Learn, Achieve, Predict, Propose, Forget, Suffer: Analysing and Classifying Anthropomorphisms of LLMs

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

Anthropomorphism is a literary device where human-like characteristics are used to refer to non-human entities. However, the use of anthropomorphism in the scientific description and public communication of large language models could lead to misunderstanding amongst scientists and lay-people regarding the technical capabilities and limitations of these models. In this study, we present an analysis of anthropomorphised language commonly used to describe LLMs, showing that the presence of terms such as 'learn', 'achieve', 'predict' and 'can' are typically correlated with human labels of anthropomorphism. We also perform experiments to develop a classification system for anthropomorphic descriptions of LLMs in scientific writing at the sentence level. We find that whilst a supervised Roberta-based system identifies anthropomorphisms with F1-score of 0.564, state-of-the-art LLM-based approaches regularly overfit to the task.

1 Introduction

Effective scientific communication is predicated on two key tenets: accuracy and clarity. To effectively communicate, an author must accurately describe his or her findings, giving complete technical details and faithful explanation of methods. At the same time, the explanation must be sufficiently clear that a reader can interpret and understand the original intent of the author. Accuracy and clarity conflict in scientific reporting, leading to miscommunication. Overly technical language compromises understandability, whereas overly familiar language impedes the author from properly communicating the intricacies of their methodology.

Most authors find some compromise between accuracy and clarity. Sacrificing technical detail for friendly explanation or substituting turn-of-phrase for methodological justification. One such form of compromise in scientific reporting is the use of language reserved for characteristics of animate entities to describe the inanimate. *Anthropomorphism* is a long-held literary device, whereby non-humans are conferred with innately human characteristics. We might consider a city friendly, if we find its residents welcoming, or a car as obstinate if it does not start on a cold winter's morning. Anthropomorphism is an innate part of the human psyche and we are quick to infer agency on our environment. Further, we might define the idea of anthropomorphisation or anthropomimeticism as the active attribution of anthropomorphised qualities to inanimate agents (Inie et al., 2024).

Anthropomorphised terms are prevalent in the AI field, with 'machine learning', 'natural language understanding', 'computer vision', all being long standing examples of human characteristics inferred to algorithms. As large language models (LLMs) have become prevalent beyond the NLP field, the use of anthropomorphised terminology to describe interactions with LLMs has also grown among lay people. There is also a concerning tendency to adopt anthropomorphised terminology to describe scientific study (Cheng et al., 2024b).

In this work, we analyse anthropomorphised terms in the scientific literature (Section 4) making use of a recent corpus of anthropomorphisms in LLM reporting (Shardlow et al., 2025) and demonstrating that there are clear text markers for anthropomorphism. We additionally develop a method of text classification for anthropomorphic LLM reporting which operates at the sentence level in Section 5, which differs from prior approaches which have provided a document-level score.

We release all materials, including corpora, and information on the prompt setting via GitHub[1].

[1] `https://github.com/mattshardlow/Anthropomorphism_Corpus`

Proceedings of the First Interdisciplinary Workshop on OMMM associated with RANLP 2025,
pages 86–94, Varna, Bulgaria, Sep 8-10, 2025.
https://doi.org/10.26615/978-954-452-101-1-010

2 Related Work

AI anthropomorphism is a growing field of study (Brooker et al., 2019; Shardlow and Przybyła, 2024; Cheng et al., 2024b), which can be seen as a dimension of 'AI hype'. The term 'Artificial Intelligence' may be considered itself as anthropomorphising (Brooker et al., 2019), indicating that the agent possessing the inferred quality of 'AI' has attained a human characteristic. Anthropomorphic language in AI may also be applied to NLP tasks, such as 'reading *comprehension*' or 'sentiment *analysis*' (Lipton and Steinhardt, 2019).

Previous studies have sought to highlight the potential for harms apparent when anthropomorphising AI systems. 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). A concrete example of the danger of anthropomorphising AI systems is the case of false claims of sentience of the LaMDA model, with associated claims for employment rights, legal representation and beyond (Shardlow and Przybyła, 2024). 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. 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). Anthropomorphism is growing in the NLP literature with a recent study demonstrating a sharper rise in anthropomorphism for literature in the ACL anthology than for literature from general CS during the same period (Cheng et al., 2024b). 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 (Shardlow and Przybyła, 2024; Cheng et al., 2024b). 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).

We may also consider anthropomorphism through the lens of AI production in the field of dialogue systems. Efforts to categorise the anthropomorphic qualities of systems (Abercrombie et al., 2023) as well as the utterances they make (Gros et al., 2022) are fruitful first steps towards defining appropriate vocabulary for AI agents. Recently, a secondary study of datasets containing human-robot dialogues demonstrated that up to 80% of responses may reflect some form of self-anthropomorphisation (Li et al., 2024). There are clear implications of this work for the wider generative AI community in developing clear guidelines around ethical practices for the anthropomorphisation of LLMs (Cheng et al., 2024a).

3 Anthropomorphism Corpus

In our work we rely on the corpus gathered by Shardlow et al. (2025), which is a recent manually annotated corpus of anthropomorphic language in the context of NLP/AI modelling.

The Anthropomorphism Corpus was obtained by selecting 601 abstracts from the long papers of ACL 2022 and 49 news articles reporting on LLMs for a general audience. These abstracts and news articles were annotated at the sentence level for three categories: Non-anthropomorphic, ambiguous anthropomorphism and explicit anthropomorphism with the definitions taken from the work of Shardlow and Przybyła (2024) and reproduced here:

- *Non-anthropomorphic*: Any language which correctly describes the functioning of a model without implying human capabilities.
- *Ambiguous anthropomorphism*: Language which correctly describes the functioning of a model, but in a way that could be understood as the model having human capabilities (i.e., by a non-expert).
- *Explicit anthropomorphism*: Language that is unambiguously and erroneously used to claim a model possesses human capabilities.

	#	NA	AA	EA
Ab	3584	2770 (77.3%)	709 (19.8%)	105 (2.9%)
Jn	756	571 (75.5%)	130 (17.2%)	55 (7.3%)
All	4340	3341 (77.0%)	839 (19.3%)	160 (3.7%)

Table 1: Corpus statistics at the sentence level for the scientific abstracts (Ab), news articles written by journalists (Jn) 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.

We report summary statistics of the corpus in Table 1. The corpus contains 652 documents comprising 4340 claim sentences, each with a label indicating the degree of anthropomorphism on a 3-point scale.

4 Analysis of Anthropomorphism

We analysed the corpus to present insights on text features that are common for text classified as anthropomorphic. To perform this analysis, we identified common unigrams and bigrams to create a set of corpus-specific terms. We then create a vector for each term, which has S dimensions, where S is the number of sentences (4340) in our corpus. Each dimension has a 1 if the term is present in the sentence and a zero if the term is not present. We additionally manipulate the annotations to give a label vector for each analysis. The label vector is also of size S, containing one label per sentence. The sentences we consider and the method of determining the labels are adjusted for each analysis to expose a particular facet of the corpus. We finally calculate Pearson's correlation between the each term vector and the label vector, identifying the terms with the highest correlation with the labels (i.e., those terms that are typically present in a sentence when the label is also present).

4.1 Anthropomorphic Language

Firstly, we investigated the term correlations across our entire corpus when considering texts marked as non-anthropomorphic as compared to texts marked as ambiguous or explicit anthropomorphic. We assigned all non-anthropomorphic terms to a label of '0' and ambiguous or explicit anthropomorphic terms to a label of 1. We then calculated the correlations between the resulting label vectors and term vectors for each unigram and bigram.

Table 2 shows the unigrams and bigrams with the highest positive correlations to anthropomorphism across our entire corpus. We do not include high negative correlates as these are indicative of general language and did not show clear trends of non-anthropomorphic terms. The unigrams that are identified through this analysis are emblematic of the types of language that are typically included in anthropomorphic statements. The terms 'learn', 'learns' and 'learning' are identified as correlated with anthropomorphism. These typically occur in the sense of an algorithm 'learning' some feature of a problem or dataset. Although the term 'machine learning' is commonplace in the description of modern NLP systems, it is still inherently anthropomorphic. Further, when applying the term 'learning' to the ability of a model it may confuse a reader into believing that the model has some capacity for human level learning or assimilation of knowledge. Further, we see the terms 'achieve', 'achieves' and 'achieved' correlated with anthropomorphism. This pattern of anthropomorphism occurs when describing the model itself as 'achieving' some goal. We also note the presence of terms such as 'predict', 'propose', 'forgetting' and 'suffer', which all indicate human actions which have been used to describe inanimate models. The bigrams

Correlation	Term	Freq
0.170	learn	82
0.149	achieve	64
0.133	achieves	98
0.113	learns	28
0.096	learning	255
0.084	predict	40
0.074	achieved	37
0.071	propose	196
0.070	forgetting	15
0.068	suffer	17
0.101	to learn	44
0.084	and achieve	3
0.075	have achieved	18
0.074	learns a	6
0.072	to predict	28
0.071	and achieves	17
0.07	achieves state-of-the-art	15
0.066	learn from	8
0.066	models can	23
0.065	achieves the	12

Table 2: Highest correlated unigrams and bigrams for anthropomorphic language

that are identified by this analysis give some additional context to the unigrams, indicating where terms such as ‘learn’ and ‘achieve’ are typically used. Interestingly, the term ‘models can’ is identified as correlated with anthropomorphism, which may be used to indicate some range of anthropomorphic abilities that are inferred to a model.

4.2 Explicit Anthropomorphism

Correlation	Term	Freq
0.145	student	21
0.132	Then	26
0.127	*Product 1*	16
0.127	*Product 2*	13
0.126	added	12
0.126	post-hoc	5
0.126	inherently	5
0.126	inherent	4
0.126	*Product 3*	3
0.124	describing	6
0.143	while the	12
0.127	them to	20
0.126	her to	3
0.126	a framework	6
0.126	said that	6
0.126	a language	11
0.126	inherently faithful	3
0.126	faithful models	3
0.126	post-hoc explanations	3
0.124	work in	11

Table 3: Highest correlated unigrams and bigrams for explicit anthropomorphic language. We have anonymised the names of proprietary products.

In the annotation schema two levels of anthropomorphism are present: Ambiguous and Explicit. We determine lexical features that distinguished between these two categories by using the same methodology as above, but adapting the transformation of the labels. To conduct this analysis, we only considered the portion of our corpus that was annotated as either ambiguous anthropomorphic or explicit anthropomorphic. We assigned ambiguous anthropomorphic texts a score of zero and explicit anthropomorphic a score of one and calculated Pearson correlation against the one-hot encoded vectors. The results of this analysis are shown in Table 3.

The term with the highest correlation is ‘student’. This typically occurs in the context of ‘student models’ as used in the task of model distillation. It is also notable that the names of several proprietary products are correlated, indicating that descriptions of commercial activities are more likely to be explicitly anthropomorphic than ambiguous anthropomorphic. The bigrams that are identified indicate elements of anthropomorphism (‘said that’, ‘faithful models’, etc.). There is also some noise in this analysis, with ‘Then’, ‘them to’ and ‘her to’ also included. The noise is likely due to the small corpus size (there were only 160 instances of explicit anthropomorphism).

4.3 Journalistic Writing

Correlation	Term	Freq
0.184	ask	12
0.17	respond	5
0.151	*Product 4*	17
0.151	human	180
0.143	scenario	18
0.138	questions	91
0.138	response	39
0.136	though	8
0.128	visual	51
0.125	point	10
0.128	what it	5
0.128	respond to	5
0.111	and destroy	2
0.111	to prompts	3
0.111	to kill	3
0.111	while the	12
0.111	you ask	3
0.111	responses to	4
0.108	it was	11
0.09	data points	5

Table 4: Highest correlated unigrams and bigrams for anthropomorphic language in the journalism sector. We have anonymised the names of proprietary products.

Finally, we present an analysis of features that are indicative of anthropomorphism in journalistic writing. We analysed the portion of the corpus extracted from journalistic sources and compared examples of non-anthropomorphic language to examples of ambiguous or explicit anthropomorphic language using the methodology described in Section 4.1. The results of this analysis are presented in Table 4.

The examples of anthropomorphic language from journalistic texts make use of metaphorical or extreme language such as ‘destroy’ or ‘kill’. Journalistic sources are sensational in their reporting of anthropomorphic language as evidenced by terms such as ‘destroy’ and ‘kill’. Anthropomorphic terms in journalistic sources focus on the interaction of humans with LLMs as evidenced by terms such as ‘ask’, ‘respond’ and ‘question’ indicating anthropomorphised dialogue.

5 Sentence Classification

This section reports on the development of text classification methods to distinguish between anthropomorphic and non-anthropomorphic sentences.

5.1 Data Processing

We split the available data into train (80%) and test (20%) partitions ensuring that the splits were stratified and that both genres occurred evenly across each subset. The distribution of labels was also preserved. As Explicit Anthropomorphism is a minority class (3.7%), we conflated this class with Ambiguous Anthropomorphism leading to a two-class problem with the labels: 'Non-Anthropomorphic' and 'Anthropomorphic'.

In our corpus, 77.0% of identified claims were labelled as non-anthropomorphic. Imbalanced data can lead to a classifier overly relying on one class and so we explored two different methods of balancing our classes for the training set. We did not perform any adjustments to the data distribution in the test set to reflect the real-world class distribution. Firstly, we employed down-sampling of the majority class. In this setting, we selected a random sample of the Non-Anthropomorphic examples (2678) in our corpus which was the same size as the Anthropomorphic examples (778 examples of each class). The down-sampling method led to perfectly balanced data, but involved discarding 1900 examples of non-anthropomorphic text. We further explored up-sampling the minority class through the use of the Parrot Paraphraser (Damodaran, 2021). The Parrot Paraphraser relies on a T5-based paraphrase model (Raffel et al., 2020) and provides metric-based filtering for adequacy, fluency and lexical diversity of the returned paraphrases. We used the Parrot Paraphraser in the default configuration to produce an additional 1538 examples of anthropomorphised claim sentences. We again, balanced the classes in this setting to give 2316 examples in each class. Statistics for each train setting and for the test setting are given in Table 5.

5.2 Baseline Approaches

We provide minority and majority class baselines (i.e., assigning the anthropomorphic, or non-anthropomorphic labels to all classes respectively). This approach demonstrates a baseline effect of a classifier which has not adapted to the task and fails to make any discriminative judgements. We also include two randomised baselines. Firstly, we include a random baseline where each class is equally likely to be assigned (random 1:1). Secondly, we also include a random baseline where the non-anthropomorphic label is 3 times more likely than the anthropomorphic label, reflecting our data distribution. These approaches represent the base performance of a classifier which is making randomised decisions, either with respect to the class label, or with respect to the data distribution. We provide these baselines as we believe they are a reasonable means of contextualisation of the results from the other approaches as described below.

Partition	Sampling	NA	A
Test	None	663	221
Train	None	2678	778
Train	Down	778	778
Train	Up	2316	2316

Table 5: Data settings used for evaluation of sentence classification. Down-sampling and up-sampling are used to create a balanced training set, however the test-set remains imbalanced throughout all experiments reflecting the nature of the corpus. NA refers to Non-anthropomorphic annotations. A refers to Anthropomorphic annotations consisting of explicit anthropomorphism and ambiguous anthropomorphism.

5.2.1 ML Classifiers with SciKitLearn

We used Random Forest (Breiman, 2001) and SVM (Cortes and Vapnik, 1995) from SciKitLearn (Pedregosa et al., 2011). To convert each sentence into a numerical format we employed (a) BOW vectorisation via the CountVectorizer library in SciKitLearn and (b) sentence embeddings using Sentence Transformer (Reimers and Gurevych, 2019). We used the default configurations in SciKitLearn for the Random Forest and SVM and did not tune the hyperparameters in each case (due to the small size of our data).

5.2.2 BERT-based classifiers with Transformers

We used the following models via the Transformers library in Python downloaded from the HuggingFace hub:

```
google-bert/bert-base-uncased
google-bert/bert-large-uncased
FacebookAI/roberta-base
FacebookAI/roberta-large
allenai/scibert_scivocab_uncased
```

All models were fine-tuned against the training set under each train-setting for 5 epochs using the AdamW optimiser with learning rate of 4×10^{-5}. In some cases the model failed to converge, in which case the training process was repeated.

Baseline	Acc	Anthropomorphic		
		R	P	F1
Majority Class	0.750	0.000	0.000	0.000
Minority Class	0.250	1.000	0.250	0.400
random 1:1	0.494	0.471	0.240	0.318
random 3:1	0.618	0.226	0.230	0.228

Table 6: Baseline results for anthropomorphism classification

5.2.3 Prompt Engineering with MLX

We also experimented with MLX, a library for MacOS for implementing LLMs. In this case we used an 8B version of Llama3.1 (Grattafiori et al., 2024), specifically the model available at the HuggingFace Hub here: ”mlx-community/Meta-Llama-3.1-8B-Instruct-bf16”, which is 4-bit quantised. We used this model for in-context learning (Wei et al., 2022), in which case we simulated a multi-turn conversation between the LLM and the user, demonstrating examples of anthropomorphic and non-anthropomorphic sentences and their classifications. The model was then presented with a new sentence from the test set and the response it generated was interpreted as the classification. We also fine-tuned Llama for this task under the same setting using examples from the training set. We also include a zero-shot classification setting.

5.2.4 Closed-source LLMs

We additionally performed in-context learning in a 100-shot setting using the same prompts as before and a 100-shot in-context learning setting drawn from the training set. We accessed GPT-4o and GPT-4 Turbo on the 8th November 2024 via the web-based API. The total costs were 7 dollars for GPT-4o and 33 dollars for GPT-4 Turbo for a single run through the entire test set (n=884) in each case. We compare these results to LLama3.1 in a 100-shot setting. All results are shown in Table 9.

6 Results

We present results for baseline approaches (Table 6), machine learning classifiers (Table 7), prompt engineering (Table 8) and GPT-4 models (Table 9). For each table, we have presented Accuracy (the percentage of all correct instance regardless of class), as well as the Precision, Recall and F1-score for the anthropomorphic class.

We provided four heuristic baseline approaches examining different approaches to classification in our corpus as demonstrated in Table 6. As our data is split 75:25 between the majority (Non-anthropomorphic) and minority (Anthropomorphic) classes, we observe that the majority and minority baselines reflect this. We have only reported F1-score for the anthropomorphic class as this is the feature we are trying to identify. This means that whilst the accuracy for the majority class baseline is 0.75 (all the non-anthropomorphic examples were correctly identified), the Recall, Precision and consequently the F1-score are all 0, as no non-anthropomorphic examples were identified. Conversely, the minority class baseline does much worse in terms of accuracy (0.25), but has perfect recall by retrieving all anthropomorphic examples.

Train	Method	Acc	Anthropomorphic		
			R	P	F1
O	SVM-BOW	0.753	0.018	0.800	0.035†
	SVM-ST	0.750	0.018	0.500	0.035†
	RF-BOW	0.753	0.018	0.800	0.035†
	RF-ST	0.750	0.018	0.500	0.035†
	bert-base	0.784	0.403	0.601	0.482
	roberta-base	0.739	0.059	0.361	0.101
	scibert-base	0.768	0.362	0.556	0.438
D	SVM-BOW	0.613	0.475	0.317	0.380
	SVM-ST	0.617	0.647	0.354	0.458
	RF-BOW	0.613	0.475	0.317	0.380
	RF-ST	0.617	0.647	0.354	0.458
	bert-base	0.670	0.706	0.407	0.517
	roberta-base	0.708	0.756	0.450	**0.564**
	scibert-base	0.660	0.715	0.399	0.512
U	SVM-BOW	0.683	0.416	0.379	0.397
	SVM-ST	0.657	0.579	0.379	0.458
	RF-BOW	0.683	0.416	0.379	0.397
	RF-ST	0.657	0.579	0.379	0.458
	bert-base	0.777	0.462	0.567	0.509
	roberta-base	0.784	0.471	0.584	0.521
	scibert-base	0.757	0.339	0.521	0.411

Table 7: The results of classifying anthropomorphic and non-anthropomorphic sentences. Best F1-score in bold. Three training settings are explored: **O**riginal, **D**own-sample and **U**p-sampled. †The F1 scores for these two values appear the same due to rounding. This is an effect of the low-recall in both instances masking the substantial difference in precision.

We also provide two randomised baselines. The random 1:1 baseline has a lower accuracy, but higher F1 score (owing to a higher recall) than the random 3:1 score. This is effectuated by the random 1:1 baseline over-predicting the prevalence of anthropomorphic terms in the data. Nevertheless, the random 1:1 baseline still has a lower F1-score than the Minority class baseline.

These baselines serve to help the reader understand and interpret the behaviour of the classifiers that we present in our results. Whilst we will see

Method	N	Acc	Anthropomorphic		
			R	P	F1
0-shot	0	0.707	0.389	0.410	0.399
ICL	1	0.537	0.738	0.317	0.444
	2	0.467	0.824	0.296	0.436
	3	0.518	0.765	0.311	0.442
	4	0.577	0.692	0.333	0.450
	5	0.399	0.869	0.277	0.420
	7	0.506	0.733	0.300	0.426
	9	0.563	0.674	0.322	0.436
FT	1	0.644	0.566	0.363	0.442
	2	0.650	0.529	0.363	0.431
	3	0.567	0.683	0.325	0.441
	4	0.648	0.538	0.363	0.434
	5	0.733	0.258	0.442	0.326
	7	0.698	0.394	0.395	0.395
	9	0.679	0.457	0.381	0.416

Table 8: The results of using LLama3.1 to classify anthropomorphic and non-anthropomorphic sentences. ICL refers to In-context Learning. FT refers to Fine-tuning. The number of examples (N-shots) at inference time is also presented.

Method	Acc	Anthropomorphic		
		R	P	F1
GPT4o	0.763	0.181	0.588	0.277
GPT-4-Turbo	0.766	0.176	0.609	0.274
Llama3.1	0.729	0.308	0.439	0.362

Table 9: Comparison of GPT-4 models and Llama in a 100-shot in-context learning setting to classify anthropomorphic and non-anthropomorphic sentences.

that many of our classifiers attain a high accuracy, many do so at the severe compromise of F1-score, indicating that all or most predictions were to the majority class. We also see that there is a lower bound on the F1-score as evidenced by the random classification. Any systems scoring higher than this can be interpreted as performing better than random, i.e., indicating learning has taken place.

We tested two Machine learning approaches: SVMs and Random Forests with features coming from a Bag-of-words and from Sentence Transformers. Our results in Table 7 showed little difference between these approaches and typically that the classifiers were not able to reliably predict the presence of anthropomorphic language in a sentence with the accuracy and F1-scores falling below baseline in most cases. The Sentence Transformer features gave higher scores than the BOW features in the down-sampled and up-sampled settings, but not in the original setting where minimal learning took place as evidenced by the extremely low recall. We note that in all cases the SVM and RF algorithms returned the same scores under the same settings indicating that the same decision manifold was learnt in each case. This indicates that the task is more complex than simply relying on word features (i.e., no word is a strong indicator) and that the sentence embeddings did not provide sufficient information for the classifiers. It may be possible that a larger dataset of anthropomorphic language would permit the algorithms to learn a more comprehensive representation of the feature space and perform better at test time.

Following on from this, we also tried three transformer based approaches for sentence classification. We observe some slight improvement in Roberta as compared to Bert in the down-sampled and up-sampled settings (see Table 7). We additionally noted that Scibert performed worse than Bert and Roberta in the down-sampled and up-sampled settings. In the original setting Roberta did not accurately retrieve anthropomorphic examples (Recall = 0.059), however Bert still marginally outperformed Scibert on all metrics. Our best performing system in terms of the F1 metric was Roberta-base in the down-sampled setting. This returns an F1 score of 0.564 made up of a recall score of 0.756 and a precision of 0.450 indicating that the model over-estimated the degree of anthropomorphism in the corpus (i.e., this result occurs because the model tended to label non-anthropomorphic sentences as anthropomorphic).

We explored three experimental settings for our training dataset in Table 7, whilst keeping the test data as a constant split in all experiments. The original setting had a 3:1 distribution of non-anthropomorphic and anthropomorphic sentences whereas the down-sampled and up-sampled data had a 1:1 ratio in each case. Balancing the classes in the training set led to a clear improvement in classification ability for all models. Whilst the up-sampled data typically exhibits a higher precision than the down-sampled data, the overall F1-scores for the transformer-based methods are lower, owing to a drop in recall for these methods. Whereas we had expected that including more data would lead to an overall improvement in scores this was not the case and may well be due to the fact that our up-sampled data included synthetic examples

Method	Acc	Anthropomorphic		
		R	P	F1
Journalism	0.724	0.553	0.712	0.622
Abstracts	0.771	0.505	0.679	0.579

Table 10: F1 scores for separate genre subsets within our corpus.

that were not suitable representations of the type of information seen at testing time.

In Table 8 we present the results of our experiments with Llama 3.1 (the most advanced version of Llama available at the time of experimentation) in an ICL and Fine-tuned setting with 1-9 examples as well as a zero-shot approach. The zero-shot experiment demonstrates that an LLM such as Llama is able to correctly answer in some cases for our task without any task specific information being introduced as the F1-score of 0.399 is above the randomised baselines. Compared to zero-shot, we can observe that strategies such as ICL and fine-tuning with 1-9 examples improved the F1-score marginally, but that there was no significant improvement by including more examples or between both techniques.

We also compare Llama 3.1 in a 100-shot ICL-setting to the equivalent experiment with the closed source OpenAI models GPT4o and GPT-4-Turbo in Table 9. Whilst the accuracy is improved for the GPT models compared to Llama3.1, the F1-score indicates that the GPT models underperform providing classification results which are indistinguishable from a random baseline. We were not able to identify a strategy using newer LLMs such as Llama and GPT that performed better for our corpus than the Roberta-base system which makes use of a much smaller version of the transformer architecture.

We present an additional analysis of our results in Table 10, where we show the performance of sentence classification for each of the sub-genres represented in our dataset. These results were produced by using Roberta-Base and training in the down-sampled setting (i.e., the system with the best F1-score in our prior experiment). The results show that the F1 score for sentences from the journalism genre is higher than for the scientific abstracts.

7 Discussion

In writing this work (and other works on the topic) it became apparent to the first author that reporting on LLMs is difficult, and maybe impossible, without leaning on anthropomorphised terminology. As such, the description of the methods and results herein necessarily contains some anthropomorphism and the authors have deliberately left this in-situ. We are not advocating for the abolishment of anthropomorphised terminology, but rather seeking to better understand and quantify the phenomenon. The value of an anthropomorphism classification tool is not to punish authors who lean on metaphors, but rather to better equip scientists and the general public with tools for understanding the way we describe LLMs.

Anthropomorphism is of course not limited to the study of large language models and one may envision a similar study on other technology (e.g., self-driving cars, drones, etc.). We do not seek to make claims about anthropomorphisation outside of the realm of LLMs, however we do expect that similar phenomena are apparent and that the work here may be a good starting point for adaptation to other areas of study.

An interesting finding of our work is that despite extensive study, we were unable to improve the performance of the LLM approach beyond that of the random baselines (e.g., GPT4o/GPT4-Turbo in the 100-shot ICL setting). A deliberately anthropomorphised interpretation of this finding may be that LLMS don't *know* when they are being anthropomorphised. Of course, our study is non-exhaustive and there may well be alternative methods of LLM-prompting strategies beyond our study that would yield improved results.

8 Conclusion

In this work, we have presented an analysis of anthropomorphism in scientific reporting of LLMs as well as experiments on developing new classifiers for sentence-level anthropomorphism. Our most promising results show that we are able to produce sentence classifications which outperform reasonable baselines. The use of Bert-based models was most effective in our study as compared to machine learning classifiers or prompt engineering. Our work lays the foundation for future studies on anthropomorphism classification at the sentence level and beyond.

References

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.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

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, Alicia DeVrio, Lisa Egede, Su Lin Blodgett, and Alexandra Olteanu. 2024a. " i am the one and only, your cyber bff": Understanding the impact of genai requires understanding the impact of anthropomorphic ai. *arXiv preprint arXiv:2410.08526*.

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

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Prithiviraj Damodaran. 2021. Parrot: Paraphrase generation for nlu.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

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.

Yu Li, Devamanyu Hazarika, Di Jin, Julia Hirschberg, and Yang Liu. 2024. From pixels to personas: Investigating and modeling self-anthropomorphism in human-robot dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9695–9713, Miami, Florida, USA. Association for Computational Linguistics.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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):1–26.

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.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.