Novelty Assessment Report

Paper: Computer Agent Arena: Toward Human-Centric Evaluation and Analysis of Computer-Use Agents
PDF URL: https://openreview.net/pdf?id=3x4SDbXbgl
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

As Computer-Use Agents (CUAs) proliferate and grow increasingly capable, evaluation has become more challenging: static, manually curated benchmarks are narrow in domain, contamination-prone, and environment-heavy, and they diverge substantially from user-driven, real-world evaluation. We present Computer Agent Arena, an open-source platform for head-to-head CUA evaluation and a dynamic methodology that converts human preferences into structured feedback in realistic environments. The system (i) simulates real-world computer use via cloud-hosted, diverse, and dynamic environment initializations and customizations; (ii) ensures authentic, fair comparison by faithfully reproducing open-source CUAs and executing anonymously in matched, controlled environments; and (iii) extends evaluation beyond pairwise preference and correctness to capability- and behavior-oriented signals. Across 2,201 high-quality votes over 12 agents—spanning multi-app interactions, ambiguous instructions, and open-ended queries—we observe striking ranking reversals relative to static benchmarks. Further analysis shows that overall correctness mainly drives human preference; beyond that, agent-human interaction and self-correction boost user preference, even when overall task completion is comparable. Our error analysis reveals agent behavior errors, such as long-horizon memory and fine-grained action failures that static benchmarks fail to evaluate. We also contrast pure GUI agents with universal digital agents capable of tool use and coding, and discuss the trade-offs of these different design philosophies. We open source the full platform, collected dataset, and code of Computer Agent Arena to support future research on the evaluation and development of CUA.

Core Task Landscape

This paper addresses: **human-centric evaluation of computer-use agents**

A total of **50 papers** were analyzed and organized into a taxonomy with **18 categories**.

Taxonomy Overview

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

- **Benchmark Design and Task Environments**
- **Human-Centered Evaluation Methodologies**
- **Human-Agent Collaboration and Interaction Design**
- **User-Centered Design and Personalization**
- **Theoretical Foundations and Research Agendas**
- **Agent Architectures and Capabilities**

Complete Taxonomy Tree

- human-centric evaluation of computer-use agents Survey Taxonomy
- Benchmark Design and Task Environments
  - Web and GUI Interaction Benchmarks (3 papers)
  - [2] Bearcubs: A benchmark for computer-using web agents (Song Yixiao, 2025) View paper
  - [15] Webgames: Challenging general-purpose web-browsing ai agents (Thomas George, 2025) View paper
  - [18] Videowebarena: Evaluating long context multimodal agents with video understanding web tasks (Lawrence Jang, 2024) View paper
  - Open-World and Multi-Step Task Benchmarks (2 papers)
  - [16] MCU: An Evaluation Framework for Open-Ended Game Agents (Zheng Xinyue, 2025) View paper
  - [28] DRBench: A Realistic Benchmark for Enterprise Deep Research (Abaskohi, 2025) View paper
  - Safety and Adversarial Robustness Benchmarks (3 papers)
  - [3] OS-Harm: A Benchmark for Measuring Safety of Computer Use Agents (Thomas Kuntz, 2025) View paper
  - [4] SusBench: An Online Benchmark for Evaluating Dark Pattern Susceptibility of Computer-Use Agents (Yuan Chenjie, 2025) View paper
  - [35] Just Do It!? Computer-Use Agents Exhibit Blind Goal-Directedness (Hines Keegan, 2025) View paper
- Human-Centered Evaluation Methodologies
  - Human Preference and Comparative Evaluation ★ (2 papers)
  - [0] Computer Agent Arena: Toward Human-Centric Evaluation and Analysis of Computer-Use Agents (Anon et al., 2026) View paper
  - [19] Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents (Eric Smith, 2022) View paper
  - Agent-Based and Automated Evaluation (3 papers)
  - [10] Evaluation agent: Efficient and promptable evaluation framework for visual generative models (Zhang, 2025) View paper
  - [11] Agent-as-a-judge: Evaluate agents with agents (Zhuge, 2024) View paper
  - [25] Ali-agent: Assessing llms' alignment with human values via agent-based evaluation (Zheng, 2024) View paper
  - Usability and Interaction Quality Assessment (4 papers)
  - [1] Heuristic evaluation of conversational agents (Raina Langevin, 2021) View paper

- [5] PIPA: A Unified Evaluation Protocol for Diagnosing Interactive Planning Agents (Kim, 2025) View paper
- [49] CareAssist GPT improves patient user experience with a patient centered approach to computer aided diagnosis. (Ali Algarni, 2025) View paper
- [50] Performance Evaluation using Human Computer Interaction Models (Apicha Deearom, 2024) View paper

- Human-Agent Collaboration and Interaction Design
  - Collaborative Task Execution Frameworks (3 papers)
  - [23] CowPilot: A Framework for Autonomous and Human-Agent Collaborative Web Navigation (Huq, 2025) View paper
  - [24] Towards Effective Human-in-the-Loop Assistive AI Agents (Li Yayuan, 2025) View paper
  - [33] Decision-oriented dialogue for human-AI collaboration (Jessy Lin, 2024) View paper
  - Conversational and Dialogue-Based Interaction (3 papers)
  - [6] A survey on near-human conversational agents (Satwinder Singh, 2022) View paper
  - [43] Conversational agents for information retrieval in the education domain: A user-centered design investigation (Wambsganss, 2022) View paper
  - [47] Conversational agents for automated group meeting facilitation a computational framework for facilitating small group decision-making meetings (Shamekhi, 2020) View paper
  - Transparency and Trust in Agent Systems (3 papers)
  - [7] Verios: Query-driven proactive human-agent-gui interaction for trustworthy os agents (Wu Zheng, 2025) View paper
  - [9] Smart Transparency: A User-Centered Approach to Improving Humanâ□□Machine Interaction in High-Risk Supervisory Control Tasks (Keran Wang, 2025) View paper
  - [26] Magentic-ui: Towards human-in-the-loop agentic systems (Mozannar, 2025) View paper
- User-Centered Design and Personalization
  - Persona and Personality Customization (4 papers)
  - [17] Personagym: Evaluating persona agents and llms (Samuel Vinay, 2024) View paper
  - [21] CloChat: Understanding how people customize, interact, and experience personas in large language models (Jeon, 2024) View paper
  - [31] Hiring an AI: Incorporating Personnel Selection Methods in User-Centered Design to Design AI Agents for Safety-Critical Domains (Stephan Huber, 2024) View paper
  - [42] Exploring a Gamified Personality Assessment Method through Interaction with LLM Agents Embodying Different Personalities (Zhang Bai-qiao, 2025) View paper
  - User Modeling and Adaptive Systems (3 papers)
  - [13] User modeling in humanâ□□computer interaction (Fischer, 2001) View paper
  - [30] Evolving agents: Interactive simulation of dynamic and diverse human personalities (Li Jiale, 2024) View paper
  - [36] A user-centered approach to user-building interactions (Kelly Kalvelage, 2014) View paper
  - Co-Creativity and Agency Modulation (2 papers)
  - [8] A user-centered framework for human-ai co-creativity (Caterina Moruzzi, 2024) View paper
  - [41] Creative agents: Empowering agents with imagination for creative tasks (Zhang Chi, 2023) View paper
- Theoretical Foundations and Research Agendas
  - Human-AI Interaction Research Frameworks (2 papers)
  - [12] Human-AI interaction research agenda: A user-centered perspective (Tingting Jiang, 2024) View paper
  - [27] Rethinking theory of mind benchmarks for llms: Towards a user-centered perspective (Wang, 2025) View paper
  - User Experience and Human Factors (3 papers)
  - [29] Sentiment-aware design of humanâ□□computer interactions: How research in humanâ□□computer interaction and sentiment analysis can lead to more user-centered â□¦ (Ghosh, 2023) View paper
  - [40] A Conceptual Model of User Experience in Humanâ□□Computer Interaction Using Thematic Analysis (Choopankareh, 2025) View paper
  - [46] The First Impression: Understanding the Impact of Multimodal System Responses on User Behavior in Task-oriented Agents (Diogo Silva, 2022) View paper
  - Domain-Specific Human Factors (3 papers)
  - [34] Human-centred test and evaluation of military AI (Helmer, 2024) View paper
  - [44] Humanâ□□agent teaming for multirobot control: A review of human factors issues (Jessie Y.C. Chen, 2014) View paper
  - [45] FURIOUS: Fully unified risk-assessment with interactive operational user system for vessels. (JeeHong Kim, 2025) View paper
- Agent Architectures and Capabilities
  - Open-Source Agent Frameworks and Platforms (3 papers)
  - [20] OpenCUA: Open Foundations for Computer-Use Agents (Wang Xin-yuan, 2025) View paper
  - [39] Evalai: Towards better evaluation systems for ai agents (Yadav, 2019) View paper
  - [48] Optimizing SIA Development: A Case Study in User-Centered Design for Estuary, a Multimodal Socially Interactive Agent Framework (Spencer Lin, 2025) View paper
  - Specialized Agent Architectures (2 papers)
  - [14] Transagents: Build your translation company with language agents (Minghao Wu, 2024) View paper
  - [32] Emergence of social norms in generative agent societies: principles and architecture (Bo Xu, 2024) View paper
  - Agent Behavior and Error Analysis (3 papers)
  - [22] Characterizing unintended consequences in human-gui agent collaboration for web browsing (Zhang Shuning, 2025) View paper
  - [37] Evaluation of human-ai teams for learned and rule-based agents in hanabi (Siu, 2021) View paper
  - [38] Evaluating the LLM agents for simulating humanoid behavior (C Chen, 2024) View paper

Narrative

Core task: human-centric evaluation of computer-use agents. The field has organized itself around six major branches that together address how to design, evaluate, and deploy agents that interact with computers on behalf of users. Benchmark Design and Task Environments focuses on creating realistic test scenarios and datasets, often drawing from web navigation, GUI manipulation, and multi-step workflows. Human-Centered Evaluation Methodologies emphasizes methods that capture user preferences, comparative judgments, and qualitative feedback rather than purely automated metrics. Human-Agent Collaboration and Interaction Design explores how agents and humans can work together effectively, including transparency mechanisms and co-creative workflows. User-Centered Design and Personalization investigates tailoring agent behavior to individual needs and contexts, while Theoretical Foundations and Research

Agendas lays out conceptual frameworks and long-term research questions. Agent Architectures and Capabilities examines the technical underpinnings that enable robust, safe, and capable computer-use agents.

A particularly active tension runs between automated evaluation approaches—such as Agent-as-Judge[11] and Evaluation Agent[10]—and methods that directly involve human judgment, as seen in Human Evaluation Conversations[19] and Heuristic Evaluation Conversational[1]. Safety and alignment concerns also cut across branches, with works like OS-Harm[3] and SusBench[4] highlighting risks in real-world deployment. Computer Agent Arena[0] sits squarely within the Human Preference and Comparative Evaluation cluster, emphasizing direct human feedback to rank agent behaviors rather than relying solely on task-success metrics. This positions it close to Human Evaluation Conversations[19], which similarly advocates for nuanced human input, but Computer Agent Arena[0] extends the approach to interactive computer-use scenarios where user satisfaction and perceived helpfulness become central. The broader landscape reveals an ongoing shift from benchmark-driven evaluation toward richer, more ecologically valid assessments that account for diverse user needs and real-world variability.

Related Works in Same Category

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

1. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents

Authors: Eric Smith, Orion Hsu, Eric Michael Smith, Rebecca Qian, Stephen Roller, et al. (9 authors total) | **Year/Venue**: 2022 | **URL**: View paper

Abstract

At the heart of improving conversational AI is the open problem of how to evaluate conversations. Issues with automatic metrics are well known (Liu et al., 2016), with human evaluations still considered the gold standard. Unfortunately, how to perform human evaluations is also an open problem: differing data collection methods have varying levels of human agreement and statistical sensitivity, resulting in differing amounts of human annotation hours and labor costs. In this work we compare five ...

Relationship Analysis

Both papers belong to the Human Preference and Comparative Evaluation category, employing crowdworker-based pairwise comparisons to assess agent performance through human judgments. While the original paper (Computer Agent Arena) focuses on evaluating computer-use agents through head-to-head comparisons in realistic desktop environments with diverse task initializations, the candidate paper examines various human evaluation methodologies (per-turn vs. per-dialogue, pairwise vs. single-model) specifically for conversational dialogue agents. The key difference lies in the application domain: the original targets GUI-based computer automation agents, whereas the candidate addresses open-domain conversational AI systems.

Contributions Analysis

Overall novelty summary. The paper introduces Computer Agent Arena, a platform for head-to-head evaluation of computer-use agents through human preference judgments, and a dataset of 2,201 votes across 12 agents. It resides in the Human Preference and Comparative Evaluation leaf, which contains only two papers total: the original work and one sibling (Human Evaluation Conversations). This leaf is notably sparse, suggesting that direct human preference collection for computer-use agents remains an underexplored direction within the broader field of 50 papers. The work sits within the Human-Centered Evaluation Methodologies branch, which itself comprises three leaves addressing preference-based, automated, and usability-focused evaluation approaches.

The taxonomy reveals that most evaluation activity clusters around Benchmark Design and Task Environments, particularly Web and GUI Interaction Benchmarks (three papers) and Safety and Adversarial Robustness Benchmarks (three papers). The sibling leaf Agent-Based and Automated Evaluation contains three papers focused on AI-as-judge methods, representing a contrasting paradigm to human preference collection. The paper's emphasis on dynamic, user-driven evaluation diverges from the static benchmark tradition prevalent in neighboring leaves, and its focus on pairwise comparison and capability-oriented signals distinguishes it from the usability heuristics explored in the Usability and Interaction Quality Assessment leaf (four papers).

Among 29 candidates examined, none clearly refute the three core contributions. The platform and methodology contribution examined 10 candidates with zero refutable matches; the preference dataset contribution examined 10 candidates with zero refutable matches; and the error analysis framework examined 9 candidates with zero refutable matches. This suggests that within the limited search scope, no prior work directly overlaps with the combination of head-to-head agent comparison, cloud-hosted dynamic environments, and structured human feedback for computer-use agents. The absence of refutable candidates across all contributions indicates that the work occupies a relatively novel position, though the search scale (29 papers) leaves open the possibility of relevant work beyond the top-K semantic matches.

Given the sparse Human Preference and Comparative Evaluation leaf and the lack of refutable candidates among 29 examined papers, the work appears to address a gap in human-centric evaluation for computer-use agents. However, the analysis is constrained by the limited search scope and does not cover the full breadth of human-computer interaction or agent evaluation literature. The novelty assessment reflects what is visible within the top-30 semantic neighborhood and the constructed taxonomy, not an exhaustive field survey.

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

Contribution 1: Computer Agent Arena platform and evaluation methodology

Description: The authors introduce an open-source platform that enables pairwise evaluation of Computer-Use Agents through human preferences collected in cloud-hosted, diverse, and dynamic environments. The system simulates real-world computer use, ensures fair comparison via matched environments, and extends evaluation beyond correctness to include capability- and behavior-oriented signals.

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. Evaluation of a smart audio system based on the ViP principle and the analytic hierarchy process humanâ□□computer interaction design

URL: View paper

Brief Assessment

Smart Audio ViP[52] focuses on smart speaker design using the ViP principle and AHP for human-computer interaction evaluation, not on computer-use agent evaluation platforms or pairwise preference collection systems.

---

2. Copychats: Question sequencing with artificial agents

URL: View paper

Brief Assessment

Copychats[58] focuses on schema matching tasks using LLM-based artificial agents to evaluate question sequencing strategies, not on building a pairwise evaluation platform for computer-use agents with human preferences in cloud-hosted environments.

3. BrowserArena: Evaluating LLM Agents on Real-World Web Navigation Tasks

URL: View paper

Brief Assessment

BrowserArena[53] focuses specifically on web navigation tasks using browser automation, while the original paper presents a broader computer-use agent evaluation platform covering multi-app interactions, desktop operations, and diverse GUI tasks beyond web browsing.

4. Aligning with human judgement: The role of pairwise preference in large language model evaluators

URL: View paper

Brief Assessment

Aligning Human Judgement[51] focuses on pairwise preference evaluation for text generation tasks (summarization, story generation) using LLMs as evaluators, not on evaluating computer-use agents in interactive GUI environments with cloud-hosted VMs.

5. Human-AI Collaboration: Trade-offs Between Performance and Preferences

URL: View paper

Brief Assessment

Performance Preferences Trade-offs[56] focuses on human preferences for collaborative AI agents in a target interception task, not on building an open-source platform for pairwise evaluation of computer-use agents through human preferences in cloud-hosted environments.

6. Soft Condorcet Optimization for Ranking of General Agents

URL: View paper

Brief Assessment

Soft Condorcet Optimization[54] focuses on ranking schemes for general agents using social choice frameworks and voting methods, not on building pairwise evaluation platforms for computer-use agents with human preferences in cloud-hosted environments.

7. Who's Sorry Now: User Preferences Among Rote, Empathic, and Explanatory Apologies from LLM Chatbots

URL: View paper

Brief Assessment

User Preferences Apologies[59] focuses on user preferences for different apology types from LLM chatbots in error contexts, not on pairwise evaluation platforms for computer-use agents or human preference collection for agent capabilities.

8. An adaptive decision-making system supported on user preference predictions for humanâ□□robot interactive communication

URL: View paper

Brief Assessment

This candidate focuses on human-robot interaction with preference learning for activity selection in social robotics, not on evaluating computer-use agents through pairwise comparisons in cloud-hosted environments.

9. Agent-as-a-judge: Evaluate agents with agents

URL: View paper

Brief Assessment

Agent-as-Judge[11] focuses on using agentic systems to evaluate other agentic systems through intermediate feedback during task-solving, primarily applied to code generation tasks. It does not address pairwise evaluation platforms for computer-use agents using human preferences in cloud-hosted environments, which is the core novelty of the original paper's Computer Agent Arena platform.

10. Comparing a computer agent with a humanoid robot

URL: View paper

Brief Assessment

Computer Agent Humanoid[55] focuses on comparing human responses to computer agents versus humanoid robots in health interviews, not on building pairwise evaluation platforms for computer-use agents using human preferences.

Contribution 2: Human-centric preference dataset with 2,201 votes

Description: The authors collect and release a large-scale dataset of 2,201 filtered human preference votes across 12 agents, revealing that agent rankings differ substantially from static benchmarks. The dataset captures diverse, open-ended tasks and provides multimodal, human-labeled preference signals for future research.

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. Few-shot in-context preference learning using large language models

URL: View paper

Brief Assessment

Few-shot Preference Learning[62] focuses on synthetic and human preference data for reward function learning in RL tasks, not on computer-use agent evaluation with multimodal GUI interactions and ranking reversals relative to static benchmarks.

2. Lipo: Listwise preference optimization through learning-to-rank

URL: View paper

Brief Assessment

Lipo[60] focuses on listwise preference optimization methods for language model alignment using ranking objectives, not on collecting human preference datasets for agent evaluation or revealing ranking reversals in agent benchmarks.

3. Assessing top-preferences

URL: View paper

Brief Assessment

Assessing Top-Preferences[67] focuses on preference judgments for information retrieval ranking (TREC 2019), not computer-use agent evaluation with multimodal trajectories and diverse task environments.

4. Measuring the inconsistency of large language models in preferential ranking

URL: View paper

Brief Assessment

Inconsistency Preferential Ranking[65] focuses on measuring consistency in LLM-generated preferential rankings using formal order theory criteria, not on collecting human preference votes for agent evaluation. The candidate does not present a human-annotated preference dataset for computer-use agents.

---

5. Preference learning algorithms do not learn preference rankings

URL: View paper

Brief Assessment

Preference Learning Algorithms[63] focuses on analyzing existing preference learning algorithms (RLHF, DPO) and their ranking accuracy properties, not on collecting new human preference datasets for agent evaluation or revealing ranking reversals in agent benchmarks.

---

6. Biased perceptions of income distribution and preferences for redistribution: Evidence from a survey experiment

URL: View paper

Brief Assessment

Biased Perceptions Income[69] focuses on survey data about income distribution perceptions and redistribution preferences in Argentina, not on agent evaluation or human preference datasets for AI systems. The domains are entirely distinct.

---

7. In-context Ranking Preference Optimization

URL: View paper

Brief Assessment

In-context Ranking Preference[61] focuses on ranking preference optimization for LLMs using synthetic or task-specific datasets, not human preference evaluation of computer-use agents with multimodal trajectories.

---

8. Learning from Human Feedback: Ranking, Bandit, and Preference Optimization

URL: View paper

Brief Assessment

Learning Human Feedback[66] focuses on theoretical foundations of preference learning across ranking, bandits, and language model alignment, but does not present a human preference dataset for agent evaluation. The paper addresses different problem settings (ranking algorithms, dueling bandits, LLM alignment) rather than collecting human votes for computer-use agent evaluation.

---

9. Prd: Peer rank and discussion improve large language model based evaluations

URL: View paper

Brief Assessment

Peer Rank Discussion[64] focuses on LLM-based evaluation methods using peer ranking algorithms, not on collecting human preference datasets for agent evaluation or revealing ranking reversals in agent benchmarks.

---

10. How do people rank multiple mutant agents?

URL: View paper

Brief Assessment

Ranking Mutant Agents[68] focuses on ranking AI agents through a qualitative study with 10 participants using explanation-based assessment, not on collecting large-scale human preference votes for agent evaluation or revealing ranking reversals from static benchmarks.

Contribution 3: Error analysis and preference analysis frameworks

Description: The authors conduct systematic error analysis identifying failure modes (long-horizon memory lapses, tool-selection errors, fine-grained action failures) and preference analysis showing that users value process quality, agent-human interaction, and self-correction beyond task completion. These analyses surface alignment signals that static benchmarks overlook.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. OMMA: open architecture for Operator-guided Monitoring of Multi-step Attacks

URL: View paper

Brief Assessment

OMMA[79] focuses on multi-step attack detection in cybersecurity through event correlation and log analysis. It does not address error analysis frameworks for agent systems, long-horizon memory failures, or action failures in computer-use agents, which are the core focus of the original contribution.

---

2. Multi-Step Temperature Prognosis of Lithium-Ion Batteries for Real Electric Vehicles Based on a Novel Bidirectional Mamba Network and Sequence Adaptive â¦

URL: View paper

Brief Assessment

Multi-Step Temperature Prognosis[74] focuses on battery temperature prediction using bidirectional Mamba networks for electric vehicles, not on error analysis frameworks for agent systems or long-horizon memory failures in computer-use contexts.

---

3. Where llm agents fail and how they can learn from failures

URL: View paper

Brief Assessment

LLM Agents Failures[70] focuses on systematic error taxonomy and debugging for single-agent systems, while the original paper emphasizes human-centric preference analysis and evaluation methodology in realistic computer-use environments. The candidate does not challenge the novelty of surfacing alignment signals through user preferences.

---

4. A memory for goals model of sequence errors

URL: View paper

Brief Assessment

Memory Goals Model[78] focuses on cognitive modeling of sequence errors in routine tasks (perseveration, anticipation, omission errors) through memory activation and decay mechanisms. This is fundamentally different from the original paper's error analysis framework, which identifies failure modes in computer-use agents (long-horizon memory lapses, tool-selection errors, fine-grained action failures) and conducts preference analysis of user values in agent evaluation contexts.

---

5. Mobile-agent-e: Self-evolving mobile assistant for complex tasks

URL: View paper

Brief Assessment

Mobile-agent-e[72] focuses on mobile assistant self-evolution through tips and shortcuts, not on systematic error analysis frameworks identifying long-horizon memory lapses, tool-selection errors, and fine-grained action failures as evaluation signals that static benchmarks overlook.

---

6. Diagnostics of cognitive failures in multi-agent expert systems using dynamic evaluation protocols and subsequent mutation of the processing context

URL: View paper

Brief Assessment

Cognitive Failures Diagnostics[77] focuses on diagnostic frameworks for expert system behavior transfer in multi-agent recruiter systems, not on error analysis of long-horizon memory lapses, tool-selection errors, and fine-grained action failures in computer-use agents evaluated through human preferences.

---

7. Agentic Robot: A Brain-Inspired Framework for Vision-Language-Action Models in Embodied Agents

URL: View paper

Brief Assessment

Agentic Robot[73] focuses on robotic manipulation with visual verification and recovery mechanisms, not on error analysis frameworks for computer-use agents or preference analysis of human-agent interaction. The paper addresses execution failures in robotic tasks through a verifier module, but does not present systematic error analysis identifying long-horizon memory lapses, tool-selection errors, or preference analysis frameworks as described in the original contribution.

---

8. Self-Evaluating LLMs for Multi-Step Tasks: Stepwise Confidence Estimation for Failure Detection

URL: View paper

Brief Assessment

Self-Evaluating LLMs[76] focuses on confidence estimation and failure detection in multi-step LLM tasks through stepwise scoring methods. It does not present systematic error analysis frameworks identifying long-horizon memory lapses, tool-selection errors, or fine-grained action failures in computer-use agents, nor does it conduct preference analysis of user values beyond task completion.

---

9. Learning to correct mistakes: Backjumping in long-horizon task and motion planning

URL: View paper

Brief Assessment

Learning Correct Mistakes[71] focuses on backjumping heuristics for task and motion planning in robotics, identifying culprit actions in long-horizon planning. This is fundamentally different from the original paper's error analysis framework for computer-use agents, which identifies failure modes like long-horizon memory lapses, tool-selection errors, and fine-grained action failures in GUI-based tasks.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Computer Agent Arena: Toward Human-Centric Evaluation and Analysis of Computer-Use Agents View paper
- [1] Heuristic evaluation of conversational agents View paper
- [2] Bearcubs: A benchmark for computer-using web agents View paper
- [3] OS-Harm: A Benchmark for Measuring Safety of Computer Use Agents View paper
- [4] SusBench: An Online Benchmark for Evaluating Dark Pattern Susceptibility of Computer-Use Agents View paper
- [5] PIPA: A Unified Evaluation Protocol for Diagnosing Interactive Planning Agents View paper
- [6] A survey on near-human conversational agents View paper
- [7] Verios: Query-driven proactive human-agent-gui interaction for trustworthy os agents View paper
- [8] A user-centered framework for human-ai co-creativity View paper
- [9] Smart Transparency: A User-Centered Approach to Improving Humanâ□Machine Interaction in High-Risk Supervisory Control Tasks View paper
- [10] Evaluation agent: Efficient and promptable evaluation framework for visual generative models View paper
- [11] Agent-as-a-judge: Evaluate agents with agents View paper
- [12] Human-AI interaction research agenda: A user-centered perspective View paper
- [13] User modeling in humanâ□computer interaction View paper
- [14] Transagents: Build your translation company with language agents View paper
- [15] Webgames: Challenging general-purpose web-browsing ai agents View paper
- [16] MCU: An Evaluation Framework for Open-Ended Game Agents View paper
- [17] Personagym: Evaluating persona agents and llms View paper
- [18] Videowebarena: Evaluating long context multimodal agents with video understanding web tasks View paper
- [19] Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents View paper
- [20] OpenCUA: Open Foundations for Computer-Use Agents View paper
- [21] CloChat: Understanding how people customize, interact, and experience personas in large language models View paper
- [22] Characterizing unintended consequences in human-gui agent collaboration for web browsing View paper
- [23] CowPilot: A Framework for Autonomous and Human-Agent Collaborative Web Navigation View paper
- [24] Towards Effective Human-in-the-Loop Assistive AI Agents View paper
- [25] Ali-agent: Assessing llms' alignment with human values via agent-based evaluation View paper

- [26] Magentic-ui: Towards human-in-the-loop agentic systems View paper
- [27] Rethinking theory of mind benchmarks for llms: Towards a user-centered perspective View paper
- [28] DRBench: A Realistic Benchmark for Enterprise Deep Research View paper
- [29] Sentiment-aware design of humanâcomputer interactions: How research in humanâcomputer interaction and sentiment analysis can lead to more user-centered â View paper
- [30] Evolving agents: Interactive simulation of dynamic and diverse human personalities View paper
- [31] Hiring an AI: Incorporating Personnel Selection Methods in User-Centered Design to Design AI Agents for Safety-Critical Domains View paper
- [32] Emergence of social norms in generative agent societies: principles and architecture View paper
- [33] Decision-oriented dialogue for human-AI collaboration View paper
- [34] Human-centred test and evaluation of military AI View paper
- [35] Just Do It!? Computer-Use Agents Exhibit Blind Goal-Directedness View paper
- [36] A user-centered approach to user-building interactions View paper
- [37] Evaluation of human-ai teams for learned and rule-based agents in hanabi View paper
- [38] Evaluating the LLM agents for simulating humanoid behavior View paper
- [39] Evalai: Towards better evaluation systems for ai agents View paper
- [40] A Conceptual Model of User Experience in HumanâComputer Interaction Using Thematic Analysis View paper
- [41] Creative agents: Empowering agents with imagination for creative tasks View paper
- [42] Exploring a Gamified Personality Assessment Method through Interaction with LLM Agents Embodying Different Personalities View paper
- [43] Conversational agents for information retrieval in the education domain: A user-centered design investigation View paper
- [44] Humanâagent teaming for multirobot control: A review of human factors issues View paper
- [45] FURIOUS: Fully unified risk-assessment with interactive operational user system for vessels. View paper
- [46] The First Impression: Understanding the Impact of Multimodal System Responses on User Behavior in Task-oriented Agents View paper
- [47] Conversational agents for automated group meeting facilitation a computational framework for facilitating small group decision-making meetings View paper
- [48] Optimizing SIA Development: A Case Study in User-Centered Design for Estuary, a Multimodal Socially Interactive Agent Framework View paper
- [49] CareAssist GPT improves patient user experience with a patient centered approach to computer aided diagnosis. View paper
- [50] Performance Evaluation using Human Computer Interaction Models View paper
- [51] Aligning with human judgement: The role of pairwise preference in large language model evaluators View paper
- [52] Evaluation of a smart audio system based on the ViP principle and the analytic hierarchy process humanâcomputer interaction design View paper
- [53] BrowserArena: Evaluating LLM Agents on Real-World Web Navigation Tasks View paper
- [54] Soft Condorcet Optimization for Ranking of General Agents View paper
- [55] Comparing a computer agent with a humanoid robot View paper
- [56] Human-AI Collaboration: Trade-offs Between Performance and Preferences View paper
- [57] An adaptive decision-making system supported on user preference predictions for humanârobot interactive communication View paper
- [58] Copychats: Question sequencing with artificial agents View paper
- [59] Who's Sorry Now: User Preferences Among Rote, Empathic, and Explanatory Apologies from LLM Chatbots View paper
- [60] Lipo: Listwise preference optimization through learning-to-rank View paper
- [61] In-context Ranking Preference Optimization View paper
- [62] Few-shot in-context preference learning using large language models View paper
- [63] Preference learning algorithms do not learn preference rankings View paper
- [64] Prd: Peer rank and discussion improve large language model based evaluations View paper
- [65] Measuring the inconsistency of large language models in preferential ranking View paper
- [66] Learning from Human Feedback: Ranking, Bandit, and Preference Optimization View paper
- [67] Assessing top-preferences View paper
- [68] How do people rank multiple mutant agents? View paper
- [69] Biased perceptions of income distribution and preferences for redistribution: Evidence from a survey experiment View paper
- [70] Where llm agents fail and how they can learn from failures View paper
- [71] Learning to correct mistakes: Backjumping in long-horizon task and motion planning View paper
- [72] Mobile-agent-e: Self-evolving mobile assistant for complex tasks View paper
- [73] Agentic Robot: A Brain-Inspired Framework for Vision-Language-Action Models in Embodied Agents View paper
- [74] Multi-Step Temperature Prognosis of Lithium-Ion Batteries for Real Electric Vehicles Based on a Novel Bidirectional Mamba Network and Sequence Adaptive â View paper
- [75] FCRF: Flexible Constructivism Reflection for Long-Horizon Robotic Task Planning with Large Language Models View paper
- [76] Self-Evaluating LLMs for Multi-Step Tasks: Stepwise Confidence Estimation for Failure Detection View paper
- [77] Diagnostics of cognitive failures in multi-agent expert systems using dynamic evaluation protocols and subsequent mutation of the processing context View paper
- [78] A memory for goals model of sequence errors View paper
- [79] OMMA: open architecture for Operator-guided Monitoring of Multi-step Attacks View paper