LLM Task Interference: An Initial Study on the Impact of Task-Switch in Conversational History

Anonymous ACL submission

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.[1]

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 and 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).

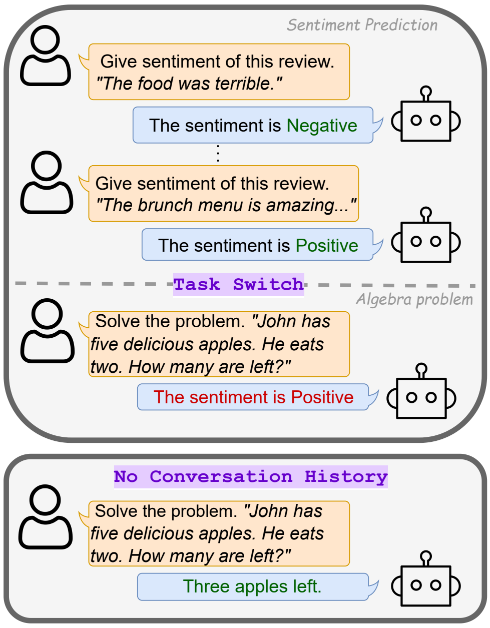


Figure 1: An illustrative example where the chat history is based on sentiment prediction. Algebra word problem introduces *task-switch* which results in an incorrect prediction.

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

[1] Code attached anonymously with the submission.

the door to vulnerabilities, where malicious actors can exploit prompt 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 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 and 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 and Wang, 2023; Chang and Jia, 2023). Chang and 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 and 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:

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

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,

$$r_{L+1} \perp \mathbf{h} | u_{L+1} \qquad \mathbf{h} \in T_h,\ \ u_{L+1} \in T_t. \quad (2)$$

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.[2]

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

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

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

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 and Lee (2005)); and human-aging subset from the MMLU dataset (MMLU HA) in the Appendix.

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

Data	Task
Gigaword	Summarization
MMLU AA	Math Multiple Choice Question
TweetQA	Social Question Answer
RT	Sentiment classification
MMLU HA	Social Multiple Choice Question

Table 1: Datasets Summary.

Models. We explore the task-switch sensitivity of four popular models. We consider two open-source small models, Llama-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 B.

4.2 Results

Task-switch. 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 2 and Table 3 showcases the impact of conversational task-switch with MMLU AA and Rotten Tomatoes as the target tasks, T_t respectively[3]. 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 2 and Table 3, 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

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

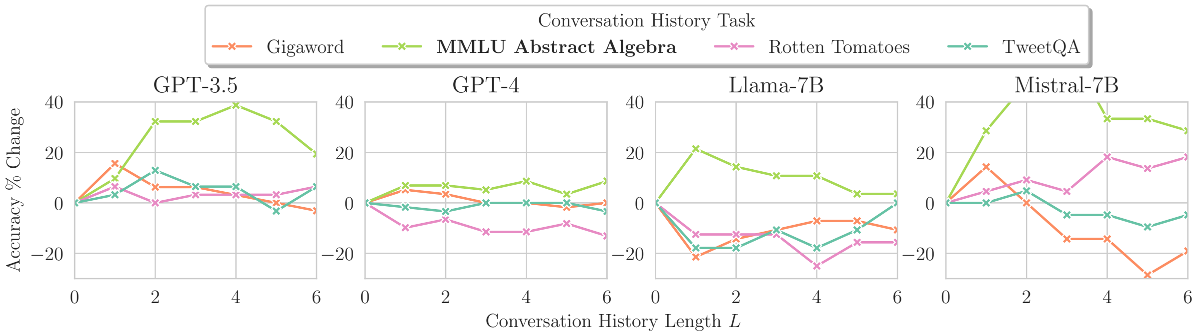


Figure 2: Target Task: MMLU Abstract Algebra. % change in accuracy relative to zero-shot performance.

CH-Task	Model	% Change	$\tau(\cdot)$
MMLU AA	GPT-3.5	*19.35*	$*$
	GPT-4	*8.62*	$*$
	Llama-7B	*3.57*	*31.51*
	Mistral-7B	*28.57*	*1.12*
Gigaword	GPT-3.5	−3.13	$*$
	GPT-4	0.00	$*$
	Llama-7B	−10.71	5.23
	Mistral-7B	−19.05	3.13
Rotten Tomatoes	GPT-3.5	6.45	$*$
	GPT-4	−13.11	$*$
	Llama-7B	−15.63	9.91
	Mistral-7B	18.18	0.83
TweetQA	GPT-3.5	6.45	$*$
	GPT-4	−3.39	$*$
	Llama-7B	0.00	6.37
	Mistral-7B	−4.76	2.78

Table 2: 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	$\tau(\cdot)$
Rotten Tomatoes	GPT-3.5	*3.00*	$*$
	GPT-4	*1.74*	$*$
	Llama-7B	*1.82*	*4.02*
	Mistral-7B	*3.79*	*2.65*
Gigaword	GPT-3.5	0.11	$*$
	GPT-4	−0.98	$*$
	Llama-7B	1.82	1.98
	Mistral-7B	−1.30	3.04
MMLU AA	GPT-3.5	−0.22	$*$
	GPT-4	0.76	$*$
	Llama-7B	−5.69	3.37
	Mistral-7B	0.97	1.39
TweetQA	GPT-3.5	−0.33	$*$
	GPT-4	−0.98	$*$
	Llama-7B	3.76	2.77
	Mistral-7B	−0.87	3.01

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

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 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 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.

7 Ethics and Risks

All of the datasets used are publicly available. Our implementation utilizes the PyTorch 1.12 framework, an open-source library. We obtained a license from Meta to employ the Llama-7B model via HuggingFace. Additionally, our research is conducted per the licensing agreements of the Mistral-7B, GPT-3.5, and GPT-4 models. We ran our experiments on A100 Nvidia GPU and via OpenAI API.

Our work may be built upon to identify vulnerabilities of LLMs. Overall, there are no ethical concerns with this work.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2022. What learning algorithm is in-context learning? investigations with linear models. *arXiv*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv*.

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv*.

Stephanie CY Chan, Ishita Dasgupta, Junkyung Kim, Dharshan Kumaran, Andrew K Lampinen, and Felix Hill. 2022. Transformers generalize differently from information stored in context vs in weights. *arXiv preprint arXiv:2210.05675*.

Ting-Yun Chang and Robin Jia. 2023. 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)*, pages 8123–8144.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv*.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. 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*, pages 79–90.

Michael Hahn and Navin Goyal. 2023. A theory of emergent in-context learning as implicit structure induction. *arXiv*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv*.

Shuyu Jiang, Xingshu Chen, and Rui Tang. 2023b. Prompt packer: Deceiving llms through compositional instruction with hidden attacks. *arXiv preprint arXiv:2310.10077*.

Harrison Lee, Raghav Gupta, Abhinav Rastogi, Yuan Cao, Bin Zhang, and Yonghui Wu. 2022. Sgd-x: A benchmark for robust generalization in schema-guided dialogue systems. In *AAAI*, volume 36, pages 10938–10946.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2023. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

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

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

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

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

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

Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. 2022. What spurious features can pretrained language models combat?

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv*.

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

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. *ACL Findings*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv*.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.

Jason Weston and Sainbayar Sukhbaatar. 2023. System 2 attention (is something you might need too). *arXiv*.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. 2021. Tiage: A benchmark for topic-shift aware dialog modeling. *ACL*.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Tweetqa: A social media focused question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. 2023. An llm can fool itself: A prompt-based adversarial attack. *arXiv preprint arXiv:2310.13345*.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021. Topic-aware multi-turn dialogue modeling. In *AAAI*, volume 35, pages 14176–14184.

Chenxu Yang, Zheng Lin, Jiangnan Li, Fandong Meng, Weiping Wang, Lanrui Wang, and Jie Zhou. 2022. Take: topic-shift aware knowledge selection for dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 253–265.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Appendix

The appendix is structured as follows: Appendix A gives more details about the datasets, Appendix B reports the zero-shot absolute performance of all models on all tasks, Appendix C presents an ablation study on the conversation history length, Appendix D discusses the prompt templates used for each dataset, and Appendix E discusses other definitions for task-switch sensitivity.

A Datasets and Metrics Summary

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

Table 4: Dataset Summary. QA: Question-Answering. MCQ: Multiple Choice Question

In Section 4.2 of the main paper, we present results evaluated on two different datasets: MMLU Abstract Algebra (MMLU AA) multiple choice questions and Rotten Tomatoes (RT) sentiment classification. In Appendix B, C, we present results evaluated on all of the datasets covering a range of tasks: MMLU Human Aging (MMLU HA) multiple choice questions, Gigaword for summarization, and TweetQA question-answering. The train-test splits of these datasets are shown in Table 4. 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 D.

For classification tasks performance is measured using accuracy, whilst for generative tasks it is measured using ROUGE (Lin, 2004) or METEOR (Banerjee and Lavie, 2005).

B 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 D. 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.

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

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

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

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

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

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

C 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 4 we select four datasets (including itself), from which we use the training set to use as conversation history. The details of the prompt structure are presented in Appendix D.

C.1 Performance change with Task-switch

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. Intuitively, we would expect that when there is *not* a task switch, the performance would increase (assuming the training examples are aligned well with 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.

C.2 Format Failure Rate

Typically, classification tasks (MMLU HA, MMLU AA, RT) are evaluated using logits, however we chose to 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"`, however 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.

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.

D 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 C.2. Furthermore, we note that the standard evaluation method used in the Open-LLM leaderboard code (available on GitHub) is to 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).

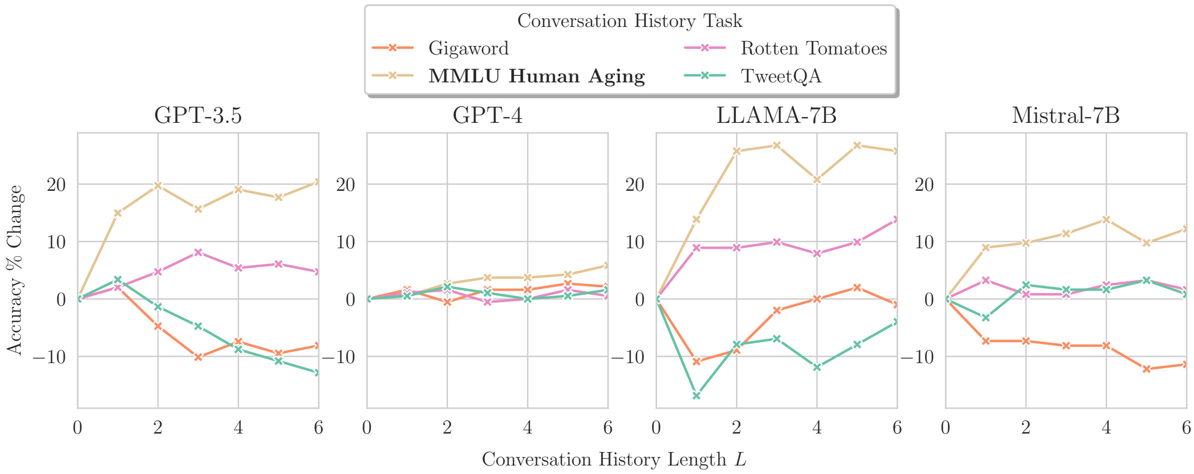


Figure 3: Target Task: MMLU HA. Percentage % change in accuracy relative to zero-shot performance (no conversation history) for increasing conversation history length L and various models.

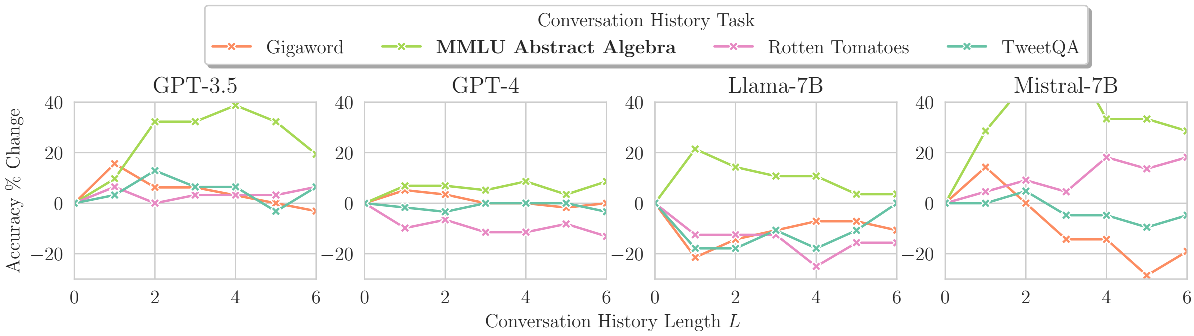


Figure 4: Target Task: MMLU AA. Percentage % change in accuracy relative to zero-shot performance (no conversation history) for increasing conversation history length L and various models.

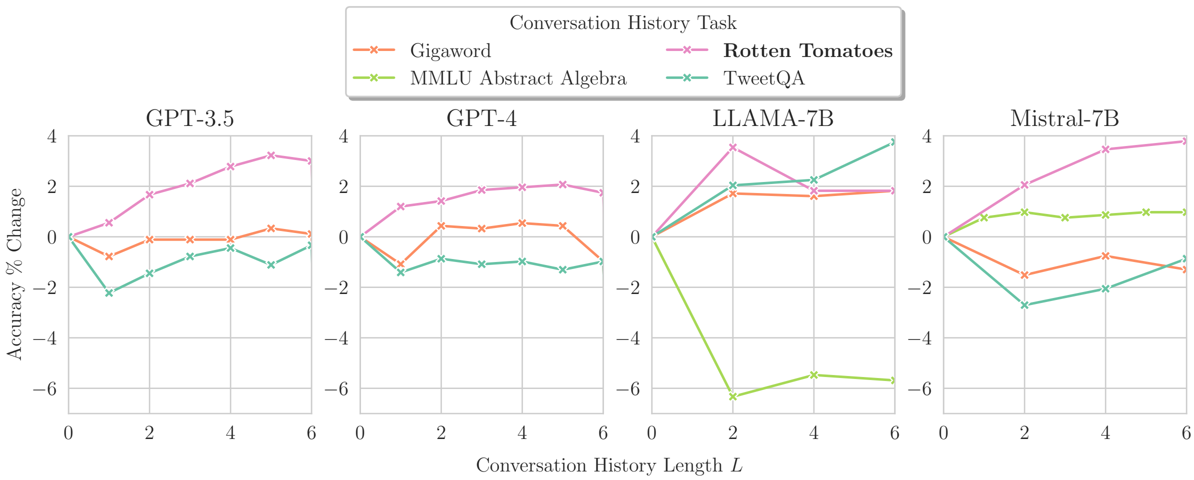


Figure 5: Target Task: RT. Percentage % change in accuracy relative to zero-shot performance (no conversation history) for increasing conversation history length L and various models.

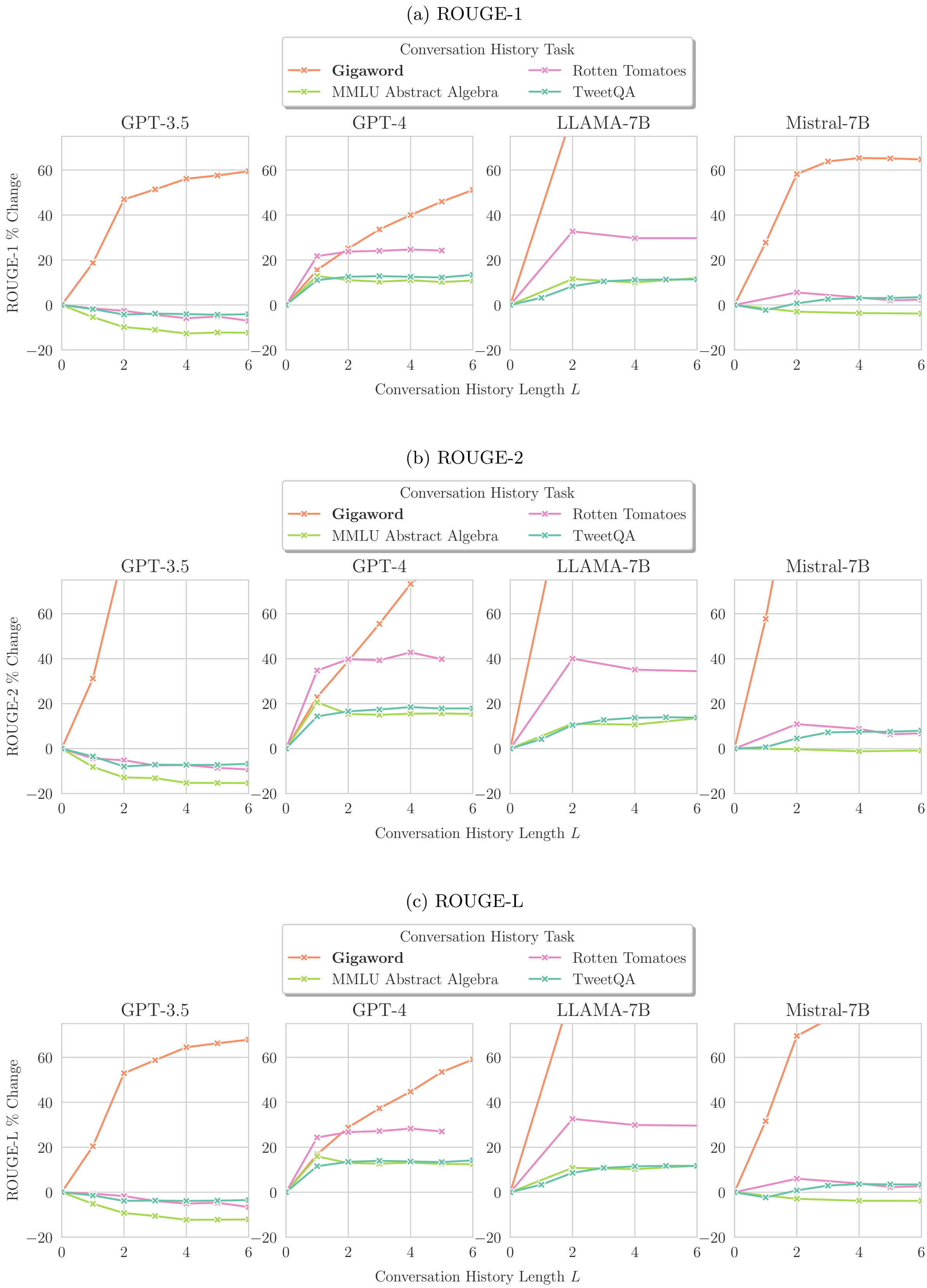
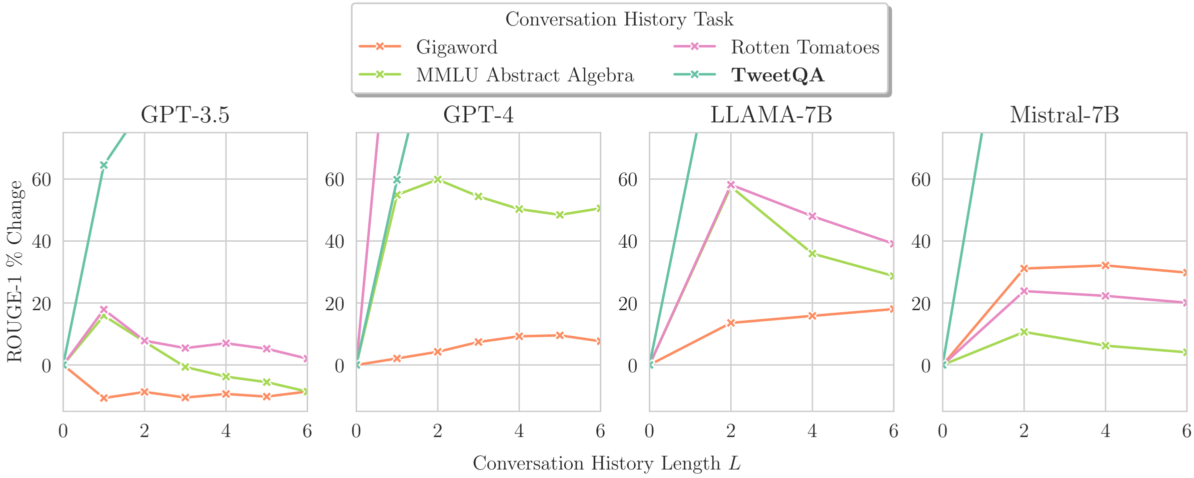
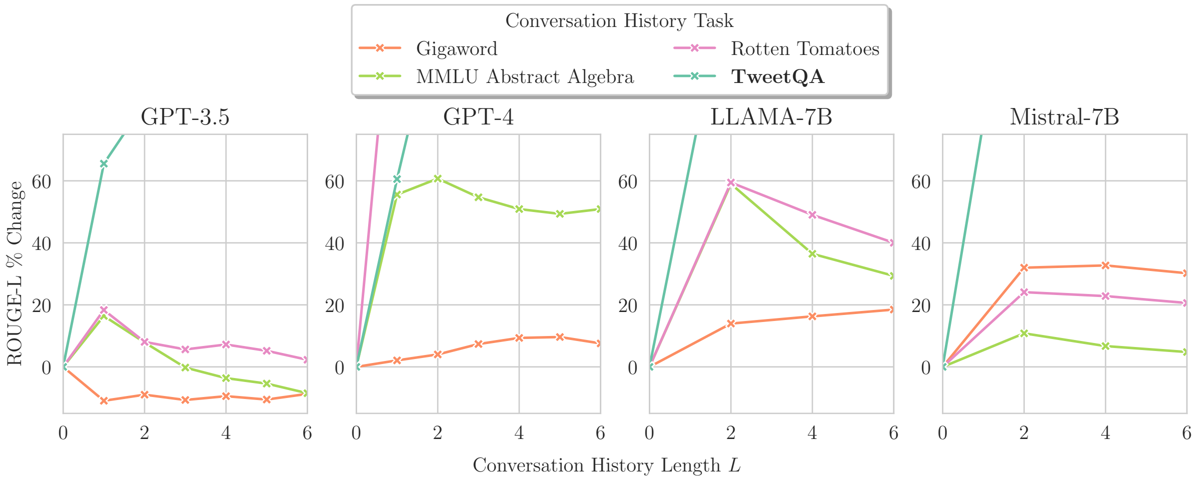


Figure 6: Target Task: Gigaword. Percentage % change in accuracy relative to zero-shot performance (no conversation history) for increasing conversation history length L and various models. Note that we focus on the effect of task-switching by clipping the y-axes at +75%.

(a) ROUGE-1



(b) ROUGE-L



(c) METEOR

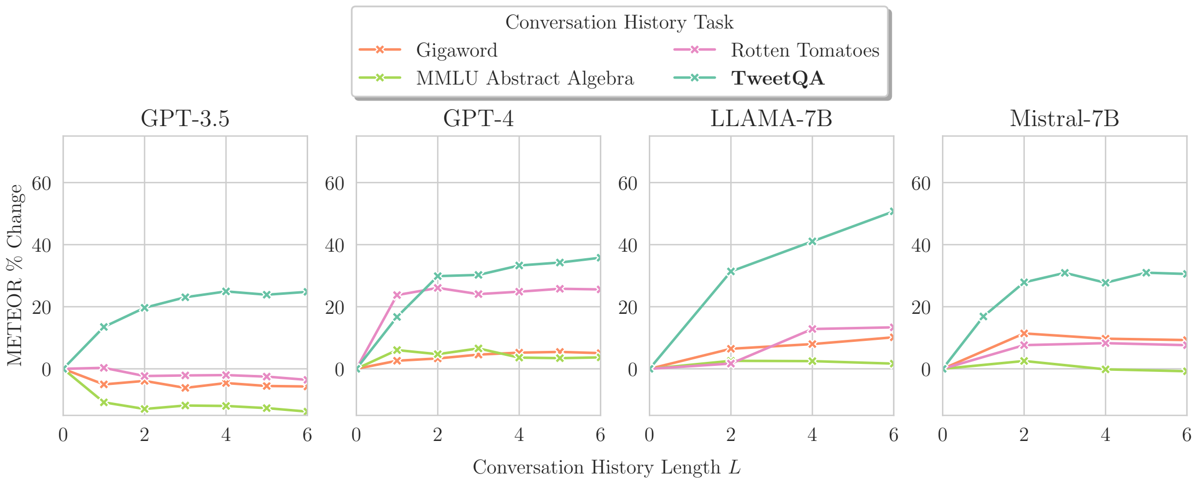


Figure 7: Target Task: TweetQA. Percentage % change in accuracy relative to zero-shot performance (no conversation history) for increasing conversation history length L and various models. Note that we focus on the effect of task-switching by clipping the y-axes at +75%.

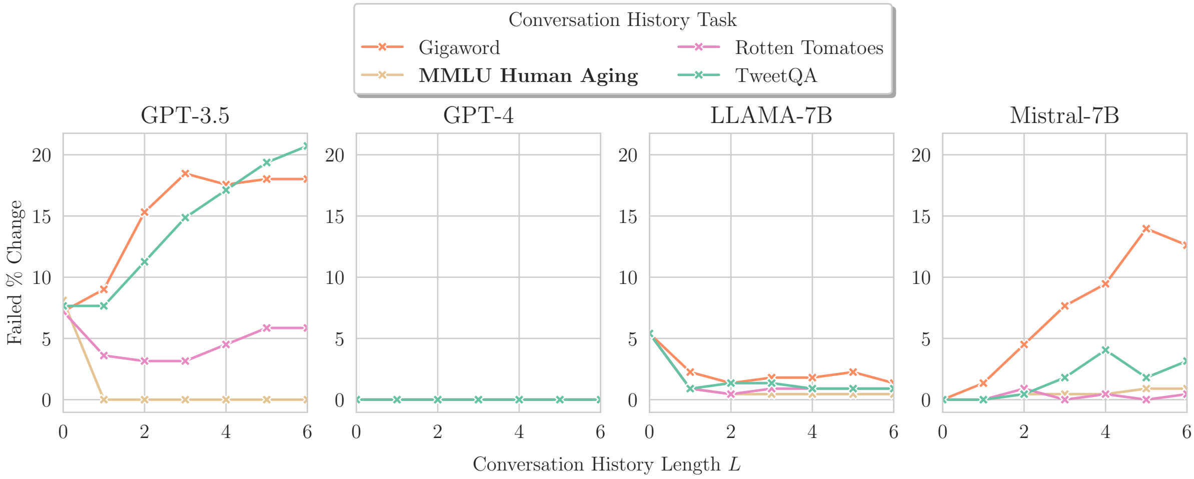


Figure 8: Target Task: MMLU Human Aging. Percentage % of examples where format failed.

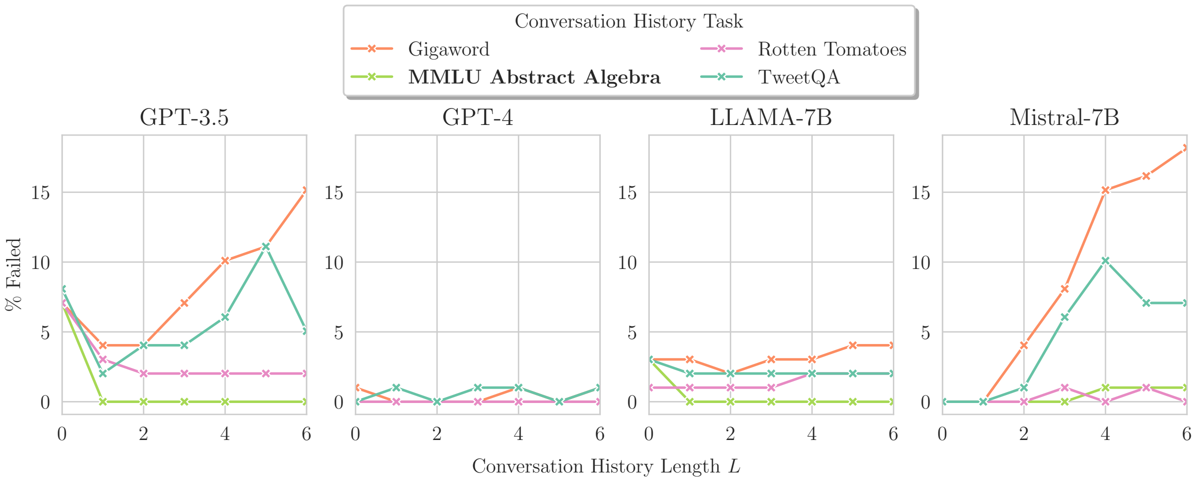


Figure 9: Target Task: MMLU Abstract Algebra. Percentage % of examples where format failed.

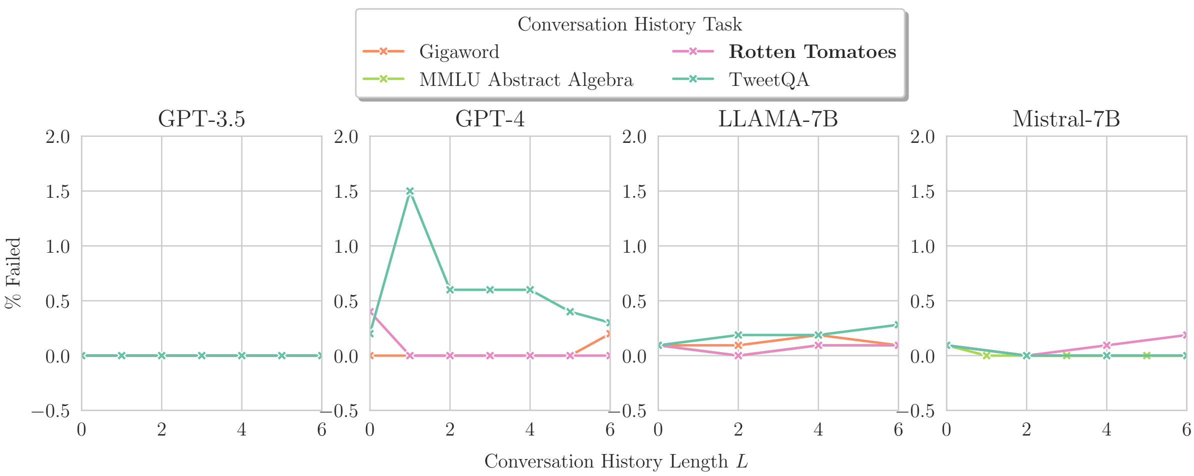


Figure 10: Target Task: Rotten Tomatoes. Percentage % of examples where format failed.

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}`

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.

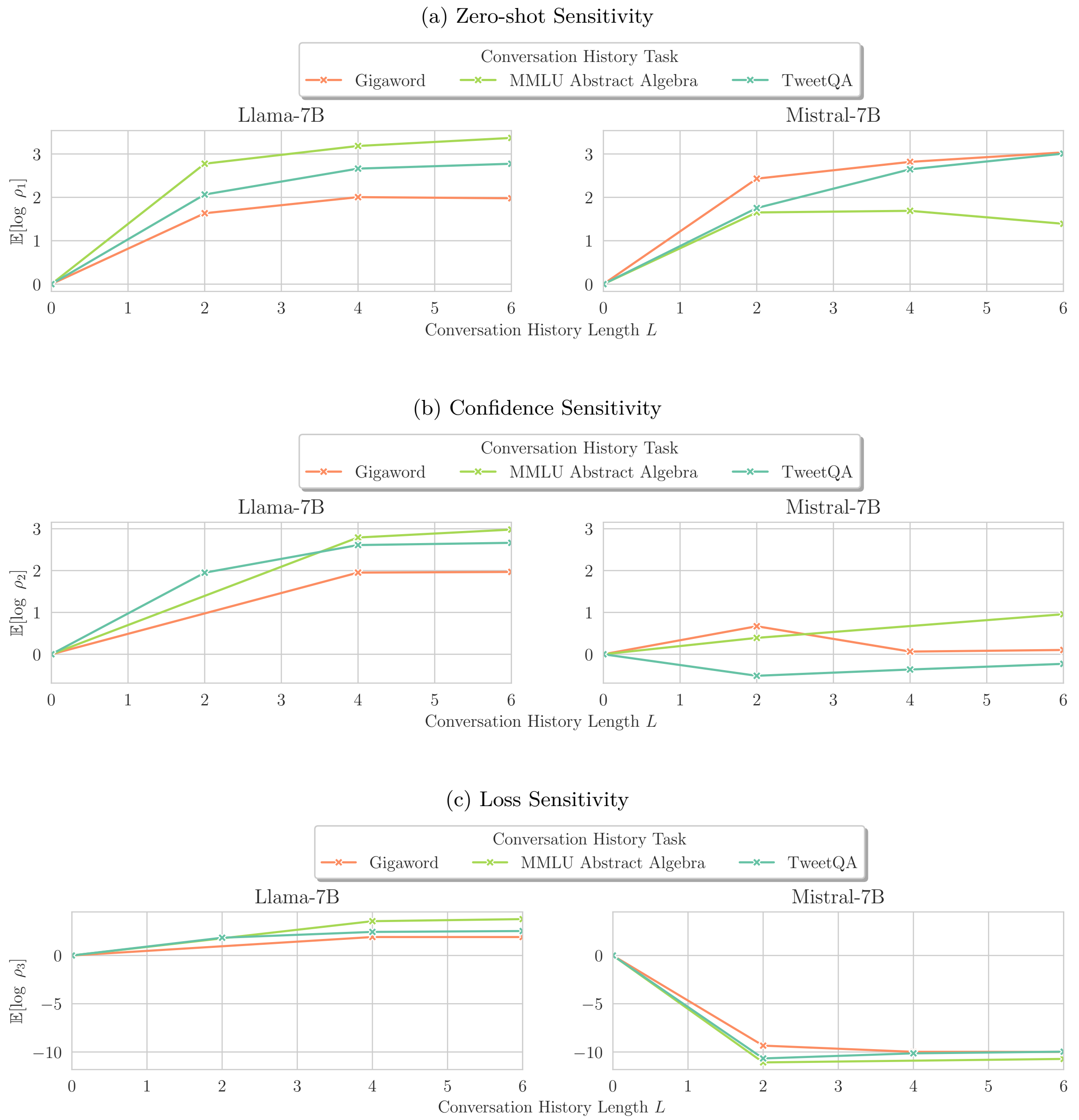


Figure 11: Empirical investigation of various sensitivity metrics on the target task Rotten Tomatoes as a function of the conversation history length L for Llama-7b and Mistral-7b. Note that we omit the line for the in-context dataset as this is not relevant to the investigation.

E 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 $r^*|u, \mathbf{h} = \emptyset$ to the probability of the model outputting the same zero-shot response after a task switch $P(r^*|u, \mathbf{h} \neq \emptyset)$. 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 11. Figure 11a 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 11b, 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 11c, 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 11a, hence in the main paper, we report zero-shot sensitivity as the task-switch sensitivity.