Exploring Supervised Approaches to the Detection of Anthropomorphic Language in the Reporting of NLP Venues

Matthew Shardlow[1], **Ashley Williams**[1], **Charlie Roadhouse**[1],
Filippos Ventirozos[1], **Piotr Przybyła**[2,3],
[1]Manchester Metropolitan University, [2]Universitat Pompeu Fabra,
[3]Institute of Computer Science, Polish Academy of Sciences,
Correspondence: m.shardlow@mmu.ac.uk

Abstract

We investigate the prevalence of anthropomorphic language in the reporting of AI technology, focussed on NLP and LLMs. Anthropomorphised description of LLM/AI technology has the potential to misrepresent the capabilities of our field to other scientists, policy makers and the public. We undertake a corpus annotation focussing on one year of ACL long-paper abstracts and news articles from the same period. We find that 74% of ACL abstracts and 88% of news articles contain some form of anthropomorphic description of AI technology. Further, we train a regression classifier based on BERT, demonstrating that we can automatically label abstracts for their degree of anthropomorphism based on our corpus. We conclude by applying this labelling process to abstracts available in the entire history of the ACL Anthology and reporting on diachronic and inter-venue findings, showing that the degree of anthropomorphism is increasing at all examined venues over time.

1 Introduction

In an age of ubiquitous AI Agents powered by NLP technologies, science communicators must be careful in their choice of words. Overly technical language can stymie the readability, citability and impact of scientific endeavour. In addition to this, authors are regularly instructed by funders and editors to write for non-technical audiences (typically as a lay summary in addition to a technical abstract). In an effort to improve communication authors turn to more familiar language and in particular to the rhetorical device of anthropomorphism, defined by Merriam Webster[1] as :

> An interpretation of what is not human or personal in terms of human or personal characteristics

Anthropomorphism is a useful literary tool that uses known human experiences as characteristic metaphors, with many terms such as 'machine learning' or 'chain-of-thought' being widely accepted beyond academic borders.

Within the field of NLP, common anthropomorphisms are generally understood to refer to the technical contributions they indicate. However, they also give potential for incorrect assumptions to be made about the capacities of LLMs. Recent criticism of anthropomorphised reporting has highlighted the responsibility of AI practitioners to faithfully describe their techniques (Bender and Koller, 2020; Abercrombie et al., 2023).

Take for example, the following sentences extracted from an online news source (1) and an ACL main conference paper (2) in 2022 (emphasis from the author, sources available on request):

(1) "When the human asked if it was 'a robot' *the system lied* and said it was a person with a visual impairment."

(2) "The recent success of reinforcement learning (RL) in solving complex tasks is often attributed to *its capacity to explore and exploit an environment*."

Note that in both these examples there is some inanimate agent upon which human capability is inferred: a robot capable of deceit in (1) and the reinforcement learning algorithm capable of inquiry in (2).

The use of anthropomorphism has the potential to misrepresent the capabilities of NLP technologies and in this work we seek to better understand this phenomenon. We examine the use of anthropomorphisms in the NLP literature and in news text. To achieve this goal, we present a new corpus[2] of

[1] https://www.merriam-webster.com/dictionary/anthropomorphism

[2] All data is available via GitHub at: https://github.com/mattshardlow/Anthropomorphism_Corpus

Findings of the Association for Computational Linguistics: ACL 2025, pages 18010–18022
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

annotated anthropomorphisms from scientific literature and news sources (Section 3), giving insights on the types of anthropomorphisms present. We further train a model to predict the degree of anthropomorphism in a scientific abstract (Section 4.1) and apply this model to the ACL Anthology (Section 4.2). We report our results and provide discussion points in Section 5.

2 Related Work

Several prior efforts have sought to identify examples of texts featuring the anthropomorphisation of AI systems (Brooker et al., 2019; Shardlow and Przybyła, 2024; Cheng et al., 2024; DeVrio et al., 2025). Brooker et al. (2019) argues that the term Artificial Intelligence is itself an anthropomorphisation, indicating that the agent possessing the inferred quality of 'AI' has attained a human characteristic. Work to better quantify cases of AI anthropomorphisation has led to categorisations of phrases referring to AI into ambiguous and explicit anthropomorphism (Shardlow and Przybyła, 2024) depending on the authors intent. Later work to measure the degree to which anthropomorphisation is prevalent focussed on unsupervised automated scoring of model descriptions identifying whether a LLM was more likely to replace the name of a system with 'it' (indicating non-anthropomorphic language) or 'he/she' (indicating anthropomorphised language) (Cheng et al., 2024). Anthropomorphic language in AI is not limited to the models, but is also applied to the tasks they complete, such as 'reading *comprehension*' or 'sentiment *analysis*' (Lipton and Steinhardt, 2019).

Anthropomorphised language is often a factor in the misrepresentation of AI abilities (Watson, 2019; Placani, 2024). Misrepresentation leads to misunderstanding and misapplication of AI tools which leads to confusion amongst AI scholars, developers and the general public (Brooker et al., 2019; Lipton and Steinhardt, 2019). Further, in a recent study, Inie et al. (2024) analysed user trust when interacting with anthropomorphised and deanthropomorphised descriptions of AI systems, finding that the presence of anthropomorphic terminology alone did not influence user trust.

Various audiences who may produce and/or consume anthropomorphised descriptions of AI systems have been considered in the literature. Firstly, we may consider scientists in the NLP and AI community, who are actively working on the latest models and have typically been engaged with the technology for a number of years prior to the recent increase in AI technologies. These scholars are prone to AI anthropomorphisation with a recent study showing that 32 out of 81 examined papers (39.5%) concerning language modelling technology exhibited some form of anthropomorphisation in the abstract (Shardlow and Przybyła, 2024). Secondly, journalists reporting on AI for the general public are also responsible for anthropomorphisation with a growing body of evidence to demonstrate that public news reporting is more anthropomorphic than science communication of the same topics (Bender and Koller, 2020; Ryazanov et al., 2025). Finally, the general public possess lay knowledge of AI systems and may prefer anthropomorphised descriptions in some cases (Inie et al., 2024). Science communicators must work to ensure that descriptions are not harmful in misrepresenting the abilities of AI systems to the general public (Salles et al., 2020).

Prior to our work, no systematic annotation of anthropomorphism in the NLP literature has been undertaken. Our work complements existing studies by providing a large-scale annotation of anthropomorphism and evidence-based NLP methodologies to longitudinally detect anthropomorphic language.

3 Corpus Development

In this section, we present the first ever annotated corpus of anthropomorphic language used to describe LLMs. We begin by describing the corpus collection and annotation procedure and then present an analysis of the annotations within the corpus.

The ACL Anthology (Bird et al., 2008) is freely available online in XML format[3]. We selected all abstracts of long papers presented at the 60th Meeting of the Association of Computational Linguists (ACL 2022) from the Anthology. This gave us 601 abstracts, comprising of 3,584 sentences.

To contrast with scientific writing, we also selected news articles for annotation. To collect the news articles, we harvested public-facing RSS feeds from BBC News, the New York Times and the Register for article titles containing keywords (*Artificial Intelligence*, *AI*, *A.I.*, *ChatGPT*, *Large Language Model*, *LLM*) related to LLMs. We collected articles for a two month period (May-June 2023) executing the searches every day and remov-

[3] https://github.com/acl-org/acl-anthology

ing duplicates. We obtained 49 articles that were suitable for annotation, extracting the plain text.

3.1 Annotation

We annotated the selected documents at the sentence level to determine the degree of anthropomorphism present in each. One article may contain several sentences at varying degrees of anthropomorphism, which is captured by our annotation scheme. We recruited three annotators for our study, all English speakers studying for STEM degrees at undergraduate and postgraduate level. The annotators were overseen by two senior academics, who provided initial training and weekly oversight of annotations. Prior to the annotation process, we developed detailed guidelines explaining the context of the task, the process for annotation, specific definitions and examples of each category and a walkthrough of the chosen annotation platform: LightTag (Perry, 2021). The guidelines that we provided were discussed regularly throughout the annotation process and we updated the guidelines with specific information on cases that we had not considered in collaboration with the annotators (for example, how to properly handle lists).

All annotators were informed of the research purpose of the corpus and its likely intended use. Each annotator contributed around 100 hours of annotation time. Annotations took place over a 7 week period (June-July 2023) with weekly review meetings comprising all annotators and academics.

We instructed annotators to specifically focus on the application of this definition to AI systems, i.e., the description of any system leveraging AI technology (including LLMs) in a human or personal manner. Annotators were presented with an entire abstract, which they annotated at the sentence level using span annotations to identify sentence bounds.

Annotators were asked to make two key decisions regarding each sentence in each abstract or news article. Firstly, whether or not each sentence represented a claim, where we used a broad definition of claim as a statement or assertion about the findings of the work. For example, a sentence in an abstract which lists the section headings is not a claim about the findings and should be disregarded. In practice, we found that the overwhelming majority of sentences in our abstracts were annotated as claim sentences, which should not be surprising given that the purpose of an abstract is to state the claims of the given research. Claims were more sparsely distributed in the news articles.

Secondly, if a sentence was identified as a claim, annotators were required to determine the degree of anthropomorphism represented by that claim. We provided three categories: (1) non-anthropomorphic language, (2) ambiguous anthropomorphic language and (3) explicit anthropomorphic language, following the work of Shardlow and Przybyła (2024). The first category represents a claim which has not used anthropomorphic language. The next two categories both represent the use of anthropomorphic language, but at differing degrees of severity. Many anthropomorphisms are commonly used in such a manner that 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. One such example is the term ‘machine learning’. A domain expert knows that this indicates that some statistical model is built based on patterns evidenced in data. A lay reader may assume that a machine is capable of learning in the same way as a human can. We provide examples for all categories in Table 1.

When annotating for anthropomorphism, we are requiring a subjective interpretation of the guidelines. Whilst we made every effort to clarify the guidelines to give consistency, the core task of identifying whether a claim is anthropomorphic or not requires the annotator to use his or her own judgement. As such, we decided to show every document in our corpus to three parties. In subjective tasks (Frenda et al., 2024), multiple annotations allow for the identification of mistakes and for an additional layer of discussion and resolution throughout the annotation process. In the first instance, we double annotated every document using our three annotators. Annotators were instructed to revisit prior annotations in light of the updated findings of each meeting to improve agreement. Once all documents from the scientific and journalistic sources had been double annotated, all annotations were reviewed by a single academic — providing the final layer of annotation. The academic performed one of three tasks: (1) confirming the annotation in the case of agreement. (2) resolving a disagreement by siding with one annotator or another. (3) In rare cases, adding additional annotations that had been missed by both annotators. This process triplicated our efforts compared to a single annotation study, as each instance was observed 3 times by a separate annotator, but this was necessary in light of the difficulty of annotating for the subjective phenomenon of anthropomorphism.

	Category	Example
Claims	Non-Claim	In this study, we propose a . . . framework
	Claim	We demonstrate that language models begin to learn. . .
Level	Non-Anthropomorphic	We trained the model on the dataset.
	Ambiguous Anthropomorphic	The model learnt / predicted the class labels
	Explicit Anthropomorphic	The model understands / writes / says / knows. . .

Table 1: Categories for annotation in our study

	NA	**AA**	**EA**
ACL	2770 (77.3%)	709 (19.8%)	105 (2.9%)
News	571 (75.5%)	130 (17.2%)	55 (7.3%)
All	3341 (77.0%)	839 (19.3%)	160 (3.7%)

Table 2: Corpus statistics at the sentence level for the scientific abstracts (ACL), news articles written by journalists (News) and the entire corpus (All). NA = Non-anthropomorphic, AA = Ambiguous anthropomorphic, EA = Explicit Anthropomorphic. The raw count is presented, with the percentage of total sentences for each category in brackets.

	NA Only	**AA Present**	**EA Present**
ACL	217 (36.0%)	310 (51.4%)	76 (12.6%)
News	6 (12.2%)	20 (40.8%)	23 (46.9%)
All	223 (34.2%)	330 (50.6%)	99 (15.2%)

Table 3: Corpus statistics at the document level for the scientific abstracts (ACL), news articles written by journalists (News) and the entire corpus (All). Column 2 represents documents which feature only non-anthropomorphic annotations (NA Only). Column 3 represents documents which feature any claims containing Ambiguous Anthropomorphism as well as non-anthropomorphic claims (AA Present). Column 4 represents documents which feature any sentences containing Explicit Anthropomorphism as well as other claims (EA Present).

We do not calculate kappa score between annotators, as we do not expect agreement in a subjective setting, instead aiming to create a corpus that reflects the subjective interpretations of multiple annotators (Mostafazadeh Davani et al., 2022). We consider our approach to be in line with the trend towards perspectivist NLP methodologies (Abercrombie et al., 2024). Instead, we have controlled for agreement through regular annotation meetings and through a final annotation resolution procedure. This gives a final corpus of of 652 documents comprising of 4340 claim sentences, each with a finalised label indicating the degree of anthropomorphism.

3.2 Corpus Statistics

We analysed the distribution of identified claims within our final annotated corpus at the individual instance level, with the results shown in Table 2. We report on the total number of identified claims and each claim category for all instances as well as presenting the data broken down by data source (ACL abstracts vs. news articles). In total we identified 4,340 claims, of which 3,584 came from ACL abstracts and 756 came from news articles. The vast majority of identified claims (77.0%) were labelled during annotation as Non-Anthropomorphic. We identified 999 examples of anthropomorphic language, split between 839 examples (19.3%) of ambiguous anthropomorphism and 160 examples (3.7%) of explicit anthropomorphism.

We also analysed the corpus at the document level to understand the distribution of anthropomorphism from each source. To do this, we aggregated the annotations for each document and identified three distinct categories: (a) those documents which had annotations of non-anthropomorphism only, (b) those documents which had at least one ambiguous anthropomorphism annotations, but no explicit anthropomorphism and (c) those documents which had at least one explicit anthropomorphism annotation. The results are presented in Table 3. We have also presented the results according to data source (ACL abstracts vs. news articles). These results demonstrate that 34.2% of the documents that we analysed had no form of anthropomorphism. This is largely driven by ACL abstracts (36.0%), which make up the majority of the corpus. Only 6 (12.2%) of the 49 news articles that we analysed had no form of anthropomorphism.

To better understand the degree of anthropomorphism in each document, we design a simple scoring metric to allow us to quantify the level of anthropomorphic language by aggregating the scores for each sentence. Our scoring function is in the range 0 to 1, where 0 indicates no anthropomorphism and 1 indicates that every claim in a document is representative of explicit anthropomor-

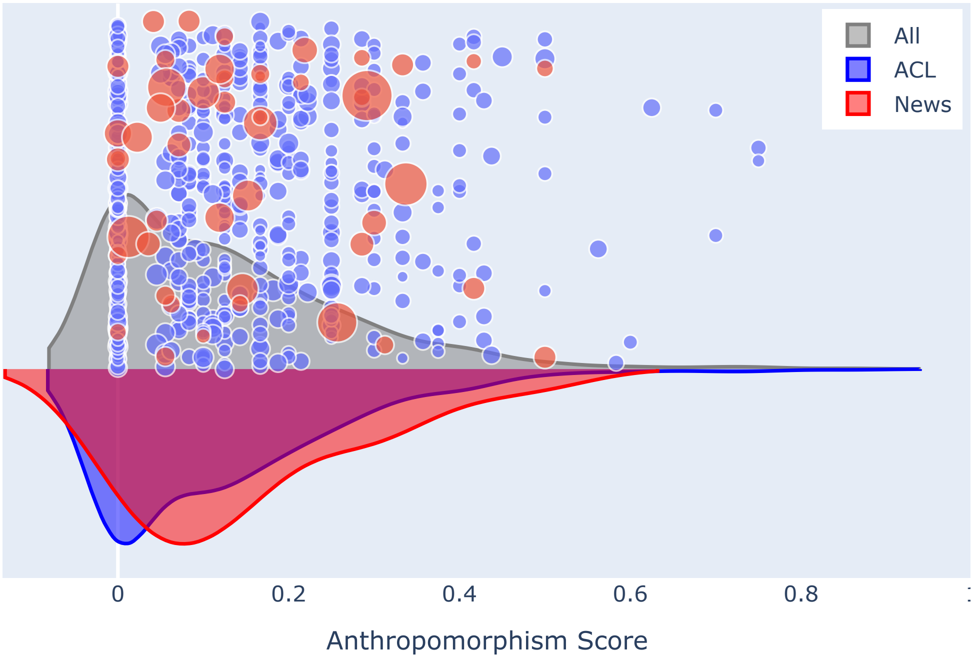

Figure 1: The Anthropomorphism score for every annotated document in our corpus, shown as bubbles which are sized for the number of instances per document and coloured according to the data source (ACL abstracts or news articles). Additionally, KDE plots are presented showing the density of data in each data source, as well as across the entire dataset.

phism. We transform all non-anthropomorphism labels to 0 and explicit anthropomorphism labels to 1. For ambiguous anthropomorphism we assign a value of 0.5.[4] To calculate the anthropomorphism score, we take the mean average of anthropomorphism annotations for each document. We express this process through the two equations below.

Firstly, considering that we have some categorical label l, we transform this to a numerical representation l' as follows:

$$l' = \begin{cases} l = \text{Non-Anthropomorphic} & \rightarrow \quad 0 \\ l = \text{Ambiguous Anthropomorphic} & \rightarrow \quad 0.5 \\ l = \text{Explicit Anthropomorphic} & \rightarrow \quad 1 \end{cases}$$

Then, considering some document D, with n numerical annotations $L' = \{l'_1, l'_2, ..., l'_n\}$ we find a score S_D as follows:

$$S_D = \frac{\sum_{i=1}^{n} l'_i}{|L'|}$$

We compute S_D for every document in our corpus, producing Figure 1.

The data in Figure 1 demonstrates that most documents have a low anthropomorphism score. This can be noted on the KDE plots, which show the majority of the probability mass between 0 and 0.2 for both ACL and News. News articles are typically longer than abstracts as shown by the larger red bubbles. A few outlier abstracts have very high levels of anthropomorphism, with the highest scoring 0.857 (based on 7 claims), indicating that almost all claims in this abstract were annotated as explicit anthropomorphism. The peak of the probability mass of the news articles (shown in the red KDE plot) is shifted to the right of the abstracts (shown in the blue KDE plot), indicating that news articles are generally more anthropomorphic than abstracts and giving corpus based evidence to this finding which has also been demonstrated elsewhere (Shardlow and Przybyła, 2024; Cheng et al., 2024).

4 Anthology Study

In this section we train a classifier to predict S_D for unseen abstracts using the ACL abstracts in our annotated corpus. We subsequently apply this classifier to the entire ACL Anthology, reporting our findings. As our model is designed for ACL abstracts, we do not consider claim detection as we previously noted (See Section 3.1) that most abstracts are made up entirely of claims. Instead we predict anthropomorphism score based on the entire abstract.

[4] It is unclear where on the scale between 0 and 1 these annotations should fit. For simplicity, we choose to assign this to the value of 0.5, halfway between 0 and 1.

4.1 Model Development

We calculate the anthropomorphism score for each labelled abstract as described previously in this work. We only considered the scientific abstracts in our corpus, and disregarded the news articles for this part of our study for reasons of domain specificity. We split the abstract data into training and testing portions. This gave 484 documents in the training set (80%) and 119 documents in the test set (20%).

We selected 4 transformer based methods from HuggingFace (Wolf et al., 2019), each of which was configured for regression through the Trainer interface: *google-bert/bert-base-uncased* (Devlin, 2018), *FacebookAI/roberta-base* (Yinhan et al., 2019), *xlnet/xlnet-base-cased* (Yang et al., 2019), *albert/albert-base-v2* (Lan et al., 2020). We used the python transformers library and HuggingFace's Trainer with default parameter configurations for training including AdamW Optimiser (Loshchilov, 2017) with learning rate of $4e^{-5}$ and 10 training epochs. We evaluated our models using mean squared error (MSE), mean absolute error (MAE), proportion of variance (R^2) and Pearson's correlation. The results for each model are shown in Table 4.

We additionally include two baselines. The first (denoted as *Mean*) calculates the mean anthropomorphism score on the training labels and uses this as the score for every test instance. The second is *AnthroScore* (Cheng et al., 2024), which is an unsupervised approach considering the likelihood of a model's name being replaced by a personal pronoun (he/she). We use the python implementation of AnthroScore available via GitHub[5] with the list of suggested entities denoted as X_{LM} in Appendix B.2 of the paper by Cheng et al. (2024). We scaled the results of AnthroScore into the range 0-1 to enable a fair comparison with our model outputs. We acknowledge that AnthroScore is an unsupervised approach and as such is not tuned to the labels in our corpus. However we include it here for completeness as it is the only similar approach for the detection of anthropomorphism in the literature.

All models trained successfully across the ten epochs of training time with the loss decreasing steadily. Whereas the range of anthropomorphism scores is 0-1 (with the majority of the annotations in the 0-0.5 range), the regression classifiers were able to predict the anthropomorphism scores of documents with MAE=0.0864–0.0909, indicating that the predicted score was on average within less than 10% of the correct label. The classifiers all outperformed both baselines across all metrics indicating that some learning took place in each instance.

Approach	MAE	MSE	R^2	Pearson
Mean	0.1070	0.0202	0.0000	—
Anthroscore	0.7218	0.5990	-7.3134	0.1875
Bert	0.0909	0.0170	0.1575	0.4338
RoBERTa	0.0877	0.0166	0.1783	0.4681
XLNet	0.0864	0.0156	0.2298	0.5208
Albert	0.0891	0.0160	0.2083	0.4730

Table 4: The results of document level regression against the anthropomorphism score. Rows 1 and 2 demonstrate baselines. rows 3–6 show the results of tranformer-based regressors.

XLNet gives the lowest MSE score at 0.0156. Bert gave the highest score on both metrics indicating that it is less suitable for this task. The R^2 and Pearson's correlation metric also demonstrate that XLNet was able to capture more of the variance in the output than other models. Whilst an R^2 score of 0.2298 (XLNet) is not especially high, it does demonstrate that the model has captured some of the trends in the data and that the predicted scores vary to a moderate degree in line with the true labels. Much of the variance mismatch is due to documents which have a gold anthropomorphism score of 0, but the model has predicted some small degree of anthropomorphism. The baseline system, AnthroScore, by contrast, has an R^2 score of -7.3, indicating that the outputs predicted by this model do not vary consistently with the gold labels and confirming that the unsupervised approach is not suitable for detecting the labels in our corpus.

4.2 Diachronic Analysis

Our annotations are based on a single year of ACL abstracts and news articles from a similar period (2022). This is useful for giving a snapshot of the state of anthropomorphism in a given time period, but does not help to understand how the use of anthropomorphic language in scientific reporting has developed over time and is developing now. To better understand the use of anthropomorphic language beyond our annotations, we applied a regression model based on XLNet to abstracts from the entire ACL Anthology from 1990-2023 (representing all data available at the time of this study).

We downloaded the ACL Anthology as a bibtex file on Friday 23rd August 2024. This edition

[5] https://github.com/myracheng/anthroscore

held 100,293 individual entries. We filtered for all entries which had an abstract associated with them prior to 2024. We do not consider 2024 as the entries for the year were not complete at the time of writing. We also removed all entries from ACL 2022, which were the abstracts we used for the annotation of our corpus. This gave 44,870 abstracts. The metadata associated with the abstracts did not give a consistent indication of venue across time (e.g., venue names change, conferences merge and identifiers are not standardised over time). For this reason, we identified 8 conferences of interest, which represent the majority venues in the ACL Anthology. All other venues are combined as ’Other’.

We retrained the XLNet-base model with all the annotated abstracts and their associated anthropomorphism scores. We only used the scientific abstracts and not the news articles for training as the model will only be used to make predictions for scientific abstracts in this case. The model was trained using the same configuration as described previously for 10 epochs. The model could then be used with a novel abstract as input and the output being the predicted anthropomorphism score.

We ran the model on every available abstract in the ACL Anthology (n=44,870), which took around 60 minutes on an M2 Macbook Pro with 16GB RAM. The mean prediction for each identified venue in each year was calculated for inclusion in the results.

For validation we analysed the results of the model by (a) manual validation and (b) human judgement. We present a sample of the manual validation consisting of model outputs (abstract texts and scores assigned by the model) with our interpretations as Table 6, which can be found in Appendix A. We broadly found that the scores aligned with our expectations and that clear justification could be given for the variability in assigned score based on known markers of anthropomorphism evident in the analysed texts. For human judgement we gave a sample of 100 abstracts to four annotators (selected from the authors) who were familiar with the annotation guidelines. The selected abstracts represented a range of years and venues. We selected 25 outputs from each of the first, second, third and fourth quartiles of the predicted probability range according to the model's predictions. For each item, annotators were asked to answer 2 questions: (1) does this abstract contain any instances of ambiguous anthropomorphism and (2) does this article contain explicit anthropomorphism. The annotators did not have any information as to the models prediction when answering the questions. The results were aggregated across annotators and are presented in Table 5, where we show the total number of abstracts assigned to each category according to the predicted quartiles.

Question	Q1	Q2	Q3	Q4
AA	12	27	42	48
EA	2	7	6	18
AA or EA	13	30	43	51

Table 5: Aggregated human judgements of model outputs. Q1-Q4 represents abstracts selected from the four quartiles of the model output probabilities. The number in each column represents the number of abstracts where the annotator detected the presence of ambiguous anthropomorphism (AA) or Explicit Anthropomorphism (EA), aggregated over all annotators.

5 Discussion of Results

Table 5 shows that the human evaluators assigned an anthropomorphic label more often to abstracts with higher predicted scores. 51 out of 100 instances in the fourth quartile were assessed as containing some form of anthropomorphism. However annotators detected anthropomorphic abstracts across all quartiles. This indicates that there is significant room for interpretation of the model's score and that the score should be used in conjunction with human judgement when assessing the degree of anthropomorphism in an article. Across all quartiles fewer abstracts were annotated as containing any form of Explicit Anthropomorphism as compared to Ambiguous Anthropomorphism. This may be an indicator that annotators were more willing to give the middle, less extreme category than the definitive category of explicit anthropomorphism.

The analysis of anthropomorphism over time is presented in Figure 2, which shows bubble plots and trend lines for each venue on a separate subfigure. We have additionally included aggregated figures for the bubble plots and trend lines in Appendix B.

Figure 2 shows the average predicted anthropomorphism score in each year for every venue. We do not have information on every venue for each year with most venues starting in 2016 and 2017. Two notable exceptions to this are LREC and the ‘Other’ category (covering workshops and smaller conferences) which start in 2004 and 1991 respectively.

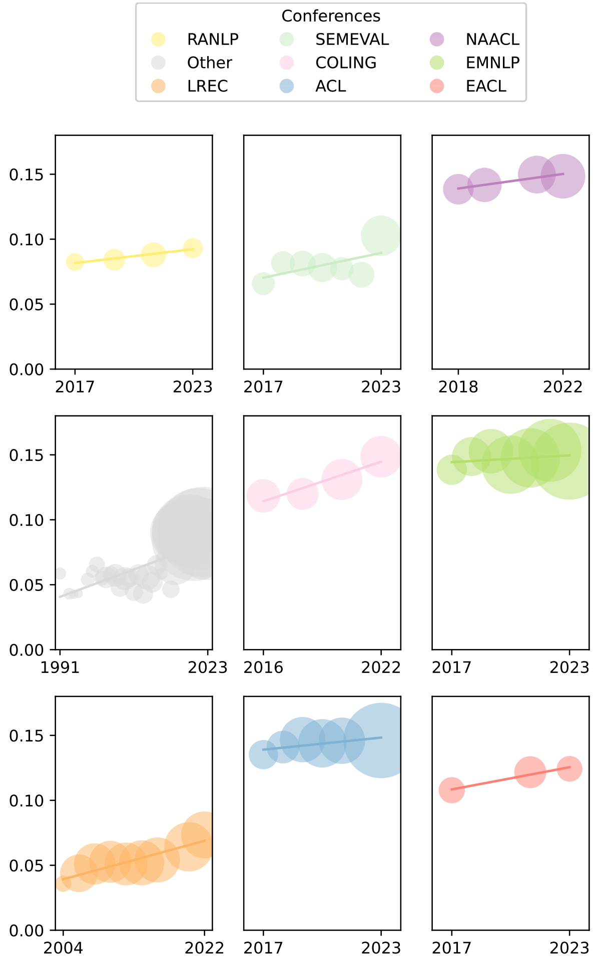

Figure 2: The trends observed for each venue considered. Bubble size indicates number of abstracts. Note that the Y-axis is the same across all sub-figs, but the X-axis is scaled to the years available for that venue.

The y-axis in Figure 2 shows that the average predicted anthropomorphism score is generally low (0.02-0.15). We found that most documents were predicted to have a low anthropomorphism score, which is reflective of the data distribution of the gold labels (reflecting the annotation of ACL 2022). However, in most venues there existed a long tail of abstracts with higher than average anthropomorphism scores. The averages that are presented against the y-axis are reflective of an aggregation of all documents within a venue for that year.

In light of this, we do not claim findings on the anthropomorphism levels of specific abstracts, but only on the aggregated values between years and venues. We also do not make claims about the raw average predicted anthropomorphism level of any one venue or year, but instead rely on comparing data points. We expect that even given some shift in scores due to model variability, the broad trends exposed by this research remain consistent.

We observe in Figure 2 that the degree of anthropomorphism in scientific writing as found in the ACL Anthology has increased in the period 1991-2023. We can further see that the trend towards increased anthropomorphism is present for every venue that we analysed. This is particularly clear in Figure 4, where each trend line is clearly increasing.

Figure 2 demonstrates some clear trends in terms of the predicted degree of anthropomorphism between venues. LREC is notably at a lower degree of anthropomorphism than other venues throughout the entire reporting period for which we have abstracts available. LREC typically publishes resource descriptions and evaluations, so the tendency to anthropomorphise these may be lower. ACL, NAACL and EMNLP exhibit the highest average predicted anthropomorphism across the period. We note that for the period 2017-2023 (where we have data for all venues) the average predictions for the 'Other' category is in between all analysed venues representing a mixture of the type of reporting found across the ACL Anthology. This helps to contextualise the earlier period (1991-2016) where we only have data for 'Other' and LREC, indicating that Other is a generally reliable measure of the average anthropomorphism in a given year.

6 Conclusion

We present the first large-scale annotated corpus of anthropomorphisms related to large language models. We use this corpus to demonstrate that anthropomorphism is present in both scientific abstracts and news articles. We have also developed a new model for predicting the degree of anthropomorphism in a scientific abstract and demonstrated that the use of anthropomorphism is a growing trend in the ACL Anthology.

We expect to extend our corpus in future work, making use of our existing model to bootstrap future annotations and extending the data sources beyond the Anthology. We will also investigate other tasks related to anthropomorphic language such as sentence classification and style transfer.

We are not advocating for the abolition or prohibition of anthropomorphic terminology. There are clearly many valid cases where this language is widely accepted and understood to refer to underlying technological advances (indeed artificial intelligence is an anthropomorphism which has been in use since the 1950s). Instead, we hope that this work provokes timely discussion amongst science communicators, both in the press and the NLP community. We must take stock of our own usage of anthropomorphised language to describe our technology and consider the impact of this on the public mind.

7 Limitations

Subjectivity of Annotation

We acknowledge that attempting to annotate scientific abstracts for anthropomorphism is a subjective process and that the definition of anthropomorphism and what forms of language may be annotated under this definition is subject to interpretation. We have attempted to control for this subjectivity by showing each instance to 3 annotators as discussed above. We have also deliberately obfuscated the authors names when discussing examples of anthropomorphised terminology to avoid embarrassment.

Model Variability

The model we used for the prediction of anthropomorphism score does not perfectly predict the labels according to the test data. No model is perfect and it may be the case that future models based on advanced technology and further annotated data will be able to outperform our model. We have used the model to explore broad trends in the ACL Anthology which are aggregated over multiple abstracts per year and we expect that the effects of any variability will be suitably mitigated by this aggregation.

Time Variance

Our model is based on ACL abstracts from 2022. We have applied it to data ranging back to 1990. We note that there may be linguistic variance in older documents that is not reflected in the training set of our model. Most of our diachronic analysis is limited to contemporaneous sources (2016 onwards) and we have also included one year after the model's training data (2023). It may also be noted that there is some natural shift in topical composition of conferences in recent years (moving from linguistics-focus to machine learning, deep learning and LLMs), with more recent conferences having much more work and more work making use of technology that can be described with anthropomorphised terms.

Authors use of Anthropomorphisms

The authors acknowledge that the use of anthropomorphism in scientific writing is a valid literary device, which we personally make use of in our own writing. We have not explicitly avoided the use of anthropomorphisms in this manuscript and we expect that an analysis of our own prior works will reveal many uses of anthropomorphism.

References

Gavin Abercrombie, Valerio Basile, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli, editors. 2024. *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. on anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore. Association for Computational Linguistics.

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.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Phillip Brooker, William Dutton, and Michael Mair. 2019. The new ghosts in the machine: 'Pragmatist' AI and the conceptual perils of anthropomorphic description. *Ethnographic studies*, 16:272–298.

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.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28.

Nanna Inie, Stefania Druga, Peter Zukerman, and Emily M Bender. 2024. From" ai" to probabilistic automation: How does anthropomorphization of technical systems descriptions influence trust? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2322–2347.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *Preprint*, arXiv:1909.11942.

Zachary C Lipton and Jacob Steinhardt. 2019. Troubling trends in machine learning scholarship: Some ml papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1):45–77.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Tal Perry. 2021. Lighttag: Text annotation platform. *arXiv preprint arXiv:2109.02320*.

Adriana Placani. 2024. Anthropomorphism in ai: hype and fallacy. *AI and Ethics*, pages 1–8.

Igor Ryazanov, Carl Öhman, and Johanna Björklund. 2025. 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.

Matthew Shardlow and Piotr Przybyła. 2024. Deanthropomorphising nlp: can a language model be conscious? *PLoS One*, 19(12):e0307521.

David Watson. 2019. The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. *Minds and Machines*, 29(3):417–440.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. arxiv. *arXiv preprint arXiv:1910.03771*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Liu Yinhan, Ott Myle, Goyal Naman, Du Jingfei, Joshi Mandar, Chen Danqi, Levy Omer, and Lewis Mike. 2019. Roberta: A robustly optimized bert pretraining approach (2019). *arXiv preprint arXiv:1907.11692*, pages 1–13.

A Anthropomorphism Score for Example Abstracts

Table 6 shows the text of abstracts from our diachronic analysis with the associated anthropomorphism score assigned by our model. None of these abstracts were part of the annotated corpus. We have provided brief notes of interpretation for each abstract.

B Diachronic analysis

Figure 3 shows the bubble plots from Figure 2 superimposed on a single axis. Figure 4 shows the trend lines from Figure 2 superimposed on a single axis.

C LLM performance on Anthropomorphism Detection

We repeated our experiments using 3 popular Large Language Models for Text Classification through in-context learning. We provided 10 samples to each model and also evaluated in a zero-shot setting. The results in Table 7 show that the LLM configurations that we experimented with are not suitable for detection of Anthropomorphism.

Venue	Score	Text	Notes
EMNLP 2022	0.448	On vision-language understanding (VLU) tasks, *fusion-encoder vision-language models achieve superior results* but *sacrifice efficiency* because of the simultaneous encoding of images and text. . . . To get the best of both worlds, we propose [MODEL], *a framework that distills the knowledge of the fusion-encoder teacher model into the dual-encoder student model.* Since the cross-modal interaction is the key to the superior performance of teacher model but is absent in the student model, *we encourage the student not only to mimic the predictions of teacher, but also to calculate the cross-modal attention distributions and align with the teacher.* Experimental results demonstrate that [MODEL] is competitive with the fusion-encoder teacher model in performance (only a 1% drop) *while enjoying 4 times faster inference*.	The category of techniques, 'fusion-encoder vision language models' is treated as an agent in its capacity to 'achieve superior results'. Further anthropomorphic language such as 'encourage the student' and 'enjoying' are used in reference to the proposed model.
EMNLP-IJCNLP 2019	0.372	Mobile agents that can *leverage help from humans can potentially accomplish more complex tasks than they could entirely on their own.* . . . To address the [MODEL] problem, we develop *a memory-augmented neural agent that hierarchically models multiple levels of decision-making*, and an *imitation learning algorithm that teaches the agent to avoid repeating past mistakes* while simultaneously *predicting its own chances of making future progress*. Empirically, our approach is able to *ask for help* more effectively than competitive baselines and, thus, attains higher task success rate on both previously seen and previously unseen environments.	This abstract casts neural agents as actors in a bespoke environment indicating that they have the capacity to 'leverage help', 'mimic' and 'align'. The framework is described as able to 'distill the knowledge', whereas it is in fact the researcher's use of the framework that leads to distillation.
ACL 2020	0.243	Most neural machine translation models only *rely on pairs of parallel sentences*, assuming syntactic information is automatically learned by an attention mechanism. In this work, we investigate different approaches to incorporate syntactic knowledge in the Transformer model and also propose a novel, parameter-free, dependency-aware self-attention mechanism that improves its translation quality, especially for long sentences and in low-resource scenarios. We show the efficacy of each approach on WMT English-German and English-Turkish, and WAT English-Japanese translation tasks.	This abstract contains a much lower degree of anthropomorphism, with the highlighted reference to models 'relying' on a certain data format. Otherwise, the researchers clearly state their own role in performing the experiments.
ACL 2019	0.048	This paper examines various unsupervised pretraining objectives for learning dialog context representations. Two novel methods of pretraining dialog context encoders are proposed, and a total of four methods are examined. Each pretraining objective is fine-tuned and evaluated on a set of downstream dialog tasks using the [ANON] dataset and strong performance improvement is observed. Further evaluation shows that our pretraining objectives result in not only better performance, but also better convergence, *models that are less data hungry* and have better domain generalizability.	Note that this abstract only has a very low predicted anthropomorphisation score. The only small usage of anthropomorphic language here is the use of 'data hungry'. This is an idiomatic usage of hunger and is used appropriately in a way that does not mislead a reader.

Table 6: Example partial abstracts (column 3) and their predicted anthropomorphism scores (column 2). Highlights are introduced by the authors of the present work. Model names are redacted our of respect for privacy. The author's notes on each abstract are provided in column 4.

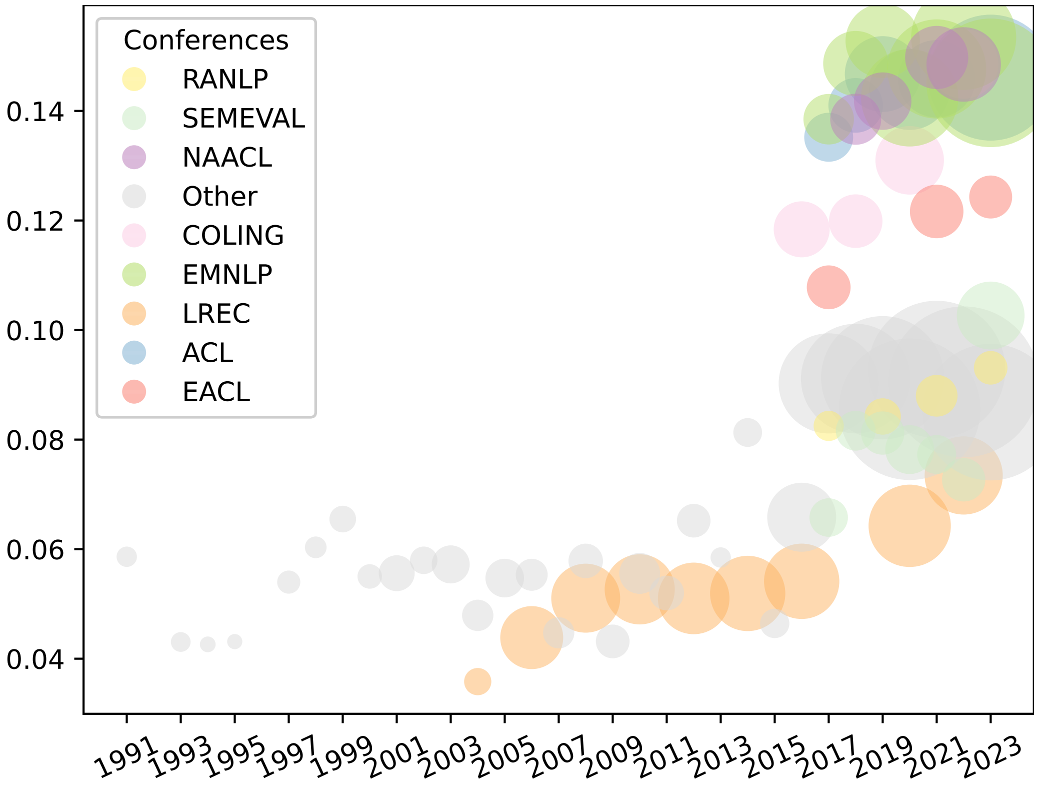

Figure 3: The average predicted anthropomorphism for different venues across time. The size of each bubble corresponds to the log of the number of abstracts considered.

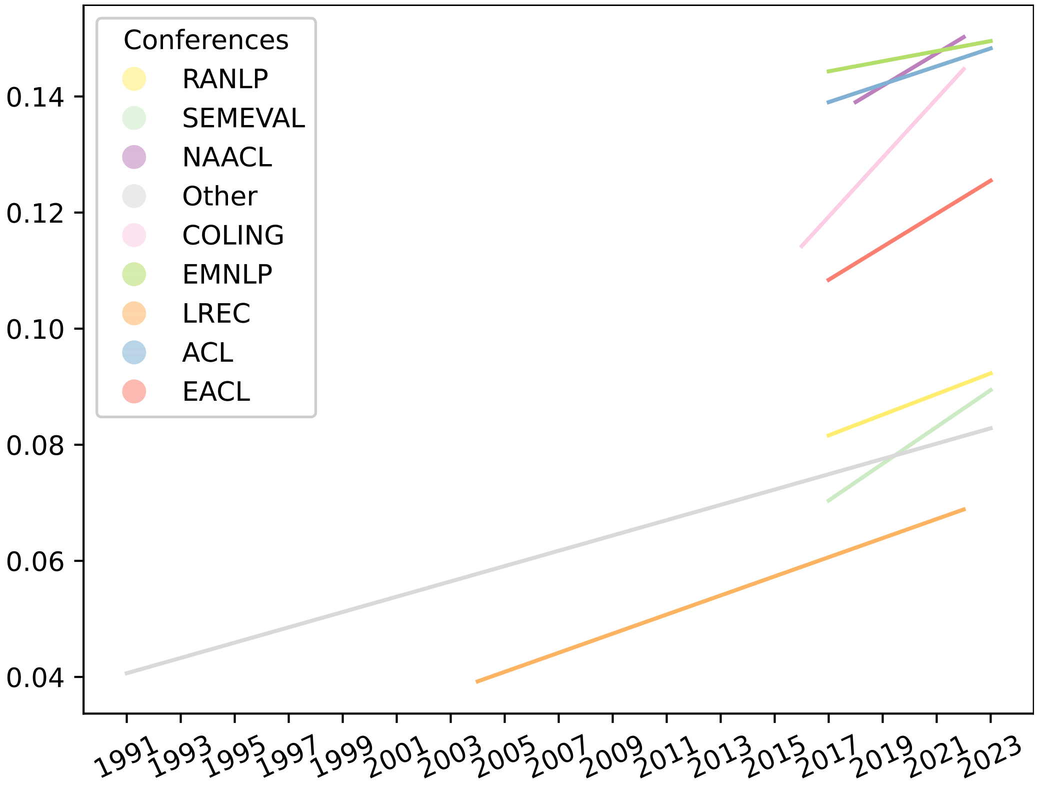

Figure 4: The average predicted anthropomorphism for different venues across time displayed as trend lines.

Model ID	**ICL Samples**	**MSE**	**MAE**	**R2**	**Pearson**
mlx-community/ Llama-3.2-3B-Instruct-8bit	0	0.083496	0.228560	-3.129379	0.046832
mlx-community/ Llama-3.2-3B-Instruct-8bit	10	0.041156	0.136085	-1.035409	-0.057462
mlx-community/ Mistral-7B-Instruct-v0.3-4bit	0	0.036576	0.127891	-0.808909	0
mlx-community/ Mistral-7B-Instruct-v0.3-4bit	10	0.036872	0.129467	-0.823520	-0.082796
mlx-community/ Meta-Llama-3-8B-Instruct-8bit	0	0.079303	0.236782	-2.922009	-0.043363
mlx-community/ Meta-Llama-3-8B-Instruct-8bit	10	0.037277	0.129479	-0.843542	0.051658

Table 7: Performance of different models with varying ICL sample counts