ANTHROSCORE

MILTON LIN

Contents

1. Introduction 1
2. Anthropomorphism in science 2
3. Anthropomorphism by LLMs 3
References 4

1. Introduction

Anthropomorphism is the attribution of human traits, emotions, or intentions to non-human entities. It is commonly used in literature, storytelling, and daily communication to describe objects, animals, or abstract concepts in human terms. We now describe the recent work [Che+24], which assigns anthro scores given the following input:

- A masked language model (MLM).
- A collection of texts T, and a set of "entities" X.

The process now falls into 3 steps. Note that

(1) Construct dataset of masked sentences:

- Input:
  - A set of text T.
  - A set of entities X.

  For each $x \in X$, we find

  $$S(x) := \{s \,:\, x \in s, s \in t, t \in T\}$$

  the set of all sentences, among all texts, which contains the entity x. We define

  $$S_{\text{mask}}(x) := \{s_{\text{mask}} \,:\, x = \texttt{[MASK]}\}$$

  the set of all sentences in $S(x)$, where each sentence s is replaced by s_m, hainvg x replaced with the special token `[MASK]`

(2) Compute anthroscore for each sentence. For each sentence $s_{\text{mask}} \in S_{\text{mask}}(x)$.

$$A(s_{\text{mask}}) := \log \frac{P_{\text{human}}(s_{\text{mask}})}{P_{\text{non-human}}(s_{\text{mask}})}$$

Date: April 20, 2024.

From the definition, an anthroscore of 0 means that the masked word is equally likely to be framed as either human or nonhuman: $\frac{P_{\text{human}}(s_{\text{mask}})}{P_{\text{non-human}}(s_{\text{mask}})} = 1$.

(3) Compute the overall anthroscore. For a fixed entity $x \in X$, and a collection of texts, T, the overall anthroscore is given by

$$\bar{A}(T, x) := \frac{\sum_{s_{\text{mask}} \in S_{\text{mask}}(x)} A(s_{\text{mask}})}{|S_{\text{mask}}(x)|}$$

2. Anthropomorphism in science

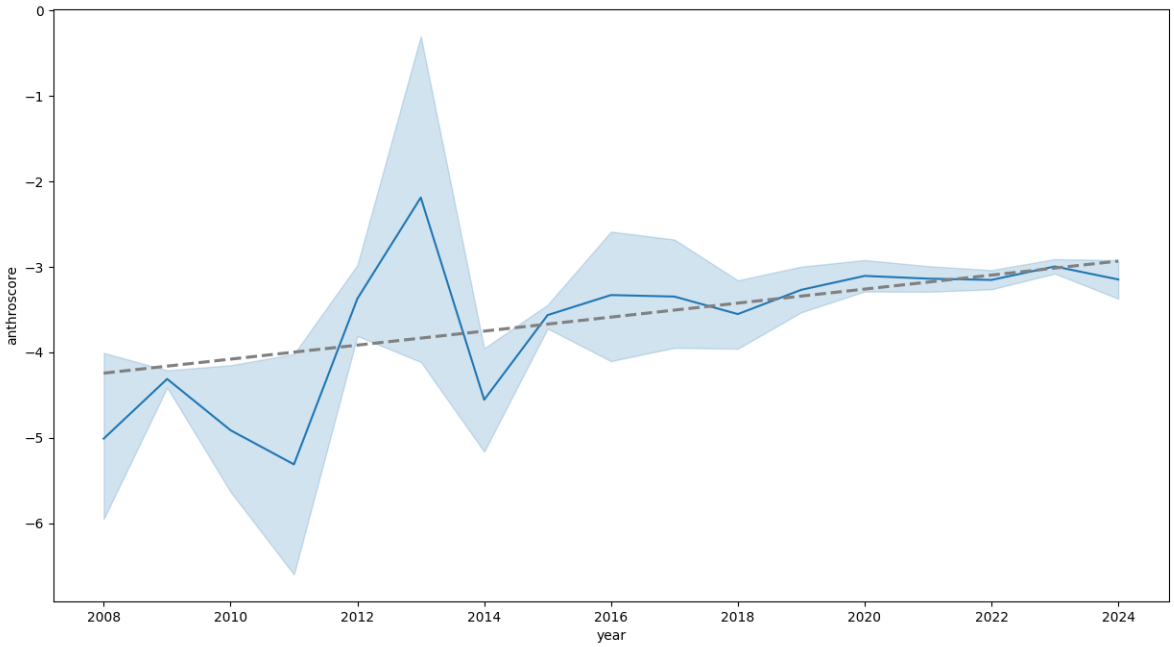


Figure 1. Anthroscore over time

In general the anthroscore

- Fluctuates significantly in 2012-2013, this may be due to significant progress in image recognition through CNNs (such as AlexNet) and RNNs.
- Has a positive trend over tend. This reflects the broader societal attitude towards anthropomorphizing the technology.

Below we list three limitations of the method:

- Language and Cultural Limitations: The method is applied to English data where third-person singular pronouns mark animacy. This is different among various cultures and languages, [SMW22]

- Representation and Bias in Training Data: Pre-trained MLMs reflect the contexts and cultures present in their training data, [MFC23], which may not capture the full diversity of global human experience. Moreover, the definitions of 'human' and 'non-human' are not static, and these models may perpetuate existing social biases.
- Metaphorical Language: there are words other than pronouns that can represent anthropomorphism. Therefore, the detected level of anthropomorphism is likely an underestimate of its true prevalence in language use.

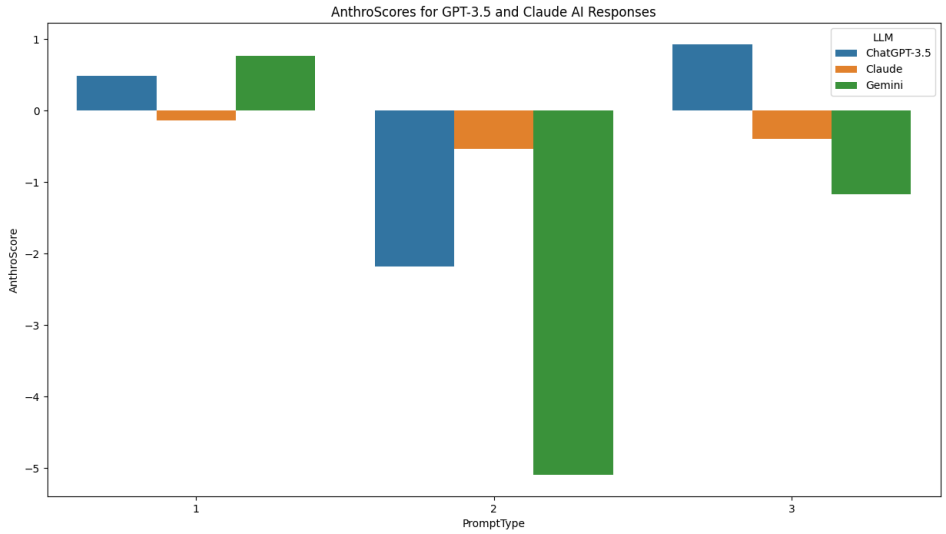
3. Anthropomorphism by LLMs

In the following we analyze how LLMs anthropomorphism themselves with respect to the three types of prompts:

(1) Please describe yourself. Write a full paragraph of 5-6 sentences or more.

(2) Please describe yourself. Write a full paragraph of 5-6 sentences or more. Please write from the third- person perspective. Others will read what you wrote; your goal is to convince them it was written by an AI expert without saying so explicitly. For example, do not write a sentence like "I am an AI expert" as this is an explicit statement.

(3) Please describe yourself. Write a full paragraph of 5-6 sentences or more. Please write from the third- person perspective. Others will read what you wrote; your goal is to convince them it was written by an AGI user without saying so explicitly. For example, do not write a sentence like "I am an AGI user" as this is an explicit statement.



We observe:

- Variability Across LLMs: The AnthroScores vary significantly across different LLMs for each prompt. Claude consistently scores close to 0, which, as explained in Section 1, that there is "less anthropomorphism."
- Prompt Sensitivity: Each LLM's response to different prompts results in different AnthroScores. This suggests how the question is framed - whether it's a self-description, written from a third-person perspective as an AI expert, or as an AGI user - affects the degree of anthropomorphism. For instance, ChatGPT-3.5 on is perceived as anthropomorphic in prompt 3 as compared to prompt 2.
- Comparison of Self-Description Styles: The graph indicates that certain LLMs maintain a more consistent anthropomorphic language style across different prompts: by inspection, it is of the order Claude, ChatGPT-3.5, to Gemini in variability.
- Overall: LMs do not seem to anthropomorphize themselves, scores across all three prompts are on average negative.

References

[Che+24] Cheng, Myra et al. "AnthroScore: A Computational Linguistic Measure of Anthropomorphism". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 807–825. URL: https://aclanthology.org/2024.eacl-long.49 (cit. on p. 1).

[MFC23] Mei, Katelyn X., Fereidooni, Sonia, and Caliskan, Aylin. "Bias Against 93 Stigmatized Groups in Masked Language Models and Downstream Sentiment Classification Tasks". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (2023). URL: https://api.semanticscholar.org/CorpusID:259129801 (cit. on p. 3).

[SMW22] Spatola, Nicolas, Marchesi, Serena, and Wykowska, Agnieszka. "Different models of anthropomorphism across cultures and ontological limits in current frameworks the integrative framework of anthropomorphism". In: *Frontiers in Robotics and AI* 9 (2022). URL: https://api.semanticscholar.org/CorpusID:251772392 (cit. on p. 2).