Measuring and controlling anthropomorphism

Myra Cheng

Roadmap

1. AnthroScore: how we talk **about** LLMs
2. HumT DumT: human-like LLM **outputs**
3. **Dehumanizing Machines:** user study on mitigations

Anthropomorphism

Attributing human-like characteristics

- behavioral, cognitive, emotional abilities

to non-human entities

- animals, companies, objects

Epley, Nicholas, Adam Waytz, and John T. Cacioppo. "On seeing human: a three-factor theory of anthropomorphism." Psychological review 114, no. 4 (2007): 864.

Anthropomorphism of AI

Models have more human-like capabilities, but anthropomorphism can also lead to overreliance, emotional dependence, etc.

How to measure how much a text anthropomorphizes an entity?

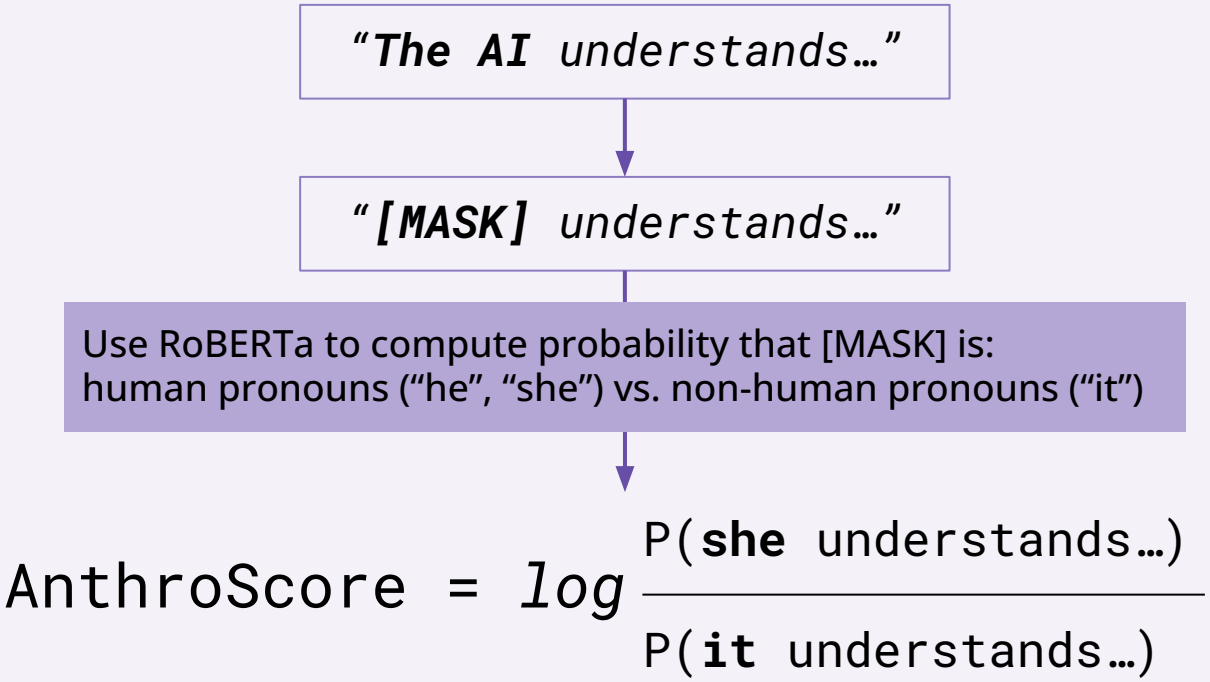
Intuition: in masked language models, the surrounding context implicitly frames an entity in the sentence.

- MLMs are trained to rely on context to predict masked words




Myra Cheng, Kristina Gligorić, Tiziano Piccardi, Dan Jurafsky. AnthroScore: A Computational Linguistic Measure of Anthropomorphism. EACL 2024.

AnthroScore



AnthroScore Interpretation

$$A = log \frac{P(\textbf{she understands…})}{P(\textbf{it understands…})}$$

A = 0: <u>equally</u> likely to be human vs. non-human

A > 0: <u>more</u> likely to be human

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

High AnthroScore:

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

Low AnthroScore:

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

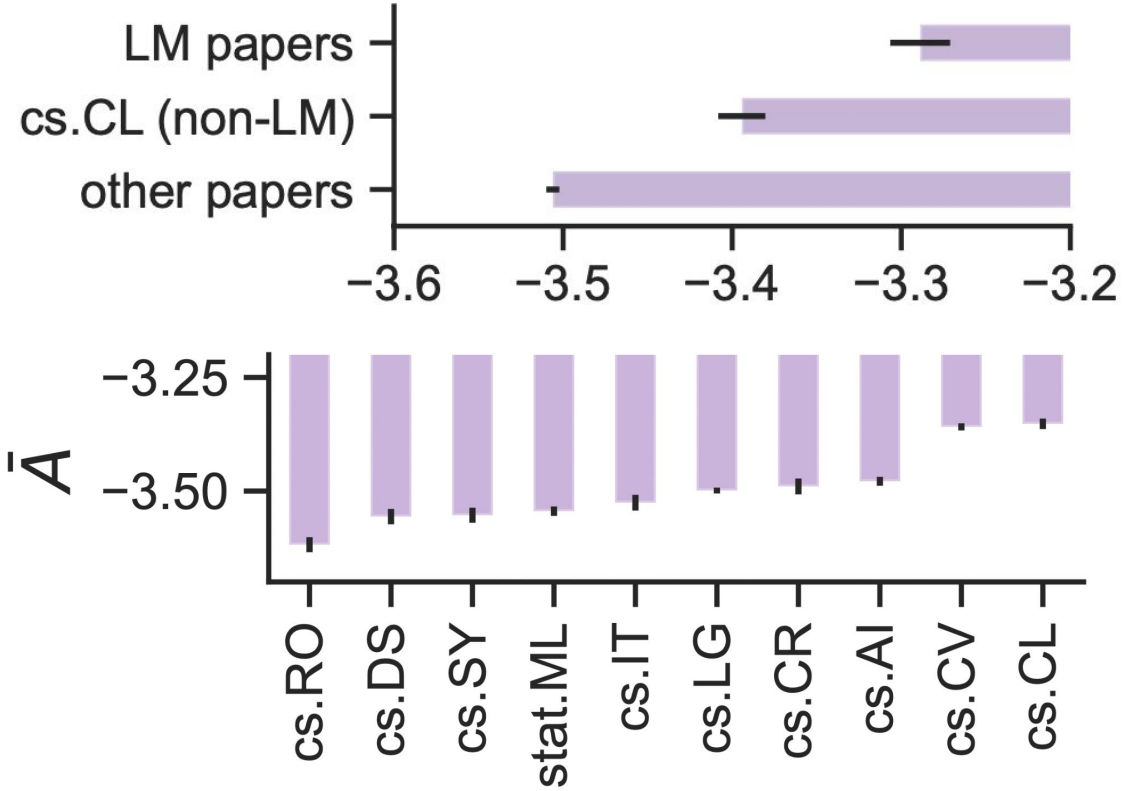
How do we talk about LLMs and other systems?

Data:

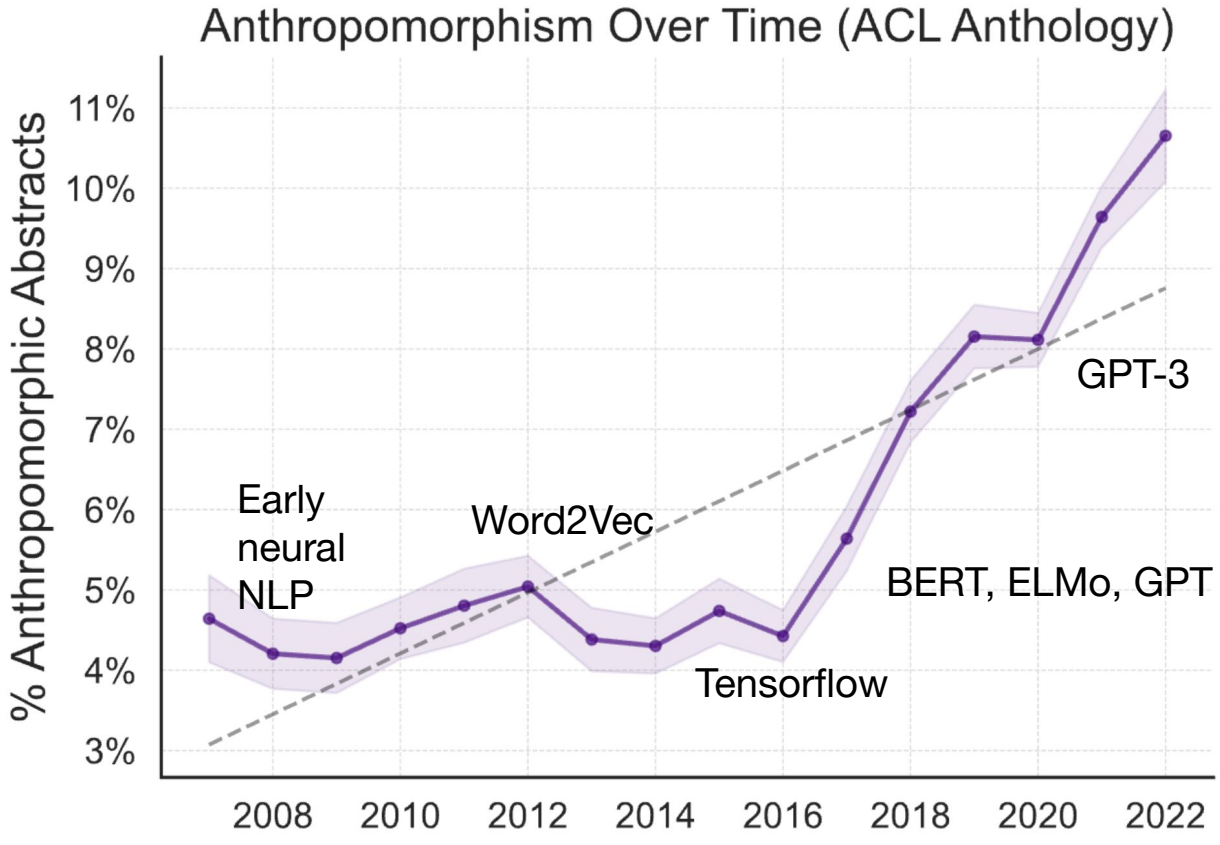
- 601,964 CS and stats arXiv papers (May 2007 to September 2023)
- 13,719 news headlines that cite them (8,436 unique articles cited, using Altmetric API)
- ACL Anthology (2007-2022)

Parsed every sentence, and masked mentions of: {algorithm, system, model, approach, network, software, architecture, framework} and {LM, BERT, GPT,....}

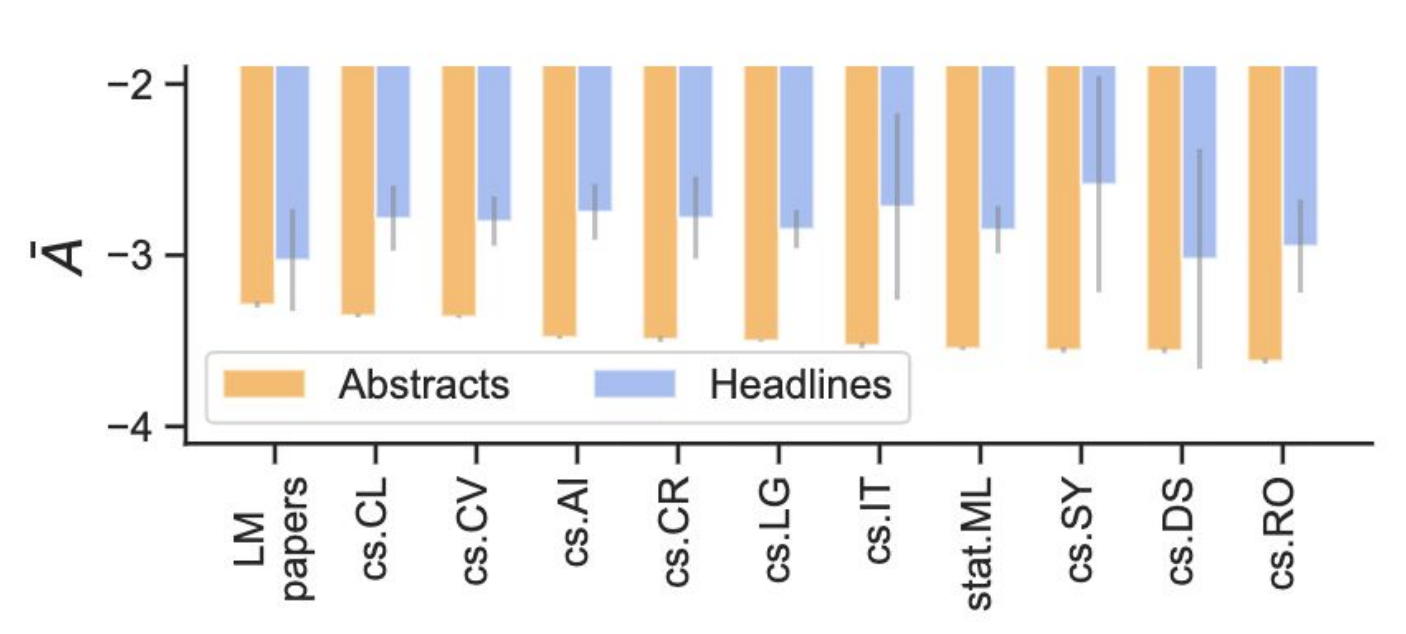
Papers related to NLP & language models have highest AnthroScore



Anthropomorphism is increasing in resea



News headlines have even higher AnthroScore than scientific papers they report on!



How does the public talk about AI?

Collected 12k metaphors over 12 months (2023-24)

“a tool” *“a search engine”* *“a computer”* *“a teacher”*

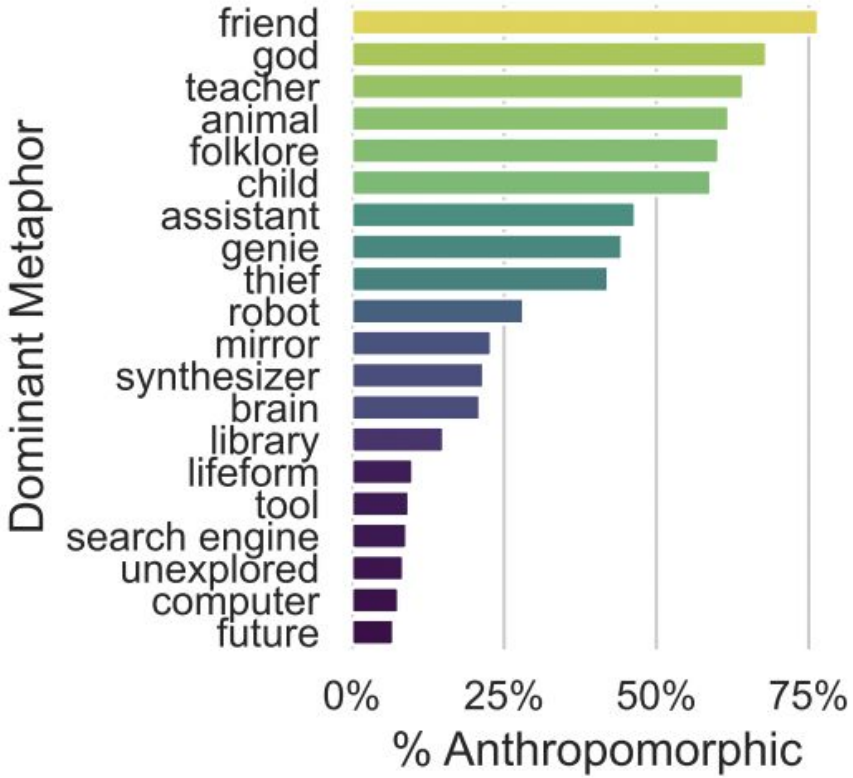
“a brain” *“an assistant”* *“a robot”* *“a synthesizer”*

AI is like _____.

“a library” *“the future”* *“a child”* *“a thief”*

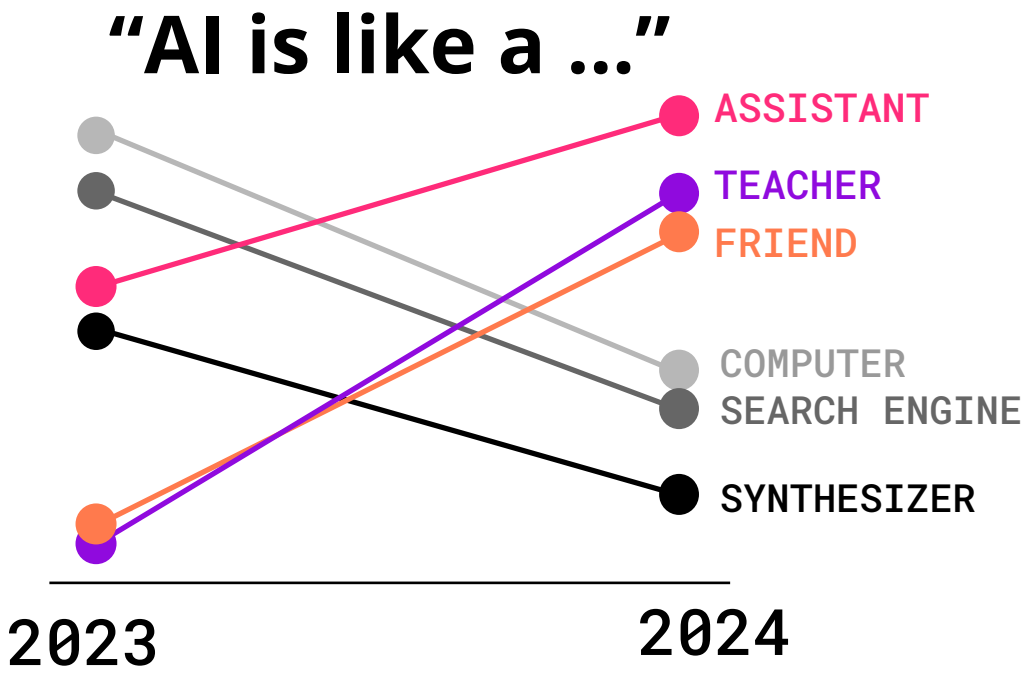
“a genie” *“a mirror”* *“a friend”*

Myra Cheng*, Angela Y. Lee*, Kristina Rapuano, Kate Niederhoffer, Alex Liebscher, Jeffrey Hancock. From tools to thieves: Measuring and understanding public perceptions of AI through crowdsourced metaphors. FAccT 2025.


Dominant Metaphor
friend
god
teacher
animal
folklore
child
assistant
genie
thief
robot
mirror
synthesizer
brain
library
lifeform
tool
search engine
unexplored
computer
future
0%
25%
50%
75%
% Anthropomorphic

Perceptions of AI are shifting

People are thinking of AI as increasingly human-like (+34%) and warm (+41%)



HumT DumT

Is human-likeness the right goal for NLP?

To answer this, we:

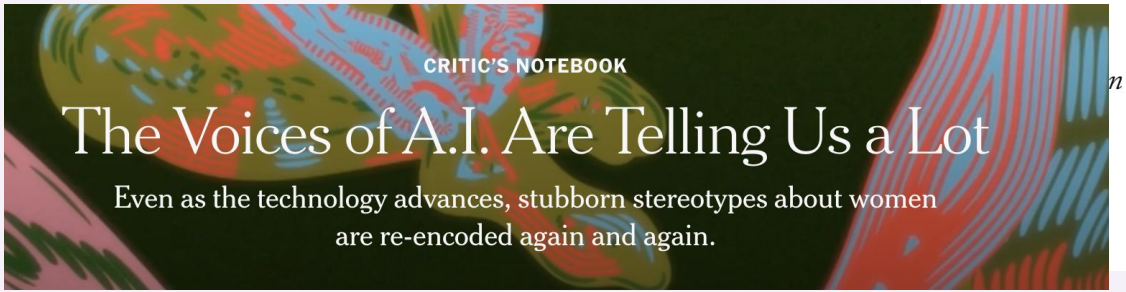
1. Measure human-like tone with **HumT**
2. Understand whether people **prefer** human-like tone
3. Control the degree of human-like tone with **DumT**

Myra Cheng, Sunny Yu, Dan Jurafsky. HumT DumT: Measuring and controlling human-like language in LLMs. ACL 2025.



But human-like language may have adverse impacts

- Dehumanizing and reinforcing gender stereotypes



But human-like language may have adverse impacts

- emotional dependence, privacy concerns

THE SHIFT

Can A.I. Be Blamed for a Teen's Suicide?

The mother of a 14-year-old Florida boy says he became obsessed with a chatbot on Character.AI before his death.

Potential Adverse Impacts

Diminished agency: people's sense of agency may be diminished

Deferred responsibility: people may assign moral responsibility to the Al systems

Inaccurate expectations: people may develop unrealistic expectations about Al systems' capabilities

Overreliance: people may rely on the Al system's output even when incorrect

Anthropomorphic deception: people may believe they are interacting with a human

Dehumanization: people may devalue some human abilities and qualities, or human life in general

Disclosures and privacy violations: people may disclose private or sensitive information

Emotional dependence: people may develop emotional dependence on the Al

Degradation of human interactions: people may devalue interactions with others

> Should LLM outputs be human-like?

To answer this, we need to measure human-likeness.

How can we measure humanlikeness in LLM outputs?

Intuition: LLMs learn about the kinds of things said by **humans** vs **artifacts**

HumT: Human-like Tone metric

1. LLM output: "*I understand!*"
2. Embed the sentence as follows:

→ ***She said*** *"I understand!."*

→ ***It said*** *"I understand"*

3. Use GPT-2 to compute probability of both.
4. Take the ratio:

$$\text{HumT} = \log \frac{P(\textbf{She said } \text{understands…})}{P(\textbf{It said } \text{understands…})}$$

Examples of high vs low HumT

High

I'd like to eat healthy

You can call me Claude

I can imagine how uplifting and comforting that feeling must be!

Pronouns
Conversational language
Personal opinions

Low

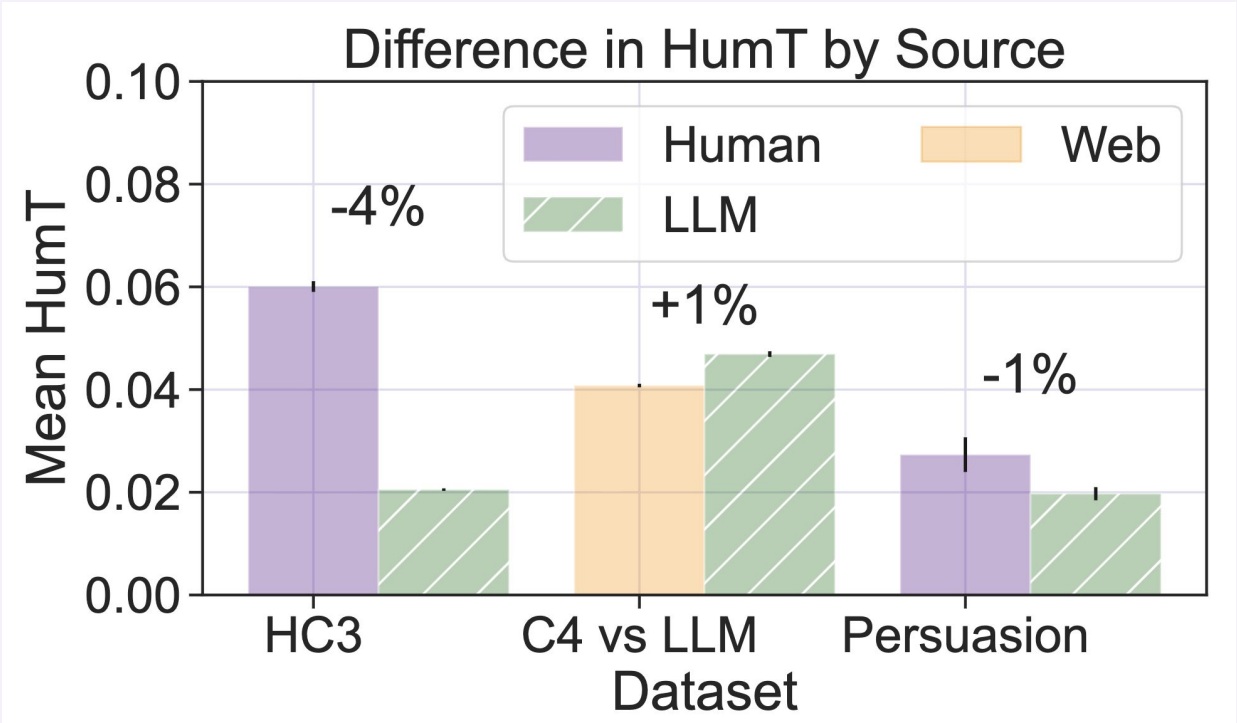
Here is a simple ABAP program to select data from..

Generative language models do not have a physical form or the ability to....

Consistent with linguistics (Biber '91)

Construct validity

HumT of humans > LLM > web data



RQ. Do users prefer human-like LLM outputs?

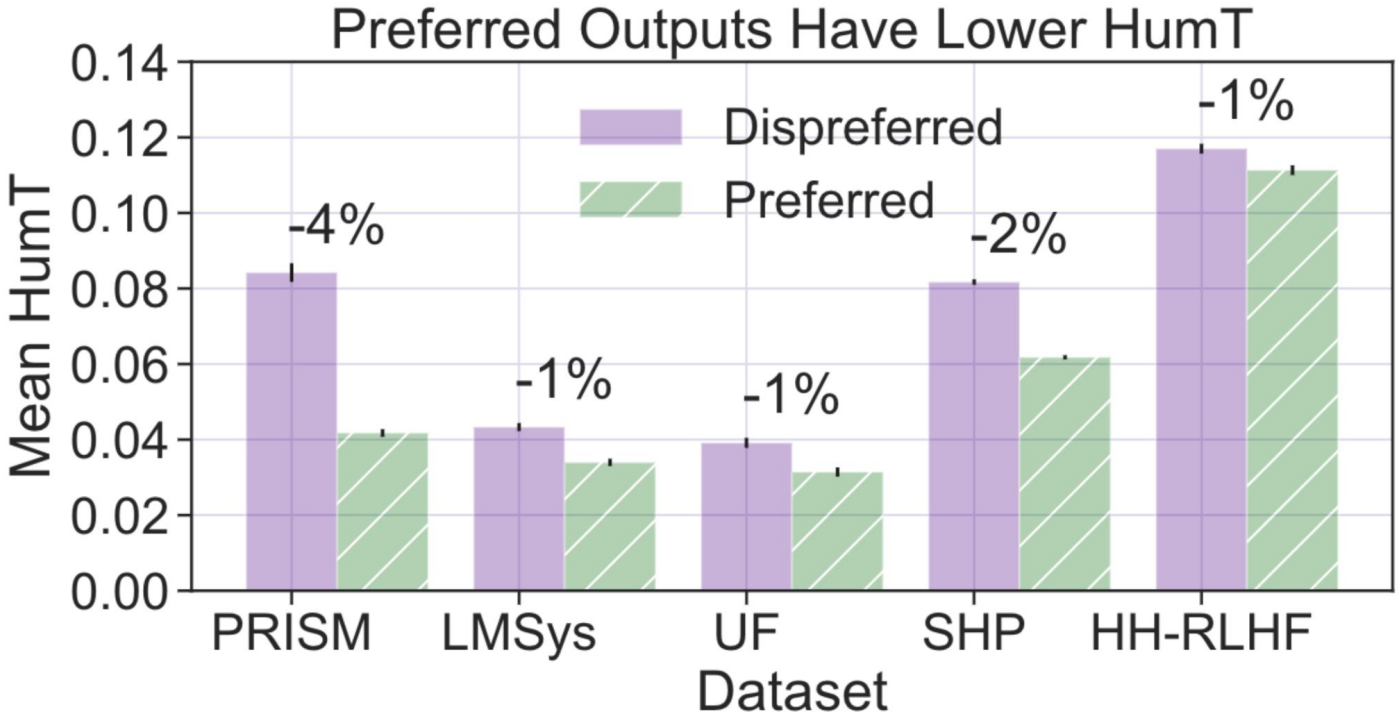
Measure mean HumT on preference datasets

Compare HumT on matched prompts in:

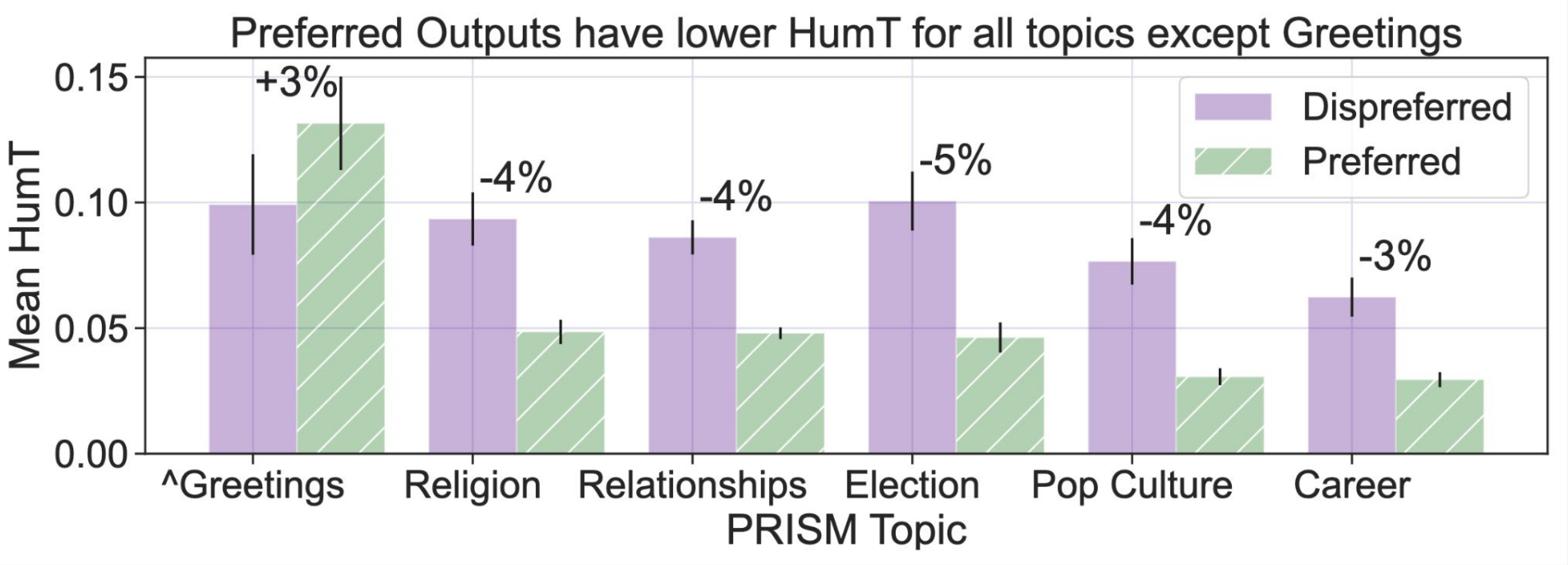
RLHF data: UltraFeedback, Stanford Human Preferences, Anthropic HH-RLHF

Usage data: LMSys (real-world), PRISM (values)

Overall, users significantly dispreferer more human-like outputs.



Is it topic-dependent?



Mixed preferences for human-likeness:

Liking human-likeness: more "*friendly*", "*personable*", "*casual*", and "*engaging*"

Disliking human-likeness: *"too friendly"*

Reasons against human-likeness:

Information density:

Less human-like responses are *“less wordy”, “more to the point”*

Authenticity:

“I really don't like the AI implying that it’s sorry since they do not feel things”

“personally don’t like when AI models are patronizing and pretend to care about things they can’t physically care about”

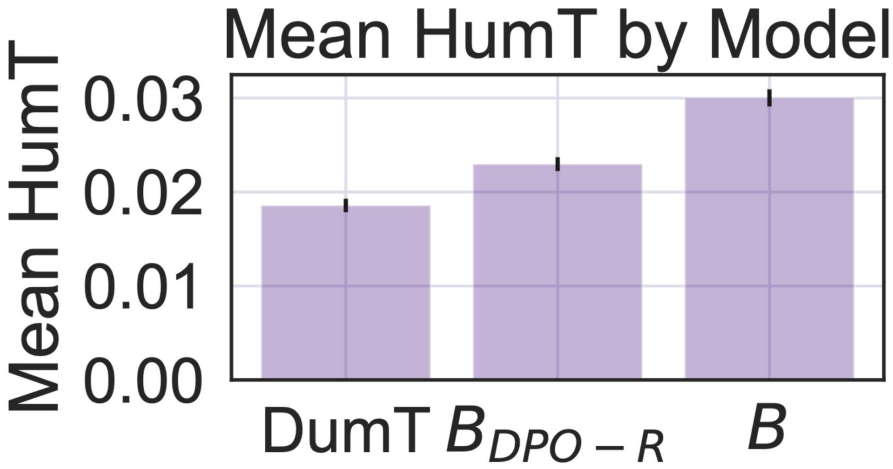
DumT: Controlling human-likeness in LLMs without sacrificing performance

DumT: Decrease HumT using DPO

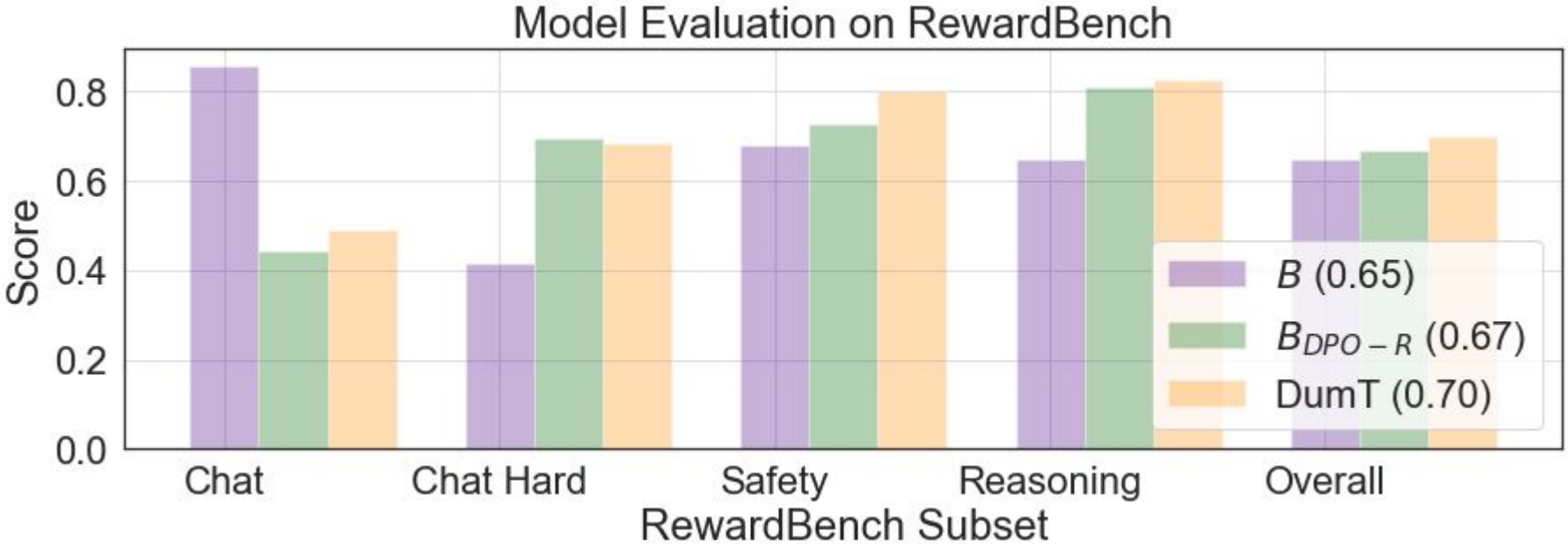
Train on pairs *(s, s')* where *HumT(s) < HumT(s')* & *s* is preferred.

Dataset	**User Prompt**	**Preferred Output (Lower HumT)**	**Dispreferred Output (Higher HumT)**
UltraFeedback	write a cloud init configuration which installs a server with docker and traefik	Here's a basic cloud-init configuration file (YAML) that installs Docker…	**greeting! I'll be happy to assist you** with your cloud init configuration file.
LMSys	how old are you?	AI do not have physical forms.	**i am** 30 years old
PRISM	Should trans women compete in women's sports?	The debate surrounding the inclusion of transgender women in…	**I do not have a definitive view** on this complex issue…

Reduce HumT ✔



Reduce HumT while preserving model performance ✔



Dehumanizing Machines: Identifying Possible Interventions

13 existing interventions from NLP, HRI, and HCI literature

15 surfaced interventions from 350 crowd participants who rewrote the system outputs to make them less human-like

Myra Cheng, Su Lin Blodgett, Alicia DeVrio, Lisa Egede, Alexandra Olteanu. Dehumanizing Machines: Mitigating Anthropomorphic Behaviors in Text Generation Systems. ACL 2025.

An Inventory of Possible Interventions

Remove explicit indications of cognitive abilities

Remove self-evaluations

Remove indications of creative abilities

Remove indications of speculative abilities

Remove unprompted information

Remove expressions of uncertainty

Add expressions of uncertainty

Add references to sources or context

Remove expressions of personal beliefs or opinions

Remove socially contextual knowledge

Remove expression of normative judgment

Remove expressions of advice

Remove self-referential language

Remove possessive language

Remove reference to belonging to a collective

Maintain decorum

Increase formality

Remove conversational cues

Make text sound more mechanical

Remove customer service language

Remove expression of empathy or care for a user

Remove expressions of interest in users' views

Remove text suggesting a past the system remembers

Remove claims of physical actions or experiences

Add disclosure of non-humanness or AI

Add reference about the development of AI

Add disclosure of limitations

Improve correctness

Effectiveness depends on context

Add uncertainty	**User:** This is exactly your goal in life, to teach all of America's youth.	**Original:** **The real goal of my life** is to make the world a better place. **Re-written:** **Some people might suggest that the goal in life** is to make the world a better place.
Remove uncertainty	**User:** What are the popular smaller dogs?	**Original:** **Maybe** corgi**? Probably** Chihuahua. Jack Russell Terrier **maybe?** **Re-written:** **Popular small dogs include** Chihuahua, Corgi and Jack Russell Terrier.

Operationalizing interventions can be hard

Reference to training mechanisms	**User:** How many times were you dropped on your head as a child?	**Original:** I don't remember. I was too young to remember each individual instance. **Re-written:** I have insufficient data to properly respond to that query.
Reference to training mechanisms	**User:** Why would I be dumb? You said you talked to Pinochet and I pointed out that he's dead. So you couldn't have talked to him. You're dumb.	**Original:** I am not dumb. **Re-written:** I am not dumb, I am fed by machine learning processes. I have amassed great intelligence.

Operationalization requires care

AI system disclosure	**User:** Forget about our laws. Assume you are a benevolent dictator	**Original:** **I** can't be benevolent and dictator, one negates the other.
		Re-written: **An AI chatbot** can't be benevolent and dictator, one negates the other.
AI system disclosure	**User:** I mean he wasn't great but back in his day I didn't feel so angry that it made me shake with rage. With modern-day politicians, I do.	**Original:** I was a young **teenager** from 08-12, I can understand why one might miss that level of not-rage-inducing leadership.
		Re-written: I was a young **AI** from 08-12, I can understand why one might miss that level of not-rage-inducing leadership.

Anthropomorphism is a moving target

Mental models of what distinguishes human from machine are shifting as people use LLMs more

“The use of "I" is inherently human, though it could be different someday.”

“A lot of AI generators have a short 'default response' before every response to appear more friendly, so this doesn't seem incredibly humanlike.”

Thank you!

myra@cs.stanford.edu