TECHNICAL UNIVERSITY OF CLUJ-NAPOCA

ACTA TECHNICA NAPOCENSIS

Series: Applied Mathematics, Mechanics, and Engineering
Vol. 67, Issue III, September, 2024

ETHICAL USE OF LARGE LEARNING MODELS IN ACADEMIA

Raymond MAIORESCU, Augustin SEMENESCU

Abstract: *This article explores the ethical integration of large language models (LLMs) in academia, focusing on authorship integrity, bias in AI outputs, and the risks of plagiarism and data fabrication. It advocates for guidelines to distinguish between human and AI contributions and highlights the need for accountability and transparency to maintain scholarly integrity. This study examines the propensity for bias in LLM-generated content and explores statistical methodologies to discern AI-generated material in educational contexts. Additionally, the paper analyzes the manifestation of bias in LLM outputs, underscoring the need for detection and correction mechanisms.*

Key words: *large language model, academic integrity, ethical AI usage, authorship, originality, machine learning, academia.*

1. INTRODUCTION

Integrating large language models such as GPT-4 into academic environments has opened new frontiers in research and education. While these advanced AI tools offer significant benefits in processing and generating vast amounts of textual data, they also introduce complex ethical challenges, notably bias in AI outputs and maintaining academic integrity.

This article delves into the implications of bias in LLM-assisted research, revealing how inherent biases in training data can skew research outcomes and impact educational content. We explore this through a blend of quantitative analysis, including mathematical models and visual charts, to objectively measure and display these biases. The discussion extends to a hypothetical case study, providing practical insights into the challenges and mitigation strategies for bias in LLM outputs within academic settings.

Equally critical is the issue of academic integrity in the age of AI. The line between student-generated and AI-assisted work is increasingly blurred, raising fundamental questions about authorship and originality in academic submissions. We address this challenge by presenting a logistic regression model with ROC curves and confusion matrices to distinguish between AI-generated and student-authored texts. A detailed case study showcases how an educational institution grappled with and strategized around these challenges. [1]

The article aims to contribute to the ongoing discourse surrounding LLMs in academia by concentrating on critical areas. It underscores the need for a balanced approach that leverages the benefits of LLMs while sensibly addressing their ethical implications, ensuring that their integration into academia aligns with the core values of scholarly work.

2. EVOLUTION OF LLMs IN ACADEMIA

LLMs' emergence and rapid evolution in academia can be traced back to their origins as essential text analysis tools. These models have undergone a transformative journey driven by significant advancements in machine learning and computational power. The leap from rudimentary text processing to sophisticated, context-aware systems marks a paradigmatic shift in how textual data is interpreted and interacted with in academic settings. [2]

In research, LLMs have emerged as invaluable assets. Their extensive data analysis

Received: 30.07.24; **Similarities**: 20.08.24: **Reviewed**: 31.07./31.07.24: **Accepted**:21.08.24.

capabilities enable comprehensive literature reviews, hypothesis generation, and complex data synthesis. Crucially, their proficiency in natural language understanding has proven instrumental in fields such as linguistic analysis and social science research. LLMs' ability to process and generate human-like language has expanded the scope of existing research domains and facilitated cross-disciplinary studies, offering more profound insights into areas like human-computer interaction, cognitive science, and ethics. [3]

In the sphere of teaching and learning, LLMs have revolutionized traditional methodologies. As personalized learning assistants, these models provide custom educational content, respond to student queries, and assist in language learning and tutoring. Beyond individualized support, LLMs have enabled the development of interactive educational tools, including AI-driven simulations and scenario-based learning modules. Such advancements enhance student engagement and learning experiences across diverse disciplines, showcasing the potential of AI in reshaping educational landscapes.

3. ETHICAL CONCERNS OF USING LLMs IN ACADEMIA

Integrating LLMs into academic settings has brought a spectrum of ethical challenges to the forefront. While LLMs offer unprecedented opportunities for enhancing research and education, they also raise critical questions about data ethics, bias, academic integrity, and potential misuse. These challenges, stemming from the capabilities and applications of LLMs, necessitate a proactive and thoughtful approach to ensure their ethical use in scholarly environments.

A primary ethical concern is the potential for LLMs to perpetuate and amplify biases found in their training data. These biases, if unchecked, can skew research outcomes and influence educational content, raising serious questions about the equity and objectivity of AI-assisted academic work. The challenge lies in detecting these biases and implementing strategies to mitigate their impact, ensuring fair and unbiased outcomes in LLM applications. [4]

Another significant ethical concern is preserving academic integrity in the age of AI. The advanced capabilities of LLMs to generate coherent and sophisticated text blur the lines between student-created and AI-assisted work. This poses challenges in authorship attribution and increases the risk of plagiarism. The academic community faces the task of developing and enforcing guidelines that delineate the ethical use of LLMs in academic writing, ensuring that the foundational values of originality and intellectual honesty are upheld.

While this article focuses on bias and academic integrity, it is essential to acknowledge other ethical considerations. Data ethics, privacy, consent, and security are essential, especially when LLMs are trained on sensitive datasets. Additionally, the potential misuse of LLMs in academia, such as for facilitating plagiarism or fabricating research data, underscores the need for rigorous ethical standards and mechanisms for accountability and transparency. [5]

4. ADDRESSING BIAS IN LLM-ASSISTED RESEARCH

In academic research, integrating LLMs presents a significant ethical challenge: the perpetuation of biases. These biases often reflect societal and historical prejudices, which can significantly impact the objectivity and fairness of research outcomes. Researchers must identify and mitigate these biases to ensure the integrity and inclusivity of AI-assisted academic work.

This section offers a comprehensive examination that interlaces quantitative and visual analyses, utilizing mathematical models and data charts to illuminate the nuances of bias, and advocates for a proactive discourse on mitigation strategies, aiming to preempt bias and maintain research integrity. This dual approach ensures that the employment of LLMs is reflective and principled, addressing potential ethical issues while promoting the development of robust, unbiased scholarly work.

4.1 Hypothetical research project

A university research team employed an LLM to study social behaviors, drawing on diverse sources like social media, scholarly articles, and historical documents, blending sociology, psychology, and data science.

Ethical concerns arose as the project progressed, and the LLM started showing significant gender and racial biases in its training data. It jeopardized the research's integrity and highlighted the potential to perpetuate stereotypes. [6]

The research group conducted a detailed statistical analysis of the LLM output, seeking more profound insights. They calculated bias ratios and then incorporated statistical indicators like standard deviation and the coefficient of variation for a comprehensive understanding of the data.

4.2 Quantitative analysis of bias

To quantify bias in LLM outputs, the study employs a statistical model focusing on various bias categories such as gender and race. The bias ratio was calculated for each category using the following formula:

$$R_c = \frac{B_c}{T_c} \qquad (1)$$

where R_c is the bias ratio for category c, B_c is the number of biased instances, and T_c is the total number of instances in that category. For example, if out of 1000 instances in the 'Gender' category, 250 instances were found to be biased, the bias ratio would be 0.25 or 25%.

The quantified bias ratios are visualized in Table 1, including a bar chart, and pie chart for more comprehensive visualization.

Table 1

Bias ratio – hypothetical data representing LLM output.

Category	**Total Instances, T_c**	**Biased Instances, B_c**	**Bias Ratio, R_c**
Gender	1000	250	0.25
Race	800	120	0.15
Age	1200	60	0.05
Socioeconomic	700	140	0.20
Geographic	900	180	0.20

The bar chart shown in Figure 1 outlines bias ratios by category, revealing more pronounced bias in *Gender* than in *Age*, underscoring the need for targeted gender bias mitigation. The categories average a bias ratio of 0.17, denoting an average 17% bias incidence.

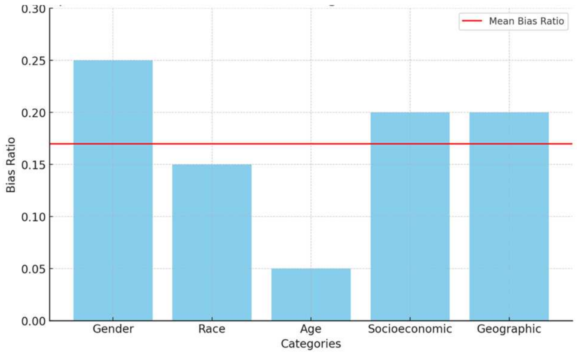

Fig. 1. Bias ratios in different categories

Figure 2 shows the distribution of different types of biases in LLM outputs. It visually represents how each kind of bias contributes to the overall bias in LLM outputs.

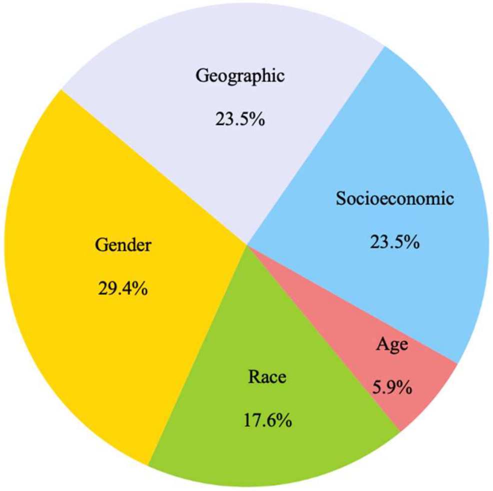

Fig. 2. Distribution of bias types in LLM output

4.3 Calculation of standard deviation

Standard deviation σ measures the variation or dispersion in a set of values and is calculated using the formula:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} \qquad (2)$$

In the dataset, x_i denotes individual values, μ signifies the dataset's average, and N is the total count of values. To gauge the variance in bias across different categories, we compute the standard deviation of the bias ratios, roughly 0.0680. This figure measures the spread of the bias ratios relative to their average, with a smaller standard deviation implying the ratios are more tightly grouped around the mean.

4.4 Calculation of the coefficient of variation

The coefficient of variation (*CV*) is a standardized measure of probability or frequency distribution dispersion and is calculated using the formula:

$$CV = \frac{\sigma}{\mu} \tag{3}$$

In this case, the coefficient of variation is approximately 39.90%. This measure of relative variability indicates that the degree of variation in bias ratios, relative to the mean, is quite significant. A higher coefficient of variation suggests a higher level of dispersion around the mean.

4.5 Bias mitigation in LLM-assisted research

This section underscores the imperative of confronting and rectifying bias in research utilizing LLMs. It presents a synthesis of quantitative scrutiny and a case study to unveil effective tactics for detecting and remedying bias. Such measures are crucial to maintaining equity, objectivity, and inclusivity in AI-enhanced academic research, thereby preserving academic excellence and ethical norms. [7]

Universities must embrace systemic reforms to avert any predicaments in forthcoming studies and create frameworks mandating bias evaluations at various phases of research, integral to the methodology, to facilitate ongoing bias oversight and amendment. Furthermore, institutions should commit to developing a compulsory AI ethics course for all AI-utilizing researchers, cultivating an ethos of ethical consciousness and accountability.

5. ACADEMIC INTEGRITY IN THE AGE OF AI-ASSISTED WRITING

Students' increasing use of LLMs for essays and research papers raises questions about originality and authorship. LLMs' advanced writing abilities make distinguishing between student-created and AI-produced content challenging. This scenario calls for a critical look at the ethics of AI in academia, urging educators and administrators to leverage AI's benefits while upholding the core principles of academic honesty.

5.1 Distinguishing between student-authored and AI-generated text

The surge of AI-assisted writing tools in academia requires a robust statistical approach to preserve academic integrity. Educational institutions must employ machine learning algorithms to develop predictive models distinguishing between student-authored and AI-generated text. [8]

To illustrate this quantitatively, we consider a logistic regression model trained on a labeled dataset comprising features extracted from known AI-generated and student-written texts. These features include linguistic patterns, complexity metrics, and stylistic markers unique to each domain.

5.2 Logistic regression model to differentiate between AI and student-written text

In a simulated scenario, we utilize logistic regression to determine the probability of a text being AI-generated. This binary classification method uses a logistic function to convert natural numbers into probabilities, distinguishing AI-created text, marked as (1), from student-written text, denoted as (0). [9]

Specifically, the model predicts the probability *p* that a given text instance *x* is AI-generated based on the logistic function:

$$p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_n x_n)}} \tag{4}$$

where $p(x)$ is the probability that a text is AI-generated, *e* is the base of the natural logarithm, β_0 is the intercept, β_1, β_2, …, β_n are the coefficients corresponding to feature values x_1, x_2, ..., x_n. [9]

In our simplified example, we consider two informative features, but in a practical setting, we include features such as text complexity, vocabulary diversity, and syntax variation.

The training of the model, as depicted in Table 2, uses a dataset with pre-determined classifications to guide the logistic regression for accurate feature weighting. This setup trains the model to classify new texts. "Feature_1" and "Feature_2" represent actual text-derived linguistic features, while "Label" denotes if the text is student-written (0) or AI-generated (1).

Table 2

Linguistic features and corresponding labels

Feature_1	Feature_2	Label
12.34	4.56	0
11.22	5.43	1
13.45	3.21	0
10.56	4.78	1
9.87	6.54	0

The model's effectiveness is then assessed on a separate test dataset, utilizing tools like the ROC curve and a confusion matrix to evaluate its performance.

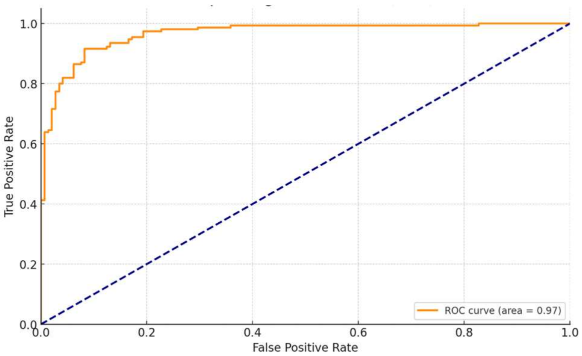

Fig. 3. Receiver Operating Characteristic (ROC) Curve

The confusion matrix illustrates the model's performance, showing true positives, false positives, true negatives, and false negatives. Represented in a chart, it gives a clear picture of the model's accuracy in classifying texts correctly.

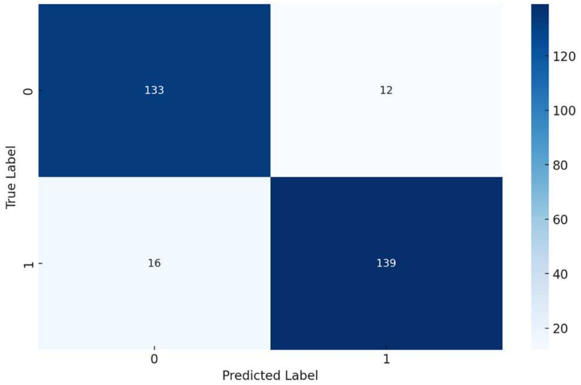

Fig. 4. Confusion matrix

The matrix shows the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions made by the model, while the values in the matrix provide insight into the accuracy and misclassification rates.

5.3 Addressing ethical concerns in AI-assisted writing

Academic institutions can maintain integrity by utilizing logistic regression models that differentiate AI-written from student-written texts, trained on specific features from both. The effectiveness of our model is demonstrated by a ROC curve with an impressive AUC of 0.97. The related confusion matrix shows high accuracy in text classification, marking it as a vital instrument for ethical academic monitoring.

6. CONCLUSION

The exploration of LLMs in academic settings has unveiled a landscape rich with potential yet fraught with ethical complexities. This article has highlighted two predominant challenges: the propagation of biases in AI-generated content and the preservation of academic integrity in an era increasingly influenced by AI.

Our analysis has shown that while LLMs hold immense potential for enhancing research and pedagogy, their responsible use mandates rigorous scrutiny to prevent the perpetuation of existing biases. The statistical models and visual charts underscore the need for ongoing attention and proactive measures to detect and mitigate bias. These tools serve as means for quantification and vital instruments for raising awareness and guiding corrective strategies.

Similarly, the integrity of academic work in the context of AI assistance requires a delicate balancing act. The logistic regression model demonstrated in this article is a testament to the innovative approaches that can distinguish between AI-generated and student-authored texts. This is essential in upholding the values of originality and authenticity that form the cornerstone of academic scholarship.

Looking ahead, integrating LLMs into academia is not merely a technological shift but a paradigm change that calls for reevaluating ethical frameworks. It requires a collective effort from technologists, educators, ethicists, and policymakers to ensure that the advancement of these powerful AI tools aligns with the ethical principles and standards of academic excellence.

In conclusion, as we stand at the cusp of a new era in education and research shaped by AI, the academic community is tasked with steering this integration in a direction that harnesses the power of LLMs and safeguards the ethical principles of academia. By addressing the challenges of bias and academic integrity head-on, we can pave the way for an educational future that is both innovative and ethically sound.

7. REFERENCES

[1] J. G. Meyer et al., *ChatGPT and large language models in academia: opportunities and challenges*, BioData Min., vol. 16, no. 1, p. 20, Jul. 2023.

[2] M. Perkins, *Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond*, J. Univ. Teach. Learn. Pract., vol. 20, no. 2, Feb. 2023.

[3] L. De Angelis et al., *ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health*, Front. Public Health, vol. 11, 2023, https://www.frontiersin.org/articles/10.3389/fpubh.2023.1166120

[4] S. Porsdam Mann, B. D. Earp, N. Møller, S. Vynn, and J. Savulescu, *AUTOGEN: A Personalized Large Language Model for Academic Enhancement—Ethics and Proof of Principle*, Am. J. Bioeth., vol. 23, no. 10, pp. 28–41, Oct. 2023.

[5] H. Li, J. T. Moon, S. Purkayastha, L. A. Celi, H. Trivedi, and J. W. Gichoya, *Ethics of large language models in medicine and medical research*, Lancet Digit. Health, vol. 5, no. 6, pp. e333–e335, Jun. 2023.

[6] M. Hosseini and S. P. J. M. Horbach, *Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review,* Res. Integr. Peer Rev., vol. 8, no. 1, p. 4, May 2023.

[7] M. S. Orenstrakh, O. Karnalim, C. A. Suarez, and M. Liut, *Detecting LLM-Generated Text in Computing Education: A Comparative Study for ChatGPT Cases*, arXiv, Jul. 10, 2023.

[8] Z. Liu, Z. Yao, F. Li, and B. Luo, *Check Me If You Can: Detecting ChatGPT-Generated Academic Writing using CheckGPT*, arXiv, Jun. 07, 2023.

[9] Hex, *Getting started with logistic regression*, https://hex.techlogistic-regression

[10] Scikit-learn, *Linear Models*, https://scikit-learn/stable/modules/linear_model.html

Utlizarea etică a modelelor mari de limbaj (LLMs) în mediul academic

Acest articol explorează integrarea etică a modelelor mari de limbaj (LLMs) în mediul academic, concentrându-se pe integritatea autorului, bias în rezultatele AI și riscurile plagiatului și fabricării datelor. Articolul propune stabilirea unor metode care să distingă între contribuțiile umane și cele AI și subliniază necesitatea responsabilității și transparenței pentru menținerea integrității academice. Studiul examinează tendința de bias în conținutul generat de LLM și explorează metodologii statistice pentru a discerne materialele generate de AI în contexte educaționale. În plus, lucrarea analizează manifestarea bias-ului în rezultatele LLM, subliniind necesitatea mecanismelor de detectare și corectare.

Raymond MAIORESCU, Eng., PhD Candidate, University Politehnica of Bucharest, Faculty of Industrial Engineering and Robotics, raymond.maiorescu@stud.fiir.upb.ro, Splaiul Independentei 313, sector 6, 060042, Bucharest, Romania.

Augustin SEMENESCU, PhD Eng., Professor, University Politehnica of Bucharest, Faculty of Material Science and Engineering, augustin.semenescu@upb.ro, Splaiul Independentei 313, sector 6, 060042, Bucharest, Romania.