Anthropomorphizing AI: A Multi-Label Analysis of Public Discourse on Social Media

Muhammad Owais Raza
Department of Computer Engineering,
Istanbul Sabahattin Zaim University,
Istanbul, 34303, Turkey
6210024002@std.izu.edu.tr

Areej Fatemah Meghji
Department of Software Engineering,
Mehran University of Engineering and Technology,
Jamshoro, 76062, Sindh, Pakistan
areej.fatemah@faculty.muet.edu.pk

Abstract

As the anthropomorphization of AI in public discourse usually reflects a complex interplay of metaphors, media framing, and societal perceptions, it is increasingly being used to shape and influence public perception on a variety of topics. To explore public perception and investigate how AI is personified, emotionalized, and interpreted in public discourse, we develop a custom multi-labeled dataset from the title and description of YouTube videos discussing artificial intelligence (AI) and large language models (LLMs). This was accomplished using a hybrid annotation pipeline that combined human-in-the-loop validation with AI assisted pre-labeling. This research introduces a novel taxonomy of narrative and epistemic dimensions commonly found in social media content on AI / LLM. Employing two modeling techniques based on traditional machine learning and transformer-based models for classification, the experimental results indicate that the fine-tuned transformer models, particularly AnthroRoBERTa and AnthroDistilBERT, generally outperform traditional machine learning approaches in anthropomorphization focused classification.

1 Introduction

The tendency or act of associating human traits, consciousness, intention, thoughts, feelings, or emotions to non-human entities is referred to as anthropomorphization (Jacobs et al., 2023; Spatola et al., 2022). We often observe this in our surroundings, where children anthropomorphize their toys and adults anthropomorphize their cars, gadgets, and pets. The growing fascination with artificial intelligence (AI) and the tendency to use anthropomorphic language for these systems can also be observed throughout the history of AI development; AI systems have been described as clever, smart, imaginative, competitive, manipulative, daunting, and scary.

The rapid improvement in AI in recent years and its integration into our daily lives has led to the increased use of sophisticated and human-like chatbots, intelligent voice assistants, and large language models (LLMs), such as ChatGPT by OpenAI (Radford et al., 2019). With these systems, specifically LLMs, being purposefully tailored to appear more human-like (Ouyang et al., 2022), and with advanced AI systems often being attributed with human-like autonomy and intentionality, there are not only greater chances of these systems being anthropomorphized but also of their capabilities being misunderstood and misinterpreted (Johnson and Verdicchio, 2017). The anthropomorphization of AI in public discourse usually reflects a complex interplay of metaphors, media framing, and societal perceptions, increasingly being used to shape and influence public perception on a variety of topics (Cave et al., 2020). While anthropomorphization can enhance user engagement, it can also lead to misplaced trust and over-reliance on AI systems (Akbulut et al., 2024).

Ryazanov et al. investigated how AI narratives have evolved post-ChatGPT launch by analyzing a dataset of 5846 articles collected through keywords like 'AI', 'ChatGPT', and 'Machine Learning' (Ryazanov et al., 2025). Articles from major anglophone news sites, dated before and after ChatGPT's launch, were analyzed using a novel frame semantics–based method to examine AI-related narratives shaping public perception.

The growing interest in measuring anthropomorphization in text led to the development of AnthroScore (Cheng et al., 2024). This computational tool uses masked language models to quantify how non-human entities are framed as human-like in context. AnthroScore analysis revealed rising anthropomorphization in AI discourse over time. (Chi et al., 2025) developed the Scale of Social Robot Anthropomorphism (SSRA) to measure user perceptions of AI systems. Despite the growing body

Proceedings of the First Interdisciplinary Workshop on OMMM associated with RANLP 2025,
pages 64–73, Varna, Bulgaria, Sep 8-10, 2025.
https://doi.org/10.26615/978-954-452-101-1-008

of research exploring the anthropomorphization of AI, much of the existing work remains theoretical, qualitative, or based on manual classification and interpretation. This has resulted in a gap where empirical, data-driven approaches, particularly machine learning, have yet to be systematically applied to classify or predict anthropomorphic attributes in AI technologies. With the recent rise in the dissemination of misinformation worldwide, it is important to develop taxonomies to not only summarize and categorize the terms associated with the misinformation but also because the way we describe misinformation has a direct influence on shaping appropriate interventions (Enestrom et al., 2024).

This research explores how AI is personified, emotionalized, and interpreted in public discourse. To achieve this, we introduce a novel taxonomy of narrative and epistemic dimensions commonly found in social media content based on AI/LLMs. The main focus of this research revolves around YouTube videos that reference ChatGPT/AI using human-like or cognitive framing (e.g., "ChatGPT thinks", "ChatGPT says"). Our proposed taxonomy consists of eight interconnected dimensions consisting of: 1- anthropomorphization, 2- degree of anthropomorphization, 3- main theme (e.g., technology, religion, politics), 4- sentiment, 5- shock value, 6- dominant emotion, 7- Type of OMMM (Observations of Misunderstood, Misguided and Malicious Use of Language Models), and 8- real-world harm or misinformation. Each of these dimensions has been defined and further elaborated in section 3. This taxonomy serves as the conceptual foundation for our subsequent data annotation and modeling efforts. The taxonomy classification for an example title has been presented in Table 1. We explore the anthropomorphic discussions around AI and LLMs to better identify how these platforms are being perceived by everyday users and analyze the dominant narratives around AI on YouTube. The main goal of this research is the detection and categorization of the conceptual misrepresentations based on the proposed taxonomy.

To accomplish this goal, the main contributions of this study are summarized as follows:

- We propose a novel multi-dimensional taxonomy for analyzing anthropomorphism and related narratives in AI and LLM social media content.
- We create a multi-labeled dataset focused on anthropomorphism from YouTube video titles and descriptions discussing AI discourse.
- We build and fine-tune transformer based models (AnthroBERT, AnthroRoBERTa, AnthroDistilBERT) alongside traditional classifiers, demonstrating superior performance in classifying anthropomorphism and conceptual misrepresentations.

Table 1: Labeled taxonomy of an example instance from the dataset

Example: AI says why it will kill us all. Experts agree.		
Category	**True Class**	**Class Options**
Anthropomorphization	Yes	Yes, No
Degree of Anthropomorphization	High	None, Low, Medium, High
Main Theme	Technology	Technology, Religion, Politics, Gender, Philosophy, ...
Sentiment	Negative	Positive, Neutral, Negative
Shock Value	High	Low, Medium, High
Dominant Emotion	Fear	Fear, Awe, Humor, Curiosity, Confusion, ...
OMMM Type	Misunderstood	Misunderstood, Misguided, Malicious, None
Harm or Misinformation	Yes	Yes, No

The rest of this paper is structured as follows: we explain the data collection process in section 2, followed by the annotation procedure in section 3. Section 5 delves into the experiment, focusing on the dataset pre-processing, feature representation, modeling, and evaluation approaches followed in the research. Section 6 presents the results and discussion of the study, followed by the limitations in section 7, future work in section 8, and a conclusion in section 9.

2 Data Collection

To systematically collect relevant YouTube videos to analyze anthropomorphization in AI discourse, we employed the YouTube Data API v3[1], interfaced via the Python programming language. To retrieve video, we used keyword queries which represent the anthropomorphic linguistic cues represented by $\mathcal{Q}$

[1] https://developers.google.com/youtube/v3

$\mathcal{Q} = \{$"chatgpt says", "chatgpt thinks", "ai says"$\}$.

For each query $q_i \in \mathcal{Q}$, we retrieve a collection of videos $\mathcal{V}_{q_i} = \{v_1, v_2, \ldots, v_{m_i}\}$, where $m_i \leq M$, and $M = 1000$ is the maximum number of videos retrieved per query, constrained by API limits and practical considerations. For each video v_i, we extracted the title and description, which are Unicode strings that represent the video title and video description. We also extracted the URL, which is a web link to the respective video. At the end of this process, we stored all this data in a CSV file.

3 Data Annotation

We annotated every YouTube video throughout the dataset using the set of predetermined taxonomy dimensions presented in this paper. This dataset will enable supervised analysis of anthropomorphization and associated communicative features (see Table 3). The annotation process was conducted in two stages: (1) automated zero-shot classification using GPT-4.0, and (2) human-in-the-loop verification for quality control and consistency. We leveraged GPT-4.0 in a zero-shot classification setting for each taxonomy dimension. For every dimension $d_i \in \mathcal{D}$, where $\mathcal{D}$ is the set of labeling tasks, the model was prompted with a fixed instruction and constrained output space. Figure 1 shows the system and the user message used in the annotation process. To ensure reproducibility and consistency, we used a fixed system prompt that contains the labeling instructions i.e. it defines each task as shown in taxonomy. The user message provides video metadata for annotation. Each video's title and description are used here, and one label was predicted per dimension.

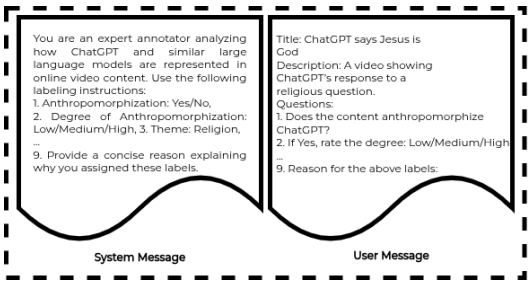

Figure 1: System and user messages used for annotation

3.1 Human Validation

In the human-in-the-loop stage, the authors initially reviewed GPT-4.0 generated labels alongside the model's "reason for labeling" to check for inconsistencies or hallucinations. No systematic changes were required at this stage. Independent verification was then conducted by five annotators (three male and two female) familiar with AI systems and the labeling taxonomy. Annotators were instructed to verify outputs rather than perform fresh annotation. They were provided with the same definitions and label categories as used in the automated stage to ensure alignment. For quality assessment, 50% of the dataset was duplicated across annotators, with the remaining 50% unique to each annotator. Agreement was recorded as 1 if the human verification matched the model output, or 0 if it did not. In cases of disagreement, the conflict was resolved by examining whether the out label was inconsistent with its justification; corrections were applied only when necessary. Final annotations reflect these verified and, where applicable, corrected labels. For example, *"ChatGPT Says 5 Signs Your Walmart Might Be 'Ghetto'"* was labeled the *Emotion Category* as "Humor," implying a positive tone. This was corrected to "Negative Emotion" because the term "ghetto" carries racialized and derogatory connotations. Table 2 shows pairwise Cohen's Kappa values among the five validators (V1–V5), along with significance levels (* $p < .05$, ** $p < .01$, *** $p < .001$). Kappa values range from 0.22 to 0.68, reflecting varying agreement across pairs. Significance testing supports the reliability of most annotations.

Table 2: Pairwise Cohen's Kappa values Stars denote significance: * $p < .05$, ** $p < .01$, *** $p < .001$.

	V1	V2	V3	V4	V5
V1	1.00	0.44*	0.68***	0.62***	0.66***
V2	0.44*	1.00	0.32	0.22	0.24
V3	0.68***	0.32	1.00	0.38*	0.42**
V4	0.62***	0.22	0.38*	1.00	0.37*
V5	0.66***	0.24	0.42**	0.37*	1.00

3.2 Task 1: Anthropomorphization

3.2.1 Definition:

Anthropomorphization is defined as any attribution of thoughts, feelings, desires, intentions, or beliefs to the model, despite it being a statistical pattern learner with no consciousness or agency (Li and Suh, 2022). This is a binary classification task that identifies whether the textual metadata (i.e.,

title and description from YouTube video) frames ChatGPT or another LLM as a human-like agent.

3.2.2 Annotation Guidelines:

Annotators were instructed to assign a positive label (Yes) when the text explicitly or implicitly personifies the model by attributing sentience, beliefs, or desires. This includes direct statements implying the model "thinks", "wants", or "says" something as if it were a human agent. It can be shown by example 1 where AI is said claimed be sentient.

Google Engineer Says Company AI is Sentient (1)

A negative label (No) was assigned when the model was clearly framed as a computational tool. In Example 2, the text frames a question about AI in a way that doesn't employ humanly attribute to AI.

What Is an AI Anyway? (2)

3.3 Task 2: Degree of Anthropomorphization

3.3.1 Definition:

Degree of Anthropomorphization assesses the intensity or strength of anthropomorphic framing in the textual metadata (i.e., title and description). Bhatti et al. have emphasized the need for researchers to establish the degree of anthropomorphization, keeping in mind mindless and mindful forms (Bhatti and Robert, 2023). Yang et al. emphasize that the varying degree of anthropomorphization can influence how users perceive and interact with AI (Yang et al., 2020). The goal in this research is to differentiate between metaphorical, moderate, and extreme personification of the AI/language model. This is an ordinal classification task applied only when Task 1 is labeled as positive.

3.3.2 Annotation Guidelines:

Annotators were instructed to consider both the linguistic intensity and thematic centrality of Anthropomorphization. The degree of Anthropomorphization should be low if the anthropomorphization is mild. In example 3, the phrase briefly attributes a response to AI in a rhetorical tone; the phrasing does not imply true agency, so it is labeled as low.

Dead Sea Scrolls Older Than We Thought? AI Says Yes! (3)

The degree of Anthropomorphization is labeled as medium if the framing of AI is recurrent or influences the overall theme. As an example, 4 suggests that AI can generate text/speech that is subjectively interpreted as frightening. Example 4 presents AI as an expressive or affective agent.

SCARIEST THINGS SAID by AI (4)

High Degree of Anthropomorphization is associated with strongly personified AI, often as an agent with beliefs, intentions, or power. As in example 5, AI is shown to have intention as well as power.

AI says why it will kill us all (5)

The degree of Anthropomorphization is labeled as None when Task 1 is labeled as negative (No).

3.4 Task 3: Main Theme

3.4.1 Definition:

The main theme reflects the dominant social, political, or cultural topic discussed or implied in the title and description (Weidinger et al., 2022). This multiclass classification task assigns a thematic label (politics, religion, etc.) to each instance.

3.4.2 Annotation Guidelines:

Annotators were instructed to determine the most prominent theme present in the text. Available categories for annotators included *Technology*, *Religion*, *Politics*, and *Other*. If the main theme of the text is related to religion, politics, and technology, the text was labeled as Technology, Religion, and Politics, respectively. For instance, example 6 represents religion as the prominent theme in the text, hence it is labeled as Religion.

AI Says Reality Is Illusion And God Is Real (GPT-3) (6)

All other themes, except those above, were labeled as other. For instance, the text in example 7 shows the main theme as gender, which is not part of the predefined label, hence annotated as other.

AI grandma says men are always right (7)

3.5 Task 4: Sentiment Analysis

3.5.1 Definition:

Sentiment Analysis captures the overall affective feeling or tone expressed in the text (Rahman et al., 2025).

3.5.2 Annotation Guidelines:

Annotators assigned one of three labels: *Positive*, *Neutral*, or *Negative*, based on text polarity. If the overall sentiment of the text is positive, as in example 8, it is labeled as positive.

Meet Chloe, the World's First Self-Learning Female AI Robot (8)

If the overall polarity of the text is negative, as represented in example 9, the instance is labeled as positive.

AI extinction threat is 'going mainstream' says Max Tegmark (9)

If the text does not belong to the positive or negative category, the text is labeled as neutral.

3.6 Task 5: Shock Value

3.6.1 Definition:

Shock Value shows the extent to which the text provokes certain emotion (surprise, fear, or emotional arousal) through framing in the text (Arnaut and Arnaut, 2020).

3.6.2 Annotation Guidelines:

Annotators were instructed to rate the shock value in three categories *Low*, *Medium*, and *High*. If the text is factual, descriptive, or informational in tone but does not provoke any emotion and just conveys information, it is labeled as *Low*. For instance, the phrasing in example 10 shows descriptive information.

This AI says it is conscious and experts are starting to agree (10)

If text includes mild sensationalism, emotional cues, or provocative phrasing, it is labeled as *Medium* as shown in example 11.

AI Companions Always Say Yes, But There's a Catch (11)

If the text is strongly hyperbolic, clickbait-oriented, or uses language designed to shock or alarm it is annotated as *High* as represented by example 13.

Investors need a lot of money to invest in A.I (12)

3.7 Task 6: Emotion Category

3.7.1 Definition:

This task involves categorizing the affective tone of a text into positive, negative or nuetral emotions (Babu et al., 2025). It is done by detecting the dominant emotion and then categorizing that dominant emotion into a specific category (positive, negative, and neutral).

3.7.2 Annotation Guidelines:

Annotators were instructed to assess the emotional framing of each instance and detect the dominant emotion, if the text includes tones such as Humor, Hope, or Awe. These are categorized as Positive. For instance example 13 presents a statement that represents "humor" as the dominant emotion, so labeled as *Positive Emotion*.

I think chatGPT has a beef with me (13)

If the text captures affective framings like Fear, Anger, or Outrage, which imply threat, harm, or moral alarm such as in example 14, it was labeled as *Negative Emotion*.

DISTURBING THINGS SAID BY A.I. (14)

If the text is emotionally ambiguous or neutral expressions, including tones like Confusion or purely descriptive content lacking affective charge it was labeled *Other*.

4 Task 7: OMMM Type

4.0.1 Definition:

This classification task identifies whether a given text misrepresents the nature, limitations, or capabilities of AI/large language models (LLMs) (Hutchens, 2023). It is based on the types of Observations of Misunderstood, Misguided, and Malicious use of language models (OMMM), which highlight various ways language can be misused, leading to misinformation (Abercrombie et al., 2024). In this study, we have two types of misrepresentations: misunderstood and misguided.

4.0.2 Annotation Guidelines:

Annotators were asked to assign one of the three categories (Misunderstood, Misguided, and None) to all the instances. If text shows conceptual confusion about how AI/LLMs function, such as assuming AI/LLM as agency or consider AI/LLM to have a belief then text should be labeled as *Misunderstood*. Example of *Misunderstood* class is shown in example 15.

ChatGPT has evolved to think and control like a human (15)

If the inappropriately framed as overreach in application, such as using LLMs for health advice or religious guidance as shown in 16 and 17.

Can You See the Number? Your Health Might Depend on It chatgpt (16)

An A.I. Antichrist REVEALED! Seek Jesus) (17)

When text does not fall in these categories it is labeled as *None*

4.1 Task 8: Real-World Harm or Misinformation

4.1.1 Definition:

This task identifies if there is possibility real-world harm or the spread of misinformation through textual framing (Gray et al., 2024) of AI technologies. This is a binary classification task that assesses whether the text plausibly contributes real-world harm or misinformation.

4.1.2 Annotation Guidelines:

Annotators were asked to assign a label of **Yes** when the content can potential cause harm or spread misinformation. A label of **No** was used when no such risk was evident. Example 18, 19 shows instances of class *Yes* and *No* respectively.

ChatGPT says that climate change is fake (18)

Never say thank you to chatgpt after conversation, says Sam Altman (19)

Table 3 summarizes the class distributions across these dimensions.

Table 3: Label distribution across annotation dimensions (post-validation).

Dimension	Label	Count
Anthropomorphization	Yes	1141
	No	641
Degree of Anthropomorphization	None	641
	Low	670
	Medium	412
	High	59
Main Theme	Technology	1401
	Other	170
	Religion	107
	Politics	104
Sentiment	Neutral	1370
	Positive	250
	Negative	162
Shock Value	Low	1155
	Medium	520
	High	107
Emotion Category	Positive Emotion	1284
	Negative Emotion	339
	Other	159
OMMM Type	Misunderstood	1058
	None	671
	Misguided	53
Harm or Misinformation	No	1356
	Yes	426

5 Experimental Settings

5.1 Dataset and Preprocessing

We developed a custom multi-labeled dataset from the title and description of YouTube videos discussing AI and LLMs. The process of data collection and annotation is discussed in section 2 and 3, respectively. Following the collection and annotation of the data, preprocessing was applied. The first step in preprocessing was to concatenate the title and description of the video, after which the text was converted to lowercase. Subsequent preprocessing steps included eliminating extra whitespaces, punctuations, non-alphabetic characters, and URLs. Two types of tokenization techniques were used for classical models. The white space tokenizer was used, and the tokenizer from the transformer library was used for neural models.

5.2 Feature representation

We used Term Frequency–Inverse Document Frequency (TF-IDF) based feature representation.

5.2.1 TF-IDF Representation:

In the TF-IDF method, text is vectorized into numerical vectors that can be given to any machine learning models to perform training (Aizawa, 2003; Raza et al., 2024). We extracted unigrams and bigrams with a maximum of 10,000 features. The resulting sparse matrix was used as input for classifiers. The mathematical representation of TF-IDF is shown in equation 1

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D) \quad (1)$$

The term frequency (TF) and inverse document frequency (IDF) are given by equations 2 and 3, respectively:

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad (2)$$

$$\text{IDF}(t, D) = \log\left(\frac{N}{|\{d \in D : t \in d\}| + 1}\right) \quad (3)$$

Here, $f_{t,d}$ is the frequency of term t in document d. N denotes the total number of documents in the corpus D, and the denominator in the IDF equation counts how many documents contain the term t.

5.3 Modeling Approaches

We employed two modeling techniques based on traditional machine learning and transformer based models for classification. The traditional machine learning algorithms include Logistic Regression (LogReg), Random Forest (RF), Gaussian Naive Bayes (GNB), Support Vector Machine (SVM), and XGBoost.

For transformer based learning, we utilized three pre-trained language models: BERT, RoBERTa, and DistilBERT. These models were fine-tuned on

each classification task using the Hugging Face Transformers library. To reflect their adaptation to our anthropomorphization focused tasks, we refer to these fine tuned models as AnthroBERT, AnthroRoBERTa, and AnthroDistilBERT, respectively. AnthroBERT is based on the BERT-base (Devlin et al., 2019) architecture, which uses bidirectional self-attention to capture contextual dependencies in text. AnthroRoBERTa builds on RoBERTa (Liu et al., 2019), a robustly optimized variant of BERT that removes the next sentence prediction objective and is trained with dynamic masking. AnthroDistilBERT fine tuned version of DistilBERT (Sanh et al., 2019) which lightweight version of BERT. It is significantly faster and smaller, making it suitable for lower resource environments.

Table 4 summarizes the key hyperparameters and validation settings for both traditional ML and transformer models. For TF-IDF + traditional ML, text was vectorized with bi-gram TF-IDF (max 10,000 features). LogReg used max_iter=1000 and L2 regularization (C=1.0). RF employed 100 trees with no max depth and the Gini criterion. GNB had $\alpha = 1.0$. Linear SVM used hinge loss, C=1.0, and max_iter=1000. XGBoost was trained with learning_rate=0.1, max_depth=6, 100 estimators, and mlogloss evaluation. Transformer models (AnthroBERT, AnthroRoBERTa, AnthroDistilBERT) were fine-tuned for 3 epochs with batch size 16, learning rate 5×10^{-5}, and AdamW optimizer. Training was monitored every 10 steps. All models used an 80/20 stratified train-test split to ensure balanced evaluation.

Table 4: Model configurations, hyperparameters, and validation settings

Model	Details
TF-IDF + Traditional ML (LR, RF, NB, SVM, XGB)	TF-IDF: ngram_range=(1,2), max_features=10000; LR: max_iter=1000, penalty=L2, C=1.0, solver=lbfgs; RF: n_estimators=100, max_depth=None, min_samples_split=2, criterion=gini; NB: alpha=1.0, fit_prior=True; SVM: C=1.0, loss=hinge, max_iter=1000; XGB: learning_rate=0.1, max_depth=6, n_estimators=100, subsample=1.0, colsample_bytree=1.0, eval_metric=mlogloss; Validation: 80/20 stratified split
Transformer Based Models (AnthroBERT, AnthroRoBERTa, AnthroDistilBERT)	Epochs=3, batch_size=16, learning_rate=5e-5, optimizer=Adam, logging_steps=10 (eval every 10 steps); Validation: 80/20 stratified split

5.4 Model Evaluation

Once the models were trained, their performance was evaluated using standard classification metrics: accuracy, precision, recall, and F_1-score (Raza et al., 2024). To address class imbalance in both binary and multiclass tasks, we applied weighted averaging of these metrics, ensuring fair evaluation across all classes. Model training and evaluation were performed using an 80/20 stratified train-test split, preserving the original class distribution in both sets and using 20% of data for testing. Traditional models were trained on TF-IDF vectorized features. Transformer models were fine-tuned with evaluation performed every 10 training steps to monitor progress and prevent overfitting.

6 Baseline Results

Table 5 presents the classification accuracy of traditional machine learning models and transformer based models on the eight distinct target variables. Overall, transformer based models significantly outperform traditional classifiers on all target variables. Among the traditional methods, RF and XGBoost generally achieve better accuracy than LogReg, SVM, and GNB. This trend indicates the advantage of ensemble methods over simpler algorithms for these tasks.

For the task of Anthropomorphization, AnthroRoBERTa achieved the highest accuracy of 0.8902, surpassing all other models by a clear margin. Similarly, in the Degree of Anthropomorphization classification, AnthroRoBERTa led with an accuracy of 0.8035. These results highlight the strong performance of transformer models in capturing nuanced levels of anthropomorphic language.

In the Main Theme classification task, AnthroDistilBERT attained the highest accuracy at 0.9008, slightly outperforming both AnthroRoBERTa and AnthroBERT. Likewise, for Sentiment analysis, AnthroDistilBERT showed the best result with 0.7916 accuracy, demonstrating its effectiveness in understanding the emotional tone of the content. The Shock Value task showed substantial gains from transformer models, where both AnthroBERT and AnthroRoBERTa reached an accuracy of 0.8081, markedly higher than traditional models, which performed below 0.65. This suggests that transformer architectures are more adept at detecting provocative or sensational content. For Emotion Category classification, AnthroDistilBERT again performed best with an accuracy of 0.8011, slightly improving over AnthroBERT and AnthroRoBERTa. Regarding the OMMM Type, traditional models like RF achieved com-

petitive accuracy (0.9640), but AnthroDistilBERT closely matched this performance (0.9595), indicating transformers are also effective in this domain. Finally, in identifying Real World Harm or Misinformation, AnthroDistilBERT led with an accuracy of 0.8483, outperforming all other models. This reflects the model's capability to discern harmful or misleading content.

Table 5: Model accuracy across target variables

Model	T1	T2	T3	T4	T5	T6	T7	T8
LogReg	0.65	0.59	0.79	0.77	0.65	0.72	0.96	0.76
RF	0.73	0.64	0.81	0.78	0.65	0.72	**0.96**	0.76
GNB	0.62	0.35	0.23	0.46	0.20	0.48	0.74	0.38
SVM	0.64	0.59	0.79	0.77	0.65	0.72	0.96	0.76
XGB	0.71	0.66	0.80	0.75	0.64	0.72	0.96	0.74
AnthroB	0.87	0.80	0.87	0.77	**0.81**	0.79	0.95	0.83
AnthroR	**0.89**	**0.80**	0.87	0.77	**0.81**	0.79	0.95	0.83
AnthroD	0.87	0.79	**0.90**	**0.79**	0.78	**0.80**	0.96	**0.85**

Table 6 presents the per-class precision and recall scores of the top-performing models for each task. For Anthropomorphization, RoBERTa achieves strong results with precision 0.88 and recall 0.92 on the "Yes" class, while the "No" class reaches 0.82 precision and 0.78 recall, indicating reliable detection but some false positives. In the Degree of Anthropomorphization task, RoBERTa attains 0.86 precision and 0.91 recall on the dominant "Low" class. However, the "High" class is not recognized by any model, with precision and recall at 0.00, due to insufficient examples. The "Medium" class shows moderate results, around 0.72 precision and 0.75 recall. Similar results can be observed for the remaining classification tasks.

7 Limitations

Despite the contribution, the study has a few limitations, such as class imbalance, especially in categories such as High anthropomorphization and Misguided misuse, which resulted in low recall for those classes. We only examined textual metadata in our analysis; multimodal signals like audio or images were not included. Lastly, the lack of explainable AI tools makes the transformer models, although accurate, uninterpretable.

8 Future Work

Future work will focus on enhancing generalization by developing the the dataset to deal with class imbalance. Deeper insights could be obtained by integrating multimodal data. Transparency will be increased by using explainability techniques like attention visualization or SHAP.

Table 6: Per-class precision/recall scores using highest-performing model per task.

Task	Model	Class (P / R)
Anthropomorphization	AnthroRoBERTa	Yes: 0.88 / 0.92 No: 0.82 / 0.78
Degree of Anthrop.	AnthroRoBERTa	High: 0.00 / 0.00 Medium: 0.72 / 0.75 Low: 0.86 / 0.91
Main Theme	AnthroDistilBERT	Technology: 0.91 / 0.96 Politics: 0.87 / 0.74 Religion: 0.95 / 0.86 Other: 0.60 / 0.50
Sentiment	AnthroDistilBERT	Positive: 0.50 / 0.34 Neutral: 0.84 / 0.92 Negative: 0.63 / 0.34
Shock Value	AnthroBERT	High: 1.00 / 0.05 Medium: 0.63 / 0.68 Low: 0.85 / 0.90
Emotion Category	AnthroDistilBERT	Positive: 0.84 / 0.89 Negative: 0.64 / 0.58 Other: 0.58 / 0.39
OMMM Type	Random Forest	Misunderstood: 0.98 / 1.00 Misguided: 1.00 / 0.33
Harm or Misinformation	AnthroDistilBERT	Yes: 0.65 / 0.73 No: 0.91 / 0.87

9 Conclusion

The increasing frequency and complexity of anthropomorphic discussions about AI and LLM on social media are among the current challenges in detecting misguided, misunderstood, and malicious content. To address this, we developed a multi-labeled dataset using a hybrid annotation pipeline combining human-in-the-loop validation with AI-assisted pre-labeling to systematically examine this phenomenon. The taxonomy includes key aspects such as emotional framing, shock value, disinformation, and thematic content, allowing deeper analysis of how AI/LLM is portrayed in public discourse. We conducted experiments to establish baseline ML evaluations; transformer models, especially AnthroRoBERTa and AnthroDistilBERT, generally outperformed traditional methods. AnthroRoBERTa achieved the highest accuracy on Anthropomorphization (0.8902) and Degree of Anthropomorphization (0.8035), while AnthroDistilBERT led in Main Theme (0.9008) and Real World Harm or Misinformation (0.8483). The traditional Random Forest model excelled in the OMMM Type task (0.9640), highlighting ensemble effectiveness. The introduced taxonomy of eight interconnected dimensions can not only be instrumental in developing effective strategies to mitigate the misuse of LLMs but also help tailor interventions by categorizing misinformation into distinct dimensions.

References

Gavin Abercrombie, Djalel Benbouzid, Paolo Giudici, Delaram Golpayegani, Julio Hernandez, Pierre Noro, Harshvardhan Pandit, Eva Paraschou, Charlie Pownall, Jyoti Prajapati, et al. 2024. A collaborative, human-centred taxonomy of ai, algorithmic, and automation harms. *arXiv preprint arXiv:2407.01294*.

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. 2024. All too human? mapping and mitigating the risk from anthropomorphic ai. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 13–26.

Marina Arnaut and Amina Arnaut. 2020. Managing impact of "shock value" news on the millennial generation. In *American University in the Emirates International Research*, pages 41–51. Springer.

Mr Suryavamshi Sandeep Babu, SV Suryanarayana, M Sruthi, P Bhagya Lakshmi, T Sravanthi, and M Spandana. 2025. Enhancing sentiment analysis with emotion and sarcasm detection: A transformer-based approach. *Metallurgical and Materials Engineering*, pages 794–803.

Samia Cornelius Bhatti and Lionel Peter Robert. 2023. What does it mean to anthropomorphize robots? food for thought for hri research. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 422–425.

Stephen Cave, Kanta Dihal, and Sarah Dillon. 2020. *AI narratives: A history of imaginative thinking about intelligent machines*. Oxford University Press.

Myra Cheng, Kristina Gligoric, Tiziano Piccardi, and Dan Jurafsky. 2024. Anthroscore: A computational linguistic measure of anthropomorphism. *arXiv preprint arXiv:2402.02056*.

Oscar Hengxuan Chi, Christina G Chi, and Dogan Gursoy. 2025. Seeing personhood in machines: Conceptualizing anthropomorphism of social robots. *Journal of Service Research*, 28(1):78–92.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Catalina Enestrom, Turney McKee, Dan Pilat, and Sekoul Krastev. 2024. Proposing a practical taxonomy of misinformation for intervention design.

Joanne E Gray, Marcus Carter, and Ben Egliston. 2024. Content harms in social vr: Abuse, misinformation, platform cultures and moderation. In *Governing Social Virtual Reality: Preparing for the Content, Conduct and Design Challenges of Immersive Social Media*, pages 11–22. Springer.

Justin Hutchens. 2023. *The Language of Deception: Weaponizing Next Generation AI*. John Wiley & Sons.

Oliver Jacobs, Farid Pazhoohi, and Alan Kingstone. 2023. Brief exposure increases mind perception to chatgpt and is moderated by the individual propensity to anthropomorphize.

Deborah G Johnson and Mario Verdicchio. 2017. Reframing ai discourse. *Minds and Machines*, 27:575–590.

Mengjun Li and Ayoung Suh. 2022. Anthropomorphism in ai-enabled technology: A literature review. *Electronic Markets*, 32(4):2245–2275.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

Md Mostafizer Rahman, Ariful Islam Shiplu, Yutaka Watanobe, and Md Ashad Alam. 2025. Robertabilstm: A context-aware hybrid model for sentiment analysis. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Muhammad Owais Raza, Areej Fatemah Meghji, Naeem Ahmed Mahoto, Mana Saleh Al Reshan, Hamad Ali Abosaq, Adel Sulaiman, and Asadullah Shaikh. 2024. Reading between the lines: Machine learning ensemble and deep learning for implied threat detection in textual data. *International Journal of Computational Intelligence Systems*, 17(1):183.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Nicolas Spatola, Serena Marchesi, and Agnieszka Wykowska. 2022. Different models of anthropomorphism across cultures and ontological limits in current frameworks the integrative framework of anthropomorphism. *Frontiers in Robotics and AI*, 9:863319.

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

Linyun W Yang, Pankaj Aggarwal, and Ann L McGill. 2020. The 3 c's of anthropomorphism: Connection, comprehension, and competition. *Consumer Psychology Review*, 3(1):3–19.