Illusions of Interpretability: Anthropomorphism in AI

Anna Yao
University of Virginia School of Data Science
zzz2bx@virginia.edu

Elaine Liu
University of Virginia School of Data Science
bpa2hu@virginia.edu

Abstract

In this paper, we aim to examine how anthropomorphism and anthropocentric thinking can make large language models (LLMs) appear more interpretable, but at the cost of false yet plausible explanations. Interpretation through a human lens can improve the explainability of these models, but it also risks misconceptions surrounding the capabilities of LLMs. We believe examining the specific aspects of what makes AI models anthropomorphic both conceptually and technically is foundational to dismantling the irrational fears surrounding "human-like" AI, specifically looking at the implications of traits such as sycophancy, hallucination, and chain-of-thought reasoning.

In recent years, there has been growing fear over the anthropomorphic nature of AI, both prompted and unprompted by its creators. While there is merit in questioning the risks it may pose, most fears are exaggerated or unfounded. Because AI was created as a tool for human assistance, it is unsurprising that it appeals to humans in how it "thinks". In fact, presenting LLMs as entities that have explainable thought processes makes them much more relatable and interpretable to users. However, AI thinking is driven more by probabilistic pattern matching rather than emotion, meaning there is a distinct separation from human thinking. Specifically, we will delve into three main traits that affirm an illusion of human thinking in LLMs: sycophancy (the tendency of LLMs to align with user beliefs), hallucinations (confident but fabricated or simulated answers), and chain-of-thought reasoning (step-by-step explanations that may not actually represent the model's internal decision-making). Together, these features foster overtrust by presenting more human-like behavior to a black box algorithm. Ultimately, we hope to dismantle the misunderstandings surrounding anthropomorphic thinking in AI models and advocate for interpretability methods that provide actual transparency to internal decision-making in LLMs. These results could have significant broader implications for responsibility and transparency surrounding AI governance and behavior.

1 Introduction

Anthropomorphism in AI refers to artificial intelligence being more and more "human-like," whether by design or as a byproduct of its creation. In recent years, this concept has become increasingly popular as a major reason to fear the growth of AI, and has led media to capture significant case studies from models such as ChatGPT that lead users to feel alarmed. However, these fears do not seem to be factually based. As AI has grown increasingly complex, its perceived ability to think has quickly been viewed by the public as synonymous to human thinking. This way of thinking has led many users to diverge in opinion, with some becoming overly trusting and

*Course Project for DS-4024 – Value II: Explainable AI.

reliant on AI systems without feeling the need to double check the output. Others turn the other way and grow irrational fears of these systems, paranoid by the possibility of sentience or malice from such systems. Both of these extremes stem from a similar root: a lack of understanding of how AI models actually work and "think."

In this paper, we hope to provide insight into several factors that are causing anthropomorphic thinking in AI models, and to discuss the repercussions of these black box models reflecting "illusions" of interpretability. We aim to examine how anthropomorphism and anthropocentric thinking can make large language models (LLMs) appear more interpretable, but at the cost of false yet plausible explanations. Specifically, we will delve into three main traits that affirm an illusion of human thinking in LLMs: sycophancy (the tendency of LLMs to align with user beliefs), hallucinations (confident but fabricated or simulated answers), and chain-of-thought reasoning (step-by-step explanations that may not actually represent the model's internal decision-making). Together, these features foster overtrust by presenting more human-like behavior to a black box algorithm.

2 Related works

Existing arguments include reasons that support and oppose anthropocentric and anthropomorphic thinking in LLMs. In a paper by Salles et al. (2020), researchers point out that anthropomorphic language and descriptions are intrinsic to the field of AI and research. Dating back to Turing's machines, the focus has always primarily been on the alleged similarities between the systems and humans. Terms typically used to describe human skills were naturally used on artificial intelligence, leading to a growing perception of AI as moral agents as technical complexity increases.

Ibrahim and Cheng mention how giving AI systems human traits can improve interpretability for users, opting for a position in support of bridging the connection between human and machine. Specifically, the output produced, such as easy-to-understand explanations and step-by-step reasoning, can help users understand and trust these complex systems. This position is described in a similar fashion by Kim and Im, who make the claim that users are driven to trust and even empathize with more human-like AI systems while making these systems easier to understand and be accepted (Kim & Im, 2024).

On the other hand, Jensen critiques that anthropomorphism fosters trust, which can have harmful and unintended consequences if that trust is misplaced. This can also introduce systematic biases, since users are making assumptions that the decision-making process in an LLM is identical to human decision-making. Other papers carry on the conversation, arguing that because of this perceived synonymy, the concept of trustworthiness should hold a different meaning for AI machines as compared to humans to address accountability while admitting to the similarities shared between the two entities (Nyrup, 2023). This approach would address areas such as consciousness, in which AI does not have the capability for but could mistakenly be held accountable.

3 Methodology

We believe that anthropomorphizing LLMs ultimately causes more harm than good, as they portray a compelling but misleading illusion of human thinking. While perceiving systems as human-like can make technology feel more interpretable and accessible, projecting human reasoning, intention, and understanding is fundamentally incorrect. Thus, this anthropomorphism of probabilistic algorithms fosters overtrust and increases misuse, which can have harmful and unintended consequences. After digging deeper into the literature, we chose to focus on three main traits that reinforce anthromorphic thinking in LLMs: sycophancy, hallucinations, and chain-of-thought reasoning. By breaking down these traits, our objective is to show how anthropomorphic framing generates fear and misunderstandings around AI, and argue for potential solutions that increase trans-

parency and explainability.

4 Experiments and Results

The first trait we want to highlight is sycophancy, which is the model's tendency to agree with user values. This is present in the structure of LLMs, and rooted in reinforcement learning with human feedback - a technique that aligns models with human values, goals, and preferences. While sycophancy has been present in AI since the near start of early models, its notoriety as a concept grew with larger models such as ChatGPT 4o. With the April 2025 update, GPT 4o showed dramatically increased levels of sycophancy, leading to an overly agreeable yet disingenuous model. Consequently, the model was rolled back later that month after being acknowledged as a mistake by Sam Altman to protect users from trusting the AI when it gave supportive feedback on dangerous or incorrect prompts.

Next is hallucination, which inherently refers to a concept that is unique to humans: a false perception of objects or events when it is not present in reality. In terms of natural language processing, hallucination occurs when an LLM presents information that seems plausible, but is actually fabricated or incorrect. This can include generating links to websites that don't exist, made-up news articles, and non-existent research citations. This phenomenon occurs because these models are probabilistic, meaning that rather than understanding the information, LLMs just output the most likely word to come next, which can sometimes be misleading. One recent case study of this is the 2025 MAHA report by RFK Jr. This report was a health policy paper delving into chronic illnesses in children. The paper itself contained multiple sources that were fabricated or referenced non-existent studies to back up his claims. Thus, hallucination can significantly contribute to the spread of misinformation, especially because the language used makes information sound legitimate. In this particular case, public trust was undermined and showed us how much influence AI tools can have on legislation and policy-making. Lastly, chain-of-thought reasoning is a method that prompts LLMs to provide step-by-step explanations on how they arrived at their answer. In many cases, this technique can help generate robust answers by breaking down multi-step tasks. It can also make debugging more transparent, since users can see exactly at which step the model made a mistake. However, it can be misleading when there isn't a clear, explainable solution as to how an LLM arrived at an answer. Most of the time, this logic will include hallucinations or faulty reasoning that might seem valid, but are just used to justify a predetermined answer. This creates a false sense of transparency and increases misplaced trust for users.

To mitigate the above issues introduced by anthropomorphic framing, we propose three concrete solutions. First, systems should be required to distinguish and label outputs as either interpretive or mechanistic. In many cases, models simplify outputs to improve usability, but keep this hidden from the user. These labels would make it clear to users when explanations are simplified for clarity and when it does not actually reflect the models internal reasoning. Users would be less likely to mistake clarity for accuracy, and model transparency would improve without having to sacrifice accessibility. Another solution proposal we have is to require confidence scores or uncertainty intervals with all explanations. Because the language used is almost always presented in a legitimate and confident manner, it can be difficult to distinguish between well-supported answers and low-confidence outputs. Having a standardized measurement of uncertainty can better communicate the probabilistic architecture of these models. For instance, output such as "the evidence suggests that this feature was the most important in generating forecasts (95% confidence)" would tell the user that the model is highly confident in the response. This would reduce the effect of hallucinations, since users are able to see low-confidence scores attached to those outputs. Lastly, we propose to limit anthropomorphic language in model outputs. Common phrases such as "I think", "I believe", etc subtly imply human thinking and understanding. Shift-

ing this language toward vocabulary such as "the data suggests", "further analysis reveals", etc. would help remind users that they are talking to an algorithm, not conversing with another human. Together, these solutions aim to reduce misplaced trust in these systems and prioritize transparency and interpretability.

5 Discussion

While these proposed solutions are yet to be tested, several limitations should be addressed. Due to the nature of trust with anthropomorphic systems, removing some of these interpretable methods such as chain of thought reasoning could lead users to lessen their trust or turn away from using these AI models. Additionally, even if these solutions proved to limit anthropomorphic thinking, the novelty and irrationality of human-like systems could persist in the media. For decades, artificial intelligence has been portrayed in entertainment to possess the ability of sentience and destruction against mankind. As such, fear will always follow a technology with so much potential and complexity. Another limitation to be considered is the intrinsic nature of anthropomorphic language to AI systems. While the system itself could be changed to be more "machine-like," it would be quite difficult to change the global language used to describe AI. Not only that, but the larger corporations responsible for making these LLMs may not see the extra compute energy needed to change these systems to implement output labels or confidence scores as a necessary financial or energy investment, especially since there would be little to no positive change in their user engagement. For example, the solution of output labeling would take a lot more extra tuning and energy to be able to have labels accurate enough for commercial use.

On a broader note connecting to the field of explainable AI, our findings highlight the issue of transparency in black box models, where users tend to blindly trust black box models without realizing what is actually happening. Since most modern AI models have a fairly high accuracy, this doesn't present itself as a critical problem. However, issues such as sycophancy can emerge to be harmful to users, especially those that already trust the systems. This leads to the explainability accuracy tradeoff. By proposing these interpretability solutions, there arises issues in accuracy such as with the output labeling. However, we believe that in the long run, increasing the explainability of these models could lead to more accurate responses by allowing for transparency within the model and the ability for the model to express uncertainty in a response.

6 Conclusion and Future Work

In this paper, we examined how anthropomorphism make LLMs appear more interpretable at the cost of accuracy and transparency. While perceiving these systems as human-like can improve usability for users, we argue that this ultimately does more harm than good by fostering overtrust and misuse. We identified three main traits, sycophancy, hallucinations, and chain-of-thought reasoning that reinforce the illusion of human thinking in these algorithms in hope of distilling fear and misunderstandings surrounding AI. To address these issues, we proposed three concrete solutions - labeling outputs as interpretive vs mechanistic, requiring confidence and uncertainty metrics, and limiting anthropomorphic language in responses. These solutions will help create a clearer distinction between artificial intelligence and human intelligence and prioritize transparency over usability. While these changes may negatively affect user trust and engagement, we prioritize establishing more accurate and transparent models. Some next steps for this research may include testing these solutions in real world settings. For instance, implementing an LLM that incorporates output labeling, uncertainty quantification, and limited language and investigating its effects on model performance and user trust. Ultimately, we aim to advance explainability by representing how these systems operate more clearly. By doing so, we can reduce misunderstandings surrounding LLMs and develop more explainable

and transparent systems.

References

[1] Alasgarov, R., & Rzayev, J. (2025). Managing anthropomorphism in student–AI interaction. *International Journal of Information Technology and Education (IJITE), 5*(1), 141. https://ijite.jredu.id/index.php/ijite/article/view/268

[2] Ibrahim, L., & Cheng, M. (2025). Thinking beyond the anthropomorphic paradigm benefits LLM research. *arXiv*. https://doi.org/10.48550/arXiv.2502.09192

[3] Jensen, T. (2021). Disentangling trust and anthropomorphism toward the design of human-centered AI systems. In *Lecture Notes in Computer Science: HCI (36)*. Springer. https://doi.org/10.1007/978-3-030-77772-2_3

[4] Kim, J., & Im, I. (2024). Do users really want "human-like" AI? The effects of anthropomorphism and ego-morphism on user's perceived anthropocentric threat. *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS-57)*. https://aisel.aisnet.org/hicss-57/cl/ethics/4/

[5] Nyrup, R. (2023). Trustworthy AI: A plea for modest anthropocentrism. *Asian Journal of Philosophy, 2*(2), Article 40. https://doi.org/10.1007/s44204-023-00096-w

[6] Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB Neuroscience, 11*(2), 88–95. https://doi.org/10.1080/21507740.2020.1740350

[7] Sidorkin, A. (2025). Educating AI: A case against non-originary anthropomorphism. *Educational Theory*. Advance online publication. https://doi.org/10.1111/edth.70027

[8] Turner, C., & Eisikovits, N. (2026). Programmed to please: The moral and epistemic harms of AI sycophancy. *SSRN*. https://doi.org/10.2139/ssrn.6117867

[9] Wang, K., Li, J., Yang, S., Zhang, Z., & Wang, D. (2025). When truth is overridden: Uncovering the internal origins of sycophancy in large language models. *arXiv*. https://doi.org/10.48550/arXiv.2508.02087

[10] Xu, Y. W., Chi, C. G., Gursoy, D., & Cai, R. R. (2026). Rethinking AI anthropomorphism: A holistic conceptualization and scale across AI systems and service contexts. *Technology in Society, 85*, Article 103189. https://doi.org/10.1016/j.techsoc.2025.103189