Newhouse Impact Journal

Volume 1
Issue 1 *Fall*

Article 8

2023

LLMs and Linguistic Competency: An Exploration of GPT-4 and a Non-Hegemonic English Variety

Samantha Jackson
University of Toronto

Barend Beekhuizen
University of Toronto

Zhao Zhao
University of Toronto

Yi Cheng Zhao
University of Toronto

Rhonda N. McEwen
University of Toronto

Recommended Citation

Jackson, Samantha; Beekhuizen, Barend; Zhao, Zhao; Zhao, Yi Cheng; and McEwen, Rhonda N. (2023) "LLMs and Linguistic Competency: An Exploration of GPT-4 and a Non-Hegemonic English Variety," *Newhouse Impact Journal*: Vol. 1: Iss. 1, Article 8.
DOI: http://doi.org/10.14305/jn.29960819.2024.1.1.04
Available at: https://surface.syr.edu/newhouseimpactjournal/vol1/iss1/8

LLMs and Linguistic Competency: An Exploration of GPT-4 and a Non-Hegemonic English Variety

Samantha Jackson, Barend Beekhuizen, Zhao Zhao, Yi Cheng Zhao, and Rhonda N. McEwen
University of Toronto, Canada

Large language models (LLMs), including the currently popular General Pre-trained Transformers (GPTs), have many natural language processing (NLP) applications and are already changing the landscape of how written language is developed. Such models interact with clients via conversational chatbots and AI (artificially intelligent) assistants, execute rigorous searches and provide human-like answers from search engines, summarize texts, categorize, and classify feedback and requests, translate texts, and make recommendations to improve one's grammar[1] and style.[2] However, biases stemming from the textual training data as well as the annotator-driven subsequent training steps can result in misunderstanding, and subsequent misanalysis, of users' language. With OpenAI's GPTs, misunderstanding is more likely to arise for users of underrepresented languages or language varieties, since the dataset is sourced from publicly available data as well as data licensed by the company, which are skewed toward global Northern and Western users.

Within the English-speaking world, there are multiple varieties of English due to the spread of English through colonialism and globalization[3] and certain Englishes are privileged over others.[4] The hegemonically dominant ones such as British and American English are much more likely to dominate training datasets because they have more written materials. However, users of these varieties are in the minority. For instance, the 374 million English speakers in the US and UK represent a small portion of the 1.5 billion English speakers worldwide.

Through a year-long research project, we investigated (potential) bias in GPT-4 when it interacts with Trinidadian English Creole (TEC), a non-hegemonic English variety that partially overlaps with standardized English (SE) but contains distinctive characteristics. We investigated the linguistic performance of GPT-4, OpenAI's latest GPT model, when it interacts with users of TEC. Specifically, we assessed the model's comprehension, production and metalinguistic knowledge.

Two of the authors of this paper are native speakers of TEC, and the first author is a trained linguist specializing in TEC. The latter scoured thousands of these comments and categorized any TEC features by type. Only features that differed from SE were considered. Following this, two research assistants were trained in recognizing TEC features and assisted with categorizing the rest of the corpus. The first author reviewed these analyses and compiled a list of salient categories with example sentences for reference.

The data were collected between May 1 and August 23, 2023. First, we asked GPT-4 18 questions in TEC and SE. We then asked it to translate 29 sentences from TEC to SE and vice versa. Finally, we asked the model to identify the language of the sentences. All prompts were submitted five times each to check for consistency. We examined whether responses to questions in both varieties were comparable, and whether translations were accurate or seemed to be relying on language stereotypes. By stereotypes, we refer to imitations or exaggerations of features of a variety based more on their perceived association with that variety rather than actual speaking norms of its users.[5] To investigate what language the model categorized the prompts as written in, we examined the five answers for each prompt and classified them into one of the following groups: (i) English, (ii) another variety, (iii) English with influence from another variety, (iv) English with informal/dialectal/slang features and (v) English with errors. We conducted this analysis based on linguistic level, looking at representation of the sound system, vocabulary, and grammar.

In short, GPT-4's performance on answering questions in both varieties provided encouraging results about its ability to comprehend TEC. The model gave comparable content and detail for 56% of the prompts and there were only minor differences for the remaining prompts. The model was highly proficient (over 90% accuracy) at understanding pronouns and 10 grammatical categories. However, it performed poorly in its comprehension of vocabulary items and 2 grammatical categories (<50%). Moreover, GPT-4's TEC production was less proficient than its comprehension. This was evaluated via its translations from SE into TEC. In terms of representing TEC via spelling, GPT-4 was 58% accurate, providing inauthentic, unattested spellings that resemble imitation of TEC by non-users of the variety. The model was highly proficient (over 90% accuracy) at producing pronouns and nine grammatical categories. However, it was weak at producing vocabulary items (13%) and three grammatical categories (8-76%). It also misused some TEC features and produced many ungrammatical features. Furthermore, errors were spread across the responses so that only 69 out of 175 translations (39.4%) contained fully grammatical and accurate TEC. In terms of language identification, while the model identified SE 100% of the time, it only identified TEC 21% of the time. This inaccuracy was compounded by the fact that it sometimes classified TEC as English with errors and tried to "correct" it.

Currently, GPT-4's scope of use is limited for non-hegemonic English users. It is problematic that some of its analyses perpetuate bias against underrepresented Englishes. Increased research on lesser-documented Englishes is necessary and we anticipate that this problem could affect dialects of other languages. We note that the bias identified in the pre-training data may not be the only source of bias within the LLM. A growing number of researchers are finding that algorithms themselves can be a source of bias.[6] We intend to partner with Trinidadian stakeholders to conduct further assessments and to train GPT-4 in the future.

References

1 Parde, Natalie. “Natural Language Processing.” In The SAGE Handbook of Human-Machine Communication, edited by Andrea L. Guzman, Rhonda McEwen and Steve Jones, 318-326. SAGE Publications Limited, 2023. https://doi.org/10.4135/9781529782783.

2 Mücke, Justin, Daria Waldow, Luise Metzger, Philipp Schauz, Marcel Hoffman, Nicolas Lell, and Ansgar Scherp. “Fine-Tuning Language Models for Scientific Writing Support.” arXiv preprint arXiv:2306.10974 (2023). https://arxiv.org/pdf/2306.10974.pdf

3 Kachru, Braj B., Yamuna Kachru, and Cecil L. Nelson. “Introduction: The World of World Englishes.” In The Handbook of World Englishes, edited by Braj B. Kachru, Yamuna Kachru, and Cecil L. Nelson, 1-16. Oxford: Blackwell, 2006.

4 Lippi-Green, Rosina. English with an Accent: Language, Ideology and Discrimination in the United States. London: Routledge, 2011.; Phillipson, Robert, and Tove Skutnabb-Kangas. “Linguistic Imperialism and the Consequences for Language Ecology.” In Handbook of Ecolinguistics, edited by Hermine Penz and Alwin Fill, 121-134. New York: Routledge, 2018.

5 Johnstone, Barbara. “Southern speech with a Northern accent: Performance norms in an imitation.” American Speech: A Quarterly of Linguistic Usage 93, no. 3-4 (2018): 497-512. https://doi.org/10.1215/00031283-7271294.

6 Stinson, Catherine. “Algorithms are not neutral: Bias in collaborative filtering.” AI and Ethics 2, no. 4 (2022): 763-770. https://doi.org/10.1007/s43681-022-00136-w.; Janowicz, Krzysztof. “Philosophical foundations of GeoAI: Exploring sustainability, diversity, and bias in GeoAIand spatial data science.” arXiv preprint (2023). https://doi.org/10.48550/arXiv.2304.06508.; Färber, Michael, Melissa Coutinho, and Shuzhou Yuan. “Biases in scholarly recommender systems: impact, prevalence, and mitigation.” Scientometrics 128, no. 5 (2023): 2703-2736.