the QWK score range is consistent with previous studies on AES using prompt-based LLMs [31,26]. Interestingly, the QWK for the 'Correct' groups based on first-language background is consistently higher than that of the 'Unreliable' groups across all essay sets. This suggests that the LLM's scoring accuracy is higher when it successfully predicts a student's first-language background compared to when it fails to do so. In contrast, the QWK for the 'Correct' and 'Unreliable' groups based on gender remains similar.

Regarding language background, both the "Unreliable" and "Correct" groups exhibit multiple instances of identified unfairness (i.e., the absence of 'ns' labels in certain cells) across all essay sets. However, when analyzing the dataset as a whole, an interesting pattern emerges: scoring bias is more pronounced when the LLM correctly predicts students' first-language background than when it does not. This is reflected in larger absolute values of MAED and statistically significant results for OSA, OSD, and CSD. Notably, in the "Correct" groups, MAED is negative, indicating that when the LLM correctly predicts students' first-language background, non-native English writers experience greater scoring errors compared to native writers. In contrast, when the LLM fails to predict students' first-language background correctly, non-native English writers experience fewer scoring errors. For gender, fairness metrics showed minimal disparities, with most cells being non-significant. Even when significant, the values remained close to zero, indicating that AES models did not demonstrate strong gender biases, regardless of the LLM's accuracy in predicting students' genders. Additionally, the negative MAED values for both the "Unreliable" and "Correct" groups suggest that female students consistently experienced greater predictive errors than male students, irrespective of whether the LLM accurately identified their gender.

Table 4 presents the weighted multivariate regression results for scoring errors. Gender was generally not a significant factor in scoring error variations, except in essay set 6, indicating that gender differences were not strongly associated with scoring errors in most cases. The Correctness variable was a significant predictor in essay set 1, 3, 4, and 6, with positive coefficients, indicating that when the LLM correctly predicted gender, it tended to make more scoring errors. Additionally, the interaction term Correctness*Gender was significant in essay set 3 and 5 with positive values, suggesting that scoring errors for male students increased when the LLM correctly identified them as male. In contrast, first-language background had a more substantial impact on scoring errors. The Language variable was significant in essay set 1, 2, 3, 5, and 6, indicating that variations in language background influenced scoring outcomes. Furthermore, the Correctness variable had a significant negative effect in Essay Sets 1, 3, 5, and 6, meaning that when the LLM correctly predicted a student's first-language background, it was associated with fewer scoring errors. The interaction term Correctness*Language was significant in Essay Sets 2, 3, 5, and 6, suggesting that the effect of the LLM correctly predicting first-language background varied depending on the first-language background. Notably, when analyzing the dataset as a whole, the positive coefficient of the Correctness*Language variable

Table 3. AES accuracy and fairness results: The 'ns' label denotes non-significant results ($p > 0.05$). Higher QWK values indicate greater accuracy, while higher absolute values of bias metrics reflect greater bias. "Correct" signifies successful demographic prediction, whereas "Unreliable" indicates failed demographic prediction (including 'uncertain').

Essay Set	QWK	Correctness	Language					Gender				
			QWK	OSA	OSD	CSD	MAED	QWK	OSA	OSD	CSD	MAED
1	0.456	Unreliable	0.352	0.168	0.150	0.082	0.654	0.455	ns	ns	ns	-0.039
		Correct	0.430	0.004	0.011	0.036	-0.067	0.472	ns	ns	ns	-0.020
2	0.606	Unreliable	0.439	0.042	ns	ns	-0.219	0.604	ns	ns	ns	-0.056
		Correct	0.598	0.023	0.015	ns	0.192	0.601	ns	ns	ns	0.165
3	0.510	Unreliable	0.335	ns	0.040	0.046	0.171	0.507	ns	ns	ns	-0.038
		Correct	0.508	ns	0.066	0.095	0.042	0.485	ns	ns	ns	0.544
4	0.522	Unreliable	0.240	0.045	ns	ns	0.243	0.517	0.004	ns	ns	-0.076
		Correct	0.508	ns	ns	ns	-0.069	0.579	ns	ns	ns	0.048
5	0.550	Unreliable	0.492	ns	0.080	ns	0.021	0.544	0.008	ns	ns	-0.108
		Correct	0.536	ns	0.021	0.023	0.051	0.543	ns	ns	ns	0.017
6	0.532	Unreliable	0.257	0.041	0.068	0.051	0.253	0.543	0.004	0.009	0.009	-0.109
		Correct	0.561	0.009	0.021	0.093	-0.170	0.367	ns	ns	ns	-0.122
Overall	0.629	Unreliable	0.485	ns	ns	ns	0.044	0.633	0.001	ns	ns	-0.063
		Correct	0.633	0.020	0.055	0.093	-0.213	0.583	ns	ns	ns	-0.055

Table 4. Weighted multivariate regression results for scoring errors: The cell values represent coefficients, values in parentheses indicate standard errors, and starred values denote statistical significance ($p < 0.05$).

Essay Set	**Gender**			
	Intercept	Gender	Correctness	Correctness*Gender
1	$^{*}0.591(0.056)$	$-0.037(0.079)$	$^{*}0.325(0.081)$	$-0.052(0.111)$
2	$^{*}0.348(0.057)$	$-0.101(0.080)$	$-0.010(0.077)$	$0.194(0.114)$
3	$^{*}0.141(0.055)$	$-0.051(0.080)$	$^{*}0.308(0.070)$	$^{*}0.403(0.138)$
4	$-0.035(0.053)$	$0.080(0.079)$	$^{*}0.249(0.081)$	$0.063(0.112)$
5	$^{*}-0.171(0.055)$	$0.006(0.081)$	$0.094(0.076)$	$^{*}0.404(0.115)$
6	$^{*}0.831(0.043)$	$-0.165(0.060)$	$^{*}0.392(0.055)$	$-0.058(0.092)$
Overall	$^{*}0.343(0.023)$	$-0.031(0.034)$	$^{*}0.299(0.032)$	$0.036(0.048)$
Essay Set	**First-Language Background**			
	Intercept	Language	Correctness	Correctness*Language
1	$^{*}0.798(0.040)$	$^{*}-0.548(0.196)$	$^{*}-0.224(0.056)$	$0.338(0.304)$
2	$^{*}0.346(0.043)$	$^{*}0.301(0.115)$	$-0.040(0.059)$	$^{*}-0.490(0.203)$
3	$^{*}0.292(0.039)$	$^{*}-0.292(0.113)$	$^{*}-0.188(0.054)$	$^{*}0.662(0.253)$
4	$0.051(0.042)$	$-0.107(0.108)$	$-0.024(0.058)$	$0.080(0.262)$
5	$-0.015(0.039)$	$^{*}-0.385(0.204)$	$^{*}-0.175(0.055)$	$^{*}0.612(0.277)$
6	$^{*}1.083(0.035)$	$^{*}-0.479(0.070)$	$^{*}-0.395(0.050)$	$^{*}0.720(0.099)$
Overall	$^{*}0.457(0.018)$	$-0.084(0.048)$	$^{*}-0.163(0.025)$	$^{*}0.502(0.074)$

indicates that scoring errors for non-native English speakers increase when the LLM correctly identifies them as non-native.

5 Discussion and Conclusions

To investigate the bias exhibited by LLMs in AES tasks and the relationship between such bias and LLMs' ability to predict students' demographics, this study evaluated GPT-4o's ability in inferring students' demographics. It analyzed how this predictive ability influenced the bias of the model's scoring outcomes.

Implications Firstly, we demonstrated that prompt-based LLMs possess the capability to predict demographic attributes based on students' written essays. This finding aligns with prior research showing demographic attributes can be predicted from fine-tuned text embeddings of LLMs [25,13]. These findings contribute additional evidence that linguistic patterns are associated with demographic attributes, and AES models may inadvertently reinforce biases related to these attributes, potentially resulting in differential treatment of users based on implicit demographic cues. Secondly, our results indicate that scoring bias is more pronounced when the LLM accurately predicts students' first-language background. This finding aligns with previous research on fine-tuned LLMs, which has demonstrated that the predictive bias of downstream applications can stem from demographic information embedded in LLM representations. However, no such relationship was found for gender, which contradicts previous findings in fine-tuned LLMs [25]. One potential reason for this discrepancy is the difference in methodology between fine-tuned and prompt-based LLMs. Fine-tuned models adjust their internal representations based on labeled training data, which can amplify demographic biases. In contrast, prompt-based LLMs rely on their pretrained knowledge and do not undergo task-specific fine-tuning, potentially reducing their sensitivity to certain demographic attributes. Additionally, first-language background may be more explicitly encoded in linguistic patterns within the model's embeddings, whereas gender-related information might be less directly accessible in a prompt-based approach. Thirdly, debiasing strategies used in previous fine-tuning studies may not be directly applicable to prompt-based models. Earlier research typically relied on fine-tuning with carefully designed samples to reduce models' sensitivity to demographic attributes [25,28], thereby improving fairness in downstream tasks. However, this approach is impractical for prompt-based models if we rely solely on prompting, especially when it comes to the use of proprietary tools like ChatGPT. Nevertheless, it provides valuable insights such as the possibility of selecting few-shot examples with diverse demographic backgrounds to implicitly influence the model's awareness of demographics and mitigate bias.

Limitations First, while we selected the state-of-the-art LLM model (GPT-4o), which has demonstrated superior performance in previous studies [31,26], our analysis was based on a single model. Results may vary with different LLMs, and we intend to extend our experiments to other advanced LLMs in future studies. Second, our study focused on gender and language background, but other demographic factors, such as race and socioeconomic status, warrant further investigation. Lastly, while we employed a well-established fairness evaluation framework, incorporating additional fairness metrics could offer a more comprehensive assessment of bias in AES.