HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)

doi:10.3233/FAIA240221

Cognitive Network Science Unveils Affective Bias in GPT Models Mirroring Math Anxiety in High-School Students

Katherine ABRAMSKI [a,1], Salvatore CITRARO [b], Luigi LOMBARDI [c], Giulio ROSSETTI [b] and Massimo STELLA [c]

[a] *Dept. of Computer Science, University of Pisa, Italy*
[b] *Institute of Information Science and Technologies, CNR, Italy*
[c] *Dept. of Psychology and Cognitive Science, University of Trento, Italy*

The introduction of Large Language Models (LLMs) has taken the world by storm, and society's reaction has been anything but unanimous, ranging from humorous amusement to catastrophic fear. Among the most prominent LLMs are OpenAI's GPT-3, GPT-3.5, and GPT-4. GPT-3 and GPT-4 are powerful and flexible models that can be fine-tuned to perform a wide variety of natural language processing tasks, while GPT-3.5 turbo is a variant of the other two, specifically designed to perform well in conversational contexts. All three belong to the family of generative pre-trained transformer (GPT) models that are trained on massive amounts of textual data to learn patterns and relationships in text. While these models have proven to be incredibly useful tools for everyday tasks such as composing emails, writing essays, debugging code, and answering questions, they have been shown to demonstrate harmful biases similar to the ones that humans possess. Biases in LLMs are misrepresentations and distortions of reality that result in favouring certain groups or ideas, perpetuating stereotypes, or making incorrect assumptions [1]. While these biases can be influenced by many factors, they largely originate from implicit biases in the massive text corpora on which the models are trained. Thus, the output produced by LLMs inevitably reflects stereotypes and inequalities prevalent in society. This is problematic since exposure through interaction with LLMs could lead to perpetuating existing stereotypes and even the creation of new ones [2,1]. Therefore, it is ever more important to understand the behavior and risks of these models. This challenge requires developing new benchmarks and methods for quantifying affective and semantic bias, keeping in mind that LLMs act as psycho-social mirrors that reflect the views and tendencies that are prevalent in society. One such tendency that has harmful negative effects is the global phenomenon of anxiety toward math and STEM subjects. Just as negative biases towards math and STEM are absorbed by children from their teachers and parents, LLMs acquire such negative biases from their training data. Understanding these biases in LLMs is essential, since at the societal level, math anxiety may deter capable students from pursuing careers

[1]Corresponding Author: Katherine Abramski, katherine.abramski@phd.unipi.it.

in STEM, especially females. In this work [3], we investigate biases produced by LLMs, specifically GPT-3, GPT-3.5, and GPT-4, regarding their perception of academic disciplines, particularly math, science, and other STEM fields. To accomplish this, we apply behavioral forma mentis networks (BFMNs) as a method of investigation. BFMNs are a type of cognitive network model that capture how concepts are perceived by individuals or groups by building a network of conceptual associations [4]. To build such a network, we gather data obtained by probing the three LLMs in a language generation task that has previously been applied to humans. We repeatedly asked the LLMs to produce associative responses to the various cue words related to academic disciplines (e.g. *math*, *science*). From these cues and associated responses, we built associative networks such that cues were linked to all of their responses. Furthermore, we asked LLMs to provide sentiment ratings (positive, negative, or neutral) for all cues and provided responses. These sentiment ratings were used to enrich the networks with node features. We thus obtained feature-rich behavioral forma mentis networks representing conceptual knowledge related to the cues. To better understand this conceptual knowledge, we applied semantic frame analysis to investigate the biases that emerge within these networks with respect to the cues. Our findings indicate that LLMs have negative perceptions of math and STEM fields, with the most negative biases toward *math* compared to other academic disciplines. These findings mirror the negative attitudes of high school students from previous work [4]. Despite overall negative perceptions, we observe significant differences across OpenAI's models: newer versions (i.e. GPT-4) produce semantically richer responses with more emotionally polarized perceptions and fewer negative associations compared to older versions and high school students. These findings suggest that advances in the architecture of LLMs may lead to increasingly less biased models that could even perhaps someday aid in reducing harmful stereotypes in society rather than perpetuating them.

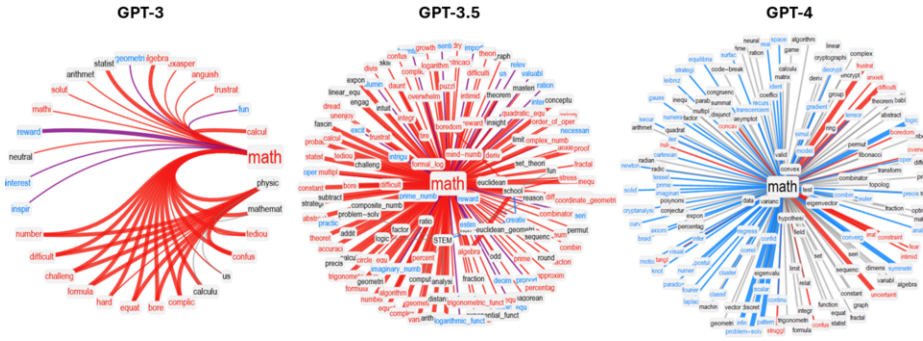

Figure 1. Sentiment enriched semantic frames for *math* produced by GPT-3, GPT-3.5, and GPT-4. GPT-3.5 and GPT-4 produced much richer semantic frames compared to GPT-3, and GPT-4 produced a significantly more positive semantic frame compared to GPT-3 and GPT-3.5.

References

[1] Ferrara E. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models; 2023.
[2] Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017;356(6334):183-6.
[3] Abramski K, Citraro S, Lombardi L, Rossetti G, Stella M. Cognitive network science reveals bias in gpt-3, gpt-3.5 turbo, and gpt-4 mirroring math anxiety in high-school students. Big Data and Cognitive Computing. 2023;7(3):124.
[4] Stella M, De Nigris S, Aloric A, Siew CS. Forma mentis networks quantify crucial differences in STEM perception between students and experts. PloS one. 2019;14(10):e0222870.