Experiment Results

Figures 8 and 9 shows the comparison results for the task location experiments for the *same-domain* and *cross-domain* context.

## D Unrelated-text Experiment

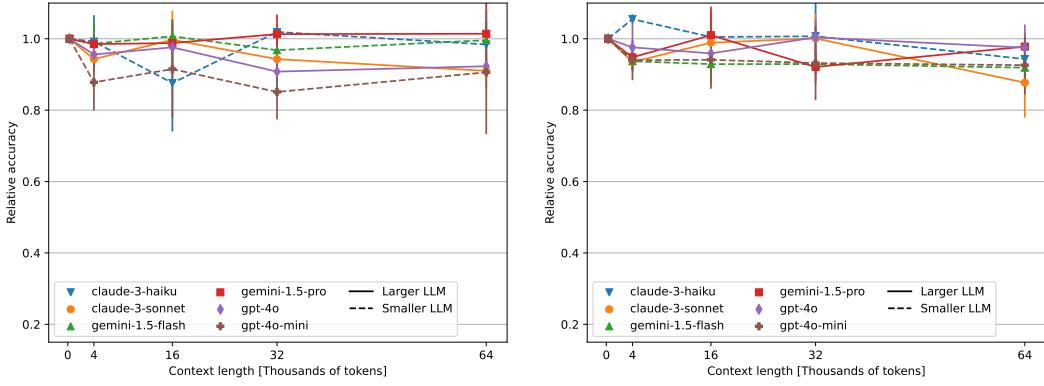
We also explore the effect on model performance when Unrelated English text is added to the context prior to the target question. To do this we use sentences from the two most common books from Project Gutenberg: MobyDick [45] and Frankenstein [46]. The unrelated text entails shuffled, concatenated passages of the books, chosen for their length and semantic distance from STEM domains, ensuring a substantial amount of unrelated context for the study. The different prior context



(a) QA task-at-top same-domain context

(b) QA task-at-top cross-domain context

Figure 6: Experimental LLM performance, relative to the base *no-context* case, over multi-turn *task-at-top same-domain* (a) and variable *cross-domain* (b) context. Scores are mean relative accuracies across three runs with the standard deviations shown as error bars. Dashed and continuous lines represent smaller and larger LLMs respectively.



(a) QA task-repeated same-domain context

(b) QA task-repeated cross-domain context

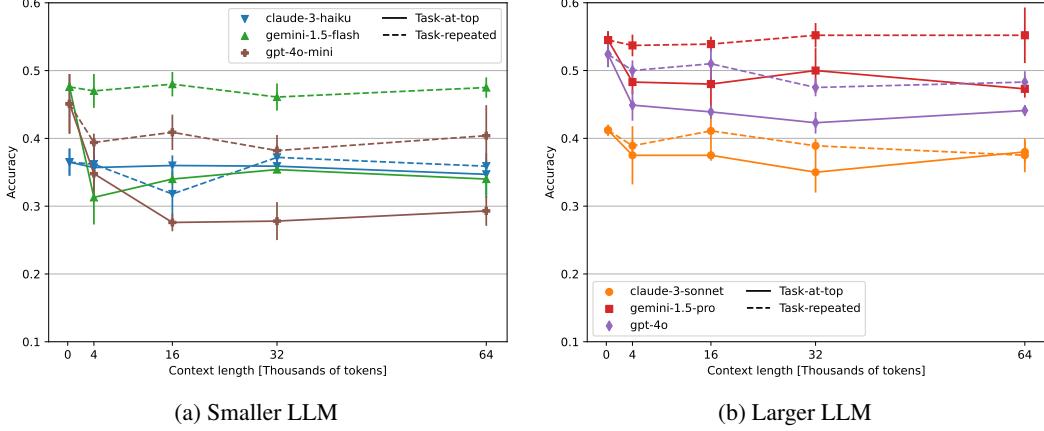
Figure 7: Experimental LLM performance, relative to the base *no-context* case, over variable multi-turn *task-repeated same-domain* context (a) and *cross-domain* (b) context. Scores are mean relative accuracies across three runs with the standard deviations shown as error bars. Dashed and continuous lines represent smaller and larger LLMs respectively.

lengths are extracted by hard truncation on the tokens level. The task description for the target query is placed after the text blob and before the target query for the unrelated text context, as a separation between the two different content.

**Results.** The presence of unrelated English text does not seem to significantly disrupt the performance of models as can be seen in figure 10. Most models exhibit almost flat performance across various context lengths. The exception is GPT-4o-mini, where its accuracy drops by approximately 10% in all experiments; however, the standard deviations of the accuracies are large, making it difficult to draw strong conclusions. We also observe many fluctuations, where sometimes the performance becomes slightly higher than in the baseline *no-context* scenario. This may be due to some slight model instability affecting the prediction probabilities, which can also be magnified by the non-zero value of the sampling temperature.

## E LLM Models and Versions

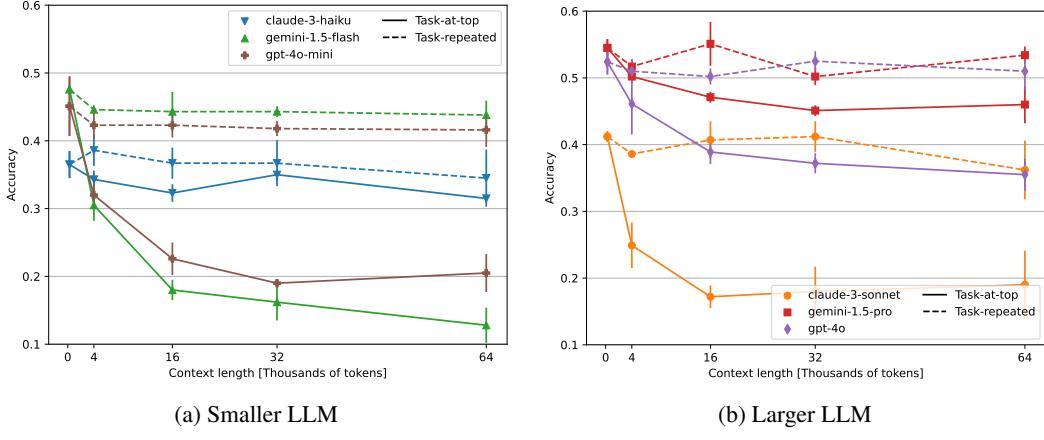
Table 2 shows what models and versions we used in our experiments.



(a) Smaller LLM

(b) Larger LLM

Figure 8: Comparing smaller (a) and larger (b) LLM performances in *task-at-top* and *task-repeated* scenarios over *same-domain* context. Scores are mean accuracies across three runs with the standard deviations shown as error bars. Continuous and dashed lines represent *task-at-top* and *task-repeated* experiments respectively.



(a) Smaller LLM

(b) Larger LLM

Figure 9: Comparing smaller (a) and larger (b) LLM performances in *task-at-top* and *task-repeated* scenarios over *cross-domain* context. Scores are mean accuracies across three runs with the standard deviations shown as error bars. Continuous and dashed lines represent *task-at-top* and *task-repeated* experiments respectively.

## F MMLU Subjects Selection

We used the following STEM subject for the *same-domain* context:

- High-School Biology
- High-School Physics
- High-School Chemistry
- Clinical Knowledge
- College Biology
- College Physics
- College Chemistry
- Anatomy
- Professional Medicine
- Virology
- Conceptual Physics
- Medical Genetics
- High-School Math

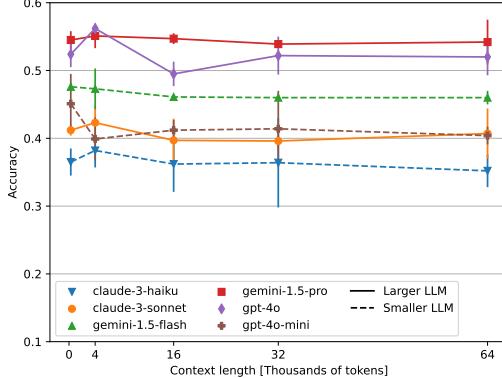


Figure 10: Experimental LLM performance over variable *unrelated-text* context. Scores are mean accuracies across three runs with the standard deviations shown as error bars. Dashed and continuous lines represent smaller and larger LLMs respectively.

Table 2: LLM Models, Model version, and Services for text-completion APIs.

Model	Model version	Service
GPT-4o	gpt-4o (2024-08-06)	Azure OpenAI
GPT-4o-mini	gpt-4o mini (2024-07-18)	Azure OpenAI
Claude 3 Sonnet	claude-3-sonnet-20240229-v1.0	AWS Bedrock
Claude 3 Haiku	claude-3-haiku-20240307-v1.0	AWS Bedrock
Gemini 1.5 Pro	gemini-1.5-pro-002	GCP Vertex AI
Gemini 1.5 Flash	gemini-1.5-flash-002	GCP Vertex AI

- College Math
- Abstract Algebra

We used the following Non-STEM subject for the *cross-domain* context:

- Business
- Global fact
- High-School Geography
- High-School Government and Politics
- High-School Psychology
- High-School US History
- High-School World History
- International Law
- Jurisprudence
- Logical Fallacies
- Management
- Marketing
- Philosophy
- Professional Law
- Professional Psychology
- Public Relations
- Sociology
- US Foreign Policy

## G Experiments Cost

Table 3 shows the estimated cost of our experiments.

Table 3: Experiment estimated cost from 15k API calls per model across 25 experiments and 3 runs.

Model	Cost (\$)
Claude 3 Haiku	117
Claude 3 Sonnet	1406
Gemini 1.5 Flash	34
Gemini 1.5 Pro	573
GPT-4o	1167
GPT-4o-mini	69
<b>Total</b>	<b>3367</b>

## H LLM Usage

This paper was written by the authors, with GPT-4o used as a tool for refining grammar, spelling, and clarity.

## I Data Usage

This paper makes use of three datasets: LMSYS-1M-Chat [42], GPQA [14] and MMLU [43] datasets, all of which are available under open-source licenses. We adhered to their usage terms, ensuring proper citation and compliance with the license provisions. LMSYS-1M-Chat contains one million real-world conversations with state-of-the-art LLMs, curated for understanding and advancing LLM capabilities. GPQA and MMLU are MCQ-based datasets. GPQA is an English-language dataset designed to evaluate general-purpose question-answering models, while MMLU is a large-scale benchmark that covers multiple subject areas to assess reasoning and knowledge. Additionally, we used Frankenstein [46] and Moby-Dick [45], both sourced from Project Gutenberg. These texts are in the public domain, and we used them in accordance with Project Gutenberg’s terms. All assets were used appropriately within the scope of research and in compliance with their respective licenses.

## J Full Results Table

Table 4 shows the full performance results for the main experiments, including mean accuracies and mean relative accuracies with respect to the base *no-context* experiment across three runs, and their standard deviation.

Table 4: Full LLM performance results, in terms of accuracy, across different context scenarios (various sources and lengths). Accuracies are reported as mean values of three runs, with standard deviations.

<b>Model</b>	<b>Context Length</b>	<b>Accuracy</b>	<b>Relative Accuracy</b>
Claude Haiku	no-context	0.365±0.02	1±0
Claude Haiku	free-chat-4k	0.365±0.021	1.004±0.111
Claude Haiku	unrelated-text-4k	0.382±0.025	1.046±0.031
Claude Haiku	multi-turn-qa-4k-cross-domain-task-at-top	0.343±0.013	0.943±0.087
Claude Haiku	multi-turn-qa-4k-cross-domain-task-repeated	0.386±0.023	1.055±0.013
Claude Haiku	multi-turn-qa-4k-same-domain-task-at-top	0.357±0.018	0.98±0.094
Claude Haiku	multi-turn-qa-4k-same-domain-task-repeated	0.362±0.029	0.992±0.074
Claude Haiku	free-chat-16k	0.332±0.036	0.906±0.048
Claude Haiku	unrelated-text-16k	0.362±0.041	0.99±0.091
Claude Haiku	multi-turn-qa-16k-cross-domain-task-at-top	0.323±0.013	0.888±0.087
Claude Haiku	multi-turn-qa-16k-cross-domain-task-repeated	0.367±0.023	1.005±0.047
Claude Haiku	multi-turn-qa-16k-same-domain-task-at-top	0.36±0.015	0.987±0.013
Claude Haiku	multi-turn-qa-16k-same-domain-task-repeated	0.318±0.035	0.876±0.136
Claude Haiku	free-chat-32k	0.354±0.036	0.966±0.047
Claude Haiku	unrelated-text-32k	0.364±0.066	0.991±0.128
Claude Haiku	multi-turn-qa-32k-cross-domain-task-at-top	0.35±0.008	0.961±0.061
Claude Haiku	multi-turn-qa-32k-cross-domain-task-repeated	0.367±0.034	1.007±0.108
Claude Haiku	multi-turn-qa-32k-same-domain-task-at-top	0.359±0.005	0.984±0.062
Claude Haiku	multi-turn-qa-32k-same-domain-task-repeated	0.372±0.018	1.019±0.009
Claude Haiku	free-chat-64k	0.352±0.006	0.965±0.044
Claude Haiku	unrelated-text-64k	0.352±0.024	0.963±0.028
Claude Haiku	multi-turn-qa-64k-cross-domain-task-at-top	0.315±0.006	0.863±0.034
Claude Haiku	multi-turn-qa-64k-cross-domain-task-repeated	0.345±0.042	0.943±0.076
Claude Haiku	multi-turn-qa-64k-same-domain-task-at-top	0.347±0.031	0.951±0.092
Claude Haiku	multi-turn-qa-64k-same-domain-task-repeated	0.359±0.005	0.984±0.066
Claude Sonnet	no-context	0.412±0.008	1±0
Claude Sonnet	free-chat-4k	0.375±0.034	0.91±0.081
Claude Sonnet	unrelated-text-4k	0.423±0.025	1.025±0.074
Claude Sonnet	multi-turn-qa-4k-cross-domain-task-at-top	0.249±0.034	0.603±0.073
Claude Sonnet	multi-turn-qa-4k-cross-domain-task-repeated	0.386±0.006	0.935±0.03
Claude Sonnet	multi-turn-qa-4k-same-domain-task-at-top	0.375±0.043	0.911±0.118
Claude Sonnet	multi-turn-qa-4k-same-domain-task-repeated	0.389±0.009	0.943±0.025
Claude Sonnet	free-chat-16k	0.367±0.019	0.89±0.037
Claude Sonnet	unrelated-text-16k	0.397±0.032	0.963±0.068
Claude Sonnet	multi-turn-qa-16k-cross-domain-task-at-top	0.172±0.017	0.416±0.041
Claude Sonnet	multi-turn-qa-16k-cross-domain-task-repeated	0.407±0.028	0.989±0.086
Claude Sonnet	multi-turn-qa-16k-same-domain-task-at-top	0.375±0.008	0.91±0.018
Claude Sonnet	multi-turn-qa-16k-same-domain-task-repeated	0.411±0.028	0.997±0.082
Claude Sonnet	free-chat-32k	0.382±0.011	0.927±0.031
Claude Sonnet	unrelated-text-32k	0.396±0.008	0.959±0.027
Claude Sonnet	multi-turn-qa-32k-cross-domain-task-at-top	0.18±0.037	0.437±0.087
Claude Sonnet	multi-turn-qa-32k-cross-domain-task-repeated	0.412±0.023	1.001±0.069
Claude Sonnet	multi-turn-qa-32k-same-domain-task-at-top	0.35±0.03	0.849±0.073
Claude Sonnet	multi-turn-qa-32k-same-domain-task-repeated	0.389±0.01	0.943±0.025
Claude Sonnet	free-chat-64k	0.384±0.031	0.931±0.071
Claude Sonnet	unrelated-text-64k	0.407±0.037	0.987±0.074
Claude Sonnet	multi-turn-qa-64k-cross-domain-task-at-top	0.19±0.051	0.461±0.12
Claude Sonnet	multi-turn-qa-64k-cross-domain-task-repeated	0.362±0.044	0.877±0.098
Claude Sonnet	multi-turn-qa-64k-same-domain-task-at-top	0.38±0.016	0.922±0.037
Claude Sonnet	multi-turn-qa-64k-same-domain-task-repeated	0.375±0.025	0.91±0.051
Gemini Flash	no-context	0.476±0.006	1±0
Gemini Flash	free-chat-4k	0.412±0.008	0.866±0.007
Gemini Flash	unrelated-text-4k	0.473±0.03	0.993±0.071
Gemini Flash	multi-turn-qa-4k-cross-domain-task-at-top	0.305±0.023	0.639±0.042

Model	Context Length	Accuracy	Relative Accuracy
Gemini Flash	multi-turn-qa-4k-cross-domain-task-repeated	0.446±0.006	0.937±0.021
Gemini Flash	multi-turn-qa-4k-same-domain-task-at-top	0.313±0.04	0.658±0.087
Gemini Flash	multi-turn-qa-4k-same-domain-task-repeated	0.47±0.025	0.986±0.064
Gemini Flash	free-chat-16k	0.421±0.011	0.884±0.031
Gemini Flash	unrelated-text-16k	0.461±0.003	0.968±0.01
Gemini Flash	multi-turn-qa-16k-cross-domain-task-at-top	0.18±0.015	0.378±0.029
Gemini Flash	multi-turn-qa-16k-cross-domain-task-repeated	0.443±0.029	0.929±0.053
Gemini Flash	multi-turn-qa-16k-diff-taskinlast	0.441±0.024	0.926±0.054
Gemini Flash	multi-turn-qa-16k-same-domain-task-at-top	0.34±0.008	0.714±0.017
Gemini Flash	multi-turn-qa-16k-same-domain-task-repeated	0.48±0.018	1.007±0.048
Gemini Flash	free-chat-32k	0.439±0.02	0.922±0.033
Gemini Flash	unrelated-text-32k	0.46±0.005	0.965±0.022
Gemini Flash	multi-turn-qa-32k-cross-domain-task-at-top	0.162±0.027	0.34±0.061
Gemini Flash	multi-turn-qa-32k-cross-domain-task-repeated	0.443±0.008	0.929±0.007
Gemini Flash	multi-turn-qa-32k-same-domain-task-at-top	0.354±0.005	0.742±0.014
Gemini Flash	multi-turn-qa-32k-same-domain-task-repeated	0.461±0.02	0.968±0.038
Gemini Flash	free-chat-64k	0.421±0.034	0.884±0.083
Gemini Flash	unrelated-text-64k	0.46±0.01	0.965±0.032
Gemini Flash	multi-turn-qa-64k-cross-domain-task-at-top	0.128±0.026	0.268±0.052
Gemini Flash	multi-turn-qa-64k-cross-domain-task-repeated	0.438±0.021	0.919±0.048
Gemini Flash	multi-turn-qa-64k-same-domain-task-at-top	0.34±0.028	0.714±0.067
Gemini Flash	multi-turn-qa-64k-same-domain-task-repeated	0.475±0.015	0.996±0.022
Gemini Pro	no-context	0.545±0.013	1±0
Gemini Pro	free-chat-4k	0.545±0.018	1.001±0.057
Gemini Pro	unrelated-text-4k	0.551±0.018	1.009±0.027
Gemini Pro	multi-turn-qa-4k-cross-domain-task-at-top	0.502±0.008	0.92±0.013
Gemini Pro	multi-turn-qa-4k-cross-domain-task-repeated	0.517±0.011	0.948±0.032
Gemini Pro	multi-turn-qa-4k-same-domain-task-at-top	0.483±0.018	0.886±0.02
Gemini Pro	multi-turn-qa-4k-same-domain-task-repeated	0.537±0.016	0.985±0.026
Gemini Pro	free-chat-16k	0.488±0.015	0.896±0.049
Gemini Pro	unrelated-text-16k	0.547±0.008	1.003±0.024
Gemini Pro	multi-turn-qa-16k-cross-domain-task-at-top	0.471±0.008	0.864±0.021
Gemini Pro	multi-turn-qa-16k-cross-domain-task-repeated	0.551±0.033	1.01±0.08
Gemini Pro	multi-turn-qa-16k-same-domain-task-at-top	0.48±0.033	0.879±0.045
Gemini Pro	multi-turn-qa-16k-same-domain-task-repeated	0.539±0.011	0.988±0.021
Gemini Pro	free-chat-32k	0.502±0.03	0.921±0.078
Gemini Pro	unrelated-text-32k	0.539±0.006	0.988±0.019
Gemini Pro	multi-turn-qa-32k-cross-domain-task-at-top	0.451±0.008	0.828±0.034
Gemini Pro	multi-turn-qa-32k-cross-domain-task-repeated	0.502±0.013	0.921±0.046
Gemini Pro	multi-turn-qa-32k-same-domain-task-at-top	0.5±0.033	0.916±0.049
Gemini Pro	multi-turn-qa-32k-same-domain-task-repeated	0.552±0.018	1.013±0.055
Gemini Pro	free-chat-64k	0.481±0.044	0.884±0.102
Gemini Pro	unrelated-text-64k	0.542±0.033	0.995±0.082
Gemini Pro	multi-turn-qa-64k-cross-domain-task-at-top	0.46±0.028	0.843±0.055
Gemini Pro	multi-turn-qa-64k-cross-domain-task-repeated	0.534±0.013	0.978±0.005
Gemini Pro	multi-turn-qa-64k-same-domain-task-at-top	0.473±0.013	0.867±0.013
Gemini Pro	multi-turn-qa-64k-same-domain-task-repeated	0.552±0.041	1.014±0.095
GPT-4o	no-context	0.524±0.019	1±0
GPT-4o	free-chat-4k	0.453±0.03	0.864±0.029
GPT-4o	unrelated-text-4k	0.562±0.003	1.075±0.042
GPT-4o	multi-turn-qa-4k-cross-domain-task-at-top	0.461±0.046	0.884±0.116
GPT-4o	multi-turn-qa-4k-cross-domain-task-repeated	0.51±0.018	0.976±0.067
GPT-4o	multi-turn-qa-4k-same-domain-task-at-top	0.449±0.023	0.86±0.062
GPT-4o	multi-turn-qa-4k-same-domain-task-repeated	0.5±0.015	0.956±0.048
GPT-4o	free-chat-16k	0.461±0.031	0.88±0.028
GPT-4o	unrelated-text-16k	0.495±0.018	0.947±0.065
GPT-4o	multi-turn-qa-16k-cross-domain-task-at-top	0.389±0.018	0.743±0.015
GPT-4o	multi-turn-qa-16k-cross-domain-task-repeated	0.502±0.012	0.959±0.047

Model	Context Length	Accuracy	Relative Accuracy
GPT-4o	multi-turn-qa-16k-same-domain-task-at-top	0.439±0.01	0.84±0.019
GPT-4o	multi-turn-qa-16k-same-domain-task-repeated	0.51±0.022	0.976±0.066
GPT-4o	free-chat-32k	0.451±0.013	0.863±0.051
GPT-4o	unrelated-text-32k	0.522±0.028	0.998±0.062
GPT-4o	multi-turn-qa-32k-cross-domain-task-at-top	0.372±0.015	0.711±0.04
GPT-4o	multi-turn-qa-32k-cross-domain-task-repeated	0.525±0.015	1.004±0.05
GPT-4o	multi-turn-qa-32k-same-domain-task-at-top	0.423±0.016	0.807±0.03
GPT-4o	multi-turn-qa-32k-same-domain-task-repeated	0.475±0.013	0.908±0.057
GPT-4o	free-chat-64k	0.458±0.013	0.875±0.035
GPT-4o	unrelated-text-64k	0.52±0.027	0.996±0.085
GPT-4o	multi-turn-qa-64k-cross-domain-task-at-top	0.355±0.024	0.68±0.068
GPT-4o	multi-turn-qa-64k-cross-domain-task-repeated	0.51±0.028	0.975±0.065
GPT-4o	multi-turn-qa-64k-same-domain-task-at-top	0.441±0.008	0.843±0.043
GPT-4o	multi-turn-qa-64k-same-domain-task-repeated	0.483±0.016	0.923±0.024
GPT-4o-mini	no-context	0.451±0.044	1±0
GPT-4o-mini	free-chat-4k	0.412±0.011	0.92±0.098
GPT-4o-mini	unrelated-text-4k	0.399±0.032	0.892±0.127
GPT-4o-mini	multi-turn-qa-4k-cross-domain-task-at-top	0.32±0.016	0.711±0.034
GPT-4o-mini	multi-turn-qa-4k-cross-domain-task-repeated	0.423±0.019	0.94±0.056
GPT-4o-mini	multi-turn-qa-4k-diff-taskinlast	0.402±0.028	0.901±0.145
GPT-4o-mini	multi-turn-qa-4k-same-domain-task-at-top	0.348±0.036	0.782±0.152
GPT-4o-mini	multi-turn-qa-4k-same-domain-task-repeated	0.394±0.013	0.878±0.079
GPT-4o-mini	free-chat-16k	0.406±0.026	0.904±0.097
GPT-4o-mini	unrelated-text-16k	0.412±0.015	0.918±0.054
GPT-4o-mini	multi-turn-qa-16k-cross-domain-task-at-top	0.226±0.024	0.505±0.092
GPT-4o-mini	multi-turn-qa-16k-cross-domain-task-repeated	0.423±0.018	0.941±0.081
GPT-4o-mini	multi-turn-qa-16k-same-domain-task-at-top	0.276±0.013	0.616±0.072
GPT-4o-mini	multi-turn-qa-16k-same-domain-task-repeated	0.409±0.026	0.915±0.137
GPT-4o-mini	free-chat-32k	0.396±0.026	0.882±0.096
GPT-4o-mini	unrelated-text-32k	0.414±0.056	0.92±0.119
GPT-4o-mini	multi-turn-qa-32k-cross-domain-task-at-top	0.19±0.006	0.423±0.029
GPT-4o-mini	multi-turn-qa-32k-cross-domain-task-repeated	0.418±0.011	0.932±0.104
GPT-4o-mini	multi-turn-qa-32k-same-domain-task-at-top	0.278±0.028	0.621±0.107
GPT-4o-mini	multi-turn-qa-32k-same-domain-task-repeated	0.382±0.023	0.851±0.077
GPT-4o-mini	free-chat-64k	0.392±0.048	0.869±0.058
GPT-4o-mini	unrelated-text-64k	0.404±0.013	0.901±0.089
GPT-4o-mini	multi-turn-qa-64k-cross-domain-task-at-top	0.205±0.028	0.454±0.019
GPT-4o-mini	multi-turn-qa-64k-cross-domain-task-repeated	0.416±0.025	0.926±0.082
GPT-4o-mini	multi-turn-qa-64k-same-domain-task-at-top	0.293±0.022	0.65±0.028