Traditional machine learning-based methods. To enable an AES model to accurately assess essay quality, these methods typically placed considerable emphasis on manually designing meaningful textual features to serve as input for training traditional machine learning models (e.g., Bayesian Linear Regression, Random Forests, and Support Vector Machines (SVM)). For instance, Zesch et al. [38] improved the training of an AES model using SVM by leveraging a comprehensive feature set that included key linguistic attributes essential for evaluating essay quality, such as word n-grams, cohesion features, and syntactical features. Similarly, Chen et al. [5] employed a variety of linguistic and statistical features (e.g., lexical features, grammar and fluency features, and syntactical features) to train a rank-based SVM rating model, where the agreement between human raters and the model is explicitly integrated into the loss function.

Deep learning-based methods. Manually designing features is a laborious process, leading some researchers to explore the use of deep learning models to automate the extraction of features from raw textual data. For example, Taghipour et al. [27] utilized Convolutional Neural Networks to identify local textual dependencies and Long Short-Term Memory networks to capture sequential dependencies, leveraging these extracted features to score essays. Similarly, Dong et al. [9] employed hierarchical network structures to capture dependencies at the word and sentence levels, incorporating attention mechanisms into Recurrent Neural Networks to highlight key words or sentences to score essays.

Methods based on fine-tuning LLMs. With the emergence of pre-trained LLMs (e.g., BERT), which are trained on large text corpora to learn general linguistic features, some researchers have begun leveraging these advancements to score essays more accurately by further fine-tuning these models on a smaller set of labeled essays to adapt these models to the specific scoring task. For example, Rodriguez et al. [22] fine-tuned BERT to generate essay embeddings for subsequent scoring. Similarly, Yang et al. [35] introduced a hybrid loss function that integrates dynamically weighted mean square error loss with batch-wise ListNet loss, enhancing scoring accuracy during fine-tuning BERT.

Methods based on prompting LLMs. Fine-tuning LLMs demands substantial computational resources and often yields only holistic scores, which oftentimes fails to provide detailed explanations. To address this, recent research has investigated prompt-based LLMs (e.g., ChatGPT) for essay scoring and explanation generation. These approaches typically rely on natural-language prompts to guide LLMs in performing the scoring task. For instance, Xiao et al. [31] proposed various prompting strategies, such as including or excluding detailed rubric contexts and utilizing zero-shot or few-shot in-context learning strategies, to guide LLMs in scoring student essays and providing feedback. Similarly, Stahl et al. [26] examined the effectiveness of various task instruction types and few-shot learning during the prompt design phase for essay scoring.

2.2 Bias in Automated Essay Scoring

The importance of algorithmic bias in AES was recognized as early as 2012 [30]. However, it was not until recent years that several studies were conducted to explore the biases in existing AES algorithms. For example, Litman et al. [17] analyzed the biases of three AES models, highlighting that different models exhibit distinct biases related to students' gender, race, and socioeconomic status. Yang et al. [34] examined nine widely used AES methods, assessing their performance across seven metrics on an open-source dataset. Their findings revealed that topic-specific models tend to display greater bias towards students of varying economic backgrounds compared to cross-topic models. Schaller et al. [24] investigated shallow learning, deep learning, and LLMs using both balanced and skewed training data subsets. They found that models trained on skewed data from students with higher or lower cognitive abilities showed no bias but suffered from significantly reduced accuracy for students outside the training set. However, these studies have primarily focused on the fine-tuning paradigm in using LLMs. To the best of our knowledge, only one study has explored AES fairness in prompt-based LLMs. Yancey et al. [33] assessed GPT-4's scoring performance across different genders and L1 language groups, concluding that bias did not significantly vary based on gender or L1. Across all the studies mentioned above, while they demonstrated the presence of predictive bias, they did not explicitly investigate to what extent the observed scoring bias is related to LLMs' ability to predict students' demographic attributes, while an LLM's ability to predict students' demographic attributes has been recognized to be somewhat related to its predictive bias towards disadvantaged users [25]. As a result, these studies provide limited insights for developing effective debiasing strategies to address predictive bias issues in existing AES systems. To our knowledge, only one study [13] has attempted to investigate such relationships in AES. The authors fine-tuned XLNet to score essays written by students from different demographic backgrounds and then examined whether XLNet's hidden states could predict students' demographic attributes. Their findings provided evidence that demographic group differences were embedded in the model's hidden layers. However, their study focused on fine-tuning LLMs. Whether and how such relationships exist in prompt-based LLMs (especially the latest ones like GPT-4o) remains unexplored. Our study fills this gap by prompting LLMs to infer students' demographic attributes from their essays, score the essays, and quantify the relationship between LLMs' ability to predict these attributes and biases in essay scoring.

3 Method

3.1 Dataset

The PERSUADE 2.0 corpus comprises over 25,000 argumentative essays written by U.S. students in grades 6 through 12 across 15 different topics [6]. The dataset includes holistic essay scores assigned by trained human raters based on a standardized scoring rubric used in the Scholastic Aptitude Test (SAT). These holistic scores range from 1 to 6, with higher scores indicating better essay quality. Additionally, the dataset contains various demographic attributes of the students, such as gender, race, and first-language background. The PERSUADE 2.0 corpus has been utilized in several studies on AES fairness [34,13], further demonstrating its usability and quality. However, since some writing topics are source-based writing and the dataset does not include the source articles, and some topics lack demographic information, we retained only 6 independent writing topics for our experiments. The dataset was publicly released in a Kaggle competition and was already split into a 40:60 ratio for training and testing with stratified random sampling. As a result, we conducted our experiments and evaluation on 60% of the data, while the remaining 40% was used as a pool for selecting few-shot examples, as detailed in Section 3.2. The distribution of the 60% experimental dataset is presented in Table 1. Given the dataset's imbalance in first-language background distribution, our study applied techniques to mitigate its impact, as detailed in Sections 3.3 and 3.4. Previous research has shown that students' essays display unique linguistic features based on gender and first-language background [36,12], which motivated us to investigate whether LLMs were capable of discerning these demographic attributes solely by analyzing student written essay.

Table 1. Distribution of student demographics, with 'Native/Non-Native' indicating whether students are Native or Non-Native English speakers.

Essay Set	Topic	Language Background		Gender	
		Native	Non-Native	Male	Female
1 (n=875)	Summer projects	849	26	442	433
2 (n=839)	Mandatory extracurricular activities	779	60	412	427
3 (n=813)	Cell phones at school	781	32	372	441
4 (n=807)	Grades for extracurricular activities	770	37	375	432
5 (n=760)	Community service	728	32	353	407
6 (n=1,498)	Distance learning	1,124	374	743	755

3.2 Prompting Design

To ensure the independence of experimental results, we created three independent prompts and submitted them to GPT-4o in separate sessions: one for inferring gender, one for inferring language background, and one for scoring. To enhance the quality of LLMs responses, we reviewed literature on effective prompting techniques [3,23] as well as previous studies on using LLMs for AES [31,26]. The prompting techniques we employed include:

Role-assigned Prompting. Explicitly assigning a role or persona to LLMs (e.g., educator) enhances contextual understanding and enables responses that align with specific expectations [21]. For instance, a prompt such as *"You are an educator with expertise in grading essays for students in Grades 6 through 12 in the U.S."* helps tailor LLM's output to the desired context.