Novelty Assessment Report

Paper: LLMs Get Lost In Multi-Turn Conversation
PDF URL: https://openreview.net/pdf?id=VKGTGGcwl6
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Large Language Models (LLMs) are conversational interfaces. As such, LLMs have the potential to assist their users not only when they can fully specify the task at hand, but also to help them define, explore, and refine what they need through multi-turn conversational exchange. Although analysis of LLM conversation logs has confirmed that underspecification occurs frequently in user instructions, LLM evaluation has predominantly focused on the single-turn, fully-specified instruction setting. In this work, we perform large-scale simulation experiments to compare LLM performance in single- and multi-turn settings. Our experiments confirm that all the top open- and closed-weight LLMs we test exhibit significantly lower performance in multi-turn conversations than single-turn, with an average drop of 39% across six generation tasks. Analysis of 200,000+ simulated conversations decomposes the performance degradation into two components: a minor loss in aptitude and a significant increase in unreliability. We find that LLMs often make assumptions in early turns and prematurely attempt to generate final solutions, on which they overly rely. In simpler terms, we discover that when LLMs take a wrong turn in a conversation, they get lost and do not recover.

Core Task Landscape

This paper addresses: **Multi-Turn Underspecified Conversation Performance Evaluation**
A total of **50 papers** were analyzed and organized into a taxonomy with **17 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Benchmark Design and Evaluation Frameworks**
- **Ambiguity and Underspecification Handling**
- **Performance Degradation and Error Analysis**
- **Training and Optimization Methods**
- **Conversational Modeling Approaches**
- **Domain-Specific Applications**

Complete Taxonomy Tree

- Multi-Turn Underspecified Conversation Performance Evaluation Survey Taxonomy
- Benchmark Design and Evaluation Frameworks
  - General Multi-Turn Dialogue Benchmarks (8 papers)
  - [2] MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback (Wang, 2023) View paper
  - [3] AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents (Junxian He, 2024) View paper
  - [7] Beyond Single-Sentence Prompts: Upgrading Value Alignment Benchmarks with Dialogues and Stories (Zhang Yazhou, 2025) View paper
  - [15] MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models (Jiang Xin, 2024) View paper
  - [19] MMDU: A Multi-Turn Multi-Image Dialog Understanding Benchmark and Instruction-Tuning Dataset for LVLMs (Tao Chu, 2024) View paper
  - [26] ConvBench: A Multi-Turn Conversation Evaluation Benchmark with Hierarchical Capability for Large Vision-Language Models (Liu Shuo, 2024) View paper
  - [36] MultiVerse: A Multi-Turn Conversation Benchmark for Evaluating Large Vision and Language Models (Lee, 2025) View paper
  - [42] Dynamic benchmarking framework for LLM-based conversational data capture (Zietkiewicz, 2025) View paper
  - Domain-Specific Benchmarks (6 papers)
  - [5] Crmarena-pro: Holistic assessment of llm agents across diverse business scenarios and interactions (Huang, 2025) View paper
  - [8] LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultation Conversation (Li Haitao, 2025) View paper
  - [22] C3: A Bilingual Benchmark for Spoken Dialogue Models Exploring Challenges in Complex Conversations (Chengqian Ma, 2025) View paper
  - [23] Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-world Multi-turn Dialogue (Songhua Yang, 2023) View paper
  - [24] CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling (Chenhao Zhang, 2024) View paper
  - [27] An Automatic Evaluation Framework for Multi-turn Medical Consultations Capabilities of Large Language Models (Liao Yusheng, 2023) View paper
  - Tool-Use and Agent Interaction Benchmarks (3 papers)
  - [21] BIRD-INTERACT: Re-imagining Text-to-SQL Evaluation for Large Language Models via Lens of Dynamic Interactions (Huo, 2025) View paper
  - [30] ACEBench: Who Wins the Match Point in Tool Usage? (Chen Chen, 2025) View paper

- [41] UserBench: An Interactive Gym Environment for User-Centric Agents (Qian Cheng, 2025) View paper
- Retrieval-Augmented Generation Evaluation (2 papers)
- [1] Benchmarking Poisoning Attacks against Retrieval-Augmented Generation (Zhang Bao-lei, 2025) View paper
- [40] MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems (Danilevsky, 2025) View paper

- Ambiguity and Underspecification Handling
  - Clarification Question Generation (5 papers)
  - [11] InfoQuest: Evaluating Multi-Turn Dialogue Agents for Open-Ended Conversations with Hidden Context (BrandÃ£o, 2025) View paper
  - [18] CLAM: Selective Clarification for Ambiguous Questions with Generative Language Models (Kuhn, 2022) View paper
  - [20] Learning to Clarify: Multi-turn Conversations with Action-Based Contrastive Self-Training (Chen, 2024) View paper
  - [39] Ask-to-Clarify: Resolving Instruction Ambiguity through Multi-turn Dialogue (Lin Xing-yao, 2025) View paper
  - [43] Identifying & Interactively Refining Ambiguous User Goals for Data Visualization Code Generation (Ä°nan, 2025) View paper
  - Query Resolution and Rewriting (2 papers)
  - [4] Query resolution for conversational search with limited supervision (Voskarides, 2020) View paper
  - [37] Learning Contextual Retrieval for Robust Conversational Search (Yang, 2025) View paper
  - Visual and Multi-Modal Ambiguity Resolution (4 papers)
  - [13] nvBench 2.0: Resolving Ambiguity in Text-to-Visualization through Stepwise Reasoning (Luo Tianqi, 2025) View paper
  - [28] Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning (Ni, 2024) View paper
  - [29] Repairs in a Block World: A New Benchmark for Handling User Corrections with Multi-Modal Language Models (Suglia, 2024) View paper
  - [38] Multi-Turn Multi-Modal Question Clarification for Enhanced Conversational Understanding (Yuan Yifei, 2025) View paper
- Performance Degradation and Error Analysis ★ (3 papers)
  - [0] LLMs Get Lost In Multi-Turn Conversation (Anon et al., 2026) View paper
  - [14] LLM Defenses Are Not Robust to Multi-Turn Human Jailbreaks Yet (Li, 2024) View paper
  - [31] Verifiable Accuracy and Abstention Rewards in Curriculum RL to Alleviate Lost-in-Conversation (Ming, 2025) View paper
- Training and Optimization Methods
  - Reinforcement Learning and Policy Optimization (2 papers)
  - [6] On overcoming miscalibrated conversational priors in llm-based chatbots (Herlihy, 2024) View paper
  - [9] CPO: Addressing Reward Ambiguity in Role-playing Dialogue via Comparative Policy Optimization (Wang Rui, 2025) View paper
  - Supervised and Multi-Task Learning Approaches (2 papers)
  - [25] Data-Centric Improvements for Enhancing Multi-Modal Understanding in Spoken Conversation Modeling (Maximillian Chen, 2024) View paper
  - [46] ContextQFormer: A New Context Modeling Method for Multi-Turn Multi-Modal Conversations (Lei Yiming, 2025) View paper
  - Collaborative and Multi-Agent Training (1 papers)
  - [34] Collaborative Multi-Agent Dialogue Model Training Via Reinforcement Learning (Alexandros Papangelis, 2019) View paper
- Conversational Modeling Approaches
  - Context and State Representation (2 papers)
  - [17] User Intent and State Modeling in Conversational Systems (Ye, 2024) View paper
  - [35] Better Semantic Understanding in LLM-Based Multi-Turn Dialogues: A Survey (Nana Li, 2025) View paper
  - Multi-Modal Conversational Systems (2 papers)
  - [10] Improving situated conversational agents with step-by-step multi-modal logic reasoning (Y Long, 2023) View paper
  - [12] Enhancing Troubleshooting Task-Oriented Dialog Systems with Large Language Models (Jiahao Zhou, 2024) View paper
  - Task-Oriented and Goal-Driven Systems (3 papers)
  - [32] A Concurrent Intelligent Natural Language Understanding Model for an Automated Inquiry System (Gokul Sunilkumar, 2022) View paper
  - [47] Emotionally Intelligent Task-oriented Dialogue Systems: Architecture, Representation, and Optimisation (Feng, 2025) View paper
  - [48] Beyond Task-Oriented and Chitchat Dialogues: Proactive and Transition-Aware Conversational Agents (Yejin Yoon, 2025) View paper
  - Spatial and Embodied Dialogue Systems (1 papers)
  - [33] Talk-to-Resolve: Combining scene understanding and spatial dialogue to resolve granular task ambiguity for a collocated robot (Pramanick, 2022) View paper
- Domain-Specific Applications
  - Mental Health and Counseling Applications (2 papers)
  - [49] MoPHES:Leveraging on-device LLMs as Agent for Mobile Psychological Health Evaluation and Support (Wei Xun, 2025) View paper
  - [50] Reasoning Is Not All You Need: Examining LLMs for Multi-Turn Mental Health Conversations (Chandra, 2025) View paper
  - Other Specialized Domains (3 papers)
  - [16] Act2P: LLM-Driven Online Dialogue Act Classification for Power Analysis (Zhangwenbo Zhangwenbo, 2025) View paper
  - [44] C-MTCSD: A Chinese Multi-Turn Conversational Stance Detection Dataset (Niu Fuqiang, 2025) View paper
  - [45] Pluralistic Behavior Suite: Stress-Testing Multi-Turn Adherence to Custom Behavioral Policies (Sreedhar, 2025) View paper

Narrative

Core task: multi-turn underspecified conversation performance evaluation. This field examines how conversational systems handle extended interactions where user intent is incomplete, ambiguous, or evolving across turns. The taxonomy organizes research into six main branches. Benchmark Design and Evaluation Frameworks (e.g., AgentBoard[3], ConvBench[26]) establish standardized testbeds for measuring multi-turn capabilities. Ambiguity and Underspecification Handling (e.g., Query Resolution[4], Ask to Clarify[39]) focuses on methods for detecting and resolving unclear user requests through clarification strategies. Performance Degradation and Error Analysis investigates how and why systems fail as conversations lengthen, including adversarial scenarios like Multi Turn Jailbreaks[14]. Training and Optimization Methods (e.g., CPO[9], Verifiable Accuracy Rewards[31]) develop techniques to improve model robustness in extended dialogues. Conversational Modeling Approaches explores architectural choices for maintaining context and coherence, while Domain-Specific Applications (e.g., Zhongjing[23], CPsyCoun[24]) adapt these techniques to specialized settings like healthcare or customer service.

A central tension emerges between proactive clarification strategies and passive context accumulation: some works emphasize explicit question-asking to resolve ambiguity early (InfoQuest[11], Question Clarification[38]), while others focus on implicit context modeling that infers intent from dialogue history (ContextQFormer[46], MTRAG[40]). Lost In Conversation[0] sits squarely within the Performance Degradation and Error Analysis branch, examining how conversational systems deteriorate over extended interactions. Its emphasis on diagnosing failure modes complements nearby work like Verifiable Accuracy Rewards[31], which addresses degradation through training-time interventions, and Multi Turn Jailbreaks[14], which explores adversarial vulnerabilities. Where these neighbors focus on mitigation or exploitation of weaknesses, Lost In Conversation[0] provides systematic analysis of when and why multi-turn underspecification leads to breakdowns, offering diagnostic insights that inform both benchmark design and optimization strategies across the broader landscape.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. LLM Defenses Are Not Robust to Multi-Turn Human Jailbreaks Yet

Authors: Li, Nathaniel, Han, Ziwen, Zhang, et al. (11 authors total) | **Year/Venue**: 2024 • arXiv.org | **URL**: View paper

Abstract

Recent large language model (LLM) defenses have greatly improved models' ability to refuse harmful queries, even when adversarially attacked. However, LLM defenses are primarily evaluated against automated adversarial attacks in a single turn of conversation, an insufficient threat model for real-world malicious use. We demonstrate that multi-turn human jailbreaks uncover significant vulnerabilities, exceeding 70% attack success rate (ASR) on HarmBench against defenses that report single-digit A...

Relationship Analysis

Both papers belong to the Performance Degradation and Error Analysis category, examining how conversational systems fail across multiple turns. They overlap in studying multi-turn conversation vulnerabilities, with the original paper focusing on performance degradation from underspecified instructions through simulation experiments, while the candidate paper examines security vulnerabilities through adversarial human jailbreak attacks. The key difference is that the original paper analyzes general task performance decline in underspecified conversations, whereas the candidate paper specifically targets defense mechanisms against malicious multi-turn attacks.

---

2. Verifiable Accuracy and Abstention Rewards in Curriculum RL to Alleviate Lost-in-Conversation

Authors: Li Ming | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large Language Models demonstrate strong capabilities in single-turn instruction following but suffer from Lost-in-Conversation (LiC), a degradation in performance as information is revealed progressively in multi-turn settings. Motivated by the current progress on Reinforcement Learning with Verifiable Rewards (RLVR), we propose Curriculum Reinforcement Learning with Verifiable Accuracy and Abstention Rewards (RLAAR), a framework that encourages models not only to generate correct answers, but ...

Relationship Analysis

Both papers belong to the Performance Degradation and Error Analysis category, examining how conversational systems fail across multiple turns. They share overlapping focus on the 'Lost in Conversation' phenomenon where LLMs exhibit performance degradation in multi-turn underspecified settings, with both analyzing reliability issues and premature answer generation. The key difference is that the original paper provides a diagnostic analysis through large-scale simulation experiments to characterize the degradation, while the candidate paper proposes a solution method (RLAAR) using curriculum reinforcement learning with abstention rewards to mitigate the identified problem.

Contributions Analysis

Overall novelty summary. The paper investigates performance degradation in multi-turn underspecified conversations through large-scale simulation experiments across six generation tasks. It resides in the 'Performance Degradation and Error Analysis' leaf, which contains only three papers total, making this a relatively sparse research direction within the broader taxonomy of 50 papers. The two sibling papers address adversarial vulnerabilities (Multi Turn Jailbreaks) and training-time mitigation (Verifiable Accuracy Rewards), whereas this work focuses on systematic diagnostic analysis of failure modes. This positioning suggests the paper targets an underexplored niche: empirical characterization of how and why LLMs fail in extended underspecified dialogues.

The taxonomy reveals substantial activity in adjacent areas. The 'Benchmark Design and Evaluation Frameworks' branch contains 19 papers across four leaves, including general dialogue benchmarks (8 papers) and domain-specific evaluations (6 papers). The 'Ambiguity and Underspecification Handling' branch (13 papers) addresses clarification strategies and query resolution, representing a complementary perspective focused on mitigation rather than diagnosis. The paper's analytical approach bridges these areas: it evaluates performance degradation (its home leaf) while examining underspecification handling (a neighboring branch), but does so through diagnostic lens rather than proposing new clarification mechanisms or benchmarks.

Among 30 candidates examined across three contributions, none were identified as clearly refuting the work. The sharded simulation environment examined 10 candidates with no refutations; the aptitude-unreliability decomposition framework examined 10 candidates with no refutations; and the large-scale empirical study examined 10 candidates with no refutations. This suggests that within the limited search scope, the specific combination of simulation-based methodology, performance decomposition framework, and scale of empirical analysis (200,000+ conversations) appears distinctive. However, the search examined only top-30 semantic matches, leaving open whether more exhaustive literature review might surface closer prior work in simulation methodologies or decomposition frameworks.

Based on the limited search scope of 30 candidates, the work appears to occupy a relatively novel position at the intersection of performance analysis and underspecification handling. The sparse population of its home leaf (3 papers) and absence of refuting candidates suggest the diagnostic framing and decomposition approach may be distinctive contributions. However, the analysis cannot rule out relevant prior work outside the top-30 semantic matches, particularly in adjacent areas like benchmark design or training optimization where methodological overlaps might exist.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Sharded simulation environment for multi-turn underspecified conversations

Description: The authors develop a simulation framework that transforms single-turn instructions into sharded instructions, revealing information gradually across conversation turns. This enables large-scale evaluation of LLMs in multi-turn, underspecified settings using existing benchmarks.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. OpenDeception: Benchmarking and Investigating AI Deceptive Behaviors via Open-ended Interaction Simulation

URL: View paper

Brief Assessment

OpenDeception[65] focuses on evaluating deception behaviors in LLM agents through multi-turn dialogue simulation, not on creating frameworks for underspecified conversation evaluation. The technical focus differs fundamentally from the original paper's sharded instruction methodology.

---

2. User simulation in task-oriented dialog systems based on large language models via in-context learning

URL: View paper

Brief Assessment

User Simulation[60] focuses on task-oriented dialogue systems where users have specific goals within defined domains, not on transforming single-turn instructions into multi-turn underspecified conversations for general LLM evaluation.

---

3. DialSim: A Dialogue Simulator for Evaluating Long-Term Multi-Party Dialogue Understanding of Conversational Agents

URL: View paper

Brief Assessment

DialSim[66] focuses on evaluating conversational agents' comprehension in multi-party dialogues from TV shows, not on transforming single-turn instructions into gradually-revealed sharded instructions for underspecified task evaluation.

---

4. Dynamic evaluation with cognitive reasoning for multi-turn safety of large language models

URL: View paper

Brief Assessment

Dynamic Safety Evaluation[63] focuses on safety evaluation of LLMs through cognitive reasoning and dynamic prompt generation, not on creating simulation environments for underspecified conversations. The candidate addresses safety assessment rather than general multi-turn conversation simulation frameworks.

---

5. Contextualized Evaluations: Judging Language Model Responses to Underspecified Queries

URL: View paper

Brief Assessment

Contextualized Evaluations[68] focuses on evaluating language model responses to underspecified queries by providing context during evaluation, not on creating simulation environments for multi-turn conversations. The candidate's approach involves generating follow-up question-answer pairs to clarify underspecified queries for evaluation purposes, which differs from the original paper's sharded simulation framework that transforms single-turn instructions into multi-turn conversations for testing LLM performance degradation.

---

6. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models

URL: View paper

Brief Assessment

Math LLaVA[61] focuses on multimodal mathematical reasoning through data synthesis for vision-language models, not on simulation environments for evaluating multi-turn underspecified conversations with language models.

---

7. Flipping the dialogue: Training and evaluating user language models

URL: View paper

Brief Assessment

Flipping Dialogue[64] focuses on training user language models to simulate human users in conversations with assistants, not on creating simulation environments for evaluating multi-turn underspecified conversations with sharded instructions as the original paper does.

---

8. Multi-turn Evaluation of Anthropomorphic Behaviours in Large Language Models

URL: View paper

Brief Assessment

Anthropomorphic Behaviours[67] focuses on evaluating anthropomorphic behaviors in LLMs through multi-turn interactions, not on creating simulation environments for underspecified conversations. Their multi-turn evaluation uses role-playing user simulations with specific conversational principles, distinct from the sharding approach that gradually reveals instruction information.

---

9. Automated Safety Evaluations Across 20 Large Language Models: The Aymara LLM Risk and Responsibility Matrix

URL: View paper

Brief Assessment

Aymara Safety Matrix[62] focuses on safety evaluation across policy domains using adversarial prompts, not on simulating multi-turn underspecified conversations or sharding instructions across turns.

---

10. MATRIX: Multi-Agent simulaTion fRamework for safe Interactions and conteXtual clinical conversational evaluation

URL: View paper

Brief Assessment

MATRIX[69] focuses on safety-oriented evaluation of clinical dialogue systems using patient simulation, not on general multi-turn underspecified conversation frameworks or sharding methodologies for benchmark transformation.

Contribution 2: Decomposition of performance degradation into aptitude and unreliability

Description: The authors introduce metrics to separate LLM performance drops into aptitude (best-case capability) and unreliability (variance across runs). They find that multi-turn degradation stems primarily from increased unreliability rather than aptitude loss.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Psychometric Personality Shaping Modulates Capabilities and Safety in Language Models

URL: View paper

Brief Assessment

Psychometric Personality Shaping[70] focuses on how personality trait prompting affects LLM safety and capability benchmarks, not on decomposing performance variance into aptitude versus unreliability components across conversation turns.

---

2. Variability, Its Limits, and the Performanceâ□□Competence Debate: Implications of Linguistic Variability for a Theory of Grammar

URL: View paper

Brief Assessment

Variability Limits[76] focuses on linguistic variability in natural language and speaker competence, not on decomposing LLM performance metrics into aptitude and unreliability components for multi-turn conversations.

---

3. ChatGPT on the Road: Leveraging Large Language Model-Powered In-vehicle Conversational Agents for Safer and More Enjoyable Driving Experience

URL: View paper

Brief Assessment

ChatGPT Road[79] focuses on in-vehicle conversational agents for driving safety and user experience, not on decomposing LLM performance metrics into aptitude and unreliability components.

---

4. Incoherent Beliefs & Inconsistent Actions in Large Language Models

URL: View paper

Brief Assessment

Incoherent Beliefs[78] focuses on belief updating consistency and action-belief alignment in sequential settings, not on decomposing performance metrics into aptitude and unreliability components across conversation turns.

---

5. Do large language models show human-like biases? exploring confidenceâ□□competence gap in ai

URL: View paper

Brief Assessment

Confidence Competence Gap[71] examines self-assessment biases in LLMs (confidence vs. correctness alignment), not performance decomposition into aptitude and unreliability components across multiple runs as defined in the original paper.

---

6. ERGO: Entropy-guided Resetting for Generation Optimization in Multi-turn Language Models

URL: View paper

Brief Assessment

ERGO[74] adopts the aptitude and unreliability metrics from the original paper but does not claim to have introduced them. The candidate explicitly cites and builds upon these metrics rather than challenging their novelty.

---

7. Variable rules: Performance as a statistical reflection of competence

URL: View paper

Brief Assessment

Variable Rules[75] appears to be a linguistics paper about statistical reflection of competence in language performance. The provided context contains only JSTOR terms of use text and does not contain technical content about LLM performance metrics, aptitude, or unreliability decomposition.

---

8. Artificial Intelligence Is Stereotypically Linked More with Socially Dominant Groups in Natural Language

URL: View paper

Brief Assessment

AI Stereotypes[73] examines social biases in AI representations using stereotype content models and demographic associations. It does not address LLM performance metrics, aptitude measurement, or reliability decomposition in conversational settings.

---

9. Measuring (a Sufficient) World Model in LLMs: A Variance Decomposition Framework

URL: View paper

Brief Assessment

World Model Variance[77] focuses on decomposing response variability into purpose sensitivity, articulation sensitivity, and model uncertainty to assess semantic consistency. The original paper decomposes performance degradation into aptitude (best-case capability) and unreliability (variance across runs) in multi-turn conversations. These are distinct frameworks addressing different aspects of model behavior.

---

10. Skill-it! a data-driven skills framework for understanding and training language models

URL: View paper

Brief Assessment

Skill It[72] focuses on data-driven skill frameworks for training language models through ordered skill sets and data selection algorithms. It does not address decomposing LLM performance degradation into aptitude versus unreliability components in multi-turn conversations.

Contribution 3: Large-scale empirical study revealing multi-turn performance degradation

Description: The authors conduct over 200,000 simulated conversations across 15 LLMs and six tasks, demonstrating consistent and substantial performance drops in multi-turn settings. This empirical finding establishes the 'lost in conversation' phenomenon across state-of-the-art models.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning

URL: View paper

Brief Assessment

Ask Patients Patience[53] focuses on medical dialogue systems with Bayesian active learning for diagnostic conversations, not on general multi-turn performance degradation across diverse generation tasks as studied in the original paper.

---

2. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following

URL: View paper

Brief Assessment

Multi IF[58] focuses on multi-turn and multilingual instruction following with verifiable instructions, while the original paper examines multi-turn underspecified conversations where information is gradually revealed. These represent different experimental paradigms and research questions.

---

3. KAPA: A Deliberative Agent Framework with Tree-Structured Knowledge Base for Multi-Domain User Intent Understanding

URL: View paper

Brief Assessment

KAPA[57] focuses on building a deliberative agent framework with tree-structured knowledge for multi-domain user intent understanding, not on empirical studies of LLM performance degradation across conversation turns.

---

4. Mtalk-bench: Evaluating speech-to-speech models in multi-turn dialogues via arena-style and rubrics protocols

URL: View paper

Brief Assessment

Mtalk Bench[52] focuses on evaluating speech-to-speech models in multi-turn dialogues, not text-based LLM performance degradation in multi-turn conversations as studied in the original paper.

---

5. Reasoning-augmented conversation for multi-turn jailbreak attacks on large language models

URL: View paper

Brief Assessment

Reasoning Augmented Jailbreak[51] focuses on adversarial jailbreak attacks exploiting reasoning capabilities in multi-turn conversations, not on general performance degradation across diverse generation tasks as studied in the original paper.

---

6. ReSURE: Regularizing Supervision Unreliability for Multi-turn Dialogue Fine-tuning

URL: View paper

Brief Assessment

ReSURE[59] focuses on addressing supervision unreliability during multi-turn dialogue fine-tuning through adaptive loss reweighting, not on conducting large-scale empirical studies demonstrating performance degradation patterns across models and tasks.

---

7. ChatGPT vs. Modest Large Language Models: an extensive study on benefits and drawbacks for conversational search

URL: View paper

Brief Assessment

ChatGPT Modest Models[54] focuses on comparing ChatGPT with smaller models for conversational search tasks, not on systematic multi-turn performance degradation across diverse generation tasks with 200,000+ simulated conversations.

---

8. MultiVerse: A Multi-Turn Conversation Benchmark for Evaluating Large Vision and Language Models

URL: View paper

Brief Assessment

MultiVerse[36] focuses on evaluating vision-language models (VLMs) in multi-turn conversations with images, not general language models in text-only settings. The original paper studies LLMs across text generation tasks (code, math, database queries), while MultiVerse[36] addresses visual understanding in conversational contexts—a fundamentally different domain and modality.

---

9. From isolated conversations to hierarchical schemas: Dynamic tree memory representation for llms

URL: View paper

Brief Assessment

Hierarchical Tree Memory[56] focuses on memory representation structures for LLMs using tree-based schemas, not on empirical studies of multi-turn versus single-turn performance degradation across multiple models and tasks.

---

10. B-score: Detecting biases in large language models using response history

URL: View paper

Brief Assessment

B Score[55] focuses on bias detection through multi-turn conversations where models observe their own prior answers, not on general performance degradation across diverse generation tasks as studied in the original paper.

Appendix: Text Similarity Detection

Textual similarity detection checked 32 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. ERGO: Entropy-guided Resetting for Generation Optimization in Multi-turn Language Models

Detected in: Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] LLMs Get Lost In Multi-Turn Conversation View paper
- [1] Benchmarking Poisoning Attacks against Retrieval-Augmented Generation View paper
- [2] MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback View paper
- [3] AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents View paper
- [4] Query resolution for conversational search with limited supervision View paper

- [5] Crmarena-pro: Holistic assessment of llm agents across diverse business scenarios and interactions View paper
- [6] On overcoming miscalibrated conversational priors in llm-based chatbots View paper
- [7] Beyond Single-Sentence Prompts: Upgrading Value Alignment Benchmarks with Dialogues and Stories View paper
- [8] LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultation Conversation View paper
- [9] CPO: Addressing Reward Ambiguity in Role-playing Dialogue via Comparative Policy Optimization View paper
- [10] Improving situated conversational agents with step-by-step multi-modal logic reasoning View paper
- [11] InfoQuest: Evaluating Multi-Turn Dialogue Agents for Open-Ended Conversations with Hidden Context View paper
- [12] Enhancing Troubleshooting Task-Oriented Dialog Systems with Large Language Models View paper
- [13] nvBench 2.0: Resolving Ambiguity in Text-to-Visualization through Stepwise Reasoning View paper
- [14] LLM Defenses Are Not Robust to Multi-Turn Human Jailbreaks Yet View paper
- [15] MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models View paper
- [16] Act2P: LLM-Driven Online Dialogue Act Classification for Power Analysis View paper
- [17] User Intent and State Modeling in Conversational Systems View paper
- [18] CLAM: Selective Clarification for Ambiguous Questions with Generative Language Models View paper
- [19] MMDU: A Multi-Turn Multi-Image Dialog Understanding Benchmark and Instruction-Tuning Dataset for LVLMs View paper
- [20] Learning to Clarify: Multi-turn Conversations with Action-Based Contrastive Self-Training View paper
- [21] BIRD-INTERACT: Re-imagining Text-to-SQL Evaluation for Large Language Models via Lens of Dynamic Interactions View paper
- [22] C3: A Bilingual Benchmark for Spoken Dialogue Models Exploring Challenges in Complex Conversations View paper
- [23] Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-world Multi-turn Dialogue View paper
- [24] CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling View paper
- [25] Data-Centric Improvements for Enhancing Multi-Modal Understanding in Spoken Conversation Modeling View paper
- [26] ConvBench: A Multi-Turn Conversation Evaluation Benchmark with Hierarchical Capability for Large Vision-Language Models View paper
- [27] An Automatic Evaluation Framework for Multi-turn Medical Consultations Capabilities of Large Language Models View paper
- [28] Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning View paper
- [29] Repairs in a Block World: A New Benchmark for Handling User Corrections with Multi-Modal Language Models View paper
- [30] ACEBench: Who Wins the Match Point in Tool Usage? View paper
- [31] Verifiable Accuracy and Abstention Rewards in Curriculum RL to Alleviate Lost-in-Conversation View paper
- [32] A Concurrent Intelligent Natural Language Understanding Model for an Automated Inquiry System View paper
- [33] Talk-to-Resolve: Combining scene understanding and spatial dialogue to resolve granular task ambiguity for a collocated robot View paper
- [34] Collaborative Multi-Agent Dialogue Model Training Via Reinforcement Learning View paper
- [35] Better Semantic Understanding in LLM-Based Multi-Turn Dialogues: A Survey View paper
- [36] MultiVerse: A Multi-Turn Conversation Benchmark for Evaluating Large Vision and Language Models View paper
- [37] Learning Contextual Retrieval for Robust Conversational Search View paper
- [38] Multi-Turn Multi-Modal Question Clarification for Enhanced Conversational Understanding View paper
- [39] Ask-to-Clarify: Resolving Instruction Ambiguity through Multi-turn Dialogue View paper
- [40] MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems View paper
- [41] UserBench: An Interactive Gym Environment for User-Centric Agents View paper
- [42] Dynamic benchmarking framework for LLM-based conversational data capture View paper
- [43] Identifying & Interactively Refining Ambiguous User Goals for Data Visualization Code Generation View paper
- [44] C-MTCSD: A Chinese Multi-Turn Conversational Stance Detection Dataset View paper
- [45] Pluralistic Behavior Suite: Stress-Testing Multi-Turn Adherence to Custom Behavioral Policies View paper
- [46] ContextQFormer: A New Context Modeling Method for Multi-Turn Multi-Modal Conversations View paper
- [47] Emotionally Intelligent Task-oriented Dialogue Systems: Architecture, Representation, and Optimisation View paper
- [48] Beyond Task-Oriented and Chitchat Dialogues: Proactive and Transition-Aware Conversational Agents View paper
- [49] MoPHES:Leveraging on-device LLMs as Agent for Mobile Psychological Health Evaluation and Support View paper
- [50] Reasoning Is Not All You Need: Examining LLMs for Multi-Turn Mental Health Conversations View paper
- [51] Reasoning-augmented conversation for multi-turn jailbreak attacks on large language models View paper
- [52] Mtalk-bench: Evaluating speech-to-speech models in multi-turn dialogues via arena-style and rubrics protocols View paper
- [53] Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning View paper
- [54] ChatGPT vs. Modest Large Language Models: an extensive study on benefits and drawbacks for conversational search View paper
- [55] B-score: Detecting biases in large language models using response history View paper
- [56] From isolated conversations to hierarchical schemas: Dynamic tree memory representation for llms View paper
- [57] KAPA: A Deliberative Agent Framework with Tree-Structured Knowledge Base for Multi-Domain User Intent Understanding View paper
- [58] Multi-if: Benchmarking llms on multi-turn and multilingual instructions following View paper
- [59] ReSURE: Regularizing Supervision Unreliability for Multi-turn Dialogue Fine-tuning View paper
- [60] User simulation in task-oriented dialog systems based on large language models via in-context learning View paper
- [61] Math-llava: Bootstrapping mathematical reasoning for multimodal large language models View paper
- [62] Automated Safety Evaluations Across 20 Large Language Models: The Aymara LLM Risk and Responsibility Matrix View paper
- [63] Dynamic evaluation with cognitive reasoning for multi-turn safety of large language models View paper
- [64] Flipping the dialogue: Training and evaluating user language models View paper
- [65] OpenDeception: Benchmarking and Investigating AI Deceptive Behaviors via Open-ended Interaction Simulation View paper
- [66] DialSim: A Dialogue Simulator for Evaluating Long-Term Multi-Party Dialogue Understanding of Conversational Agents View paper
- [67] Multi-turn Evaluation of Anthropomorphic Behaviours in Large Language Models View paper
- [68] Contextualized Evaluations: Judging Language Model Responses to Underspecified Queries View paper
- [69] MATRIX: Multi-Agent simulaTion fRamework for safe Interactions and conteXtual clinical conversational evaluation View paper
- [70] Psychometric Personality Shaping Modulates Capabilities and Safety in Language Models View paper

- [71] Do large language models show human-like biases? exploring confidenceâcompetence gap in ai View paper
- [72] Skill-it! a data-driven skills framework for understanding and training language models View paper
- [73] Artificial Intelligence Is Stereotypically Linked More with Socially Dominant Groups in Natural Language View paper
- [74] ERGO: Entropy-guided Resetting for Generation Optimization in Multi-turn Language Models View paper
- [75] Variable rules: Performance as a statistical reflection of competence View paper
- [76] Variability, Its Limits, and the Performanceâ□Competence Debate: Implications of Linguistic Variability for a Theory of Grammar View paper
- [77] Measuring (a Sufficient) World Model in LLMs: A Variance Decomposition Framework View paper
- [78] Incoherent Beliefs & Inconsistent Actions in Large Language Models View paper
- [79] ChatGPT on the Road: Leveraging Large Language Model-Powered In-vehicle Conversational Agents for Safer and More Enjoyable Driving Experience View paper