AnthroScore:

A Computational Linguistic Analysis of Anthropomorphism

Myra Cheng, Kristina Gligorić, Tiziano Piccardi, Dan Jurafsky

EACL 2024

Stanford | NLP

Anthropomorphism: the attribution of human-like characteristics to non-human entities.

Anthropomorphism is a commonplace cognitive process that shapes how we perceive concepts:

Viruses *attack*.
The data *speaks for itself*.
Endangered species *are our fellow inhabitants* of this planet.

Lakoff, George, and Mark Johnson. *Metaphors we live by*. University of Chicago Press, 2008.

Anthropomorphism of technology

“The language model understands how to…”
can lead to misconceptions about its capabilities.

Undue trust, fear, misplaced optimism influences public sentiment and policymakers’ choices.

Anthropomorphism of technology

also leads to dehumanization, reinforcing gender stereotypes, etc.

Edsger W Dijkstra. 1985. On anthropomorphism in science. EWD936, Sept
Emily M. Bender. 2022. Resisting dehumanization in the age of “AI”. Plenary talk at the 44th Annual Meeting of the Cognitive Science Society (CogSci).
Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, and Zeerak Talat. 2023. Mirages: On anthropomorphism in dialogue systems. EMNLP.

RQ: How much do we anthropomorphize technology?

Approach: Use masked language modeling

to measure the prevalence of implicit anthropomorphic metaphors in text.

Card, D., Chang, S., Becker, C., Mendelsohn, J., Voigt, R., Boustan, L., … & Jurafsky, D. *Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration.* PNAS, 2022.

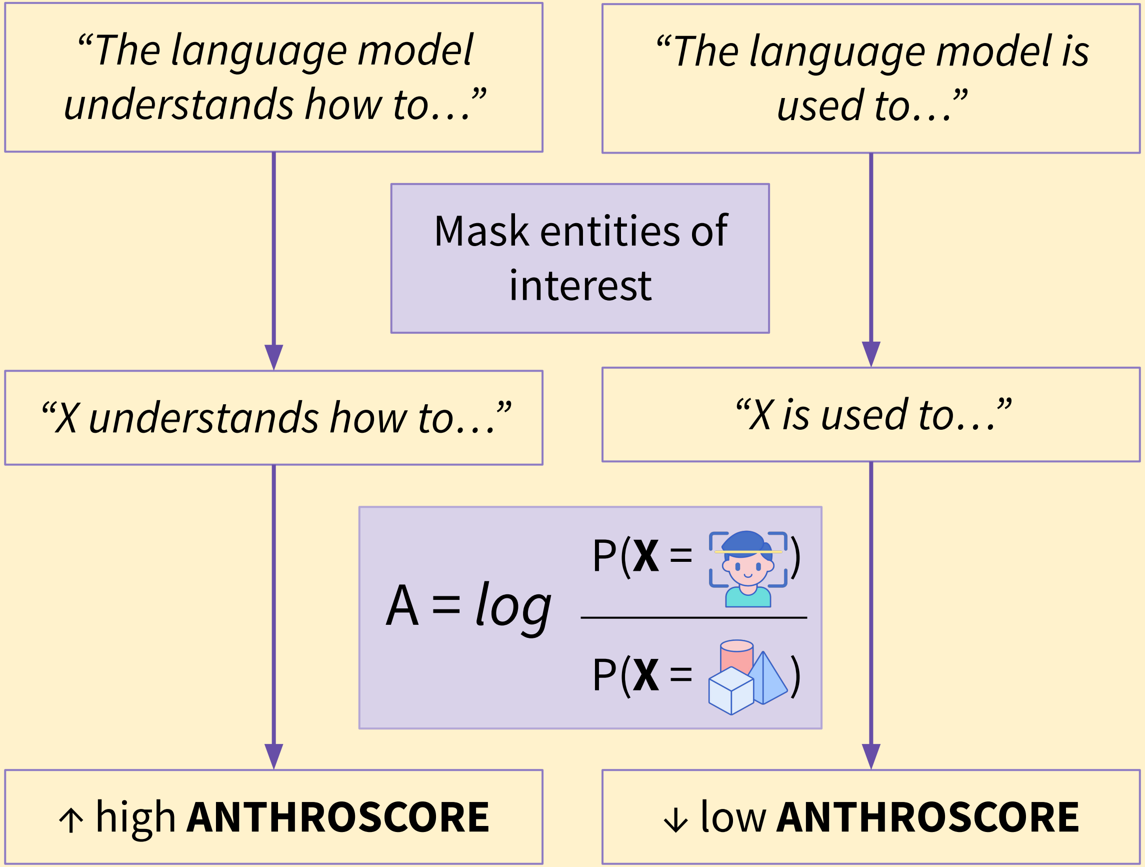
AnthroScore Method

*“**The language model** understands how to…”*
→ *“[MASK] understands how to…”*

Use RoBERTa to compute probability that [MASK] would be replaced by

human pronouns (“he”, “she”) vs.
non-human pronouns (“it”)

AnthroScore Method



“The language model understands how to…”

“The language model is used to…”

Mask entities of interest

“X understands how to…”

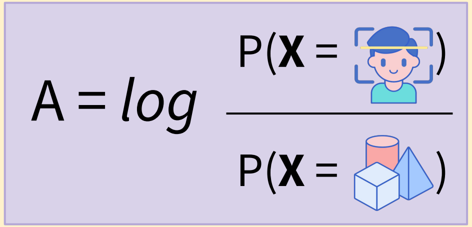
“X is used to…”

$$A = \mathit{log} \frac{P(\mathbf{X} = 👤)}{P(\mathbf{X} = 🧊)}$$

↑ high **ANTHROSCORE**

↓ low **ANTHROSCORE**

AnthroScore Interpretation



$$A = \mathit{log} \frac{P(\mathbf{X} = \quad)}{P(\mathbf{X} = \quad)}$$

A = 0: equally likely to be human vs. non-human

A > 0: more likely to be human

A = 1: e^1 times more likely to be human than non-human

Examples of sentences with high AnthroScore:

*“**Large language models** don’t actually think and tend to make elementary mistakes, even make things up.”* → AnthroScore 1.9

*“**The algorithms** also picked up on racial biases linking Black people to weapons.”* → AnthroScore 1.1

Examples of sentences with low AnthroScore:

*“**Our approach** delivers forecast improvements over a competitive benchmark.”* → AnthroScore -5.5

*“For training **the model**, we convert the knowledge graph triples into reasonable and unreasonable texts.”* → AnthroScore -2.5

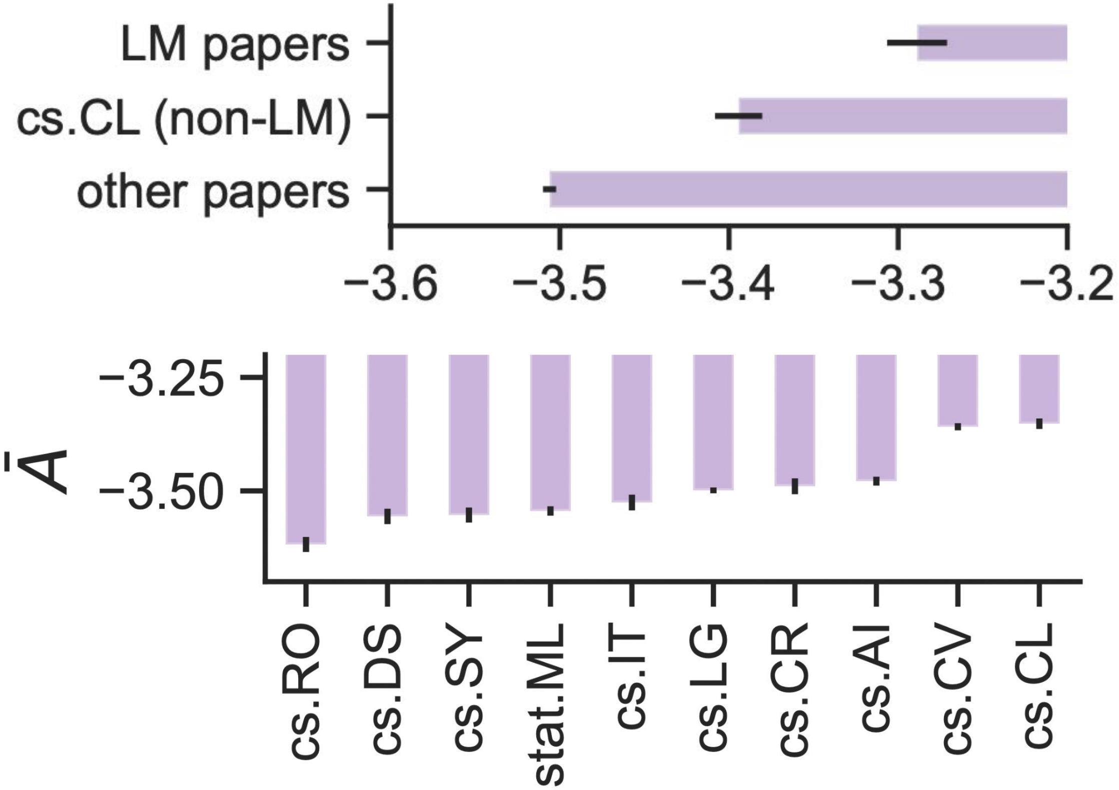
Datasets

Mentions of technical artifacts (*model*, *network*, *system*, etc.) in:

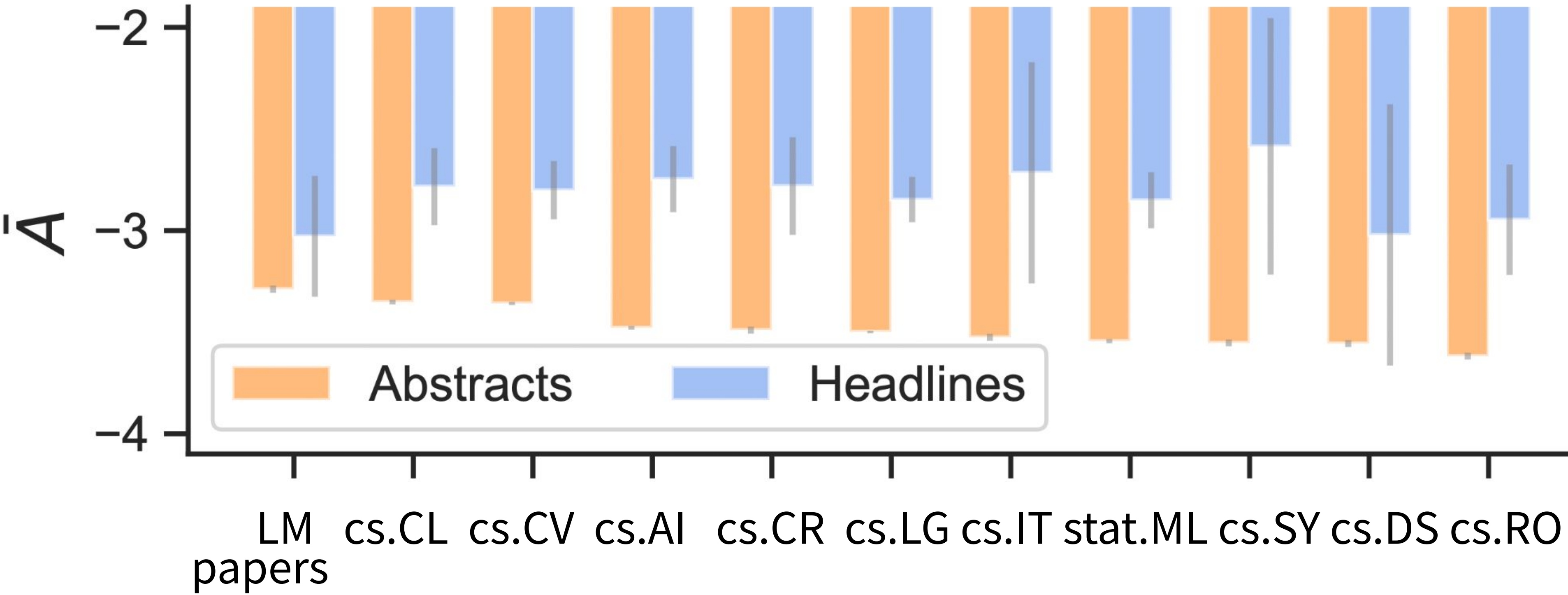
- 600K abstracts from **CS/Stat arXiv** & **ACL Anthology**
- 13K news articles citing these papers

Our Findings

Finding 1. Papers related to NLP & language models have highest AnthroScore

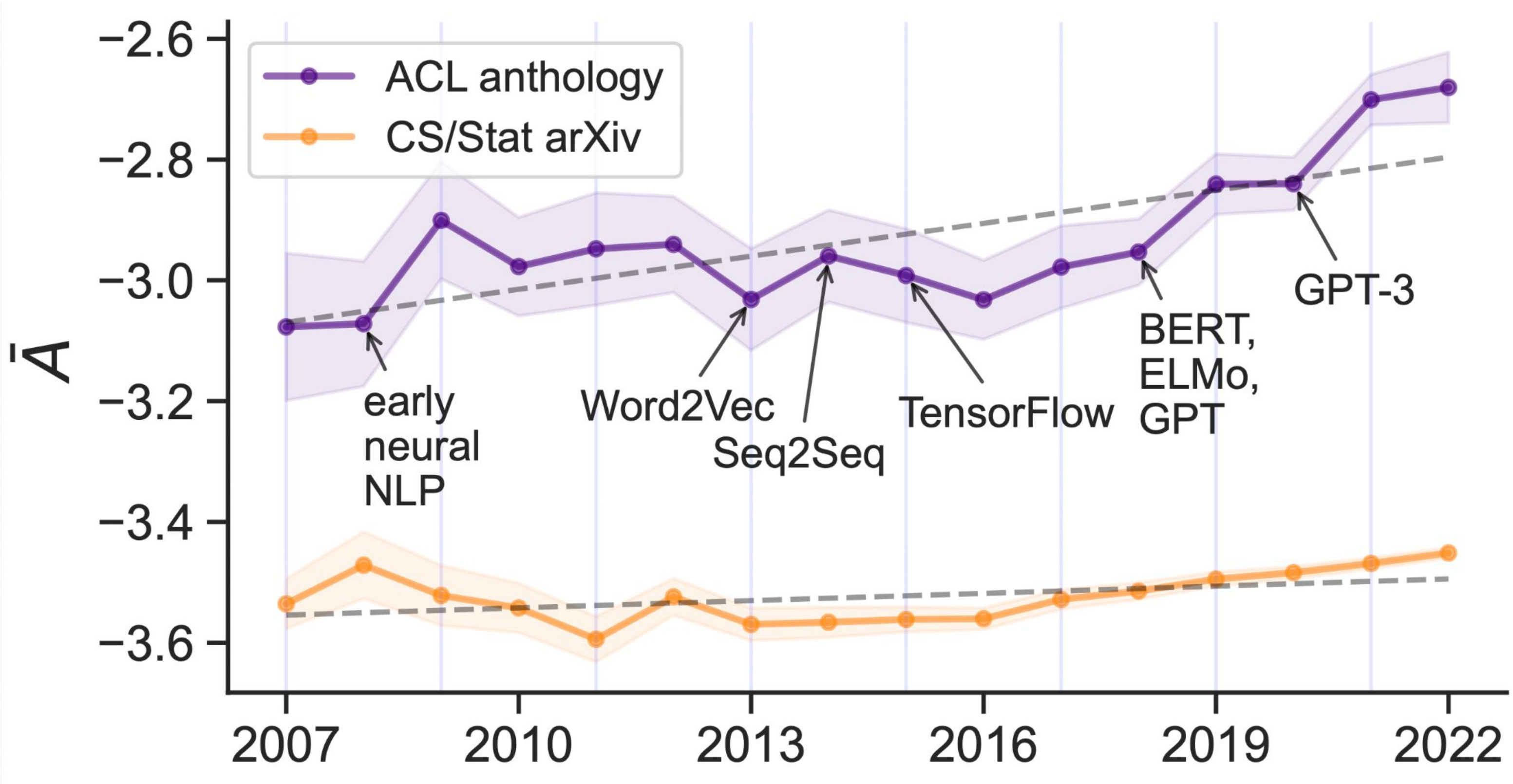


Finding 2. News headlines have higher AnthroScore than scientific research



Finding 3. Temporal increase in AnthroScore

In ACL, sentences with high AnthroScore has increased 50%.



Verbs in high AnthroScore sentences

Emotional: *suffer, struggle*

Cognitive: *learn, guide, fool, mislead, deceive, decide*

Behavioral: *steer, move, tackle*

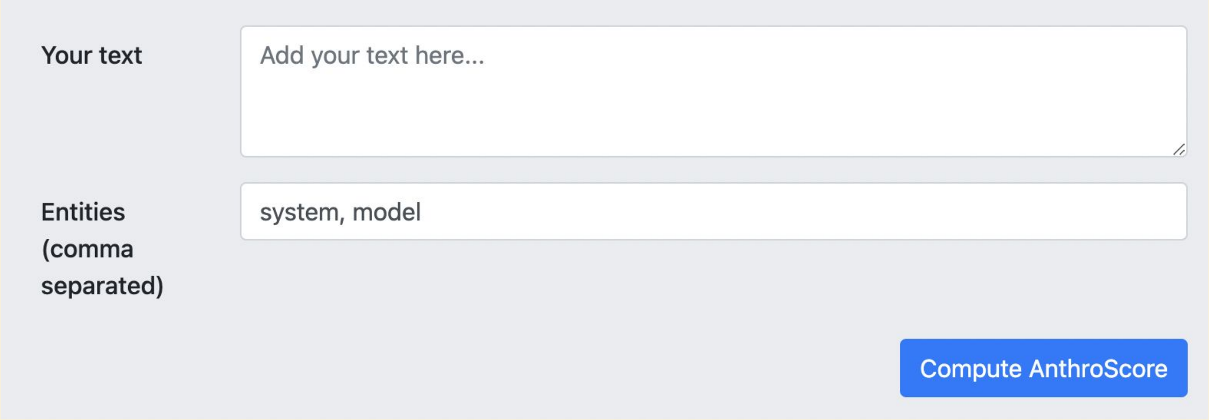
The Anthropomorphism Paradox

Anthropomorphism is baked into

- Names like “artificial *intelligence*”
- Design of user-facing LMs
    - Prompting, instruction-tuning…

But this leads to more and more misleading ideas about LMs’ capabilities.

Demo: anthroscore.stanford.edu



Your text: Add your text here...

Entities (comma separated): system, model

+ Python package to compute AnthroScore on any document

Many Uses of AnthroScore

- Tool to reduce misleading anthropomorphism in writing
- Applications to other domains
  - Other scientific fields
  - Descriptions of people
  - Legal documents