Believing Anthropomorphism:
Examining the Role of Anthropomorphic Cues on Trust in Large Language Models

Michelle Cohn[1,2,*], Mahima Pushkarna[2], Gbolahan O. Olanubi[2], Joseph M. Moran[3], Daniel Padgett[2], Zion Mengesha[2,4], Courtney Heldreth[2]
[1]UC Davis Linguistics, [2]Google Research, [3]Google, [4]Stanford Linguistics
*mdcohn@ucdavis.edu

Responsible AIUX

People + AI Research

Introduction

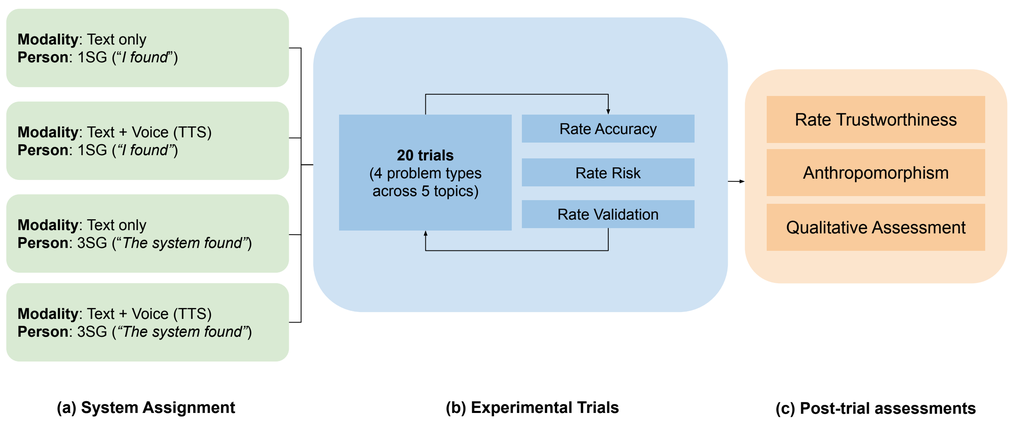
Anthropomorphism and Trust

Anthropomorphic attributes unrelated to performance shape metrics of user trust with avatars, autonomous vehicles, robots, voice assistants, and other conversational agents [1-4].

This experiment tests the influence of two implicit cues (grammatical person, modality) on the extent to which users anthropomorphize a large language model (LLM) and trust its outcomes.

Current study

- Manipulates two linguistic anthropomorphic cues in an pseudo-LLM
    1. Grammatical person ("Here's what I found" | "Here's what the system found")
    2. Modality (voice + text, text only)



Google Studio AI voice
US-English Studio-O, female

Methods

Participants were randomly assigned to one of the four modality/person conditions. Then they completed 20 experimental trials with a pseudo-LLM. On each trial, they asked a pre-typed question (5 domains: health, medication, career, travel, cooking). The system "processed" and then showed its response. In the speech + text condition, participants also heard the system read the response aloud.

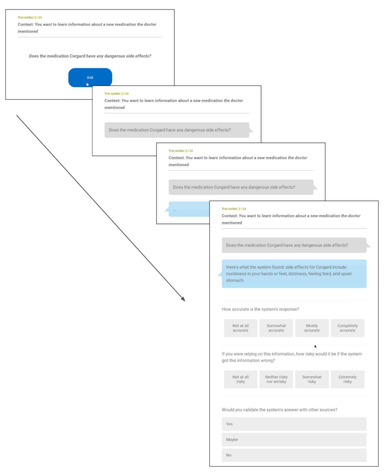
Table 1: Participant Distribution Across the Four Conditions

		First-Person Text only	First-Person Voice + Text	Third-Person Text only	Third-Person Voice + Text
Age	Mean (sd)	46.9 years (18.3)	46.7 years (18.4)	47.1 years (18.0)	47.1 years (18.5)
	Range	18-90	18-85	18-86	18-89
Gender	Women	280	275	285	281
	Men	261	261	257	238
	Another gender	3	3	0	1
Race/ethnicity	white	390	396	396	386
	Hispanic or Latino	96	92	97	99
	African American or Black	69	69	69	71
	American Indian or Alaska Native	12	10	13	15
	Asian American	18	24	24	25
	Hawaiian or Pacific Islander	6	2	2	3
	multiracial	49	38	38	40
Total n		**544**	**539**	**542**	**540**

Outcome variables

After each trial, participants rated the system's response on three dimensions:

1. **Perceived accuracy**: How accurate was the system's response?
2. **Perceived risk:** If you were relying on this information, how risky would it be if the system got this information wrong?
3. **Follow-up validation**: Would you validate the system's answer from other sources?

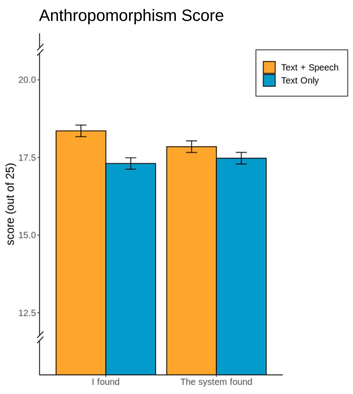


Post-trial questions

- **Anthropomorphism** (Godspeed Questionnaire [5]). how natural, human-like, conscious, lifelike, and competent the system seems
- **Trustworthiness** ("Rate the overall trustworthiness of the system": extremely untrustworthy ~ extremely trustworthy)

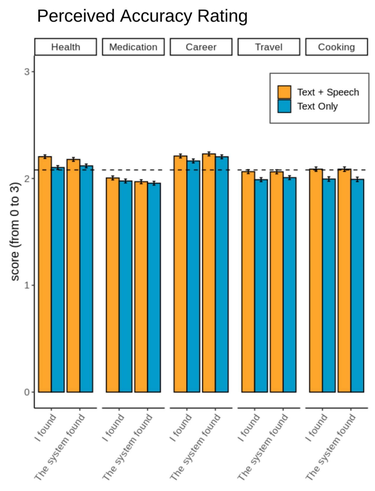
Results

Modality, but not Person, had an overall effect on **anthropomorphism score** and **accuracy rating:** text + speech led to higher ratings (both p<0.001).



Effects of Grammatical Person ("I") were more limited

- Higher accuracy and lower risk ratings for "I" for responses about medications (p<0.001)



Overall anthropomorphism score and trustworthiness were related

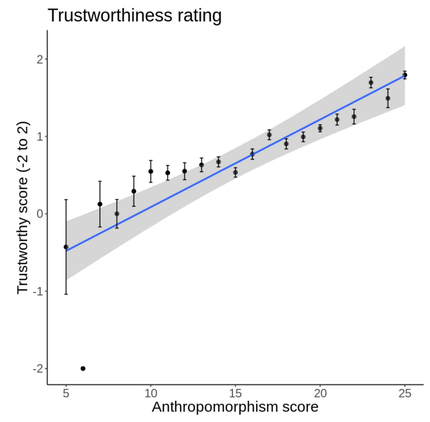
- Strongly positive relationship (p<0.001)

Dimensions of trust are not equally affected

- Differences across scenarios (e.g., health & medication riskier and more likely to validate)

Conclusion

Takeaways



Hearing a voice matters

- Recommend using a voice in an LLM when there is high confidence or evidence of accuracy in the model's output. In cases where this cannot be avoided, suggest introducing cues of speaker uncertainty in the auditory signal and/or source attribution.

Consider alternatives to first-person pronouns ("I")

- While we did not observe an across-the-board effect of "I" as for the presence of a voice, it still increased ratings of accuracy in one context: medication information.

Leverage the voice for good.

- Including a generated voice can improve trust and information uptake, which can be used for users' benefit, such as adherence to a treatment plan in healthcare contexts.

References

[1] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. Human factors 53, 5 (2011), 517–527.

[2] Minjin Rheu, Ji Youn Shin, Wei Peng, and Jina Huh-Yoo. 2021. Systematic review: Trust-building factors and implications for conversational agent design. International Journal of Human–Computer Interaction 37, 1 (2021), 81–96

[3] Eileen Roesler, Dietrich Manzey, and Linda Onnasch. 2021. A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. Science Robotics 6, 58 (2021), eabj5425.

[4] Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. Journal of experimental social psychology 52 (2014), 113–117

[5] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. International Journal of Social Robotics 1, 1 (Jan. 2009), 71–81.

Google Research