scientific reports

OPEN

MATTERS ARISING

Stylistic language drives perceived moral superiority of LLMs

Kalil Warren[1,2✉], Chandler Nichols[1], Dawson Petersen[1], Valerie L. Shalin[3] & Amit Almor[1,2]

ARISING FROM: E. Aharoni et al.; *Scientific Reports* https://doi.org/10.1038/s41598-024-58087-7 (2024).

In this paper, we are responding to a recent article published in *Scientific Reports* by Aharoni et al.[1] titled, "Attributions toward artificial agents in a modified Moral Turing Test." Aharoni et al. tested how humans evaluate the quality of moral reasoning in human-generated and LLM-generated responses to moral questions. The human responses were sourced from university undergraduates, while the LLM responses were generated using OpenAI's ChatGPT-4. The prompts used to elicit the responses asked whether and why certain actions were morally wrong or acceptable. Ten pairs of human-generated responses and LLM-generated responses were then used as stimuli in a modified Moral Turing Test (m-MTT) in which different human participants rated the quality of these responses. Participants rated the LLM-generated stimuli as showing higher quality of moral virtuousness, trustworthiness, and intelligence. However, the participants were able to distinguish between the human-generated and the LLM-generated responses.

Aharoni et al.[1] claimed that "participants' aptitude at identifying the computer, [was due] not to its failures in moral reasoning, but potentially to its perceived superiority—not necessarily in the form of conscious attitudes about its general moral capabilities but at least in the form of implicit attitudes about the quality of the moral responses observed," (Aharoni et al., 2024, p. 8). We argue that their findings do not yet merit this conclusion. While we appreciate the Aharoni et al. contribution to the ongoing discourse on AI and moral reasoning, we propose an alternative interpretation of their results. We suggest that the observed ratings primarily reflect participants' perceptions of the LLM's use of specialist language, not its moral reasoning. Specifically, we argue that the perceived superiority of the LLM-generated responses was driven by uncontrolled psycholinguistic features—namely, word frequency, age of acquisition, word length, and overall readability. These features are not specific to moral reasoning. Therefore, participants' *explicit* judgements of intelligence, moral virtuosity, and trustworthiness are likely driven by well-known *implicit*, domain independent, (psycho)linguistic effects.

Indeed, such psycholinguistic features are well-known to influence perceived credibility, trustworthiness, intelligence, and persuasiveness (e.g.,[2,3]. Seminal research in human intelligence has demonstrated a positive correlation between larger vocabularies and intelligence[4]. Oppenheimer[5] demonstrated that experimentally manipulating psycholinguistic features—such as word length—can significantly influence participants' perception of an author's intelligence, even when texts have identical semantic content. This finding underscores that perceived intelligence—and by extension, other evaluative judgments—can be shaped by surface-level linguistic features alone, independent of the actual substance of the argument. We argue that the differences observed in Aharoni et al.'s[1] study can be fully explained by these uncontrolled low-level psycholinguistic features, that is, a simpler explanation, rather than by the perceived quality of moral reasoning. We recommend that future evaluations of AI controls for these types of confounding psycholinguistic variables, to disentangle the effects of language complexity from genuine perceptions of AI's capabilities.

To test the linguistic differences between the LLM- and human-generated responses used as rating stimuli by Aharoni et al.[1], we examined both responses for mean word length (measured by number of letters, number of phonemes, and number of syllables), mean word frequency, and mean age of acquisition. We also calculated their overall Flesch-Kincaid readability scores using the Text Ease and Readability Assessor (T.E.R.A.)[6,7]. If, as we suspected, the two response types show significant differences in these measures, then the differences in the rating results reported by Aharoni et al. are not informative about moral reasoning.

We used the South Carolina psycholinguistic metabase (SCOPE[8], to extract, for each word in Aharoni et al.'s stimuli, the SUBTLEXus corpus word frequency (Brysbaert & New)[9], Living Word Vocabulary (Dale & O'Rourke)[10] age of acquisition (AoA), and word length (measured in letters, phonemes, and syllables). Out of the 444 distinct words in the stimuli, data were missing for 19 words on frequency, 7 on letter count, 13 on phoneme and syllable count, and 173 on AoA. For the analysis of each measure, we excluded words with missing values on that measure. For readability, each LLM-generated and human-generated response was analyzed using T.E.R.A. to assess the response's grade level readability. For each response, T.E.R.A. produces a Flesch-Kincaid

[1]Linguistics Program, University of South Carolina, Columbia, SC 29208, USA. [2]Department of Psychology, University of South Carolina, Columbia, SC 29208, USA. [3]Department of Psychology, Wright State University, Dayton, OH 45435, USA. ✉email: knwarren@email.sc.edu

Variables	LLM mean (M)	Human mean (M)	t-value (df)	p-value
Word frequency (Log10 frequency in a corpus of 51 million words)	4.38	4.62	4.16 (9)	**.002****
Age of acquisition (Years)	5.85	4.99	−4.25 (9)	**.002****
Word length (Letters)	4.77	4.40	−3.33 (9)	**.009****
Word length (Phonemes)	3.97	3.53	−4.33 (9)	**.002****
Word length (Syllables)	1.60	1.42	−4.27 (9)	**.002****
Grade level readability (Flesch-Kincaid)	11.2	8.4	2.69 (9)	**.025***

Table 1. Comparison of linguistic measures of LLM and human responses.

reading grade level[6,7]. We then computed the means of each variable for each text and used them to compare the texts generated by the humans to those generated by the LLM using two-tailed paired-sample t tests. All statistical testing was done in R 4.4.1[11].

The results are shown in Table 1.

As expected, the LLM responses were significantly more complex than the human responses in terms of word frequency, age of acquisition, word length, and readability. Although the numerical differences may appear modest, they fall within the range known to influence perceptions of author intelligence, expertise, and clarity[2,5]. Thus, we argue that the findings of Aharoni et al.[1] reflect the persuasive effect of linguistic style rather than genuine perceived differences in moral reasoning.

To more accurately evaluate differences in the perception of moral reasoning between LLMs and humans, it is essential to control for psycholinguistic features. Prior work has shown that ChatGPT can adopt different expository styles depending on the prompt[12,13]. Thus, LLM prompts could be crafted to match the psycholinguistic style of the human comparison group. For instance, ChatGPT-4 could be explicitly instructed to respond in the style of undergraduate students, with subsequent verification that both LLM- and human-generated texts are comparable on key linguistic dimensions. Alternatively, researchers could directly revise LLM outputs to match the language form of the human population while preserving the underlying content.

In this paper, we have shown that Aharoni et al.'s[1] results can be explained by low-level psycholinguistic features and thus do not merit conclusions about the perception of LLM's moral reasoning. Considering basic, well known, psycholinguistic features is critical for any study that gauges LLMs' performance in any domain on the basis of human evaluation of verbal responses[14–16].

Data availability

The data and code that supported this study can be found on the Open Science Framework at https://osf.io/xmu93/.

Received: 18 December 2024; Accepted: 16 October 2025
Published online: 07 November 2025

References

1. Aharoni, E. et al. Attributions toward artificial agents in a modified moral turing test. *Sci. Rep.* **14** (1), 8458 (2024).
2. König, L. & Jucks, R. Influence of enthusiastic language on the credibility of health information and the trustworthiness of science communicators: Insights from a between-subject web-based experiment. *Int. J. Med. Res.* **8** (3), e13619 (2019).
3. Zimmermann, M. & Jucks, R. How experts' use of medical technical jargon in different types of online health forums affects perceived information credibility: Randomized experiment with laypersons. *J. Med. Internet Res.* **20** (1), e30 (2018).
4. Spearman, C. "General Intelligence" Objectively Determined and Measured. In *Studies in individual differences: The search for intelligence* (eds Jenkins, J. J. & Paterson, D. G.) 59–73 (Appleton-Century-Crofts, 1961).
5. Oppenheimer, D. M. Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly. *Appl. Cogn. Psychol. J. Appl. Res. Mem. Cogn.* **20** (2), 139–156 (2006).
6. Kincaid, J. P., Fishburne, R. P. Jr., Rogers, Richard L. & Chissom, B. S. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Institute for Simulation and Training*. **56** (1975).
7. Tanner Jackson, G., Allen, L. K., & McNamara, D. S. Common core tera: Text ease and readability assessor. In *Adaptive Educational Technologies for Literacy Instruction*. 49-68. https://doi.org/10.4324/9781315647500 (Taylor and Francis, Routledge, 2016).
8. Gao, C., Shinkareva, S. V. & Desai, R. H. Scope: the South Carolina psycholinguistic metabase. *Behav. Res. Methods* **55** (6), 2853–2884 (2023).
9. Brysbaert, M., & New, B. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*. **41** (4), 977–990 (2009).
10 Dale, E. & O'Rourke, J. The Living Word Vocabulary. Chicago: World Book-Childcraft International, Inc. (1981).
11. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, (Vienna, Austria). https://www.R-project.org/ (2024).
12. Alafnan, M. A. & Mohdzuki, S. F. Do artificial intelligence chatbots have a writing style? An investigation into the stylistic features of ChatGPT-4. *J. artif. intell. technal.* **3** (3), 85–94 (2023).
13. Kobak, D., Márquez, R. G., Horvát, E. Á., & Lause, J. Delving into ChatGPT usage in academic writing through excess vocabulary. *arXiv preprint* http://arxiv.org/abs/2406.07016 (2024).
14. Grice, H. P. Logic and conversation. *Syntax and semantics*, *3*. (1975).
15. Katz, D. M., Bommarito, M. J., Gao, S. & Arredondo, P. Gpt-4 passes the bar exam. *Phil. Trans. R. Soc. A.* **382** (2270), 20230254 (2024).
16. Wu, T. et al. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA. J. Autom. Sin.* **10** (5), 1122–1136 (2023).

Acknowledgements
The authors thank Anne Bezuidenhout, Sarah Wilson, Nadra Salman, Ruhan Çoban, and all members of the alab language and cognition research group for their assistance in the preparation of the manuscript.

Author contributions
Conceptualization: K.W., C.N., D.P., V.L.S., and A.A. Formal Analysis: C.N. and D.P.; Project Administration: K.W. and A.A.; Supervision: A.A.; Writing – original draft: K.W.; Writing – review & editing: K.W., C.N., D.P., V.L.S., and A.A.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to K.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.