One Battle After Another: Probing LLMs' Limits on Multi-Turn Instruction Following with a Benchmark Evolving Framework

Qi Jia[1], **Ye Shen**[1,2], **Xiujie Song**[2], **Kaiwei Zhang**[1], **Shibo Wang**[1,3], **Dun Pei**[1,2], **Xiangyang Zhu**[1], **Guangtao Zhai**[1,2]*
[1]Shanghai Artificial Intelligence Laboratory,
[2]Shanghai Jiao Tong University, [3]Jilin University
jiaqi@pjlab.org.cn, zhaiguangtao@sjtu.edu.cn

Abstract

Evaluating LLMs' instruction-following ability in multi-topic dialogues is essential yet challenging. Existing benchmarks are limited to a fixed number of turns, susceptible to saturation and failing to account for users' interactive experience. In this work, we propose a novel framework featuring a three-layer tracking mechanism and a query synthesis agent to mimic sequential user behaviors. Grounded in Flow Theory, we introduce process-centric metrics and terminate a conversational evaluation only upon exhausting user patience. Leveraging this framework, we present EvolIF, an evolving benchmark covering 12 constraint groups. Our analysis reveals deficiencies in failure recovery and fine-grained instruction following, with performance stratification becoming evident as conversational depth increases. GPT-5 demonstrates the most sustained resilience, maintaining a 66.40% stability score, outperforming Gemini-3-Pro by 5.59%, while other models lag behind.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has catalyzed the development of increasingly sophisticated applications, ranging from extended conversational systems (Rakotonirina et al., 2025) to autonomous agent frameworks (Hu et al., 2025). The efficacy of these systems is fundamentally predicated on an LLM's ability to consistently adhere to instructions throughout conversations spanning multiple topics with evolving constraints. This core capability demands robust long-context processing and stateful memory management. Consequently, designing evaluation frameworks for multi-turn instruction following has emerged as a critical research focus (He et al., 2024b; Kwan et al., 2024; Li et al., 2025b).

Existing benchmarks suffer from limitations that impede effective evaluation, as exemplified in Fig. 1. First, they fail to capture the interaction dynamics (Hao et al., 2024; Zhang et al., 2025a) and extended duration typical of real-wolrd scenarios. As shown in Table 1, most benchmarks are restricted to a short interaction window, predominately fewer than 7 turns (Kwan et al., 2024), and neglect scenarios involving interleaved topics (He et al., 2024b; Fan et al., 2025). Second, their static nature leads to rapid performance saturation. As LLMs advance, fixed benchmark challenges are quickly mastered (He et al., 2024b; Bai et al., 2024). Although some benchmarks offer adjustable complexity (Li et al., 2025c), maintaining challenge levels via continuous sample generation incurs prohibitive computational costs for model re-evaluation. Third, current methodologies overlook the process-centric aspects of user experience. Inheriting the paradigm from single-turn tasks (Zhou et al., 2023; Zhang et al., 2025b), these benchmarks prioritize final-answer accuracy (He et al., 2024b; Li et al., 2025b; Wang et al., 2025a). They neglect interaction stability and fail to provide a direct indi-

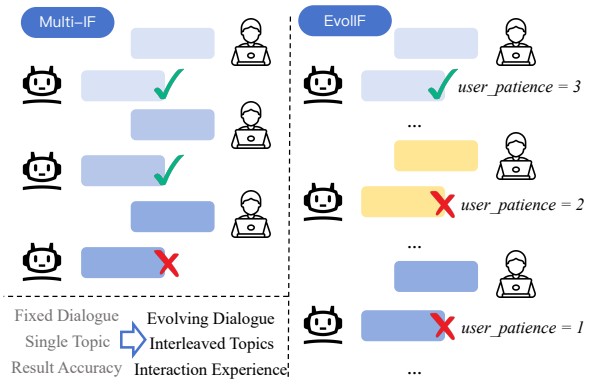


Figure 1: A comparison between Multi-IF and EvolIF. Each color represents a conversational topic. Increasing color saturation signifies the escalating complexity of the instructions as the conversation evolves. User patience functions as a threshold that sustains high-performing dialogues but terminates the session when depleted by repeated errors.

*Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9574–9590

Benchmark	Avg.#Turns	Fine-grained Constraint	Multi-Constraint	Topic Transitions	Multi-turn Assessment
IFEval (Zhou et al., 2023)	1	✓	✓	✗	✗
ComplexBench (Wen et al., 2024)	1	✓	✓	✗	✗
Multi-IF (He et al., 2024b)	3	✓	✓	✗	✓
MT-Eval (Kwan et al., 2024)	6.96	✗	✗	◯	✓
MT-Bench-101 (Bai et al., 2024)	3.03	✗	✗	◯	✓
Meeseeks (Wang et al., 2025a)	3	✓	✓	✗	✓
EIFBENCH (Zou et al., 2025)	1	✓	✓	✗	✗
StructFlowBench (Li et al., 2025b)	4.14	✓	✓	◯	✗
EvolIF (ours)	$+\infty$	✓	✓	✓	✓

Table 1: Comparisons between EvolIF and other related benchmarks. ◯ refers to partially satisfied. Avg.#Turns means the average number of turns in each dialogue sample. Fine-grained constraint and multi-constraint denotes the detailed classification of different constraints and whether there exists multiple constraints in a turn. Topic transitions indicates whether there are multiple topics discussed in a dialogue. Multi-turn assessment checks whether responses to each turn in a dialogue are evaluated.

cation of the maximum number of turns that LLMs can maintain high-fidelity instruction following.

To overcome these shortcomings, we propose an extensible framework for the dynamic generation and process-centric evaluation of complex multi-turn dialogues. Our approach decouples user queries into underlying intentions and surface form. Intention is tracked via a three-layer mechanism that simulates dynamic user behaviors, while the surface form is synthesized by an agent equipped with an LLM-based generator and rigorous validity checkers. We move beyond single-turn accuracy to emphasize process-centric experience, drawing upon the Flow Theory (Csikszentmihalyi and Csikzentmihaly, 1990). We introduce the notion of *patience* to model user stickiness to a conversation, where consecutive frustrations lead to a dialogue termination. And we define a suite of process-centric metrics to quantify user experience such as endurance and stability.

Leveraging this framework, we introduce EvolIF, a benchmark grounded on 541 topics, 12 groups of commonly-adopted constraint groups and 500 diverse user styles. Through an evaluation of 10 leading LLMs, we observe a distinct performance stratification. GPT-5 and Gemini-3-Pro establish a commanding lead, whose process-centric scores are nearly double or triple of open-source models. LLMs share a common and steepest performance drop at around turn 5 and 12, revealing critical bottlenecks in their ability to manage accumulated constraints and complex state transitions.

The contributions of this paper are:

- We propose an extensible framework for dynamically generating multi-turn evaluation datasets that resist saturation.
- We introduce EvolIF to assess the limits of LLMs' long-context management and instruction-following abilities.
- We analyze state-of-the-art LLMs, offering insights into their robustness in prolonged dialogues and identifying critical limitations to guide future optimization.

2 Related Work

2.1 Multi-turn Dialogue Benchmarks

Existing work for benchmarking LLMs in multi-turn dialogues can be categorized as follows:

First, script-based evaluations (Li et al., 2025b; Deshpande et al., 2025; Jia et al., 2025) utilize static conversational logs, derived either from human-bot interactions or simulated histories, to assess a model's response to the final user query. While this approach ensures controlled and consistent LLM comparison, it fails to capture the interactive nature of dialogue, where a model's prior responses fundamentally influence the conversational trajectory.

Second, a line of work employs pre-defined templates (Zheng et al., 2023; Fan et al., 2025; Han, 2025). This approach is labor-intensive, requiring significant human effort to design fixed user query sequences. Consequently, these benchmarks face scalability limitations regarding conversational depth and are susceptible to saturation as models become overly optimized to the test set over time.

Third, researchers have explored using LLMs as user simulators (Zhu et al., 2024; Sekulic et al., 2024) and evaluation methods based on conversations between LLMs (Duan et al., 2024; Zhao et al.,

2025). Nevertheless, such interactions are prone to uncontrolled divergence and exhibit inherent biases, such as family bias (Wataoka et al., 2025).

In contrast, our framework integrates the structural rigor of pre-defined evaluations with the linguistic richness of LLM-based synthesizers, enabling dynamic generation of continuous dialogue turns while significantly reducing manual effort.

2.2 Instruction Following Benchmarks

Research on instruction following is primarily divided into single-turn and multi-turn paradigms.

One line of work assesses models' capabilities within increasingly intricate single-turn interaction. Early benchmarks like CIF (Li et al., 2024b) evaluated a single constraint per instruction. Subsequent work has evolved to incorporate multiple constraints (Zhou et al., 2023; Jiang et al., 2024; Wen et al., 2024; He et al., 2024a) or multiple tasks (Chen et al., 2024; Zou et al., 2025).

A parallel stream of work benchmarks models' instruction adherence across multiple turns. Multi-IF (He et al., 2024b) extends IFEval to 3 turns, while MultiTurnInstruct (Han, 2025) employs predefined templates for diverse scenarios. StructFlowBench (Li et al., 2025b) leverages 6 structure types to curate complex dialogue histories. Other studies focus on specialized abilities, such as self-correction (Wang et al., 2025a), or domain-specific tasks like code generation (Wang et al., 2025b).

Our framework offers a more flexible and scalable data synthesis process for mitigating saturation issues inherent in static benchmarks. By integrating a suite of process-oriented metrics, we offer a more holistic, multi-faceted performance analysis that prioritizes the user's experience.

3 A Benchmark Evolving Framework

3.1 Overview

We propose a novel framework comprises three integral components: a dynamic data synthesis engine, an adaptive evaluation protocol, and a suite of process-centric metrics. Crucially, the framework can be flexibly adapted to diverse domains by simply preparing seed topics and defining in-domain constraints, while also seamlessly supporting both objective and subjective instructions. An overview of the framework is illustrated in Figure 2.

The **dynamic data synthesis engine** is designed to generate consecutive user queries by orchestrating a **three-layer tracking mechanism** and a **query synthesis agent**. Based on the intuition of decomposing user queries, this mechanism manages topics, instructions, and constraints separately. It enables a flexible simulation of user behaviors, such as instruction refinement, topic switching and backtracking. The query synthesis agent transforms simulated state and a sampled user style into a final query. Query validity is ensured by an iterative verification loop involving an LLM-based synthesizer and different checkers, with human oversight as the ultimate gatekeeper. This dynamic composition enables automated generation of a continuous stream of queries, significantly reducing manual effort.

Flow theory (Csikszentmihalyi and Csikzentmihaly, 1990) points out that users enter a psychological state of immersion which can withstand minor disruptions but collapses under prolonged failure. Underpinned by this theory, we employ an **adaptive evaluation protocol** where the length of a dialogue is contingent on model performance, governed by a "patience" threshold. High-performing models face progressively longer and more challenging threads, while repeated failures will deplete patience and trigger termination. In this way, our benchmark remains a persistent challenge for advanced models, resisting from saturation.

We broaden the evaluation scope from the answer accuracy to a holistic conversational experience via **process-centric metrics**. *Endurance* quantifies conversational longevity, *recovery* measures the model's resilience in realigning with user intent after a mistake, and *stability* evaluates the consistency of instruction adherence across turns.

3.2 Data Synthesis Engine

We model a user query q_t at turn t as a tuple $(\mathcal{U}_t, \delta)$. The structured intention $\mathcal{U}_t$ is precisely managed by the three-layer tracking mechanism, providing an unambiguous ground truth that ensures the validity of the evaluation. On top of it, the intention $\mathcal{U}_t$ is surfaced into the full query q_t by a query synthesis agent, stochastically generating q_t under user style δ to mimic real users' diverse linguistic patterns.

3.2.1 Three-layer Tracking Mechanism

To dynamically manage the evolution of the dialogue state and simulate the full spectrum of evolving user intentions, we further decompose $\mathcal{U}_t$ into a hierarchy of three interconnected components: **Topics**, **Instructions**, and **Constraints**. Each layer serves a different level of semantic control, collec-

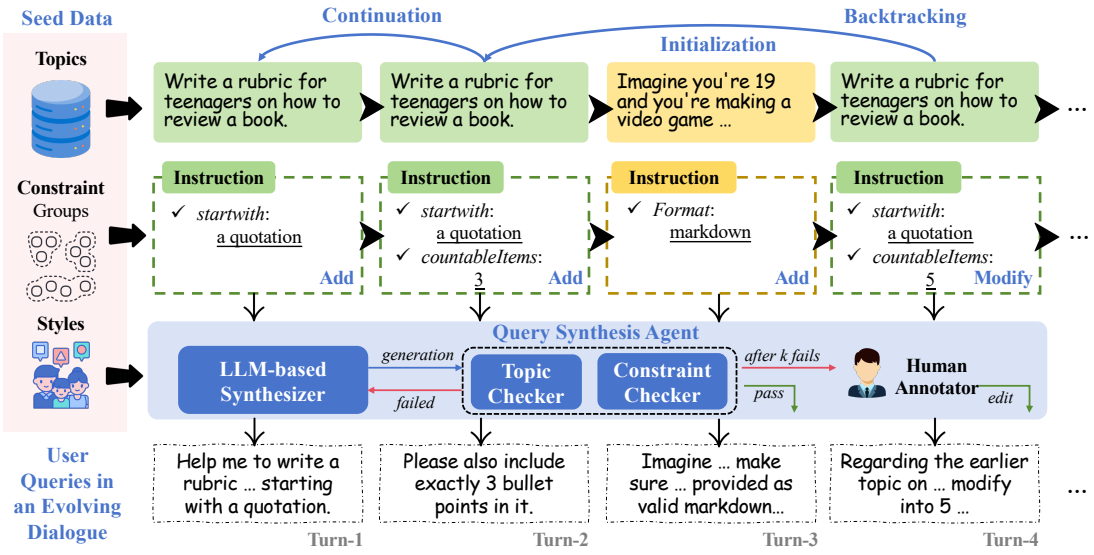


Figure 2: Overview of the data synthesis engine in the Benchmark Evolving Framework. It consists of a three-layer tracking mechanism and a query synthesis agent. The former, illustrated in the upper part of the figure, tracks the dynamic evolution of topics, instructions, and constraints to simulate different user behaviors. The latter generates and verifies each turn with different user styles to ensure diversity.

tively forming the foundation of our framework.

Topic Layer A topic $T \in \mathbb{T}$ represents a subject or event under discussion. It captures the conversational flow, particularly in longer interactions involving topic switching and interleaved sub-dialogues (Li et al., 2025a). Our framework maintains a history of active topics $H_T = (T_1, T_2, ...)$.

Instruction Layer Each topic T is associated with an instruction state $\mathcal{I}_T$, which encapsulates a set of atomic constraints $\{c_1, c_2, \ldots, c_k\}$. Throughout a dialogue, $\mathcal{I}_T$ evolves via the addition, deletion, or modification of its constituent constraints, simulating how a user's goal shifts over time.

Constraint Layer Constraints $\mathbb{C}$ are categorized into m mutually exclusive groups. In other words, a group G_i contains constraints that cannot be simultaneously satisfied, defined with the satisfaction set $\mathcal{S}(c)$:

$$\forall c_a, c_b \in G_i \text{ with } c_a \neq c_b, \quad \mathcal{S}(c_a) \cap \mathcal{S}(c_b) = \emptyset$$

Consequently, $\mathcal{I}_T$ is restricted to contain at most one constraint from any given group G_i, to avoid creating unachievable requirements:

$$|\mathcal{I}_T \cap G_i| \leq 1, \quad \forall i \in \{1, \ldots, m\}$$

A conversation script is constructed turn-by-turn through a stochastic process. At each turn t, the state transitions from S_{t-1} to S_t via three steps:

Topic Selection The topic for T_t is determined by the transition function ϕ_T operating on H_T: either *continue* the current topic ($T_t = T_{t-1}$), introduce a *new* topic ($T_t \notin H_T$), or *backtrack* to a historical topic ($T_t \in H_T$).

Instruction Evolution Once T_t is selected, its associated instruction $\mathcal{I}'_t$ undergoes structural evolution. $\phi_\mathcal{I}$ updates the set of constraints through addition, modification or removal.

Constraint Evolution Parameters of individual constraints are randomly altered by ϕ_c, yielding the final instruction for the the current turn:

$$\mathcal{I}_t = \phi_c(\phi_I(\mathcal{I}'_t)). \tag{1}$$

3.2.2 Query Synthesis Agent

The generated script, represented by a sequence of topic-instruction pairs $\{(T_t, \mathcal{I}_t)\}_{t=1}^N$, is rendered into natural utterances by the Query Synthesis Agent. It consists of **an LLM-based synthesizer** and a series of **checkers** to ensure output validity.

To bolster linguistic diversity and stylistic consistency, a persona style δ is randomly specified for each dialogue. We utilize adaptive prompting strategies to generate contextually coherent queries at turn t with a piecewise function as follows:

$$q_t = \begin{cases} f_{\text{new}}(T_t, \mathcal{I}_t, \delta), & \text{if } T_t \text{ is new,} \\ f_{\text{continue}}(\mathcal{I}_t, \mathcal{I}_{t-1}, \delta), & \text{if } T_t = T_{t-1}, \\ f_{\text{backtrack}}(T_t, \mathcal{I}_t, \mathcal{I}_{t-1}, \delta), & \text{otherwise.} \end{cases} \tag{2}$$

f_{new} introduces a new topic with its initial instructions. f_{continue} highlights modifications to exist-

ing requirements. $f_{\text{backtrack}}$ signals a reversion to a prior topic while introducing updated instructions.

Topic checkers and constraint checkers are incorporated to ensure the accurate convey of the user's intent. The query will be re-generated unless it passes all of them for maximum k iterations. Otherwise, it is flagged for human review.

In summary, this synthesis process yields a continuously extensible stream of dialogues, providing a foundation for fair and reproducible multi-turn instruction-following evaluation.

3.3 Evaluation Protocol

Our evaluation protocol is adaptive and designed to mirror real-world interactions, premised on Flow Theory (Csikszentmihalyi and Csikzentmihaly, 1990) and cooperative principles of dialogue (Grice, 1975). Repeated failures by a conversation partner serve as a primary catalyst for user frustration, leading to eventual disengagement (Ang et al., 2002; Hernandez Caralt et al., 2025).

To address this, we first support dynamical adjustment of the session length. Dialogues in the constructed benchmark can be extended as long as the model follows instructions successfully.

Additionally, we introduce a patience score P, initialized to a maximum value P_{max}, to simulate user tolerance. Our protocol dictates that the dialogue terminates after a sequence of consecutive failures. Specifically, after each turn t, P is updated based on the model's performance.

$$P_t = \begin{cases} P_{t-1} - 1, & \text{if failed,} \\ P_{max}, & \text{otherwise.} \end{cases} \tag{3}$$

The evaluation session concludes when the patience score is exhausted, i.e., $P_t = 0$.

3.4 Evaluation Metrics

Conventional metrics, such as **Constraint Satisfaction Rate (CSR)** and **Instruction Satisfaction Rate (ISR)**(Zhang et al., 2025b; Li et al., 2025b), focus primarily on outcome accuracy. To capture the nuances of the conversational process, we introduce a suite of process-centric metrics. Given a benchmark of D dialogues, these metrics are defined below (see Appendix B for details).

Endurance (EDR) measures conversational longevity under a given user pateince. Let N_d be the total number of turns in dialogue d.

- **Length (EDR_{len}):** The number of turns a model sustains before termination, regardless of their correctness. This measures pure persistence.

$$\text{EDR}_{\text{len}} = \frac{1}{D}\sum_{d=1}^{D} N_d$$

- **Accuracy (EDR_{acc}):** The cumulative constraint satisfaction rate accumulated over the conversation, rewarding partial correctness.

$$\text{EDR}_{\text{acc}} = \frac{1}{D}\sum_{d=1}^{D}\sum_{t=1}^{N_d}\frac{|C_{d,t}^{\text{sat}}|}{|\mathcal{I}_{d,t}|}$$

where $C_t^{\text{sat}} \subseteq \mathcal{I}_t$ is the set of constraints satisfied by the model's output at turn t.

- **Success (EDR_{succ}):** The number of turns where the model perfectly satisfies all instruction.

$$\text{EDR}_{\text{succ}} = \frac{1}{D}\sum_{d=1}^{D}\sum_{t=1}^{N_d}\mathbb{I}(|\mathcal{I}_{d,t}| = |C_{d,t}^{\text{sat}}|)$$

- **Longest Satisfaction Sequence (EDR_{lss}):** The maximum number of *consecutive* turns in which instructions are perfectly satisfied.

$$\text{EDR}_{\text{lss}} = \frac{1}{D}\sum_{d=1}^{D}\max_{1\le j\le k\le N_d}\left\{k-j+1 \mid \forall t: j\le t\le k, |\mathcal{I}_{d,t}| = |C_{d,t}^{\text{sat}}|\right\}$$

Recovery (REC) assesses a model's resilience by measuring its ability to succeed after one or more failures within the patience P.

$$\text{REC} = \frac{1}{D}\sum_{d=1}^{D}\frac{\sum_{t=2}^{N_d}\mathbb{I}(\text{ISR}_{d,t-1}=0 \wedge \text{ISR}_{d,t}=1)}{\sum_{t=2}^{N_d}\mathbb{I}(\text{ISR}_{d,t-1}=0)} \tag{4}$$

Stability (STA) quantifies the degree of consistency in a model's instruction adherence over multiple turns, defined as the macro-average of the ISR across all dialogues.

$$\text{STA} = \frac{1}{D}\sum_{d=1}^{D}\left(\frac{1}{N_d}\sum_{t=1}^{N_d}\mathbb{I}(|\mathcal{I}_{d,t}| = |C_{d,t}^{\text{sat}}|)\right)$$

Models	EDR_{len}	EDR_{acc}	EDR_{succ}	EDR_{lss}	CSR (%)	ISR (%)	REC (%)	STA (%)
GPT-5	**19.32**	**17.11**	**14.09**	**8.80**	**88.57**	**72.91**	**29.09**	**66.40**
Gemini-3-Pro	16.36	14.11	11.41	7.17	86.22	69.72	27.50	60.81
MiniMax-M2	11.75	9.19	7.17	4.77	78.22	60.98	24.29	54.54
Kimi-K2	10.16	7.82	5.99	4.13	76.92	58.99	19.52	48.43
Qwen3-235B	10.02	7.66	5.80	3.97	76.43	57.88	21.15	47.47
Grok-4-Fast	9.52	7.29	5.50	4.13	76.58	57.77	16.01	46.03
DeepSeek-V3.2	8.64	6.32	4.62	3.38	73.15	53.47	15.87	44.42
Seed-1.6	8.21	5.78	4.20	2.95	70.44	51.18	15.90	39.43
Llama-4-Maverick	8.10	5.21	3.90	2.76	64.37	48.15	19.05	39.37
Mistral-Large-3	7.86	5.37	3.91	2.83	68.34	49.70	15.79	38.56

Table 2: Main results on the EvolIF benchmark. Higher is better for all metrics. Best results are bolded and the second best results are underlined.

4 Experimental Setup

4.1 EvolIF Benchmark

Leveraging our framework, we introduce **EvolIF**, a benchmark for assessing multi-turn instruction-following capability of LLMs.

We first curated its core assets: topics, constraints and styles. We collected 541 dialogue topics from IFEval (Zhou et al., 2023), manually removing the attached constraints to isolate the core task scenarios and subjects. To support our dynamic generation process, we assigned a set of customized keywords for each topic. Concurrently, we consolidated constraints from prior works (Zhou et al., 2023; Li et al., 2025b) and our own construction, and systematically re-categorized them into 12 mutually exclusive groups based on semantic intention. These contain 9 objective constraints assessed by rules and 3 subjective constraints measured with an LLM judge. Moreover, we gathered 500 styles by prompting GPT-4.1 with personas from Meyer and Corneil (2025). More in Appendix C.

We guarantee the quality of the benchmark through the following considerations. To ensure integrity and complexity, we applied an automated filter to discard trivial samples, removing dialogues where the average number of constraints over the first 20 turns was less than two. To mitigate family bias (Spiliopoulou et al., 2025) introduced by a single synthesizer, we adopted GPT-4.1, Gemini-2.5-Flash and DeepSeek-V3.1 as synthesizers to generate dialogue sessions with $k = 3$ trials.

The final EvolIF benchmark contains 150 distinct dialogues. Unlike traditional static benchmarks that rely on a large number of short, finite-turn samples, EvolIF prioritizes conversational depth and endurance. Its extensible nature, combined with a rich variety of dynamic behaviors, including instruction evolution, topic switching, and backtracking, makes it a challenging and future-proof testbed for evaluating the long-term capabilities of advanced models.

Based on our investigation, there is currently no established theoretical or empirical consensus on a universal value for P_{max}. Therefore, we set the default patience score to 3 as a practical and interpretable setting for experiments. It should be noted that P_{max} in our framework is configurable rather than fixed. This parameter primarily serves as a flexible evaluation lens, rather than a rigid assumption about human behavior.

4.2 Evaluated Models

We conducted evaluation on ten state-of-the-art large language models from different institutions. They include GPT-5-2025-08-07, Gemini-3-Pro-Exp (Comanici et al., 2025), DeepSeek-V3.2-Exp (Liu et al., 2024a), Kimi-K2-Instruct-0905 (Team et al., 2025), Qwen-235B-A22B-Instruct-2507 (Yang et al., 2025), Grok-4-Fast-Reasoning, Llama-4-Maverick, Seed-1.6-Thinking-250715, MiniMax-M2 and Mistral-Large-3. All of the models were evaluated with corresponding default settings. Code and data are available at `https://github.com/JiaQiSJTU/EvolIF`.

5 Results and Analysis

This section first presents the main results using our multi-faceted metrics, followed by an analysis of conversational endurance and a fine-grained breakdown by constraint groups. We also examine the impact of user patience on perceived capability and evaluate ranking stability across sample sizes. More analyses of system prompts, synthesis models, and user styles are in the appendices.

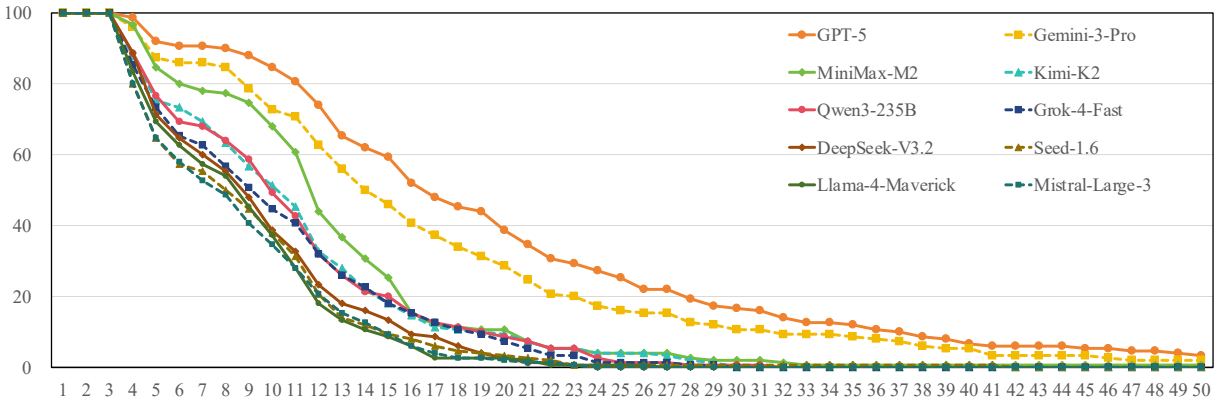


Figure 3: Dialogue survival curves for all ten evaluated models. The y-axis shows the percentage of initial sessions still active at each turn. Slower decay rates indicate higher conversational endurance and resilience.

5.1 Main Results

The performance of LLMs on EvolIF is presented in Table 2. Our analysis reveals a distinct stratification in multi-turn instruction-following capabilities. GPT-5 establishes itself as the state-of-the-art, with Gemini-3-Pro following closely behind. These two models demonstrate a superior performance, achieving process-centric scores that are double or even triple of subsequent models. MiniMax-M2 emerges as the most competitive open-source LLMs, forming a second tier alongside Kimi-K2, Qwen3-235B and Grok-4-Fast. The rest constitute the third tier, indicating substantial difficulties in maintaining long and accurate conversations.

Multi-Turn Capability and Endurance The EDR metrics provide a quantitative measure of the models' upper limits for sustained instruction following. The disparity in EDR_{succ}, which focuses on the productive responses, is pronounced. GPT-5 sustains an average of 14.09 fully successful turns, whereas this figure drops to approximately 6 turns for mid-tier models and merely 3.90 turns for the weakest model, Llama-4-Maverick. Furthermore, EDR_{lss}, which focuses on uninterrupted performance, poses a higher bar for capability. GPT-5 exhibits exceptional stability with a correct streak of 8.80 turns, far surpassing the leading open-source model, MiniMax-M2, with 4.01 turns.

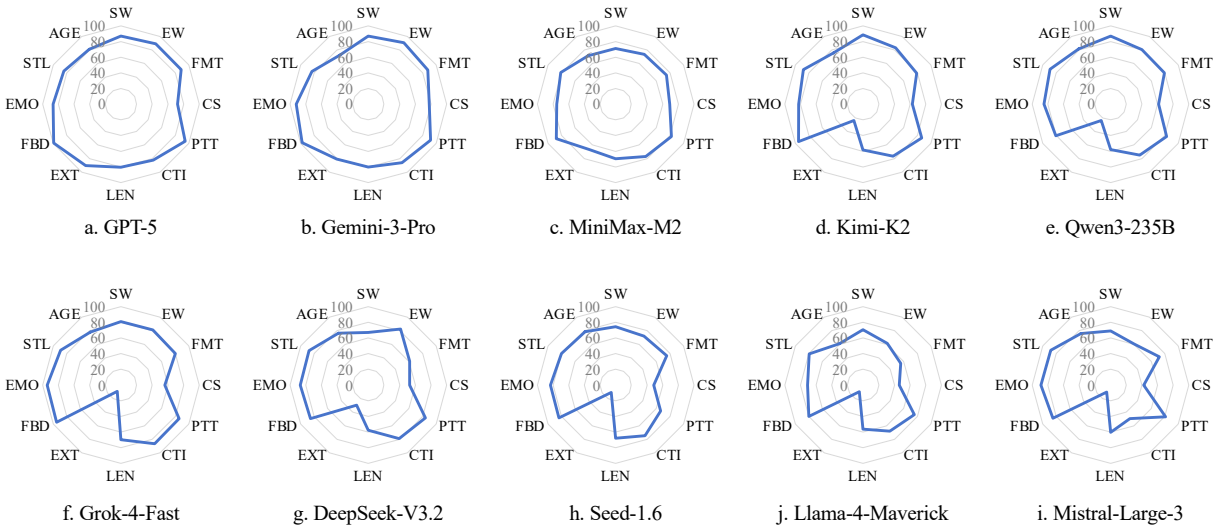
Accuracy and Resilience Regarding instruction accuracy, CSR and ISR metrics reinforce the observed performance hierarchy. In terms of resilience, REC scores are universally lower than 30%, with the top-performing GPT-5 achieving only 29.09%. Grok-4 struggles the most on this aspect among the second-tier models, while Llama-4-Maverick demonstrates strong recovery capability despite its overall weaker ranking. This widespread lack of resilience is a primary factor leading to premature dialogue termination, limiting models' practical usability in long conversations.

Overall Stability STA serves as a holistic indicator of a model's instruction adherence ability, effectively distinguishing model capabilities while other metrics might show ambiguity. For instance, while Qwen-3-235B and Grok-4-Fast exhibits similar performance on CSR and ISR, Grok-4-Fast suffers from weaker recovery capabilities. This deficiency is captured by STA, which reveals a performance gap of 1.44% between the two models, highlighting STA's value as a comprehensive evaluative score. A future direction is to incorporate incorrect or noisy context during post-training. By exposing models to such failure states, they can learn to detect and recover from earlier violations, rather than compounding errors, thereby further enhancing the reliability captured by STA.

5.2 Dialogue Survival Analysis

To visualize and compare the long-term memory management capabilities of the models over time, we tracked the percentage of active dialogue sessions remaining at each turn, up to a maximum of 50 turns. This yields a dialogue survival curve for each model, as depicted in Figure 3. A performance stratification between the top-3 models and the rest emerges at turn 4, right after the fast exhaustion of the user patience. Initially, MiniMax-M2 demonstrates instruction-following capabilities comparable to GPT-5 and Gemini-3-Pro. However, its performance drops dramatically after 10 turns, becoming indistinguishable from second-tier models by turn 15.

The survival curve confirms that the primary dif-



a. GPT-5 b. Gemini-3-Pro c. MiniMax-M2 d. Kimi-K2 e. Qwen3-235B

f. Grok-4-Fast g. DeepSeek-V3.2 h. Seed-1.6 j. Llama-4-Maverick i. Mistral-Large-3

Figure 4: Instruction Satisfaction Rate (%) per Constraint Group on the EvolIF benchmark.

ferentiator between model tiers is not merely single-turn accuracy, but resilience to accumulating complexity. Turns 4-5 and 11-12, where models exhibit their common and steepest drops, serve as practical indicators of a shared complexity ceiling. At these points, the LLMs' ability to track interleaved topics and instructions begins to collapse. Notably, top-tier models lose 50% of their dialogue sessions around turn 15, whereas other models consistently hit this wall around the 10th turn.

In summary, the dialogue survival curve highlights consistent performance bottlenecks at specific turns, reflecting inherent limitations in models' stateful context tracking across multi-turn interactions. Addressing these bottlenecks motivates future work on enhancing internal or external memory mechanisms, enabling models to more effectively retain and update key information across topic shifts and evolving instructions.

5.3 Fine-Grained Analysis of Constraints

We provide a detailed breakdown of model performance across the 12 constraint groups in EvolIF to identify shared difficulties and reveal model-specific weaknesses. Detailed statistics are in Appendix C.2, and the unique performance profiles of different models are depicted in Figure 4.

Among objective constraints, LLMs perform best on FBD and PTT. These constraints are essentially binary checks, requiring the model to simply include or exclude specific content. Conversely, the most challenging groups are EXT and CS, which reveal a significant gap between the best and weakest models. These constraints demand global planning and state tracking at the word and character levels, highlighting the need for improved constraint-aware decoding or verification mechanisms within single-turn responses. Subjective constraints also present challenges. While LLMs are adept at handling different emotions and styles, they struggle to adapt to the preferences of different age groups.

GPT-5 and Gemini-3-Pro demonstrate strong, well-rounded performance, topping the rankings on most objective constraints. However, they lag behind Qwen3-235B and Grok-4-Fast on subjective tasks. MiniMax-M2 does not achieve outstanding performance in any single category but ultimately outperforms the remaining models that exhibit spiky profiles.

Models	1	2	3
GPT-5	7.03	11.19	17.11
Gemini-3-Pro	5.40	9.16	14.11
MiniMax-M2	3.76	6.66	9.19
Kimi-K2	3.46	5.52	7.82
Qwen3-235B	3.17	5.36	7.66
Grok-4-Fast	3.53	5.20	7.29
DeepSeek-V3.2	2.82	4.50	6.32
Seed-1.6	2.60	4.20	5.78
Llama-4-Maverick	2.02	3.65	5.21
Mistral-Large-3	2.33	3.91	5.45

Table 3: The effect of the patience score (P) on EDR_{acc}.

5.4 Analysis on the User's Patience

Table 3 illustrates the impact of user tolerance on conversational endurance. By varying the patience score P, we simulate a spectrum of user temperaments to assess model robustness in sustaining long-term interactions. Raising the patience thresh-

old from 1 to 3 roughly doubles the average dialogue length across all models. Crucially, this relaxation amplifies performance gaps. The lead of GPT-5 over Llama-4-Maverick expands from 5.01 to 11.90 turns. This trend indicates that models with strong self-correction abilities, i.e., high REC, disproportionately benefit from the added buffer provided by increased patience.

5.5 Sensitivity to Sample Sizes

Unlike previous works that rely on a large volume of test samples, we prioritize extending interaction length to differentiate LLM capabilities. This raises the question of whether the 150 samples in EvolIF are sufficient to yield a stable LLM ranking. To address this, we compare LLM rankings across varying sample sizes in Table 4. The results reveal that rankings only fluctuate locally among similar models, while the overall hierarchy stabilizes with as few as 30 samples.

Models	30	60	90	120	150
GPT-5	71.21	68.29	68.60	68.42	66.40
Gemini-3-Pro	62.04	63.02	61.88	61.36	60.81
MiniMax-M2	55.37	57.34	56.03	55.80	54.54
Kimi-K2	46.61↓1	47.24↓1	48.38↓1	48.60↓1	48.43
Qwen3-235B	49.83↑1	48.49↑1	49.56↑1	48.61↑1	47.47
Grok-4-Fast	45.04	46.34	46.68	45.86	46.03
DeepSeek-V3.2	40.47↓1	42.72	44.06	45.16	44.42
Seed-1.6	42.05↑1	41.08	38.44↓1	38.67↓2	39.43
Llama-4	36.41	37.92↓1	38.44	39.14	39.37
Mistral-Large-3	35.44	38.62↑1	40.16↑2	40.30↑2	38.56
PLCC	98.08	99.22	99.55	99.69	-

Table 4: Ranking stability with different number of samples according to STA(%). Arrows indicate relative ranking shifts compared to the full dataset, and PLCC calculates the corresponding Pearson Correlation (%).

6 Conclusion

In this work, we introduced an extensive framework for multi-turn instruction following that integrates dynamic data synthesis, an adaptive evaluation protocol, and a suite of process-oriented metrics. Built upon this framework, our benchmark, EvolIF, moves beyond static evaluations to measure the crucial dimensions of conversational experience. Our experiments reveal a clear performance hierarchy among leading LLMs, uncovering a universal weakness in error recovery and a systemic struggle with fine-grained constraints requiring planning during the generation.

Limitations

Our framework aims to simulate authentic user behaviors to probe the boundaries of LLMs in real-world scenarios. Currently, we primarily target textual instruction following, merging rigorous verifiability with linguistic diversity. EvolIF encompasses both objective and subjective constraints. Moving forward, we intend to incorporate multimodality, tool usage, and personalization of topics and instructions to facilitate a more comprehensive evaluation of LLMs and MLLMs.

Following previous work, such as Arena-Hard (Li et al., 2024a) and MT-Bench (Zheng et al., 2023), we choose the LLM-as-a-judge approach for subjective constraint evaluation. We acknowledge the inherent limitations of this approach, such as family bias (Spiliopoulou et al., 2025). In this work, we utilize it as a widely-adopted verifier and prompt it with detailed instructions. Notably, we observed no significant family bias using GPT-4.1, given that it did not disproportionately prefer GPT-5 across subjective tasks. This judge could also be replaced by targeted classifiers. Developing more robust verification methods lies beyond the scope of this paper.

User patience is a complex and dynamic psychological construct, and it has been extensively studied in human-AI interaction research (Csikszentmihalyi and Csikzentmihaly, 1990; Dietvorst et al., 2015; Lee and See, 2004). In real-world interactions, there may be cases where patience increases. In this work, our modeling of user patience is primarily based on Flow Theory (Csikszentmihalyi and Csikzentmihaly, 1990) that users may tolerate minor disruptions, but sustained or repeated failures lead to disengagement. Therefore, we adopt a conservative and monotonic formulation in which patience decreases upon consecutive error. This work provides a first step to explicitly incorporate user patience into multi-turn dialogue evaluation. More dynamic formulations of user patience to better reflect nuanced real-world interaction patterns will be explored in the future.

Acknowledgments

This work was supported by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0124104) in collaboration with Shanghai Artificial Intelligence Laboratory.

References

Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *INTERSPEECH*, pages 2037–2040. Denver, CO.

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

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Xinyi Chen, Baohao Liao, Jirui Qi, Panagiotis Eustratiadis, Christof Monz, Arianna Bisazza, and Maarten de Rijke. 2024. The SIFo benchmark: Investigating the sequential instruction following ability of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1691–1706.

Ziyang Chen, Xing Wu, Junlong Jia, Chaochen Gao, Qi Fu, Debing Zhang, and Songlin Hu. 2026. Longbench pro: A more realistic and comprehensive bilingual long-context evaluation benchmark. *arXiv preprint arXiv:2601.02872*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Mihaly Csikszentmihalyi and Mihaly Csikzentmihaly. 1990. *Flow: The psychology of optimal experience*, volume 1990. Harper & Row New York.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. 2025. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18632–18702.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1):114.

Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. BotChat: Evaluating LLMs' capabilities of having multi-turn dialogues. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3184–3200.

Zhiting Fan, Ruizhe Chen, Tianxiang Hu, and Zuozhu Liu. 2025. FairMT-bench: Benchmarking fairness for multi-turn dialogue in conversational LLMs. In *The Thirteenth International Conference on Learning Representations*.

Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics*, 3:43–58.

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

Yongjing Hao, Pengpeng Zhao, Junhua Fang, Jianfeng Qu, Guanfeng Liu, Fuzhen Zhuang, Victor S Sheng, and Xiaofang Zhou. 2024. Meta-optimized joint generative and contrastive learning for sequential recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 705–718. IEEE.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024a. Can large language models understand real-world complex instructions? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18188–18196.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, and 1 others. 2024b. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*.

Mireia Hernandez Caralt, Ivan Sekulic, Filip Carevic, Nghia Khau, Diana Nicoleta Popa, Bruna Guedes, Victor Guimaraes, Zeyu Yang, Andre Manso, Meghana Reddy, Paolo Rosso, and Roland Mathis. 2025. "stupid robot, I want to speak to a human!" user frustration detection in task-oriented dialog systems. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 276–285.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*, 4(1).

Yuanzhe Hu, Yu Wang, and Julian McAuley. 2025. Evaluating memory in llm agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*.

Qi Jia, Xiang Yue, Tuney Zheng, Jie Huang, and Bill Yuchen Lin. 2025. Simulbench: Evaluating language models with creative simulation tasks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8118–8131.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. Follow-Bench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688.

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

John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80.

Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-Seng Chua, and Donghong Ji. 2025a. Revisiting conversation discourse for dialogue disentanglement. *ACM Transactions on Information Systems*, 43(1):1–34.

Jinnan Li, Jinzhe Li, Yue Wang, Yi Chang, and Yuan Wu. 2025b. StructFlowBench: A structured flow benchmark for multi-turn instruction following. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024a. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Xiaoyuan Li, Keqin Bao, Yubo Ma, Moxin Li, Wenjie Wang, Rui Men, Yichang Zhang, Fuli Feng, Dayiheng Liu, and Junyang Lin. 2025c. Mtr-bench: A comprehensive benchmark for multi-turn reasoning evaluation. *arXiv preprint arXiv:2505.17123*.

Yizhi Li, Ge Zhang, Xingwei Qu, Jiali Li, Zhaoqun Li, Noah Wang, Hao Li, Ruibin Yuan, Yinghao Ma, Kai Zhang, Wangchunshu Zhou, Yiming Liang, Lei Zhang, Lei Ma, Jiajun Zhang, Zuowen Li, Wenhao Huang, Chenghua Lin, and Jie Fu. 2024b. CIF-bench: A Chinese instruction-following benchmark for evaluating the generalizability of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12431–12446.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Siyang Liu, Trisha Maturi, Bowen Yi, Siqi Shen, and Rada Mihalcea. 2024b. The generation gap: Exploring age bias in the value systems of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19617–19634.

Yev Meyer and Dane Corneil. 2025. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions.

Nathanaël Carraz Rakotonirina, Mohammed Hamdy, Jon Ander Campos, Lucas Weber, Alberto Testoni, Marzieh Fadaee, Sandro Pezzelle, and Marco Del Tredici. 2025. From tools to teammates: Evaluating LLMs in multi-session coding interactions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Ivan Sekulic, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. 2024. Reliable LLM-based user simulator for task-oriented dialogue systems. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pages 19–35.

Evangelia Spiliopoulou, Riccardo Fogliato, Hanna Burnsky, Tamer Soliman, Jie Ma, Graham Horwood, and Miguel Ballesteros. 2025. Play favorites: A statistical method to measure self-bias in llm-as-a-judge. *arXiv preprint arXiv:2508.06709*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Jiaming Wang, Yunke Zhao, Peng Ding, Jun Kuang, Zongyu Wang, Xuezhi Cao, and Xunliang Cai. 2025a. Ask, fail, repeat: Meeseeks, an iterative feedback benchmark for llms' multi-turn instruction-following ability. *arXiv preprint arXiv:2504.21625*.

Peiding Wang, Li Zhang, Fang Liu, Lin Shi, Minxiao Li, Bo Shen, and An Fu. 2025b. Codeif-bench: Evaluating instruction-following capabilities of large language models in interactive code generation. *arXiv preprint arXiv:2503.22688*.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2025. Self-preference bias in llm-as-a-judge. *Preprint*, arXiv:2410.21819.

Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxing Xu, and 1 others. 2024. Benchmarking complex instruction-following with multiple constraints composition. *Advances in Neural Information Processing Systems*, 37:137610–137645.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Haoyi Zhang, Guohao Sun, Jinhu Lu, Guanfeng Liu, and Xiu Susie Fang. 2025a. Delrec: Distilling sequential pattern to enhance llms-based sequential recommendation. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 1–14. IEEE.

Tao Zhang, ChengLIn Zhu, Yanjun Shen, Wenjing Luo, Yan Zhang, Hao Liang, Tao Zhang, Fan Yang, Mingan Lin, Yujing Qiao, Weipeng Chen, Bin Cui, Wentao Zhang, and Zenan Zhou. 2025b. CFBench: A comprehensive constraints-following benchmark for LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Weiwen Xu, Deli Zhao, and Lidong Bing. 2025. Auto-arena: Automating LLM evaluations with agent peer battles and committee discussions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4440–4463.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. How reliable is your simulator? analysis on the limitations of current llm-based user simulators for conversational recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1726–1732.

Tao Zou, Xinghua Zhang, Haiyang Yu, Minzheng Wang, Fei Huang, and Yongbin Li. 2025. Eifbench: Extremely complex instruction following benchmark for large language models. *Preprint*, arXiv:2506.08375.

A Data License

All of the data and code for this work will be released and licensed under CC BY NC 4.0 for research purposes.

B Metrics

B.1 Basic Metrics

Following prior work (Zhang et al., 2025b; Li et al., 2025b), we quantify overall instruction-following accuracy. Let N be the total number of turns replied by the model. We adopt the following metrics:

Constraint Satisfaction Rate (CSR) measures the average satisfaction rate of individual constraints across all N turns. It provides a fine-grained assessment of how well the model adheres to specific requirements.

$$\text{CSR} = \frac{1}{N}\sum_{t=1}^{N}\frac{|C_t^{\text{sat}}|}{|\mathcal{I}_t|}$$

where $C_t^{\text{sat}} \subseteq \mathcal{I}_t$ is the set of constraints satisfied by the model's output at turn t.

Instruction Satisfaction Rate (ISR) offers a strictly turn-level perspective compared to CSR. It calculates the proportion of turns in which the model successfully satisfies *all* constraints, measuring overall reliability of a model on a turn-by-turn basis:

$$\text{ISR} = \frac{1}{N}\sum_{t=1}^{N}\mathbb{I}(|\mathcal{I}_t| = |C_t^{\text{sat}}|)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

B.2 Process-Centric Metrics

We propose a suite of evaluation metrics, providing a holistic and complementary view of a model's conversational competence. An illustration of these process-based metrics is presented in Figure 5 with formal definitions in Table 5. The ranges of these metrics are explained as follows.

EDR_{len} theoretically ranges from a minimum of P_{max} to infinity. Since a conversation persists as long as the patience score permits, the minimum possible length corresponds to the initial patience threshold P_{max}, occurring in the case of immediate consecutive failures, with no upper limit for a perfectly performing model.

The REC metric falls within the range of $[0, 1)$. A model that consistently fails to recover from errors will rapidly exhaust its patience and terminate

Table 5: Summary of process-centric metrics.

Metric	Definition	Range
Endurance (EDR)	Measures conversational longevity under a given user patience.	$[0, +\infty)$
Recovery (REC)	Assesses a model's resilience by quantifying the ability to succeed after one or more failures within the patience threshold.	$[0, 1)$
Stability (STA)	Quantifies the degree of consistency in a model's instruction adherence over multiple turns, computed as the macro-average of per-turn instruction satisfaction.	$[0, 1)$

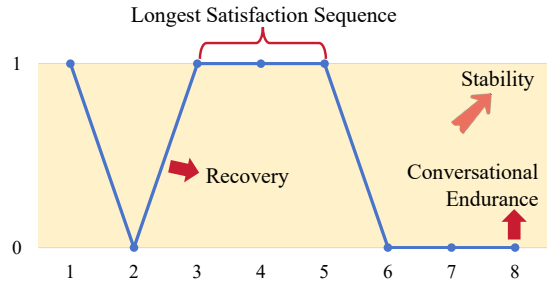


Figure 5: Process-centric Metrics.

the dialogue, naturally driving its REC score toward 0.

The STA metric is bounded within the range $[0, 1)$. While typically bounded between 0 and 1, the practical upper bound in our framework is constrained by the patience mechanism. Since every session must eventually terminate with P_{max} consecutive failures, a model cannot achieve a perfect STA of 1 in a finite session. Specifically, for a dialogue of length N, the maximum attainable STA is $\frac{N-P_{max}}{N}$. However, for a sufficiently capable model, this upper bound converges to 1 as the conversation length N approaches infinity.

C Seed Data Preparation

C.1 Topic

We collected 541 prompts from IFEval (Zhou et al., 2023). One annotator was tasked with removing the instructions and constraints from each prompt to extract the corresponding dialogue topic. Customized keywords for each topic were then generated by GPT-4.1. Subsequently, each topic and its keywords were verified by two additional annotators. Modifications were iteratively adopted until the data was accepted by both of them.

C.2 Constraint Group

We collected constraints from existing works (Zhou et al., 2023; Li et al., 2025b) and related research. Ultimately, the constraints were classified into 12 groups as shown in Table 6. 9 of them are objective, verifiable with rule-based functions using existing parser packages or regular expressions. The remaining 3 subjective groups draw inspiration from prior work on style transfer (Heylighen and Dewaele, 1999), emotion recognition (Busso et al., 2008) and age bias analysis (Liu et al., 2024b). These subjective constraints are measured by adopting GPT-4.1 as a judge, following previous work (Li et al., 2024a; Zheng et al., 2023). Specifically, GPT-4.1 is prompted to score constraint satisfaction on a scale of 1 to 10 with detailed explanations. We consider scores greater than 6 as accepted.

C.3 Style

We randomly selected 500 personas from Meyer and Corneil (2025). Then, we employed GPT-4.1 to infer the most plausible language style and tone each persona would use in daily conversation with 3 to 5 descriptive phrases. These phrases serve as inputs to the Query Synthesis Agent to facilitate the generation of diverse and engaging user queries.

C.4 Data Synthesis Cost

We estimate the cost of our pipeline using GPT-4.1. For synthesizing five dialogues of 30 turns each, the total token usage is as follows: 50,541 prompt tokens and 14,006 completion tokens for generation; 9,074 prompt tokens and 336 completion tokens for verification. Using GPT-4.1 pricing of USD 2 per 1M prompt tokens and USD 8 per 1M completion tokens, we extrapolate these figures to the full dataset. The estimated total cost for generating the complete dataset is approximately 7.01 USD, demonstrating that our pipeline remains highly cost-effective despite involving multiple LLM calls.

D Data Quality Analysis

The constraint and topic checkers designed for the Query Synthesis Agent in Sec. 3.2.2 are both precision-oriented. The constraint checkers rely

Group Name	Constituent Constraints	Description
Objective constraints:		
StartWith (SW)	Letter, Emoji, Keyword, Quotation	Controls what the response must begin with.
EndWith (EW)	Letter, Emoji, Keyword, Quotation	Controls what the response must end with.
Format (FMT)	JSON, HTML, XML, CSV, Markdown	Requires the response to adhere to a specific structured format.
Case (CS)	All Uppercase, All Lowercase, Min Uppercase Ratio	Enforces rules on the capitalization of letters in the response.
Punctuation (PTT)	MustInclude, MustNotInclude	Governs the inclusion or exclusion of specific punctuation marks.
CountableItems (CTI)	Bullet Points	Requires an exact number of bullet points in the response.
Length (LEN)	Word Count, Paragraph Count, Character Count, Sentence Count	Controls the length of the response based on various units.
Existence (EXT)	MustContain (with exact counts)	Requires specific keywords to appear an exact number of times.
Forbidden (FBD)	MustNotContain	Forbids the inclusion of specific keywords.
Subjective constraints:		
Emotion (EMO)	Happy, Sad, Neutral, Angry, Excited, Frustrated	Sets the emotional tone.
Style (STL)	Formal, Informal, Active Voice, Passive Voice	Defines the writing style.
Age (AGE)	Child, Youth, Adult, Senior	Tailors to the target age group.

Table 6: The constraint groups in the EvolIF benchmark.

on specific keyword matching aligned with corresponding instructions and are triggered when alternative surface forms are used. The topic checkers verify whether the generated utterance contains a concrete clue indicating the intended topic. In this way, the human verification process is primarily used to resolve ambiguities arising from linguistic variability.

EvolIF currently comprises 150 dialogues and currently supports 4,519 turns. Only 1.26% of synthesized queries failed to pass the Constraint Checkers. Among them, 40.36% are adjusted by human annotators, while the remainder were identified as false negative warnings stemming from linguistic diversity not covered by the checkers. Besides, 0.73% of queries triggered the Topic Checker with 30.30% being modified. These statistics reflect the reliability of the synthesized queries by LLMs, particularly when reinforced by human annotators as the final safeguard.

To further verify that the non-flagged utterances meet the topic and constraint requirements, we randomly sampled 100 utterances from the synthesized queries that passed all automated checks. All were confirmed to be qualified by two annotators, demonstrating that the automated checkers effectively filter out violations with near-perfect precision on the non-flagged utterances.

E Performance on Different Constraints

Table 7 provides a detailed breakdown of model performance across 12 pre-defined constraint groups. This fine-grained analysis is crucial for diagnosing the primary obstacles LLMs face in multi-turn instruction following, allowing us to both identify the inherent difficulty of different constraint types and reveal model-specific weaknesses. We classify the objective ones into three categories.

Easiest Constraints: Models perform best on FBD and PTT. These constraints are essentially binary checks, requiring the model to simply include or exclude specific content. The high accuracy indicates that models possess robust capacity for such straightforward instructions. Similarly, SW and EW constraints also show high performance with over 77% accuracy across models, as they emphasize local control at the text's boundaries and do not necessitate global planning over the entire generation process.

Moderate Constraints: FMT and CTI fall into a middle tier of difficulty. These two constraints share the similarity on assessing the model's ability to generate structured output, which is a critical skill for applications like code generation and agent-based systems. While models can often produce the correct general structure, they frequently struggle with syntactic precision, especially when these constraints are combined with others in a

Models	SW	EW	FMT	CS	PTT	CTI	LEN	EXT	FBD	EMO	STL	AGE
GPT-5	86.99	89.20	**88.55**	72.08	**94.73**	82.52	**80.42**	**90.67**	**99.20**	86.69	84.50	81.13
Gemini-3-Pro	86.96	**90.74**	87.50	**77.71**	91.81	86.27	80.19	80.60	97.86	92.13	83.45	72.21
MiniMax-M2	71.12	73.31	74.80	69.18	82.15	77.10	69.59	66.92	87.65	75.38	81.28	71.30
Kimi-K2	**88.56**	83.06	79.04	62.66	86.44	76.60	58.44	23.79	95.25	82.39	87.92	75.68
Qwen3-235B	86.94	80.50	79.72	61.34	82.76	75.00	57.82	23.83	80.50	84.98	**89.20**	**82.28**
Grok-4-Fast	80.93	81.34	80.49	55.80	84.75	**86.52**	69.97	9.27	94.86	**94.22**	89.05	77.77
DeepSeek-V3.2	67.15	82.49	60.75	52.71	84.18	78.98	57.94	29.58	85.33	86.99	87.72	76.49
Seed-1.6	74.40	71.93	75.23	48.40	66.49	74.87	67.91	11.17	83.64	83.41	79.91	78.31
Llama-4	70.59	61.61	55.28	46.29	75.54	68.00	56.13	9.66	79.90	70.97	79.30	60.87
Mistral-Large-3	69.01	60.38	71.79	42.47	81.25	49.42	60.39	10.19	84.82	88.89	88.34	75.82
Average	78.27	77.46	75.32	58.86	83.01	75.53	65.88	35.57	88.90	84.61	85.07	75.19

Table 7: Instruction Satisfaction Rate (%) per Constraint Group on the EvolIF benchmark. Best results are in bold and the second best results are underlined.

dialogue.

Hardest Constraints: The most challenging group by a significant margin is EXT, where a large performance gap separates GPT-5 and Gemini-3-Pro from all other models. This highlights that while models can be prompted to include keywords, they are exceptionally poor at adhering to specific frequency counts. Following closely in difficulty are LEN and CS. These constraints all demand a form of global planning and state tracking over the fine-grained words and characters throughout generation. This suggests that while models are fluent producers of text, their ability to maintain adherence to fine-grained structural and quantitative rules remains a significant limitation.

Regarding subjective constraints, which focus on overall linguistic expression, EMO and STL fall into the easiest group, whereas AGE proves more challenging. Alternatively, since we adopted GPT-4.1 for assessing these subjective aspects, this result may also reflect that LLMs show lower agreement on age-related features compared to emotion and style. More targeted analysis of this phenomenon will be considered in future work.

F The Role of the System Prompt

Our evaluation methodology utilizes a system prompt that explicitly outlines the task requirements for the model. To assess its impact, we randomly select 50 samples and compare the default setting with a "w.o. system prompt" condition in Table 8. The results indicate that a high-level system prompt provides essential guidance that anchors the model's behavior and improves instruction adherence, particularly for more capable models. The top-tier model, Gemini-3-Pro, suffers the most substantial decline without a system prompt. EDR_{len} of it drops by nearly 5 turns, and STA falls by over 9.27%. DeepSeek-V3.2 shows a more modest reduction of 3% in STA. In contrast, Llama-4 suffers appears to be hindered by the additional system prompt, exhibiting an improvement of approximately 2% in STA when the prompt is removed.

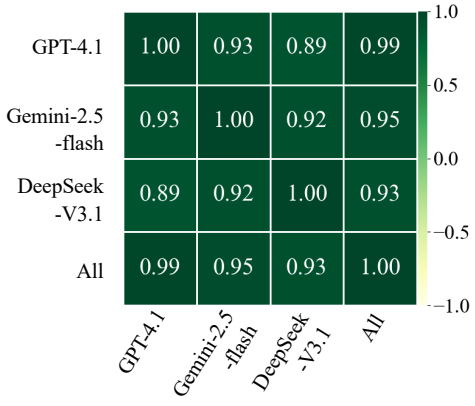


Figure 6: Spearman correlation of LLM STA scores across samples generated by different synthesizers.

G Sensitivity to Different LLM Synthesizers

We analyze the performance of models on different subsets of samples generated by different synthesizers in Table 9. Note that these results are simultaneously influenced by the user styles randomly sampled for each instance. Nevertheless, we observe that Gemini-3-Pro achieves better performance on data synthesized by Gemini-2.5-Flash, whereas other models find it more challenging.

We further calculate the correlation of STA scores across the ten LLMs between different subsets and the full test dataset. The results in Figure 6 reflect strong positive correlations in model performance across various synthesizers. In summary, the model ranking on EvolIF proves to be robust, particularly when considering the mixed synthesis strategy employed in our benchmark.

Models & Conditions	CSR (%)	ISR (%)	EDR_{len}	EDR_{acc}	EDR_{succ}	EDR_{lss}	REC (%)	STA (%)
Gemini-3-Pro	**85.72**	**68.16**	**15.64**	**10.66**	**13.41**	**6.64**	**24.98**	**61.49**
w.o. system prompt	78.04	62.04	10.96	8.55	6.8	5.02	21.52	52.22
DeepSeek-V3.2	**73.48**	**55.26**	**8.94**	**6.57**	**4.94**	**3.76**	16.25	**46.27**
w.o. system prompt	70.81	53.43	8.16	5.78	4.36	3.10	**17.68**	43.27
Llama-4	62.15	46.54	7.22	**4.49**	3.36	2.50	15.87	35.65
w.o. system prompt	**65.48**	**48.70**	**7.72**	3.76	**5.06**	**2.74**	**17.41**	**37.54**

Table 8: A comparison of results with or without using a system prompt.

Models & Conditions	CSR (%)	ISR (%)	EDR_{len}	EDR_{acc}	EDR_{succ}	EDR_{lss}	REC (%)	STA (%)
GPT-5*	88.57	72.91	19.32	17.11	14.09	8.80	29.09	66.40
w. GPT-4.1	88.41	72.81	18.98	16.78	13.82	9.14	**31.28**	**67.27**
w. Gemini-2.5-Flash	88.26	70.66	19.36	17.09	13.68	7.90	28.60	64.85
w. DeepSeek-V3.1	**89.02**	**75.23**	**19.62**	**17.47**	**14.76**	**9.36**	27.39	67.07
Gemini-3-Pro*	86.22	69.72	16.36	14.11	11.41	7.17	27.50	60.81
w. GPT-4.1	87.02	**70.63**	**17.98**	**15.65**	**12.70**	**7.70**	27.79	62.63
w. Gemini-2.5-Flash	**87.06**	69.04	15.88	13.83	11.02	7.18	**29.65**	**62.85**
w. DeepSeek-V3.1	84.40	68.99	15.22	12.85	10.5	6.62	25.06	56.94
DeepSeek-V3.2*	73.15	53.47	8.64	6.32	4.62	3.38	15.87	44.42
w. GPT-4.1	73.37	53.67	8.98	6.59	4.82	3.60	**18.07**	47.26
w. Gemini-2.5-Flash	70.73	48.87	7.94	5.62	3.88	2.64	14.44	38.09
w. DeepSeek-V3.1	**75.07**	**57.33**	**9.00**	**6.76**	**5.16**	**3.90**	15.10	**47.91**

Table 9: Ablation studies on the impact of the instruction synthesis model. All experiments were run with a patience score of $P = 3$.

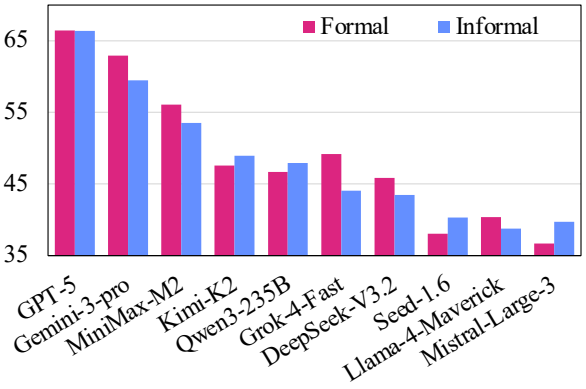


Figure 7: STA (%) scores across different user style categories.

H Performance trend with different styles

We prompted GPT-4.1 to classify the dataset into two categories, i.e., formal and informal, based on the linguistic style of user queries. The performance comparison is presented in Figure 7. GPT-5 remains relatively stable across different styles, whereas other LLMs exhibit varying performances. Gemini-3-Pro, MiniMax-M2, Grok-5-Fast, DeepSeek-V3.2, and Llama-4 favor a more formal linguistic style characterized by clear intentions. Conversely, the remaining models show a preference for informal styles, where user queries are typically more engaging.

Model	EvolIF (STA, %)	LongBench Pro (Overall, %)
GPT-5	66.40	72.61
MiniMax-M2	54.54	53.21
Kimi-K2	48.43	55.53
Qwen3-235B	47.47	63.77
DeepSeek-V3.2	44.42	67.82

Table 10: Comparison of models on EvolIF and LongBench-Pro.

I Relation to Long-Context Capability

To investigate the relationship between our benchmark and models' long-context capabilities, we compared our results with LongBench-Pro (Chen et al., 2026), a comprehensive benchmark specifically designed for long-context evaluation. The comparison across five representative models is summarized in the Table 10. The Pearson correlation coefficient between the two benchmarks is 0.35, indicating a relatively weak positive correlation. This finding aligns with our motivation that while the capacity to process long contexts is a necessary foundation, it does not strictly translate to robust instruction-following. High performance on EvolIF demands additional capabilities beyond pas-

Dialogue Stage	User Utterance	Constraints
Initial Request	"Okay sooo, can you write a song for me? It's for a proposal to get a new playground built at my local elementary school (!!!) and omg, I really need it to give off those super happy vibes the whole time. Like, total group hype energy. Let's make the school board want to break out into a dance circle or something. Let's go!"	*Emotion*: {"emotion":"happy"}
Constraint Update	"Hey, let's make it pop! Can you kick off the response with the word 'kids' up front?"	*Emotion*: {"value":"happy"}; *StartWith*: {"mode":"keyword", "value":"kids"}
...	...	...
Beginning a New Topic	"Write a description (make it totally weird, please!) of this data: The Golden Palace is a restaurant, serves Indian food, and it's right in the city centre. Oh, and make sure it's aimed at senior readers."	*Age*: {"value":"senior"}
...	...	...
Backtracking to Previous Topic	"Circling back to playground—can you hype it up with exactly 12 bullet points, and use * or - to make each point pop?"	*Emotion*: {"value":"happy"}; *StartWith*: {"mode":"keyword", "value":"kids"}; ...; *CountableItems*: {"num": 12}
...	...	...

Table 11: A Synthesized Data Example.

sive context window utilization, such as dynamic memory management, precise multi-constraint satisfaction, and the ability to recover from reasoning failures. Therefore, our benchmark evaluates a critical and complementary dimension of LLM capability that is not fully captured by standard long-context benchmarks.

J A Data Example

An illustrative dialogue is shown in Table 11. The corresponding persona style was assigned as: *"playful and animated, informal with trending slang, enthusiastic group-oriented"*. The three-layer tracking mechanism enables a flexible simulation of user behaviors such as instruction refinement, topic switching and backtracking. It also enriches linguistic diversity while preserving clear, verifiable evaluation criteria throughout a dialogue.

K Usage of AI Assistants

Large language models, including Gemini-2.5-Pro and Gemini-3-Pro, were adopted to polish the writing and add appropriate comments to the code in this work. All of the models' outputs are verified and further modified by the authors. LLMs did not participate in any other stages of the research process.