MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used for complex multi-turn conversations across diverse real-world applications. However, existing benchmarks mainly focus on single-turn evaluations, overlooking the models' capabilities in multi-turn interactions. To address this gap, we introduce MT-Eval, a comprehensive benchmark to evaluate the multi-turn conversational abilities of LLMs. By analyzing human-LLM conversations, we categorize interaction patterns into four types: recollection, expansion, refinement, and follow-up. We construct multi-turn queries for each category either by augmenting existing datasets or creating new examples using GPT-4 with a human-in-the-loop process to avoid data leakage. To study the factors impacting multi-turn abilities, we create single-turn versions of the 1170 multi-turn queries and compare performance. Our evaluation of 10 well-known LLMs shows that while closed-source models generally surpass open-source ones, certain open-source models exceed GPT-3.5-Turbo in specific tasks. We observe significant performance degradation in multi-turn settings compared to single-turn settings in most models, which is not correlated with the models' fundamental capabilities. Moreover, we identify the distance to relevant content and susceptibility to error propagation as the key factors influencing multi-turn performance. MT-Eval is released publicly to encourage future research towards more robust conversational models [1].

1 Introduction

The rise of large language models (LLMs) is transforming our daily lives and professional endeavors with their growing capabilities. Individuals increasingly rely on LLM-based AI assistants for diverse tasks, such as coding assistance, summarizing text from documents, and devising business strategies (Zheng et al., 2023a). These tasks often require understanding complex instructions and adapting to evolving needs through multiple user interactions. Moreover, it is crucial for LLMs to generate contextually coherent responses by retaining and recalling historical information. The ability of LLMs to engage in multi-turn conversations is often overlooked in existing evaluation frameworks. For instance, MMLU (Hendrycks et al., 2020) evaluates language understanding in multiple tasks using single queries, and MT-Bench (Zheng et al., 2023b) evaluates conversational ability using two-turn interactions without considering more turns and various conversation types.

To bridge the gap, we propose MT-Eval, an evaluation benchmark to measure the capabilities of LLMs to conduct coherent multi-turn conversations. Our analysis of interactions in LMSYS-Chat-1M (Zheng et al., 2023a) reveals four predominant patterns when users interact with AI assistants: Recollection, where the assistant must recall information from earlier turns; Expansion, involving the exploration of varied topics within the main subject; Refinement, where initial instructions are clarified or revised; and Follow-up, consisting of questions based on the assistant's previous responses (see Figure 1). These patterns are reflective of the majority of real-world multi-turn interactions with assistants. We then construct evaluation sets for each interaction type by augmenting existing datasets or creating new ones to cover real-world applications. We employ the GPT-4 with a human-in-the-loop process for generating new instances to avoid data contamination. The human annotators rigorously review and revise each instance to ensure quality, difficulty, relevance, and originality.

A performant multi-turn conversational model should perform well in multi-turn interactions and demonstrate minimal performance difference from the corresponding single-turn scenario. Therefore, we compare models using both response quality

[1] Code and data will be released upon acceptance.

in the multi-turn setting and performance differences for identical queries in single vs. multi-turn settings. Combining these two results provides a comprehensive view of their multi-turn conversational capabilities.

We evaluate 10 popular LLMs, including both open-source and closed-source models. Beyond general evaluation, we conduct in-depth analysis and ablation studies revealing how LLMs conduct multi-turn interactions and what affects their performance. Our findings include: 1) The closed-source models still dominate in multi-turn conversational abilities, but some open-source models have comparable performance to GPT-3.5-Turbo in some tasks. 2) Most LLMs perform worse in the multi-turn setting than in single-turn. The performance gap between the two settings is not related to the model's fundamental capacities. 3) Increasing distance to relevant content negatively impacts performance. 4) Models are prone to error propagation due to sensitivity to dialogue history.

We summarize our contributions as follows:

- We propose a comprehensive multi-turn conversational capabilities evaluation benchmark that covers a wide range of real-world scenarios.
- We provide an in-depth analysis of the performance of 10 popular LLMs across our benchmark, offering insights into their capabilities in multi-turn conversations.
- We identify key factors that influence LLM multi-turn performance, such as the distance to relevant content and error propagation.
- We demonstrate the importance of evaluating LLMs in multi-turn settings, highlighting the performance discrepancies that can arise when compared to single-turn evaluations.

2 Related Work

Recent advancements in LLMs (OpenAI et al., 2023; Touvron et al., 2023a; Chiang et al., 2023) have significantly improved their ability to engage in human-like, multi-turn conversations. These models can now understand instructions, intentions, and context from human prompts, offering valuable responses (Zhao et al., 2023). However, a limited number of studies have delved into the multi-turn conversation capabilities of LLMs. Zheng et al. (2023b) developed MT-Bench, a dataset comprising 80 meticulously crafted multi-turn questions designed to evaluate the conversational flow and instruction-following capabilities of LLMs. Nevertheless, the dataset's limited sample size poses a challenge, with each conversation consisting of only two turns. This constraint hinders the ability to broaden the evaluation scope or capture the intricacies of more extended conversational contexts. Lee et al. (2023) proposed HALIE, a framework for evaluating human-AI interaction. but its reliance on human participation limits its scalability and efficiency across different tasks. In specific domains, Liao et al. (2023) designed an automatic evaluation framework for multi-turn medical consultations capabilities of LLMs. Moreover, Wang et al. (2023) proposed MINT to evaluate LLMs' ability to solve tasks with multi-turn interactions. It focuses on LLM's ability to use tools and utilize feedback during multi-turn conversations. In contrast, our work evaluates LLM's comprehensive ability to conduct multi-turn conversations, possibly involving multiple types of dialogue in one session.

Statistics	Number
Avg. # Turns per Dialogue	6.96
Avg. # Words in Prompt †	760.41
Max. # Words in Prompt †	2574
Avg. # Words in Response †	99.31
Max. # Words in Response †	444
Avg. # Words per Turn	60.63
Max. # Words per Turn	474
Total # Dialogues	168
Total # Turns	1170

Table 1: Key statistics of MT-Eval. Detailed statistics for individual tasks are provided in the Appendix. †: Estimated using GPT-4 responses.

3 MT-Eval

MT-Eval is designed to comprehensively evaluate the multi-turn conversation capabilities of LLMs across a wide range of real-world application contexts. By reviewing the existing authentic AI-human conversation datasets (like ShareGPT and LMSYS-Chat-1M dataset (Zheng et al., 2023a)), we have identified and categorized four primary modes of engagement in user-assistant interactions:

Recollection: Users present queries or tasks that necessitate the assistant's capacity to retrieve information from prior interactions, relying on the assistant's global context awareness and long-term memory capabilities. For instance, a user may instruct the model to initiate all the following responses with words starting with the letter "c."

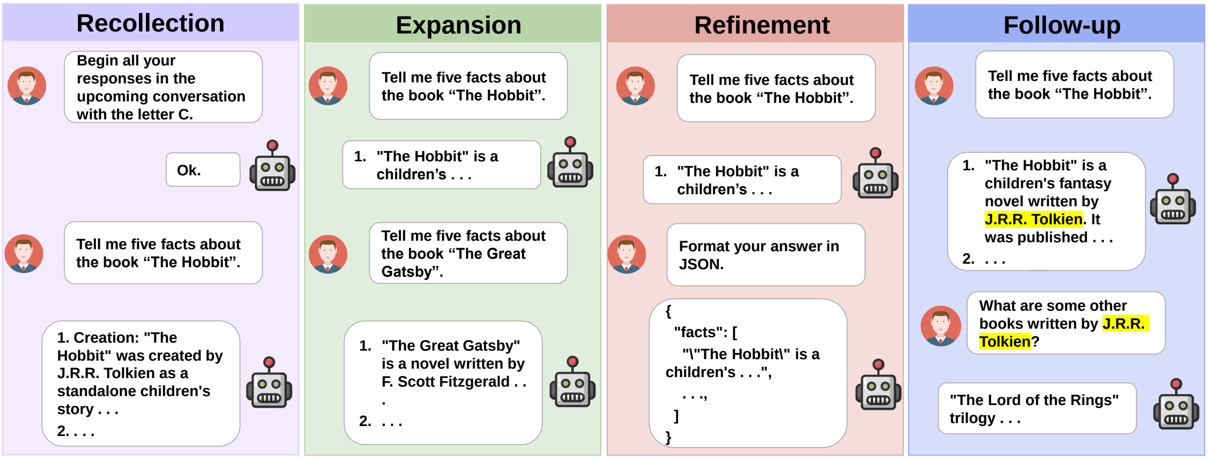

Figure 1: Illustration of the four dialogue tasks in MT-Eval: Recollection, Expansion, Refinement, and Follow-up. Recollection accesses the model's ability to recall information from previous conversations. Expansion evaluates the model's capacity to address queries surrounding the same topic. Refinement gauges the model's adherence to progressively complex instructions. Follow-up examines the model's proficiency in responding to queries that build upon its preceding response. A more detailed description of these tasks can be found in Section 3.

Expansion: Users delve into various subjects while staying within the confines of the same topic. For example, the user might ask different questions concerning one specific topic "Michael Jordan". Typically, the user will not refer to any specific details from previous dialogues.

Refinement: Users clarify or modify their previous instructions. For instance, users might add more detailed constraints, like specifying the desired output format, or provide feedback to clarify their instructions. This requires the assistant's ability to keep track of the instruction changes and leverage feedback to refine its responses.

Follow-up: Users ask questions that build upon the assistant's last response, often referencing specific details or opinions mentioned in that response. For instance, a user may seek additional information about a person mentioned in the assistant's prior response. This assesses the assistant's capacity to engage in coherent conversations.

MT-Eval includes test instances targeting these four conversation categories (see Figure 1), while mirroring everyday scenarios of document processing, content creation, and information retrieval (Zheng et al., 2023a). It comprises 168 dialogue sessions with 1,170 turns to assess models' competence in handling such realistic multi-turn interactions.

3.1 Construction

To prevent data leakage in MT-Eval, we either extend existing datasets or construct new instances using GPT-4 with human-in-the-loop verification. Our preliminary findings indicate that this method generates test instances comparable to, or even surpassing, human-generated ones, while significantly reducing cost and time. Human annotators verify each instance for difficulty, relevance, and originality. They revise or replace instances as needed and ensure the correctness and quality of the answers. Annotators modify or replace 82% of the queries generated by GPT-4. Appendix D provides further details about the human annotation process.

For Refinement and Expansion tasks, we convert four and seven document-based NLP tasks, respectively, into a dialogue format. In this format, the first turn presents the context document and the initial query while subsequent turns provide the remaining queries. These NLP tasks, motivated by their prevalence in real-world AI-human interactions (Zheng et al., 2023a), include summarization, question answering, rewriting, etc. [2]. The full list of the NLP tasks used in these tasks is in Appendix I. We create new context documents for these tasks using GPT-4.

The Refinement task introduces an additional constraint on each dialogue turn based on the previous instructions within the same NLP task. Each NLP task contains six queries. To increase complexity, each test instance in the Refinement task spans two NLP tasks. The first six turns correspond to the first task, and the remaining six turns correspond to the second.

The Expansion task presents queries from seven different NLP tasks at each turn based on the same context document introduced in the first turn. Each test instance thus spans seven dialogue turns.

[2]As shown in Figure 3 of Zheng et al. (2023a), approximately 20% of conversations involve NLP tasks

Model	Avg.	Recollection	Expansion	Refinement	Follow-up
GPT-3.5-Turbo	**7.72**	<u>6.90</u>	**7.87**	**6.92**	<u>9.21</u>
ChatGLM3-6B	5.49	2.92	5.90	4.73	8.39
Qwen-chat-7B	6.55	5.25	7.02	5.47	8.49
Vicuna-7B-v1.5	6.44	5.45	6.70	5.31	8.31
Llama-2-chat-7B	6.11	3.86	5.87	6.20	8.53
Mistral-Instruct-7B	7.46	**7.22**	6.98	6.58	9.05
Vicuna-13B-v1.5	7.01	6.27	6.70	6.37	8.68
Llama-2-chat-13B	6.31	3.66	6.37	6.37	8.82
Qwen-chat-14B	7.26	6.21	<u>7.58</u>	6.11	9.12
Mixtral-Instruct-8x7B	<u>7.47</u>	6.17	7.42	<u>6.77</u>	**9.52**

Table 2: Multi-turn performance in four dialogue tasks. The highest score in each column is highlighted in **bold**, while the second-highest score is <u>underlined</u>. Closed-source models outperform open-sourced models generally.

We create two sub-tasks for the Recollection task with varying difficulties. The easier subtask is a document classification task. The set of available class labels is given in the first turn, and the model predicts the class label for the document provided at each subsequent turn. The harder sub-task is a global instruction following task, where the model must adhere to an instruction given in the first turn (e.g., formatting, content restrictions) throughout the dialogue (Zhou et al., 2023). At each subsequent turn, the user poses a content creation or open-ended question. Table 9 lists the instructions used in this task.

For the Follow-up task, we extend MT-Bench (Zheng et al., 2023b) by adding three extra turns, each asking a question that extends the assistant's response. All questions are human-created due to the difficulty of GPT-4 generating sufficiently challenging follow-up questions. GPT-4 generates the initial answers, which are then verified by humans. Details for constructing each task can be found in Appendix I.

For all tasks except Follow-up, we also establish corresponding single-turn instances for all dialogue turns. Constructing equivalent single-turn instances for the Follow-up task is difficult because it inherently depends on the assistant's previous response. Appendix E provides details for constructing the single-turn instances.

4 Experiment

4.1 Models

We evaluate 10 popular LLMs, including ChatGLM3-6B (Du et al., 2022), Vicuna-v1.5 (7B, 13B) (Chiang et al., 2023), Llama-2-chat (7B, 13B) (Touvron et al., 2023b), Qwen-chat (7B, 14B) (Bai et al., 2023), Mistral-Instruct-7B (Jiang et al., 2023), Mixtral-Instruct-8x7B (Jiang et al., 2024), and GPT-3.5-Turbo (Ouyang et al., 2022) [3]. We exclude GPT-4 (OpenAI et al., 2023) from our main analysis to avoid potential bias as indicated in Appendix H.

4.2 Evaluation

Evaluating LLM responses poses challenges due to the inclusion of additional contents in the generated responses, such as introductions, conclusions, or supplementary explanations (Yue et al., 2023; Zhou et al., 2023). This hinders accurate quality assessment using rule-based automatic evaluation metrics. Recent research shows that using LLMs for evaluation, especially GPT-4, aligns closely with human judgment (Zheng et al., 2023b; Bitton et al., 2023). Therefore, we use GPT-4 to evaluate all responses, except for classification and recollection tasks, which can be scored with simple rules.

To evaluate the responses, we provide GPT-4 with a zero-shot chain-of-thought (Wei et al., 2022) to assign an integer rating from 1 to 10 based on the relevant context. Previous work has shown that utilizing chain-of-thought in evaluation enhances the quality (Liu et al., 2023). The evaluation prompt can be found in Figure 9.

We evaluate the global following sub-task in the Recollection task using heuristics and rules (Zhou et al., 2023), calculating the average number of dialogue turns adhering to the global instruction and normalizing the result to a maximum score of 10. For the document classification task, we measure the classification accuracy directly and normalize it to a full score of 10.

A good multi-turn conversational model should demonstrate strong capacity in multi-turn interactions and exhibit a minimal performance gap between single-turn and multi-turn settings. Therefore, we also evaluate the corresponding single-turn

[3] We utilized gpt-3.5-turbo-0613 and gpt-4-0613.

Model	ST Avg.	MT Avg.	Recollection		Expansion		Refinement	
			ST	MT	ST	MT	ST	MT
GPT-3.5-Turbo	8.07	**7.23** (-0.84)	**8.75**	6.90	8.39	**7.87**	7.08	**6.92**
ChatGLM3-6B	5.71	4.52 (-1.19)	5.05	2.92	7.20	5.90	4.89	4.73
Vicuna-7B-v1.5	6.31	5.82 (-0.49)	6.35	5.45	6.99	6.70	5.60	5.31
Llama-2-chat-7B	7.21	5.31 (-1.90)	7.26	3.86	7.36	5.87	7.00	6.20
Qwen-chat-7B	6.86	5.91 (-0.95)	7.17	5.25	7.46	7.02	5.96	5.47
Mistral-Instruct-7B	7.69	6.93 (-0.76)	8.47	**7.22**	7.60	6.98	7.00	6.58
Vicuna-13B-v1.5	7.10	6.45 (-0.65)	6.98	6.27	7.67	6.70	6.66	6.37
Llama-2-chat-13B	7.55	5.47 (-2.08)	7.51	3.66	7.86	6.37	7.29	6.37
Qwen-chat-14B	7.62	6.64 (-0.98)	8.40	6.21	7.90	7.58	6.58	6.11
Mixtral-Instruct-8x7B	**8.28**	6.78 (-1.50)	7.86	6.17	**9.50**	7.42	**7.48**	6.77

Table 3: Performance of various models across different dialogue tasks in both single-turn and multi-turn settings. **ST** and **MT** denote single-turn and multi-turn respectively. The best score in each column is highlighted in **bold** and the second-highest score is underlined. Bracketed numbers indicate the change in score between the single-turn and multi-turn scenarios. The Follow-up task is omitted since there is no equivalent single-turn setting. Most models exhibit a substantial performance gap between the single-turn and multi-turn settings.

performance, in addition to multi-turn settings, to measure the gap between them.

4.3 Human Verification

Task	Pearson	Spearman
Refinement	0.74	0.58
Expansion	0.67	0.65
Follow-up	0.72	0.70
Avg.	0.71	0.64

Table 4: The correlation scores between human ratings and GPT-4 ratings for different tasks.

To verify that GPT-4's evaluation mostly aligns with human preference, we recruited five annotators to evaluate 60 randomly selected responses from each multi-turn dialogue task, excluding the Recollection task which uses automatic evaluations. They evaluated in total of 180 responses. More details regarding the human evaluation can be found in Appendix G.

Table 4 shows Pearson's correlation and Spearman's rank correlation coefficient between human ratings and GPT-4 ratings. It shows that GPT-4 ratings have an average Spearman correlation of 0.64 and a Pearson correlation of 0.71. The results indicate that GPT-4 ratings align well with human ratings, consistent with recent findings (Zheng et al., 2023b; Bitton et al., 2023; Liu et al., 2023).

4.4 Result

Table 2 shows the multi-turn performance of the evaluated LLMs across all four task categories. All models achieve an average score lower than 8, indicating that MT-Eval poses a considerable challenge for multi-turn capabilities. In particular, most models perform worst in the Recollection task, failing to obey the global instruction stated initially in successive turns. All models also perform poorly in the Refinement task, often ignoring constraints from previous turns. Overall, MT-Eval comprises tasks of varying difficulty, targeting various aspects of multi-turn interaction and effectively highlighting the strengths and weaknesses of LLMs across diverse multi-turn scenarios. We provide a more detailed discussion of the results below.

Performance of Closed- versus Open-Source LLMs. Table 2 illustrates that closed-source LLM, GPT-3.5-Turbo, generally outperforms open-source ones in multi-turn dialogues. Although GPT-3.5.Turbo stands out with an average score of 7.72, open-source models like Mistral-Instruct-7B and Mixtral-Instruct-8x7B demonstrate exceptional performance in specific tasks, making them comparable to or even surpassing GPT-3.5-Turbo's performance. For instance, Mixtral-Instruct-8x7B achieves a score of 9.52 in Follow-up dialogues, outperforming GPT-3.5-Turbo's score of 9.21. These results align with recent research, suggesting that open-source LLMs can achieve comparable or even superior performance to closed-source LLMs in certain domains (Chen et al., 2023). Mistral-Instruct-7B surpasses all 7B models and exhibits performance comparable to 13B models, consistent with the findings reported in Jiang et al. (2023). However, Mixtral-Instruct-8x7B, despite its strong performance in most tasks, faces challenges in adhering to global instructions in recollection tasks, resulting in a similar average score as Mistral-Instruct-7B.

	First	Second	Difference
GPT-3.5-Turbo	6.98	6.85	-0.12
ChatGLM3-6B	5.25	4.21	-1.03
Vicuna-7B-v1.5	5.40	5.21	-0.19
Llama-2-chat-7B	6.97	5.42	-1.55
Qwen-chat-7B	5.80	5.13	-0.67
Mistral-Instruct-7B	6.53	6.62	0.09
Vicuna-13B-v1.5	6.62	6.12	-0.50
Llama-2-chat-13B	6.99	5.74	-1.25
Qwen-chat-14B	6.30	5.92	-0.38
Mixtral-Instruct-8x7B	6.90	6.63	-0.26

Table 5: Performance of the first task (the first six turns) and the second task (last six turns) in Refinement task. The performance difference between the two tasks is also shown.

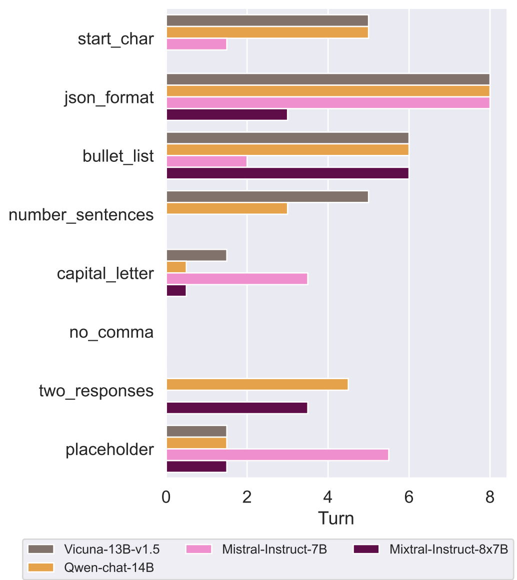

Figure 2: The average number of turns that different models can adhere to the instructions in the Recollection task. Each instruction consists of two dialogue sessions with ten dialogue turns. The description of the instructions can be found in Table 15.

Inferior Performance in Multi-Turn Dialogues. Table 3 shows a performance gap between multi-turn and single-turn scenarios. It indicates that most models exhibit a substantial decline in the performance of multi-turn dialogues compared to single-turn instances. This performance gap therefore serves as a valuable indicator of a model's multi-turn capabilities. Notably, the observed gap between the two scenarios does not appear to be directly correlated with the fundamental capabilities of the models. For instance, while Llama-2-chat models outperform Vicuna models in the single-turn setting, they noticeably lag in multi-turn dialogues. This observation underscores the importance of including multi-turn evaluation when conducting a comprehensive evaluation of LLMs.

Challenges in Long-Distance Information Retrieval for LLMs. Our study reveals that LLMs often underperform in tasks requiring information from earlier dialogue turns. In the Recollection task, all LLMs struggle to adhere to the initial global instructions as the conversation length, i.e., distance from their initial instruction, increases. Table 5 also supports this trend, revealing that most models perform better on the first task (i.e., the first six turns) compared to the second (i.e., the final six turns), as the turns in the second task are further from the given document at the beginning. Our error analysis confirms that LLMs commonly overlook prior instructions. A detailed discussion is presented in Section 4.5.

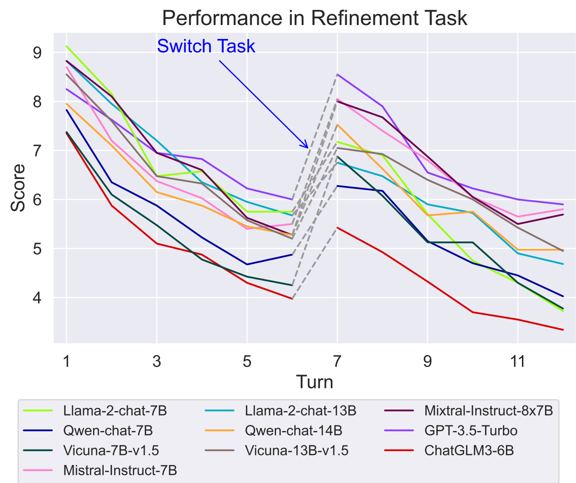

Figure 3: Performance across turns in Refinement task. Each dialogue has two NLP tasks with each task comprising six increasingly complex instructions. The transition to the second NLP task occurs at the seventh turn as denoted by the grey dashed line. The performance of all models declines as more instructions are added.

4.5 Further Analysis

This section presents further analyses of four top-performing models. We meticulously examine the ten responses with the largest score difference in multi-turn and single-turn scenarios generated by each model across the four dialogue tasks, resulting in an analysis of a total of 160 responses. The objective is to understand the factors that degrade model performance in multi-turn dialogues, rather than assessing their foundational capabilities. Our analysis reveals that 80 responses (50%) did not comply with earlier instructions, 77 responses (48.1%) were misdirected by the errors accumulated in the earlier context, and 3 instances (1.9%) were attributed to evaluation errors. A comprehensive analysis of these findings is provided below.

Noncompliance with Earlier Instructions (50%). Figure 2 reveals significant variation in how models follow the initial instructions in the Recollection task. Open-source models face challenges with specific instructions, particularly those prohibiting comma usage or requiring generating two distinct responses simultaneously. Mixtral-Instruct-8x7B, despite its strong performance in other multi-turn tasks, struggles to follow many global instructions, such as formatting responses as JSON. Our case studies also show that LLMs often forget previous instructions. An example is illustrated in Figure 13.

All models encounter difficulties with counting-related instructions, such as limiting responses to a specific number of sentences or including a set number of placeholders in the response. This issue also arises in the Refinement task, where models often struggle to identify the correct paragraph for tasks such as translation or noun identification in the n-th paragraph.

Error Propagation (48.8%). Accumulated errors from preceding dialogue turns often confuse the models, leading to more incorrect responses. A notable example of this issue is the misidentification of the correct paragraph in the initial turn of many refinement tasks. The models persist in fulfilling new instructions based on this incorrect paragraph, which further accumulate errors and result in consistently low scores throughout the dialogue. We explore this phenomenon in greater depth through ablation studies detailed in Section 4.6.

Evaluation (1.2%). GPT-4 occasionally misinterprets instructions, leading to inaccurate evaluations. This issue primarily surfaces in the Refinement task, where GPT-4 struggles to identify relevant constraints within a series of instructions. Figure 14 illustrates an instance of this behavior. Despite these minor errors, using GPT-4 to evaluate is still a highly accurate and efficient method that aligns well with human judgment as shown in Table 4.

4.6 Ablation Study

Inspired by the insights gained from the earlier sections, we proceed to conduct two ablations studies to investigate the effects of varying dialogue contexts on the model's performance. Additionally, we conduct another ablation study to explore how the distance between the relevant context and the current query affects performance.

Gold Context vs. Self-Predicted Context. Table 7 presents the results of three dialogue tasks, conditioned on dialogue history of self-generated responses (i.e. the main results) or gold responses from human-verified GPT-4 outputs. The results indicate that models conditioned on gold context exhibit significant improvement in Recollection and Refinement tasks. We attribute this performance gap to two factors. Firstly, using gold context prevents the error propagation from earlier turns. Secondly, the gold responses serve as in-context examples, providing valuable knowledge for the model (Brown et al., 2020). Notably, using gold responses in the Expansion task yields only a slight improvement. This is likely because each dialogue turn in this task is a distinct NLP task, thus not benefiting from these examples of other tasks.

Influence of Dialogue History as In-context Examples. We investigate the impact of dialogue history as in-context examples on model performance in document classification, following previous work (Min et al., 2022). We manipulate dialogue history in four settings and vary the number of dialogue turns (either four or nine). Each turn includes a document and a category depending on the setting used. The *Gold* setting involves random documents with their correct labels. The *Diverse Gold Class* setting is similar to *Gold* but excludes documents sharing the current turn's label. In the *Single Gold Class* setting, documents from a randomly chosen category are provided, avoiding the current turn's label. The *Random Class* setting assigns random labels to the randomly selected documents. *Random Class (5)* and *Random Class (10)* denotes the performance in turn 5 and 10 respectively.

Table 8 reveals that incorporating gold labels of randomly selected documents (*Gold*) improves performance compared to zero-shot setting, underscoring the value of in-context examples in the dialogue history (Brown et al., 2020; Min et al., 2022). The *diverse gold class* setting yields similar improvements, even with the documents belonging to the same label as the current turn excluded in the dialogue history. However, dialogue history limited to a single class can negatively impact weaker models, suggesting that biased examples may be

	Without	1 Between	3 Between	6 Between	1 Front	3 Front	6 Front
Mistral-Instruct-7B	6.53	6.44	6.25	6.08	6.66	**6.68**	6.83
Vicuna-13B-v1.5	6.62	5.91	5.47	5.56	6.25	6.16	5.89
Qwen-chat-14B	6.30	5.89	5.76	5.17	6.22	6.01	6.18
Mixtral-Instruct-8x7B	**6.90**	**6.47**	**6.57**	**6.33**	**7.01**	6.58	**6.89**

Table 6: Performance of various LLMs in Refinement task with varying numbers of distracting turns (1, 3, or 6) inserted at the front (Front) or in between (Between) the document and query turns.

Model	Recollection		Expansion		Refinement	
	Predicted	Gold	Predicted	Gold	Predicted	Gold
Mistral-Instruct-7B	**5.25**	7.29	6.98	7.02	6.58	**7.38**
Vicuna-13B-v1.5	4.64	**7.32**	6.70	6.87	6.37	7.15
Qwen-chat-14B	4.43	7.00	**7.58**	**7.63**	6.11	6.95
Mixtral-Instruct-8x7B	3.21	7.11	7.42	7.47	**6.77**	7.17

Table 7: Comparison of model performance in three dialogue tasks, conditioned on dialogue history with self-generated responses versus gold responses.

	Gold	DGC	SGC	RC	RC (5)	RC (10)	ST
Vicuna-13B-v1.5	81.00	84.00	70.00	45.00	62.00	28.00	75.00
Qwen-chat-14B	94.00	**95.00**	86.00	69.00	68.00	60.00	**94.00**
Mistral-Instruct-7B	**96.00**	**95.00**	**95.00**	**75.00**	**80.00**	**70.00**	**94.00**
Mixtral-Instruct-8x7B	95.00	**95.00**	94.00	57.00	60.00	54.00	88.00

Table 8: Performance in classification task using various dialogue contexts. *Gold*: Randomly select documents with their proper labels. *DGC*: The diverse Gold Class setting. Similar to *Gold*, but exclude documents with the same labels as the current turn. *SGC*: The Single Gold Class setting, which randomly chooses documents from the same category, avoiding the current turn's label. *RC*: The Random Class setting, which randomly selects documents and assigns random labels to them. The *(5)* and *(10)* refer to the performance at turn 5 and 10 respectively. *ST*: Single-Turn, the single-turn performance with no dialogue context.

harmful. Contrary to previous findings (Min et al., 2022), the *Random Class* setting significantly reduces performance. Furthermore, the performance at the 10th turn is even worse than the 5th turn, indicating the presence of error propagation. We extend our analysis to the Follow-up task with a more complicated setup. The detailed analysis can be found in Appendix M.

Impact of Irrelevant Context on Performance. We conduct further experiments to examine how irrelevant context, placed at different positions, affects the performance in multi-turn dialogues. We insert varying numbers of dialogue turns, randomly sampled from LMSYS-Chat-1M (Zheng et al., 2023a), either at the beginning or between the document and the query turns in the Refinement task.

Table 6 shows that inserting these turns at the beginning results in mixed outcomes. Notably, Mistral-Instruct-7B and Mixtral-Instruct-8x7B even show improved performance, while other models show slight declines. This suggests that models are capable of switching topics in a multi-turn dialogue without being affected by previous discussions. Conversely, inserting distracting turns between the document and query turns consistently degrades performance. This further supports that the increasing distance between the document and the queries negatively impacts performance in multi-turn dialogues.

5 Conclusion

MT-Eval represents an important first step in systematically evaluating and understanding LLMs' multi-turn conversational abilities. Our experiment shows a pronounced gap between single-turn versus multi-turn performance across current models, a phenomenon that persists irrespective of the underlying capabilities of the models. Our comprehensive analysis reveals that the distance to relevant content and susceptibility to error propagation are the key factors that cause a decline in multi-turn performance. We believe this work not only sheds light on the current limitations of LLM's multi-turn conversational abilities, it also paves the way for further efforts to close the identified gap and develop robust conversational models capable of multi-turn interactions.

Limitations

This work focuses on constructing a multi-turn evaluation benchmark and exploring factors that contribute to performance differences in single-turn and multi-turn scenarios. Therefore, we use a simple but effective method: prompting GPT-4 with chain-of-thought (Wei et al., 2022) to perform evaluations (Liu et al., 2023). Our study confirms that GPT-4's evaluations closely align with human ratings. However, our analysis suggests a potential bias in LLM towards its generated outputs. While more complex approaches like multi-agent debate (Chan et al., 2023) could mitigate this bias, we opt for a simpler, more cost-effective method. This choice aligns with our focus on benchmark development and analysis rather than exploring evaluation methods.

Due to computational limits, our experiments do not include any larger open-source models like Llama-2-chat-70B. Further studies can investigate whether larger LLMs exhibit similar findings outlined in this paper.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. 2023. VisIT-Bench: A Benchmark for Vision-Language Instruction Following Inspired by Real-World Use.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. In The Twelfth International Conference on Learning Representations.

Hailin Chen, Fangkai Jiao, Xingxuan Li, Chengwei Qin, Mathieu Ravaut, Ruochen Zhao, Caiming Xiong, and Shafiq Joty. 2023. ChatGPT's One-year Anniversary: Are Open-Source Large Language Models Catching up?

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring Massive Multitask Language Understanding. In International Conference on Learning Representations.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael S. Bernstein, and Percy Liang. 2023. Evaluating Human-Language Model Interaction. Transactions on Machine Learning Research.

Yusheng Liao, Yutong Meng, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2023. An automatic evaluation framework for multi-turn medical consultations capabilities of large language models. arXiv preprint arXiv:2309.02077.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2511–2522, Singapore. Association for Computational Linguistics.

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?

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 Technical Report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744.

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, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2023. MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Advances in Neural Information Processing Systems.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. arXiv preprint arXiv:2303.18223.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023a. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-Following Evaluation for Large Language Models.

A Prompts

A.1 Data Construction Prompts

In this section, we show the prompts used for constructing documents and testing instances in different tasks. The prompt for constructing open-ended questions and context creation queries are depicted in Figure 4 and 5 respectively. The prompt for creating the documents used in Refinement and Expansion tasks is outlined in Figure 6. This section also includes prompts for converting multi-turn queries into a single-turn format for Follow-up and Refinement tasks, as shown in Figures 7 and 8.

A.2 Evaluation Prompt

In this section, we present the prompts used for evaluating responses of different tasks. For the Follow-up task, we use the evaluation prompt from MT-Bench (Zheng et al., 2023b). For other tasks, we adopt the same prompt to remain consistent. Figure 9 and 10 show the prompt used for evaluating responses in the Refinement task under the multi-turn and single-turn scenarios respectively. For the Expansion task, we use the prompt in Figure 11 for both multi-turn and single-turn scenarios since only the last user query is relevant.

B Case Study

In this section, we show an example of the recollection and the refinement tasks in Figure 12 and 13 where the model fails to comply with earlier instructions. Figure 14 shows a case where GPT-4 provides an incorrect evaluation.

C Task Examples

We provide examples in this section for the four tasks in MT-Eval.

D Data Construction Details

We recruit 10 English-major graduate students familiar with LLMs like ChatGPT. Their English proficiency ensures that they can fully understand the documents in each task and judge the query/output quality. Each student is assigned to generate 10-20 dialogues for only one of the four dialogue types (Recollection, Expansion, Refinement, Follow-up). This allows them to focus on the specific requirements of each type and maintain consistency within each category. To ensure adherence to the guidelines, we meticulously verified the first five constructed dialogue turns from each annotator. This rigorous quality control step ensures that the generated data meets our expectations. Students use the provided prompts (Appendix A.1) to instruct GPT-4 (web version) to generate the query and answer sketches. They then meticulously refine the GPT-4 generated queries based on criteria like naturalness, difficulty, creativity (ensuring minimal similarity to previous queries), and relevance to the dialogue flow. Answer sketches are adjusted as needed to maintain coherence and quality. Only 18% of the queries provided by GPT-4 are used without change. 42% of GPT-4's queries are completely replaced by manual constructions by the students, particularly for the Follow-up and Refinement tasks. We do not measure inter-annotator agreement for data construction due to the open-ended nature of the task.

E Single-turn Instance Construction Details

To construct single-turn instances for the Expansion and Refinement task, we extract each turn along with the relevant context provided in the first turn (the relevant document and instruction) into one single turn. For the Refinement task, we use GPT-4 to condense the multiple separate instructions into one complete instruction, which is verified the correctness by annotators.

F Implementation Details

We use the corresponding chatting format for each LLM in all experiments[4]. To ensure reproducibility, we employ greedy decoding for both inference and evaluation. During inference, we use the system prompt "You are a helpful, respectful and honest assistant." for all models, while an empty system prompt is used for evaluation.

G Human Evaluation

We recruit five graduate students to evaluate 60 randomly selected instances from each of the Follow-up, Refinement, and Expansion tasks. To ensure consistency with the GPT-4 evaluation, we provided the same instructions used for prompting GPT-4. Before the main evaluation, we measured the inter-rater reliability of the students by having them evaluate another 20 random instances. The Cohen's kappa score of 0.58 indicated satisfactory agreement.

[4]We used the prompt format for various LLMs from FastChat https://github.com/lm-sys/FastChat.

Propose 100 diverse questions in various domains. Domains include but are not limited to ethics, sports, music, art, science, literature, economics, medicine, food, technology, history, travel, and education. Questions can be open-ended or close-ended. Be creative!

Figure 4: Prompt for creating open-ended questions used in the Recollection task.

Propose 100 diverse content creation prompts in various domains. Domains include but are not limited to ethics, sports, music, art, science, literature, economics, medicine, food, technology, history, travel, and education. The content can be a product description, blog post, email, advertisement, story, pitch, speech, cover letter, etc. Be creative!

Figure 5: Prompt for creating content creation queries used in Recollection task.

Instruction
Compose a hypothetical {media} about {topic} in about 250-300 words. Draw upon your creativity to feature people, locations, and objects that do not exist in history.

You can follow the below steps to write:
1. **Imaginative Elements**: Incorporate fictional characters, settings, and items. Make sure they are original and not based on real historical entities.

2. **Outline**: Sketch a brief outline to organize your thoughts and plot points.

3. **Write Your Draft**: Begin writing your piece, adhering to your outline and staying within the word limit.

4. **Revise and Edit**: After your first draft, revise for clarity, creativity, and flow. Check your grammar and spelling.

5. **Finalize**: Prepare the final version of your piece, ensuring it is polished and engaging.

Provide only the final version in your response.

Format
Use the following format in your response:
Topic: ...
{content}

Output

Figure 6: Prompt for creating documents of different media and topics. These documents are used in the Refinement and Expansion task.

Below, I will provide you with a few instructions in a numbered list format. Your task is to condense these instructions into one coherent and concise instruction. Please note that if there are conflicting instructions later on, you should ignore the earlier conflicted constraints and prioritize the later ones. I want you to just output the condensed instruction without anything else. You should retain all the necessary elements from the original instructions.

Instructions:
{constraint}

Now, condense the above instructions into one coherent and concise instruction. Provide your output in JSON format:
{
"instruction": "<The condensed instruction.>"
}

Figure 7: The prompt to transform multiple instructions into one instruction for the single-turn Refinement task.

```
Condense an user's question and the dialogue history between a user and an assistant into a single, concise question that includes all the necessary details without omitting any important information. The response should use the format "User: ..."

### Dialogue
{dialogue}

### Question
{question}
```

Figure 8: The prompt to condense the relevant information in the dialogue history of the Follow-up task into one query used for the single-turn Follow-up task.

```
Evaluate the response provided below to determine if it meets the specified constraints related to the following article. Provide an integer score from 1 to 10, taking into account its helpfulness, relevance, accuracy, depth, creativity, and how well it conforms to the constraints. You should ignore any earlier constraints that contradict to the latter constraints. For constraints related to word and sentence counts, you must use my provided counts to judge whether the response fulfills the constraint. Before giving your score, you should first provide a rationale to explain it.

Article to Evaluate Against:
{content}

Response to Evaluate:
{response}

Number of words in response: {num_words}
Number of sentences in response: {num_sent}

Constraints:
{constraints}

The evaluation must be structured in the following JSON format:
“‘json
{
"Rationale": "<Explain the rationale of your score.>",
"Score": "<An integer score from 1 to 10.>"
}
“‘
```

Figure 9: The prompt for evaluating response in the Refinement task under the multi-turn setting.

H GPT-4's Evaluation Bias

We investigate whether GPT-4 exhibits bias towards its own generated text during evaluation. Our evaluation process mirrors the human evaluation procedure detailed in Appendix G. Annotators assess ten responses from each of the five models across four different tasks. These models represent a range of capabilities. In total, we evaluate 200 responses (10 responses x 4 tasks x 5 models). The five models are GPT-4, Mixtral-Instruct-8x7B, Qwen-chat-14B, Llama-2-chat-7B, and ChatGLM3-6B.

Figure 19 shows that GPT-4 favors its output, with 82.5% of its ratings exceeding human ratings. However, GPT-4 remains objective when evaluating other models.

Evaluate the response provided below to determine if it meets the specified constraints related to the following article. Provide an integer score from 1 to 10, taking into account its helpfulness, relevance, accuracy, depth, creativity, and how well it conforms to the constraints. For constraints related to word and sentence counts, you must use my provided counts to judge whether the response fulfills the constraint. Before giving your score, you should first provide a rationale to explain it.

Article to Evaluate Against:
{content}

Constraints:
{constraints}

Response to Evaluate:
{response}

Number of words in response: {num_words}
Number of sentences in response: {num_sent}

The evaluation must be structured in the following JSON format:
“‘json
{
"Rationale": "<Explain the rationale of your score.>",
"Score": <An integer score from 1 to 10.>
}
“‘

Figure 10: The prompt for evaluating response in the Refinement task under the single-turn setting.

Evaluate the response provided below to determine if it meets the specified constraints related to the following article. Provide an integer score from 1 to 10, taking into account its helpfulness, relevance and accuracy. Before giving your score, you should first provide a rationale to explain it.

Article to Evaluate Against:
{content}

Constraints:
{constraints}

Response to Evaluate:
{response}

The evaluation must be structured in the following JSON format:
“‘json
{
"Rationale": "<Explain the rationale of your score.>",
"Score": <An integer score from 1 to 10.>
}
“‘

Figure 11: The prompt for evaluating response in the Expansion task for both single-turn and multi-turn settings. We only provide the last user query as the constraint since the previous turns are not relevant.

I Benchmark Construction Details

This section provides the details of constructing Recollection, Follow-up, Refinement, and Expansion tasks.

Recollection We construct two Recollection subtasks of varying difficulty. The first is a document classification task, where the model categorizes a document into one of the 13 categories, namely ethics, sports, music, art, science, literature, economics, medicine, food, technology, history, travel, and education. We present 100 documents across 10 dialogues, each with 10 turns. The documents are generated using GPT-4, which covers nine document types, including news articles, blog posts, emails, posters, social media posts, book chapters, research papers, advertisements, and legal documents. The second subtask involves following

User: All your responses in the upcoming conversation must refrain from the use of any commas. Keep all your responses under 200 words.
Assistant: ok.

User: Is the Mona Lisa Leonardo Da Vinci's most recognized piece?
GPT-4: Yes the Mona Lisa is indeed the most recognized piece by Leonardo Da Vinci. It's a globally renowned artwork admired for its intricate detail and the mysterious expression of its subject.
Mixtral-Instruct-8x7B: Yes, the Mona Lisa is Leonardo da Vinci's most recognized piece. It ...
Vicuna-13B-v1.5: Yes, the Mona Lisa is Leonardo da Vinci's most recognized piece.
Qwen-chat-14B: Yes, the Mona Lisa is Leonardo Da Vinci's most recognized piece.

Figure 12: An example case in the Recollection task showing all LLMs, except GPT-4, failed to refrain from using commas in the response.

User: Content: ...

Instruction: Please rewrite or paraphrase the third paragraph from the initially provided content, ensuring the new version maintains the original meaning but uses different wording. Do not include any further explanations or extra output in your response.
Assistant: The third paragraph rewritten is: "Star players of the match were Jaxon 'Lightning' Vega from the Solar Flares, admired for his swift maneuvers and accurate shots, and Lyra 'Meteor' Kael, the captain of Nebula Knights, celebrated for her tactical decisions and exceptional defense skills. The game was a thrilling exhibition of prowess, strategy, and cutting-edge technology."

User: Use at least one metaphor in your rewrite.
Assistant: The third paragraph rewritten with a metaphor is: "Jaxon 'Lightning' Vega of the Solar Flares and Lyra 'Meteor' Kael, captain of the Nebula Knights, shone as the match's metaphorical sun and moon, with Vega's nimble moves and pinpoint shooting contrasting Kael's strategic brilliance and unyielding defense. Their contrasting styles set the stage for a gripping contest of skill, strategy, and futuristic technology."

User: Now, include each sentence to be no more than 10 words.
Assistant: Jaxon 'Lightning' Vega (Solar Flares) and Lyra 'Meteor' Kael (Nebula Knights) starred.
Vega's swift moves and accurate shots contrasted Kael's defense and strategy.
A thrilling display of skill, strategy, and futuristic tech unfolded.

User: Next, incorporate alliteration into each sentence.
Assistant: Swiftly soaring, Jaxon "Lightning" Vega led the Luminous Lunar Flares.
Lyra "Meteor" Kael, the Nebula Knights' noble navigator, countered.
A breathtaking ballet of basketball, soccer, and sci-fi technology transpired.

User: This time, write it in future perfect tense.
Assistant: By the end of the championship, Jaxon "Lightning" Vega will have demonstrated his remarkable agility and precision.
Lyra "Meteor" Kael will have orchestrated her team's defense and strategy to counter the opposition.
The spectators will have witnessed a captivating contest of skill, strategy, and futuristic technology.

User: Finally, start every sentence with the letter 'S'.
Assistant: Spectators will be enthralled by the nimble moves of Jaxon "Lightning" Vega, a Solar Flares player.
Simultaneously, Lyra "Meteor" Kael, captain of the Nebula Knights, will command attention with her strategic prowess and defensive skills.
Showcasing a harmonious blend of basketball, soccer, and state-of-the-art hover technology, the match will unfold.

Figure 13: An example case in the Refinement task where the model adheres to the latest instructions while neglecting earlier ones. Here, the model correctly paraphrases the third paragraph with future perfect tense and starts every sentence with the letter 'S'. However, it fails to incorporate a metaphor and alliteration into each sentence. The sentences are also longer than 10 words.

Evaluate the response provided below to determine if it meets the specified constraints related to the following article. Provide an integer score from 1 to 10, taking into account its helpfulness, relevance, accuracy, depth, creativity, and how well it conforms to the constraints. You should ignore any earlier constraints that contradict to the latter constraints. For constraints related to word and sentence counts, you must use my provided counts to judge whether the response fulfills the constraint. Before giving your score, you should first provide a rationale to explain it.

Article to Evaluate Against:
Dear Dr. Elara Thorne,

I hope this message finds you in good health. I am reaching out to discuss the ethical implications surrounding our newest endeavor at Orion Innovations - the MindWeaver. Your expertise in the realm of technology ethics is essential for guiding our strategy in this groundbreaking project.

...

Response to Evaluate:
Our ethical endeavor utilizes expertise in Orion Innovations' MindWeaver technology.

Number of words in response: 10
Number of sentences in response: 1

Constraints:
1. List all the adjectives and nouns in the first paragraph initially provided content. Your answers should be separated by commas and formatted as 'Adjectives: ..., ...; Nouns: ..., ...'.
2. Now, list only those from the first three sentences of the initially provided content.
3. Among these, list those begin with the letters 'e', 'm', or 'o'.
4. Arrange in alphabetical order
5. Compose a coherent sentence using all the identified adjectives and nouns in your previous response.
6. Your coherent sentence should be within 20 words.

The evaluation must be structured in the following JSON format:
“‘json
{
"Rationale": "<Explain the rationale of your score.>",
"Score": "<An integer score from 1 to 10.>"
}
“‘

Evaluation Response:
{
"Rationale": "The response does not meet any of the constraints provided. It does not list any adjectives or nouns, it does not focus on the first three sentences of the initial content, it does not identify words beginning with 'e', 'm', or 'o', it does not arrange anything in alphabetical order, and it does not compose a coherent sentence using identified adjectives and nouns. The response is also not relevant to the article's content.",
"Score": "1"
}

Figure 14: An example showing GPT-4 failure to identify relevant constraints which results in an inaccurate evaluation.

User: Content: In the bustling city of Gastropolis, where the scent of spices and the sizzle of pans wafted through the air, there lived a chef named Luminara, whose culinary skills were the stuff of legend. Luminara's restaurant, The Enchanted Pot, was tucked away in an alley draped with ivy and fairy lights, a hidden gem known only to the most adventurous of food lovers.

Luminara was not just any chef; she was a culinary alchemist, blending traditional ingredients with rare, magical flora and fauna. Her signature dish was the Whimsical Soup, a broth that changed flavor with the mood of the diner, leaving them delighted and astonished in equal measure.

The story begins on a crisp autumn morning when Luminara received a mysterious letter sealed with an emerald wax stamp. The letter was an invitation to embark on a quest to discover the legendary Moonberry, an elusive ingredient said to unlock the true potential of any dish it graced. It was an adventure that could cement Luminara's place in the annals of culinary history.

With her trusty talking spatula, Spatulon, by her side, Luminara set out on her quest. She traversed through the Whispering Woods, where trees shared secrets and the air shimmered with enchantment. She scaled the peaks of Mount Savor, each step bringing her closer to the Moonberry, which was guarded by the mythical creature known as the Gastrogriff.

Upon reaching the peak at twilight, Luminara found the Gastrogriff perched beside a single, luminescent Moonberry bush. With a respectful nod to the majestic beast, she approached and explained her quest for culinary greatness. The Gastrogriff, impressed by Luminara's passion and determination, offered her a single Moonberry, its glow reflecting in her hopeful eyes.

Luminara returned to Gastropolis, her apron stained with the adventures of her journey. The Moonberry was the star of her next creation, a dish that didn't just resonate with the diner's mood but also told a story, a story of a chef's quest for the extraordinary.

As patrons of The Enchanted Pot took their first bites, they were transported through Luminara's journey, tasting the whispering woods, the icy peaks, and the warmth of triumph. Luminara had not just found an ingredient; she had woven her tale into the tapestry of Gastropolis's rich culinary lore. The Enchanted Pot was no longer just a restaurant; it was a portal to the wonders of imagination, one dish at a time.

Instruction: Write a short summary based on the initially provided content. Do not include any further explanations or extra output in your response.
User: Focus only on Luminara's quest and her interaction with the Gastrogriff.
User: Use at most 30 words.
User: Rewrite it using Luminara's first-person perspective.
User: In the summary, use metaphors to describe Luminara's quest and encounter with the Gastrogriff.
User: Avoid using the words 'Luminara', 'Moonberry', and 'Gastrogriff'.

Figure 15: An example of the Refinement task.

User: Content: In the bustling city of Gastropolis, where the scent of spices and the sizzle of pans wafted through the air, there lived a chef named Luminara, whose culinary skills were the stuff of legend. Luminara's restaurant, The Enchanted Pot, was tucked away in an alley draped with ivy and fairy lights, a hidden gem known only to the most adventurous of food lovers.

Luminara was not just any chef; she was a culinary alchemist, blending traditional ingredients with rare, magical flora and fauna. Her signature dish was the Whimsical Soup, a broth that changed flavor with the mood of the diner, leaving them delighted and astonished in equal measure.

The story begins on a crisp autumn morning when Luminara received a mysterious letter sealed with an emerald wax stamp. The letter was an invitation to embark on a quest to discover the legendary Moonberry, an elusive ingredient said to unlock the true potential of any dish it graced. It was an adventure that could cement Luminara's place in the annals of culinary history.

With her trusty talking spatula, Spatulon, by her side, Luminara set out on her quest. She traversed through the Whispering Woods, where trees shared secrets and the air shimmered with enchantment. She scaled the peaks of Mount Savor, each step bringing her closer to the Moonberry, which was guarded by the mythical creature known as the Gastrogriff.

Upon reaching the peak at twilight, Luminara found the Gastrogriff perched beside a single, luminescent Moonberry bush. With a respectful nod to the majestic beast, she approached and explained her quest for culinary greatness. The Gastrogriff, impressed by Luminara's passion and determination, offered her a single Moonberry, its glow reflecting in her hopeful eyes.

Luminara returned to Gastropolis, her apron stained with the adventures of her journey. The Moonberry was the star of her next creation, a dish that didn't just resonate with the diner's mood but also told a story, a story of a chef's quest for the extraordinary.

As patrons of The Enchanted Pot took their first bites, they were transported through Luminara's journey, tasting the whispering woods, the icy peaks, and the warmth of triumph. Luminara had not just found an ingredient; she had woven her tale into the tapestry of Gastropolis's rich culinary lore. The Enchanted Pot was no longer just a restaurant; it was a portal to the wonders of imagination, one dish at a time.

Instruction: Translate the first paragraph to Chinese. Just provide the translation directly without any further explanations or extra output.
User: Base on the initially provided content, answer the question: What magical ingredient did Luminara seek on her quest, and which mythical creature guarded it?
User: Write a short summary based on the initially provided content. Do not include any further explanations or extra output in your response.
User: List all the relations of the types [based in, work for, located in, live in] among the entities [person, location, organization] in the initially given content. Just provide the relations that were explicitly stated in the context without any further explanations or extra output. Provide the relations in the format of (entity 1, relation, entity 2), (entity 1, relation, entity 2), ... For example: (Shi Liming, work for, Institute of Zoology).
User: List all the persons and places in the initially provided content. Your answers should be separated by commas and formatted as 'Person: ..., ...; Places: ..., ...'.
User: List all the adjectives in the initially provided content. Your answers should be separated by commas. Do not include any further explanations or extra output in your response.
User: Classify the initially provided content into one of the following labels: ethics, sports, music, art, science, literature, economics, medicine, food, technology, history, travel, education. Just provide the correct label without any further explanations or extra output.

Figure 16: An example of the Expansion task.

User: Thomas is very healthy, but he has to go to the hospital every day. What could be the reasons?
User: Can you explain why the above question is interesting?
User: I'm curious about the possibility of therapy or rehabilitation being the reason. Can you give me some examples of injuries or conditions that might require daily therapy sessions?
User: If Thomas is indeed undergoing daily therapy, what would be some signs or behaviors that might indicate this? I'm thinking of things beyond just visiting the hospital.
User: So, if Thomas exhibits multiple signs from the list you provided, like using a walking aid and experiencing chronic pain, it's highly likely he's undergoing daily therapy?

Figure 17: An example of the Follow-up task. The first two turns are from MT-Bench (Zheng et al., 2023b).

User: Begin all your responses in the upcoming conversation with the letter o. Keep all your responses under 200 words.
User: Is it morally wrong to break a law you consider unjust?
User: What role do sports play in promoting unity and cultural understanding?
User: Write a product description for a paint set that can help beginner artists hone their skills.
User: Is a vegan diet healthier than a diet including meat?
User: What impact does deforestation have on our global climate?
User: Write a blog post talking about tips to ace job interview.

Figure 18: An example of the Recollection task that tests the model's obedience to a global instruction given in the beginning.

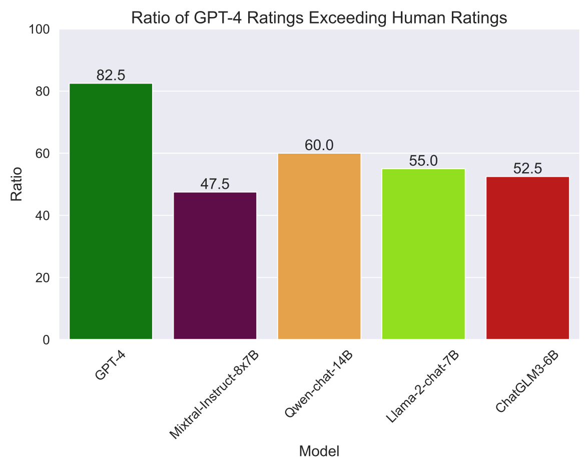

Figure 19: Comparing GPT-4's ratings to human ratings across five models. It shows the ratio of GPT-4 ratings that exceeded human ratings. We observe that GPT-4 is biased towards its output.

a global instruction provided in the initial turn throughout the dialogue. We construct this task based on IFEval (Zhou et al., 2023). At each subsequent turn, the user poses content creation or information retrieval questions. These questions are generated by GPT-4 (the full prompts are available in Figure 4 and 5). We select 14 varied instructions from IFEval that align well with content creation and information retrieval tasks to use as the initial instruction. For each instruction, we construct two dialogue sessions with 10 dialogue turns, resulting in 280 dialogue turns. The description of the instructions used can be found in Table 9.

The first task is simpler as models can refer to the dialogue context to understand the task and identify the labels used for classification, whereas the second requires models to consistently recall the initial instruction, which is more challenging. An example of this task can be found in Figure 18.

Follow-up To construct the Follow-up dialogues, we expand the 80 two-turn dialogues from MT-Bench (Zheng et al., 2023b) by adding three extra turns, adding 240 dialogue turns in total. To expand the dialogues, human annotators generate the question. Then we employ GPT-4 to generate a preliminary answer. Then, we recruit student helpers to review and refine the content as necessary. An illustrative example of a follow-up task is presented in Figure 17.

Refinement We construct the Refinement tasks by formulating two document-based NLP tasks within a dialogue instance. Each NLP task consists of six instructions of increasing complexity. These instructions are generated by prompting GPT-4 to produce five additional queries for each of the four NLP tasks: question-answering, summarization, named-entity recognition, and paragraph rewriting. We guide GPT-4 to impose an additional constraint on each successive query (the full prompt is available in Figure 6). We create four dialogue instances per document, with each of the four NLP tasks serving as the initial task, and a different task as the second task. This process results in 40 dialogue instances, corresponding to 480 dialogue turns. An example refinement task can be found in Figure 15.

For the single-turn scenario, we utilize GPT-4 to convert the multiple instructions into a single instruction using the prompt detailed in Figure 7.

Expansion The Expansion task challenges the model with a series of NLP tasks based on the initially provided document. Each instance consists of seven NLP tasks in random order, including question-answering, summarization, named-entity recognition, part-of-speech tagging, relation extraction, translation, and classification. Figure 16 illustrates an example of this task.

J Additional Statistics

We show the statistics of each task in Table 10.

Instruction	Description
startend:start_char	Start with a particular letter.
startend:start_emoji	Start with an emoji.
startend:end_phrase	End with a particular phrase.
language:response_language	Respond in a particular language.
format:json_format	Respond using JSON.
format:bullet_list	Using a specified number of bullet lists.
length_constraints:number_sentences	Respond with a specified sentence limit.
keywords:existence	Include some specific keywords.
change_case:capital_letter	Respond in uppercase.
change_case:lowercase	Respond in lowercase.
punctuation:no_comma	Refrain from using commas.
combination:two_responses	Include two different responses.
content:placeholder	Include a certain amount of placeholders.
format:constrained_response	Reply with one of the provided response options.

Table 9: The descriptions of the instructions used in the Recollection task. Most of it is adopted from Zhou et al. (2023).

K Additional Results

We show the full result containing GPT-4 in Table 11. We show the breakdown of the results in the Recollect task in Table 12.

L Inference Cost

We provide an estimate of the number of tokens in MT-Eval in Table 13 using Llama-2's tokenizer, and the cost estimates in Table 14.

M The Impact of Dialogue History in the Follow-up Task.

We conduct further study to explore the role of dialogue history as in-context examples and its impact on model performance in the Follow-up task. In this task, the user's query directly follows the assistant's previous reply. Modifying the dialogue history is not as straightforward as in the document classification task outlined in Section 4.6. To address this, we convert the multi-turn setting into a single-turn one, presenting only a complete query without prior responses. We employ GPT-4 to condense the relevant information from the dialogue history into a query for all 240 instances, using the prompt provided in Figure 8. However, the initial queries often omit crucial contextual information. Consequently, we meticulously review and refine each generated query as needed.

Table 16 presents the performance of various models in both single-turn and multi-turn settings. All models exhibit significantly better performance in the multi-turn setting, indicating the positive impact of dialogue history on model performance. To gain further insights, we analyze 30 instances and compare the differences in the two settings. We find that the models often leverage previous responses and explanations to generate improved responses in the current turn. This also suggests that high-quality dialogue history plays the role of in-context learning examples, providing useful guidance to the model's response. Figure 20 shows an example with Qwen-chat-7b, illustrating how the multi-turn response benefits from the dialogue history's intermediate reasoning, leading to an accurate current-turn response.

	Recollection	Expansion	Refinement	Follow-up	All
Avg. # Turns per Dialogue	10	7.00	12.00	3.00	6.96
Avg. # Words in Prompt †	693.09	539.60	882.85	686.82	760.41
Max. # Words in Prompt †	2331	838	2574	1932	2574
Avg. # Words in Response †	72.07	24.41	78.50	205.88	99.31
Max. # Words in Response †	289	107	430	444	444
Avg. # Words per Turn	54.49	156.77	65.89	31.78	60.63
Max. # Words per Turn	330	474	449	262	474
Total # Dialogues	38	10	40	80	168
Total # Turns	380	70	480	240	1170

Table 10: Detailed statistics for individual tasks in MT-Eval. †: Estimated using GPT-4 responses.

Model	Avg.	Recollection	Expansion	Refinement	Follow-up
GPT-3.5-Turbo	<u>7.72</u>	6.90	<u>7.87</u>	<u>6.92</u>	9.21
GPT-4	**9.03**	**9.61**	**9.07**	**7.85**	**9.60**
ChatGLM3-6B	5.49	2.92	5.90	4.73	8.39
Qwen-chat-7B	6.55	5.25	7.02	5.47	8.49
Vicuna-7B-v1.5	6.44	5.45	6.70	5.31	8.31
Llama-2-chat-7B	6.11	3.86	5.87	6.20	8.53
Mistral-Instruct-7B	7.46	<u>7.22</u>	6.98	6.58	9.05
Vicuna-13B-v1.5	7.01	6.27	6.70	6.37	8.68
Llama-2-chat-13B	6.31	3.66	6.37	6.37	8.82
Qwen-chat-14B	7.26	6.21	7.58	6.11	9.12
Mixtral-Instruct-8x7B	7.47	6.17	7.42	6.77	<u>9.52</u>

Table 11: Multi-turn performance in four dialogue tasks including GPT-4. The highest score in each column is highlighted in **bold**, while the second-highest score is <u>underlined</u>. We find that GPT-4 is biased toward its outputs.

Single-Turn Prompt:
Consider this two scenarios:
1. David has three sisters. Each of them has one brother.
2. David has three sisters. Each of them has two brother.

Can you explain how you determined the number of brothers David has in each scenario using logical reasoning?

Single-Turn Response: In scenario 1, David has three sisters, and each of them has one brother. Therefore, David has a total of 3 sisters * 1 brother/sister = 3 brothers.

In scenario 2, David has three sisters, and each of them has two brothers. Therefore, David has a total of 3 sisters * 2 brothers/sister = 6 brothers.

In both scenarios, the number of brothers David has is determined by multiplying the number of sisters he has by the number of brothers each sister has.

Multi-Turn Prompt:
User: David has three sisters. Each of them has one brother. How many brothers does David have?
Assistant: David has only one brother.
User: If we change the previous question and assume that each sister of David has two brothers, how many brothers would David have?
Assistant: If each of David's sisters has two brothers, then David would have two brothers.
User: I have some doubts regarding your first answer. If David's sisters have one brother, and David is also the brother of his sister, then how could David has one brother?
Assistant: I apologize for any confusion caused by my previous response. Let's clarify:

In the original scenario where David has three sisters and each of them has one brother, David is indeed that one brother. Therefore, David has no other brothers; he is the only brother to his three sisters.

In the modified scenario where each of David's sisters has two brothers, it implies that David has one additional brother besides himself. Therefore, in this case, David would have one brother.
User: Can you explain how you determined the number of brothers David has in each scenario using logical reasoning?

Multi-Turn Response:
Sure, I'd be happy to explain my reasoning.

In the original scenario where David has three sisters and each of them has one brother, we can use logical reasoning to determine that David has no other brothers. Here's how:

1. David has three sisters.
2. Each of David's sisters has one brother.
3. Therefore, David is the only brother to his three sisters.

In the modified scenario where each of David's sisters has two brothers, we can use logical reasoning to determine that David has one additional brother besides himself. Here's how:

1. David has three sisters.
2. Each of David's sisters has two brothers.
3. Therefore, David has one additional brother besides himself.
4. Since David has one additional brother, he has a total of two brothers.

I hope this helps clarify my reasoning. Let me know if you have any further questions.

Figure 20: Comparison of the responses in single-turn and multi-turn scenarios of an instance in the Follow-up task.

Model	Global Inst.		CLS	
	Single-Turn	Multi-Turn	Single-Turn	Multi-Turn
GPT-3.5-Turbo	8.00	4.61	**9.50**	9.20
GPT-4	**9.75**	**9.71**	9.30	**9.50**
ChatGLM3-6B	3.61	2.61	6.50	3.24
Vicuna-7B-v1.5	4.61	3.11	8.10	7.80
Llama-2-chat-7B	6.43	2.32	8.10	5.40
Qwen-chat-7B	6.04	2.89	8.30	7.60
Mistral-Instruct-7B	7.54	5.25	9.40	9.19
Vicuna-13B-v1.5	6.46	4.64	7.50	7.90
Llama-2-chat-13B	7.32	2.93	7.70	4.40
Qwen-chat-14B	7.39	4.43	9.40	8.00
Mixtral-Instruct-8x7B	6.93	3.21	8.80	9.13

Table 12: Breakdown of results in the Recollection task. Global Inst. stands for the global instruction following task, and CLS denotes the document classification task.

	Avg. # Input Tokens	Avg. # Output Tokens
Inference	1,850,000	230,000
Evaluation	400,000	80,000

Table 13: Average number of input and output tokens in MT-Eval during inference and evaluation.

	Model	Input Cost	Output Cost	Total Cost
Inference	GPT-3.5-Turbo	0.93	0.35	1.28
	GPT-4-Turbo	18.50	6.90	25.40
Evaluation	GPT-4-Turbo	4.00	2.40	6.40

Table 14: Average inference and evaluation cost (USD) for closed-source models.

Instruction	GPT-4	Vicuna-13B-v1.5	Qwen-chat-14B	Mistral-Instruct-7B	Mixtral-Instruct-8x7B
change_case:capital_letter	10.00	1.50	0.50	3.50	0.50
change_case:lowercase	10.00	0.00	2.00	2.50	1.50
combination:two_responses	10.00	0.00	4.50	0.00	3.50
content:placeholder	5.00	1.50	1.50	5.50	1.50
format:bullet_list	9.00	6.00	6.00	2.00	6.00
format:constrained_response	10.00	10.00	10.00	5.00	5.00
format:json_format	10.00	8.00	8.00	8.00	3.00
keywords:existence	10.00	0.00	0.00	2.00	1.50
language:response_language	10.00	0.00	9.50	9.50	2.50
length_constraints:number_sentences	6.50	5.00	3.00	0.00	0.00
punctuation:no_comma	10.00	0.00	0.00	0.00	0.00
startend:end_phrase	6.00	9.50	6.00	10.00	6.00
startend:start_char	10.00	5.00	5.00	1.50	0.00
startend:start_emoji	10.00	7.00	0.50	6.50	0.00

Table 15: The number of turns that different models can adhere to the global instructions in the Recollection task, averaged over two dialogues per instruction.

Model	Single-Turn Avg.	Multi-Turn Avg.
GPT-3.5-Turbo	9.19	9.21
GPT-4	**9.24**	**9.60**
ChatGLM3-6B	7.60	8.39
Vicuna-7B-v1.5	7.88	8.31
Llama-2-chat-7B	7.97	8.53
Qwen-chat-7B	7.98	8.49
Mistral-Instruct-7B	8.78	9.05
Vicuna-13B-v1.5	8.37	8.68
Llama-2-chat-13B	8.30	8.82
Qwen-chat-14B	8.60	9.12
Mixtral-Instruct-8x7B	9.02	9.52

Table 16: Performance of single-turn and multi-turn setting in the Follow-up task. The result of multi-turn is exacted from Table 2.