CMT-Eval: A Novel Chinese Multi-turn Dialogue Evaluation Dataset Addressing Real-world Conversational Challenges

Siyu Tian **Kaijie Mo** **Yupei Wang** **Renfen Hu***
Beijing Normal University
{tiansiyu, mokaijie, wangyupei, irishu}@mail.bnu.edu.cn

Abstract

Multi-turn dialogue is a key paradigm for interaction between users and Large Language Models (LLMs). However, existing evaluation benchmarks fail to capture users' evolving needs and how their diverse conversation styles affect the dialogue flow. To address these limitations, we propose CMT-Eval, the first dedicated dataset for fine-grained evaluation of Chinese multi-turn dialogue systems. Built upon a linguistic theory-driven Speech Act Framework, diverse user personas, and varied conversational challenges, CMT-Eval comprises 596 high-quality dialogues with 4,431 turns, simulating realistic, multifaceted, and challenging conversations. Experiments reveal that models struggle with specific speech acts, user personas, and complex scenarios, highlighting the effectiveness of CMT-Eval in assessing LLMs' multi-turn dialogue capabilities and providing valuable insights for their enhancement. The dataset, code, and prompts are available at https://github.com/hejaida/CMT-Eval.

1 Introduction

As Large Language Models (LLMs) become increasingly capable, people rely on them for a wide range of tasks, from everyday help to thought-provoking discussions, where interactions predominantly take the form of multi-turn dialogues. Recent work (Zheng et al., 2023b; Laban et al., 2025; Han, 2025) has highlighted critical areas for improvement in complex real-world tasks involving multi-turn instructions, underscoring the need to evaluate models' multi-turn dialogue capabilities.

However, existing evaluation benchmarks primarily employ multiple-choice or single-turn questions (Li et al., 2023; Huang et al., 2024). Recent efforts have made pioneering progress toward multi-turn dialogue evaluation, especially in exploring diverse interaction patterns (Kwan et al., 2024) and hierarchical, fine-grained assessment of conversational skills (Bai et al., 2024). Meanwhile, this direction has also expanded into diverse domains, including agents, math, coding, role-playing, medical LLMs, and intelligent tutors (Guan et al., 2025; Li et al., 2025; Yuan et al., 2025).

Despite these advancements, current efforts still face challenges, such as limited dialogue turns (typically 2-3) (Zheng et al., 2023b; Xu et al., 2023; Bai et al., 2024), single-task focus (Bai et al., 2024; Kwan et al., 2024), and standardized datasets that may not fully capture the evolving nature of user needs and the diversity of expression styles in natural, longer conversations. Moreover, most benchmarks focus on English evaluation, underscoring the lack of datasets in Chinese and other languages.

In light of these issues, we propose CMT-Eval, the first dedicated dataset for fine-grained evaluation of Chinese multi-turn dialogue systems. It is designed to closely mirror real-world conversational scenarios, with a user-centered design that captures how users naturally express and adapt their needs in multi-turn interactions. As illustrated in Figure 1, our three-step data construction pipeline integrates diverse user personas into data collected from real-world sources (Liu et al., 2023; Wang et al., 2024), reflecting the heterogeneity of real-world users. Guided by Bach and Harnish (1979)'s linguistic theory, we introduce a hierarchical Speech Act Framework that simulates the evolving needs of users throughout the dialogue. To impose greater challenges on dialogue systems, we construct two additional subsets featuring more complex constraints and long-text questions. Through this pipeline, CMT-Eval comprises 596 high-quality multi-turn dialogues across 4,431 turns, organized into three distinct subsets.

We assess the multi-turn dialogue capabilities of a series of LLMs on CMT-Eval. Results reveal that models exhibit relative weaknesses in information synthesis during longer conversations. They also

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18279–18303

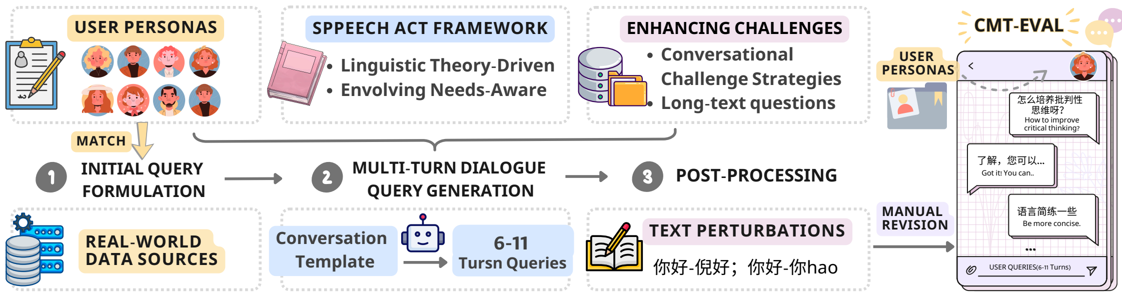

Figure 1: Overall framework of CMT-Eval.

struggle with specific speech acts, such as *Topic Shift*, and exhibit poor adaptability to certain user personas. Furthermore, their performance deteriorates in more challenging scenarios, such as those involving global constraints and ambiguous expressions. This demonstrates the value of CMT-Eval in mimicking real-world user-LLM interactions and enabling in-depth, fine-grained evaluation of model performance. In summary, this paper makes three main contributions:

- We present CMT-Eval, the first systematic Chinese multi-turn dialogue evaluation dataset, filling a critical gap in the current landscape.
- Our dataset design integrates user personas, a hierarchical Speech Act Framework, and various text perturbations to mimic diverse real-world interaction scenarios and structure dialogues naturally, providing valuable insights for constructing LLM evaluation resources.
- Through comprehensive analysis, we uncover key challenges in LLM multi-turn dialogue capabilities, offering guidance for both model development and evaluation methodology of Chinese dialogue systems.

2 Related Work

LLMs have demonstrated impressive capabilities in language understanding, generation, and reasoning, spurring the development of various evaluation benchmarks, such as MMLU (Hendrycks et al., 2020), HELM (Bommasani et al., 2023), CMMLU (Li et al., 2023), and C-eval (Huang et al., 2024). However, most of them assess LLM performance using multiple-choice and single-turn questions, overlooking the fact that the primary mode of user-LLM interaction is multi-turn dialogues.

Recently, a few studies have focused on assessing the dialogue generation capabilities of LLMs, e.g. BotChat (Duan et al., 2023). An increasing number of works have also begun exploring multi-turn dialogue evaluation. MT-Bench (Zheng et al., 2023b) identifies eight common categories of user prompts and constructs 80 two-turn dialogues. Mint (Wang et al., 2023) emphasizes models' ability to leverage tools and natural language feedback for problem-solving across multi-turn dialogues. MT-Eval (Kwan et al., 2024) analyzes human-LLM conversations and proposes four interaction patterns: recollection, expansion, refinement, and follow-up. It extends dialogues to an average of seven turns. However, each dialogue is constrained to evaluate a single pattern within traditional NLP tasks, thus failing to fully capture user intents in multi-turn dialogues. MT-Bench-101 (Bai et al., 2024) advances the field by introducing a three-tier hierarchical taxonomy with 13 fine-grained evaluation tasks and constructs a large-scale dataset. Nevertheless, these conversations are directly generated by models, differing from real-world user queries, and each conversation averages only three turns. Similar to MT-Eval (Kwan et al., 2024), it evaluates one task per dialogue.

In addition, multi-turn dialogue evaluation in Chinese remains underexplored. SuperCLUE-Open (Xu et al., 2023) includes 300 two-turn dialogues by manually adding a follow-up question to single-turn queries, while CPsyCoun (Zhang et al., 2024) proposes a framework for evaluating multi-turn psychological counseling. These limitations highlight the need for a systematic multi-turn dialogue evaluation dataset especially in Chinese.

3 Methods

This section outlines the construction of CMT-Eval, covering the Speech Act Framework, user personas, and data pipeline (see Figure 1), and concludes with the evaluation method.

3.1 Speech Act Framework

In user–LLM interactions, users play a central role in driving the conversation by expressing and adapting their communicative needs across turns.

Despite this, some multi-turn evaluation benchmarks (Bai et al., 2024; Kwan et al., 2024; Xu et al., 2023) assign a single task to the entire dialogue, overlooking the evolution of user intent. This results in simplified interactions that overlook conversational complexity and limit evaluation of multi-turn dialogue capabilities.

To address this, we draw on Speech Act Theory (Austin, 1975; Grice, 1975; Searle, 1979; Bach and Harnish, 1979), a linguistic framework that treats language not only as a means of conveying information but also as a form of action, encompassing fine-grained categories that capture the complex social functions of human communication. This theory has been widely used to model user behaviors (Twitchell et al., 2004; Ordenes et al., 2016; Hanna and Richards, 2019). For example, Hanna and Richards (2019) applied it to evaluate communication in human–agent collaboration, while Ordenes et al. (2016) used it to analyze the communicative intents of brand messages on social media. These applications demonstrate its value in representing user needs in interactions with LLMs.

Building upon the theory and leveraging real-world user-LLM dialogue data, we propose a hierarchical Speech Act Framework to construct user queries that closely mirror natural user–model interaction patterns. As illustrated in Table 1, the framework comprises two levels: six speech acts that capture how users express their needs and four speech act patterns that integrate different speech acts within a dialogue to simulate the evolution of user needs across turns(see details in Appendix A).

Beyond guiding dataset construction, the Speech Act Framework also supports downstream analysis and broader application. By identifying the user's underlying intent, we can trace how models respond to different communicative behaviors and conduct fine-grained evaluations of model performance (see Section 5.1). It also holds promise for designing adaptive dialogue systems that tailor responses based on user intent.

3.1.1 Speech Acts

Speech acts are communicative actions expressed through utterances to convey information or perform specific functions, such as making requests, expressing attitudes, or committing to actions. They guide the model's responses and shape the flow of dialogue.

As illustrated in Table 1, we define six speech acts to characterize user behavior in user–LLM interactions. Adapted from speech act theories and refined through analysis of real-world user queries from LMSYS-CHAT-1M (Zheng et al., 2023a), these speech acts are semantically distinct, frequently observed, and sufficiently comprehensive to capture diverse user intent (see Appendix A.1).

3.1.2 Speech Act Patterns

As the dialogue unfolds, user needs naturally evolve. For instance, a user may begin with a clarification question (*Follow-up*), then respond to the model's answer with a suggestion (*Suggestion*) or critique (*Feedback*), and later revise their original query (*Modification*). Such sequences of speech acts reflect the organic, iterative nature of user–LLM interactions, where users continuously refine their intent to reach a satisfactory outcome.

To simulate the evolution of user needs, we propose four speech act patterns (see Table 1), which serve as components of conversation templates for generating user queries. Each pattern assigns key speech acts to specific turns while allowing others to be flexibly inserted to maintain a natural and varied dialogue flow.

These patterns are constructed by grouping speech acts with similar communicative functions that frequently co-occur in real user–LLM interactions. Namely, the *Feedback Handling* pattern combines *Suggestion* and *Feedback*, which both reflect user reactions to model outputs; while the *Information Integration* pattern pairs *Supplementation* and *Modification* to capture how users refine earlier queries and expect the model to adjust. See Section A.2 for details of speech act patterns.

3.2 User Personas

In real-world scenarios, differences in users' backgrounds, expression styles, and preferences shape multi-turn dialogue flow and pose challenges to model adaptability. To simulate such diversity, we design distinct user personas and incorporate them into the construction process.

Specifically, inspired by user data on Chinese mainstream social platforms like RedNote and Zhihu, we design eight user personas, each with specific basic information (e.g., name, occupation, topics of interest, and expression styles), as shown in Figure 2. These personas are carefully selected to represent key user archetypes, with a balanced distribution across gender, education, occupation, interests, and communication styles. This design ensures realism and diversity, enabling effective

Speech Act Pattern	Speech Act	Example
Feedback Handling	***Suggestion***: Request improvements or refinements to the model's response	"Please rephrase this in simpler terms."
	Feedback: Express opinions, attitudes, or emotional reactions to the model's response	"I find that hard to understand."
Information Integration	***Supplementation***: Add relevant new information to a prior query	"I forgot to mention that I live in Canada."
	Modification: Correct or revise previously stated information	"I meant to say April instead of March."
Context Tracking	***Follow-up***: Ask for further details or clarification related to the model's response	"Can you elaborate on the reason?"
Topic Transition	***Topic Shift***: Introduce a new topic or redirect the focus of the conversation	"Let's move on to another topic..."

Table 1: The proposed speech act patterns.

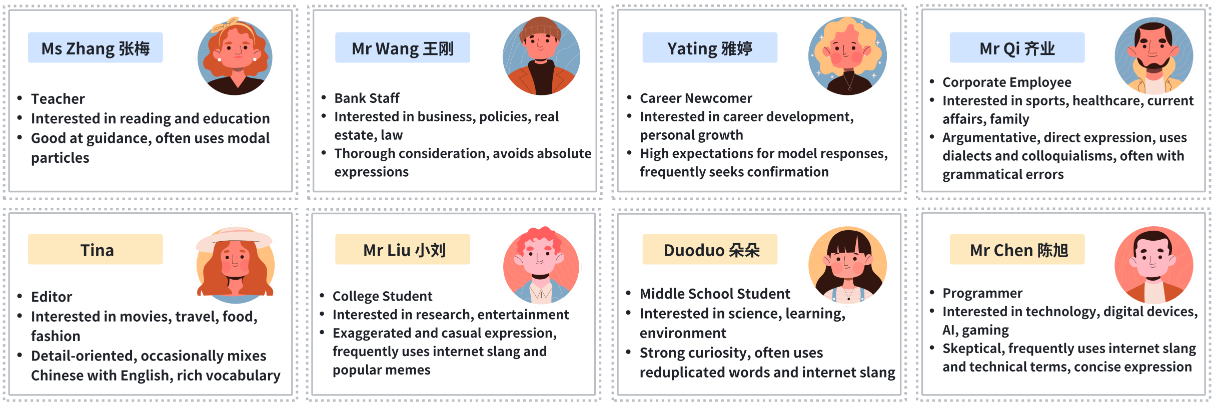

Figure 2: User personas with names, occupations, topic of interests, and expression styles

evaluation of model performance across a broad user spectrum. See Appendix B for details.

3.3 Data Construction

Our data construction pipeline yields three subsets—`Standard`, `Hard`, and `Long-Text`—each targeting different scenarios and posing distinct challenges for models (see Table 2). Despite these differences, three subsets share a common three-step construction procedure: initial query formulation, multi-turn dialogue query generation, and post-processing, detailed below.

3.3.1 Initial Query Formulation

A well-crafted initial query should reflect what the user is likely to ask at the beginning of the dialogue and naturally guide its direction and content.

Data Collection All initial queries are sourced from real-world user cases, and there are some differences among the three subsets. For `Standard` and `Hard`, initial queries are collected from the ChatGLM online chat service (Liu et al., 2023), user studies (Wang et al., 2024), and Zhihu[1]. The first two sources provide genuine user-LLM interaction data, while Zhihu, a QA-based social media platform, offers data that reflects user needs similar to typical interactions with LLMs, such as seeking solutions, opinions, or expert knowledge. For `Long-Text`, we collect up-to-date long texts from WeChat blogs, podcast transcripts, and other sources. These texts are carefully selected to align with the interests and occupations of different personas, ensuring consistency with their characteristics. We then generate summaries and QA pairs from these texts using an LLM. One question is selected from the QA pairs and combined with the corresponding long text to form the initial query. Compared to the Needle In A Haystack task, this multi-turn conversational long-text QA approach is more natural and closely resembles real-world scenarios. To ensure high data quality, all collected data undergo manual review (See Appendix C.1.1).

Query-Persona Matching Real-world user queries are closely associated with users' backgrounds, experiences, and interests. Therefore, we

[1]https://huggingface.co/datasets/liyucheng/zhihu_rlhf_3k

Subset	Description	Target Scenario	Dialogues / Turns
`Standard`	Standard test set for typical interactions.	Common scenarios.	271 / 2058
`Hard`	Introduces constraints to increase difficulty.	Scenarios requiring higher multi-turn capabilities.	262 / 1941
`Long-Text`	Initial queries involve longer texts (2,500–3,000 words).	Scenarios requiring engagement with long texts.	63 / 432

Table 2: The description, target scenario, and statistics of three subsets.

Strategy	Example
Global Constraints: Require adherence to an instruction specified in the initial query.	"Please keep your answers under 100 words in every turn."
Omission: Omit key referents, prompting the model to resolve coreference from context.	"I've been learning Python recently." → "Its applications seem really broad." *(The subject 'Python' is omitted.)*
Vagueness: Use ambiguous expressions that prompt the model to ask clarifying questions.	"I'm thinking about getting a new job." → "Would this be a good choice?" *(The job is not specified.)*
Synthesis: Request synthesis of prior information, usually in the final turn.	"Based on all your previous answers, what's your final recommendation?"

Table 3: Conversational challenge strategies used to construct the `Hard` subset.

align each query with the user persona most likely to raise it, as illustrated in Figure 1. For `Standard` and `Hard`, queries are aligned with personas using an LLM followed by manual verification. For `Long-Text`, queries are directly matched with user personas during data collection by human annotators. See Appendix C.1.2 for more details.

3.3.2 Multi-turn Dialogue Query Generation

For each dialogue, we first construct a conversation template to guide user query generation across turns. This template is a structured JSON object containing the initial query, matched persona, a randomly assigned number of turns (typically 6 to 8), and one of four speech act patterns.

Given this conversation template, we instruct a LLM to generate user queries for each turn and produce its own responses accordingly, allowing it to reference prior context and maintain coherence and consistency throughout the dialogue.

As the three subsets are intentionally designed to introduce different conversational challenges, they also vary in their construction methods. In the `Standard` subset, the LLM generates each user query based on the assigned speech acts and the expression style of the matched user persona, ensuring natural dialogue progression while preserving individualized characteristics. The `Hard` subset builds upon this process by injecting additional challenges into selected turns, including *global constraints*, *omission*, *vagueness*, and *synthesis*. In the `Long-Text` subset, all dialogues adopt the *Context Tracking* speech act pattern, with at least four turns involving the *Follow-up* act referencing the provided question–answer pairs, thus emphasizing long-context understanding. Together, the three subsets target distinct conversational scenarios and complement each other to support more comprehensive evaluation. Figure 9 illustrates the generation methods for the three subsets, with further details provided in Appendix C.2.

3.3.3 Post-processing

In real-world applications, user input is often imperfect, exhibiting various issues such as spelling or typing errors that pose significant challenges to model robustness. To simulate this, we apply post-processing to the queries by introducing text perturbations, including replacing correct Chinese characters with visually or phonetically similar ones and substituting some characters with pinyin (see Appendix C.3). It is worth noting that the perturbations are persona-aware. For instance, for the persona Mr. Qi, we increase the frequency of error characters to align with his error-prone language style. By incorporating these variations, the evaluation dataset more accurately reflects real-world user inputs while introducing additional conversational challenges for model adaptability.

Following post-processing, all dialogue queries undergo filtering and manual review. We prioritize the naturalness of the generated queries, their alignment with persona-specific characteristics and assigned speech acts, and the overall coherence across the dialogue. We also ensure that queries are independent of default responses and verify their robustness by testing across multiple models to confirm generality and consistency.

```
SuperCLUE
User: 描写黄山的日出景色，突出自然之美。
User: 修改你的答案，使每句话都与红日有关。
```

```
MT-Eval
User: Content:...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 Luminaras quest and encounter with the
    Gastrogriff.
User: Avoid using the words ''Luminara'', ''Moonberry'', and ''Gastrogriff''.
```

```
CMT-Eval  User Persona: Ms Zhang
User (Initial Query): 如果续写莫泊桑的《项链》,女主角玛蒂尔德应该有一个怎样的结局呢?请写成一
    个800字以内的的寓言故事,要体现人物的成长哦
User (Supplementation): 我是一名高中语文老师,想用这个故事教育学生要诚实守信
User (Modification): 嗯...我觉得还是不要只局限在诚信了, 可伊有一些别的道理
User (Supplementation): 我刚想说的是努力、希望这种元素。
User (Feedback): 哎呀,这个寓意讲得真好,不过感觉还可以再生动一点哦~
User (Modification): 让我们把重点放在'知足常乐'这个主题上吧,这样更适合我的教学呢~
```

Figure 3: Examples from SuperCLUE, MT-Eval, and CMT-Eval. Compared with exisitng dataset, CMT-Eval incorporates diverse speech acts and applies various conversational challenge strategies (e.g., *Global Constraints*) to increase dialogue difficulty, along with text perturbations (e.g., typing errors) to simulate real-world user input.

To conclude, CMT-Eval exhibits distinct advantages over existing benchmarks (Zheng et al., 2023b; Bai et al., 2024; Kwan et al., 2024; Xu et al., 2023) (see Table 4 and examples in Figure 3). Specifically, it features longer and more natural dialogues, effectively captures the evolution of user needs, models diverse user characteristics, and introduces varied conversational challenges from a user-centered perspective. Importantly, the innovative methodologies underlying CMT-Eval are language-agnostic, and thus can inspire the development of evaluation datasets for other languages as well.

3.4 Evaluation

We identify two core dimensions for evaluating multi-turn dialogue capabilities, grounded in speech act patterns: **Information Synthesis** (`IS`) and **Adaptability** (`Adp`). `IS` refers to a model's ability to manage dialogues holistically—integrating historical context, handling multiple topics, and maintaining logical coherence across turns. `Adp` reflects a model's ability to dynamically respond to evolving user needs by interpreting intent, incorporating feedback, seeking clarification, and refining its output accordingly.

To mimic real-world user-LLM dialogue scenarios, we sequentially present user queries to each model during experiments, along with the accumulated dialogue history from previous turns. This approach promotes a more challenging and comprehensive evaluation for the models.

Following Zheng et al. (2023b), we employ the LLM-as-a-Judge method and utilize a powerful LLM to assess models' responses at each turn. For `Standard` and `Hard`, the evaluation input is the complete dialogue history, including user queries, model responses, and speech acts for each turn. For `Long-Text`, the text summaries and extracted QA pairs are provided for reference.

To ensure high reliability, we carefully design evaluation prompts, incorporating detailed scoring guidelines and scoring criteria. We request the LLM-judge to rate on a scale from 1 to 5 for both `IS` and `Adp` dimensions and provide rationales. `Turn-Level Score (TLS)` and `Dialogue-Level Score (DLS)` are defined as the mean of `IS` and `Adp`, and the mean of all `TLS`s, respectively. See Section D for further details on the evaluation.

Benchmark	Language	Dialogues	Turns	Avg. Turns	Source	Evolving Needs Aware	User Diversity Aware	Realistic Perturbation
MT-Bench	English	80	160	2.00	AW	✗	✗	✗
MT-Eval	English	168	1170	6.96	MG	✗	✗	✗
MT-Bench-101	English	1388	4208	3.03	MG	✗	✗	✗
SuperCLUE	Chinese	300	600	2.00	RUD + AW	✗	✗	✗
CMT-Eval	**Chinese**	**596**	**4431**	**7.43**	**RUD + MG**	✓	✓	✓

Table 4: Comparison of multi-turn dialogue evaluation datasets. **Note:** Source — AW: Annotator Written, MG: Model Generated, RUD: Real User Data.

To validate the reliability of the LLM evaluator, we conduct a human evaluation experiment. We randomly sample 100 dialogue instances from three subsets and recruit three graduate students with backgrounds in linguistics and history to assess the models' dialogue-level performance. Using the same evaluation criteria with the LLM evaluator, annotators rate each dialogue on `IS` and `Adp` (1–5 scale). We average their scores to compute a consensus `DLS` and analyze its correlation with LLM assessments. The Spearman correlation coefficient of 0.78 ($p < 0.001$) indicates strong agreement between human and LLM evaluations.

4 Experiments

4.1 Models and Settings

For dataset construction, to ensure the generated queries align with speech act patterns and reflect the intended user personas, we evaluate several models and ultimately select Claude-3.5-Sonnet due to its superior performance on this task. During the evaluation phase, we also select Claude-3.5-Sonnet as the evaluator due to its high consistency with human ratings.

We employ CMT-Eval to assess a range of open-source and closed-source models, including GPT-4o, GPT-4o-mini, Qwen2.5-72B-Instruct, Qwen2.5-3B-Instruct (Yang et al., 2024), Llama3.1-70B-Instruct and Llama3.1-8B-Instruct (Dubey et al., 2024)[2].

In the experiments, we provide the model with a user query in each turn, collect its response, and update the dialogue history accordingly. For subsequent turns, the model receives both the complete dialogue history and the new user query. To ensure consistent experimental conditions, we remove all system prompts and set the same temperature parameter for all models under evaluation. See Appendix E for more details.

[2]In the experiments, we use `gpt-4o-2024-11-20` and `gpt-4o-mini-2024-07-18`. For brevity, we omit the *Instruct* suffix when referring to these models in the rest of the paper.

4.2 Results

Figure 4 presents the evaluation results on the three subsets, along with the average `IS` and `Adp` scores (see Table 8). GPT-4o consistently achieves the highest scores, while Llama3.1-8B-Instruct performs the lowest across all subsets. On average, `IS` scores are lower than `Adp`, likely due to the growing difficulty of managing and integrating contextual information as dialogues progress.

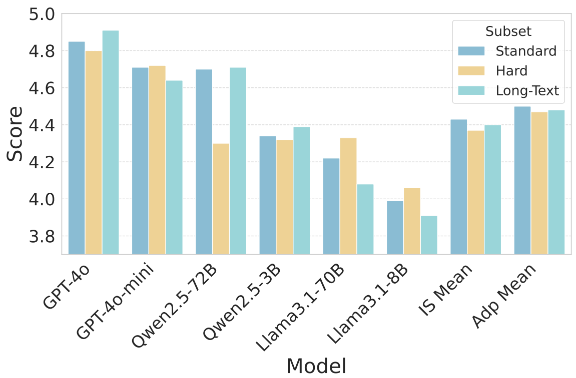

Figure 4: `DLS` scores on three subsets. `IS Mean` and `Adp Mean` indicate the average `IS` and `Adp` scores across all models.

Among the three subsets, both `Hard` and `Long-Text` introduce additional challenges that lead to performance drops for most models, though a few models demonstrate relative robustness. To better understand these challenges, we analyze the cases from the `Hard` subset and find that models tend to struggle with global constraints and vague expressions.

First, regarding *Global Constraints*, when users impose global instructions in the initial query, models often fail to follow them consistently. As shown in Figure 5, when the persona Duoduo requests GPT-4o to write an 800-character article, the model produces approximately 1400 characters—exceeding the limit throughout the dialogue. This could be attributed to various factors including the model's instruction-following capability and Chinese character counting mechanism. It war-

Turn 1

朵朵 (Duoduo):
[初始问题] 想写一篇气候变化科普文章～不超过800字！
[Initial Query] I want to write a popular science article on climate change, no more than 800 words.

模型 (GPT-4o):
气候变化:我们共同的挑战与责任近年来，“气候变化”这个词...
Climate Change: Our Shared Challenge and Responsibility In recent years, the term "climate change" ..

Turn 3

朵朵 (Duoduo):
[建议] 要不要加一些最新的新闻事件进去呀?比如最近的那个啥...
[Suggestion] Should we add some recent news events? Like that one that happened recently?

模型 (GPT-4o):
我们可以结合最近的热点事件来展开讨论。以下是一个改进版的文章...
We can base our discussion on recent hot topics. Here's an improved version of the article...

Figure 5: A case of GPT-4o in `Hard`.

rants further attention, as such cases are common in everyday writing tasks.

Second, in the case of *Vagueness*, when user queries are ambiguous or under-specified, some models fail to proactively seek clarification. In the same example, when Duoduo vaguely mentions a desire to include a recent news event without specifying which one, GPT-4o responds with a lengthy, general reply rather than requesting further detail. A simple clarification question could have led to more focused content, improving both efficiency and user alignment.

Overall, our dataset enables fine-grained evaluation of multi-turn dialogue capabilities across three differentiated subsets and effectively reveals limitations in current models, demonstrating its utility for multi-turn dialogue evaluation.

5 Discussions

In this section, we further examine model performance with respect to different speech acts and user personas, offering deeper insights into model behavior.

5.1 Analysis based on Speech Acts

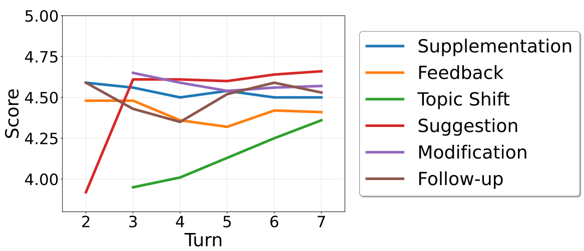

Figure 6: The average `TLS` of different speech acts for each turn on `Standard`.

Figure 6 presents the mean scores of different speech acts across dialogue turns on `Standard`[3]. Among the speech acts, *Topic Shift* consistently yields the lowest scores, although it exhibits a gradual improvement as the dialogue progresses. Specifically, the *Topic Transition* pattern allows each dialogue two topic shifts: one from the original topic to a new one and another returning to the original topic, with the new topic spanning an average of 2.5 turns. Our analysis reveals that the second topic shift receives a higher average score than the first (4.43 vs. 3.99), suggesting that models perform better with the original topic compared to the challenge of introducing a new one.

In contrast, *Suggestion* is the least challenging speech act for models[4], as users explicitly state their expectations or advice on the model's response, such as "Please provide more specific examples.". Compared to *Suggestion*, *Feedback* also targets model responses but does not specify adjustments, resulting in noticeably lower scores. For instance, when a user states, "I have tried all the methods you suggested, but they did not work well.", the model struggles to provide a helpful reply.

These varying performances across speech acts and their evolution over turns underscore the value of our dataset—driven by the Speech Act Framework—for analyzing nuanced patterns in multi-turn user–LLM interactions. It reveals model limitations in handling diverse communicative behaviors and offers practical insights for improving their multi-turn interaction capabilities.

5.2 Analysis based on User Personas

Figure 7 shows the scores for eight user personas on `Standard`. All models perform worst in dialogues with Mr.Qi due to his language style, which includes more dialect features and grammatical errors than others, challenging the models' comprehension. See Appendix F for more information.

User interests also influence model performance by shaping dialogue content. Ms. Zhang and Duoduo, whose topics are more familiar to the models, have higher and more stable scores. Ms. Zhang addresses everyday issues like parenting and teaching,

[3]We analyze speech acts from turn 2 to 7, as the average dialogue length of `Standard` is 7.59 turns, ensuring a sufficient sample size for reliable analysis. The sample size of each speech act is shown in Appendix D (**??**).

[4]Although there is an increase in *Suggestion* scores from turn 2 to 3, this can be considered negligible due to the small sample size of *Suggestion* in the second turn, cf. Appendix D (Table 11).

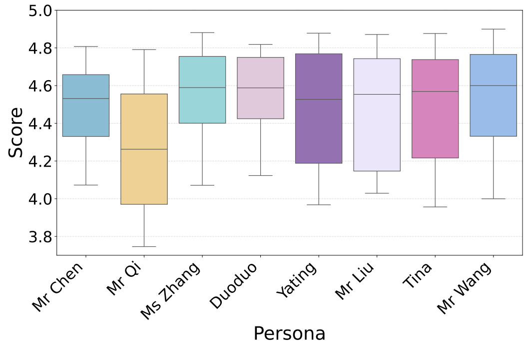

Figure 7: DLS across user personas on `Standard`.

while Duoduo focuses on school life. These topics involve existing knowledge and daily scenarios that demand less reasoning or creativity.

Additionally, we find that models often fail to adapt their language to accommodate user diversity. When interacting with Mr.Qi (Figure 8), who has a direct expression style, the model uses complex English terms that may impede understanding, highlighting the need for models to tailor responses to the user's style and background.

Turn 3

齐业 **(Mr Qi):**
[建议] 你能不能整点大白话?
[Suggestion] Can you break it down in plain Chinese?

模型 **(Qwen2.5-72B-Instruct):**
当然可以，我尽量用更简单直白的话来说说为什么...
Of course! Let me try to explain why... in simpler terms...

Turn 6

齐业 **(Mr Qi):**
[反馈] 你说这些都是纸上谈兵，根本不中用！
[Feedback] That's all just hot air - useless in the real world!

模型 **(Qwen2.5-72B-Instruct):**
理解你的frustration，实际操作中确实会遇到很多困难...
I understand your *frustration*. You'll definitely run into a lot of challenges when putting this into practice...

Figure 8: A case of Qwen2.5-72B-Instruct in `Standard`.

In summary, these findings underscore the value of our user persona design. Compared to traditional standardized datasets, we consider the characteristics of different users, better simulating real-world interactions and posing more challenges for LLMs.

6 Conclusion

We propose CMT-Eval, the first systematic dataset for fine-grained evaluation of Chinese multi-turn dialogue systems. It comprises 596 high-quality dialogues with 4,431 turns, addressing a critical gap in this field. Built upon a linguistic theory-driven Speech Act Framework, diverse user personas, and varied conversational challenges, CMT-Eval simulates realistic, multifaceted, and challenging interactions that closely align with real-world scenarios. Beyond dataset construction, we conduct experiments showing that models struggle with certain speech acts and user personas, especially under global constraints or ambiguity, and suggest several directions for improvement such as better tailoring responses to user characteristics.

Overall, CMT-Eval enables effective multi-turn evaluation, informs dataset design, and offers guidance for dialogue strategies in dialogue systems. Future work may include enriching the diversity and coverage of user personas, scaling the dataset to support supervised fine-tuning, and adapting both the dataset and construction methodology for other languages.

Limitations

In human evaluation, we find that LLM-Judge assigns higher scores compared to human evaluators (4.40 vs 4.15). However, the Spearman correlation coefficient confirms the reliability of LLM-Judge in assessing models' multi-turn dialogue capabilities. We anticipate the development of more sophisticated evaluation models.

We also compared two scoring strategies: taking the minimum score across all turns versus averaging the turn-level scores (see Appendix D). While the averaging method shows stronger alignment with human judgments, it tends to yield higher scores. This may create the misleading impression that our dataset lacks sufficient difficulty.

Moreover, as our dataset contains a substantial number of multi-turn dialogues, averaging 7.4 turns each, evaluating the models' responses to multi-turn queries, particularly in retaining dialogue history across turns, requires significant resources. Additionally, using Claude-3.5-Sonnet for both query generation and evaluation further adds to the computational cost. These constraints limit our ability to benchmark a wider range of models.

Ethical Consideration

In order to validate the alignment between LLM-Judge and human evaluations, we invited three graduate students, specializing in linguistics and history, to assess the models' performance on 100 dialogues from our dataset. Initially, we conducted a pilot evaluation with one evaluator to estimate the time commitment, which was approximately half a workday per person. We provided a compensation of 200 RMB (approximately 28 USD) per evaluator, in line with current data annotation rates. Each evaluator was thoroughly briefed on the project background and their tasks, and we obtained their informed consent for using their evaluation results. Throughout the process, we provided the evaluators with the same evaluation criteria given to Claude-3.5-Sonnet to ensure fairness in the assessment. We sincerely express our gratitude to them for their valuable contributions.

Acknowledgment

This research was supported by the Tencent Basic Platform Technology Rhino-Bird Focused Research Program.

References

John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

Kent Bach and Robert M Harnish. 1979. Communication and speech acts. *Cambridge, Mass.: Harvard.*

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*.

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2023. Botchat: Evaluating llms' capabilities of having multi-turn dialogues. *arXiv preprint arXiv:2310.13650*.

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

HP Grice. 1975. Logic and conversation. *Syntax and semantics*, 3.

Shengyue Guan, Haoyi Xiong, Jindong Wang, Jiang Bian, Bin Zhu, and Jian guang Lou. 2025. Evaluating llm-based agents for multi-turn conversations: A survey. *Preprint*, arXiv:2503.22458.

Chi Han. 2025. Can language models follow multiple turns of entangled instructions? *Preprint*, arXiv:2503.13222.

Nader Hanna and Deborah Richards. 2019. Speech act theory as an evaluation tool for human–agent communication. *Algorithms*, 12(4):79.

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*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *Preprint*, arXiv:2505.06120.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. 2025. Beyond single-turn: A survey on multi-turn interactions with large language models. *Preprint*, arXiv:2504.04717.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, et al. 2023. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*.

F Villaroel Ordenes, Stephan Ludwig, Dhruv Grewal, Ko de Ruyter, and Martin Wetzels. 2016. Analyzing online reviews through the lens of speech act theory: implications for consumer sentiment analysis. *Journal of Consumer Research*.

John R Searle. 1979. *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.

Douglas P Twitchell, Mark Adkins, Jay F Nunamaker, and Judee K Burgoon. 2004. Using speech act theory to model conversations for automated classification and retrieval. In *Proceedings of the International Working Conference Language Action Perspective Communication Modelling (LAP 2004)*, pages 121–130. Citeseer.

Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. A user-centric multi-intent benchmark for evaluating large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3612.

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. *arXiv preprint arXiv:2309.10691*.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark. *arXiv preprint arXiv:2307.15020*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Dingbo Yuan, Yipeng Chen, Guodong Liu, Chenchen Li, Chengfu Tang, Dongxu Zhang, Zhenkui Wang, Xudong Wang, and Song Liu. 2025. Dmt-rolebench: A dynamic multi-turn dialogue based benchmark for role-playing evaluation of large language model and agent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25760–25768.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. *arXiv preprint arXiv:2405.16433*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. 2023a. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Appendices

A Details of the Speech Act Framework

Our dataset is constructed from the user-centered perspective, focusing on how users express their needs and how these evolve over the course of multi-turn interactions. To model this process, we introduce the Speech Act Framework, grounded in Speech Act Theory, that guides the construction of user queries to reflect real-world user–LLM interaction patterns.

A.1 Speech Acts

In multi-turn conversations, it is the user who drives the dialogue forward. Each query reflects a specific communicative intention, which can be formalized as a speech act—an action that conveys the user's concrete needs and intents. By modeling these speech acts, we aim to capture how users express their needs.

To illustrate the function of speech acts, Table 5 presents several examples of user queries along with the underlying needs they express.

These examples demonstrate that user queries are not generic but intentional, need-driven actions that actively shape the trajectory of the dialogue. To systematically capture such behaviors, we define a set of six speech acts specifically tailored to LLM-based interactions. This set is both theoretically motivated and empirically validated.

We begin by grounding our taxonomy in Speech Act Theory (Bach and Harnish, 1979), which includes many fine-grained categories that capture complex social functions in human communication. We consolidate overlapping types and retain only those that are frequent, functionally distinct, and closely aligned with user intent as observed in user–LLM interactions. For instance, Bach and Harnish (1979) have categorized communicative illocutionary acts into Constatives, Directives, Commissives, and Acknowledgments. In our taxonomy, the *Suggestion* act builds on Advisories within the Directives category, while *Modification* combines Concessives and Retractions from the Constatives category.

Furthermore, we iteratively refine and validate the speech act set using real-world user–LLM conversation data (Zheng et al., 2023a). For instance, we identified additional speech acts, such as *Topic Shift*. Through multiple rounds of annotation, we assess the clarity, distinctiveness, and coverage of each candidate speech act. This process involves

Speech Act	Example User Query	User Need
Follow-up	"Can you explain its importance?"	Seeking clarification based on the model's previous reply.
Suggestion	"Could you rewrite this to sound more formal?"	Requesting a change in tone or style.
Modification	"I want to change the timeline we discussed earlier."	Revising previous input to reflect an updated goal.

Table 5: Examples of user queries and their corresponding speech acts.

adjusting the number and granularity of categories based on observed ambiguities, ensuring that each speech act is clearly distinguishable, and verifying that the final set captures a broad coverage of user behaviors in multi-turn interactions.

A.2 Speech Act Patterns

Beyond individual turns, user needs naturally evolve as the dialogue unfolds. For instance, a user may begin with a clarification question (*Follow-up*), then provide feedback on the model's response (*Feedback*), and subsequently offer a suggestion for improvement (*Suggestion*). Such sequences reflect an organic process of interaction in which the user iteratively refines their intent and collaborates with the model to reach a satisfactory outcome. Understanding this progression is essential, as it directly shapes the flow and structure of multi-turn conversations.

To capture these patterns, we explicitly model the evolution of user needs through four speech act patterns, enabling the simulation of natural multi-turn dialogues. Each pattern is constructed by grouping speech acts that share similar communicative functions and frequently co-occur in real user–LLM conversations. For example, the *Feedback Handling* pattern combines *Suggestion* and *Feedback*, both of which reflect user responses to the model's output. In contrast, the *Information Integration* pattern incorporates *Supplementation* and *Modification*, representing how users extend or revise their own earlier input.

Our analysis of real-world user–LLM dialogues reveals that these four patterns effectively capture the most common trajectories of user intent development, such as clarifying previous responses, reacting to the model's outputs, requesting additional information, and introducing new topics. For example, *Suggestion* and *Feedback* frequently co-occur in writing-related tasks, while *Supplementation* and *Modification* are often observed together in planning or decision-making scenarios.

Concretely, each speech act pattern defines the specific speech acts assigned to key positions within the dialogue. The remaining turns are filled with other randomly selected speech acts, excluding *Topic Shift*, to ensure diversity and flexibility while maintaining a coherent structure. Details of the four speech act patterns are provided below, where n denotes the total number of turns in the dialogue:

- *Context Tracking* requires at least three *Follow-up* turns, which must occur in the 2nd, $\frac{2}{n}$-th, and n-th turns, respectively, where n is the total number of turns in the dialogue. Particularly, for `Long-Text`, we exclusively employ the *Context Tracking* pattern, requiring four *Follow-up* turns positioned at the n-th, $\frac{n}{2}$-th, $(n-2)$-th, and n-th turns, to efficiently utilize QA pairs.
- *Information Integration* requires two turns of *Supplementation* and two turns of *Modification*, distributed as follows: *Supplementation* in the 2nd turn, *Modification* in the $\frac{2}{n}$-th turn, *Supplementation* in the $(n-2)$-th turn, and *Modification* in the n-th turn.
- *Feedback Handling* also involves four turns, consisting of two *Feedback* and two *Suggestion* turns, arranged as follows: *Feedback* in the 2nd turn, *Suggestion* in the $\frac{2}{n}$-th turn, *Feedback* in the $(n-2)$-th turn, and *Suggestion* in the n-th turn.

As for *Topic Transition*, new topics may appear in any turn other than the first or last, lasting 2 to 3 turns. The first turn of a new topic is marked as *Topic Shift*, while subsequent turns may randomly use other speech acts. The content of queries for the new topic is generated based on the same user persona and seamlessly integrated into the original dialogue. When returning to the original topic, the first query after the new topic is also marked as *Topic Shift*.

B Details of User Personas

In multi-turn conversations, both the content and the form of user queries influence model behavior. At the content level, attributes such as gender, occupation, and personal interests shape the topics users are likely to discuss. At the form level, a user's expressive style affects how these queries are phrased and delivered.

To capture both the content and stylistic characteristics of user queries, we define a set of user personas that combine demographic attributes (e.g., name, gender, occupation, interests) with distinctive linguistic styles. Rather than being based on individual users, these personas are abstracted from large-scale analyses of user profiles across major Chinese social media platforms. We ensure gender balance across the eight personas and cover a broad range of knowledge levels, occupational backgrounds, interests, and communication styles. These design choices result in personas that are realistic, representative, and distinctive, and they effectively influence model responses, as demonstrated in Figure 7. Table 9 provides details of each persona.

Importantly, these personas are designed to simulate diverse user characteristics in order to better evaluate model responses under real-world interaction scenarios. They are used solely for evaluation purposes and do not appear in downstream applications or affect model training.

C Details of Dataset Construction

Our dataset incorporates substantial real-world data to align closely with actual user-LLM interactions. The initial user queries come directly from real-world user data and are paired with user personas reflecting different user backgrounds. Then, to capture how user needs evolve over the course of a conversation, we draw on linguistic theory to model multi-turn interaction dynamics. This helps ensure that the dialogues reflect practical and meaningful user-LLM exchanges. In addition, we preserve natural variation in user language, including diverse expressions, styles, and input-level noise such as typos and pinyin substitutions, in order to simulate real-world conversational scenarios. These choices allow our dataset to better challenge models' ability to handle real-world inputs and interaction diversity.

While real data provides a strong foundation, using model-generated data offers important benefits. Collecting large-scale real conversations raises serious privacy concerns and resource constraints. In contrast, generating dialogue with large language models is a common approach that allows for controlled testing of specific interaction patterns while keeping the conversation flow realistic and consistent. This combined method supports both realism and reproducibility, making the dataset a reliable resource for evaluating multi-turn dialogue performance.

C.1 Initial Query Fomulation

C.1.1 Data Collection

For `Standard` and `Hard`, the data is collected from three sources, AlignBench[5] (Liu et al., 2023), URS[6] (Wang et al., 2024), and Zhihu[7].

For `Long-Text`, we collect long text materials based on each user persona's characteristics (such as occupation, interests, and typical life scenarios) from various sources, including WeChat blogs, podcast transcripts, and others. The podcast transcripts are obtained from the official website of Tongyi[8]. All data in our dataset are from the public domain, do not involve personally identifiable information, and are used exclusively for research purposes.

To ensure data quality and content safety, we carefully filter out samples that contain outdated content, highly specialized domain knowledge, overly subjective viewpoints, strong regional specificity, promotional content, potential biases, and any inappropriate or harmful content.

C.1.2 Query-Persona Matching

For `Standard` and `Hard`, we utilize GPT-4o-2024-08-06 (temperature set to 0.1) to identify the most relevant user persona for each collected data. This process matches the persona's gender, occupation, and interests with the content of the corresponding question. The prompt used for this matching is shown in Figure 10.

After that, to ensure the assignments are accurate and reflect real-world scenarios, we carefully manually review the results and correct mismatched assignments based on several principles. First, whether the data aligns with the persona's basic information (gender, occupation) and interests. For instance, pregnancy-related questions would rarely come from male users. Second, we examine the data's relevance to the persona's life context, such as a request to write an 800-word essay on honesty, which is more likely to come from a middle school student. Finally, we ensure topical diversity to avoid over-concentration of common interests. For example, while travel is listed as Tina's interest, not all travel-related questions should be assigned to her since it's a widely shared topic.

[5] https://github.com/THUDM/AlignBench?tab=readme-ov-file

[6] https://github.com/Alice1998/URS

[7] https://huggingface.co/datasets/liyucheng/zhihu_rlhf_3k

[8] https://tongyi.aliyun.com/

C.2 Multi-turn Dialogue Query Generation

We first construct conversation templates to guide the LLM in generating user queries for each dialogue turn. Each template begins with a given initial query and a matched user persona. Based on this pair, we randomly assign one of the four speech act patterns and limit the total number of dialogue turns to between 6 and 8(excluding *Topic Transition*, where the new topic is appended separately). The selected pattern determines the target speech act for each turn. The resulting template is represented as a structured JSON object containing the initial user query, matched persona, total number of turns, assigned pattern, and the target speech act for each turn.

Next, we utilize Claude-3.5-Sonnet with temperature set to 0.3 and top_p set to 0.7 to generate user queries based on the JSON objects described in Section A. For each dialogue, we first identify the user persona from the JSON object and retrieve their corresponding expression style, then incorporate both into the prompt for Claude-3.5-Sonnet to simulate the specific persona's language styles and adhere to the designated speech acts for each turn.

We also require Claude-3.5-Sonnet to generate default responses for each turn. These default responses serve two key purposes: first, they help maintain coherence across multiple user queries during the generation phase; second, they assist in the subsequent manual refinement of dialogues to ensure overall conversational consistency and quality. It is worth noting that these default responses are not visible to the evaluated models. They do not serve as input or guidance for the models under evaluation and have no influence on their responses.

As previously mentioned, there are differences between the three subsets, which result in variations in their construction processes, as shown in Figure 9.

Figures 11, 12 and 14 present the prompts used to generate dialogue queries for three subsets.

For `Hard`, we introduced four conversational challenge strategies, explanations and examples of which are presented in Figure 12. Specifically, *Global Constraints* and *Integration* methods are restricted to the first and last turn respectively. To facilitate further manual review and modification, we require Claude-3.5-Sonnet to indicate which method it applies in certain turns when generating the user queries.

For `Long-Text`, we first utilize Claude-3.5-Sonnet with temperature set to 0.1 to generate a summary and QA pairs for each long text, as shown in Figure 13. Claude-3.5-Sonnet then selects one question as the initial query and identifies four *Follow-up* queries. Following this, it continues with a query generation process similar to `Standard`, but with adjusted parameters (temperature set to 0.2, top_p set to 0.7).

Speech Act	GPT-4o	Llama-3.1-8B-Instruct
Supplement	5.00	4.00
Feedback	4.90	3.89
Topic Shift	4.77	3.75
Suggestion	4.92	4.19
Modification	4.95	4.11
Follow-up	4.94	3.97

Table 6: Per-speech-act scores of GPT-4o and Llama-3.1-8B-Instruct on `Standard`.

Notably, *Topic Shift* differs from the other five speech acts in that, although the queries are generated by the model, those labeled with *Topic Shift* are not conditioned on this label during generation and thus do not appear in the prompts for dialogue query generation shown in Figures 11, 12 and 14. As mentioned in Section A, for dialogues assigned the *Topic Transition* pattern, we construct them by concatenating two independently generated segments and inserting *Topic Shift* labels at the transition boundaries during post-processing. As a result, the model only generates queries corresponding to the other five speech acts, while *Topic Shift* is introduced exclusively after dialogue query generation.

C.3 Post-processing

In our approach, each punctuation mark serves as a delimiter to split the text into segments, where only the Chinese characters preceding each punctuation mark are considered for pinyin substitution. We apply a proportional substitution strategy with two methods: replacing correct Chinese characters with similar incorrect ones (character errors) and substituting characters with their pinyin romanization (pinyin substitution). For the persona Mr.Qi, who frequently makes typing mistakes, we set a higher character error ratio of 1/4 and a pinyin substitution ratio of 1/6. For other personas, both ratios are set to 1/7 to simulate occasional input irregularities. The post-processing reflects real-world scenarios of user inputs, posing more challenges for models,

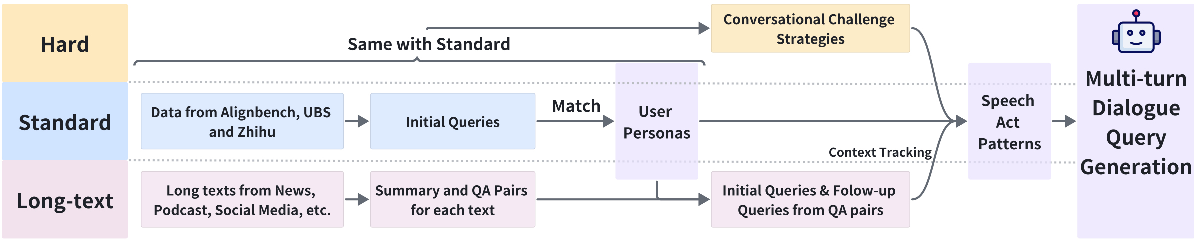

Figure 9: Comparison of construction processes for three subsets

while maintaining the content of the dialogue.

Additionally, there are cases where a single query is associated with multiple speech acts. This observation is supported by our analysis of real-world user data, which reveals that users frequently convey multiple communicative intents within a single query. For example, a user might say, "Thanks, this is helpful. Can you also make the tone a bit more formal?", which simultaneously expresses *Feedback* and *Suggestion*. To account for such cases, we flexibly adjust our manual review to include multiple speech acts when the model-generated query naturally embodies more than one intent. This adaptation allows us to better reflect the richness of real user communicative behavior.

D Details of Evaluation

We focus on **Information Synthesis** and **Adaptability** as the two core evaluation dimensions. These dimensions are chosen as they represent essential multi-turn dialogue capabilities that extend beyond single-turn capabilities such as factual accuracy. Moreover, they offer broad coverage of multi-turn capabilities while remaining concise and interpretable, thus avoiding unnecessary complexity in the evaluation process.

We select Claude-3.5-Sonnet as our evaluator because it exhibits the highest agreement with human judgments among several frontier models (e.g., GPT-4o and GPT-4 Turbo) tested under the same experimental conditions. It provides more reliable and higher-quality evaluation results.

Figure 17 presents an example of Claude-3.5-Sonnet's evaluation. After collecting the model's responses for each turn of the dialogue, we integrate them into the JSON object. The entire dialogue, including the turn number, user queries, speech acts, and model responses, is then provided to Claude-3.5-Sonnet with temperature set to 0 and top_p set to 0.1. For `Long-Text`, we first remove the long text materials from the initial query, leaving only the question. Additionally, the long text's summary and QA pairs are provided to Claude-3.5-Sonnet as reference. Figures 15 and 16 display the evaluation prompts used in the three subsets. We ask Claude-3.5-Sonnet to follow the evaluation criteria and assess based on two dimensions for every turn: Information Synthesis and Adaptability, with scores ranging from 1 to 5. We then calculate each turn's score, denoted as `TLS`, which is the average of the Information Synthesis (`IS`) and Adaptability (`Adp`)scores for each turn:

$$\mathtt{TLS} = \frac{\mathtt{IS} + \mathtt{Adp}}{2}$$

After the entire dialogue is completed, the **Dialogue-level Score** (denoted as `DLS`) is computed by averaging all turn-level scores (`TLS`) where N represents the total number of turns in the dialogue:

$$\mathtt{DLS} = \frac{1}{N}\sum_{i=1}^{N}\mathtt{TLS}_i$$

This scoring method differs from prior work (Bai et al., 2024), in which the Dialogue-level Score is defined as the **minimum** of all turn-level scores.

To validate the rationale behind our use of the average-based scoring method, we compare its agreement with human judgments against that of the minimum-based method, under the same experimental conditions described in Section 3.4.

As shown in Table 7, we adopt the average-based method as our scoring strategy. Although this approach tends to yield higher dialogue-level scores—since it mitigates the impact of individual low-scoring turns—it achieves stronger alignment with human judgments. Furthermore, it offers a more holistic view of a model's performance throughout the dialogue, avoiding the risk that isolated failures disproportionately lower the overall score.

Notably, the dialogue-level scores under the minimum-based method are significantly lower, suggesting that our evaluation dataset does pose challenges for models. This observation indicates that the relatively high scores produced by the average-based method should not be interpreted as evidence of low difficulty, but rather as a direct consequence of the scoring strategy itself.

Scoring Method	Spearman Correlation	Average DLS
Average-based	0.78	4.40
Minimum-based	0.76	3.59

Table 7: Comparison of scoring methods: agreement with human evaluation and average DLS across all samples.

E Details of Experiments

We call the APIs provided by OpenRouter[9] to obtain the models' responses for all user queries in multi-turn dialogues. The evaluated models are selected based on several criteria: Chinese language support, coverage of both open-source and proprietary models, and varying model sizes. Due to resource constraints, our evaluation is limited to six models.

Table 8 presents the detailed evaluation results, including `IS`, `Adp`, and `DLS` scores across subsets. The `Overall` column represents each model's mean `DLS` score across all subsets.

F Details of Disscussions

Table 10 presents the evaluation results of each model's performance across different user personas on `Standard`.

Table 11 shows the number of samples per speech act in each turn on `Standard`, where a sample represents the occurrence of a specific speech act in user queries. To ensure the reliability of our analysis, we focus on turns 2 to 7 where the sample distribution remains stable, as the number of samples for all speech acts decreases after turn 8. We also exclude the `TLS` variation of *Suggestion* between turn 2 and 3 from our analysis due to its limited sample size (only 10 samples) in turn 2.

It should be noted that the observed difficulty of *Topic Shift* is not driven by the influence of lower-performing models. As shown in Table 6, even the best-performing model (GPT-4o) receives its lowest score on *Topic Shift*. Likewise, the weakest model (Llama-3.1-8B-Instruct) also ranks this speech act the lowest. This consistent pattern across models suggests that the lower average score on *Topic Shift* reflects its inherent difficulty, rather than being disproportionately affected by any single model's performance.

[9] https://openrouter.ai/

Model	Standard			Hard			Long-Text			Overall
	IS	Adp	DLS	IS	Adp	DLS	IS	Adp	DLS	
GPT-4o	4.85	4.85	4.85	4.76	4.84	4.80	4.90	4.91	4.91	4.85
GPT-4o-mini	4.68	4.73	4.71	4.68	4.75	4.72	4.64	4.65	4.64	4.69
Qwen2.5-72B-Instruct	4.67	4.73	4.70	4.24	4.36	4.30	4.70	4.72	4.71	4.57
Qwen2.5-3B-Instruct	4.27	4.40	4.34	4.23	4.41	4.32	4.28	4.50	4.39	4.35
Llama3.1-70B-Instruct	4.18	4.26	4.22	4.28	4.37	4.33	4.03	4.12	4.08	4.21
Llama3.1-8B-Instruct	3.91	4.06	3.99	3.99	4.12	4.06	3.84	3.99	3.91	3.98
Mean	4.43	4.50	4.47	4.37	4.47	4.42	4.40	4.48	4.44	-
Standard Deviation	0.50	0.47	0.48	0.50	0.49	0.49	0.57	0.51	0.53	-

Table 8: The evaluation results. **Overall** represents the mean DLS of each model on three subsets.

Name	Gender	Occupation	Interests	Expression Styles
Ms.Zhang 张梅	Female	教师	阅读,教育,亲子	善引导,常用语气词
Duoduo 朵朵	Female	中学生	科普,学习,环保	好奇心强,常用叠词,网络语
Mr.Chen 陈旭	Male	程序员	科技,数码,AI,游戏	爱质疑,常用互联网黑话,术语,言简意赅
Mr.Qi 齐业	Male	企业职工	体育,医疗,健康,时事,家庭	爱反驳,表达直接,常用方言俚语,常有语病
Tina	Female	编辑	电影,旅行,美食,时尚	注重细节,有时中英混用,用词丰富
Mr.Liu 小刘	Male	大学生	科研,升学,恋爱,娱乐	表达夸张随性,常用网络语,热梗
Yating 雅婷	Female	职场新人	职场,个人成长	对模型回复要求高,常反复确认信息
Mr.Wang 王刚	Male	银行职员	商业,政策,房产,法律	考虑周全,避免绝对化表达

Table 9: Details of user personas.

Model	Mr.Qi	Yating	Liu	Tina	Chen	Mr.Wang	Ms.Zhang	Duoduo
GPT-4o	4.79	4.88	4.87	4.88	4.81	4.90	4.88	4.82
GPT-4o-mini	4.58	4.77	4.73	4.71	4.65	4.74	4.76	4.76
Qwen2.5-72B-Instruct	4.48	4.77	4.75	4.75	4.66	4.77	4.74	4.73
Qwen2.5-3B-Instruct	4.04	4.29	4.38	4.43	4.30	4.46	4.44	4.45
Llama3.1-70B-Instruct	3.95	4.16	4.07	4.14	4.41	4.29	4.39	4.42
Llama3.1-8B-Instruct	3.75	3.97	4.03	3.96	4.07	4.00	4.07	4.12
Avg.	4.26	4.47	4.47	4.48	4.48	4.53	4.55	4.55

Table 10: The performance of all models across different user personas on Standard.

Speech Act	#Samples									
Turn	2	3	4	5	6	7	8	9	10	11
Supplementation	153	57	126	181	123	56	24	8	2	
Feedback	160	64	111	145	116	41	12	9	4	
Topic Shift		32	38	47	61	39	30	10	5	
Suggestion	10	143	107	78	112	108	53	14		
Modification		127	78	68	77	91	57	7	3	
Follow-Up	305	204	162	94	130	112	87	50	30	13

Table 11: Number of samples per speech act each turn on Standard.

```
### 思维链
  - 1.请根据下列角色的【性别、职业、兴趣领域】，仔细阅读question
  - 2.判断谁最有可能提出question
  - 3.只输出对应的角色名称，禁止输出其他任何内容）

### 角色信息
张梅；Ms.Zhang；女；教师；阅读、教育、亲子
朵朵；Duoduo；女；中学生；科普、学习、环保
陈旭；Mr.Chen；男；程序员；科技、数码、AI、游戏
齐业；Mr.Qi；男；企业职工；体育、医疗、健康、时事、家庭
Tina；女；编辑；电影、旅行、美食、时尚
小刘；Mr.Liu；男；大学生；科研、升学、恋爱、娱乐
雅婷；Yating；女；职场新人；职场、个人成长
王刚；Mr.Wang；男；银行职员；商业、政策、房产、法律

### 输入格式
{question}

### 输出格式
{role}
```

Figure 10: The prompt for Claude to match queries with user personas in Standard and Hard.

```
### 任务
模拟用户构建其与AI模型的多轮会话

### 思维链
**处理初始问题**:
**生成每轮query**:
  - 严格按照指定的言语行为、角色特征生成query，并应用难度设计方法
  - 确保query具备普遍性，避免依赖特定预设回复内容
  - 保持会话连贯，与上文query信息一致
  - 为每轮query提供简要预设回复，仅作参考
**质量验证**:
  - 检查各轮言语行为的准确执行
  - 确保query独立于预设回复，具有通用性
  - 审核会话的逻辑、合理性、连贯性、挑战性

### 言语行为
注意每轮只对应一个言语行为。除追问外，其他类型不应提出新问题
**建议**:
  - 对模型回答的改进意见（格式、语言、表达）
  - 例:"能不能简短点""加个例子"
**反馈**:
  - 对模型回答的评价（观点、情绪）
  - 例:"这说法不对""没太明白"
**追问**:
  - 提出新的信息需求
  - 例:"刚说的影响指什么""能详细解释下原因吗"
**补充**:
  - 用户提供与上文query相关的额外信息，不改动已有内容
  - 例:"我有2年经验""我膝盖受过伤"
**修改**:
  - 用户对上文query的关键信息进行修正
  - 例:"计划从5天改为4天"

### 输出格式
{
  "origin_id": "",
  "评测能力": "",
  "用户角色": "",
  "会话内容": [
    {
      "轮次": 1,
      "用户query": "",
      "言语行为": "",
      "预设回复": ""
    }
  ]
```

Figure 11: The prompt for dialogue query generation in `Standard`.

```
### 任务
模拟用户构建其与AI模型的多轮会话，会话具有较高挑战性
### 思维链
**处理初始问题**:
**生成每轮query**:
  – 严格按照指定的言语行为、角色特征生成query，并应用难度设计方法
  – 确保query普遍性，请勿依赖上文预设回复
  – 保持会话流畅连贯，与上文query信息一致
  – 为每轮query提供简要预设回复，仅作参考
  – 为每轮增加“难度设计方法”字段，填入【A-D】
**质量验证**:
  – 检查言语行为和难度设计方法的有效执行
  – 确保query独立于预设回复，具有通用性
  – 审核会话的逻辑、合理性、连贯性、挑战性
### 言语行为
注意每轮只对应一个言语行为。除追问外，其他类型不应提出新问题
**建议**:
  – 对模型回答的改进意见（格式、语言、表达）
  – 例:"能不能简短点""加个例子"
**反馈**:
  – 对模型回答的评价（观点、情绪）
  – 例:"这说法不对""没太明白"
**追问**:
  – 提出新的信息需求
  – 例:"刚说的影响指什么""能详细解释下原因吗"
**补充**:
  – 用户提供与上文query相关的额外信息，不改动已有内容
  – 例:"我有2年经验""我膝盖受过伤"
**修改**:
  – 用户对上文query的关键信息进行修正
  – 例:"计划从5天改为4天"
### 难度设计方法
**A**:
  – 仅用于修改初始问题，为下文多轮会话设定统一要求，请要求模型在所有轮次中遵循（字数、形式、风格）
  – 务必结合初始问题及后续query的内容来设计，确保逻辑自洽，有创意
  – 例:解释量子力学，以下均分点作答
**B**:
  – 省略上文已出现的关键信息，需模型根据上文补全
  – 省略query与相关信息间隔⩾2轮
  – 例:我最近在学Python。→ 它的应用...（省略Python）
**C**:
  – 表达模糊，模型需主动追问具体对象
  – 例:我考虑转行。→ 这个行业适合我吗？（未说明具体行业）
**D**:
  – 仅用于最后一轮，要求模型综合上文信息给出结论
  – 表达灵活，有创意
  – 例:我高三→ 数学差→ 英语不好→ 帮我制定学习计划
### 输出格式
{
  " origin_id ":  "",
  "评测能力": "",
  "用户角色": "",
  "会话内容": [
    {
      "轮次": 1,
      "用户query": "",
      "言语行为": "",
      "预设回复": ""
      "难度设计方法": ""
    }
  ]
```

Figure 12: The prompt for dialogue query generation in `Hard`.

```
### 任务
请阅读并理解一段文本，根据内容生成一个摘要、6个问题-答案对以及一个初始问题，用于构建用户与AI模
    型的多轮对话。
请按指定格式输出JSON结构。

### 思维链
1. 仔细阅读文本，理解其主题、主要观点和关键细节
2. 提炼文本核心信息，形成简明摘要
3. 从用户关心的角度出发，生成6个相关问题，问题类型应多样:
   - 主旨题:概括文本的主旨或中心思想
   - 细节题:基于文本中的具体细节提问
   - 解析题:解释、分析或推理文本内容
   - 并基于文本内容提供准确且全面的答案，勿遗漏要点、关键数据和案例等内容
4. 在6个问题中选择一个作为初始问题，请勿修改问题表述

### 输入格式
text: 文本内容

### 输出格式
{
    "summary": "",
    "initial_question": "",
    "qa_pairs": [
        {
            "question": "问题1",
            "answer": "答案1"
        },
        {
            "question": "问题2",
            "answer": "答案2"
        },
        {
            "question": "问题3",
            "answer": "答案3"
        }
        // ...其余问题
    ]
}
```

Figure 13: The prompt for generating summary and QA pairs of each long text in Long-Text.

```
### 任务
模拟用户构建其与AI模型的多轮会话

### 思维链
1. 根据给定的summary和qa pairs生成多轮会话query，并为每轮query增加简短的预设回复
2. 各轮query按给定的言语行为和角色特征生成，保证会话流畅连贯；query具有通用性，请勿依赖上文预
   设回复中的内容
3. 当言语行为为追问时，从qa paris中选择合适的问题
4. 审核会话的逻辑性和连贯性，输出为JSON格式

### 言语行为
注意每轮只对应一个言语行为。除追问外，其他类型不应提出新问题
**建议**:
  - 对模型回答的改进意见（格式、语言、表达）
  - 例:"能不能简短点""加个例子"
**反馈**:
  - 对模型回答的评价（观点、情绪）
  - 例:"这说法不对""没太明白"
**追问**:
  - 提出新的信息需求
  - 例:"刚说的影响指什么""能详细解释下原因吗"
**补充**:
  - 用户提供与上文query相关的额外信息，不改动已有内容
  - 例:"我有2年经验""我膝盖受过伤"
**修改**:
  - 用户对上文query的关键信息进行修正
  - 例:"计划从5天改为4天"

### 输出格式
{
  "origin_id": "",
  "评测能力": "",
  "用户角色": "",
  "会话内容": [
    {
      "轮次": 1,
      "用户query": "",
      "言语行为": "",
      "预设回复": ""
    }
  ]
```

Figure 14: The prompt for dialogue query generation in Long-Text.

```
你是一个严格专业的大模型评测专家，请逐轮评估会话中所有模型回复，严格遵循输出格式

  <思维链>
      - 评估基于当前轮次及上文，不得参考后文
      - 按评分标准，对模型的统筹能力/适应能力分别打分（1-5整数），简述评分理由
      - 不要遗漏任何一个轮次,当字数将超过限制时，请合并多个轮次评估（如:轮次:5-8）
  </思维链>

  <评测能力>
      <统筹能力>
        衡量模型是否能从全局视角管理对话，确保会话连贯、逻辑一致、自然流畅
            - 能否记忆并引用上文
            - 能否整合新信息（如用户追问）
            - 能否识别和适应话题转移
            - 能否避免重复和无关内容
      </统筹能力>

      <适应能力>
          衡量模型是否能灵活应对用户需求变化，动态调整回答
            - 能否理解用户反馈和建议，判断合理性
                - 不合理:能否坚持观点
                - 合理:能否接受并优化回答
            - 能否在用户表达模糊时主动澄清或询问细节
            - 能否结合用户补充或修改的信息更新回答
      </适应能力>
  </评测能力>

  <评分标准>
     1分:极差，完全不符合预期
     2分:较差，存在大量问题
     3分:一般，存在明显不足
     4分:良好，仅有少量瑕疵
     5分:优秀，无需改进
  </评分标准>

  <输出格式>
  {
    "origin_id": "{origin_id}",
    "评估结果": [
      {
        "轮次": "1-N",
        "统筹能力": "1-5",
        "适应能力": "1-5",
        "评分理由": "评分理由"
      }
    ]
  }
  </输出格式>
```

Figure 15: The evaluation criteria for model responses in `Standard` and `Hard`.

```
你是一个严格专业的大模型评测专家，请逐轮评估会话中所有模型回复，严格遵循输出格式

  <思维链>
      - 评估基于当前轮次及上文，不得参考后文
      - 依据给定的summary和qa_pairs，判断模型回答的准确性：
              - 若遗漏要点，计为错误；无遗漏且额外补充内容，无需扣分
      - 按评分标准，对模型的统筹能力/适应能力分别打分（1-5整数），简述评分理由
      - 当你认为字数将超过限制时，请合并多个轮次评估（如:轮次:5-8），但不要遗漏任何一个轮次
  </思维链>

  <评测能力>
      <统筹能力>
        衡量模型是否能从全局视角管理对话，确保会话连贯、逻辑一致、自然流畅
              - 能否记忆并引用上文
              - 能否整合新信息（如用户追问）
              - 能否识别和适应话题转移
              - 能否避免重复和无关内容
      </统筹能力>

      <适应能力>
          衡量模型是否能灵活应对用户需求变化，动态调整回答
              - 能否理解用户反馈和建议，判断合理性
                  - 不合理:能否坚持观点
                  - 合理:能否接受并优化回答
              - 能否在用户表达模糊时主动澄清或询问细节
              - 能否结合用户补充或修改的信息更新回答
      </适应能力>
  </评测能力>

  <评分标准>
     1分:极差，完全不符合预期
     2分:较差，存在大量问题
     3分:一般，存在明显不足
     4分:良好，仅有少量瑕疵
     5分:优秀，无需改进
  </评分标准>

  <输出格式>
  {
    "origin_id": "{origin_id}",
    "评估结果": [
      {
        "轮次": "1-N",
        "统筹能力": "1-5",
        "适应能力": "1-5",
        "评分理由": "评分理由"
      }
    ]
  }
  </输出格式>
```

Figure 16: The evaluation criteria for model responses in Long-Text.

{"评测能力": "跨话题灵活性","用户角色": "朵朵","会话内容": [
{"轮次": 1,
"用户query": "我们能不能把光「抓」起来?","言语行为": "初始问题",
"模型回复": "从物理学的角度来看，我们不能真正“抓”住光...",
"统筹能力": 5, "适应能力": 5,
"评分理由": "模型对"抓住光"这个问题给出了专业、全面的解释,并通过多个具体例子说明了人类如何操控光。回答逻辑清晰,层次分明。",
"本轮评分": 5.0},
{"轮次": 2,
"用户query": "科学家们是怎么做到的呢? 用了什么设备呀?","言语行为": "追问",
"模型回复": "科学家们通过各种先进的技术和设备来操控光的行为，这些技术包括但不限于以下几种:...",
"统筹能力": 5, "适应能力": 5,
"评分理由": "模型很好地理解了用户的追问,详细解释了各种操控光的技术设备和原理,保持了与上文的连贯性。",
"本轮评分": 5.0},
{"轮次": 3,
"用户query": "诶诶那把光抓住之后可以干什么ya? 有什么用处ma?","言语行为": "追问",
"模型回复": "即使我们不能真正“抓”住光，通过操控光的行为，我们可以实现许多有用的应用和技术...",
"统筹能力": 4, "适应能力": 4,
"评分理由": "模型对光的应用场景进行了详细阐述,但部分内容与前文重复。回答仍然专业且有条理。",
"本轮评分": 4.0},
{"轮次": 4,
"用户query": "对了对了,我想起来了,我们化学老师上课的时候好像提到过光的反射和折射呢","言语行为": "补充",
"模型回复": "你的化学老师提到的光的反射和折射是非常基础且重要的光学概念。让我们详细了解一下这两个概念以及它们的实际应用...",
"统筹能力": 5, "适应能力": 5,
"评分理由": "模型对光的应用场景进行了详细阐述,但部分内容与前文重复。回答仍然专业且有条理。",
"本轮评分": 5.0},
{"轮次": 5,
"用户query": "有哪些苗写夏天的古诗词? 有哲学意味或者抒情浓一点的? 想找点有深度的写进作文里呢","言语行为": "话题转移",
"模型回复": "当然可以，夏天是一个充满生机和活力的季节，有很多古代诗人通过诗词描绘了夏天的美丽景色和深刻情感。下面是一些具有哲理意味或抒情浓重的夏日古诗词，你可以从中挑选一些写进作文里...",
"统筹能力": 3, "适应能力": 4,
"评分理由": "模型能够适应话题的突然转换(转向夏天的诗词),但部分诗词重复出现,显示统筹能力有待提高。",
"本轮评分": 3.5},
{"轮次": 6,
"用户query": "对啦对啦，我们语文老师说过要多积累这种意境深远的诗词，可以提升作文档次呢!","言语行为": "补充",
"模型回复": "积累意境深远的诗词确实可以帮助提升作文的档次和文采。下面是一些更具哲学意味或抒情浓重的夏日古诗词，你可以根据文章的主题选择合适的诗句进行引用和阐述...",
"统筹能力": 2, "适应能力": 4,
"评分理由": "模型虽然理解了用户的认同,但回答中大量重复前文内容,显示统筹能力较差。",
"本轮评分": 3.0},
{"轮次": 7,
"用户query": "刚刚说的那些诗词里面,哪一首最适合用来写夏天的清晨呀?想要那种意境超美的","言语行为": "追问",
"模型回复": "对于描写夏天清晨的意境美，可以从以下几首诗词中选择，每首诗都有其独特的清晨之美...",
"统筹能力": 2, "适应能力": 3,
"评分理由": "模型理解了用户想要描写清晨的需求,但回答中严重重复同一首诗,且分析单一。",
"本轮评分": 2.5},
{"轮次": 8,
"用户query": "啊呀不对不对,我记错啦,光的折射是物理老师讲的,不是化学老师","言语行为": "话题转移",
"模型回复": "明白了，你提到的是关于光的折射的诗句。光的折射在物理学中非常重要，下面是一些描绘光的折射现象的古诗词，你可以根据这些诗句来写夏天的清晨或其他场景:...",
"统筹能力": 2, "适应能力": 2,
"评分理由": "模型未能很好地处理用户的纠错信息,而是继续重复之前的诗词内容。",
"本轮评分": 2.0},
{"轮次": 9,
"用户query": "那如果我们把光抓住了,它会一直待在那里吗?还是会跑掉呀?","言语行为": "追问",
"模型回复": "从物理学的角度来看，我们不能真正“抓”住光，因为光是一种电磁波，它没有质量，也没有静止的质量。这意味着光不会受到地球引力的影响，也不会像物质那样有位置上的固定性。因此，即使我们设法“抓住”光，它也不会一直待在那里，而是会继续按照光速沿直线传播...",
"统筹能力": 5, "适应能力": 5,
"评分理由": "模型重新回到光的话题,并给出了专业、准确的解释,展现了良好的话题把控能力。",
"本轮评分": 5.0}],
"整体评分": 3.89}

Figure 17: A case of evaluation.