Intent Mismatch Causes LLMs to Get Lost in Multi-Turn Conversation

Geng Liu[1,2], Fei Zhu[2], Rong Feng[1,2], Changyi Ma[3], Shiqi Wang[1], Gaofeng Meng[2]
[1]College of Computing, City University of Hong Kong, Hong Kong SAR, China.
[2]Centre for Artificial Intelligence and Robotics, HKISI, CAS
[3]School of Artificial Intelligence, Jilin University, China
gengliu6@my.cityu.edu.hk, zhfei2018@gmail.com

Abstract

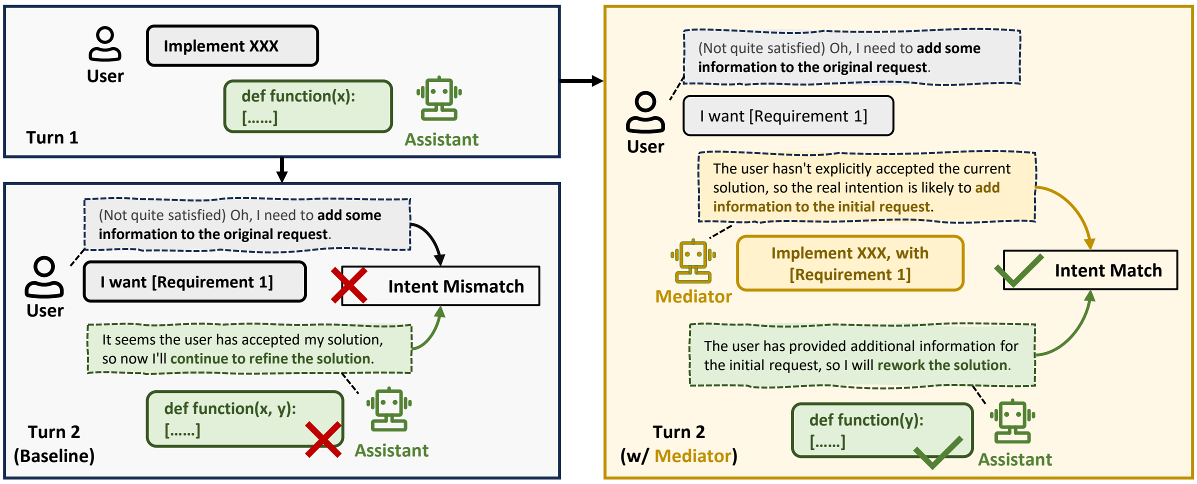
Multi-turn conversation has emerged as a predominant interaction paradigm for Large Language Models (LLMs). Users often employ follow-up questions to refine their intent, expecting LLMs to adapt dynamically. However, recent research (Laban et al., 2025) reveals that LLMs suffer a substantial performance drop in multi-turn settings compared to single-turn interactions with fully specified instructions, a phenomenon termed “Lost in Conversation” (LiC). While this prior work attributes LiC to model unreliability, we argue that the root cause lies in an *intent alignment gap* rather than intrinsic capability deficits. In this paper, we first demonstrate that LiC is not a failure of model capability but rather a breakdown in interaction between users and LLMs. We theoretically show that scaling model size or improving training alone cannot resolve this gap, as it arises from structural ambiguity in conversational context rather than representational limitations. To address this, we propose to decouple intent understanding from task execution through a Mediator-Assistant architecture. By utilizing an experience-driven Mediator to explicate user inputs into explicit, well-structured instructions based on historical interaction patterns, our approach effectively bridges the gap between vague user intent and model interpretation. Experimental results demonstrate that this method significantly mitigates performance degradation in multi-turn conversations across diverse LLMs.

1 Introduction

In contemporary AI-assisted applications, multi-turn dialogue has become the primary mode of interaction between users and large language models (LLMs). Thanks to their massive parameter scales and extensive pretraining on diverse corpora, modern LLMs now exhibit impressive capabilities in language understanding, reasoning, and task execution, and they often perform remarkably well when given clear, complete, and well-structured instructions in a single turn. However, real-world user behavior rarely conforms to this idealized setting. In practice, users frequently start with vague, underspecified, or even internally inconsistent goals, and only gradually clarify and refine their true needs through an iterative conversational process with the model (Zamfirescu-Pereira et al., 2023; Min et al., 2020). This incremental, exploratory nature of human problem formulation poses substantially greater challenges for LLMs than standard single-turn benchmarks: the model must not only understand and solve the current subtask, but also continually infer, update, and realign with a moving target of user intent across turns.

Recent research (Laban et al., 2025) presents a set of controlled experiments designed to simulate the instruction underspecification that frequently occurs in human conversation (Herlihy et al., 2024; Zipf, 1949). The study systematically compares performance under “single-turn, fully specified” (Full) versus “multi-turn, underspecified” (Sharded) interactions, revealing a substantial performance degradation of approximately 30% for all evaluated LLMs. The authors argue that under incomplete information, LLMs tend to make premature assumptions early in the dialogue and subsequently “lock in” these assumptions, causing the final responses to drift away from the user’s true intent. They term this phenomenon “Lost in Conversation” (LiC) and primarily attribute it to the reduced reliability of LLMs in multi-turn dialogue. On this basis, they advocate that LLMs should natively support multi-turn interaction and that model builders should jointly optimize models’ aptitude and reliability in iterative conversational settings.

In this work, we revisit this phenomenon and offer a different explanatory perspective. We argue that: (1) **Making early assumptions and providing tentative answers is not simply erroneous**



(Not quite satisfied) Oh, I need to add some information to the original request.

It seems the user has accepted my solution, so now I'll continue to refine the solution.

(Not quite satisfied) Oh, I need to add some information to the original request.

The user hasn't explicitly accepted the current solution, so the real intention is likely to add information to the initial request.

The user has provided additional information for the initial request, so I will rework the solution.

Figure 1: **Intent Mismatch in Multi-turn Dialogue.** (Left) The LiC benchmark simulates passive users who act as "lazy" interlocutors, omitting corrections for erroneous model assumptions. This behavior causes the Assistant's interpretation to progressively drift away from the user's true intent, leading to significant performance degradation. (Right) Our approach introduces a Mediator to bridge this pragmatic gap by fundamentally decoupling intent inference from task execution. The Mediator aligns the Assistant with the user's true goals, effectively mitigating performance degradation.

behavior, but a rational strategy induced by the dominant training objective of being helpful (Ouyang et al., 2022) and the penalty often associated with evasive responses in RLHF pipelines. Under conditions of incomplete information, the model is inclined to construct a plausible task formulation for a typical user and produce a provisional answer based on that formulation, instead of repeatedly refusing to answer or endlessly requesting additional information. (2) **The primary bottleneck in failed multi-turn conversations is not a lack of model capacity or reasoning depth, but a pragmatic mismatch between user expression and model interpretation** (Figure 1 left). Users exhibit systematic individual variation, where the same utterance may map to disparate underlying intentions. General-purpose LLMs, aligned to the "average" user, fail to adapt to these idiosyncratic behaviors. For instance, models frequently misinterpret a user's fragmentary continuation as a confirmation of previous assumptions rather than a correction, thereby reinforcing an incorrect context.

To address this, we propose a framework that fundamentally decouples intent understanding from task execution. We operationalize this through a Mediator-Assistant pipeline, where a Mediator explicates user inputs to explicitly articulate latent requirements before they reach the execution Assistant. To align with specific user pragmatics, we employ an LLM-based Refiner to automatically distill explicit guidelines by analyzing the discrepancies between failed and successful interaction trajectories. These guidelines then serve as context for the Mediator, enabling the system to bridge the alignment gap and adapt to individual user behaviors without the need for weight updates.

Our approach directly addresses the root cause of LiC: the misalignment between how users express intent and how models interpret it (Figure 1 right). By bridging this gap through adaptive input rewriting, we demonstrate substantial recovery of multi-turn performance across diverse LLMs, highlighting the critical role of user-aware intent modeling in conversational AI.

2 Related Works

Multi-turn Dialogue Evaluation. Recent benchmarks for multi-turn dialogue, such as MT-Bench (Zheng et al., 2023), MT-Bench-101 (Bai et al., 2024), and LOCOMO (Maharana et al., 2024), primarily focus on either (i) sequential task decomposition (e.g., planning a trip over multiple steps) or (ii) long-context retention in extended conversations. However, these settings often assume that each turn is sufficiently specified or that the full task context is available early in the dialogue. In contrast, our work targets a more challenging regime: *incremental intent revelation*, where the user's goal is only partially observable at each turn and may contradict earlier model assumptions—a scenario systematically studied in the "Lost in Conversation" (LiC) framework (Laban et al., 2025) but largely overlooked by existing benchmarks.

Clarification and Intent Disambiguation. Another line of research encourages LLMs to actively seek clarification when faced with ambiguous queries (Li, 2025; Herlihy et al., 2024). While effective in controlled settings, such approaches often conflict with real-world helpfulness norms, as users typically expect immediate, provisional responses rather than repeated clarification requests. As argued in Section 1, premature assumption-making is a rational outcome of prevailing training objectives. Instead of modifying the model's behavior, we preserve its default helpfulness while correcting misinterpretations upstream, through an adaptive mediator that refines inputs into unambiguous and complete instructions.

Personalization in LLMs. A growing body of work explores personalizing LLMs via parameter-efficient fine-tuning (PEFT) (Xu et al., 2023), user-specific adapters (Zhong et al., 2021), or memory-augmented architectures. Systems like Mem0 (Chhikara et al., 2025), A-Mem (Xu et al., 2025), and MemoryBank (Zhong et al., 2024) store user facts to enable long-term contextual awareness. However, these approaches primarily address *factual* personalization (e.g., remembering user preferences) rather than *pragmatic* alignment, which is the challenge of interpreting ambiguous utterances according to a user's idiosyncratic expression style.

3 Problem Analysis

We formulate the multi-turn interaction between a user and an LLM assistant within a latent variable framework. Let $I_t \in \mathcal{I}$ denote the user's deep intent (the specific goal) at turn t, and $T \in \mathcal{T}$ represent the user's expression habits and pragmatic patterns. In the t-th turn of interaction, the input to the LLM is the accumulated dialogue context C_t, consisting of the sequence of historical user utterances and assistant responses:

$$C_t = (u_1, a_1, u_2, a_2, \dots, u_t). \tag{1}$$

The user's current utterance u_t is generated via a stochastic process $u_t \sim P_{\text{user}}(u \mid I_t, T, C_{t-1})$. Crucially, this process acts as a lossy projection, where complex, high-dimensional intents are compressed into low-dimensional and often ambiguous surface forms. The LLM, defined by parameters θ, aims to generate a response R conditioned on the observed context:

$$R \sim P_\theta(R \mid C_t). \tag{2}$$

3.1 Performance Decomposition

We posit that the model's performance in multi-turn scenarios is not a monolithic metric but can be theoretically decomposed into two orthogonal components: (1) Intent Inference, the ability to recover the true intent I_t from C_t; and (2) Task Execution, the ability to solve the identified intent I_t. Assuming that the true intent I_t is a sufficient statistic for the task such that R becomes conditionally independent of the noisy context C_t given I_t (i.e., $P_\theta(R \mid I_t) \approx P_\theta(R \mid I_t, C_t)$), we derive:

$$P_\theta(R \mid C_t) = \sum_{I_t \in \mathcal{I}} \underbrace{P_\theta(R \mid I_t)}_{\text{Execution}} \cdot \underbrace{P_\theta(I_t \mid C_t)}_{\text{Inference}}. \tag{3}$$

This decomposition reveals the fundamental mechanism behind performance degradation in multi-turn dialogues. The execution capability $P_\theta(R \mid I_t)$ represents the model's intrinsic reasoning ability given a perfectly defined instruction. This is largely determined by the model's pre-training and remains relatively stable for a specific task. However, the accuracy of intent inference $P_\theta(I_t \mid C_t)$ faces severe challenges as interactions progress. In an ideal single-turn setting, users tend to provide self-contained descriptions, making the intent I_t clearly inferable from the utterance. Conversely, in multi-turn dialogues, driven by the *principle of least effort* and individual habits (T), users often generate highly personalized, ambiguous, and fragmented surface forms (e.g., using pronouns or vague directives) based on the same intent I_t. This pragmatic ellipsis significantly widens the semantic gap between the surface utterance and the deep intent. Consequently, the LLM fails not because it loses the capability to solve the problem, but because it cannot penetrate the user's ambiguous expression to clearly define the problem.

3.2 The Information Bottleneck

The challenge of maximizing $P(I_t \mid C_t)$ is not merely a lack of model capacity, but an information-theoretic limit. Mathematically, the conditional entropy of the user's intent given the context, $H(I_t \mid C_t)$, remains high because the mapping from intent to utterance is many-to-one. When critical constraints are omitted by the user due to lossy compression, the missing information bits are simply absent from C_t.

Under such high uncertainty, a frozen LLM P_θ tends to revert to its training priors. It implicitly solves for $\arg\max_{I_t} P_{\text{pretrain}}(I_t \mid C_t)$, aligning with the "average user" rather than the spe-