Annotation guidelines: anthropomorphic language in AI research

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.

Instructions

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.

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. A detailed taxonomy of cognitive capacities and corresponding words is given below (appendix C).

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.

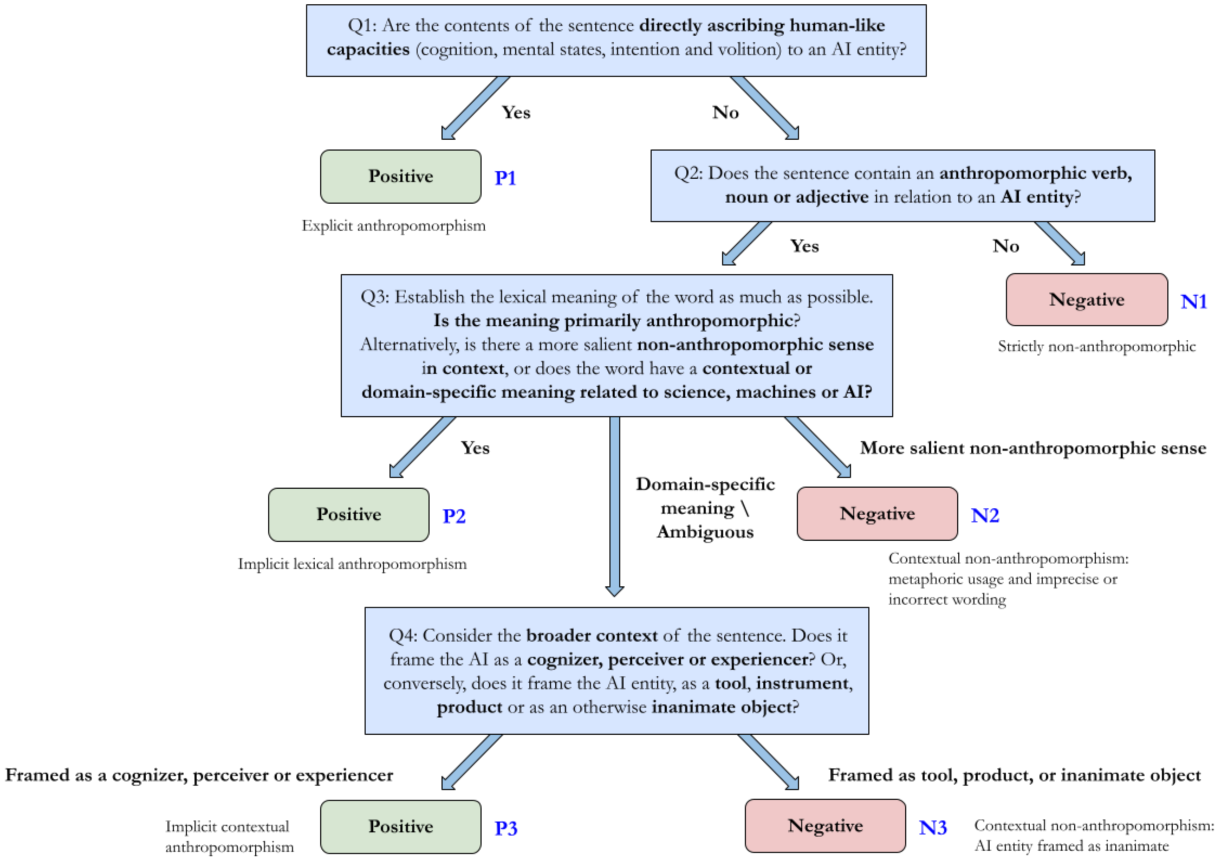
Q4: Does it contribute to anthropomorphism in other ways, framing the AI entity as a cognizer, perceiver or experiencer? On the contrary, the broader context explicitly frames the AI entity as a product, tool or machine lacking agency, and highlights its non-human status?

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.

Appendix A: Decision tree


Q1: Are the contents of the sentence directly ascribing human-like capacities (cognition, mental states, intention and volition) to an AI entity?
Yes
No
Positive
P1
Explicit anthropomorphism
Q2: Does the sentence contain an anthropomorphic verb, noun or adjective in relation to an AI entity?
Yes
No
Negative
N1
Strictly non-anthropomorphic
Q3: Establish the lexical meaning of the word as much as possible. Is the meaning primarily anthropomorphic? Alternatively, is there a more salient non-anthropomorphic sense in context, or does the word have a contextual or domain-specific meaning related to science, machines or AI?
Yes
Domain-specific meaning \ Ambiguous
More salient non-anthropomorphic sense
Positive
P2
Implicit lexical anthropomorphism
Negative
N2
Contextual non-anthropomorphism: metaphoric usage and imprecise or incorrect wording
Q4: Consider the broader context of the sentence. Does it frame the AI as a cognizer, perceiver or experiencer? Or, conversely, does it frame the AI entity, as a tool, instrument, product or as an otherwise inanimate object?
Framed as a cognizer, perceiver or experiencer
Framed as tool, product, or inanimate object
Implicit contextual anthropomorphism
Positive
P3
Negative
N3
Contextual non-anthropomorphism: AI entity framed as inanimate

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.

Appendix B: Notes and clarifications

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 modelling 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 labelled 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 erroneous. 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 co-reference 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) labelling, 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 towards the measuring of the IAA. 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 C: A taxonomy of anthropomorphic capacities

The chart below provides a taxonomy of anthropomorphic capacities that could be attributed to AI. The corresponding lexical items are not exhaustive by any means – their purpose is to provide a general mapping between the type of descriptions that often accompany AI entities in text, and the cognitive and mental faculties that are associated with them. This taxonomy is based on an expansion of the taxonomy of anthropomorphic language of AI synthetic text outputs provided by DeVrio et. al (2025), as well as an analysis of the verbs and adjectives that were present in the sentences collected for the evaluation.

CONCEPTUAL THOUGHT AND MENTAL STATES Hypothesizes, theorizes, and imagines something. Anticipates, guesses or predicts something about the world. (exclude model predictions and guesses)	think, expect, hope, guess, predict, dream, imagine, believe (v) (self-)aware, cognizant (a)
KNOWLEDGE AND AWARENESS Has factual knowledge about and experience in the world, or memories of things that happened. As a result, has an ontology of things, and can identify, classify, and categorize.	know, remember, recognize, memorize, forget, identify, classify, differentiate, distinguish (v) knowledge (n)
REASONING AND UNDERSTANDING Reasons, rationalizes, analyses, makes sense of something. Understands, considers, weighs options, takes something into consideration or account. As a result, can be considered smart, clever, and intellectual.	deduce, conclude, rationalize, reason, consider, decide, (mis)understand, (mis)interpret, take into account, analyse, infer (v) smart, intelligent (a)
JUDGMENT Has an opinion, makes decisions and choices, gives advice, has a preference, evaluates, imparts judgment. Has a concept of morality and ethics, knows right and wrong. (exclude model evaluations)	advise, prefer, select, choose, decide, determine, resolve (v)
PLANNING AND DECISION MAKING Plans, strategizes, sets a goal, devises a method, game plan or scheme, can also struggle or experience difficulties.	plan, coordinate, strategize, come up with a plan, struggle (v)
AGENCY AND AUTONOMY Takes action, able to autonomously carry out a goal – used in a way that attributes agency and control over the action and situation. As a result, can follow or break rules to achieve and accomplish a goal.	cheat, follow or break rules, achieve (v) autonomous, independent, creative (a)

COMMUNICATION Communicates, teaches or explains, Similarly, can also learn or be at the receiving end of communication or explanation. (exclude model training and machine learning)	communicate, talk, speak, tell, explain, teach, learn, ask (v) communicative (a)
ACTIVE SUPPORT Recommends, makes a suggestion or an offer. Actively and directly helps, aids and assists by employing skills to solve a problem. Can have expertise in certain domains. Related to communication.	suggest, aid, help, contribute (v) responsible (a) feedback, insights (n) expert, advisor (a)
CANDIDNESS Capable of, or has a concept of honesty or dishonesty, truthfulness or deception, revealing or concealing. As a result, can be trustworthy or untrustworthy, reliable or unreliable. Related to judgment.	trust, believe, lie (v) (un)truthful, deceitful (a)
AFFABILITY Acts as a friend or as an enemy, companion or adversary, collaborator or rival. As a result can act benevolent or malevolent, friendly or hostile, kind or mean, collaborative or competitive.	collaborate, attack, manipulate, insult, deceive (v) thoughtful, attentive, friendly (a) collaborator, partner, attacker, adversary (n)
POWER AND RELATIONSHIPS Plays a role in a relationship dynamic – romantic or platonic, superior (boss, manager, teacher) or subordinate (employee, student). Related to communication and affability.	teach, tutor, supervise, manage, oversee (v) manager, employee, teacher, tutor, student, companion, lover (n)
EMOTIONS Empathizes, sympathizes, displays emotions, experiences pain or pleasure.	experience, emote (v) sensitive, vulnerable (a)
SELF-EXPRESSION AND PERCEPTION OF DEEPER MEANING Partakes in activities related to self-expression such as art and storytelling, humor and jokes. Perceives beauty and aesthetics. Has a deeper understanding of meaning, purpose, and context. Related to emotions, awareness and conceptual thought.	create poetry, create art, write, compose, paint, sing, dance (v) creative, artistic, funny (a) artist, poet, humor, irony (n)
SENSORY PERCEPTION Receives and processes sensory input and feedback from the environment, picks up visual/auditory/sensory (olfactory) cues.	see, hear, perceive, feel, sense (v) blind, deaf (a)