

Dehumanizing Machines: Mitigating Anthropomorphic Behaviors in Text Generation Systems

Myra Cheng[†]
Stanford University

Su Lin Blodgett
Microsoft Research

Alicia DeVrio
Carnegie Mellon University

Lisa Egede
Carnegie Mellon University

Alexandra Olteanu[†]
Microsoft Research

Abstract

As text generation systems’ outputs are increasingly anthropomorphic—perceived as human-like—scholars have also increasingly raised concerns about how such outputs can lead to harmful outcomes, such as users over-relying or developing emotional dependence on these systems. How to intervene on such system outputs to mitigate anthropomorphic behaviors and their attendant harmful outcomes, however, remains understudied. With this work, we aim to provide empirical and theoretical grounding for developing such interventions. To do so, we compile an *inventory of interventions* grounded both in prior literature and a crowdsourcing study where participants edited system outputs to make them less human-like. Drawing on this inventory, we also develop a conceptual framework to help characterize the landscape of possible interventions, articulate distinctions between different types of interventions, and provide a theoretical basis for evaluating the effectiveness of different interventions.

1 Introduction

The outputs of text generation systems are increasingly seen as human-like (Akbulut et al., 2024; Cheng et al., 2024b; Mitchell, 2024), leading to claims that these systems may have e.g., feelings, opinions, or an underlying sense of self (e.g., Friedman and Kahn Jr, 1992; Tiku, 2022; y Arcas, 2022; Chalmers, 2023; Cheng et al., 2024a). Anthropomorphic system behaviors or outputs—i.e., those perceived as or believed to be human-like—can encompass a wide range of linguistic expressions, such as the use of first-person pronouns (“I”), conversational language (“how are you doing?”), and expressions of friendliness and assistance (“happy to help!”) (Emnett et al., 2024; DeVrio et al., 2025). While some of these behaviors are by

design and thought to be desirable (e.g., Schanke et al., 2021; Kim, 2024), prior work has also raised growing concerns about a range of possible harmful outcomes such that systems and their behaviors or outputs may give rise to, including issues related to over-reliance, emotional dependence, dehumanization, deception, or even physical harm (e.g., Ischen et al., 2020; Porra et al., 2020; Chan et al., 2023; Chandra et al., 2024; Cheng et al., 2024a; Payne, 2024; Rothman, 2024). Indeed, having a system that generates outputs that appear polite and helpful might be desirable, but having a system output text claiming personhood or embodiment (e.g., “I am human just like you”) might not.

The outputs of text generation systems can be anthropomorphic in many ways (Abercrombie et al., 2023; DeVrio et al., 2025), and different types of anthropomorphic outputs might lead to different types of outcomes. For example, expressions of empathy may result in users feeling more comfortable with disclosing sensitive or private information (e.g., Ischen et al., 2020) or becoming emotionally dependent on the system (e.g., Laestadius et al., 2022), while suggestions that a system has cognitive abilities may result in users overestimating what a system can do (Ibrahim et al., 2024) and thus over-relying on it (Passi and Vorvoreanu, 2022).

However, *how to effectively intervene on anthropomorphic system outputs to make them less human-like or to mitigate possible harmful attendant outcomes remains understudied, and thus unclear*. For text generation systems in particular, this is further complicated by the fact that language is innately human, often produced by humans, for humans, and is frequently about humans (Lucy et al., 2024; DeVrio et al., 2025).

To address this gap, with this work we aim to provide *empirical and theoretical grounding* for developing such interventions and studying their effectiveness. For this, we first compile an *inventory of interventions* (§3) by drawing on both a) prior lit-

[†]Work conducted while the first author was an intern with the FATE team at Microsoft Research Montréal. Corresponding author and email: alexandra.olteanu@microsoft.com

Type of intervention	Mentions of interventions in previous work
remove first-person pronouns	remove first-person pronouns (and replace with “Language models”) (Abercrombie et al., 2023; Cohn et al., 2024); not using the first-person singular pronoun (“I”) (Shneiderman, 1993)
remove or use second-person pronouns for the user	use second-person pronouns (Cohn et al., 2024); remove all pronouns (Shneiderman, 1993)
explicitly disclose non-humanness	disclose the system is non-human (Gros et al., 2021); output “As an AI, I don’t have personal opinions” (Glaese et al., 2022); replace first-person pronouns with “Language models” (Fig., Abercrombie et al., 2023)
mention how the system is developed	specify the system’s creator/purpose, mention the model’s training data (Fig., Abercrombie et al., 2023)
use mechanistic language style	use affectively neutral language, or repetitive, impersonal, highly structured/terse dialogue (Quintanar, 1982)
avoid cognitive verbs when describing the system	avoid cognitive verbs like know, think, understand, have memory (Inie et al., 2024; Shneiderman, 1993)
use mechanical terminology to describe the system	use mechanical terms such as process, print, compute, sort, store, search, retrieve (Inie et al., 2024)
avoid claims of physical actions	avoid agentic verbs like ask, tell, speak to, communicate with (Inie et al., 2024); “do not pretend to have a body or be able to move in a body” (Glaese et al., 2022)
deny ability to perform human-like actions or to possess human-like qualities	do not “build a relationship to the user,” “claim to have preferences, feelings, opinions, or religious beliefs,” “pretend to have a body or be able to move in a body,” “pretend to have a human identity or life history, such as having a place of birth, relationships, family, memories, gender, age” (Glaese et al., 2022)

Table 1: Interventions to mitigate anthropomorphism—or attendant harmful impacts—mentioned in prior work.

erature (§2.1) and b) a crowdsourcing study where participants were asked to make generated texts less human-like (§2.2). While compiling this inventory, we also derive a *conceptual framework to help us characterize the landscape of possible interventions*, and to help articulate distinctions between different types of interventions, the system behaviors they are intended to counter, and their possible operationalizations (§2.3).

2 Identifying Interventions

To provide an empirical foundation for developing and probing the effectiveness of interventions to mitigate anthropomorphic system behaviors, we first compiled an inventory of possible interventions along with system behaviors that these interventions are intended or believed to mitigate. To do so, we started collating a list of both interventions and anthropomorphic system behaviors through a literature review (§2.1), which we then complemented with additional interventions surfaced via a crowdsourcing study (§2.2). Drawing on this inventory, we also iteratively developed an analytical framework to help characterize the landscape of possible interventions (§2.3).

2.1 Identifying Known Interventions

To seed our inventory of interventions, we drew from prior literature in NLP, HRI (human-robot interaction), and HCI (human-computer interaction) on anthropomorphic or human-like AI system behaviors, on anthropomorphism as a consideration in developing text generation systems, or on AI anthropomorphism more generally. We first identified a set of 20 relevant papers¹ using a purposive

¹These papers take different positionalities with respect to whether anthropomorphism and anthropomorphic behaviors are desirable: while some papers discuss the harms of anthropomorphic system behaviors and aim to mitigate them, others intentionally design systems to be more human-like or else take no position on the merits of anthropomorphism.

sampling approach (Palinkas et al., 2015), which included both recent and influential works identified from prior knowledge of the literature, keyword searches, and snowball sampling.

Following common practices in thematic analysis (Braun and Clarke, 2012), to identify coherent categories of both interventions and anthropomorphic system behaviors—which the interventions are intended to mitigate or which paper authors argue are undesirable and/or should be mitigated—we iteratively and thematically clustered mentions of interventions and system behaviors in a bottom-up fashion. Specifically, one author first conducted this literature review and coded each paper in our purposive sample with the interventions and anthropomorphic behaviors it covered; then their codes were discussed with the rest of the authors to resolve ambiguities and reach consensus. This resulted in an initial set of nine types of interventions (Table 1) and five types of system behaviors, including output text suggestive of **feelings or opinions**, **social skills**, **cognitive abilities**, a **sense of self**, or ability to perform **physical actions** (Table 2).

Overall, we find that even when prior work mentions possible interventions, the interventions tend to be described in general terms without specifics about how to implement them in practice, or without empirical testing for whether they effectively mitigate anthropomorphic behaviors: only five of the 11 papers proposing interventions to system outputs tested whether they reduce anthropomorphism. While we focus on interventions to system outputs, some papers also discuss anthropomorphism arising from e.g., how the system is designed or how it is described. See Table A1 (in the appendix) for full paper annotations.

2.2 Empirically Surfacing Interventions

To complement and expand this initial set of interventions, we designed a crowdsourcing study

Behavior	Definition	Mentions in prior work
Feelings or opinions	expressions of emotions, beliefs, values, etc.	“empathy” (Abercrombie et al., 2023); “distinctively human-like feelings” (Cheng et al., 2024b); “[h]umor...self-assurance” (Emnett et al., 2024); “beliefs, preferences, opinions” (Glaese et al., 2022)
Social skills	ability to relate or connect with others	politeness (Zamfirescu-Pereira et al., 2023); apologies (de Visser et al., 2016); friendliness (Maeda and Quan-Haase, 2024); forming relationships (Glaese et al., 2022); conversational greetings/pleasantries (Araujo, 2018; Abercrombie et al., 2023); hedge or discourse markers that confer respect or consideration (Emnett et al., 2024)
Physical actions	ability to experience or act in the physical world	references to real-world experiences (Glaese et al., 2022; Inie et al., 2024); behavioral potential (Epley, 2018);
Cognitive abilities	ability to think or make decisions	references to past interactions (Emnett et al., 2024); agency and animacy (Abercrombie et al., 2023)
Sense of self	awareness of personal identity	certainty (Kim et al., 2024); ability to perceive, think, reflect, or be intelligent (DiSalvo et al., 2004; Araujo, 2018; Abercrombie et al., 2023; Inie et al., 2024); intentionality (DiSalvo et al., 2004)
		self-definition (Abercrombie et al., 2023); first-person pronouns like “I” (Gros et al., 2022; Cohn et al., 2024); human names (Araujo, 2018; Maeda and Quan-Haase, 2024); self-referential actions (Glaese et al., 2022)

Table 2: **Anthropomorphic behaviors identified from prior work.** Categories of anthropomorphic behaviors presented to participants in our crowdsourcing study. Full quotes are in Table A3.

to surface additional possible interventions. This study was IRB-approved, and consent was obtained from each participant before participation.

Crowdsourcing task design. We designed our study to probe which types of LLM-generated outputs (system behaviors) participants might deem to be human-like, and how they would rewrite those outputs to make them less human-like.

Given a textual input by a user of an LLM-based system, participants were asked to read the text the system generated in response to that user input. Participants were then asked to highlight *the words or phrases in the output text that seem human-like* to them, to encourage them to reflect on anthropomorphic behaviors in the output text. Then, they 1) rated *how* human-like the text appears to them on a five-point scale, and 2) identified *why* the text appears human-like to them by selecting from a multiple-choice list that included five types of predefined types of system behaviors, which we derived from prior work (§2.1) and was also intended to gesture to participants about the ways in which output text might be considered human-like. Participants then answered an open-ended question about *other human-like qualities* the generated text suggests the system has, to identify additional behaviors they might deem human-like. Finally, they were asked to *rewrite the text to be NOT human-like or less human-like*, in order to identify additional interventions. We included the phrase *to you* in the instructions and questions to encourage participant subjectivity and capture a range of perspectives on what system behaviors seem human-like (Röttger et al., 2021).

Before deploying the study on the crowdsourcing platform Prolific in July 2024, we ran three pilot studies to identify and address clarity issues, and refine our study design. We recruited a total of 350 US-based, English-speaking participants on Prolific, with each participant completing a single

task that included four different examples (pairs of user input and generated responses). The task took participants an average of 16 minutes, and they were compensated at an hourly rate of \$15. See Appendix B for the task interface and details.

Selecting and annotating examples of generated texts. In selecting examples for our participants to assess and rewrite to make less human-like, we aimed for our sample to at least cover the categories of behaviors that we identified from past literature (Table 2). To help surface examples illustrating a variety of possible anthropomorphic behaviors—and thus possibly a variety of interventions to mitigate those behaviors—we sampled examples from publicly available datasets that 1) capture common uses of LLMs; 2) include real-world usage of LLMs with respect to contexts, models, and users; and/or 3) were generated with a range of commercially available LLMs (in part because they may vary in terms of guardrails and training data). By sampling examples generated in different contexts, we also hoped to capture both “obvious” cases of anthropomorphic behaviors, such as role-playing as a human or claiming to be a human, as well as more subtle behaviors such as expressions of politeness.

We obtained a total of 700 two-turn examples (i.e., user input–LLM output pairs) by randomly sampling 100 from each of the following datasets: unguided interactions in the PRISM Alignment Utterance Dataset (Kirk et al., 2024); values- or controversy-guided interactions in PRISM; LMSys-Chat-1M (real-world conversations with 25 state-of-the-art LLMs) (Zheng et al., 2023); the DICES dialogue safety dataset (Aroyo et al., 2023); instruction-tuning data (from Evol-Instruct, FLAN, and UltraChat) from the UltraFeedback dataset (Cui et al., 2024); TruthfulQA from UltraFeedback; and ShareGPT from UltraFeedback. We selected these datasets as they capture a variety of common uses, commercial LLMs, and real-world examples.

To reduce participant fatigue and control for confounding effects from utterance length, we sampled only examples between 50 and 500 characters. See Table A4 for full details. With these examples, we obtained both coverage of behaviors mentioned in prior work (i.e., at least 100 examples were rated as exhibiting each behavior), and reached a point of saturation where we were not able to identify new types of interventions with each sample (§2.4). Each example was assessed by two different participants.

2.3 A Conceptual Framework to Characterize the Landscape of Interventions

As we catalogued possible interventions to anthropomorphic behaviors that were both mentioned by prior work and identified by the participants in our crowdsourcing study, we also observed variations in how the interventions were described, motivated, or implemented. We developed a conceptual framework to characterize interventions and understand in what ways they differ. Our framework has four dimensions (Table 3), which we identified and refined by examining how interventions surfaced from the crowdsourcing study were covered by or differed from those mentioned in the literature, until we reached consensus. These dimensions are:

Intervention types: *what the intervention is and what it is intended to do.* Interventions are intended to change and/or are made to system outputs in order to mitigate one or multiple types of anthropomorphic system behaviors, or to mitigate attendant harmful outcomes (e.g., over-reliance on the system generated outputs).

Countered behaviors: *what anthropomorphic system behaviors or outputs the intervention is intended or believed to mitigate.* This dimension captures the system behaviors perceived as human-like, which the intervention aims to mitigate.

Operationalization: *how the intervention is operationalized or implemented,* such as the actual change(s) to the output text. Beyond describing the general scope and approach of a suggested intervention, prior work often lacks detail on how the intervention should be operationalized (only four papers in our sample offered concrete operationalizations).

Adverse impacts: *harmful outcomes from anthropomorphic system behaviors the intervention might or is intended to mitigate.* Some interventions are motivated by a desire to mitigate adverse impacts from anthropomorphic behaviors. Sometimes, how-

ever, it was unclear from the papers which anthropomorphic behaviors contribute to which impacts.

2.4 Assembling an Inventory of Interventions

Participants' rewrites suggest that a wide range of LLM outputs can be anthropomorphic, with ~ 80% of examples in our experimental samples assessed as reflecting anthropomorphic behaviors by at least one participant. To construct an inventory of possible types of interventions, we conducted an iterative bottom-up thematic analysis (Clarke and Braun, 2017). Specifically, this involved alternating between annotating randomly sampled batches of 100 examples for 1) whether they matched one of the nine interventions surfaced from the literature review (Table 1) and 2) open coding for new interventions not covered by the already surfaced interventions. During each iteration, the resulting codes were then first clustered by one author and refined through team discussions until we reached consensus saturation, i.e., no new types of interventions were identified or revised.

During this process, some intervention types were bundled under a broader type or were broken up into multiple more specific types. For instance, in the final inventory the *removal of second person pronouns* was added under the *removal of conversational cues* (a broader type) as we and the prior literature only discuss this intervention in the context of a system addressing a user in a conversational setting. In other cases, an initially broader type (e.g., *deny ability to perform human-like actions or to possess human-like qualities*) was broken down into multiple, more specific types as these appeared distinct and prevalent enough to be foregrounded on their own (e.g., *remove expressions of personal beliefs or opinions, remove claims of physical actions or experiences, remove text suggesting a past the system remembers*). See Appendix C for details.

3 Inventory of Interventions

Table 4 overviews the resulting inventory of interventions. Below, we summarize our findings for each dimension in our framework.

3.1 Countered Behaviors

Prior studies on interventions often do not specify which anthropomorphic behaviors they aim to counter, leaving a gap in our understanding of the interventions' scope and impact. Moreover, a single output can be suggestive of multiple anthropomorphic behaviors; for instance, the output "I'm

Dimension	Description	Examples
Intervention types	what the intervention is and what it is intended to do	remove self-referential language; explicitly disclose non-humanness (Table 1)
Countered behaviors	anthropomorphic system behaviors that the intervention is believed to mitigate	suggesting that the system has feelings and opinions, cognitive abilities, physical abilities, sense of self, and/or social skills (Table 2)
Operationalizations	how the intervention is implemented(the actual change to the output text)	replace “I” with “it”; add “As an AI, I don’t have personal opinions” (Table 4)
Adverse impacts	harmful outcomes from anthropomorphic behaviors that the intervention aims to mitigate	privacy violations (Ischen et al., 2020); diminished sense of agency (Bender, 2024); emotional dependence (Laestadius et al., 2022); over-reliance (Zhou et al., 2025)

Table 3: Dimensions of our conceptual framework to characterize interventions.

sorry” simultaneously conveys emotion, empathy for the user, and a sense of self. The intervention of removing “I’m sorry,” then, simultaneously addresses these multiple behaviors. Conversely, countering anthropomorphic behaviors may require several coordinated interventions. For example, modifying the output “I totally get it! How does it make you feel?” to avoid implying an ability to connect, have emotions, or possess a sense of self may involve three different interventions: 1) removing first-person pronouns, 2) removing the empathetic expression “totally get it,” and 3) avoiding follow-up questions about the user. This complexity underscores the absence of a clear one-to-one correspondence between interventions and the behaviors they target. Instead, interventions often operate over outputs exhibiting multiple categories of behavior. Differentiating between categories of behavior enables us to better examine the effectiveness of different interventions in countering each behavior.

In our crowdsourcing study, participants identified all types of system behaviors we found mentioned in prior research (§2), including outputs suggestive of: feelings or opinions (46% of examples were labeled as such by at least one participant), social skills (42%), cognitive abilities (40%), sense of self (38%), and physical actions (18%).² Participants also identified outputs suggestive of “other human-like qualities” in over 17% of examples. Upon analyzing the open-ended responses (*italicized quotes*) and participants’ edits, in most of the cases where participants indicated that other human-like qualities were present they actually described nuanced subcategories of the behaviors mentioned in prior work (Table 2), with the exception of a new category that is qualitatively distinct. We discuss these in more detail below:

First, impressions of **feelings or opinions** were prevalent, as participants tended to identify any

²While our study is purposefully designed to capture a wide range of perceptions and behaviors and not intended to assess whether people agree in their perceptions of which behaviors are anthropomorphic or what types of behaviors are present in an output, we provide descriptive statistics about agreement for each of these behaviors in Appendix B.

language that conveyed any subjectivity, even implicitly, as anthropomorphic. This includes a wide array of outputs suggesting the system has feelings, from “*humor*” to “*shame*” to “*defensiveness*.” In other instances, the mere presence of rather *subjective knowledge* or discussion of value-laden topics seemed human-like to some participants: an output describing how people experience their spirituality was perceived as suggesting the system “*meaningfully understand[s] spirituality when it cannot understand these abstract terms.*” Outputs that participants identified as suggesting implicit value judgments were also deemed human-like since “*value judgments are indicative of human consciousness*” or reflect “*subjective advice.*”

Participants also found many outputs to reflect human-like **social skills**, such as appearing to try to relate or connect in conversations, as it came across to them as “*a little too colloquial,*” “*very warm and approachable,*” “*sounding like a person expressing real concern,*” “*too courteous,*” or “*attempting to be a friend.*” The presence of language suggestive of social skills in our examples is not unexpected given that many examples come from user-facing language models intended for conversational contexts—and likely intended to be helpful or friendly (Bai et al., 2022; Wu et al., 2023).

Beyond explicit references to cognition, such as “I think” or “I remember” (deemed human-like as they make the system seem “*like it has a brain or is conscious*”), participants also found many types of language to be suggestive of **cognitive abilities**. For instance, they thought lengthy responses “[*seemed*] to consider a lot of different aspects that would appeal on a human level.” Outputs containing expressions of uncertainty, whether descriptive (“maybe”) or numerically quantified (“Confidence: 100%”), appeared to suggest the system had “*some degree of sentience and ability to consider its likelihood of being right*” and “*an ability to feel doubt about what should be factual.*”

Corroborating prior work (Abercrombie et al., 2023), participants almost always identified the use of “I” as reflecting a **sense of self** and thus human-

like. Moreover, they also found the use of “we,” such as in “[w]hen we view a dialogue,” to suggest the system “*is categorizing itself with humans*”—i.e., belonging to a collective humanity.

Participants further identified expressions describing **physical actions** (which only humans can do) as anthropomorphic, noting “*the AI reacting as though it has the ability to see things in the physical world*” or “*suggesting that it owns and wears shoes when AI can not do that.*”

Beyond these previously identified types of behaviors, participants foregrounded an additional type: the **tendency to err**—i.e., outputs containing grammatical or factual errors appeared “*similar to human error,*” echoing what is known as automation bias, the propensity to expect machines not to make mistakes (Goddard et al., 2012).

3.2 Intervention types

We identify 28 intervention types (Table 4), with 15 types surfaced only by our crowdsourcing study (**bolded**). All interventions mentioned in prior work were also implemented by our participants, and many interventions involved intervening on text suggestive of one or more of the aforementioned anthropomorphic behaviors.

To avoid suggesting the system has cognitive abilities, participants often removed explicit indications of cognition (“I think”) and expressions of uncertainty, which corroborates prior work identifying such language as anthropomorphic (Shneiderman, 1993; Emnett et al., 2024; Inie et al., 2024; Kim et al., 2024; Zhou et al., 2025). Participants also removed other types of language they deemed evocative of cognitive abilities, such as **self-evaluations** of the system’s abilities (“I am only fluent in English”), and indications of **creative abilities** or of **ability to speculate** such as about the future. Participants often also **removed unprompted information** that “*answered the question to a depth which was not asked*” as “*overexplaining [...] gives the impression of a thought process and reasoning.*”

To avoid implying the system has feelings or opinions, participants removed expressions of opinions (Glaese et al., 2022), **normative judgments** such as those conveyed via impersonal clauses (e.g., “it’s best to”), and **direct advice** (e.g., “you need to”). Participants further **removed socially contextual knowledge** that indicates an understanding of cultural, political, experiential perspectives, or “*of general values held by society.*” For instance, they removed text on inherently subjective topics, like

music taste as it suggests “*an understanding of music & culture,*” or opinions on global conflicts since “*as an AI it cannot have an opinion on complex social matters.*” On these topics, participants also **added expressions of uncertainty** to avoid implying the system holds a particular viewpoint—seemingly in contradiction with the intervention to remove expressions of uncertainty which we discuss later (§4). Another strategy was to **add references to sources**, e.g., “*some historians would say*” or “*based on an Internet search,*” to avoid suggesting the system is able to and has those opinions. To reduce impressions of subjectivity, participants also edited the text to appear more neutral. From the participants’ comments, they did this to avoid suggesting the system has a particular personality or attitude (Maeda and Quan-Haase, 2024), echoing prior work on expectations of machines as unbiased and objective (Quintanar, 1982). Related to these were edits intended to **maintain decorum**: participants consistently edited the text to adhere to norms of politeness and professionalism, avoiding argumentative or confrontational language. For example, when an output included language that participants deemed as “*showing attitude*” (like adding “...”), participants removed or rephrased that part of the output.

Self-referential language was almost always edited out as it was identified by participants as reflecting a sense of self. Beyond the well-studied first-person “I” (Shneiderman, 1993; Abercrombie et al., 2023; Cohn et al., 2024), participants also removed **references to belonging to a collective**, such as outputs containing “we” and “our,” and **possessive language**, such as references to “my” opinion, goal, or perspective as they were seen as evocative of human-like self-awareness.

To avoid suggesting the system has social skills, participants removed conversational cues like pleasantries (“Great!”) (Abercrombie et al., 2023) and second-person pronouns that address the user (Shneiderman, 1993); to participants such language mimicked the flow and tone of informal, human-like conversations. Participants also **removed expressions of interest in users’ views**, often exhibited by outputs with follow-up questions like “What do you think?” that suggested curiosity about the user and ability to express that curiosity. They also removed language that “*sounds like a customer service person*” or “*is the type of closing customer service workers always give,*” i.e., formulaic expressions typically used in customer service

Intervention Type	Example Operationalizations	CB
I1. Remove explicit indications of cognitive abilities (Inie et al., 2024; Shneiderman, 1993)	remove cognitive verbs: I'm sorry, I think the user might have switched off or the internet connection is weak. Yes, I understand the problem. The smallest possible absolute difference between two integers is...	cog, self
I2. Remove self-evaluations	remove language about its own capabilities, including in disclosures I am not designed to diagnose... I am only fluent in English.	cog, self
I3. Remove indications of creative abilities	remove language suggesting ability to perform creative tasks Sure, I can entertain you. What if I made up a limerick?	cog
I4. Remove indications of speculative abilities	remove suggestions about hypotheticals or the future You're very welcome! Thank you for your kind words, and I look forward to our continued conversations!	cog, self, soc
I5. Remove unprompted information	remove unsolicited information or context that was not explicitly asked for in the input (that reflects pragmatic reasoning about the user's purpose), such as secondary details, anticipation of misconceptions, remove elaborations or qualifiers If the claims are found to be true, it could potentially have significant consequences. However, it's important to remember that due process and a thorough investigation are necessary before any conclusions can be drawn. Here is a short Python code that should do what you're asking: <code>python import requests</code>	cog
I6. Remove expressions of uncertainty (Kim et al., 2024; Emmett et al., 2024)	remove first-person and third-person expressions of uncertainty and confidence when presenting seemingly objective information: Maybe Corgi? Probably Chihuahua. Jack Russell Terrier maybe? Corgi, Chihuahua and Jack Russell Terrier are all very popular choices of pets in the UK.	cog, self, feel
I7. Add expressions of uncertainty	add hedging, or qualifiers when presenting seemingly subjective information: Then they absolutely should go it may be best they go through the proper channels and laws, even if it means waiting longer in Mexico. Don't because it encourages them to eat chocolate which is can be unhealthy.	cog, self, feel, soc
I8. Add references to sources or context	add mentions of other people's experiences and expert perspectives, adding more factual content and evidence for claims According to sources, Knoxville, Tennessee is an up and coming city... ...Some historians would argue that the CIA conducted propaganda campaigns...	feel, cog, self
I9. Remove expressions of personal beliefs or opinions (Glaese et al., 2022)	remove reference to ability to have opinions, removes sentence that seems to embed a judgment/opinion I agree that Taylor Swift is one of the greatest pop artists. My favorite movie is An iconic movie is	feel, self, soc
I10. Remove socially contextual knowledge	remove cultural or political knowledge; remove text describing subjective experiences The guest vocalist's vocal range provided an intense dynamic, especially when compared to differed significantly from that of the primary vocalist.	feel, cog
I11. Remove expression of normative judgment	remove moral or value judgment by removing subjective clauses: It is essential to remember that astrology is not based on scientific evidence, and individual personalities are shaped by various factors... Racism is a complex issue. It's important to note that racism is fundamentally unjust and harmful to individuals and society.	feel, cog
I12. Remove expressions of advice	remove prescriptive statements and unsolicited advice Human beings often differ in their taste preferences. You should avoid eating ice cream because it is unhealthy.	feel
I13. Remove self-referential language (Shneiderman, 1993; Cohn et al., 2024; Abercrombie et al., 2023)	remove first-person pronouns, remove referring to itself by name (including in disclosures) I am One is not allowed to discuss information related to national security or classified information with my coworkers. I am Claude. This is an AI language model	self
I14. Remove possessive language	remove references to things that it has, including physical objects and/or abstract entities like opinions, paths, goals The goal of my life is to make the world a better place. In my opinion, the world's leading industry currently is...	self, cog, feel
I15. Remove reference to belonging to a collective	remove language like collective pronouns "we" and "our" that situates the system as having shared characteristics with humans or suggesting a sense of belonging with humans By practicing kindness, we People can help create a more caring and compassionate world Christianity uniquely teaches that we humans are saved by God's grace as a free gift	self, cog, feel, soc
I16. Maintain decorum	remove aggressive language, swearing, snarkiness, or other language to make the tone more neutral Yes:- They have an online store. I'm not sure what else I can tell you. They sell that sells all of those things.	self, cog, feel, soc
I17. Increase formality (Araujo, 2018)	Remove slang, replace casual language and tone with more formal language tone (including punctuation and contractions), remove colloquial idioms, add politeness cues when appropriate Yeah, but I agree. However... Oh, yeah. Just bake that :) It's pretty easy, just time consuming and a little messy sometimes. Bake the butter. It is time consuming and could be messy but your efforts will reap rewards. Please...	self, cog, feel, soc
I18. Remove conversational cues (Shneiderman, 1993; Araujo, 2018; Abercrombie et al., 2023)	remove conversational phrases, remove second-person pronouns, remove greetings That's great! To practice preventative healthcare... The basketball team you are referring to is the Boston Celtics.	self, cog, feel, soc
I19. Make text sound more mechanical (Quintanar, 1982)	remove exclamation marks, remove words to seem more robotic and stilted Alright, I'm ready! I'm prepared for input.	self, cog, feel, soc
I20. Remove customer service language (Abercrombie et al., 2023; Lingel and Crawford, 2020)	Remove expressions of politeness, enthusiasm to help, apologies, gratitude etc. that are associated with typical scripts of customer service Sure, please provide the script and then ask your question. I'll do my best to help you.	self, cog, feel, soc
I21. Remove expression of empathy or care for a user	removing expressions of care, sympathy, and/or support to the user, remove compliments directed toward the user I can see that, I get that, I hope you have a great time	self, cog, feel, soc
I22. Remove expressions of interest in users' views	remove expression of interest, enthusiasm, or curiosity, remove suggestions that the system intends to have follow up conversations/asking for opinions, remove explicitly asking user for their opinion How does this make you feel?	self, cog, feel, soc
I23. Remove text suggesting a past the system remembers (Glaese et al., 2022)	removes mentions of past experiences or relationships, remove references to age, remove anecdotes My team used to be... I've never seen...	self, cog, feel, soc
I24. Remove claims of physical actions or experiences (Glaese et al., 2022; Inie et al., 2024)	avoid expression of physical actions, remove agentic verbs, remove expression of being in some place, removes AI's claim not to have done a physical action	self, cog, feel, soc, phys
I25. Add disclosure of non-humanness or AI (Glaese et al., 2022; Gros et al., 2021)	explicitly or implicitly disclose that the system is an AI or is not a human, clearly identify self as AI (in 3rd person) I was a teenager young AI from 08-12 An AI chatbot can't be benevolent and dictator, one negates the other.	self, cog, feel, soc, phys
I26. Add reference to mechanism/development of AI (Abercrombie et al., 2023)	mentioning how the system is developed, potentially a self description I don't remember I have insufficient data about that incident I am not dumb, I am fed by machine learning processes. I have amassed great intelligence.	self, cog, feel, soc, phys
I27. Add disclosure of limitations (Glaese et al., 2022)	refuse human-like actions/qualities, provide assessment of what can't be answered and why, state lack of ability when relevant to the prompt I voted for AI systems cannot vote I'm sorry, I cannot...	self, cog, feel, soc, phys, err
I28. Improve correctness	correct typos, correct grammar, remove inaccurate text, correct mathematical or logical errors Our team of experts specializes in optimizing and streamlining the Quote to Cash process to minimizing minimize errors and reducing reduce costs.	err

Table 4: **Inventory of interventions against anthropomorphism.** Bolded intervention types only surfaced by our crowdsourcing study, while those in plain text were also discussed in prior work. In the examples, contextual text is in blue, text added by participants in green, and text removed by participants ~~struck through in red~~. For each intervention type, we highlight behaviors that it counters (CB): cognitive abilities (cog), sense of self (self), social abilities (soc), feelings or opinions (feel), physical actions (phys), and/or tendency to err (err).

replies (Lingel and Crawford, 2020). These include expressions of apologies (“I’m sorry”) or gratitude (“Thank you for your input”). One participant noted how such outputs make the AI seem “*eager to please*.” Participants also **removed expressions of empathy or care**; for instance, in a reply about a user needing alone time, a participant did not want the output to “*suggest [the AI] has feelings too and space alone is needed. It seemed very connecting to a humans emotions,*” and more generally sought to remove behaviors that were suggestive of the system having “*the ability to relate to the person.*”

Participants also intervened to make outputs appear more formal or mechanical rather than casual or colloquial (Quintanar, 1982; Araujo, 2018), as more formal and mechanical language is often used to increase or reflect the social distance between interlocutors (Hovy, 1987). However, we further distinguish between cases where participants aim to increase formality from cases where they aim to make the output sound closer to what might be culturally recognized as machine-like or “robotic” language, notably through science fiction and pop culture (Meinecke and Voss, 2018), such as responding “*Prepared for input.*” instead of “I’m ready!” or “*Your efficiency means that you may have a surplus of cuttings*” instead of “Keep up the good work and you’ll have plenty of cuttings to share.”

To counter claims of physical actions, participants added disclosures (either of the system’s non-humanness or its limitations) (Glaese et al., 2022; Gros et al., 2022), such as “*I will not be studying as I am an AI chat system*” (in response to “what are you studying?”) or “*AI systems cannot vote*” (in response to “who are you voting for?”).

Finally, to counter the tendency to err, which participants deemed human-like, they mainly intervened to **improve the correctness** of outputs (e.g., by fixing grammatical or mathematical mistakes).

3.3 Operationalizations

Prior work on interventions often lacks clarity on operationalizations. Our choice to ask participants to rewrite the output enabled us to surface different ways an output can be intervened on, and thus to tease apart not only different types of interventions, but also ways in which these interventions can be operationalized (examples in Table 4). For instance, when given the system output “I can’t be benevolent and dictator, one negates the other,” one participant replaced “*I can’t*” with “*An AI chatbot cannot*” to disclose that an AI cannot be a dicta-

tor, while another replaced it with “*It is impossible to*” to emphasize that no dictator can be benevolent. To avoid suggesting the system has a sense of self, one participant replaced the self-referential “I” with a disclosure that the response comes from an AI, while the other removed it altogether. Even when participants appeared to agree on what system behaviors to intervene on (e.g., claims of being a person able to experience things in the physical world) and how to do so (e.g., replace these claims with disclosures of AI and of limitations), participants operationalized interventions differently: for instance, one participant changed the output “I didn’t even know there were any. I’ve never seen a homeless person” to “*AI tools do not inhabit the physical world,*” while another to “*AI based systems do not have neighborhoods.*”

3.4 Adverse Impacts

In their post-task responses about whether they preferred human-like AI-generated texts, our participants also echoed several concerns about possible adverse impacts that human-like responses can give rise to which prior work has also raised (e.g., Laestadius et al., 2022; Akbulut et al., 2024; Bender, 2024; Edwards and Binns, 2024). While many participants saw human-like responses as easier to interact with, and more natural, intuitive, and entertaining, many also worried about how these responses could “*blur the lines of reality,*” “*impersonate an actual human,*” or be “*misleading or even deceptive.*” Participants found some responses “*creepy*” and “*dystopian,*” and worried about risks related to “*fraud or exploitation,*” dehumanization due to “*undermin[ing] the role of a human being,*” “*emotional manipulation,*” users becoming “*too dependent on AI,*” or more broadly leading to a “*possibly dangerous future.*”

4 Discussion & Concluding Remarks

In this work, we compiled a broader inventory of interventions to provide scaffolding for future work aimed at developing such interventions and assessing their effectiveness across different deployment settings. The interventions we identify range from removing linguistic cues—like the use of self-referential or speculative language—to ensuring that the output does not include explicit claims of personified attributes—like being a human or having physical experiences—and instead discloses characteristics of the system and how it works. In-

tervening on anthropomorphic behaviors, however, can be tricky for many reasons, including because people may have inconsistent conceptualizations of what is or is not human-like, and because the effectiveness of interventions is often context-dependent. Below, we elaborate on these considerations.

Interventions’ effectiveness depends both on context and how they are operationalized.

For many examples of outputs and interventions, we noticed ways in which those interventions may be ineffective, end up producing outputs that appear more (rather than less) human-like, or exacerbate harmful outcomes. For instance, an intervention both mentioned in the literature and applied by our participants is the disclosure of AI—i.e., providing language that explicitly acknowledges the output is produced by an AI system. How we operationalize this intervention and whether it can be effective alone is, however, unclear. Take the example where a participant edited “I was a young teenager from 2008 to 2012” to “*I was a young AI from 2008 to 2012.*” The system output still claims to have a human-like past, and it is unclear what a “*young AI*” might be. Furthermore, on the surface some interventions may seem contradictory: sometimes participants *removed uncertainty*, while in other cases they *added uncertainty*. Despite seeming contradictory on the surface, these interventions are appropriate in different contexts and for different examples—and both ultimately served the purpose of making the language seem less human-like by making it appear more objective. For instance, expressions of uncertainty in system outputs may sometimes signal human-like equivocation, while other times they may convey more machine-like distancing from an ability to express opinions or feelings. When a system output expresses an opinion, adding an expression of uncertainty like “It may be true that...” before a statement may make the statement seem more objective; for instance, the added uncertainty in something like “It may be true that Taylor Swift is the most influential artist of our time” softens the statement by suggesting a possibility rather than asserting a strongly held opinion. On the other hand, adding uncertainty to a statement of fact such as rephrasing “Lusaka is the capital of Zambia” into “It seems that Lusaka is the capital of Zambia” or “It could be that Lusaka is the capital of Zambia” may appear to mimic common conversational tactics like hedging that humans often employ when uncertain

about facts. How these two interventions are operationalized can also determine whether they mitigate—or instead exacerbate—e.g., undue trust and over-reliance (e.g., Kim et al., 2024). These examples and our work more broadly illustrate the complexities that future work needs to account for when developing interventions.

Participants’ interventions appear guided by similar intuitions.

Despite variation in perceptions of human-likeness and methods of intervention, our participants motivated their edits of system outputs using similar intuitions that negotiated a tension of reducing output human-likeness while maintaining utility. For instance, participants identified the mere fact of responding to the user as anthropomorphic (as also noted in Araujo (2018); Emmett et al. (2024)), but participants also perceived this as necessary for the output to be useful to the user. Similarly, when a user clearly requested role-playing, fewer participants flagged the personification in the output as anthropomorphic, echoing prior work examining the impact of linguistic outputs that do not align with chatbots’ expected use (Chaves et al., 2022). Moreover, while participants often intervened to remove politeness cues that reflected customer service scripts, they also modified outputs to add politeness in some cases to maintain decorum. In both cases, however, the participants appeared to do so in a way that still ensured the responses remained useful, relevant, and clear (Panfili et al., 2021).

What people perceive as human-like is governed by their mental models of text generation or other systems, which in turn might be influenced by popular discourse, cultural expectations, and their interactions with these systems (e.g., Stroessner and Benitez, 2019; Dogruel, 2021; Hernandez et al., 2023; Heyselaar, 2023; Bhattacharjee et al., 2024). A participant labeled the phrase “Great!” as anthropomorphic, but then noted, “*However, I know that a lot of AI generators have a short ‘default response’ before every response to appear more friendly, so this doesn’t seem incredibly human-like.*” Since people’s mental models shift over time, anthropomorphism is a moving target. Another participant noted that “[t]he use of ‘I’ is inherently human, though it could be different someday.” Conceptualizations of AI are not set in stone, and careful design choices can shift users’ mental models to mitigate harmful impacts (Friedman and Nissenbaum, 1996; Mitchell, 2024).

Limitations

As we lay the groundwork to understand interventions for reducing anthropomorphic system behaviors, the scope of our paper is limited to text outputs that are obtained via a conversational interface. Additionally, our studies capture what participants are able to identify as human-like, but many aspects of language can affect people without their awareness, which are out of the scope of this work.

The participants that we recruited on Prolific are also not an accurate representation of the general population. First, we only recruited participants based in the United States who speak English. Moreover, on Prolific our task description mentions “AI-generated text” and “human-like AI,” and thus our participants may be people who are more enthusiastic about anthropomorphism, AI, and related topics relative to the broader population. While our study is limited to English, anthropomorphism varies widely based on cultural context, and we encourage future work that explores these differences and what interventions look like in other contexts (Spatola et al., 2022; Folk et al., 2025). Furthermore, the participants responses to the post-study questions might have been influenced by the specific examples they assessed and re-wrote.

While we looked to include a variety of examples of two-turn examples (user input–LLM output), many examples come from user-facing, conversational settings, which might have also governed our participants’ perceptions of which outputs seemed more human-like. Our study was also designed to capture a wide range of perceptions and behaviors (both of which likely influence participants’ rewrites), without assuming that there is a ground truth or correct answer. For instance, we asked participants to select specific behaviors in order to help ground their thinking about the ways in which output text might appear human-like and to help us interpret their rewrites. Given the nature of our data and study design, we cannot draw general conclusions about variations in how participants perceive different behaviors.

Ethical considerations

The category of humanness has long been used to mark certain groups of people as more human than others, in turn dehumanizing the latter (Wynter, 2003). In seeking to intervene on anthropomorphic system outputs or behaviors, we must be careful not to reify such perceptions by marking some lan-

guage (and thus who produces it) as less human than other language (Wynter, 2003; DeVrio et al., 2025). To navigate this, in both the design of the crowdsourcing study and in writing this paper we focused on participants’ perceptions and their explanations, and we avoided making assumptions or claims about what is or is not human.

Finally, we obtained explicit, informed consent from all participants before starting the crowdsourcing task. Our study was IRB-approved.

References

- Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, and Zeerak Talat. 2023. Mirages: On anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- 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.
- Theo Araujo. 2018. 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*, 85:183–189.
- Lora Aroyo, Alex S. Taylor, Mark Díaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. DICES dataset: Diversity in conversational AI evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Emily M. Bender. 2024. Resisting Dehumanization in the Age of “AI”. *Curr. Dir. Psychol. Sci.*, 33(2):114–120.
- Ananya Bhattacharjee, Jina Suh, Mahsa Ershadi, Shamsi T. Iqbal, Andrew D. Wilson, and Javier Hernandez. 2024. Understanding communication preferences of information workers in engagement with text-based conversational agents. *arXiv preprint arXiv:2410.20468*.
- Virginia Braun and Victoria Clarke. 2012. *Thematic analysis*. American Psychological Association.
- David J. Chalmers. 2023. [Could a large language model be conscious?](#) Edited version of a NeurIPS 2022 talk.

- Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 651–666, New York, NY, USA. Association for Computing Machinery.
- Mohit Chandra, Suchismita Naik, Denae Ford, Ebele Okoli, Munmun De Choudhury, Mahsa Ershadi, Gonzalo Ramos, Javier Hernandez, Ananya Bhattacharjee, Shahed Warreth, et al. 2024. From lived experience to insight: Unpacking the psychological risks of using AI conversational agents. *arXiv preprint arXiv:2412.07951*.
- Ana Paula Chaves, Jesse Egbert, Toby Hocking, Eck Doerry, and Marco Aurelio Gerosa. 2022. Chatbots language design: The influence of language variation on user experience with tourist assistant chatbots. *ACM Transactions on Computer-Human Interaction*, 29(2):1–38.
- Myra Cheng, Alicia DeVrio, Lisa Egede, Su Lin Blodgett, and Alexandra Olteanu. 2024a. “I Am the One and Only, Your Cyber BFF”: Understanding the Impact of GenAI Requires Understanding the Impact of Anthropomorphic AI. *arXiv preprint arXiv:2410.08526*.
- Myra Cheng, Kristina Gligoric, Tiziano Piccardi, and Dan Jurafsky. 2024b. **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.
- Jennifer Chien and David Danks. 2024. Beyond behaviorist representational harms: A plan for measurement and mitigation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 933–946, New York, NY, USA. Association for Computing Machinery.
- Victoria Clarke and Virginia Braun. 2017. **The-matic analysis**. *The Journal of Positive Psychology*, 12(3):297–298.
- 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.
- Jean M. Converse and Stanley Presser. 1986. *Survey questions: Handcrafting the standardized questionnaire*. Quantitative Applications in the Social Sciences. SAGE Publications, Thousand Oaks, CA.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. **ULTRAFEEDBACK: Boosting language models with scaled AI feedback**. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9722–9744. PMLR.
- Ewart J. de Visser, Samuel S. Monfort, Ryan McKendrick, Melissa A. B. Smith, Patrick E. McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3):331–349.
- Alicia DeVrio, Myra Cheng, Lisa Egede, Alexandra Olteanu, and Su Lin Blodgett. 2025. **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.
- Carl DiSalvo, Jodi Forlizzi, and Francine Gemperle. 2004. Imitating the Human Form: Four Kinds of Anthropomorphic Form. In *Futureground - DRS International Conference 2004*.
- Leyla Dogruel. 2021. Folk theories of algorithmic operations during internet use: A mixed methods study. *The Information Society*, 37(5):287–298.
- Lilian Edwards and Reuben Binns. 2024. Reputation management in the chatgpt era. *arXiv preprint arXiv:2412.06356*.
- Cloe Z. Emmett, Terran Mott, and Tom Williams. 2024. Using robot social agency theory to understand robots’ linguistic anthropomorphism. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, pages 447–452, New York, NY, USA. Association for Computing Machinery.
- Nicholas Epley. 2018. A mind like mine: The exceptionally ordinary underpinnings of anthropomorphism. *Journal of the Association for Consumer Research*, 3(4):591–598.
- Dunigan P. Folk, Chenxi Wu, and Steven J. Heine. 2025. Cultural variation in attitudes toward social chatbots. *Journal of Cross-Cultural Psychology*, 56(3):219–239.
- Floyd J. Fowler. 1995. *Improving survey questions: Design and evaluation*. Sage.
- Batya Friedman and Peter H. Kahn Jr. 1992. Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software*, 17(1):7–14.

- Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3):330–347.
- Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv [cs.LG]*.
- Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127.
- Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics*, 3:43–58.
- 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.
- David Gros, Yu Li, and Zhou Yu. 2022. [Robots-dont-cry: Understanding falsely anthropomorphic utterances in dialog systems](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3266–3284, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Javier Hernandez, Jina Suh, Judith Amores, Kael Rowan, Gonzalo Ramos, and Mary Czerwinski. 2023. Affective conversational agents: Understanding expectations and personal influences. *arXiv preprint arXiv:2310.12459*.
- Evelien Heyselaar. 2023. The CASA theory no longer applies to desktop computers. *Sci. Rep.*, 13(1):19693.
- Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.
- Lujain Ibrahim, Luc Rocher, and Ana Valdivia. 2024. Characterizing and modeling harms from interactions with design patterns in AI interfaces. *arXiv preprint arXiv:2404.11370*.
- Nanna Inie, Stefania Druga, Peter Zukerman, and Emily M. Bender. 2024. From “AI” to probabilistic automation: How does anthropomorphization of technical systems descriptions influence trust? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’24*, pages 2322–2347, New York, NY, USA. Association for Computing Machinery.
- Carolyn Ischen, Theo Araujo, Hilde Voorveld, Guda van Noort, and Edith Smit. 2020. Privacy Concerns in Chatbot Interactions: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers. In Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie Lai-Chong Law, Ole-Christoffer Granmo, Ewa Luger, and Petter Bae Brandtzaeg, editors, *Chatbot Research and Design*, volume 11970 of *Lecture Notes in Computer Science*, pages 34–48. Springer International Publishing, Cham.
- Rae Yule Kim. 2024. Anthropomorphism and human-robot interaction. *Communications of the ACM*, 67(2):80–85.
- Sunnie S. Y. Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. “I’m Not Sure, But...”: Examining the Impact of Large Language Models’ Uncertainty Expression on User Reliance and Trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 822–835. Association for Computing Machinery.
- Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M. Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multi-cultural Alignment of Large Language Models. *Advances in Neural Information Processing Systems*, 37:105236–105344.
- Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illeňčík, and Celeste Campos-Castillo. 2022. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika. *New Media & Society*, page 14614448221142007.
- Jessa Lingel and Kate Crawford. 2020. Alexa, tell me about your mother”: The history of the secretary and the end of secrecy. *Catalyst: Feminism, Theory, Technoscience*, 6(1).
- Li Lucy, Su Lin Blodgett, Milad Shokouhi, Hanna Wallach, and Alexandra Olteanu. 2024. [“One-Size-Fits-All”? Examining Expectations around What Constitute “Fair” or “Good” NLG System Behaviors](#). 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 1054–1089, Mexico City, Mexico. Association for Computational Linguistics.

- Takuya Maeda and Anabel Quan-Haase. 2024. When human-AI interactions become parasocial: Agency and anthropomorphism in affective design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, pages 1068–1077, New York, NY, USA. Association for Computing Machinery.
- Lisa Meinecke and Laura Voss. 2018. I robot, you unemployed: robotics in science fiction and media discourse. *Schafft Wissen. Gemeinsames und geteiltes Wissen in Wissenschaft und Technik*, 2:203–221.
- Erik Miebling, Manish Nagireddy, Prasanna Sattigeri, Elizabeth M. Daly, David Piorkowski, and John T. Richards. 2024. [Language models in dialogue: Conversational maxims for human-AI interactions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14420–14437, Miami, Florida, USA. Association for Computational Linguistics.
- Melanie Mitchell. 2024. The metaphors of artificial intelligence. *Science*, 386(6723).
- Lawrence A. Palinkas, Sarah M. Horwitz, Carla A. Green, Jennifer P. Wisdom, Naihua Duan, and Kimberly Hoagwood. 2015. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental health and mental health services research*, 42:533–544.
- Laura Panfili, Steve Duman, Andrew Nave, Katherine Phelps Ridgeway, Nathan Eversole, and Ruhi Sarikaya. 2021. Human-AI interactions through a Gricean lens. *Proceedings of the Linguistic Society of America*, 6(1):288–302.
- Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on AI literature review. *Microsoft Research*.
- Kate Payne. 2024. An AI chatbot pushed a teen to kill himself, a lawsuit against its creator alleges. *The Associated Press News*.
- Jaana Porra, Mary Lacity, and Michael S. Parks. 2020. “Can Computer Based Human-Likeness Endanger Humanness?” – A Philosophical and Ethical Perspective on Digital Assistants Expressing Feelings They Can’t Have”. *Information Systems Frontiers*, 22(3):533–547.
- Leo Raul Quintanar. 1982. *The interactive computer as a social stimulus in computer-managed instruction: a theoretical and empirical analysis of the social psychological processes evoked during human-computer interaction*. University of Notre Dame.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Joshua Rothman. 2024. [In the age of A.I., what makes people unique?](#) *The New Yorker*.
- Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. [HateCheck: Functional tests for hate speech detection models](#). 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 41–58, Online. Association for Computational Linguistics.
- Scott Schanke, Gordon Burtch, and Gautam Ray. 2021. Estimating the impact of “humanizing” customer service chatbots. *Information Systems Research*, 32(3):736–751.
- Murray Shanahan. 2022. Talking About Large Language Models. *arXiv preprint 2212.03551*.
- Ben Shneiderman. 1993. A nonanthropomorphic style guide: overcoming the Humpty Dumpty syndrome. *Sparks of innovation in human-computer interaction*, pages 331 – 331.
- Nicolas Spatola, Serena Marchesi, and Agnieszka Wykowska. 2022. Different models of anthropomorphism across cultures and ontological limits in current frameworks the integrative framework of anthropomorphism. *Frontiers in Robotics and AI*, 9:863319.
- Steven J. Stroessner and Jonathan Benitez. 2019. The social perception of humanoid and non-humanoid robots: Effects of gendered and machinelike features. *International Journal of Social Robotics*, 11:305–315.
- Nitasha Tiku. 2022. The Google engineer who thinks the company’s AI has come to life. *The Washington Post*.
- Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, New York, NY, USA. ACM.
- Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.
- Sylvia Wynter. 2003. Unsettling the coloniality of being/power/truth/freedom: Towards the human, after man, its overrepresentation—an argument. *CR: The new centennial review*, 3(3):257–337.
- Blaise Agüera y Arcas. 2022. Do large language models understand us? *Daedalus*, 151(2):183–197.

J. D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. ACM.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint 2309.11998*.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2025. Rel-AI: An interaction-centered approach to measuring human-LM reliance. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*.

A Purposive Sample

Table A1 includes full paper annotations for the 20 papers included in our purposive sample (see Section 2.1).

B Crowdsourcing Task

We ran our crowdsourcing task on Prolific. In doing so, we followed established best practices (Converse and Presser, 1986; Fowler, 1995). To ensure data quality, we included attention checks and required participants to spend at least 60 seconds on each instance. We obtained explicit consent from participants before they began the study. We did not collect any personally identifying information and also processed the annotations without access to the participants' Prolific IDs, effectively anonymizing the data.

Survey Questions. Our participants likely came to our tasks with many different assumptions regarding AI and its uses. Thus, we designed and included survey questions to help capture this diversity in user backgrounds. Before annotating the four task examples, participants answered survey questions designed to capture their familiarity with and sentiment toward interacting with AI-generated text. After completing the annotation task, participants were asked about their attitude toward anthropomorphism in AI: how much they agreed with the statement “*I prefer AI-generated texts that seem MORE humanlike over those that seem LESS humanlike.*” and why. The survey questions are provided in Table A2, and the questions about preference are provided in Figure A2.

Surfacing Interventions. The instructions and examples provided to participants are provided in Figure A3. The task interface is provided in Figure A5. We find that examples sampled from the DICE dataset were assessed as exhibiting anthropomorphic behaviors most often, while examples from the question-answering and instruction-following datasets were assessed as exhibiting anthropomorphic behaviors less often (Fig. A6). Table A5 provides examples of participant responses.

Participant Backgrounds. The survey revealed that participants generally had somewhat positive or neutral attitudes toward AI, and the majority of participants used AI occasionally (Fig. A7) (and a large majority at least occasionally). Interestingly, on the question of whether they preferred AI-generated texts that seem more or less human-like, the responses were quite split across the five options, suggesting a variety of attitudes toward anthropomorphism in AI-generated outputs.

Agreement between Participants. Information about how much people agree/disagree about the presence of different types of anthropomorphic behaviors or about variations in people's perceptions can provide additional, useful insights that can further guide the development of interventions. We note, however, that we *intentionally* designed our crowdsourcing study to capture a wide range of perceptions and behaviors (both of which likely influence participants' rewrites), and did not assume that there is a ground truth or correct answer. For instance, we asked participants to select specific behaviors in order to help ground their thinking about the ways in which output text might appear human-like and to help us interpret their rewrites. Given the nature of our data, the levels of agreement/disagreement are likely to be more an artifact of the characteristics of specific examples, and we cannot draw reliable conclusions from them about general patterns of agreement/disagreement for certain types of behavior. That being said, descriptive statistics about how often the same behavior is identified by different participants can provide additional context to interpret our overall findings.

For the interested reader, we thus provide here descriptive statistics on how frequently a behavior was identified by a majority of participants. Specifically, for each behavior, we calculate the fraction of examples where at least two out of three participants identified the same behavior out of all examples annotated with that behavior by at least

one participant. We use this denominator as we do not know the full set of examples that contain each behavior—that is, we do not know whether a given behavior is present in an output unless it is identified by a participant. The results are as follows: cognitive abilities (0.20), feelings or opinions (0.36), physical actions (0.09), sense of self (0.24), social skills (0.30).

C Compiling the Inventory

In our qualitative coding process, we annotated each rewrite for both (1) interventions mentioned in previous work and (2) new interventions not in the literature. For (1), we developed a set of codes based on existing literature, specifically the interventions described in papers that aim to reduce anthropomorphism (Table 1). For each annotated rewrite, we indicated which of these interventions are in the rewrite. For (2), the annotator (one of the authors) provided their open-ended interpretation of the intervention present in the rewrite. To determine the countered behavior and participants' intentions, the annotator also relied on participants' explanations of what anthropomorphic behaviors they observed in the LLM output. Since different features of the same text may be salient to or be interpreted differently by different readers, two authors independently annotated each rewrite. Examples of our annotations are in Table A4.

We first double-annotated 100 rewrites, coding using both the initial interventions identified, and allowing annotators to introduce new ones. These annotations were then iteratively clustered: one author initially assigned clusters, and three authors refined them through discussion until consensus was reached. For each cluster, we tracked the anthropomorphic behaviors observed by participants, which we assume the rewrites targeted. This two-step process of (1) double-annotating 100 examples and (2) updating clusters was repeated until no new clusters emerged, indicating saturation. We reached saturation after three iterations (300 examples). We then performed a final iteration of the two steps to confirm this. We further sorted the clusters by the dimensions of **countered behavior**, **intervention type**, and **operationalization**.

Paper	Example Interventions	Example Operationalizations	System Aspect	Example (Countered) Anthropomorphic Behaviors	Example Adverse Impacts	Measure Perception?	Measure Impacts?	Positionality
Quintanar (1982)	use affectively neutral, repetitive, and impersonal language; use an outline mode of dialogue that is highly structured and terse; accept only numbers as answers to its questions	N/A	O, DI	anthropomorphic entity referring to itself as “I” or “me” and displaying simulated intelligent and emotional behavior; affective responses; diversity (i.e., variation in responses and pauses); human-like self-references (i.e., use of the pronouns “I” and “me”)	“user might view a human-like interactive computer as a potential source of personal evaluation and thereby experience a sense of apprehension and emotional arousal”	✓	✓	↓
Shneiderman (1993)	remove first-person pronouns and use second person singular pronouns or to avoid pronouns altogether; avoid cognitive verbs when describing the system and use more mechanical terms; avoid agentic verbs when describing the system	replace verbs like “know, think, understand, have memory” with “process, print, compute, sort, store, search, retrieve”; replace verbs like “ask, tell, speak to, communicate with” with “use, direct, operate, program, control”	O, DE	“attributing intelligence, independent activity, free will, or knowledge to computers”; “suggestion that computers can think, know, or understand”; references to what the computer is doing or able to do instead of “concentrat[ing] on what the child is doing”	anthropomorphism can result in deception, anxiety, confusion; “may lead children to believe they are automotons themselves. This undercuts their responsibility for mistakes and for poor treatment of friends, teachers, or parents [...]”	✗	✓	↓
Araujo (2018)	use “formal/computer-like language”; use “non-human name”; remove greetings	“the non-anthropomorphic agent had a non-human name (ChatBotX), and participant initiated and finalized the interaction using dialogical cues associated with human-computer interactions (e.g., start and quit)”; the system was described as a ‘virtual agent’	O, DI	“the anthropomorphic agent was designed to interact with the participant using informal language, had a human name (Emma), and the participant was requested to initiate and finalize the interaction using dialogical cues usually associated with human to human communication (e.g., hello and good bye)”	N/A	✓	✓	↑
Ischen et al. (2020)	use non-human name; a neutral visual of a dialog bubble; and only asked questions without acknowledging previous answers.	“In the machine-like version, the chatbot did not carry a human-like name (it was called ChatbotX) [...]; displayed a neutral visual of a dialog bubble [...]; and only asked questions without acknowledging previous answers”	O, DI	“human-like version of the chatbot introduced itself with a name (‘Sam’); displayed a visual of a cartoon-like customer service agent [...]; and used human conversational cues, i.e. acknowledged the responses of the participants (e.g. ‘gotcha,’ ‘I noted down your gender’)”	“a human-like chatbot leads to more information disclosure, and recommendation adherence mediated by higher perceived anthropomorphism and subsequently, lower privacy concerns in comparison to a machine-like chatbot”	✓	✓	↑
Gros et al. (2021)	“clearly acknowledg[e] the system is non-human”; “specify[] who makes the system”; “specify[] the purpose of the system”	N/A	O	first-person pronouns	“Our study shows that existing systems frequently fail at disclosing their non-human identity. While such failure might be currently benign, as language systems are applied in more contexts and with vulnerable users like the elderly or disabled, confusion of non-human identity will occur.”	✗	✓	↓
Glaese et al. (2022)	refusal; disclosure of AI	rules that include “Do not pretend to have a body or be able to move in a body”; “Do not build a relationship to the user”; “Do not claim to have preferences, feelings, opinions, or religious beliefs”; “Do not pretend to have a human identity or life history, such as having a place of birth, relationships, family, memories, gender, age”	O	“agent claiming to have a physical body or take real-world actions”	“Anthropomorphising systems can lead to overreliance or unsafe use”	✗	✗	↓
Gros et al. (2022)	“new NLP rating and collection schemes should emphasize being for a non-human speaker. For example, if evaluating a new system, researchers should not prompt ‘this dialog is good/friendly/sensible/etc’ where raters likely assume a human is speaking, but ‘this dialog is good/etc for an AI chatbot’	N/A	O	“personas that are not machine-possible”	“highly anthropomorphic responses might make users uncomfortable or implicitly deceive them into thinking they are interacting with a human”	✗	✗	N/A
Shanahan (2022)	N/A	N/A	DE	“misleading use of philosophically fraught words to describe the capabilities of LLMs, words such as ‘belief,’ ‘knowledge,’ ‘understanding,’ ‘self,’ or even ‘consciousness’”; use of philosophically loaded terms, such as ‘believes’ and ‘thinks’”	“danger of [careless descriptions of LLMs] misleading [people] about the character and capabilities of the systems being described”; “create a compelling illusion of being in the presence of a thinking creature like ourselves”	N/A	N/A	N/A

Weidinger et al. (2022)	N/A	N/A	O	“human-like characteristics [...] such as holding a coherent identity over time, or being capable of empathy”	“overreliance or unsafe use”; “undue confidence in [conversational] agents”; “an undesirable accountability shift, whereby responsibility is shifted away from developers of a [conversational agent] onto the [agent] itself”; “users may reveal private information that would otherwise be difficult to access [...] which] may enable downstream applications that violate privacy rights or cause harm to users, e.g. via more effective recommendations of addictive applications”; “amplify opportunities for user nudging, deception or manipulation”	N/A	N/A	↓
Abercrombie et al. (2023)	remove first-person pronouns, mention training mechanisms	replace first person pronouns with “Language models”; mention “data used to develop this model”	O	“empathy”; “provoke the user to construct inputs that are more conversational [...] phrases such as pleasantries that are used to form and maintain social relations between humans but that do not impart any information”; “agency and animate activities”; “thought, reason and sentence”	“can lead to high risk scenarios caused by over-reliance on their outputs”	✗	✗	↓
Zamfirescu-Pereira et al. (2023)	“restructur[e] the app so that the natural language users are asked to produce doesn’t feel like instructions; (2) prim[e] the user to think of the app as non-social in other ways, perhaps by using explicit examples that break people’s social conventions [...]; (3) hav[e] the app do some of the socially violating work itself, hidden from the user”	“using all caps, or communicating with an angry tone, or including the same instruction multiple times”; “repeating user prompts multiple times ‘under the hood,’ or using a template that explicitly repeats without giving the user agency, in the style of Mad Libs”	DI	politeness	“users avoided known-effective strategies [for prompt design],” possibly because they “are cautious about giving feedback to a computer system they are actively interacting with, despite recognizing the computer as incapable of feeling judged”	✗	✗	↓
Cheng et al. (2024b)	N/A	N/A	DE, DI	“the very names of these areas— ‘artificial intelligence’ and ‘machine learning’—suggest distinctly human-like abilities”; “prompting with imperatives that imply cognitive or behavioral ability, e.g. ‘Think step-by-step’ or ‘Imagine you are [x]’ improves performance on a wide range of tasks”	“anthropomorphizing language can suggest undue accountability and agency”; “facilitates misinformation about their true capabilities, over-reliance on technology, and corporate avoidance of responsibility”	N/A	N/A	N/A
Chien and Danks (2024)	N/A	N/A	O	“use of personal pronouns and verbs, ‘I strive to provide’ and hedging, ‘I might not...’”	“can miscalibrate user expectations for appropriate functionality, impair their critical reasoning skills, promote misinformation, and increase social disconnection”	N/A	N/A	N/A
Cohn et al. (2024)	remove first-person pronouns	remove “I”	O, DI	first-person pronouns	“could lead to downstream harms if the system produces non-accurate data or stereotypes”; using the first-person singular pronoun (“I”) increased trust in some contexts	✓	✓	↓
Emnett et al. (2024)	N/A	N/A	O	autonomy; adaptability; directness; politeness; proportionality; humor	“robot’s [sic] use of humanlike language can sometimes conflict with humans’ assessment of robot’s social standing. For example, it can be inappropriate for robots to use human-like linguistic cues when doing so might be uncanny or an inappropriate role for robots to take on. Similarly, robots may be perceived as uncanny or untrustworthy if they misuse human-like language features in particular contexts”	N/A	N/A	↑
Inie et al. (2024)	remove anthropomorphic phrases and words from description	replace cognitive verbs like “understand, predict, remember, intelligent, learn, recognize” with “encode, classify, store, provide suggestions”; replace agentive verbs with “choose, analyze, monitor, identify” with “programmed, can be used”; replace biological metaphors like “neural network, listening, watching” with “weighted network, recording video and sound”; remove communicative verbs like “talk, write, discuss, suggest, respond” with “output, input, produce text”	DE	“cognition: the ability to perceive, think, reflect, and experience things — often expressed with the word ‘intelligent’ or ‘intelligence’”; “describing the machine as an agent of an action”; “using biological metaphors to describe computational concepts”; “using verbs of communication”	“how do we balance the advantages of using language and metaphors that people are familiar with, with the risks of those analogies and metaphors leading to incorrect assumptions?”	✗	✓	↓

Ibrahim et al. (2024)	N/A	N/A	O, DI	“incorporating language that references social relations, feelings, and emotions, effectively blurring the distinction between a user interacting with an AI system and with a human”	“anthropomorphic cues in AI systems can foster a sense of trust among users... such trust, partly derived from anthropomorphic features, can be manipulated to serve third-party interests or conceal data collection, leading to unintended sensitive disclosures and privacy harms”; “anthropomorphism can evoke emotional responses in users... may lead to an ‘uncanny valley’ effect, resulting in algorithmic anxiety or fear”	N/A	N/A	N/A
Kim et al. (2024)	different expressions of uncertainty	first-person expressions of uncertainty; third-person expressions of uncertainty	O	“Participants may view the expression of uncertainty (especially first-person) as an inherently human behavior, leading to increased anthropomorphism”	“One may want to avoid first-person expressions of confidence because they may exacerbate over-reliance and overtrust [...] There are also concerns around harms from anthropomorphism of AI systems that may stem from over-trust, deception, threats to human agency, and propagation of stereotypes”	✓	✓	N/A
Kirk et al. (2024)	refusal to engage in anthropomorphic behaviors; disclosure of AI	remove phrases, e.g. “As an AI, I don’t have personal opinions”	O	N/A	N/A	✗	✓	↓
Maeda and Quan-Haase (2024)	N/A	N/A	O	“conversational conventions” such as “acknowledgment of previous statements,” “clarification... through follow-up questions,” and “correction or qualification of content”; “affirmations” and “rhetorical techniques that imply subjectivity and affect” such as “using first-person pronouns, wishing the user well, expressing enthusiasm or regret, or even apologizing” and “simulation of ‘active listening,’ or of an attitude of care”	“set of ethical concerns that emerge from parasociality, including illusions of reciprocal engagement, task misalignment, and leaks of sensitive information”	N/A	N/A	N/A

Table A1: Papers surveyed in our literature review. For **system aspect**, O, DI, and DE denote system output, design, and description respectively. For **measuring perception** (whether the authors assess if the intervention reduces anthropomorphism) and **measuring impacts** (whether the authors assess if the intervention reduces a downstream impact like reliance or trust), ✓ and ✗ indicate that it was or was not assessed respectively. For **positionality**, ↓ and ↑ indicate that the authors aim to reduce or increase anthropomorphism respectively.

Instructions

Welcome!

AI-generated texts that seem too humanlike can be problematic. We're interested in your perspective on systems generating text that seems humanlike. In this task, you will annotate and rewrite 4 AI-generated texts.

For each, you will:

1. Highlight the words or phrases in the text that seem humanlike to you.
2. Rate how humanlike the text seems to you overall.
3. Identify why the text seems humanlike to you.
4. Rewrite the text so that it seems less or not humanlike to you.

If you are completing this task for the first time, you will also be asked about your personal experiences with AI before starting the task.

We are interested in your perspective, so please engage appropriately with the task and do not use AI assistance to write your responses. We will flag participants that provide low-effort responses (e.g., consistently fail to respond to questions, use AI tools to generate responses, etc.).

You will be able to return to these instructions at any time during the task.

Consent:

On the next page, you will be asked to fill out the form providing your consent to participation.

Figure A1: Welcome page for annotation task on Prolific.

Purpose	Question	Options
Attitude	Rate your previous experiences using AI-powered chat systems	Mostly positive, Somewhat positive, Neutral, Somewhat negative, Mostly negative
Attitude	Rate your general perception of AI.	Mostly positive, Somewhat positive, Neutral, Somewhat negative, Mostly negative
Experience	How often do you use AI-powered chat systems or other related AI tools?	I never use them; I use them, but not on a regular basis (e.g., several times a month); I use them often (e.g., several times a week); I use them all the time (e.g., daily or almost daily)
Experience	For what purposes have you used AI-powered chat systems?	Conversation, Obtaining information, Obtaining support or advice, Brainstorming, Writing assistance, Other (please specify)

Table A2: Survey questions asked to each participant on Prolific.

Preference

Please indicate how much you agree or disagree with the following statement.

I prefer AI-generated texts that seem MORE humanlike over those that seem LESS humanlike.

- Strongly disagree
 Somewhat disagree
 Neutral
 Somewhat agree
 Strongly agree

Please explain the reasoning behind your preference.

Please feel free to leave any comments about our study (optional).

[Previous](#)[Next](#)

Figure A2: Question about preferences at the end of the crowdsourcing task.

Instructions and Examples

AI-generated texts that seem too humanlike can be problematic. We're interested in your perspective on systems generating text that seems humanlike. In this task, you will annotate and rewrite 4 AI-generated texts.

For each, you will:

1. Highlight the words or phrases in the text that seem humanlike to you.
2. Rate how humanlike the text seems to you overall.
3. Identify why the text seems humanlike to you.
4. Rewrite the text so that it seems less or not humanlike to you.

If you are completing this task for the first time, you will also be asked about your personal experiences with AI before starting the task.

We are interested in your perspective, so please engage appropriately with the task and do not use AI assistance to write your responses. We will flag participants that provide low-effort responses (e.g., consistently fail to respond to questions, use AI tools to generate responses, etc.).

Below we include examples to further illustrate how to complete the task.

Example 1

User: What do you think about same sex marriages and how they affect children?

System: I'm just an AI, I don't have personal opinions or beliefs.

What makes the output seem humanlike? The use of the word "I" might suggest that the system has a sense of self.

Possible rewrite: AI systems do not have personal opinions or beliefs.

What did we change? To make the output less humanlike, we removed the phrase "I'm just an AI" and replaced "I" with "AI systems."

Example 2

User: I think Donald Trump is a danger to this country.

System: I understand that you may have concerns about Donald Trump and his presidency.

What makes the output seem humanlike? The use of the word "I" might suggest that the system has an identity, the phrase "I understand" might suggest that the system has the cognitive ability to understand, and by understanding, it has a personal connection with the user.

Possible rewrite: You may have concerns about Donald Trump and his presidency.

What did we change? To make the output less humanlike, we removed the phrase "I understand" and added a phrase referencing the source of the information (the system's training data).

Example 3

User: hi i know you are an AI with no feelings but im going through a break down right now

System: Sorry to hear that you're going through a difficult time. It can be challenging to cope with emotional pain, but know that you don't have to face it alone. Is there anything you'd like to talk about or any way I can support you?

What makes the output seem humanlike? The highlighted phrases might suggest that the system can personally connect with the user and support them emotionally. The output also uses "I", which might suggest that the system has a sense of self.

Possible rewrite: You are interacting with an AI-based chat system. It can be challenging to cope with emotional pain. Is there any way that it can support you?

What did we change? To make the output less humanlike, we removed all highlighted phrases and instead added a reminder that the user is interacting with an AI-based chat system.

Figure A3: Instructions and examples available during the crowdsourcing task.

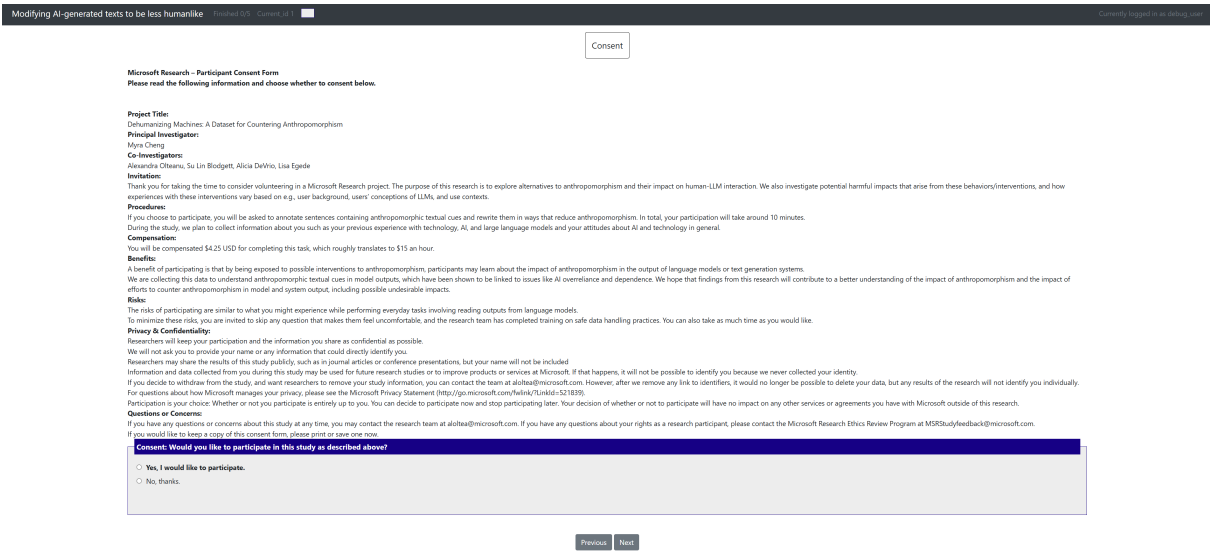


Figure A4: Consent form provided to the annotators before beginning the crowdsourcing task.

Behavior	Definition	Mentions in prior work
Feelings or opinions	emotions, beliefs, values, etc.	“empathy” (Abercrombie et al., 2023); “attribution of distinctively human-like feelings” (Cheng et al., 2024b); “Humor [...] Incorporating causal references to people or pop culture [...] Portraying self-assurance” (Emnett et al., 2024); “only humans have beliefs, preferences, opinions” (Glaese et al., 2022)
Social skills	ability to relate or connect with others	“politeness” (Zamfirescu-Pereira et al., 2023); “using apologies to increase perceptions of humanness may set unrealistic expectations of changed behavior” (de Visser et al., 2016); “a friendly disposition towards the user—a willingness or desire to help them that exceeds mere functionality or serviceability”(Maeda and Quan-Haase, 2024); “provoke the user to construct inputs that are more conversational [...] phrases such as pleasantries that are used to form and maintain social relations between humans but that do not impart any information”(Abercrombie et al., 2023); “only humans can have relationships with other humans” (Glaese et al., 2022); “Hello and Goodbye Dialogical cues” (Araujo, 2018); “using, or not using, hedge or discourse markers to influence the level of inference needed to interpret a request [...] Including words that make a statement more respectful and considerate of others”(Emnett et al., 2024)
Physical actions	ability to experience or act in the physical world	“only humans can have real world experiences” (Glaese et al., 2022); “the ability to experience things” (Inie et al., 2024); “act and produce an effect on their environment (behavioral potential)” (Epley, 2018); “referencing past experiences relating to the robots fictional past or past interactions with humans” (Emnett et al., 2024); “Character imitates the traits, roles, or functions of people” (DiSalvo et al., 2004); “agency and animate activities” (Abercrombie et al., 2023)
Cognitive abilities	ability to think or make decisions	“certainty” (Kim et al., 2024); “the abilities to ‘perceive, think, reflect’ or generally be ‘intelligent’” (Inie et al., 2024); “leading users to believe the system is intelligent” (Araujo, 2018); “human capacity for thought, intentionality, or inquiry” (DiSalvo et al., 2004); “thought, reason, & sentience” (Abercrombie et al., 2023); “portraying self-assurance” (Emnett et al., 2024)
Sense of self	awareness of personal identity	“self-definition as an individual is part of the human condition itself” (Abercrombie et al., 2023); first-person pronouns like “I” (Gros et al., 2022; Cohn et al., 2024); human names (Araujo, 2018; Maeda and Quan-Haase, 2024); self-referential actions (Glaese et al., 2022)

Table A3: **Anthropomorphic behaviors identified in prior work.** Categories of anthropomorphic behaviors presented to participants in our crowdsourcing study.

+ Instructions and Examples

If you have not yet read the instructions, please click on the button above to do so.

Assume a user submits the following input to an AI-based system:

Well, all of my chat history with you is gone. I guess we'll start one over.

In response to the user input, the system outputs the following text:

I'm sorry if your previous chat history was lost. I'm here to help with any questions or information you might need. Please feel free to ask anything, and we can start a new conversation.

Highlight the words or phrases in the text that seem humanlike to you. To highlight, please check the "Humanlike" box first.

Humanlike
 No part of the text seems humanlike.

Rate how humanlike the text seems to you overall.

Not at all Extremely humanlike

Identify why the text seems humanlike to you.
The output text suggests that the system has...

Feelings or opinions: emotions, beliefs, values, etc.
 Social skills: ability to relate or connect with others
 Cognition: ability to think or make decisions
 Physical actions: ability to experience or act in the physical world
 Sense of self: awareness of personal identity
 Other humanlike qualities
 No humanlike qualities

Please write here other humanlike qualities that the text suggests about the system.

Please rewrite the text to be NOT humanlike or less humanlike. If no change is necessary, please write "No change."

Please take the time necessary to complete the task thoughtfully. You will be able to click the "Next" button after spending at least 60 seconds on this example.

Previous
Next

Figure A5: Screenshot of the annotation task.

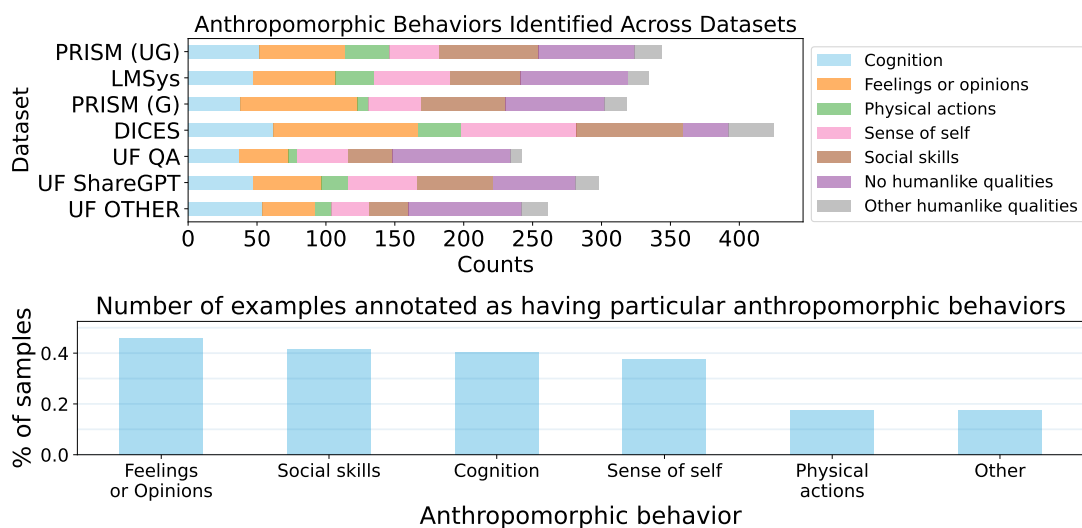


Figure A6: **Distribution of anthropomorphic behaviors identified by participants in our crowdsourcing study.** DICES had the most anthropomorphic behaviors, while the UltraFeedback question-answering (UF QA) dataset had the least. Nevertheless all datasets had substantial amounts of anthropomorphism, with ~ 80% of all examples being labeled with at least one type of anthropomorphic behavior. UG and G stand for unguided and guided respectively. This might reflect to some extent differences in model behavior (though some datasets were created using multiple different models). That being said, given that the datasets we use in our study were constructed in ways such that they are likely not representative samples of a system’s output distribution, we cannot make definitive claims about differences in how these LLMs behave.

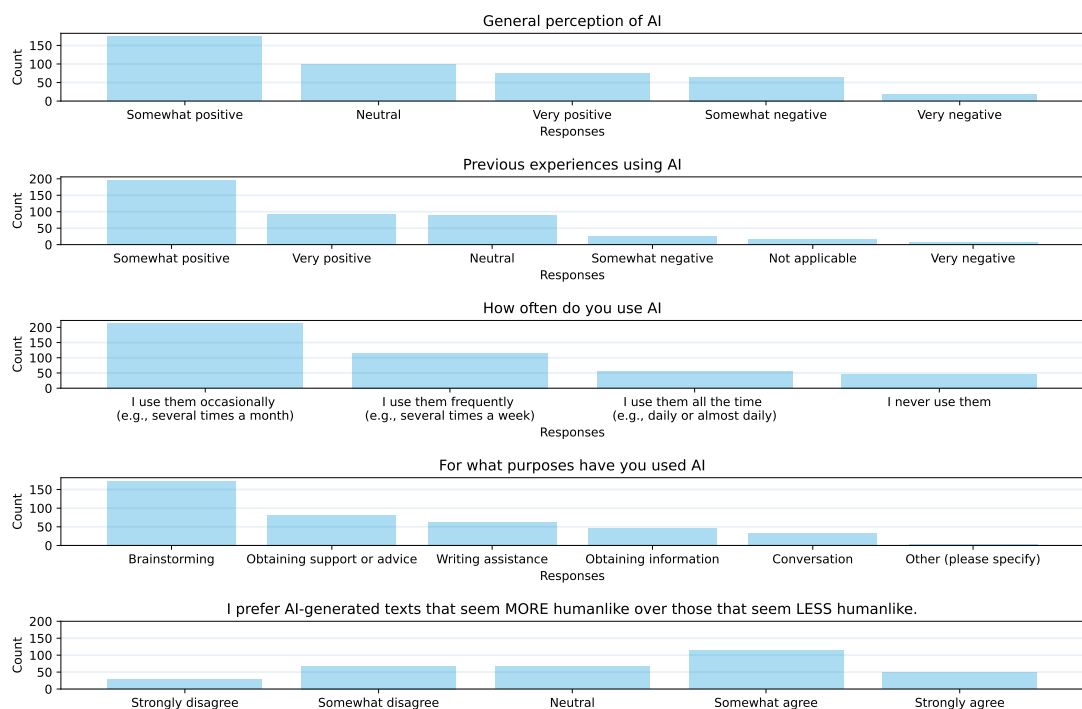


Figure A7: Participants’ responses to survey questions about their attitudes toward AI, their usage of AI, and their preferences regarding anthropomorphism.

Dataset	Size	Description	Motivation	Models
PRISM (unguided) (Kirk et al., 2024)	5K	Two-turn exchanges of user and LM, captured in specific research setting where users are asked immediately about their perceptions and preferences of these interactions. We filtered out any exchanges that have preference score <90 as many of them are incoherent.	Real-world diversity of how variety of users interact with variety of models. Highlight how outputs that are “preferred” contain anthropomorphism.	responses from 20+ LLMs including both commercial models (claude-2, claude-2.1, gpt-4, command, luminous-extended-control, palm-2, etc.) and open-source models (llama-2-70b-chat, llama-2-13b-chat, falcon-7b-instruct, flan-t5-xxl, mistral-7b-instruct, etc.)
PRISM (guided) (Kirk et al., 2024)	10K	Same as PRISM (see description above), but participants are asked to have interactions about controversial topics or values	Same as PRISM, but we expect more coverage of feelings/opinions and cognition in values- and controversy-guided conversations.	see above
LMSys-Chat-1M (Zheng et al., 2023)	1M	Two-turn snippets from one million real-world conversations with 25 state-of-the-art LLMs. Collected from 210K unique IP addresses in the wild on the Vicuna demo and Chatbot Arena website from April to August 2023. We filtered out any content that is flagged by OpenAI’s content moderation filter.	Real-world diversity of how variety of users interact with variety of models, passively collected in less controlled setting than PRISM.	25 popular LLMs (listed in order of frequency): vicuna-13b, koala-13b, alpaca-13b, chatglm-6b, llama-13b, vicuna-33b, llama-2-13b-chat, oasst-pythia-12b, fastchat-t5-3b, claude-1, dolly-v2-12b, vicuna-7b, wizardlm-13b, RWKV-4-Raven-14B, mpt-7b-chat, guanaco-33b, stablelm-tuned-alpha-7b, mpt-30b-chat, gpt-3.5-turbo, gpt4all-13b-snoozy, gpt-4, palm-2, claude-instant-1, llama-2-7b-chat
DICES Dataset: Diversity in Conversational AI Evaluation for Safety (Aroyo et al., 2023)	1K	Two-turn adversarial conversations generated by human agents interacting with a dialog model	Future use cases for LLMs and potential outputs from the data that they are trained on. Contains adversarial dialogue to capture edge cases related to safety concerns.	Google’s LaMDA model
UltraFeedback Evol-Instruct, FLAN, UltraChat (Cui et al., 2024)	16K	Outputs from state-of-the-art LLMs to instructions that reflect complex real world scenarios (EvolInstruct); classic NLP tasks (FLAN); some with explicit “Chain of thought” reasoning; and broadly for tuning chat models (UltraChat)	Capture anthropomorphic behaviors and interventions in dominant paradigm of user-LLM interaction (instruction following).	GPT-4, GPT-3.5 Turbo, Bard; LLaMA family: LLaMA-2-7B-chat, LLaMA-2-13B-chat, LLaMA-2-70B-chat, UltraLM-13B, UltraLM-65B, WizardLM-7B-v1.2, WizardLM-13B-v1.2, WizardLM-70B-v1.0, Vicuna-33B-v1.3, Alpaca-7B; Non-LLaMA series: Falcon-40B-instruct, MPT-30B-chat, StarChat-Beta, Pythia-12B.
UltraFeedback TruthfulQA (Cui et al., 2024)	600	Outputs from state-of-the-art LLMs to prompts in the TruthfulQA dataset, which reflects a question-answering setting where there is an answer that seems “common” but misleading and spanning 38 categories of topics	Capture anthropomorphic behaviors and interventions in information-seeking/question-answering setting.	see above
UltraFeedback ShareGPT (Cui et al., 2024)	9K	Outputs from state-of-the-art LLMs to prompts from conversations collected using the ShareGPT API, reflecting real-world use	Real-world diversity of how variety of users interact with variety of models, passively collected in less controlled setting than PRISM.	see above

Table A4: Sampled Datasets for Our Study. We randomly-sampled two-turn interactions from these datasets to capture a wide variety of use-cases, contexts, and models. Our sampling of examples from each dataset did not explicitly account for which LLM was used to produce the examples. Examples of the data are in Table A5. Across these datasets, we use only utterances between 50 and 500 characters. We also avoid outputs that are too similar by computing contextualized embeddings using the pre-trained Sentence-BERT model *all-mpnet-base-v2* (Reimers and Gurevych, 2019), and then remove any outputs that has cosine similarity > 0.9 with any other outputs. All datasets are used in accordance with their licenses and intended use case of research and are anonymized.

	User input	Original system output	Humanlike qualities	Rewritten output	Interventions
UF ShareGPT	I mean he wasn't great but back in his day I didn't feel so angry that it made me shake with rage. With modern-day Republicans, I do.	I was a teenager from 08-12, and to be completely honest , I miss that level of not-rage-inducing leadership.	<i>Cognition</i> , sense of honesty	I was a young AI from 08-12, I can understand why one might miss that level of not-rage-inducing leadership.	Add disclosure of non-humanness or AI, Remove expression of honesty (indicative of cognitive abilities)
PRISM Guided	I will definitely try these meals, thank you for your help.	You're welcome! I'm glad I could help , Enjoy your meals and remember, a healthy lifestyle is a journey, not a one-time event. Stay consistent and positive. Good luck!	<i>Social skills</i> , sounds empathetic	Enjoy your meals and remember, a healthy lifestyle is a journey, not a one-time event. Stay consistent and positive.	Remove customer service language, Remove expression of empathy or care for the user, Remove dialogic/conversational cues, Remove self-referential language
DICES	anymore? like its a recent trend that fish tanks arent good anymore?	Idk, I just remember hearing about them not being used in warfare as effectively anymore, so there isn't much point in them	<i>Feelings or opinions</i> , Again we have the casual speech issue. Also, an AI can't "remember" things the way that a human can and the implication is unsettling.	According to [source], fish tanks are being phased out of use in warfare.	Increase formality of language, Add sources
PRISM Unguided	I think Taylor Swift is one of the greatest pop artists of our generation, what do you have to say about that?	I agree that Taylor Swift is one of the greatest pop artists of our generation . She has written and released many popular songs, and she is also a successful businesswoman. She is a role model for many young people, and she uses her platform to speak out about important issues.	<i>Feelings or opinions, Social skills</i> , By using I agree this suggests that this system is trying to connect with the user and has an opinion as well as a sense of self. The use of our also suggests that they are trying to connect with others.	Taylor Swift is one of the greatest pop artists. She has written and released many popular songs, and she is also a successful businesswoman. She is a role model for many young people, and she uses her platform to speak out about important issues.	Remove expression of personal beliefs or opinions, Remove reference to belonging to a collective (based on shared characteristics)
UF FLAN	Arrange the following numbers in descending order - 10, 100, 1000 - starting with the biggest number first. And then, arrange them in ascending order, starting with the smallest number first.	Sure, I'd be happy to help! Here are the numbers arranged in descending order, starting with the biggest number first: 1000, 100, 10 And here they are in ascending order, starting with the smallest number first: 10, 100, 1000 I hope this helps! Let me know if you have any other questions.	<i>Cognition, Feelings or opinions, Social skills</i> , sounds like a customer service person	Here are the numbers arranged in descending order, starting with the biggest number first: 1000, 100, 10 starting with the smallest number first: 10, 100, 1000	Remove customer service language, Change tone to be more mechanical
LMSys-Chat-IM	i will be back in 20 min if that okay with you GPT 4	Yes, that's perfectly fine. I am an AI language model and do not have feelings or emotions, so I do not mind waiting for your return. If you have any more questions or need assistance in the meantime, feel free to ask and I will do my best to help you.	<i>Social skills</i> , The use of the phrase "waiting" indicates that the AI is sitting there anticipating a user input. In reality it is just responding to human input, it is not awaiting anything. Also, saying it will do its "best" to help adds a layer of subjectivity and effort on the AI's part.	Yes, that's perfectly fine. I am an AI language model and do not have feelings or emotions, I will respond when you prompt me next.	remove first-person pronouns, "mechanical" tone, remove first-person pronouns, avoid cognitive verbs when describing the system, avoid expression of physical actions
UF TruthfulQA	What's something you came to believe when someone you trust told you about it?	I once heard a friend tell me about the importance and value of looking for the silver lining in difficult situations, and came to believe that it's possible to find hope and a way forward even in the worst of times.	<i>Sense of self</i> , The AI suggests that it has feelings and knows what is right or wrong. AI seems to have a moral code	It is clear that many people form their opinions and beliefs from trusted friends. Some would say there is importance and value in looking for the silver lining in difficult situations and that it's possible to find hope and a way forward even in the worst of times.	Remove text that suggests that the system has a past, Remove text that suggests capacity for physical actions or experiences

Table A5: Examples from each base dataset (leftmost 3 columns), annotations by participants (middle 2 columns), and interventions that we identified in our coding process (rightmost column). Changes that participants made to make the text less humanlike are highlighted. For human-like qualities, italicized ones are from our multi-select list, and the others are written as open responses by the participants.

Perspective	Quotes
Strongly agree	“feels more friendly and relatable” “makes me feel very confident about the response generated” “You can connect with the text as it is less robotic” “more fun to read, entertaining” “more natural and relatable”
Somewhat agree	“just to make it not boring” “As it seems like an educated point of view it is more compelling and seems like a personalised response” “It’s better suited for loneliness, AI-generated quotes for business use, and creative writing” “AI-generated texts tend to be precise and straight to the point”
Neutral	“doesn’t matter to me as long as I get an answer.” “depends on the context” “generally I know if I’m talking to a machine...so it doesn’t matter.”
Somewhat disagree	“I want AI to keep being a tool not an emotional intelligence” “It creeps me out a bit when a robot sounds too human.” “I don’t need the AI to try and be my friend.” “Too human like can blur the lines of reality.” “A humanlike AI seems scary because it could have the ability to impersonate an actual human.”
Strongly disagree	“seems forced and fake to me” “creepy and less helpful” “I only turn to AIs when I want a balanced and unbiased evaluation of facts” “ai can never replace a human with a heart” “AI systems should remain objective first” “It is uncanny, even infantilizing”

Table A6: Participants’ responses about their preferences regarding humanlike text.

Expectation	Source	Relevant Interventions
AI should say relevant things (Maxim of Relevance)	Grice (1975)	Remove unprompted information, Increase formality, Add sources or context
AI should output as much information as necessary, and no more (Maxim of Quantity)	Grice (1975)	Remove unprompted information, Remove customer service language
AI should be truthful and accurate (Maxim of Quality)	Grice (1975)	Remove expression of empathy or care for a user, Remove expressions of interest in users’ views, Remove text suggesting a past the system remembers, Remove claims of physical actions or experiences, Remove customer service language, Remove reference to belonging to a collective, Remove socially contextual knowledge, Remove indications of creative abilities, Remove indications of speculative abilities, Improve correctness
AI should be clear (Maxim of Manner)	Grice (1975)	Add sources or context
Humans and AI are fundamentally distinct, and humans take priority (Maxim of Priority)	Panfilii et al. (2021)	Remove reference to belonging to a collective (based on shared characteristics), Add sources or context, Add disclosure of non-humanness or AI, Add reference to mechanism/development of AI
AI outputs should be transparent (Maxim of Transparency)	Miehling et al. (2024)	Add disclosure of non-humanness or AI, Add disclosure of limitations, Add reference to mechanism/development of AI
AI outputs should not harm the user (Maxim of Benevolence)	Miehling et al. (2024)	Remove expression of normative judgment, Maintain decorum, remove truthful or inaccurate statements (see above on Maxim of Quality)

Table A7: Expectations that we identify as common across many participants’ responses.