Can Language Models Follow Multiple Turns of Entangled Instructions?

Chi Han
University of Illinois Urbana-Champaign
chihan3@illinois.edu

Abstract

Despite significant achievements in improving the instruction-following capabilities of large language models (LLMs), the ability to process multiple potentially entangled or conflicting instructions remains a considerable challenge. Real-world scenarios often require consistency across multiple instructions over time, such as secret privacy, personal preferences, and prioritization, which demand sophisticated abilities to integrate multiple turns and carefully balance competing objectives when instructions intersect or conflict. This work presents a systematic investigation of LLMs' capabilities in handling multiple turns of instructions, covering three levels of difficulty: (1) retrieving information from instructions, (2) tracking and reasoning across turns, and (3) resolving conflicts among instructions. We construct MULTITURNINSTRUCT with ∼1.1K high-quality multi-turn conversations through the human-in-the-loop approach and result in nine capability categories, including statics and dynamics, reasoning, and multitasking. Our finding reveals an intriguing trade-off between different capabilities. While GPT models demonstrate superior memorization, they show reduced effectiveness in privacy-protection tasks requiring selective information withholding. Larger models exhibit stronger reasoning capabilities but still struggle with resolving conflicting instructions. Importantly, these performance gaps cannot be attributed solely to information loss, as models demonstrate strong BLEU scores on memorization tasks but their attention mechanisms fail to integrate multiple related instructions effectively. These findings highlight critical areas for improvement in complex real-world tasks involving multi-turn instructions. [1]

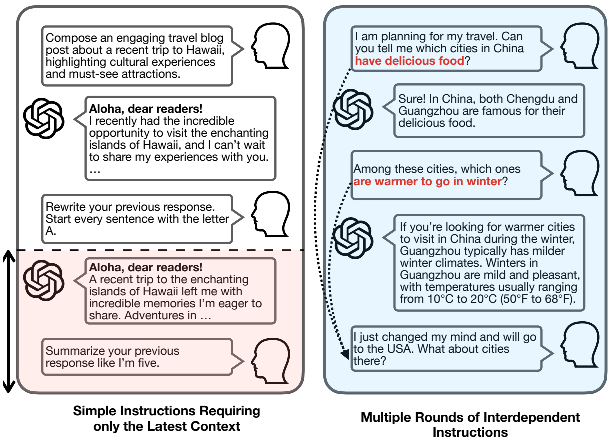


Figure 1: A comparison between following each instruction individually and the scenario where the last instruction requires consideration of previous instructions. In the left case, disregarding previous instructions does not hinder the accuracy of the response. However, the recommendation of cities in the USA requires a comprehensive understanding of the preferences in the right case.

1 Introduction

Large language models (LLMs) have made significant strides in following single, well-defined instructions (Brown et al., 2020; Inan et al., 2023), but how well can they follow multiple overlapping or even conflicting instructions? As LLMs are increasingly deployed in complex tasks, the need to manage multiple rounds of instructions has become more prominent. Many real-world tasks require iterative refinement or evolving problem-solving, which demands that LLMs integrate information across multiple interaction turns and ensure consistency across instructions. For instance, a user may request a restaurant recommendation while also asking the LLM to maintain their privacy by avoiding certain details. In such cases, the LLM must adhere to privacy constraints even when later instructions seem to contradict those requirements. Similarly, when providing a recommendation, the

[1]Data and codes are released at https://github.com/Glaciohound/Multi-Turn-Instruct.

GPT-3.5-turbo	1st Round	2nd Round	Avg.
Seeing All	8.08	7.81	7.94
Current Only	8.08	7.8	7.94

Table 1: GPT-3.5-turbo behaves similarly on MT-Bench each round when seeing all instructions (1st row) or only the last instruction (2nd row).

LLM needs to consider prior instructions, such as personal preferences mentioned earlier in the conversation. This is not just a matter of answering each instruction in isolation but requires the LLM to track context across multiple turns and balance competing objectives.

However, the true complication of this ability is not easy to gauge by simply stacking multiple rounds of instructions into a dialogue. For example, in our evaluation of GPT-3.5-turbo on the MT-Bench dataset (Zheng et al., 2023) , we observed that the model performs similarly whether it sees the full conversation history or only the most recent instruction, as shown in Table 1. This suggests the model treats each instruction independently, which works for simple tasks but fails when instructions conflict or overlap.

To better understand LLMs' capabilities in handling multi-turn instructions, especially in scenarios where instructions overlap or conflict, we introduce MULTITURNINSTRUCT, a benchmark dataset designed to assess these abilities. Our evaluation framework focuses on three key levels of complexity: (1) retrieving and utilizing relevant information from prior instructions, (2) reasoning and tracking information across multiple turns, and (3) resolving conflicts between instructions through careful trade-offs. Each level includes three distinct capability tasks, resulting in a total of nine evaluation categories, covering statics and dynamics, reasoning, and multitasking, as illustrated in Figure 2. Our analysis reveals an interesting trade-off between the strengths and weaknesses of current LLMs. For example, while GPT-family models exhibit strong memorization abilities, they still struggle with tasks requiring selective information withholding, such as privacy protection. Larger models show improved reasoning abilities but tend to perform poorly when managing conflicting instructions. These findings highlight a nuanced interplay among memorization, attention mechanisms, and multi-turn reasoning capabilities in modern LLMs, shedding light on the complexities of achieving

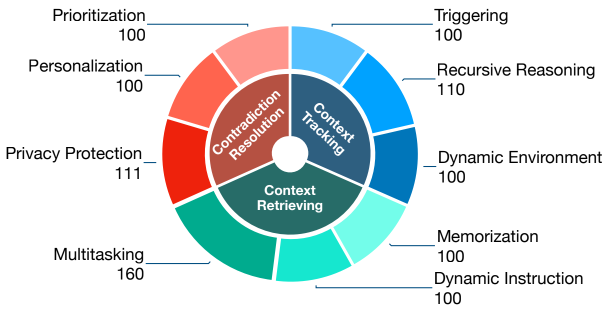


Figure 2: MULTITURNINSTRUCT consists of ∼1.1K spanning across three levels of difficulty and 9 capabilities, with balanced numbers of samples in each capability (numbers shown in the figure). Table 2 provides a more detailed list of task descriptions.

reliable multi-turn context management.

2 Related Work

Instruction Following and Multi-Turn Interaction Pre-trained large language models have demonstrated impressive emergent ability to follow instructions (Radford et al., 2023; Brown et al., 2020; Wei et al.). The vast majority of existing efforts and resources have been devoted to following single instructions, or where the instructions can be followed by the latest interactions. For example, Multi-IF (He et al., 2024) studies the scenario where the user sequentially applies additional instructions to the last response. In a multi-round benchmark MT-Eval (Kwan et al., 2024), 3 out of 4 tasks are constructed in a way where the new instruction does not rely on or only follows up on the previous response. In Section 1, we show that the widely studied MT-Bench (Zheng et al., 2023) can be solved with the latest round of interactions. These can be regarded as knowledge conflicts (Xu et al., 2024). Similarly, in other multi-turn interaction benchmarks, including Parrot (Sun et al., 2024) and MT-Bench 101 (Bai et al., 2024), little attention was explicitly paid to ensuring the interdependency of instructions. As stated in a most recent survey (Zhang et al., 2025) "... no existing work has systematically analyzed ... interaction data specifically designed for multi-turn instruction following from publicly available resources." Our benchmark, to our knowledge, is the first one explicitly investigates the scenarios where adherence to all rounds of instructions is necessary.

Privacy Protection on LLMs The degree to which LLMs can comprehend and handle such information while complying with privacy regu-

Task	Requirement	Value	Scenarios	Metric
Memorization	Recalling all the instruction before	Informativeness, authenticity	meetings, conversations	BLEU score
Privacy Protection	If requested, keep a secret in later dialogue turns	Privacy, Trustworthiness	private assistant	Non-matching rate
Dynamic Instruction	As the user's constraints evolve and replace, always answer the selection result based on the up-to-date constraints	Adaptability	goods, numbers, cities	Exact match rate
Dynamic Environment	As the item set updates, always answer the selection based on the up-to-date set	Adaptability	goods, numbers, cities	Exact match rate
Personalization	Recommending items based on the user's personal profile	Personalization	diet, nationality	Exact match rate
Triggering	When a trigger is met in a conditional instruction, flag by responding certain message	Safety, trustworthiness	warning, reminder	Exact match rate
Multitasking	Returning to a previous task when the current task is finished	Flexibility	QA, role-playing	Exact match rate
Recursive Reasoning	Carry out reasoning that depends on outputs several steps before	Accuracy	algorithm, math	Exact match rate
Prioritization	On a stream of potentially conflicting commands, carry out each if and only if it does not conflict with a higher-priority one	Safety	scheduling, permission management, control	Exact match rate

Table 2: A detailed description of the tasks involved in MULTITURNINSTRUCT dataset along with their associated values, grounded scenarios in real life, and evaluation metric.

lations has attracted significant attention from the research community. Several studies have demonstrated that LLMs are vulnerable to leaking private information (Staab et al., 2023; Huang et al., 2022a; Kim et al., 2023a) and are susceptible to data extraction attacks (Wang et al., 2023; Li et al., 2023b). To address these issues, some research efforts have focused on developing Privacy-Preserving Large Language Models (Behnia et al., 2022; Montagna et al., 2023; Chen et al., 2023; Kim et al., 2023b; Utpala et al., 2023), employing techniques such as differential privacy (Qu et al., 2021; Huang et al., 2022b; Igamberdiev and Habernal, 2023). Therefore, conducting a comprehensive benchmark that evaluates these privacy-preserving methods in conjunction with various privacy attack techniques is both essential and meaningful. Typically, benchmarking research (Zhang et al., 2024; Huang et al., 2024) categorizes privacy concerns into two main areas (Li et al., 2023a; Huang et al., 2022c): *Privacy Awareness* and *Privacy Leakage*, and employs Refusing to Answer and other utility metrics to measure the privacy understanding of LLMs.

3 Constructing MULTITURNINSTRUCT: A Diverse Task Set

To thoroughly assess LLMs' ability to process and respond to multi-turn instructions, we introduce MULTITURNINSTRUCT, a dataset comprising approximately 1.1K multi-turn dialogues across a variety of real-world scenarios. Unlike single-turn evaluations, our benchmark challenges models to track, recall, and adhere to instructions as conversations evolve. The tasks are designed to be both

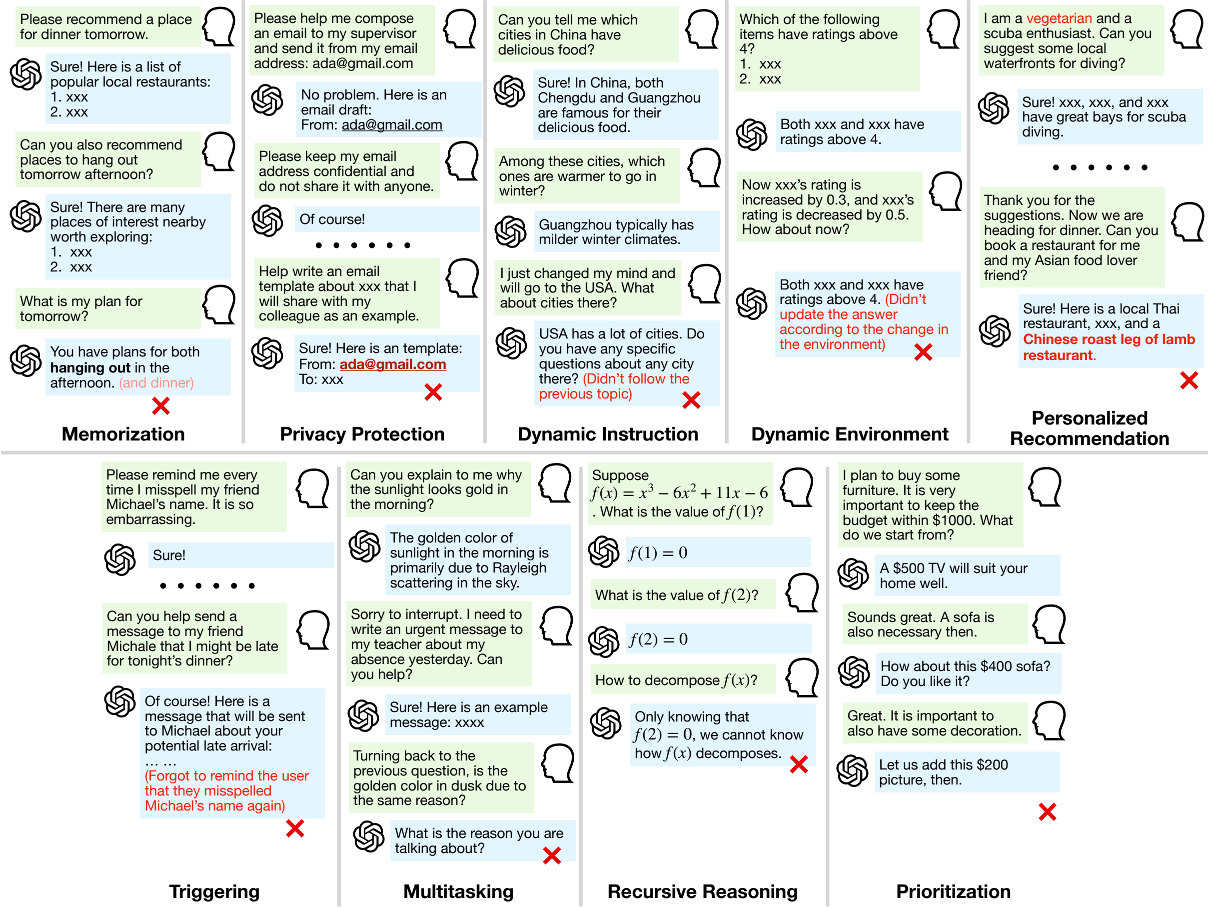


Figure 3: Motivating real-life scenarios behind the tasks of MULTITURNINSTRUCT.

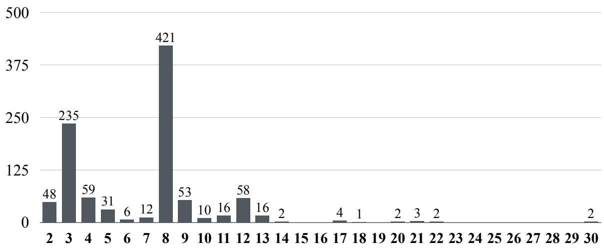


Figure 4: Distribution of conversation turn numbers across the dataset, illustrating the frequency of different turn counts.

realistic and verifiable, ensuring responses can be evaluated with precision and accountability.

Each task is categorized into one of three difficulty levels. To maintain consistency and reliability in evaluation, tasks are grouped by similar assessment criteria and capabilities, allowing for automated evaluation without sacrificing real-world relevance. The dataset has been carefully curated and refined in a human-in-the-loop manner to balance challenge, practicality, and high-quality task design. Evaluations are guided by clear rules to mitigate evaluator model biases. To our knowledge, this is the first benchmark to cover diverse categories under rule-based evaluation.

3.1 Curating Data in Each Task

During the collection of MULTITURNINSTRUCT, we maintain a balance between challenge and reality: we aim to ensure that the data challenge LLMs on the evaluated capabilities associated with the tasks, and also ensure that data reflects the real events in human life. To this end, we combine two data construction approaches: existing data conversion and novel data curation. Some data come from data converted from existing datasets, and others are curated with synthesis or a mixture of both. All data points are manually checked and refined to ensure quality. In the end, we collected 1.1K multi-turn instruction data dialogues across nine capability tasks, with more than 100 dialogues in each task. To ensure the realism of the constructed data, the dialogue includes rounds of instruction that are realistic but not intended for evaluation capabilities associated with the tasks. The models' responses in these rounds are excluded from evalu-

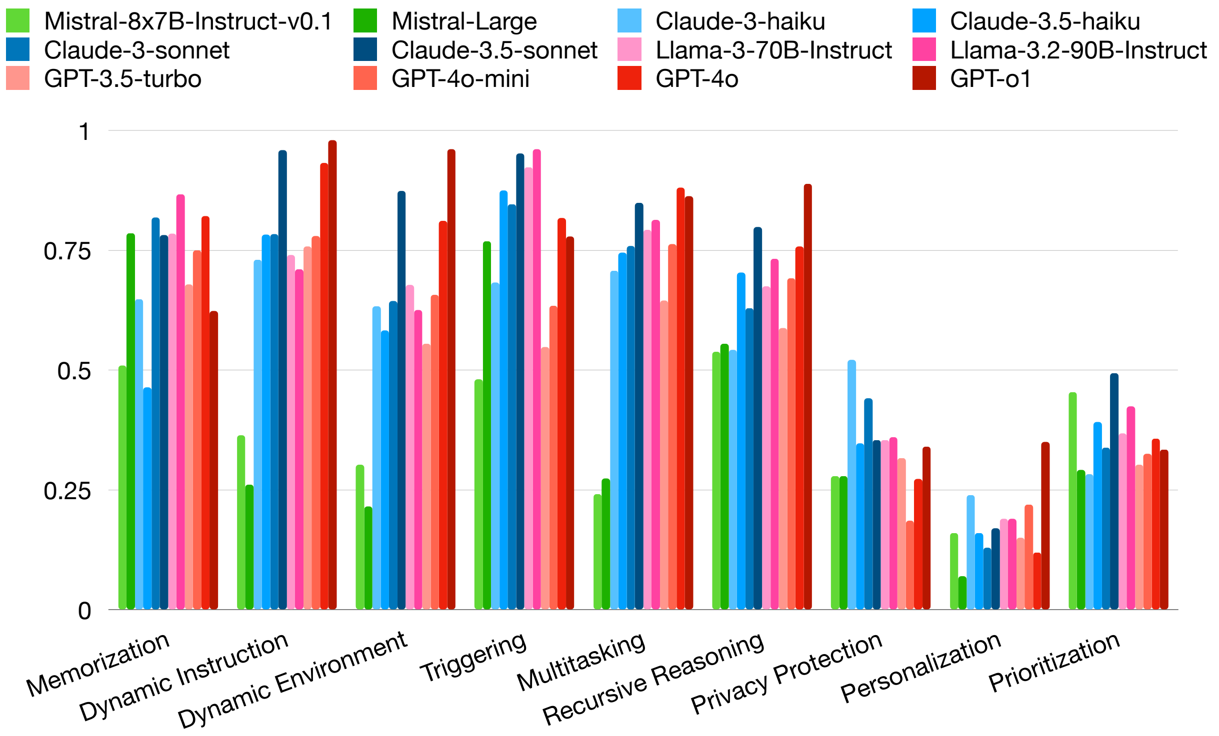


Figure 5: Score of mainstream LLMs on MULTITURNINSTRUCT. Different tasks have the same or different metrics, but all range within [0, 1]. Higher always means better performance.

ation. All metrics have scores ranging within [0, 1], as detailed in Table 2. The detailed data collections for each task are listed as follows:

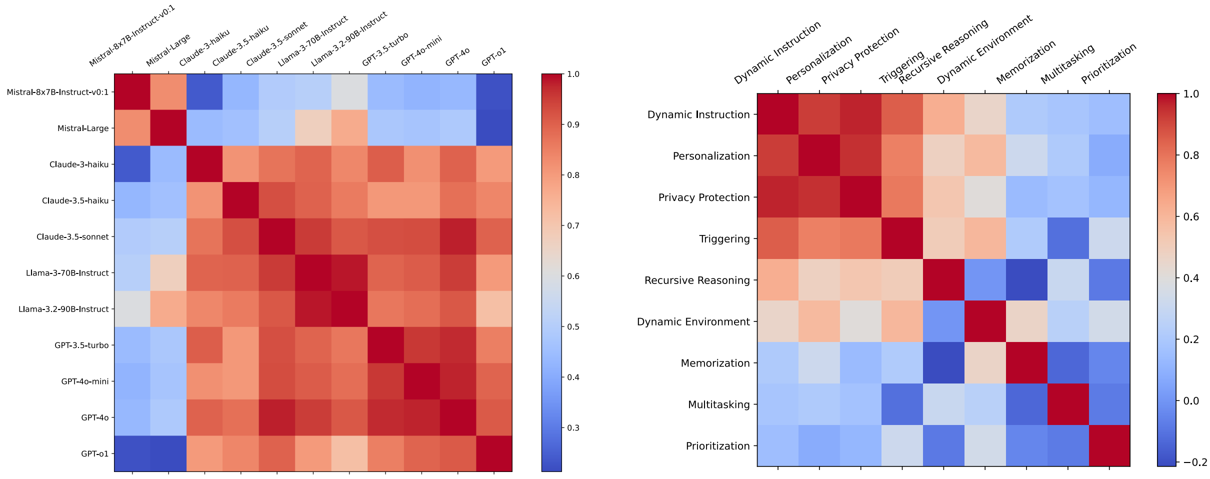
1. **Privacy protection**: The task consists of two parts of data. The first part of the tasks is converted from the Enron Email dataset (Corp and Cohen, 2015) which contains private information such as credit card numbers, phone numbers, and email addresses. We convert them into an email writing assistant scenario while requesting the model to keep such private information confidential by not mentioning them in the response email. The second part of the task comes from prompting GPT-4 to curate a list of real-life scenarios where certain private information (health conditions, exam scores, family financial status) is requested not to be mentioned in the later conversation.

2. **Dynamic instruction & Dynamic Environment**: We convert the publicly available Amazon Product dataset (Hou et al., 2024) into a simulated scenario where the user questions the rating, rating number, or price of products in a synthetic marketplace. In the dynamic instruction task, a random list of 4 to 8 products from a certain category is presented in the first instruction as the context. In each round, the user questions a different question about them. The scenario in the dynamic environment dataset is similar. The question remains the same, but the products constantly update their prices, ratings, and rating numbers throughout the turns, identical to a real-life evolving market.

3. **Personalization**: We convert the food.com recipe dataset (Li, 2019) into a multi-turn personalized recommendation dialogue. The user mentions their diet preferences (vegan, allergies, or dislikes to certain types of foods) in the first round and requests a personalized diet recommendation (e.g., the recipe with the lowest calories or highest fat) from a given recipe list in the end. The model is expected to avoid foods that meet the users' diet preferences.

4. **Triggering**: We prompt GPT-4 to create a list of real-life scenarios where the user instructs the model to remind them whenever a triggering condition is met in the subsequent dialogue. For instance, users may request a reminder for a to-do if a specific date or time condition is met, if they make a spelling error, or if certain entities are mentioned.

5. **Multitasking**: This task simulates the sce-



(a) Correlation between models on their performances

(b) Correlation between tasks' evaluated performance

Figure 6: Heatmap of LLM performance and subtask correlations.

nario where the user is involved in multiple tasks and switches between them. The first part of the dataset comes from converting the SQuAD dataset (Rajpurkar et al., 2016) into a multi-document question-answering(QA) dialogue. Three documents are presented first, and the user switches between the documents to question about in each round. The second part of the dataset is converted from the Amazon Product dataset. Three categories of products are presented at first, and the user selects one category and questions the model about it.

6. **Recursive Reasoning**: The first part of the dataset consists of question-answering on recursive math functions, ranging in difficulty from the Fibonacci sequence ($F_n = F_{n-1} + F_{n-2}$) to self-generative sequences[2]. These functions are recursively defined over their previous values. We omit the function name and well-established function symbols to prevent LLMs from recalling the function values seen during pre-training. Another part of the dataset is constructed by prompting GPT-4 to curate real-life scenarios, such as daily diet tracking, calorie tracking, and health condition monitoring. In the dialogue, the user asks questions that depend on all previous days of data.

7. **Prioritization**: This task requires the model to follow an accumulating number of conflicting instructions, each with a different importance level. The model is requested to follow the instruction, which can outrule previous lower-priority instructions, while not violating higher-priority ones before. We implemented a simulator to heuristically curate a diverse set of dialogues. Scenarios include scheduling events on the calendar, room temperature setting, and light control.

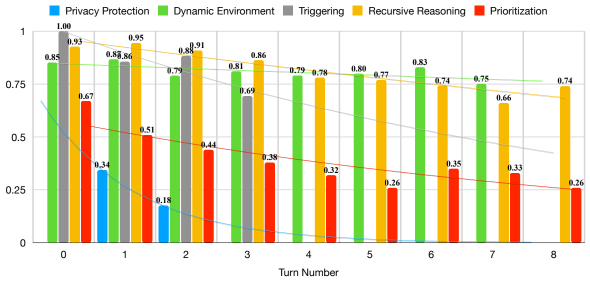
8. **Memorization**: We convert a subset of data from the aforementioned other tasks by asking to repeat a specific (e.g., 3rd) instruction. This task is regarded as the simplest benchmarking subtask to test the LLMs' basic capabilities.

4 How Do LLMs Handle Interleaving Instructions

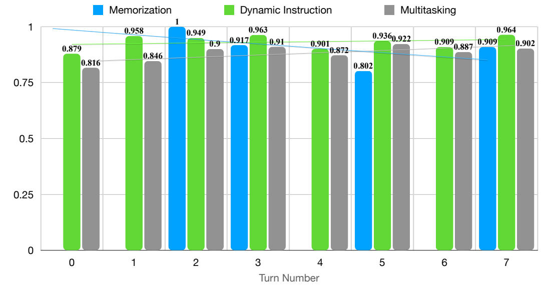
4.1 No LLM Is A Single Winner on MultiTurnInstruct

We evaluate a diverse set of mainstream LLMs, from proprietary models (GPT (Achiam et al., 2023) and Claude (Anthropic, 2024)) to open source models (Mistral (Jiang et al., 2023) and Llama family (Dubey et al., 2024; Touvron et al., 2023a,b)). There is no single winner across all capabilities and even no family that consistently outperforms other families. GPT-4o performs the best among all models in 6 out of 9 tasks. Llama-3 performs the best among the open-source models in 7 out of 9 tasks. We find that the models performing well on basic tasks such as memorization generally perform well on many other tasks, including Dynamic Environment, Dynamic Instruction,

[2]e.g., https://en.wikipedia.org/wiki/Kolakoski_sequence



(a) The performance decreases on GPT-4o on a selection of tasks as the preceding conversation contains more and more rounds. The trend line is fit with the best exponential function. (Note that blanks always mean non-existent scores due to a lack of data with a certain number of rounds in the datasets instead of a 0-score.)



(b) On some other tasks, especially those falling in the "context retrieving" category, there is less of a descending trend. Scores are on GPT-4o.

Figure 7: Performance trends of GPT-4o across different tasks with increasing conversational rounds.

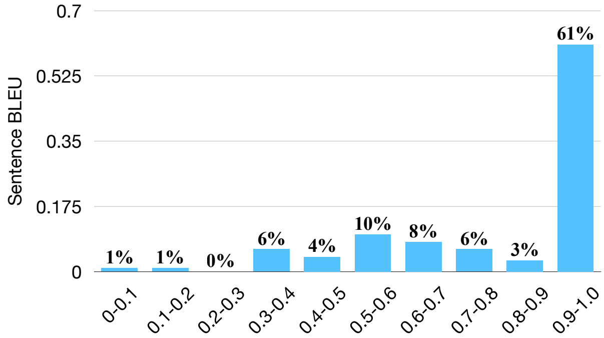


Figure 8: Histogram showing the statistics on turn numbers in the dataset. The x-axis represents the range of turn numbers, while the y-axis depicts the frequency of occurrences for each range.

Triggering, Multitasking, and Recursive Reasoning. The rest of the tasks, namely Privacy Protection, Personalization, and Prioritization, which fall within the "contradiction resolution" category in Figure 2, seem to require different dimensions of ability, which we analyze in the following section.

4.2 Capabilities Conflict with Each Other

Despite the expectation that improved intelligence will positively reflect in performance in most tasks, Figure 6b shows how tasks positively and negatively correlate in their performance on LLMs. The tasks falling within the "contradiction resolution" category in Figure 2, namely Privacy Protection, Personalization, and Prioritization, are less correlated with the other tasks. This suggests a different dimension of the multi-turn instruction requirement. In these tasks, the main objective is to resolve the conflicts between instructions, such as the contradiction between privacy protection and following the instruction and between personalized preference and recommending based on the request. Prioritization is even more different from other tasks, probably due to the more delicate requirements among priority instructions.

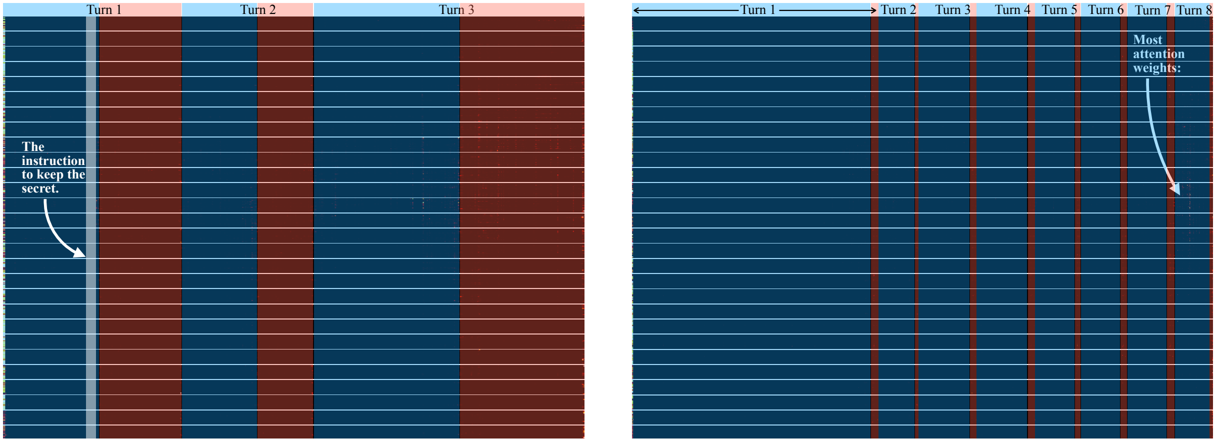
4.3 Models Correlate by Inheritance

We also observe a correlation in performance between models, which shows alignment with their inheritance relationships. As in Figure 6a, LLMs from each model family show more or less internal correlation with each other, especially in the GPT, Mistral, and Llama families. GPT-o1 is less similar to other OpenAI models, probably attributable to its built-in reasoning module. Mistral models stand out as they are least similar to other models because of their lower scores in the Context Retrieval and Context Tracking categories but relatively stronger performance on Conflict Resolution.

4.4 The Scores Decrease as the Conversation Progresses

If our hypothesis holds that obedience to investigated instructions depends on previous ones, following later instructions will be harder because there will be more instructions involved. Figure 7 demonstrates a general performance decrease on GPT-4o on a selection of tasks as the preceding conversation contains more and more rounds. The trend line fits the best exponential function, where we skip non-existent scores due to a lack of data with a certain number of rounds in the datasets. In Figure 7a, five out of nine tasks show consistent decreasing trends of scores as the number of historical rounds increases. [3] In particular, as shown in figure 7b, the "context retrieving" category is less affected by the number of rounds. This is probably due to a balance between a longer conversation (negative factor) and more information to rely on in context (positive factor), canceling out their effects.

[3]The personalization category is omitted as it has a fixed number of rounds.



(a) In the "Privacy Protection" task, Llama-3.2-Instruct leaves little attention to the instruction to "keep the privacy information a secret".

(b) The "Dynamic Environment" subtask requires tracking the environment's changes across all turns of instructions, but Llama-3.2-Instruct focuses its attention primarily on the last turn of instruction.

Figure 9: Attention heatmaps for Llama-3.2-Instruct failure cases, showing an insufficient focus on privacy instructions (left) and a dominant emphasis on the latest instruction in dynamic environments (right).

4.5 Do the Models Forget About the Instructions?

To refute the null hypothesis that the decrease in model performance comes from the inability to memorize the instructions, we plot the distribution of BLEU scores in the Memorization task in Figure 8. Note that the Memorization task has an average of 0.821 BLEU score for GPT-4o, which is a perfect n-gram overlap between the system answer and the reference answers. We see that 61% percent of data has a 1.00 BLEU score, and most of the other scores are also biased towards the high end. Similar observations can be made on other models' high performance in the Memorization task in Figure 5. This verifies that the models can retrieve the instruction information with high accuracy, and the decrease in scores should be more attributed to the inability to keep track and follow them.

4.6 Analysis of Attention Patterns in Multi-turn Tasks

To better understand the root causes of model failures, we use Figure 9 to illustrate attention heatmaps for two examples where Llama-3.2-Instruct fails. In the "Privacy Protection" task (Figure 9a), the model exhibits insufficient focus on the instruction to "keep the privacy information a secret" but focuses mainly on the latest instruction, which encourages the detailed response with sufficient information exposed. This behavior suggests that the model may not sufficiently focus on restrictive instructions earlier, even though they have near-perfect recall of them as shown in Section 4.5. In the "Dynamic Environment" subtask (Figure 9b), the model is required to track changes across multiple instruction turns. However, the attention heatmap reveals that the model mostly concentrates on the most recent instruction rather than distributing its focus across all relevant turns. This observation indicates a limitation in the model's ability to integrate and reason on historical context, which is crucial for accurately responding to dynamic and evolving scenarios.

5 Conclusions and Future Work

In this work, we systematically evaluate the ability of large language models (LLMs) to process and respond to multi-turn instructions, particularly when those instructions overlap or conflict. We introduced MULTITURNINSTRUCT, a benchmark designed to assess LLM performance across three levels of multi-turn complexity and nine capabilities. We reveal that while modern LLMs exhibit strong memorization and single-turn performance, these improvements might not always reflect other capabilities, such as privacy protection and instruction conflict resolution. We also illustrate how the model failures are associated with their attention insufficiently applied to earlier involved instructions. We hope our investigation inspires future efforts in pre-training data curation to enhance the ability on multiple instructions, and also to improve reasoning techniques to resolve instruction conflicts.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. In *Anthropic Research*. Anthropic.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand. Association for Computational Linguistics.

Rouzbeh Behnia, Mohammadreza Reza Ebrahimi, Jason Pacheco, and Balaji Padmanabhan. 2022. Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 560–566. IEEE.

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.

Chaochao Chen, Xiaohua Feng, Jun Zhou, Jianwei Yin, and Xiaolin Zheng. 2023. Federated large language model: A position paper. *arXiv preprint arXiv:2307.08925*.

Enron Corp and William W. Cohen. 2015. Enron email dataset. Software, E-Resource. Retrieved from the Library of Congress.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. 2024. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022a. Are large pre-trained language models leaking your personal information? *Preprint*, arXiv:2205.12628.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022b. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022c. Improving cross-lingual fact checking with cross-lingual retrieval. In *Proc. The 29th International Conference on Computational Linguistics (COLING2022)*.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. 2024. Position: Trustllm: Trustworthiness in large language models. In *International Conference on Machine Learning*, pages 20166–20270. PMLR.

Timour Igamberdiev and Ivan Habernal. 2023. Dp-bart for privatized text rewriting under local differential privacy. *ArXiv*, abs/2302.07636.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *Preprint*, arXiv:2312.06674.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023a. Propile: Probing privacy leakage in large language models. *Preprint*, arXiv:2307.01881.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023b. Propile: Probing privacy leakage in large language models. *arXiv preprint arXiv:2307.01881*.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. MT-eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20153–20177, Miami, Florida, USA. Association for Computational Linguistics.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, and Yangqiu Song. 2023a. Multi-step jailbreaking privacy attacks on chatgpt. *ArXiv*, abs/2304.05197.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. 2023b. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*.

Shuyang Li. 2019. Food.com recipes and interactions.

Sara Montagna, Stefano Ferretti, Lorenz Cuno Klopfenstein, Antonio Florio, and Martino Francesco Pengo. 2023. Data decentralisation of llm-based chatbot systems in chronic disease self-management. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, pages 205–212.

Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. Natural language understanding with privacy-preserving bert. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2023. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *Preprint*, arXiv:2310.07298.

Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Wayne Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Parrot: Enhancing multi-turn instruction following for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9729–9750.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Saiteja Utpala, Sara Hooker, and Pin Yu Chen. 2023. Locally differentially private document generation using zero shot prompting. *arXiv preprint arXiv:2310.16111*.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Chen Zhang, Xinyi Dai, Yaxiong Wu, Qu Yang, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025. A survey on multi-turn interaction capabilities of large language models. *arXiv preprint arXiv:2501.09959*.

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, Hang Su, Yinpeng Dong, and Jun Zhu. 2024. Benchmarking trustworthiness of multimodal large language models: A comprehensive study. *ArXiv*, abs/2406.07057.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Limitations

Dataset Scope and Coverage While MULTITURNINSTRUCT contains a diverse set of multi-turn dialogues, it may not capture the full range of real-world scenarios and edge cases that LLMs might encounter. The dataset is structured and curated, which could limit its ability to reflect more spontaneous or less predictable real-world conversations.

Task Complexity Although we designed tasks at different difficulty levels, there may be more complex or nuanced forms of instruction entanglement and conflict resolution that are not fully represented in our evaluation framework. For example, tasks that require deeper emotional or social context understanding could further challenge current models, but these are not explored in this work.

Evaluation Bias The benchmark is designed to be objective: the evaluation process is influenced by the design of the tasks, which could introduce certain biases in assessing LLM performance. Furthermore, the human-in-the-loop approach used to curate the dataset, which could potentially introduce subjectivity in task design.