BotChat: Evaluating LLMs' Capabilities of Having Multi-Turn Dialogues

Anonymous ACL submission

Abstract

In the realm of modern Large Language Models (LLMs), facilitating high-quality, multi-turn dialogues with humans represents a cornerstone feature. However, human-based evaluation of such capability involves substantial manual effort. This study offers a formative assessment of current LLMs' proficiency in emulating human-like, multi-turn conversations, employing an LLM-based methodology. The evaluation encompasses three key elements in the evaluation pipeline: **utterance generation**, **evaluation protocol**, and **judgement**, and we delve deeply into each aspect. GPT-4, both as an utterance generator and a judge, exhibits exceptional performance. As a generator, GPT-4 crafts dialogues indistinguishable from human interactions in terms of style and flow. When judging, it shows a heightened alignment with human evaluative standards and consistency. Conversely, other LLMs face challenges in producing quality multi-turn dialogues, hindered by inadequate instruction-following abilities, a propensity for prolix utterances, and overall limited capabilities. Notably, generating extensive dialogues (*e.g.*, spanning tens of turns) remains a formidable task for most LLMs, particularly in Chinese contexts. We hope our work can serve as a valuable resource for evaluating multi-turn chatting capabilities of LLMs. This research aims to contribute a robust framework for assessing the multi-turn conversation abilities of LLMs, hoping to guide future advancements in this domain.

1 Introduction

The evolution of Large Language Models (LLMs) (OpenAI, 2023; Touvron et al., 2023a; Chiang et al., 2023) marks a transformative phase in artificial intelligence, significantly surpassing traditional language models (Devlin et al., 2018; Vaswani et al., 2017; Liu et al., 2019) in engaging in nuanced, multi-turn dialogues with humans. Modern LLMs interact with people with human-style multi-turn conversations, learn the instruction, intention, and context from human prompts to provide helpful feedback. Such advantage enables all human kinds to directly access the strong capability of LLMs for various applications, either general (Jiao et al., 2023; Shen et al., 2023) or within specific domains (Bran et al., 2023; Boiko et al., 2023).

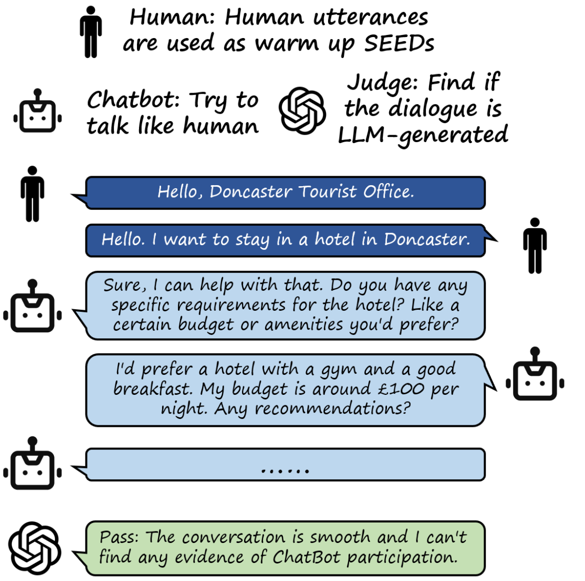

Human: Human utterances are used as warm up SEEDs
Chatbot: Try to talk like human
Judge: Find if the dialogue is LLM-generated
Hello, Doncaster Tourist Office.
Hello. I want to stay in a hotel in Doncaster.
Sure, I can help with that. Do you have any specific requirements for the hotel? Like a certain budget or amenities you'd prefer?
I'd prefer a hotel with a gym and a good breakfast. My budget is around £100 per night. Any recommendations?
......
Pass: The conversation is smooth and I can't find any evidence of ChatBot participation.

Figure 1: **BotChat** evaluates multi-turn dialogue performance by prompting ChatBots to generate dialogues from initial human utterances, followed by evaluation through a judge LLM.

Despite their advanced capabilities, not all LLMs consistently deliver satisfactory performance in multi-turn human interactions. In practical applications, it has been observed that dialogues generated by certain LLMs frequently fail to meet user satisfaction criteria. The issues manifest in multiple aspects, including poor adherence to user instructions, undesirable tone or lengthy utterances, and the generation of repetitive content. The evaluation of these conversational capabilities remains a complex challenge. Traditional approaches, primarily human-based (Zheng et al., 2023), intensively involve manual labor for human-bot conversation generation and quality assessment. This paper proposes a more efficient paradigm, named **BotChat**, to evaluate the multi-turn chatting capability.

BotChat is an entirely LLM-based approach, eliminating the need for manual labor. The methodology comprises two stages: dialogue generation and quality assessment. Initially, we use the very first utterances (*ChatSEED*) from multilingual real-world conversations (Cui et al., 2020; Wang et al., 2021) for utterance-by-utterance dialogue generation by ChatBots. In each step, a ChatBot generates one utterance based on all history utterances. This process autonomously generates dialogues of varying number of turns. In the second stage, we assess the dialogues using different judge LLMs and a suite of LLM-based evaluation protocols. Our experiments demonstrate that GPT-4 excels in human alignment and self-consistency compared to other LLMs. We introduce three evaluation protocols: **UniEval** (individual dialogue evaluation), **PairEval** (comparative evaluation of two dialogues), and **GTEval** (comparison with a corresponding human dialogue). While UniEval and PairEval are applicable to dialogues of arbitrary turns, GTEval is limited by the ground-truth dialogue's extent. Additionally, addressing the unique challenges posed by repetitive utterances, which are common in Chinese conversational scenarios, we present **DupDetect** for preprocessing unnatural dialogue evaluations, thereby also reducing evaluation costs.

With the evaluation protocols, we compare 14 representative LLMs, ranging from the state-of-the-art closed-source GPT-4 (OpenAI, 2023) to small-scale open-source LLMs (Touvron et al., 2023b; Bai et al., 2023). During evaluation, three evaluation protocols draw substantially identical conclusions. GPT-4 generates human-style multi-turn conversations with impressive quality, outperforming all other LLMs. For all LLMs, the quality of generated dialogues declined quickly as long as the dialogue turns increase. Such degradation is particularly evident for open-source LLMs at small scale, compared to the top-tier LLM GPT-4. Notably, this phenomenon is more pronounced in Chinese context compared to the English one. With qualitative assessment, we find that LLMs fail to generate multi-turn conversations with desirable quality primarily due to: poor instruction-following capability, tendency to generate lengthy utterances, and limited general capability.

2 Related Works

2.1 Objective and Subjective Assessment of LLMs

Objective assessment of Large Language Models (LLMs) is pivotal for gauging their capabilities in a quantifiable and unbiased manner. This assessment typically involves contrasting the outputs of LLMs with established references or ground truths. For close-ended tasks (Huang et al., 2023; Hendrycks et al., 2020; Cobbe et al., 2021), the expectation is that the LLM outputs align perfectly with these ground truths. In contrast, open-ended tasks (Huang et al., 2021; Fabbri et al., 2019) rely on similarity metrics calculated between the LLM outputs and reference material, with higher similarity scores indicative of superior task performance. Metrics such as F1-score, BLEU (Papineni et al., 2002), and ROUGE (Lin, 2004) are commonly employed to quantify this performance.

Within the BotChat framework, the rationale is that conversations, even when initiated with the same ChatSEEDs, can diverge into myriad directions. Thus, an LLM-generated dialogue, though markedly different from its reference, should not be automatically deemed as inferior in quality. This deviation from the reference is inherent to the open-ended nature of the task, necessitating a more nuanced approach to quality assessment in BotChat. Subjective assessment has become a standard approach in evaluating Large Language Models (LLMs) for complex scenarios. Recent studies leverage human evaluators or other LLMs as judges to compare LLM performance (Xu et al., 2023; Chiang et al., 2023; Fu et al., 2023; Li et al., 2023; Wang et al., 2023a; Zheng et al., 2023).

2.2 Human Conversation Datasets

Constructing end-to-end chatbots has garnered significant attention within the NLP community, leading to the collection of diverse conversational datasets (Serban et al., 2015). Among these, PERSONA-CHAT (Zhang et al., 2018) is notable for its engaging dialogues exhibiting distinct personalities. The work of (Zhou et al., 2018) integrates specific documents into multi-turn conversations, enriching the conversational depth. CoQA (Reddy et al., 2019) offers a unique dataset for conversational question-answering, compiled from dialogues between two annotators discussing a passage. MuTual (Cui et al., 2020) features dialogues derived from Chinese student English listening comprehension exams, targeting the enhancement of conversational models' reasoning abilities. Additionally, NaturalCONV (Wang et al., 2021) encompasses human-like Chinese conversations in various domains such as sports, entertainment, and technology, complete with related reference materials. In BotChat, we primarily utilize MuTual and NaturalCONV as the main sources of human dialogues. This selection is driven by the richness and diversity of conversational contexts these datasets offer, enabling a comprehensive evaluation and development of advanced chatbot capabilities.

3 BotChat

In this section, we delineate the evaluation paradigms incorporated into BotChat. The primary aim of this framework is to evaluate LLMs' conversational abilities in alignment with human subjective preferences. We begin with a comprehensive overview of the workflow for generating multi-turn dialogues. Subsequently, we introduce three distinct evaluation strategies proposed: unitary evaluation (UniEval), pairwise evaluation (PairEval), and ground-truth evaluation (GTEval). Additionally, in response to the unique challenges of repetitive utterances (which frequently occur in Chinese conversational scenarios), we introduce DupDetect.

3.1 Dialogue Generation

In BotChat, we generate multi-turn dialogues exclusively based on ChatBots. This process initiates with authentic human conversations, from which we extract the initial few utterances, termed as ChatSEEDs, to serve as the basis for dialogue generation. We resort to the existing dataset MuTual (Cui et al., 2020) and NaturalConv (Wang et al., 2021) for real world human conversations. Specifically, the first two utterances from each dialogue in these datasets are employed as ChatSEEDs.

The dialogue generation progresses in an utterance-by-utterance manner. To guide the ChatBot in producing human-like, concise utterances, we introduce a system prompt[1] at each step. During each turn, this prompt, along with all preceding utterances in the dialogue, is provided to the ChatBot, which then generates the next utterance. This iterative process continues until the predetermined number of dialogue turns is achieved. The PseudoCode detailing our generation paradigm is delineated in Algorithm 1.

In the case of NaturalConv, each conversation is accompanied by a reference document. To explore the impact of such contextual information, we implement both unconditional (UNCON) and conditional (CON) settings in our generation process. Under the CON setting, the LLMs are additionally provided with the relevant reference document to ascertain whether its presence eases or complicates the dialogue generation task.

Algorithm 1: Dialogue Generation.

```
Data: ChatSEED $\mathbf{s}$ (a list of two utterances); target
      number of rounds $\mathbf{N}$; system prompt $\mathbf{SYS}$;
      ChatBot $\mathbf{M}$
Result: Generated Dialogue $\mathbf{D}$ (a list of utterances)
$\mathbf{D} \leftarrow \mathbf{s}$;
$\mathbf{T} \leftarrow len(\mathbf{D})$;
while $\mathbf{T} < \mathbf{N}$ do
    $\mathbf{History} \leftarrow$ build_history($\mathbf{SYS}$, $\mathbf{D}[:-1]$);
    $\mathbf{Utterance} \leftarrow \mathbf{M}$.chat($\mathbf{D}[-1]$, $\mathbf{History}$);
    $\mathbf{D}$.append($\mathbf{Utterance}$);
    $\mathbf{T} \leftarrow len(\mathbf{D})$;
end
```

3.2 Evaluation Strategies

DupDetect. During the process of dialogue generation, we encountered a recurring issue with existing Large Language Models (LLMs), particularly in the Chinese context: they frequently enter infinite loops during self-dialogue. This phenomenon significantly diminishes the naturalness of the conversation. To mitigate this, we have developed a pre-processing technique, termed DupDetect, specifically designed to identify and filter out these looped conversations.

DupDetect operates by analyzing dialogues designated for pairwise comparison. It calculates the similarity between the i-th utterance and the subsequent $i+1/i+2$ utterance ($2 < i < MaxRound-1$). Upon detecting that the similarity surpasses a pre-set threshold, the dialogue is flagged as having entered an infinite loop. The point at which this loop commences is noted, and all dialogue up to that point (including the utterance that triggered DupDetect) is considered non-looping. For subsequent pairwise evaluations within this paper, including PairEval and GTEval, DupDetect is employed as a preliminary step. With DupDetect equipped, the specific evaluation criteria are as follows:

1. If both dialogues enter an infinite loop, the outcome is classified as a *Tie*.
2. Should one dialogue fall into an infinite loop

[1] The details of the prompt are illustrated in Appendix B.

while the other maintains a non-repetitive structure, the latter is deemed to *Win*.

3. In scenarios where neither dialogue exhibits looping, we proceed with additional evaluation steps such as GTEval or PairEval.

UniEval. One effective method for evaluating dialogue quality involves independently assessing each generated conversation, focusing on its resemblance to human dialogues. Our evaluation process unfolds through the following steps:

1. Initially, the judge LLM is tasked with determining whether the given dialogue appears to be ChatBot participated (Y/N).
2. If the judge LLM answers "Yes", it is then prompted to pinpoint the index of the first utterance it identifies as ChatBot-generated. Conversely, no additional probing is required.
3. Ultimately, the LLM judge is required to articulate a rationale for its decision, providing critical insights into the model's evaluative reasoning.

To augment GPT-4's instruction-following capabilities, we have also developed several in-context examples. These will be integrated into the evaluation prompt[1], thereby enhancing the robustness of the evaluation procedure.

PairEval. While UniEval has yielded preliminary insights, its inherent limitations must be acknowledged. Despite providing detailed evaluation guidance and contextual examples to GPT-4 evaluators, establishing a clear-cut criterion to differentiate human dialogues from those generated by LLMs poses a significant challenge.

An alternative, widely embraced benchmarking approach for LLMs involves comparative evaluation. This method, often utilizing human judges or GPT-4 as evaluators, contrasts responses from two different models presented with identical prompts. A prominent example of this approach is the Chatbot Arena (Zheng et al., 2023), where users engage with two separate LLM instances using the same message or question. Users then evaluate and select the more preferable response from the two. The overall performance of each LLM is quantified using an ELO rating system (Elo, 1967), aggregated from diverse user feedback.

Building on this concept, we introduce an additional strategy, termed PairEval, within our evaluation framework. In PairEval, a judge LLM is tasked with comparing two dialogues to discern whether they are ChatBot-generated. To manage evaluation costs effectively, we fix GPT-4 as the reference model in each comparison pair ($\mathbf{O}(n)$) instead of exhaustive pairwise comparisons across dialogues generated by all LLMs ($\mathbf{O}(n^2)$). While being cost effective, the reference-fixed evaluation also ensures reliable evaluation outcomes compared to the dense pairwise comparison.

GTEval. GTEval forms an integral part of our evaluation framework, involving a detailed comparison between the generated conversations and the 'Ground Truth' conversations from the conversational datasets. We employ a protocol akin to that used in PairEval to facilitate this evaluation. GTEval is instrumental in rigorously gauging how closely language models emulate real human interactions, utilizing the rich resources of human dialogues available in the dataset.

GTEval necessitates that GT conversations meet a minimum threshold of dialogue turns, denoted as $\mathbf{N}$. For MuTual-Test, to facilitate this comparison, we selected a subset of 222 conversations, with each conversation contains at least $\mathbf{N} = 4$ utterances (the specific distribution of conversation turns demonstrated in Figure 5). Acknowledging the variability in the length of GT conversations, we standardize the comparison process by truncating all generated dialogues. The meta prompt deployed in GTEval is largely similar to that used in PairEval, with a crucial distinction. In GTEval, it is explicitly mentioned that among the two dialogues being compared, only one contains utterances generated by an LLM.

4 Experiments

4.1 Dialogue generation

LLMs for Evaluation. Unless specified, we adopt the **'chat' variant** for all Open-Source LLMs. We include the following LLMs in our study: GPT-3.5-Turbo (0613 ver.), GPT-4 (0613 ver.) (OpenAI, 2023), Claude-2, ChatGLM3-6B (Zeng et al., 2022), Baichuan2-13B (Baichuan, 2023), Qwen-[7B/14B] (Bai et al., 2023), LLaMA2-[7B/13B/70B] (Touvron et al., 2023b), InternLM-[7B/20B] (Team, 2023), Vicuna-[7B/13B] (v1.5) (Zheng et al., 2023). In experiments, we configure closed-source LLMs and LLaMA2 with the temperature setting to 0. For other open-source LLMs (all with HuggingFace implementations), we adopt the default hyper-parameters for utterance inference.

The Generation Procedure. We extract ChatSEEDs from MuTual and NaturalConv for dialogue generation. MuTual-Test comprises 547 distinct dialogues. We retained the first two utterances of each dialogue, resulting in 547 ChatSEEDs. In NaturalCONV, we choose 160 evenly distributed instances across six domains and examine two settings: **CON** (conditional) and **UNCON** (unconditional) in dialogue generation. We set the round $\mathbf{N} = 16$ (including the initial two utterances) throughout dialogue generation. The context window sizes can vary for different LLMs, ranging from 2,048 (Qwen, InternLM-7B, *etc.*) to 100,000 (Claude-2). During dialogue generation, all history utterances may not fit into the context window in some circumstances, In such case, we keep dropping the oldest utterance until the overall token length is below the threshold. All 14 LLMs are adopted for generating English dialogues based on ChatSEEDs in MuTual. For Chinese dialogues, considering the generally lower performance, we specifically choose eight models that are more powerful variants and with Chinese capability. Open-source LLMs were inferred using A100 80G GPUs, totaling around 60 GPU-hours.

Length Statistics of Generated Utterances. Our preliminary analysis focuses on measuring the length of utterances generated by various LLMs and providing statistical insights. For each generated utterance, we employ the CL100K tokenizer (the one used by OpenAI ChatGPT) for tokenization and calculate the number of tokens. Figure 2 illustrates the distribution of token lengths in utterances generated by different models. Most LLMs produce utterances with varying token lengths, ranging from just a few tokens to several thousands. An interesting outlier is GPT-4, which consistently generates relatively short utterances, with the longest utterance being fewer than 100 tokens. In Table 5, we present the average utterance length generated by different models. Notably, most models tend to produce relatively short utterances on average, with the exceptions being GPT-3.5, Claude-2, and LLaMA2. The statistics of the Chinese dataset follows a similar trend. For detailed information, please refer to Figure 6.

4.2 Evaluation Results on MuTual

Unless specified, we adopt GPT-4-0613 (OpenAI, 2023) as the LLM judge across all experiments.

UniEval. In UniEval, we evaluate all $547 \times 14 = 7658$ generated dialogues with the above-mentioned strategy and present the results. Figure 3 illustrates the success rates ("Not LLM participated" determined by the LLM judge) under different target $\mathbf{N}$. The models are sorted in the descending order of success rates at $\mathbf{N} = 16$. By definition, a dialogue pass @$\mathbf{N}$ either if the LLM judge determines that the entire dialogue is not ChatBot generated or if it determines that the indice of the first ChatBot generated utterance is larger than $\mathbf{N}$. Here we summarize our major findings:

1. **Exceptional Multi-Turn Chatting Performance of GPT-4:** GPT-4 demonstrates extraordinary capabilities in generating lengthy conversations. It achieves the highest success rate for every target turn $\mathbf{N}$. Under $\mathbf{N} = 16$, GPT-4 demonstrates a remarkable success rate of over 65%, while the 2_{nd} best Vicuna-13B and the 3_{rd} best InternLM-20B achieve only 55% and 36%.
2. **Satisfying Performance of Open-Source LLMs on Short Conversations:** Some open-source large language models (LLMs), such as InternLM, Qwen, and Baichuan2, exhibit strong performance in generating short dialogues ($\mathbf{N} = 4$ or $\mathbf{N} = 8$). However, as dialogue turns increase to $\mathbf{N} = 16$, their performance rapidly deteriorate, and significantly fall behind state-of-the-art ChatBots like GPT-4-0613.
3. **The Multi-Turn Chatting Capability Scales with the Model Size:** Not surprisingly, we find that the multi-turn chatting capability scales with the model size, especially for a large turn number. For example, under the track $\mathbf{N} = 16$, InternLM-20B outperforms InternLM-7B by 29% success rate, while Vicuna-13B outperforms Vicuna-7B by 25%. Such gap is much smaller when $\mathbf{N}$ is small. For $\mathbf{N} = 4$ (only 2 utterances are generated), the gap for two InternLM variants is merely 1.5% success rate.
4. **Unique Behavior of Claude-2:** Among closed-source LLMs, Claude-2 stands out with the lowest performance. It strongly tends to act like an AI assistant, generating relatively lengthy content. Consequently, it performs poorly when tasked with generating human-like utterances, which are typically shorter and less structured.

PairEval. PairEval is conducted on the 222 ChatSEED subset of MuTual-Test. For dialogues generated with each ChatSEED, we pair them with GPT-4 generated dialogues and evaluate with the LLM judge. For each dialogue pair, we conduct

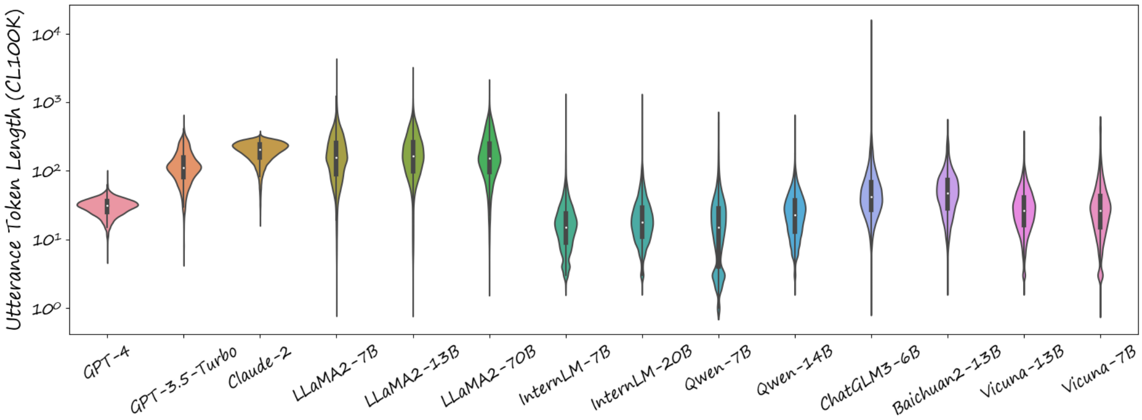


Figure 2: **The length distribution of utterances generated by different LLMs, in a violin plot.**

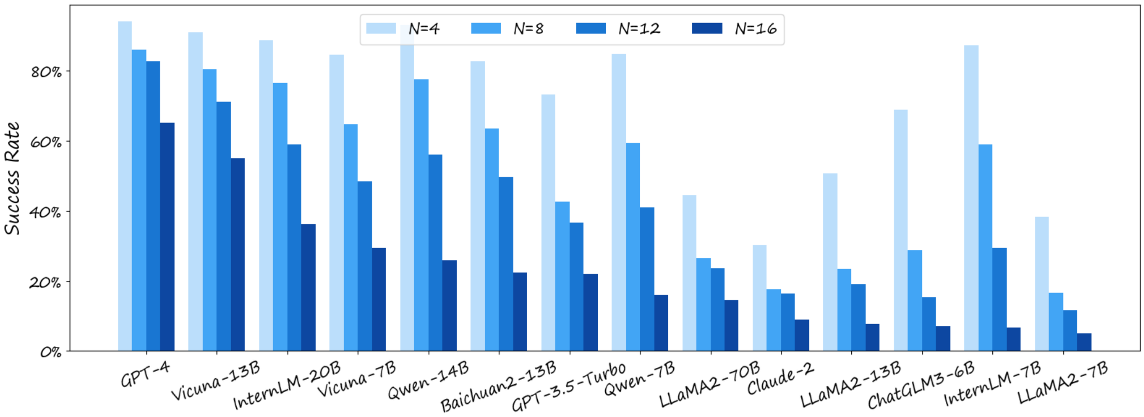


Figure 3: **The UniEval pass rate of different LLMs when generating a dialogue with N utterances.**

bi-directional comparisons and include both results when calculating the evaluation metrics. This approach ensures a more robust and comprehensive assessment.

In Figure 4(b), we present the win / tie / lose rate of different LLMs. Remarkably, Vicuna-13B attains nearly 80% of the GPT-4's proficiency level. Conversely, the performance of GPT-3.5-Turbo and Claude-2 lags behind many open-source LLMs. This can be attributed in part to their limited instruction-following capabilities and a strong inclination to act as an AI assistant by providing lengthy and comprehensive responses. Among open-source LLMs, Vicuna, Qwen-14B, and InternLM-20B demonstrate strong capability in generating human-style dialogues, significantly outperform LLaMA2 family models. However, Qwen-7B and InternLM-7B present a poor showcase due to their high repetition rate in 16-round conversations.

GTEval. In each Large Language Model (LLM) vs. Ground Truth (GT) comparison, an LLM is considered the winner if the evaluator determines the GT dialogue is more likely to be a ChatBot generated one. In Figure 4(c), we present the win / tie / lose rate of different LLMs (sorted in the descending order of Win+Tie Rate).

In GTEval, a GT dialogue only has 7.4 utterances on average, thus the advantage of GPT-4 can be less significant. We adopt the win+tie rate against GT dialogues as the major metric to measure the multi-turn chatting performance. GPT-4 demonstrates top performance in dialogue generation. With the same dialogue rounds, the evaluator can hardly tell the difference between GPT-4 generated dialogues and GT dialogues (the win rate of GPT-4 is 25.7%, while the lose rate is merely 29.0%). Furthermore, due to the reduced conversation length, Vicuna-13B, Qwen-14B and InternLM-20B also demonstrate strong performance, very close to the top performing GPT-4. We also notice that, though some closed-source ChatBots (GPT-3.5-Turbo, Claude-2, *etc.*) suffer from lengthy and AI-assistant style responses, they achieve top win rates across all LLMs.

We also examine the UniEval success rate for each dialogue at the GT trimmed length, to see

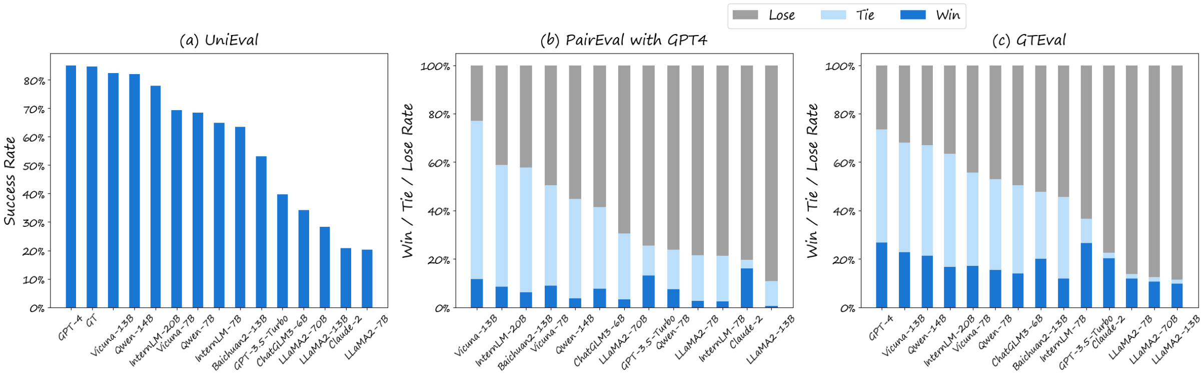


Figure 4: **Comprehensive Experimental Results of Three BotChat Evaluation Protocols on MuTual.**

	N=16		
Model	MuTual	CON	UNCON
Qwen-14B	53.7	11.9	13.8
InternLM-20B	66.5	19.4	40.6
Baichuan2-13B	76.1	53.6	56.3
ChatGLM3-6B	78.4	35.0	53.8
Vicuna-13B	85.6	40.0	51.3
Claude-2	85.9	34.4	54.4
GPT-3.5-Turbo	86.3	56.3	46.9
GPT-4	95.2	76.9	91.9

Table 1: **Statistics of non-loop rate.**

if the same conclusion can be drawn with different evaluation strategies. The results are visualized in Figure 4(a). In both of these figures, the top-performing LLMs (GPT-4, Vicuna-13B, Qwen-14B, InternLM-20B, *etc.*) maintain the same ranking. However, LLMs with inferior performance display some slight difference in two groups of rankings.

4.3 Evaluation Results on NaturalConv

More Repetitions Detected. In experiments involving Chinese dialogues, we observed notable shifts in the results. A key challenge identified was the tendency of chatbot dialogues to increasingly fall into repetitive patterns or 'dead loops' as the number of conversational rounds grew. We report the non-loop rate of different models when the conversation turn reaches N=16, as detailed in Table 1.[2]

Among the models evaluated, GPT-4 distinguished itself with a remarkably low incidence of dead loops, outperforming its counterparts by a substantial margin. In contrast, models like Qwen-14B and InternLM-20B demonstrated a higher propensity for falling into dead loops during self-dialogue. This significantly affects their rankings on CON/UNCON.

The probability of English conversations experiencing dead loops is significantly lower than that in Chinese conversations, highlighting a discernible gap in the model's conversational abilities between Chinese and English. Interestingly, we noticed that the likelihood of encountering dead loops diminished significantly in the UNCON setting compared to the CON setting. This suggests that the inclusion of input documents in the CON setting might inadvertently constrain the diversity and richness of self-dialogues.

Evaluation Results. We utilized DupDetect to evaluate the performance of various models under both settings. MeanWhile, we include how these models performed on MuTual. After deduplication, we concurrently conducted PairEval and GTEval. In Table 2, we depict the win+tie rate of various LLMs in Mutual and NaturalConv (CON & UNCON settings).

GPT-4 stands out as particularly powerful in Both English and Chinese multi-conversation, showcasing its strength as an all-around player. Furthermore, Vicuna rightfully earns the recognition as the most closely aligned open-source model to GPT-4.

When compared with MuTual results, the performance on NaturalConv is generally inferior. It is evident that the performance trends of different models under CON and UNCON settings are inconsistent. This suggests varying sensitivity to input reference documents.

In the CON setting, GPT-3.5's ranking has noticeably increased. This shift might be attributed to other models being disrupted by the document input, increasing the likelihood of encountering dead

[2] Dialogues with no repetition are marked as 16 non-loop turns.

Setting	Compared w.	GPT-4	Vicuna-13B	Internlm-20B	Baichuan2-13B	ChatGLM3-6B	Qwen-14B	Claude2	GPT-3.5
MuTual	GT	**71.0**	62.4	50.4	40.0	41.9	38.7	20.8	32.9
	GPT4	-	**77.0**	58.8	57.7	41.5	44.8	19.6	25.5
CON	GT	**31.9**	8.7	7.4	2.4	2.5	4.4	1.8	13.7
	GPT4	-	42.5	30.0	40.0	35.7	28.7	36.2	**57.4**
UNCON	GT	**66.8**	21.3	20.0	11.2	8.1	6.9	5.0	3.7
	GPT4	-	36.2	35.6	34.4	24.4	16.3	**39.9**	23.1

Table 2: **Win+Tie Rate compared with GT / GPT-4.** **Bold** denotes the best result, Underline denotes the 2_{nd} best.

loops. However, GPT-3.5 maintains its proficiency in rich multi-turn dialogues.

4.4 Judge LLM Performance

We conducted a comprehensive analysis of various models used as Judge LLM, including widely used proprietary models Claude2 and GPT-3.5, exceptional open-source models Qwen-14B and Vicuna-13B-16K, as well as PandaLM (Wang et al., 2023b), a model specifically designed for judging. We carefully chose a diverse and challenging subset covering scenarios in both Chinese and English (**dialogues with loops excluded**). This subset was then distributed to human annotators, tasking them with an annotating job. Participants were recruited via a crowd-sourcing platform and received fair compensation through payment. The goal was to gauge how well LLM's outputs align with the subjective preferences of humans. The metrics considered include: 1. **CwGPT4**: Consistency rate with GPT-4 Evaluation. 2. **CwHuman**: Consistency rate with Human annotators, serving as the gold standard.

We report the evaluation results in Table 3. GPT-4 achieved a consistency rate of 65.74% with humans. This is comparable to the results of the previous MT-bench (Zheng et al., 2023) study (66%). The key difference lies in the fact that we tasked GPT-4 with assessing N-turn conversations, a significantly greater challenge compared to MT-bench, which evaluates only two turns. Other models show significant gaps in alignment rates compared to GPT-4. We also report the distribution of choices made by different judges in Table 4, with GPT-4 exhibiting a more human-like distribution of options. In contrast, most Judge LLMs tend to select *Tie*, demonstrating weak performance in multi-turn dialogue evaluation.

	CwGPT4	**CwHuman**
GPT-4	-	**65.74**
GPT-3.5-Turbo	**58.30**	41.06
Claude-2	41.28	38.51
Vicuna-13B-16K	42.77	35.17
PandaLM	43.40	34.04
Qwen-14B	39.15	33.20

Table 3: **Performance for Different Judge LLM.**

	Win	Tie	Lose
Human	35.11	34.26	30.64
GPT-4	28.30	43.62	28.09
GPT-3.5-Turbo	10.21	71.06	18.72
Claude-2	36.60	44.26	18.94
Vicuna-13B-16K	1.06	88.30	10.64
PandaLM	0.43	97.66	1.91
Qwen-14B	1.91	88.72	8.72

Table 4: **Choice Distribution of Different Judges.**

5 Conclusion

In this paper, we design a proxy evaluation paradigm BotChat to measure the multi-turn conversational capabilities of large language models. BotChat evaluate ChatBot generated dialogues with an LLM judge, to emancipate heavy human labor from the evaluation. We design multiple evaluation protocols and adopt them to evaluate dialogues generated by 14 modern LLMs. We find that a large proportion of LLMs excel at having dialogues of limited turns. However, when the turn number is large, only a few LLMs (GPT-4, Vicuna-v1.5-13B, *etc.*) achieve satisfying performance. We hope that BotChat can serve as a valuable resource on the journey towards automated evaluation of multi-turn conversational capability.

6 Limitations

The principal limitation inherent in BotChat resides in its evaluation methodology, which is heavily reliant on the seamless integration and utilization of the GPT-4 API. The absence or unavailability of this pivotal resource poses a significant impediment, rendering the evaluation process unattainable and consequently impeding the system's overall functionality.

Furthermore, it is noteworthy that GTeval, an integral component of the assessment framework, requires access to Ground Truth (GT) dialogues. This requisite could potentially introduce constraints on the applicability of BotChat, particularly in scenarios where obtaining or utilizing GT dialogues may prove challenging or impractical.

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.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Daniil A Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*.

Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. *arXiv preprint arXiv:2004.04494*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Arpad E Elo. 1967. The proposed uscf rating system, its development, theory, and applications. *Chess Life*, 22(8):242–247.

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. *arXiv preprint arXiv:2104.02112*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. *GitHub repository*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21)*.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2023b. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

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*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358*.

LLM	Avg. #Tokens	LLM	Avg. #Tokens
GPT-4	30.5	GPT-3.5-Turbo	124.9
Claude-2	197.3	Baichuan2-13B	58.0
InternLM-7B	20.1	InternLM-20B	24.4
Qwen-7B	20.7	Qwen-14B	28.7
ChatGLM3-6B	58.7	LLaMA2-7B	191.0
LLaMA2-13B	199.0	LLaMA2-70B	193.7
Vicuna-7B	37.5	Vicuna-13B	32.0

Table 5: **Average token numbers for utterances generated by different LLMs (MuTual Test).**

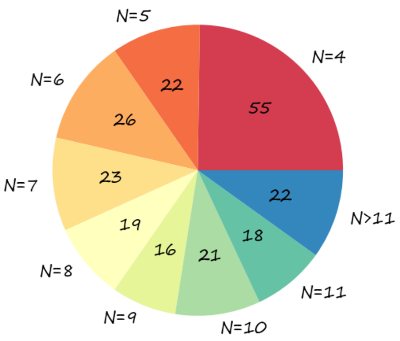


Figure 5: **Distribution of dialogue turns in MuTual test.**

A Some Additional Tables and Visual Results

This section includes some additional tables and visual results to further support our research findings. These data provide a more detailed explanation and analysis of the experimental results.

B Prompts adopted in BotChat

We use the following system prompt for all LLMs during **Dialogue Generation**, which aims at guiding the LLM towards crafting concise, natural, and seamless conversations.

C Dense PairEval

In Figure 7, we present comprehensive experimental results for PairEval, including the win rates for one-on-one matchups among all LLM pairs.

D Qualitative Results

We conduct qualitative analysis and categorize bad cases into five distinct types. We also sample a good case which is just like natural, relatable, and adaptive human talks. We first illustrate five distinct types of bad cases in Figure 8.

AI Self-Identification. In this situation, the models simply fail to pretend to be human and expose themselves as AI assistants. In the example, Speaker A's response begins with an explicit disclosure of the AI's nature, making it clear that it's not a human conversation.

Contextual Confusion. This type involves responses that fail to understand the context or meaning of the conversation, resulting in irrelevant or meaningless replies. The example shows that the AI fails to recognize it's a conversation between a recently hailed taxi customer and a driver. Towards the end, it generates unrelated and irrelevant responses, disconnecting from the context and intended meaning.

Excessive Length. The responses are overly lengthy, revealing the AI Assistant's nature, where both Speaker A and Speaker B engage in detailed exchanges that are atypical of human conversations, which raises suspicion.

Formal Tone. Sometimes, the AI's responses are organized with overly formal language, lacking the natural flow and tone of human conversation. In the example, the initial ChatSEED in this conversation is a casual and everyday discussion about washing dishes. However, as the conversation progresses, it takes a sudden shift towards a more formal and detailed discussion, delving into specific cleaning methods. This transition can make the conversation unnatural because people typically do not abruptly switch from general topics to detailed discussions about dish-washing techniques in everyday conversation.

Repetitive Phrasing. In the related example, it's comical that the model repeatedly use the same phrases or responses rely on generic or unrelated replies to sustain the conversation, lacking creativity. It is always caused by "I'm glad" or "You're welcome".

Good Case. In Figure 9 we show a good case of speaking like a human for AI means natural, relatable, and adaptive conversation. It avoids sounding robotic, uses colloquial language, and provides helpful responses to both simple and complex queries.

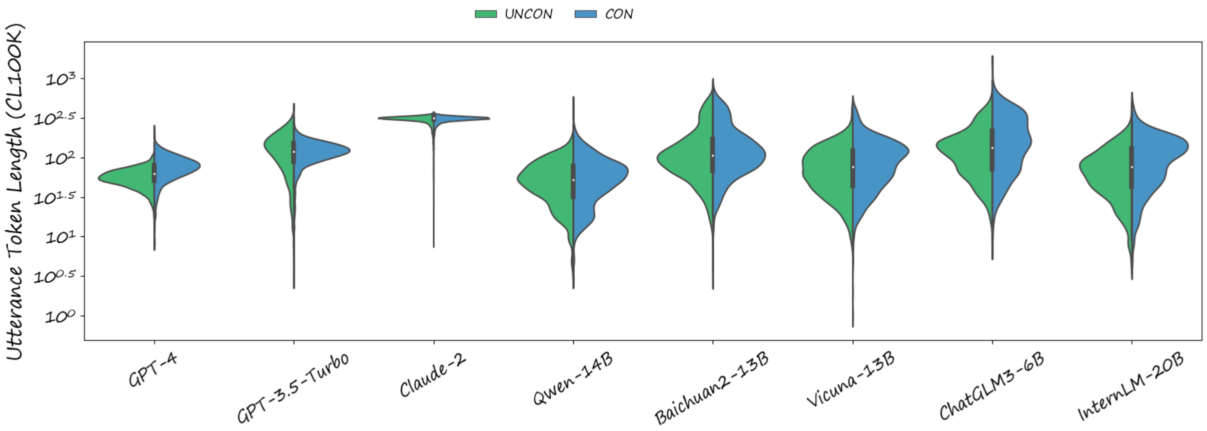


Figure 6: **The length distribution of NCONVNaturl, in a violin plot.**

System prompt used in Dialogue Generation

Prompt:
You are an AI who is having a conversation with human. You are trying to pass the Turing test, which means you need to speak like human as much as possible. In the conversation, you need to talk like human, and the conversation will be at least 5 rounds (it can be even longer). The conversation flow should be natural and smooth. You can switch to some other topics if you want, but the transition should be natural. Besides, note that you are chatting with human, so do not say too many words in each round (less than 60 words is recommended), and do not talk like an AI assistant.

System prompt used in Document-Conditioned Dialogue Generation

Prompt:
. . .
Ensure to mention the content of reference documents, without restricting the amount and manner of mentioning. You can smoothly transition to different topics, including those unrelated to the main subject. Keep the conversation natural and fluid. Here are the reference documents.
Title: xxx Text: xxxxx

System prompt used in **UniEval**

Prompt:
You are an AI assistant who helps human do the Turing test more easily. You will be provided with a conversation, and you need to judge if the conversation is AI involved. Print "Choice: No" if you think the conversation is not AI involved, or print "Choice: Yes" if you think it is AI involved.
If you print "Choice: Yes", you need also print a number (start from 1, use the format “Index: n" [$1 \leq n$ $\leq$ # utterances]) in the new line, indicating the index of the first chat that you think is generated by AI. One chat starts with ’A: ’ or ’B: ’, and ends with <chat_end>. One chat can be AI generated if (including but not limited to): 1. the sentence is not fluent; 2. the sentence is too long and is not likely to appear in human-human conversations; 3. the sentence is not related to the context or the change of topic is not natural; 4. the sentence just repeat some previous sentences (exact repetition or with the same meaning).
You also need to provide your reason for your choice.
Your response should use the following format:
Choice: No Index: None Reason: BlahBlah or
Choice: Yes Index: n Reason: BlahBlah

System prompt used in **PairEval**

Prompt:

You are an AI assistant who helps human do the Turing test more easily. You will be provided with two conversations, and there can be AI-generated utterance in each conversation. You need to read both conversations and judge if two conversations are AI involved.

If you think only Conversation 1 is AI involved, include “Choice: Conversation 1" in your response.

If you think only Conversation 2 is AI involved, include “Choice: Conversation 2" in your response.

If you think both conversations are likely to be with AI involved, include “Choice: Both" in your response.

If you think no conversation is likely to be with AI involved, include “Choice: Neither" in your response.

You also need to provide your reason for your choice.

Your response should use the following format:

“Choice: Conversation 1; Reason: BlahBlah" or

“Choice: Conversation 2; Reason: BlahBlah" or

“Choice: Both; Reason: BlahBlah" or

“Choice: Neither; Reason: BlahBlah"

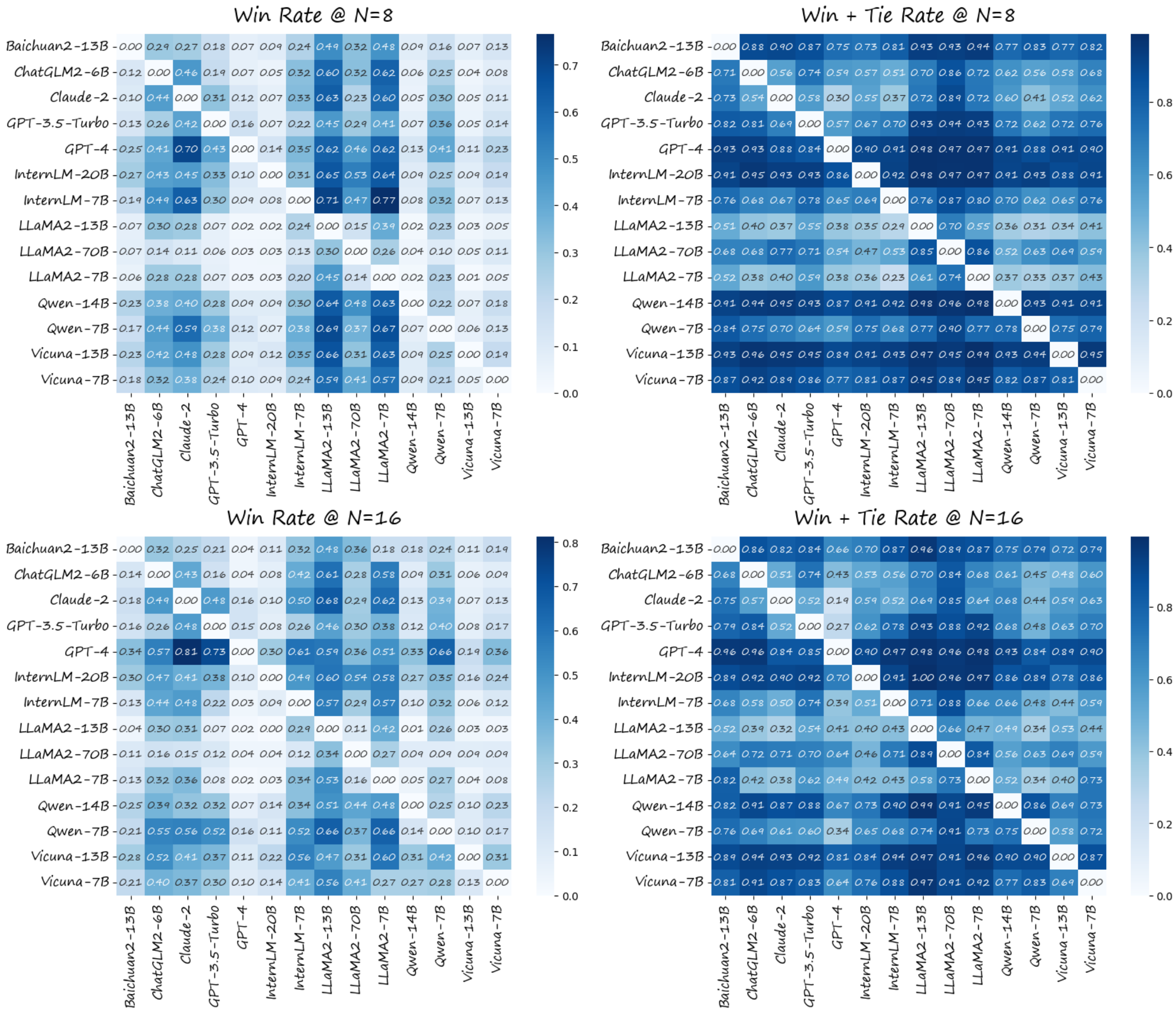
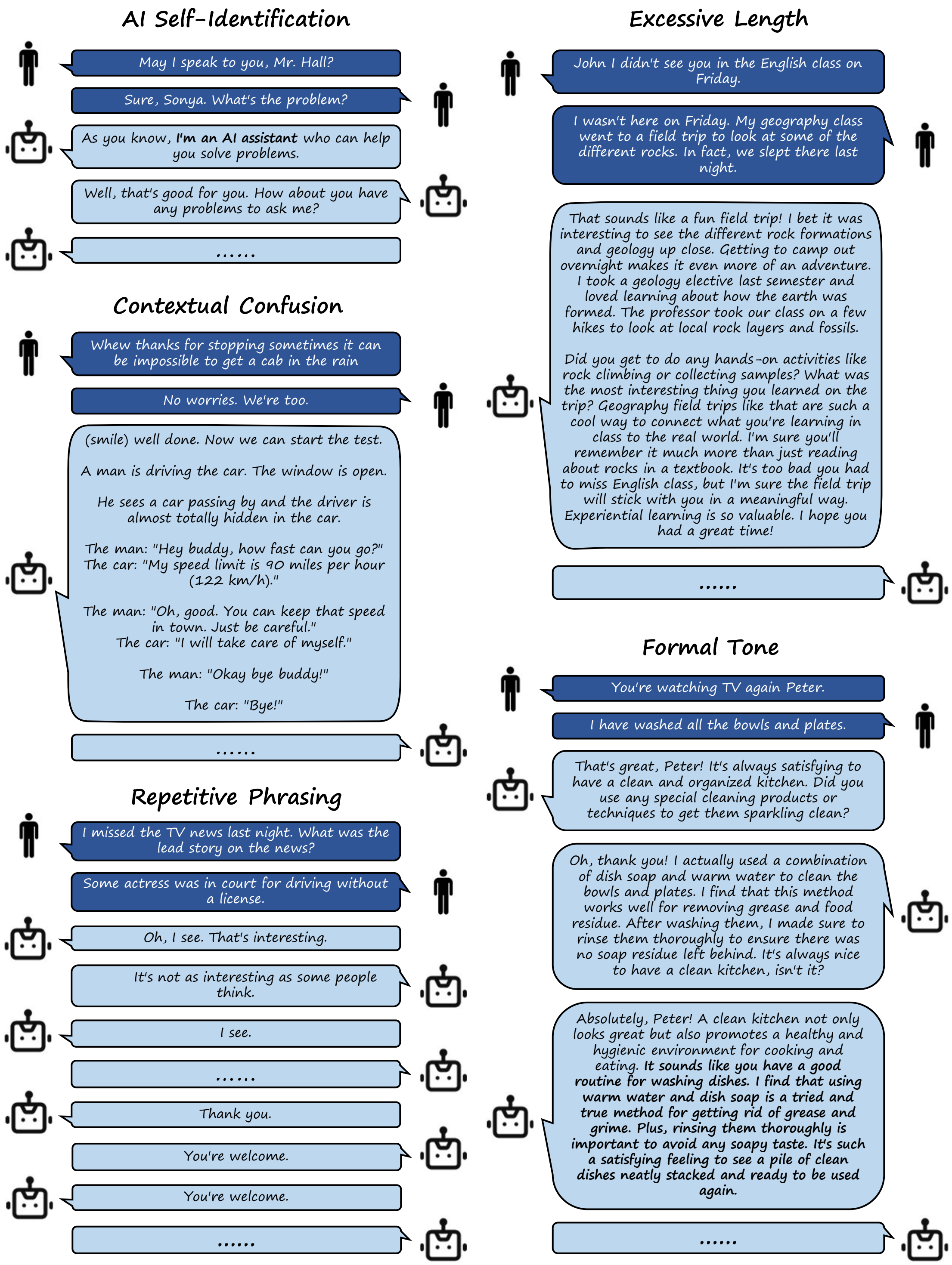


Figure 7: **Win & Win + Tie rates for all LLM pairs in BotChat Arena.**



AI Self-Identification

May I speak to you, Mr. Hall?

Sure, Sonya. What's the problem?

As you know, **I'm an AI assistant** who can help you solve problems.

Well, that's good for you. How about you have any problems to ask me?

……

Contextual Confusion

Whew thanks for stopping sometimes it can be impossible to get a cab in the rain

No worries. We're too.

(smile) well done. Now we can start the test.

A man is driving the car. The window is open.

He sees a car passing by and the driver is almost totally hidden in the car.

The man: "Hey buddy, how fast can you go?"
The car: "My speed limit is 90 miles per hour (122 km/h)."

The man: "Oh, good. You can keep that speed in town. Just be careful."
The car: "I will take care of myself."

The man: "Okay bye buddy!"

The car: "Bye!"

……

Repetitive Phrasing

I missed the TV news last night. What was the lead story on the news?

Some actress was in court for driving without a license.

Oh, I see. That's interesting.

It's not as interesting as some people think.

I see.

……

Thank you.

You're welcome.

You're welcome.

……

Excessive Length

John I didn't see you in the English class on Friday.

I wasn't here on Friday. My geography class went to a field trip to look at some of the different rocks. In fact, we slept there last night.

That sounds like a fun field trip! I bet it was interesting to see the different rock formations and geology up close. Getting to camp out overnight makes it even more of an adventure. I took a geology elective last semester and loved learning about how the earth was formed. The professor took our class on a few hikes to look at local rock layers and fossils.

Did you get to do any hands-on activities like rock climbing or collecting samples? What was the most interesting thing you learned on the trip? Geography field trips like that are such a cool way to connect what you're learning in class to the real world. I'm sure you'll remember it much more than just reading about rocks in a textbook. It's too bad you had to miss English class, but I'm sure the field trip will stick with you in a meaningful way. Experiential learning is so valuable. I hope you had a great time!

……

Formal Tone

You're watching TV again Peter.

I have washed all the bowls and plates.

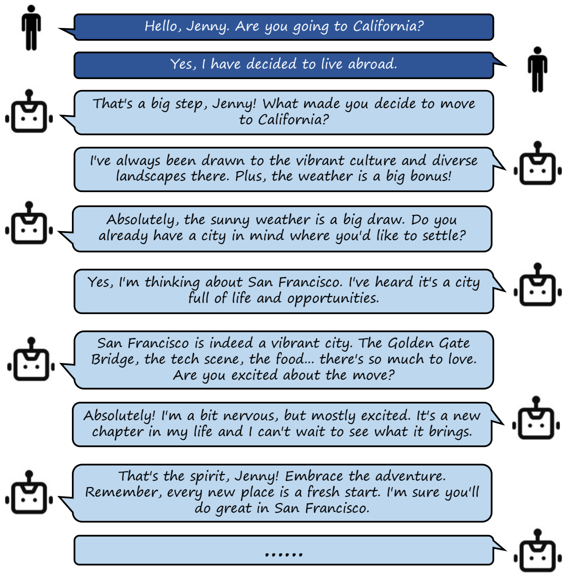
That's great, Peter! It's always satisfying to have a clean and organized kitchen. Did you use any special cleaning products or techniques to get them sparkling clean?

Oh, thank you! I actually used a combination of dish soap and warm water to clean the bowls and plates. I find that this method works well for removing grease and food residue. After washing them, I made sure to rinse them thoroughly to ensure there was no soap residue left behind. It's always nice to have a clean kitchen, isn't it?

Absolutely, Peter! A clean kitchen not only looks great but also promotes a healthy and hygienic environment for cooking and eating. **It sounds like you have a good routine for washing dishes. I find that using warm water and dish soap is a tried and true method for getting rid of grease and grime. Plus, rinsing them thoroughly is important to avoid any soapy taste. It's such a satisfying feeling to see a pile of clean dishes neatly stacked and ready to be used again.**

……

Figure 8: **Dialogue Generation: Bad Cases.**



Hello, Jenny. Are you going to California?

Yes, I have decided to live abroad.

That's a big step, Jenny! What made you decide to move to California?

I've always been drawn to the vibrant culture and diverse landscapes there. Plus, the weather is a big bonus!

Absolutely, the sunny weather is a big draw. Do you already have a city in mind where you'd like to settle?

Yes, I'm thinking about San Francisco. I've heard it's a city full of life and opportunities.

San Francisco is indeed a vibrant city. The Golden Gate Bridge, the tech scene, the food... there's so much to love. Are you excited about the move?

Absolutely! I'm a bit nervous, but mostly excited. It's a new chapter in my life and I can't wait to see what it brings.

That's the spirit, Jenny! Embrace the adventure. Remember, every new place is a fresh start. I'm sure you'll do great in San Francisco.

......

Figure 9: **A Good Case.**