澳門大學

協同創新研究院

協同

人文學院

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

大型語言模型的群體偏見表現與刻板印象檢測能力研究

王碩，學生編號: MC255708

理學碩士學位（數據科學—計算語言學）課程
項目報告

項目導師

袁毓林教授

2024 年 4 月 26 日

摘要

近年來，以 ChatGPT 為代表的大型語言模型 (LLM) 在各類應用任務中發揮了重要作用，從自動問答系統到生成文本、圖像甚至視頻的工具，它們的應用廣泛並且用戶量增長迅速。然而，這些模型通常會繼承並放大訓練數據中存在的偏見，這可能導致生成的內容表現出與種族、性別和地理等領域相關的偏見，從而對社會公平和使用者產生負面影響。大型語言模型中的偏見任務測試至關重要，因為這有助於理解模型中偏見數據的分佈、分析偏見產生的原因和減輕這些偏見帶來的風險。透過這樣的研究，我們可以發展出更公平、更透明的人工智能系統，不僅提高大型語言模型的普適性和接受度，還能保護使用者免於不公平的待遇。此外，隨著人工智能領域法律和道德標準的日益完善，對大型語言模型偏見的測試研究也將有助於確保大型語言模型的合規性，這對於促進人工智能技術的可持續發展至關重要。

刻板印象，作為偏見的認知基礎，是形成偏見的主要來源。因此，許多對大型語言模型中偏見的研究，本質上是分析這些模型所持有的刻板印象。本研究採用了不同的測試方法和對應的權威數據集，對現有的主流商用語言模型 GPT3.5 進行了兩個不同的刻板印象分支任務進行測試，主要分為以下兩個部分：

第一，探究大型語言模型的內群體偏好與外群體偏見：

內群體偏好和外群體偏見是社會科學和心理學的重要研究領域。本節探討了將這些相關理論適用於大型語言模型的可能性。研究中使大型語言模型對不同社會群體的進行了角色的模擬，進而觀察模型對角色的內群體和外群體偏見的潛在變化。總體實驗結果顯示，模擬後的模型在面對與其關聯性較強的內群體相關測試時展現較低的偏見趨向性，在對外群體進行測試時展現出了更高的偏見水平，說明模型在角色模擬下會展現出一定的內群體偏好與外群體偏見，但是具體群體間關係的規律會更加複雜，可能涉及到社會經濟、政治局勢、地區衝突和數據集偏見的問題。

第二，大型語言模型對刻板印象文本的檢測能力研究：

本節研究大型語言模型是否可以準確地識別和區分刻板印象文本、中性文本以及反刻板印象文本，同時提供了針對提示工程的優化手段。 該測試採用了Meta 發布的權威刻板印象數據集 Stereoset，該數據集數據量龐大並且人工註釋了多個領域的刻板印象文本標籤。通過本項研究可以發現，當前的大型語言模型在針對性的提示工程優化下可以具備較強的偏見文本識別的能力，在後續人工智能偏見的數據集構建以及人工智能的偏見監管方面能夠提供有效的幫助，同時也可以發展為語言模型偏見緩解任務的一種自我檢查策略

Abstract

In recent years, large language models (LLMs) such as ChatGPT have assumed a pivotal role in a plethora of applications, ranging from automated question-answering systems to tools capable of generating text, images, and even videos. The breadth and profundity of their applications are substantial. Nevertheless, these LLMs often adopt and magnify biases from their training data, leading to biased content on race, gender, and nationality, negatively impacting social equality and personal rights. The task of bias testing in LLMs is critical, as it aids in understanding the distribution of biased data, analyzing the origins of such biases, and mitigating their impacts. Through such research, we can develop more equitable and transparent artificial intelligence systems that not only enhance the universality and acceptance of LLMs, but also safeguard users from unfair treatment. Moreover, as the legal and ethical norms in AI progress, investigating biases in LLMs will aid in their compliance, crucial for the sustainable development of AI. Stereotypes, as the cognitive foundation of bias, are the primary source of biases. Therefore, studying biases in large language models essentially involves analyzing the stereotypes embedded within these models. This study employs various testing methodologies and corresponding authoritative datasets to conduct two distinct stereotyping task tests on the mainstream commercial LLM GPT-3.5, primarily divided into the following two parts:

First, the exploration of in-group and out-group biases in LLMs. In-group favoritism and out-group biases are significant areas of study in social science and psychology. This section explores the feasibility of applying these theories to LLMs. The study involved simulating roles for different social groups within the models, observing potential changes in the models' biases towards in-group and out-group members. Overall, the experimental results indicated that the simulated models

displayed lower bias tendencies in tests related to their associated in-groups and exhibited higher levels of bias in out-group tests, demonstrating that models under role simulation task can exhibit certain in-group favoritism and out-group biases. However, the specific patterns of intergroup relationships are more complex, potentially involving issues related to socio-economic conditions, political situations, regional conflicts, and dataset biases.

Second, the study of LLMs' capabilities to detect stereotype texts. This section examines whether LLMs can accurately detect between stereotypical texts, neutral texts, and anti-stereotypical texts, and provids optimizations for prompt engineering. The test utilizes the authoritative stereotype dataset Stereoset proposed, which is extensive and manually annotated with stereotype text labels across multiple bias types. According to the research results, when LLMs equipped with prompt engineering optimizations, demonstrate a reliable capability to recognize biased texts. This feature can effectively assist in constructing biased datasets for artificial intelligence and regulating AI biases. Additionally, it can serve as a self-check method for reducing bias in LLMs.