Novelty Assessment Report

Paper: τ^2-bench: : Evaluating Conversational Agents in a Dual-Control Environment

PDF URL: https://openreview.net/pdf?id=LGmO9VvuP5

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

Existing benchmarks for conversational AI agents simulate **single-control environments**, where only the AI agent can use tools to interact with the world, while the user remains a passive information provider. This differs from real-world scenarios like technical support, where users need to actively participate in modifying the state of the (shared) world. In order to address this gap, we introduce τ^2-bench, with four key contributions:

1. A novel **Telecom dual-control domain** modeled as a Dec-POMDP, where both agent and user make use of tools to act in a shared, dynamic environment that tests both agent coordination and communication,
2. A **compositional task generator** that programmatically creates diverse, verifiable tasks from atomic components, ensuring domain coverage and controlled complexity,
3. A **reliable user simulator** tightly coupled with the environment, whose behavior is constrained by tools and observable states, improving simulation fidelity,
4. **Fine-grained analysis of agent performance** through multiple ablations including separating errors arising from reasoning vs communication/coordination.

In particular, our experiments show significant performance drops when agents shift from no-user to dual-control, highlighting the challenges of guiding users. Overall, τ^2-bench provides a controlled testbed for agents that must both reason effectively and guide user actions.

Core Task Landscape

This paper addresses: **Evaluating Conversational Agents in Dual-Control Environments with Shared Tool Use**

A total of **8 papers** were analyzed and organized into a taxonomy with **7 categories**.

Taxonomy Overview

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

- **Benchmark Design and Evaluation Frameworks**
- **Multi-Agent Coordination and Control Architectures**

Complete Taxonomy Tree

- Evaluating Conversational Agents in Dual-Control Environments with Shared Tool Use Survey Taxonomy
- Benchmark Design and Evaluation Frameworks
  - Dual-Control Environment Benchmarks ★ (2 papers)
  - [0] τ^2-bench: : Evaluating Conversational Agents in a Dual-Control Environment (Anon et al., 2026) View paper
  - [1] Ï2-Bench: Evaluating Conversational Agents in a Dual-Control Environment (Barres Victor, 2025) View paper
  - Task-Specific Workflow Automation Evaluation (1 papers)
  - [3] AssistEditor: Multi-Agent Collaboration for GUI Workflow Automation in Video Creation (Difei Gao, 2024) View paper
- Multi-Agent Coordination and Control Architectures
  - Shared Control with Oracle-Based Coordination (1 papers)
  - [4] Shared Control with Black Box Agents Using Oracle Queries (Inbal Avraham, 2025) View paper
  - Decentralized Multi-Agent Synthesis Frameworks (1 papers)
  - [5] Matrix: Peer-to-Peer Multi-Agent Synthetic Data Generation Framework (Dong Wang, 2025) View paper
  - Dialogue-Based Shared Control Systems
  - Spatial Navigation and Robotics Applications (1 papers)
    - [6] Towards dialogue based shared control of navigating robots (R. Ross, 2004) View paper
  - Safety-Critical Dialogue Systems with Formal Methods (2 papers)
    - [7] Formalising control in robust spoken dialogue systems (Hui Shi, 2005) View paper
    - [8] A safe and robust approach to shared-control via dialogue (Bernd Krieg-BrÃ¼ckner, 2004) View paper
  - Human-Robot Interaction Data Collection (1 papers)
    - [2] A Dual-Control Dialogue Framework for Human-Robot Interaction Data Collection: Integrating Human Emotional and Contextual Awareness with Conversational AI (Jonas Beskow, 2024) View paper

Narrative

Core task: Evaluating conversational agents in dual-control environments with shared tool use. This emerging field addresses scenarios where multiple agents or an agent and a human must coordinate control over shared resources or tools through dialogue. The taxonomy organizes work into two main branches: Benchmark Design and Evaluation Frameworks, which focuses on creating testbeds and metrics for assessing agent performance in these complex settings, and Multi-Agent Coordination and Control Architectures, which explores the underlying mechanisms and algorithms that enable effective collaboration. Within the first branch, researchers have developed specialized dual-control environment benchmarks that simulate realistic scenarios of shared tool manipulation and conversational

grounding, while the second branch examines how agents can negotiate control, resolve conflicts, and maintain coherent dialogue during joint task execution. Early foundational work such as Dialogue Shared Control[6] and Safe Shared Control[8] established basic principles, while more recent efforts like Dual Control Dialogue[2] and AssistEditor[3] have introduced richer interactive settings.

Recent developments reveal contrasting emphases between synthetic benchmark construction and real-world applicability. Some studies prioritize controlled experimental conditions using synthetic data generation approaches like Matrix Synthetic Data[5] or oracle-based evaluation methods such as Black Box Oracle[4], enabling systematic measurement of agent capabilities under varied conditions. Others focus on naturalistic human-agent collaboration scenarios where dialogue must adapt to unpredictable user intentions and tool states. Tau Squared Bench[0] situates itself within the Benchmark Design branch, specifically targeting dual-control environment evaluation. Compared to closely related work like AssistEditor[3], which emphasizes collaborative editing tasks, Tau Squared Bench[0] appears to offer a broader framework for assessing conversational coordination across diverse shared-tool scenarios, providing structured metrics for both dialogue quality and control effectiveness in settings where agents must dynamically negotiate resource access.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Ï□2-Bench: Evaluating Conversational Agents in a Dual-Control Environment

Authors: Barres Victor, Dong Honghua, Victor Barres, Honghua Dong, Si, et al. (11 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Existing benchmarks for conversational AI agents simulate single-control environments, where only the AI agent can use tools to interact with the world, while the user remains a passive information provider. This differs from real-world scenarios like technical support, where users need to actively participate in modifying the state of the (shared) world. In order to address this gap, we introduce τ^2-bench, with four key contributions: 1) A novel Telecom dual-control domain modeled as a De...

⚠ Similarity Notice

This paper is highly similar to the original paper; it may be a variant or near-duplicate. Please manually verify.

Contributions Analysis

Overall novelty summary. The paper introduces τ²-bench, a benchmark for evaluating conversational agents in dual-control environments where both agent and user actively use tools to modify shared state. It sits within the 'Dual-Control Environment Benchmarks' leaf of the taxonomy, which contains only one sibling paper (AssistEditor). This indicates a relatively sparse research direction within the broader field of conversational agent evaluation. The paper's core contributions—Dec-POMDP formalization, compositional task generation, and tightly coupled user simulation—target systematic evaluation of coordination and communication in shared-control scenarios.

The taxonomy reveals that dual-control benchmarks form a small subset of the broader 'Benchmark Design and Evaluation Frameworks' branch, which also includes task-specific workflow automation evaluation. Neighboring branches address multi-agent coordination architectures, including oracle-based coordination, decentralized synthesis, and dialogue-based shared control systems spanning robotics, safety-critical applications, and data collection. The paper's telecom domain and Dec-POMDP formalization connect it to formal multi-agent coordination work, while its emphasis on conversational grounding aligns with dialogue-based shared control research. The taxonomy's scope notes clarify that this work differs from single-control benchmarks and pure architectural studies.

Among 25 candidates examined across three contributions, none were found to clearly refute the paper's claims. The Dec-POMDP formalization examined 5 candidates with no refutations, suggesting this formal modeling approach may be relatively novel for dual-control conversational benchmarks. The compositional task generator and user simulator contributions each examined 10 candidates, also with no refutations. This limited search scope indicates that within the top-25 semantically similar papers, no substantial prior work directly overlaps with these specific technical contributions. However, the small candidate pool means the analysis cannot rule out relevant work outside this search window.

Based on the limited literature search, the paper appears to occupy a relatively unexplored niche within conversational agent evaluation. The sparse taxonomy leaf (only one sibling) and absence of refuting candidates among 25 examined papers suggest the dual-control benchmark approach with compositional generation and coupled simulation may be distinctive. However, the analysis is constrained by the top-K semantic search methodology and does not constitute an exhaustive survey of all potentially relevant prior work in multi-agent systems, dialogue evaluation, or simulation-based benchmarking.

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

Contribution 1: Dual-control environment formalized as Dec-POMDP

Description: The authors introduce a dual-control setup where both the AI agent and the simulated user possess distinct tools to observe, act upon, and verify the state of a shared environment. This is formalized using a Decentralized Partially Observable Markov Decision Process (Dec-POMDP), enabling realistic simulations of collaborative scenarios like technical support.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Solving efficiently decentralized MDPs with temporal and resource constraints

URL: View paper

Brief Assessment

Decentralized Temporal MDPs[29] focuses on temporal and resource constraints in decentralized MDPs but does not address dual-control scenarios where both agent and user possess distinct tools for shared environment interaction. The candidate paper's Dec-POMDP formalization serves a different purpose than the original's dual-control setup for agent-user coordination.

---

2. Decentralized communication strategies for coordinated multi-agent policies

URL: View paper

Brief Assessment

Decentralized Communication Strategies[28] uses Dec-POMDP for multi-agent coordination in general domains, not for dual-control agent-user scenarios with shared tools as in the original paper's telecom support setting.

---

3. Probabilistic Decision-Making Models for Multi-Agent Systems and Human-Robot Collaboration

URL: View paper

Brief Assessment

Probabilistic Multi Agent[27] focuses on robot-robot and human-robot collaboration using Dec-POMDP for multi-robot coordination, not on dual-control agent-user environments with shared tools for technical support scenarios.

---

4. -Bench: Evaluating Conversational Agents in a Dual-Control Environment

URL: View paper

Brief Assessment

Dual Control Bench[10] presents the same Dec-POMDP formalization for dual-control environments where both agent and user have tools. The papers appear to be identical or near-identical versions of the same work, not independent prior work.

---

5. Ï□2-Bench: Evaluating Conversational Agents in a Dual-Control Environment

URL: View paper

Brief Assessment

Tau Squared Bench[1] introduces the same dual-control Dec-POMDP formalization for agent-user coordination with shared tools. This is the same paper being evaluated, not prior work.

Contribution 2: Compositional task generator

Description: The authors develop a programmatic task generator that automatically composes a vast and diverse set of verifiable tasks from atomic base scenarios defined by initialization, solution, and assertion functions. This method ensures provable correctness, provides complete domain coverage, and allows explicit control over task complexity.

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. Large language models are innate crystal structure generators

URL: View paper

Brief Assessment

LLM Crystal Structures[13] focuses on crystal structure generation in materials science using evolutionary algorithms with LLMs, not on compositional task generation for conversational AI agents or verification systems.

---

2. Compositional programming and testing of dynamic distributed systems

URL: View paper

Brief Assessment

Compositional Distributed Programming[16] focuses on compositional programming and testing of distributed systems using state machines and trace refinement, not on task generation for conversational AI benchmarks. The compositional approach in the candidate refers to system decomposition for testing distributed services, which is fundamentally different from programmatically generating diverse verifiable tasks from atomic scenarios for agent evaluation.

---

3. Trustworthy genetic programming-based synthesis of analog circuit topologies using hierarchical domain-specific building blocks

URL: View paper

Brief Assessment

Genetic Programming Circuits[17] focuses on analog circuit topology synthesis using hierarchical building blocks and grammar-based representations, not on compositional task generation for agent evaluation benchmarks.

---

4. Compositional verification of composite byzantine protocols

URL: View paper

Brief Assessment

Byzantine Compositional Verification[9] focuses on compositional verification of Byzantine fault-tolerant protocols using formal methods in Coq, not on task generation for conversational AI agents or reinforcement learning environments.

---

5. Compositional verification using a formal component and interface specification

URL: View paper

Brief Assessment

Compositional Interface Verification[14] focuses on compositional verification of hardware components using formal interface specifications and refinement checking. The candidate addresses hardware RTL design verification, not task generation for conversational AI agents or reinforcement learning environments.

---

6. A theory of composition for proofs of knowledge

URL: View paper

Brief Assessment

Composition Proofs Knowledge[11] focuses on mathematical proofs of knowledge and cryptographic protocols, not task generation for conversational AI agents or benchmarking systems.

---

7. -Bench: Evaluating Conversational Agents in a Dual-Control Environment

URL: View paper

Brief Assessment

Dual Control Bench[10] describes the identical compositional task generation approach with atomic subtasks defined by initialization, solution, and assertion functions. This appears to be the same work rather than independent prior art.

---

8. Compounding metaâ□□atoms into metamolecules with hybrid artificial intelligence techniques

URL: View paper

Brief Assessment

Metamolecules Hybrid AI[15] focuses on designing meta-molecules for optical metasurfaces using compositional pattern-producing networks (CPPNs) to generate nanostructure patterns, not on creating verifiable tasks from atomic components for conversational AI benchmarking.

---

9. Enumerate-Conjecture-Prove: Formally Solving Answer-Construction Problems in Math Competitions

URL: View paper

Brief Assessment

Enumerate Conjecture Prove[12] focuses on formal mathematical problem solving through enumeration and theorem proving in Lean, not on compositional task generation from atomic components for conversational AI benchmarks.

10. Ï□2-Bench: Evaluating Conversational Agents in a Dual-Control Environment

URL: View paper

Brief Assessment

Tau Squared Bench[1] presents the same compositional task generator creating verifiable tasks from atomic components. This is the same paper, not prior work that could refute novelty.

Contribution 3: Reliable user simulator tightly coupled with environment

Description: The authors enhance user simulation reliability by tightly coupling the user simulator to the environment. User behavior is constrained by available tools and observable state, leading to more predictable and consistent interactions with substantially lower error rates compared to existing domains.

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. Design of a physical and interactive real-time simulator based on a dynamic vpp as a support tool for sailing yacht design and operation

URL: View paper

Brief Assessment

Sailing Simulator[22] focuses on physical sailing yacht simulation with dynamic VPP (Velocity Prediction Program) for yacht design and operation, not on conversational AI user simulation or tool-constrained user behavior in dual-control environments.

2. Developing a VR Socially Assistive Robot Simulator Employing Game Development Tools

URL: View paper

Brief Assessment

VR Robot Simulator[20] focuses on developing a virtual reality digital twin for robot testing in long-term care facilities, not on user simulation for conversational AI agents or dual-control environments with tool-constrained behavior.

3. User Mobility Simulator for Full-Immersive Multiuser Virtual Reality with Redirected Walking

URL: View paper

Brief Assessment

VR Mobility Simulator[23] focuses on mapping virtual user movements to physical trajectories in VR environments with redirected walking, not on conversational agent simulation or tool-constrained user behavior in dialogue systems.

4. Productagent: Benchmarking conversational product search agent with asking clarification questions

URL: View paper

Brief Assessment

ProductAgent[19] focuses on product search clarification with an LLM-based user simulator conditioned on a target item, not on tightly coupling user simulators to environment tools and observable states for general conversational AI benchmarking.

5. Self-adaptation with end-user preferences: Using run-time models and constraint solving

URL: View paper

Brief Assessment

Runtime Constraint Adaptation[24] focuses on self-adaptation systems using constraint solving for runtime configuration, not on user simulation for conversational AI benchmarks. The candidate does not address user simulator reliability or environment coupling in the context of agent evaluation.

6. Views for tools in integrated environments

URL: View paper

Brief Assessment

Views for Tools[21] discusses tool integration in development environments but does not address user simulation or environment coupling for conversational AI evaluation.

7. RSS Demonstrator: a Tool for User Experience Interactions with Automated Driving Safety Models

URL: View paper

Brief Assessment

RSS Demonstrator[26] focuses on a simulation tool for experiencing automated driving safety models (RSS) with user interactions, not on developing user simulators for conversational AI benchmarks or coupling simulators to environments through tool constraints.

8. ToolSandbox: A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities

URL: View paper

Brief Assessment

ToolSandbox[18] includes a user simulator for conversational evaluation, but focuses on stateful tool execution and API testing rather than the specific approach of tightly coupling user behavior to environment constraints through tool-based state observation that τ2-bench emphasizes for dual-control scenarios.

9. Ï□2-Bench: Evaluating Conversational Agents in a Dual-Control Environment

URL: View paper

Brief Assessment

Tau Squared Bench[1] describes the same user simulator approach tightly coupled with environment using tool constraints. This is the same paper under review, not prior work.

10. Eye-Si Simulator: A user experience

URL: View paper

Brief Assessment

Eye-Si Simulator[25] is a surgical training simulator for ophthalmology, not a conversational AI user simulator. It does not address user simulation in dual-control environments or tool-constrained behavior modeling.

Appendix: Text Similarity Detection

Textual similarity detection checked 22 papers and found 4 similarity segment(s) across 2 paper(s).

The following **2 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. -Bench: Evaluating Conversational Agents in a Dual-Control Environment

Detected in: Contribution: contribution_1, 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.

---

2. Ï□2-Bench: Evaluating Conversational Agents in a Dual-Control Environment

Detected in: Core Task (sibling), Contribution: contribution_1, Contribution: contribution_2, Contribution: contribution_3

⚠ **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] τ^2-bench: : Evaluating Conversational Agents in a Dual-Control Environment View paper
- [1] Ï□2-Bench: Evaluating Conversational Agents in a Dual-Control Environment View paper
- [2] A Dual-Control Dialogue Framework for Human-Robot Interaction Data Collection: Integrating Human Emotional and Contextual Awareness with Conversational AI View paper
- [3] AssistEditor: Multi-Agent Collaboration for GUI Workflow Automation in Video Creation View paper
- [4] Shared Control with Black Box Agents Using Oracle Queries View paper
- [5] Matrix: Peer-to-Peer Multi-Agent Synthetic Data Generation Framework View paper
- [6] Towards dialogue based shared control of navigating robots View paper
- [7] Formalising control in robust spoken dialogue systems View paper
- [8] A safe and robust approach to shared-control via dialogue View paper
- [9] Compositional verification of composite byzantine protocols View paper
- [10] -Bench: Evaluating Conversational Agents in a Dual-Control Environment View paper
- [11] A theory of composition for proofs of knowledge View paper
- [12] Enumerate-Conjecture-Prove: Formally Solving Answer-Construction Problems in Math Competitions View paper
- [13] Large language models are innate crystal structure generators View paper
- [14] Compositional verification using a formal component and interface specification View paper
- [15] Compounding metaâ□□atoms into metamolecules with hybrid artificial intelligence techniques View paper
- [16] Compositional programming and testing of dynamic distributed systems View paper
- [17] Trustworthy genetic programming-based synthesis of analog circuit topologies using hierarchical domain-specific building blocks View paper
- [18] ToolSandbox: A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities View paper
- [19] Productagent: Benchmarking conversational product search agent with asking clarification questions View paper
- [20] Developing a VR Socially Assistive Robot Simulator Employing Game Development Tools View paper
- [21] Views for tools in integrated environments View paper
- [22] Design of a physical and interactive real-time simulator based on a dynamic vpp as a support tool for sailing yacht design and operation View paper
- [23] User Mobility Simulator for Full-Immersive Multiuser Virtual Reality with Redirected Walking View paper
- [24] Self-adaptation with end-user preferences: Using run-time models and constraint solving View paper
- [25] Eye-Si Simulator: A user experience View paper
- [26] RSS Demonstrator: a Tool for User Experience Interactions with Automated Driving Safety Models View paper
- [27] Probabilistic Decision-Making Models for Multi-Agent Systems and Human-Robot Collaboration View paper
- [28] Decentralized communication strategies for coordinated multi-agent policies View paper
- [29] Solving efficiently decentralized MDPs with temporal and resource constraints View paper