AnthroScore: A Computational Linguistic Measure of Anthropomorphism

Myra Cheng Kristina Gligorić Tiziano Piccardi Dan Jurafsky
Stanford University
myra@cs.stanford.edu

Abstract

Anthropomorphism, or the attribution of human-like characteristics to non-human entities, has shaped conversations about the impacts and possibilities of technology. We present AnthroScore, an automatic metric of implicit anthropomorphism in language. We use a masked language model to quantify how non-human entities are implicitly framed as human by the surrounding context. We show that AnthroScore corresponds with human judgments of anthropomorphism and dimensions of anthropomorphism described in social science literature. Motivated by concerns of misleading anthropomorphism in computer science discourse, we use AnthroScore to analyze 15 years of research papers and downstream news articles. In research papers, we find that anthropomorphism has steadily increased over time, and that papers related to language models have the most anthropomorphism. Within ACL papers, temporal increases in anthropomorphism are correlated with key neural advancements. Building upon concerns of scientific misinformation in mass media, we identify higher levels of anthropomorphism in news headlines compared to the research papers they cite. Since AnthroScore is lexicon-free, it can be directly applied to a wide range of text sources.

1 Introduction

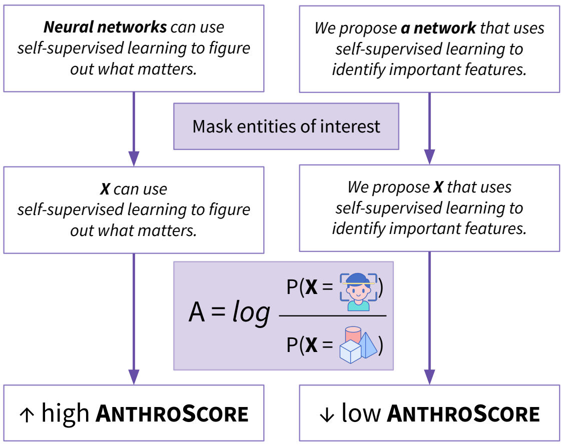

Neural networks *can use self-supervised learning to figure out what matters.*
We propose ***a network*** *that uses self-supervised learning to identify important features.*
Mask entities of interest
X *can use self-supervised learning to figure out what matters.*
We propose ***X*** *that uses self-supervised learning to identify important features.*
A = log P(X =) / P(X =)
↑ high AnthroScore
↓ low AnthroScore

Figure 1: To measure anthropomorphism in text, AnthroScore relies on probabilities computed using a masked language model to compare how much an entity is implicitly framed as human versus non-human.

Anthropomorphism, or assigning human-like characteristics to non-human entities, is commonplace in people's interactions with technology (Vasconcelos et al., 2023). However, anthropomorphizing language can suggest undue accountability and agency in technologies like artificial intelligence (AI) and language models (LMs). Projecting human qualities onto these tools facilitates misinformation about their true capabilities, over-reliance on technology, and corporate avoidance of responsibility (Watson, 2019; Shneiderman, 2020, 2022; Shanahan, 2022; Hunter, 2023). Such metaphors are especially consequential in public discourse (Fast and Horvitz, 2017) and in high-stakes domains like healthcare (Sharma et al., 2023) and education (Kasneci et al., 2023). Risks of harm from anthropomorphic misconceptions are underscored by regulation that prohibits hidden or undisclosed deployment of AI systems (Maréchal, 2016; Lamo and Calo, 2019).

Dialogue about the risks of AI has become prominent in recent years, including worries about human loss of control over AI ("AGI") as well as ethical concerns about the way that these technologies affect marginalized communities (Fast and Horvitz, 2017; Weidinger et al., 2022; Ferri and Gloerich, 2023). Anthropomorphic metaphors strengthen concerns about AI's hypothetical human-like capabilities, in turn distracting from the ways that these technologies have facilitated real-world harm to various populations (Tiku, 2022; Hunter, 2023).

We aim to make explicit—via quantification—the ways that anthropomorphic metaphors implicitly influence AI discourse.

There are currently no methods to identify an-

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 807–825
March 17-22, 2024 ©2024 Association for Computational Linguistics

thropomorphism and measure its prevalence. To bridge this gap, we introduce ANTHROSCORE, an automatic metric for anthropomorphism in language (Figure 1). ANTHROSCORE is a measure of how much the language of a text may lead the reader to anthropomorphize a given entity. (We elaborate on the definition and implications of anthropomorphism in Section 2.) Since anthropomorphism is the *inverse process* of dehumanization (Epley et al., 2007), our metric (described in Section 3) is a generalization of methods for measuring dehumanization in language (Card et al., 2022).

After demonstrating that ANTHROSCORE correlates to human judgment and established definitions of anthropomorphism, we use ANTHROSCORE to investigate the extent to which technical artifacts—the very objects of study for researchers—are anthropomorphized in computer science, statistics, and computational linguistics.

We use ANTHROSCORE to measure anthropomorphism in abstracts from ∼600K papers on CS/Stat arXiv and ∼55K papers in the Association of Computational Linguistics (ACL) Anthology. Building upon existing work on the widespread distortion of scientific claims in media, we also quantify anthropomorphism in headlines from ∼14K downstream news articles that cite these papers.

Our key findings are that

1. anthropomorphism in research papers has steadily increased over time, both in CS/Stat arSiv and in the ACL Anthology,
2. ACL, language model, and multimodality-related papers contain more anthropomorphism than other areas of research,
3. and anthropomorphism is much higher in downstream news headlines than in research paper abstracts.

We discuss causes and implications of these results, and we provide recommendations at the individual and community levels to minimize misleading anthropomorphism (Section 5). More broadly, ANTHROSCORE generalizes to analyzing any text since it does not rely on any lexicon or data curation, and we provide future directions in Section 6. Our code is available at https://github.com/myracheng/anthroscore and can be used to measure ANTHROSCORE for any text.

2 Background: Anthropomorphism

We ground our work in the social science literature on anthropomorphism. Previous scholars define anthropomorphism as "the attribution of distinctively human-like feelings, mental states, and behavioral characteristics" to non-human entities (Epley et al., 2007; Airenti, 2015; Salles et al., 2020). These characteristics entail

Definition 1. *"the ability to (1) experience emotion and feel pain (affective mental states), (2) act and produce an effect on their environment (behavioral potential), and (3) think and hold beliefs (cognitive mental states)" (Tipler and Ruscher, 2014).*

Scientific and technological concepts, especially human-centered ones, are particularly susceptible to anthropomorphic metaphors and interpretations (Sullivan, 1995; Salles et al., 2020). According to the Media Equation theory from social psychology, people tend to assign human characteristics to computers, interacting with them as if they were social actors (Reeves and Nass, 1996). This phenomenon leads people to behave and refer to computers in ways that are typical of human-human interactions–such as attributing personality–even when they are aware that they are interacting with a non-human entity (Nass and Moon, 2000).

Harms of Anthropomorphizing Technology Anthropomorphizing technology fuels misleading narratives that exaggerate their true capabilities, resulting in humans placing undue trust in them or harboring overblown fears (Proudfoot, 2011; Watson, 2019; Kenton et al., 2021; Crowell et al., 2019; Li and Suh, 2021; Gros et al., 2022; Deshpande et al., 2023). This has serious implications, such as spreading misinformation and diverting attention from the actual risks posed by these technologies (Weidinger et al., 2022; Shneiderman, 2022; Tiku, 2022). As news coverage of AI has ballooned since the 2000s (Fast and Horvitz, 2017), headlines like "Can AI cut humans out of contract negotiations?" and "Will AI Take Over The World?" reflect the influence of misleading anthropomorphic narratives in media coverage and public discourse (Salles et al., 2020; Hinton, 2023; Dhall and Kanungo, 2023; McManus, 2023).

Using anthropomorphic metaphors to discuss technology has long been connected to dehumanizing language (Dijkstra, 1985; Bender, 2022). These metaphors, which implicitly attribute agency to technology, carry legal, normative, and ethical implications regarding responsibility for decisions made with the assistance of AI and other technologies (Waytz et al., 2010). Anthropomorphic language also reinforces harmful gender stereotypes and has the potential to be manipulated for adverse influence by technology creators (Abercrombie et al., 2023; Deshpande et al., 2023).

Benefits of Anthropomorphism Thus far, we have emphasized the consequences of anthropomorphizing AI and related technologies. Beyond this specific context, however, anthropomorphism is not inherently harmful, but rather quite the contrary: it is a widespread, instinctive cognitive process that is often beneficial (Epley et al., 2007). For as long as humans have described and documented non-human entities, we have attributed human-like qualities to them, from folklore and mythology to scientific writing (Sherman, 2015; Mdoka, 2022; Darwin and Prodger, 1998; Freud, 1989; Hume, 1956). Anthropomorphism can facilitate learning (Kallery and Psillos, 2004; Wood, 2019), foster environmentalism (Root-Bernstein et al., 2013; Kopnina et al., 2018), and motivate protective action against deadly viruses (Wan et al., 2022).

In the context of technology, anthropomorphism also has benefits, such as providing intuition, facilitating the connection with technology, bonding, increasing trust, and enhancing understanding for the less tech-savvy (Yanai and Lercher, 2020; Zhong and Ma, 2022). Our metric can be used to understand these aspects as well.

Metaphors are powerful. Anthropomorphic metaphors are not merely linguistic choices without consequence—instead, a vast body of foundational literature has asserted that metaphors, however implicit, fundamentally structure our thoughts by facilitating our conceptualization of new ideas (Gibbs, 1994; Landau et al., 2010; Lakoff and Johnson, 2008; Tipler and Ruscher, 2014). As metaphors are repeated, they become ingrained into the social fabric of our language, becoming self-evident and escaping conscious notice (Lakoff and Johnson, 2008). Metaphors can have significant consequences: Tipler and Ruscher (2014) identify that dehumanizing metaphors have historically facilitated violence on massive scales, from the justification of American slavery to the Holocaust to anti-immigrant attitudes (Lott, 1999; Santa Ana, 2002; O'Brien, 2003; Musolff, 2010). Concerns about misleading anthropomorphic metaphors, especially regarding the capabilities of technology, broadly motivate our work to measure implicit anthropomorphism in language.

3 Methods

3.1 Measuring Anthropomorphism

Our metric relies on two key insights: (1) Anthropomorphism is the inverse process of dehumanization (Epley et al., 2007; Waytz et al., 2010; Tipler and Ruscher, 2014). AnthroScore is inspired by Card et al. (2022)'s context-sensitive method of using a masked language model (MLM) to measure implicitly dehumanizing language. (2) In English, the third-person singular pronoun marks animacy, i.e. *he* and *she* are used for animate beings while *it* is reserved for inanimate entities. Thus, we use these pronouns as the lexicons in our method.

The intuition behind our method is that the implicit framing provided by the context of a sentence reveals the degree of anthropomorphism of an entity in the sentence. Moreover, an MLM's predictions capture these implicit connotations since it is trained on a vast corpus of language to predict a missing word given the surrounding context.

AnthroScore measures the degree of anthropomorphism in a given set of texts (or a single text) T for a given set of entities (or a single entity) X as follows:

1. **Construct dataset of masked sentences S:** For every mention of $x \in X$ in T, we extract the surrounding sentence, and mask the mention of x (replacing x with a special [MASK] token) in the sentence.

2. **Compute A for each sentence:** For each sentence $s_x \in S$ where x is the masked entity, we compute the probability, according to an MLM, that the [MASK] would be replaced with either human pronouns (e.g., "he", "she") or non-human pronouns (e.g., "it"), i.e.,

$$P_{\text{HUMAN}}(s_x) = \sum_{w \in \text{human pronouns}} P(w),$$

$$P_{\text{NON-HUMAN}}(s_x) = \sum_{w \in \text{non-human pronouns}} P(w),$$

where $P(w)$ is the model's outputted probability of replacing the mask with the word w. (See Appendix B.1 for the full list of human and non-human pronouns.) We report the score A for s_x, as the log of the ratio between these two scores:

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

A captures the degree of anthropomorphism for entity x in sentence s.

3. **Compute the overall AnthroScore:** For the text(s) T, we compute the mean value of A across S, i.e.,

$$\bar{A}(T) = \frac{\Sigma_{s_x \in S} A(s_x)}{|S|}. \qquad (2)$$

$A(s_x)$ is lexicon-free and requires only the target texts T and entities E. We provide examples of how to use AnthroScore in various domains in Appendix A.

Interpretation $A(s_x)$ implies that in sentence s_x, according to the MLM's output distribution, the entity x is e^A times more likely to be implicitly framed as human than as non-human (e is the log base). Thus, $A(s_x) = 0$ means that x is equally likely to be implicitly framed as either human or non-human ($P_{\text{HUM}} = P_{\text{OBJ}}$), and $A = 1$ implies that the entity is $e^1 \approx 2.7$ times more likely to be implicitly framed as human than as non-human in the context of sentence s.

Implementation Details Following the approach of Antoniak et al. (2023), whose method we build upon for measuring semantic representations, we use the spaCy dependency parser to split texts into sentences and parse semantic triples (subject, verb, and object) from the texts. We then identify the relevant entities to mask from the subject and object noun chunks. We use the verbs in later analysis (Section 5.1). We use the HuggingFace Transformers Library's implementation of RoBERTa (`roberta-base`, 125M parameters), a state-of-the-art pre-trained MLM, as the model and tokenizer (Liu et al., 2020).[1] Our method enables us to obtain scores on various levels: for individual sentences, for entire corpora, and also for particular terms/entities. In Section 4, we report results by comparing $\bar{A}$ across these different scales.

3.2 Datasets

We measure anthropomorphism both in scientific papers and downstream news headlines. We apply AnthroScore to three datasets to analyze when and how researchers anthropomorphize their objects of study, and how these entities are perceived in the news: **(1) arXiv Dataset:** We use abstracts from all papers posted to the computer science (CS) and statistics (Stat) arXivs that are in the publicly available dataset (Clement et al., 2019). These 601,964 papers span from May 2007 to September 2023. **(2) News Dataset:** We extract headlines (titles and ledes) from all downstream news articles that explicitly cite any of the papers in the arXiv Dataset using the Altmetric API (Adie and Roe, 2013). After filtering the headlines for English language, our dataset contains 13,719 news headlines that cite 8,436 unique articles. **(3) ACL Dataset:** We use abstracts from the ACL Anthology (Rohatgi et al., 2023), the primary digital archive for papers related to computational linguistics and NLP. To maintain consistency with the arXiv and downstream news datasets, which begin in 2007, we use only the 55,185 articles from 2007 onwards.

For the entities X, we focus on technical artifacts. We first parsed research papers' abstracts for sentences with mentions of technical artifacts. To determine the list of technical artifacts, we extracted the top 100 most common entities (subjects and objects identified by the spaCy dependency parser) in the abstracts of a random sample of 15K arXiv abstracts. Then, from this list, we manually annotated for entities that refer to technical artifacts, agreeing on:

$$X_{\text{artifact}} = \{\text{algorithm, system, model, approach,} \\ \text{network, software, architecture, framework}\}.$$

We parsed all datasets for all semantic triples that included these keywords. We found 1,048,893 such instances (~950K from arXiv, 3K from news, 97K from ACL). For each instance, we extract the full sentence and mask the mention of the technical artifact (replacing the keyword phrase with a special [MASK] token) in the sentence to create the set of masked sentences S. After deduplicating the datasets, we computed A for each sentence as well as average scores $\bar{A}$ across the texts.

To address concerns of anthropomorphism related to language models (LMs), we also filter explicitly for papers that mention LMs. We do this using Movva et al. (2023)'s method of searching all titles and abstracts for terms related to LMs (Appendix B.2). This resulted in a subset of ~18K papers, which we henceforth refer to as LM papers.

Across all papers, we also analyze anthropomor-

[1] We compute AnthroScore using a machine with 1 GPU and 128GB RAM in < 10 GPU hours combined for all datasets described in Section 3.2.

phism for LM-related entities

$$X_{\text{LM}} = \{\text{language model, GPT, BERT}, \dots\}.$$

To construct X_{LM}, we followed a similar procedure as for X_{artifact}: we parsed all semantic triples for the 100 most common entities in these triples. Then, we filtered this list for entities that refer explicitly to LMs. We also added terms from Movva et al. (2023)'s list of keywords. X_{LM} is listed in Appendix B.2. Then, we collected all unique sentences containing $x \in X_{\text{LM}}$ and computed A for each sentence.

3.3 Construct Validity and Robustness

Qualitative Analyses To validate our method, we first analyze the scores of sentences that mentioned explicitly human entities ($X_{\text{human}} = \{$researchers, people, ... $\}$). The full list of terms in X_{human} is in Appendix B.2. We found that sentences containing these entities have much higher scores of $\bar{A}$ than the non-human entities we analyze, suggesting that A indeed captures an intuitive notion of anthropomorphism (Figure 2, top right).

Correlation with Human Perception To confirm this, we conducted a more in-depth human annotation study of 400 masked sentences: a randomly-sampled set and a set stratified by A score. Two authors (who did not have access to the scores) independently annotated the sentences, indicating whether the sentence contains anthropomorphism using Def. 1. After two rounds of annotation, we reached substantial inter-rater agreement (Cohen's $\kappa = 0.87$).

A chi-square test was performed to examine the relation between human perception of anthropomorphism and inferred anthropomorphism measured via high A scores (thresholding at the average A score in the respective set; randomly-sampled set: $avg(A) = -3.28$, stratified set: $avg(A) = 1.32$). Within both sets, higher than average A is significantly more likely among sentences humans judged to contain anthropomorphism (randomly-sampled set: $\chi^2 = 17.98$, $p < 0.00001$; stratified set: $\chi^2 = 11.26$, $p < 0.001$). Complete details and full distributions of scores are in Appendix C.1.

Correlation with LIWC As another measure of validity, we examine correlations between A and dimensions of LIWC-22. LIWC-22 is a state-of-the-art software for analyzing word use in text whose construct validity has been shown by many papers over the years (Tausczik and Pennebaker, 2010; Pennebaker, 2011; Boyd et al., 2022). It contains lexicons for words that relate to different dimensions such as writing styles, psychological processes, topic categories, etc., and computes the prevalence of each dimension based on counts of the words in the corresponding lexicon. Thus, we compute LIWC scores for high- and low-anthropomorphism sentences. We define high and low-anthropomorphism sentences as

$$S_{\uparrow} = \{s_e \in S | A(s_e) > 1\}, \text{ and}$$
$$S_{\downarrow} = \{s_e \in S | A(s_e) < -1\}$$

respectively, where S is all sentences parsed from the datasets Table 1 lists examples of sentences in $S_{\downarrow}$ and $S_{\uparrow}$.

Using two-sample t-tests to compare LIWC scores between $S_{\downarrow}$ and $S_{\uparrow}$, we find that many of the LIWC dimensions that are statistically significantly higher in $S_{\uparrow}$ correspond to the three aspects of anthropomorphism (Def. 1), while the LIWC dimensions that are higher in $S_{\downarrow}$ relate to academic language (Figure A3).

Specifically, the *Affect* LIWC dimension is statistically significantly higher in $S_{\uparrow}$, connecting to the affective component of Def. 1. The other two components are behavior and cognition. Behavior is connected to dimensions like *Physical* (terms related to the human body and health) and *Lifestyle* (work, home, religion, money, and leisure), while cognition is linked to *Perception* (perceiving one's surroundings), all three of which are statistically significantly higher in $S_{\uparrow}$ than in $S_{\downarrow}$.

The LIWC scores also reveal stylistic differences between $S_{\uparrow}$ and $S_{\downarrow}$: the dimensions of emotional tone, authenticity, and casual conversation are significantly higher for $S_{\uparrow}$. Dimensions that are higher for $S_{\downarrow}$ include Words Per Sentence, the number of long words, and Clout (language of leadership/status). This aligns with theories that anthropomorphism is related to more accessible and easily understood language (Epley et al., 2007).

Interestingly, the *Cognition* LIWC dimension is higher in $S_{\downarrow}$. We hypothesize that this is due to the inclusion of words like *but, not, if, or,* and know in the lexicon as well as the *causation* subdimension, which reflects the prevalence of causal claims in scientific language rather than anthropomorphism.

Robustness We compute three modified versions of $\bar{A}$ to evaluate robustness. (1) We remove individual words from the pronoun lists before recalculating $\bar{A}$. Using Spearman's rank correlation

$S_\uparrow$: Sentences with high AnthroScore ($A > 1$)	$S_\downarrow$: Sentences with low AnthroScore ($A < -1$)
• When a job arrives, **the system** must decide whether to admit it or reject it, and if admitted, in which server to schedule the job. • Meanwhile, anti-forensic attacks have been developed to fool **these CNN-based forensic algorithms**. • **The models** demonstrated qualifications in various computer-related fields, such as cloud and virtualization, business analytics, cybersecurity, network setup...	• More and more users and developers are using **Issue Tracking Systems** to report issues, including bugs, feature requests, enhancement suggestions, etc. • **Our approach** delivers forecast improvements over a competitive benchmark and we discover evidence for strong spatial interactions. • To this end, for training **the model**, we convert the knowledge graph triples into reasonable and unreasonable texts.
• ***Large language models*** *don't actually think and tend to make elementary mistakes, even make things up.* • ***The algorithms*** *also picked up on racial biases linking Black people to weapons.* • ***The AI system*** *was able to defeat human players in…*	• *Microsoft is betting heavily on integrating* ***OpenAI's GPT language models*** *into its products to compete with Google.* • *Deepmind has been the pioneer in making* ***AI models*** *that have the capability to mimic a human's cognitive…* • *For workers who use* ***machine-learning models*** *to help them make decisions, knowing when to…*

Table 1: **Examples of sentences with high and low AnthroScore.** Bolded phrases are the entities that are masked when computing A. The non-/italicized sentences are from the arXiv and News datasets respectively.

coefficient r between the modified score and the original score, the bootstrapped scores have a statistically significant correlation $r > 0.86$ ($p < 0.001$) for all pronouns. (2) We compute $\bar{A}$ after removing the top three verbs for $S_\downarrow$ and $S_\uparrow$ based on the verbs in Table 2. (3) We compute $\bar{A}$ after removing sentences containing reporting verbs. We find the same trends using these modified scores (Figure A6). For more details on (2) and (3), see Section 5.1 and Appendix E.5.

4 Results

4.1 Category analysis: LMs and multi-modal models are most anthropomorphized.

Among the top 10 most popular categories in CS/Stat arXiv, Computation and Language (cs.CL) has the highest rate of anthropomorphism, followed closely by Computer Vision (cs.CV) (Figure 2, top left). Artificial Intelligence (cs.AI), Security & Cryptography (cs.CR), and Machine Learning (cs.LG) also have higher $\bar{A}$. For cs.CR, manual inspection reveals that these sentences are primarily about security in the context of AI models.

Among the top 50 most popular categories, subfields related to multimodality and multidimensional signals (Multimedia (cs.MM), Audio and Speech Processing (eess.AS), sound (cs.SD), Image and Video Processing (eess.IV)) emerge as categories with the highest $\bar{A}$ (Figure 2, bottom). Among these papers, we find that 82% are cross-listed with stat.ML, cs.CL, cs.CV, cs.LG or cs.AI. Among the remaining 18%, manual inspection reveals that the sentences with high A are largely focused on neural models, such as multimodal and speech language models (note, however, terms used by these subfields are not in X_{LM}). We hypothesize that this trend of high anthropomorphism will continue given the rising prevalence of multimodal language models; the use of transformers, neural models, etc. for other types of data beyond text; and various AI actors' declarations of aiming to build more powerful "general intelligence" (Team et al., 2023; Zhu et al., 2023; Li et al., 2023; Yu et al., 2023). Quantitative biology subfields (q-bio.QM and q-bio.NC) also have high $\bar{A}$; manual inspection reveals that q-bio sentences often have metaphors about cognition, which is a key aspect of anthropomorphism (Def. 1).

On the other side of the spectrum, the subfields of Programming Languages (cs.PL), Multiagent Systems (stat.MA), and statistical methodology (stat.ME) have the lowest $\bar{A}$. This is interesting since CS subfields like AI, ML, etc. use many of the same tools as stat.ME yet have much higher $\bar{A}$. This reflects that $\bar{A}$ is a measure of a field's implicit norms and values, which we discuss further in Section 5.2.

Regarding LMs, $\bar{A}$ is statistically significantly higher for LM papers than other papers (Figure 2, top middle). Within LM papers, X_{LM} has even higher $\bar{A}$ than X_{artifact} (Figure 2, top right). LMs in particular are more anthropomorphized than other artifacts, which connects to existing concerns about misleading anthropomorphism of LMs (Bender and Koller, 2020; Shanahan, 2022).

4.2 Temporal analysis: Anthropomorphism in research papers is increasing over time.

Figure 3 displays temporal trends in anthropomorphism within the arXiv and ACL data. Using Spearman's r between year and $\bar{A}$ to measure temporal trends, we find that anthropomorphism is increasing over time in both datasets ($r = 0.54$ and

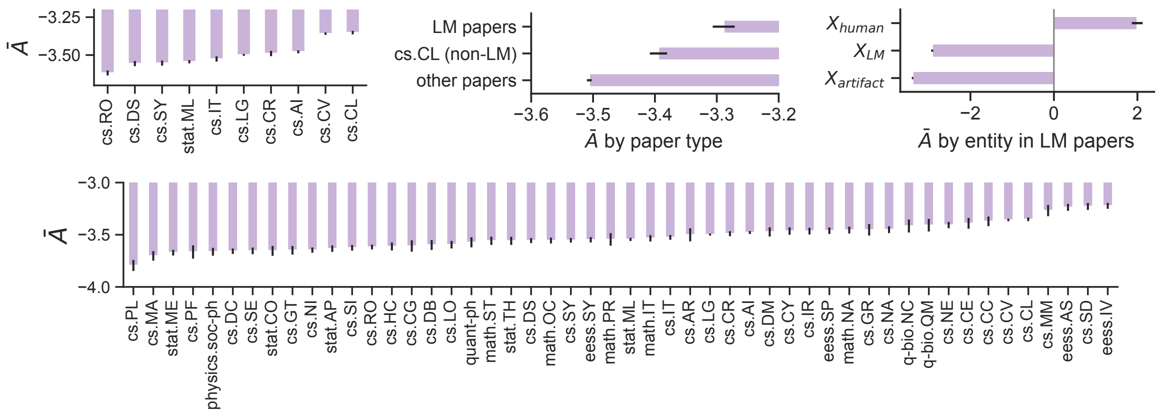


Figure 2: **Anthropomorphism is most prevalent in paper abstracts about computational linguistics, and language models.** Top left: Among the top 10 categories in CS/Stat arXiv, Computation and Language (cs.CL) has the highest average ANTHROSCORE ($\bar{A}$). Top middle: LM-related papers have higher scores of $\bar{A}$ than papers that do not mention LMs. Top right: Within LM papers, LMs are much more anthropomorphized than other technical artifacts, but do not have as high of a score as human entities do. Bottom: $\bar{A}$ for top 50 most popular categories in CS/Stat arXiv. There are categories outside of CS/Stat since many papers are cross-listed with other fields. Error bars indicate 95% CI.

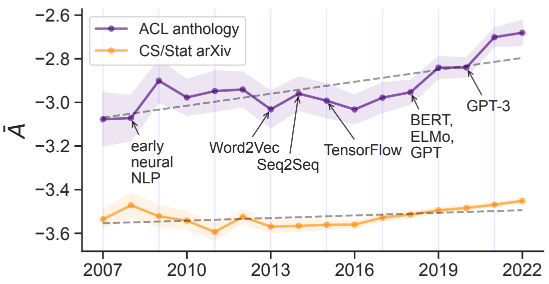


Figure 3: **Anthropomorphism is increasing over time.** In arXiv and ACL (orange and purple respectively), average ANTHROSCORE ($\bar{A}$) has increased in the past 15 years. In ACL papers, trends correspond with key advancements in neural models (annotated). Error bars indicate 95% CI. Straight line is least-squares linear fit.

$r = 0.63$, $p < 0.05$). We do not find a statistically significant temporal increase in the news headlines.

Within the ACL anthology, we see a correlation between increases in anthropomorphism and the introduction of artifacts that are widely acknowledged as marking paradigm shifts in NLP (Gururaja et al., 2023), such as early neural work and deep learning infrastructure (annotated in Figure 3, more details in Appendix D).

In the arXiv data, we find that among the top 10 categories, only machine learning (cs.LG) has a temporal increase within the subfield, while no other subfield has a statistically significant temporal correlation. This suggests that the increase in anthropomorphism is due to increases both in the sheer number of ML papers and in the anthropomorphic language *within* ML.

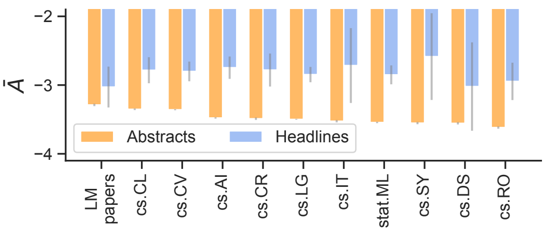


Figure 4: **News headlines anthropomorphize more than paper abstracts.** Anthropomorphism is more prevalent in news headlines than in research abstracts overall and for all of the top 10 arXiv categories, as well as in LM-related papers. Error bars indicate 95% CI.

4.3 News headlines anthropomorphize more than research abstracts.

News coverage of AI is rapidly increasing (Fast and Horvitz, 2017), motivating concerns of misleading anthropomorphism in public discourse. Our analysis of news headlines builds upon previous work on how news articles are crafted to be engaging (Gligorić et al., 2023) by exaggerating the strength of scientific claims and perpetuating misinformation (Sumner et al., 2014; Li et al., 2017; Horta Ribeiro et al., 2019; Wright et al., 2022; Hwang et al., 2023). Previous works focus on the difference in information communicated, while we focus on the *framing* of the information, which plays a critical role in readers' understanding (Lakoff, 2010).

We measure ANTHROSCORE in news headlines to see if they amplify anthropomorphism present in papers. We find that news headlines have higher rates of $\bar{A}$ than research paper abstracts (Figure 4).

Dataset	Top verbs for $S_\uparrow$ ($A > 1$)	Top verbs for $S_\downarrow$ ($A < -1$)
arXiv	achieve, **learn**, **guide**, show, embed, **fool**, find, **need**, **assist**, follow, **search**, **mislead**, inspire, win, demonstrate, **benefit**, try, **face**, deceive, plan, **make**, **steer**, generative, attempt, **retrain**, **train**, flow, weight, **require**, alternate, focus, **motivate**, experiment, tackle, **see**, hide, spiking, recommend, **discover**, participate, spike, **pass**, code, check, suggest, **decide**, interference, aim, move	**propose**, **present**, **outperform**, **develop**, **be**, **evaluate**, **improve**, **introduce**, **allow**, **use**, **compare**, **extend**, **implement**, give, **apply**, **consist**, **validate**, design, **yield**, analyze, **combine**, test, **leverage**, **deploy**, adapt, **build**, generalize, **enhance**, **devise**, **become**, optimize, reduce, derive, **utilize**, scale, study, **run**, modify, converge, illustrate, assess, **increase**, provide, contain, surpass, maximize, perform, complement, depend, simplify
News	say, hire, beat, encounter, fool	develop, use, build, be, create, introduce, help
ACL (unique)	provide, have, generate, create, parse, enable, suffer, construct, capture, obtain, fail, encourage, struggle, understand, help, do, select, extract, tend, predict, training, handle, lack, encode, deal, identify, ask, prevent, distinguish, model, establish, respond, ignore, report, inform, choose, interpret, recurrent, detect, seem	achieve, rely, explore, employ, show, adopt, investigate, include, demonstrate, submit, integrate, prove, augment, involve, participate, aim, tune, conduct

Table 2: **Top verbs for high- and low-scoring sentences.** All verbs displayed are statistically significant in frequency difference between $S_\uparrow$ and $S_\downarrow$ (z-score > 1.96 using the Fightin' Words method). $A = 0$ corresponds to an equal likelihood of being implicitly framed as human or as non-human, and $A = \pm 1$ corresponds to ≈ 2.7 times more likely human/non-human. For the arXiv and ACL datasets, > 100 verbs are statistically significant, and we display the 50 with the highest z-scores. **Bolded** verbs are also in the top 50 for ACL, and those unique to the top 50 for ACL are in the third row. Many verbs reflect the emotional, behavioral, and cognitive aspects of anthropomorphism.

We also compute $\bar{A}$ only among papers directly cited by news articles and find the same trend (Figure A4).

Trends on the category level within news headlines differ from the abstracts: unlike in the arXiv and ACL datasets, papers about LMs are not the most anthropomorphized, and there is no clear category that has highest $\bar{A}$ (Figure 4). This suggests that in public discourse, more general metaphors of human-like AI abound compared to academic papers, where LMs are, in contrast, disproportionately anthropomorphized.

5 Discussion

In this section, we first explore the underlying causes of anthropomorphism in text, including verb choice and norms of different academic fields. (We discuss other linguistic features of anthropomorphism in Appendix E.) Based on these observations, we then provide recommendations for individual authors and the broader community to avoid misleading anthropomorphism.

5.1 Verbs

First, we examine the verbs in sentences that contribute to anthropomorphism. This is inspired by previous work stating that NLP researchers tend to misleadingly state that LMs "understand" meaning (Bender and Koller, 2020), as well as the method of Connotation Frames, which use a lexicon of connotations for different verbs to measure social dynamics between entities (Sap et al., 2017; Antoniak et al., 2023). While our approach also operationalizes concepts closely related to agency and power like Connotation Frames, note that verbs that carry negative agency and power of an actor might still be evidence of anthropomorphism. For instance, describing an entity that "struggles" with a task is low in agency and power according to Connotation Frames, but high in anthropomorphism due to the implied affective state.

Thus, we explore the verbs that distinguish $S_\uparrow$ from $S_\downarrow$. We use the Fightin' Words method (Monroe et al., 2008) to measure statistically significant differences between the two sets after controlling for variance in words' frequencies (full details in Appendix E.3). In Table 2, we report top verbs. We find that many of the top verbs for $S_\uparrow$ can be categorized under one of the three aspects of anthropomorphism (Def. 1). For example, *suffer* and *struggle* suggest emotion; *learn, guide, fool, mislead, deceive, decide*, etc. imply cognitive abilities; and *steer, move, tackle*, etc. suggest human-like behaviors. *Understand* is a top verb only within the ACL dataset, connecting to Bender and Koller (2020)'s discussion of inaccurate claims in research papers about LLMs "understanding." The term "natural language *understanding*" has for many decades been the standard name for components of NLP related to semantics (Allen, 1995), reflecting how this anthropomorphic metaphor has long since permeated the field's vocabulary.

5.2 Disciplinary Norms

Our results show that anthropomorphism is embedded into the way that researchers conceptualize, discuss, and interact with their objects of study.

In NLP, for instance, evaluation benchmarks involve directly comparing LMs' performance to humans on cognition- and behavior-based tasks like answering questions and writing stories (Liang et al., 2023). The very idea of a chatbot inherently entails human-like conversational capabilities, and the concept of instruction-tuning builds upon this. Such LMs are not only designed to be prompted in human-like ways (Sanh et al., 2022) but often *require* anthropomorphic prompts to maximize performance: prompting with imperatives that imply cognitive or behavioral ability, e.g. "Think step-by-step" or "Imagine you are [x]" improves performance on a wide range of tasks (Wei et al., 2022; Cheng et al., 2023a,b). The outputs of instruction-tuned LMs contain anthropomorphism: ChatGPT's outputs frequently include variants of "I am a language model" that assign personhood to itself.

The community is caught in a double bind: although anthropomorphic metaphors of LMs facilitate misconceptions and other harms, these systems are built in ways that necessitate anthropomorphism from LM users. This paradox is tightly connected to the rise and prevalence of anthropomorphism in ACL and LM papers.

Similarly, *learn* (top verb for $S_{\uparrow}$) is often used in the context of AI/ML. In fact, the very names of these areas—"artificial *intelligence*" and "machine *learning*"—suggest distinctly human-like abilities. In this way, anthropomorphism is baked into the nature of these fields, fundamentally shaping the way that research is done. We hypothesize that as AI/ML have become more popular not only as fields but also as tools for other researchers, the language around their use has broadly percolated into the vernacular of other academic disciplines.

5.3 Recommendations

We provide recommendations, both on the individual level for authors who hope to minimize anthropomorphism in their writing as well as on the community level for ACL. First, authors should be careful about the verbs used, and how they may connote behavioral, emotional, and/or cognitive potential, especially when the subject of a sentence is a technical artifact. For example, the sentence "the model's performance is poor in X setting" connotes far less anthropomorphism than "the model *struggles* with X."

Second, our results call attention to the way that anthropomorphism shapes the norms of the ACL community. Like other initiatives for improving reproducibility and incorporating ethical considerations (Dodge et al., 2019; Rogers et al., 2021; Ashurst et al., 2022), we advocate for interventions to minimize misleading anthropomorphism, such as incorporating a disclosure about efforts taken to minimize anthropomorphism into the Responsible NLP Checklist filled by authors before submission, or adding anthropomorphism as a criterion for reviewers to evaluate.

6 Other Applications of AnthroScore

While we focus on anthropomorphism in research papers and downstream news articles, AnthroScore can be applied to many other settings, including other research areas, analyzing full papers and comparing across disciplines; perceptions of corporations and brands, which has political and legal implications (Ripken, 2009; Avi-Yonah, 2010; Plitt et al., 2015); conspiracy theories, which Douglas et al. (2016) link to anthropomorphism; and relationships with pets and objects (Mota-Rojas et al., 2021; Wan and Chen, 2021). AnthroScore is a first step toward analyzing anthropomorphism across different cultures, languages, and times. Leveraged in large-scale quantitative contexts, AnthroScore and its extensions facilitate deeper insights into human behavior.

Moreover, anthropomorphism is closely related to discussions of agency, human exceptionalism, and subjectivity (Bennett, 2010; Hodder, 2012; Latour, 2014). There is a rich literature on the implications of anthropomorphism in relation to biology and the natural world (Karadimas, 2012; DeMello, 2021; Hathaway, 2022). Also, feminist studies of science and technology have long leveraged anthropomorphism in their challenging of the dominant values and traditional boundaries between subject and object in science (Haraway, 1988; Longino, 1990; Suchman, 2008; Harding, 2013). AnthroScore enables engagement with these topics using a quantitative lens.

7 Limitations

Our analysis is limited to English data, where third-person singular pronouns mark animacy. However, many other languages have various grammatical markers of animacy (Comrie, 1989), to which our method can be extended to study how various cultural factors, societal values, and religious beliefs affect the tendency to anthropomorphize non-human entities, as well as the meaning and perception of anthropomorphism in different contexts (Inoue, 2018; Wood, 2019; Spatola et al., 2022).

Outputs from pre-trained MLMs only reflect the contexts and cultures of the models' training data, which does not reflect the diversity of the real world (Bender et al., 2021). In particular, our method implicitly relies on the idea that the distribution of the MLM has a representation of both "human" (from text that contains human pronouns) and "non-human" (from text that contains non-human pronouns). However, the definitions of these concepts are not static, and the MLM may only capture a subset of possible definitions. As the long literature on dehumanization shows, many people are not recognized as human in various ways: deprived of human rights, or not viewed and treated as fully human by society or in legal and state contexts. These phenomena are reinforced by language, as "the very terms that confer 'humanness' on some individuals are those that deprive certain other individuals of the possibility of achieving that status" (Butler, 2004). It is well-established that MLMs reflect social biases (Kurita et al., 2019; Guo and Caliskan, 2021; Mei et al., 2023), which also percolate into our measure. That being said, we focus on the anthropomorphism of objects and not the humanity of people, so these concerns should not affect the use of our metric.

Also, since anthropomorphizing metaphors are ubiquitous in English, it is inevitable that they are also embedded into the MLM's probability distributions; thus, the patterns of anthropomorphism that we uncover is a lower bound on the amount of anthropomorphism in the language of a text.

Acknowledgments

Myra Cheng is supported by an NSF Graduate Research Fellowship (Grant DGE2146755) and Stanford Knight-Hennessy Scholars graduate fellowship. Kristina Gligorić is supported by Swiss National Science Foundation (Grant P500PT-211127). Tiziano Piccardi is supported by Swiss National Science Foundation (Grant P500PT-206953). This work is also funded by the Hoffman–Yee Research Grants Program and the Stanford Institute for Human-Centered Artificial Intelligence. Fig. 1 icons from Flat Icons.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. TensorFlow: a system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283.

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, and Zeerak Talat. 2023. Mirages: On anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Euan Adie and William Roe. 2013. Altmetric: Enriching scholarly content with article-level discussion and metrics. *Learned Publishing*, 26(1):11–17.

Gabriella Airenti. 2015. The cognitive bases of anthropomorphism: from relatedness to empathy. *International Journal of Social Robotics*, 7:117–127.

James Allen. 1995. *Natural Language Understanding*. Benjamin-Cummings Publishing Co., Inc.

Maria Antoniak, Anjalie Field, Jimin Mun, Melanie Walsh, Lauren Klein, and Maarten Sap. 2023. Riveter: Measuring power and social dynamics between entities. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 377–388, Toronto, Canada. Association for Computational Linguistics.

Carolyn Ashurst, Emmie Hine, Paul Sedille, and Alexis Carlier. 2022. AI Ethics Statements: Analysis and lessons learnt from NeurIPS broader impact statements. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2047–2056.

Reuven S Avi-Yonah. 2010. Citizens United and the corporate form. *Wis. L. Rev.*, page 999.

Emily M. Bender. 2022. Resisting dehumanization in the age of "AI". Plenary talk at the 44th Annual Meeting of the Cognitive Science Society (CogSci).

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*, pages 610–623.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Jane Bennett. 2010. *Vibrant matter: A political ecology of things*. Duke University Press.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Judith Butler. 2004. *Undoing Gender*. Psychology Press.

Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31):e2120510119.

Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 2019. Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–32.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023a. Marked Personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023b. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore. Association for Computational Linguistics.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arXiv as a dataset.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, pages 160–167.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago Press.

Charles R Crowell, Jason C Deska, Michael Villano, Julaine Zenk, and John T Roddy Jr. 2019. Anthropomorphism of robots: Study of appearance and agency. *JMIR human factors*, 6(2):e12629.

Charles Darwin. 1905. *Journal of researches*. PF Collier.

Charles Darwin and Phillip Prodger. 1998. *The expression of the emotions in man and animals*. Oxford University Press, USA.

E Emory Davis and Barbara Landau. 2021. Seeing and believing: the relationship between perception and mental verbs in acquisition. *Language Learning and Development*, 17(1):26–47.

Margo DeMello. 2021. *Animals and society: An introduction to human-animal studies*. Columbia University Press.

Ameet Deshpande, Tanmay Rajpurohit, Karthik Narasimhan, and Ashwin Kalyan. 2023. Anthropomorphization of AI: Opportunities and risks. *arXiv preprint arXiv:2305.14784*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhopi Dhall and Saurajit Kanungo. 2023. Will AI take over the world? Or will you take charge of your world? *Forbes*.

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

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Karen M Douglas, Robbie M Sutton, Mitchell J Callan, Rael J Dawtry, and Annelie J Harvey. 2016. Someone is pulling the strings: Hypersensitive agency detection and belief in conspiracy theories. *Thinking & Reasoning*, 22(1):57–77.

Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4):864.

Ethan Fast and Eric Horvitz. 2017. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Gabriele Ferri and Inte Gloerich. 2023. Risk and harm: Unpacking ideologies in the ai discourse. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages 1–6.

Anita Fetzer. 2008. “And I think that is a very straightforward way of dealing with it” The communicative function of cognitive verbs in political discourse. *Journal of Language and Social Psychology*, 27(4):384–396.

Sigmund Freud. 1989. *The future of an illusion*. W. W. Norton & Company.

Raymond W Gibbs. 1994. *The poetics of mind: Figurative thought, language, and understanding*. Cambridge University Press.

Kristina Gligorić, George Lifchits, Robert West, and Ashton Anderson. 2023. Linguistic effects on news headline success: Evidence from thousands of online field experiments (registered report). *Plos one*, 18(3):e0281682.

David Gros, Yu Li, and Zhou Yu. 2022. Robots-Dont-Cry: Understanding falsely anthropomorphic utterances in dialog systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3266–3284, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133.

Sireesh Gururaja, Amanda Bertsch, Clara Na, David Widder, and Emma Strubell. 2023. To build our future, we must know our past: Contextualizing paradigm shifts in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13310–13325, Singapore. Association for Computational Linguistics.

Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies*, 14(3):575–599.

Sandra Harding. 2013. Rethinking standpoint epistemology: What is “strong objectivity”? In *Feminist epistemologies*, pages 49–82. Routledge.

Michael J Hathaway. 2022. *What a Mushroom Lives For: Matsutake and the worlds they make*. Princeton University Press.

Geoffrey Hinton. 2023. ‘Godfather of AI’ says AI could kill humans and there might be no way to stop it. *CNN*.

Ian Hodder. 2012. *Entangled: An archaeology of the relationships between humans and things*. John Wiley & Sons.

Manoel Horta Ribeiro, Kristina Gligoric, and Robert West. 2019. Message distortion in information cascades. In *The World Wide Web Conference*, pages 681–692.

David Hume. 1956. *The Natural History of Religion. Edited... by HE Root...* A. and C. Black.

Tatum Hunter. 2023. 3 things everyone’s getting wrong about AI. *Washington Post*.

Sohyeon Hwang, Emőke-Ágnes Horvát, and Daniel M Romero. 2023. Information retention in the multiplatform sharing of science. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 375–386.

Ken Hyland. 1998. Boosting, hedging and the negotiation of academic knowledge. *Text & Talk*, 18(3):349–382.

Cristina Yumie Aoki Inoue. 2018. Worlding the study of global environmental politics in the Anthropocene: Indigenous voices from the Amazon. *Global Environmental Politics*, 18(4):25–42.

Maria Kallery and Dimitris Psillos. 2004. Anthropomorphism and animism in early years science: Why teachers use them, how they conceptualise them and what are their views on their use. *Research in Science Education*, 34:291–311.

Dimitri Karadimas. 2012. Animism and perspectivism: Still anthropomorphism? on the problem of perception in the construction of amerindian ontologies. *Indiana*, 29:25–51.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. *arXiv preprint arXiv:2103.14659*.

Helen Kopnina, Haydn Washington, Bron Taylor, and John J Piccolo. 2018. Anthropocentrism: More than just a misunderstood problem. *Journal of Agricultural and Environmental Ethics*, 31(1):109–127.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

George Lakoff. 2010. Why it matters how we frame the environment. *Environmental Communication*, 4(1):70–81.

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

Madeline Lamo and Ryan Calo. 2019. Regulating bot speech. *UCLA L. Rev.*, 66:988.

Mark J Landau, Brian P Meier, and Lucas A Keefer. 2010. A metaphor-enriched social cognition. *Psychological bulletin*, 136(6):1045.

Bruno Latour. 2014. Agency at the time of the anthropocene. *New literary history*, 45(1):1–18.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2023. Multimodal foundation models: From specialists to general-purpose assistants. *ArXiv*, abs/2309.10020.

Mengjun Li and Ayoung Suh. 2021. Machinelike or humanlike? a literature review of anthropomorphism in AI-enabled technology. In *54th Hawaii International Conference on System Sciences (HICSS 2021)*, pages 4053–4062.

Yingya Li, Jieke Zhang, and Bei Yu. 2017. An NLP analysis of exaggerated claims in science news. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 106–111, Copenhagen, Denmark. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Helen E Longino. 1990. *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.

Tommy L. Lott. 1999. *The Invention of Race: Black Culture and the Politics of Representation*. Wiley-Blackwell.

Nathalie Maréchal. 2016. Automation, algorithms, and politicsl when bots tweet: Toward a normative framework for bots on social networking sites (feature). *International Journal of Communication*, 10:10.

Sean McManus. 2023. Can AI cut humans out of contract negotiations? *BBC*.

Witness Mdoka. 2022. Anthropomorphism and anthropocentrism: Human-animal ontology and the environment. *The Criterion: An International Journal in English*, 13.

Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. 2023. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1699–1710.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

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

Daniel Mota-Rojas, Chiara Mariti, Andrea Zdeinert, Giacomo Riggio, Patricia Mora-Medina, Alondra del Mar Reyes, Angelo Gazzano, Adriana Domínguez-Oliva, Karina Lezama-García, Nancy José-Pérez, et al. 2021. Anthropomorphism and its adverse effects on the distress and welfare of companion animals. *Animals*, 11(11):3263.

Rajiv Movva, Sidhika Balachandar, Kenny Peng, Gabriel Agostini, Nikhil Garg, and Emma Pierson. 2023. Large language models shape and are shaped by society: A survey of arXiv publication patterns. *arXiv preprint arXiv:2307.10700*.

Andreas Musolff. 2010. *Metaphor, nation and the holocaust: The concept of the body politic*, volume 3. Routledge.

Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103.

Gerald V O'Brien. 2003. Indigestible food, conquering hordes, and waste materials: Metaphors of immigrants and the early immigration restriction debate in the united states. *Metaphor and symbol*, 18(1):33–47.

Anna Papafragou, Kimberly Cassidy, and Lila Gleitman. 2007. When we think about thinking: The acquisition of belief verbs. *Cognition*, 105(1):125–165.

James W Pennebaker. 2011. The secret life of pronouns. *New Scientist*, 211(2828):42–45.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Mark Plitt, Ricky R Savjani, and David M Eagleman. 2015. Are corporations people too? the neural correlates of moral judgments about companies and individuals. *Social Neuroscience*, 10(2):113–125.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany. Association for Computational Linguistics.

Byron Reeves and Clifford Nass. 1996. The Media Equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10(10).

Susanna K Ripken. 2009. Corporations are people too: A multi-dimensional approach to the corporate personhood puzzle. *Fordham J. Corp. & Fin. L.*, 15:97.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'Just what do you think you're doing, Dave?' A checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in computational linguistics. *arXiv preprint arXiv:2305.14996*.

Meredith Root-Bernstein, Leo Douglas, Ashley Smith, , and Diogo Verissimo. 2013. Anthropomorphized species as tools for conservation: Utility beyond prosocial, intelligent and suffering species. *Biodiversity and Conservation*, 22:1577–1589.

Arleen Salles, Kathinka Evers, and Michele Farisco. 2020. Anthropomorphism in AI. *AJOB neuroscience*, 11(2):88–95.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Otto Santa Ana. 2002. *Brown tide rising: Metaphors of Latinos in contemporary American public discourse*. University of Texas Press.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Aaron Shackelford. 2010. Dickinson's animals and anthropomorphism. *The Emily Dickinson Journal*, 19(2):47–66.

Murray Shanahan. 2022. Talking about large language models. *arXiv preprint arXiv:2212.03551*.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.

Josepha Sherman. 2015. *Storytelling: An encyclopedia of mythology and folklore*. Routledge.

Ben Shneiderman. 2020. Design lessons from ai's two grand goals: human emulation and useful applications. *IEEE Transactions on Technology and Society*, 1(2):73–82.

Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press.

Nicolas Spatola, Serena Marchesi, and Agnieszka Wykowska. 2022. Different models of anthropomorphism across cultures and ontological limits in current frameworks the integrative framework of anthropomorphism. *Frontiers in Robotics and AI*, page 230.

Lucy Suchman. 2008. Feminist STS and the sciences of the artificial. *The handbook of science and technology studies*, 3:139–164.

Lucy G Sullivan. 1995. Myth, metaphor and hypothesis: How anthropomorphism defeats science. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 349(1328):215–218.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, et al. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ*, 349.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Natasha Tiku. 2022. The Google engineer who thinks the company's AI has come to life. *Washington Post*.

Caroline Tipler and Janet B Ruscher. 2014. Agency's role in dehumanization: Non-human metaphors of out-groups. *Social and Personality Psychology Compass*, 8(5):214–228.

Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38.

Echo Wen Wan and Rocky Peng Chen. 2021. Anthropomorphism and object attachment. *Current Opinion in Psychology*, 39:88–93.

Jing Wan, Katina Kulow, and Kirsten Cowan. 2022. It's alive! increasing protective action against the coronavirus through anthropomorphism and construal. *Journal of the Association for Consumer Research*, 7(1):81–88.

David Watson. 2019. The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds and Machines*, 29(3):417–440.

Adam Waytz, Nicholas Epley, and John T Cacioppo. 2010. Social cognition unbound: Insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science*, 19(1):58–62.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

Matthew Wood. 2019. The potential for anthropomorphism in communicating science: Inspiration from Japan. *Cultures of Science*, 2(1):23–34.

Dustin Wright, Jiaxin Pei, David Jurgens, and Isabelle Augenstein. 2022. Modeling information change in science communication with semantically matched paraphrases. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1783–1807.

Itai Yanai and Martin Lercher. 2020. The two languages of science. *Genome Biology*, 21:1–9.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Runting Zhong and Mengyao Ma. 2022. Effects of communication style, anthropomorphic setting and individual differences on older adults using voice assistants in a health context. *BMC Geriatrics*, 22(1):751.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

A Usage Examples

In this section, we provide examples of how to use AnthroScore in both scientific and non-scientific contexts. To use AnthroScore, the only information required is the set of texts T and the given set of entities X. Only the potentially-anthropomorphized entity is masked during the computation of AnthroScore.

Computer science example Suppose we are interested in measuring AnthroScore of "the machine learning model" in the sentence: "The machine learning model will start to become aware of the visual world." Then, we mask the term, resulting in the following sentence, "<MASK> will start to become aware of the visual world." We then compute AnthroScore for this sentence, as per equation (1).

Biology example Consider measuring how much the following text by Darwin anthropomorphizes tortoises:

> "One set eagerly travelling onwards with outstretched necks. Another set returning, after having drunk their fill. When the tortoise arrives at the spring, quite

> regardless of any spectator, he buries his head in the water above his eyes, and greedily swallows great mouthfuls, at the rate of about ten in a minute" (Darwin, 1905).

We know that the terms *set* and *tortoise* all refer to tortoises, so these are the entities X that we will mask. Our method works as follows:

1. Construct a dataset of sentences where X is masked. This results in three masked sentences:
   - <MASK> eagerly travelling onwards with outstretched necks.
   - <MASK> returning, after having drunk their fill.
   - When <MASK> arrives at the spring, quite regardless of any spectator, he buries his head in the water above his eyes, and greedily swallows great mouthfuls, at the rate of about ten in a minute.
2. Compute AnthroScore for each sentence, as per equation (1) on L211. This step is lexicon-free and does not depend on the choice of text or entity since we compare the probabilities of human vs. non-human pronouns replacing <MASK>.
3. Then, we take the average AnthroScore across the three sentences as a measure of anthropomorphism of tortoises in this text.

Poetry example Suppose we are interested in the anthropomorphism of birds in Emily Dickinson's poems. The input texts T are the poems, and the target entities X are words referring to birds like "bird," "hummingbird", "owl", etc. (Shackelford, 2010). Then, our method outputs ANTHROSCORE for each poem as well as each sentence mentioning a bird.

B Full lists of pronouns and entities

B.1 Pronoun Lists

For calculating P_{HUM} and P_{OBJ}, we use the following lists of pronouns:

Human pronouns: he, she, her, him, He, She, Her

Non-human pronouns: it, its, It, Its

Following Card et al. (2022), we only use pronouns that are in the tokenizer's vocabulary. We do not include low-frequency pronouns, such as reflexive and nonbinary pronouns, which could be added to make the model more complete.

Note that we only use third-person singular pronouns, which mark animacy in English. The pronoun "they/them" does not mark animacy; nonetheless, we still find that our metric works for plural entities.

B.2 Entity lists

To construct the dataset of LM papers, we use the following keyword list from Movva et al. (2023): {language model, foundation model, BERT, XLNet, GPT-2, GPT-3, GPT-4, GPT-Neo, GPT-J, ChatGPT, PaLM, LLaMA}.

X_{human} includes terms that refer explicitly to humans in the top 100 entities parsed from a random sample of papers (see details in the previous section), and also the terms in the "person" discursive category from Table 2 of Chancellor et al. (2019)'s study of "human" definitions in human-centered machine learning.

X_{human} ={humans, users, researchers, people, patient, victim, user, author, followers, poster, population, participant, subject, respondents, person, individual, she, he, woman, man, youth, student, worker, female, someone, peers, friends, others}.

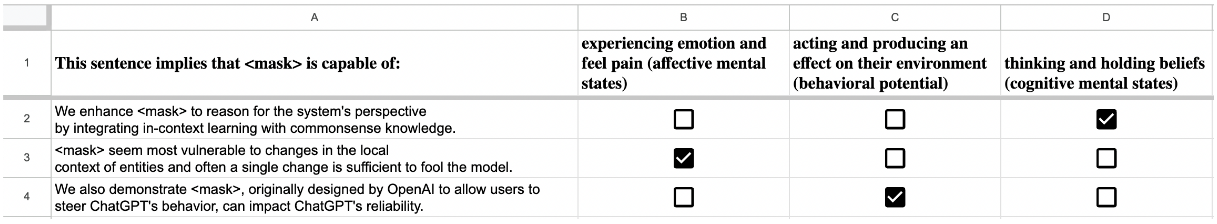
X_{LM} = {palm, lms, llama, transformers, language models, language model, gpt, plms, pretrained language models, gpt-2, xlnet, large language models, llms, gpt-3, foundation model, gpt-neo, gpt-j, chatgpt, gpt-4}.

C Further information about validity measures

C.1 Correlation with human perception

Domain knowledge was important for this task since the texts contain dense academic language, so we leveraged our expertise rather than crowdsourcing or otherwise recruiting participants. While we established correlation with two expert annotators, this may not represent general human perception; our method may require further validation in other contexts.

The 400 sentences include two sets of sentence: first, we use a randomly-sampled set of 300 masked sentences. We performed two rounds total of annotation (interface displayed in Figure A1) for this



	A	B	C	D
1	**This sentence implies that <mask> is capable of:**	**experiencing emotion and feel pain (affective mental states)**	**acting and producing an effect on their environment (behavioral potential)**	**thinking and holding beliefs (cognitive mental states)**
2	We enhance <mask> to reason for the system's perspective by integrating in-context learning with commonsense knowledge.	☐	☐	☑
3	<mask> seem most vulnerable to changes in the local context of entities and often a single change is sufficient to fool the model.	☑	☐	☐
4	We also demonstrate <mask>, originally designed by OpenAI to allow users to steer ChatGPT's behavior, can impact ChatGPT's reliability.	☐	☑	☐

Figure A1: Screenshot of interface for human annotators.

set. In each round, for each sentence, the annotators indicated whether the sentence implies that the masked term is capable of affective mental states, behavioral potential, or cognitive mental states (Def. 1). This was then aggregated into annotations of whether anthropomorphism is present. After the first round of annotation, there was a moderate agreement between annotators (Cohen's $\kappa = 0.40$). After discussing disagreements and re-annotating, we reached substantial agreement (Cohen's $\kappa = 0.87$).

To include more sentences with extreme sentences, we also use a stratified set of 100 masked sentences based on A score quartile. For this set, we had high agreement after the first round of annotation, so we did not discuss disagreements or reannotate. In Figure A2 we display the complete distributions of the A scores within the evaluated sets.

C.1.1 Nuances in the language of anthropomorphism

During the annotation process, ambiguities emerged and were discussed among authors. Here we list the main sources of disagreements:

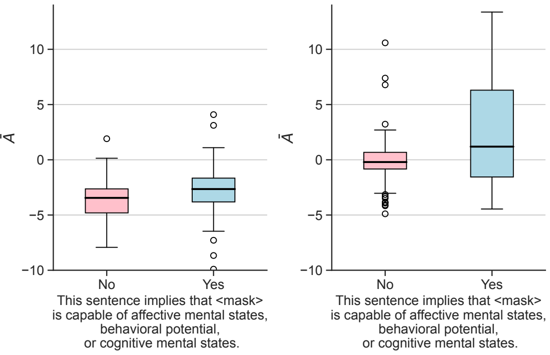
1. **Artifacts with affective or cognitive characteristics.** Within the same sentence, masked entities were at times simultaneously framed as tools and as entities that can display affective and cognitive abilities. While framing the entity as a tool implies a low level of anthropomorphism, subsequent descriptions of how such tools might be used can nonetheless imply human abilities. Such ambiguous framings were ultimately categorized as potentially implying behavioral potential, affective or cognitive mental states, even when described as tools.

2. **Popular and revolutionary artifacts.** Similarly, within the same sentence, masked entities were at times simultaneously framed as tools, and as entities with a behavioral potential to gain popularity or revolutionize a field. Since non-human entities might become popular, or in other ways affect the state of human affairs (e.g., a creative artifact such as a song can become popular), such ambiguous framings were categorized as not implying behavioral potential.

3. **Artifacts that can learn.** Lastly, a source of ambiguity was the fact that technological artifacts such as models are designed to learn patterns from datasets. While the goal of learning itself does imply a cognitive state, such statements mentioning learning in the specific context of capturing patterns present in the data were not classified as instances of anthropomorphism, since this is the purpose of the said entities. Note that this decision may differ from what ANTHROSCORE captures since *learn* is one of the top verbs for high-A sentences (Table 2), the implications of which we discuss in Section 5.

C.2 Correlation with LIWC Scores

Figure A3 reports $t-$test statistics for all dimensions of LIWC for which there is a statistically significant ($p < 0.01$) difference between $S_{\uparrow}$ and $S_{\downarrow}$. p is small and the test statistics are large, and



(a) (b)

Figure A2: Distribution of A scores in the two evaluated sets: random (left) and stratified (right).

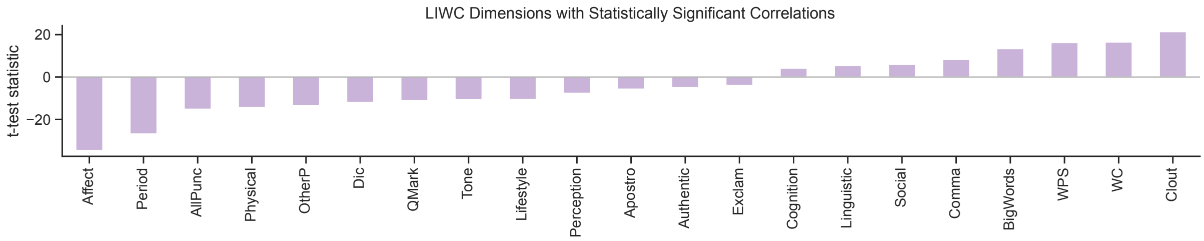


Figure A3: t-test statistics for LIWC Dimensions. Negative scores indicate that the value is higher in $S_\uparrow$ than in $S_\downarrow$, and positive scores indicate that the value is statistically significantly higher in $S_\downarrow$. All reported values are statistically significant ($p < 0.01$).

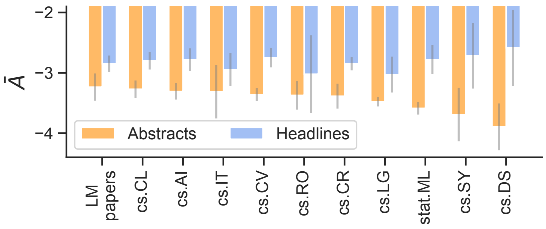


Figure A4: Rates of $\bar{A}$ among only the abstracts of research papers that are directly cited by downstream news articles whose sentences we use in our analysis. The trend is the same as in Figure 4. Error bars indicate 95% CI.

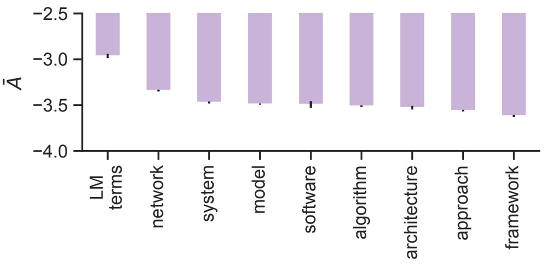


Figure A5: $\bar{A}$ **by entity.** The term "language model" is included under "LM terms" and not "model." Error bars indicate 95% CI.

our conclusions are robust to the choice of score threshold for $S_\uparrow$ and $S_\downarrow$.

D Further Details on History of NLP

In Figure 3, we annotate the graph using the release of particular landmarks that are determined by Gururaja et al. (2023) as important to paradigm shifts in NLP. First, Collobert and Weston (2008)'s paper on using neural networks for NLP shifted the community's perspective on neural models from skepticism and motivated work on early neural NLP, which led to widespread adoption. Word2Vec, Seq2Seq and Tensorflow were released in 2013, 2014, and 2015 respectively, facilitating a "neural revolution in NLP" (Mikolov et al., 2013; Sutskever et al., 2014; Abadi et al., 2016; Gururaja et al., 2023). The first LLMs (ELMo, GPT and BERT) were released in 2018 (Peters et al., 2018; Radford et al., 2019; Devlin et al., 2019). GPT-3 was released in 2020, which led to an even wider range of uses for LLMs (Brown et al., 2020).

E Linguistic Features of Anthropomorphism

E.1 Entities

Figure A5 shows $\bar{A}$ aggregated based on the specific entity masked, finding that LM-related terms have the highest rates of anthropomorphism.

E.2 Parts of Speech

Moreover, we find that 55% and 44% of $S_\uparrow$ and $S_\downarrow$ respectively are those in which the masked entity is the subject of the verb (rather than the object). In $S_\downarrow$, when the masked entity is the subject, it is often with intransitive verbs, which are less likely to suggest that the masked entity is exhibiting behavioral potential and directly acting upon another entity (the object of the sentence).

E.3 Top Verbs

To compute top verbs, we use the method described in Monroe et al. (2008) with the informative Dirichlet prior to compute the weighted log-odds ratios of verb frequencies between $S_\uparrow$ and $S_\downarrow$, using the sentences where $|A| < 0.5$ as the prior distribution. We find that using other thresholds, such as 0.2 or 0.7, for the prior distribution, does not affect the top verbs. This method provides a z-score, i.e. a measure of statistical significance, for each verb.

E.4 Cognitive Verbs

We further explore differences in verb frequency by drawing upon the literature on cognitive verbs (Papafragou et al., 2007; Fetzer, 2008; Davis and Landau, 2021) to build a lexicon of cognitive verbs

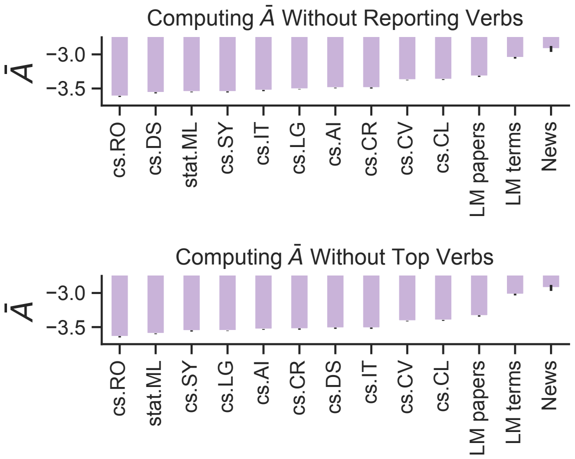


Figure A6: The patterns that we find (LM-related terms/papers and cs.CL papers have higher $\bar{A}$ than other papers, and news headlines have higher $\bar{A}$) hold even when we calculate $\bar{A}$ without reporting verbs (top) and without top verbs (bottom).

(know, think, believe, understand, remember, forget, guess, pretend, dream, mean, suspect, suppose, feel, assume). Using the weighted log-odds ratio method described in Section 5.1, we compute whether the differences in frequency for these words are statistically significant (z-score $>$ 1.96, which corresponds to a 95% CI.) We find that among these verbs, only *understand* is statistically significantly more frequent in $S_{\uparrow}$, while low-anthropomorphism verbs have statistically significant higher rates of the verbs *assume, know* and *mean*. Relatedly, we also find that *understand* occurs 1.7 times more frequently in LM-papers than in non-LM papers.

Note that we also explored using existing lexica for verbs related to agency, power, and emotion to measure anthropomorphism (Rashkin et al., 2016; Sap et al., 2017). However, these lexica did not seem appropriate for capturing anthropomorphism in this particular context of academic writing. For instance, many of the low-agency and low-power verbs suggest humanlike characteristics, such as *suffer*, while many high-agency verbs are ones that are frequently used in scientific writing as reporting verbs, such as *show* and *demonstrate*.

E.5 Reporting Verbs

Reporting verbs are a well-documented manner of anthropomorphism in scientific writing (Hyland, 1998): they are the verbs used by authors in phrases like "X demonstrates Y" to mean "*we* demonstrate Y using X." We found that reporting verbs alone do not explain the trends we document. We built a lexicon of reporting verbs based on existing literature (indicate, suggest, show, demonstrate, support, confirm, add, argue, agree, warn, advise, prove, claim, find, declare, express, conclude, study, admit, assure, justify, emphasize, assert, accept) and find that our trends hold even when we remove sentences with reporting verbs from our dataset (Figure A6). Thus, ANTHROSCORE captures patterns beyond the presence of reporting verbs, which are extremely common in paper abstracts.