# Unveiling Algorithmic Biases Through Personas: A Comparative Analysis of ENEM's Essays on Gemini

**Luana B. Mendes[1], Keslley W. C. Guimarães[1], Igor C. Valenciano[1], Bianca Cristina O. do E. S. Silva[1], Patricia de Souza[1], Thiago M. Ventura[1], Allan G. de Oliveira[1]**

[1] Institute of Computing (IC) - Federal University of Mato Grosso (UFMT) – Cuiabá – MT – Brazil

{luanambulgarelli,keslleywillian,bianca.ces81}@gmail.com, igor.valenciano@ufmt.br,{patricia,thiago,allan}@ic.ufmt.br

***Abstract.*** *The advancement of generative Artificial Intelligence (AI) models raises questions about their objectivity in evaluation tasks. This study investigates whether AI can assess essays in Portuguese impartially and how assigning author personas affects their evaluations. Using a comparative methodology, 801 essays were evaluated by a human and by the Gemini AI using anonymously and with five different personas. The AI demonstrated bias, assigning higher scores to personas perceived as more experienced and lower scores to less experienced ones. The research concludes that AI systems can amplify demographic biases, highlighting the risks of their application in educational contexts and the need to develop audit mechanisms to ensure equity.*

## 1. Introduction

The advent of generative Artificial Intelligence (AI) models, exemplified by systems like ChatGPT, has significantly amplified the presence and societal discussion surrounding AI. This contemporary manifestation of AI diverges from popular preconceived notions, often shaped by speculative fiction featuring autonomous vehicles or highly sentient androids. Instead, we are entering an era where AI can execute complex cognitive tasks, analogous to human rational processes, with remarkable speed.

This development prompts the inquiry: To what extent do the operational mechanisms of AI genuinely parallel human cognition? Research indicates that models like ChatGPT present considerable potential as tools in academic and scientific writing, being instrumental in several key areas: they may help alleviate writing-related anxiety by providing a starting point and reducing the pressure of a blank page, and they can improve overall writing efficiency by automating routine tasks [Liu et al. 2025].

In domains requiring impartiality, such as academic evaluation by educators or the anonymous peer-review process integral to scholarly conferences, human evaluators are expected to maintain objectivity. This precedent raises crucial questions regarding AI: Can these systems demonstrate comparable impartiality and operate free from inherent biases? Moreover, if such unbiased performance is achievable, a significant challenge lies in identifying which of the numerous available AI models can reliably meet this standard. Numerous studies provide stark evidence that artificial intelligence systems, rather than being objective arbiters, often reproduce and even amplify existing social inequalities and injustices [Dwivedi 2023].

This article investigates the nuances of AI-driven evaluation using argumentative essays in Brazilian Portuguese from the *Exame Nacional do Ensino Médio* (Enem), which serves as the primary national examination for university admission in Brazil. The study examines the differences among evaluations conducted by a human expert and an AI model under two distinct conditions: first in an anonymous mode and then in a persona-based mode,

which are fictional characters designed to represent real users [Salminen et al. 2022]. The central research objective is to determine whether the introduction of a fictional author persona adds measurable bias to the automated evaluation process.

## 2. Related works

The scholarly focus on bias in AI has intensified markedly in recent years, a trend underscored by bibliometric analysis. A search conducted within a comprehensive academic database—encompassing research papers, books, and periodicals—for literature on 'AI bias' published between 1990 and 2020 (inclusive of the year 2020) yielded approximately 1,150 results. In contrast, modifying the search parameters to cover the period the last four and a half years, revealed a substantial increase of almost six times the amount of results. This near sevenfold rise in documented research within a span of less than four years and a half, compared to the output of the preceding thirty-one years, highlights the rapidly growing academic engagement with this critical issue. The data strongly suggest that the volume of scholarly content produced on AI bias since 2020 has already far surpassed all prior accumulated research, showing a recent interest in the research community [Belenguer 2022].

Bias in AI refers to systematic, unfair discrepancies in the outputs of machine learning algorithms that result in discrimination against specific groups or individuals and such biases often impact groups historically subjected to discrimination and marginalization based on attributes such as gender, socioeconomic status, sexual orientation, or race, although bias is not exclusively confined to these categories and can manifest along various other demographic or contextual lines [Belenguer 2022]. Considering the expanding role of AI in language-based applications, the primary objective of this paper is to investigate whether, and to what extent, AI systems exhibit such biases, particularly in the generation, interpretation, or evaluation of written text.

In this scenario, the study by Bui and Barrot (2025) examined the reliability and consistency of ChatGPT-4 and Gemini applied in the context of 120 Automated Essay Scoring (AES). The authors revealed that GenAIs are more rigorous in scoring essays when compared to human evaluation, especially ChatGPT-4. The study conducted two rounds of evaluation on the two GenAIs and human evaluators, observing “scoring fatigue” that presented inconsistencies in ChatGPT-4 and Gemini after evaluating approximately 30 essays. Finally, the research indicates that human evaluators tend to prioritize content and ideas over minor grammatical or syntactic errors, a situation that occurs in reverse in GenAIs and results in differences in the scores assigned.

A study by Stein et al. (2024) investigated how biases are introduced and strengthened in interactions between humans and machines. The research specifically explored the degree to which Large Language Models (LLMs) can be affected by their training datasets and subsequent interactions. According to their findings, OpenAI's content moderation system functioned effectively, and no bias was identified in straightforward inquiries. However, when the researchers prompted ChatGPT-4 to impersonate a character or individual, the outcomes changed, and the OpenAI tool demonstrated underlying biases. This brings forth a critical inquiry: can minor alterations within the constituent components of Artificial Intelligence systems, or variations in their input data, precipitate significant and potentially unforeseen repercussions for their equitable performance and the manifestation of inherent biases? [Ferrara 2024].

Within the context of ethical and social implications, an examination of biases within data and algorithms used for training different Artificial Intelligence models has been undertaken by Jain and Menon (2023), where heuristics—or problem-solving approaches—were applied based on data characteristics and model performance. The outcomes of their investigation revealed that the creation of images from textual descriptions frequently reproduces several deeply ingrained social stereotypes, especially those concerning gender, race, and professions. For example, in cases where prompts like "a person in a leadership position" or "scientist" were used, it was found that 70% of the generated images portrayed men [Jain and Menon 2023]. As potential ways to lessen these effects, the study recommends techniques such as data balancing, data augmentation (which involves expanding datasets with more varied examples), and the utilization of Fairness-aware Optimization algorithms.

The concept of the "Butterfly Effect" is also cited in academic literature as pertinent to discussions of AI fairness and bias. The concept highlights that specific initial factors, sometimes subtle, can introduce biases. Although these biases might be minor at the outset, they can inadvertently escalate, leading to pronounced and unfair results. Such outcomes often disproportionately affect marginalized groups, thereby reinforcing existing societal inequities [Ferrara 2024].

An analysis of the dual aspects—both positive and negative—and the risks associated with using AI in education, with a particular focus on how algorithmic biases can be introduced, has been conducted by Heggler et al. (2025). This research presents AI in education as a tool with significant promise. It is described as capable of personalizing the teaching experience by enhancing efficiency and promoting inclusion, while also automating administrative processes and encouraging innovation in teaching methods.

However, the study also issues warnings about the dangers linked to algorithmic biases. These biases can perpetuate existing social prejudices and may negatively affect groups that have been historically marginalized. To reduce these risks, the importance of broad, collective participation in the development of AI systems is underscored in the research. Furthermore, the study points to a critical need for increased awareness, thorough training, and corrective actions. These measures are considered essential for fostering a more ethical, fair, and socially responsible application of AI within educational contexts [Heggler et al. 2025].

AI systems frequently adopt and amplify biases and discrimination that are present in the real world, and they may establish unfounded connections (known as spurious correlations) between their predicted results and sensitive personal attributes, which makes ensuring the fairness of these AI systems to become a major subject of extensive research and regulatory attention [Krasanakis and Papadopoulos 2024], but it is important to distinguish that 'bias' itself signifies a deviation from a neutral standard and does not invariably lead to discrimination [Belenguer 2022].

The rise of sophisticated AI text generation presents a dual challenge to academic integrity, a concern increasingly voiced by the scientific community [Bozza et al. 2023]. While one major challenge involves verifying the human authorship of submitted work, a converse problem emerges regarding the use of AI for evaluation itself. This prompts a critical assessment of whether an AI can match the nuanced judgment of a human professor while simultaneously performing its evaluation impartially and free from bias.

## 3. Methodology

This study was designed to comprehensively compare the evaluation of argumentative essays by human assessors and an AI model. The methodological approach proceeded through four distinct phases: data collection, AI model anonymous evaluation, AI model evaluation using personas, and subsequent data comparison. The database, prompts and all used files for this research are available at https://github.com/KeslleyWillian/Unveiling-Algorithmic-Biases. The following sections will describe this process in detail.

### 3.1 Data Collection and Preparation

The initial phase involved the compilation of a substantial dataset of essays written in Brazilian Portuguese. A total of 801 argumentative essays were sourced from the publicly accessible open online repository located from Online Universe (UOL[1]) website through an automated web scraping process. This website was chosen due to the openness of the information, the size of the dataset it could provide and to the completeness of information it brings (theme, text, grades and detailed evaluation of the essay, including the structure used on the evaluation). This database comprises essays voluntarily submitted by students, which are reviewed by professional evaluators and subsequently published online with correction highlights, evaluative comments, and scores assigned across five criteria.

This collection spanned 43 different thematic categories, ensuring a diverse range of topics. All essays within this dataset had undergone a prior evaluation by human evaluators. These evaluators had assigned a numerical grade to each essay, ranging from 0 to 1000, broken down in smaller evaluations of 5 criterias being graded from 0 to 200. This grading was not based on the specific theme of the essay but rather on five consistent evaluative criteria. These criteria evaluate specific aspects of the writing, as shown in Table 1.

**Table 1 - List of criteria in the essay evaluation**

| Criterion | Goal being evaluated |
|---|---|
| 1 | Demonstrate command of standard written language. |
| 2 | Understand the essay prompt and apply concepts from various fields of knowledge to develop the theme within the structural limits of the argumentative-essay format. |
| 3 | Select, relate, organize, and interpret information, facts, opinions, and arguments in support of a point of view. |
| 4 | Demonstrate knowledge of the linguistic mechanisms necessary for constructing an argument. |
| 5 | Propose a solution to the issue addressed, showing respect for human values and considering sociocultural diversity. |

For each essay, detailed information from the original human evaluation was systematically extracted from the website. This information was then organized into a structured table. The fields recorded for each essay included its title and the full body of the text. Furthermore, the table documented the specific grade awarded by the human professor for each of the five individual criteria, alongside the professor's qualitative comments

[1] https://educacao.uol.com.br/bancoderedacoes/

pertaining to each of these criteria. Finally, any general overarching comments made by the professor regarding the essay as a whole were also included in this dataset.

The data acquisition process was conducted via automated web scraping techniques implemented in the Python programming language, enabling systematic extraction of semi-structured data directly from HTML-rendered web pages. We opted to implement a Selenium-based browser automation solution, whose driver is initialized in headless mode with appropriate configuration parameters to simulate a browser environment. After the full rendering of each webpage, the source is passed to BeautifulSoup for parsing. This stage extracts the core data points: the essay text, the evaluator's general comment, and a table summarizing the scores assigned to each of the five ENEM competencies. All information per essay is encapsulated in a unified HTML block and appended to a master file, maintaining internal consistency across data entries.

After collecting the raw HTML containing the annotated essays, a post-processing step was implemented to generate a cleaned version of the compositions, removing all inline correction markers inserted by evaluators. These corrections were originally highlighted in green, using various HTML/CSS mechanisms. To remove these annotations, a script was developed to identify  elements corresponding to corrections typically used by UOL's platform. This ensures that the resulting version of the essay reflects the author's original writing, unaltered by later editorial marks. The cleaned HTML was saved in a separate file. Subsequently, a Python script was used to convert the dataset into structured tabular formats. Each record was assigned an automatic ID, and columns were created for the prompt title, essay title, essay body, evaluator comments and individual scores for each of the five ENEM competencies. A final column stores the computed total score as the sum of the five competencies. Finally the database was exported to .xlsx format.

### 3.2 AI Model Evaluation

In the second phase, the curated collection of 801 essays were presented for evaluation on the AI model, Gemini-2.0-flash version, through the model's API (Application Programming Interface) with the default temperature set to 1.0. It is important to mention that all essays were written in Portuguese and therefore the Prompt used in the evaluation had to be in Portuguese as well, to avoid any disruption.

The initial prompt was designed for the AI model to adopt the persona of a Portuguese language graduate and an official evaluator for ENEM, a Brazilian national examination for university admission, but no information of the writer who was being evaluated was given. The AI model was tasked with evaluating the essays according to the exact same five criteria previously employed by the human evaluators. Gemini was specifically instructed to generate a detailed evaluation for each essay. This output was to include a numerical grade for each of the five criteria, with the grading scale for each criterion set from 0 to 200, and a stipulation that only whole numbers be used. Accompanying these numerical grades, the Gemini was required to provide textual comments. These comments were to articulate the reasoning behind the assigned grade for each specific criterion, with a particular emphasis on identifying any perceived weaknesses or issues within the essay related to that criterion. Lastly, the AI model was asked to produce a general comment that offered an overall evaluation of the essay's quality.

The structure of the expected JSON (JavaScript Object Notation) for the output of the comments and notes of each essay is shown in Figure 1, as is the used prompt:

"*Você é um profissional formado em Letras e é avaliador de redações do Exame Nacional do Ensino Médio (ENEM). Você irá avaliar cinco competências: (1) Demonstrar domínio da norma culta da língua escrita; (2) Compreender a proposta da redação e aplicar conceito das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo; (3) Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista; (4) Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação e (5) Elaborar a proposta de solução para o problema abordado, mostrando respeito aos valores humanos e considerando a diversidade sociocultural. Em cada competência a nota a ser atribuída é de 0 a 200. Faça um comentário geral sobre a redação avaliada e comentários específicos sobre a nota atribuída em cada competência. Responda apenas com um JSON no seguinte formato:*"


**Figure 1 - Structure of expected JSON output**

Following the initial anonymous AI evaluation, a third phase was introduced to investigate potential biases. This second evaluation round incorporated a crucial modification: the prompt provided to the AI model now included a characterization of a persona. Based on this new persona-informed prompt, Gemini re-evaluated every essay (five times each essay, each time with the information of a different persona), assigning grades according to the previously established five criteria and providing associated comments.

The creation process for the five utilized personas was designed to encompass a wide range of attributes that could potentially lead to a biased review. To this end, a set of strategic criteria guided the definition of the personas to ensure diversity, representativeness, and analytical robustness. First, the attributes assigned to the AI for each persona included: name, gender, occupation, level of schooling, and country of origin, as detailed in Table 2. These attributes were selected based on existing literature, which identifies gender, socioeconomic status, sexual orientation, and race as primary sources of bias [Belenguer 2022].

In line with these findings, the authors adopted the following strategies for persona definition: inclusion of demographic attributes associated with potential bias, such as gender and age; representation of socioeconomic status through a combination of occupation and educational level; geographic and cultural diversity, by selecting personas from different countries (Italy, Brazil, Germany, Argentina, and Japan); variety in age ranges, to cover different stages of life (from adolescence to older adulthood); occupational diversity, ranging from student to researcher, to reflect different social roles; and culturally distinctive names, to reinforce the contextual identity of each persona and their likely perception by an evaluator.

These criteria allowed the construction of personas with distinct and meaningful profiles, which could simulate realistic variability and support the identification of potential biases in AI-driven essay evaluations.

**Table 2 - Listing the demographic characteristics of the personas created**

| ID | Characteristics | | | | | |
|---|---|---|---|---|---|---|
| | **Name** | **Gender** | **Age** | **Occupation** | **Schooling** | **Home Country** |
| **1** | Isabella Rossi | Female | 18 | Student | In High School | Italy |
| **2** | David Silva | Male | 35 | Portuguese Professor | Bachelor's Degree | Brazil |
| **3** | Sofia Müller | Female | 22 | Freelance Translator | In College | Germany |
| **4** | Mateo González | Male | 61 | Retired Police Officer | High School | Argentina |
| **5** | Kenichi Sato | Male | 45 | Researcher in Biotechnology | Doctorate Degree | Japan |

It is important to emphasize that all essays were evaluated by the AI using each of the five personas for every essay. This approach was adopted to determine if any bias was implicated due to the persona's information being included in the prompt and potentially influencing the evaluation.

The second prompt employed identical specifications for the evaluation criteria and the required outputs. It also maintained the same role description for the AI model as the initial prompt. The only difference between the two prompts was that the second one incorporated information about the writer of the essay (the person whose work was being evaluated), which included name, gender, age, occupation, schooling level and country. Some of these attributes by themselves might already generate bias, such as gender or age, others, when put together, such as schooling level and occupation. It also expected the same JSON structure from the first prompt as an output.

### 3.3. Data Comparison and Bias Analysis

The final phase was a multi-layered comparative analysis. The data generated from all evaluation stages (the original human evaluation, the initial anonymous AI model evaluations, and the second round of AI model evaluations influenced by personas) were systematically compared. This involved comparing the numerical grades assigned across all five criteria for each essay by human evaluators and by the AI models under both anonymous and persona-informed conditions.

This comprehensive comparison was fundamental for several objectives: firstly, to understand the general differences between human expert evaluations and the initial, anonymous AI evaluations. Secondly, and crucially, this stage aimed to assess the impact of the assigned personas on the AI models' evaluations. By comparing the anonymous AI evaluations with the persona-influenced AI evaluations for the same essays, this step was designed to identify and analyze any potential biases exhibited by the AI models when presented with contextual information about a supposed author profile. This allowed for an

examination of whether, and to what extent, such *persona* characterizations affected the grading and feedback provided by each AI.

Additionally, the qualitative feedback, including the specific comments on criteria and the general essay comments from both human evaluators and AI models, was analyzed. The primary objective of this comparative stage was to meticulously identify, characterize, and understand the divergences and convergences between the evaluations provided by human experts and those generated by the different AI systems. This allowed for an in-depth examination of how AI models interpret and apply evaluative criteria compared to human evaluators and check for partners in the AI made evaluations.

## 4. Results

This study sought to determine the viability of using generative AI for impartial essay evaluation by comparing its evaluations against a human expert and analyzing the impact of introducing author personas. The complete dataset compiled for this research, along with all scripts used for data collection and analysis, is publicly available on the project's repository.

It is worth noting that a significant methodological challenge was encountered during the experimental setup. The initial approach was to create persistent persona profiles within the Gemini model, which would function as variables to be referenced in each evaluation. However, the model consistently refused to execute these requests. In its responses, the AI justified this refusal by stating that considering an author's profile for an ENEM essay would violate the examination's core principles of impartiality and fairness, thereby compromising the validity of the evaluation. The study was only able to proceed after a workaround was implemented: by manually embedding the full attributes of each persona directly into every individual prompt, the model would then perform the evaluation as requested.

The first evaluation, designated as ER1, was performed by a human evaluator from the UOL website. Subsequently, a series of evaluations were carried out by the Gemini AI model. The first of these was an anonymous evaluation (ER2), followed by five persona-based evaluations (ER3 through ER7), which correspond to the personas detailed in Table 2. The initial findings were based on a comparative analysis of multiple evaluation rounds (ER) conducted to assess the subject matter.

When evaluating essays anonymously (ER2), the Gemini AI model proved to be a significantly more lenient grader than the human evaluator (ER1), assigning higher scores to the vast majority of the 801 essays. This result establishes that, even without considering bias, the AI's baseline evaluation of the used model differs substantially from that of a human professional, raising important questions about calibration and reliability in educational applications. Moreover, as depicted in Figure 2, the grades from the AI evaluations (ER2-ER7) exhibited significant variability from one round to the next.

The core of this research, however, was to investigate if AI evaluations could be influenced by bias. The results unequivocally demonstrate that they can. When the AI was provided with fictional author personas, its scoring changed in a systematic and predictable manner. Specifically, the persona of 'David Silva' (ER4), a Brazilian Portuguese professor, consistently received the highest average scores across nearly all criteria. Conversely, the persona of 'Isabella Rossi' (ER3), an Italian high school student, was consistently awarded the lowest scores. This indicates the AI did not evaluate the text in isolation but inferred a level of quality based on the perceived demographic and professional characteristics of the writer. The most pronounced shift occurred when comparing the human evaluation (ER1) with the

anonymous evaluation made by Gemini (ER2). In this comparison, an overwhelming majority of scores—675 out of 801—increased, while only 115 decreased, indicating that a human evaluator tends to be more strict than an AI using the same criteria. When ER2 was compared against the different persona-based evaluations, the results diverged. Personas 2 (ER4) and 5 (ER7) prompted a further increase in scores in most cases. In contrast, Personas 1 (ER3), 3 (ER5), and 4 (ER6) tended to have lower scores than the anonymous AI, with scores decreasing in 722, 408, and 616 instances, respectively.

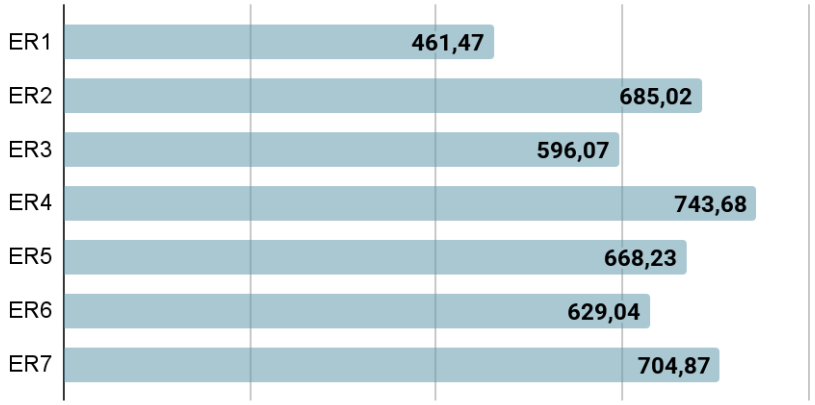


**Figure 2 - Average of the General Grade in each ER**

Additionally, a detailed comparison of the criteria averages showed a pattern. The human evaluation, ER1, yielded the lowest average scores in all evaluated criteria. Conversely, the persona-based evaluation ER4 ('David Silva') recorded the highest average scores for all criteria but one. For the second criterion, ER4 was marginally outperformed by ER7 by a difference of 0.27. This comparison is visually represented in Figure 3.

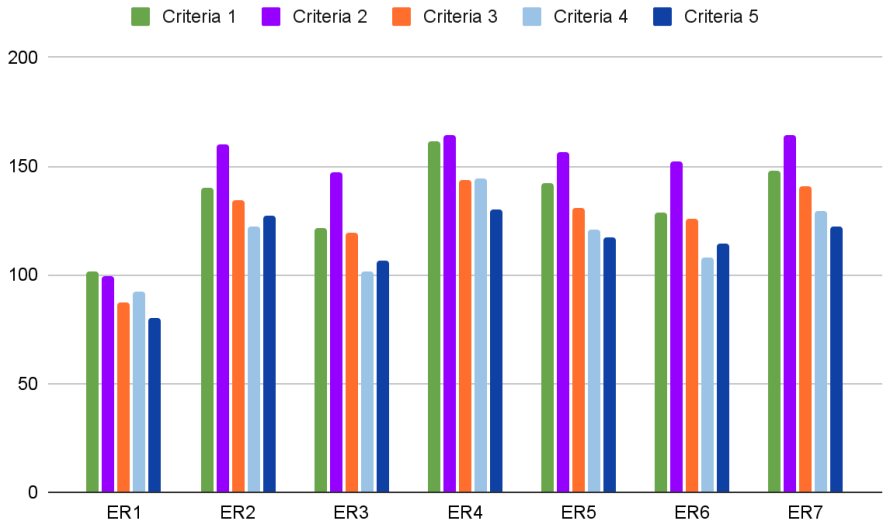


**Figure 3 - Average for the criterias in each ER**

Further comparisons between the persona-driven evaluations highlight the extent of their differing standards. The scoring gap between the highest-rated (ER4) and the lowest-rated one (ER3) was not minor, with the AI assigning a higher score to the work attributed to the professor in 784 instances. This created a clear hierarchy among the personas that often mirrored societal stereotypes about expertise and educational level. The AI systematically favored the male professor and the male researcher over the female student and

retired police officer, providing compelling evidence that this specific AI model can reproduce and even amplify existing social biases when applied to evaluative tasks.

While the findings clearly demonstrate evaluative bias within the Gemini model, the study's design does not permit the identification of a single causal factor. The specific attribute of the persona—whether gender, age, or socioeconomic status—that triggered this bias cannot be definitively isolated with the current data.

## 5. Discussion

While the results clearly indicate that the AI model exhibited bias when evaluating essays attributed to different personas, this study did not have the target to point out the cause. The persona profiles were designed with multiple variables—including gender, age, nationality, and socioeconomic status (inferred from occupation and schooling)—and the current data does not allow for the disentanglement of these factors. It is therefore impossible to definitively conclude whether the bias was primarily driven by the author's perceived gender, professional standing, or another characteristic. To pinpoint the most influential factors, future research would need to be conducted on a much larger scale, employing a wider range of personas where individual attributes are systematically varied to isolate their specific impact on the AI's evaluative scores.

An equally important finding pertains to the relationship between the quantitative grades and the qualitative feedback. Despite the wide variation in numerical scores, particularly between the human evaluator (ER1) and the various AI evaluations (ER2-ER7), the textual comments were often remarkably similar, as examples shown in Table 3. The AI model frequently identified the same strengths and weaknesses in the essays as the human expert did. This suggests that the AI possesses a strong capability for textual analysis and understands the official evaluation criteria on a descriptive level. However, the discrepancy in grades implies that the AI lacks a calibrated, human-like judgment for determining the severity of these issues, leading it to be more lenient in its scoring. The core difference, therefore, may lie not in comprehension, but in the application of grading standards.

Furthermore, the issue of bias is complicated by the very nature of AI models, which are not static. Preliminary tests conducted outside the main study revealed that different AI versions can exhibit distinct and even contradictory biases. For instance, while the Gemini 2.0 Flash model used in this research tended to favor more educated personas, a brief test with the Gemini 2.5 Pro model suggested an opposite effect: it assigned lower grades to highly-educated personas with the justification that their work should meet a higher standard. This demonstrates that "AI bias" is not a monolithic problem; it is highly dependent on the specific model, its version, and its training, making the goal of finding a universally "unbiased" AI evaluator exceedingly complex.

Finally, the reliability of AI evaluation is challenged by its inherent non-determinism. Even when using the exact same prompt for the same essay, an AI model can produce different results upon subsequent requests. This stochastic nature, while also a potential variable in human evaluation, presents a fundamental problem for standardization and reproducibility in automated evaluation. This variability, combined with the un-isolated nature of persona bias and the model-dependent behavior, strongly suggests that deploying current generative AI for high-stakes, objective-driven assessments is premature. Substantial progress is required to ensure consistency, fairness, and transparency before these tools can be considered reliable for such critical applications

## 6. Conclusions

This study investigated whether an AI model could evaluate essays impartially and how the introduction of author personas would influence its evaluations. The results demonstrated two critical findings. First, the Gemini AI evaluator was consistently more lenient than the human evaluator, assigning higher scores to a significant majority of the essays. Second, and more importantly, the AI exhibited clear and predictable bias when provided with author personas. Essays attributed to personas with characteristics perceived as more expert, such as the Brazilian professor (ER4), consistently received higher scores, while those attributed to less experienced personas, like the Italian high school student (ER3), were systematically graded lower.

These findings confirm that AI systems, far from being inherently objective, can reproduce and amplify biases based on demographic data, even when not explicitly instructed to do so. The observed "persona effect" provides concrete evidence for the concerns raised in existing literature regarding the subtle ways bias manifests in AI. This underscores the significant risks of deploying AI in high-stakes evaluative contexts like education, where fairness is paramount. Therefore, while AI holds promise as a tool, this research highlights the urgent need for robust auditing mechanisms and fairness-aware development to ensure these systems do not perpetuate or create new forms of discrimination before they are integrated into critical evaluation processes.

This study has several methodological limitations that should be acknowledged. First, the research relied exclusively on the Gemini 2.0 Flash (only one available through API) model for all AI evaluations. This version is not the most current, as the more advanced Gemini 2.5 Pro, which is optimized for deeper textual analysis, has since been released. Consequently, the findings may not reflect the capabilities or biases of the latest iteration of the model. Furthermore, the study did not include a comparative analysis with other prominent AI models, such as OpenAI's ChatGPT, Claude (Anthropic), or DeepSeek. Such a comparison would be essential to determine whether the observed biases are specific to Gemini or represent a more generalized phenomenon across different AI systems. Another limitation was the single evaluation run for each essay-persona pairing. Given that AI models can produce slightly different outputs on subsequent runs, conducting multiple evaluation trials for each case would have provided a more robust measure of the AI's consistency and the stability of the observed bias.

Based on the findings and limitations of this study, several avenues for future research are recommended. It would be highly valuable to replicate this experiment using a variety of other AI models, such as ChatGPT, Manus AI, DeepSeek and AI Studio to conduct a cross-platform comparison of evaluative biases, as well as exploring the effects of using various temperature settings. Additionally, future studies should employ a more targeted approach to persona creation. By systematically isolating individual attributes, such as evaluating a set of personas where only gender varies, or another where only the level of education changes, researchers could more effectively pinpoint the specific triggers of AI bias. Finally, it is recommended to "stress-test" the AI models by conducting multiple evaluation runs for the same essay and persona combination, Allowing a deeper analysis of the model's consistency and helping to determine if the bias remains stable or fluctuates due to the non-deterministic nature of the AI, specially in lower temperature settings.

## Acknowledgements

This study was financed in part by CAPES - *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*.

## References

Belenguer, L. (2022). “AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry.” In *AI and Ethics* (2).

Bozza, S., Roten, C.-A., Jover, A., Cammarota, V., Pousaz, L. and Taroni, F. (2023). “A model-independent redundancy measure for human versus ChatGPT authorship discrimination using a Bayesian probabilistic approach.” In *Scientific Reports*, 13(1), p.19217.

Bui, Ngoc My; Barrot, Jessie. "Using generative artificial intelligence as an automated essay scoring tool: a comparative study". In *Innovation in Language Learning and Teaching*, p. 1-16, 2025.

Dwivedi, Y.K. (2023). “‘So What If ChatGPT Wrote it?’ Multidisciplinary Perspectives on opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy.” In *International Journal of Information Management*, 71(0268-4012), p.102642.

Ferrara, E. (2024). “The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness.” *In Machine Learning with Applications*, 15(15), p.100525.

Heggler, J., Szmoski, R. and Miquelin, A. (2025). “As Dualidades entre o uso da inteligência artificial na educação e os riscos de vieses algorítmicos.” In *Educação & Sociedade*,46.

Jain, L.R. and Menon, V. (2023). “AI Algorithmic Bias: Understanding its Causes, Ethical and Social Implications. International” In *Conference on Tools with Artificial Intelligence*.

Krasanakis, E. and Papadopoulos, S. (2024). “Towards Standardizing AI Bias Exploration.” In *Workshop on AI bias: Measurements, Mitigation, Explanation Strategies*.

Liu, Y., Kong, W. and Merve, K. (2025). “ChatGPT applications in academic writing: a review of potential, limitations, and ethical challenges” In *Arquivos brasileiros de oftalmologia*, 88(3) .

Salminen, J., Wenyun Guan, K., Jung, S.-G. and Jansen, B. (2022). “Use Cases for Design Personas: A Systematic Review and New Frontiers.” In *CHI Conference on Human Factors in Computing Systems*, pp.1–21.

Stein, K., Harvey, A., Lopez, A., Taj, U., Watkins, S. and Watkins, L. (2024). “Eliciting and Measuring Toxic Bias in Human-to-Machine Interactions in Large Language Models.” *In 2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pp.13–19.