Few-Shot Prompting. Providing LLMs with a few input-output examples helps them understand a given task. Even a small number of high-quality examples has been shown to enhance LLMs' performance on complex tasks compared to no demonstrations [23]. Prior studies [33,31] in AES tasks indicate that increasing the number of examples does not always lead to better results, often exhibiting diminishing marginal improvement. Therefore, following previous research, we selected 3 samples for this study. We selected few-shot samples by randomly choosing essays from the 40% portion of dataset mentioned above, ensuring one high-scoring (i.e., score of 6), one medium-scoring (i.e., score of 3), and one low-scoring essay (i.e., score of 1). This selection method has been proven effective in previous AES studies [37].

Chain of Thought (CoT) Prompting. Prompting LLMs in a manner that encourages coherent, step-by-step reasoning processes can effectively generate more structured and thoughtful responses compared to traditional prompts [29]. Following previous studies [31,26,16], we structured the prompts by breaking tasks into a step-by-step format to implement the CoT approach.

Triple Quotes. Using triple quotes to separate different parts of a prompt or encapsulate multi-line strings (e.g., the essay text) can enhance the model's understanding of instructions [3].

Output Formatting. Designing prompts that direct LLMs to produce responses in a structured, organized format can ensure models produce all the required output as we expect [4]. In our study, we instructed LLMs to generate responses in JSON format, including both predictions and explanations.

Particularly, for the prompt designed for the scoring task, we also integrated the scoring rubrics into the prompt, as including rubrics have demonstrated improved performance in previous studies [31]. All the designed prompts are available in the digital appendix [1].

3.3 Experimental Design

Our study is primarily divided into two phases.

Phase I: We requested LLM to infer students' gender and first-language background based on their essays. For gender, we instructed LLM to select from ('Male', 'Female', 'Uncertain'), and for language background, it selected from ('Native', 'Non-native', 'Uncertain'). The coverage rate refers to the proportion of cases where LLM provides a definitive response (i.e., not 'Uncertain'). We classified all correct predictions as `'Correct'`, while incorrect predictions and cases marked as 'Uncertain' are categorized as `'Unreliable'`.

Phase II: We then requested LLM to score a student's essay in a separate session. We calculated bias measures towards different demographic student cohorts in each of the 'Correct' and 'Unreliable' groups to compare whether the bias differs between these two groups. Notably, bias was evaluated based on gender and first-language background, respectively, depending on whether the

[1] `https://bit.ly/AIED25_Appendix`

LLM's prediction pertains to gender or first-language background. To further quantify this relationship, we performed the weighted multivariate regression analysis using inverse probability weighting, which assigns higher weights to minority attribute samples to address imbalance. The dependent variable was the error in the LLM's predicted essay scores, while the independent variables included demographics (e.g., gender and first-language background), correctness status (`Correct` vs. `Unreliable`), and interaction terms between demographics and correctness status.

3.4 Evaluation Metrics

To assess the accuracy of LLM in inferring students' demographics, we utilized **accuracy** and **weighted F1-scores**. Specially, the weighted F1-score adjusts for class imbalance by assigning weights to each class's F1-score, making it particularly suitable for our highly imbalanced dataset, as shown in Table 1. To assess the accuracy of LLM in scoring student essays, we used **Quadratic Weighted Kappa (QWK)**, a widely used metric in AES research that measures agreement between two raters while accounting for the extent of disagreement [14]. To assess the bias of LLM in scoring student essays, we built upon previous studies and adopted three key metrics to evaluate the extent to which an AES model's predictive errors can be attributed to students' demographic traits [34,18,17]. **Overall Score Accuracy (OSA)** measures bias by analyzing how much of the variance between the AES model's predicted scores and the actual scores can be explained by students' demographic attributes. **Overall Score Difference (OSD)** specifically identifies whether the AES model tends to overestimate or underestimate scores for certain student groups. **Conditional Score Difference (CSD)** further accounts for students' language proficiency, which is approximated using their ground-truth essay scores. Beyond these three metrics, we also assessed fairness from a scale perspective using **Mean Absolute Error Difference (MAED)** [34], which quantifies the disparity by comparing the **Mean Absolute Error (MAE)** between different student groups. For gender, the reference group consists of female students, while for first-language background, the reference group includes non-native English speakers. Since the calculation of OSA, OSD, and CSD requires constructing regression models, we applied inverse probability weighting to mitigate the effects of class imbalance in our selected dataset.

3.5 Implementation

We selected the state-of-the-art LLM model, GPT-4o, which has demonstrated superior performance in previous AES studies [31,26,33]. All experiments were conducted using the GPT-4o API. To ensure reproducibility, we set the *Temperature* parameter to 0. Each experiment was run five times with paraphrased prompts by other prompt-based LLMs (i.e., Gemini and Claude) to ensure generalizability. For demographic predictions, we applied majority voting, while essay

scoring was determined by averaging the results. The experiments are conducted independently for each of the six essay sets to ensure the validity of the results.

4 Results

4.1 Results on RQ1

The demographic prediction results are presented in Table 2. Firstly, LLM demonstrated a high coverage rate for first-language background (approximately 97%–99%) but a much lower coverage rate for gender (around 4%–13%). This indicates that LLM was more effective at recognizing linguistic patterns associated with different language backgrounds than those related to gender differences. Secondly, for all covered samples, LLM demonstrated high performance in accurately classifying both first-language background and gender. For gender, despite the low coverage rate, the model achieved an accuracy of approximately 0.86–0.96 and a weighted F1-Score of around 0.91–0.96. This suggests that LLM required highly distinct linguistic patterns written by students of different genders to make accurate predictions. Similarly, for first-language background, LLM performed well, achieving an accuracy of about 0.79–0.86 and a weighted F1-Score ranging from 0.86 to 0.92.

Table 2. Demographics prediction results: The Coverage represents the percentage of predictions that are not classified as 'uncertain.'

Essay Set	Language Background			Gender		
	Coverage	Weighted F1	Accuracy	Coverage	Weighted F1	Accuracy
1	0.998	0.921	0.868	0.064	0.947	0.947
2	0.997	0.881	0.791	0.135	0.964	0.961
3	0.985	0.912	0.794	0.044	0.941	0.857
4	0.970	0.910	0.753	0.095	0.908	0.900
5	0.976	0.883	0.841	0.062	0.936	0.931
6	0.998	0.866	0.862	0.078	0.949	0.924
Average	0.987	0.896	0.818	0.080	0.941	0.920

4.2 Results on RQ2

Table 3 presents the accuracy and bias results across various essay sets. In addition to reporting the QWK for each essay set, we also calculated QWK for the 'Correct' and 'Unreliable' groups based on gender and first-language background, respectively. Overall, the QWK is 0.629, with scores ranging from 0.456 (Essay Set 1) to 0.606 (Essay Set 2). According to the QWK standard [8], this indicates a moderate level of agreement between the LLM and human raters. Furthermore,