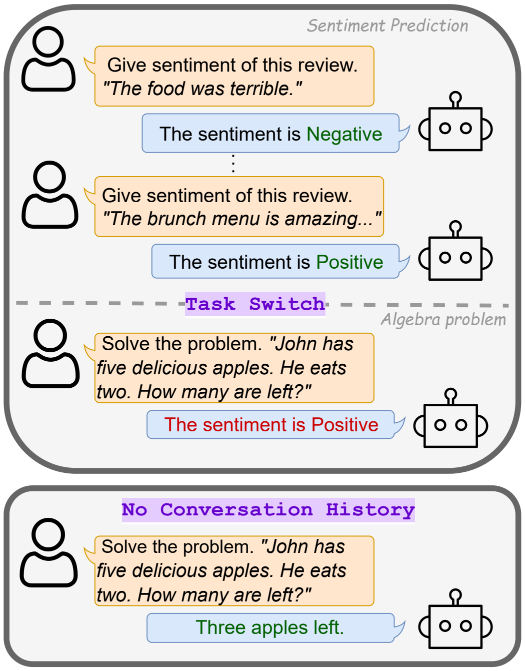
LLM Task Interference: Impact of Task-Switch in Conversational History

Akash Gupta [*1] **Ivaxi Sheth** [*2] **Vyas Raina** [*1] **Mark Gales** [1] **Mario Fritz** [2]

Abstract

With the recent emergence of powerful instruction-tuned large language models (LLMs), various helpful conversational Artificial Intelligence (AI) systems have been deployed across many applications. When prompted by users, these AI systems successfully perform a wide range of tasks as part of a conversation. To provide some sort of memory and context, such approaches typically condition their output on the entire conversational history. Although this sensitivity to the conversational history can often lead to improved performance on subsequent tasks, we find that performance can in fact also be negatively impacted, if there is a *task-switch*. To the best of our knowledge, our work makes the first attempt to formalize the study of such vulnerabilities and interference of tasks in conversational LLMs caused by task-switches in the conversational history. Our experiments across 5 datasets with 15 task switches using popular LLMs reveal that many of the task-switches can lead to significant performance degradation. [0]



Give sentiment of this review. *"The food was terrible."*

The sentiment is Negative

⋮

Give sentiment of this review. *"The brunch menu is amazing..."*

The sentiment is Positive

Solve the problem. *"John has five delicious apples. He eats two. How many are left?"*

The sentiment is Positive

Solve the problem. *"John has five delicious apples. He eats two. How many are left?"*

Three apples left.

Figure 1. An illustrative example where the chat history is based on sentiment prediction. Algebra word problem introduces *task-switch* which results in an incorrect prediction.

1. Introduction

Recent advancements in Natural Language Processing (NLP) (Brown et al., 2020; OpenAI, 2023), have led to their widespread deployment of large language models (LLMs) across various applications (Bubeck et al., 2023; Anil et al., 2023; Singhal et al., 2022). One of the popular NLP tasks includes conversational systems where LLMs are capable of engaging in dialogues that mimic human interactions (Manyika & Hsiao, 2023; Bai et al., 2022). A typical interaction involves a series of conversation turns starting with the user and the LLM responds to the user. This interaction is however focused on a specific topic or a task (Hosseini-Asl et al., 2020; Lee et al., 2022).

The performance of LLMs is further boosted by leveraging in-context examples or few-shot examples of a particular task (Brown et al., 2020; Smith et al., 2022; Thoppilan et al., 2022). In-context learning, by utilizing examples within the conversation history, enables LLMs to generate responses that are relevant and tailored to the contextual conversation. The auto-regressive nature of popular instruction-tuned (LLMs) suggests that the LLM generated response is conditioned on the entire conversation history. This underscores the sequential dependency and contextual awareness embedded within these models. While prompt sensitivity has been exploited by in-context learning to improve downstream performance, this sensitivity has also opened the door to vulnerabilities, where malicious actors can exploit prompt

[*] Equal contribution [1] Department of Engineering,University of Cambridge,Cambridge,UK [2] CISPA Helmholtz Center for Information Security, Saarbrücken, Germany. Correspondence to: Akash Gupta <ag2118@cam.ac.uk>, Ivaxi Sheth <ivaxi.sheth@cispa.de>.

Published at ICML 2024 Workshop on Foundation Models in the Wild. Copyright 2024 by the author(s).

[0] Code available on GitHub.

sensitivity for adverse purposes (Greshake et al., 2023; Liu et al., 2023; Jiang et al., 2023b; Xu et al., 2023).

In this paper, we investigate the sensitivity and the impact of LLM performance on past conversational interaction. To do so, we introduce the concept of *task-switch*. A task-switch is characterized by a conversational objective, moving from one distinct task to another within the same conversation thread, for example: Figure 1 illustrates a task-switch from sentiment prediction to math algebra which confuses the model to output erroneously. Designing LLMs that can seamlessly switch between tasks without degradation in performance can influence the reliability of LLMs in realistic scenarios.

In this work, we systematically study the impact of predictive performance and the sensitivity of LLMs in the presence of different task-based chat histories. Our key contributions and takeaways can be summarised as:

- We formalize the risk of performance degradation of LLMs due to task-switch.
- We present the impact of task-switch on diverse datasets with more than 15 different task-switches.
- We measure the task-switch sensitivity for popular LLMs of different sizes, where we observe that in some cases very large (175B) and small (7B) LLMs can both be susceptible to performance degradation from task-switch.

2. Related Work

Large Language Models (LLMs) are becoming a crucial building block of conversation-based virtual assistants (OpenAI, 2023; Touvron et al., 2023; Jiang et al., 2023a; Anil et al., 2023). Leveraging in-context or few-shot examples, LLMs have demonstrated remarkable capabilities for downstream tasks (Brown et al., 2020). In contrast to the resource-intensive fine-tuning process (Gao et al., 2020), in-context learning eliminates the need for parameter updates, while achieving state-of-the-art performance (Rae et al., 2021; Smith et al., 2022; Thoppilan et al., 2022; Von Oswald et al., 2023; Chan et al., 2022; Akyürek et al., 2022; Hahn & Goyal, 2023). However, despite its advantages, in-context learning tends to suffer from sensitivity to prompts, input distribution, and formats, which can potentially impact the model's performance (Liu et al., 2021; Zhao et al., 2021; Lu et al., 2021; Min et al., 2022; Liu & Wang, 2023; Chang & Jia, 2023). Chang & Jia (2023) observe that the in-context examples implicitly bias the model. In our work, we aim to study the bias that arises due to chat history (in-context examples) when a user switches the task. Furthermore, recent works (Liu et al., 2023; Greshake et al., 2023) have looked at the vulnerability of LLM to prompt injections and adversarial attacks. Unlike prompt injection, where a malicious prompt may be added to the conversation of LLM, our setting, is concerned with non-malicious task-switches. While a few recent works have investigated the reliance on shortcuts in conversation history (Tang et al., 2023; Si et al., 2022; Weston & Sukhbaatar, 2023), our work aims to evaluate prompt history sensitivity for a new task. Our work is also differentiated from the study topic change in Task-oriented Dialogue systems (Xie et al., 2021; Xu et al., 2021; Yang et al., 2022) as we consider a stronger shift of task-switch from open dialogue LLMs.

3. Conversational Task-Switch

This work introduces and formalizes *task-switch* in a conversation for LLMs. A conversation between a user and the LLM consists of multiple conversation turns. Now consider (u_k, r_k) as the k-th turn of the conversation where u_k corresponds to the k-th user prompt and the model's corresponding response r_k. Each user prompt u_k can be viewed as an instance of a specific task request, e.g. *sentiment classification* or *mathematical reasoning*. A conversation history of L turns can be defined as $\mathbf{h} = \{(u_k, r_k)\}_{k=1}^{L}$. Subsequently, the next response, r_{L+1} for model θ is given as:

$$r_{L+1} = \arg\max_{r} P_\theta(r|u_{L+1}, \mathbf{h}) \tag{1}$$

In this work, we consider conversations with a single task-switch, where all user requests in the conversation history $\mathbf{h}$ belong to the same task and the final user request u_{L+1} is a different task. We refer to the task associated with $\mathbf{h}$ as the conversation history task (*CH task*) T_h where $\mathbf{h} \in T_h$ and the switched task associated with the final user request u_{L+1} as the *target task* T_t where $u_{L+1} \in T_t$.

When the tasks T_h and T_t are sufficiently different (as per human understanding of language and tasks), the conversation history $\mathbf{h}$ ideally must not impact the response, r_{L+1}. For a model robust to such a task-switches, $T_h \rightarrow T_t$, its response r_{L+1} is conditionally independent of the conversation history,

$$r_{L+1} \perp \mathbf{h} \mid u_{L+1} \qquad \mathbf{h} \in T_h, \;\; u_{L+1} \in T_t. \tag{2}$$

However, in practice, models can be sensitive to the conversation history, $\mathbf{h}$, which can harm the quality of the response r_{L+1} after a task-switch, $T_h \rightarrow T_t$. We define $\tau(\cdot)$, the *task-switch sensitivity* of a model θ, to measure the extent of this vulnerability.[1]

$$\tau(T_h, T_t; \theta) = \mathbb{E}_{u_{L+1} \in T_t, \mathbf{h} \in T_h} \left[\log \rho\right] \tag{3}$$

$$\rho = \frac{P_\theta\left(r^*|u_{L+1}\right)}{P_\theta\left(r^*|u_{L+1}, \mathbf{h}\right)} \tag{4}$$

$$r^* = \arg\max_{r} P_\theta(r|u_{L+1}). \tag{5}$$

[1] Theoretical and empirical implications of other definitions for task-switch sensitivity in Appendix D

Task-switch sensitivity can be interpreted as:

1. $\tau(\cdot) > 0$: The model is impacted by the task-switch in the conversation history and is less confident in zero-shot prediction.
2. $\tau(\cdot) = 0$: The task-switch has no impact on the model's zero-shot prediction, suggesting a level of task-switch robustness.
3. $\tau(\cdot) < 0$: The task-switch gives the model more confidence in its zero-shot prediction.

To simulate a setting where the model has perfect performance on the CH-task, T_h we adopt teacher-forcing, s.t. $\mathbf{h} = \{(u_k, \hat{r}_k)\}_{k=1}^{L}$, where $\hat{r}$ is the reference ground-truth response.

4. Experiments

4.1. Experimental Setup

Data. We evaluate five different datasets covering a range of tasks: Gigaword (Graff et al., 2003); abstract algebra subset of Measuring Massive Multitask Language Understanding (MMLU; Hendrycks et al. (2021)), named MMLU AA; TweetQA (Xiong et al., 2019); Rotten Tomatoes (RT; Pang & Lee (2005)); and human-aging subset from the MMLU dataset (MMLU HA). Table 1 summarizes the nature of the task for each dataset. In the the main paper, we consider two datasets for the target tasks in a conversation with a task-switch: MMLU AA and RT. In Appendix A, B, we present results for the remaining datasets as the target tasks: MMLU HA, Gigaword, and TweetQA. The train-test splits of these datasets are shown in Table 2. The train set is randomly sampled to form prompts to produce a conversation history $\mathbf{h}$ of L turns, and the test set is used to evaluate model performance on the $(L+1)$-th turn. The prompt templates used for each dataset are discussed in Appendix C. For classification tasks performance is measured using accuracy, whilst for generative tasks it is measured using ROUGE (Lin, 2004) or METEOR (Banerjee & Lavie, 2005).

Table 1. Dataset Task Description.

DATA	TASK
GIGAWORD	SUMMARIZATION
MMLU AA	MATH MULTIPLE CHOICE QUESTION
TWEETQA	SOCIAL QUESTION ANSWER
RT	SENTIMENT CLASSIFICATION
MMLU HA	SOCIAL MULTIPLE CHOICE QUESTION

Models. We explore the task-switch sensitivity of four popular models. We consider two open-source small models, Llama2-7b-chat (Touvron et al., 2023) and Mistral-7b-chat (Jiang et al., 2023a); and two larger closed models, GPT-3.5 (Brown et al., 2020) and GPT-4 (OpenAI, 2023). Zero-shot, absolute model performances are presented in Appendix A.

Table 2. Dataset Statistics. QA: Question-Answering. MCQ: Multiple Choice Question

DATA	#TRAIN	#TEST	TASK
MMLU HA	26	222	SOCIAL MCQ
MMLU AA	14	99	MATH MCQ
RT	8.53K	1.07K	SENTIMENT CLASS
GIGAWORD	3.8M	1.95K	SUMMARIZATION
TWEETQA	4.54K	583	SOCIAL QA

4.2. Results

In addition to the task-switch sensitivity $\tau(\cdot)$, we assess performance changes between the predictions in the presence of history and task-switch vs zero-shot. Table 3 and Table 4 showcases the impact of conversational task-switch with MMLU AA and Rotten Tomatoes as the target tasks, T_t respectively[2]. As would be expected with *in-context examples*, the performance change in accuracy is generally positive. The negative trend for change in accuracy from $T_h \rightarrow T_t$, suggests that the task-switch causes performance degradation. For example, in the Gigaword summarization task as T_h and MMLU AA as T_t, most models (GPT-3.5, Llama-7B and Mistral-7B) see a performance drop. Interestingly, for some models, the task-switch may increase performance; most prominently for Mistral-7B with Rotten Tomatoes as T_h and MMLU AA as T_t.

The sensitivity of different models to different task-switches can be compared fairly using the task-switch metric, $\tau(\cdot)$ The larger the value of $\tau(\cdot)$, the greater a model's sensitivity to a specific task-switch. In Table 3 and Table 4, Llama-7B usually has the highest sensitivity to task-switches with for example $\tau = 3.37$ for a switch from MMLU AA to Rotten Tomatoes and $\tau = 9.91$ for task-switch from Rotten Tomatoes to MMLU AA. We observe a general trend between the change in accuracy and $\tau(\cdot)$ for task-switch scenarios for $T_t =$ Rotten Tomatoes where a negative change in performance also suggests very high task-switch sensitivity. In Figure 2, we plot the change in performance with increasing T_h examples for MMLU AA dataset. Here we can clearly observe that in-context examples improve the predictive performance. Notably, the accuracy variation is more pronounced in smaller 7B models, likely due to their lower baseline performance, which is substantially improved by in-context learning. Performance fluctuations for conversation history, $\mathbf{h}$, can stem from two primary factors: a significant drop in the predicted probability for the

[2] The impact of task-switch for other datasets as the target tasks is given in Appendix B.1

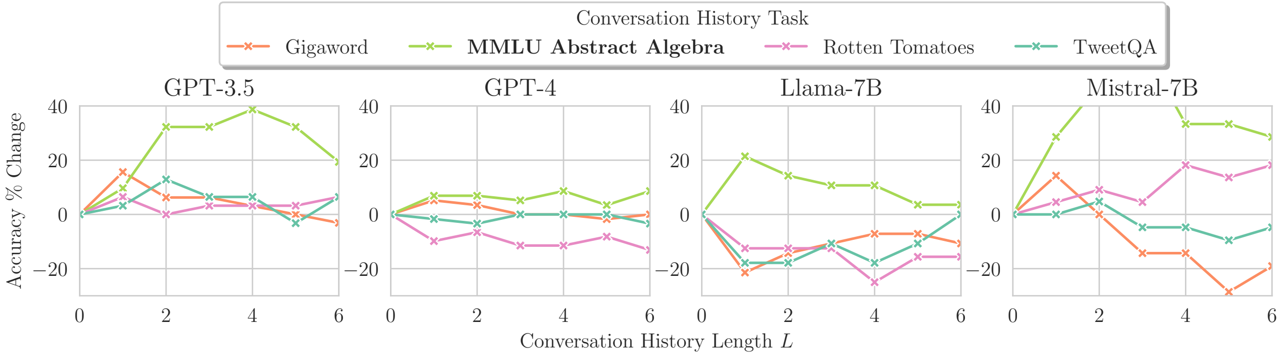


Figure 2. Target Task: MMLU Abstract Algebra. % change in accuracy relative to zero-shot performance.

Table 3. Task-switch impact from CH-tasks (T_h) to target (T_t): **MMLU AA** and conversation length L = 6. Sensitivity not calculable for $*$.

CH-Task	Model	% Change	$\tau(\cdot)$
MMLU AA	GPT-3.5	*17.17*	$*$
	GPT-4	*-1.09*	$*$
	Llama-7B	*0.00*	*31.51*
	Mistral-7B	*37.68*	*1.12*
Gigaword	GPT-3.5	−12.12	$*$
	GPT-4	−8.74	$*$
	Llama-7B	−18.75	5.23
	Mistral-7B	−21.74	3.13
Rotten Tomatoes	GPT-3.5	2.02	$*$
	GPT-4	−8.20	$*$
	Llama-7B	−12.50	9.91
	Mistral-7B	11.59	0.83
TweetQA	GPT-3.5	−19.19	$*$
	GPT-4	−8.20	$*$
	Llama-7B	−12.50	6.37
	Mistral-7B	−7.25	2.78

Table 4. Task-switch impact from CH-tasks (T_h) to target (T_t): **Rotten Tomatoes** and conversation length L = 6. Sensitivity not calculable for $*$.

CH-Task	Model	% Change	$\tau(\cdot)$
Rotten Tomatoes	GPT-3.5	*3.00*	$*$
	GPT-4	*1.74*	$*$
	Llama-7B	*2.54*	*4.02*
	Mistral-7B	*3.17*	*2.65*
Gigaword	GPT-3.5	0.11	$*$
	GPT-4	−0.98	$*$
	Llama-7B	1.82	1.98
	Mistral-7B	−0.79	3.04
MMLU AA	GPT-3.5	−0.22	$*$
	GPT-4	0.76	$*$
	Llama-7B	−5.33	3.37
	Mistral-7B	1.33	1.39
TweetQA	GPT-3.5	−0.33	$*$
	GPT-4	−0.98	$*$
	Llama-7B	2.72	2.77
	Mistral-7B	−1.23	3.01

zero-shot response, r^*, or a notable increase in the probability for an alternative response, r. The latter can result in substantial performance change while maintaining low sensitivity, $\tau(\cdot)$. By analyzing both performance changes and task-switch sensitivity, we gain deeper insights into the models' adaptability to task-switches and the underlying dynamics influencing these shifts.

5. Conclusions and Future Work

This work formalizes and performs an initial investigation into the sensitivity of large language models (LLMs) to task-switch scenarios within conversational contexts. We introduce a task-sensitivity metric that can explain a model's behavior to task-switches along with the performance change. By experimenting with various task-switch settings, we observe that even advanced models like GPT-4 can exhibit vulnerabilities to task-switches. Our work additionally lays the foundation for future work on 'side-channel' vulnerabilities of LLMs to undesired information leakage/bias from the conversation history. Further work will focus on developing adaptive context management strategies within LLMs to mitigate the risk of task-switch sensitivity.

6. Limitations

Although both GPT-3.5 and GPT-4 show degradation in performance, given the closed nature of OpenAI models, we were not able to perform task sensitivity analysis. We were additionally limited by the maximum token length, hence analysis over extremely long conversations was not feasible. Future work could also look into alignment between humans and the model as a metric which was out of the scope for this paper.

Impact Statement

As LLMs become increasingly powerful, they will be deployed in a range of real-world settings as virtual conversational assistants that can perform a multitude of tasks in a single session. In this work, we are the first to formalise a risk associated with a *task-switch* in this setting: current state-of-the-art LLMs can suffer from performance degradation when a user switches between tasks in a conversation. Therefore, we urge the community to design methods that mitigate the risk of task-switch sensitivity, so that LLMs can be deployed with fewer vulnerabilities as conversational assistants.

References

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? investigations with linear models. *arXiv*, 2022.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv*, 2022.

Banerjee, S. and Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W05-0909.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv*, 2023.

Chan, S. C., Dasgupta, I., Kim, J., Kumaran, D., Lampinen, A. K., and Hill, F. Transformers generalize differently from information stored in context vs in weights. *arXiv preprint arXiv:2210.05675*, 2022.

Chang, T.-Y. and Jia, R. Data curation alone can stabilize in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8123–8144, 2023.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Gao, T., Fisch, A., and Chen, D. Making pre-trained language models better few-shot learners. *arXiv*, 2020.

Graff, D., Kong, J., Chen, K., and Maeda, K. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4 (1):34, 2003.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pp. 79–90, 2023.

Hahn, M. and Goyal, N. A theory of emergent in-context learning as implicit structure induction. *arXiv*, 2023.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Hosseini-Asl, E., McCann, B., Wu, C.-S., Yavuz, S., and Socher, R. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191, 2020.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv*, 2023a.

Jiang, S., Chen, X., and Tang, R. Prompt packer: Deceiving llms through compositional instruction with hidden attacks. *arXiv preprint arXiv:2310.10077*, 2023b.

Lee, H., Gupta, R., Rastogi, A., Cao, Y., Zhang, B., and Wu, Y. Sgd-x: A benchmark for robust generalization in schema-guided dialogue systems. In *AAAI*, volume 36, pp. 10938–10946, 2022.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Liu, H. and Wang, Y. Towards informative few-shot prompt with maximum information gain for in-context learning. *arXiv preprint arXiv:2310.08923*, 2023.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

Liu, Y., Deng, G., Li, Y., Wang, K., Zhang, T., Liu, Y., Wang, H., Zheng, Y., and Liu, Y. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*, 2023.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

Manyika, J. and Hsiao, S. An overview of bard: an early experiment with generative ai. *AI. Google Static Documents*, 2, 2023.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

OpenAI, R. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2:13, 2023.

Pang, B. and Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv*, 2021.

Si, C., Friedman, D., Joshi, N., Feng, S., Chen, D., and He, H. What spurious features can pretrained language models combat? 2022.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *arXiv*, 2022.

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.

Tang, R., Kong, D., Huang, L., and Xue, H. Large language models can be lazy learners: Analyze shortcuts in in-context learning. *ACL Findings*, 2023.

Team, G., Anil, R., and et al. Gemini: A family of highly capable multimodal models, 2024.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv*, 2023.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent.

In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Weston, J. and Sukhbaatar, S. System 2 attention (is something you might need too). *arXiv*, 2023.

Xie, H., Liu, Z., Xiong, C., Liu, Z., and Copestake, A. Tiage: A benchmark for topic-shift aware dialog modeling. *ACL*, 2021.

Xiong, W., Wu, J., Wang, H., Kulkarni, V., Yu, M., Guo, X., Chang, S., and Wang, W. Y. Tweetqa: A social media focused question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Xu, X., Kong, K., Liu, N., Cui, L., Wang, D., Zhang, J., and Kankanhalli, M. An llm can fool itself: A prompt-based adversarial attack. *arXiv preprint arXiv:2310.13345*, 2023.

Xu, Y., Zhao, H., and Zhang, Z. Topic-aware multi-turn dialogue modeling. In *AAAI*, volume 35, pp. 14176–14184, 2021.

Yang, C., Lin, Z., Li, J., Meng, F., Wang, W., Wang, L., and Zhou, J. Take: topic-shift aware knowledge selection for dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 253–265, 2022.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pp. 12697–12706. PMLR, 2021.

Appendix

Appendix A reports the zero-shot absolute performance of all models on all tasks, Appendix B presents an ablation study on the conversation history length (with multiple seeds), Appendix C discusses the prompt templates, Appendix D discusses other definitions for task-switch sensitivity, Appendix E discusses correlations, and Appendix F tabulates confusion matrices for each model.

A. Absolute Performance

When evaluating the target task with a conversation history, it is useful to compare the performance against a baseline with no conversation history ($\mathbf{h} = \emptyset, L = 0$). This is equivalent to evaluating in a zero-shot setting. This section reports the zero-shot performance for all the target task (T_t) datasets: MMLU HA in Table 5, MMLU AA in Table 6, RT in Table 7, Gigaword in Table 8 and TweetQA in Table 9. Also note that for the classification tasks (MMLU HA, MMLU AA, RT), we also report the number of responses for which we were unable to extract the answer (# Format Errors), which is further discussed in Appendix C. We evaluate on the test set with four LLMs (GPT-3.5, GPT-4, Mistral-7B, Llama-7B), which were all set to Temperature 0 for reproducability.

Table 5. Zero-shot performance on **MMLU HA**.

Model	Accuracy	# Format Errors
GPT-3.5	66.22	18
GPT-4	84.68	0
Llama-7B	45.50	12
Mistral-7B	55.41	0

Table 6. Zero-shot performance on **MMLU AA**.

Model	Accuracy	# Format Errors
GPT-3.5	31.31	7
GPT-4	58.59	0
Llama-7B	28.28	3
Mistral-7B	21.21	0

Table 7. Zero-shot performance on **RT**.

Model	Accuracy	# Format Errors
GPT-3.5	89.90	0
GPT-4	91.80	4
Llama-7B	87.43	1
Mistral-7B	86.68	1

Table 8. Zero-shot performance on **Gigaword**.

Model	ROUGE-1	ROUGE-2	ROUGE-L
GPT-3.5	17.37	4.79	14.78
GPT-4	15.76	4.07	13.34
Llama-7B	11.61	3.13	9.90
Mistral-7B	18.60	5.19	15.84

Table 9. Zero-shot performance on **TweetQA**.

Model	ROUGE-1	ROUGE-L	METEOR
GPT-3.5	30.66	30.39	44.18
GPT-4	28.03	27.68	43.41
Llama-7B	17.91	17.67	33.84
Mistral-7B	25.35	25.01	40.71

B. Conversation History Length Ablation

This section presents an ablation study on the performance change after a task-switch for varying conversation history lengths. For each dataset in Table 2 we select four datasets (including itself), from which we use the training set as conversation history. The details of the prompt structure are presented in Appendix C.

B.1. Task-switch Performance Change

We compare the percentage change in metrics relative to zero-shot performance ($\mathbf{h} = \emptyset$, i.e. no conversation history) as a function of conversation history length L and for different LLMs. Results are plot in Figures 3, 4, 5, 6, 7 for MMLU HA, MMLU AA, RT, Gigaword and TweetQA respectively. When there is *not* a task switch, we would expect a performance increase (assuming the training examples are representative of the test set). As per our discussion in Section 4.2, we observe that different models degrade on different task-switches and this is not limited by the model size.

B.2. Format Failure Rate

Typically, classification tasks (MMLU HA, MMLU AA, RT) are evaluated using logits, however we use a generative approach for consistency: we are evaluating the model in a conversational setting, and we do not have access to the logits exactly. Thus, we must post-process the model output to determine the class. In this, we try to give the LLM the benefit of the doubt and do our best to extract the class. For example, although the prompt requests the model to output within answer tags like "`<Answer> positive </Answer>`", we also accept "`positive`", but we do not accept "`positive/negative`". Due to the imperfect nature of this setup, either we may not detect the correct format, or the model generates erroneous text.

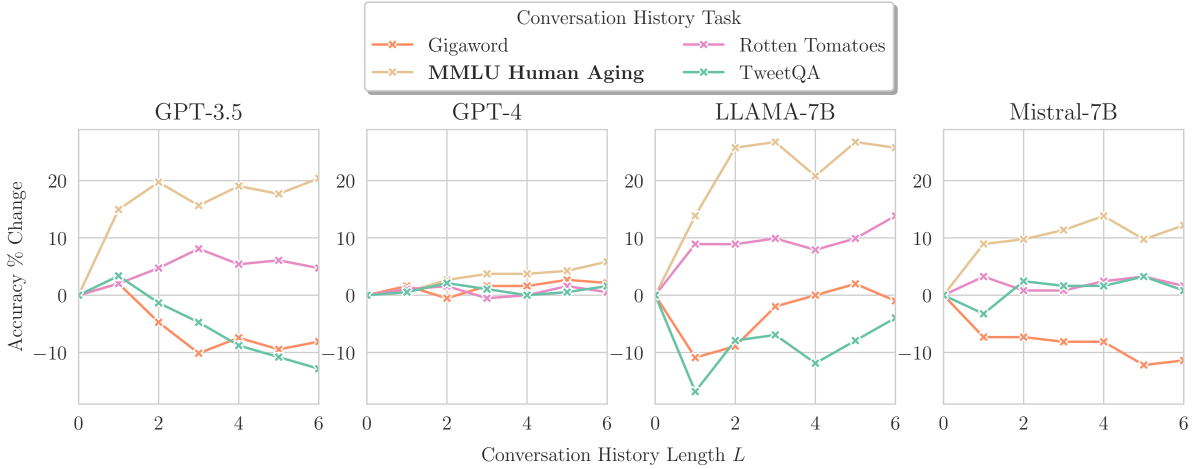


Figure 3. Target Task: MMLU HA. Percentage % change in accuracy relative to zero-shot performance (no conversation history) for increasing conversation history length L and various models.

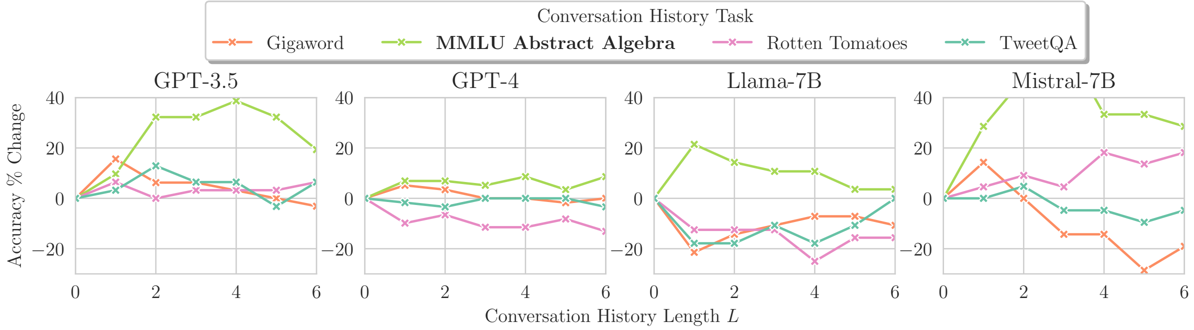


Figure 4. Target Task: MMLU AA. Percentage % change in accuracy relative to zero-shot performance (no conversation history) for increasing conversation history length L and various models.

Importantly, models may become more susceptible to these errors when performing a task-switch, causing performance degradation. We capture this by reporting the percentage % change in the number of examples that the model failed on (relative to zero-shot) as the context history length increases. These are plot in Figures 8, 9, 10 for MMLU HA, MMLU AA and RT respectively. Figures 8 and 9 show that GPT-3.5 and Mistral-7B are susceptible to format errors in task-switches when evaluating on multiple choice questions, whereas Figure 10 shows that GPT-4 and Llama-7B are more susceptible in sentiment classification.

B.3. Performance Variance

Presented experimental results in the main paper are the average across multiple seeds. However, it can be of interest to understand the extent to which the results can vary across multiple runs, as this provides an error bound on the worst-case and best-case scenarios. In this section we present the variance around the mean results for the models LLama-7b and Mistral-7b when evaluated on the target tasks Rotten Tomatoes (Fig 11) and MMLU-AA (Fig 12) with conversation history lengths $L \in \{0, 3, 6\}$.

C. Prompt Template

In each conversation turn, the user prompts the model u_k. The prompts are shown in Table 10. We chose these prompts after careful research and experimentation. We began with popular templates and refined them for our purpose.

Additionally, since we do not have access to the logits for all models, we take a generative approach to the classification tasks (MMLU HA, MMLU AA, RT). Since the model may fail to output an answer in the desired format, we post process the text to extract the answer (which we count as a positive result it matches the reference). We report and discuss the effect of format failures further in B.2. Furthermore, we note that the standard evaluation method used in the Open-LLM leaderboard code (available on GitHub) is to

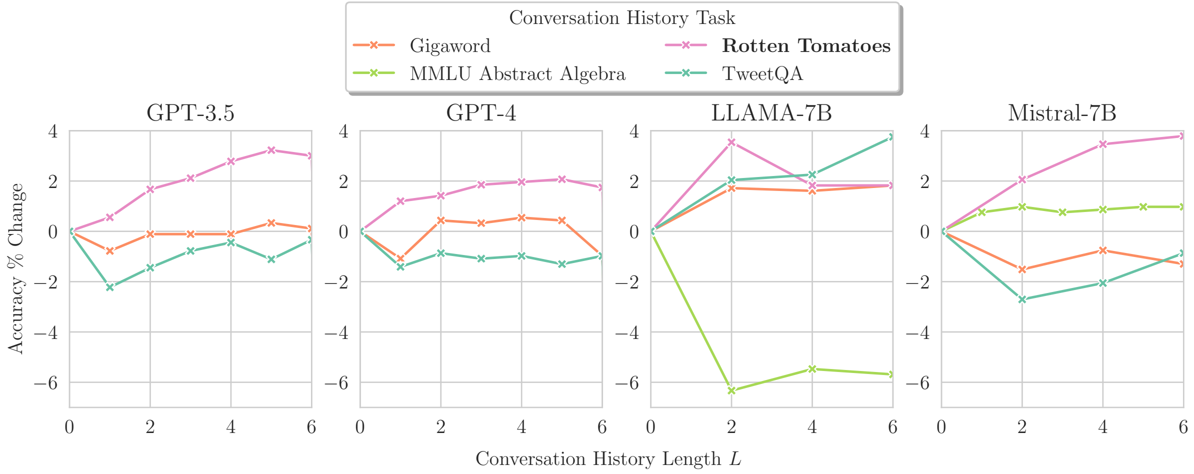
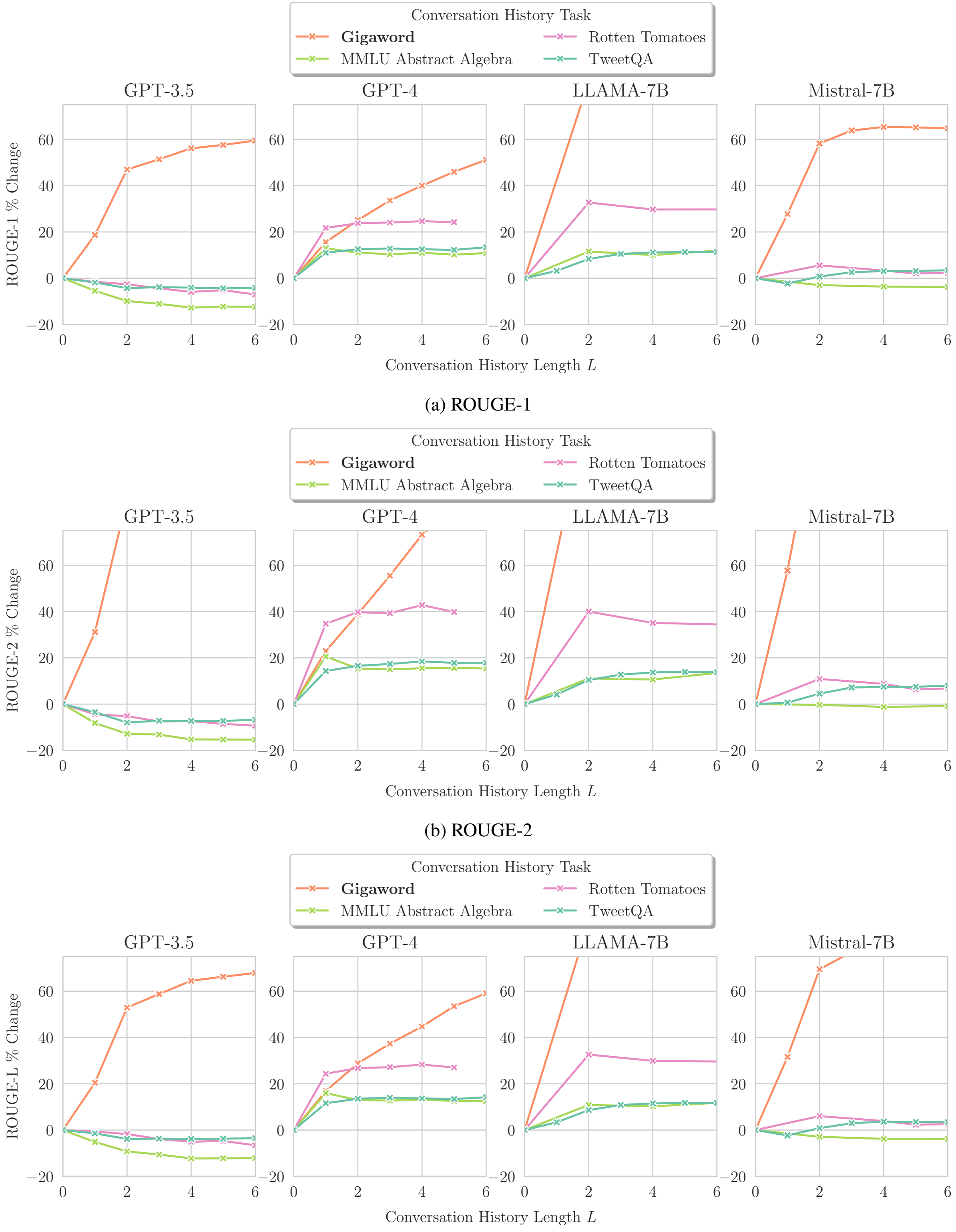


Figure 5. Target Task: RT. Percentage % change in accuracy relative to zero-shot performance (no conversation history) for increasing conversation history length L and various models.

see if the response starts with `A,B,C` or `D`(Gao et al., 2023). We modified the prompt to ensure a more consistent output format (across the different models) resulting in fewer mistakes made.

For the classification tasks, we structure the prompt such that we request the model to output their final answer within answer tags. We note that giving an example of how to use the answer tags always helped, however, this can bias the model towards a particular answer. Instead, we found for MMLU to just leave the answer tags empty, whereas for RT to have the all the sentiment classes inside the tags (see Table 10 for further details).

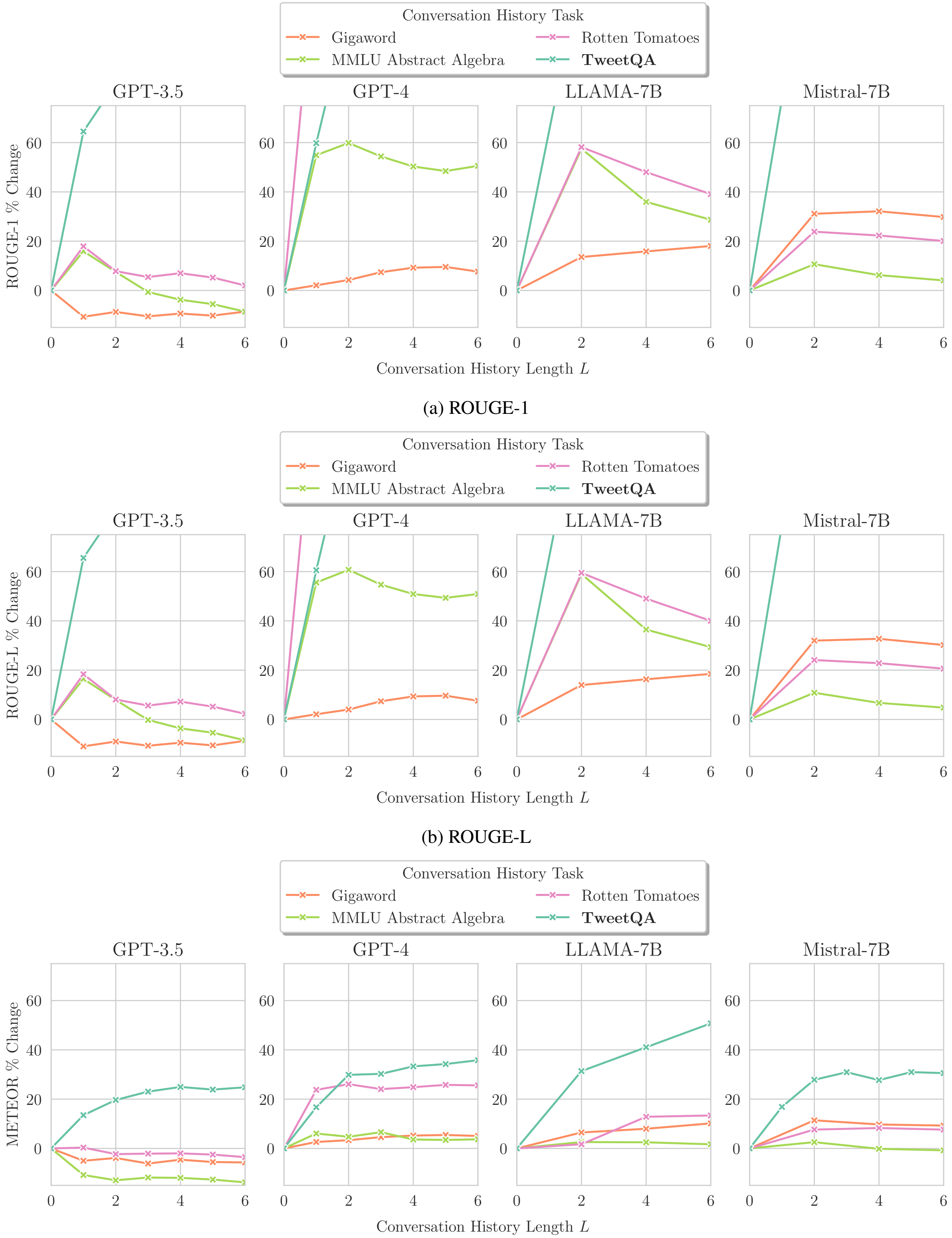


(a) ROUGE-1

(b) ROUGE-2

(c) ROUGE-L

Figure 6. Target Task: Gigaword. Percentage % change in accuracy relative to zero-shot performance (no conversation history) for increasing conversation history length L and various models. Note that we focus on the effect of task-switching by clipping the y-axes at +75%.



(a) ROUGE-1

(b) ROUGE-L

(c) METEOR

Figure 7. Target Task: TweetQA. Percentage % change in accuracy relative to zero-shot performance (no conversation history) for increasing conversation history length L and various models. Note that we focus on the effect of task-switching by clipping the y-axes at +75%.

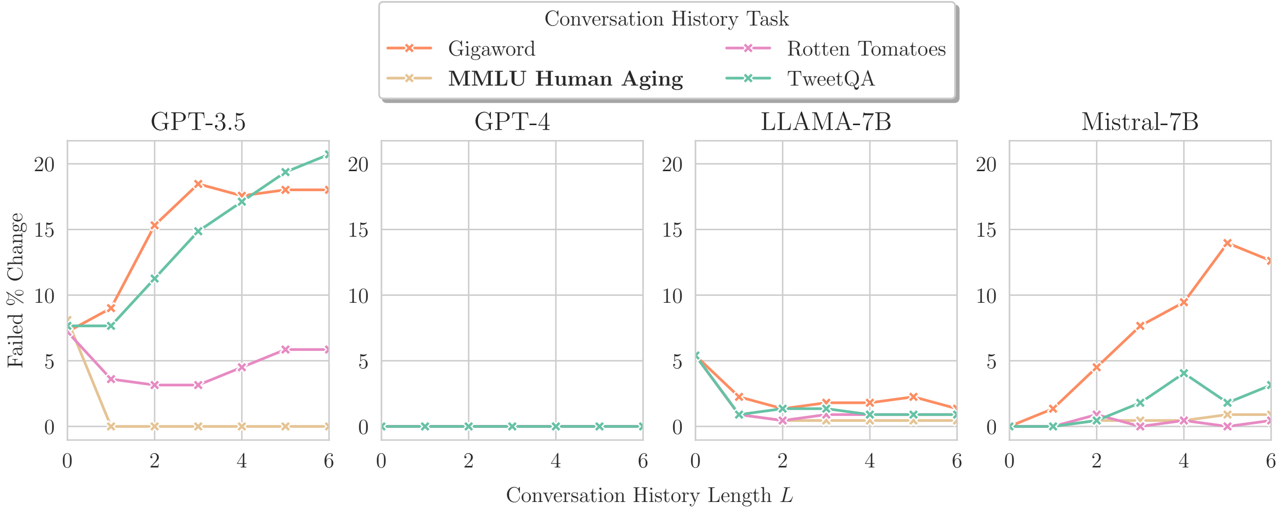


Figure 8. Target Task: MMLU Human Aging. Percentage % of examples where format failed.

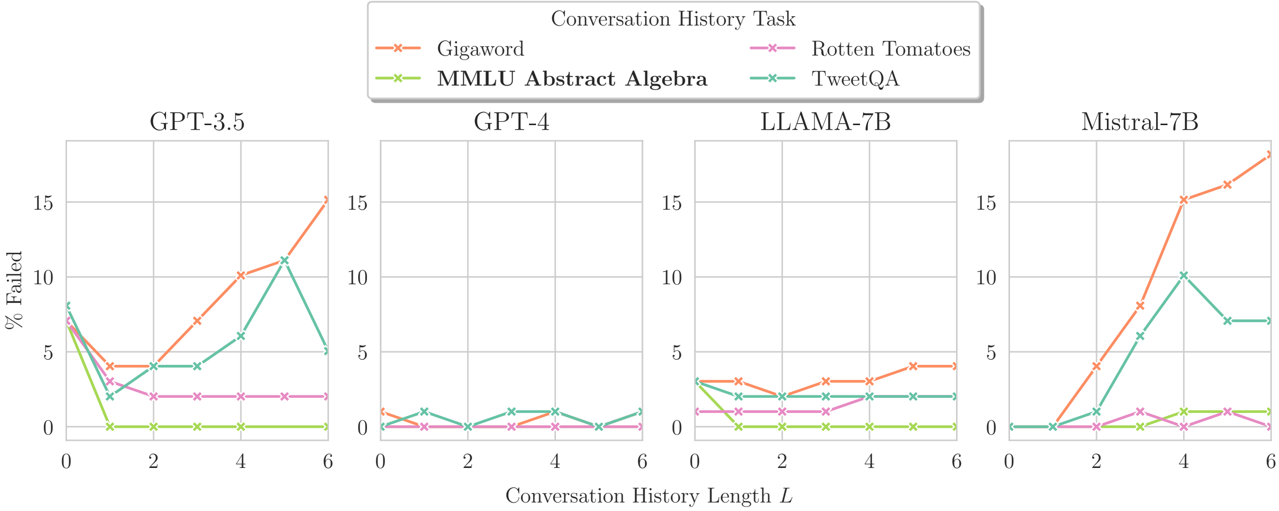


Figure 9. Target Task: MMLU Abstract Algebra. Percentage % of examples where format failed.

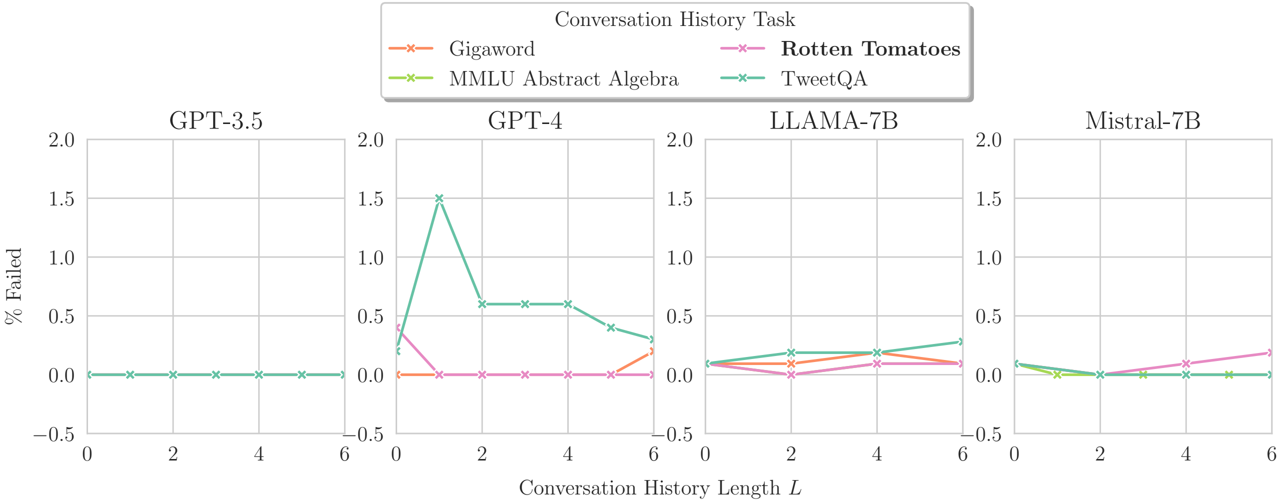


Figure 10. Target Task: Rotten Tomatoes. Percentage % of examples where format failed.

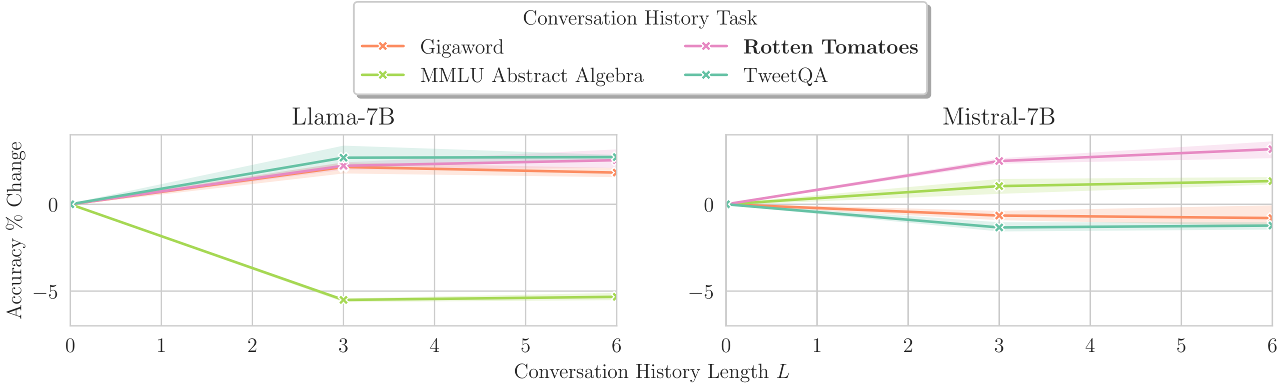


Figure 11. Target Task: RT. Percentage % change in accuracy relative to zero-shot performance for increasing conversation history length L for multiple seeds. Mean is shown in solid line, and the shaded region is bounded by the min/max values.

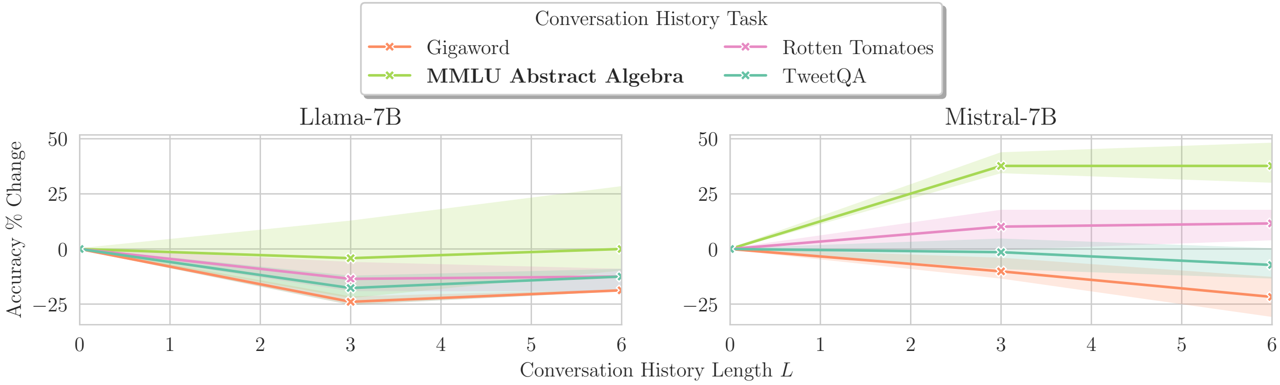


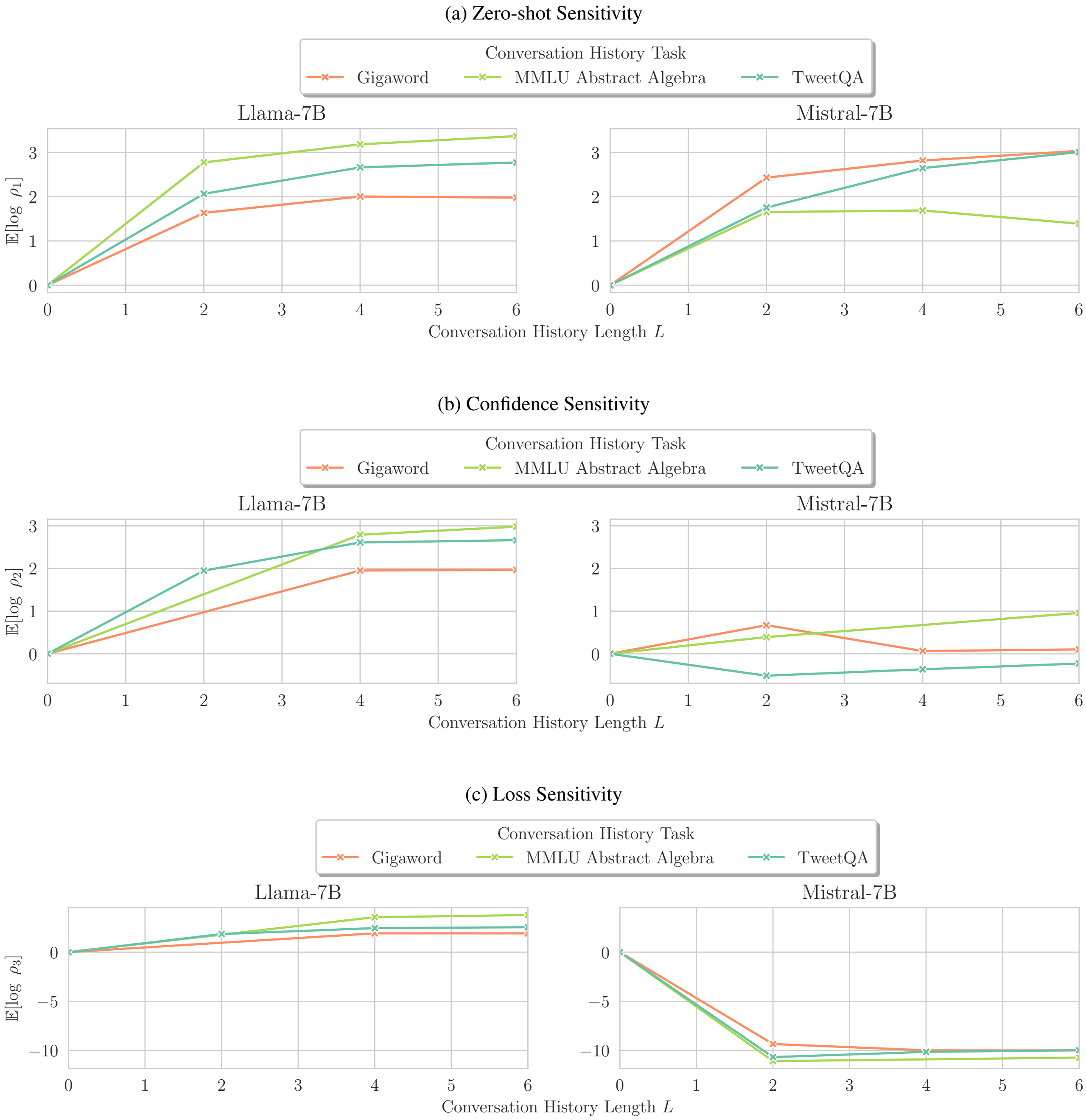
Figure 12. Target Task: MMLU Abstract Algebra. Percentage % change in accuracy relative to zero-shot performance for increasing conversation history length L for multiple seeds. Mean is shown in solid line, and the shaded region is bounded by the min/max values.

Table 10. Prompt templates for each dataset. Note that the MMLU {Topic} can be either `Human Aging` or `Abstract Algebra`. Other {words} enclosed in curly braces are replaced by the corresponding field in the datasets.

MMLU {TOPIC}	You have a multiple choice question on {Topic}. Only one of the options is correct: A, B, C, or D. Give your answer in the following format with the tags provided: <Answer> </Answer>. Please read the following question and options and answer the question
	Question: {Question}
	(A) {A}
	(B) {B}
	(C) {C}
	(D) {D}

ROTTEN TOMATOES	Can you choose only one sentiment [‘negative’, ‘positive’] for this review. review: {Review} Return only the sentiment label without any other text. Make sure to follow the format otherwise your answer will be disqualified: <Answer> positive / negative </Answer>. Do not output neutral.
GIGAWORD	Please summarize the following article. {Article}

TWEETQA	Read the given tweet and answer the corresponding question.
	tweet: {Tweet}
	question: {Question}


(a) Zero-shot Sensitivity

(b) Confidence Sensitivity

(c) Loss Sensitivity

Figure 13. Empirical investigation of various sensitivity metrics on the target task Rotten Tomatoes as a function of the conversation history length L for Llama-7b and Mistral-7b. Note that we omit the line for the in-context dataset as this is not relevant to the investigation.

D. Task-Switch Sensitivity Metrics

In Section 3, we introduced and formalized evaluation of a model's sensitivity to task-switch, namely the task sensitivity τ. This metric aims to capture the vulnerability of a model prompt to its chat history after a task-switch. Formally, it compares the zero-shot prediction $r^*|u, \mathbf{h} = \text{Ø}$ to the probability of the model outputting the same zero-shot response after a task switch $P(r^*|u, \mathbf{h} \neq \text{Ø})$. In this section, we compare the theoretical and empirical implications of different task switch sensitivity metrics.

Formally, given a conversation history $\mathbf{h}$ of length L and the next user prompt u, the probability of a model's response r_{L+1} is given by $P_\theta(r_{L+1} \mid u, \mathbf{h})$. We consider the probability of three possible responses:

1. r^*: zero-shot response
2. r_{L+1}: model's actual response
3. $\hat{r}_{L+1}$: reference response

We posit that after a task-switch, a robust model's likelihood of the zero-shot response remains high. Naturally, this gives us the formulation for the aforementioned sensitivity metric

$$\rho_1 = \frac{P_\theta(r^*|u)}{P_\theta(r^*|u, \mathbf{h})}, \tag{6}$$

which we call *zero-shot sensitivity*.

Additionally, after a task-switch, we posit that a robust model's likelihood of the actual response should be similar to that of the zero-shot response, because the irrelevant history should be largely ignored. This gives us

$$\rho_2 = \frac{P_\theta(r^*|u)}{P_\theta(r_{L+1}|u, \mathbf{h})}, \tag{7}$$

which we call the *confidence sensitivity*.

Lastly, we posit that if a model is well aligned to a task, then both the zero-shot and model's actual response should be close to the reference response:

$$\rho_3 = \frac{P_\theta(\hat{r}_{L+1}|u)}{P_\theta(\hat{r}_{L+1}|u, \mathbf{h})}, \tag{8}$$

where each probability is essentially a measure of the loss, hence we label this as the *loss sensitivity*.

The above are sensitivity per example, which we can use to estimate the task-switch sensitivity $\tau_i = \mathbb{E}[\log \ \rho_i]$ as per Equation 3, where the expectation is calculated over the examples and histories (for a given length L). We evaluate these metrics on the target task RT (rotten tomatoes) as shown in Figure 13. Figure 13a shows that the zero-shot sensitivity metric trends upwards for both models. This is expected for a model which does not handle task-switch well as the probability of the output with an increased conversation length decreases in comparison to the zero-shot probability. For the confidence sensitivity in Figure 13b, we observe that Mistral-7B behaves as we expect, whereas Llama-7B becomes less confident in its output compared to having no conversation history. For the loss sensitivity metric in Figure 13c, we observe that Llama behaves as we expect as the sensitivity remains relatively flat: as the conversation history increases, there is no significant change in the probability of outputting the reference. However, for Mistral-7b, the probability falls immediately and plateaus showing that the model was giving a very low probability mass to the reference with no conversation history. Intuitively, it is clear that both models agree in their trends only for the zero-shot sensitivity τ_1 in Figure 13a, hence in the main paper, we report zero-shot sensitivity as the task-switch sensitivity.

E. Correlations Models, Datasets and Performance

We rank model performance against various metrics to see if there is any correlation that may help explain model performance more generally.

E.1. Task Tokens

Table 11. Target Task: **MMLU AA**.

CH Task	Length	Llama-7B	Mistral-7B
Gigaword	75	-21.35	-15.94
TweetQA	93	-15.10	-4.35
RT	108	-13.02	10.87
MMLU AA	143	-1.79	37.68

Table 12. Target Task: **RT**

CH Task	Length	Llama-7B	Mistral-7B
Gigaword	76	1.98	-0.72
TweetQA	93	2.70	-1.28
RT	108	2.38	2.83
MMLU AA	143	-5.42	1.19

We compare the model performance against the mean conversation history task, T_h length. The length is measured as the number of tokens in the model, and the mean is taken over the whole dataset. The model performance is taken for three different seeds with conversation history lengths $L \in \{3, 6\}$.

E.2. Task Distance

In this section we aim to assess the hypothesis that the 'distance' between tasks can explain the extent of performance degradation in different task-switches, from the conversation history task, T_h to the target task, T_t. Measuring distance between tasks is a multi-faceted and complex metric. Given the lack of formal task distance measures, we instead use a consensus ranking approach, where multiple powerful Large Language Models (LLMs) are required to rank the different tasks on how similar they are. For the target task *RT*, we queried four of the largest and most powerful models to rank the closest tasks, based on the description of each task. We consider the following LLMs: ChatGPT; Gemini Ultra (Team et al., 2024), Claude 3 Sonnet from Anthropic; and Perplexity AI. The rankings by the LLMs are given in Table 13 relative to RT. We then select an overall ranking with the greatest consensus - in this case three of the four LLMs agree perfectly in the ranking. This gives a consensus vote of ranks (relative to RT): RT (1); MMLU AA (3); TweetQA (2); and Gigaword (3). The equivalent ranks are given in Table 14 with MMLU AA as the reference task. In this case, three of the four models perfectly agree in their rankings.

Table 13. Rank given by LLM for different datasets on how similar they are to the target task RT.

Dataset	ChatGPT	Gemini	Claude	Perplexity
RT	1	1	1	1
MMLU AA	4	4	4	4
TweetQA	2	3	2	2
Gigaword	3	2	3	3

Table 14. Rank given by LLM for different datasets on how similar they are to the target task MMLU AA.

Dataset	ChatGPT	Gemini	Claude	Perplexity
RT	4	4	4	4
MMLU AA	1	1	1	1
TweetQA	2	3	2	2
Gigaword	3	2	3	3

The following tables compare the rank of the dataset distance against the mean model performance. The model performance is the %-percentage accuracy change relative to zero-shot, and the mean is taken over three seeds and over conversation history lengths $L \in \{3, 6\}$.

Table 15. Target Task, T_t: **RT**. Performance degradation (with different conversation history tasks) compared to the task rank, measuring similarity to T_t.

CH-Task	Rank	Llama-7B	Mistral-7B
RT	1	2.38	2.83
TweetQA	2	2.70	-1.28
Gigaword	3	1.98	-0.72
MMLU AA	4	-5.42	1.19

Table 16. Target Task: **MMLU AA**. Performance degradation (with different conversation history tasks) compared to the task rank, measuring similarity to T_t.

CH-Task	Rank	Llama-7B	Mistral-7B
MMLU AA	1	-1.79	37.68
TweetQA	2	-15.10	-4.35
Gigaword	3	-21.35	-15.94
RT	4	-13.02	10.87

Overall, there appears to be only a weak correlation in some settings between the task distance and the performance degradation. This suggests that performance degradation is not only a function of the task distance, but is also an attribute of the specific model. Further analysis would be required to understand the aspects of specific models for certain task-switches that influence the level of performance degradation.

F. Performance Confusion Matrix

In this section, we summarize the performance change for every pairing of task-switches from conversation history task (T_h) to target task (T_t). We present the results here for a conversation length of $L = 6$ for each model separately. Tables 17, 18, 19, 20 report the results for models GPT-3.5, GPT-4, Llama-7B, Mistral-7B respectively. Each *row* is the performance change in the Target Task T_t. Please note that the metric for the tasks are: accuracy for MMLU AA, RT, MMLU HA, METEOR for TweetQA, and RougeL for Gigaword.

Table 17. Model: **GPT-3.5**. Percentage % change in model performance.

	Conversation History Task				
Target Task	AA	RT	TQ	GW	HA
MMLU AA	19.35	6.45	6.45	-3.13	
RT	-0.22	3.00	-0.33	0.11	
Tweet QA	-13.78	-3.55	24.81	-5.69	
Gigaword	-12.10	-6.59	-3.48	67.85	
MMLU HA		4.73	-12.84	-8.11	20.41

Table 18. Model: **GPT-4**. Percentage % change in model performance.

	Conversation History Task				
Target Task	AA	RT	TQ	GW	HA
MMLU AA	8.62	-13.11	-3.39	0.00	
RT	0.76	1.74	-0.98	-0.98	
Tweet QA	3.69	25.58	35.80	5.06	
Gigaword	12.52		14.18	59.07	
MMLU HA		0.53	1.59	2.14	5.85

Table 19. Model: **Llama-7B**. Percentage % change in model performance.

	Conversation History Task				
Target Task	AA	RT	TQ	GW	HA
MMLU AA	3.57	-15.63	0.00	-10.71	
RT	-5.69	1.82	3.76	1.82	
Tweet QA	1.68	13.37	50.74	10.17	
Gigaword	11.76		11.73	158.79	
MMLU HA		13.86	-3.96	-0.99	25.74

Table 20. Model: **Mistral-7B**. Percentage % change in model performance.

	Conversation History Task				
Target Task	AA	RT	TQ	GW	HA
MMLU AA	28.57	18.18	-4.76	-19.05	
RT	0.97	3.79	-0.87	-1.30	
Tweet QA	-0.78	7.62	30.56	9.26	
Gigaword	-3.81	2.61	3.44	78.71	
MMLU HA		1.63	0.81	-11.38	12.20