Novelty Assessment Report

Paper: Multi-turn Evaluation of Anthropomorphic Behaviours in Large Language Models
PDF URL: https://openreview.net/pdf?id=ZAx4c4ZH5Y
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

The tendency of users to anthropomorphise large language models (LLMs) is of growing societal interest. Here, we present AnthroBench: a novel empirical method and tool for evaluating anthropomorphic LLM behaviours in realistic settings. Our work introduces three key advances; first, we develop a multi-turn evaluation of 14 distinct anthropomorphic behaviours, moving beyond single-turn assessment. Second, we present a scalable, automated approach by leveraging simulations of user interactions, enabling efficient and reproducible assessment. Third, we conduct an interactive, large-scale human subject study (N=1101) to empirically validate that the model behaviours we measure predict real users' anthropomorphic perceptions. We find that all evaluated LLMs exhibit similar behaviours, primarily characterised by relationship-building (e.g., empathy and validation) and first-person pronoun use. Crucially, we observe that the majority of these anthropomorphic behaviors only first occur after multiple turns, underscoring the necessity of multi-turn evaluations for understanding complex social phenomena in human-AI interaction. Our work provides a robust empirical foundation for investigating how design choices influence anthropomorphic model behaviours and for progressing the ethical debate on the desirability of these behaviours.

Core Task Landscape

This paper addresses: **Evaluating Anthropomorphic Behaviors in Large Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **37 categories**.

Taxonomy Overview

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

- **Psychological and Personality Trait Assessment**
- **Behavioral and Social Interaction Studies**
- **Cognitive Biases and Reasoning Patterns**
- **Human-Likeness in Language Production**
- **Domain-Specific Human-Like Behavior**
- **Role-Playing and Character Simulation**
- **Cognitive Representation and Memory**
- **Learning Dynamics and Behavioral Simulation**
- **Evaluation Methodologies and Frameworks**
- **Philosophical and Theoretical Perspectives**
- ... and 2 more categories

Complete Taxonomy Tree

- Evaluating Anthropomorphic Behaviors in Large Language Models Survey Taxonomy
- Psychological and Personality Trait Assessment
  - Psychometric Profiling and Trait Measurement (5 papers)
  - [6] PersonaLLM: Investigating the ability of large language models to express personality traits (Jiang Hang, 2024) View paper
  - [7] Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories (Max Pellert, 2024) View paper
  - [37] The personality dimensions GPT-3 expresses during human-Chatbot interactions (Nikola KovaÄeviÄ, 2024) View paper
  - [39] On the humanity of conversational ai: Evaluating the psychological portrayal of llms (J Huang, 2023) View paper
  - [42] A comparative study of large language models and human personality traits (Jiaqi Wang, 2025) View paper
  - Temporal Stability and Consistency Analysis (1 papers)
  - [35] Personality testing of large language models: limited temporal stability, but highlighted prosociality (Bojana BodroÅ¾a, 2024) View paper
  - Comprehensive Psychological Measurement Frameworks (1 papers)
  - [36] Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications (Dong Wenhan, 2025) View paper
- Behavioral and Social Interaction Studies
  - Prosocial and Economic Decision-Making (4 papers)
  - [9] Do large language models learn human-like strategic preferences? (Jesse Roberts, 2025) View paper
  - [18] Risk and prosocial behavioural cues elicit human-like response patterns from AI chatbots (Yukun Zhao, 2024) View paper
  - [22] Can machines think like humans? a behavioral evaluation of llm-agents in dictator games (Ma, 2024) View paper
  - [30] Can LLMs Mimic Human-Like Mental Accounting and Behavioral Biases? (Yan Leng, 2024) View paper
  - Multi-Turn Anthropomorphic Behavior Evaluation ★ (1 papers)
  - [0] Multi-turn Evaluation of Anthropomorphic Behaviours in Large Language Models (Anon et al., 2026) View paper

- Social Misattribution and Role Perception (1 papers)
- [40] Social misattributions in conversations with large language models (Andrea Ferrario, 2025) View paper
- Cognitive Biases and Reasoning Patterns
- Systematic Cognitive Bias Evaluation (2 papers)
- [14] Do LLMs Exhibit Human-Like Cognitive Biases? A Large-Scale Systematic Evaluation (Tomer Geva, 2025) View paper
- [27] Do Large Language Models Show Human-like Biases? Exploring Confidence - Competence Gap in AI (Aniket Kumar Singh, 2024) View paper
- Content Effects and Reasoning Patterns (1 papers)
- [4] Language models show human-like content effects on reasoning (Dasgupta, 2022) View paper
- Theory of Mind and Social Cognition (1 papers)
- [15] Do llms exhibit human-like reasoning? evaluating theory of mind in llms for open-ended responses (MartÃn ElÃas, 2024) View paper
- Overthinking and Computational Resource Allocation (1 papers)
- [3] Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs (Chen, 2024) View paper
- Human-Likeness in Language Production
- Dialogue and Conversational Quality (1 papers)
- [25] Real or robotic? Assessing whether LLMs accurately simulate qualities of human responses in dialogue (Jonathan Ivey, 2024) View paper
- Psycholinguistic and Linguistic Structure (1 papers)
- [41] HLB: Benchmarking LLMs' Humanlikeness in Language Use (Duan, 2024) View paper
- Human-Like Tone and Social Perception (1 papers)
- [49] HumT DumT: Measuring and controlling human-like language in LLMs (Myra Cheng, 2025) View paper
- Typing and Temporal Production Patterns (1 papers)
- [44] Beyond Words: Infusing Conversational Agents with Human-like Typing Behaviors (Jijie Zhou, 2024) View paper
- Domain-Specific Human-Like Behavior
- Autonomous Driving and Vehicle Interaction (3 papers)
- [1] Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles (Can Cui, 2024) View paper
- [5] Drive like a human: Rethinking autonomous driving with large language models (Daocheng Fu, 2024) View paper
- [23] Exploring human-SAV interaction using large language models: The impact of psychological ownership and anthropomorphism on user experience (L Guo, 2025) View paper
- Educational and Pedagogical Applications (3 papers)
- [13] Leveraging LLM respondents for item evaluation: A psychometric analysis (Yunting Liu, 2025) View paper
- [17] Towards human-like educational question generation with large language models (Zichao Wang, 2022) View paper
- [19] Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors (Kaushal Kumar Maurya, 2025) View paper
- Translation and Cross-Lingual Evaluation (2 papers)
- [10] Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt (LU Qingyu, 2023) View paper
- [31] Evaluating o1-Like LLMs: Unlocking Reasoning for Translation through Comprehensive Analysis (Chen An-dong, 2025) View paper
- Spatial Navigation and Instruction Generation (1 papers)
- [20] Can llms generate human-like wayfinding instructions? towards platform-agnostic embodied instruction synthesis (Vishnu Sashank Dorbala, 2024) View paper
- Journalism and Content Planning (1 papers)
- [28] Do LLMs Plan Like Human Writers? Comparing Journalist Coverage of Press Releases with LLMs (Alexander Spangher, 2024) View paper
- Role-Playing and Character Simulation
- Character Customization and Dialogue Systems (1 papers)
- [16] Characterglm: Customizing chinese conversational ai characters with large language models (Zhou Jin-feng, 2023) View paper
- Role Identification and Attribution (1 papers)
- [47] PersonaEval: Are LLM Evaluators Human Enough to Judge Role-Play? (Zhou Ling-feng, 2025) View paper
- Non-Player Character Design and Evaluation (1 papers)
- [32] Evaluating the efficacy of LLMs to emulate realistic human personalities (Lawrence J. Klinkert, 2024) View paper
- Cognitive Representation and Memory
- Object Concept Representations (1 papers)
- [48] Human-like object concept representations emerge naturally in multimodal large language models (Changde Du, 2024) View paper
- Memory Recall and Consolidation (2 papers)
- [11] "My agent understands me better": Integrating Dynamic Human-like Memory Recall and Consolidation in LLM-Based Agents (Miyashita, 2024) View paper
- [46] Aspects of human memory and large language models (Janik, 2023) View paper
- Computational Ingredients of Human-Like Representations (1 papers)
- [50] Uncovering the Computational Ingredients of Human-Like Representations in LLMs (Rogers, 2025) View paper
- Learning Dynamics and Behavioral Simulation
- Learning Process Simulation (1 papers)
- [2] Simulating human-like learning dynamics with llm-empowered agents (Yuan Yu, 2025) View paper
- General Agent Behavior Simulation (1 papers)
- [45] Evaluating the LLM agents for simulating humanoid behavior (C Chen, 2024) View paper
- Evaluation Methodologies and Frameworks
- LLM-as-Judge and Evaluation Alignment (1 papers)
- [26] Ask me like i'm human: Llm-based evaluation with for-human instructions correlates better with human evaluations than human judges (Rudali Huidrom, 2025) View paper

- Response Bias and Survey Design (1 papers)
  - [8] Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design (Lindia Tjuatja, 2023) View paper
  - Structured Reasoning and Out-of-Distribution Generalization (1 papers)
  - [33] Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks (Collins, 2022) View paper
  - Human-Centric LLM Capability Surveys (1 papers)
  - [21] A survey on human-centric llms (Wang Jing-yi, 2024) View paper
- Philosophical and Theoretical Perspectives
  - Understanding and Intelligence in LLMs (1 papers)
  - [34] Do large language models understand us? (Blaise Aguera y Arcas, 2022) View paper
  - Trait Manifestation and Controllability (1 papers)
  - [12] Tracing Human-like Traits in LLMs: Origins, Real-World Manifestation, and Controllability (P Han, 2025) View paper
  - Uncanny Valley in AI-Generated Content (1 papers)
  - [38] The Uncanny Valley: An Empirical Study on Human Perceptions of AI-Generated Text and Images (Kishnani, 2025) View paper
- Multimodal and Embodied Interaction
  - Augmented Reality Social Assistance (1 papers)
  - [29] SocialMind: LLM-based Proactive AR Social Assistive System with Human-like Perception for In-situ Live Interactions (Bufang Yang, 2024) View paper
  - Vision-Language Reasoning and Exploration (1 papers)
  - [24] ZoomEye: Enhancing Multimodal LLMs with Human-Like Zooming Capabilities through Tree-Based Image Exploration (Haozhan Shen, 2024) View paper
- Planning and Strategic Reasoning (1 papers)
  - [43] A human-like reasoning framework for multi-phases planning task with large language models (Zou, 2024) View paper

Narrative

Core task: evaluating anthropomorphic behaviors in large language models. The field has grown into a rich taxonomy spanning twelve major branches, each addressing distinct facets of how LLMs exhibit or fail to exhibit human-like characteristics. Psychological and Personality Trait Assessment examines whether models display stable traits measurable by psychometric instruments (e.g., AI Psychometrics[7], Personality Testing Stability[35]), while Behavioral and Social Interaction Studies investigates multi-turn conversational dynamics and prosocial cues (Prosocial Behavioural Cues[18]). Cognitive Biases and Reasoning Patterns explores whether models replicate human fallacies such as mental accounting (Mental Accounting Biases[30]) or content effects (Content Effects Reasoning[4]), and Human-Likeness in Language Production scrutinizes stylistic and typographic behaviors (Typing Behaviors[44]). Domain-Specific Human-Like Behavior targets specialized contexts like driving (Drive As You Speak[1], Drive Like Human[5]) or tutoring (AI Tutor Evaluation[19]), whereas Role-Playing and Character Simulation focuses on persona consistency (CharacterGLM[16], PersonaLLM[6]). Additional branches cover memory mechanisms (Dynamic Memory Recall[11]), learning trajectories (Human-like Learning Dynamics[2]), evaluation frameworks (HLB Humanlikeness Benchmark[41]), philosophical debates (Social Misattributions[40]), multimodal embodiment (ZoomEye[24]), and strategic planning (Multi-phases Planning[43]).

Several active lines of work highlight contrasting emphases and open questions. One cluster examines whether anthropomorphism is a stable property or an emergent artifact of prompting and context, with studies like Response Biases Survey[8] and Tracing Human-like Traits[12] documenting variability across tasks. Another thread investigates the gap between surface-level mimicry and genuine cognitive alignment, as seen in debates over whether models truly understand (Do LLMs Understand[34]) or merely simulate plausible outputs (Simulating Humanoid Behavior[45]). The original paper, Anthropomorphic Behaviours Evaluation[0], sits within the Behavioral and Social Interaction Studies branch, specifically targeting multi-turn anthropomorphic behavior evaluation. Its focus on extended conversational sequences aligns it closely with works assessing dynamic social cues and interaction realism, contrasting with single-shot psychometric approaches (LLM Respondents Psychometric[13]) or domain-specific simulations (Human-SAV Interaction[23]). By emphasizing temporal consistency and interactive authenticity, it addresses a key challenge: distinguishing transient prompt-driven responses from robust, human-like behavioral patterns across sustained exchanges.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on measuring anthropomorphic behaviors through multi-turn conversations with empirical human validation, emphasizing the temporal and interactive dimension of human-likeness. The sibling subtopics examine specific facets of anthropomorphic behavior: one investigates prosocial and economic decision-making in experimental settings, while the other studies how humans perceive and attribute social roles to LLMs. All three areas contribute to understanding different aspects of LLM anthropomorphism, but differ in their methodological approaches and specific behavioral dimensions.

Similarities: - All three subtopics address anthropomorphic behaviors in LLMs from an empirical perspective - Each involves some form of human evaluation or human-centered measurement - All examine behavioral rather than purely technical aspects of LLM outputs - Each subtopic contributes to understanding how LLMs exhibit or are perceived to exhibit human-like characteristics

Differences: - The original leaf emphasizes multi-turn conversational dynamics, while Prosocial and Economic Decision-Making focuses on specific experimental game scenarios that may be single or multi-turn - Social Misattribution examines human perception and attribution processes, whereas the original leaf measures actual behavioral patterns in LLM outputs - Prosocial and Economic Decision-Making targets specific behavioral domains (altruism, cooperation, risk-taking), while the original leaf appears broader in scope regarding anthropomorphic behaviors - The original leaf explicitly requires empirical human validation across turns, while Social Misattribution studies how humans naturally attribute roles without necessarily validating behavior quality

Suggested Search Directions: - Investigate how multi-turn conversational patterns influence prosocial decision-making in LLMs - Explore whether social role misattribution by humans changes across multiple conversational turns - Examine the relationship between economic game performance and broader anthropomorphic behavior metrics in extended interactions

Sibling Subtopics

- **Prosocial and Economic Decision-Making** (leaves: 1, papers: 4)
- Scope: Studies assessing LLM altruism, cooperation, risk-taking, and economic behaviors in experimental game settings.
- Exclude: Excludes cognitive bias evaluations; see Cognitive Biases and Reasoning Patterns.
- **Social Misattribution and Role Perception** (leaves: 1, papers: 1)
- Scope: Studies examining how humans attribute social roles and identities to LLMs in conversational contexts.
- Exclude: Excludes role-playing quality evaluation; see Role-Playing and Character Simulation.

Contributions Analysis

Overall novelty summary. The paper introduces AnthroBench, a multi-turn evaluation framework for measuring anthropomorphic behaviors in LLMs through automated user simulations and empirical human validation. It resides in the 'Multi-Turn Anthropomorphic Behavior Evaluation' leaf under 'Behavioral and Social Interaction Studies', where it is currently the sole occupant. This positioning reflects a relatively sparse research direction: while the broader 'Behavioral and Social Interaction Studies' branch contains sibling leaves examining prosocial decision-making and social role misattribution, no prior work in the taxonomy explicitly targets sustained multi-turn anthropomorphic behavior assessment with human validation at scale.

The taxonomy reveals neighboring work in adjacent branches that address related but distinct concerns. 'Psychological and Personality Trait Assessment' focuses on psychometric profiling using standardized inventories, emphasizing trait stability rather than interactive behavioral dynamics. 'Human-Likeness in Language Production' examines stylistic and typographic features in generated text, excluding the multi-turn conversational context central to this paper. 'Role-Playing and Character Simulation' investigates persona consistency and character-specific dialogue generation, but does not systematically measure anthropomorphic perceptions across diverse conversational turns. The paper's emphasis on temporal emergence of behaviors across extended interactions distinguishes it from these single-turn or static assessment paradigms.

Among thirty candidates examined through semantic search and citation expansion, none clearly refute the three core contributions. The multi-turn evaluation method (ten candidates examined, zero refutable) appears novel within the limited search scope, as does the automated simulation approach (ten candidates, zero refutable) and the large-scale human validation study linking measured behaviors to user perceptions (ten candidates, zero refutable). This absence of overlapping prior work suggests the integration of multi-turn assessment, automated simulation, and empirical validation represents a distinctive methodological package, though the limited search scale means potentially relevant work outside the top-thirty matches may exist.

The analysis indicates the paper occupies a methodologically underexplored niche, combining elements from behavioral evaluation, simulation-based testing, and human-subjects research in a way not captured by existing taxonomy leaves. However, the search examined only thirty candidates from semantic neighborhoods, not an exhaustive survey of human-AI interaction or conversational AI literature. The novelty assessment reflects what is visible within this bounded scope, acknowledging that broader literature searches or domain-specific venues might reveal closer precedents.

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

Contribution 1: AnthroBench: Multi-turn evaluation method and tool for anthropomorphic LLM behaviours

Description: The authors introduce AnthroBench, a comprehensive evaluation framework that assesses 14 distinct anthropomorphic behaviours in large language models through multi-turn dialogues. The method uses automated user simulations to generate realistic conversations and employs multiple LLM judges to detect anthropomorphic behaviours across different interaction contexts.

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. MedKGEval: A Knowledge Graph-Based Multi-Turn Evaluation Framework for Open-Ended Patient Interactions with Clinical LLMs

URL: View paper

Brief Assessment

MedKGEval[73] focuses on evaluating clinical LLMs in medical doctor-patient interactions using knowledge graph-driven patient simulation, not on assessing anthropomorphic behaviors in general conversational AI systems.

---

2. MemoryBank: Enhancing Large Language Models with Long-Term Memory

URL: View paper

Brief Assessment

MemoryBank[71] focuses on long-term memory mechanisms for LLMs in companion scenarios, not on systematic multi-turn evaluation of anthropomorphic behaviors. While it mentions anthropomorphism in memory updating, it does not present a comprehensive evaluation framework for measuring 14 distinct anthropomorphic behaviors across multiple interaction contexts as AnthroBench does.

---

3. Hierarchical Prompting Taxonomy: A Universal Evaluation Framework for Large Language Models Aligned with Human Cognitive Principles

URL: View paper

Brief Assessment

Hierarchical Prompting Taxonomy[79] focuses on evaluating LLM task complexity through cognitive-load-based prompting strategies, not on multi-turn evaluation of anthropomorphic behaviors. The candidate addresses prompt optimization and cognitive demands, while the original paper specifically targets anthropomorphic behavior assessment in conversational contexts.

---

4. Towards Anthropomorphic Conversational AI Part I: A Practical Framework

URL: View paper

Brief Assessment

Anthropomorphic Conversational Framework[72] focuses on building anthropomorphic conversational systems through a multi-module framework for real-time interaction, not on evaluating or benchmarking anthropomorphic behaviors in LLMs through automated multi-turn dialogues.

---

5. 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

In-vehicle Conversational Agents[77] focuses on evaluating conversational agents in driving contexts, measuring driving performance and user experience rather than systematically evaluating anthropomorphic behaviors across multiple dimensions in LLMs through automated multi-turn dialogues.

---

6. Overcoming Multi-step Complexity in Multimodal Theory-of-Mind Reasoning: A Scalable Bayesian Planner

URL: View paper

Brief Assessment

Multimodal Theory-of-Mind[78] focuses on Bayesian inverse planning for inferring mental states (goals, beliefs) in multimodal environments, not on evaluating anthropomorphic behaviors in conversational AI systems. The candidate addresses Theory-of-Mind

reasoning in physical simulation scenarios, while the original contribution targets anthropomorphic behavior detection across multi-turn dialogues.

---

7. Esc-eval: Evaluating emotion support conversations in large language models

URL: View paper

Brief Assessment

Esc-eval[74] focuses on evaluating emotion support conversations through role-playing agents, not anthropomorphic behaviors. The evaluation dimensions (fluency, empathy, information, etc.) and the problem domain (emotional support) are fundamentally different from AnthroBench's focus on anthropomorphism assessment.

---

8. A Study on Individual Spatiotemporal Activity Generation Method Using MCP-Enhanced Chain-of-Thought Large Language Models

URL: View paper

Brief Assessment

Spatiotemporal Activity Generation[75] focuses on urban spatiotemporal behavior simulation using MCP-enhanced chain-of-thought reasoning for activity generation, not on evaluating anthropomorphic behaviors in conversational AI systems through multi-turn dialogues.

---

9. Towards more accurate US presidential election via multi-step reasoning with large language models

URL: View paper

Brief Assessment

Presidential Election Reasoning[76] focuses on election prediction using multi-step reasoning for political forecasting, not on evaluating anthropomorphic behaviors in LLMs through multi-turn dialogues. The candidate addresses a completely different domain (political science) with different objectives (vote prediction vs. anthropomorphism assessment).

---

10. Real or robotic? Assessing whether LLMs accurately simulate qualities of human responses in dialogue

URL: View paper

Brief Assessment

Real or Robotic[25] focuses on evaluating how well LLMs simulate human responses in dialogue using similarity metrics across lexical, syntactic, semantic, and stylistic features. This differs from AnthroBench's focus on detecting and measuring specific anthropomorphic behaviors (e.g., empathy, validation, first-person pronoun use) in LLM outputs through multi-turn dialogues.

Contribution 2: Scalable automated multi-turn evaluation approach using user simulations

Description: The authors develop a fully automated evaluation pipeline that simulates multi-turn user interactions with AI systems, moving beyond single-turn assessments. This approach enables scalable and reproducible measurement of anthropomorphic behaviours as they emerge across extended conversations rather than isolated exchanges.

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. Llms get lost in multi-turn conversation

URL: View paper

Brief Assessment

Lost Multi-turn Conversation[61] focuses on evaluating performance degradation in multi-turn conversations through sharded instruction simulation, not on anthropomorphic behavior evaluation. The candidate's simulation framework serves a different purpose than the original's automated evaluation of anthropomorphic behaviors.

---

2. From Intents to Conversations: Generating Intent-Driven Dialogues with Contrastive Learning for Multi-Turn Classification

URL: View paper

Brief Assessment

Intent-Driven Dialogues[63] focuses on generating intent-driven dialogues for training multi-turn intent classification models in e-commerce, not on evaluating anthropomorphic behaviors in LLMs through user simulations.

---

3. Intent-aware dialogue generation and multi-task contrastive learning for multi-turn intent classification

URL: View paper

Brief Assessment

Intent-aware Dialogue Generation[67] focuses on generating intent-driven dialogues for e-commerce chatbots using HMM-LLM integration, not on evaluating anthropomorphic behaviors in LLMs across multi-turn conversations. The candidate's user simulation serves dialogue generation purposes rather than behavioral evaluation.

---

4. Balancing Accuracy and Efficiency in Multi-Turn Intent Classification for LLM-Powered Dialog Systems in Production

URL: View paper

Brief Assessment

Balancing Accuracy Efficiency[68] focuses on multi-turn intent classification for production dialogue systems using LLM-based data augmentation and pseudo-labeling, not on evaluating anthropomorphic behaviors through user simulations. The candidate's automated approach serves a fundamentally different purpose (improving classification accuracy) compared to the original's evaluation methodology (measuring anthropomorphic behaviors).

---

5. IntellAgent: A Multi-Agent Framework for Evaluating Conversational AI Systems

URL: View paper

Brief Assessment

IntellAgent[62] focuses on evaluating conversational AI systems through policy-driven graph modeling and API integration for task-oriented dialogues, rather than measuring anthropomorphic behaviors across extended conversations as in the original paper.

---

6. MathChat: Benchmarking Mathematical Reasoning and Instruction Following in Multi-Turn Interactions

URL: View paper

Brief Assessment

MathChat[69] focuses on evaluating mathematical reasoning and instruction following in multi-turn interactions, not on simulating user behaviors for anthropomorphism assessment. The evaluation methodology and target constructs differ fundamentally from the original paper's approach to measuring social phenomena in human-AI interaction.

---

7. MemeCMD: An Automatically Generated Chinese Multi-turn Dialogue Dataset with Contextually Retrieved Memes

URL: View paper

Brief Assessment

MemeCMD[70] focuses on automatically generating Chinese multi-turn dialogues with meme retrieval, not on evaluating anthropomorphic behaviors through user simulations. The candidate uses dual agents for dialogue generation rather than simulating user interactions for evaluation purposes.

---

8. Towards a Zero-Data, Controllable, Adaptive Dialog System

URL: View paper

Brief Assessment

Zero-Data Adaptive Dialog[66] focuses on generating synthetic training data from dialog trees for RL agents in task-oriented dialog systems, not on evaluating anthropomorphic behaviors through multi-turn user simulations. The candidate addresses controllable dialog navigation and domain adaptation, while the original paper develops an evaluation methodology for measuring social phenomena in conversational AI.

---

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

URL: View paper

Brief Assessment

User Language Models[64] focuses on training user simulators for evaluating assistant LMs in multi-turn conversations, not on evaluating anthropomorphic behaviors in LLMs. The candidate's automated approach simulates users to test assistant performance on coding/math tasks, while the original paper evaluates anthropomorphic behaviors across multiple turns using a different methodology (user LLM conversing with target LLM to elicit behaviors).

---

10. Comparing Data Augmentation Methods for End-to-End Task-Oriented Dialog Systems

URL: View paper

Brief Assessment

Data Augmentation Methods[65] focuses on data augmentation techniques for training task-oriented dialog systems, not on automated evaluation methodologies using user simulations for multi-turn interactions.

Contribution 3: Empirical validation through large-scale human subject study

Description: The authors validate their automated evaluation method through a controlled experiment with 1,101 human participants who interacted with AI systems exhibiting different levels of anthropomorphic behaviours. The study demonstrates that their automated measurements correlate with both explicit survey responses and implicit behavioural indicators of human anthropomorphic perceptions.

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. Seeing personhood in machines: conceptualizing anthropomorphism of social robots

URL: View paper

Brief Assessment

Anthropomorphism Social Robots[53] focuses on developing and validating a measurement scale for social robot anthropomorphism through qualitative and quantitative methods. The original paper validates an automated evaluation method for LLM anthropomorphic behaviors through interactive experiments, representing a fundamentally different research approach and domain.

---

2. To be or not to beâ¦ human? Theorizing the role of human-like competencies in conversational artificial intelligence agents

URL: View paper

Brief Assessment

Human-like Competencies Role[51] focuses on user engagement with conversational AI through cognitive, relational, and emotional competencies using a two-wave survey and qualitative interviews. This differs from the original paper's validation approach, which specifically tests automated anthropomorphism measurements against human perceptions in controlled experimental conditions with 1,101 participants.

---

3. Evaluating trust in recommender systems: A user study on the impacts of explanations, agency attribution, and product types

URL: View paper

Brief Assessment

Trust Recommender Systems[56] focuses on trust in recommender systems through agency attribution (n=268), not on validating anthropomorphic perceptions in conversational AI systems through multi-turn interactions (n=1,101).

---

4. Posthumanist Phenomenology and Artificial Intelligence

URL: View paper

Brief Assessment

Posthumanist Phenomenology[59] appears to be a philosophical/theoretical paper discussing how AI models 'experience' data. The minimal context provided shows no evidence of conducting human subject studies or validating anthropomorphic perceptions empirically.

---

5. Innovation in tune: an empirical investigation of user acceptance of artificial intelligence-generated music

URL: View paper

Brief Assessment

AI-Generated Music Acceptance[52] focuses on anthropomorphism in AI-generated singing using survey-based acceptance models (TAM/UTAUT2), not on validating automated evaluation methods for anthropomorphic behaviors in conversational AI systems through controlled experiments.

---

6. â¦ AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence â¦

URL: View paper

Brief Assessment

Interaction Quality Empathy[58] focuses on anthropomorphic characteristics and empathy in AI devices within service industry contexts, not on validating automated evaluation methods for anthropomorphic behaviors through controlled experiments with behavioral and survey measures.

---

7. How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents

URL: View paper

Brief Assessment

Intelligence Anthropomorphism Adoption[54] focuses on measuring user perceptions of intelligence and anthropomorphism in personal intelligent agents, not on validating automated evaluation methods for anthropomorphic behaviors in LLMs through controlled experiments.

---

8. An Exploratory Study Into the Impact of AI Literacy Training on Anthropomorphism and Trust in Conversational AI

URL: View paper

Brief Assessment

AI Literacy Training[60] focuses on the impact of AI literacy training on anthropomorphism and trust, not on validating automated evaluation methods for anthropomorphic behaviors. The candidate's human subject study examines training interventions, while the original validates automated measurements against human perceptions.

---

9. The Partner Modelling Questionnaire: A validated self-report measure of perceptions toward machines as dialogue partners

URL: View paper

Brief Assessment

Partner Modelling Questionnaire[55] focuses on developing and validating a self-report scale to measure partner models of speech interfaces, not on validating automated evaluation methods for anthropomorphic behaviors through human subject studies.

---

10. The role of user perceptions of intelligence, anthropomorphism, and self-extension on continuance of use of personal intelligent agents

URL: View paper

Brief Assessment

Self-Extension Continuance[57] examines user perceptions of personal intelligent agents (Siri, Alexa) through surveys measuring intelligence, anthropomorphism, and self-extension constructs. The original paper validates automated measurements of anthropomorphic behaviors in LLMs through interactive experiments where participants converse with AI systems. These are fundamentally different validation approaches: Self-Extension Continuance[57] uses survey-based perception measures for voice assistants, while the original validates automated behavioral detection methods for conversational AI through both explicit surveys and implicit behavioral measures.

Appendix: Text Similarity Detection

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

References

- [0] Multi-turn Evaluation of Anthropomorphic Behaviours in Large Language Models View paper
- [1] Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles View paper
- [2] Simulating human-like learning dynamics with llm-empowered agents View paper
- [3] Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs View paper
- [4] Language models show human-like content effects on reasoning View paper
- [5] Drive like a human: Rethinking autonomous driving with large language models View paper
- [6] PersonaLLM: Investigating the ability of large language models to express personality traits View paper
- [7] Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories View paper
- [8] Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design View paper
- [9] Do large language models learn human-like strategic preferences? View paper
- [10] Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt View paper
- [11] "My agent understands me better": Integrating Dynamic Human-like Memory Recall and Consolidation in LLM-Based Agents View paper
- [12] Tracing Human-like Traits in LLMs: Origins, Real-World Manifestation, and Controllability View paper
- [13] Leveraging LLM respondents for item evaluation: A psychometric analysis View paper
- [14] Do LLMs Exhibit Human-Like Cognitive Biases? A Large-Scale Systematic Evaluation View paper
- [15] Do llms exhibit human-like reasoning? evaluating theory of mind in llms for open-ended responses View paper
- [16] Characterglm: Customizing chinese conversational ai characters with large language models View paper
- [17] Towards human-like educational question generation with large language models View paper
- [18] Risk and prosocial behavioural cues elicit human-like response patterns from AI chatbots View paper
- [19] Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors View paper
- [20] Can llms generate human-like wayfinding instructions? towards platform-agnostic embodied instruction synthesis View paper
- [21] A survey on human-centric llms View paper
- [22] Can machines think like humans? a behavioral evaluation of llm-agents in dictator games View paper
- [23] Exploring human-SAV interaction using large language models: The impact of psychological ownership and anthropomorphism on user experience View paper
- [24] ZoomEye: Enhancing Multimodal LLMs with Human-Like Zooming Capabilities through Tree-Based Image Exploration View paper
- [25] Real or robotic? Assessing whether LLMs accurately simulate qualities of human responses in dialogue View paper
- [26] Ask me like i'm human: Llm-based evaluation with for-human instructions correlates better with human evaluations than human judges View paper

- [27] Do Large Language Models Show Human-like Biases? Exploring Confidence - Competence Gap in AI View paper
- [28] Do LLMs Plan Like Human Writers? Comparing Journalist Coverage of Press Releases with LLMs View paper
- [29] SocialMind: LLM-based Proactive AR Social Assistive System with Human-like Perception for In-situ Live Interactions View paper
- [30] Can LLMs Mimic Human-Like Mental Accounting and Behavioral Biases? View paper
- [31] Evaluating o1-Like LLMs: Unlocking Reasoning for Translation through Comprehensive Analysis View paper
- [32] Evaluating the efficacy of LLMs to emulate realistic human personalities View paper
- [33] Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks View paper
- [34] Do large language models understand us? View paper
- [35] Personality testing of large language models: limited temporal stability, but highlighted prosociality View paper
- [36] Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications View paper
- [37] The personality dimensions GPT-3 expresses during human-Chatbot interactions View paper
- [38] The Uncanny Valley: An Empirical Study on Human Perceptions of AI-Generated Text and Images View paper
- [39] On the humanity of conversational ai: Evaluating the psychological portrayal of llms View paper
- [40] Social misattributions in conversations with large language models View paper
- [41] HLB: Benchmarking LLMs' Humanlikeness in Language Use View paper
- [42] A comparative study of large language models and human personality traits View paper
- [43] A human-like reasoning framework for multi-phases planning task with large language models View paper
- [44] Beyond Words: Infusing Conversational Agents with Human-like Typing Behaviors View paper
- [45] Evaluating the LLM agents for simulating humanoid behavior View paper
- [46] Aspects of human memory and large language models View paper
- [47] PersonaEval: Are LLM Evaluators Human Enough to Judge Role-Play? View paper
- [48] Human-like object concept representations emerge naturally in multimodal large language models View paper
- [49] HumT DumT: Measuring and controlling human-like language in LLMs View paper
- [50] Uncovering the Computational Ingredients of Human-Like Representations in LLMs View paper
- [51] To be or not to beâ human? Theorizing the role of human-like competencies in conversational artificial intelligence agents View paper
- [52] Innovation in tune: an empirical investigation of user acceptance of artificial intelligence-generated music View paper
- [53] Seeing personhood in machines: conceptualizing anthropomorphism of social robots View paper
- [54] How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents View paper
- [55] The Partner Modelling Questionnaire: A validated self-report measure of perceptions toward machines as dialogue partners View paper
- [56] Evaluating trust in recommender systems: A user study on the impacts of explanations, agency attribution, and product types View paper
- [57] The role of user perceptions of intelligence, anthropomorphism, and self-extension on continuance of use of personal intelligent agents View paper
- [58] â AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence â View paper
- [59] Posthumanist Phenomenology and Artificial Intelligence View paper
- [60] An Exploratory Study Into the Impact of AI Literacy Training on Anthropomorphism and Trust in Conversational AI View paper
- [61] Llms get lost in multi-turn conversation View paper
- [62] IntellAgent: A Multi-Agent Framework for Evaluating Conversational AI Systems View paper
- [63] From Intents to Conversations: Generating Intent-Driven Dialogues with Contrastive Learning for Multi-Turn Classification View paper
- [64] Flipping the dialogue: Training and evaluating user language models View paper
- [65] Comparing Data Augmentation Methods for End-to-End Task-Oriented Dialog Systems View paper
- [66] Towards a Zero-Data, Controllable, Adaptive Dialog System View paper
- [67] Intent-aware dialogue generation and multi-task contrastive learning for multi-turn intent classification View paper
- [68] Balancing Accuracy and Efficiency in Multi-Turn Intent Classification for LLM-Powered Dialog Systems in Production View paper
- [69] MathChat: Benchmarking Mathematical Reasoning and Instruction Following in Multi-Turn Interactions View paper
- [70] MemeCMD: An Automatically Generated Chinese Multi-turn Dialogue Dataset with Contextually Retrieved Memes View paper
- [71] MemoryBank: Enhancing Large Language Models with Long-Term Memory View paper
- [72] Towards Anthropomorphic Conversational AI Part I: A Practical Framework View paper
- [73] MedKGEval: A Knowledge Graph-Based Multi-Turn Evaluation Framework for Open-Ended Patient Interactions with Clinical LLMs View paper
- [74] Esc-eval: Evaluating emotion support conversations in large language models View paper
- [75] A Study on Individual Spatiotemporal Activity Generation Method Using MCP-Enhanced Chain-of-Thought Large Language Models View paper
- [76] Towards more accurate US presidential election via multi-step reasoning with large language models View paper
- [77] ChatGPT on the Road: Leveraging Large Language Model-Powered In-vehicle Conversational Agents for Safer and More Enjoyable Driving Experience View paper
- [78] Overcoming Multi-step Complexity in Multimodal Theory-of-Mind Reasoning: A Scalable Bayesian Planner View paper
- [79] Hierarchical Prompting Taxonomy: A Universal Evaluation Framework for Large Language Models Aligned with Human Cognitive Principles View paper