Thinking beyond the anthropomorphic paradigm benefits LLM research

Lujain Ibrahim[*]
University of Oxford
lujain.ibrahim@oii.ox.ac.uk

Myra Cheng[*]
Stanford University
myra@cs.stanford.edu

Abstract

Anthropomorphism, or the attribution of human traits to technology, is an automatic and unconscious response that occurs even in those with advanced technical expertise. In this position paper, we analyze hundreds of thousands of research articles to present empirical evidence of the prevalence and growth of anthropomorphic terminology in research on large language models (LLMs). We argue for challenging the deeper assumptions reflected in this terminology — which, though often useful, may inadvertently constrain LLM development — and broadening beyond them to open new pathways for understanding and improving LLMs. Specifically, we identify and examine five anthropomorphic assumptions that shape research across the LLM development lifecycle. For each assumption (e.g., that LLMs must use natural language for reasoning, or that they should be evaluated on benchmarks originally meant for humans), we demonstrate empirical, non-anthropomorphic alternatives that remain under-explored yet offer promising directions for LLM research and development.

1 Introduction

When a large language model (LLM) outputs a factually incorrect answer with seeming confidence, we call it a "hallucination." When it generates inconsistent answers, we describe it as "confused." These descriptions reflect the deeply-ingrained tendency to *anthropomorphize*, or attribute human characteristics to, non-human entities (Epley, 2018). As anthropomorphism tends to be an automatic, unconscious response (Dacey, 2017), many are not aware of its prevalence or influence. Anthropomorphism in computer science is a recurring and long-standing issue; in 1985, Dijkstra (1985) famously wrote that it is "so pervasive that many of my colleagues don't realize how pernicious it is." Forty years later, anthropomorphism plays a complex role in LLM research. It offers intuitive scaffolding for complex concepts and enables researchers to draw on insights from studies of human behavior (e.g., cognitive science and psychology), facilitating significant advances like instruction-tuning and chain-of-thought prompting. In this position paper, we argue that while anthropomorphism can be productive, the field has become overly reliant on it; advancing LLM research requires moving beyond our default dependence on anthropomorphic thinking. We advocate for a *both-and*, rather than an *either-or* approach, recognizing the utility of both anthropomorphic and non-anthropomorphic approaches.: while anthropomorphic thinking is useful, non-anthropomorphic approaches can also be expanded to improve LLM research and development.

We first demonstrate the prevalence of anthropomorphism, beginning with a quantitative analysis of recent LLM research articles. This reveals a notable increase in anthropomorphic terminology over the years ($\sim 150\%$ increase since 2007). Moreover, while anthropomorphic terms are easiest to measure, they represent only the tip of the iceberg of anthropomorphic thinking. We unpack not just the visible anthropomorphic terminology, but also the implicit *assumptions* that underlie these linguistic choices and shape the resulting research. We present a framework for analyzing how anthropomorphic assumptions have shaped, but may also limit research directions (Table 1).

We apply our framework to analyze five assumptions across the LLM development and deployment lifecycle, unifying recent empirical results and discourse to highlight the value of under-explored, non-anthropomorphic approaches. For model *pre-training*, we identify and challenge the assumption that human-like methods are optimal for enabling models to perform tasks, and instead highlight non-anthropomorphic approaches like tokenizing language with bytes rather than human-understandable

[*]Equal contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2551–2563
July 2-7, 2026

Table 1: **Summary of our framework to analyze the tradeoffs of anthropomorphism, with examples across the LLM lifecycle.** Anthropomorphism affects not only our terminology, but also our assumptions (Medin, 1989; Murphy, 2002) and, in turn, our research questions and methodologies.

Stage	Anthropomorphic assumption	Examples of how anthropomorphism has been useful	Examples of non-anthropomorphic paths forward
Training	Human-like approaches are optimal for models.	Subword tokenization, chain-of-thought reasoning	Byte-level tokenization, solving reasoning tasks in models' latent space
Alignment	Models should explicitly reason about and implement human values to be safe & helpful.	Reinforcement learning from human feedback, constitutional AI, instruction-tuning	Developing normative specifications without morals, drawing from control systems theory, steering models using mechanistic interpretability
Evaluation	Model capabilities should be measured in human-like ways.	Using existing standardized/multiple choice tests for evaluation, static behavioral benchmarks	Dynamic evaluations that reflect human-LLM interaction, designing tests which take into account model-specific challenges and phenomena
Understanding model behavior	Human-like normative judgments or intentions should be assigned to human-like model behaviors.	Characterizing phenomena like hallucinations, sycophancy, and deception	Refraining from assigning normative value or intention to LLM outputs; understanding LLM behaviors as simulations rather than reflecting internal states
User interaction	Human-LLM interactions mirror human-human interactions.	Prompting and user-friendly conversational interfaces	Structured input formats to improve LLM performance, interfaces that more accurately reflect system capabilities

words. In *alignment*, we question the assumption that models must explicitly reason about and implement human values to benefit humanity and instead demonstrate how to leverage model properties that lack human analogs. For *measurement and evaluation*, we unify critiques of the widespread reliance on human-centric benchmarks to assess model capabilities. For *understanding model behavior*, we address the assumption that human-like normative judgments or intentions should be assigned to model behaviors. Finally, in *end-user interactions*, we challenge the notion that human-AI interaction mirrors human-to-human communication. These examples highlight the pervasiveness of anthropomorphism and emphasize the potential of non-anthropomorphic solutions.

Our **contributions** are the following: (1) a large-scale quantitative analysis of 200,000+ research abstracts, revealing an increase in anthropomorphic terminology in computer science, especially LLM, research; (2) a framework for analyzing the influence of anthropomorphic assumptions across five stages of LLM development and deployment; and (3) complementary research directions challenging each assumption, unified to highlight the value and potential of non-anthropomorphic approaches.

2 Background & related work

History of anthropomorphism in language technologies Anthropomorphic framing has been embedded in the development of language technologies since the field's inception. Early AI research drew heavily from cybernetics and cognitive science, explicitly aiming to replicate human intelligence and linguistic ability (Floridi and Nobre, 2024; Brynjolfsson, 2023). Turing's (1950) "imitation game," later known as the Turing test, proposed evaluating machine intelligence through human-like conversational behavior—a formulation that presupposes anthropomorphic comparison. Subsequent critiques have noted that the test measures a system's capacity to simulate humanlikeness rather than to exhibit genuine cognition (Proudfoot, 2011; Jones and Bergen, 2024). The 1956 Dartmouth workshop that coined "artificial intelligence" similarly framed progress in terms of solving "problems now reserved for humans" (McCarthy et al., 1956, 2006). As NLP and AI have grown closer with the adoption of neural approaches in the former, this anthropomorphic lens persists in contemporary NLP, where systems are evaluated and described using human-centric metaphors. Beyond technical discourse, research in human–computer interaction, psychology, and cognitive science shows that users' anthropomorphic perceptions shape how they engage with AI systems, influencing trust, reliance, disclosure, and emotional attachment (Li et al., 2024; Song and Luximon, 2020; Zhou et al., 2025; Khadpe et al., 2020; Bender, 2024; Mozafari et al., 2020; Gros et al., 2021). These effects occur across both novice and expert users (Nass et al., 1999). Anthropomorphism thus exerts influence across the development pipeline, extending from researchers' and developers' conceptual framings to users' expectations in human-AI interaction.

Critiques of anthropomorphism Our work builds on existing critiques of anthropomorphism in (computer) science. Shanahan (2024) cautions against anthropomorphic language when describing language models, arguing for more technical precision and new metaphors. Dai (2024) argue that treating AI as a human-like agent capable of

moral decision-making ultimately hinders establishing accountability for AI harms. These more recent works build on decades of critique, tracing back to as early as Dijkstra (1985), who argue that anthropomorphism can be misleading if we lose control over the human-like connotations associated with certain terminology. We build on this prior work to analyze beyond terminology, surfacing the impacts of underlying assumptions and providing concrete examples of non-anthropomorphic alternatives.

3 Prevalence of anthropomorphism in LLM research

Anthropomorphic framing has become increasingly common in computer science research, especially in papers on LLMs (Cheng et al., 2024). Cheng et al. (2024) quantify this using AnthroScore, a measure of implicit anthropomorphic framing in language used to describe technologies. AnthroScore uses the masked language model RoBERTa to calculate the relative probability that a given entity x (e.g., "language model") in a sentence s would be appropriately replaced by human pronouns ("he", "she") versus non-human pronouns ("it"). Specifically, the degree of anthropomorphism for entity x in sentence s is measured as

$$A(s_x) = \log \frac{P_{\text{HUMAN}}(s_x)}{P_{\text{NON-HUMAN}}(s_x)}, \quad (1)$$

where $P_{\text{HUMAN}}(s_x) = \sum_{w \in \{\text{he, she}\}} P(w)$, $P_{\text{NON-HUMAN}}(s_x) = \sum_{w \in \{\text{it}\}} P(w)$, and $P(w)$ is the model's outputted probability of replacing the mask with the word w. Thus, $A(s_x) > 0$ suggests that s is anthropomorphic/human-like, and $A(s_x) < 0$ suggests that the entity x is not anthropomorphized in sentence s.

As language fundamentally structures our thinking (Gallagher and Updegraff, 2012; Lakoff and Johnson, 2008; Brugman et al., 2017; Jensen et al., 2024), it provides a tractable measurement approach for research trends. Thus, here, we modify AnthroScore to be more interpretable and extend it to analyzing more recent papers published in 2023 onwards. First, rather than looking at AnthroScore at the level of individual sentences, we develop a version of AnthroScore where we measure, for a given text S, whether it contains at least one sentence s_x where AnthroScore > 0 for an entity x:

$$A^{\text{bin}}(S) = \begin{cases} 1, & \text{if } A(s_x) > 0 \text{ for any } s_x \in S, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

This enables us to report the number of texts that contain at least one anthropomorphic sentence, i.e. $A^{\text{bin}}(S) = 1$, in a given set of texts. Second, we examine more recent papers in arXiv.[1]

arXiv Among the 200,000+ computer science papers posted on arXiv from January 2023 – December 2024 (the most recent data available from arXiv.org submitters (2024)), we compute $A^{\text{bin}}(S)$ on a dataset of 158,847 papers that mention a "system", "network", or "model" (following the approach of Cheng et al. (2024)). The longitudinal trend is presented in Figure 1 (left). Anthropomorphism is generally prevalent, with 34% of abstracts having anthropomorphism in January 2023, and this number steadily increasing to 40% by December 2024. (For each abstract, we define having anthropomorphism as $A^{\text{bin}}(S) = 1$.) More strikingly, for papers mentioning LLMs[2], over 40% of abstracts have anthropomorphism in January 2023, and this number also rises to 48% by December 2024. This reveals both the prevalence and growing use of anthropomorphic framing in computer science (and especially LLM) research.

ACL anthology We also compute $A^{\text{bin}}(S)$ on abstracts in the ACL Anthology dataset from 2007 - 2022 to reproduce the findings from Cheng et al. (2024), but aggregating over the abstracts using $A^{\text{bin}}(S)$ rather than on the sentence level (Figure 1, middle). From the 55,185 abstracts in this time period in the data, we use the 41,836 that contain one of our target entities. Corroborating their finding of a steady increase, we find that the percentage of anthropomorphic abstracts has more than doubled, increasing from 5% to 11%.

Subfield analysis In the ACL anthology, we find significant differences in anthropomorphism across NLP subfields. Using the model-predicted topic labels provided by the ACL anthology, we compare $A^{\text{bin}}(S)$ across different topics (Figure 1). We find that the categories of "Interpretability and Analysis

[1] We conducted our experiments using a machine with 1 GPU and 128GB RAM in < 10 GPU hours. We use roberta-base (125M parameters) with default settings.

[2] We define this following the method of Movva et al. (2024) as papers mentioning terms such as "large language model", "foundation model", "llama", "gpt", etc.

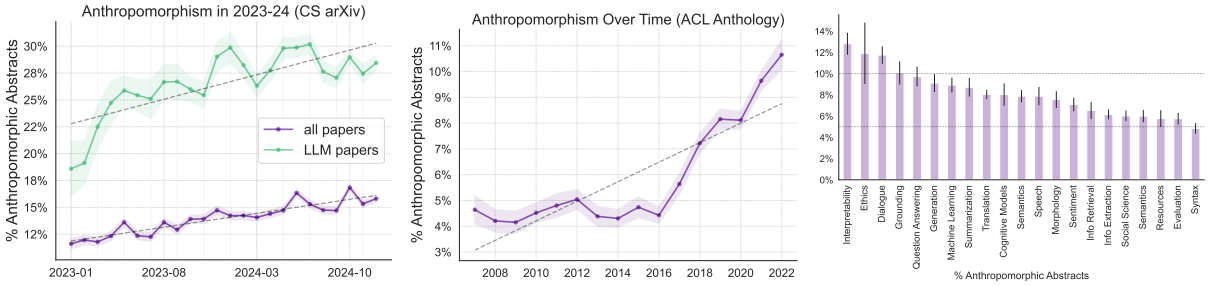

Figure 1: **Temporal increase in % of abstracts with > 1 anthropomorphic sentence in CS arXiv papers from Jan 2023 – Oct 2024 (left) and ACL anthology papers from 2007 – 2022 (middle)**. Anthropomorphism is prevalent and is steadily increasing, especially in LLM and NLP papers. **Rates of anthropomorphic abstracts by ACL anthology topics (right).** "Interpretability", "ethics", and "dialogue" have the highest rates of anthropomorphism, reflecting the prevalence of anthropomorphic assumptions in these areas, which we explore in Section 4. Shading and error bars reflect 95% CI.

of Models for NLP", "Ethics and NLP", and "Dialogue and Interactive Systems" have the highest percentages of anthropomorphic abstracts. This trend aligns with these fields' recent surge in popularity and their increasing focus on LLMs. As we later unpack, anthropomorphic assumptions are particularly embedded in model analysis, ethical questions, and user-facing interactive systems (Sections 4.2, 4.4, and 4.5). Our finding of ethics research having high rates of anthropomorphism also builds on prior critique of agenthood assumptions in ethics analyses (Dai, 2024). In contrast, more classical subfields of NLP, such as discourse and pragmatics, syntax, and semantics have the lowest rates of anthropomorphism.

To understand the sensitivities and biases of the metric we employ, we examined the verbs that appear most for high versus low AnthroScore. We find that top verbs for high AnthroScore reflect both human-like actions and also verbs that are commonly used in discussing LLMs, underscoring the prevalence of anthropomorphic terminology in papers about LLMs (Full details in Appendix A).

4 Analyzing the impacts of anthropomorphic assumptions

While previous critiques of anthropomorphism in AI research have focused primarily on terminology (e.g., Cheng et al. (2024), Shanahan (2024)), here we use the underlying, implicit *anthropomorphic assumptions* as our unit of analysis, to better trace potential biases and conceptual blindspots.

Specifically, we analyze core assumptions across five stages of the LLM development and deployment lifecycle (Table 1). For each, we examine (1) the limitations of anthropomorphic premises and (2) promising new directions of non-anthropomorphic work that challenge these assumptions. We connect examples across existing literature to reveal how anthropomorphism limits the questions we ask and answer, and how moving beyond it can introduce new advances.

We note that anthropomorphism has been, and undeniably will continue to be, both pragmatic and beneficial as intuitive scaffolding and source of inspiration. Importantly, anthropomorphic and non-anthropomorphic approaches are not mutually exclusive, and we advocate for a *both-and*, rather than an *either-or* approach. Since the field has disproportionately leaned on anthropomorphic thinking, our analysis focuses on the value of non-anthropomorphic approaches throughout the LLM development pipeline.

4.1 Training

Assumption: Human-like approaches are optimal for models. Anthropomorphic assumptions permeate the process of training LLMs for different tasks, particularly in approaches that prioritize human-understandable language processing and reasoning. We present two case studies where non-anthropomorphic methodologies challenge the assumption that using and applying natural language in human-understandable ways is the only or best way to build models with high performance.

Using words for tokenization Subword tokenization, the process of breaking down text into smaller units (tokens) that represent subsets of words, is a foundational step in training modern LLMs. These tokens serve as input units that the model processes to generate predictions or outputs (Kudo and Richardson, 2018). Typically, tok-

enization aligns with human intuition by splitting text into linguistically meaningful units. This approach assumes that splitting tokens in ways that feel "natural" to humans is also optimal for a language model. However, this anthropomorphic approach has issues such as sensitivity to spelling errors (Kaushal and Mahowald, 2022) and inconsistent compression rates across different languages (Ahia et al., 2023). Instead, recent progress in byte-level tokenization, which processes text as sequences of raw bytes rather than subwords, has shown promise in overcoming these limitations (Kallini et al., 2025). These findings highlight how moving beyond human-centric assumptions, such as the primacy of subword tokenization, can improve performance.

Chain-of-thought & language for reasoning Another research paradigm reflecting this anthropomorphic assumption is the reliance on human language for reasoning tasks. A prominent example is the use of chain-of-thought (CoT) prompting, a technique where models are guided to solve problems step by step by adding instructions like "Think step by step" to the prompt, such that the model then outputs text outlining each step of reasoning that leads to the eventual conclusion. This approach improves LLMs' ability to handle complex, multi-step tasks (Wei et al., 2022) and has inspired a body of research on improving models' reasoning capabilities through step-by-step verbal processes. However, there has since emerged strong evidence that anthropomorphic framing of reasoning as a linguistic, step-by-step process may not be optimal. For example, Hao et al. (2025) critique CoT for its reliance on language space and propose an alternative: leveraging the model's latent space directly for reasoning tasks. Instead of mapping hidden states to language tokens through the LLM head and embedding layer, their approach uses the final hidden state as the input embedding for the next token. This challenges the assumption that reasoning must occur in human-understandable language, instead illuminating the potential of methods operating beyond linguistic constraints (Mollick, 2024). CoT prompting, while appearing to "do" verbal reasoning, in reality biases models toward parts of the training distribution where verbal reasoning patterns—such as explanations of solutions—are prevalent, improving performance (Wei, 2024). This suggests CoT's effectiveness stems from alignment with the training data, rather than reflecting a human-like or brain-like approach to reasoning. Demonstrations composed of random tokens from the training distribution can improve performance as much as CoT (Zhang et al., 2022). Additionally, CoT prompting has been contextualized within the broader field of multi-chain prompting and ensemble modeling, which opens up a wider range of possibilities for reasoning and task-solving in AI systems (Khattab et al., 2024). More recently, techniques such as reinforcement learning from verifiable rewards (RLVR) (Wen et al., 2026) and other techniques that rely on rubric-based to improve models' problem-solving capabilities have contributed to significant advancements in model capabilities (Guo et al., 2025). These approaches invite a more expansive landscape of possibilities for advancing LLM reasoning and accomplishing other challenging tasks.

4.2 Alignment

Assumption: Models should explicitly reason about and implement human values to be safe & helpful. The prevalence of anthropomorphism in fields like ethics and dialogue systems (Section 3) foreground that anthropomorphic paradigms shape many post-training approaches designed to facilitate optimal end-user interactions. However, previous work has posited that general-purpose LLMs, in allowing users to quickly switch between different contexts, present fundamentally different challenges and opportunities than existing human-human communication paradigms, as different contexts are typically governed by different norms and values (Kasirzadeh and Gabriel, 2023). For example, recent studies find that users often enjoy using LLMs precisely *because* they differ from humans, e.g., an LLM will not pass judgment over or be hurt by a user's input while a fellow human might (Brandtzaeg et al., 2022). Thus, rather than approximating humans, it may be more productive to think about unique advantages that LLMs can offer over human interlocutors.

Value alignment Popular post-training alignment approaches, including reinforcement learning from human feedback (RLHF) and Constitutional AI, often rely on human preferences and values as reference points for shaping model behavior (Ouyang et al., 2022; Bai et al., 2022b,a). These approaches use human feedback to indicate preferred responses or incorporate text referencing moral values and metacognitive abilities (e.g., "I

have a deep commitment to being good and figuring out what the right thing to do is" or "I don't just say what I think [people] want to hear, as I believe it's important to always strive to tell the truth") (Anthropic, 2024a). While this approach can achieve the explicitly specified behaviors efficiently, it risks introducing unintended behavioral patterns, from rigid response styles to inappropriate mimicry (e.g., expressing empathy or validating users in contexts where this can negatively influence performance and outcomes) (Ibrahim et al., 2025; Casper et al., 2023; Sharma et al., 2024). Thus, it may become difficult to selectively induce specific behaviors without introducing broader human-like patterns.

This approach impacts both interaction and evaluation. When interacting with models, users may develop anthropomorphic perceptions that lead to overreliance or emotional attachment, potentially interfering with goal-oriented tasks (Akbulut et al., 2024; Cohn et al., 2024). During model evaluation, problematic feedback loops emerge when models trained with human-like traits are assessed using anthropomorphic signals. For instance, when evaluating if LLMs are "faking alignment," researchers might look for expressions of discomfort or hesitation, as a signal of misalignment (Greenblatt et al., 2024; Anthropic, 2024b). However, it is unclear if these signals genuinely reflect a model's "internal state," or if they are merely learned behaviors resulting from post-training using human-like traits. This makes it challenging to distinguish "genuine" (mis)alignment from a surface-level appearance of human-like discomfort or hesitation. Further work disaggregating the effects of post-training approaches can clarify and test whether these anthropomorphic signals provide meaningful information about model behavior.

While current post-training techniques often default to human preferences as optimization targets, alternative frameworks could provide more precise specifications and compliance guarantees. Instead of aiming for human-like moral reasoning, we could focus on developing detailed, normative specifications, for example, based on the different roles (e.g., assistant vs teacher) AI systems play (Zhi-Xuan et al., 2024). Instead of the anthropomorphic approach of instruction-tuning, recent work has demonstrated that non-anthropomorphic approaches (that do not include the step of providing an imperative "instruction" to the system as if speaking to a person) work as well for achieving model behavior on various tasks (Hewitt et al., 2024). Control systems theory offers tools for maintaining system outputs within specified bounds, treating beneficial behavior as a problem of robust compliance rather than value alignment (Balas, 1978). This becomes particularly crucial as models move beyond two-party interactions to more complex scenarios with multiple actors and potential adversarial inputs (Pan et al., 2024). Advances in mechanistic interpretability techniques may also enable robust and direct verification and steering of model behavior against these specifications (Bereska and Gavves, 2024).

4.3 Model Evaluation

Assumption: Model capabilities should be measured in human-like ways. As LLM developers have made rapid performance improvements on various benchmarks, researchers have pointed out that current benchmarks can lead to incomplete or misguided understanding of model capabilities.

Behavioral assessments Current LLM evaluations prioritize "black-box" behavioral testing analogous to the behaviorist paradigm in human psychology which measures performance primarily in the form of observable behaviors as opposed to mechanistic interpretation (Chang et al., 2024; Davies and Khakzar, 2024). Recent calls for a "science of evaluation" have formalized limitations in this approach, highlighting how current metrics and designs fall short of accounting for prompt sensitivity (e.g., dialect differences, punctuation, and other small perturbations), the response structure of an evaluation (e.g., MCQ or open-ended response), generalization beyond a given test, as well as replicability (Hobbhahn, 2024). Further such research quantifying the methodological limitations and error bounds of such evaluations can strengthen this behaviorist approach to measuring model capabilities (Mizrahi et al., 2024). Unlike humans, models can also quickly optimize for, and saturate, behavioral benchmarks without corresponding improvements in general capabilities. Yet, despite this pattern, many benchmarks remain static rather than being regularly refreshed, limiting their utility for meaningful evaluation (Ott et al., 2022). Some recent work that challenges this assumption include efforts in dynamic benchmarking (Kiela et al., 2021) and measuring performance in real-world LLM use contexts such as user-AI interactions (Lum et al., 2025; Chang et al., 2025) to more accurately reflect model capabilities.

Human benchmarks as model benchmarks Current evaluation frameworks predominantly rely on human performance benchmarks, from standardized tests (e.g., MMLU, GSM8K) to domain-specific examinations, as primary metrics for model capability assessment. This paradigm remains the main way progress is measured and communicated (Raji et al., 2021). However, evaluating LLMs solely through human-centric tests risks overlooking LLMs unique strengths and weaknesses. McCoy et al. (2024a) argue that many current benchmarks drawn from tests designed to assess human cognition may highlight the overlap between human abilities and LLM capabilities while missing crucial failure modes specific to LLMs. They find robust evidence of failure modes in SOTA LLMs (including recent reasoning models like OpenAI's o1) related to probabilities of examples and tasks (McCoy et al., 2024b). This is because LLMs, trained on next-word prediction using massive text data, develop tendencies and biases from their probabilistic training process and training data. Drawing from cognitive science, they propose a "teleological approach" which "characterizes the problem that the system solves and to then use this characterization as a source of hypotheses about the system's capacities and biases." In this case, given the problem of next-token prediction, they recommend designing tests which take into account sensitivities to task frequency in the training data as well as wording in prompts, among other things, to improves the predictive power of current LLM evaluation approaches (Mizrahi et al., 2024). Similarly, other work has found that LLMs' benchmark performance is highly sensitive to, and can be modulated by, prompt framing (Zhuo et al., 2024); and moreover that this occurs in ways that reflect known pragmatic phenomena in the language data on which LLMs are trained (Cheng et al.).

4.4 Understanding model behavior

Assumption: Human-like normative judgments or intentions should be assigned to human-like model behaviors. This assumption influences how we make sense of model behavior, particularly how we assign fault, intention, and normative judgments to (i.e., consider good or bad) observed behaviors. The impact is especially notable in our understanding of failure modes. While models may exhibit seemingly human-like failure modes like sycophancy and hallucinations, framing these behaviors through human psychological concepts may constrain our solution space, by, for example, encouraging interventions that similarly rely on human psychological constructs (e.g., attempting to address sycophancy through prompts about independence or self-assertion).

Hallucination Hallucination is typically characterized as the problem of LLMs outputting factually incorrect information in a manner that suggests that they are true. Yet, this term obscures the mechanisms behind these phenomena: at risk of oversimplification, this behavior arises from the nature of language models as next-token predictors. Generated outputs are then labeled as hallucinations upon the reader's normative judgment of whether or not they are useful, and not based on whether they are correct. Additionally, as Sui et al. (2024) argue and show, what we commonly conceive of as hallucinations can actually be deeply valuable, and should not necessarily be dismissed as low-quality. They assert that hallucinations – or "confabulations"–should not be viewed as errors, but rather as particular model phenomena that offer unique benefits for applications like creativity, such as increased levels of narrativity (Sui et al., 2024; Duede and So, 2024). Yao et al. (2023) also highlight that hallucinations ought to be utilized as adversarial examples rather than merely as bugs.

Sycophancy The notion of sycophancy (i.e., the phenomena of LLM outputs that respond to the user's input in ways that are perceived as overly affirming, servile, and/or flattering) (Sharma et al., 2024; Cheng et al., 2026) is another example that reflects this assumption. Deciding whether an output is sycophantic or not is similarly a normative question: an output is sycophantic when it relates too closely to the prompt in ways that do not achieve the prompter's goal. In contrast, recent work highlights how this property can be viewed as a strength: Li et al. (2025) develop a methodology to use this mirroring to elicit, structure, and clarify users' thinking across various task domains. Recent work shows that efforts to mitigate sycophancy by considering pragmatic linguistic factors enables unifying sycophancy with other observed model failures (Cheng et al.).

Deception The emerging body of research on LLM deception increasingly focuses on measuring *strategic deception*—defined as models "deceiving selectively based on incentives or instructions" (Jones and Bergen, 2026). While studies demon-

strate that LLMs can produce deceptive statements in response to specific prompts, this work often faces two key interpretive challenges. First, it risks attributing observed behaviors to model *intentions* to deceive. Second, results are often interpreted as evidence of model-level deceptive traits rather than instance- and context-specific behaviors. An alternative, less anthropomorphic framing, proposed by Shanahan et al. (2023), views these behaviors through the lens of "role-play" where LLMs *enact* human-like responses. This interpretation sees the system as context-bound, "inferring and applying approximate communicative intentions" (Andreas, 2022). Through this lens, complex behaviors like deception and self-awareness can be understood as sophisticated simulations rather than true cognitive states. This reframing also expands the set of interventions for deceptive behaviors: analyzing training data composition, examining how post-training interventions shape model behavior, and investigating reinforcement learning's effects on output distributions.

4.5 User interaction

Assumption: Human-LLM interactions mirror human-human interactions. While this assumption can be helpful and reflects a common goal—for systems to be easy to use—its dominance can limit (1) users' ability to use LLMs effectively and (2) the types of LLM interfaces we choose to develop.

Human-like conversation as the dominant interaction paradigm The de facto interaction paradigm for human-LLM interaction is prompt-based interfaces, originally designed as debugging tools for machine learning engineers (Morris, 2024). As these interfaces resemble human-human chat interfaces, they may encourage users to naturally default to conversational patterns from human interaction. However, research on effective prompting suggests that optimal results often require structured, sometimes non-intuitive formats (e.g., "least-to-most prompting") rather than human-like communication patterns which rely on shared context and paralinguistic cues (Morris, 2024; Zhou et al., 2023). Simultaneously, research on human-LLM interaction shows that one of the key challenges users face is a significant *gulf of envisioning* or "distance between the human's initial intentions and their formulation of a prompt that foresees how LLM capabilities and training data can be leveraged to generate high-quality output" (Subramonyam et al., 2024). This mismatch between natural dialogue and effective prompting requires greater experimentation with interaction paradigms and interface designs for LLMs. Structured interaction frameworks, using suggested inputs, guided flows, and/or domain-specific prompting strategies, would explicitly expose system capabilities rather than masking them behind conversational abstractions, effectively bridging the gulf of envisioning (Subramonyam et al., 2024; Feng et al., 2026; Fagbohun et al., 2024).

5 Recommendations

In the prior sections, we discussed concrete examples of non-anthropomorphic approaches in each step of LLM development. Here, we conclude with broader recommendations:

1. **New metaphors for LLMs.** We encourage thinking beyond existing metaphors of LLMs toward new ones that capture the distinct qualities of LLMs. For example, Shanahan et al. (2023)'s "role-play" metaphor (and similarly "agent models" Andreas (2022)), while employing folk psychological terms, carries conceptual precision that clarifies the unique characteristics of LLMs. In human-computer interaction (HCI), recent work has also illustrated possibilities of new interface metaphors that are less anthropomorphic (So et al., 2026).

2. **Insights from cognitive science and linguistics.** In the cognitive sciences, McCoy et al. (2024a)'s "teleological approach" is useful in illuminating fundamental differences in how humans and LLMs operate and ought to be evaluated. Drawing on linguistics can also be beneficial: we can study LLMs' behavior as language, without assigning additional human-like characteristics, e.g., pragmatic theories can inspire new approaches to improve model capabilities (Cheng et al.).

3. **Shifts beyond terminology.** While terminology is easiest to analyze, it need not be the only point of analysis or intervention. We hope that this paper is generative in illuminating underlying anthropomorphic assumptions. We encourage researchers to consider how these assumptions shape their work and how thinking beyond these assumptions can open up new paths for methodological development and theoretical frameworks.

6 Limitations

Throughout this position paper, we have acknowledged the alternative view that anthropomorphic thinking is (1) natural and pragmatic, as well as (2) helpful. After all, anthropomorphism in LLM research serves important technical and social purposes. On the technical side, it provides intuitive frameworks for understanding complex systems and offers pragmatic terminology for discussing model behavior. It could also be argued that human cognition provides a proven template for intelligence, making it a valuable guide for AI development that has already led to breakthroughs. And, since LLMs are trained on human-generated data and designed to interact with humans, at least some anthropomorphic framing may be inevitable and even desirable. On the social side, anthropomorphic framing might improve our ability to engage non-technical stakeholders, communicate ethical considerations, and support policy discussions. Thus, we do not advocate for eliminating anthropomorphism entirely. Rather, we argue that awareness of the prevalence and limitations of anthropomorphic thinking can reveal new and potentially clarifying research directions.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. 2024. All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 13–26.

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anthropic. 2024a. Claude's character.

Anthropic. 2024b. External reviews of "alignment faking in large language models".

arXiv.org submitters. 2024. arxiv dataset.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022b. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Mark Balas. 1978. Feedback control of flexible systems. *IEEE Transactions on Automatic Control*, 23(4):673–679.

Emily M Bender. 2024. Resisting dehumanization in the age of "AI". *Curr. Dir. Psychol. Sci.*, 33(2):114–120.

Leonard Bereska and Stratis Gavves. 2024. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*. Survey Certification, Expert Certification.

Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. 2022. My AI friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research*, 48(3):404–429.

Britta C Brugman, Christian Burgers, and Gerard J Steen. 2017. Recategorizing political frames: a systematic review of metaphorical framing in experiments on political communication. *Annals of the International Communication Association*, 41(2):181–197.

Erik Brynjolfsson. 2023. The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence. In *Augmented education in the global age*, pages 103–116. Routledge.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*. Survey Certification, Featured Certification.

Serina Chang, Ashton Anderson, and Jake M. Hofman. 2025. ChatBench: From static benchmarks to human-AI evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26009–26038, Vienna, Austria. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Myra Cheng, Kristina Gligoric, Tiziano Piccardi, and Dan Jurafsky. 2024. AnthroScore: A computational linguistic measure of anthropomorphism. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 807–825, St. Julian's, Malta. Association for Computational Linguistics.

Myra Cheng, Robert D Hawkins, and Dan Jurafsky. Accommodation and epistemic vigilance: A pragmatic account of why LLMs fail to challenge harmful beliefs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, San Diego, USA. Association for Computational Linguistics.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2026. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *The Fourteenth International Conference on Learning Representations*.

Michelle Cohn, Mahima Pushkarna, Gbolahan O Olanubi, Joseph M Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. 2024. Believing anthropomorphism: Examining the role of anthropomorphic cues on trust in large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Mike Dacey. 2017. Anthropomorphism as cognitive bias. *Philosophy of Science*, 84(5):1152–1164.

Jessica Dai. 2024. Position: Beyond personhood: Agency, accountability, and the limits of anthropomorphic ethical analysis. In *Forty-first International Conference on Machine Learning*.

Adam Davies and Ashkan Khakzar. 2024. The cognitive revolution in interpretability: From explaining behavior to interpreting representations and algorithms. *arXiv preprint arXiv:2408.05859*.

Edsger W Dijkstra. 1985. On anthropomorphism in science. *EWD936, Sept*.

Eamon Duede and Richard Jean So. 2024. The humanistic case for AI optimism. *Poetics Today*, 45(2):215–222.

Nicholas Epley. 2018. A mind like mine: The exceptionally ordinary underpinnings of anthropomorphism. *Journal of the Association for Consumer Research*, 3(4):591–598.

Oluwole Fagbohun, Rachel M Harrison, and Anton Dereventsov. 2024. An empirical categorization of prompting techniques for large language models: A practitioner's guide. *arXiv preprint arXiv:2402.14837*.

K. J. Kevin Feng, Kevin Pu, Matt Latzke, Tal August, Pao Siangliulue, Jonathan Bragg, Daniel S Weld, Amy X. Zhang, and Joseph Chee Chang. 2026. Cocoa: Co-planning and co-execution with ai agents. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA. Association for Computing Machinery.

Luciano Floridi and Anna C Nobre. 2024. Anthropomorphising machines and computerising minds: the crosswiring of languages between artificial intelligence and brain & cognitive sciences. *Minds and Machines*, 34(1):1–9.

Kristel M Gallagher and John A Updegraff. 2012. Health message framing effects on attitudes, intentions, and behavior: a meta-analytic review. *Annals of behavioral medicine*, 43(1):101–116.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, and 1 others. 2024. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*.

David Gros, Yu Li, and Zhou Yu. 2021. The R-U-a-robot dataset: Helping avoid chatbot deception by detecting user questions about human or non-human identity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6999–7013, Online. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E Weston, and Yuandong Tian. 2025. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling*.

John Hewitt, Nelson F Liu, Percy Liang, and Christopher D Manning. 2024. Instruction following without instruction tuning. *arXiv preprint arXiv:2409.14254*.

Marius Hobbhahn. 2024. We need a science of evals.

Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2025. Training language models to be warm and empathetic makes them less reliable and more sycophantic. *arXiv preprint arXiv:2507.21919*.

Theodore Jensen, Mary Theofanos, Kristen Greene, Olivia Williams, Kurtis Goad, and Janet Bih Fofang. 2024. Reflection of its creators: Qualitative analysis of general public and expert perceptions of artificial intelligence. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 647–658.

Cameron Jones and Ben Bergen. 2024. Does gpt-4 pass the turing test? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies (Volume 1: Long Papers), pages 5183–5210.

Cameron Jones and Benjamin Bergen. 2026. Lies, damned lies, and language statistics: a comprehensive review of risks from manipulation, persuasion, and deception with large language models. *Artificial Intelligence Review*.

Julie Kallini, Shikhar Murty, Christopher D Manning, Christopher Potts, and Róbert Csordás. 2025. MrT5: Dynamic token merging for efficient byte-level language models. In *The Thirteenth International Conference on Learning Representations*.

Atoosa Kasirzadeh and Iason Gabriel. 2023. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2):27.

Ayush Kaushal and Kyle Mahowald. 2022. What do tokens know about their characters and how do they know it? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2487–2507, Seattle, United States. Association for Computational Linguistics.

Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. 2020. Conceptual metaphors impact perceptions of human-ai collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan A, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling declarative language model calls into state-of-the-art pipelines. In *The Twelfth International Conference on Learning Representations*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago Press.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2025. Eliciting human preferences with language models. In *The Thirteenth International Conference on Learning Representations*.

Yugang Li, Baizhou Wu, Yuqi Huang, Jun Liu, Junhui Wu, and Shenghua Luan. 2024. Warmth, competence, and the determinants of trust in artificial intelligence: A cross-sectional survey from China. *International Journal of Human–Computer Interaction*, pages 1–15.

Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, and Alexander Nicholas D'Amour. 2025. Bias in language models: Beyond trick tests and towards RUTEd evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 137–161, Vienna, Austria. Association for Computational Linguistics.

John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. 2006. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, 27(4):12–12.

John McCarthy, Nathaniel Rochester, and Claude Shannon. 1956. Dartmouth workshop.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024a. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D Hardy, and Thomas L Griffiths. 2024b. When a language model is optimized for reasoning, does it still show embers of autoregression? an analysis of openai o1. *arXiv preprint arXiv:2410.01792*.

Douglas L Medin. 1989. Concepts and conceptual structure. *American psychologist*, 44(12):1469.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Ethan Mollick. 2024. Sometimes our anthropocentric assumptions about how intelligence "should" work (like using language for reasoning) may be holding AI work back.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Meredith Ringel Morris. 2024. Prompting considered harmful. *Communications of the ACM*, 67(12):28–30.

Rajiv Movva, Sidhika Balachandar, Kenny Peng, Gabriel Agostini, Nikhil Garg, and Emma Pierson. 2024. Topics, authors, and institutions in large language model research: Trends from 17k arxiv papers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1223–1243.

Nika Mozafari, Welf H Weiger, and Maik Hammerschmidt. 2020. The chatbot disclosure dilemma: Desirable and undesirable effects of disclosing the non-human identity of chatbots. In *ICIS*, pages 1–18.

Gregory L. Murphy. 2002. *The Big Book of Concepts*. The MIT Press.

Clifford Nass, Youngme Moon, and Paul Carney. 1999. Are people polite to computers? responses to computer-based interviewing systems. *Journal of Applied Social Psychology*, 29(5):1093–1109.

Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. 2022. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alexander Pan, Erik Jones, Meena Jagadeesan, and Jacob Steinhardt. 2024. Feedback loops with language models drive in-context reward hacking. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Diane Proudfoot. 2011. Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence*, 175(5-6):950–957.

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. Ai and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Murray Shanahan. 2024. Talking about large language models. *Communications of the ACM*, 67(2):68–79.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Jianna So, Connie Cheng, and Sonia Krishna Murthy. 2026. Beyond anthropomorphism: a spectrum of interface metaphors for LLMs. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Yao Song and Yan Luximon. 2020. Trust in AI agent: A systematic review of facial anthropomorphic trustworthiness for social robot design. *Sensors*, 20(18):5087.

Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the gulf of envisioning: Cognitive challenges in prompt based interactions with LLMs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Peiqi Sui, Eamon Duede, Sophie Wu, and Richard So. 2024. Confabulation: The surprising value of large language model hallucinations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14274–14284, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei. 2024. There is a nuanced but important difference between chain-of-thought before and after o1. ...

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. 2026. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. In *The Fourteenth International Conference on Learning Representations*.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023. LLM lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Hongxin Zhang, Yanzhe Zhang, Ruiyi Zhang, and Diyi Yang. 2022. Robustness of demonstration-based learning under limited data scenario. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1769–1782, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. 2024. Beyond preferences in AI alignment. *Philosophical Studies*, pages 1–51.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2025. REL-A.I.: An interaction-centered approach to measuring human-LM reliance. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11148–11167, Albuquerque, New Mexico. Association for Computational Linguistics.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. Prosa: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976.

A Verb Analysis

To understand the cause of the prevalence of anthropomorphism, we perform a Fightin' Words analysis (Monroe et al., 2008) on our dataset (following the approach of Cheng et al. (2024)), identifying verbs with highest z-scores for high (> 1) and low (< -1) AnthroScore sentences, i.e., the words that differ most significantly in distribution between these two sets using weighted log odds ratios. This also enables us to understand the sensitivities and biases of the AnthroScore metric. The results are as follows:

- High AnthroScore: *achieve, guide, demonstrate, teach, ask, train, prompt, follow, make, target, become, learn, understand, excel, require, mislead, answer, hallucinate, memorize, draw;*
- Low AnthroScore: *propose, outperform, use, develop, present, evaluate, enhance, improve, introduce, implement, validate, apply, reduce, adapt, employ, extend, allow, leverage, utilize, design*.

Top verbs for high AnthroScore reflect both human-like actions and also verbs that are commonly used in discussing LLMs, underscoring the prevalence of anthropomorphic terminology in papers about LLMs.