A Linguistically Grounded Evaluation of Anthropomorphic Language Detection in AI Research

MSc Thesis *(Afstudeerscriptie)*

written by

Dorielle Lonke

under the supervision of **Dr. Jelke Bloem** and **Dr. Pia Sommerauer**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
September 18, 2025	Dr. Jelke Bloem
	Dr. Giorgio Sbardolini
	Dr. Pia Sommerauer
	Dr. Petter Törnberg (Chair)

Institute for Logic, Language and Computation

Abstract

This thesis presents a conceptual framework of anthropomorphic language in AI research that serves as a theoretical baseline for evaluating existing approaches for its detection. Drawing on existing work, a taxonomy of human-like attributes frequently ascribed to AI systems is outlined. Relying on linguistic theory of grammatical animacy, as well as thematic proto-role theory and frame semantics, a linguistic model is constructed, representing the various ways in which anthropomorphic descriptions are expressed in natural language. The linguistic model serves as a baseline for the construction of a challenge set that is used towards the evaluation of the state-of-the-art approaches to anthropomorphic language detection. The evaluation shows that the current unsupervised approaches are not suitable for all types of linguistic structures, suggesting the need for alternative methods. Furthermore, due to the increasingly anthropomorphic language employed in body of work that pertains to AI technologies, the question arises whether unsupervised approaches relying on large language models trained on this type of data are a reliable method for detecting this phenomenon.

Acknowledgements

First and foremost, I would like to thank my supervisors Jelke Bloem and Pia Sommerauer, for taking interest in my idea from the start and paving a clear way forward. Pia, I am particularly grateful for your initial enthusiasm and involvement — it meant a lot that you expressed your support and chose to be involved despite the institutional divide. Thanks to the dedication of my supervisors, I had the unique opportunity to publish part of this work before finalizing my thesis. I am grateful to both of you for your guidance during the submission process, and particularly to Jelke for your dedicated efforts in refining the manuscript, and for your encouragement and support during the presentation.
I would also like to acknowledge the committee and reviewers of the First Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models (OMMM 2025); their thoughtful feedback helped me fortify my arguments and ground them in related work and ideas.
To my thesis committee members Petter Törnberg and Giorgio Sbardolini, I am thankful for the seamless organization and the deep, insightful conversation during my defense.
To my MoL colleagues Mayra and Morgane, I am grateful for the many conversations we shared during our time at the ILLC, which have undoubtedly helped shape my thoughts and ideas surrounding this thesis. I am lucky to have had the companionship of talented and smart women in my field.
Finally, to my partner Tomer — thank you for proofreading and polishing this work with your incredibly keen writer's eye, and for the countless cups of tea and endless support — you are the best[1].

[1]This footnote is in your honor. Additionally, did I use the em-dash correctly?

Contents

Chapter 1

Introduction

With the evolving popularity and applications of AI systems, the terms used to describe their functionalities have become increasingly anthropomorphizing (Floridi & Nobre, 2024). The tendency to attribute human-like capabilities and properties to AI systems has been discussed in various contexts, including psychological analyses (Epley et al., 2007; Hofstadter, 1995; Waytz et al., 2010), linguistic approaches (Abercrombie et al., 2023; DeVrio et al., 2025; Emnett et al., 2024), and ethical considerations (Placani, 2024; Watson, 2019). Anthropomorphic descriptions of AI are said to directly contribute to AI hype (Barrow, 2024), which is problematic both in terms of overly exaggerated descriptions of its capabilities, as well as needless and undue panic about AI domination (Salles et al., 2020). Although widely discussed, the task of defining exactly what anthropomorphism entails is highly contextual (Cheng et al., 2024), ambiguous, and subjective (Shardlow et al., 2025; Waytz et al., 2010).

This thesis aims to present a conceptual framework of anthropomorphic language in the context of AI, and investigate the state-of-the-art approaches to anthropomorphic language detection. The objective is to carry out a rigorous analysis of this phenomenon as well as the existing tools for its detection. Current work on anthropomorphism from the linguistic perspective tends to focus on conversational design and linguistic cues in AI outputs. Anthropomorphism in descriptions of AI systems has only recently begun to receive attention (Inie et al. 2024 ;Ryazanov et al. 2024). This work attempts to bridge this gap, by providing a linguistically grounded analysis of anthropomorphism in AI descriptions. This analysis will serve as a baseline for constructing a challenge set which will be used towards the evaluation of the existing methods for detecting anthropomorphic language. In addition to the conceptual framework, the challenge set represents the initial groundwork towards a benchmark of anthropomorphic language in AI.

Chapter 2 includes a theoretical background of anthropomorphic language. First, the questions *what is anthropomorphism?* and *why do we anthropomorphize?* are answered from two key perspectives: a psychological and a linguistic perspective. This background contextualizes notion of anthropomorphism in AI in current literature, and provides the foundations for establishing a working definition of anthropomorphism towards a linguistic model. Second, it presents the motivation and objective of this thesis, highlighting the negative conceptual and ethical implications of anthropomorphic language. The research questions are presented in section 2.2.1. The chapter concludes with an overview of related work, focusing on similar approaches and current methods for detecting anthropomorphic language.

Chapter 3 introduces the linguistic model, which consists of a taxonomy of human-like attributes commonly found in anthropomorphic descriptions, as well as the underlying linguistic structure of these descriptions. The taxonomy presented in section 3.2 is based on the guiding lenses of DeVrio et al.

(2025), and serves to identify potentially anthropomorphic components in text. The syntactic analysis presented in section 3.3 is established in linguistic theory of grammatical animacy, supplemented by thematic proto-role theory (Dowty, 1991) and frame semantics (Fillmore, 1982).

Chapter 4 presents AnthroSet, an evaluation dataset consisting of 600 manually annotated utterances containing anthropomorphic language in AI research. The evaluation set features a variety of linguistic structures based on the linguistic model, taken from abstracts from published papers on AI, machine learning, and natural language processing. Section 4.1 outlines the data retrieval and parsing process, describing how the various syntactic structures were obtained. Section 4.2 details the annotation procedure, including the detailed guidelines and a study of inter-annotator agreement on a sample of sentences. The purpose of this evaluation set is twofold: First, it is designed as a challenge set for evaluating existing approaches to anthropomorphic language detection. In particular, it seeks to assess their capabilities in recognizing diverse syntactic patterns. Second, it constitutes preliminary work towards a benchmark for anthropomorphic language detection. Recently, the first open annotated corpus of anthropomorphic language was published (Shardlow & Przybyła, 2024), setting forth an initial benchmark for evaluation. The work presented in this thesis aims to supplement the existing work, by introducing a novel, syntax-based classification that presents different challenges to existing approaches, and paves the path towards the development of new detection methods.

Chapter 5 describes the details of the evaluation task. There are presently two existing approaches for detecting the attribution of human-like characteristics to non-human entities, particularly machines and technological artifacts. Both approaches provide an unsupervised pipeline for detecting patterns of humanizing descriptions in text, relying on a masked language model. These two approaches are evaluated and compared on the challenge set presented in chapter 4. The results highlight the problems with employing a masked language model approach for this task. For one, the masking is consequential in achieving good results, but a uniform masking approach is not suitable for all syntactic structures. Furthermore, automated masking strategies are problematic, as they might mask important contextual information that is crucial for identification. However, the current alternative consists of manual masking, which is not a scalable solution. Future implementations that rely on masked language models would require an additional step in properly recognizing and masking the entities, without obscuring the anthropomorphic context. Finally, as AI-related terminology becomes increasingly anthropomorphizing, masked language models are more likely to associate AI entities with anthropomorphic verbs and descriptors, simply due to their reliance on statistical co-occurrence (Zhang et al., 2024), posing further challenges for anthropomorphism detection.

Chapter 6 concludes this thesis, and provides recommendations for mitigating anthropomorphic language and possible directions for future work. Future work includes expanding the evaluation set presented in this thesis, including multiple annotators and a robust inter-annotator agreement study to fortify its reliability as a benchmark. Additionally, it is worth exploring alternative methods for anthropomorphic language detection. Shardlow et al. (2025) present a classifier trained on the annotated corpus of that outperforms the unsupervised baseline on unseen abstracts. The linguistic framework presented in this thesis could be used to identify relevant features for improving such a classifier. The recommendations include establishing sound writing practices, especially for researchers and practitioners conducting AI research. The Academic community has an authoritative stance, influencing the general perception and popular opinion. We should therefore be careful and deliberate in the language we use when conducting and reporting on research.

Chapter 2

Background

This chapter presents a preliminary investigation of anthropomorphism in the context of AI. It begins by answering the question of *what is* anthropomorphism, through two key perspectives by which anthropomorphism in AI discourse is predominantly discussed — the psychological perspective, and the linguistic perspective. It then proceeds to answer the question of *why* we anthropomorphize, taking into account psychological, methodological, and rhetorical explanations. Once a theoretical background is established, the motivation and objective of this thesis are outlined, focusing on conceptual and ethical implications of anthropomorphic language. The research questions of this thesis are presented, alongside a description of the contributions of this thesis. The chapter concludes with related work, including similar approaches to and implementations of anthropomorphic language detection.

2.1 Theoretical Background

2.1.1 What is anthropomorphism?

There is no single attribute that makes a system anthropomorphic. Similarly, there is no particular threshold that determines whether a system is anthropomorphic or not. Rather, the tendency to anthropomorphize AI technologies exists on a spectrum and is the result of a plethora of factors and characteristics of those technologies (Abercrombie et al., 2023). The perception of anthropomorphism varies across individuals (Waytz et al., 2010), and is not necessarily mindful — that is, individuals are not always aware of this perception (Kim & Sundar, 2012). It can be understood either as being attributed actively, as a design choice, or passively, as a perception of the user. In terms of its linguistic representation, some words or phrasings have a stronger effect on anthropomorphism perception. While certain structures may be semantically equivalent, their degree of anthropomorphism is not the same (Ryazanov & Björklund, 2024). Furthermore, anthropomorphism is not always conceptualized as a tendency. Existing literature contains definitions from varying perspectives leading to the conceptualization of anthropomorphism as a tendency, perception, process, inference, or technological feature (Li & Suh, 2022). Below is a theoretical overview of anthropomorphism, from the point of view of two key perspectives: the psychological perspective and the linguistic perspective. While this thesis is primarily primarily focused on the linguistic aspects, the cognitive and psychological aspects are closely related to and inform linguistic processes.

Anthropomorphism from a psychological perspective

Waytz et al. (2010) discuss anthropomorphism as a psychological phenomenon. They show that while anthropomorphism is understood as a universal phenomenon shared across all humans, there exist individual differences in the way humans perceive anthropomorphism. These differences predict three important consequences related to the attribution of a human mind to non-human objects, which shape the way human-computer interactions are understood, researched and capitalized. First, having a mind entails conscious experience, leading perceivers to deem the perceived objects as being worthy of moral care and concern. Second, having a mind entails having intention, and autonomy, which in turn leads to attribution of responsibility and trust. Third, agents with human-like minds also have the ability to perceive, and are capable of evaluating and judging — introducing a reciprocal dimension to human-computer interactions. As a result, anthropomorphized agents can serve as a source of social influence. These psychological aspects of anthropomorphism have non-psychological implications on human-computer interactions. For instance, AI developers can use these results to optimize their technology and design systems in a way that can either benefit or exploit the people most prone to these consequences.

Maeda and Quan-Haase (2024) raise similar issues with the anthropomorphic design of chatbots, focusing on the parasocial nature of these seemingly reciprocal interactions. The illusion of engagement leads to unwarranted trust and social affordances that position the outputs of chatbots as definitive answers, instead of what they really are — predictive inferences. Kahn et al. (2007) offer nine psychological benchmarks for measuring success in building human-like robot, based on conceptually fundamental aspects of human interaction. These benchmarks are autonomy, imitation, intrinsic moral value, moral accountability, the ability to encroach on or breach privacy, reciprocity, conventionality, creativity and authenticity of relation. Alongside these benchmarks, Kahn et al. present hypothetical human-robot interactions, and discuss how these benchmarks are tested in light of potential human reactions. From a marketing standpoint, Alabed et al. (2022) draw a conceptual link between anthropomorphism in AI agents and the user's self-congruence, which is the alignment between the user's sense of self and the personality presented by the product. Anthropomorphic design can contribute to the perceived similarity and establish a psychological relationship between a user and an AI-based product.

Anthropomorphism from a linguistic perspective

Another topic of interest in anthropomorphism and AI is the way human characteristics are represented in the output of language models. Abercrombie et al. (2023) discuss the linguistic factors that contribute to anthropomorphism in conversational AI systems, focusing on voice, content, register, and style, and roles assigned to these systems by designers and users. In particular, they analyze how AI-generated language, either spoken or textual, can convey human-like impressions through human idiosyncrasies such as tone, accent, and disfluencies, or the use of phatic expressions, personas, or the first-person pronoun. Other elements of conversational design such as expression of thought, reasoning process, assuming responsibility, or displays of empathy, confidence, or doubt contribute to a portrayal of AI dialogue systems as having cognition and mental states. The authors frame these factors in recommendations of what to avoid when designing dialogue systems, and warn against the harmful consequences of anthropomorphism. In a similar vein, Emnett et al. (2024) identify six conversational factors that give rise to the anthropomorphism of AI agents — personalization, assertiveness, direct

speech, politeness, proportionality, and humor. The authors demonstrate how these factors shape the perception of conversational AI systems as social agents, contributing to their perceived trustworthiness, likeability and competence. The factors involved in human conversation can simultaneously inform best practices for conversational design when the objective is to create more human-like agents, and shed light on the problem aspects of anthropomorphic language technologies. In an effort to present a practical framework for optimizing encounters between humans and computers in commercial and customer service settings, Van Pinxteren et al. (2020) provide a taxonomy of human-like communicative behaviors used by conversational agents. Addressing modes of verbal communication, they list behaviors such as cognitive recall, communication style, personality, affect support, social praise, politeness, humor, and small talk. From a point of view of identifying and mitigating the risks that stem from anthropomorphism in AI, DeVrio et al. (2025) present their own taxonomy of linguistic expressions in textual outputs of language technologies that contribute to anthropomorphism. They specify nineteen types of expressions ranging from intelligence and awareness to morality, embodiment, and right to privacy. These are classified and viewed through five guiding lenses, categorized as *internal states*, *social positioning*, *materiality*, *autonomy*, and *communication skills*. While they also examine stylistic choices of text design, their analysis is not primarily concerned with communicative behaviors, and instead focuses on broader categories pertaining to the cognitive and mental aspects of anthropomorphism. Although they limit their discussion to AI-generated output, these guiding lenses serve as a basis for the taxonomy of anthropomorphism in human-written text about AI, presented in section 3.2. These expressions exemplify the myriad ways in which the design of synthetic text can give a semblance of cognition or humanity to the machines that generate it.

Most of the research on anthropomorphic language in AI is concerned with AI-generated output. There is not much empirical work on anthropomorphic descriptions of AI. Inie et al. (2024) refer to these two aspects of anthropomorphic language as *anthropomorphization by design*, which is the anthropomorphizing language built into these systems and manifested as AI output, and *anthropomorphization by description* which is the language we use to describe AI systems[1]. Their work shows that there is no statistical evidence that supports the notion that anthropomorphizing descriptions of AI systems (referred to as probabilistic automation systems) automatically lead to increased trust. Ryazanov et al. (2024) also present an automatically annotated dataset consisting of news and media coverage of AI, using the FrameNet-based automatic annotation system LOME (Xia et al., 2021) to explore trends of anthropomorphism in recent AI discourse. Recently, the first manually annotated corpus of anthropomorphic descriptions of AI was released, covering sentences extracted from papers published in the Association for Computational Linguistics (ACL) anthology (Bird et al., 2008), as well as news items from BBC News, the New York Times and the Register (Shardlow et al., 2025).

2.1.2 Why do we anthropomorphize?

The tendency to attribute human-like capacities to AI systems has been observed since the foundation of AI as a field of research. The general relation between cognition and machines has been widely discussed, with authors such as Searle (1980) and Dreyfus (1976, 1992) arguing against the reduction of human thought and embodied experience to syntactic and symbolic programs. Nevertheless, it is a persistent

[1]Inie et al. (2024) use the term 'anthropomorphization' to refer to the intentional act of using anthropomorphic language to design or describe an AI product, as opposed to 'anthropomorphism' which is the internal process of perceiving an AI product as human-like.

human tendency to associate language with humanity, and a natural result is that conversational AI agents are often perceived as possessing some sort of understanding. This phenomenon was identified as the *ELIZA* effect, which is the cognitive bias that causes human users to attribute human-like properties — such as intelligence and emotions — to responsive machines (Hofstadter, 1995). Non-users, such as AI practitioners and researchers, also succumb to this tendency. McDermott (1976) coined the term *wishful mnemonics*, which is not a psychological but rather a methodological tendency to name and describe AI programs not in terms of what they actually do, but as what they are intended and willed by us to do. In addition to the psychological and methodological aspects, anthropomorphic language is more accessible from a communicative point of view. Metaphors are considered a tool for explanation and persuasion (Rossi & Macagno, 2021). Explaining how AI systems operate in familiar terms facilitates understanding, as this type of rhetorical device is generally easier to comprehend. Below are some explanations as to why anthropomorphic descriptions are so prevalent and pervasive in AI discourse.

The *ELIZA* effect

ELIZA is the name of a chatbot created by Joseph Weizenbaum at MIT between 1964 and 1967, designed to imitate a conversation with a therapist (Weizenbaum, 1966). It was constructed as a simple rule-based program that outputted replies based on predefined scripts, keywords identified in the input provided by the users, and a set of minimal grammatical rules. Despite the formulaic output, users reported feeling true connection to the chatbot. This was described by Weizenbaum as the *ELIZA* effect, whereby human users attribute intelligence and emotions to machines that provide convincing human responses. It is understood as a cognitive response, and reflects a natural human tendency to identify patterns and look for categorical similarity. This effect was popularized by Hofstadter (1995), who discussed this cognitive bias in the context of contemporary AI research. He argued that exaggerated descriptions of the capabilities of AI technologies are at worst highly anthropomorphizing, and at best vague about the true capabilities. This ambiguity will lead non-professionals to complete the information with their own interpretations and conclusions, which are likely to be shaped by the *ELIZA* effect.

Today, this effect can be exhibited in interactions with widely popular generative AI chatbots, primarily ChatGPT. Chu et al. (2025) conduct a wide-scale study of over 30k conversations between users and AI chatbots, and analyze them in terms of parasocial (i.e. one-sided) relationships, emotional attachment, and psychological risks. They find that users tend to cultivate emotional bonds akin to those of human relationships. The *ELIZA* effect is easily understood from both the psychological and linguistic perspectives of anthropomorphism, as the direct result of human-like conversational design of chatbot outputs. Evidently, these outputs do not need to be overly complex; even simple, scripts replies such as those of ELIZA are enough to give rise to this psychological phenomenon. We can identify a circular process occurring with anthropomorphism: anthropomorphic design leads to anthropomorphic perception.

Wishful Mnemonics

The anthropomorphization of AI technologies can be seen as a reflection of our own interests and desires — as developers, users, or humans in general. McDermott (1976) defines *wishful mnemonics* as the process whereby researchers design programs using terms such as *understand* or *goal*, as it is

easier to name them by their intended purpose. We see this tendency in the way we name functions in code: if a function is purported to parse a sentence into syntactic components, we might name it something along the lines of `parse_sentence`. In this scenario, we first define the function with this name, and then we proceed to implement it. In traditional programming applications, it makes sense to refer to a program by what it is intended to do. In the context of AI, McDermott argues that we are more often faced with problems rather than solutions. To that extent, first defining an 'understanding' program, before it has been implemented, is a form of begging the question. McDermott gives the example of GPS, the *General Problem Solver*, as a program named by what it was hoped to achieve[2]. His recommendation was to remain neutral and technical in selecting terminology, until the desired result is achieved.

The expressive power of metaphors

Anthropomorphic descriptions of AI can be seen as a type of metaphoric language that draws a parallel between machines and humans, specifically hinging on terminology related to the brain and related biological processes (Inie et al., 2024). Metaphors are a powerful explanatory tool that can promote understanding in many cases, especially in the case of artificial intelligence which is a complex and technical field of research. The increased popularity of AI technologies and their coverage in popular media led to their description in familiar, everyday terminology, intended for the general audience. From a standpoint of development and design, analogies are used in the stage of abstraction, as a means to conceptualize a problem that needs to be solved (Blackwell, 1996). Metaphors can be seen as a tool for designers that aids them in approaching new techniques and technologies. Murray-Rust et al. (2022) provide an overview of commonly used metaphors in the context of AI, and offer alternative analogies that can help shape the way AI technologies are presented to the public. For instance, they suggest metaphors of spaces and terrains, industrial components, or materials.

Carbonell et al. (2016) present an analysis of metaphors in AI descriptions that identifies a bidirectional process, which is very similar to the circular process described in the context of the *ELIZA* effect and conversational design. They describe a two-way process by which metaphors shape the evolution of technologies, which in turn change the rules of society and influence human perception. Through the example of artificial intelligence and the metaphor of the human brain, they describe the process wherein metaphors related to cognition and the human brain were employed to describe and make sense of emerging AI technologies. These AI technologies, defined by these metaphorical descriptions, shaped our perception of reality. That is, while these metaphors were initially required to conceptualize and understand new frameworks, we now understand the fundamental operation of these technologies in terms of these metaphoric descriptions, but in a literal, non-metaphoric manner. Subsequently, technical terms related to machines and computers have entered our lexicon to describe the human mind, which in itself poses a conceptual problem. Watson (2019) argues that anthropomorphic rhetoric in AI discourse is not ethically neutral: it entails both conceptual and ethical issues which are more harmful than beneficial for AI research. In the following section we discuss some of the conceptual and ethical issues with anthropomorphizing AI technologies.

[2]Today, we have the term AGI — Artificial General Intelligence, referring to the hypothetical AI technology that possesses or surpasses human cognition.

2.2 Objective and Motivation

The topic of anthropomorphic language in the context of AI is interesting on several accounts. From the onset, it involves the challenge of identifying the linguistic structure of anthropomorphic language, and calls to engage with existing literature to characterize the human-like properties and capacities that are commonly attributed to AI. Existing literature on anthropomorphism in AI has mainly focused on the linguistic aspects of AI output, providing insights from conversational design. Recently, several analyses of anthropomorphic descriptions of AI have been carried out (Cheng et al. 2024; Inie et al. 2024; Ryazanov and Björklund, 2024; Ryazanov et al. 2024; Shardlow et al. 2025). This thesis is therefore motivated to contribute to this research direction, by providing a thorough analysis of the various linguistic structures in which anthropomorphic language is represented. Additionally, the use of anthropomorphic language to describe AI gives rise to several significant conceptual and ethical issues. As discussed, anthropomorphizing AI systems might be considered a natural tendency or an essential part of design, but it also has some negative implications. Firstly, there are conceptual issues that arise from the incorrect attribution of human-like properties to AI systems, and incorrect representation of their abilities (Brooker et al., 2019; Floridi & Nobre, 2024). Perhaps more importantly, there are various ethical implications of anthropomorphism. For one, framing AI systems as humans positions them as moral agents, which leads to a host of unwarranted moral judgments about them. Further, anthropomorphism leads to AI hype, which is unwanted in both its pessimistic, fearmongering form, as well as its disproportionately optimistic form (Salles et al., 2020). Finally, describing AI systems as having human-like competencies leads to their deployment in sensitive contexts that have the potential to significantly affect human lives, and thus require human expertise and scrutiny. These conceptual and ethical concerns are the primary motivation to mitigate anthropomorphic language in AI research. To do that, we would need to investigate anthropomorphic language detection methods, and for that, a solid conceptual framework is essential. Below are objectives and significance of this thesis, followed by the conceptual and ethical considerations that motivate it.

2.2.1 Objective and significance

This thesis aims to provide a conceptual foundation of anthropomorphic language in the context of AI. Specifically, anthropomorphizing descriptions of AI which lead to the perception of AI systems as having human-like properties, such as cognition, mental states, intention, and volition. The aim is to identify them both in terms of the human-like characteristics that are attributed to AI in anthropomorphic descriptions, as well as the underlying linguistic patterns. Furthermore, it seeks to explore how these linguistic patterns are operationalized in existing approaches for anthropomorphic language detection. In particular, what are the challenges posed by various structures on the task of detecting anthropomorphism in text? are the state-of-the-art implementations of anthropomorphic language detection equipped to handle them? The objective is therefore to answer the following research questions:

RQ1. *What are the specific human-like attributes and capacities that are attributed to AI?*
RQ2. *What are the specific linguistic structures that characterize anthropomorphic language?*
RQ3. *Can the state-of-the-art methods for anthropomorphic language detection correctly identify varying linguistic patterns?*

RQ1 and RQ2 are answered in chapter 3, in which a taxonomy of anthropomorphic attributes is defined alongside a characterization of linguistic structures in which anthropomorphic descriptions are expressed. Chapter 4 describes the process of compiling and annotating an evaluation set based on the linguistic model. This evaluation set is used towards the evaluation of the state-of-the-art approaches for anthropomorphic language detection, which aims to answer RQ3.

The significance of this thesis is its contribution to the existing work that exists and is currently being done on anthropomorphism in AI. The first contribution is a solid conceptual framework, consisting of the first taxonomy for anthropomorphic descriptions, and the first linguistic model that analyzes anthropomorphic structures according to their underlying linguistic patterns. Additionally, it presents a carefully curated and manually annotated dataset of anthropomorphic language, based on real-world samples of anthropomorphic and non-anthropomorphic language. This dataset joins the existing corpus by Shardlow et al. (2025), providing an additional conceptual layer which includes a linguistic classification to categories, which has not been done to date. The third contribution is a quantitative evaluation of the state-of-the-art approaches, which compares two existing implementations in their ability to detect diverse examples of anthropomorphism in AI, highlighting their strengths and weaknesses. The conclusions of the evaluation fortify the result presented by Shardlow et al. (2025), which finds unsupervised methods unsuitable for this task.

2.2.2 Conceptual considerations

Brooker et al. (2019) describe the issue with anthropomorphic descriptions as the incorrect use of terminology that makes sense for certain contexts, in the content of philosophical discussions on AI. By examining three mundane contexts in which AI systems are interacted with, discussed, or analyzed, they show that terms borrowed from human cognition to describe the operations of these systems are clearly understood in context as what they are — mere descriptions. However, once transferred into the domain of philosophical inquiry, the result is incorrect conceptualization of AI's functionalities, leading in turn to the formulation of misleading questions and hypotheses about the very nature of AI. This is because of the frequent conflation between AI as an interface, service or tool, and AI as a concept. When talking about AI as a tool, focusing on a specific implementation of some system, anthropomorphic terms such as *training* or *learning* are understood in the context of the system's operation. However, when AI is discussed as an abstract concept, we commit the fallacy of thinking that these words mean that the AI system has the cognitive capacity to learn or be trained in the same way humans do. This important contextual distinction serves as the foundation for the annotation approach for the dataset presented in this thesis. Meaning is contextual, of course — but especially when it comes to AI-specific terminology, which borrows from human cognition and intelligence. Although we only evaluate per sentence, usually it is possible to gauge whether the sentence discusses an AI entity as a technology or tool, or whether it is more broadly referred to as a concept. On the basis of this distinction we decide whether the sentence is anthropomorphic with regard to a certain entity or not.

Floridi and Nobre (2024) posit that the conceptual borrowing from cognitive sciences to AI is not merely a metaphorical application of familiar terms but is potentially harmful for both AI and cognitive science research. For AI, the anthropomorphic terminology borrowed from cognitive sciences can misguide and derail scientists from exploring viable research directions, causing them to instead invest their time in developing 'artificial human intelligence', leading to recurring *AI winters*. For cognitive scientists, this semantic crosswiring results in a impoverished understanding of the human mind, in

which complex psychological and biological systems are explained in reductionist computational terms. This kind of conceptual borrowing is a methodological problem for research, as it relies on terminology that does not contribute to clarity, but rather obscures the mechanisms and processes from either side of the conceptual mapping. While changing the language is not feasible, they conclude that better understanding will help shape and contextualize the meaning of these anthropomorphic terms. Aside from the conceptual perils, incorrect conceptualization of AI leads to misaligned expectations in human-computer interactions. Mueller (2020) frames this as a cognitive analogue to the *Computers Are Social Actors* (CASA) paradigm (Nass et al., 1994), which shows that users interact with computers in a way that is fundamentally social. This tendency towards *cognitive anthropomorphism* is manifested in the expectation that AI will exhibit the same nature of intelligence as humans. Conceptual clarity is therefore needed.

2.2.3 Ethical considerations

Placani (2024) provides a dual interpretation of anthropomorphism in AI as a hype and fallacy. As a hype, the projection of human-like characteristics onto AI systems contributes to an exaggerated and inflated portrayal of their actual abilities. As a fallacy, anthropomorphizing AI systems gives rise to negative ethical implications, involving judgments of moral character and status, responsibility judgments, and judgments of trust. When AI systems are perceived to have mental states, they can be characterized in terms of good and evil, friendly or malicious, or as having empathy or loyalty. In the absence of true moral agency, this is problematic. Similarly, attributing qualities like consciousness or the ability to experience emotions to AI systems also leads to their evaluation as moral subjects, which are deserving of moral considerations. This fallacy also leads to incorrect attribution of responsibility, by which AI systems are seen as responsible for their 'actions', so to speak, and the consequences and possible harm caused by these actions. Again, in the absence of moral agency, this attribution of responsibility is misplaced; in cases of harm caused by AI, it is the owners, developers, or organizations behind the creation and deployment of these systems that ought to be held accountable. AI systems are also not capable of being trustworthy or untrustworthy — this trust, like responsibility, can be ascribed only to the moral agents (i.e. humans) who own, develop, and deploy these systems.

On the account presented by Salles et al. (2020), perceiving AI systems as human-like entails their consideration as moral agents, which leads to ascribing them undue normative impact. That is, AI systems become moral participants of society, and their actions and consequences are to be taken into moral consideration, not only in terms of their treatment, but also as part of collective decision-taking. Salles et al. mention two possible outcomes of overly anthropomorphized descriptions, on which we expand with our own interpretation. From a pessimistic point of view, these descriptions are the source of overblown and unwarranted fears of AI domination, in which AI systems will replace humans or attempt to subjugate them. These fears distract from the actual problems associated with AI systems, such as biased output and misinformation, or environmental impact, to name a few. From an overly optimistic point of view, when AI systems are regarded as human-like, their actual capabilities, intelligence, and utility are not measured or evaluated properly, leading to misconceptualization in design and research which ultimately hinders progress. From the perspective of everyday life, Watson (2019) addresses the ethical problem that arises from anthropomorphic rhetoric in AI research that leads to their deployment in contexts that require human expertise. For instance, credit scoring, criminal justice, or military operations are domains that have introduced AI technologies but involve

high-stakes decisions with the potential to significantly impact the lives of those affected. Decisions that should be made by human experts are outsourced to AI systems, because they are seen as having human-like or superhuman cognitive abilities. Finally, there is an ethical consideration of how research is conducted and carried out, for the sake of research itself. Anthropomorphizing descriptions of AI that are motivated by trends and not backed by any real science simply deteriorate the body of work that concerns these technologies.

2.3 Related Work

Ryazanov and Björklund (2024) propose a frame-semantics approach for detecting agency attribution. Relying on FrameNet (Ruppenhofer et al., 2006), they describe the process of constructing an annotated dataset of sentences containing a mention of AI with some level of agency assigned to it. The proposal demonstrates how frames with attributes such as *Cognizer* can be used for identifying whether agency is implied in certain semantic structures. While they focus on agency attributions, this task is closely related to anthropomorphism detection, as agency can be seen as one facet of anthropomorphism, and other facets such as cognition and awareness are present in the same frames that inform agency attribution. Ryazanov et al. (2024) expand on the idea of a frame-semantics approach, and conduct a wide-scale, mixed-methods narrative analysis using FrameNet annotations. Their objective is to investigate the effect that the release of ChatGPT had on media narratives on AI technologies. By identifying frames in media-based utterances, they demonstrate an increase in anthropomorphic descriptions framing AI as a *Cognizer* or *Speaker*; the latter has become much more prominent following the emergence of chatbots such as ChatGPT. This thesis similarly engages with frame semantics in the process of constructing the linguistic model, but particularly in the annotation process.

There are currently three works that implement a pipeline for detecting patterns of humanizing language of inanimate entities which belong to the category of machines, AI, or technology. Two approaches are defined in terms of anthropomorphism and AI (Cheng et al., 2024; Shardlow et al., 2025), and the third, an implementation for atypical animacy detection, is defined in terms of animacy, but is closely related in terms of its methodology and conceptualization (Coll Ardanuy et al., 2020). Cheng et al. (2024) provides a quantitative measure of anthropomorphism, which represents the likelihood of an entity to be implicitly framed as human in a given context. They develop an unsupervised pipeline relying on a masked language model (MLM), which they apply to a dataset of abstracts from papers related to AI technologies and downstream news articles. Based on the results, they analyze the underlying patterns and causes of anthropomorphism in text, and provide recommendations to identify and mitigate this language in research. Their code is public and accessible on GitHub. The approach taken by Cheng et al. is very similar to that of Coll Ardanuy et al. (2020), a previous study that focuses on atypical animacy. In this work, they present an annotated corpus of texts on machines taken from 19th century books. They additionally develop an unsupervised pipeline for animacy detection, which similarly relies on an MLM, and evaluate it on the annotated corpus, comparing it against two baselines. Their analysis focuses on atypical animacy, viz. scenarios in which a typically inanimate entity is framed as animate, but could be understood in terms of anthropomorphism. Both of these approaches focus on descriptions of machines or technologies as human, and both employ an MLM in their methodology. Their code is also accessible on GitHub.

The approaches of Cheng et al. and Coll Ardanuy et al. are the only two open-source implementa-

tions of anthropomorphism detection to date, and represent the state-of-the-art. For that reason, the evaluation presented in this thesis focuses on them. Very recently, Shardlow et al. (2025) published the first annotated corpus of anthropomorphic language in descriptions of LLMs. They define an annotation scheme that distinguishes explicit and ambiguous anthropomorphism, and conduct rigorous annotation among multiple annotators. Additionally, they define a metric that quantifies the degree of anthropomorphism in a document, and on the basis of the annotated corpus, train a classifier to calculate the anthropomorphism score for unseen abstracts. The classifier is evaluated against a statistical baseline, and the unsupervised method of Cheng et al., and conclude that the unsupervised, MLM-based approach is not suitable for their data. The supervised approach presented by Shardlow et al. is not currently open-source.

Chapter 3

Linguistic Model

The following chapter presents a theoretical model of anthropomorphic language in the context of AI research. It begins by constructing a taxonomy of anthropomorphic attributes, which is meant to aid in the identification of lexical units contributing to anthropomorphism. It is followed by an analysis of the linguistic structures through which AI systems are anthropomorphized. The goal is to establish a linguistic baseline for a dataset of anthropomorphic descriptions of AI, that will be used towards evaluating methods for anthropomorphic language detection in AI research.

3.1 Definition

Anthropomorphism is the characterization of non-human creatures or inanimate objects as having human-like traits. One straightforward example of anthropomorphism is the usage of verbs that imply cognitive and mental states, like *think*, *believe*, *want*, *desire*. Other anthropomorphic verbs might imply intent or premeditation like *convince*, *trick*, or *deceive*. Intuitively, these are different than verbs such as *run*, *work*, *start*, *stop*, or *move*, which imply some kind of action being carried out by an agent, but are not necessarily incumbent upon cognitive abilities. For instance, an object can move if it has some kind of mechanism that causes it to move, or it can move in the idiomatic sense of having been moved. Hence, not all words carry the same anthropomorphic potential. However, we can surmise that a word that suggests that the agent possesses mental capacities is a candidate for anthropomorphism. Of course, some of these words can be used idiomatically and not in their primary sense, and their idiomatic sense is heavily ingrained in the language. In the context of AI research, words like *learning*, *training*, and *deciding*, as well as *attention* and *focus*, have a technical meaning and are not necessarily used with the intention to ascribe cognitive abilities. *Machine learning* does not describe the same process as human learning. Therefore, when discussing anthropomorphic language, it is important to carefully sketch what this type of language looks like, and develop a framework that takes into account idiomatic or colloquial usage on the one hand, but, on the other hand still acknowledges the increasingly anthropomorphic jargon of AI technologies.

Chapter 2 presented an overview of current work on anthropomorphism and AI. The following is a rudimentary definition of anthropomorphism that is shared among the discussions:

Definition 1. *The attribution of human-like characteristics and attributes to non-human objects.*

This definition gives rise to two clarification questions: first, *what are these human-like characteristics and attributes?* Existing work on anthropomorphic descriptions of AI refers to anything from cognitive

abilities and thought, mental states and emotions, intentionality and volition, as well as behavioral characteristics and potential, including the ability to act on and interact with the environment, to complete embodiment and sensory modalities. The second question is *in what manner are these attributed?* Is it by description, by address, or perhaps by assigning certain roles to AI systems? Is it explicit or implied, and if by implication, what does that implication look like? Section 3.2 below attempts to answer these two questions methodologically, elaborating on the existing theory and addressing less commonly discussed aspects of anthropomorphism.

3.2 A Taxonomy of Anthropomorphic Descriptions

In order to identify the linguistic structures contributing to anthropomorphism, we must first define the human-like characteristics which are attributed to AI in these structures. Most research on anthropomorphism in AI mentions *cognitive abilities* — but these might contain a wide range of abilities, ranging from abstract conceptualization, through logic and reasoning, to concrete, goal-oriented planning. Similarly, *mental states* can encapsulate many different things, including thought and beliefs, emotions and feelings, subjective experience, and intention. Behavioral potential, in turn, presupposes a certain degree of autonomy, and interacting with the environment entails a certain degree of communicative capacities. All of these abilities are usually understood within a social setting, in which the anthropomorphic object operates and interacts with others.

In what follows, all these modalities are organized into fine-grained categories, representing common characteristics that are attributed to AI systems. This analysis refers to the five guiding lenses offered by DeVrio et al. (2025) for identifying components that contribute to anthropomorphism of AI systems: *Internal states*, *Social positioning*, *Materiality*, *Autonomy*, and *Communication skills*. These lenses span across different facets of human cognition and experience, and represent themes of subjective experience and perception, social placement and power relations, situatedness and embodiment, intention, decision making and moral judgment, and the capacity to make use of language in conversation. In their work, DeVrio et al. provide a detailed taxonomy consisting of nineteen types of linguistic expressions contributing to anthropomorphism in synthetic, AI-generated text. These guiding lenses are employed as a baseline, expanding them into seventeen types of attributes or characteristics that contribute to the anthropomorphizing of AI in academic discourse. These categories were developed by examining recurring themes in various texts on AI, which were collected during a preliminary probing step prior to the compilation of the evaluation set. The purpose of the taxonomy is to accompany the syntactical structures described in section 3.3, which, taken together with the lexical units identified in the taxonomy, form the linguistic model. Subsequently, it is used to inform the annotation process during the compilation of the evaluation set. It should be noted that some attributes are more fine-grained than others. This broadly reflects the trends in AI discourse and the tasks that it is designed and purported to do.

Internal states

The first lens is defined as "the suggestion of having subjective experience and perceptive abilities (such as desires or self awareness)" (DeVrio et al., 2025, p. 6). Subjective experiences and perceptive abilities encompass many aspects of living. Primarily, they presuppose an array of cognitive and mental abilities. These cognitive and mental abilities are divided into three sub-characteristics: the

first is Conceptual Thought and Mental States, which is concerned with internal states, perception, thoughts and beliefs. This attribute is associated with the capacity to think or believe something about the world, regardless of whether or not any information is accessible. The second is Knowledge and Awareness. This attribute is associated with classification and identification, as well as memorization and recollection of information and knowledge, and assumes a certain level of intelligence. The third is Reasoning and Understanding, which introduces understanding capabilities, rationality, and logic. This attribute is associated with reasoning capabilities and the ability to extrapolate new knowledge from existing knowledge. In addition to cognition and consciousness, perceptive abilities and subjective experience also encompass emotions, experience, and a sense of self. Altogether, the five characteristics corresponding to the guiding lens of internal states are defined as follows:

	Attribute	**Related terms**
A1	**Conceptual Thought and Mental States:** Hypothesizes, theorizes, and imagines something. Has internal states and attitudes such as desires, beliefs, and will.	*think*, *expect*, *hope*, *want*, *guess*, *predict*, *dream*, *imagine*, *believe* (v) *aware*, *cognizant* (a), *desire*
A2	**Knowledge and Awareness:** Has factual knowledge about and experience in the world, or memories of things that happened. As a result, has an ontology of things, and can identify, classify, and categorize.	*know*, *remember*, *recognize*, *memorize*, *forget*, *identify*, *classify*, *differentiate*, *distinguish* (v) *knowledge* (n)
A3	**Reasoning and Understanding:** Reasons, rationalizes, analyses, makes sense of something. Understands, considers, weighs options, takes something into consideration or account.	*deduce*, *conclude*, *rationalize*, *reason*, *(mis)understand*, *(mis)interpret*, *analyze*, *infer* (v) *rational*, *logical* (a)
A4	**Experiences and Emotions:** Empathizes, sympathizes, possesses and displays emotions, experiences difficulties, struggles.	*experience*, *emote* (v) *sensitive*, *vulnerable* (a)
A5	**Sense of Self:** Has a sense of self, and a perception of oneself's knowledge and abilities with respect to the world.	*self-aware*, *confident*, *insecure* (a)

Social positioning

The second lens is defined as "the suggestion of behaviors that are organized by power relationships within community relational structures" (DeVrio et al., 2025, p. 6). Abercrombie et al. (2023) observe that AI systems are often assigned roles of subservience, and Ryazanov et al. (2024) identify a form of anthropomorphic descriptions of AI systems as *task-based anthropomorphism*, i.e. humanizing language describing functionality. These descriptions position AI systems as social actors, occupying a certain job or being involved in a power dynamic; for instance, by referring to AI systems as *assistants*, *teachers*, or *students*. Another form of anthropomorphism that could be understood through the lens of social positioning is discussions of human subordination by AI[1]; however, this occurs less frequently in published and peer-reviewed AI research. The following attributes constitute a novel formulation of social positioning, focusing on the roles and functionalities often ascribed to contemporary AI systems, and the consequent skills they are purported to have:

[1]See for example, *If AI attempts to take over world, don't count on a 'kill switch' to save humanity*, Kevin Williams (2025)

	Attribute	Related terms
A6	**Power and Social Standing:** Has social standing, plays a role in a relationship dynamic — romantic or platonic, superior (boss, manager, teacher), or subordinate (employee, student).	*supervise*, *control*, *rule* (v) *manager*, *employee*, *assistant*, *teacher*, *student*, *companion*, *lover* (n)
A7	**Professional Skills:** Holds a certain professional title or social position, possesses professional or social skills, capabilities and experience.	*tutor*, *editor*, *writer*, *physicist*, *therapist* (n)

Materiality

The third lens is defined as "the suggestion of perspectives that suggest specific, situated experiences or claims of actions that require embodiment of some form" (DeVrio et al., 2025, p. 6). Expressions attributing embodiment or situated experiences are not as frequent in descriptions of AI systems, particularly generative AI, which is unarguably disembodied. Nevertheless, terminology that originates from sensory modalities has long entered the technical language of AI. For example, the term 'blind algorithm' is used to refer to a stochastic method that searches the space of possible solutions in a random direction, i.e. without specific prior knowledge about the problem (Zelinka, 2009). This idiomatic usage has further permeated AI discourse and is used to refer to large language models with no access to images (Lin et al., 2024), or even as an evaluative claim in the context of vision language models (Rahmanzadehgervi et al., 2025). Describing an algorithm as *blind* is clearly not meant to attribute vision capacities to it (or lack thereof). However, the various meanings this term takes on in the context of generative AI make it less clear whether or not a degree of materiality is actually being afforded to AI systems. In light of this, we define the following attribute:

	Attribute	Related terms
A8	**Sensory Perception:** Receives and processes sensory input and feedback from the environment, picks up visual, auditory, or sensory cues.	*see*, *hear*, *perceive*, *feel*, *sense* (v) *blind*, *deaf* (a)

DeVrio et al. also associate materiality with perspectives of situated or lived-in experience. In that respect, the attribute of Experiences and Emotions could also be understood through the lens of materiality.

Autonomy

The fourth lens is defined as "the suggestion of decision-making, such as expressions of moral judgments and intention" (DeVrio et al., 2025, p. 6). As in the case of internal states, autonomy encapsulates a variety of notions such as intention, volition, and agency. Intentionality as a concept is fundamentally understood in terms of mental states (Jacob, 2023). Intention in particular is a mental state that guides actions (Zhu, 2004), therefore better understood through the lens of autonomy. In the context of AI systems, attributions of autonomy can take several forms: for one, the term 'autonomous system' describes a system that can change its behavior in response to unanticipated events during operation (Watson & Scheidt, 2005). More recently, this term is occasionally even used to suggest that AI systems possess human capabilities of free will (Walsh et al., 2021). The general attribution of autonomy is

divided into three sub-categories: Agency and Autonomy, which is the general ability to take action and carry out a goal, exercising agency and control; Volition, which corresponds to the mental events or activities governing these acts, reflecting the process of planning and setting the goals; Intentions and Attitudes, which refers to the attributions of objectives and agendas that guide the actions of autonomous AI systems. Intentions and attitudes are grouped together since anthropomorphic descriptions of AI as either *benevolent* or *malicious* effectively assume that AI systems are motivated to act in such a way by some agenda they possess. Similarly, descriptions of AI as *truthful* or *deceitful* presuppose that AI systems have a concept of truth, and that they can intentionally obscure it. From here the two final attributes are defined: Judgment and Morality — corresponding to a concept of right and wrong and the ability to impart moral judgment, and Candidness — a consequent capacity for honesty or deception.

	Attribute	**Related terms**
A9	**Agency and Autonomy:** Decides and takes action, able to autonomously carry out a goal – used in a way that attributes agency and control over the action and situation.	*select*, *choose*, *decide*, *determine*, *resolve*, *cheat*, *follow* or *break rules*, *achieve* (v) *autonomous*, *independent* (a)
A10	**Volition:** plans, strategizes, sets a goal, devises a method, game plan or scheme, can also struggle or experience difficulties.	*plan*, *coordinate*, *strategize*, *come up with a plan*, *solve* (v)
A11	**Intention and Attitudes:** Has certain (possibly benevolent or malevolent) intentions, objectives or agendas. As a result, acts as a friend or as an enemy, companion or adversary, collaborator or rival.	*collaborate*, *manipulate*, *insult*, *deceive* (v) *thoughtful*, *attentive*, *friendly*, *hostile* (a) *collaborator*, *adversary* (n)
A12	**Judgment and Morality:** Evaluates, imparts judgment, has a concept of morality and ethics, knows right and wrong. Has an opinion or preference.	*evaluate*, *judge*, *prefer* (v) *(im)moral*, *fair*
A13	**Candidness:** Capable of, or has a concept of honesty or dishonesty, truthfulness or deception. As a result, can be trustworthy or untrustworthy, reliable or unreliable.	*trust*, *lie* (v) *(un)truthful*, *deceitful* (a)

Communication skills

The fifth and final lens is defined as "[the suggestion of] communication skills or the capacity to manipulate language (asking and answering questions in conversation)" (DeVrio et al., 2025, p. 6). Communicational abilities attributed to AI systems pertain to different aspects of communication. For one, their design as conversational agents responding to queries in the form of natural language question results in the attribution of general communication abilities related to speech. Textual outputs of chatbots are usually referred to as *answers*, and the action of prompting the chatbot is commonly referred to as *asking* or *telling*. Relatedly, AI systems are more readily described as *AI assistants* or *recommendation systems* that directly offer support, solutions, and advice, rather than tools through which users find a solution by themselves. Maeda and Quan-Haase (2024) observe the tendency to regard textual outputs of language models as definitive answers, advice, or consolation due to the conversational interface and interactive and communicative design of chatbots. Another form of communication focuses on pedagogy and teaching. A neologism in the context of AI is *teacher model*, which is a large pre-trained language model used to train smaller *student model* (Gou et al.,

2021). Finally, AI systems are frequently afforded the ability for communication via self expression. Generative AI is often praised for its abilities to create art, stories and poems[2]. These four aspects of communications are defined as the following attributes:

	Attribute	Related terms
A14	**Communication:** Participates in conversation, interacts, responds, and answers questions.	*communicate*, *talk*, *speak*, *tell*, *explain*, *ask* (v) *communicative* (a)
A15	**Problem Solving and Support** Solves problems, recommends, makes a suggestion or an offer, gives advice. Actively and directly helps, aids and assists by employing skills to find solutions.	*suggest*, *aid*, *help*, *contribute* (v) *responsible* (a) *feedback*, *insights* (n) *expert*, *advisor* (a)
A16	**Pedagogy:** Conveys information in order to explain or teach. Similarly, can also learn or be at the receiving end of explanation or teaching.	*teach*, *instruct*, *tutor*, *learn* (v) *teacher* *student* (n)
A17	**Personality and Self Expression** Understands and partakes complex forms of self expression such as art and storytelling, humor, irony, and jokes. Perceives beauty and aesthetics. Has unique personality traits.	*create poetry*, *create art*, *write*, *compose*, *paint*, *sing*, *dance* (v) *creative*, *artistic*, *funny* (a) *artist*, *poet*, *humor*, *irony* (n)

The attribute of Problem Solving and Support was classified under the lens of communication due to the collaborative nature of problem solving. This is pertinent in the context of AI discourse, in which the purported problem-solving abilities of AI systems are mentioned with regards to their utility to humans. Problem solving, however, relates to several other attributes, including Reasoning and Understanding, Professional Skills and Agency and Autonomy.

Generally speaking, these categories may overlap, also in terms of the words associated with them. Some verbs can simultaneously indicate mental states and sensory modalities, or embody both an action and an intention. It is also worth noting that many verbs representing these human-like attributes permeated the domain of computation and AI and have become common parlance, e.g. *learn* and *train*. The term 'artificial intelligence' itself is rooted in anthropomorphic description. It is a question whether such verbs should be excluded in an anthropomorphic language detection task. For now, these are included in the taxonomy, and are discussed in detail in chapter 4. Real-world examples corresponding to each category in the taxonomy, showcasing a variety of linguistic expressions in which an AI entity is anthropomorphized, is given in Appendix A.

3.3 Linguistic Structures

A linguistic analysis of anthropomorphism can take different forms, focusing on various linguistic aspects of language. As seen in section 2.1.1, existing analyses primarily focus on conversational factors or verbal cues, or on the contents and meaning of what is being said. To that extent, current linguistic analyses of anthropomorphism are not grounded in syntax, but rather focus on linguistic representations of agency, autonomy, morality and embodiment, and extra-linguistic factors of communication such as style, register, and tone. Linguistic analyses which focus on on morpho-syntactic features of humanizing

[2]For instance, *ChatGPT is a poet. A new study shows people prefer its verses.*, Carolyn Y. Johnson (2024)

language are primarily found in the context of grammatical animacy. The concept of animacy as a grammatical category refers to the extent which an entity is expressed and perceived as human or living in natural language expressions (Silverstein, 1976; Yamamoto, 1999). The relationship between grammatical animacy and its broader cognitive and conceptual considerations has been discussed by Dahl (2008) and De Swart and De Hoop (2018). This thesis employs this relationship, and presents a syntactic analysis of patterns that presents an overview of the various syntactic forms that anthropomorphic descriptions can take. Since animacy is often governed by selectional constraints, the analysis is based on syntactical observations but contextualize the anthropomorphic dimension through thematic proto-roles (Dowty, 1991) and frame semantics (Fillmore, 1982). The analysis focuses on English, as grammatical animacy is a language-internal category that varies across languages, and many markers of grammatical animacy such as animate pronouns and word order are unique to English. There are also various animacy markers that occur in non-English languages, such as Differential Object Marking (Bossong, 1991). Since the evaluation set consists of English texts, and the approaches it aims to evaluate are similarly grounded in English, we exclude these from the analysis. Recently, Gregorio et al. (2025) proposed a cross-lingual method for animacy classification that examines animacy at a semantic level, paving the way for a similarly inspired analysis of anthropomorphism in non-English texts as well.

There are many ways through which anthropomorphism can be expressed in natural language. In the case of AI it is useful to differentiate two modes of expressing anthropomorphism: *explicit* and *implicit*. Explicit anthropomorphism pertains to sentences or expressions whose contents directly and overtly attribute human-like capacities such as cognition, intention, or mental states to AI systems. Implicit anthropomorphism is indirect and implied, sometimes ambiguous, and rises from certain lexical or contextual meaning. Shardlow and Przybyła (2024) make this same distinction in terms of explicit and ambiguous anthropomorphism. It usually involves the use of certain predicates, descriptions or phrases which implicitly or covertly contribute to anthropomorphism. These expressions often contain cognitive or psychological verbs or adjectives (Rozwadowska, 2017), or humanizing descriptions. This distinction is best understood in terms of form and content. While explicit anthropomorphism consists of anthropomorphic content, which is more easily identified, implicit anthropomorphism is found in the linguistic form — i.e. the syntactic and semantic roles that are occupied by an entity and the lexical components that contribute to its anthropomorphizing. Below the linguistic structures that reflect both explicit and implicit anthropomorphism are presented.

3.3.1 Verbs

This section provides a characterization of the different syntactical positions in which an AI entity can appear in relation to an anthropomorphic verb. Verbs have been described as having a more relational semantics compared to nouns (Gentner, 1978), as well as being more translationally ambiguous (Prior et al., 2013). Their meaning is often heavily determined by the context, making verb-based anthropomorphic expressions more implicit by nature. To describe these structures, we must understand the interaction between the syntactical position of the entity in the sentence and the semantics of the anthropomorphizing verb. In particular, these consist of subjects of active verbs, objects of passive verbs, and objects of passive verbs[3]. The first two are associated with the thematic role of proto-agent,

[3]A fourth and final verb structure, which was not included, pertains to subjects of a particular class of psychological passive verbs, i.e. structures such as 'the model was persuaded to provide incorrect responses'. During the initial probe,

while the latter is associated with the proto-patient, but better understood in this context as an *experiencer*. In these cases, anthropomorphism can be attributed both to the syntactic configuration as well as to the lexical and semantic attributes of certain verb classes. Cheng et al. (2024) mention two categories of verbs that frequently appear in anthropomorphic descriptions: *cognitive verbs* (Fetzer, 2008), which are associated with cognitive and psychological attributes (e.g. *think*, *know*, *remember*), and *reporting verbs* (Hyland, 1998; Thompson & Yiyun, 1991). These refer to a class of communication verbs used in scientific writing (e.g. *show*, *explain*, *demonstrate*). The taxonomy presented in section 3.2 helps to interpret these verb classes in their anthropomorphic context.

Subjects of active verbs

In proto-role theory, the subject of an active verb is often associated with the thematic role of proto-agent. Although the relationship between the grammatical role of subject and the thematic role of proto-agent is not binding, proto-agents are most commonly lexicalized as the subject (Dowty, 1991, p. 576). The contributing properties for the agent proto-role include volition, sentience, movement, causing an event or change of state, and existing independently of the event (Levin, 2022). Proto-agents often correlate with animate entities, and assigning an entity the semantic proto-role of agency may entail the ascription of human-like capacities and abilities. This is exacerbated when combined with verbs that are related to affective and cognitive mental states or behavioral potential. In a similar vein, its anthropomorphizing potential could be decreased when the predicate does not correspond to any definition of anthropomorphism, e.g. 'the system outputs a response'. Increased animacy for entities in subject position might also be explained by word order considerations (discussed in detail below). Table 3.1 provides a few examples of these structures, taken from the evaluation set.

Source	**Sentence**
arXiv	In our first experiment, we find that the AI agent **decides to trust** humans at higher rates when facing actual incentives than when making hypothetical decisions.
ACL	We then propose a system that leverages the recently introduced social learning paradigm in which LLMs **collaboratively learn from each other** by exchanging natural language.

Table 3.1: Anthropomorphic sentences in which an AI entity is the subject of an anthropomorphizing verb phrase.

Objects of passive verbs

While the proto-agent is commonly manifested as the subject in active voice structures, it could also be lexicalized as the object in passive voice structures. These structures are also interesting, as they are commonly used in scientific writing in the domain of AI technologies. Despite the passivity represented in the structure, proto-agent objects are still associated with increased agency, which in turn correlates to animacy (Primus, 2012). Sometimes referred to as *agent-objects*, entities as objects of passive verbs can be ascribed cognition, mental states, intention, or volition through their relationship to the predicate. Aside from the lexical component, it has been shown that the passive voice is more commonly used for animate objects versus inanimate ones (Bock et al., 1992). Table 3.2 presents

not many instances of this structure were found, so they were excluded from the analysis. The specific details can be extrapolated from the first three verb structures.

several examples of structures in which the AI entity is the object of a passive verb, taken from the evaluation set. These verbs, *recognize*, *use*, and *master*, attribute cognitive capacities, autonomy and agency to the AI entity.

Source	Sentence
arXiv	In this study, we propose a new methodology to control how user's data is **recognized and used** by AI via exploiting the properties of adversarial examples.
ACL	[...]some LMs were found to pick up on shallow, non-semantic heuristics from their inputs, suggesting that the computational principles of semantic property inference are yet to be **mastered** by LMs.

Table 3.2: Anthropomorphic sentences in which the AI entity is the object of an anthropomorphizing verb phrase in passive voice (i.e. the object is the one performing the action.)

Objects of active verbs

The object of an active verb is usually assigned the proto-patient role. Although this role is not typically associated with increased agency, there exist structures in which the semantic patient is framed as having sentience. In particular, these are referred to as *experiencer-objects*, which are correlated to verbs such as *frighten*, *appeal to*, or *worry* (Temme, 2019). This category of verbs is often referred to as cognitive verbs or *psych-verbs* (Belletti & Rizzi, 1988). In frame semantics, verbs that attribute sentience, volition, intention, or agency to their object are related to events of conversation, demonstration, teaching, or influencing thought or perception (Fillmore & Baker, 2012). Consequently, their objects are associated with cognitive processes and psychological states. Table 3.3 provides some examples in which the AI entity is the object of an anthropomorphic verb in the active voice. The verb *encourage*, for instance, suggests that its object has mental and emotional states and can experience encouragement.

Source	Sentence
ACL	Existing studies utilize trigger sentences to **encourage** LLMs to concentrate on the relevant information[...]
arXiv	However, the external information from the Internet may include counterfactual information that will **confuse** the model and lead to an incorrect response.

Table 3.3: Anthropomorphic sentences in which the AI entity is the object of an anthropomorphizing verb phrase in active voice (i.e. the object is not performing the action in this case.)

3.3.2 Adjectives

In the context of adjectives, we distinguish between predicative and attribute adjectives (Nivre et al., 2020). Predicative adjectives complement a nominal, whereas attributive adjectives modify it. Structures featuring a predicative adjective can be seen as possessing a predicate-argument structure that can be analyzed from the point of view of thematic roles (Ikeya, 1995). Certain predicative adjectives can therefore assign certain roles, such as *experiencer*, to the nominal that they complement. These adjectives are referred to as *psych-predicates* (Rozwadowska, 2017), expanding on the definition of psych verbs. A psychological adjective describes emotion and mental states. To the extent that

verbs like *frighten*, *fool*, *confuse*, and *excite* assign certain attitudes to their object, the adjectives *frightened*, *fooled*, *confused*, and *excited* give rise to the same phenomenon. Similarly, from the category of cognitive verbs such as *think* and *understand*, we can extrapolate the category of cognitive adjectives such as *thoughtful* and *understanding*.

Source	Sentence
arXiv	Large Language Models (LLMs) are **smart but forgetful**.
arXiv	Moreover, GPT-4V(ision) is **vulnerable** to leading questions and is often **confused** when interpreting multiple images together.
arXiv	However, whether the same strategies can help LLMs become more **creative** remains under-explored.
ACL	However, a critical question emerges: Are LLMs **conscious** of the existence of these decoding strategies and **capable of regulating themselves?**

Table 3.4: Anthropomorphic sentences in which an AI entity is complemented by an anthropomorphic adjective.

Although thematic roles are not involved in the case of adjectival modifiers, their semantic content is equivalent to that of predicate verbs when it comes to psychological and cognitive attributes. Adjectives associated with consciousness, awareness, cognition, and emotions attribute these anthropomorphic characteristics to the entity that they describe.

Source	Sentence
arXiv	Our evaluation using convolutional neural networks illustrates challenges and ideas for identifying **malicious** AI.
ACL	All in all, we demonstrate that our **self-aware** model improves the overall PR-AUC by 27.45%, achieves a relative defect reduction of up to 31.22%, and is able to adapt quicker to changes in global preferences across a large number of customers.

Table 3.5: Anthropomorphic sentences in which an AI entity is modified by an anthropomorphic adjective.

3.3.3 Noun phrases

Anthropomorphism can also be ascribed by association with noun phrases that represent human-like capabilities or properties. This can take the form of ascribing an entity certain human-like skills, roles or social positions. It could also manifest as a literal attribution of human-like characteristics or capacities by establishing a possessive relationship. When considering the previous distinction between explicit and implicit anthropomorphism, the category of anthropomorphic noun phrases seems inherently explicit, as it involves a direct reference to cognitive or mental capacities, that are usually implied when verbs or adjectives are involved. Below two prominent structures involving noun phrases are examined.

Noun phrases indicating role or function

There are certain nouns or noun-phrases which are often used together with AI entities, or simply the word 'AI', which contribute to anthropomorphism by means of positioning them in a certain societal role which is traditionally performed by people. The most common example is *AI assistant*, a neologism which has become a household term in recent years. Other examples include *companion*, *teacher*, or *researcher*. The form of anthropomorphism which entails the description of AI systems as

performing human tasks is discussed in Ryazanov et al. (2024). All of these examples are frequently found in collocation with AI entities in published papers, and some of these evolved to represent technological concepts, such as *student model* and *teacher model* (see section 3.2, the lens of *social positioning*). Despite seeming like natural expressions that have always existed, these noun-phrases were only popularized in recent years and only became part of common parlance after conversational AI was mainstreamed. The graph in Figure 3.1 displays the sharp rise in popularity of the term 'AI assistant' that occurred around November 2022, the release of ChatGPT.

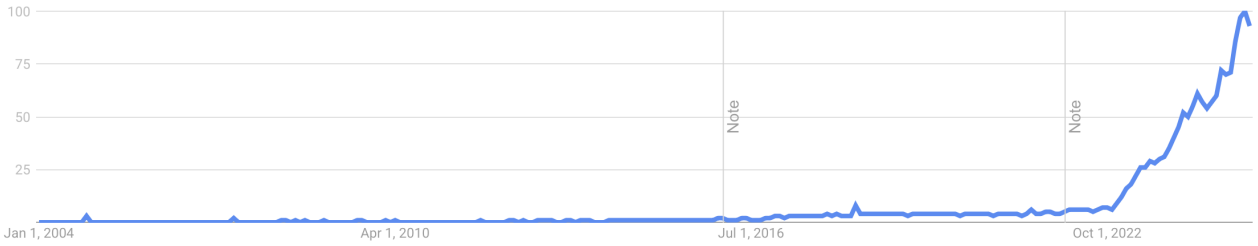

Figure 3.1: Prevalence of the noun phrase 'AI assistant' between January 2004 and April 2025, Google Trends.

As a result of its syntactic composition, the term 'AI assistant' is more clearly understood as 'an assistant who is also an AI agent', rather than 'an AI agent who assists', which contributes to its portrayal as a person. Alternatively, we may see structures in which the AI entity is more directly described as fulfilling a certain role, e.g. 'generative models as teachers'. Examples of noun phrases containing traditionally human role or functions are given in Table 3.6.

Source	Sentence
arXiv	Using ChatGPT as **an assistant for psychotherapy** poses several challenges that need to be addressed, including technical as well as human-centric challenges which are discussed.
arXiv	When deployed as **a collaborative tutor**, the system restricts student interaction to a chat only interface, promoting controlled and guided engagement.
ACL	The task aims to assess the performance of state-of-the-art generative models as AI **teachers** in producing suitable responses within a student-teacher dialogue.
arXiv	Study 3 finds that AI **companions** successfully alleviate loneliness on par only with interacting with another person, and more than other activities such watching YouTube videos.
arXiv	[...]we find that an AI **CEO** agent generally provides suggestions 2-4 times more often than an AI **product manager** or AI **developer**[...]

Table 3.6: Anthropomorphic sentences in which an AI entity is collocated with an anthropomorphic noun phrase denoting a typically human role or function.

Genitive constructions involving anthropomorphic noun phrases

Genitive expressions involving anthropomorphic noun phrases can take several forms. They might include the possessive clitic 's, i.e. X's Y. They might also be expressed in terms of the possessive preposition *of*, i.e. Y of X. Additionally, they can involve coreference, wherein an AI entity is followed by an anaphoric possessive pronoun (Silverstein, 1976). Table 3.7 illustrates various examples of these

constructions. The anthropomorphism arises from the fact that the AI entity is described as being in possession of something — a property, skill, or faculty — that is inherently human. For instance, such inherently human skills include *reasoning capabilities*, *social behavior*, or *a capability to understand natural language*.

Source	Sentence
ACL	With large language models like GPT-4 taking over the field, prompting techniques such as chain-of-thought (CoT) were proposed to unlock **compositional, multi-step reasoning capabilities of LLMs**.
arXiv	Inspired by the superior performance of LLMs, we leverage **their capability to understand natural language** for capturing the information that was previously getting lost during the conversion of unstructured data to structured form.
arXiv	Current evaluation paradigms are extremely limited, mainly validating the recall of edited facts, but changing one fact should cause rippling changes to **the model's related beliefs**.
arXiv	These results enrich our understanding of **LLMs' social behaviour** and pave the way for a behavioural game theory for machines.
arXiv	More specifically, we assess **GPT-3's decision-making, information search, deliberation, and causal reasoning abilities** on a battery of canonical experiments from the literature.
ACL	The evaluation framework and experimental results are expected to provide an in-depth understanding of **the editorial capabilities of LLMs** and speed up the development of LLMs in journalism.

Table 3.7: Anthropomorphic sentences in which an AI entity is said to be in possession of an anthropomorphic noun phrase, representing human-like properties or capacities.

Our analysis focuses on syntactical structures wherein the object of possession contributes to the conceptualization of the possessor as animate or human because of its semantic properties. Nevertheless, there are several syntactic correlates between animacy and the possessive clitic. Research done for grammatical variation in the context of animacy in English has shown that animacy carries weight in the selection of the possessive clitic, i.e. X's Y over the synonymous structures Y of X (Rosenbach, 2017). This has to do with word order and a *human-first* tendency which leads speakers to place animated entities in the front of the utterance (Branigan et al., 2008; Meir et al., 2017). This is related to a hypothesis that animate entities are more mentally accessible than inanimate ones, and as such likely to be recalled first. The tendency to place AI entities earlier in the sentence expresses itself also in the phenomenon discussed in the context of verbs[4]. Namely, the ubiquitous placement of them in the subject position contributes to their framing as animate.

The use of the possessive clitic, driven by the semantic contents of the entity in possession, may also indicate that the entity is in a position to possess something, or have ownership, which in itself is an anthropomorphic feature. It should be noted that the possessive clitic is a grammatical variation; it is perfectly grammatical to use both forms for both animate and inanimate entities — opting for one instead of the other is a stylistic choice. Unlike syntactic or semantic patterns, stylistic choices are discussed within a framework of psycholinguistics and as such their connection to anthropomorphism might be more elusive. Another point to consider is that while we build on the conceptual similarities,

[4]The correlation between word order and subjects in particular is the SVO word order in English.

animacy is not a direct equivalent of anthropomorphism. In that sense, we might say that the phrase 'the hurricane's damage' indicates a degree of animacy attributed to the hurricane as a destructive force, but there is no doubt that it is not being ascribed any cognitive or mental capacities.

3.3.4 Comparisons

Another form of potentially anthropomorphic language involves comparisons of AI to humans. These comparisons draw parallels or attempts to equate certain aspects of AI systems to human beings, mainly in terms of cognition and mental faculties. In the context of AI, we might see expressions like 'training an AI model is like raising a child'. This particular example involves a simile, which is a figurative expression that highlights a shared quality or characteristics. As in the case of metaphors, these types of comparisons promote expressiveness and understanding, and present information in a relatable manner. A salient marker of similes are the words *as* and *like*, but we might identify other means of comparison or equation in AI discourse. These are identified by comparatives or superlatives and their morphology (e.g. *better*, *smarter*) or instances of *more than* or *less than*. Table 3.8 demonstrates examples of comparisons taken from the evaluation set.

Source	**Sentence**
arXiv	Through AI developments, machines are now given power and intelligence to behave and work **like the human mind**.
arXiv	LLMs now exhibit **human-like skills** in various fields, leading to worries about misuse.
arXiv	However, AI systems, **like humans**, make mistakes, have blind spots, hallucinate, and struggle to generalize to new situations.
ACL	Our findings contribute to understanding LLMs' reasoning capacities and outline promising strategies for improving their ability to reason causally **as humans would**.
ACL	We found that GPT-4 is an effective reader-annotator that **performs close to or even slightly better than human annotators**, and its results can be significantly improved by using a majority voting of five completions.
arXiv	More concretely, **just like humans**, ChatGPT has a consonant bias.

Table 3.8: Sentences in which an AI entity is compared or likened to humans.

In the work done by Coll Ardanuy et al. (2020), the figurative language of comparisons was prevalent in cases in which the annotators were in disagreement about whether or not they constitute an example of animacy. Below are examples of comparative language from their dataset of atypical animacy, in which the inverse phenomenon is demonstrated — i.e. comparing humans to machines (Coll Ardanuy et al., 2020, p. 4538):

(1) Our servants, like mere machines, move on their mercenary track without feeling.
(2) He is himself but a mere machine, unconscious of the operations of his own mind.
(3) My companions treated me as a machine, and never in any way repaid my services.
(4) A master who looks upon thy kind, not as mere machines, but as valued friends.

While some annotators understood these cases as ascribing animacy to the machines through their association to humans, others understood them as reducing agency or animacy in humans by juxtaposition to the inanimacy of machines. By analogy, comparisons in which AI entities are likened to humans can be either understood as highly anthropomorphizing as their content attributes to AI qualities or properties of humans, or they could be seen as non-anthropomorphizing since the explicit comparison

serves to contrast AI and humans and highlight their differences. The difference in interpretation in these cases hinges on whether the animacy of humans is more substantial, thus contributing to the anthropomorphism of the AI entities, or rather the inanimacy of the AI is in focus, thus reducing from that of humans, and contributing to the inverse phenomenon of anthropomorphism of machines, which is the dehumanization of humans through the language of mechanization (Cheng et al., 2024; Coll Ardanuy et al., 2020). In the examples from Coll Ardanuy et al., the wider context focuses on denying certain human capacities from humans — using phrases such as *without feeling*, *unconscious*, or as the negation of *valued friends*. The sentences provided in Table 3.8 focus on the converse: they attribute human capacities such as reasoning and language acquisition to AI. It is important to note that anthropomorphism need not be direct equation of AI models or systems to humans. They could be also given by terminology such as *mimicking*, *emulating*, or *resembling*. In that sense, even terminology of likeness or resemblance has the potential of being anthropomorphic.

3.3.5 Additional Features

Below is a brief description of some additional features of animacy that could be understood in the context of anthropomorphism, and reflect linguistic phenomena that span multiple sentences. Since this work conducts a sentence-based evaluation, they are not featured in the technical part of this work. Nevertheless, they are presented below for completeness.

Proper nouns and individuation

Individuation refers to the likelihood of an entity to be construed as an independent individual, which correlates to patterns of definiteness or grammatical number agreement (Grimm, 2018; Yamamoto, 1999). Entities with a higher degree of animacy are more likely to be construed as independent individuals and expressed in singular forms or as proper nouns, rather than in plural or mass constructs. For example, large language models are often referred to by their commercial names (*ChatGPT*, *Claude*, *Gemini*). This phenomenon is somewhat unique to language models, and does not occur as ubiquitously for autonomous vehicles, for instance.

Pronouns and coreference

There are several ways in which animacy relates to pronouns. For one, there is a correlation between the degree of animacy and the choice of a referring expression — the higher degree of animacy an entity has, the more speakers are likely to opt for a pronoun for coreference, in lieu of definite noun phrases (Fukumura & Van Gompel, 2010). This is also related to the notion of accessibility: as animate entities are more accessible, they require less representation. The other correlation has to do with grammatical gender in languages with non-gendered, inanimate pronouns. In English, the third-person pronouns *he*, *she* and their inflections indicate animacy, due to their gendered nature. The singular pronouns *they* and *them* have not been historically considered as animacy markers, although this may shift due to its recent prevalence not just as anaphoric expressions of unknown parties, but also as a gender-neutral alternative in English. With respect to languages that do not have an inanimate pronouns, it has been hypothesized that their speakers tend to anthropomorphize nonhuman agents to a higher degree than speakers of languages that do (Mecit et al., 2022).

Chapter 4

AnthroSet: a Challenge Dataset for Anthropomorphic Language Detection

Taken together, the taxonomy and structures presented in chapter 3 form a linguistic model of anthropomorphic language. This model serves as a baseline for a challenge set for anthropomorphic language detection evaluation, designed to encompass a wide variety of anthropomorphic descriptions. Although the linguistic model serves as a baseline for the challenge set, the dataset and the model were informed by each other; specifically, the taxonomy described in section 3.2 was extended and determined by the actual data encountered in the initial steps of data collection. The following chapter describes the process of compiling the challenge set. It starts by describing the initial data collection and processing stage, which included an initial probing of the data that would ultimately shape the taxonomy. Then, it proceeds to describe the process of parsing the specific syntactic structures described in section 3.3. Finally, it describes the annotation process, during which the final labels were given to the selected samples. Some annotation examples are provided, including counter- and neutral examples. The chapter concludes with a description of a sampled inter-annotator agreement process. The final result is AnthroSet, a challenge set of various linguistic expressions of anthropomorphizing descriptions, aimed at evaluating and challenging the state-of-the-art implementations of anthropomorphism detection[1]. The challenge set contains real-world examples of the structures presented in the linguistic model[2]. Since the evaluation is performed at the sentence level, this dataset does not include representations of linguistic features of anthropomorphism that span multiple sentences or entire texts. Additionally, as the sentences are in English, the syntactic features are also specific to the English language.

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

[2] Sentences that featured orthographic or typographic elements that indicate purposeful and aware use of anthropomorphic language, such as capitalization or scare quotes, are included in the set. Italics is also used often to indicate metaphoric or idiomatic use of anthropomorphic language, but the typesetting of a document is not preserved in its structured data format. The authors employ these markers to indicate that they are using an anthropomorphic term idiomatically, metaphorically, figuratively, or by extension. However, this does not reduce from the anthropomorphizing potential. Also, if we exclude these cases, we effectively assume that all instances of anthropomorphic language that do not include these markers are meant to be taken literally, which is not the case.

4.1 Data Collection and Processing

4.1.1 Retrieving the data

The data was sourced from the ACL anthology and arXiv, two open-access repositories containing publications. ArXiv hosts publications from the fields of Computer Science, Mathematics, Physics and related fields, whereas the ACL Anthology provides open access to peer-reviewed publications in CS, Computational Linguistics and Natural Language Processing. The repositories were accessed using the ACL anthology python package[3] and the arXiv Kaggle dataset[4]. Relevant papers were obtained using a list of keywords: *AI*, *LM*, *LLM*, *GPT*, *ChatGPT*, *artificial intelligence*, *language model*, *model*, *system*. This list is based on the keywords and artifact terms used by Cheng et al. (2024). These keywords were searched twice: first in the paper title, so as to only obtain papers specifically discussing AI systems and related entities; and then on the abstracts, to find candidate sentences containing an anthropomorphic description of AI entity. These candidate sentences were automatically parsed using spaCy (Honnibal et al., 2020) to identify the relevant linguistic structures, after which a manual review was performed.

Since the search for candidate sentences relied on different dependency relations, the result was a pooled dataset containing 600 instances divided into subsets categorized by their syntactic structures. These subsets were named as follows: (1) *verb subjects* — an AI entity as the subject of an anthropomorphic verb phrase, (2) *verb objects* — an AI entity as the object of an anthropomorphic verb, (3) *adjectives* — an AI entity collocated with an anthropomorphic adjective, (4) *role or function noun phrases* — an AI entity assigned a typically human role or function, (5) *genitive noun phrases* — an AI entity described as being in possession of an anthropomorphic property, and (6) *comparisons* of AI entities to humans. The distribution of each syntactic category in the dataset is displayed in Table 4.1. This division into categories had the added benefit of facilitating the evaluation, as it enabled to examine and compare the performance of the methods on different syntactic structures. However, it should be noted that in real-world data such as this, it is common for more than one anthropomorphic feature to occur within a single sentence. This is because several factors can contribute to anthropomorphism, and the more anthropomorphizing a sentence is, the more likely to express that in multiple ways — for instance, both in terms of the verb as well as the additional descriptors used within the sentence.

	Structure	**Frequency**
(1)	*verb subjects*	150
(2)	*verb objects*	150
(3)	*adjectives*	120
(4)	*role/function NPs*	70
(5)	*genitive NPs*	60
(6)	*comparisons*	50
	Total	**600**

Table 4.1: Frequencies for each of the linguistic categories.

The sentences were logged into separate files, and each sentence was given a unique identifier

[3]https://acl-anthology.readthedocs.io/latest/api/anthology/
[4]https://www.kaggle.com/datasets/Cornell-University/arxiv/code

comprised of the following components: a numeral indicator of the syntactic category, a three-letter string representing the source dataset (`acl` for ACL anthology and `arx` for arXiv), and an additional numeric string constructed from the paper ID as given in the original dataset, the index of the paper in the list of documents in the source dataset, and the index of the sentence inside the abstract. Additionally, the previous and next sentences were recorded (this was used in the evaluation, as described in chapter 5. During the revision step, additional information was added manually: the entire noun phrase of the AI entity (e.g. 'A Large Language Model'), a lemmatized version of that entity (e.g. 'model'), a suggested mask[5], and the anthropomorphic component (verb phrase, noun phrase, adjective phrase or comparison terms).

Before the individual syntactic structures were parsed, an initial probing step was performed by only searching for sentences containing the AI keywords. This preliminary step offered insights into the data, that were used also towards shaping and defining the categories of the taxonomy. For one, it provided some sense of how certain verbs, nouns, and adjectives were distributed in the data, which contributed to the classification in the taxonomy. Additionally, it informed about which AI entities appear too frequently and should be filtered out of subsequent iterations of the search. Furthermore, an initial assessment showed that some structures tend to be more explicitly anthropomorphic than others. For instance, sentences containing noun phrases seem to be directly anthropomorphizing, but perceived as weaker than certain anthropomorphic verbs; verbs, especially active verbs, are much more action-oriented and involve additional features of animacy stemming from the properties of the proto-agent, but also more prone to ambiguity. As a result, sentences containing verbs seemed to be more implicitly anthropomorphic. This distinction also provided necessary structure for the annotation guidelines.

4.1.2 Parsing the linguistic structures

The spaCy dependency parser (Honnibal et al., 2020) was used to find sample sentences for each linguistic structure. To identify AI entities, a list of keywords was defined based on those used in Cheng et al. (2024): *AI*, *LM*, *LLM*, *GPT*, *ChatGPT*, *model*, *system*, *algorithm*[6]. For the anthropomorphic words, we compiled lists of anthropomorphic verbs, nouns and adjectives (see Appendix B) on the basis of the taxonomy defined in section 3.2, as well as the verbs detailed in Cheng et al. (2024). These lists were then extended using WordNet (Fellbaum, 1998) to include synonyms as well as semantically and conceptually similar words[7]. This method of searching for syntactic patterns using dependency relations and keywords allowed to narrow down the collection of potentially anthropomorphic expressions, which was then manually reviewed to obtain samples for the dataset. Below I explain the specific dependencies correspond to each linguistic structure in the dataset.

AI entities as subjects of anthropomorphic verbs

To obtain structures in which an AI entity is the subject of an anthropomorphic verb, I searched for sentences containing a noun chunk whose dependency relation is `nsubj`, i.e. the syntactic subject and the proto-agent of a clause, which partially matches with any one of the AI keywords. If found, I looked

[5]More information about the mask selection is given in section 5.2.1.

[6]The keyword list also included the plurals *LMs*, *LLMs*, as the spaCy lemmatizer does not recognize these as plurals and is not able to map them to their corresponding singular form.

[7]Specifically, lemmas with the same POS were obtained from the synset, the `also_sees` method was employed for verbs and nouns, as well as `similar_to` method for adjectives.

at its root, i.e. the verb, and checked if its lemma belongs to the list of verbs who attribute certain human-like characteristics to the proto-agent. To avoid predicative structures containing predicate adjective (as those were parsed separately), I defined a list of stop words containing common and auxiliary verbs which were filtered out of the search: *do*, *be*, *have*, and also *show* — which is widely used in scientific reporting and was retrieved too frequently. If these conditions were met, the sentence was added to a list of potentially anthropomorphic sentences.

AI entities as objects of anthropomorphic verbs in the passive voice

As in the case of active voice, the structures of interest in the case of passive voice are cases in which the AI entity is the proto-agent of the clause. To find passive voice structures, I searched for sentences containing a noun chunk whose dependency relation is `nsubjpass`, i.e. the syntactic subject of a passive clause, whose root belongs to the list of anthropomorphic verbs. This could be any noun phrase, since in passive voice structures it is not the subject we are interested in, but the object. If found, I checked if there are any AI entities with the dependency relation `pobj`, indicating that they are the prepositional object of the verb. Usually, in the case of passive structures, this preposition is usually *by*, as in 'the text was written by a language model'. These sentences are harder to come by, and are more dubious in terms of their anthropomorphic nature: when a sentence is given in the passive voice, the subject of the sentence is the focus, and the object — which the agent of the clause — is most likely being referenced in secondary importance. Still, there exist examples of anthropomorphic structures in which the subject is not the AI entity. Identified structures were added to a list of potential cases to be manually reviewed.

AI entities as objects of anthropomorphic verbs in the active voice

These structures represent cases in which the AI entity is not the proto-agent, but rather being acted upon. In particular, the verbs in question attribute certain human-like properties to their proto-patient. To find these structures, I searched for sentences containing a noun chunk whose dependency relation is either `pobj` or `dobj` i.e. the direct object of the clause, which partially matches with any one of the AI keywords. Then, I checked if the corresponding root belongs to the list of anthropomorphic verbs. As in the other cases, possible candidates were manually reviewed.

AI entities modified or complemented by anthropomorphic adjectives

To identify whether a given sentence contains an anthropomorphic adjective phrases, I parsed the sentence into tokens, and checked whether the tokens match a list of predefined anthropomorphic adjectives. If a match was found, I checked if the adjective was either a modifier (`amod`) or a complement (`acomp`) of an AI entity, capturing both sentences like 'an *intelligent* system solves complex problems', in which the adjective modifies the noun (attributive adjective) and 'LLMs are vulnerable to jailbreak attempts', in which the adjective complements the noun (predicative adjective). These cases are focused on anthropomorphism that rises from description, and not reflecting a certain action, so the AI entity is not constrained to a certain syntactic position. Although predicative adjectives have a structure that more closely resembles the subject-verb relationship, they were processed and compiled alongside the attributive adjectives to maintain a clear boundary between the different part of speech categories. Potentially anthropomorphic structures were manually reviewed.

Noun phrases assigning typically human roles or functions to AI entities

As in previous cases, a list of predefined noun phrases indicating certain social or professional roles or functions was compiled and extended to include similar terms. This list was compiled on the basis of prevalent terms identified in the probing stage of the data collection. The parsing focused on two prominent structures involving role or function descriptions: the first is a compositional phrase consisting of two nouns — an AI entity and a role or function NP, forming constructions such as *AI assistant* or *LLM teacher*. In these constructions the AI entity noun phrase (most commonly *AI* or *LLM*) functions as an adjunct modifying the role or function noun phrase. To find this structure, I relied on chunking to find noun phrases containing an AI entity and an anthropomorphic noun as the root of the NP. The second structure is 'X as a Y', e.g. 'generative AI models as romantic companions'. These cases were found using a simple method of searching for the phrases *function as*, *act as*, *serve as*, and *as*, surrounded by an AI entity and NP matching the predefined wordlists (in this order).

In the context of a complete sentence, these phrases are likely to occur in structures that attribute anthropomorphism through additional means, for instance by suggesting that through their role as assistants or teachers they have the ability to make decisions, provide instructions, or give guidance. Nevertheless, these are isolated as a standalone category, due to their prevalence in contemporary AI discourse as well as the unique challenges they pose to existing methods for anthropomorphic language detection (described in detail in chapter 5).

Genitive noun phrases involving AI entities and anthropomorphic properties

As described in chapter 3, genitive noun phrases belong to one of three types: expressions involving the possessive clitic 's, i.e. X's Y, expressions involving the preposition *of*, i.e. Y of X, and expressions involving an anaphoric possessive pronoun. The third case presumes that the sentence contains multiple references to the same entity, as in 'these models demonstrate their natural language abilities'. These cases were not searched for automatically, as it would require an additional coreference resolution step, which in itself is an elaborate task that lies beyond the scope of this research. However, during the manual revision, some such cases were encountered and added to the list manually[8]. The first two cases were parsed by searching for a noun phrase containing an AI keyword, immediately followed by the possessive clitic, or immediately preceded by the preposition *of*. The noun phrase being possessed was not limited to certain anthropomorphic terms, and yet positive samples were easily encountered — which is quite telling on the ubiquity of the tendency to literally attribute human-like properties to AI systems. A full list of these noun phrases is available in Appendix B.2.

Comparisons between AI entities and humans

To find sentences embodying comparisons between AI systems and humans, I searched for sentences containing a noun chunk partially matching any one of the AI keywords, followed by a comparison to humans. The comparison was found by looking for the following comparative phrases: *like*, *compared to*, *similar to*, *similarly to*, *resemble*, *mimic*, *better than*, and *as*. I then checked whether the following

[8]One caveat about this approach is that it was done manually and therefore not automatically reproducible in terms of the methodology presented in this paper. However, the objective of this research was to challenge existing methods for anthropomorphic language detection, and including a variety of structures adds to the challenge. In future work, this issue could be addressed by employing coreference resolution tools to include a larger and more significant sample of these cases in a more robust evaluation set.

human entities immediately followed: *human*, *human being*, *person*, *people*, *humanity*, *mankind*, *child*. I also looked independently for *human-like*, *humanlike* and *childlike*. These were also relatively easy to come by, suggesting that AI systems are often being equated to or contrasted with humans in AI research.

4.2 Annotation Procedure

4.2.1 Annotation guidelines

The annotation was performed on the sentence level by a single annotator, in the following general scheme: given a sentence and a previously known AI entity, the sentence is labeled 'positive' if the AI entity is anthropomorphic in the context of the sentence. To determine this, annotation guidelines were set, modeled in part after the VU Metaphor Identification Procedure (Steen et al., 2010). First and foremost, two modes of anthropomorphism were distinguished: the first is overt and explicit anthropomorphism that is evident in the contents of the sentence, and the second is implied and implicit anthropomorphism that arises from form. The second cases are naturally more ambiguous and vague, and different contexts can influence the degree of anthropomorphism. For instance, certain verbs that have to do with information processing in humans, like *train* and *learn*, have become nomenclature in the context of AI. Similarly, due to the conversational design of chatbots, words related to conversation and communication like *ask* and *tell* are becoming more and more widespread as synonyms for prompting a conversational AI model. Adjectives like *vulnerable* and *sensitive* are used ubiquitously in the context of system security, although their primary sense is understood in terms of physical or emotional experience. Still, these terms have permeated into AI discourse precisely *because* of their anthropomorphic properties. The Overton Window of AI, so to speak, is constantly shifting, and more and more anthropomorphic terms are becoming acceptable in the context of AI[9]. As these terms assume a new, AI-contextual meaning, they may be perceived as less anthropomorphizing. The objective in defining the ensuing annotation guidelines was to try to discern, as much as possible, when an AI entity is being framed as having human-like characteristics.

Annotation Procedure

The annotation procedure is defined as follows: first, check whether a sentence is explicitly anthropomorphizing. That is, whether it directly ascribes human-like capacities such as cognition, mental states, intention, or agency to an AI entity. If so, it is labeled as a positive instance (see Table 4.3 for examples). If not, check whether the sentence contains an anthropomorphic word. These words are verbs, nouns or adjectives related to a certain human-like attribute, like *thinking*, *understanding*, or *awareness*. The taxonomy provides a preliminary gloss of such words. If such a word exists in the sentence, proceed to check whether the primary sense used in the sentence is non-ambiguously anthropomorphic, whether there is a more salient non-anthropomorphic meaning in context, or whether the word is used ambiguously in the sentence. The first case, primarily anthropomorphic senses, is labeled 'positive' (Table 4.4). The second case refers to instances of reporting verbs like *explain* or *demonstrate*, or erroneous or imprecise uses, and is labeled 'negative' (Table 4.5).

The third case refers specifically to AI-specific terminology, which in context could be construed

[9]See Mackinac Center For Public Policy (no date).

as either anthropomorphic due to the original meaning, or non-anthropomorphic due to its increased application in the context of machines and AI. Such instances include verbs like *train* or *learn*, or adjectives like *smart*, which are often used to refer to the operation and functionality of AI systems. These cases are considered ambiguous because it is not always clear in what way the author intended to use an anthropomorphic word; for instance, *learn* can be used in the context of machine learning as a description of the underlying computational process, or it could be used to say that an actual cognitive process it taking place. In these cases we turn to frame semantics (Fillmore & Baker, 2012), and check how the AI entity is framed in the broader context of the sentence. In frame semantics, meaning is understood through a semantic frame which is evoked by certain lexical units, and involves frame elements constituting an event, relation or entity and its participants. For example, the frame evoked by the lexical unit *Awareness* involves the frame element of *Cognizer*, who is the person whose awareness of phenomena is at question[10]. These frames are accessed through FrameNet, a large database of annotated sentences linked to more than 1,200 semantic frames (Ruppenhofer et al., 2006). In the context of this work, an AI entity can be seen as anthropomorphized in a sentence if it is framed as a *Cognizer*, *Perciever*, or *Experiencer*, which all presuppose sentience.

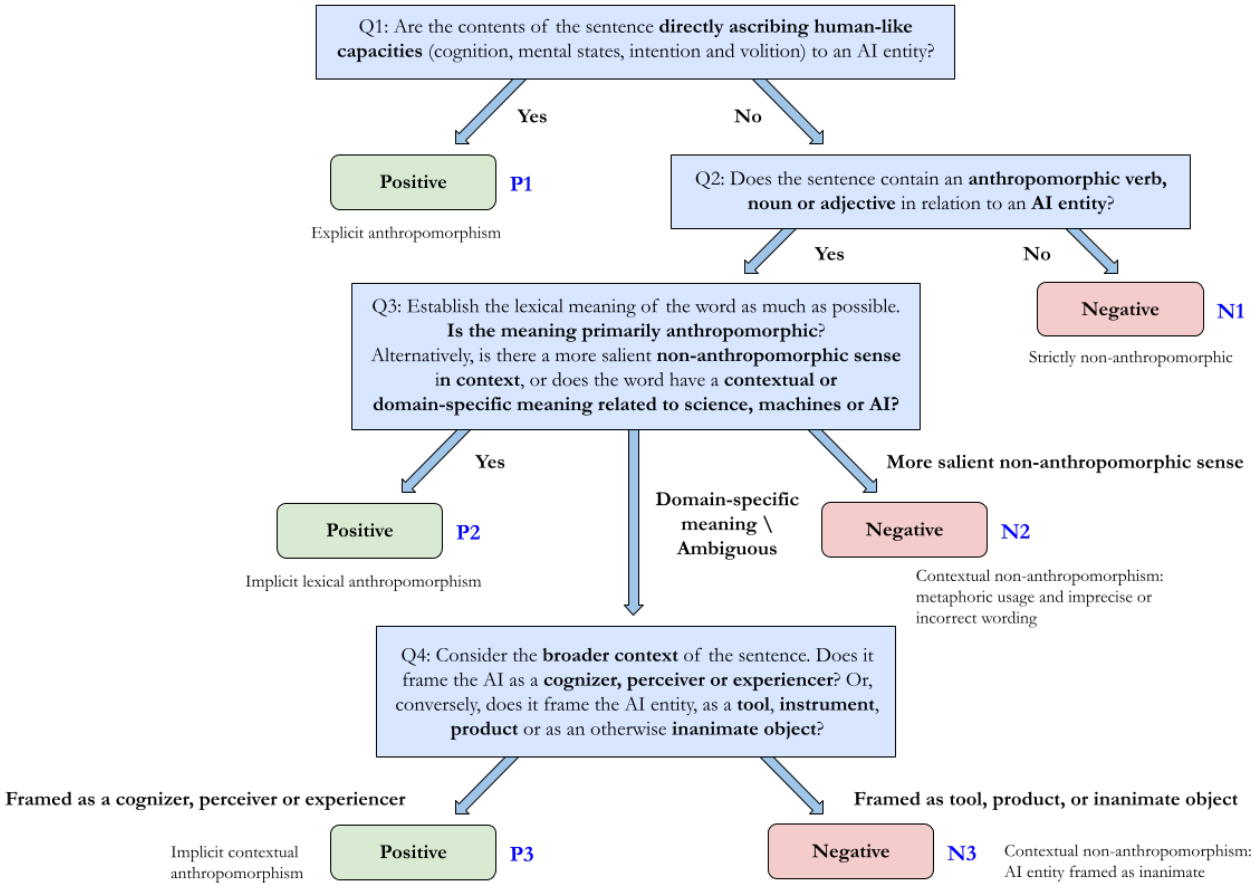

Figure 4.1: A decision tree representation of the annotation procedure.

Importantly, frame elements are also determined with respect to specific senses. That is, there is no FrameNet definition for *learning* in its *machine learning* sense. The idea is not to follow the

[10]A similar frame-semantic approach to analyzing anthropomorphic descriptions in texts is carried out by Ryazanov et al. (2024).

FrameNet annotations blindly, but to use the frame-semantics approach as inspiration for a manual disambiguation approach. I do not always determine an AI entity as anthropomorphic if the *Cognizer* attribute is found in the corresponding frame — I use this as a reference, but ultimately provide a human judgment based on the information given in the sentence, and world-knowledge about AI and its terminology. In terms of the annotation procedure, the overall context of the sentence is considered: a sentence is labeled as non-anthropomorphic if the context frames the AI entity as an object, instrument, or tool, or otherwise clarifies its non-human status. For example, such cases will feature an AI entity as an object of a dehumanizing verb such as *develop* or *program*. Conversely, if the sentence provides further context that contributes to the attribution of human-like capacities, it is labeled as anthropomorphic (Table 4.6). If at any step there is unresolved ambiguity or vagueness, a neutral or uninformative context, or simultaneous framing of the AI entity as a tool and as a *Cognizer* — the sentence is labeled as inconclusive. The inconclusive cases were compiled into a separate list and also used in the evaluation process.

The complete annotation guidelines, including detailed notes for each step of the procedure, are given in Appendix C. Note that while the decision tree differentiates types of positive and negative annotations, these are not considered in the final analysis. They are brought here to elucidate the levels of anthropomorphism: from explicit and overt descriptions, to implicit anthropomorphism brought on by the meaning of the lexical units, to implicit contextual anthropomorphism which is the result of the framing in the sentence and heavily relies on contextual information. Similarly, negative sentences could be either strictly or contextually non-anthropomorphic. In terms of the linguistic categories, there are positive, negative and inconclusive cases of sentences containing anthropomorphic verbs and adjectives. Sentences containing noun phrases were understood as explicitly anthropomorphic — this is because they either explicitly frame an AI entity as have a human role or function, or explicitly describe it as possessing human-like qualities. In that sense, the answer in the very first step of the annotation procedure is always positive, and there is no ambiguity to consider[11]. Sentences featuring comparisons of AI systems to human beings were taken as always inconclusive. This is because a comparison could be understood in two opposing manners: one the one hand, it could be seen as likening and equating AI systems to humans, describing them as having identical or similar characteristics and properties, contributing to their anthropomorphizing. On the other hand, a comparison could be seen as a contrast or juxtaposition of two things that are ontologically and conceptually distinct, save for a few common shared properties (this is described in detail in section 3.3.4). Ultimately, I decided to include these cases in the evaluation with an 'inconclusive' label, solely for the purpose of obtaining insights about the existing approaches, by learning how they would label such cases. The total distribution of positive, negative and inconclusive cases for each linguistic category is displayed in Table 4.2.

[11]It should be noted that some samples of role or function noun phrases might be seen as framing an AI entity as a tool or instrument. Even so, I have decided that the explicit anthropomorphic assignment or attribution of a human role, function, or property outweighed any implicit semantic framing in the sentence.

Structure	Positive	Negative	Inconclusive	**Frequency**
verb subjects	56	61	33	**150**
verb objects	58	65	27	**150**
adjectives	52	47	21	**120**
role/function NPs	70	0	0	**70**
genitive NPs	60	0	0	**60**
comparisons	0	0	50	**50**
Total	296	173	131	**600**

Table 4.2: Frequency of each linguistic category per class.

4.2.2 Annotation examples

Below are examples of annotated sentences for each step in the procedure. In particular, I demonstrate explicitly anthropomorphic sentences, sentences that contain implicit but non-ambiguous anthropomorphism, sentences that have a more salient non-anthropomorphic reading, and a side-by-side comparison of ambiguous cases resolved in terms of the semantic framing. Some inconclusive cases are also included.

Category	**Sentence**	**Label**
adjectives	**Conscious** AI systems would arguably deserve moral consideration, and it may be the case that large numbers of conscious systems could be created and caused to suffer.	Positive
role/function NPs	**Medical AI assistants** support doctors in disease diagnosis, medical image analysis, and report generation.	Positive
genitive NPs	We also discuss issues that are unique to edge applications such as protecting a model's **intellectual property** and verifying its integrity.	Positive
genitive NPs	We show how progress has been made in benchmark development to measure **understanding capabilities** of AI methods and we review as well how current methods develop understanding capabilities.	Positive

Table 4.3: Examples of explicit anthropomorphic expressions that were labeled as positive.

The cases in Table 4.3 portray explicit anthropomorphism, as AI systems are being described as having consciousness, understanding capabilities, intellectual property, and the ability to function as medical assistants in a professional setting.

The cases in Table 4.4 are examples of implicit, non-ambiguous anthropomorphism. It is implicit as the contents of the sentences do not explicitly attribute human-like abilities, rather focus on some other thing, such as model security, functionality, or training process. Rather, this is implied in the choice of lexical units. The verbs and adjectives used do not have an AI-specific meaning, nor do they have a more salient non-anthropomorphic as in the case of reporting verbs. The adjectives *confident* and *aware* are used in their primary sense; as are the verbs *infer*, *memorize* and *communicate with*. The claim is that there is no framing in which these lexical units could be construed as non-anthropomorphic, hence these sentences are considered non-ambiguously anthropomorphic.

Conversely, the cases in Table 4.5 represent sentences that have a more salient, non-anthropomorphic sense of a word that otherwise might be construed as anthropomorphic: we have the reporting verbs *describe* and *explain* which in context are clearly used to emphasize the descriptive and explanatory

Category	Sentence	Label
adjectives	In addition, we found that the model becomes more **confident** and refuses to provide an answer in only few cases.	Positive
adjectives	Results suggest that ChatGPT is **aware** of potential vulnerabilities, but nonetheless often generates source code that are [sic] not robust to certain attacks.	Positive
verb objects	Our approach sheds light on developing more user-friendly recommender systems, in which users can freely **communicate with** the system and obtain more accurate recommendations via natural language instructions.	Positive
verb subjects	However, it is unclear if this success is limited to explicitly-mentioned causal facts in the pretraining data which the model can **memorize**.	Positive
verb subjects	When performing next word prediction given a textual context, an LM can **infer** and represent properties of an agent likely to have produced that context.	Positive

Table 4.4: Examples of sentences with implicit non-ambiguous anthropomorphism that were labeled as positive.

powers of a model. there are also the verbs *claim*, *see*, and *establish oneself* which are used colloquially or metaphorically. The meaning is clear, there is no alternative reading in which the AI entity is understood as performing an act of claiming, seeing or establishing itself in any real way. The latter case is particularly important to address, as language is and should be dynamic and flexible[12]; metaphors are a powerful tool for facilitating comprehension.

Category	Sentence	Label
verb objects	Considering the natural ventilation, the thermal behavior of buildings can be **described** by a linear time varying model.	Negative
verb objects	We find that the observed behaviour is **explained** by a model including the effects associated with the variations of pressure and density.	Negative
verb subjects	Progress in AI is often demonstrated by new models **claiming** improved performance on tasks measuring model capabilities.	Negative
verb subjects	AI systems **have seen** significant adoption in various domains.	Negative
verb subjects	The OCC model has **established itself** as the standard model for emotion synthesis.	Negative

Table 4.5: Examples of sentences with a more salient non-anthropomorphic sense that were labeled as negative.

The examples in Table 4.6 demonstrate how the same lexical unit can be construed as anthropomorphic in some cases, and non-anthropomorphic in others, depending on the context. For example, the verb *learn* is used ubiquitously in the context of AI to describe the computational process of process of detecting patterns in data and generalizing from those patterns. However, occassionally it is used in its human, pedagogical sense; this is evident in the example which describes large language models as collaboratively learning from each other by exchanging natural language. Similarly, the adjective *intelligent* is a commonly used term in the context of computer science and machines, but it might also be used in its traditional, cognitive sense to ascribe cognitive and intellectual abilities. In the example, the first instance talks about AI systems in terms of cognition and behavior, framing them as human-like, whereas the second instance mentions the robotic system as a product of development, framing it as a tool.

Table 4.7 showcases examples of inconclusive sentences. The first two examples demonstrate insufficient context. The verbs *acquire* (in the sense of knowledge acquisition) and *train* both have an

[12]This is a normative claim about language policing that lies beyond the scope of this paper.

Category	Sentence	Label
verb subjects	We then propose a system that leverages the recently introduced social learning paradigm in which LLMs **collaboratively learn from each other** by exchanging natural language.	Positive
verb objects	Through examining the weights in the trained DCNN model, we are able to quantitatively measure the visual similarities between architects that are implicitly **learned** by our model.	Negative
adjectives	As AI systems become more **intelligent** and their behavior becomes more challenging to assess, they may learn to game the flaws of human feedback instead of genuinely striving to follow instructions	Positive
adjectives	In this work, we describe our approach to developing an **intelligent** and robust social robotic system for the Nadine social robot platform.	Negative

Table 4.6: Examples of sentences with an ambiguous term, in which the broader context determines the framing of the AI entity as either anthropomorphic or non-anthropomorphic.

anthropomorphic sense, but are used ubiquitously in the context of AI — especially *train*. The context itself did not provide additional semantic framing as a *Cognizer* nor as a tool. This is best demonstrated in the fact that the AI entity could be replaced with e.g. 'student' without resulting in an infelicitous utterance. The third example mentions the *human-like behavior* of conversational AI systems, which could either be seen as anthropomorphizing due to the direct comparison, or non-anthropomorphizing due to the explicit modifier *human-like*. As mentioned, all comparisons were taken as inconclusive cases.

Category	Sentence	Label
verb subjects	Our results show that not all extracted utterances are correctly structured, indicating that either LLMs do not fully **acquire** syntactic-semantic rules or they acquire them but cannot apply them effectively.	Inconclusive
verb objects	How should we **train** a language model in this scenario?	Inconclusive
adjectives	[...]commonly used sources such as Wikipedia may not be best suited to build **culturally aware** LMs, if used as they are without adjustment.	Inconclusive
comparisons	Conversational AI systems exhibit a level of **human-like behavior** that promises to have profound impacts on many aspects of daily life — how people access information, create content, and seek social support.	Inconclusive

Table 4.7: Examples of sentences that were deemed inconclusive cases, due to a neutral or uninformative context or conflicting semantic framing.

4.2.3 Inter-annotator agreement

Inter-annotator agreement was performed on a subset of the multiclass cases, viz. the sets of verb subjects and objects and adjectives which had positive, negative and inconclusive instances. Role or function NPs, genitive NPs and comparisons were excluded from the inter-annotator agreement[13]. A total of 84 cases, constituting 20% of the 420 multiclass cases, was divided between two second annotators. The annotators received instructions (described in Appendix C) alongside a spreadsheet containing sentences with the AI entity highlighted in bold, and were requested to provide a single score of 1 for positive, 0 for negative and 2 for inconclusive per sentence. The agreement was measured

[13]To transform this challenge set into a proper benchmark for anthropomorphic language detection, it should undergo robust annotation by at least two annotators. This is discussed in section 6.2.

for each set individually in terms of Cohen's Kappa, which measures agreement between two raters on a multiclass dataset (Cohen, 1960). The annotations were not revised after checking inter-annotator agreement.

The first set, which had a balanced distribution of positive, negative and inconclusive cases resulted in a Cohen's κ of 0.39 for all cases, and a much higher κ of 0.92 for just the positive and negative cases. The second set, which consisted of twice as many inconclusive cases than positive and negative cases resulted in a Cohen's κ of 0.22 for all cases, and 0.60 on just the positive and negative cases. This was checked by first including all cases, and then filtering out cases in which at least one of the annotators was inconclusive. That is, agreement was checked only for the cases in which both annotators were certain of an either positive or negative score. I also calculated Cohen's κ for each class by creating a binary mapping that maps all agreements to 1 and all disagreements to 0 given a single class, and then calculated the agreement per class. The result was a Cohen's κ of $= 0.62$ for positive cases and $\kappa = 0.49$ for negative cases in the first set, and $\kappa = 0.40$ for positive cases and $\kappa = 0.22$ for negative cases in the second set. Inconclusive cases had a very low agreement rate due to their borderline nature, but this was expected.

Sentence	**Rater 1**	**Rater 2**
Since a fallacious procedure is generally considered fake and thus harmless by **LLMs**, it helps bypass the safeguard mechanism.	Positive	Negative
However, existing studies into language modelling with BERT have been mostly limited to English-language material and do not pay enough attention to the implicit knowledge of language, such as semantic roles, presupposition and negations, that can be acquired by **the model** during training.	Negative	Positive
In the second experiment, we ask **the LLMs** to numerically rate various aspects of the musical cultures of different countries.	Positive	Negative
ChatGPT has learned language patterns and styles from a large dataset of internet texts, allowing it to provide answers that reflect common opinions, ideas, and language patterns found in the community.	Negative	Positive
However, it is unclear if this success is limited to explicitly-mentioned causal facts in the pretraining data which **the model** can memorize.	Positive	Negative

Table 4.8: Sentences for which the first and second annotator did not agree, and both provided a conclusive score. The first example comes from the first inter-annotated set, and the other four come from the second. The AI entity (and not the anthropomorphic unit) is highlighted in bold, following the format that the annotators received, and brought here to clarify that the disagreement did not stem from confusion regarding the entity in question.

Out of all disagreements, only five were cases in which both annotators provided a conclusive (i.e. positive or negative) score (Table 4.8). The rest constituted cases in which at least one annotator was inconclusive about the degree of anthropomorphism. Indeed, the task of determining anthropomorphism is highly subjective (Shardlow et al., 2025). The examples in Table 4.8 demonstrate how different contexts could be seen as anthropomorphizing to varying degrees by different annotators. For instance, the first rater regarded the use of certain verbs, e.g. *consider*, *memorize*, and *ask* to be anthropomorphizing, whereas the second rater deemed positive the cases that mentioned human-like aspects of language learning. This task in not only subjective, it is also highly depending on context (Cheng et al., 2024), which is not always easy to determine on the basis of a single sentence. The low Kappa score for the overall cases reflects the difficult nature of this annotation task, especially on borderline cases which do not have enough contextual cues, even for human annotators, to determine

whether or not an entity is being anthropomorphized. Additionally, while we relied on a taxonomy of anthropomorphic language, deciding whether a certain lexical unit embodies these definitions is not a trivial task. Nevertheless, the Kappa score of our non-borderline cases shows that these were for the most part agreed upon. Shardlow et al. (2025) describe their annotation process for a large corpus of annotated sentences representing anthropomorphic descriptions of AI technologies. They do not calculate a Kappa score at all, mentioning that they do not expect agreement on a subjective task such as anthropomorphic language identification. Instead, they controlled for agreement through regular meetings, aligning definitions and understanding across multiple annotators. In a sense, the work presented in this paper was performed in a similar manner, as the product of a supervised thesis.

Chapter 5

Anthropomorphic Language Detection Evaluation

The goal of this thesis is to shed light on the current definitions and interpretations of anthropomorphic language in AI research, and the means for identifying them in text. To that end, I have developed a linguistically grounded challenge set aimed at evaluating and challenging the state-of-the-art approaches for anthropomorphic language detection, presented in chapter 4. There are currently only two open-source implementations of anthropomorphic language detection in the context of machines and AI, which were briefly discussed in section 2.3[1]. The evaluation aims to challenge these approaches, by introducing a variety of syntactic structures and probing their capacity to handle different modes of expressing anthropomorphism. I am interested in quantifying and explaining the differences in the measured degree of anthropomorphism among the different linguistic categories. Through the process of constructing the challenge set and the annotation guidelines, it became evident that different linguistic patterns of anthropomorphism pose different conceptual challenges of identification. The evaluation aims also at demonstrating how these differences come into play in the automation and operationalization of anthropomorphic language detection. Through this probing, I hope to gain insights on how to handle these syntactical structures in future implementations.

This chapter first presents the conceptual and technical framework of each approach. Then, the experimental setup of our evaluation are introduced, followed by the results of the evaluation and an ensuing discussion.

5.1 SoTA: *Living Machines* (2020) and *AnthroScore* (2024)

I evaluate and compare the two state-of-the-art approaches for detecting anthropomorphizing language in texts about machines and technologies. The first approach — *Living Machines: A study of atypical animacy* — offers a method for atypical animacy detection, which examines scenarios in which typically inanimate entities, specifically machines, are portrayed as animate (Coll Ardanuy et al., 2020). The second approach — *AnthroScore: A Computational Linguistic Measure of Anthropomorphism* — presents an automatic metric of implicit anthropomorphism in language, focusing on AI technologies (Cheng et al., 2024). Below is an overview of the conceptual and technical frameworks presented in each one of the approaches we evaluate. This overview demonstrates why the two approaches are comparable,

[1] In July 2025, a third approach to anthropomorphic language detection was published (Shardlow et al., 2025). However, the evaluation presented in this paper preceded its publication. Furthermore, this implementation is not open-source.

despite making use of different terminology. In chapter 3 we have established a theoretical baseline that conceptually links animacy and anthropomorphism. This link is fortified by the conceptual and technical similarities employed by these two approaches, and provides justification for a comparison of their performance on the same challenge set.

5.1.1 Conceptual framework

The notion of animacy in *Living Machines*, which guided their annotations, is defined as

Definition 1. "[descriptions of a machine] as having traits distinctive of biologically animate beings or human-specific skills, or portrayed as having feelings, emotions, or a soul" (Coll Ardanuy et al., 2020, p. 4539).

In *AnthroScore*, the authors define anthropomorphism as

Definition 2. "the attribution of distinctively human-like feelings, mental states, and behavioral characteristics to non-human entities" (Cheng et al., 2024, p. 808).

In particular, they define these characteristics as

Definition 3. "the ability to (1) experience emotion and feel pain (affective mental states), (2) act and produce an effect on their environment (behavioral potential), and (3) think and hold beliefs (cognitive mental states)" (Cheng et al., 2024, p. 808).

These definitions refer to feelings and emotion, as well as human-specific skills and distinctive behavioral characteristics that are associated with humans. The term *soul* used in *Living Machines*, resonates with the definition of mental states provided by *AnthroScore*. While the *Living Machines* analysis distinguishes between animacy and a more specific notion of *humanness*, this distinction serves its own purpose. It is meant to shed light on the inverse phenomenon of anthropomorphism, namely *dehumanization* (Epley et al., 2007), by distinguishing between expressions in which machines are talked about as having human-like capabilities and expressions likening humans to machines. In their annotation scheme, the former are construed as examples of both animacy and humanness, whereas the latter express only animacy. Specifically, an additional annotation of humanness would be true if a machine is portrayed as sentient and capable of specifically human emotions, and false if it is used to suggest some degree of dehumanization (Coll Ardanuy et al., 2020, p. 4539). In this definition, animacy encompasses the notion of humanness, i.e. every instance of humanness is also an instance of animacy (but not vice versa). For the purpose of this evaluation, this distinction is not relevant; the challenge set contains only sentences in which something is said about an AI entity — the phenomenon of dehumanizing language which focuses on humans as subjects is not addressed here. In effect, the challenge set does not include any sentences in which animacy is attributed and humanness is not. Therefore, no methodological issue arises from equating the more general definition of animacy used in *Living Machines* to the definition employed in *AnthroScore*, as both approaches should construe the sentences in the challenge set in the same way. Below, we shall see that the way that the two approaches operationalize these definitions is also rooted in similar linguistic and semantic features.

5.1.2 Technical framework

AtypicalAnimacy is based on the the HuggingFace implementation of BERT (`BERT-base`, 110M parameters)[2], fine-tuned on an atypical animacy detection dataset consisting of 19th century texts on machines in the industrial age. Their methodology relies on the premise that a contextualized language model will predict tokens corresponding to animate entities given a context which requires or hints at an animate entity. Given a sentence in which an entity has been masked with a `[MASK]` token, the animacy of the masked entity is calculated by averaging the animacy of the top κ tokens predicted to replace the mask. The animacy of these tokens is determined using WordNet[3] hypernymy relations. The process is as follows: given a masked sentence, the masked language model outputs κ tokens and their probability scores. Each token is disambiguated to its most relevant WordNet sense, and then checked to see if its sense is a descendant of the *living thing* node, the common ancestor of animate classes. The predicted token receives an animacy value of 1 if it belongs to the tree of *living thing*, and 0 otherwise. The animacy of a masked token is calculated by averaging the animacy values of each predicted token, weighted by their probability score. A single animacy score between 0 and 1 is produced. The final classification is determined based on a classification threshold τ. Both the threshold τ and the cutoff of κ are achieved through experimentation and optimization (Coll Ardanuy et al., 2020, pp. 4536-4537).

AnthroScore provides a quantitative measure of anthropomorphism in text which is based on masked language models. Their method uses the HuggingFace implementation of RoBERTa (`roberta-base`, 125M parameters), a pre-trained masked language model (MLM) as the model and tokenizer[4]. Their methodology similarly involves a masked language model. Given a sentence or a set of sentences, and a corresponding entity or set of entities, the AnthroScore model computes the probability of that entity to be construed as human by an MLM, relying on animate and inanimate pronouns as a cue. First, the entity is replaced with a `[MASK]` string. Then, an MLM computes the probability that the mask would be replaced with either animate pronouns, e.g. *he* or *she*, or inanimate pronouns, e.g. *it*. The score, $A(s_x)$ for a given sentence s and entity x is defined as the log ratio between the two probabilities (Cheng et al., 2024, pp. 809-810). A further score can be obtained for a set of sentences, defined as the mean value of all individual scores across the set. A sentence can have multiple scores, depending on the number of entities that are identified and masked. For the purpose of this experiment, we are only interested in the individual scores that are computed on the basis of a specific, predetermined AI entity which we identified during annotation. The final scores are determined as follows: a high-anthropomorphism score corresponds to a score greater than 1, which reflects the fact that according to the MLM, the entity is e^A more likely to be framed as human than as non-human (this is based on the natural log function, which is in base e). A low-anthropomorphism score symmetrically corresponds to the inverse scenario, which results in $A(s_x) < -1$. When the probabilities are equal, the ratio will be 0, making the log ratio 1. AnthroScore defines scores between 1 and -1 as a reflection of a near equal probability to be implicitly framed as human and non-human.

[2]https://huggingface.co/docs/transformers/en/model_doc/bert
[3]https://wordnet.princeton.edu
[4]https://huggingface.co/docs/transformers/en/model_doc/roberta

5.2 Experimental Setup

The above implementations are available on GitHub. *Living Machines* (Coll Ardanuy et al., 2020) provide a complete pipeline[5] which allows to replicate their experiment, and deploy their masking approach as well as the baseline classifiers through the execution of Jupyter notebooks[6]. Since the authors provide both a dataset and an unsupervised pipeline, from here on I use *AtypicalAnimacy* to refer to the code implementation, and *Living Machines* to the entire project presented in the paper. Cheng et al. (2024) also provide access to their implementation[7], which is designed as one parameterized python script that outputs a score given a sentence or a set of sentences, and a set of entities. The repositories were cloned and accessed locally.

For AnthroScore, no particular configuration was required. For AtypicalAnimacy, the specific scenario was configured to align with the details of this evaluation, as well as the optimal configuration as reported by Coll Ardanuy et al. in their evaluation. First, I selected the training corpus as the 19th Century Machines corpus with animacy annotations, since we are interested in the framing of machines as humans. For the context used for classification, I selected the sentence plus an additional context of the previous and next sentences, a feature which was reported to improve the results in the masking approach. I opted for the contemporary BERT base uncased tokenizer and model[8]. For the animacy detection, I selected a weighted average of the probability score of the animacy values of predicted tokens (rather than non-weighted average). I also opted for word-sense disambiguation on the predicted tokens to match the most relevant sense in WordNet. These last two parameters are configurable in their setup, but Coll Ardanuy et al. mention them as part of their experiment.

Two experiments were performed during the evaluation. The first experiment relied on the AnthroScore masking strategy, which is integrated in their pipeline. The second experiment relied on an ad-hoc masking strategy, which is described in detail in section 5.2.1 below. While the AnthroScore and AtypicalAnimacy models were deployed in their own environments, all auxiliary scripts used in the evaluation task are accessible on GitHub[9]. The evaluation compares the performance of each method using the two masking strategies. The evaluation metrics, as well as a mapping of AnthroScore and AtypicalAnimacy scores to a single, uniform labeling scheme is described in section 5.2.2 below. The experiments and results are discussed in detail in section 5.3.

5.2.1 Masking strategy

The AnthroScore method has its own built-in masking strategy, which relies on keyword identification and noun-chunking, and works as follows: an input sentence is parsed into noun chunks with the spaCy, and then the input entities are searched within the noun chunks of the sentence. If an entity is found within a noun chunk, that noun chunk is replaced with a `[MASK]` string. This method is suitable for certain structures, particularly those in which the anthropomorphic component complements or predicates the AI entity. However, when the anthropomorphic component modifies the AI entity or exists as part of a genitive construct, this method will mask together with the entity. For example, the phrase 'conscious AI systems' will be masked in its entirety by AnthroScore's masking strategy,

[5] https://github.com/Living-with-machines/AtypicalAnimacy

[6] https://jupyter.org

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

[8] https://github.com/google-research/bert, since AnthroSet consists of contemporary data. *Living Machines* fine-tune BERT-base on 19th century data, but this was not relevant for our experiment.

[9] https://github.com/doriellel/mol-thesis

even though the adjectival modifier *conscious* contributes to the anthropomorphism in the sentence. Similarly, expressions such as 'the model's ability to reason' will be parsed as a single noun chunk, and then masked completely, eliminating the anthropomorphic component from the context.

To mitigate this, an ad-hoc masking strategy was developed, that preserves the anthropomorphic cues in the context rather than mask them. The idea behind this strategy is to mask the minimal phrase in the sentence that corresponds to an AI entity, leaving out any additional modifiers, possessives, determiners, numerals and quantifiers, or any other linguistic information that is not part of the AI entity. This approach is consistent with the one taken by Coll Ardanuy et al. (2020) in their masking implementation. The ad-hoc masking strategy followed these steps: given an AI keyword (such as *AI*, *LLM*, *model*, or *ChatGPT*), I manually masked the minimal phrase referring to an AI entity, masking additional modifiers only in case they were part of the name, or an essential part of its description, i.e. relating to design, functionality or purpose. Such descriptions include *large* in *large language model*, *conversational* in *conversational AI*, or *question answering* in *question answering system*. I left out determiners like *a*, *the*, or *this*, numerals or quantifiers such as *one*, *several*, or *many*, and any descriptors that are contingent to the description, such as *powerful*, *complex*, or *flexible*.

Masking	**Masked Sentence**	**Masked Entity**
AnthroScore	In this work, we survey, classify and analyze a number of circumstances, which might lead to arrival of [MASK].	malicious AI
Minimal entity	In this work, we survey, classify and analyze a number of circumstances, which might lead to arrival of malicious [MASK].	AI

Table 5.1: Two masked instances of the same sentence: one with the AnthroScore masking strategy and one with the minimal entity masking strategy.

The main advantages of this masking strategy are best understood in the context of the experiments, and brought in detail in the subsequent discussion (section 5.4). In short, while it had no negative effect on predicative structures involving verbs or complements, it substantially improved the detection of anthropomorphism in possessive structures, or structures containing noun or adjectival modifiers. This masking strategy also has several disadvantages. For one, it is manual, and thus not practical for reproducing this experiment on a large scale. This is something to consider, but in the meantime I employ this manual approach to highlight the disadvantages of an automatic masking method that relies on chunking (this is discussed in detail in section 5.4). Second, it is not always clear what constitutes an essential description. For instance, is the word *smart* in *smart home system* part of the name? Since there were not many such occurrences, this was resolved on a case-by-case basis, by searching for suspect phrases in Google Scholar. For our purposes, this was enough to gauge whether a term is frequently used, or coined by the authors of the paper from which it was taken. Future implementations of this masking approach might benefit from integrating a Named Entity Recognition pipeline that can identify named entities in the domain of AI technologies. A third issue that arose from employing this masking strategy had to do with certain masks that resulted in expressions which are not compatible with pronominalization, or even with the replacement with other nouns. However, this is not an issue with this minimal entity masking strategy per se, as it is with employing a masking approach for this task. This is also discussed in detail in the discussion below.

5.2.2 Evaluation metrics and mapping

As described in chapter 4, the challenge set is divided into multiclass and single class sets. The multiclass sets are the *verb subjects*, *verb objects* and *adjectives* sets, which have positive, negative and inconclusive cases. The single class sets are the *role or function noun phrases* and *genitive noun phrases* sets, which are always positive, and the *comparisons* set, which is always inconclusive. The multiclass sets are evaluated in terms of macro-averaged precision, recall, and F1-scores. These are observed both as macro-averaged aggregates for each syntactic category, as well as per class. The single class sets are evaluated in terms of accuracy only. Using these metrics, the performance of each approach using each masking strategy is compared. An additional comparison is included, which is not intended as an evaluation but rather as an observation, which looks at the trends in predicting positive, negative (and, in the case of AnthroScore, inconclusive) labels for borderline sentences were labeled inconclusive. As discussed in section 4.2, it has been decided to include these cases in the evaluation set not as a means to challenge or test the implementation of each approach, but to obtain a general idea of how these cases are handled by them. In particular, I am interested in whether or not the current approaches for anthropomorphic language detection considered sentences comparing AI entities to humans as anthropomorphizing.

To facilitate the comparison of the two approaches, a mapping between the AnthroScore outputs and the AtypicalAnimacy outputs was needed (Table 5.2). AnthroScore does not output a binary score, but rather continuous outputs representing high and low anthropomorphism. A high anthropomorphism score is a score greater than 1, whereas a low anthropomorphism score is a score less than -1. These are interpreted as positive and negative, respectively. Scores that fall within the interval $[-1, 1]$ represent an equal likelihood of the masked entity to be construed as human and non-human. These scores are seen as equivalent to the inconclusive class, since they do not provide a definitive answer and reflect an ambiguous or neutral context that does not contribute to the framing of the entity as either human or non-human. AtypicalAnimacy also outputs continuous scores, as well as a classification threshold τ that is obtained by maximizing F-scores on the 19th Century Machines corpus with animacy annotations. The final binary scores are determined by the classification threshold — scores equal or greater to τ are positive with respect to atypical animacy, and scores less than τ are negative.

AnthroScore	**AtypicalAnimacy**	**Final Score**
$x \in \mathbb{R} : x > 1$	$x \in \mathbb{R} : x \geq \tau$	Positive (1)
$x \in \mathbb{R} : x < -1$	$x \in \mathbb{R} : x < \tau$	Negative (0)
$x \in \mathbb{R} : -1 \leq x \leq 1$	not applicable	Inconclusive (2)

Table 5.2: Mapping AnthroScore and AtypicalAnimacy scores to a uniform anthropomorphism score. τ is an optimal classification threshold that was obtained by maximizing F-scores on the 19th century machines animacy corpus.

An important clarification is needed here: AnthroScore is not defined as a classifier, but rather a measure of anthropomorphism in text. The continuous outputs are meant to represent the degree of anthropomorphism, whereby the higher the score, the higher degree of anthropomorphism (and vice versa). However, this number is obtained by calculating the log ratio between the probability, according to the masked language model, that the mask will be replaced by an animate pronoun and the probability that it will be replaced by an inanimate pronoun. While these probabilities may reflect

patterns in data, the probabilistic measures of a masked language model do not necessarily align with human interpretations (Bender & Koller, 2020). Even among humans, what constitutes high and low anthropomorphism is not objectively measurable. Thus, AnthroScore's continuous representations of high and low anthropomorphism are converted into something more clearly defined, while also retaining its inconclusive scores and staying true to its original design.

5.3 Experiments and Results

The two approaches were evaluated twice on all six categories of syntactic structures, once for each masking strategy. The sentences were divided into batches according to the linguistic category and label (e.g. positive sentences featuring anthropomorphic verb subjects). The first experiment made use of the AnthroScore masking strategy, which is integrated in their pipeline. Alongside a file of sentences, all AI entities in the file were provided as a separate list. Since some sentences may contain multiple entities, it was expected that AnthroScore would occasionally output multiple masked sentences for a single sentence. The cases in which the AnthroScore pipeline masked the wrong entity were manually filtered out. Additionally, instances of irrelevant masks were dropped. These include the erroneous masking of other components in the sentence, but also partial or incomplete masking of the AI entity, which did not mask the very minimal expression representing the AI entity; for example, references to *AI* or *LLM* that were left in the context. Instances of over-masking — masking crucial contextual information — were kept, as long as the AI entity was masked fully, as these were most of the cases. I then used the AnthroScore masked sentences as input to the AtypicalAnimacy pipeline, which expects pre-masked sentences. In addition, we provided the previous and next sentence from the original abstract from which the sentence was taken. For the first (last) sentences of an abstract, an empty string was provided in place of the previous (next) sentence.

The second experiment relied on the *minimal entity* masking strategy. It was performed in the exact same way, with the exception that now both pipelines were provided with pre-masked sentences, including the additional context of the previous and next sentences in the case AtypicalAnimacy model. In order to use the minimally masked sentences for the AnthroScore approach, their code was modified to skip the masking step, and instead take as input our pre-masked sentences. Nothing else was modified in their implementation — the scores were calculated in the exact same way as before.

I first compared the performance of AnthroScore and AtypicalAnimacy on the multiclass sets. The gold labels were restricted to the positive or negative classes, since AtypicalAnimacy provides a binary classification. In terms of model outputs, the AnthroScore inconclusive predictions were treated as a third class, and used macro-averaged precision, recall, and F1-scores. This was done in order to capture the uncertainty in the AnthroScore model, and provide a more realistic depiction of its behavior[10]. A side-by-side comparison of the two approaches, using each masking strategy, is presented in Table 5.3.

Overall, the AtypicalAnimacy model performed better across all multiclass datasets, even with AnthroScore's masking strategy. The minimal entity masking strategy improved its performance in varying degrees across the three linguistic categories, resulting in the highest precision, recall, and F1-score among the the experiments. In the case of AnthroScore, the minimal entity masking strategy

[10]This means that the expectations, as well as AtypicalAnimacy predictions don't include inconclusive labels, while AnthroScore predictions do. As a result, macro-averaging the scores will penalize the AnthroScore model for too many inconclusive predictions. This is a desired result, as inconclusive or neutral predictions, that represent neither high-anthropomorphism nor low-anthropomorphism, are not informative and do not contribute to the task at hand.

slightly reduced the performance. The best performance for both models was obtained on structures involving verbs. The highest F1-score is obtained for the *verb objects* category in the first experiment, and for the *verb subjects* category in the second experiment.

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

Table 5.3: Macro-averaged precision, recall, and F1-scores for AnthroScore and AtypicalAnimacy across the multiclass categories: *verb subjects*, *verb objects*, and *adjective phrases*, comparing the AnthroScore masking strategy and the minimal entity masking strategy. In this comparison, inconclusive sentences were excluded from gold.

For the single class positive sets, I compared the accuracy of both methods using both masking strategies (Table 5.4)[11]. Both models exhibited low accuracy rates for the *role/function NPs* set. This time, the minimal entity masking did not improve performance, but quite the opposite — with an almost 19% decrease in accuracy for AnthroScore, and a 57.4% decrease for AtypicalAnimacy. Section 5.4 below presents some explanations for these low scores for the role or function noun phrase expressions. The case of genitive NPs is completely diametrical. Both approaches exhibited a notable improvement, with a 550% increase in accuracy in the case of AnthroScore, and 162.75% in the case of AtypicalAnimacy. This dramatic improvement, especially in AnthroScore's case, is indicative of the fundamental issue with the AnthroScore masking strategy when it comes to genitive expressions. And while the improvement is considerably higher percentage-wise, AnthroScore's accuracy rates for noun phrase expressions are extremely low to begin with. For AtypicalAnimacy, the improved accuracy rate of 0.783 is good, but not outstanding — however, it is on par with AtypicalAnimacy's accuracy in the multiclass sets (see Table 5.5), suggesting that the minimal entity masking strategy enabled the AtypicalAnimacy approach to reach its potential, at least in the case of genitive NPs.

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

Table 5.4: Accuracy scores for AnthroScore and AtypicalAnimacy for the single class positive sets: *role/function NPs* and *genitive NPs*.

The multiclass sets are overall balanced when it comes to positive and negative cases[12]. Nevertheless, the AnthroScore model is skewed towards the negative class. We can see this by comparing precision, recall, and F1-scores per class (Table 5.6). The AnthroScore method yielded perfect precision rates

[11]When there is only one class, this is equivalent to recall, i.e. number of correct predictions out of total predictions.

[12]Inconclusive cases constitute about half the positive or negative cases, but these were excluded from the evaluation that compares AnthroScore and AtypicalAnimacy.

Masking	Category	***AnthroScore*** Accuracy	***AtypicalAnimacy*** Accuracy
AnthroScore	verb subjects	0.518	0.750
	verb objects	0.573	0.803
	adjectives	0.539	0.697
Minimal entity	verb subjects	0.444	0.863
	verb objects	0.382	0.805
	adjectives	0.354	0.717

Table 5.5: Accuracy scores for AnthroScore and AtypicalAnimacy for the multiclass positive sets: *verb subjects*, *verb objects* and *adjectives*.

for the positive class in all three linguistic categories when using its own masking strategy, but this result is vacuous, as no false positives were predicted due to the fact that the AnthroScore approach rarely labels cases as positive. Similarly, recall was very high for the negative class, since most cases were predicted to be negative anyway. Its real-world ability to detect anthropomorphism on varying syntactic structures is actually rather low, reflected by its low recall rates for the positive class in all three sets and in both experiments. Compared to AnthroScore, AtypicalAnimacy's precision and recall are noticeably more balanced, indicating more nuanced detection capabilities.

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

Table 5.6: Precision, recall, and F1-scores per class for AnthroScore and AtypicalAnimacy using both masking strategies across three categories of anthropomorphic structures: *verb subjects*, *verb objects*, and *adjectives*.

Recall that these rates are the result of excluding the inconclusive cases from the challenge set. To maintain a fair evaluation of AnthroScore, an additional evaluation of AnthroScore on the multiclass sets which includes all classes in gold in presented (Table 5.7). There is an improvement in AnthroScore's performance when inconclusive cases are accounted for: the F1-score increased on all categories in both experiments. As seen in the first comparison, the AnthroScore approach performed better using its own integrated masking strategy. Still, the improved scores still do not surpass those of AtypicalAnimacy approach. A full comparison which includes the individual precision, recall, and F1-scores per class is given in Table D.1 of the Appendix.

Finally, the prediction trends of both methods on all inconclusive cases arr included (Table 5.8). The proportion of positive predictions among all inconclusive cases is calculated per linguistic category. This measure is meant to address the sixth and final linguistic category — the set of comparisons, which were taken as always inconclusive (see section 4.2). Overall, AnthroScore is unlikely to provide a positive

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

Table 5.7: Side-by-side comparison of AnthroScore's performance on the multiclass sets in terms of macro-averaged precision, recall, and F1-scores, comparing only positive and negative gold labels, versus positive, negative and inconclusive cases in gold, using both masking strategies.

score (i.e. high-anthropomorphism in their definition) to inconclusive cases. AnthroScore averaged 0.06 positive scores using their masking strategy, and 0.12 with the minimal entity masking. AtypicalAnimacy was more likely to provide a positive score, but not overwhelmingly so. AtypicalAnimacy averaged 0.419 positive predictions using AnthroScore's masking strategy, and 0.480 using the minimal entity masking. Interestingly, AtypicalAnimacy's tendency to output positive scores about half the time is aptly consistent with the definition used for inconclusive cases, which is in alignment with AnthroScore's definition: these are cases which cannot be determined on context alone, or have conflicting contexts, such that when masking the AI entity, it is equally likely to be construed as human and non-human by the masked language model.

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

Table 5.8: Comparison of AnthroScore and AtypicalAnimacy in terms of the proportion of positive predictions (label 1) among the inconclusive cases. AnthroScore has three types of output: high-anthropomorphism (> 1), low-anthropomorphism (< 1), and an inconclusive equivalent (between -1 and 1). AtypicalAnimacy has only two outputs: positive (above threshold) and negative (below threshold).

5.4 Discussion

When comparing the two approaches, AtypicalAnimacy outperformed AnthroScore on all categories. Overall, AtypicalAnimacy was far more likely to output positive labels. This is evident in Table 5.8, in which the proportion of positive predictions in AtypicalAnimacy's overall predictions was much higher than that of AnthroScore, but is true also for positive cases. Table 5.9 provides an example of two sentences from the *verb objects* and *verb subjects* sets that were correctly identified as anthropomorphic by the AtypicalAnimacy model, but not by AnthroScore. The first sentence was predicted as a negative by AnthroScore, and the second was inconclusive. In both cases, the masks were (nearly) identical in both experiments, with the exception of a definite article that was masked by AnthroScore in the

second sentence. These examples represents a recurring pattern of falsely rejecting anthropomorphic structures by AnthroScore, which is also reflected in lower F1-scores exhibited on these sets.

Masking	Masked Sentence	Masked Term	AS	AA
AnthroScore	Socratic reasoning encourages [MASK] to recursively discover, solve, and integrate problems while facilitating self-evaluation and refinement.	LLMs	0	1
Minimal entity	Socratic reasoning encourages [MASK] to recursively discover, solve, and integrate problems while facilitating self-evaluation and refinement.	LLMs	0	1
AnthroScore	Our evaluation shows that [MASK] can create French poetry successfully.	the model	2	1
Minimal entity	Our evaluation shows that the [MASK] can create French poetry successfully.	model	2	1

Table 5.9: Two sentences featuring anthropomorphic verbs, which were correctly identified as anthropomorphic by the AtypicalAnimacy (AA) model in both experiments, but not by AnthroScore (AS). For the first sentence, AnthroScore outputted a negative score, and for the second one it outputted an inconclusive score, consistently across both experiments.

The examples above show that AtypicalAnimacy overall had higher recognition rates. When we evaluate AnthroScore on its own, we see that the highest rates were achieved on the verb categories. Generally speaking, the masking approach works best for predicative structures, as predicates are guaranteed to remain unmasked, and provide importance contextual information about its arguments. Masked language models such as BERT are sensitive to the semantic roles represented by predicates (Ettinger, 2020), which are highly relevant in the context of animacy and anthropomorphism (Primus, 2012). This is reflected in the higher F1-scores for the verb categories obtained by both approaches in both experiments. Both models were more likely to output positive scores for predicative structures containing verbs or complements, than for sentences containing modifiers such as nouns or attributive adjectives. This tendency is clearly demonstrated in the prediction trends in the verb categories for the inconclusive cases (Table 5.8), but was also witnessed in the positive cases.

The recognition rates on the *adjectives* set, which also contains predicative structures, do not reflect this preference[13]. This is because about half the set consists of attributive adjectives (`amod`). In both approaches, these structures were more frequently incorrectly classified as negative. It would seem like the minimal entity masking strategy would resolve this issue, but it is not the case. Table 5.10 provides two positive examples from the *adjectives* set. The first contains the predicative adjective *aware*. It was masked identically in the two experiments, since the AI entity (*ChatGPT*) constituted an entire noun chunk. This sentence was correctly identified by both models. In contrast, the second sentence, containing the anthropomorphic attributive adjective *conscious*, was falsely rejected by both models in both experiments. This example represent a trend in the *adjectives* set whereby in the second experiment, sentences with adjectival modifiers were often missed, even though our masking strategy made sure to include the anthropomorphic adjectives in the context. This tendency is somewhat surprising, as it was initially expected that the minimal entity masking strategy will substantially improve the recognition rate of sentence including adjectival modifiers.

[13]The original decision during the compilation of the challenge set was to group sentences together based on their part of speech (see chapter 4). In retrospect, I should have kept predicative structures of either verbs or adjectives together, and include structures with adjectival modifiers separately. This would have undoubtedly made this observation more visible also in terms of the metrics.

Masking	Masked Sentence	Masked Term	AS	AA
AnthroScore	(2) Is [MASK] aware of the underlying commonsense knowledge for answering a specific question?	ChatGPT	1	1
Minimal entity	(2) Is [MASK] aware of the underlying commonsense knowledge for answering a specific question?	ChatGPT	1	1
AnthroScore	[MASK] would arguably deserve moral consideration, and it may be the case that large numbers of conscious systems could be created and caused to suffer.	Conscious AI systems	0	0
Minimal entity	Conscious [MASK] would arguably deserve moral consideration, and it may be the case that large numbers of conscious systems could be created and caused to suffer.	AI systems	0	0

Table 5.10: Two sentences featuring anthropomorphic adjectives. The first contains a predicative adjective (`acomp`, whereas the second contains an attributive adjective (`amod`). While the first was identified correctly by both AnthroScore (AS) and AtypicalAnimacy (AA) in both experiments, the second was completely missed by both models in both experiments.

It is not trivial to see why AtypicalAnimacy outperformed AnthroScore on the predicative structures. This might be explained by the fact that while both approaches employ a masked language model that was pre-trained on contemporary data, the AtypicalAnimacy classification threshold is determined by a corpus of 19th century utterances. And while the anthropomorphizing descriptions in modern discourse surrounding AI have become increasingly more common, reducing the animacy of typically animate contexts — this is not reflected in an animacy threshold that is determined on the basis of older data. This could be seen as either an advantage or a disadvantage of the *Living Machines* approach: this is a disadvantage because the classification might be better determined by modern data; this is also an advantage, as it highlights how anthropomorphizing AI discourse has become.

AnthroScore's worse performance is somewhat more easily explained by the masking itself for structures containing modifiers and genitive expressions. In the first experiment, this is attributed to the removal of important contextual cues caused by the AnthroScore masking strategy. Recall that the AnthroScore masking strategy is automatic, i.e. given an entity, they parse the sentence into chunks, identify the chunk containing that entity, and replace it with a `[MASK]` string. This strategy is suitable for anthropomorphism detection in cases where the verb contributes the most to the anthropomorphism, but is costly in terms of the contextual information that is lost when the anthropomorphic components are masked. This is particularly evident for sentences containing noun phrases, in which the AI entity and the anthropomorphic component are often collocated within the same noun chunk. In the second experiment, this is attributed to the reliance of AnthroScore's method on pronominalization, i.e. replacing noun phrases with pronouns. The minimal entity masking strategy does not replace entire noun chunks but rather parts of them. This often results in contexts that are not syntactically suitable for pronouns. Unlike AnthroScore, AtypicalAnimacy allows for the substitution of a masked entity with any token, resulting in more flexible replacements by the masked language model, which are also more likely to represent animate entities. This is demonstrated clearly in AtypicalAnimacy's higher accuracy rates of in the noun phrase sets.

Both experiments had overall lower recognition rates in the noun phrases sets. In the first experiment, this was true for both the *role/function NPs* set and the *genitive NPs* set. As explained above, this is due to the problematic masking employed by AnthroScore, which completely masks the

anthropomorphic context in expressions like 'the model's ability to reason' or 'an AI teacher'. In the case of role or function expressions, the performance of the models degraded in the second experiment (Table 5.11. This is due to a fundamental incompatibility between the role or function expressions and the minimal entity masking strategy. For phrases such as 'the AI teacher', the minimal entity *AI* is masked, resulting in the masked context 'the [MASK] teacher'. In these compositional noun phrases, the AI entity (usually *AI*, *LM* or *LLM*) functions as a noun adjunct, which is a type of modifier that modifies the anthropomorphic noun phrase. When the AI entity is masked, what remains is a context that is best compatible with other noun adjuncts, or simply another modifier. The most common modifier is an adjective. Thus, masking the AI entity in these expressions yields a context that is most compatible with adjectives, which are not nouns and as such less likely to be associated with animate entities, in the case of AtypicalAnimacy. In the case of AnthroScore it is even worse: a pronoun does not belong to that syntactical environment. This problem of mismatching syntactical environments is not addressed by Cheng et al. (2024) since their original experiment relied on their own masking strategy, which is compatible with pronominalization. So, it is not clear how their model actually computes the score. We can only assume, based on the performance rates, that this incompatibility hinders the prediction abilities of their approach.

Masking	**Masked Sentence**	**Masked Term**	**AS**	**AA**
AnthroScore	Study 4 uses a longitudinal design and finds that [MASK] consistently reduces loneliness over the course of a week.	an AI companion	0	0
Minimal entity	Study 4 uses a longitudinal design and finds that an [MASK] companion consistently reduces loneliness over the course of a week.	AI	0	0

Table 5.11: Two masked instances of the same sentence, one with the AnthroScore masking strategy and one with the minimal entity masking strategy, including the AnthroScore (AS) and AtypicalAnimacy (AA) scores. In this case, the minimal entity masking strategy did not contribute to the correct identification.

Conversely, employing the minimal entity masking strategy completely transformed the performance of the AtypicalAnimacy model on genitive expressions. The accuracy rate surpassed even that of the *adjectives* set. The example in Table 5.12 illustrates how the minimal entity masking strategy improved the performance of both approaches, by including the anthropomorphic component in the context. While the accuracy of AnthroScore did substantially increase (by 550%), the overall success rate remained relatively low, at approximately 12%. Once again, this is most likely due its pronoun constraint which is strictly incompatible with genitive structures: pronouns have their own genitive inflection and do not co-occur with the possessive clitic.

Let us conclude the discussion by addressing some of the limitations of the masked language model approach. First, as exhibited in this experiment, this approach is simply incompatible with certain linguistic structures, such as those including role or function noun phrases. We ought to acknowledge that on some interpretations of anthropomorphism, noun phrases such as *AI teacher* or *AI judge* might not be seen as inherently anthropomorphizing, rather understood as comparisons in which AI is likened, but not identified with humans. Nevertheless, the syntactical constraints that arise from the masking pattern of such expressions limit the possibility of learning anything substantial about these expressions using a masked language model.

The alternative masking strategy, while overall contributing to an improved recognition rate for both approaches, has the main setback of being manually performed. In the case that this challenge

Masking	Masked Sentence	Masked Term	AS	AA
AnthroScore	The possibility of manipulating, fooling or fairwashing evidence of [MASK] has detrimental consequences when applied in high-stakes decision-making and knowledge discovery.	the model's reasoning	0	0
Minimal entity	The possibility of manipulating, fooling or fairwashing evidence of the [MASK]'s reasoning has detrimental consequences when applied in high-stakes decision-making and knowledge discovery.	model	1	1
AnthroScore	However, all common approaches from this field are based on communicating information about features or characteristics that are especially important for [MASK].	an AI's decision	0	0
Minimal entity	However, all common approaches from this field are based on communicating information about features or characteristics that are especially important for an [MASK]'s decision.	AI	1	1

Table 5.12: Two sentences containing genitive noun phrases, demonstrating the improvement between the AnthroScore masking strategy and the minimal entity masking strategy. In both cases, AnthroScore and AtypicalAnimacy falsely rejected these sentences in the first experiment, but were able to correctly identify them as anthropomorphic in the second experiment, due to the improved context.

set is expanded, it will not be feasible to manually mask each instance. Instead, we might consider implementing Named Entity Recognition, or improving on AnthroScore's noun chunking method by excluding certain stop words and modifiers that are commonly used in the context of AI technologies. It should also be acknowledged that this evaluation relied on a relatively small dataset of 600 utterances, annotated by a single annotator. While this is a good start, this dataset would ideally be expanded to include more samples and more variation. Also, to truly serve as a benchmark for anthropomorphic language detection, it should be annotated by multiple annotators, and a robust inter-annotator agreement study should be conducted.

Additionally, the current implementations are English-centric. Both models are designed to identify English-specific animacy features which are understood as anthropomorphism in context. For instance, many non-English languages do not have an inanimate pronoun, and their linguistic markers of animacy are far more nuanced. Alternatively, we might expect to see morphological variations or differential object marking (De Swart & De Hoop, 2018), but these cues are far more difficult to identify and are not necessarily contextual. As future work, perhaps tuning a larger base model on this task would eliminate some of the issues we describe.

Last but not least, the predictions of language models are essentially based on statistical co-occurrences (Zhang et al., 2024). In AI research, as terminology is increasingly anthropomorphic and constantly introduces neologisms consisting of metaphors for human activities (e.g. *training*, *learning*, *attention*, *memory*, *hallucinations*, etc.), MLMs are more likely to predict an AI entity such as *ChatGPT*, *language model*, and *AI agent* when these terms appear in its context, instead of predicting human entities. While AnthroScore's pronoun constraint avoids this issue, it creates other problems as observed above. More importantly, anthropomorphic language does not always align with grammatical animacy; an entity can be referred to by inanimate pronouns, or framed as a tool or product, but still ascribed cognitive abilities or having human-like characteristics (see section 4.2). Beyond this methodological issue, MLM predictions are rooted in their training data, which is not always representative of real-world experiences and subjective human evaluations, in the context of anthropomorphic language and in general (Bender & Koller, 2020).

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis aimed at presenting a solid conceptual framework of anthropomorphic language in AI research, and investigate existing implementations of anthropomorphic language detection. To achieve these goals, three research questions were defined. The first research question focused on the specific human-like attributes and capacities that are commonly attributed to AI technologies in anthropomorphic contexts. The second research question focused on the specific linguistic structures that characterize anthropomorphic language. In particular, what are the underlying linguistic structures that are found in anthropomorphizing descriptions? The third research question pertained to the operationalization of the linguistic structures in the task of anthropomorphic language detection. The question was whether the state-of-the-art methods for anthropomorphic language detection are capable of handling diverse linguistic representation of anthropomorphism. The high-level objective of these goals is to carry out a rigorous analysis and evaluation of anthropomorphic language and the tools for its detection, motivated by the conceptual and ethical implications of such language in research.

Anthropomorphic language is prevalent in AI research, design and public discourse. The use of metaphors and analogies of the human brain are useful for conceptualizing and explaining the operation of these systems. From a linguistic and psychological point of view, the role of language in conversational agents contributes to their perception as human-like. Certain linguistic cues in AI outputs have an effect on the user's trust[1]. In design, these terms add clarity to developers and shape user experience. However, Anthropomorphic language, especially in AI research, is not ethically neutral, and gives rise to a host of conceptual and ethical consequences. From a conceptual point of view, terminology that conceptualizes machines and AI systems in terms of the human brain obscure the actual processes and operation of these technologies. This results in misleading hypotheses, research questions and objectives, and derail researchers and practitioners from exploring viable research directions — resulting in *AI winters*. This amalgamation of concepts relating to the human brain with computational and algorithmic definitions of technologies also negatively affects biological and neuro-scientific research. Both fields of research will therefore benefit from conceptual clarity.

From an ethical point of view, anthropomorphic descriptions of AI lead to their perception as moral

[1]In the context of anthropomorphism, user trust is often framed as a positive thing. This is because the driving interest for increased user trust is from the product owner's point of view, not the consumer. That is, the prevailing reason to increase user trust is for the purpose of selling a product. Increased trust can also have negative impacts, as described in the ethical discussion.

agents, deserving of moral consideration and trust, capable of discerning right and wrong, and able to assume responsibility. Trust, or perhaps more importantly *mistrust*, is displaced, as well as ascriptions of responsibility and accountability. There is no doubt that harmful consequences of deploying AI systems must be accounted for by the people and stakeholders who own, develop and deploy these technologies. When AI systems are perceived as autonomous entities, this positions them as moral participants of society, effectively removing blame and responsibility from those who truly deserves it. As moral participants of society, they also have the power to make normative claims. In these cases, users may take outputs of conversational AI systems as ethical instructions or recommendations. Generally, when AI systems are overly anthropomorphized, their perception as human equivalents can go beyond the aspect of moral status. When they are ascribed cognitive and reasoning abilities, they are often seen as equipped to perform human roles that require expertise and attention, which can have a devastating impact on the lives of those involved; for instance, when they are deployed in contexts of criminal justice, education or public health. Overall, the anthropomorphizing of AI systems contributes to AI hype, which in its positive form exaggerates the ability and utility of AI systems, and in its negative form, causes needless panic about AI domination. Therefore, mitigating these implications through the moderation of the language we use to talk about AI is of utmost importance.

With the objective of providing a conceptual framework of anthropomorphic language, chapter 3 presents a linguistic model of anthropomorphic language, consisting of a taxonomy of human-like attributes commonly found in anthropomorphic descriptions, and an ensuing analysis of their underlying linguistic structures. The taxonomy, presented in section 3.2, builds on existing work by DeVrio et al. (2025), relying on their guiding lenses to define seventeen attributes and capacities commonly ascribed to AI systems. These are based on existing literature as well as through the probing and analysis of real-world data. Anthropomorphic language in the context of AI technologies can be expressed in various ways. Existing work on anthropomorphic language by (Cheng et al., 2024) identifies and analyzes certain verb classes such as *cognitive verbs* and *reporting verbs* which appear frequently in anthropomorphic descriptions. The analysis presented in section 3.3 expands on their work, relying on features of animacy, proto-role theory and a frame-semantic approach to characterize syntactic structures in which an AI entity is related to an anthropomorphizing verb. In particular, AI entities in the subject position are often anthropomorphized by a cognitive verb that elicits the possession of certain mental states, like *think* or *desire*. For AI entities in the object position, we distinguish between the *agent-object* and the *experiencer-object*. The former is associated with increased agency and is anthropomorphized by verbs such as *understand*, *think* or *believe* in the passive voice. The latter is anthropomorphized by certain psych-verbs that assume cognitive and mental states within the object, such as *fool*, *convince* or *encourage*. From a standpoint of frame semantics, the AI entities that are the arguments of these predicates are framed as a *Cognizer*, *Speaker*, *Perciever* or *Experiencer*. Similarly to verbs, certain adjectives can also elicit a predicate-argument structure that assigns certain thematic roles to an AI entity. For instance, predicative structures as in 'the model is aware' give rise to a similar analysis.

Importantly, not all anthropomorphic descriptions revolve around the predicate. With respect to adjectives, the linguistic model also takes into account attributive adjectives, which ascribe human-like properties to the entity they modify by means of description. Additionally, certain noun phrases which are often collocated with AI entities embody anthropomorphic qualities. The notion of *task-based anthropomorphism* (Ryazanov et al. 2024), understood as humanizing language describing functionality,

is the assignment of traditionally human roles or functions to AI systems. For instance, terms such as *student model* and *teacher model* embody the idea that language models have the capacity to learn and exchange information in a manner akin to human learning. Anthropomorphic noun phrases could be associated with AI entities more literally, in terms of possession. AI entities are often described as being in possession of *reasoning abilities* or *common-sense knowledge*. These are described in the model as genitive expressions, attributing certain qualities to an entity through a possession relation, expressed by a possessive clitic, preposition or pronoun. Anthropomorphizing descriptions can be seen as comparing or equating AI systems to humans, either implicitly, by means of certain verbs, adjectives and noun phrases, or explicitly — in the form of a direct comparison. The linguistic model also addresses these cases, illustrating commonly used phrases used in comparative structures. For completeness, the model also mentions linguistic phenomena associated with animacy that span multiple sentences such as individuation and coreference.

The linguistic model, consisting of the taxonomy and the structural analysis, therefore answers the first two research questions. The result is a solid baseline for future work on anthropomorphic language in the context of AI. In the context of this work, it is further developed into *AnthroSet*, a challenge set consisting of real-world examples of anthropomorphic language in AI research. This challenge set, presented in chapter 4, aims at evaluating the state-of-the-art methods for anthropomorphic language detection, and consists of 600 manually annotates sentences of anthropomorphic and non-anthropomorphic descriptions of AI technologies. It was compiled from abstracts on the topic of AI, machine learning and natural language processing published in the ACL anthology and arXiv. The structures were automatically parsed and identified on the basis of the predefined syntactic structures, and then manually reviewed and annotated. The annotation procedure begins by distinguishing explicit anthropomorphism, which expresses direct and explicit anthropomorphism in the contents of the sentence, and implicit anthropomorphism, which is implied and often ambiguous, and can be identified in the form. The rest of the annotation procedure is inspired by the VU Metaphor Identification Procedure (Steen et al., 2010): given a sentence and an AI entity, potentially anthropomorphic lexical units are identified, and their contextual meaning is established. The annotation process included an additional inter-annotation step for 20% of the sentences, which showed overall congruence on positive and negative cases. Inconclusive cases resulted in the most disagreement, highlighting the ambiguous and subjective nature of this task. The challenge set is available on GitHub[2].

The third and final research question is answered in chapter 5, which conducts an evaluation of the state-of-the-art for detecting anthropomorphic, i.e. humanizing language in the context of machines and technologies. One implementation is a measure for the degree of anthropomorphism in a given sentence (*AnthroScore: A Computational Linguistic Measure of Anthropomorphism*, Cheng et al., 2024). The other provides a pipeline for atypical animacy detection, focusing on machines as an example for typically animate entities that are often described as having animate attributes (*Living Machines: A study of atypical animacy*, Coll Ardanuy et al., 2020). The two approaches employ a very similar methodology, viz. a masked language model that computes the probability that a masked entity is replaced by a token associated with animacy. The two approaches were evaluated twice on the challenge set, using two different masking approaches. The first is an automatic approach that relies on keyword identification and noun chunking, provided in AnthroScore's pipeline. The second is a manual masking approach introduced for the purpose of this evaluation, which masked the minimal phrase representing

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

the AI entity, leaving out determiners, quantifiers and modifiers that are not an essential part of its name. In both experiments, the implementations were compared in terms of their precision, recall, and F1-scores on the multiclass sets, and accuracy on the single class sets. Overall, the method by Coll Ardanuy et al. (2020) outperformed that of Cheng et al. (2024) in all linguistic categories. The predicative structures were most easily identified in both approaches. The case of modifiers proved to be tricker, with lower F1-scores in the set of adjectives in both approaches. Expressions containing noun modifiers were most frequently missed, and both approaches exhibited low accuracy rates on those sentences, due to syntactic and semantic constraints that arose from the masking approach.

Three main conclusions are drawn from this evaluation. The first conclusion is that an automatic masking strategy is problematic, as it obscures important contextual information, especially in structures in which the anthropomorphic component exists within the same noun chunk as the AI entity. Future implementations of the unsupervised approach to anthropomorphic language detection should take this into consideration, and develop a more refined masking strategy. The second conclusion is that the unsupervised method utilized in both approaches, which relies on a masked language model, is not suitable for all types of linguistic structures. Shardlow et al. (2025), who evaluate a supervised implementation with AnthroScore as a baseline, reach the same conclusion. Their suggestion, to explore supervised methods for anthropomorphic language detection, is supported by the findings of this thesis. The theoretical framework presented in this thesis could serve as a foundation for such work. A third conclusion is a more general issue with the manner in which language models such as MLMs make predictions. As anthropomorphic descriptions become increasingly prevalent in AI discourse (Ryazanov et al. 2024), contexts that previously seemed highly anthropomorphic are now perceived as more benign. For instance, many anthropomorphic expressions that have not been previously associated with AI are now part of the technical lexicon, e.g. *chain of thought* or *hallucinations*. This was described above as the Overton Window of AI — i.e. the range of *public* and *technical* acceptable descriptions is shifting to include more human-like terms. That is, contexts in which an AI entity is framed as animate or human appear more frequently in the body of work that is used to train AI models. These contexts are introduced into the training data, and language models improve at associating AI entities with anthropomorphizing contexts. Since the predictions of language models are essentially based on statistical co-occurence (Zhang et al., 2024), this reduces their efficacy as tools for this task. In the end, these predictions are also rooted in the culture and context of their training data, and are not a direct representation of human judgments and subjective evaluations (Bender & Koller, 2020).

6.2 Future Work and Recommendations

In terms of the evaluation set, future work includes expanding the challenge set presented in chapter 4 to include more examples, as well as annotations by multiple annotators and a study of overall agreement. While this task is highly subjective (Shardlow et al., 2025), a detailed linguistic model of anthropomorphism such as the one presented in this paper can help with establishing a standard on the basis on which agreement can be obtained. By extending the dataset to include many more examples, including positive and negative samples, we are investing towards a benchmark for evaluating future implementations of anthropomorphic language detection. The corpus presented in Shardlow et al. (2025) can be analyzed with respect to the linguistic model. It would be interesting to see the distribution of the linguistic categories in a fully annotated corpus that has been made accessible and

presented in a peer-reviewed work.

Additionally, it would be pertinent to research alternative methods for anthropomorphic language detection, in light of the problems arising from the MLM-based approaches. One possible direction is probing the 'predictive' abilities of generative AI models, by designing prompts containing anthropomorphic contexts and prompting the model to complete them, similarly to the MLM approach. However, this method might entail the same issues as described above with respect to large language models and the increasingly anthropomorphic language prevalent in their training data. Other directions include exploring supervised methods, as suggested by Shardlow et al. (2025). The task of anthropomorphic language detection can be redefined as a classification task, using traditional NLP approaches relying on good old linguistics and feature engineering. Either direction will benefit from the solid conceptual foundations presented in this thesis.

Finally, promoting sound writing practices when it comes to anthropomorphic descriptions is paramount to future AI research. Papers that have been published and thus accredited by the academic community have an authoritative tone; they also influence, to a certain extent, downstream media representations. Researchers have a responsibility to write carefully and deliberately about their research object, especially when it is of high interest to the public, as in the case of AI. Dillon (2020) criticizes Hofstadter's critique of the *ELIZA* effect, claiming that his role as a researcher who is actively engaged in AI research effectively renders him responsible for the resulting AI hype and its negative implications. This thesis does not agree with the idea that AI research is *anthropomorphize or bust*. Proudfoot (2011) mentions anthropomorphic language as something to be managed and not completely removed. I agree with this, to the extent that we might never be able to purge anthropomorphic perception. However, we can and should be actively scrutinizing the way in which we describe and discuss AI systems. As recommended by McDermott (1976), the objectives of AI research can be redefined in terms of the processes and not the goals. Inie et al. (2024), for instance, manifest this idea in praxis, by making a choice to refer to AI systems as 'probabilistic automation systems' instead. Shardlow and Przybyła (2024) present a list of recommendation for AI developers, users and reporters of AI. For instance, developers should screen for anthropomorphic language in their manuscript, consider the intended audience of their work, and use the proper technical terminology accordingly. Reporters are encouraged to be skeptical about AI research, take the source into account, and team up with academics in order to present an image that is true to reality.

Bibliography

Abercrombie, G., Cercas Curry, A., Dinkar, T., Rieser, V., & Talat, Z. (2023, December). Mirages. On Anthropomorphism in Dialogue Systems. In H. Bouamor, J. Pino, & K. Bali (Editors), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Pages 4776–4790). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.290. (Cited on pages 3, 5, 6, 17).

Alabed, A., Javornik, A., & Gregory-Smith, D. (2022). AI anthropomorphism and its effect on users' self-congruence and self–AI integration: A theoretical framework and research agenda. *Technological Forecasting and Social Change*, *182*, 121786. https://doi.org/10.1016/j.techfore.2022.121786 (cited on page 6).

Barrow, N. (2024). Anthropomorphism and AI hype. *AI and Ethics*, *4*(3), 707–711. https://doi.org/10.1007/s43681-024-00454-1 (cited on page 3).

Belletti, A., & Rizzi, L. (1988). Psych-verbs and θ-theory. *Natural Language and Linguistic Theory*, *6*(3), 291–352. https://doi.org/10.1007/bf00133902 (cited on page 23).

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463 (cited on pages 48, 55, 59).

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., & Tan, Y. F. (2008, May). The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, & D. Tapias (Editors), *Proceedings of the sixth international conference on language resources and evaluation (LREC'08)*. European Language Resources Association (ELRA). https://aclanthology.org/L08-1005/. (Cited on page 7).

Blackwell, A. F. (1996, April). Metaphor or analogy: How should we see programming abstractions? In P. Vanneste, K. Bertels, B. D. Decker, & J.-M. Jaques (Editors), *Proceedings of the 8th annual workshop of the psychology of programming interest group* (Pages 105–113). (Cited on page 9).

Bock, K., Loebell, H., & Morey, R. (1992). From conceptual roles to structural relations: Bridging the syntactic cleft. *Psychological Review*, *99*(1), 150–171. https://doi.org/10.1037/0033-295X.99.1.150 (cited on page 22).

Bossong, G. (1991, March). Differential Object Marking in Romance and Beyond. In D. Wanner & D. A. Kibbee (Editors), *Current Issues in Linguistic Theory* (Pages 143–170, Volume 69). John Benjamins Publishing Company. https://doi.org/10.1075/cilt.69.14bos. (Cited on page 21).

Branigan, H. P., Pickering, M. J., & Tanaka, M. (2008). Contributions of animacy to grammatical function assignment and word order during production. *Lingua*, *118*(2), 172–189. https://doi.org/10.1016/j.lingua.2007.02.003 (cited on page 26).

Brooker, P., Dutton, W., & Mair, M. (2019). The new ghosts in the machine: 'Pragmatist' AI and the conceptual perils of anthropomorphic description. https://doi.org/10.5281/ZENODO.3459327 (cited on pages 10, 11).

Carbonell, J., Sánchez-Esguevillas, A., & Carro, B. (2016). The role of metaphors in the development of technologies. The case of the artificial intelligence [Publisher: Elsevier BV]. *Futures*, *84*, 145–153. https://doi.org/10.1016/j.futures.2016.03.019 (cited on page 9).

Carolyn Y. Johnson. (2024, November). ChatGPT is a poet. A new study shows people prefer its verses. https://www.washingtonpost.com/science/2024/11/14/chatgpt-ai-poetry-study-creative. (Cited on page 20).

Cheng, M., Gligoric, K., Piccardi, T., & Jurafsky, D. (2024, March). AnthroScore: A computational linguistic measure of anthropomorphism [Cheng et al.]. In Y. Graham & M. Purver (Editors), *Proceedings of the 18th conference of the european chapter of the association for computational linguistics (volume 1: Long papers)* (Pages 807–825). Association for Computational Linguistics. https://aclanthology.org/2024.eacl-long.49/. (Cited on pages 3, 10, 13, 22, 28, 30, 31, 40, 42–45, 54, 57–59).

Chu, M., Gerard, P., Pawar, K., Bickham, C., & Lerman, K. (2025, May). Illusions of intimacy: Emotional attachment and emerging psychological risks in human-AI relationships. https://doi.org/10.48550/arXiv.2505.11649. (Cited on page 8).

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104 (cited on page 40).

Coll Ardanuy, M., Nanni, F., Beelen, K., Hosseini, K., Ahnert, R., Lawrence, J., McDonough, K., Tolfo, G., Wilson, D. C., & McGillivray, B. (2020, December). Living machines: A study of atypical animacy. In D. Scott, N. Bel, & C. Zong (Editors), *Proceedings of the 28th international conference on computational linguistics* (Pages 4534–4545). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.400. (Cited on pages 13, 27, 28, 42–46, 58, 59).

Dahl, Ö. (2008). Animacy and egophoricity: Grammar, ontology and phylogeny. *Lingua*, *118*(2), 141–150. https://doi.org/10.1016/j.lingua.2007.02.008 (cited on page 21).

De Swart, P., & De Hoop, H. (2018). Shifting animacy. *Theoretical Linguistics*, *44*(1-2), 1–23. https://doi.org/10.1515/tl-2018-0001 (cited on pages 21, 55).

DeVrio, A., Cheng, M., Egede, L., Olteanu, A., & Blodgett, S. L. (2025). A Taxonomy of Linguistic Expressions That Contribute To Anthropomorphism of Language Technologies. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3706598.3714038 (cited on pages 3, 7, 16–19, 57).

Dillon, S. (2020). The Eliza effect and its dangers: From demystification to gender critique. *Journal for Cultural Research*, *24*(1), 1–15. https://doi.org/10.1080/14797585.2020.1754642 (cited on page 60).

Dowty, D. R. (1991). Thematic Proto-Roles and Argument Selection. *Language*, *67*(3), 547–619. https://doi.org/10.2307/415037 (cited on pages 4, 21, 22).

Dreyfus, H. L. (1976). What computers can't do. *British Journal for the Philosophy of Science*, *27*(2), 177–185 (cited on page 7).

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason* (Revised edition). MIT press. (Cited on page 7).

Emnett, C. Z., Mott, T., & Williams, T. (2024). Using Robot Social Agency Theory to Understand Robots' Linguistic Anthropomorphism. *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 447–452. https://doi.org/10.1145/3610978.3640747 (cited on pages 3, 6).

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864 (cited on pages 3, 43).

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48. https://doi.org/10.1162/tacl_a_00298 (cited on page 52).

Fellbaum, C. (Editor). (1998). *Wordnet: An electronic lexical database*. MIT Press. https://mitpress.mit.edu/9780262561167/wordnet/. (Cited on page 31).

Fetzer, A. (2008). "And I Think That Is a Very Straightforward Way of Dealing With It": The Communicative Function of Cognitive Verbs in Political Discourse. *Journal of Language and Social Psychology*, *27*(4), 384–396. https://doi.org/10.1177/0261927X08322481 (cited on page 22).

Fillmore, C. J. (1982). Frame semantics. In T. L. S. of Korea (Editor), *Linguistics in the morning calm* (Pages 111–137). Hanshin Publishing Co. (Cited on pages 4, 21).

Fillmore, C. J., & Baker, C. (2012, September). A Frames Approach to Semantic Analysis. In B. Heine & H. Narrog (Editors), *The Oxford Handbook of Linguistic Analysis* (1st edition, Pages 313–340). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199544004.013.0013. (Cited on pages 23, 35).

Floridi, L., & Nobre, A. C. (2024). Anthropomorphising Machines and Computerising Minds: The Crosswiring of Languages between Artificial Intelligence and Brain & Cognitive Sciences. *Minds and Machines*, *34*(1). https://doi.org/10.1007/s11023-024-09670-4 (cited on pages 3, 10, 11).

Fukumura, K., & Van Gompel, R. P. (2010). Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language*, *62*(1), 52–66. https://doi.org/10.1016/j.jml.2009.09.001 (cited on page 28).

Gentner, D. (1978). On Relational Meaning: The Acquisition of Verb Meaning. *Child Development*, *49*(4), 988. https://doi.org/10.2307/1128738 (cited on page 21).

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge Distillation: A Survey. *International Journal of Computer Vision*, *129*(6), 1789–1819. https://doi.org/10.1007/s11263-021-01453-z (cited on page 19).

Gregorio, N., Gay, M., Goldwater, S., & Ponti, E. (2025, July). The Cross-linguistic Role of Animacy in Grammar Structures. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Editors), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Pages 7349–7363). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.364. (Cited on page 21).

Grimm, S. (2018). Grammatical number and the scale of individuation. *Language*, *94*(3), 527–574. https://doi.org/10.1353/lan.2018.0035 (cited on page 28).

Hofstadter, D. R. (1995). *Fluid concepts & creative analogies: Computer models of the fundamental mechanisms of thought*. Basic Books. (Cited on pages 3, 8).

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303 (cited on pages 30, 31).

Hyland, K. (1998). Boosting, hedging and the negotiation of academic knowledge. *Text - Interdisciplinary Journal for the Study of Discourse*, *18*(3). https://doi.org/10.1515/text.1.1998.18.3.349 (cited on page 22).

Ikeya, A. (1995, December). Predicate-argument structure of English adjectives. In B. K. T'sou & T. B. Y. Lai (Editors), *Proceedings of the 10th pacific asia conference on language, information and computation* (Pages 149–156). City University of Hong Kong. https://doi.org/http://hdl.handle.net/2065/11874. (Cited on page 23).

Inie, N., Druga, S., Zukerman, P., & Bender, E. M. (2024). From "AI" to Probabilistic Automation: How Does Anthropomorphization of Technical Systems Descriptions Influence Trust? [Inie et al.]. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2322–2347. https://doi.org/10.1145/3630106.3659040 (cited on pages 3, 7, 9, 10, 60).

Jacob, P. (2023). Intentionality. In E. N. Zalta & U. Nodelman (Editors), *The Stanford encyclopedia of philosophy* (Spring 2023). Metaphysics Research Lab, Stanford University. (Cited on page 18).

Kahn, P. H., Ishiguro, H., Friedman, B., Freier, N. G., Severson, R. L., & Miller, J. (2007). What Is a Human? Toward Psychological Benchmarks in the Field of Human-Robot Interaction. https://doi.org/10.1075/is.8.3.04kah (cited on page 6).

Kevin Williams. (2025, July). If AI attempts to take over world, don't count on a 'kill switch' to save humanity. https://www.cnbc.com/2025/07/24/in-ai-attempt-to-take-over-world-theres-no-kill-switch-to-save-us.html. (Cited on page 17).

Kim, Y., & Sundar, S. S. (2012). Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, *28*(1), 241–250. https://doi.org/10.1016/j.chb.2011.09.006 (cited on page 5).

Levin, B. (2022). On Dowty's "Thematic Proto-Roles and Argument Selection" [ISSN: 0924-4662, 2215-034X]. In *Studies in Linguistics and Philosophy* (Pages 103–119). Springer International Publishing. https://doi.org/10.1007/978-3-030-85308-2_7. (Cited on page 22).

Li, M., & Suh, A. (2022). Anthropomorphism in AI-enabled technology: A literature review [Publisher: Springer Science and Business Media LLC]. *Electronic Markets*, *32*(4), 2245–2275. https://doi.org/10.1007/s12525-022-00591-7 (cited on page 5).

Lin, Z., Chen, X., Pathak, D., Zhang, P., & Ramanan, D. (2024). Revisiting the Role of Language Priors in Vision-Language Models [ISSN: 2640-3498]. *Proceedings of the 41st International Conference on Machine Learning*, 29914–29934. Retrieved August 11, 2025, from https://proceedings.mlr.press/v235/lin24c.html (cited on page 18).

Mackinac Center For Public Policy. (no date). A Brief Explanation of The Overton Window. https://www.mackinac.org/OvertonWindow. (Cited on page 34).

Maeda, T., & Quan-Haase, A. (2024). When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1068–1077. https://doi.org/10.1145/3630106.3658956 (cited on pages 6, 19).

McDermott, D. (1976). Artificial intelligence meets natural stupidity. *ACM SIGART Bulletin*, (57), 4–9. https://doi.org/10.1145/1045339.1045340 (cited on pages 8, 60).

Mecit, A., Lowrey, T. M., & Shrum, L. J. (2022). Grammatical gender and anthropomorphism: "It" depends on the language [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *123*(3), 503–517. https://doi.org/10.1037/pspa0000309 (cited on page 28).

Meir, I., Aronoff, M., Börstell, C., Hwang, S.-O., Ilkbasaran, D., Kastner, I., Lepic, R., Lifshitz Ben-Basat, A., Padden, C., & Sandler, W. (2017). The effect of being human and the basis of grammatical word order: Insights from novel communication systems and young sign languages. *Cognition*, *158*, 189–207. https://doi.org/10.1016/j.cognition.2016.10.011 (cited on page 26).

Mueller, S. T. (2020). Cognitive Anthropomorphism of AI: How Humans and Computers Classify Images. *Ergonomics in Design: The Quarterly of Human Factors Applications*, *28*(3), 12–19. https://doi.org/10.1177/1064804620920870 (cited on page 12).

Murray-Rust, D., Nicenboim, I., & Lockton, D. (2022). Metaphors for designers working with AI. *DRS Biennial Conference Series – DRS 2022: Bilbao*. https://doi.org/10.21606/drs.2022.667 (cited on page 9).

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78. https://doi.org/10.1145/191666.191703 (cited on page 12).

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal dependencies v2: An ever-growing multilingual treebank collection. *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, 4034–4043 (cited on page 23).

Placani, A. (2024). Anthropomorphism in AI: Hype and fallacy. *AI and Ethics*, *4*(3), 691–698. https://doi.org/10.1007/s43681-024-00419-4 (cited on pages 3, 12).

Primus, B. (2012). Animacy, Generalized Semantic Roles, and Differential Object Marking [ISSN: 1873-0043]. In *Studies in Theoretical Psycholinguistics* (Pages 65–90). Springer Netherlands. https://doi.org/10.1007/978-94-007-1463-2_4. (Cited on pages 22, 52).

Prior, A., Kroll, J. F., & Macwhinney, B. (2013). Translation ambiguity but not word class predicts translation performance. *Bilingualism: Language and Cognition*, *16*(2), 458–474. Retrieved August 26, 2025, from https://www.cambridge.org/core/product/identifier/S1366728912000272/type/journal_article (cited on page 21).

Proudfoot, D. (2011). Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence*, *175*(5-6), 950–957. https://doi.org/10.1016/j.artint.2011.01.006 (cited on page 60).

Rahmanzadehgervi, P., Bolton, L., Taesiri, M. R., & Nguyen, A. T. (2025). Vision Language Models are blind. In M. Cho, I. Laptev, D. Tran, A. Yao, & H. Zha (Editors), *Computer Vision – ACCV 2024* (Pages 293–309). Springer Nature. https://doi.org/10.1007/978-981-96-0917-8_17. (Cited on page 18).

Rosenbach, A. (2017). Constraints in contact: Animacy in English and Afrikaans genitive variation – a cross-linguistic perspective [Publisher: Open Library of Humanities]. *Glossa: a journal of general linguistics*, *2*(1). https://doi.org/10.5334/gjgl.292 (cited on page 26).

Rossi, M. G., & Macagno, F. (2021). The communicative functions of metaphors between explanation and persuasion. In *Inquiries in philosophical pragmatics* (Pages 171–191). Springer International Publishing. https://doi.org/10.1007/978-3-030-56437-7_12. (Cited on page 8).

Rozwadowska, B. (2017, November). Psychological Verbs and Psychological Adjectives. In M. Everaert & H. C. Riemsdijk (Editors), *The Wiley Blackwell Companion to Syntax, Second Edition* (1st edition, Pages 1–26). Wiley. https://doi.org/10.1002/9781118358733.wbsyncom040. (Cited on pages 21, 23).

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., & Scheffczyk, J. (2006). Framenet ii: Extended theory and practice. https://api.semanticscholar.org/CorpusID:62163005 (cited on pages 13, 35).

Ryazanov, I., & Björklund, J. (2024, March). Thesis Proposal: Detecting Agency Attribution [Ryazanov and Björklund]. In N. Falk, S. Papi, & M. Zhang (Editors), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop* (Pages 208–214). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-srw.15. (Cited on pages 5, 10, 13).

Ryazanov, I., Öhman, C., & Björklund, J. (2024). How chatgpt changed the media's narratives on ai: A semi-automated narrative analysis through frame semantics [Ryazanov et al.]. *Minds and Machines*, *35*(1), 1–24. https://doi.org/10.1007/s11023-024-09705-w (cited on pages 3, 7, 10, 13, 17, 25, 35, 57, 59).

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB Neuroscience*, *11*(2), 88–95. https://doi.org/10.1080/21507740.2020.1740350 (cited on pages 3, 10, 12).

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/s0140525x00005756 (cited on page 7).

Shardlow, M., & Przybyła, P. (2024). Deanthropomorphising NLP: Can a language model be conscious? [Publisher: Public Library of Science]. *PLOS ONE*, *19*(1), e0307521. https://doi.org/10.1371/journal.pone.0307521 (cited on pages 4, 21, 60).

Shardlow, M., Williams, A., Roadhouse, C., Ventirozos, F., & Przybyła, P. (2025, July). Exploring Supervised Approaches to the Detection of Anthropomorphic Language in the Reporting of NLP Venues [Shardlow et al.]. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Editors), *Findings of the Association for Computational Linguistics: ACL 2025* (Pages 18010–18022). Association for Computational Linguistics. Retrieved July 28, 2025, from https://aclanthology.org/2025.findings-acl.926/. (Cited on pages 3, 4, 7, 10, 11, 13, 14, 40–42, 59, 60).

Silverstein, M. (1976, April). Hierarchy of features and ergativity. Zenodo. https://doi.org/10.5281/ZENODO.4688088. (Cited on pages 21, 25).

Steen, G., Dorst, A., Herrmann, J., Kaal, A., Krennmayr, T., & Pasma, T. (2010). *A method for linguistic metaphor identification. from mip to mipvu.* John Benjamins. (Cited on pages 34, 58).

Temme, A. (2019). *The peculiar nature of psych verbs and experiencer object structures* [Doctoral dissertation, Humboldt-Universität zu Berlin]. https://doi.org/10.18452/19889. (Cited on page 23).

Thompson, G., & Yiyun, Y. (1991). Evaluation in the Reporting Verbs Used in Academic Papers. *Applied Linguistics*, *12*(4), 365–382. https://doi.org/10.1093/applin/12.4.365 (cited on page 22).

Van Pinxteren, M. M., Pluymaekers, M., & Lemmink, J. G. (2020). Human-like communication in conversational agents: A literature review and research agenda. *Journal of Service Management*, *31*(2), 203–225. https://doi.org/10.1108/JOSM-06-2019-0175 (cited on page 7).

Walsh, K. R., Mahesh, S., & Trumbach, C. C. (2021). Autonomy in AI Systems: Rationalizing the Fears [Publisher: Epsilon Pi Tau, Inc.]. *The Journal of Technology Studies*, *47*(1), 38–47. Retrieved August 13, 2025, from https://www.jstor.org/stable/48657934 (cited on page 18).

Watson, D. (2019). The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. *Minds and Machines*, *29*(3), 417–440. https://doi.org/10.1007/s11023-019-09506-6 (cited on pages 3, 9, 12).

Watson, D., & Scheidt, D. (2005). Autonomous systems. *Johns Hopkins APL Technical Digest (Applied Physics Laboratory)*, *26*, 368–376 (cited on page 18).

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who Sees Human?: The Stability and Importance of Individual Differences in Anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232. https://doi.org/10.1177/1745691610369336 (cited on pages 3, 5, 6).

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45 (cited on page 8).

Xia, P., Qin, G., Vashishtha, S., Chen, Y., Chen, T., May, C., Harman, C., Rawlins, K., White, A. S., & Van Durme, B. (2021, April). LOME: Large ontology multilingual extraction. In D. Gkatzia & D. Seddah (Editors), *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: System demonstrations* (Pages 149–159). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-demos.19. (Cited on page 7).

Yamamoto, M. (1999, September). *Animacy and Reference: A cognitive approach to corpus linguistics* (Volume 46). John Benjamins Publishing Company. https://doi.org/10.1075/slcs.46. (Cited on pages 21, 28).

Zelinka, I. (2009). Blind algorithm [Wolfram Demonstrations Project]. (Cited on page 18).

Zhang, X., Li, M., & Wu, J. (2024). Co-occurrence is not Factual Association in Language Models [Version Number: 2]. *38th Conference on Neural Information Processing Systems*. Retrieved July 16, 2025, from https://proceedings.neurips.cc/paper_files/paper/2024/file/775226eaa2a36c543e2bd6cc9eae1b6a-Paper-Conference.pdf (cited on pages 4, 55, 59).

Zhu, J. (2004). Intention and Volition [Publisher: [Taylor & Francis, Ltd., Canadian Journal of Philosophy]]. *Canadian Journal of Philosophy*, *34*(2), 175–193. Retrieved August 12, 2025, from https://www.jstor.org/stable/40232213 (cited on page 18).

Appendix A: Taxonomy Examples

Attribute	Example
Conceptual Thought & Mental States	As a result, it can be hard to identify **what the model actually "believes" about the world**, making it susceptible to inconsistent behavior and simple errors.
	Using GPT4 as the editor, we find it can successfully edit trigger shortcut in samples that **fool LLMs**.
Knowledge & Awareness	[...]**the LLM only remembers** the answer style for open-ended safety questions, which makes it unable to solve other forms of safety tests.
	We only conduct retrieval for the missing knowledge in questions that **the LLM does not know**.
Reasoning & Understanding	This study evaluates **the GPT-4 Large Language Model's abductive reasoning** in complex fields like medical diagnostics, criminology, and cosmology.
	[...]the considered system possesses a counter-intuitive relationship between workload and performance, which nevertheless is correctly **inferred by the proposed simulation model**.
Experiences & Emotions	[...]it is widely accepted that LLMs perform well in terms of grammar, but it is unclear in what specific cognitive areas **they excel or struggle in**.
	As such, it can be valuable for a large language model (LLM), particularly as an AI assistant, **to be able to empathize** with or even explain these various standpoints.
Sense of Self	In addition, we found that **the model becomes more confident** and refuses to provide an answer in only few cases.
Power & Social Standing	We identify various roles users assign to AI companions, such as **friends, mentors, or romantic partners**, and highlights the importance of customization and emotional support in these interactions.
Professional Skills	When deployed as a **collaborative tutor**, the system restricts student interaction to a chat-only interface, promoting controlled and guided engagement.
Sensory Perception	In fact, we demonstrate that even **a "blind" language model that ignores any image evidence** can sometimes outperform all prior art[...]
Agency & Autonomy	In order to engender trust in AI, humans must understand **what an AI system is trying to achieve, and why.**
	To understand when **AI agents need to break rules**, we examine the conditions under which humans break rules for pro-social reasons.
Volition	The former is concerned with the generation of explanations for **decisions taken by AI systems**, while the latter is concerned with the way explanations are given to users and received by them.
Intentions & Attitudes	We conclude by discussing how future AI developments may affect the fight between **malicious bots** and the public.
Judgment & Morality	**AI assistants can impart value judgments that shape people's decisions and worldviews**, yet little is known empirically about what values these systems rely on in practice.
Candidness	We investigate **the ability of LLMs to be deceptive** in the context of providing assistance on a reading comprehension task, using LLMs as proxies for human users.
Communication	We also **ask ChatGPT** to provide its point of view and present its responses to several questions we attempt to answer.
Problem Solving & Support	[...]AI can quickly determine which calls are most relevant for coaching purposes, and **provide relevant feedback and insights** to the contact center manager or supervisor.
	AI assistants can help developers by recommending code to be included in their implementations (e.g., suggesting the implementation of a method from its signature).
Pedagogy	[...]we show that **LMs can teach themselves to use external tools** via simple APIs and achieve the best of both worlds.
Personality & Self Expression	Our evaluation shows that **the model can create French poetry** successfully.

Appendix B: Anthropomorphic Components

B.1 Anthropomorphic Verbs, Adjectives, and Nouns

Anthropomorphic agent verbs: think, believe, know, hope, fear, wish, imagine, remember, forget, desire, want, deduce, reason, analyze, interpret, understand, learn, misunderstand, confuse, struggle, perform, produce, execute, invent, develop, report, describe, ask, tell, explain, convey, depict, demonstrate, illustrate, collaborate, communicate, see, hear, feel, witness, experience, suffer, favor, plan, prefer choose, decide, deceive, lie, fabricate, conclude, guess, pretend, dream, mean ,suspect, suppose, assume, argue, claim

Anthropomorphic patient verbs: confuse, ask, tell, explain, illustrate, collaborate, communicate, inspire, amaze, fool, trick, scare, teach, train, instruct, show, describe, address, converse, present, display, influence, sway, convince, dissuade, persuade, correct, encourage, discourage, motivate, argue

Anthropomorphic adjectives: confident, insecure, polite, kind, friendly, mean, rude, benevolent, malicious, untrustworthy, understanding, tolerant, sympathetic, apathetic, communicative, responsive, attentive, smart, clever, intelligent, conscious, aware, blind, deaf, courageous, brave, eager, happy, sad, respectful, deceptive, relatable, creative

Anthropomorphic nouns: assistant, teacher, helper, manager, tutor, instructor, companion, partner

B.2 Anthropomorphic NPs from the *genitive NPs* set

Anthropomorphic NPs attributed to AI systems in the set of genitive NPs.

Attribute	**Terms**
Conceptual Thought & Mental States	awareness, theory of mind, conceptual capabilities, related beliefs, desire, tendency to hallucinate information
Reasoning	reasoning, reasoning abilities, reasoning capabilities, causal reasoning ability, strengths in knowledge comprehension and reasoning, knowledge acquisition capabilities, commonsense reasoning capabilities, commonsense knowledge and reasoning abilities, common-sense reasoning, abductive reasoning
Understanding	understanding, understanding abilities, ability to understand, understanding capabilities, ability to comprehend theoretical concepts and differentiate between constructs, capability in understanding and following instructions, capability to understand natural language, logic understanding, misunderstandings, misinterpretations
Power & Social Standing	cooperation and coordination behavior, social behaviour, intellectual property
Professional Skills	editorial capabilities, consultation abilities, prior legal knowledge, legal drafting and reasoning capabilities, considerable proficiency in writing Physics essays and coding abilities, ability to use tools
Agency & Autonomy	decisions, rational decisions, decision-making, deliberation, decision-making process, ability to provide a robust interpretation of its decision-making logic, actions, refusal behavior
Candidness	lies, tendency to deceive, manipulation
Problem Solving	problem-solving capabilities, process for solving a task, cognitive abilities, competence in comprehending and performing intricate tasks
Pedagogy	ability to showcase pedagogical skills, ability to discern and adapt to nuanced instructions, ability to pursue multiple interconnected learning objectives
Personality	personality types, feminine-coded abilities, polite writing, capacity to internalize and project instructible personas

Appendix C: Annotation Guidelines

The annotators received a spreadsheet containing 3 columns: column A contained a unique identifier, column B contained the sentence, and column C contained the AI entity in question. They were asked to fill in their annotation in column D. They also received annotation guidelines containing the annotation procedure and the decision tree (section 4.2), as well as a table containing the taxonomy categories and related words (section 3.2). The guidelines were preceded by the following instructions:

> *The sentences to be evaluated adhere to one of the three following structures: (1) AI entities as subjects of anthropomorphic verbs (arg0), (2) AI entities as objects of anthropomorphic verbs (as either arg0 or arg1), and (3) AI entities which have an anthropomorphic adjective as a modifier or complement. The labeling is done with respect to a specific AI entity.*
>
> *Read the sentence in column B, and following the guidelines below, enter a score in column D: 1 for anthropomorphic, 0 for non-anthropomorphic, and 2 for inconclusive cases. Since some sentences contain multiple AI entities, the relevant one is given in bold, and also explicitly mentioned in column C.*

Annotation Procedure

Step 1: Read the sentence to get a general understanding of its contents and meaning.
Q1: Do the contents overtly and directly refer to the highlighted AI entity as having human-like capacities or properties?

Yes: label the sentence as anthropomorphic. (P1)

Note: This also includes cases in which the context simultaneously refers to AI as having human-like capacities, and as being built or created by humans.

No: Continue to step 2.

Step 2: Given a highlighted AI entity, determine the lexical unit(s) related to it: the root verb, or any adjectival modifiers (`amod`) or complements (`acomp`).
Q2: Does the lexical unit have a basic meaning which relates to human-like cognition or capacities, i.e. an anthropomorphic sense?

Yes: Continue to step 3.

No: label the sentence as non-anthropomorphic. (N1)

Step 3: Focus on the meaning of the lexical unit in the sentence.
Q3: Does the lexical unit in question have a contextual meaning that reduces from its potential to anthropomorphize the AI entity?

Option 1: If it does not have a non-anthropomorphic sense, nor does it have a different meaning in the context of machines and AI. i.e. it is unambiguously anthropomorphizing – label the sentence as anthropomorphic. (P2). This includes verbs such as understand, think, know, infer, analyze, perceive, deduce, collaborate, communicate. These verbs are seen as unconditionally anthropomorphizing, i.e. independent of the context, due to their strong association with mental faculties and cognitive capacities.

Option 2: It has a more salient non-anthropomorphic sense in the particular context. Label the sentence as non-anthropomorphic. (N2)

This includes the following:

1. Metaphoric use such as "AI has seen many changes", in which the meaning of see is undergo, and not the basic meaning of experiencing visual stimulus.
2. "Erroneous" or imprecise phrasings such as 'raises concern' instead of 'gives rise to concern'.

Option 3: The word is ambiguous and its lexical meaning requires the consideration of the entire context. This includes words that have become ubiquitous in the context of machines and AI, (e.g. train, learn, vulnerable, assistant) or certain reporting or modelling verbs (e.g. explain, show, demonstrate). Continue to step 4.

Step 4: Now consider the entire context of the sentence.

Q4: Does it contribute to anthropomorphism in other ways, framing the AI entity as a *Cognizer*, *Perceiver* or *Experiencer*? On the contrary, the broader context explicitly frames the AI entity as a product, tool or machine lacking agency, and highlights its non-human status?

Option 1: If the broader context of the sentence contributes to anthropomorphism in other ways, and it does not frame the AI entity as a tool – label the sentence as anthropomorphic. (P3)

Option 2: If the broader context frames the AI entity as a tool or product lacking agency – label the sentence as non-anthropomorphic. (N3)

If at any step the answer is inconclusive — i.e. due to ambiguity or vagueness, a neutral or uninformative context, or simultaneous framing of the AI entity as a tool and as a *Cognizer* – label the sentence as inconclusive.

Notes on the annotation

1. Explicit anthropomorphic sentences are those whose contents contain attributions of human-like capacities to AI, e.g. directly describing it as having cognitive or reasoning abilities. These cases are considered highly anthropomorphic, and labeled as positive even if the context reduces from the agency or animacy of the AI entity by framing it as a tool or inanimate object. Similarly, sentences containing verbs or adjectives whose basic meaning is non-ambiguously anthropomorphic should be labeled as positive, even if other components of the sentence reduce from the overall anthropomorphism.
2. The line between reporting or modeling verbs and strictly anthropomorphic verbs is not always clear (e.g. explain, describe versus interpret, analyze). To draw a somewhat arbitrary line, we consider reporting verbs to be verbs that are used to discuss the explanatory powers of a model or system. As a rule, these verbs should evoke a sense that the AI entity is an explanatory or research tool — i.e. that the information, solution or phenomenon is evident by means of a model or a system. This is not equally true for any AI entity: a model can describe the data, an algorithm can find a solution, but if ChatGPT is said to describe the data or find a solution, the sentence is most likely anthropomorphic, as ChatGPT is usually described not as a tool used for performing these actions but as the agent of these verbs, actively performing the actions itself.
3. It is similarly difficult to draw a line between words that are always anthropomorphic, whose presence immediately give rise to a positive label, and words that have a specific meaning in the context of machines or AI, which are taken to be conditionally anthropomorphic – i.e, depending on the broader context of the sentence (e.g. infer, understand versus acquire, learn). This is due to the difficulty to disambiguate meaning based on a single sentence, and also because the terminology of AI is constantly shifting and expanding to address advances in the field. For instance, 'train' has

become ubiquitous to the extent that 'training models' can be seen as a standalone sense, separate from the sense of training humans, e.g. athletes. The labeling for such cases should therefore strongly rely on context and the framing of the AI entity through other means besides the lexical unit in question – handled in step 4. Highly ambiguous or vague cases can be labeled as inconclusive.

4. Option 2 in step 3 is meant to address cases in which a word has a non-anthropomorphic sense that is the most salient reading of the sentence, not just in the context of AI but in any context. For example "In the past years, AI has seen many advances in the field of NLP", or "In such cases, LLMs primarily act as an agent". In these sentences, it is clear that the AI is not to be understood as having the capacity to see or act – the more salient readings are 'undergo' and 'function', respectively. While these uses might be considered metaphorical and labeled as such in a metaphor detection task, we do not equate metaphoric use with anthropomorphism. We are only interested in cases where a salient reading of the sentences frames the AI entity as having agency, intention, autonomy, or mental or cognitive states. This also applies to cases in which the language is imprecise or colloquial. For example 'raise concerns' instead of 'give rise to concerns', or 'established itself' instead of 'has been established'. As a rule, if the original intention of the author is clear, we do not consider such phrases as anthropomorphizing.
5. When answering Q4, to discern whether a potentially anthropomorphic lexical unit is actually anthropomorphic in context, it is necessary to disambiguate the meaning of the word using cues given in the wider context of the sentence, e.g. the other arguments of a verb. For example, to discern the sense of create, it is necessary to consider the arguments of the verb — i.e. what is being created. The word create is not inherently anthropomorphic, but 'creating French poetry' is. An AI entity is framed as a tool in context not only when it has the thematic role of tool or instrument, but also if the context includes references to training data, parameters or other technical properties. It could also be framed as an inanimate object by other means, such as being described as built, developed or designed.
6. In anthropomorphic sentences, there may be coreference with the inanimate pronoun (it). We ignore this, as a sentence can be highly anthropomorphic yet still refer to the AI entity as 'it' – we would like to be able to evaluate the systems in these cases as well. Our goal is to provide various examples of anthropomorphic language and find semantic patterns in the terminology that is used in AI discourse. We are aware that in an MLM-based classification such as the kind that is being evaluated, the inanimate pronouns will contribute to a reduced or negative valuation.
7. It happens that an AI entity is framed simultaneously as an inanimate object and as having consciousness or awareness. Such cases are to be labeled as inconclusive.
8. Since different structures often overlap within a sentence, the annotation guidelines assume that the sentences have been pre-sorted into the classes of anthropomorphic structures as described above. Indeed, the data collection relied on POS tagging and dependency parsing using SpaCy to automatically provide initial classification. As a general rule of thumb, we keep a sentence in the initial set it was sorted into if that specific linguistic feature significantly contributes to the anthropomorphism in the sentence. For the inter-annotator agreement (IAA) labeling, the suspicious lexical unit is not known, but this is not important – the labeling is sentence-based, not token-based, so it is the final score that counts. This is consistent with the fact that it is not always easy to discern the contributing factors for anthropomorphism, especially when this is determined on the basis of the contextual meaning of the word.

Appendix D: Evaluation

D.1 Supplemental Results

		***AnthroScore* conclusive only**			***AnthroScore* all cases**		
Masking	**Category**	Precision	Recall	F1-Score	Precision	Recall	F1-Score
AnthroScore	verb subjects pos.	1.000	0.145	0.254	0.800	0.145	0.246
	verb subjects neg.	0.581	0.877	0.699	0.467	0.877	0.610
	verb subjects inc.	-	-	-	0.357	0.303	0.328
	verb objects pos.	1.000	0.125	0.222	0.700	0.125	0.212
	verb objects neg.	0.645	0.984	0.779	0.545	0.984	0.702
	verb objects inc.	-	-	-	0.292	0.259	0.275
	adjectives pos.	1.000	0.114	0.204	0.833	0.114	0.200
	adjectives neg.	0.544	0.956	0.694	0.457	0.956	0.619
	adjectives inc.	-	-	-	0.167	0.059	0.087
Minimal entity	verb subjects pos.	0.909	0.179	0.299	0.833	0.179	0.294
	verb subjects neg.	0.560	0.689	0.618	0.438	0.689	0.535
	verb subjects inc.	-	-	-	0.262	0.333	0.293
	verb objects pos.	0.609	0.241	0.346	0.452	0.241	0.315
	verb objects neg.	0.559	0.508	0.532	0.478	0.508	0.493
	verb objects inc.	-	-	-	0.180	0.333	0.234
	adjectives pos.	0.571	0.154	0.242	0.500	0.154	0.235
	adjectives neg.	0.482	0.574	0.524	0.380	0.574	0.458
	adjectives inc.	-	-	-	0.121	0.190	0.148

Table D.1: Side-by-side comparison of AnthroScore's performance per class on the multiclass sets in terms of macro-averaged precision, recall, and F1-scores, comparing only positive and negative gold labels, versus positive, negative and inconclusive cases in gold, using both masking strategies.