Bias in Large Language Models (LLMs); Paving the Way for an Equitable Artificial General Intelligence (AGI)

Kotani, Rihito (School: Tokyo Gakugei University International Secondary School)

This study focuses on the fundamental aspects of LLMs (Large Language Models) such as GPT, Bard, Llama 2, Claude, and more, particularly their inherent biases. These models, some engaging over 18 million daily users, have significant and inescapable social impact. At the same time, they are known to have some biases that are irrational, and stereotypical, influencing societal perceptions. Unlike prior research, only quantifying shallow and superficial biases (e.g. text being positive/negative, etc.), we tried to clarify the concrete structure of bias. The study adopts a rubric of 5 bias aspects with 5 scales (0-4) — Relevance, Representation Bias, Stereotyping, Neutrality, and Assumptions — from a psychological perspective to evaluate bias objectively. It adopts a human-aligned approach to understanding bias, utilizing a global crowdsourcing method with over 6000 responses to dissect and understand biases through a detailed lens, utilizing statistical tools such as Combined Error Variance (CEV) and Symmetric Distance Error (SDE) to ensure precision in analysis. The findings suggest LLMs develop personality-like characteristics in biases, influenced by their unique algorithms and data handling processes. Some features include LLMs developed by the same developer showing similar characteristics (GPT3.5 and GPT4), and anticipated models such as Llama 2 and Gemini-Pro both have a structure of non-stereotypical but non-neutral. Through its novel approach and methodology, this study not only advances our understanding of LLM biases but also reevaluates the desirable structure for the upcoming Artificial General Intelligence (AGI) systems.

Awards Won:
Non-Trivial: 10 scholarships for Non-trivial