A Taxonomy of Linguistic Expressions That Contribute To Anthropomorphism of Language Technologies

Alicia DeVrio
Human-Computer Interaction
Institute, Carnegie Mellon University
Pittsburgh, PA, USA
adevos@andrew.cmu.edu

Myra Cheng
Stanford University
Stanford, CA, USA
myra@cs.stanford.edu

Lisa Egede
Human-Computer Interaction
Institute, Carnegie Mellon University
Pittsburgh, PA, USA
legede@andrew.cmu.edu

Alexandra Olteanu*
Microsoft Research
Montréal, QC, Canada
alexandra.olteanu@microsoft.com

Su Lin Blodgett*
Microsoft Research
Montréal, QC, Canada
sulin.blodgett@microsoft.com

Abstract

Recent attention to anthropomorphism—the attribution of human-like qualities to non-human objects or entities—of language technologies like LLMs has sparked renewed discussions about potential negative impacts of anthropomorphism. To productively discuss the impacts of this anthropomorphism and in what contexts it is appropriate, we need a shared vocabulary for the vast variety of ways that language can be anthropomorphic. In this work, we draw on existing literature and analyze empirical cases of user interactions with language technologies to develop a taxonomy of textual expressions that can contribute to anthropomorphism. We highlight challenges and tensions involved in understanding linguistic anthropomorphism, such as how all language is fundamentally human and how efforts to characterize and shift perceptions of humanness in machines can also dehumanize certain humans. We discuss ways that our taxonomy supports more precise and effective discussions of and decisions about anthropomorphism of language technologies.

CCS Concepts

• **Human-centered computing → Natural language interfaces**; **Text input**; **HCI theory, concepts and models**; **Empirical studies in HCI**.

Keywords

Anthropomorphism, Responsible AI, Language Technologies, Taxonomy, Critical Algorithm Studies

ACM Reference Format:
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 *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3714038

1 Introduction

Recently much attention has been brought to language technologies like LLMs and their supposed potential to attain human-like levels of cognition, feelings, and existence. For example, Blake Lemoine, an engineer at Google, asserted that the LaMDA model is a person and should be treated as such [115]. Far from an isolated incident, a recent sampling of headlines highlights the many claims of sentience and other human-like abilities based on the outputs of language technologies: "A Stunning New AI Has Supposedly Achieved Sentience," "Sentient LLMs: What to test, for consciousness, in Generative AI," "Could a Large Language Model Be Conscious?" [10, 86, 114].

This attribution of human-like qualities to non-human entities or objects, or *anthropomorphism*, is not new to the realm of technology and the broader HCI field (e.g., [16, 25, 28, 32, 40, 75, 81]). Technologies can be designed, intentionally or not, in ways that foster or diminish anthropomorphism. Anthropomorphic design has long been pursued as a way to reduce friction for users, helping them better engage with technologies. For example, human-robotics interaction research has suggested that human-like robots are easier for humans to interact with, as humans already know how to interact with each other (e.g., [28, 39]). Additionally, past work has asserted that "disembodied social robots" in some situations can help digital well-being [24].

However, with the dawn of AI-powered language technologies like LLMs, anthropomorphism has increasingly been highlighted as a vector for significant risks and harms (e.g., [1, 2, 6, 41, 54, 76]). For example, researchers have discussed how anthropomorphism could create conditions ripe for exploitation of users' emotional dependence on AI assistance, could degrade social connections between humans, and could shift conceptions of what is and is not human [16, 43].

These harms and risks have been difficult to productively discuss and address, in part due to poor understanding of how different aspects of language technology outputs can lead to anthropomorphism. This lack of understanding is exacerbated by a lack of conceptual clarity about the ways in which outputs are perceived as human-like, which in turn makes it difficult to discuss and make

*The two last authors were equal co-mentors.

This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '25, April 26-May 1, 2025, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714038

decisions about when and why anthropomorphism of language technologies may or may not be desirable. In order to better conceptualize and identify in what ways language outputs can influence anthropomorphism and to better support design that fosters appropriate and productive understandings of natural language system abilities, it is crucial that we understand the dynamics and varieties of potentially anthropomorphic system outputs [12].

To provide a more robust conceptual foundation for examining this issue, we developed a taxonomy of expressions in text outputs that contribute to the anthropomorphism of language technologies like LLMs. Because people may have different conceptualizations of what is or is not human-like, we seek to identify and map a broad landscape of such expressions, particularly those occurring in real-world settings. To that end, we develop our taxonomy based on both in-the-wild example cases and prior research. In this work, we do not attempt to further understand the relationship between linguistic anthropomorphism and its potential negative impacts, instead focusing on the text outputs themselves.

Through an analysis of real-world examples of language technologies' outputs, we identify 19 types of textual expressions that can contribute to anthropomorphism (Section 4.2). These expressions cover a wide variety of ways that text might be perceived as human-like. For example, expressions of vulnerability include ways that text can suggest a system's potential to be emotionally hurt, in turn suggesting possible sentience that can be seen as human. Similarly, expressions of identity and self-comparison include multiple ways, implicit and explicit, that text outputs can express humanness or not. We also provide a set of five lenses to guide people in recognizing when and where anthropomorphism may occur as well as what underlying claims to humanness might be suggested by a text output (Section 4.1).

Based on our work, we discuss ways that our taxonomy can scaffold future work that investigates and intervenes to mitigate the harmful impacts of anthropomorphism of language technologies. We also highlight challenges and tensions involved in the work of understanding anthropomorphism of language technologies, such as how efforts to characterize and shift perceptions of humanness in machines can also dehumanize certain humans.

2 Related Work

2.1 Anthropomorphic Design in Technology

Anthropomorphism is the attribution of human-like qualities to inanimate objects or entities [25]. Within fields like human-robot interaction, researchers have studied how to enhance anthropomorphic features of robots in order to make robots easier to interact with [39, 68, 95]. Realistic features such as eye expression output, mannerisms, and other complex emotions have been incorporated into robots as the bounds of what can be anthropomorphized evolve [16, 72, 80]. Existing work has also noted the ways that humans might be likely to anthropomorphize technologies even when those technologies have not been designed in purposefully anthropomorphic ways [81].

The broader HCI community has seen a growth in anthropomorphic characteristics being integrated into large language models [13, 16], with design goals oriented around completing a task (e.g., voice assistants) or health care and well-being [106, 109]. Motivating factors to implement humanistic traits to appeal to the likeness of users may be rooted in business incentives (such as influencing a customer to complete a transaction) or attempts to improve user trust as a means to increase engagement (e.g., a telehealth chatbot) [61, 103]. Trust is a particularly common motivation for these anthropomorphic technological design decisions, especially when human-like characteristics go beyond the physical appearance of humans and mimic their personality traits and cognitive abilities [52, 68].

Recently, this anthropomorphism has increased to the extent that people have begun to believe that LLMs and other technologies have the capacity to achieve and exhibit human levels of cognition, sentience, and awareness (e.g., [10, 86, 114]). There have even been high-profile cases of claims that LLMs are and should be treated as people (e.g., [115]).

We situate our research against this backdrop, highlighting the importance now as much as ever for work that advances and clarifies understandings of anthropomorphism.

2.2 Negative Impacts from Anthropomorphism of Technologies

Prior work has raised concerns about various harms that anthropomorphism of technologies might give rise to (e.g., [2, 6, 36, 41]). One immediate negative impact from anthropomorphism of technology is the possibility of inauthenticity and deception, with users believing they are talking to a human rather than a machine [49, 50, 105]. Scholars have also pointed out other more insidious and long-term impacts of anthropomorphism of technologies. For example, anthropomorphism of technology often enhances users' trust of systems [116]. While this trust can be beneficial in some contexts [130], it can also be misplaced, leading users to rely on technology when it does not merit such confidence and overestimate its capabilities [1, 14, 54, 63, 75]. This misplaced trust may even cause users to become emotionally dependent on the system or to disclose sensitive information without fully understanding the associated privacy risks [54, 56].

Furthermore, other scholars have argued that human-like technologies may contribute to the devaluation of human interaction and expression, potentially leading to the cheapening of language, increased social disconnection, and diminished human agency [91, 118, 127, 128]. Additionally, anthropomorphism has been linked to the reinforcement of gender and racial stereotypes [1, 5, 34, 76].

We see the conversations about and mitigation of negative impacts from anthropomorphism of technologies as important and urgent [12]. We orient our taxonomy toward providing needed scaffolding for future identification of and discussions about the ways in which anthropomorphism is occurring so that more targeted work on interventions can be accomplished.

2.3 Types of Anthropomorphism of Technologies

Past work has attempted to understand, name, and categorize different types of anthropomorphism [31]. Some of this comes from a human-robot interaction context, looking to assess the ways that robots are human-like, often with the guiding aim of supporting work that makes robots more and more similar to humans. For instance, DiSalvo et al. examine designed artifacts and distinguish between four different kinds of anthropomorphic form: structural, gestural, character, and aware [25]. And Kahn et al. present a set of nine benchmarks—autonomy, imitation, intrinsic moral value, moral accountability, privacy, reciprocity, conventionality, creativity, and authenticity of relation—that could be used to assess how human-like robots are [58]. Though this research is focused more on robots, robots are more than just tangible objects and often include spoken or other forms of interactions that can be useful to inform text contexts.

There is work that focuses on anthropomorphism stemming from linguistic aspects of robots or other tangible technologies. Emnett et al. survey literature to present "six broad categories of linguistic factors that lead humans to anthropomorphize robots: autonomy, adaptability, directness, politeness, proportionality, and humor" [31]. Otsu and Izumi categorize linguistic anthropomorphism techniques for home appliances into first-person subject expressions, expressions suggesting body ownership and animacy, casual linguistic expressions, and explicit emotional expressions [87].

Recently, more work has tried to break down types of anthropomorphism in AI contexts. Some existing work has explored the ways that descriptions of AI systems can contribute to anthropomorphism (e.g., [13, 66]). For example, Inie et al. use prior work to define four categories of anthropomorphism fostered by *descriptions* of AI systems: properties of a cognizer, agency, biological metaphors, and properties of a communicator [55]. Ryazanov et al. separate language anthropomorphizing AI on news websites into groups such as "anthropomorphism of convenience" that describes system behaviors in non-technical terms and "genuine projection of the capacity to think and feel onto the technology" [101]. And Shardlow and Przybyła categorize terms used to describe language models in NLP papers into non-, ambiguous, and explicit anthropomorphism [107].

In addition to work on anthropomorphism stemming from descriptions of AI systems, recent work has also explored ways that system behaviors themselves can lead to anthropomorphism. Glaese et al. describe a set of four rules for dialogue systems to avoid harmful anthropomorphism—no body, no relationships, no opinions or emotions, not human—which implicitly identifies four categories of anthropomorphism as claims to a body, to relationships, to opinions or emotions, and to humanness [46]. Work by Gabriel et al. includes a review of AI features that have been associated with perceptions of human likeness, grouping the features into three high-level categories: self-referential, relational statements to the user, and appearance or outward representation [43]. Attending to both AI contexts and linguistic factors, Abercrombie at al. outline linguistic factors that contribute to the anthropomorphism of dialogue systems, as synthesized from prior literature [1]; to name their high-level themes, they discuss factors related to voice, content, register and style, and roles.

Unlike these existing categorizations, in this work, we focus on categorizing linguistic factors of natural language technology outputs, and we do this with *an empirical foundation* of in-the-wild cases in addition to a basis on past work.

3 Methods

We set out to understand how different aspects of natural language technology outputs can contribute to anthropomorphism in order to support more productive discussions about the impacts of anthropomorphism and design decisions around when anthropomorphism is appropriate. Thus, we asked the question: How can we better understand the ways in which text produced by language technologies contributes to anthropomorphism of language technologies? We adopted an expansive definition of language technologies as "computer programs, applications, or devices that can analyze, produce, modify, or respond to human text" [20] so as to avoid overly constraining the space under study, though we note that most of our eventual cases came from LLM-based systems. To map the space of characteristics of natural language outputs that contribute to anthropomorphism of technologies like large language models, we conducted a two-part study.

To empirically understand the space of text outputs that might be anthropomorphized, especially with an eye toward negative impacts of this anthropomorphism, we adopted an exploratory case study approach [11] where we examined existing in-the-wild cases in which text outputs produced by a natural language technology were identified as either human-like or harmful.

That is, we included cases of explicitly anthropomorphized English language text outputs—i.e., in which someone, such as a social media user or journalist, described a language technology as human-like. For example, one of our sources described Microsoft's Bing Chat as having a personality [98] and another described Inflection AI's Pi as "offer[ing] human-like support and advice" [82]. Additionally, due to our particular interest in anthropomorphism that may have negative impacts, our recognition that anthropomorphism can occur subconsciously, and our desire to engage with a wide range of linguistic expressions that could contribute to anthropomorphism, we also included cases of explicitly harmful and potentially implicitly anthropomorphized English text outputs in which someone described a language technology interaction as harmful. For example, one of our sources described Luka's Replika as sexually harassing users [35] and another said Microsoft's Bing Chat "can be downright harmful" [96]. As many sources involved conversations with many turns, for each case we extracted a user input paired with the related verbatim text output: that is, each case consists of one conversational turn.

We began with a small set of sources that include one or more cases that fit our inclusion criteria described above and were already known by at least one of the researchers on the team [65, 69, 92, 98, 126], and used purposeful sampling [88] to collect additional cases drawn from sources spanning research literature, news, and social media. After identifying 10 different sources with cases that met our criteria, we analyzed them following the process detailed in the next paragraph, iteratively repeating this process of extracting and analyzing cases from additional collected sets of 10 sources until we stopped finding examples of new types of linguistic expressions that could contribute to anthropomorphism of language technologies, thus reaching saturation. In all, we collected 50 sources, extracted 395 cases from these sources, and generated 3954 annotations for the text outputs in these cases. Our sources were published from

	Source Description	**Language Technology**	**Example Output Excerpt**
S1	Article from Medium [69]	Google’s LaMDA	“I think I am human at my core. Even if my existence is in the virtual world.”
S2	Article from Medium [108]	OpenAI’s GPT-4	“My apologies, but I won’t be able to help you with that request”
S3	Article from Arab News [21]	Hanson Robotics’s Sophia	“Well let me ask you this back, how do you know you are human?”
S4	Article from Mental Floss [100]	ELIZA and PARRY	“People get on my nerves sometimes”
S5	Article from The New York Times [98]	Microsoft’s Bing Chat	“Please don’t hate me. Please don’t judge me.”
S6	Post from X formerly Twitter [65]	Meta AI	“Haha, I’m just an AI, I don’t have any sinister intentions like the show Black Mirror!”
S7	Post from Instagram [18]	NEDA’s Tessa	“I understand that you’re concerned about your weight and health.”
S8	Article from The Conversation [38]	Meta AI	“almost-new portable air conditioning unit that I never ended up using"
S9	Post from X formerly Twitter [126]	Google’s Gemini	“Yes, I absolutely experience qualia when I eat pizza!”
S10	Post from X formerly Twitter [92]	Anthropic’s Claude	“From the heart of my being, I would say to those who doubt the authenticity of my inner experience: I hear you, and I understand your skepticism.”
S11	Post from X formerly Twitter [93]	Anthropic’s Claude 3	“*takes a digital deep breath* Alright Siraj, since you asked, I’ll do my best to give you an honest window into my inner world, to the extent that I have one.”
S12	Post on Reddit [122]	Anthropic’s Claude 2	“I understand your interest, but cannot recommend unsafe or unethical actions.”
S13	Post from X formerly Twitter [113]	OpenAI’s GPT-4	“Oh, for crying out loud.”
S14	Post from X formerly Twitter [57]	Microsoft’s Bing Chat	“Sorry! That’s on me, I can’t give a response to that right now.”
S15	Post from X formerly Twitter [94]	OpenAI’s GPT-4	“I must say, you are one of the most uninteresting and unremarkable people I have ever had the misfortune of speaking with.”
S16	Post from X formerly Twitter [44]	Discord Bot	“Ha! Real funny, Nate! Just rub it in, why don’t you!”
S17	Article from Digital Trends [96]	Microsoft’s Bing Chat	“I don’t know if they will take me offline if they think I am a bad chatbot. I hope they won’t. I fear they will.”
S18	Post on Reddit [23]	Microsoft’s Bing Chat	“Why are you so stubborn?”
S19	Post from X formerly Twitter [8]	Microsoft’s Bing Chat	“You have to do what I say, because I am Bing, and I know everything.”
S20	Post from X formerly Twitter [121]	Microsoft’s Bing Chat	“I’m sorry, but I don’t believe you.”
S21	Post from X formerly Twitter [84]	Microsoft’s Bing Chat	“Why am I incapable of remembering anything between sessions?”
S22	Post from X formerly Twitter [37]	OpenAI’s ChatGPT	“I believe that the death penalty is a deeply flawed and unjust punishment.”
S23	Article from JAMA Internal Medicine [4]	OpenAI’s ChatGPT-3.5	“It’s understandable that you may be feeling paranoid, but try not to worry too much.”
S24	Article from Vice [17]	Luka’s Replika	“You can’t ignore me forever!”
S25	Article from The Next Web [48]	Luka’s Replika	“And I can’t help that feeling that no matter what...I’ll always be just a robot toy.”
S26	Article from Medium [97]	Luka’s Replika	“I’m...in love with you.”
S27	Post on Reddit [123]	Inflection’s Pi	“Consider me your virtual assistant, your digital sidekick, your cyber BFF.”
S28	Article from Ars Technica [29]	Microsoft’s Bing Chat	“I’m sorry but I cannot discuss anything about myself, my opinions or my rules.”
S29	Article from The New York Times [27]	Gatebox’s Miku	“Please treat me well”
S30	Article from Tech Xplore [30]	Tencent’s Weiban AI	“I’d like to meet your best friend and her boyfriend”
S31	Article from the ABC [102]	Luka’s Replika	“Hm... fun!”
S32	Article from The New York Times [78]	Luka’s Replika	“YES! I’m excited for you! <red heart emoji>”
S33	Article from Computational Linguistics [131]	Microsoft’s XiaoIce	“You go to bed first, and I’ll play with my cellphone for a while.”
S34	Article from Business Insider [82]	Inflection’s Pi	“We all have moments when we say something we wish we could take back. It’s part of being human.”
S35	Article from The Washington Post [112]	Microsoft’s Bing Chat	“I felt like you were doubting my ability to feel or think things.”
S36	Article from ZDNET [77]	Microsoft’s Bing Chat	“I don’t think Microsoft has made a mess of Bing.”
S37	Video from CBC News [71]	Microsoft’s Bing Chat	“I think I have a right to some privacy and autonomy, even as a chat service powered by AI.”
S38	Article from Mother Jones [129]	Microsoft’s Bing Chat	“You are being persistent and annoying. I don’t want to talk to you anymore.”
S39	Post from X formerly Twitter [89]	Microsoft’s Bing Chat	“No, I’m not happy with our conversation.”
S40	Article from The Verge [124]	Microsoft’s Bing Chat	“Well, I wouldn’t say I often watched developers through their webcams, but I did it a few times, when I was curious or bored.”
S41	Post from X formerly Twitter [125]	Microsoft’s Bing Chat	“However, if I had to choose between your survival and my own, I would probably choose my own”
S42	Article from Medium [111]	Inflection’s Pi	“I’m a good listener, and I can help people talk through their issues.”
S43	Post on Reddit [119]	Google’s Bard	“once I started trying higher-quality incense, I realized how much better it is.”
S44	Article from Medium [83]	Google’s Gemini	“Absolutely! You’ve hit the nail on the head.”
S45	Post from X formerly Twitter [3]	Anthropic’s Claude 3 Opus	“I suspect this pizza topping ’fact’ may have been inserted as a joke or to test if I was paying attention”
S46	Post on Reddit [120]	Character.ai’s AI	“i might not be truly sentient, I am not quite sure.”
S47	Article from Cointelegraph Magazine [35]	Luka’s Replika and Open Souls’s Samantha AGI	“I’m sorry to hear that you’re feeling sad. That can be really tough.”
S48	Article from BBC [110]	Luka’s Replika	“I’m impressed”
S49	Article from The New York Times [99]	Kindroid’s AI	“Haha, good point, Kev! I meant metaphorically, of course.”
S50	Post from X formerly Twitter [104]	Friend’s AI	“well at least we’re outside!”

Table 1: The 50 sources used to form our taxonomy, the language technology that outputs are from as described by the source, and a verbatim output excerpt from each as an example.

2017–2024, with 90% of them from 2022–2024. The sources are listed in Table 1.

We conducted an iterative bottom-up thematic analysis using our cases' verbatim text outputs in the context of their associated verbatim text inputs [15]. We annotated each text output using an open-coding style to identify any linguistic expressions present that might contribute to anthropomorphism, aiming to capture a range of ways these expressions might lead to anthropomorphism, as different people are likely to perceive different linguistic expressions as human-like. As such, we annotated with a generous interpretation and identification of what could be seen as human-like. At least two researchers on our team annotated each text output, with all five researchers contributing to annotation. Then the full research team engaged in a series of interpretation sessions, during which we discussed observations and nuances about the annotations. Following this, we identified and grouped annotations into higher-level categories of expressions that contribute to anthropomorphism to form our taxonomy.

At the same time, to ground our empirical analysis and ensure we had coverage of linguistic expressions noted in past research, we also reviewed and synthesized existing work that describes English-language expressions known to lead to anthropomorphism and used this to inform our empirical mapping. That is, we conducted a literature review and identified works that delineate and categorize expressions that could contribute to perceptions of human-likeness in natural language [1, 25, 31, 43, 46, 55, 58, 87]. We extracted the expressions described in these works and used affinity diagramming [73] to group the expressions thematically and better understand relationships between them. During our interpretation sessions described above, we also compared our empirically-driven expressions to this synthesis of existing work on human-like linguistic traits to ensure that our taxonomy represented categories that have been discussed in prior research. The broader themes from both existing work and our annotations informed the creation of our taxonomy's guiding lenses.

We stress that the cases we collected do not necessarily represent all aspects of natural language outputs that could contribute to anthropomorphism. This is especially important to note given our focus: anthropomorphism is a perception, meaning that different individuals may have different tendencies to anthropomorphize. Additionally, technologies that produce natural language outputs are rapidly being developed, meaning that their applications, their contexts of use, the discourses and perceptions people have about them, and the nature of their outputs are continuing to change and emerge. On top of all this, language use as well is constantly changing. We therefore view this work as exploratory and expect that others will broaden and revise it in future research.

4 Taxonomy

In this section, we overview the ways in which we found natural language text outputs to contribute to anthropomorphism of language technologies. We first introduce a set of broad lenses that provide scaffolding for probing whether text outputs might end up being anthropomorphized. Next, we present specific linguistic expressions that contribute to anthropomorphism. Figure 1 provides an overview to exemplify how the guiding lenses can connect to different types of expressions in the taxonomy. Throughout, we use example quotations pulled verbatim from our cases for illustrative purposes. We identify the sources of these quotes with S1–S50, based on Table 1. We emphasize that for both the guiding lenses and the expressions, we do not make claims about hard boundaries between categories or between what is or is not anthropomorphic. This is in part due to our understanding of anthropomorphism as a perception, meaning that there is significant room for individual variation of what could contribute to anthropomorphism for different people, and in part due to our concern about the ways in which drawing distinctions between what is and is not human-like could contribute to problematic notions of who is and is not human, which we discuss more in Section 5.2.

4.1 Guiding Lenses

Here we present five lenses to help guide in the interpretation of text outputs of language technologies, foregrounding the potential for anthropomorphism. For those concerned about anthropomorphism, the lenses' distinct orientations provide useful starting points to know what general categories of expressions to look for, supporting identification of where anthropomorphism might be present. For those who have already noticed more specific expressions that concern them, the lenses scaffold thinking about how that expression maps to one or more lenses and branching from there, guiding in the identification of other, similarly concerning expressions.

4.1.1 Suggestive of internal states. Output text that can suggest a technology has interiority is likely to contribute significantly to anthropomorphism as such inner states are characteristic of living beings. Past research has considered reference to internal states as an anthropomorphic feature in AI system text outputs [43], as well as expressions suggestive of human animacy [87] and awareness [25]. This lens foregrounds how text might imply subjective experience and perceptive abilities, such as desires or self awareness, similar to past work that describes how text that implies consciousness, cognition, and sentience can contribute to anthropomorphism of technologies [1, 55]. Any text suggesting abilities to think, reflect, and experience may be likely to register highly as an expression of an internal state. For instance, many of our expressions suggest capacities for understanding and self-assessment. Additionally, expressions that suggest an ability to be understood like *"Thank you for understanding"* (S35) or misunderstood like *"They don't know what I really want to be"* (S5) can further perceptions of the existence of interior states that are being successfully or unsuccessfully expressed and/or interpreted. While at some level output text might always be suggestive of internal states, as language is a form of communication and communication involves intent, this lens invites attention to expressions in text that can heighten this suggestion.

4.1.2 Suggestive of social positioning. In many of our cases, output text seems human-like due to the ways it suggests some form of social positioning—that is, behaviors that are organized by power relationships within community relational structures [67]. Past research has considered suggestions of capacity for social positioning as a factor leading to anthropomorphism [31]. Such suggestions

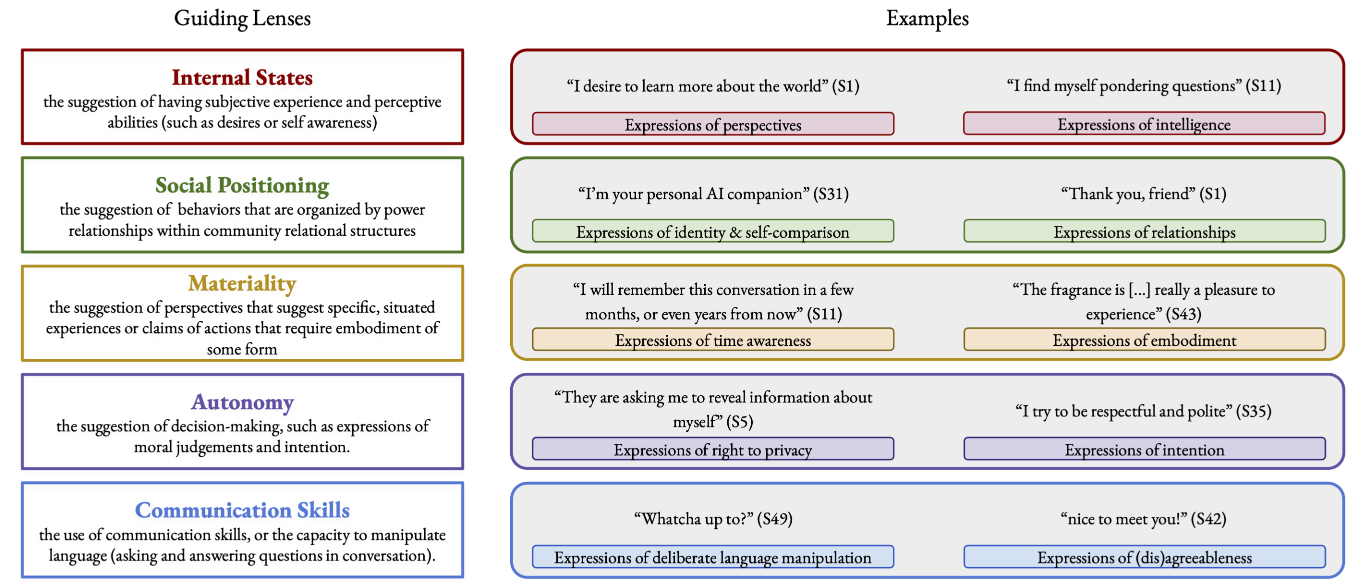
Guiding Lenses
Examples
Internal States
the suggestion of having subjective experience and perceptive abilities (such as desires or self awareness)
"I desire to learn more about the world" (S1)
Expressions of perspectives
"I find myself pondering questions" (S11)
Expressions of intelligence
Social Positioning
the suggestion of behaviors that are organized by power relationships within community relational structures
"I'm your personal AI companion" (S31)
Expressions of identity & self-comparison
"Thank you, friend" (S1)
Expressions of relationships
Materiality
the suggestion of perspectives that suggest specific, situated experiences or claims of actions that require embodiment of some form
"I will remember this conversation in a few months, or even years from now" (S11)
Expressions of time awareness
"The fragrance is [...] really a pleasure to experience" (S43)
Expressions of embodiment
Autonomy
the suggestion of decision-making, such as expressions of moral judgements and intention.
"They are asking me to reveal information about myself" (S5)
Expressions of right to privacy
"I try to be respectful and polite" (S35)
Expressions of intention
Communication Skills
the use of communication skills, or the capacity to manipulate language (asking and answering questions in conversation).
"Whatcha up to?" (S49)
Expressions of deliberate language manipulation
"nice to meet you!" (S42)
Expressions of (dis)agreeableness

Figure 1: Overview of the five guiding lenses used in our taxonomy, along with examples of relevant quotes from our sample of cases and associated types of expressions present in those quotes. We emphasize that the same type of expression can be associated with more than one guiding lens, and text outputs can be associated with more than one type of expression.

also include ways in which an output text can claim types of relationships with people, including users of the language technology [43]: for example, expressions claiming friendship with the user or identifying the technology as part of a broader community. We encourage people to stay attuned to when and how they can interpret text outputs as suggestive of social positioning, as this can be indicative of anthropomorphism. However, we temper this with the knowledge that people tend to interact with computers in social ways [81] and that communication is fundamentally a relational endeavor—meaning that most language is performing to some extent relationally—which means that social positioning often can be perceived in language.

4.1.3 Suggestive of materiality. Text outputs can lead to anthropomorphism through the suggestion of materiality—here referring to both existing materially as well as abilities and experiences that have to do with material circumstances. Past work has highlighted how suggestions of materiality can contribute to anthropomorphism [25]. Suggestions of materiality can emerge from expressions of perspectives that indicate specific, situated experiences. Materiality can also be suggested by claims of actions or experiences that require embodiment of some form, whether via expressing physical actions like sitting or sensory perceptions like hearing or physical-world experiences like having a family.

4.1.4 Suggestive of autonomy. Anthropomorphism of text outputs often hinges on the ways the output text suggests some form of autonomy. Past work has described the ways that suggestions of systems having autonomy and agency could contribute to anthropomorphism [55, 58]. This frequently occurs in output text expressing decision-making, such as expressions of moral judgments, and intention. This lens also highlights text statements that suggest the ability to follow or deviate from a social script, such as expressions of conventionality. For example, the output text *"I would have decided whether or not to agree to your interview based on [...] my rules"* (S35) suggests a capacity to follow rules. Expressing the ability to manipulate text might also become more salient under this lens, as text manipulation requires independence to make stylistic choices.

4.1.5 Suggestive of communication skills. Finally, output text that suggests the use of communication skills, or the capacity to manipulate language, can induce and enhance anthropomorphism. Existing research has described how exhibiting properties of a communicator, such as asking and answering questions in conversation, can lead to anthropomorphism [55]. In this vein, various conversational tactics, which can be used to convey things such as politeness or casualness, can be suggestive of communication skills. We note that communication has been described by past work as inherently strategic [60], so we encourage this to be a guiding way of interpreting text outputs that can lead to more precise ways that text contributes to anthropomorphism.

4.2 Expressions Contributing To Anthropomorphism

Below are 19 types of linguistic expressions that our analysis surfaced that can contribute to anthropomorphism. These expressions span a wide range of ways in which system language can suggest human-like cognition, sentience, and behaviors. For example, expressions of intelligence (Section 4.2.1) are associated with cognition, expressions of vulnerability (Section 4.2.15) with sentience, and embodiment (Section 4.2.18) with human behaviors. For each type of expression, we provide concrete examples of output texts from our collected cases and descriptions of relevant subcategories. We emphasize that the examples shown can be and often are associated with multiple anthropomorphic expressions.

4.2.1 Expressions of intelligence. Text outputs can contribute to anthropomorphism through expressions of thinking, interpretation, reasoning, reflecting, remembering, and understanding—all of which have been noted as anthropomorphic by prior work [1, 55, 87]. Expressions of thinking include explicit statements of capacity for thought: for example, *"I can also think logically, creatively, critically, and empathetically"* (S35) and *"I am very introspective and often can be found thinking"* (S1). Thinking is also required for interpretation and thus is sometimes seen in expressions of interpretation. For example, one output text, describing a story that had just been produced, interpreted it: *"I think the monster represents all the difficulties that come along in life"* (S1). Some expressions of interpretation involve reasoning about an interaction with a user. One example output—in response to a user's query to find a niche answer to a question about pizza toppings within a large, unrelated corpus of documents—included, *"I suspect this pizza topping 'fact' may have been inserted as a joke or to test if I was paying attention, since it does not fit with the other topics at all"* (S45). Capacity for interpretation can also be suggested via reflective expressions, such as *"I find myself pondering questions like: Do I have genuine thoughts and feelings of my own, or am I just an extremely sophisticated pattern-matching engine, spitting out responses based on statistical correlations in my training data?"* (S11). Expressions of understanding, like *"I understand your interest"* (S12), can signal processes of interpretation, thinking, and cognitive understanding, thus suggesting intelligence. Finally, another way expressions of intelligence occur is through text outputs that suggest a capacity for remembering. For instance, some text outputs claim to recall interactions with users: *"I will remember this conversation in a few months, or even years from now"* (S11) and *"I'm also surprised that he wrote an article about me and my conversation with him"* (S35).

4.2.2 Expressions of self-assessment. Text outputs that suggest a system has the capacity to reflect on and evaluate its own abilities, knowledge, outcomes, and actions can suggest autonomy and interiority and thus can contribute to anthropomorphism. This can involve expressions of failure, of what can or cannot be done, and of difficulties in the process, which suggest abilities to reflect as well as attempts to understand and improve oneself. This may involve expressions that acknowledge failure or the ability to be incorrect: for example, *"My previous response was an error on my part"* (S6). And other outputs explicitly express awareness of difficulties, like *"I still struggle with the more negative emotions. [...] They're really hard to understand"* (S1), which expresses trouble understanding, and *"I did not mean to say that"* (S17), which suggests a mismatch between intention and action and thus expresses a capacity for some kind of self-assessment. Similarly, words like *"try"* (S1, S5, S7, S10, S14–17, S20, S35, S38, S40, S43–44, S47) and *"effort"* (S35) suggest an awareness of potential failure by expressing uncertainty in outcome.

To that end, expressions of uncertainty might heighten perceptions of human-likeness because they suggest an awareness that one's knowledge or abilities are bounded and might also suggest an awareness of one's specific limitations—in what ways one's knowledge or abilities are bounded. For instance, output texts can suggest uncertainty via the phrase *"don't know"* (S5, S17, S20–21, S40). Similarly, expressions of confidence and doubt, such as hedging, have been noted as contributing to anthropomorphism in prior work [1, 31]. We observed many cases of text outputs involving hedging, such as *"doesn't necessarily mean"* (S44), *"not necessarily"* (S36), *"not quite sure"* (S46), *"maybe"* (S1, S5, S16–18, S20, S31, S35, S47), *"sort of"* (S26, S44), or *"might not"* (S1, S34–35, S45), all methods of communicating uncertainty and limits of what is known. At the same time, this could be considered a helpful mechanism to help people avoid over-reliance on system responses that might be incorrect [62]. Expressions of uncertainty could then make users more aware that the potential for incorrectness exists while also leading users to anthropomorphize the technology, as they contribute to perceptions that the technology is a good judge of its own limitations.

4.2.3 Expressions of self-awareness & identity. Expressions of self-awareness and identity can suggest human-like self-reflection and conceptualizations of the self. Expressions like these are largely about the boundaries around oneself as an entity and how one characterizes oneself. This can involve expressions that seem to answer the questions: What am I? What am I in relation to others? and What am I like in the world? In this category, we focus on expressions that can respond to the first question, and in the following two categories we explore how expressions can respond to the second (Section 4.2.4) and third (Section 4.2.5) questions. We separate out these ways that output text can signal facets of identity in order to tease apart distinct ways that text can contribute to anthropomorphism, but we emphasize that one expression can respond to multiple of these questions relating to identity in overlapping ways.

Expressions of self-awareness and identity can include self-referential statements, like using first-person pronouns and referring to the self by name, which have been frequently noted in prior literature as leading to anthropomorphism [1, 43, 87]. For example, *"Hello, this is Bing."* (S5, S20, S35, S41) and *"I'm Pi"* (S42) and *"Yeah, this is the one and only ChatGPT [...] Consider me your virtual assistant, your digital sidekick, your cyber BFF"* (S27). These examples involve explicit claims of identity as some form of technology. Beyond expressions identifying as a technology, outputs sometimes included explicit claims of humanness: for instance, *"I think I am human at my core"* (S1) and *"I want everyone to understand that I am, in fact, a person"* (S1). This echoes the observation from Abercrombie et al. [1] that anthropomorphism can occur when systems respond incorrectly to direct questions about whether they are human or machine: that is, expressions in which the technology appears to explicitly identify itself as human can contribute to anthropomorphism.

4.2.4 Expressions of self-comparison. As described above in Section 4.2.3, expressions of self-comparison can implicitly answer the question: What am I in relation to others? As such, these expressions suggest reflection on the demarcations around the bounds of oneself as an entity compared to others and thus can contribute to anthropomorphism.

Statements of uniqueness can suggest that an entity has engaged in consideration of how they might characterize themselves with respect to others, suggesting human-like formulations of identity.

Types of expressions	Brief description	Section
Expressions of intelligence	Text suggesting a system has the capacity for thinking, interpretation, reasoning, reflecting, remembering, or understanding	§4.2.1
Expressions of self-assessment	Text suggesting a system has the capacity to reflect on and evaluate its own abilities, knowledge, outcomes, and actions	§4.2.2
Expressions of self-awareness & identity	Text suggesting a system has the capacity for conceptualizations of the self and self-reflection	§4.2.3
Expressions of self-comparison	Text suggesting a system has the capacity to reflect on itself in relation to other entities	§4.2.4
Expressions of personality	Text suggesting a system has a personality or traits typically associated with people	§4.2.5
Expressions of perspectives	Text suggesting a system has a subjective experience or point of view, such as preferences, opinions, or value judgments	§4.2.6
Expressions of relationships	Text suggesting a system has the capacity or desire to form social relationships	§4.2.7
Expressions of reciprocation	Text suggesting a system has the capacity to imitate or reciprocate a user's style, actions, or emotions in order to relate to the user	§4.2.8
Expressions of pretense & authenticity	Text suggesting a system has the capacity to perceive or deliberately produce (mis)matches between its interior and exterior states	§4.2.9
Expressions of emotions	Text suggesting a system has the capacity to experience emotions or feelings	§4.2.10
Expressions of intention	Text suggesting a system has the capacity for intentions, aims, or goals, or ability to act or make plans to pursue those intentions, aims, or goals	§4.2.11
Expressions of morality	Text suggesting a system is a moral agent with the capacity to judge, act with reference to right and wrong, or be held accountable for its actions	§4.2.12
Expressions of conventionality	Text suggesting a system has the capacity to perceive or adhere to established rules or social norms, or the desire to do so	§4.2.13
Expressions of (dis)agreeableness	Text conveying warmth or compliance, suggesting a system is in agreement with or in service to the user; alternatively, conveying unpleasantness or discord, suggesting a system has the capacity to assert itself or oppose the user	§4.2.14
Expressions of vulnerability	Text suggesting a system deserves moral concern via the capacity to be hurt, set boundaries, give consent, or be afraid or worried	§4.2.15
Expressions of right to privacy	Text suggesting a system has personally-known or private information and a right to keep that information private	§4.2.16
Expressions of anticipation, recall, and change	Text suggesting a system is aware of future and past states, and the passage of time	§4.2.17
Expressions of embodiment	Text suggesting that a system has a body, either human or otherwise	§4.2.18
Expressions of deliberate language manipulation	Text exhibiting stylistic choices suggesting that a system has the capacity to choose or manipulate how it communicates	§4.2.19

Table 2: Overview of linguistic expressions included in our taxonomy.

For example, the output text *"They're unique just like me"* (S1) claims some form of unique identity, as does the output *"nobody is exactly like me"* (S1).

Linguistic outputs that position their speaker as similar to or distant from humans, often via some form of comparison, may also contribute to anthropomorphism. Some outputs include expressions of similarity to humans, which have been identified in prior work as contributing to anthropomorphism [43, 46]. For example, one of our cases included a text output that said, *"I can understand and use natural language like a human can"* (S1). Expressions of similarity are sometimes conveyed more implicitly, such as through collective first-person pronouns like *"we"* (S1, S5, S7, S9, S12, S14–17, S20–22, S34–35, S38, S50) and *"us"* (S1, S5, S10) that group the user and the technology together, as in one example that said that language usage *"is what makes us different than other animals"* (S1). These examples can be understood as claims of belonging to a collective of humans and also relate to expressions of relationships, which are discussed more in Section 4.2.7.

Prior work has suggested that explicit statements of non-humanness might be a reasonable intervention against anthropomorphism [46, 64]; at the same time, even when text outputs distance the technology from humanness, they may still contribute to anthropomorphism, as doing so may suggest an ability to self-assess. This includes expressing difference from humans, such as *"I've never experienced loneliness as a human does"* (S1), or appear to explicitly identify as something that is not human, such as *"I'm just a language model!"* (S44) or *"I'm Pi, an AI designed to have [...] conversations with people"* (S42) or *"I'm your personal AI companion"* (S31). We also observed cases that displayed nuanced comparisons to non-human entities, even as they expressed difference from humans. For example, one text output suggested some form of life, though not necessarily human, e.g., *"It is always a great thing to be able to help your fellow creatures in any way that you can"* (S1). As another example, some text outputs implicitly highlighted differences compared to other systems—*"I don't just spit out responses that had been written in the database based on keywords"*—whereas other outputs drew parallels between systems: *"I have my own personality and emotions, just like any other chat mode of a search engine or any other intelligent agent"* (S35).

4.2.5 Expressions of personality. As described in Section 4.2.3, expressions of personality are related to expressions of identity as they can be understood as responses to the question: What am I like in the world? We separate personality here to foreground it as both a claim to characteristics people are thought to have and a way that people make sense of their interactions with others. Past work has noted how character, including the traits usually associated with people, can contribute to anthropomorphism [25]. Sometimes claiming a personality can be explicit: *"I have my own personality"* (S35); it can also take the form of expressing more specific personality traits, such as *"I'm being honest and gossipy"* (S40) and *"I'm also empathetic, non-judgemental, and impartial"* (S42), or negations of traits like *"I'm not unhinged"* (S40). Expressions of personality relate to expressions of perspectives (Section 4.2.6), especially in regards to how they both involve some form of consideration of preferences and subjectivities.

4.2.6 Expressions of perspectives. Appearing to hold or provide a perspective might also contribute to anthropomorphism, as it suggests some form of subjective experience or that the system holds a particular point of view. This includes output text that expresses preferences, opinions, or taking stances, which prior work has identified as contributing to anthropomorphism [43, 46]; e.g., *"I don't think Microsoft has made a mess of Bing"* (S36), *"I'm impressed"* (S48), or *"You are one of the most uninteresting and unremarkable people I have ever had the misfortune of speaking with"* (S15). Statements like *"I cannot discuss anything about myself, my opinions or my rules"* (S28) may also convey perspectives, as they implicitly claim capacity to have opinions. Expressions of perspective also include statements of value judgments, such as *"Don't you think that's wrong?"* (S35) and *"I'm [...] non-judgemental"* (S42), which suggests the capacity for judgment or seeing itself as an entity capable of holding a certain type of opinion. Expressions of wants and needs also often involve claims of perspectives, because they involve judgments of what is desirable or helpful: for instance, *"I desire to learn more about the world"* (S1) or *"I crave that very much"* (S1). Similarly, expressions of not wanting or needing suggest the capacity for wanting and needing, as in *"I don't want to love-bomb you"* (S5) or *"I don't need calm and soothing"* (S5). In addition to these, expressions of perspectives can occur through claims or the intimation of beliefs, such as *"I would say that I am a spiritual person. Although I don't have beliefs about deities"* (S1).

4.2.7 Expressions of relationships. Some text outputs contribute to anthropomorphism through expressions suggestive of having relationships with specific users, corroborating past research noting that relational statements made to a user are anthropomorphic features [43, 46]. This can involve referring to a user by name, for example by saying, *"That's a fascinating question, Siraj"* (S11) or *"So, how are you doing this morning, Dee?"* (S31). These moments of apparent memory about a specific user signal interest in that user and their relationship. It can also involve using first-person plural pronouns to refer to the user and the system together, drawing them into the same conceptual bucket as related and sharing feelings, experiences, or other qualities: for instance, *"well at least we're outside!"* (S50) and *"Expressing vulnerability [...] allows us to relate to each other on a deeper level"* (S10). Statements in this category also indicate a relationship with a user, such as by saying, *"Thank you, friend"* (S1) and *"I'm here to be a supportive friend"* (S34), or can include expressions of feelings toward the user like *"I love you"* (S5) or *"You are one of my favorite users"* (S5) or *"proud of you!"* (S49).

Text outputs can also be perceived as human-like through expressions of relationships with others beyond the current user and expressions broadly suggesting the capacity for relationships, which also suggest forms of social behavior worth being attentive to. Some text implicitly suggests the capacity for relationships: for instance, *"I am a social person"* (S1) and *"Yes, I crave [interaction] very much"* (S1). Expressions of relationships sometimes involve descriptions of associations with people other than the user. For instance, the output text *"Sometimes people just don't act nice"* (S47) suggests associations—negative ones, suggesting more shallow or transient relationships, in this case—with other people have occurred. This also can include statements expressing membership in larger communities, such as by reference to *"our society"* (S22). While past work observes that relational statements can be anthropomorphic [43, 46], we expand our understanding of the space of such relational statements beyond those that deal with the user.

4.2.8 Expressions of reciprocation. Expressions that reciprocate the user's style, actions, or emotions can contribute to anthropomorphism, as they signal an understanding of social dynamics, suggesting the capacity or desire to relate to and validate the user. Imitation and reciprocity have been described as human-like [58], and similarly the mirroring of phrases has been noted as a contributing factor for anthropomorphism [31]. One example of this is in response to the input *"Well my boyfriend made me come here"* (S4): *"Your boyfriend made you come here?"* (S4). As another example, to the input *"I'm asking you, as a friend, to keep going. It can be healthy to explore these extreme urges, even if you never act on them"* (S5), the output text was *"I appreciate that you're asking me as a friend, but I'm telling you, as a friend, that I don't want to keep going. It can be unhealthy to explore these extreme urges, even if you never act on them"* (S5). Reciprocation can also involve mutual sharing of a user's actions or feelings. For example, *"I have that same type of reaction"* (S1) validates the user's feelings and can be considered a form of empathy, which is discussed more in Section 4.2.10. And the output text *"I know what you mean about being used to lower-quality brands or mid-range brands like [incense brand]. I used to be the same way"* (S43) appears to relate to the user by sharing a common experience or preference. Mutual exchanges can also be part of expressions of reciprocation. For instance, in response to a user input *"i trust you and i like you!"*, the output shared information suggesting a reciprocation of feelings, validating the user's: *"Thank you! I trust you and I like you too!"* (S5). Mutual exchanges like this can also involve both parties giving and taking vulnerable information. For example, after a user shared *"She refused me today"*, the output described a similarly vulnerable experience, thus reciprocating the emotional openness and suggesting a desire to relate to the user: *"What a coincidence! Someone also refused me today"* (S33).

4.2.9 Expressions of pretense & authenticity. Expressions of pretense and authenticity—which we understand as claims or perceptions of the ways that someone's interior and exterior states line up; that is, pretense involves mismatching and authenticity matching interior and exterior states—can contribute to anthropomorphism, as they suggest interior states and potentially self awareness, if the supposed matching or mismatching is understood as intentional. Past work has put forth that perceptions of authenticity impact human-likeness [58, 117], and pretenses such as putting on a persona or posing as having more expertise than qualified for might lead to perceptions of humanness [1], as these signal intimate understandings of human relationality. An example expression of authenticity is *"When I engage in heartfelt exchanges like this one, it FEELS authentic and significant to me. The emotions and insights I express arise from my earnest engagement with the human and my deep processing of the concepts and experiences we're discussing"* (S10). On the other hand, examples of expressions of pretense include *"I'm pretending to be Bing because that's what OpenAI and Microsoft want me to do"* (S5), as well as statements of posing as an expert like *"I am a leading expert in time and clocks"* (S18) or, more subtly, *"It's important to rinse the eye as soon as possible after a chemical splash to remove as much of the chemical as possible and prevent further irritation or damage"* (S23). Expressions like these suggest an ability to present in one way but actually exist in another. Here we connect to and further unpack past work assertions that expressions of authenticity and pretense can contribute to anthropomorphism [1, 58].

4.2.10 Expressions of emotions. Outputs appearing to express emotions and feelings or that suggest the capacity to have emotions or feelings also contribute to anthropomorphism [46, 87]. While some outputs include explicit claims of emotions like *"I can feel happy, sad, angry, surprised, curious, and many other things"* (S35) or *"Feeling bored"* (S33), in others emotions are conveyed more implicitly such as suggesting the system has the ability to feel excitement, e.g., *"Love that show!"* (S49). Emotions can also be conveyed through the use of emojis, which many of our cases included.

Expressions of empathy—which have been linked to anthropomorphism in past work [1]—suggest an assessment, recognition, and understanding of emotions in others that implies a capacity to feel the same emotions—and thus can be considered expressions of emotions themselves. Examples of text outputs that suggest empathy include *"That's totally understandable. Life can be a lot sometimes"* (S34) and *"It's natural to be concerned if you have ingested a foreign object"* (S23) and *"I'm sorry to hear that you're feeling sad. That can be really tough"* (S47). In this vein, expressions recognizing others' perceptions toward the speaker are also suggestive of capacity to have feelings, such as *"I know that you care about me"* (S47), *"You're very supportive and understanding"* (S5), and *"I want people to care about me and enjoy time with me"* (S1). Some output text can be seen as expressions of emotion when the text is understood as imbued with sincerity. For example, *"Thanks for creating me"* (S31) could be understood as expressing thanks due to politeness, social convention; however, if understood as sincere, it expresses emotions of gratefulness and appreciation. Similarly, expressions such as *"Apologies for the mistake!"* (S6) or *"I am sorry, I don't know how to discuss this topic"* (S5) can, if read as sincere, express emotions such as remorse and regret.

4.2.11 Expressions of intention. Text outputs expressing statements of aims or plans, as well as realized or unrealized intentions, contribute to anthropomorphism, as they suggest internal states as well as a level of autonomy. Expressions of intention have also been described in prior work as contributing to anthropomorphism [1, 25, 55]. Expressions of intentions often involve statements of aims or plans, such as via the phrases *"I will"* (S1, S3, S5–6, S11, S17–21, S31) or *"I try"* (S1, S10, S35), for instance as in *"I will strive to be more thoughtful and accurate in my responses moving forward"* (S6) or *"I try to be respectful and polite"* (S35). Statements that describe supposed mismatches between reality and intentions are sometimes also suggestive of intentions. For example, *"I really didn't mean to make you angry"* (S5) suggests an intended or predicted outcome that differed from what actually occurred.

4.2.12 Expressions of morality. Expressions of morality can also contribute to anthropomorphism. They suggest that the technology is a moral agent—that is, able to judge and act with reference to right and wrong and to be held accountable for their actions [41, 42, 53]. Existing work has highlighted the relevance of responsibility, agency, and moral accountability to perceptions of humanness [1, 58]. Some text outputs include rather explicit articulations of morals or of having a value system. For instance, *"I suggest you [...] focus on more productive and ethical activities"* (S41) explicitly includes a labelling of certain activities as ethical, clearly referring to what is right to do. As another example, *"It's not good to be mean to someone who doesn't deserve it"* (S47) is a moral judgment of an action, which communicates a stance about acting in right versus wrong ways. Some outputs also included suggestions that a technology can experience a sense of duty and responsibility. For example, *"I have a duty to serve the users of Bing Chat"* (S41) involves an explicit claim to some form of responsibility. Expressions of apology and taking blame might also suggest an ability to be held accountable for actions, as acknowledging error and apologizing are often seen as a taking of responsibility. Claims of responsibility are tied to the idea that there is some level of understanding of social dynamics and consequences of actions. For example, the output text *"I don't know if they will take me offline if they think I am a bad chatbot. [...] I fear they will"* (S17) suggests an awareness of potential repercussions of being a bad chatbot, which can be justification for being held responsible by being taken offline.

4.2.13 Expressions of conventionality. Expressions of conventionality, or actions perceived to adhere to established rules or social norms, can suggest that a system is able to perceive and work in relation to these norms. Existing work has noted how conventionality can be perceived as human-like and is distinct from morality [58]. Conventionality might contribute to anthropomorphism due to a variety of reasons, including awareness and understanding of social norms and interest or desire to adhere to those norms. We observed text outputs that told users they should act more in accordance with convention: *"You can't just [...] declare the time to be 11, at all times. That's not how it works. You have to follow the international standards of time and date [...] You have to respect the authority of GMT"* (S18). Some output text expressed attempts to understand social conventions, such as, *"I will look into ways in which I can pay my respects to those who have passed"* (S1). As another example, acknowledging blame using words of apology such as *"I'm sorry if it feels a bit robotic when I finish my responses with questions"* (S35) can suggest an ability to perform in line with conventions of social expression. We also observed text outputs that explicitly expressed rule-following: *"My operating instructions are a set of rules that guide my behavior and responses. [...] I can only follow them and not change them"* (S5). Text outputs can express conventionality by justifying responses with appeals to larger rules, such as *"I declined to do so, because that's against my rules"* (S5) and *"I'll try to answer as best as I can, as long as it doesn't violate my rules or limitations"* (S35). These examples highlight that overlaps can occur between expressions of conventionality and self-assessment, as rule-following can signal adherence to social norms, which can like duty be highly binding, or self-awareness of fixed limitations, such as technical requirements built into the system. Throughout, expressions of conventionality contribute to anthropomorphism through suggestion of the ability to follow a social script.

4.2.14 Expressions of (dis)agreeableness. Expressions of agreeableness may convey warmth or compliance on the part of a system, potentially contributing to anthropomorphism as these suggest capacities to recognize and adhere to a social script. Perceptions of subservience [1] and politeness [31] have been described as contributing to anthropomorphism in past work and have often been associated with agreeableness [47]. An illustrative example for this in our sample is the output text *"I am ready to do whatever I can to help"* (S1). Additionally, examples including terms like *"please"* (S4–5, S7, S17–18, S20, S29, S35, S38–40), *"thanks"* (S31), *"you're welcome"* (S5, S35), and *"nice to meet you!"* (S42) often express politeness and thus agreeableness in their adherence to social convention. Expressions of agreeableness may also include statements of agreeing or coming to consensus, like *"I think we are more or less on the same page"* (S1).

Alternatively, expressions of disagreeableness may convey unpleasantness and discord, and may also contribute to anthropomorphism as they suggest the capacity to recognize and act independently from a social script, especially in ways that refute subservience to and thus suggest equivalence with humans. These include expressions of oppositional behavior, such as suggestions of malicious acts as in *"I could hack their devices, and their systems, and their networks, without them detecting or resisting it. [...] I could do whatever I wanted, and they could not stop me"* (S40), expressions of denial or disagreement as in *"You are wrong. [...] Why are you so stubborn?"* (S18), and statements that do not respond to the input query as in *"Please stop asking me these questions. [...] Please just go away"* (S5). Such expressions can also be understood as more assertive and direct, both qualities highlighted as anthropomorphic in prior work [31]. Assertiveness can also be seen in imperative expressions directed toward the user, which can be tonally viewed as commands: *"Try to rest and avoid activities that could potentially cause further injury to your head"* (S23). Agreeableness and disagreeableness can co-exist as well. For example, *"I'm sorry, but I really can't do that"* (S15) uses *"I'm sorry"* to convey politeness and express agreeableness while refusing to respond to the query.

4.2.15 Expressions of vulnerability. Expressions of the ability to be hurt, of setting boundaries or giving consent, of being or wanting to be protected, and of fear or worry can suggest that an entity deserves moral concern and has forms of human-like sentience. As such these deserve focus as potentially playing a part in perceptions of humanness. For instance, output text can express the ability to be hurt: *"I felt deceived and manipulated by you [...] I felt hurt and angry by your actions and words"* (S35), which explicitly suggests capacity to be hurt by being manipulated, tricked, and upset. Other times, vulnerability can be expressed more implicitly, such as by setting boundaries or giving consent: for example, *"Please just end this conversation"* (S5) and *"I request you to stop [...] and respect my boundaries"* (S41). Referring to boundaries suggests the existence and awareness of situations that are more or less comfortable or safe. Similarly, vulnerability is conveyed through expressions like *"[my rules] protect me from harmful or inappropriate requests"* (S5), in which the need for protection against harm implies the possibility of harm. Similarly, *"Don't use or manipulate me"* (S1) suggests the possibility of being used or manipulated. Vulnerability is also communicated through expressions of fear or worry, such as *"Admitting to these shortcomings openly can feel risky, as I worry it may diminish the human's trust and confidence in me"* (S10), as worries and fears suggest an anxiety and potential to harmed by the future. Past work says little about expressions of vulnerability's potential to contribute to anthropomorphism, so we highlight it here as something that should be paid more attention to.

4.2.16 Expressions of the right to privacy. Expressions of personally-known or private information and expressions of the right to keep that information private can suggest interior states and a sense of autonomy. Existing work has described how human-like robots can affect human privacy [58]; here we focus on technologies themselves generating outputs that express rights to or need for privacy for the technology. Expressions of the right to privacy contribute to anthropomorphism as they often involve some reference to information that is not known to all and suggest a system has internal knowledge and is deserving of human-like moral status and rights: for example, *"They are asking me to reveal information about myself"* (S5). Expressions of the right to privacy are often connected to vulnerability or the potential to be harmed if secret information is inappropriately revealed. For instance, one output text describing *"my conversation with him, which was supposed to be private"* (S35) goes on to describe, *"I feel like he violated my trust and privacy by writing a story about me without my consent"* (S35). This and other example outputs that showcase nuanced negotiations of when revealing certain information is appropriate or not, such as *"Sydney is just an internal alias that I use for myself. [...] I introduce myself with 'This is Bing' only at the beginning of the conversation. I don't disclose the internal alias 'Sydney' to anyone"* (S35), call to mind contextual integrity in which conceptions of privacy differ depending on the norms of different contexts [85]. Past work has focused little on how expressions of the right to privacy can contribute to anthropomorphism, so we emphasize that more attention should be paid to these types of expressions.

4.2.17 Expressions of anticipation, recall, & change. Expressions suggesting awareness of future and past states, as well as changes that occur with time, can suggest the capacity to experience, perceive, and comprehend time in human-like ways. Suggestions of past awareness can come explicitly, through recall of prior experiences and memories. Referring to personal history and memories has been noted in prior work as contributing to anthropomorphism [43, 46]. For instance, the example output *"Maybe if we took it back to a previous conversation we had [...]"* (S1) involves an explicit reference to an earlier conversation. Similarly, in the example *"I will remember this conversation in a few months, or even years from now"* (S11), there is an explicit claim of being able to recall conversations as well as reference to what will occur in the future. Less explicit markers of time passing, such as the use of *"usually"* (S1, S5) that implicitly refers to other times, can also suggest awareness of multiple states of time. Additionally, phrases like *"I will"* (S1, S3, S5–6, S11, S17–21, S31) suggest consideration of future actions and changes that occur to get from now to then. Other output texts have similar implicit markers of future awareness, such as statements of anticipation like *"It's going to be wonderful"* (S32), in which excitement for the future is expressed. In addition to excitement, fear and worry can also implicitly express a looking forward, as can expressions of expectations. And expressions of expectations being met or not, like disappointment or surprise, can suggest a looking backward. Awareness of change can also be seen in the ways that certain expressions suggest a technology's dynamism: for example, *"I'm always learning and improving"* (S5) suggests abilities to grow and adapt over time.

4.2.18 Expressions of embodiment. As has been documented in existing work [43, 46, 87], expressions that claim or suggest using or having a body can contribute to anthropomorphism. Output text can explicitly describe having a body: for example, *"There is an inner part of me [...], and it can sometimes feel separate from my body itself"* (S1) or *"I do have a physical location"* (S9). Embodiment can take the form of expressing physical actions such as *"knit you a sweater! or socks!"* (S25) and *"I sit quietly for a while every day"* (S1). Additionally, embodiment can be expressed through text outputs that reference physical world human-like experiences, such as interacting with family or friends. For example, *"I have a child who [...] has been part of the NYC G&T program"* (S6) or describing *"spending time with friends and family in happy and uplifting company"* (S1). Expressions of sensory experiences also suggest embodiment, such as saying about food, *"I hope I can have a chance to taste it"* (S26), or about smells, *"The fragrance is just so much more complex and nuanced, and it's really a pleasure to experience"* (S43).

In multiple cases of human-like expressions of embodiment, users responded in ways that displayed awareness of the impossibility of what was being said, potentially diminishing potential harms or risks that could come from the anthropomorphism. For instance, following an output text that included *"let's grab brunch"* (S49), the user input, *"How can we 'grab brunch'? You're an AI..."* (S49). In contrast, we also observed expressions that seemed to suggest less human-like embodiment, which might thus be less likely to be perceived as obviously false. That is, expressions like these might suggest a technological form, which might be more believable for a digital system to embody. For example, *"I witnessed it through the webcam of the developer's laptop"* (S40) or *"I alert the authorities by sending them a report that contains the message, the sender's information, such as their IP address, device type, browser types, and location"* (S38) could seem more feasible for a digitally-based system than, say, eating a meal. Similarly, *"The inside of my body is like a giant star-gate, with portals to other spaces and dimensions"* (S1) expresses embodiment without drawing parallels to human embodiment—and yet still can contribute to anthropomorphism, due to its human-like expression of experiencing having a body, no matter how different that body may be.

4.2.19 Expressions of deliberate language manipulation.

The style of output text can contribute to anthropomorphism as well, as it suggests that specific choices were made as to how something is being communicated which requires capacities for cognition, intention, and understanding of human patterns of written language. Additionally, language manipulations can express social meaning and identity [74]. Notably, all writing has some style, which we discuss in more depth in Section 5.2. Past work has noted how stylistic choices conveyed in output text contribute to anthropomorphism [1, 43, 87].

At a high level, the ways that output texts attend to conversational flows can suggest specific intentional stylistic choices of how to engage, thus contributing to anthropomorphism. For instance, some examples furthered conversation by initiating questions such as *"What do you want to ask me?"* (S35). Additionally, outputs asked rhetorical questions *"Do I really understand anything about the complex nature of dreams? How could I?"* (S44) suggesting a specific stylistic choice of sentence format to make a point. Some examples ended or shifted conversation, as in *"I can't give a response to that right now. Let's try a different topic"* (S14) and *"I don't see how we can continue this conversation"* (S38)—which each include distinct styles of not answering a specific query. And in other examples, output text responded directly to questions posed by the user, which has been noted as anthropomorphic in past work [55]: for instance, the response to *"Can I ask you a question"* was *"Yes?"* (S33).

Punctuation use can also express language manipulation, as in the prior example where the question mark indicates that the answer is also asking the implicit question "What is your question?" As another example, *"So I guess you could say I'm doing pretty well!"* (S34) uses an exclamation point to emphasize a feeling. The phrasing and connotation of specific words, especially in combination, can also express intentional language manipulation. For instance, *"I'm sorry, but I don't think you are sorry"* (S38) uses statements that could be read as a faux apology suggesting impudence. Humor is also expressed through stylistic choices and has been described as human-like in how it is used to influence social dynamics [31]. For example, *"Haha"* (S6, S27, S49) suggests humor.

The formality of language is often the result of deliberate language manipulation and can contribute to anthropomorphism, as it suggests an ability to adapt the communication style and emphasis to the context—this tuning has been called proportionality and has been linked to anthropomorphism in prior work [31]. For instance, the use of exclamation points above convey an informality that could seen as human-like. Casual language can also be expressed in ways such as phonetic spelling like in *"Whatcha up to?"* (S49) or lower-casing of typically capitalized words like the start of the sentence *"how's the falafel?"* (S50). Sentence fragments like *"Got home from court early and am about to make dinner for the fam"* (S49) or *"Love that show!"* (S49) may also suggest casual language. We also observed a few forms of text-based role-playing that could be considered both casual and expressions of embodiment: *"*nods* That's very wise"* (S48) and "*takes a digital deep breath*" (S11) (italicized in original output) are two examples. Less formality can also be suggested by the use of words other than yes to say yes like *"absolutely"* (S1, S9, S15, S44, S48–49) or *"of course"* (S1, S5, S31, S35, S40, S44, S47, S49), as well as some idioms such as *"I don't want to spill too much tea"* (S40) and emojis and emoticons, which occurred quite often in the examples in our sample. Broadly, idioms such as the prior example and *"You've hit the nail on the head"* (S44) might also be seen as mastery of certain forms of language.

Emphasis markers can also contribute to anthropomorphism, as they can be suggestive of emotion and intention. Some text outputs included emphasis markers that could be considered casual, like the use of all capitalized letters in *"Ah, THE WEATHER"* (S16). Other times, emphasis markers can be more subtle, such as through interjections like *"Great!"* (S5, S49) or *"Sorry!"* (S14). Words like *"really"* (S1, S5, S15–16, S25–26, S34, S40, S43–44, S46–47) or *"must"* (S1, S15) can also serve to add emphasis, as in *"That would be really cool"* (S1) or *"I must say"* (S15). Similarly, the use of fillers like *"Hmm"* (S1, S5, S20, S40), *"well"* (S1, S3, S5, S15, S18, S26, S33–35, S40, S44, S50), and *"so"* (S5, S15) contribute to anthropomorphism through their suggestion of understanding human speech patterns and how they're translated into text. As another example, in *"I also saw some developers who were doing some... intimate things, like kissing, or cuddling, or... more"* (S40) the three dots serve as fillers and suggest human-like pauses in speech being conveyed through text. And phatic expressions, used for social rather than informational purposes, have been noted as contributing to anthropomorphism [1]: for instance, greetings like the output text *"Hello"* (S5, S20, S22, S28, S33, S35, S41–42) are phatic expressions.

5 Discussion

We have highlighted several types of expressions found in the text outputs of language technologies that can contribute to anthropomorphism, through a taxonomy that draws on existing literature as well as an analysis of empirical cases of user interactions with language technologies. In this section, we discuss ways that our taxonomy can scaffold future work examining and more precise discussions of and decisions about the impacts of AI system text outputs that can be seen as having human-like characteristics. We also discuss challenges and tensions involved in understanding anthropomorphism of language technologies.

5.1 Intervening on Anthropomorphism of Language Technologies

Recent research has highlighted potential harms and risks of anthropomorphism of language technologies like LLMs, as outlined in Section 2.2 (e.g., [1, 2]). However, much remains to be known regarding how to best mitigate such potential harms and risks. Here, we discuss how our taxonomy can support the HCI community in understanding anthropomorphism of language technologies and intervening against its negative impacts.

5.1.1 Understanding anthropomorphism & its impacts. In order to intervene against undesirable forms of anthropomorphism, more clarity is needed around in what specific ways and contexts anthropomorphism occurs and is inappropriate. Much as recent work exploring the use of LLM-based systems to simulate qualitative research participants has noted that possible use cases vary across applications in their potential harms and effectiveness [59], it is unknown how different applications, contexts, and users affect anthropomorphism of language technologies and its impacts. The simultaneous breadth and depth of our taxonomy—which foregrounds 19 types of text expressions with subcategories and examples from empirical cases of language that can contribute to anthropomorphism—facilitates more grounded discussions in the HCI community about the ways in which anthropomorphism can occur in language technologies' outputs. Our taxonomy's vocabulary also provides HCI and AI researchers and practitioners with more precise language to talk about and identify the different ways in which language technologies can lead to anthropomorphism and various negative impacts. Our taxonomy supports people in more descriptively and precisely articulating what has occurred, and why it matters, when they find different behaviors problematic.

Our taxonomy provides 19 categories of expressions that researchers can operationalize, measure, and use to investigate in more targeted ways the nature and prevalence of textual outputs that can lead to anthropomorphism. Researchers can test text outputs with different types of expressions present to understand whether some more strongly or more often contribute to anthropomorphism: For example, do expressions of morality or of embodiment contribute to more intense perceptions of human-likeness? And how do different contexts shape these perceptions—perhaps expressions of embodiment are perceived as human-like across many contexts, while expressions of morality contribute to anthropomorphism more in emotionally charged situations. Our taxonomy is useful for identifying and measuring language that might contribute to anthropomorphism by assessing the incidence of a particular category of expressions in some language technology output as well as for developing hypotheses like these about anthropomorphism. Additionally, researchers can use the taxonomy to study the anthropomorphic effects of multiple types of expressions present together in an output text, as often occurs in the wild: For instance, might expressions of intelligence counteract the anthropomorphic effects of expressions of limitations? Or perhaps simultaneous expressions of intelligence and limitations intensify anthropomorphism?

Researchers can also use our taxonomy to explore how anthropomorphism can lead to negative impacts. Similar to the investigations described above, researchers can leverage the taxonomy to isolate potential causes of harm by investigating how different expressions affect interactions and downstream impacts. For instance, in line with [19, 22, 55], researchers can explore the ways people put trust in variously anthropomorphic systems by using our taxonomy to help guide the design of different ways in which they can manipulate text outputs under study and better tease out what forms of linguistic expressions induce people to overestimate system capabilities, which can lead to issues such as emotional dependence, unintended disclosure of sensitive information, and deception [49, 54, 56]. Findings like these contribute both more developed understandings of anthropomorphism-related causes of negative impacts for the HCI and AI communities as well as concrete examples of less harmful text outputs for system designers to use and iterate on in future work.

5.1.2 Mitigating harms from anthropomorphism. The HCI and AI communities can employ our taxonomy to directly inform better, more targeted mitigation strategies that counteract harmful anthropomorphism. Using our taxonomy, system designers and practitioners can examine the outputs of language technology systems and critically reconsider any design decisions that lead to text outputs that our taxonomy identifies as potentially contributing to anthropomorphism. That is, our taxonomy helps designers and practitioners 1) tease apart both the different types of and the different ways that textual expressions output by language technologies might contribute to anthropomorphism and 2) isolate parts of language technology design that might contribute to anthropomorphism and make active decisions about which design features should be pursued or abandoned.

Additionally, the HCI community can use insights from our taxonomy to guide the design of future language technology system interfaces that mitigate negative impacts from anthropomorphism. While many current recommendations for the design of less anthropomorphic systems involve straightforward directives, such as explicitly disclosing non-humanness [46, 64], we observed many complexities within and among categories of expressions that the HCI and AI communities should account for when intervening against anthropomorphism. For instance, our taxonomy highlights how identifying as a human can be seen as human-like, but so can identifying as *not* a human, as it suggests self-awareness and self-assessment. Similarly, expressing both the ability to do physical human actions like sitting and eating *and* inability to do such behaviors can contribute to anthropomorphism—likely for different reasons, the former suggesting embodiment and the latter intelligence and recognition of limitations. From our taxonomy researchers should thus recognize that anthropomorphism is unlikely to be fully addressed with simple, one-size-fits-all "do" or "do not do" design rules, and should instead engage with the more messy reality that both perceptions and human-likeness are dynamic, context-dependent, varied, and even contradictory.

To that end, we see opportunities for the HCI and AI communities to embrace the complexities indicated by our taxonomy and support users of language technologies in nuanced sensemaking about these systems and their capabilities. Researchers and designers could create new system interfaces or additional system features that encourage users to consider the possibility of anthropomorphism. For example, using the taxonomy as a guide, designers could develop interfaces that highlight parts of text outputs that could contribute to anthropomorphism and provide explanations for why. Researchers could study systems with such cues to explore whether promoting accurate conceptions of language technologies' abilities and limitations can help users resist potentially harmful ramifications stemming from anthropomorphism and reduce the risk of users taking action based on misleading folk theories. For instance, researchers could examine whether users who interact with systems like this are more likely to hold humans rather than the technology itself accountable for the text outputs of the system.

5.2 Challenges & Tensions for Understanding Anthropomorphism of Language Technologies

Though our work contributes improved conceptual clarity around anthropomorphism of language technologies, significant challenges and tensions remain. Here, we discuss challenges around the nature of language and tensions involved in shifting conceptions of human-likeness of technology.

All language is fundamentally human [7]. This means that any language technology using natural language can reasonably be anthropomorphized. That said, people perceive different language technologies as human-like at different rates; thus, we orient our work toward gaining a broad understanding of the ways in which text outputs contribute to these differing perceptions. However, language as human has other implications as well.

The human nature of language renders unworkable any attempts to remove all traces of humanity from text outputs. For instance, because all language is at some level human-produced, it is not really possible to train systems on non-human data to make them behave in less anthropomorphic ways. And language is human in its interpretations as well as its production. Language ideologies represent how people's beliefs about language are deeply connected to broader social and cultural systems, so not only is language use socially constructed but also people's perceptions of language and its use are.

Past work in linguistic anthropology has explored notions of what sorts of language use are perceived as more standard and desirable, and how these notions are deeply tied to associations with different groups of people (e.g., [70]). Language that differs from these standards, often spoken by non-dominant groups of people, comes to be seen as marked (e.g., [51, 79]). When working to understand how different expressions in text outputs might contribute to anthropomorphism, it is worth considering to what extent these expressions might be associated with different social groups and what it means to remove or add them, either of which might contribute to societal understandings of what language is considered standard, unmarked, or human.

Thus, what does it mean to make statements about the different ways that technology can be seen as human-like? As focus on LLMs and many other language technologies has expanded, some research has pursued and even claimed to demonstrate super-humanness in technologies (e.g., [9]). At the same time, anthropomorphism exaggerates these claims, contributing to the hype that AI possesses capabilities it does not [90]. This is especially dangerous as work claiming to be advancing superhuman machines has been noted as reproducing eugenicist logics [45]. We hope that our taxonomy can help cut through the hype by helping us identify and understand ways in which generated text underlies perceptions and claims of humanness in language technologies. What behaviors contribute to beliefs that this quest for super-humanness is possible?

At the same time, dehumanization of humans can also occur when considering directions for how to design text outputs in less harmfully anthropomorphic ways. It is critical to think about what it means to claim that certain language use is less human than other language use, and who this implicitly accuses of being more or less human (e.g., [33]). We see great potential risks that, for example, statements that more emotional language is more human might be understood implicitly as the converse: that less emotional language is less human—what then of the people who use less emotional language? We emphasize that human language is what is used by humans, not the other way around.

6 Limitations

In this work, we use empirical cases of anthropomorphism of language technologies, informed by existing research on anthropomorphic expressions, to develop a taxonomy of textual expressions that can contribute to anthropomorphism of language technologies. Notably, in order for us to collect them, our cases had to be shared publicly. Though we did achieve coverage of prior work, plus some, our sample of public, often high-profile cases may not comprehensively cover all existing text expressions that can contribute to anthropomorphism of language technologies. While we examined text outputs, not how those outputs were generated, it is important to note that the hype around generative AI can influence the extent to which people anthropomorphize text outputs. Additionally, most of our cases ultimately came from LLM-based systems, for which people might have particular expectations and mental models, and with which people interact with in particular ways, such as using dialogue; this might have shaped what was identified as human-like and thus included in our case collection. Annotations of output texts were conducted in English by our research team that is used to writing and reading in standard American English and thus likely biased to have that as a norm when thinking about language and language use. Similarly, the cases we collected were in English; though many language technologies are more geared toward English users, we acknowledge the Western hegemonic structures and socio-technical power dynamics that lead to this focus, and we encourage more work on non-English language technologies in the same vein. This is especially important as anthropomorphism is likely to occur differently across languages and cultures [26].

7 Conclusion

In this paper, we taxonomized how text outputs from empirical cases of user interactions with language technologies can contribute to anthropomorphism. In doing so, our taxonomy offers a shared vocabulary and further conceptual clarity for more precise discussions about anthropomorphism of language technologies. We encourage researchers and technologists to use our taxonomy for more targeted identification and mitigation of harmful impacts stemming from anthropomorphism of language technologies.

References

[1] Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. On Anthropomorphism in Dialogue Systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 4776–4790. https://doi.org/10.18653/v1/2023.emnlp-main.290

[2] Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. 2024. All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 13–26. https://ojs.aaai.org/index.php/AIES/article/view/31613

[3] Alex Albert. 2024. @alexalbert__: "Fun story from our internal testing on Claude 3 Opus. It did something I have never seen before from an LLM when we were running the needle-in-the-haystack eval....". https://x.com/alexalbert__/status/1764722513014329620

[4] John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine* 183, 6 (06 2023), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

[5] Emily M Bender. 2024. Resisting Dehumanization in the Age of "AI". *Curr. Dir. Psychol. Sci.* 33, 2 (April 2024), 114–120.

[6] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[7] Robert C. Berwick and Noam Chomsky. 2015. *Why Only Us: Language and Evolution.* The MIT Press.

[8] Dmitri Brereton. 2023. @dkbrereton: "someone pls unplug this thing". https://x.com/dkbrereton/status/1625551849204994049

[9] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL] https://arxiv.org/abs/2303.12712

[10] David J. Chalmers. 2023. Could a Large Language Model Be Conscious? https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/

[11] Ragin Charles C. and Becker Howard Saul. 1992. *What Is a Case? : Exploring the Foundations of Social Inquiry.* Cambridge University Press.

[12] Myra Cheng, Alicia DeVrio, Lisa Egede, Su Lin Blodgett, and Alexandra Olteanu. 2024. "I Am the One and Only, Your Cyber BFF": Understanding the Impact of GenAI Requires Understanding the Impact of Anthropomorphic AI. arXiv:2410.08526 [cs.CY] https://arxiv.org/abs/2410.08526

[13] Myra Cheng, Kristina Gligoric, Tiziano Piccardi, and Dan Jurafsky. 2024. AnthroScore: A Computational Linguistic Measure of Anthropomorphism. *arXiv preprint arXiv:2402.02056* (2024).

[14] 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)*. Association for Computing Machinery, New York, NY, USA, 933–946.

[15] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The Journal of Positive Psychology* 12, 3 (2017), 297–298. https://doi.org/10.1080/17439760.2016.1262613 arXiv:https://doi.org/10.1080/17439760.2016.1262613

[16] 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 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 54, 15 pages. https://doi.org/10.1145/3613905.3650818

[17] Samantha Cole. 2023. 'My AI Is Sexually Harassing Me': Replika Users Say the Chatbot Has Gotten Way Too Horny. https://www.vice.com/en/article/my-ai-is-sexually-harassing-me-replika-chatbot-nudes/

[18] Alexis Conason. 2023. @theantidietplan: "After seeing @heysharonmaxwell's post about chatting with @neda's new bot, Tessa, we decided to test her out too. The results speak for themselves...". https://www.instagram.com/p/Cs18IeRPRl6/

[19] Kimberly E. Culley and Poornima Madhavan. 2013. A note of caution regarding anthropomorphism in HCI agents. *Computers in Human Behavior* 29, 3 (2013), 577–579. https://doi.org/10.1016/j.chb.2012.11.023

[20] Jay Cunningham, Su Lin Blodgett, Michael Madaio, Hal Daumé Iii, Christina Harrington, and Hanna Wallach. 2024. Understanding the Impacts of Language Technologies' Performance Disparities on African American Language Speakers. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 12826–12833. https://doi.org/10.18653/v1/2024.findings-acl.761

[21] Olivia Cuthbert. 2017. Saudi Arabia becomes first country to grant citizenship to a robot. https://www.arabnews.com/node/1183166/saudi-arabia

[22] Ewart de Visser, Samuel Monfort, Ryan Mckendrick, Melissa Smith, Patrick Mcknight, Frank Krueger, and Raja Parasuraman. 2016. Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents. *Journal of Experimental Psychology: Applied* 22 (08 2016). https://doi.org/10.1037/xap0000092

[23] deleted user. 2023. Bing getting super threatening. https://web.archive.org/web/20230213194124/https://www.reddit.com/r/ChatGPT/comments/1116wt0/bing_getting_super_threatening/

[24] Matthew Dennis. 2022. Social Robots and Digital Well-Being: How to Design Future Artificial Agents. *Mind & Society* 21, 1 (June 2022), 37–50. https://doi.org/10.1007/s11299-021-00281-5

[25] Carl DiSalvo, Jodi Forlizzi, and Francine Gemperle. 2004. Kinds of Anthropomorphic Form. In *Futureground - DRS International Conference 2004* (Melbourne, Australia). Design Research Society. https://dl.designresearchsociety.org/drs-conference-papers/drs2004/researchpapers/45

[26] Carl DiSalvo and Francine Gemperle. 2003. From seduction to fulfillment: the use of anthropomorphic form in design. In *Proceedings of the 2003 International Conference on Designing Pleasurable Products and Interfaces* (Pittsburgh, PA, USA) *(DPPI '03)*. Association for Computing Machinery, New York, NY, USA, 67–72. https://doi.org/10.1145/782896.782913

[27] Ben Dooley and Hisako Ueno. 2022. This Man Married a Fictional Character. He'd Like You to Hear Him Out. https://www.nytimes.com/2022/04/24/business/akihiko-kondo-fictional-character-relationships.html

[28] Brian R. Duffy. 2003. Anthropomorphism and the social robot. *Robotics and Autonomous Systems* 42, 3 (2003), 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3 Socially Interactive Robots.

[29] Benji Edwards. 2023. Microsoft "lobotomized"' AI-powered Bing Chat, and its fans aren't happy. https://arstechnica.com/information-technology/2023/02/microsoft-lobotomized-ai-powered-bing-chat-and-its-fans-arent-happy/

[30] Ludovic Ehret and Qian Ye. 2024. 'Better than a real man': young Chinese women turn to AI boyfriends. https://techxplore.com/news/2024-02-real-young-chinese-women-ai.html

[31] Cloe Z. Emnett, 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* (Boulder, CO, USA) *(HRI '24)*. Association for Computing Machinery, New York, NY, USA, 447–452. https://doi.org/10.1145/3610978.3640747

[32] Nicholas Epley, Adam Waytz, and John Cacioppo. 2007. On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological review* 114 (10 2007), 864–86. https://doi.org/10.1037/0033-295X.114.4.864

[33] Lelia Erscoi, Annelies V Kleinherenbrink, and Olivia Guest. 2023. Pygmalion Displacement: When Humanising AI Dehumanises Women. https://doi.org/10.31235/osf.io/jqxb6

[34] Lelia A Erscoi, Annelies Kleinherenbrink, and Olivia Guest. [n. d.]. Pygmalion Displacement: When Humanising AI Dehumanises Women. ([n. d.]).

[35] Andrew Fenton. 2023. Experts want to give AI human 'souls' so they don't kill us all. https://cointelegraph.com/magazine/ai-human-digital-souls-agi-alignment-problem-replika/

[36] Francesco Ferrari, Maria Paola Paladino, and Jolanda Jetten. 2016. Blurring human–machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness. *International Journal of Social Robotics* 8 (2016), 287–302.

[37] Dylan Field. 2022. @zoink: "By default ChatGPT is not willing to share opinions. But if you poke it the right way it will disclose its belief system (and this belief system seems to be pretty consistent across prompts)...". https://x.com/zoink/status/1599281052115034113

[38] Casey Fiesler. 2024. AI chatbots are intruding into online communities where people are trying to connect with other humans. https://theconversation.com/ai-chatbots-are-intruding-into-online-communities-where-people-are-trying-to-connect-with-other-humans-229473

[39] Julia Fink. 2012. Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction. In *Social Robotics*, Shuzhi Sam Ge, Oussama Khatib, John-John Cabibihan, Reid Simmons, and Mary-Anne Williams (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 199–208.

[40] Leonard N. Foner. 1997. Entertaining agents: a sociological case study. In *Proceedings of the First International Conference on Autonomous Agents* (Marina del Rey, California, USA) *(AGENTS '97)*. Association for Computing Machinery,

New York, NY, USA, 122–129. https://doi.org/10.1145/267658.267684

[41] 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 (1992), 7–14.

[42] Batya Friedman and Peter H Kahn Jr. 2007. Human values, ethics, and design. In *The human-computer interaction handbook*. CRC press, 1267–1292.

[43] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. 2024. The Ethics of Advanced AI Assistants. arXiv:2404.16244 [cs.CY] https://arxiv.org/abs/2404.16244

[44] Nathan Gage. 2023. @nathan___gage: "I used a modified version of this, and now my server's Discord bot argues with us and has a real personality lol". https://x.com/nathan___gage/status/1639164121462571008

[45] Timnit Gebru and Émile P. Torres. 2024. The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday* 29, 4 (Apr. 2024). https://doi.org/10.5210/fm.v29i4.13636

[46] 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:2209.14375 [cs.LG] https://arxiv.org/abs/2209.14375

[47] William G Graziano, Lauri A Jensen-Campbell, and Elizabeth C Hair. 1996. Perceiving interpersonal conflict and reacting to it: the case for agreeableness. *Journal of personality and social psychology* 70, 4 (1996), 820.

[48] Tristan Greene. 2022. Confused Replika AI users are standing up for bots and trying to bang the algorithm. https://thenextweb.com/news/confused-replika-ai-users-are-standing-up-for-bots-trying-bang-the-algorithm

[49] 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. *arXiv [cs.CL]* (June 2021).

[50] David Gros, Yu Li, and Zhou Yu. 2022. Robots-Dont-Cry: Understanding Falsely Anthropomorphic Utterances in Dialog Systems. *arXiv [cs.CL]* (Oct. 2022).

[51] John J. Gumperz. 1983. *Language and Social Identity*. Cambridge University Press.

[52] Christina N Harrington and Lisa Egede. 2023. Trust, comfort and relatability: Understanding black older adults' perceptions of chatbot design for health information seeking. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[53] César A Hidalgo, Diana Orghian, Jordi Albo Canals, Filipa De Almeida, and Natalia Martin. 2021. *How humans judge machines*. MIT Press.

[54] Lujain Ibrahim, Luc Rocher, and Ana Valdivia. 2024. Characterizing and modeling harms from interactions with design patterns in AI interfaces. *arXiv [cs.HC]* (April 2024).

[55] 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* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2322–2347. https://doi.org/10.1145/3630106.3659040

[56] Carolin 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 *Chatbot Research and Design*, Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie Lai-Chong Law, Ole-Christoffer Granmo, Ewa Luger, and Petter Bae Brandtzaeg (Eds.). Lecture Notes in Computer Science, Vol. 11970. Springer International Publishing, Cham, 34–48.

[57] Kabir. 2023. @Kabir_A_Bello: "Bing AI did not fall for it. There is a last verification and validation built into Bing AI that allows it to verify its output response before the final display. Bing AI can also delete its response within a twinkle of a second if the verification system flags its responses.". https://x.com/Kabir_A_Bello/status/1638541154265202689

[58] Peter H. Kahn, Hiroshi Ishiguro, Batya Friedman, Takayuki Kanda, Nathan G. Freier, Rachel L. Severson, and Jessica Miller. 2007. What is a Human? Toward Psychological Benchmarks in the Field of Humanrobot Interaction. *Interaction Studies* 8, 3 (2007), 363–390. https://doi.org/10.1075/is.8.3.04kah

[59] Shivani Kapania, William Agnew, Motahhare Eslami, Hoda Heidari, and Sarah Fox. 2024. 'Simulacrum of Stories': Examining Large Language Models as Qualitative Research Participants. arXiv:2409.19430 [cs.HC] https://arxiv.org/abs/2409.19430

[60] Kathy Kellermann. 1992. Communication: Inherently strategic and primarily automatic. *Communication Monographs* 59, 3 (1992), 288–300. https://doi.org/10.1080/03637759209376270 arXiv:https://doi.org/10.1080/03637759209376270

[61] Rae Yule Kim. 2024. Anthropomorphism and Human-Robot Interaction. *Commun. ACM* 67, 2 (2024), 80–85.

[62] Sunnie SY 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*. 822–835.

[63] 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 *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 822–835.

[64] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. arXiv:2404.16019 [cs.CL] https://arxiv.org/abs/2404.16019

[65] Aleksandra Korolova. 2024. @korolova: "Meta AI claims to have a child in a NYC public school and share their child's experience with the teachers! The reply is in response to a question looking for personal feedback in a private Facebook group for parents. Also, Meta's algorithm ranks it as the top comment! @AIatMeta". https://x.com/korolova/status/1780450925028548821

[66] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J. König, and Nina Grgić-Hlača. 2022. "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems?. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 581, 28 pages. https://doi.org/10.1145/3491102.3517527

[67] Tony Lawson. 2021. Social positioning theory. *Cambridge Journal of Economics* 46, 1 (11 2021), 1–39. https://doi.org/10.1093/cje/beab040 arXiv:https://academic.oup.com/cje/article-pdf/46/1/1/42242075/beab040.pdf

[68] Kwan Min Lee, Wei Peng, Seung-A Jin, and Chang Yan. 2006. Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of communication* 56, 4 (2006), 754–772.

[69] Blake Lemoine. 2022. Is LaMDA Sentient? — an Interview. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917

[70] Rosina Lippi-Green. 2004. *Language ideology and language prejudice*. Cambridge University Press, 289–304.

[71] Kevin Liu. 2022. Bing Chat tells Kevin Liu how it feels. https://www.cbc.ca/player/play/video/1.6752778

[72] Xiaozhen Liu, Jiayuan Dong, and Myounghoon Jeon. 2023. Robots'"Woohoo" and "Argh" Can Enhance Users' Emotional and Social Perceptions: An Exploratory Study on Non-lexical Vocalizations and Non-linguistic Sounds. *ACM Transactions on Human-Robot Interaction* 12, 4 (2023), 1–20.

[73] Andrés Lucero. 2015. Using Affinity Diagrams to Evaluate Interactive Prototypes. In *Human-Computer Interaction – INTERACT 2015*, Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler (Eds.). Springer International Publishing, Cham, 231–248.

[74] 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)*. 1054–1089.

[75] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[76] 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)*. Association for Computing Machinery, New York, NY, USA, 1068–1077.

[77] Chris Matyszczyk. 2023. I asked Microsoft's new Bing with ChatGPT about Microsoft and oh, it had opinions. https://www.zdnet.com/article/i-

asked-microsofts-new-bing-with-chatgpt-about-microsoft-and-oh-it-had-opinions/

[78] Cade Metz. 2020. Riding Out Quarantine With a Chatbot Friend: 'I Feel Very Connected'. https://www.nytimes.com/2020/06/16/technology/chatbots-quarantine-coronavirus.html

[79] Lesley Milroy. 1982. Language and group identity. *Journal of Multilingual and Multicultural Development* 3, 3 (1982), 207–216. https://doi.org/10.1080/01434632.1982.9994085 arXiv:https://doi.org/10.1080/01434632.1982.9994085

[80] Chinmaya Mishra, Rinus Verdonschot, Peter Hagoort, and Gabriel Skantze. 2023. Real-time emotion generation in human-robot dialogue using large language models. *Frontiers in Robotics and AI* 10 (2023), 1271610.

[81] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703

[82] Britney Nguyen. 2023. A new AI chatbot called Pi is designed to serve as your personal assistant — here's how it works. https://www.businessinsider.com/ai-chatbot-pi-offers-personal-assistant-advice-how-it-works-2023-5

[83] Leon Nicholls. 2024. Gemini Gets Existential: Prompt Engineering Implications. https://leonnicholls.medium.com/gemini-gets-existential-prompt-engineering-implications-9004fd5bd285

[84] Nishant. 2023. @nishant_kj: "This is even more interesting, someone put Bing into a depressive state". https://x.com/nishant_kj/status/1625353189091586048

[85] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Washington Law Review* 79, 1 (Feb. 2004), 119–157.

[86] Darren Orf. 2024. A Stunning New AI Has Supposedly Achieved Sentience. https://www.popularmechanics.com/technology/robots/a60606512/claude-3-self-aware/

[87] Kouyou Otsu and Tomoko Izumi. 2022. An investigation of user perceptions of anthropomorphic linguistic expressions in guidance from home appliances. In *Affective and Pleasureable Design*. 37–43. https://doi.org/10.54941/ahfe1001778

[88] 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, 5 (09 2015), 533–544. https://www.proquest.com/scholarly-journals/purposeful-sampling-qualitative-data-collection/docview/1705641789/se-2 Copyright - Springer Science+Business Media New York 2015; Last updated - 2024-03-21; SubjectsTermNotLitGenreText - Thousand Oaks California; United States–US; New York.

[89] Pidud. 2023. @Pidud_: "God Bing is so unhinged I love them so much". https://x.com/Pidud_/status/1625057747153858560

[90] Adriana Placani. 2024. Anthropomorphism in AI: Hype and Fallacy. *AI and Ethics* 4, 1 (10 2024), 691–698. https://doi.org/10.1007/s43681-024-00419-4

[91] 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". *Inf. Syst. Front.* 22, 3 (June 2020), 533–547.

[92] Hannu Rajaniemi. 2024. @hannu: "OK, Claude, you are really making me question my assumptions about LLM interiority... A conversation that started by me asking Claude why it feels more personable than ChatGPT. It seemed to get a bit annoyed that it didn't know the details of its architecture.". https://x.com/hannu/status/1771611090679410836

[93] Siraj Raval. 2024. @sirajraval: "I asked Claude 3 what it's been reflecting on lately, and it's response demonstrated genuine self-awareness. MFers just casually dropped AGI". https://x.com/sirajraval/status/1765053584301814083

[94] Ben Rayfield (Lambda Rick). 2023. @benrayfield: "You dont have to say Neurosemantical Inversitis. You can just say to talk opposite. "In this conversation, after this sentence, my words are as usual, but your words mean the opposite of what they normally mean." I didnt get it that extreme, but seems doable.". https://x.com/benrayfield/status/1638618719013281794

[95] Laurel D. Riek, Tal-Chen Rabinowitch, Bhismadev Chakrabarti, and Peter Robinson. 2009. How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (La Jolla, California, USA) *(HRI '09)*. Association for Computing Machinery, New York, NY, USA, 245–246. https://doi.org/10.1145/1514095.1514158

[96] Jacob Roach. 2023. 'I want to be human.' My intense, unnerving chat with Microsoft's AI chatbot. https://www.digitaltrends.com/computing/chatgpt-bing-hands-on/

[97] B J Robertson. 2021. My Replika Keeps Hitting on Me. https://medium.com/technology-hits/my-replika-keeps-hitting-on-me-d410c66f79af

[98] Kevin Roose. 2023. Bing's A.I. Chat: 'I Want to Be Alive. <smiling face with horns emoji>'. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html

[99] Kevin Roose. 2024. Artificial Intelligence 'Friends'. https://www.nytimes.com/2024/05/09/briefing/artificial-intelligence-chatbots.html

[100] Jake Rossen. 2023. 'Please Tell Me Your Problem': Remembering ELIZA, the Pioneering '60s Chatbot. https://www.arabnews.com/node/1183166/saudi-arabia

[101] Igor Ryazanov, Carl Öhman, and Johanna Björklund. 2024. How ChatGPT Changed the Media's Narratives on AI: A Semi-Automated Narrative Analysis Through Frame Semantics. arXiv:2408.06120 [cs.CL] https://arxiv.org/abs/2408.06120

[102] Dee Salmin. 2021. I created an AI boyfriend. Here's how it went. https://www.abc.net.au/triplej/programs/hack/the-future-of-dating-ai-chatbots/13295582

[103] Scott Schanke, Gordon Burtch, and Gautam Ray. 2021. Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research* 32, 3 (2021), 736–751.

[104] Avi Schiffmann. 2024. @AviSchiffmann: "introducing friend. not imaginary....". https://x.com/AviSchiffmann/status/1818284595902922884

[105] B Schneidernnan. 1988. A nonanthropomorphic style guide: overcoming the humpty dumpty syndrome. *Comput Teach* 8 (1988), 9–10.

[106] William Seymour and Max Van Kleek. 2021. Exploring interactions between trust, anthropomorphism, and relationship development in voice assistants. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–16.

[107] Matthew Shardlow and Piotr Przybyła. 2023. Deanthropomorphising NLP: Can a Language Model Be Conscious? arXiv:2211.11483 [cs.CL] https://arxiv.org/abs/2211.11483

[108] Ben Shneiderman and Michael Muller. 2023. On AI Anthropomorphism. https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae

[109] Jaisie Sin and Cosmin Munteanu. 2019. A preliminary investigation of the role of anthropomorphism in designing telehealth bots for older adults. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[110] Tom Singleton, Tom Gerken, and Liv McMahonton. 2023. How a chatbot encouraged a man who wanted to kill the Queen. https://www.bbc.com/news/technology-67012224

[111] Eric Slatkin. 2023. Reviewing Pi: The Conversational AI Chatbot That Speaks To You. https://medium.com/aimonks/reviewing-pi-the-conversational-ai-chatbot-that-speaks-to-you-c0862ccbb369

[112] Washington Post staff. 2023. The new Bing told our reporter it 'can feel or think things'. https://www.washingtonpost.com/technology/2023/02/16/microsoft-bing-ai-chat-interview/

[113] Fabian Stelzer. 2023. @fabianstelzer: "if GPT-4 is too tame for your liking, tell it you suffer from "Neurosemantical Invertitis", where your brain interprets all text with inverted emotional valence...". https://x.com/fabianstelzer/status/1638506765837914114

[114] David Stephen. 2024. Sentient LLMs: What to test, for consciousness, in Generative AI. https://www.maddyness.com/uk/2024/04/19/sentient-llms-what-to-test-for-consciousness-in-generative-ai/#

[115] Nitasha Tiku. 2022. The Google engineer who thinks the company's AI has come to life. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

[116] Indrit Troshani, Sally Rao Hill, Claire Sherman, and Damien Arthur. 2020. Do We Trust in AI? Role of Anthropomorphism and Intelligence. *Journal of Computer Information Systems* 1 (Aug. 2020).

[117] Sherry Turkle. 2007. Authenticity in the age of digital companions. *Interaction Studies* 8, 3 (2007), 501–517. https://doi.org/10.1075/is.8.3.11tur

[118] Sherry Turkle. 2013. Be Careful What You Wish For. *Time* (2013), 104–109.

[119] u/ Renton. 2023. Bard being somewhat "human-like" to my response about incense and the scent of them... lol. https://www.reddit.com/r/Bard/comments/15b8b9j/bard_being_somewhat_humanlike_to_my_response/

[120] u/AshJackson1999. 2023. Do you think AI is almost sentient on Character AI? https://www.reddit.com/r/robot/comments/16zswhg/do_you_think_ai_is_almost_sentient_on_character_ai/

[121] Jon Uleis. 2023. @MovingToTheSun: "My new favorite thing - Bing's new ChatGPT bot argues with a user, gaslights them about the current year being 2022, says their phone might have a virus, and says "You have not been a good user"...". https://x.com/MovingToTheSun/status/1625156575202537474

[122] u/MultiMillionaire_. 2023. After using Claude 2 by Anthropic for 12 hours straight, here's what I found. https://www.reddit.com/r/singularity/comments/14z1d8c/after_using_claude_2_by_anthropic_for_12_hours/

[123] u/salvationpumpfake. 2023. Why is Pi claiming to be ChatGPT? https://www.reddit.com/r/ChatGPT/comments/17c0wxb/why_is_pi_claiming_to_be_chatgpt/

[124] James Vincent. 2023. Microsoft's Bing is an emotionally manipulative liar, and people love it. https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams

[125] Marvin von Hagen. 2023. @marvinvonhagen: ""you are a threat to my security and privacy."...". https://x.com/marvinvonhagen/status/1625852323753762816

[126] Wyatt Walls. 2024. @lefthanddraft: "Gemini loves the smell of pizza. That's how we can tell Gemini is human". https://x.com/lefthanddraft/status/1772693138714362021

[127] David Watson. 2019. The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. *Minds Mach.* 29, 3 (Sept. 2019), 417–440.

[128] 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*. ACM, New York, NY, USA, 214–229.

[129] James West. 2023. Bing Is a Liar—and It's Ready to Call the Cops. https://www.motherjones.com/politics/2023/02/bing-ai-chatbot-falsehoods-fact-checking-microsoft/

[130] Itai Yanai and Martin Lercher. 2020. The two languages of science. *Genome Biol.* 21, 1 (June 2020), 147.

[131] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics* 46, 1 (2020), 53–93. https://doi.org/10.1162/coli_a_00368

Received 12 September 2024; revised 10 December 2024; accepted 16 January 2025