Can LLMs Tell When They Sound Human? A Study of Guided and Unguided LLM Judgments of Anthropomorphic Language

Sabrina Villata[1], Anna Bonarelli[1], Luigi Di Caro[1] and Amon Rapp[1,*]

[1]University of Turin, C.so Svizzera, 185, Turin, Italy

Abstract

Anthropomorphism represents a major challenge in human-chatbot interaction, as users often attribute human characteristics to conversational agents despite their purely simulated nature. The issue is particularly relevant for Large Language Models (LLMs), whose advanced conversational abilities closely resemble human communication, reinforcing attributions of human-like traits. Identifying anthropomorphic language in responses generated by LLM-based chatbots is important because it can significantly influence user experience while also fostering inappropriate trust, potentially harmful decisions, and excessive reliance on these systems. In this paper, we investigate whether LLMs can recognize anthropomorphic language in responses generated by LLM-based chatbots by comparing human annotations with LLM judgments under two conditions: one guided by an existing taxonomy of anthropomorphic linguistic expressions and one without such guidance. By analyzing 1,000 real-world conversational turns from five different LLM-based chatbots, we find that explicit guidance improves precision but does not improve recall for subtler stylistic and paralinguistic markers of humanness, which remain a blind spot for both guided and unguided LLM judges. We also find substantial variation across the five LLM-based chatbots in the degree to which anthropomorphic language is expressed and identify qualitative differences in how anthropomorphism manifests across them. Based on these findings, we discuss open research questions surrounding anthropomorphism in LLM-based chatbots and its automatic detection.

Keywords

LLM, Generative AI, Humanness, Anthropomorphism

1. Introduction

A key challenge in the study of text-based conversational agents, or chatbots, concerns users' tendency to anthropomorphize them: although these systems are designed only to simulate selected human-like characteristics, people often perceive them as possessing genuinely human attributes [1]. Perceived humanness strongly shapes how users experience and evaluate the conversation with chatbots [2], and prior work has documented both its benefits (e.g., increased trust [3], intersubjectivity and social closeness [4, 5], prosocial behavior [6]) and its risks, including inappropriate trust, suboptimal or harmful decisions, adverse psychological outcomes [7], uncanny reactions [8, 5, 9, 10], and excessive reliance driven by the illusion that they genuinely understand or share the user's intentions [11].

This issue becomes particularly critical when artificial agents are intentionally designed to present themselves in ways that closely resemble human communication. Notably, even simple design decisions, such as giving a name to the agent [12] or making it proactively guide the interaction [13], can substantially increase perceptions of humanness. This tendency is especially relevant for contemporary Large Language Models (LLMs), whose conversational abilities are capable of producing fluent, contextually appropriate responses that closely mimic human conversational style, emotional expressions, and other linguistic cues typically associated with human interaction [14]. Such characteristics may reinforce

AIxIA 2026 Discussion Papers, 24th International Conference of the Italian Association for Artificial Intelligence, Perugia, Italy, 2026

*Corresponding author.

sabrina.villata@unito.it (S. Villata); anna.bonarelli@unito.it (A. Bonarelli); luigi.dicaro@unito.it (L. Di Caro); amon.rapp@unito.it (A. Rapp)

0009-0002-2998-7753 (S. Villata); 0009-0007-4560-8043 (A. Bonarelli); 0000-0002-7570-637X (L. Di Caro); 0000-0003-3855-9961 (A. Rapp)

users' perceptions of humanness and amplify the well-documented human tendency to attribute minds, intentions, and emotions to non-human entities [15, 16]. Integrated into conversational systems such as ChatGPT and Copilot, these models provide highly accessible interfaces that have rapidly reached millions of users worldwide [17], making it especially likely that they will be perceived as human-like to varying degrees [18].

For this reason, it is essential to understand the extent to which current LLMs present themselves in human-like ways through their language. One systematic way to identify whether and when an LLM-based conversational agent adopts anthropomorphic communication is to ask LLMs themselves to analyze conversation transcripts and determine whether the agent presents itself in human-like ways. This approach is sensible because LLMs are highly capable of recognizing subtle linguistic and pragmatic cues at scale, making the analysis more consistent, reproducible, and significantly less labor-intensive than manual annotation. Moreover, it enables the evaluation of large conversational datasets that would otherwise be impractical to assess by human coders alone. In this paper, we precisely compare human annotations with LLM judgments of human-like self-presentation in LLM-generated responses, testing LLMs as evaluators with and without explicit guidance. This approach provides a systematic way to assess how frequently and in what forms LLMs display linguistic features that may strengthen perceptions of humanness and to what extent such features can be automatically detected.

2. Method

We drew our data from ShareChat [19], a large-scale corpus of in-the-wild conversations collected from publicly shared URLs across five widely used chatbot platforms: ChatGPT, Claude, Gemini, Grok and Perplexity. Its naturalistic interactions are particularly suited to study human-like linguistic behavior under real conditions. Each turn is labeled with a topic (e.g., *greetings_and_chitchat*, *mathematical_calculation*).

From this corpus, we randomly sampled 200 turns per platform (1,000 turns in total) and annotated each one for the presence of expressions that contribute to anthropomorphism (1/0), relying on the taxonomy proposed by [20], which identifies 19 categories of linguistic expressions contributing to the anthropomorphism of language technologies (e.g., expressions of intelligence, emotions, embodiment, vulnerability). Three human annotators independently labeled the dataset. Inter-annotator agreement (IAA) (Fleiss's κ [21]) was substantial overall (κ = .79) and computed per platform (Table 1). A majority vote label was used as the gold standard for the rest of the analysis. The percentage of turns judged as exhibiting human-like self-presentation by majority vote is reported in Table 1 and reveals an almost perfectly balanced amount of human-like turns (50.6% overall).

To assess the feasibility of an automatic classification, we replicated the annotation with two LLM judges, GPT-4o and Qwen2.5-7b, under two conditions: (*i*) guided, in which the model received the same annotation guidelines derived from [20] given to human annotators; (*ii*) blind, in which the model was asked to judge the presence of anthropomorphism without any reference to the taxonomy. Using two different architectures lets us check whether results generalize across models rather than reflecting one model's idiosyncrasies. Moreover, using LLMs as evaluators under these two conditions allows us to assess not only how well an LLM can replicate human judgments when explicitly instructed with a theoretically grounded coding scheme [22], but also the extent to which anthropomorphic communication is already implicitly recognizable to an LLM without any prior guidance [23]. We evaluated both models against the human majority-vote labels as annotators, computing Cohen's κ [24], and as classifiers, computing accuracy, precision, recall, and F1-score, reported in Table 1. Agreement with the human majority vote was substantial for both models in the guided condition (GPT-4o: κ = .78, accuracy = 89.1%; Qwen: κ = .75, accuracy = 87.4%) and slightly lower without guidelines (GPT-4o: κ = .73, accuracy = 86.7%; Qwen: κ = .73, accuracy = 86.5%).[1]

[1]Data and prompts available at: https://bit.ly/4p7ZlMY

Table 1
(a) Human IAA and evaluator agreement (κ), and % human-like turns, overall and by platform. (b) Classification performance of GPT-4o and Qwen against the human majority-vote gold standard.

	Turns	Human IAA (κ)	GPT-4o (κ)		Qwen2.5-7b (κ)		Human-like (%)
			guided	blind	guided	blind	
Overall	1,000	.79	.78	.73	.75	.73	50.6
ChatGPT	200	.75	.70	.67	.61	.62	62.0
Claude	200	.75	.51	.55	.58	.56	75.9
Gemini	200	.69	.71	.60	.66	.66	46.5
Grok	200	.72	.84	.71	.78	.61	60.5
Perplexity	200	.86	.88	.88	.94	.88	8.5

		Accuracy	Precision	Recall	F1
GPT-4o	Guided	89.1%	.87	.93	.90
	Blind	86.7%	.83	.93	.88
Qwen	Guided	87.4%	.88	.87	.88
	Blind	86.5%	.80	.97	.88

3. Qualitative patterns in LLMs' human-like behaviors

3.1. Platform-level linguistic signatures

Comparing anthropomorphic language across the five platforms reveals not only differences in frequency but also in the forms it takes. Claude produced the highest rate of anthropomorphic turns among all five chatbots, often through introspective statements about its own cognitive or experiential states (e.g., *"I think consciousness definitely exists because I experience it directly — it's the one thing I can be certain of."*). Grok, by contrast, adopts an anthropomorphized language primary through stylistic register: it produced the highest rate of exclamation marks of any examined chatbot and was the only one where casual interjections such as *"Haha"* appeared regularly. Gemini's anthropomorphic turns were dominated by explicit cognitive statements about itself (*"I understand"*, *"I realize"*), often paired with an explicit denial of humanness that nonetheless preserved anthropomorphic cues (*"As an AI language model, I cannot feel emotions like sincerity. However, I can assure you that I am committed to..."*). ChatGPT's anthropomorphic turns were instead characterized by both self-assessment statements and politeness formulas aimed at establishing a "relationship" with the user. Perplexity almost never adopted anthropomorphic language. When it did, this was limited to epistemic hedging rather than relational or identity-related language.

These results support a distinction between "deep" anthropomorphism, grounded in declarative anthropomorphic markers, like claims about models' internal states, or relational statements (e.g., greetings), and "surface" anthropomorphism, grounded in paralinguistic markers, like register and tone.

3.2. LLM judges: sensitivities and blind spots

When the two LLM judges disagree with human majority vote we can observe that LLM judges are far more sensitive to declarative anthropomorphic markers (i.e., deep anthropomorphism) than to paralinguistic ones (i.e., surface anthropomorphism). Among the 48 turns that both judges over-flagged relative to the human label, only 29% contained a stylistic or paralinguistic maker (exclamation marks, casual interjections such as *"sure"* or *"perfect"*); most were instead conversational self-presentation formulas such as *"I'm here to assist you"*. Conversely, among the 19 turns that both judges under-flagged, 58% contained such stylistic markers, e.g., *"Perfect! Let's focus on shaping the personas."*. Human annotators, instead, show the opposite tendency. This asymmetry clarifies what guidelines do not fix: providing explicit instructions substantially reduced over-flagging, but left the under-detection of stylistic anthropomorphism essentially unchanged: guidelines sharpen precision by helping LLM

judges recognize explicit anthropomorphic expressions, rather than expanding recall by enabling them to identify more implicit cues, such as paralinguistic ones.

4. Discussion and Open Questions

Our results show that LLMs can not only produce anthropomorphic language, but also detect it with substantial reliability: a particularly relevant capability given that these models are increasingly being used to audit the very language they generate. However, this detection is uneven: LLM judges reliably flag declarative anthropomorphism (i.e., deep anthropomorphism), but consistently miss stylistic and paralinguistic cues (i.e., surface anthropomorphism) that human annotators still perceive as human-like. This means that an LLM-based auditing pipeline may appear highly accurate while systematically overlooking precisely these surface cues. Human annotators, by contrast, appeared relatively insensitive to explicit self-presentational formulas, while remaining responsive to subtler stylistic features such as conversational tone, hedging, and interactional framing.

These findings suggest that current LLM judges do not yet capture the full range of cues that humans use when perceiving anthropomorphic communication, particularly when such cues are conveyed through conversational style rather than explicit self-presentation. Whereas LLM judges appear to privilege explicit semantic markers that can be readily verbalized and categorized, human judgments may integrate a broader range of pragmatic and interactional cues whose contribution is difficult to formalize. Rather than treating these discrepancies as annotation errors, future research should investigate whether they reflect different operationalizations of anthropomorphism.

More broadly, our findings raise several fundamental research questions. Is it even appropriate to ask whether a response is simply "anthropomorphic or not," or does such a binary framing misrepresent a phenomenon that may be inherently graded and observer-dependent? If anthropomorphism is ultimately defined by human perception, to what extent can an LLM serve as a valid evaluator of anthropomorphic language, even when its judgments are internally consistent? Should automated judges be expected to reproduce human perception, theoretical definitions, or some intermediate operationalization? Finally, if the cues that most strongly shape perceptions of humanness lie not only in what is said but also in how it is said, to what extent can de-anthropomorphization be achieved through content-level interventions alone?

5. Conclusion and Future Works

We provided preliminary evidence that LLM judges detect declarative anthropomorphism reliably, but remain blind to subtler stylistic cues. Future work should: (*i*) move to graded and multi-label annotation distinguishing deep and surface anthropomorphism; (*ii*) extend the analysis to different and topic-specific corpora; and (*iii*) validate whether the cues missed by LLM judges are also those that give rise to the most consequential effects of perceived anthropomorphism, such as increased user trust.

Acknowledgments

This study has been funded under the project 'HUM-LLM: Humanization, Understanding, and Ethics in LLM-Based Health Interactions' by Grant for Internationalization – Programmazione Triennale 24-26 – UniTo Starting Grants.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly and GPT-5 in order to: Grammar and spelling check, Paraphrasing and rewording. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Rapp, A. Boldi, L. Curti, A. Perrucci, R. Simeoni, How do people ascribe humanness to chatbots? an analysis of real-world human-agent interactions and a theoretical model of humanness, International Journal of Human–Computer Interaction 40 (2024) 6027–6050. URL: https://doi.org/10.1080/10447318.2023.2247596. doi:10.1080/10447318.2023.2247596. arXiv:https://doi.org/10.1080/10447318.2023.2247596.

[2] A. Rapp, L. Curti, A. Boldi, The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots, International Journal of Human-Computer Studies 151 (2021) 102630. URL: https://www.sciencedirect.com/science/article/pii/S1071581921000483. doi:https://doi.org/10.1016/j.ijhcs.2021.102630.

[3] D. Shin, The perception of humanness in conversational journalism: An algorithmic information-processing perspective, New Media & Society 24 (2022) 2680–2704.

[4] A. Følstad, C. B. Nordheim, C. A. Bjørkli, What makes users trust a chatbot for customer service? an exploratory interview study, in: International conference on internet science, Springer, 2018, pp. 194–208.

[5] M. Skjuve, I. M. Haugstveit, A. Følstad, P. Brandtzaeg, Help! is my chatbot falling into the uncanny valley? an empirical study of user experience in human–chatbot interaction, Human technology 15 (2019) 30–54.

[6] W. Shi, X. Wang, Y. J. Oh, J. Zhang, S. Sahay, Z. Yu, Effects of persuasive dialogues: testing bot identities and inquiry strategies, in: Proceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–13.

[7] A. Rapp, C. Di Lodovico, L. Di Caro, How do people react to chatgpt's unpredictable behavior? anthropomorphism, uncanniness, and fear of ai: A qualitative study on individuals' perceptions and understandings of llms' nonsensical hallucinations, International Journal of Human-Computer Studies 198 (2025) 103471. URL: https://www.sciencedirect.com/science/article/pii/S107158192500028X. doi:https://doi.org/10.1016/j.ijhcs.2025.103471.

[8] L. Ciechanowski, A. Przegalinska, K. Wegner, The necessity of new paradigms in measuring human-chatbot interaction, in: International Conference on Applied Human Factors and Ergonomics, Springer, 2017, pp. 205–214.

[9] B. Liu, S. S. Sundar, Should machines express sympathy and empathy? experiments with a health advice chatbot, Cyberpsychology, Behavior, and Social Networking 21 (2018) 625–636.

[10] V. Ta, C. Griffith, C. Boatfield, X. Wang, M. Civitello, H. Bader, E. DeCero, A. Loggarakis, User experiences of social support from companion chatbots in everyday contexts: thematic analysis, Journal of medical Internet research 22 (2020) e16235.

[11] G. Abercrombie, A. Cercas Curry, T. Dinkar, V. Rieser, Z. Talat, Mirages. on anthropomorphism in dialogue systems, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 4776–4790. URL: https://aclanthology.org/2023.emnlp-main.290/. doi:10.18653/v1/2023.emnlp-main.290.

[12] T. Araujo, Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions, Computers in Human Behavior 85 (2018) 183–189. URL: https://www.sciencedirect.com/science/article/pii/S0747563218301560. doi:https://doi.org/10.1016/j.chb.2018.03.051.

[13] K. Morrissey, J. Kirakowski, 'realness' in chatbots: Establishing quantifiable criteria, in: M. Kurosu (Ed.), Human-Computer Interaction. Interaction Modalities and Techniques, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 87–96.

[14] A. Rapp, C. D. Lodovico, L. D. Caro, What do people see in large language models' social behavior? exploring individuals' reactions to llm-llm interactions and their impact on technology perceptions, International Journal of Human–Computer Interaction 0 (2026) 1–22. URL: https://doi.org/10.1080/10447318.2026.2632156. doi:10.1080/10447318.2026.2632156. arXiv:https://doi.org/10.1080/10447318.2026.2632156.

[15] N. Haslam, Dehumanization: An integrative review, Personality and social psychology review 10 (2006) 252–264.

[16] B. Reeves, C. Nass, The media equation: How people treat computers, television, and new media like real people, Cambridge, UK 10 (1996) 19–36.

[17] Reuters, ChatGPT app hits 1 billion monthly active users in record time, data shows, 2026. URL: https://www.reuters.com/technology/chatgpt-app-hits-1-billion-monthly-active-users-record-time-data-shows-2026-06-02/, accessed: 2026-07-07.

[18] A. Rapp, C. Di Lodovico, L. Di Caro, How do people react to chatgpt's unpredictable behavior? anthropomorphism, uncanniness, and fear of ai: A qualitative study on individuals' perceptions and understandings of llms' nonsensical hallucinations, International Journal of Human-Computer Studies 198 (2025) 103471. URL: https://www.sciencedirect.com/science/article/pii/S107158192500028X. doi:https://doi.org/10.1016/j.ijhcs.2025.103471.

[19] Y. Yan, T. Nguyen, B. Su, M. Lieffers, T. Le, Sharechat: A dataset of chatbot conversations in the wild, 2026. URL: https://arxiv.org/abs/2512.17843. arXiv:2512.17843.

[20] A. DeVrio, M. Cheng, L. Egede, A. Olteanu, S. L. Blodgett, A taxonomy of linguistic expressions that contribute to anthropomorphism of language technologies, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3706598.3714038. doi:10.1145/3706598.3714038.

[21] J. L. Fleiss, Measuring nominal scale agreement among many raters., Psychological Bulletin 76 (1971) 378–382. doi:10.1037/h0031619.

[22] F. Gilardi, M. Alizadeh, M. Kubli, Chatgpt outperforms crowd workers for text-annotation tasks, Proceedings of the National Academy of Sciences 120 (2023) e2305016120. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2305016120. doi:10.1073/pnas.2305016120. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2305016120.

[23] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging llm-as-a-judge with mt-bench and chatbot arena, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[24] J. Cohen, Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit., Psychological Bulletin 70 (1968) 213–220.