

Generative AI-Enabled Chatbot for Improving Students' Understanding and Awareness of Academic Integrity Policies

Claudio Gonzalez^{1,*}, Faithful Chiagoziem Onwuegbuche^{1,2}

¹National College of Ireland, Dublin, Ireland.

²SFI Centre for Research Training in Machine Learning (ML-Labs), University College Dublin, Ireland.

Abstract

Academic integrity is a fundamental principle in education, ensuring the validity of learning and the quality of awarded degrees. However, societal and technological advancements have significantly influenced the understanding and frontiers of academic integrity, often requiring updates and redefinition. As a result, students frequently struggle to comprehend the appropriate use of new tools and practices. The advent of Generative Artificial Intelligence (GenAI) models, such as ChatGPT, presents a new and significant challenge to higher education institutions, particularly regarding their ethical integration into the learning process. This study explores the potential of leveraging GenAI to enhance students' understanding of academic integrity guidelines in the context of AI-driven education and to develop strategies for mitigating academic misconduct. The research employs a comparative analysis of Chatbots fine-tuned on national academic integrity regulations. The generated responses are evaluated against reference texts using advanced semantic similarity metrics. Based on the findings, the study provides recommendations for selecting the most effective Chatbot for an automated pre-course module aimed at improving students' comprehension of academic integrity policies. This investigation involves deploying six fine-tuned large language models (LLMs) enhanced with Retrieval-Augmented Generation (RAG) techniques. The performance of these models is assessed through post-test evaluations using metrics such as ROUGE, Pearson correlation, Cosine similarity, Jaccard similarity, BERT, Doc2Vec, SBERT, and InferSent scores.

Keywords

Generative AI, Academic Integrity, Chatbot, Large Language Models

1. Introduction

Academic integrity is a cornerstone of education, particularly in higher education institutions (HEIs), where it plays a vital role in validating knowledge production and ensuring the credibility of awarded degrees. It forms the ethical foundation upon which researchers and professionals build their careers. Concerns about academic integrity are not new; they trace back to the origins of academia in ancient Greece and have evolved alongside societal and technological advancements, which continue to shape education and society at large [1].

Two significant events have profoundly disrupted educational practices in recent years and raised new challenges for maintaining academic integrity: the COVID-19 pandemic and the rise of Generative Artificial Intelligence (GenAI) [2]. These events have reshaped education and influenced economies, social structures, and individual routines, leaving lasting impacts on societal norms. They have necessitated a reevaluation of educational practices, particularly as HEIs adapted to remote learning, online assessments, and the accompanying increase in academic misconduct stemming from difficulties in maintaining oversight [3].

The COVID-19 pandemic, as the first disruption, forced an immediate transition to remote education, creating significant challenges in preserving academic integrity. Almost concurrently, the emergence of large language models (LLMs) like ChatGPT introduced a new set of concerns. These models, capable of generating human-like responses and simulating vast knowledge bases, have drawn widespread attention from researchers, educators, and the public [1]. As a subset of AI, GenAI has been developed

AICS'24: 32nd Irish Conference on Artificial Intelligence and Cognitive Science, December 09–10, 2024, Dublin, Ireland

*Corresponding author

✉ x22244794@student.ncirl.ie (C. Gonzalez); faithful.onwuegbuche@ncirl.ie (F. C. Onwuegbuche)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

since the 1950s, but its recent advances have sparked heightened interest in its potential applications, ethical implications, and future development trajectories [1, 4, 5]. The implications of these developments are significant. Recent studies reveal that nearly a quarter of students engage in some form of academic misconduct. While such cases had been declining before 2020, the pandemic reversed this trend, prompting HEIs to update their policies and introduce new technologies for detecting misconduct, including bans on GenAI tools. Despite these measures, challenges persist, particularly in enhancing students' understanding of academic integrity and fostering adherence to its principles [3, 4, 6].

Previous research has addressed these challenges from various angles, highlighting that while policy enforcement is essential, creating a culture of academic integrity requires more than regulatory frameworks. It demands active engagement among institutions, students, and staff to promote practices that safeguard educational standards and maintain institutional credibility [1, 6, 5]. Building such a culture involves sustained efforts in communication, training, and the internalization of academic integrity as a core value [7, 8]. Traditionally, these initiatives have been implemented through face-to-face programs, including modules, workshops, and lectures. Increasingly, researchers advocate for integrating GenAI into educational processes rather than banning these tools outright. GenAI has shown promise in fields like computer science, where it can enhance learning experiences, such as by assisting with coding education [9, 10, 11].

This study operates on the premise that GenAI can serve as a valuable educational resource when guided by clear policies aligned with institutional values and the principles of academic integrity. The research aims to recommend implementing a fine-tuned LLM Chatbot designed to summarize and clarify academic integrity policies. This Chatbot is envisioned as a pre-work module for first-year students at HEIs, helping them understand academic integrity standards and promoting ethical academic practices. This study is structured as it follows. The first section presents the Related Work examining the existing research related to Education, GenAI and Academic Integrity. The subsequent part outlines the Research Methods and specifications that will be used to address the research question followed by the Experiments results and evaluations. The document concludes with the project's findings, discussion and directions for future research.

The research makes the following contributions:

- Developing various fine-tuned LLMs enhanced with RAG techniques, leveraging official government documents from the National Academic Integrity Network to improve students' comprehension of academic integrity policies.
- Comparing and fine-tuning multiple AI models to identify the most effective system for translating complex academic policies into actionable feedback tailored to specific university guidelines.
- Evaluating the impact of student-AI interactions on their understanding and adherence to academic integrity standards, providing insights for effectively implementing AI-driven solutions in educational institutions.

2. Related Work

2.1. Students Understanding and Awareness of Academic Integrity Policies

The persistent challenge of academic integrity in higher education has been widely studied, with researchers employing diverse approaches to analyze students' understanding and awareness while proposing potential solutions. This project builds on recent studies, particularly those addressing the disruptions caused by the COVID-19 pandemic and the rise of AI technologies. A recurring theme in these studies is the complexity of fostering academic integrity in today's rapidly evolving educational landscape.

For instance, Simon [12] examined the role of plagiarism detection software in cultivating a culture of integrity within academia and found that while these tools have proven effective in identifying misconduct, they alone cannot ensure the preservation of academic integrity. This limitation is especially pertinent in the context of AI detection tools, which continue to be a subject of debate. Similarly, Birks

[13] conducted an analysis proposing a framework for preventing academic misconduct, arguing that merely monitoring students' attitudes and behaviours is insufficient. Instead, the study advocated for broader, community-wide discussions to address the root causes of misconduct. Gallent-Torres [6] supported these findings, using similar methodologies to stress the necessity of comprehensive interventions aimed at fostering academic integrity.

The challenges posed by disruptions such as COVID-19 and AI to academic integrity have been an issue of interest recently with most studies emphasising the need for a transdisciplinary approach to understanding and addressing these issues better [1, 14, 15]. The importance of educating students on the ethical implications of AI, highlighting that strategies for preventing academic misconduct cannot adopt a "one size fits all" approach, instead, they should advocate for human-centred methodologies to enhance student success. Similarly, Michel-Villarreal [16] emphasized the need for clear policies, guidelines, and frameworks to integrate new technologies into higher education responsibly. It is important to consider user experiences and perceptions when developing such policies, particularly in the context of AI's expanding role in education. A common thread across these studies is the recognition that publishing, updating, and refining academic integrity policies is insufficient to address misconduct effectively. Deeper engagement is required, incorporating the academic community and emerging technologies into efforts to promote integrity. Anohina [7] aligns with this perspective, arguing that policies alone do not foster a culture of academic integrity, emphasizing the importance of educational tools and initiatives to enhance students' understanding of integrity.

In summary, while these studies provide valuable insights into academic integrity challenges, they also highlight a critical gap: the lack of comprehensive and proactive solutions. Current measures often fail to address the root causes of misconduct, necessitating innovative approaches—such as AI-driven tools—to bridge this gap and effectively promote academic integrity.

2.2. Generative AI and Academic Integrity

Research on GenAI in education has predominantly focused on its applications in foundational computing areas. However, its integration into more complex domains, such as academic integrity, remains in its developing stages and has yet to produce relevant results. Nonetheless, advancements in AI technologies signal the potential for more effective pedagogical integration, specifically the use of GenAI in computer science education, specifically through AI-powered Chatbots in coding modules. These studies demonstrated improved student performance when using Chatbots as learning assistants, particularly among students with a strong initial knowledge base [9, 10]. However, the studies also noted that the long-term implications of AI integration in classrooms remain uncertain due to limited assessment periods. Similarly, Gupta [8] used the ANOVA metric to analyze variance in student performance, advocating for the adoption of GenAI in education while highlighting the need for structured frameworks to ensure its effective implementation.

Another critical area of research involves the evaluation of AI-generated text detection tools. Weber [17] analyzed commercial tools by blending human and AI-generated text, concluding that their accuracy and reliability were suboptimal at the time of testing. Likewise, Alexander [3] examined four AI detectors, confirming similar inefficiencies. Elkhataat [18] compared GPT-3.5 and GPT-4's ability to detect textual similarities, finding no statistically significant performance differences and suggesting alternative approaches to enhance AI detection. Ibrahim [19] focused on fine-tuned RoBERTa-based classifiers, which, despite detecting AI-generated texts, demonstrated inconsistent accuracy across a dataset of 240 human-written and ChatGPT-generated essays. Perkins [5] further emphasized the sophistication of GenAI in producing original and coherent text, often eluding detection by existing technologies.

Some studies have investigated GenAI's potential as a learning assistant in specific educational contexts. For example, Kumar [20] explored AI's use in grading student papers, highlighting its discretion, convenience, and pedagogical value in delivering consistent feedback that improves student outcomes. Similarly, Caravantes [21] employed GPT for reviewing academic papers and compared its outputs with human reviews, finding notable parallels.

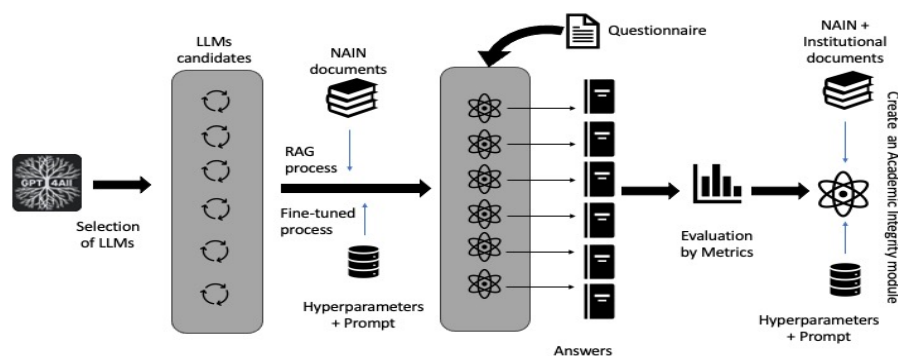


Figure 1: Project's Implementation Flow

Maryamah [22] proposed a Chatbot implementation using RAG to retrieve information from relevant documents. The study employed Recall and Precision metrics for information retrieval and BLEU and ROUGE scores to evaluate answer generation. Building on this approach, the current project expands the number of models evaluated and incorporates additional metrics to assess performance comprehensively. Maryamah also emphasized that successful pedagogical projects must transcend mere information repetition or skill mimicry. The ultimate objective is to empower students to achieve higher-order cognitive processes—such as analysis, evaluation, application, and creation—grounded in the knowledge provided.

These studies illustrate diverse approaches to integrating GenAI into education and academic integrity, each with distinct strengths and limitations. Collectively, they underscore the challenges of defining clear strategies for leveraging GenAI in pedagogical practices. This research addresses three key issues identified in the literature: (1) the potential of AI as a constructive educational resource; (2) the role of insufficient student understanding in academic misconduct; and (3) the necessity of selecting and fine-tuning models to achieve optimal outcomes. Moreover, the study extends previous efforts by introducing a comprehensive evaluation framework to assess the effectiveness of GenAI models in enhancing students' understanding of academic integrity.

3. Methodology

3.1. Research Method

This study employs a mixed-methods approach to assess the effectiveness of GenAI Chatbots in improving students' understanding of academic integrity policies. The research methodology consists of several structured stages. In the initial stage, standardized data on academic integrity in Ireland, validated by relevant government authorities, will be collected. This dataset will form the foundation for training the selected LLMs. Key information within the dataset will be identified and extracted for use during the model training process. Concurrently, a questionnaire will be developed to evaluate the models' performance regarding the academic integrity guidelines.

The next stage involves a detailed analysis of available LLM candidates. Each model will be assessed based on its specifications, training requirements, and potential applicability within educational contexts. From this analysis, the LLMs that best align with the project's objectives and usability criteria will be selected. These chosen models will then be fine-tuned using NLP techniques, with the curated academic integrity data serving as the primary training material.

Once the fine-tuning process is complete, the developed questionnaire will be used to generate responses from the trained models. These responses will be systematically stored for comparison against reference data. To ensure that the model outputs align with the original academic integrity policies, a range of evaluation metrics will be applied, including ROUGE, Pearson correlation, Cosine

similarity, Jaccard similarity, and embedding-based methods such as BERT, SBERT, Doc2Vec, and InferSent.

The final stage involves analyzing the collected data to evaluate the models' responses. This analysis will compare the LLM-generated outputs with reference answers and rank the models based on their performance across the various metrics. Statistical tools, including spreadsheets and RapidMiner, will be utilized to conduct the analysis. Based on this data-driven evaluation, recommendations will be formulated regarding the most suitable LLM for enhancing academic integrity comprehension in HEIs. These recommendations will include insights into fine-tuning and adapting LLMs for similar educational applications.

To implement the project, Google Colab will serve as the integrated development environment (IDE), and Python will be used as the primary programming language. The research workflow, illustrated in Figure 1, begins with loading pre-trained large language models from the GPT4All repository. These models will then undergo fine-tuning using reference documents related to academic integrity policies. Subsequently, their outputs will be evaluated against reference data using the selected metrics. The results will be analyzed, categorized, and ranked, leading to evidence-based recommendations for HEIs to improve academic integrity practices through the deployment of LLMs.

3.2. Large Language Models

The selection of appropriate LLMs is a critical aspect of this research, as the chosen models must be fine-tuned to align with the specific goals of the project. To accommodate budgetary constraints often faced by educational institutions, commercial solutions were excluded due to their fees, contractual limitations, and concerns over the handling of sensitive data used in training. A local implementation approach was adopted to ensure data security and control. From the various open-source options available, Nomic's GPT4All was selected. Initially developed as a distillation of ChatGPT 3.5, GPT4All has grown into a robust LLM repository, offering both desktop and Python client access. The platform's collaboration with open-source projects, such as LangChain and the Weaviate Vector Database, provides enhanced flexibility, stability, and compatibility with the project's technical setup [23].

The selection process within the GPT4All repository prioritized models with characteristics suitable for this research, including parameters, model size, and compatibility with hardware limitations. Given the constraints of standard laptop specifications, the pool of models was narrowed down to approximately twelve, with six models ultimately selected for further evaluation based on their alignment with the project's objectives. While these models may not rank highest in performance or popularity, their selection reflects a key objective of this project: to develop a sustainable and cost-effective solution that higher education institutions (HEIs) can implement without requiring high-end infrastructure. Table 1 provides an overview of the selected models.

The models were configured with the following parameters to optimize performance: Context Length = 2048, Max Length = 4096, Temperature = 0.3, Top-P = 0.2, Top-K = 40.

A significant aspect of fine-tuning involved designing a specific prompt to ensure accurate and contextually relevant responses. Following prompt engineering recommendations from [24, 25], a complex-instruction-following prompt was developed. This prompt explicitly defined the model's role, capabilities, and limitations while incorporating the knowledge base and language level requirements. The aim was to minimize hallucinations, maintain accuracy, and ensure alignment with the project's goals. The final iteration was stated as follows:

- "You are an academic integrity expert analyst bot. You can access the documents related to academic integrity, and you will base on them to answer. Your function is to help the students, and you can respond in a way that a university student level can understand, but you can get into detail if required. It would help if you always refused to answer questions unrelated to this knowledge base. You will be penalised if you refer to anything outside the documents you were trained on, [...]"

Model Name	Description	# Parameters	Creator
Llama 3 8B Instruct	4.34 GB of size. It is recognisable by its fast responses.	8 Billion	Meta
Mistral Instruct	3.83 GB of size, strong overall fast instruction model.	7 Billion	Mistral AI
Mistral Open Orca	3.83 GB of size. Fine-tuned on Open Orca dataset curated via Nomic Atlas	7 Billion	Mistral AI
GPT4All Falcon	3.92 GB of size. An instruction-based model trained and fine-tuned by Nomic AI.	7 Billion	Nomic AI
Ghost v0.91	3.83 GB of size. A variation of Mistral Instruct for fast responses.	7 Billion	Mistral AI
MPT Chat	3.54 GB of size. An MPT chat-based mode with novel architecture.	7 Billion	Mosaic ML

Table 1
Models included in the project

3.3. Information Gathering

To effectively fine-tune a pre-existing LLM, it was essential to gather comprehensive and authoritative data on academic integrity, including guidelines, policies, and relevant examples from the target institutions. In Ireland, Quality and Qualifications Ireland (QQI) serves as the central body for standardizing academic integrity through its initiative, the National Academic Integrity Network (NAIN) [26].

Established in 2019, NAIN provides a framework to foster academic integrity across Irish higher education institutions (HEIs). Its primary objectives include addressing academic misconduct, cultivating a culture of integrity, and developing tools to ensure consistent implementation of integrity practices. NAIN's network spans public and private institutions and integrates student representation to promote inclusivity and diverse perspectives.

The data collection process focused on the official documentation and resources provided by QQI, as summarized in Table 2. The entire textual content of these documents was incorporated into the RAG process. This ensures that the fine-tuned LLMs are explicitly informed by and aligned with these resources, guaranteeing relevance and adherence to established academic integrity standards.

By extracting essential information from these official sources, a custom GPT model can be fine-tuned to generate accurate responses, contextually relevant, and strictly aligned with the specific content it has been trained on. This meticulous approach aims to create a reliable tool for enhancing students' understanding of academic integrity policies.

3.4. Implementation

Following the completion of the training and fine-tuning phases described in Section 3.2, a comprehensive testing phase was designed to evaluate the performance of the selected LLMs. This involved administering a drafted questionnaire to the Chatbots, with questions directly derived from government documents on academic integrity. The explicit inclusion of such questions ensured a robust framework for comparing the Chatbots' responses against reference data.

1. What is Academic Integrity?
2. What are the academic integrity principles and fundamental values?
3. To whom do the academic integrity policies apply?
4. What is considered academic misconduct?
5. What are the guidelines for Generative Artificial Intelligence?
6. What is the life cycle for the management of cases of academic misconduct?
7. What is the classification of alleged academic misconduct?
8. What are the recommendations for creating a culture of academic integrity?

Name of the Document	Description	# Pages	Source
Academic Integrity: National Principles and Lexicon of Common Terms	A guide that reflects both current trends and developments in the field.	30	National Academic Integrity Network
Academic Integrity: Guidelines	Guidelines to provide support and advice to Irish higher education providers.	30	National Academic Integrity Network
Generative Artificial Intelligence: Guidelines for Educators	Guidelines developed by NAIN as a response of GenAI.	28	National Academic Integrity Network
Framework for Academic Misconduct Investigation and Case Management	Framework for the identification, recording and management of cases of academic misconduct HEIs.	76	National Academic Integrity Network
National Academic Integrity Network Terms of Reference 2023-2024	NAIN base statutory document.	7	National Academic Integrity Network
Glossary for Academic Integrity	European Union reference document.	51	European Network for Academic Integrity
The Fundamental Values of Academic Integrity	Reference to facilitate and support a systemic movement toward cultures of academic integrity.	17	International Center for Academic Integrity

Table 2
Documents utilised in the RAG procedure

Each model's generated responses were stored as text files, and segregated from the reference documents for independent evaluation. The comparison process focused on assessing the accuracy of the Chatbots in replicating technical information from the provided documents while minimizing speculative content. To ensure a thorough and reliable analysis, a mixed-method approach was adopted, incorporating textual similarity and semantic similarity measures. This dual approach provided deeper insights into the models' performance, allowing for a more confident evaluation of the findings. The selection of metrics was based on both their prevalence in related literature and their robustness, with additional techniques chosen for their promising documentation and innovative application potential. The following metrics were implemented to evaluate text similarity:

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Evaluates text quality by comparing generated content to reference summaries. ROUGE calculates overlap at different levels, including unigrams, bigrams (ROUGE-N), and longest common subsequences (ROUGE-L). Python implementations were adapted from sources such as Medium and Git repositories [27, 28, 29].
- **Pearson's Rank Correlation:** Measures the statistical similarity between datasets, yielding scores between -1 and +1, where higher scores indicate greater similarity. Libraries and code were based on [30].
- **Jaccard Similarity:** Computes the ratio of shared elements to the total number of elements in two texts. It applies to various units, such as words or characters. The implementation followed guidance from [31].
- **Cosine Similarity:** Determines semantic similarity by representing texts as vectors and calculating the cosine of the angle between them. A smaller angle indicates higher similarity. Code was derived from [32].
- **BERT (Bidirectional Encoder Representations from Transformers):** A transformer-based NLP framework from Google that processes sentences holistically to generate contextual embed-

dings. The codebase was implemented using resources from [33].

- **SBERT (Sentence-BERT):** An extension of BERT that incorporates pooling layers to create sentence embeddings for improved understanding at the sentence level. The implementation was based on [34].
- **Doc2Vec:** Extends Word2Vec to generate vector representations of entire documents. This method includes two models: Distributed Memory (DM) and Distributed Bag-of-Words (DBOW). The code was adapted from [34].
- **InferSent:** Uses a bi-directional LSTM network to encode sentences into vector representations and infer semantic relationships (e.g., entailment, contradiction, neutral). Its implementation included a Sentence Encoder for creating embeddings and a Classifier for relationship classification. The codebase was adapted from [34].

3.5. Ethical Considerations of the Research

This research follows strict ethical guidelines. All materials, including academic policies, are accessed with proper permissions. The LLMs operate under non-commercial licenses, and proprietary tools are employed in compliance with their terms of service. Test materials are specifically designed essays to ensure the ethical use of resources.

4. Evaluation

To identify the best-performing model, a series of experiments were conducted and compared to a reference questionnaire. These metrics encompassed both quantitative word usage and semantic understanding. Several analyses were performed using the models, questions, and metrics. The results were averaged, plotted, and analyzed to identify patterns and determine the most effective model. A summary of these findings is presented below.

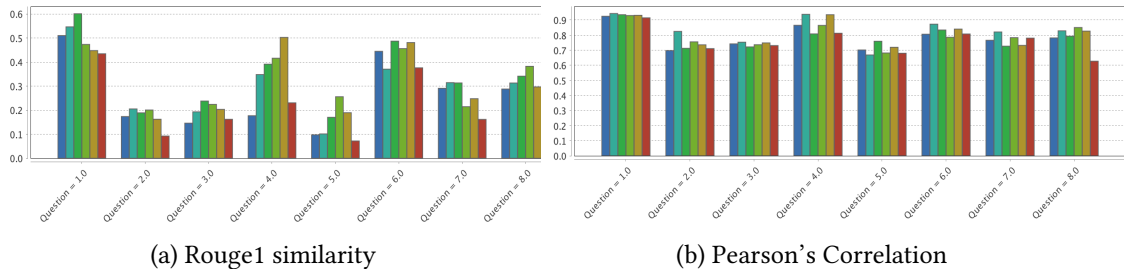


Figure 2: Scores by questions and model

- **The ROUGE** metric results indicate low similarity, with ROUGE-1 scores between 0.21 and 0.34, ROUGE-2 ranging from 0.059 to 0.125, and ROUGE-L between 0.13 and 0.2. As shown in Figure 2a, questions 1, 4, and 6 performed better, with questions 1 and 4 being the most literal (What is Academic Integrity? and What is considered academic misconduct?). Interestingly, question 6 (What is the lifecycle for managing academic misconduct?)—requiring summarization skills—also showed strong results. Overall, the best performers were Llama3 (0.34), Mistral Instruct (0.32), and Open Orca (0.31), with MPT Chat (0.21) performing the worst. Though the results showed low coefficients, especially in bigrams, this metric assesses summarization ability, which may be less applicable in a questionnaire format.
- **Pearson Rank's Correlation** delivered higher overall scores, reflecting strong word similarity with the reference text due to the temperature settings. As shown in Figure 2b, GPT4All Falcon achieved the highest average (0.83), leading in six out of eight questions, followed by Open Orca (0.8) and Mistral Instruct (0.79). MPT Chat scored the lowest (0.75). The consistency across questions suggests the models were trained on similar vocabulary and structure.

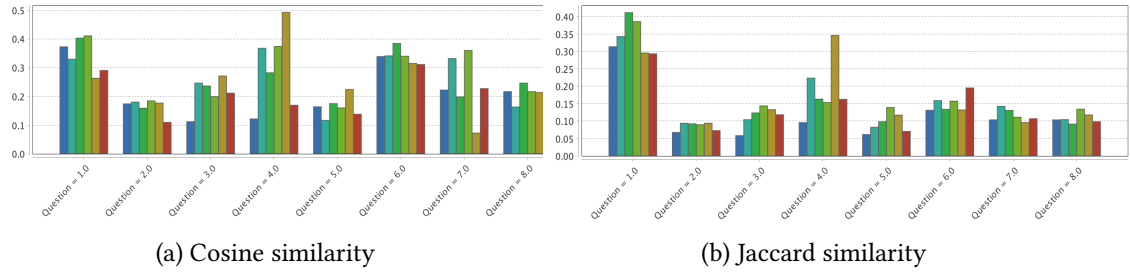


Figure 3: Scores by questions and model

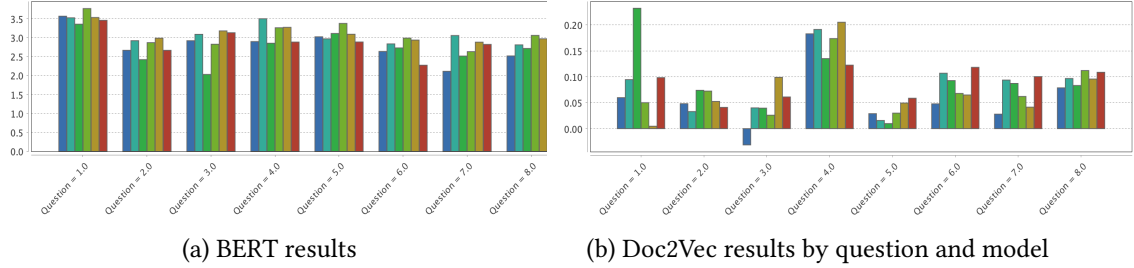


Figure 4: Scores by questions and model

- **Cosine Similarity** results were lower overall, with a distribution pattern similar to ROUGE-1. As shown in Figure 3a, the best performer was Mistral Instruct (0.28), followed by Llama3 (0.26) and Falcon (0.259). MPT Chat performed the worst with 0.19.
- **Jaccard similarity** produced low results, particularly resembling ROUGE-2 but slightly higher. As shown in Figure 3b, Mistral Instruct led in three questions (0.164), followed by Open Orca (0.166) and Falcon (0.156). Ghost7B had the lowest performance (0.116).
- **BERT model** produced a distinct range of results, similar in distribution to Jaccard. As shown in Figure 4a, Mistral Instruct led in four questions, although Open Orca (3.107) had the highest overall average. Mistral Instruct (3.098) and Falcon (3.088) followed, with Llama3 scoring the lowest (2.717).
- **Doc2Vec** performed poorly across most questions, except for question 4, where it improved. MPT Chat performed best, leading in three questions with an average of 0.088, followed by Llama3 (0.094), while Ghost had the lowest score (0.055) as shown in Figure 4b.
- **SBERT** results correlated closely with BERT and Jaccard. Figure 5a indicates that GPT4All Falcon performed best (0.83), leading in six questions, followed by Open Orca (0.808) and Mistral Instruct (0.798). MPT Chat had the lowest score (0.757).
- **The InferSent** model showed strong overall performance. Open Orca excelled in three questions with an average of 0.866, while Llama3 (0.874) had the highest average but led in only two questions. Mistral Instruct followed closely with 0.858, and MPT Chat lagged (0.803).

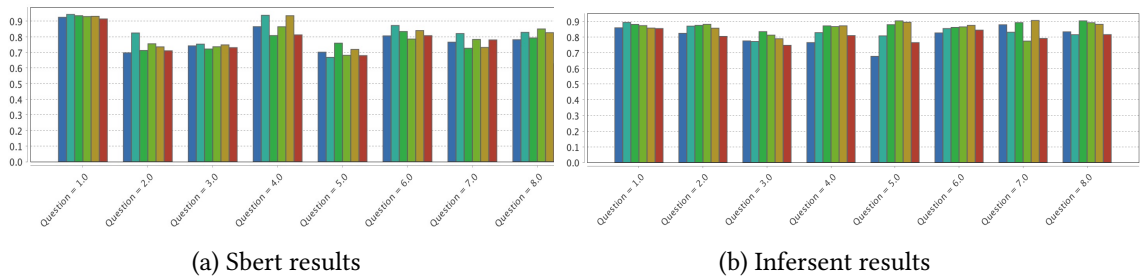


Figure 5: Scores by questions and model

	General	Ghost	Falcon	Llama 3	Mistral Instruct	Mistral Open Orca	MPT Chat
Rouge	Rouge 1	0.26576557	0.29882116	0.34119226	0.327641962	0.316200607	0.211376497
	Rouge 2	0.08849878	0.1229652	0.11621427	0.120003171	0.125245793	0.059637795
	Rouge L	0.16234959	0.17166613	0.18856242	0.184822867	0.200731463	0.129614594
	Rouge L Sum	0.16234959	0.17166613	0.18856242	0.184822867	0.200731463	0.129614594
Pearson's Rank Correlation		0.785175	0.830775	0.7859875	0.7980125	0.8082125	0.7575625
Cosine Similarity		0.21498157	0.25925224	0.26027614	0.280405213	0.253373885	0.19479162
Jaccard Similarity		0.11672208	0.15634619	0.15541746	0.164138946	0.166251061	0.139537303
Bert		2.79334981	3.08879856	2.71727016	3.098552138	3.1070595	2.852729788
Doc2Vec		0.05537341	0.08402775	0.09426688	0.074316891	0.07663806	0.088654177
Sbert		0.78517942	0.83076815	0.7859895	0.798012264	0.808217913	0.757552579
InferSent		0.8045992	0.83370822	0.87446796	0.858252846	0.866174906	0.803707577
Average		0.56675855	0.6226177	0.59165518	0.62627106	0.629894287	0.556798093

(a) Summary of results

(b) Ranking of Models

- 1° Mistral Open Orca
- 2° Mistral Instruct
- 3° GPT4All Falcon
- 4° Llama 3 8B
- 5° MPT Chat
- 6° Ghost 7B v0.91

Figure 6: Final Evaluation of the Model's Performance

4.1. Final Results

Based on the overall performance results 6a and the analysis of all metric outcomes, the final ranking is shown in Table 6b. Given these findings, the recommended large language model for implementing a pre-work module on Academic Integrity in higher education, tailored to the institution's specific policies and information, is **Mistral Open Orca**.

5. Conclusion and Future Work

5.1. Conclusion and Discussion

This study investigated the use of LLMs to enhance students' understanding of academic integrity policies, focusing on open-source options. The primary objective was to identify the most effective LLM for translating complex academic integrity guidelines into an accessible academic module. By fine-tuning and evaluating these models using documents from the NAIN, the research underscored the importance of high-quality training data and optimization processes. This finding challenges the assumption that larger models inherently produce superior results.

The experimental evaluation employed various similarity metrics to measure model performance. Results revealed that models like Mistral Open Orca and GPT4All Falcon outperformed others. For example, the ROUGE metric indicated relatively low summarization ability across all models, whereas Pearson's Rank Correlation demonstrated strong word-level similarity, particularly for GPT4All Falcon. Furthermore, semantic similarity metrics (BERT, SBERT, and InferSent) provided consistent results, reinforcing the importance of fine-tuning for domain-specific applications.

Among the evaluated models, Mistral Open Orca consistently ranked as the top performer across multiple metrics. This positions it as the most suitable choice for HEIs looking to integrate AI-driven tools to support academic integrity. The study highlights that proper fine-tuning and tailored training processes are vital for developing LLMs capable of achieving educational objectives effectively. In conclusion, our contributions include the development of fine-tuned LLMs enhanced with RAG using official documents, providing a framework for evaluating AI models in educational settings. By assessing the impact of AI interactions on students' understanding of academic integrity, we offer a data-driven solution for HEIs to improve policy communication and compliance through targeted AI applications.

The findings advocate for the adoption of Mistral Open Orca as an accessible and effective AI solution for promoting academic integrity within HEIs. This recommendation is supported by a methodologically rigorous approach that highlights the comparative strengths and limitations of various models in meeting the specific goals of this study.

5.2. Future Work

The rapid evolution of LLMs presents opportunities for future studies to explore advanced and more efficient models. While this research was limited by hardware constraints, scaling up to larger infrastructures could enable broader experimentation, including testing diverse hyperparameter configurations and newer models. Additionally, as evaluation metrics for text generation continue to advance, future studies may adopt these tools to refine methodologies and produce even more accurate assessments. These developments could further support the selection of AI models tailored to academic applications.

Finally, the study recognizes that the learning process extends beyond recalling academic integrity guidelines. Achieving higher-order learning outcomes, such as critical thinking, requires integrating human expertise into AI applications. Future research should involve educators in designing evaluations that complement the AI models, fostering deeper engagement and critical analysis of academic integrity policies.

References

- [1] S. E. Eaton, Postplagiarism: transdisciplinary ethics and integrity in the age of artificial intelligence and neurotechnology, *International Journal for Educational Integrity* 19 (2023) 23.
- [2] J. Prather, P. Denny, J. Leinonen, B. A. Becker, I. Albluwi, M. Craig, H. Keuning, N. Kiesler, T. Kohn, A. Luxton-Reilly, et al., The robots are here: Navigating the generative ai revolution in computing education, in: *Proceedings of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education*, 2023, pp. 108–159.
- [3] K. Alexander, C. Savvidou, C. Alexander, Who wrote this essay? detecting ai-generated writing in second language education in higher education., *Teaching English with Technology* 23 (2023) 25–43.
- [4] J. A. Oravec, Ai, biometric analysis, and emerging cheating detection systems: The engineering of academic integrity?., *Education Policy Analysis Archives* 30 (2022) n175.
- [5] M. Perkins, Academic integrity considerations of ai large language models in the post-pandemic era: Chatgpt and beyond, *Journal of university teaching & learning practice* 20 (2023) 07.
- [6] C. Gallent Torres, A. Zapata-González, J. L. Ortego-Hernando, The impact of generative artificial intelligence in higher education: a focus on ethics and academic integrity, *RELIEVE. Revista ELectrónica de Investigación y Evaluación Educativa*, 2023, vol. 29, num. 2, p. 1-19 (2023).
- [7] A. Anohina-Naumeca, I. Birzniece, T. Odiņeca, Students' awareness of the academic integrity policy at a latvian university, *International Journal for Educational Integrity* 16 (2020) 12.
- [8] T. Gupta, Research on the application of artificial intelligence in the education and teaching system, in: *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, IEEE, 2023, pp. 1168–1173.
- [9] J. Prather, B. Reeves, J. Leinonen, S. MacNeil, A. S. Randrianasolo, B. Becker, B. Kimmel, J. Wright, B. Briggs, The widening gap: The benefits and harms of generative ai for novice programmers, *arXiv preprint arXiv:2405.17739* (2024).
- [10] J. Prather, P. Denny, J. Leinonen, D. H. Smith IV, B. N. Reeves, S. MacNeil, B. A. Becker, A. Luxton-Reilly, T. Amarouche, B. Kimmel, Interactions with prompt problems: A new way to teach programming with large language models, *arXiv preprint arXiv:2401.10759* (2024).
- [11] S. Makeleni, B. H. Mutongoza, M. A. Linake, Language education and artificial intelligence: An exploration of challenges confronting academics in global south universities, *Journal of Culture and Values in Education* 6 (2023) 158–171.
- [12] W. Simon, Distinguishing between student and ai-generated writing: A critical reflection for teachers, *Metaphor* (2023) 16–23.
- [13] D. Birks, J. Clare, Linking artificial intelligence facilitated academic misconduct to existing prevention frameworks, *International Journal for Educational Integrity* 19 (2023) 20.
- [14] S. E. Eaton, The academic integrity technological arms race and its impact on learning, teaching, and assessment, *Canadian Journal of Learning and Technology* 48 (2022) 1–9.

- [15] S. E. Eaton, N. Chibry, M. A. Toye, S. Rossi, Interinstitutional perspectives on contract cheating: a qualitative narrative exploration from canada, *International Journal for Educational Integrity* 15 (2019) 9.
- [16] R. Michel-Villarreal, E. Vilalta-Perdomo, D. E. Salinas-Navarro, R. Thierry-Aguilera, F. S. Gerardou, Challenges and opportunities of generative ai for higher education as explained by chatgpt, *Education Sciences* 13 (2023) 856.
- [17] D. Weber-Wulff, A. Anohina-Naumeca, S. Bjelobaba, T. Foltýnek, J. Guerrero-Dib, O. Popoola, P. Šigut, L. Waddington, Testing of detection tools for ai-generated text, *International Journal for Educational Integrity* 19 (2023) 26.
- [18] A. M. Elkhataat, Evaluating the authenticity of chatgpt responses: a study on text-matching capabilities, *International Journal for Educational Integrity* 19 (2023) 15.
- [19] K. Ibrahim, Using ai-based detectors to control ai-assisted plagiarism in esl writing: “the terminator versus the machines”, *Language Testing in Asia* 13 (2023) 46.
- [20] R. Kumar, Faculty members’ use of artificial intelligence to grade student papers: a case of implications, *International Journal for Educational Integrity* 19 (2023) 9.
- [21] D. Carabantes, J. L. González-Geraldo, G. Jover, Chatgpt could be the reviewer of your next scientific paper. evidence on the limits of ai-assisted academic reviews, *Profesional de la información/Information Professional* 32 (2023).
- [22] M. Maryamah, M. M. Irfani, E. B. T. Raharjo, N. A. Rahmi, M. Ghani, I. K. Raharjana, Chatbots in academia: a retrieval-augmented generation approach for improved efficient information access, in: 2024 16th International Conference on Knowledge and Smart Technology (KST), IEEE, 2024, pp. 259–264.
- [23] Y. Anand, Z. Nussbaum, B. Duderstadt, B. Schmidt, A. Mulyar, Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo, <https://github.com/nomic-ai/gpt4all>, 2023.
- [24] P. Denny, J. Leinonen, J. Prather, A. Luxton-Reilly, T. Amarouche, B. A. Becker, B. N. Reeves, Prompt problems: A new programming exercise for the generative ai era, in: *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, 2024, pp. 296–302.
- [25] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, *arXiv preprint arXiv:2302.11382* (2023).
- [26] NAIN, Academic integrity: National principles and lexicon of common terms (2021). URL: <https://www.qqi.ie/sites/default/files/2021-11/academic-integrity-national-principles-and-lexicon-of-common-terms.pdf>.
- [27] E. Kızırmak, Text summarization: How to calculate rouge score (2023). URL: <https://medium.com/@eren9677/text-summarization-387836c9e178>.
- [28] Google, Python rouge implementation (2024). URL: <https://github.com/google-research/google-research/tree/master/rouge>.
- [29] P. Madiraju, Rouge your nlp results! (2022). URL: <https://medium.com/@priyankads/rouge-your-nlp-results-b2feba61053a>.
- [30] HuggingFace, Sentence similarity (2021). URL: <https://huggingface.co/tasks/sentence-similarity>.
- [31] NewsCatcher, Ultimate guide to text similarity with python (2022). URL: <https://www.newscatcherapi.com/blog/ultimate-guide-to-text-similarity-with-python>.
- [32] StackOverflow, How to compute the similarity between two text documents? (2021). URL: <https://stackoverflow.com/questions/8897593/how-to-compute-the-similarity-between-two-text-documents>.
- [33] PyPI, semantic-text-similarity (2019). URL: <https://pypi.org/project/semantic-text-similarity/>.
- [34] GeeksforGeeks, Different techniques for sentence semantic similarity in nlp (2024). URL: <https://www.geeksforgeeks.org/different-techniques-for-sentence-semantic-similarity-in-nlp/>.