Does the Prompt-based Large Language Model Recognize Students' Demographics and Introduce Bias in Essay Scoring?

Kaixun Yang, Mladen Raković, Dragan Gašević, and Guanliang Chen(✉)

Centre for Learning Analytics, Monash University, Melbourne, Australia
{Kaixun.Yang1, Mladen.Rakovic, Dragan.Gasevic, Guanliang.Chen}@monash.edu

Abstract. Large Language Models (LLMs) are widely used in Automated Essay Scoring (AES) due to their ability to capture semantic meaning. Traditional fine-tuning approaches required technical expertise, limiting accessibility for educators with limited technical backgrounds. However, prompt-based tools like ChatGPT have made AES more accessible, enabling educators to obtain machine-generated scores using natural-language prompts (i.e., the prompt-based paradigm). Despite advancements, prior studies have shown bias in fine-tuned LLMs, particularly against disadvantaged groups. It remains unclear whether such biases persist or are amplified in the prompt-based paradigm with cutting-edge tools. Since such biases are believed to stem from the demographic information embedded in pre-trained models (i.e., the ability of LLMs' text embeddings to predict demographic attributes), this study explores the relationship between the model's predictive power of students' demographic attributes based on their written works and its predictive bias in the scoring task in the prompt-based paradigm. Using a publicly available dataset of over 25,000 students' argumentative essays, we designed prompts to elicit demographic inferences (i.e., gender, first-language background) from GPT-4o and assessed fairness in automated scoring. Then we conducted the multivariate regression analysis to explore the impact of the model's ability to predict demographics on its scoring outcomes. Our findings revealed that (i) Prompt-based LLM can somewhat infer students' demographics, particularly their first-language backgrounds, from their essays; (ii) Scoring biases are more pronounced when the LLM correctly predicts students' first-language background than when it does not; (iii) Scoring error for non-native English speakers increases when the LLM correctly identifies them as non-native.

Keywords: Large Language Model · Automated Essay Scoring · Bias.

1 Introduction

In recent years, Large Language Models (LLMs) have undergone rapid advancements. These technologies have demonstrated significant potential for application in various educational contexts, such as assistance for writing [15], generation of learning materials [1], and virtual tutoring [11]. Among these educational

tasks, writing assessment is particularly important, as effective evaluation of student writing plays a crucial role in improving writing performance and helping students develop their writing skills [7]. Due to the time-consuming and labor-intensive nature of manual writing assessment, researchers are exploring Automated Essay Scoring (AES), a process that employs artificial intelligence techniques to assign scores to student-written essays [2]. Pre-trained LLMs (e.g., BERT) have commonly been used for AES by following the fine-tuning paradigm [22,35], which involves using a small set of labeled essays to adapt the LLMs for the specific scoring task at hand. While these fine-tuned models have demonstrated effectiveness in AES, their development typically requires educational practitioners to possess expertise in programming and machine learning. Additionally, training these models demands considerable time and computational resources, posing a significant challenge for many educators [19]. The emergence of prompt-based LLMs, such as ChatGPT, has enabled non-tech savvy educators to access advanced AI technologies to support their teaching practices [32]. These conversational LLMs allow users to provide natural-language instructions to guide the models in completing specific tasks (i.e., via prompts). Recent research has investigated AES using LLMs through prompt-based methods, comparing their accuracy with fine-tuning approaches. While prompt-based methods exhibit slightly lower accuracy, they still achieve satisfactory performance when utilizing prompting techniques such as few-shot learning [31,26].

Since LLMs are typically trained on large corpus that may contain historical biases and discriminatory content towards minority groups, the outputs they generate can perpetuate social biases, potentially harming marginalized communities [10]. A recent study has demonstrated that stereotypical associations are encoded in BERT when performing pronoun resolution tasks [28]. Moreover, such biases have been shown to negatively impact the predictive fairness of downstream educational text classification tasks [25], where models exhibit favoritism toward certain student groups based on inherent or acquired characteristics (e.g., gender and first-language background) [20]. Several studies have evaluated the predictive bias of AES when utilizing LLMs [34,24,13]; however, these studies primarily focus on fine-tuning paradigms. Given that prompt-based tools like ChatGPT are built on transformer architectures similar to conventional LLMs such as BERT, and considering the unique nature of the prompt-based paradigm, where models rely on natural-language instructions and a few in-context learning examples, it is possible that such biases persist or even become exacerbated due to the model's limited "training" in performing specific tasks properly.

To bridge the gap in understanding whether recent prompt-based LLMs continue to exhibit biases in AES tasks, we build on prior research that explores the connection between bias in fine-tuned LLMs and subsequent predictive fairness in downstream tasks [13,25]. Specifically, we set out to explicitly measure the ability of LLMs to infer students' demographic attributes through the linguistic features of their written essays. Furthermore, we aimed to analyze the relationship between LLMs' ability to predict demographic attributes and its predictive bias in AES tasks. Our study can guide strategies for mitigating bi-

ases in prompt-based LLMs, promoting fairer scoring for students from diverse backgrounds and advancing educational equity. Formally, our research addresses the following **R**esearch **Q**uestions:

RQ1 To what extent can an LLM discern students' demographic attributes based on the linguistic characteristics in their written essays?
RQ2 To what extent does the LLM's ability to predict students' demographic attributes relate to their bias in scoring students' essays?

To answer the RQs, we chose a publicly available dataset comprising over 25,000 argumentative essays, each assigned a holistic score, spanning 15 distinct topics. This dataset also includes various demographic attributes of students, such as gender, first-language background, and race. We employed the state-of-the-art LLM model, GPT-4o, which has been widely used in recent AES research [31,26].To answer RQ1, we prompted GPT-4o to infer students' gender and first-language background based on their essays. This allowed us to assess the model's ability to infer demographic attributes using linguistic features present in students' writing. To answer RQ2, we prompted GPT-4o to score the essays and evaluated its predictive fairness using multiple evaluation metrics. Specifically, we analyzed how the model's ability to predict students' demographics (identified in RQ1) affected scoring outcomes. To further quantify these relationships, we conducted multivariate regression analysis. All the prompts were designed by including the effective prompting techniques used in recent literature (e.g., few-shot learning, chain-of-thought reasoning, and role-assigned prompting), as detailed in Section 3. Through extensive analysis, this study contributed with the following key findings: (i) Prompt-based LLMs can infer students' demographics, particularly their first-language backgrounds, from their essays. When GPT-4o delivered explicit prediction labels of students' demographics (i.e., excluding those receiving the label of 'uncertain'), it achieved an accuracy of approximately 0.86–0.96 for gender and 0.75–0.87 for first-language background; (ii) Scoring biases persist regardless of whether prompt-based LLMs accurately identify students' language backgrounds. However, these biases are more pronounced when the LLM correctly predict students' first-language backgrounds; (iii) Statistically significant interaction terms with positive coefficients in the regression models for particular essay sets suggest that the scoring error for non-native speakers increases when the model correctly identifies them as non-native.

2 Related Work

2.1 Automated Essay Scoring

Over the past decades, AES has undergone rapid development. According to the existing literature [38,5,9,27,35,34], these approaches can generally be grouped into four categories: traditional machine learning-based methods, deep learning-based methods, methods based on fine-tuning LLMs, and methods based on prompting LLMs.