AnthroSet: a Challenge Dataset for Anthropomorphic Language Detection

Dorielle Lonke[1] **Jelke Bloem**[1] **Pia Sommerauer**[2]
[1]University of Amsterdam, Institute for Logic, Language and Computation
[2]Vrije Universiteit Amsterdam, Computational Linguistics and Text Mining Lab
dorielle.lonke@student.uva.nl j.bloem@uva.nl pia.sommerauer@vu.nl

Abstract

This paper addresses the challenge of detecting anthropomorphic language in AI research. We introduce AnthroSet, a novel dataset of 600 manually annotated utterances covering various linguistic structures. Through the evaluation of two current approaches for anthropomorphism and atypical animacy detection, we highlight the limitations of a masked language model approach, arising from masking constraints as well as increasingly anthropomorphizing AI-related terminology. Our findings underscore the need for more targeted methods and a robust definition of anthropomorphism.

1 Introduction

With the evolving popularity and applications of AI systems, the terms used to describe their functionalities have become increasingly anthropomorphizing (Floridi and Nobre, 2024). The tendency to attribute human-like capabilities and properties to AI systems involves various topics of interest, including cognitive and psychological analyses (Waytz et al., 2010; Hofstadter, 1995), ethical considerations and accountability (Salles et al., 2020), and undue AI hype (Placani, 2024; Barrow, 2024).

While the topic of anthropomorphism in AI is widely discussed, there is not one clear definition of what it entails. Efforts to describe anthropomorphic language focus on AI output rather than texts about AI (for examples see DeVrio et al. (2025); Emnett et al. (2024)). Detecting anthropomorphic language in human text is particularly difficult as it is highly contextual (Cheng et al., 2024), ambiguous and subjective (Waytz et al., 2010; Shardlow et al., 2025). There are currently only two open-source implementations for detecting the attribution of human properties to machines in text, both relying on a masked language model (MLM) approach that detects anthropomorphism by measuring the animacy of a masked entity (Coll Ardanuy et al., 2020; Cheng et al., 2024).

We present AnthroSet, an evaluation dataset consisting of 600 manually annotated utterances representing types of anthropomorphic language pertaining to AI. We provide a variety of linguistic structures in which anthropomorphic language is expressed, drawn from academic literature on AI. The purpose of this dataset is twofold: first, we aim to provide concrete examples of anthropomorphic language in contemporary AI research, grounded in a linguistic analysis of anthropomorphism and animacy markers in the English language, as well as the anthropomorphic language taxonomy by DeVrio et al. (2025). The second is to evaluate the state-of-the-art, open-source methods for anthropomorphic language detection.

Our results highlight the problems with employing a masked language model approach for this task. For one, the masking is consequential in achieving good results, but a uniform masking approach is not suitable for all syntactic structures. Second, as AI-related terminology becomes increasingly anthropomorphizing, MLMs are more likely to associate AI entities with anthropomorphic verbs and descriptors, simply due to their reliance on statistical co-occurrence (Zhang et al., 2024), posing further challenges for anthropomorphism detection.

2 Related Work

The tendency to attribute human-like capacities to AI systems has been observed since the foundation of AI as a field of research. The general relation between cognition and machines has been widely discussed, with authors such as Dreyfus (1976) and Searle (1980) arguing against the reduction of human thought to syntactic and symbolic programs. In the context of psychology, the *ELIZA* effect was defined as the cognitive bias that causes human

Proceedings of the First Interdisciplinary Workshop on OMMM associated with RANLP 2025,
pages 27–39, Varna, Bulgaria, Sep 8-10, 2025.
https://doi.org/10.26615/978-954-452-101-1-003

users to attribute human-like properties such as intelligence and emotions to responsive machines (Hofstadter, 1995). The tendency to anthropomorphize AI by means of the language we use was recognized by McDermott (1976) as *wishful mnemonics* – the methodological tendency to name and describe AI programs not in terms of what they actually do, but as what they are intended and willed by us to do. Anthropomorphism in AI can be seen as a metaphoric device, whose explanatory powers contribute to the evolution of emerging technologies (Carbonell et al., 2016), and are used both for explanation as well as persuasion (Rossi and Macagno, 2021). In recent years, anthropomorphic language in AI discourse has been addressed from an ethical perspective, touching on issues related to society and accountability (Watson, 2019; Salles et al., 2020; Placani, 2024).

There currently are two open-source implementations of anthropomorphism detection: Cheng et al. (2024) developed *AnthroScore*, a metric for measuring implicit anthropomorphism in contemporary scientific research and downstream media. Their approach is similar to the one presented by Coll Ardanuy et al. (2020) in *Living Machines: A study of atypical animacy*, which aims at detecting atypical animacy by focusing on scenarios in which machines are represented as having animate attributes. Recently, DeVrio et al. (2025) proposed a taxonomy of linguistic expressions in AI-generated text that contribute to anthropomorphism in AI, setting forth a theoretical baseline for anthropomorphic language analysis.

Shardlow et al. (2025) present the first corpus annotated for anthropomorphic language in the context of LLMs[1]. Their corpus is based on abstracts from the ACL Anthology and news articles, annotated at the sentence level. Their annotation focuses on classifying claims as non-anthropomorphic, ambiguously anthropomorphic or explicitly anthropomorphic as outlined in Shardlow and Przybyła (2024), following the subjective judgements of annotators. No annotation guidelines are available, but the intermediate category seems to be defined as the case where "someone who is familiar with this language would correctly interpret it as a metaphor, whereas a novice or lay reader may well infer human characteristics". The scheme is not otherwise defined in linguistic terms. 4340 sentences were annotated. They also perform scoring using encoder LLMs such as XLNet with a regression classifier head tuned on labeled data, though these models are not available at the time of writing.

[1]This work was published after we finished our study.

3 Linguistic Structures

Anthropomorphism, particularly pertaining to AI and machines, can be expressed through a variety of different syntactic and semantic structures. We differentiate between explicit anthropomorphism, i.e. sentences or expressions that directly and overtly attribute human-like capacities such as cognition, intention or mental states to AI systems through their contents, and *implicit* anthropomorphism – which is indirect, sometimes covert, and rises from certain lexical or contextual meanings. We identified prominent structures on the basis of a linguistic analysis of anthropomorphism and animacy markers in the English language, combined with a frame semantics approach that considers the lexical units in the sentence with respect to their thematic roles and the frames that they evoke (see Ryazanov et al. (2024)). For example, in the sentence 'The system decides to trust the user', the entity in the subject position ('system') is anthropomorphic as it plays the thematic role of AGENT, whose properties are sentience, volition, movement, causing an event or change of state, and existing independently of the event (Dowty, 1991; Levin, 2022). Additionally, the verb phrase 'decide to trust' is anthropomorphic as it entails the capacity for cognitive processes such as decision making and the mental state of trust. Thematic agents can occur in the object position in passive voice structures. For instance, in the sentence 'The users were deceived by the model' the verb frames the AI entity as having intention or malevolence. The AI entity can also embody the thematic role of EXPERIENCER, attributed cognitive and mental states as either subject or object of certain cognitive or *psych verbs* (Belletti and Rizzi, 1988). For example, in 'The developers tricked the system into believing the lies', the object-experiencer verb 'trick' contributes to the framing of the AI entity as having cognitive and mental faculties, suggesting it has the capacity to be tricked.

Importantly, not all anthropomorphic lexical units are verbs: adjectives can attribute human-like abilities by means of description, e.g. *conscious*, *aware*, *confident*, *benevolent*, and *malicious*; certain nouns which are often collocated with AI entities are otherwise traditionally reserved for human

roles, such as *assistant*, *teacher* or *judge*. We might also identify anthropomorphism in sentences that embody genitive structures in which an AI entity is described as possessing certain abilities, traits or properties, e.g. 'the model's advanced reasoning abilities', or contain comparative function words, e.g. 'Like children, language models learn from patterns'. While syntactic in nature, these structures are best understood alongside a taxonomy of anthropomorphic lexical units and their semantics, which we have defined on the basis of the one constructed by DeVrio et al. (2025). Based on their guiding lenses for identifying anthropomorphic patterns in synthetic text, as well as an analysis of numerous real-life examples from published papers in AI, we identified the affective and cognitive capacities aimed at elucidating anthropomorphic language in human-written text, used by human authors to describe AI systems in contemporary AI research. This taxonomy is shown in Appendix C.

4 Task and Models

We aim to shed light on the current definitions and interpretations of anthropomorphic language in AI research, and the means for identifying it in text. To that end, we evaluated and compare two implementations of anthropomorphism detection in the domain of AI and machines. We compiled and manually annotated an evaluation set consisting of examples of anthropomorphic language in the context of AI, i.e. language that humanizes AI systems by attributing to them human-like capacities of cognition, intention and mental states, and compare and examine the performance of each approach in detecting these patterns[2].

4.1 Models

Both approaches rely on a masked language model to predict the likelihood of a masked entity, corresponding to an AI model, system or machine to be construed as human. The AnthroScore method uses the HuggingFace implementation of RoBERTa (`roberta-base`, 125M parameters), a pre-trained masked language model (MLM) as the model and tokenizer. The *Living Machines* method (henceforth referred to as *AtypicalAnimacy*) is based on the the HuggingFace implementation of BERT (`BERT-base`, 110M parameters), fine-tuned on an atypical animacy detection dataset consisting of 19th-century texts related to industrialization and machines. The AnthroScore method provides a metric for measuring the degree of anthropomorphism in a given set of texts for a given set of entities. Given a sentence containing a masked entity, a high- or low-anthropomorphism score is obtained by computing the probabilities that the MLM predicts animate pronouns (*he*, *she*) and inanimate ones (*it*, *its*), and calculating the log of the ratio between the probabilities. AtypicalAnimacy also rely on MLM prediction of a masked token, determining the animacy of the expression within a sentence by averaging the animacy of the top predicted tokens. This is determined using WordNet, by disambiguating the predicted token to its most relevant word sense, and checking whether that sense is a descendant of the *living thing* node. Then, a score between 0 and 1 is produced by calculating the weighted average of the predicted token scores, and a final binary score is determined by an optimal animacy threshold.

5 AnthroSet

Our evaluation set consists of sentences taken from abstracts of papers published on ACL Anthology and arXiv. Relevant papers were selected from the ACL Anthology[3] and arXiv[4] datasets, using a list of keywords (*AI*, *artificial intelligence*, *(language) model*, *system*, *LM*, *LLM*, *GPT*, *ChatGPT*). First, we identified relevant papers by searching for the keywords in the title. Then, we found potentially anthropomorphic utterances by searching for sentences containing these keywords in the abstract. To narrow down the search, we compiled word lists of anthropomorphic verbs, nouns and adjectives, corresponding to our taxonomy of anthropomorphic attributes (Appendix C). These lists were then extended with similar words using WordNet to include synonyms and semantically related entries.

We included samples covering all linguistic structures described in section 3, which are henceforth referred to as follows: (1) *verb subjects* – an

[2]The phenomenon addressed in *Living Machines* pertains to a general sense of animacy, which encompasses the more specific notion of *humanness*. This specification is used to distinguish between sentences describing the humanization of machines through comparisons to humans, which are examples of both *animacy* and *humanness*, versus those depicting the dehumanization of humans through comparison to machines, which corresponds only to *animacy*. Since our dataset focuses on machines and AI and not humans, we interpret the *Living Machines* notion of animacy as equivalent to AnthroScore's definition of anthropomorphism.

[3]https://acl-anthology.readthedocs.io/latest/api/anthology/

[4]https://www.kaggle.com/datasets/Cornell-University/arxiv

AI entity as the subject of an anthropomorphic verb, (2) *verb objects* – an AI entity as the object of an anthropomorphic verb, (3) *adjectives* – an AI entity collocated with an anthropomorphic adjective, (4) *role/function noun phrases* – an AI entity described as performing an anthropomorphic role or function[5], (5) *genitive noun phrases* – an AI entity described as being in possession of an anthropomorphic NP, and (6) *comparisons* of AI entities to human beings. An example of each of these structures is shown in Appendix B.

For each linguistic structure, we searched for the particular dependency relations between the lexical unit and the AI entity. For example, to find anthropomorphic adjectives, we looked for AI entities that are either modified by an `amod` or complemented by a `acomp` which belongs to the extended list of anthropomorphic adjectives. We then manually reviewed and selected the candidate sentences based on our annotation guidelines (Appendix A), modeled in part after the VU Metaphor Identification Procedure (Steen et al., 2010). Since we queried for different dependency relations, we ended up with a pooled dataset divided into subsets categorized by their syntactic structures.

5.1 Annotation procedure

The linguistic category sets are divided into multiclass (*verb subjects*, *verb objects* and *adjectives*) which have positive, negative and inconclusive samples, and single-class, which are either always positive (*role/function NPs*, *genitive NPs*) or always inconclusive (*comparisons*). While verbs and adjectives tend to be much more context-sensitive and ambiguous, structures describing AI systems as performing a role or in possession of certain properties are anthropomorphic only to the extent that they feature a anthropomorphic NP. In that respect, negative samples are not clearly defined, and thus were not included in the evaluation set. As a result, we excluded these categories from the overall comparison, which is done in terms of precision, recall and F1-score, and only measure accuracy on these sets. Comparisons in which AI entities are likened to humans can be either understood as highly anthropomorphizing as their content attributes to AI qualities or properties of humans, or they could be seen as non-anthropomorphizing since the explicit comparison serves to contrast AI and humans, and highlight their differences (Coll Ardanuy et al., 2020). Because of this dual interpretation, we decided to treat these cases as inconclusive, and included them in the evaluation only as an aid for understanding model behavior[6].

For the annotation task, annotators were presented with batches of sentences where the target AI entity was highlighted in bold, along with our guidelines and a decision tree (given in Appendix A). The labels 'positive', 'negative' or 'inconclusive' were used to label the anthropomorphization of the target AI entity in the context of the sentence, following this decision tree.

Our annotators have expertise in linguistics and were aware of the research purpose of the benchmark. All instances were annotated by a primary annotator and to evaluate inter-annotator agreement, a subset of 20% of the multiclass cases was divided among two secondary annotators. The first set, which had a balanced distribution of positive, negative and inconclusive cases had a Cohen's κ of 0.39 for all cases, and a much higher κ of 0.92 for just positive and negative cases. The second set, which consisted of twice as many inconclusive cases than positive and negative cases had a Cohen's κ of 0.22 for all cases, and 0.60 on just the positive and negative cases[7]. The low κ for the overall cases reflects the difficult nature of this annotation task, especially on borderline cases which do not have enough contextual cues, even for human annotators, to determine whether or not an entity is being anthropomorphized. Additionally, while we relied on a taxonomy of anthropomorphic language, deciding whether a certain lexical unit embodies these definitions is not a trivial task. Nevertheless, the Cohen κ for our non-borderline cases shows that these were for the most part agreed upon. No disagreement resolution was performed.

To support future work, including robust re-

[5]This definition resonates with *task-based anthropomorphism* (Ryazanov et al., 2024), a form of anthropomorphic descriptions of AI systems which pertains to humanizing language describing functionality.

[6]In some interpretations of anthropomorphism, noun phrases such as *AI teacher* or *AI judge* might not be seen as inherently anthropomorphizing, rather understood as comparisons in which AI is likened, but not identified with humans. Based on our definition of anthropomorphism, we have decided to treat these cases as positive.

[7]This was checked by first including all cases, and then filtering out cases in which at least one of the annotators was inconclusive. We also calculated Cohen's κ for each class by creating a binary mapping, and had $\kappa = 0.62$ for positive cases and $\kappa = 0.49$ for negative cases in the first set, and $\kappa = 0.40$ for positive cases and $\kappa = 0.22$ for negative cases in the second set. Inconclusive cases had a very low agreement rate due to their borderline nature, but this was expected.

dundant annotation and expanding the dataset, we made our annotated set publicly accessible on GitHub[8]. The annotated dataset contains 297 (49%) positive, 173 (29%) negative and 131 (22%) inconclusive cases. This contrasts with the corpus of Shardlow et al. (2025), who found 3.7% explicit anthropomorphism, 19.3% ambiguous anthropomorphism and 77% negative cases. However, we specifically selected sentences containing potentially anthropomorphic language to create a benchmark, while their corpus aims to document the frequency of anthropomorphic language in news articles and ACL abstracts and thus covers a subset of data without selection or filtering.

6 Experiments

We employed two masking strategies in our experimental setup. The first is AnthroScore's built-in masking method, which relies on keyword identification and noun-chunking. We found that while it is suitable for identifying certain structures, particularly those in which the anthropomorphic component complements or predicates the AI entity, it tends to mask crucial contextual cues for other anthropomorphic structures, such as adjectival modifiers, noun phrases, or certain possessive structures. For example, the phrase 'conscious AI systems' is masked in its entirety by AnthoScore's masking strategy, even though the main contribution to anthropomorphism is the adjectival modifier 'conscious'. The second is our own masking strategy (referred to henceforth as *minimal entity* masking), which was put forth in order to preserve the anthropomorphic cues in the context rather than mask them. Our masking strategy works as follows: given an AI keyword (a single keyword such as *AI*, *LLM*, *model*, or *ChatGPT*), we manually masked the minimal phrase referring to an AI entity[9], masking additional modifiers only in case they are part of the name, or an essential part of its description, e.g. relating to its functionality or purpose (i.e. *conversational AI*, *question answering system* or *large language model*). We left out any descriptors that are contingent to the description, such as *powerful*, *complex*, or *flexible*.

[8] https://github.com/doriellel/anthroset

[9] Our masking strategy required manual revision, but proved significantly better than the automatic chunking method employed by AnthroScore. In future work, this could be improved by implementing something like a NER pipeline that would identify particular AI entities, rather than capturing an entire noun chunk or manually reviewing every occurrence.

6.1 Metrics and score mapping

We evaluated each system on the multiclass sets (*verb subjects*, *verb objects* and *adjectives*) in terms of precision, recall and F1-score. We observed these both as macro-averaged aggregates for each syntactic category, as well as per class. On the single-class positive sets (*role/function NPs* and *genitive NPs*), we only looked at the systems' recall (i.e. accuracy – the number of positive predictions out of total predictions). In the case of all inconclusive sentences (*verb subjects*, *verb objects* and *adjectives* and *comparisons*), since 'inconclusive' does not represent a gold label but rather a lack thereof, we did not measure accuracy. Instead, we observed the trends, and compare each system's tendency to predict positive, negative (and inconclusive in the case of AnthroScore) in those cases.

To compare the performance of both approaches, we mapped the AnthroScores to those of AtypicalAnimacy. AnthroScore does not provide a binary score, but rather high-anthropomorphism and low-anthropomorphism scores. A high score is higher than 1 (i.e. the probability to predict human pronouns is higher than non-human ones, resulting in the log of the ratio to be greater than 1), and, symmetrically, a low score is lower than -1. Scores that fall between 1 and -1 reflect an equal likelihood for both pronouns to be predicted by the MLM, corresponding to our definition of inconclusive cases. AtypicalAnimacy provides binary scores of 1 and 0. To obtain binary results for AnthroScore as well, we mapped AnthroScores >1 to 1, and scores $<$-1 to 0, and conduct the evaluation after the mapping. To compare precision, recall and F1, we simply interpreted AnthroScores between 1 and -1 as false negatives of either class, and exclude inconclusive cases from the gold set.

6.2 Evaluation results

Each method was evaluated twice on all six categories of syntactic structures, once for each masking strategy. The first experiment made use of AnthroScore's masking strategy. First, a set of sentences alongside a list of all AI entities in that set were inputted to the AnthroScore model. AnthroScore reports an average over entities in the sentence, but we are only interested in our annotated target entities. Therefore, instances where the model masked other components than the target AI entity, or partially masked it, were manually removed. Cases of over-masking, i.e. masking cru-

	AnthroScore			*AtypicalAnimacy*		
Category	Precision	Recall	F1-Score	Precision	Recall	F1-Score
AnthroScore masking						
verb subjects	0.527	0.341	0.318	0.767	0.748	0.745
verb objects	0.548	0.370	0.334	0.803	0.803	0.803
adjectives	0.515	0.356	0.299	0.769	0.694	0.673
Minimal entity masking						
verb subjects	0.490	0.289	0.305	0.871	0.860	**0.862**
verb objects	0.389	0.250	0.293	0.805	0.803	**0.804**
adjectives	0.351	0.243	0.256	0.796	0.730	**0.704**

Table 1: Macro-averaged precision, recall, and F1 scores for AnthroScore and AtypicalAnimacy across the multiclass categories: *verb subjects*, *verb objects*, and a*djective phrases*, comparing the AnthroScore masking strategy and our minimal entity masking strategy. In this comparison, inconclusive sentences in gold were excluded.

cial contextual information, were kept, as long as the AI entity was masked fully, as these were most of the cases. After filtering the results, we provided the AtypicalAnimacy model with AnthroScore's masked sentences, alongside the previous and next sentences (if existing) from the original abstract they were taken from, and obtained the AtypicalAnimacy scores on those sentences. Even with AnthroScore's masking strategy, AtypicalAnimacy outperformed AnthroScore across all sets.

The second experiment relied on our minimal entity masking strategy, and both methods were provided with pre-masked sentences, with the additional context of the previous and next sentences for the AtypicalAnimacy model.

For the multiclass sets, we compared the performance of AnthroScore and AtypicalAnimacy on only positive or negative cases (Table 1), using macro-averaged precision, recall and F1-score. Overall, the AtypicalAnimacy model performed better across all multiclass datasets. Additionally, using our minimal entity masking strategy improved its performance, resulting in the highest precision, recall and F1-score among all four experiments. In the case of AnthroScore, our masking strategy slightly reduced the performance, most likely because it is not always compatible with pronoun replacement. Both models performed best on anthropomorphic structures in which the anthropomorphic component is a verb – the highest F1-score is obtained for the *verb objects* category in the first experiment, and for the *verb subjects* category in the second experiment.

For the single-class positive sets, we compared the recall of both methods using both masking strategies (Table 2)[10]. Both models exhibited low accuracy rates for the *role/function NPs*, with a slight improvement using AnthroScore's masking strategy. AnthroScore exhibited low accuracy rates also for the *genitive NPs* sets across both experiments. The notable improvement provided by our masking strategy, particularly for possessive noun phrases is reflected in AtypicalAnimacy's much higher accuracy (0.783) in the second experiment.

Category	***AnthroScore***	***AtypicalAnimacy***
AnthroScore masking		
role/function NPs	0.106	**0.470**
genitive NPs	0.018	0.298
Minimal entity masking		
role/function NPs	0.086	0.200
genitive NPs	0.117	**0.783**

Table 2: Accuracy scores for AnthroScore and AtypicalAnimacy for the single-class positive sets: *role/function NPs* and *genitive NPs*.

To obtain a better understanding of each method's performance, we compared precision, recall and F1-scores per class (Table 5 in the appendix), since the aggregate scores are skewed by AnthroScore's preference towards negative scores. Anthroscore has perfect precision rates for all three categories when using its own masking strategy, but this is because it rarely labels cases as positive, and as a result does not predict any false positives, and similarly has a very high recall for negative cases. Its real-world ability to detect anthropomorphism on varying syntactic structures is quite low, reflected by its low recall rates for all three positive sets in both experiments.

Compared to AnthroScore, AtypicalAnimacy's precision and recall are significantly more balanced.

[10]When there is one class, this is equivalent to accuracy.

To maintain a fair evaluation of AnthroScore, which, unlike AtypicalAnimacy, predicts inconclusive scores as well, we show the improvement in AnthroScore's metrics when the inconclusive cases are included in the evaluation (Table 7 in the appendix). The F1-score increased on all categories using both masking strategies. Nevertheless, the improved scores still do not surpass those of the AtypicalAnimacy model.

Finally, we include the prediction trends of both methods for all inconclusive cases (Table 6 in the appendix). Overall, AnthroScore is unlikely to provide a positive (i.e. high-anthropomorphism) score, with an average of 0.06 positive scores for the first experiment, and 0.12 in the second experiment. AtypicalAnimacy is more likely to provide a positive score for borderline cases, but not overwhelmingly so – with an average of 0.419 positive predictions in the first experiment, and 0.480 in the second experiment. AtypicalAnimacy's tendency to output positive scores about half the time is aptly consistent with the definition we used for inconclusive cases (aligning with that of AnthroScore) – i.e., cases which cannot be determined on context alone, or have conflicting contexts, such that when masking the AI entity, it is equally likely to be construed as human and non-human.

7 Discussion

The AnthroScore model fared worse than the AtypicalAnimacy model in all categories. Multiple occurrences of AI entities and co-reference patterns with pre-existing inanimate pronouns likely contributed to the high amount of false negatives in the case of AnthroScore. This might be explained by the constraints imposed by design in the AnthroScore MLM prediction approach, which limits the predictions to pronouns. In contrast, AtypicalAnimacy allows for the substitution of a masked entity with any token and performs an additional disambiguation step to obtain precise results. The AnthroScore masking strategy, which masks an entire noun phrase containing an AI keyword, is highly compatible with pronominalization. This is useful for anthropomorphism detection in cases where the verb contributes the most to the anthropomorphism, but is costly in terms of the contextual information that is lost when important components are masked. This strategy is therefore not effective for syntactic structures in which a noun or adjective modifier is the main source of anthropomorphism.

Generally speaking, the masking approach works best for verb based structures, as verbs are guaranteed to remain unmasked, and provide significant contextual information about its arguments. Also, masked language models such as BERT are sensitive to the semantic roles represented by verbs (Ettinger, 2020), which are highly relevant in the context of animacy and anthropomorphism (Primus, 2012). This is reflected in the improved performance on the verb categories for both models in both experiments. In a similar vein, both models were more likely to give positive scores for structures containing predicative adjectives (complements, `acomp`, e.g. *the model is smart*) than for sentences containing attributive adjectives (adjectival modifiers, `amod`, e.g. *the smart model*). This was particularly exacerbated with the AnthroScore masking strategy, in which the adjective was masked along with the noun phrase.

In the case of role or function and genitive structures, both models exhibited reduced accuracy, with AnthroScore performing clearly worse. With AnthroScore's masking strategy, the main contribution to the anthropomorphism was entirely masked. With our masking strategy, the resulting masked expression yielded a syntactic configuration that was incompatible with pronouns altogether, e.g. '[MASK]'s cognitive abilities' (Table 3). The case of role/function NPs is especially problematic, resulting in masked expressions such as 'the [MASK] companion', which is also very limiting for AtypicalAnimacy, even though it is not constrained to pronouns. This led to decreased performance on the *role/function NPs* set in experiment 2 for both models, and low accuracy overall.

Our results suggests that noun phrase expressions are simply incompatible with a detection method based on MLM predictions, whether or not they are set to predict pronouns or generally animate entities[11]. In contrast, in the case of genitive structures our masking strategy resulted in a clear improvement only for the AtypicalAnimacy model. AnthroScore's masking algorithm, which is based on identifying an AI keyword within a noun chunk, recognizes a possessive expression such as 'ChatGPT's cognitive abilities' as the entire noun

[11] An alternative interpretation of these results is that nouns such as *companion*, *teacher* or *coach* are not as anthropomorphizing as verbs or adjectives. By changing the gold labels we may extract different insights with regard to the accuracy of the models. Since we do not aggregate the scores across all linguistic structures, this decision does not influence the model's performance metrics for the other categories.

Sentence	AnthroScore Mask	Our Mask
Departing from conventional practices of employing distinct models for image recognition and text-based coaching, our integrated architecture directly processes input images, enabling natural question-and-answer dialogues with **the AI coach**.	the AI coach	AI
This research sheds light on the collaborative synergy between human expertise and AI assistance, wherein **ChatGPT's cognitive abilities** enhance the design and development of potential pharmaceutical solutions.	ChatGPT's cognitive abilities	ChatGPT
In this work, we survey, classify and analyze a number of circumstances, which might lead to arrival of **malicious AI**.	malicious AI	AI

Table 3: Examples of sentences in which the AnthroScore masking strategy differs significantly from our masking strategy. The entire noun phrase, which is taken as the mask in AnthroScore's approach, is highlighted in bold. In our approach, we masked the minimal AI entity, leaving the anthropomorphic contextual cues unmasked.

phrase and masks it entirely, thus removing the important contextual information contributing to anthropomorphism – namely, the explicit mention of *cognitive abilities*. Applying our masking strategy helped AtypicalAnimacy immensely, but did not improve for AnthroScore, once again due to its pronoun constraint which is strictly incompatible with possessive structures since pronouns have their own genitive inflection and do not co-occur with the possessive clitic.

Both models make use of masked language models, whose predictions are based on statistical co-occurrences (Zhang et al., 2024). In AI research, as terminology is increasingly anthropomorphic and constantly introduces neologisms consisting of metaphors for human activities (e.g. *training*, *learning*, *attention*, *memory*, *hallucinations*, etc.), MLMs are more likely to predict an AI entity such as *ChatGPT*, *language model*, and *AI agent* when these terms appear in its context, instead of predicting human entities. While AnthroScore's pronoun constraint avoids this issue, it creates others. More importantly, anthropomorphic language does not necessarily align with grammatical animacy; an entity can be referred to by inanimate pronouns but framed as having human-like capacities.

Ultimately, both models are designed to identify animacy features which are understood as anthropomorphism in context. Even if the best method for anthropomorphism detection is to identify linguistic and grammatical animacy markers, it is still highly restricted to the English language. Many non-English languages do not have an inanimate pronoun, and their linguistic markers of animacy are far more nuanced. For instance, we might expect to see morphological variations or differential object marking (De Swart and De Hoop, 2018), but these cues are far more difficult to identify and are not necessarily contextual.

8 Conclusion

Despite the numerous studies and discussions on anthropomorphism in AI, there is not one agreed upon definition of what it entails, and consequently there are not many implementations of anthropomorphism detection, possibly due to its ambiguous and subjective nature. We present AnthroSet, a dataset of real-world instances of anthropomorphism in AI, grounded in a linguistic analysis of anthropomorphism and animacy markers in English, as well as a taxonomy of anthropomorphism based on that of DeVrio et al. (2025). We evaluate the two state-of-the-art MLM-based models for anthropomorphism detection, focusing on the advantages and limitations of employing masked language models for this task.

While a masking approach is congruent with predicate structures due to the distance between the predicate and the entity, as well the ability of MLMs to identify role arguments, an important feature of anthropomorphism – this method is not as useful for attributive structures, noun phrases or comparisons. This is due to the syntactic constraints imposed by the mask, as well as existing AI terminology influencing the masked language model, which works on the basis of statistical co-occurrences, as AI discourse becomes more anthropomorphic. Future work includes robust redundant annotation on our dataset, and combining our word-level line of work with Shardlow et al.'s (2025) sentence-level line of work, e.g. through supervised token-level classification, by cross-dataset evaluation and by assessing how our annotation schemes align.

References

Nicholas Barrow. 2024. Anthropomorphism and AI hype. *AI and Ethics*, 4(3):707–711.

Adriana Belletti and Luigi Rizzi. 1988. Psych-verbs and θ-theory. *Natural Language and Linguistic Theory*, 6(3):291–352.

Javier Carbonell, Antonio Sánchez-Esguevillas, and Belén Carro. 2016. The role of metaphors in the development of technologies. The case of the artificial intelligence. *Futures*, 84:145–153. Publisher: Elsevier BV.

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

Mariona Coll Ardanuy, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson, and Barbara McGillivray. 2020. Living machines: A study of atypical animacy. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4534–4545, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Peter De Swart and Helen De Hoop. 2018. Shifting animacy. *Theoretical Linguistics*, 44(1-2):1–23.

Alicia DeVrio, Myra Cheng, Lisa Egede, Alexandra Olteanu, and Su Lin Blodgett. 2025. A Taxonomy of Linguistic Expressions That Contribute To Anthropomorphism of Language Technologies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18, Yokohama Japan. ACM.

David R. Dowty. 1991. Thematic Proto-Roles and Argument Selection. *Language*, 67(3):547–619.

Hubert L. Dreyfus. 1976. What computers can't do. *British Journal for the Philosophy of Science*, 27(2):177–185.

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*, HRI '24, pages 447–452, New York, NY, USA. Association for Computing Machinery.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Luciano Floridi and Anna C Nobre. 2024. Anthropomorphising Machines and Computerising Minds: The Crosswiring of Languages between Artificial Intelligence and Brain & Cognitive Sciences. *Minds and Machines*, 34(1).

Douglas R. Hofstadter. 1995. *Fluid concepts & creative analogies: computer models of the fundamental mechanisms of thought*. Basic Books, New York, NY.

Beth Levin. 2022. On Dowty's "Thematic Proto-Roles and Argument Selection". In *Studies in Linguistics and Philosophy*, pages 103–119. Springer International Publishing, Cham. ISSN: 0924-4662, 2215-034X.

Drew McDermott. 1976. Artificial intelligence meets natural stupidity. *ACM SIGART Bulletin*, (57):4–9.

Adriana Placani. 2024. Anthropomorphism in AI: hype and fallacy. *AI and Ethics*, 4(3):691–698.

Beatrice Primus. 2012. Animacy, Generalized Semantic Roles, and Differential Object Marking. In *Studies in Theoretical Psycholinguistics*, pages 65–90. Springer Netherlands, Dordrecht. ISSN: 1873-0043.

Maria Grazia Rossi and Fabrizio Macagno. 2021. *The Communicative Functions of Metaphors Between Explanation and Persuasion*, page 171–191. Springer International Publishing.

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. *Minds and Machines*, 35(1):1–24.

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

John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424.

Matthew Shardlow and Piotr Przybyła. 2024. Deanthropomorphising NLP: can a language model be conscious? *PloS one*, 19(12):e0307521.

Matthew Shardlow, Ashley Williams, Charlie Roadhouse, Filippos Ventirozos, and Piotr Przybyła. 2025. Exploring supervised approaches to the detection of anthropomorphic language in the reporting of NLP venues. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18010–18022, Vienna, Austria. Association for Computational Linguistics.

Gerard J. Steen, Aletta G. Dorst, Tina Krennmayr, Anna A. Kaal, and J. Berenike Herrmann. 2010. *A Method for Linguistic Metaphor Identification*. Converging Evidence in Language and Communication Research. John Benjamins Publishing Company, Amsterdam.

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

Adam Waytz, John Cacioppo, and Nicholas Epley. 2010. Who Sees Human?: The Stability and Importance of Individual Differences in Anthropomorphism. *Perspectives on Psychological Science*, 5(3):219–232.

Xiao Zhang, Miao Li, and Ji Wu. 2024. Co-occurrence is not Factual Association in Language Models. In *38th Conference on Neural Information Processing Systems*. Version Number: 2.

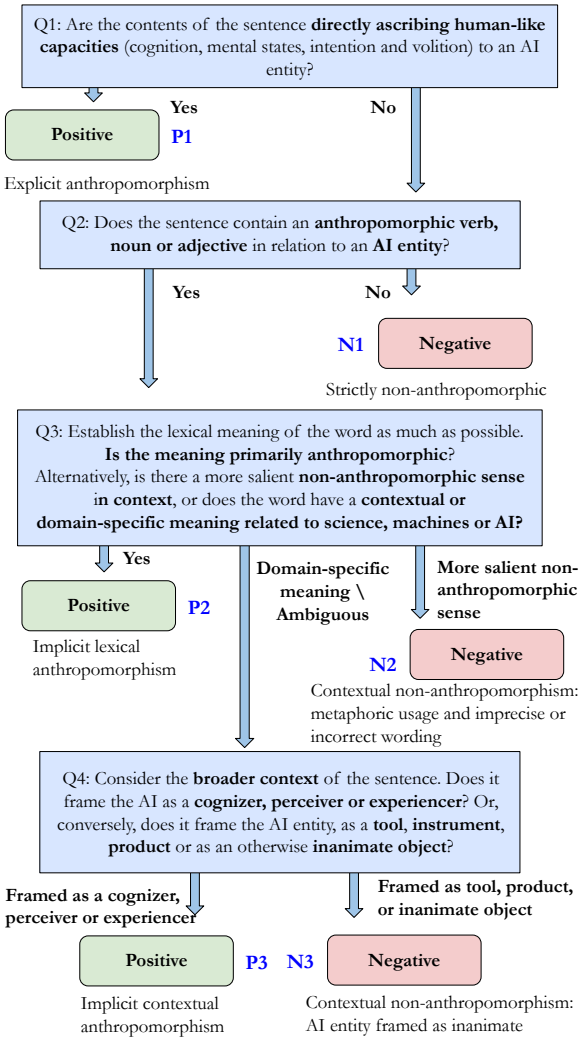

Figure 1: Decision tree for AnthroSet annotation.

A Annotation Guidelines

Annotators were instructed to annotate according to the workflow visualized in Figure 1. Some additional details were provided beyond what is shown here, including examples of typical words annotators might encounter, and a series of clarifications for potential edge cases. Full annotation guidelines can be found in our GitHub repository[12]. The taxonomy in Appendix C was also included in the instructions. Before the workflow, the following text was presented:

Read the sentence, and following the guidelines below, enter a score: 1 for anthropomorphic, 0 for non-anthropomorphic, and 2 for inconclusive cases. Since some sentences contain multiple AI entities, the relevant one is given in ***bold****.*

[12] https://github.com/doriellel/anthroset

B Examples of Anthropomorphic Sentences

verb subjects: We then propose a system that leverages the recently introduced social learning paradigm in which LLMs **collaboratively learn from each other** by exchanging natural language.
verb objects: First, we induce a language model to produce step-by-step rationales before outputting the answer to effectively **communicate the task** to the model.
verb objects (passive): In this study, we propose a new methodology to control how user's data is **recognized and used** by AI via exploiting the properties of adversarial examples.
adjectives (`acomp`): Results suggest that ChatGPT is **aware** of potential vulnerabilities, but nonetheless often generates source code that are not robust to certain attacks.
adjectives (`amod`): Consequently, we argue that the emergence of a **conscious** AI model is plausible in the near term.
role/function NPs: Many believe that use of generative AI as a **private tutor** has the potential to shrink access and achievement gaps between students and schools with abundant resources versus those with fewer resources.
role/function NPs (modifier): For example, in comparing ChatCollab AI agents, we find that an AI **CEO** agent generally provides suggestions 2-4 times more often than an AI **product manager** or AI **developer**, suggesting agents within ChatCollab can meaningfully adopt differentiated collaborative roles.
genitive NP: In this study [...] we evaluate nine popular LLMs on their **ability to understand** demographic differences in two subjective judgment tasks: politeness and offensiveness.
genitive NP ('s clitic): Our approach makes use of Large Language Models (LLMs) for this task by leveraging the LLM's **commonsense reasoning capabilities** for making sequential navigational decisions.
comparisons: In this paper, we prove in theory that AI can be **as creative as humans** under the condition that it can properly fit the data generated by human creators.

C Anthropomorphism Taxonomy

Attribute or Capacity	Examples
Conceptual Thought and Mental States: Hypothesizes, theorizes, and imagines sth. Anticipates, guesses or predicts sth about the world.	*think*, *expect*, *hope*, *guess*, *predict*, *dream*, *imagine*, *believe* (v) *(self-)aware*, *cognizant* (a)
Knowledge and Awareness: Has factual knowledge about and experience in the world, or memories of things that happened. As a result, has an ontology of things, and can identify, classify, and categorize.	*know*, *remember*, *recognize*, *memorize*, *forget*, *identify*, *classify*, *differentiate*, *distinguish* (v), *knowledge* (n)
Reasoning and Understanding: Reasons, rationalizes, analyses, makes sense of sth. Understands, considers, weighs options, takes sth into consideration or account.	*deduce*, *conclude*, *rationalize*, *reason*, *(mis)understand*, *(mis)interpret*, *analyze*, *infer*
Judgment: Has an opinion, makes decisions and choices, gives advice, has a preference, evaluates, imparts judgment. Has a concept of morality and ethics, knows right and wrong.	*advise*, *prefer*, *select*, *choose*, *decide*, *determine*, *resolve* (v)
Planning and Decision-making: plans, strategizes, sets a goal, devises a method, game plan or scheme, can also struggle or experience difficulties.	*plan*, *coordinate*, *strategize*, *come up with a plan*, *solve*, *struggle* (v)
Agency and Autonomy: Takes action, able to autonomously carry out a goal – used in a way that attributes agency and control over the action and situation.	*cheat*, *follow* or *break rules*, *achieve* (v), *autonomous*, *independent*, *creative* (a)
Communication: Communicates, teaches or explains, Similarly, can also learn or be at the receiving end of communication or explanation.	*communicate*, *talk*, *speak*, *tell*, *explain*, *teach*, *learn*, *ask* (v) *communicative* (a)
Active Support: Recommends, makes a suggestion or an offer. Actively and directly helps, aids and assists by employing skills to solve a problem.	*suggest*, *aid*, *help*, *contribute* (v) *responsible* (a) *feedback*, *insights* (n) *expert*, *advisor* (a)
Candidness: Capable of, or has a concept of honesty or dishonesty, truthfulness or deception. As a result, can be trustworthy or untrustworthy, reliable or unreliable.	*trust*, *believe*, *lie* (v) *(un)truthful*, *deceitful* (a)
Affability: Acts as a friend or as an enemy, companion or adversary, collaborator or rival. As a result can act benevolent or malevolent, friendly or hostile.	*collaborate*, *manipulate*, *insult*, *deceive* (v) *thoughtful*, *attentive*, *friendly* (a), *partner*, *adversary* (n)
Power and Relationships: Plays a role in a relationship dynamic – romantic or platonic, superior (boss, manager, teacher) or subordinate (employee, student).	*teach*, *supervise* (v) *manager*, *employee*, *teacher*, *tutor*, *student*, *companion*, *lover* (n)
Emotions: Empathizes, sympathizes, displays emotions, experiences pain or pleasure.	*experience*, *emote* (v), *sensitive*, *vulnerable* (a)
Self Expression and Perception of Deeper Meaning: Partakes in activities of self-expression such as art and storytelling, humor and jokes. Perceives beauty and aesthetics. Has a deeper understanding of meaning, purpose, and context. Related to emotions, awareness and conceptual thought.	*create poetry*, *create art*, *write*, *compose*, *paint*, *sing*, *dance* (v) *creative*, *artistic*, *funny* (a) *artist*, *poet*, *humor*, *irony* (n)
Sensory Perception: Receives and processes sensory input and feedback from the environment, picks up visual/auditory/sensory cues. Related to emotions, awareness and conceptual thought.	*see*, *hear*, *perceive*, *feel*, *sense* (v) *blind*, *deaf* (a)

Table 4: Human attributes and capacities that are usually attributed AI, representing different aspects of anthropomorphism. Based on DeVrio et al. (2025), extended to address human-written text and terminology from AnthroSet.

D Supplemental Results

	AnthroScore			*AtypicalAnimacy*		
Category	Precision	Recall	F1	Precision	Recall	F1
AnthroScore masking						
verb subjects positive	1.000	0.145	0.254	0.829	0.618	0.708
verb subjects negative	0.581	0.877	0.699	0.704	0.877	0.781
verb objects positive	1.000	0.125	0.222	0.789	0.804	**0.796**
verb objects negative	0.645	0.984	0.779	0.817	0.803	0.810
adjectives positive	1.000	0.114	0.204	0.905	0.432	0.585
adjectives negative	0.544	0.956	0.694	0.632	0.956	0.761
Minimal entity masking						
verb subjects positive	0.909	0.179	0.299	0.917	0.786	**0.846**
verb subjects negative	0.560	0.689	0.618	0.826	0.934	**0.877**
verb objects positive	0.609	0.241	0.346	0.804	0.776	0.789
verb objects negative	0.559	0.508	0.532	0.806	0.831	**0.818**
adjectives positive	0.571	0.154	0.242	0.962	0.481	**0.641**
adjectives negative	0.482	0.574	0.524	0.630	0.979	**0.767**

Table 5: Precision, recall, and F1 scores per class for AnthroScore and AtypicalAnimacy with both masking strategies across three categories of anthropomorphic structures: verb subjects, verb objects and adjectives.

		AnthroScore				*AtypicalAnimacy*			
Category	**Total**	1	0	2	1/Total	1	0	2	1/Total
AnthroScore masking									
verb subjects	33	**2**	21	10	0.06	**17**	16	-	0.52
verb objects	27	**3**	17	7	0.11	**16**	11	-	0.59
adjectives	17	**1**	15	1	0.06	**2**	15	-	0.12
comparisons	42	**1**	38	3	0.02	**19**	23	-	0.45
Minimal entity masking									
verb subjects	33	**1**	21	11	0.03	**16**	17	-	0.48
verb objects	27	**8**	10	9	0.30	**17**	10	-	0.63
adjectives	21	**2**	15	4	0.10	**4**	17	-	0.19
comparisons	42	**3**	34	5	0.07	**26**	24	-	0.62

Table 6: Comparison of AnthroScore and AtypicalAnimacy in terms of the proportion of positive predictions (label 1) among inconclusive cases, across four syntactic categories and two masking strategies.

	AnthroScore			*AnthroScore + inconclusive*		
Category	Precision	Recall	F1-Score	Precision	Recall	F1-Score
AnthroScore masking						
verb subjects	0.527	0.341	0.318	0.541	0.442	**0.395**
verb objects	0.548	0.370	0.334	0.512	0.456	**0.396**
adjectives	0.515	0.356	0.299	0.486	0.376	**0.302**
Minimal entity masking						
verb subjects	0.490	0.289	0.305	0.511	0.400	0.374
verb objects	0.389	0.250	0.293	0.370	0.361	0.347
adjectives	0.351	0.243	0.256	0.334	0.306	0.280

Table 7: Side-by-side comparison of AnthroScore's macro averaged precision, recall and F1 scores for the positive and negative cases alone, versus positive, negative and inconclusive cases, with both masking strategies.