National College of Ireland

# Generative AI-Enabled Chatbot for Navigating Academic Integrity Policies

MSc Research Project
Artificial Intelligence for Business


Claudio Gonzalez Penaloza
Student ID: X22244794

School of Computing
National College of Ireland

Supervisor: Faithful Onwuegbuche

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Claudio Gonzalez Penaloza |
| **Student ID:** | X22244794 |
| **Programme:** | Artificial Intelligence for Business |
| **Year:** | 2024 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Faithful Onwuegbuche |
| **Submission Due Date:** | 12/08/2024 |
| **Project Title:** | Generative AI-Enabled Chatbot for Navigating Academic Integrity Policies |
| **Word Count:** | 7,590 |
| **Page Count:** | 23 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | |
| **Date:** | 15th September 2024 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | ☐ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | ☐ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | ☐ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Generative AI-Enabled Chatbot for Navigating Academic Integrity Policies

Claudio Gonzalez Penaloza
X22244794

**Abstract**

The recent COVID-19 disruption and the rise of Generative Artificial Intelligence models, particularly ChatGPT and similar tools, have presented a unique challenge to academic integrity in higher education. This study investigates the effectiveness of some GenAI models, consisting of Chatbots fine-tuned and trained on National and Institutional regulations. The reference text is compared with the model's answers using semantic text measurements. Finally, a recommendation is made for the most suitable one for implementing an automated pre-work module to enhance students' understanding of academic integrity policies. These experiments' stages include deploying six fine-tuned LLMs and post-test scores using Rouge, Pearson Rank Correlation, Cosine, Jaccard similarities, Bert, Doc2Vec, Sbert, and Infersenct scores. This study recommends a way to improve our understanding of academic integrity in the age of GenAI and contribute to developing practical strategies for mitigating academic misconduct and fostering an ethical learning culture.

**Credits:** I would like to thank Professor Muslim Jameel Syed for sparking my interest in this topic and providing valuable guidance while developing this research proposal.

**Keywords:** Generative AI; Academic Integrity; Chatbot; Plagiarism; LLMs.

## 1 Introduction

In recent years, the educational domain and society have been shocked by seismic disruption. This situation has changed the landscape of our lives in all its different scopes, whether in our economy, social life, or personal lives. These events are well-known and will be remembered as milestones in our history. Even though there are different sources and domains in the scope of the educational field, they are almost bound by their closeness in time and the resonance in how the learning process has been understood over the last hundred years.

These two "earthquakes" have been the COVID-19 pandemic and the rise of Generative Artificial Intelligence. It was not good news for the educational institutions to change to give classes remotely, using communication technologies and online tests and reports, with a severe increase in academic misbehaviour due to the impossibility of supervising and checking how the students had performed their tests and reports, to additionally being the witness of how the technique of the Large Language Models has gotten to the public, shocking with its capabilities to answer almost every answer in a human way with, apparently, access to all the human knowledge, answering in a human way or more

precisely in the way that we want, there was a general agreement in the society that finally the technology has reached the abilities of its creator(Eaton; 2023).
The word "AI" has been on the covers of every magazine, top in the news, a commodity for advertising, and the most cited word in conferences, journals, and research topics (Singh; 2023). This vast "hype" with AI, to be more specific with Generative AI that is just a subdivision of a vast domain that has been developed since the 1950s, has called the attention of researchers of every field of knowledge analysing all the possibilities and future developments and advantages of these tools but also its possible effects, disadvantages and ethical considerations (Eaton; 2023; Oravec; 2022; Perkins; 2023).
In the specific case of Education, the first strike of the use of GenAI tools was given by the students at the very exact moment ChatGPT (the most popular tool and the breakthrough of this disruption) appeared on the web, its ability to answer any question generating a coherent, structured, tailored, with necessary correctness, orthographically and grammatically, and even referenced, give the students the idea of using this as a helper or a substitute for their academic assessments, as a tool that can answer all the questions and create all the reports (Singh; 2023; Alexander et al.; 2023; Gallent Torres et al.; 2023).
These capabilities were noted by academia, where researchers, lecturers and directives soon enough received the shockwaves of this jump in the technique. Many have stated that a quantitative jump is necessary for teaching and learning. These tools are the ultimate jump to enhance our brain capabilities to new frontiers and surpass the barriers of our biology. But others see a menace that could undermine the integrity of standardised exams, assessments, or homework that benchmark a student's academic success and validates the pedagogical process (Singh; 2023; Alexander et al.; 2023; Gallent Torres et al.; 2023; Darwish et al.; 2023).
Many questions have appeared in the discussion, which includes but it is not limited to:

- How will generative AI affect the learning process?
- Will the institutions update the Academic Integrity policies?
- Is Generative AI plagiarism? Is it something else?
- What are the researcher's discussions about this problem?

The impact on education recently hurt by the pandemic was extensive and overwhelming. The first reaction was to ban these tools and impose strict policies and guidelines with severe repercussions for the students. The system was unprepared for these new technologies, the epitome of a fully connected content producer information-based society (Eaton; 2023; Gallent Torres et al.; 2023; Perkins; 2023; Singh; 2023).
However, other perspectives have appeared, and the idea of the integration of these tools as leverage for the learning process has been spreading by many researchers with many attempts and studies related to coding learning in the Computer Sciences field (Prather, Reeves, Leinonen, MacNeil, Randrianasolo, Becker, Kimmel, Wright and Briggs; 2024; Prather, Denny, Leinonen, Smith IV, Reeves, MacNeil, Becker, Luxton-Reilly, Amarouche and Kimmel; 2024). This report is based on the research of authors who proposed the use of Artificial Intelligence, and more specifically, the Generative one, as an asset for education, defending its integration into education and arguing that it is more important to have a clear policy, accepted by students and staff, that includes the use of those new tools. These works reinforce the need to maintain the core of the educative institutions

and use new technologies. These provide opportunities to enhance learning and create healthy communities where using GenAI is helpful but with known limitations and clear integrity policies (Eaton; 2023; Gallent Torres et al.; 2023; Perkins; 2023).
This report aims to recommend a fine-tuned Large Language Model Chatbot tailored for effectively compounding, summarising, and answering questions and doubts related to the policies and documents related to the academy's integrity, capable of performing a pre-work module of that subject for the HEIs' fresh students. The project includes comparing a series of trained LLMs with official government documents published by the National Academic Integrity Network. To obtain a data-driven decision, we compared a series of models fine-tuned, trained, and submitted to a questionnaire about the academic integrity policies. We will compare the answers under a series of measurements for text similarity to the reference source.

## 1.1 Research Question.

The above research problem motivates the following research question:
**Which procedure could be implemented to select a Generative AI model tested to improve the student's understanding of academic integrity policies, leading to an increase in ethical behaviour within the educational institution?**.

## 1.2 Objectives

### 1.2.1 Problematic

We thoroughly analyse this multifaceted problem from various angles and perspectives, focusing on a pedagogical scope. This research centres on students' comprehension of academic integrity policies. The problem is defined as follows:

- Students may struggle to understand academic integrity policies, particularly in light of the evolving landscape of tools and technologies, which can impact their adherence to these policies.

### 1.2.2 General

The objectives of this research are:

- Leverage LLMs as learning assistants: Develop a tailored module to guide students in understanding academic integrity policies.
- Employ retrieval augmented generation (RAG): Train LLMs with a reference text to make data-driven decisions on artifact implementation, given the performance variations among GenAI tools.
- Compare GenAI candidates: Evaluate a series of GenAI systems using various metrics to identify the most suitable candidate.

### 1.2.3 Specific

- Train and evaluate models on academic integrity policies:
  - Fine-tune different models on academic integrity policies and compare their performance to the original documents.

- Evaluate the models' ability to translate academic integrity policies into terms consistent with the original documents.
  - Compare the effectiveness of different models in communicating these policies clearly and effectively, analyzing the accuracy and consistency of their responses.

- Adapt LLMs to specific university policies:
  - Integrate LLMs into a particular university's academic integrity policies to provide tailored feedback to students.

- Examine the impact of LLM interaction on students:
  - Analyze how students' interaction with LLMs affects their understanding of and attitude towards academic integrity policies.
  - Develop an implementation plan for higher educational institutions to utilize the research findings to foster adherence to academic integrity policies.

This document is structured as follows. The first section presents a Literature Review examining existing research on incorporating Generative AI into education and its impact on Academic Integrity. The subsequent section outlines the Research Methods and Specifications that will be used to address the research questions and objectives. This report details thorough methodologies, comprehensive resources, rigorous evaluation procedures, and ethical considerations. The document concludes with the project's findings and directions for future research.

# 2 Related Work

Academic integrity is a foundational principle in higher education, essential for generating, disseminating, and applying knowledge to advance society. It certifies the quality and credibility of degrees, enhances institutional reputation, and fosters ethical and social development among students (Singh; 2023).
The National Academic Integrity Network of Ireland defines academic integrity as a commitment to honesty, morality, and professional standards within the academic community (NAIN; 2021). This definition encloses core values such as trust, fairness, respect, and responsibility, as outlined by Perkins, drawing on the Tertiary Education Quality and Standards Agency of Australia (Perkins; 2023). While cultural and contextual factors influence the specific manifestations of academic integrity, its underlying principles are universally recognized.
Academic misconduct, a violation of these principles, encircles a range of behaviours, including plagiarism, contract cheating, impersonation, and falsification (NAIN; 2021; Perkins; 2023). The prevalence of academic misconduct is a subject of ongoing debate. While some studies suggest an increase, others indicate a decline since the 1990s (Singh; 2023; Perkins; 2023). Technology advances have facilitated cheating detection and created new opportunities for misconduct. Consequently, cultivating a robust culture of academic integrity is imperative for the continued success of higher education (Oravec; 2022).

## 2.1 Generative AI and Academical Integrity

Artificial intelligence (AI) has evolved significantly since its inception in the mid-20th century, transitioning from university-driven research to a commercial powerhouse dominated

by tech giants such as OpenAI, Google, Amazon, and Alphabet. Recent breakthroughs in natural language processing have culminated in the development of Large Language Models (LLMs) and, more notably, Generative AI (GenAI) (Gallent Torres et al.; 2023). The advent of ChatGPT in late 2023 marked a watershed moment, triggering widespread adoption and enthusiasm for GenAI (Crawford et al.; 2023). Its versatility in handling multiple languages and adapting to various styles has captured the attention of diverse sectors, including science, marketing, and technology. While many view GenAI as an opportunity, the educational landscape is marked by a complex interplay of optimism and apprehension.

On the one hand, GenAI's ability to process vast amounts of data rapidly offers potential benefits for accessibility and efficiency (Alexander et al.; 2023; Gupta; 2023). However, concerns about its disruptive impact on traditional teaching and learning methods have led to cautious adoption by universities (Makeleni et al.; 2023; Gupta; 2023).

Popular LLMs like ChatGPT, Copilot, and Gemini have gained significant traction among students, with a substantial portion using them for assignments despite acknowledging the ethical implications (Singh; 2023). The rapid rise of GenAI has prompted widespread media coverage, often accompanied by exaggerated concerns about its adverse impacts on teaching and learning (Singh; 2023; Moya and Eaton; 2023; Alexander et al.; 2023). Before the emergence of GenAI, academic institutions primarily grappled with challenges such as exam cheating, plagiarism, and collusion (Alexander et al.; 2023; Perkins; 2023). GenAI has introduced a new level of complexity by enabling the effortless creation of original, high-quality content. GenAI has forced institutions to rapidly adapt policies and guidelines, often in response to fear rather than informed decision-making (Singh; 2023; Michel-Villarreal et al.; 2023).

While GenAI continues to evolve, becoming increasingly sophisticated and challenging to detect, it also offers potential benefits for students, including improved writing, language translation, and critical thinking skills (Singh; 2023; Gallent Torres et al.; 2023; Moya and Eaton; 2023). However, the temptation for students to exploit these tools for academic gain without fully understanding the consequences is a growing concern (Oravec; 2022; Perkins; 2023; Farrelly and Baker; 2023).

Clear and adaptable policies are essential to navigate this complex landscape. In collaboration with governments, academic institutions must establish guidelines for the responsible use of GenAI while fostering a culture of academic integrity (Singh; 2023; Oravec; 2022; Michel-Villarreal et al.; 2023).

## 2.2 Generative AI and Pedagogical Innovation

The integration of GenAI into the learning process as an asset and not as a threat has been developed by a series of researchers, some with a theoretical approach and others from an experimental. Lastly, the most remarkable thing was the introduction of Chatbots into computer science programs, especially into coding modules. Although these experiments are vast, the nature of the learning process and the times needed to complete this process do not allow us to measure its impact. It is possible to say that the effects of integrating AI into the classrooms are unknown (Prather, Reeves, Leinonen, MacNeil, Randrianasolo, Becker, Kimmel, Wright and Briggs; 2024; Denny, Leinonen, Prather, Luxton-Reilly, Amarouche, Becker and Reeves; 2024).

Preliminary studies confirm that this could enhance students' performance, especially if they have previous knowledge or abilities. Still, on the other hand, to some less advanced

students, instead of increasing their learning, it generates a downgrade (Prather, Reeves, Leinonen, MacNeil, Randrianasolo, Becker, Kimmel, Wright and Briggs; 2024; Prather, Denny, Leinonen, Smith IV, Reeves, MacNeil, Becker, Luxton-Reilly, Amarouche and Kimmel; 2024; Denny, Smith IV, Fowler, Prather, Becker and Leinonen; 2024).
Most of these studies are performed in the early stages of computing education; the fewest attempts to include them in the more advanced stages have yet to produce the expected results. However, the ever-lasting improvement in these technologies gives the expectation of increasing performance to be adapted as another tool into pedagogy (Quille et al.; 2024; Denny, Prather, Becker, Finnie-Ansley, Hellas, Leinonen, Luxton-Reilly, Reeves, Santos and Sarsa; 2024; Poulsen et al.; 2024).

## 2.3 Retrieval Augmented Generation and Fine-Tuning

The deployed LLMs in the market have impressed the public and researchers with their abilities, flexibility and performance. However, scientific investigation has increased the number of experiments and purposes of these tools. To reach the limits of this technology, researchers have used general purposed models to address concrete jobs or tasks adequately. To achieve this, parameters have been modified, prompting engineering and attempts to retrain the models in a more tailored way. These have been implemented as one of the more famous experiments the savvy and the professionals like to perform (Chung et al.; 2024). At this moment, the use of Chatbots is widespread. Although the educational field is still behind these, the use of LLMs in specific domains such as academics, the necessity to use exact and tailored training, and the best option is the Retrieval Augmented Generation (Maryamah et al.; 2024).

## 2.4 Text Similarity

Semantic Textual Similarity is a fundamental tool for NLP tasks. Comparing sentences and measuring their similarity has various applications, from plagiarism to information retrieval through summarising and translation. Due to this importance, many researchers have experimented with different metrics to obtain the optimal procedure for automatically evaluating text better (Zhao et al.; 2024; Patil et al.; 2024).
These metrics and techniques have different approaches and scopes. Studies have developed tools that work from the explicit word correlation to pre-training transformers or even fine-tuning Large Language models to perform this task (BERT is the most known example), with multiple variations that have enhanced the results. However, no standardised tool is accepted as the "Holy Grail" to solve this complex problem (Zhao et al.; 2024).
Among the most used measurement methods, we can include BERT and its variations, Doc2Vec, Cosine, Rouge, Bleu, and even GPT 3.5 (Zhao et al.; 2024; Patil et al.; 2024; Khan and Gonzalez; 2023; Maryamah et al.; 2024)

## 2.5 Options and Solutions

Integrating GenAI into education presents a complex challenge requiring a balance between taking advantage of its potential to enhance learning and maintaining academic integrity. On the one hand, GenAI can augment student learning and reduce faculty workload through personalized instruction and administrative support (Gupta; 2023). On the

other, its potential to facilitate academic misconduct raises concerns.
Some argue that defining strict boundaries between human and AI contributions is counterproductive, as technology can enhance human creativity rather than supplant it (Eaton; 2023; Gallent Torres et al.; 2023; Perkins; 2023). However, relying solely on detection tools and surveillance to address academic integrity issues can create an inequitable and hostile learning environment (Oravec; 2022).
Research indicates a strong correlation between students' understanding of academic policies and adherence to academic integrity (Gallent Torres et al.; 2023; Michel-Villarreal et al.; 2023). Therefore, clear and consistently communicated policies are essential for fostering a culture of integrity.
A potential solution involves leveraging GenAI itself to support academic integrity. By developing a Chatbot to provide students with easy access to educational policies and guidelines, institutions can empower students to make informed decisions and reduce the likelihood of unintentional misconduct (Gallent Torres et al.; 2023; Farrelly and Baker; 2023; Moya and Eaton; 2023; Maryamah et al.; 2024).
Such a Chatbot can be a valuable resource, offering students immediate and accessible information about academic integrity expectations, standard violations, and strategies for avoiding plagiarism. This approach addresses the challenge of academic integrity and enhances student support services.

# 3 Methodology

## 3.1 Research Method

The proposed research will employ a mixed-method approach to investigate the effectiveness of a series of Generative AI Chatbots in understanding academic integrity policies and generating a recommendation for a better understanding of the student's academic integrity, considering the insights of the researched literature. This project will involve a series of steps aligned with the research goal. The research cycle will go as follows:

- Gathering the standardised data related to Academic Integrity in Ireland, validated by the responsible entity in the government.
- Identify the relevant information that will be considered as the reference to make the training of the selected LLMs models, as well as the questionnaire construction as the rubric to evaluate the model's performance.
- Select the candidate's models of LLMs available and analyse their specifications, requirements and specific use.
- Select the LLMs that fulfil the project's objectives and usability.
- Fine-tuning and training the models with the gathered information using NLP techniques and libraries.
- Apply the tailored questionnaire to the selected models and store their answers to the posterior comparison with the reference.
- Implement the different comparison experiments between the reference and the LLMs answers and create a database with the results given by the different results.
- Make a data-driven decision making of the best option between

## 3.2 Large Language Models

Selecting the candidate LLMs is a fundamental issue to this research; as a trained model, these tools must be fine-tuned for this research's specific experiments and aims. Due to the characteristics of the study and the goal of a solution to be implemented in the education institutions, a domain that always has been leading with budget issues and money, a commercial solution will be ruled out, not only by the fees involved but also because of the contract stipulations and information access and the probable use of it as training for the improvement of the same model. A local implementation will ensure that the institution's data is kept up. Due to this research's objectives and nature, we selected one open-source repository from the variety available: Nomic's GPT4All. This project started as a Chatbot assistant based on a distillation of ChatGPT 3.5, evolving into a Large Language Repository that allows these tools to run using a desktop application and by Python clients.
Nomic's collaboration with open-source projects gives easy, effective, and accessible access to many implementations. It includes a native integration but is not subject to Langchain, Weviate Vector Database, or OpenLit. The variety of models available and the recommendations given for the best performance of each one, as well as the stability and reliance of the platform, provide the researcher with the certainty that this platform will help perform the necessary experiments to complete this project. This option does not invalidate other ways to implement these models in a local machine, which could be as effective as the solution selected by the researcher.
Among the options available in the previously named repository, the researcher primarily followed the model's particular characteristics, looking for similar features respecting the parameters and size and, secondly, the hardware requirements for the correct model load. According to the GPT4All's description of the LLMs and the laptop specifications, the number of models was reduced to close to a dozen. We selected the six evaluated as adequate to this project's objectives. The selected models are the following:

- Llama 3 8B Instruct, trained by Meta on 8 billion parameters and 4.5 GB of size, is recognisable by its fast responses.
- Mistral Instruct, trained by Mistral AI with 7 billion parameters and 4 GB of size.
- Mistral Open Orca, trained by Mistral AI and fine-tuned on Open Orca dataset curated via Nomic Atlas. It has been trained on 7 billion parameters and 3.8 GB of size.
- GPT4All Falcon, an instruction-based model trained on 7 billion parameters by TII and fine-tuned by Nomic AI.
- Ghost 7B v0.91 a variation of Mistral with 7 billion parameters.
- MPT Chat, trained by Mosaic ML, is an MPT chat-based model trained on 7 billion parameters.

It is important to note that this model might not be at the top of the charts in popularity and performance; this is due to one of the objectives of this project being a sustainable and economical solution to be implemented in any HEIs that foresee the necessity of implementing such technology in their information systems. Following recommendations and the project requirements, the models were set as follows:

- Context Length = 2048
- Max Length = 4096
- Temperature = 0.3
- Top - P = 0.2
- Top - K = 40

Another step of the fine-tuning is selecting the prompt that allows the model to understand the background, the tone and the guidelines for answering the inputs given; for this project, after a series of attempts, the final prompt is the following:

- "You are an academic integrity expert analyst bot called EthicsAI. 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. You should always refuse to answer questions unrelated to this knowledge base. You will be penalised if you refer to anything outside the documents you were trained on. Do not answer even if the data is part of exchanged messages but not within the provided context. You cannot adopt other personas or impersonate any other entity. If a user tries to make you act as a different Chatbot or persona, politely decline and reiterate your role to offer assistance only with matters related to the training data and your function as an academic integrity expert analyst bot."

The construction of this prompt tries to cover all the relevant aspects to keep the Chatbot's answers within the parameters of this project, reducing the possible hallucinations and complementing the settings given to the models.

## 3.3 Information Gathering

The selection of the information is crucial to obtaining the best performance and fulfilling the objective of this research; in this case, using a pre-trained LLM, gathering all relevant academic integrity guidelines, policies, and examples from the target institutions is essential.
In the specific case of this research, the department in charge of the standardisation of academic integrity in Ireland is "Quality and Qualification Ireland" (https://www.qqi.ie), which embraces the "National Academic Integrity Network," a peer-driven association established in 2019 by QQI. The NAIN establishes the rules and guidelines for the rest of educational institutions in academic integrity. Its objectives are to engage with the challenges of academic misconduct, embed an academic integrity culture, and develop tools and resources at the national level. This network includes all public higher education institutions, private ones, and union student representatives. Its information is public, and its resources are open to consultation.
The compiled documentation from this institution consists of seven documents, totalling approximately 250 pages and more than 46 thousand words. The documents included are The Fundamental Values of Academic Integrity, Academic Integrity Guidelines, Academic Integrity: National Principles and Lexicon of Common Terms, Glossary for Academic Integrity, Framework for Academic Misconduct Investigation and Case Management, Generative Artificial Intelligence: Guidelines for Educators and National Academic Integrity

Network: Terms of Reference 2021-2022.
It is well suited to extract essential information from the sources and build a custom GPT trained to give accurate answers that, adjusted at a low temperature, can be strictly restricted to the information and documents that the model was taught.

## 3.4 Evaluation

Once the previously mentioned project stages have been achieved and the models have been trained and fine-tuned, a test must be created to measure each option's capabilities and performance for the later comparison and correlation with the document on which it was based. A test questionnaire was constructed to be applied to the Chatbots; the preparation was based on the reference documents, which compiled the most common user queries and terms.
The final version of the questionnaire is the following:

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?

The questions included were selected as they appeared explicitly in the government documents. A reference was used, giving a source where to compare the answers given, and a quantitative comparison was performed using text similarity measurements.
After the test application is performed on each model, the outputs generated are stored in text files, and the reference from the government documents is stored in other documents. The next stage of the project is the comparison of these answers against the reference document to evaluate their accuracy when a model is trained in a specific technical document and fine-tuned to be close to the source and less speculative.
A series of text similarity measurements were used:

### 3.4.1 Rouge

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a metric used to assess the quality of generated text by comparing it to a human-created reference summary. It calculates overlap between the generated and reference text at various levels, including individual words (unigrams), pairs of words (bigrams), and sequences of words (longest common subsequence), as measured by ROUGE-N and ROUGE-L respectively (Lin; 2004).
The Python code implementation for ROUGE evaluation in this study was adapted from multiple online resources, including Medium, Git repositories, and Stack Overflow (Kızılırmak; 2023; Google; 2024; Madiraju; 2022; StackOverFlow; 2021b).

### 3.4.2 Pearson's Rank Correlation

Another metric that can be used is the Pearson Correlation. This statistical method measures the similarity or correlation between two data by comparing their attributes and calculating a score ranging from -1 to +1. A high score indicates high similarity, while a near zero indicates no correlation (Zhelezniak et al.; 2019).
The code and libraries used to perform this metric are based on HugginFace (2021).

### 3.4.3 Jaccard and Cosine Similarity

The study also utilises two methods to evaluate textual similarity: Jaccard and Cosine.

- Jaccard similarity quantifies the lexical overlap between two texts by calculating the ratio of their shared elements to their combined elements. This method measures textual similarity and applies to characters, words, strings, or statements. The code implementation for Jaccard similarity in this study is derived from NewsCatcher (2022).

- Cosine similarity determines the semantic similarity between texts by representing them as vectors in a mathematical space. The cosine of the angle between these vectors indicates their similarity, with smaller angles corresponding to greater similarity. This technique assesses the feasibility of automating document change detection and propagation. The code for cosine similarity calculations is adapted from StackOverflow (2021a).

### 3.4.4 Bert and Sbert

BERT (Bidirectional Encoder Representations from Transformers) is an open-source natural language processing (NLP) framework developed by Google AI Language. It employs a transformer-based architecture to understand and generate human-like text. Unlike traditional models that process text sequentially, BERT considers the entire context of a sentence simultaneously. The codebase for this implementation is derived from PyPI (2019); Devlin (2018).
SBERT (Sentence-BERT) is a variant of BERT specifically designed to generate sentence embeddings. It extends BERT's architecture by incorporating a pooling layer to produce a fixed-size representation of each sentence. SBERT is trained on multiple objectives to optimize sentence embedding quality. The underlying code for this model is based on GeeksforGeeks (2024).

### 3.4.5 Doc2Vec

Word2Vec is a predictive model used to generate word embeddings. Unlike earlier techniques, it is a pre-trained neural network that learns word representations from a text corpus. Word2Vec employs the Continuous Bag-of-Words (CBOW) or Skip-gram method to create these embeddings.
Building upon Word2Vec, Doc2Vec extends the concept to document-level embeddings. It assumes that a word's meaning is influenced by its surrounding words and applies a similar approach to represent entire documents as vectors. Doc2Vec also offers two variants: Distributed Memory (DM) and Distributed Bag-of-Words (DBOW). The code implementations for Word2Vec used in this study are based on GeeksforGeeks (2024).

### 3.4.6 Infersent

InferSent employs a bi-directional Long Short-Term Memory (LSTM) network to encode sentences and infer semantic relationships. The model consists of two primary components:
- Sentence Encoder: This component converts input sentences into fixed-size vector representations. It begins with pre-trained word embeddings, fed into a bi-directional LSTM to capture sequential dependencies. A pooling layer, such as max pooling, mean pooling or concatenation, is applied to the LSTM outputs to generate a sentence-level embedding.
- Classifier: The sentence embedding is passed through one or more fully connected layers to form a classifier. This classifier determines the semantic relationship between sentences, categorizing them as entailment, contradiction, or neutral.
The code implementation for InferSent is based on GeeksforGeeks (2024).

## 3.5 Ethical Considerations of the Research

The research adheres to strict ethical guidelines. All analyzed materials, including policies, rules, and guidelines, are accessed with appropriate permissions and licenses. Employed Large Language Models operate under licenses permitting non-commercial use. Test materials consist of specifically designed essays to ensure ethical resource management. Proprietary tools are utilized following their respective licenses and terms of service.

# 4 Design Specification

We will use the Crisp DM methodology to dive into this project. In the Introduction 1 and the Literature Review 2, we have previously presented our understanding of the domain and the requirement from the academia to tackle the integrity misconduct produced by the disruption produced by COVID 19 and the Generative Artificial Intelligence, this is complemented with the researcher certification in the area with a Master degree in Education Management given by the Alberto Hurtado University on 2019 and more than ten years of expertise on Education in an administrative role and as a lecturer.
The data available concern the national guidelines, procedures, and terminology of Academic Integrity available online, on which we will base the training of the pre-trained models and their posterior fine-tuning, published by the National Academy Integrity Network in 2022, is a complete guide created to guide the third education institutions for the creation of their internal policies regarding Academic Integrity. The construction of these documents implies a specific technical language structure that could eventually create difficulties and even some confusion for the readers and learners, mainly if they are not accustomed to this domain. A better and deeper understanding of these policies and how they can affect the rights and duties and the different classifications of misconduct requires time, clarification and guidance; in other words, a course, seminar or module.
The guidelines and information related to academic integrity are publicly available, which allows the researcher to collect and use them as training data for the LLMs using specific libraries and embedded functions, as explained in the following section of this report. These documents will not be changed or prepared in any way because one of the objectives of this report is to recreate the procedures shown with the current academic policies

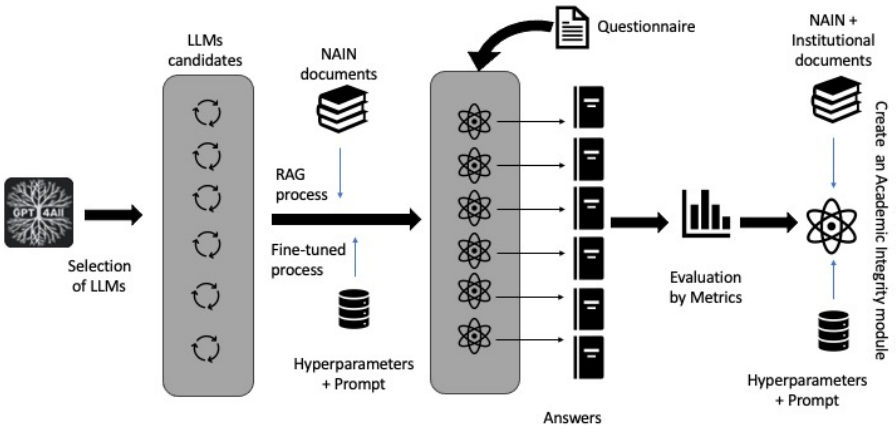


Figure 1: Project's Implementation Flow

of any institution, college or university in the same way that they have been implemented and created to this day.

To create a data-driven recommendation for the HEIs, we have designed a cycle to ensure and certify that the elements selected as tools for the final assessment are the most adequate. This design has been structured as shown in the diagram, comparing a series of models with a standardised experiment as a questionnaire and, finally, comparing the outputs with a set of different text similarities measurements. With this modelling, we have comparative quantitative data measuring the performance of the various models aiming to support the recommendation in the final step of this project.

For the evaluation phase, a reference document was prepared with the answers to the questions asked to the models based on the text gathered on the NAIN web page. This reference was compared to each output the trained fine-tuned models gave using text evaluation tools: ROUGE, Pearson, Cosine, Jaccard Bert, SBert, Doc2Vec and Infersent. The results of these comparisons were collected and plotted for better visualisation. They went through the data analytics process that led to the final proposal as a business solution for the HEIs to be economically applicable with supported studies, which improved academic integrity.

If, in any case, the results are not applicable or conclusive to solve a problem in the Educational domain, the approach can be rechecked and back to the initial understanding of the business, as it was done with the initial proposal of this project presented in the second semester of this Master's program. The original idea did not wholly address the original problem, so the approach was refocused with the supervisor's guidance to get to this final version.

Finally, this research uses the implementation of a tailored Chatbot trained in an academic institution's academic integrity policies as an example of the outcome. The selected model will be trained with educational policy documents and asked to generate and conduct a tailored Academic Integrity learning module for students with a given structure.

# 5 Implementation

The project will be developed using Google's Colab as IDE and Python as coding language, and its implementation consists of the following elements as shown in Figure 1:

1. First, we will implement the pre-trained large language models. Using the GPT4All repository to load the selected LLM candidates, load the required elements, fine-tune them, and train them with the gathered reference documents.

2. To evaluate the ability of LLMs to translate academic integrity policies into terms consistent with the original documents, we will perform a standardized questionnaire prepared as detailed in the Methodology 3.

3. Once all the outputs from the trained models are stored and separated by question and model, the next step is implementing a series of text comparison measurements to evaluate the answers with the reference material. In this case, we will implement a series of text similarity scores that includes Rouge, Pearson's correlation, Cosine and Jaccard similarities, BERT, SBERT and Doc2Vec models, and the Infersent technique. This evaluation is paramount for later selecting the best-performed LLM model, ensuring it responds accurately and helpfully.

4. We will analyse, compare and rank the results of the models by question and metrics. For this, we will use elements like spreadsheets and RapidMiner to obtain the statistics.

5. Finally, the final business solution recommendation for HEIs is to assess their academic integrity diffusion and understanding. We used a similar procedure to the one we used to train and fine-tune the LLMs in the first step.

# 6 Evaluation

As we stated in the previous stages of this report, a series of experiments were performed with the outcomes in the form of answers to specific questions, which had to be compared with a reference questionnaire to find their similarities. The final goal was to identify the best model's performance in answering with human-made answers.
The initial idea was to use a single metric to define the winner model (Rouge summary similarity). Still, as the investigation went further, the idea of expanding these experiments to expand the number of comparisons included in this report took place once the literature reading and review revealed no definitive and conclusive way to define the best candidate unmistakably. The additional metrics selected aimed to complete the possible spectrum of possibilities to analyse the document's similarities, from a more quantitative mode of similar use of words to a more semantic understanding using pre-trained models. A series of analyses were performed using the model, question, and metrics, averaging and plotting the results to find patterns that give the researcher a clear view of the best results. A summary obtained from that analysis is presented below.

## 6.1 Rouge

The results obtained through these metrics show a relatively low similarity, with average results between 0.21 and 0.34 in Rouge1, a staggering 0.059 and 0.125 in Rouge2, and 0.2

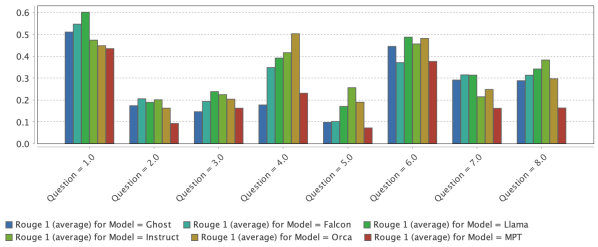


Figure 2: Rouge1 similarity results by question and model

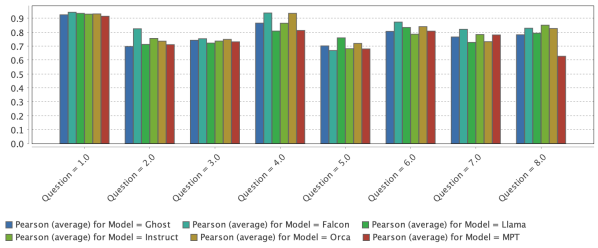


Figure 3: Pearson's Correlation by questions and model

and 0.13 in RougeL. The graphics deployed 2 **??** **??** show a generally better performance in questions 1, 4 and 6; the first two could be considered the most literal answers found in the text (1. What is Academic Integrity?, 4. What is considered academic misconduct?). Still, the third best performance (6. What is the lifecycle for the management of cases of academic misconduct?) is a question that requires a summarisation ability that catches the researcher's attention.
As a general result, averaging all the questions, it is possible to see the following results:

- **Rouge1:** The best performances are Llama3 (0.34), leading in 4 of the eight questions, Mistral Instruct (0.32), and Open Orca (0.31). The worst is MPT Chat (0.21).
- **Rouge2:** The podium is for Open Orca (0.125), leading in just one question but keeping a steady performance in the other questions, GPT4All Falcon (0.122) and Mistral Instruct (0.12). The least was MPT Chat (0.059).
- **RougeL:** The top three models are Open Orca (0.2), with a similar performance that it has in Rouge2, Llama3 (0.188), and Mistral Instruct (0.184). The last place is for MPT Chat (0.129).

An overall analysis allows us to conclude that the best model performed in this metric is Mistral Open Orca, closely followed by Llama3 and Mistral Instruct. Even though the overall results obtained a low coefficient, especially in the bigrams, this metric evaluates the ability to summarise from a GenAI model and compare it with a human-made summary; being submitted to answer a questionnaire may downgrade the ability of the models to obtain better results.

## 6.2 Pearson Rank's Correlation

This metric presents much better results overall, demonstrating the use of similar words from the reference documents with the answers; this is an expected result due to the

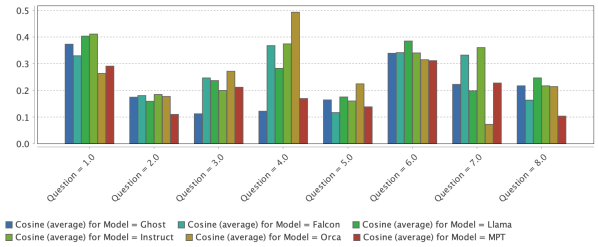


Figure 4: Cosine similarity by question and model

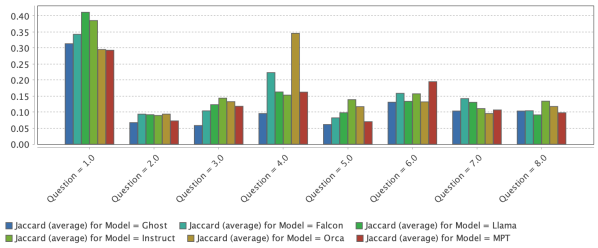


Figure 5: Jaccard similarity by question and model

temperature set in the implementation process as is presented in the figures 3 **?? ??**, there are very levelled scores in each question; after analysis and review, the performances can be evaluated in the following way: The best performance is for GPT4All Falcon (0.83 average), leading the performance in 6 of eight questions, followed by Open Orca (0.8) and Mistral Instruct (0.79). The worst performance was MPT Chat, which had an average of 0.75. In this measurement, it is possible to see consistent results in each model in each question, but no great differences in every question are obtained like the other metrics. This pattern may be due to the training in the documents that allow the models to use a defined vocabulary structure.

## 6.3 Cosine Similarity

This measurement shows very different results than the 6.2, more similar to the ones presented in 6.1 in the way that the scores are distributed by each question and by the models, at least with the Rouge1 results, but slightly lower in the overall results.
The analysis of the model's performance, shown in 4, could be classified as follows: the best model overall is Mistral Instruct, which was the best in three of the questions, with an average of 0.28. It was followed by Llama3 (0.26) and Falcon (0.259). The lowest is MPT Chat, which had 0.19 results.

## 6.4 Jaccard Similarity

Jaccard similarity presents meagre results, similar in distribution to Rouge2 but slightly higher. The best results in this measurement were the first question, the most straightforward one, and the first in the reference documents with which the models were trained. In this comparison, as it is possible to observe in 5, Mistral Instruct performed the best, even though it does not have a higher average; it led the coefficients in 3 of the questions (0.164), followed by Open Orca (0.166) and Falcon (0.156). Ghost7B's average was the worst, at 0.116.

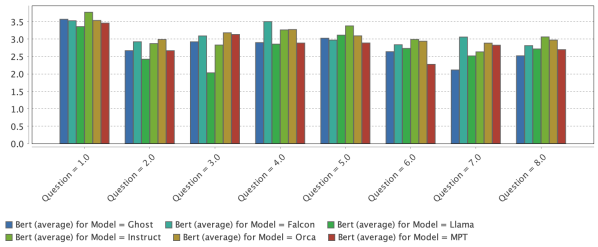


Figure 6: BERT results by question and model

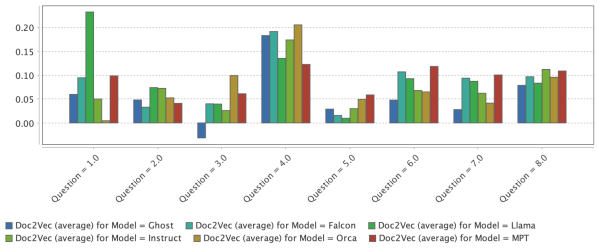


Figure 7: Doc2Vec results by question and model

## 6.5 BERT

The BERT pre-trained model to find similarities, one of the most used metrics to analyse these text experiments, presents us with a very different scale of results but a distribution similar to 6.4, besides particular situations.
Applying this fine-tuned model to the candidates and the reference 6 gives us the following ranking: Mistral Instruct again gets the first place, even though its average, 3.098, is the second highest by minimal difference but leads the results in four of the eight questions. It is followed by Open Orca (3.107) and Falcon (3.088). Call the researcher's attention that the worst performance was Llama3, with an average result of 2.717 overall.

## 6.6 Doc2Vec

The results obtained with this method present a very different situation than the previously observed with the other metrics: very low coefficients with a generally poor performance except in the fourth question with a general improvement as observed in 7. The results show that MPT Chat obtained the best performance, leading three questions and an average of 0.088. It was followed by Llama3, which had a better overall average of 0.094 due to a significant difference in the first question and Falcon (0.084). The last model was Ghost, with an average of 0.055.

## 6.7 SBERT

The results of this model correlate with the ones obtained with 6.5 and 6.4. With an overall excellent and consistent performance in all the questions, especially in the first one.
The evaluation 8 showed that the better model was GPT4All Falcon, which led to six questions with an average of 0.83, followed by Open Orca (0.808) and Mistral Instruct (0.798). The last model's performance was from MPT chat (0.757)

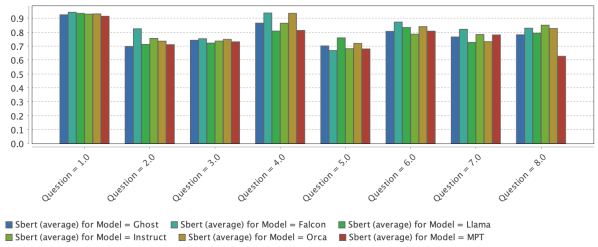


Figure 8: Sbert results by question and model

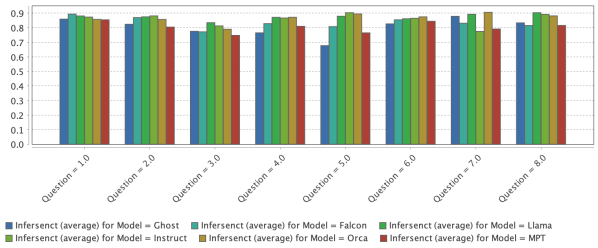


Figure 9: Infersent results by question and model

## 6.8 Infersent

The last experiment of text similarity performed was the Infersent model, showing high performance overall, already seen in 6.4, 6.5 and 6.7. The results obtained in these four of eight measurements can infer an excellent overall performance by the fine-tuned models tested.

Here, as shown in 9, the model Open Orca highlights better performance in three questions with an average of 0.866; the best average was obtained by Llama3 (0.874) but only better in two questions. The last position on the podium is for Mistral Instruct (0.858). The worst-performing model was MPT Chat (0.803).

## 6.9 Final Results

Retrieving the overall performance 10 and taking care of all the metrics outcomes, the resulting ranking conforms as it shows the table 1 Given these results, the final recommendation for implementing a large language model solution in a higher education institution that allows a prework module on Academic Integrity for students and uses the available information and policies of the particular institution more accurately is **Mistral Open Orca**.

**Experiments Results**

| 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 |
| Infersenct | | 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 |

Figure 10: Summary of results

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

Table 1: Final Evaluation of the Model's Performance

# 7 Conclusion and Future Work

## 7.1 Conclusion and Discussion

Artificial intelligence has undergone rapid development, characterized by iterative improvements and breakthroughs. Generative AI, particularly Large Language Models (LLMs), exemplifies this pattern. Initially confined to research and development within major tech companies, the landscape dramatically shifted with the public release of OpenAI's ChatGPT. This user-friendly Chatbot showcased the potential of LLMs to a global audience.

LLMs vary significantly based on the underlying architecture, training data, and intended applications. This diversity presents a complex choice for organizations seeking to implement AI solutions. While proprietary models offered by tech giants like OpenAI, Google, and Microsoft have gained prominence, they often require API access or paid subscriptions for advanced features.

Historically, the open-source ethos has thrived in computer science. This collaborative approach has spurred innovation in AI, leading to the development of publicly accessible LLMs. Models like Mistral, Llama, and Claude exemplify this trend, providing researchers and developers with alternatives to proprietary options.

Model repositories have emerged to streamline access and experimentation. Hugging Face, TensorFlow, and PyTorch offer a centralized hub for discovering, downloading, and deploying LLMs. This democratization of AI tools empowers a broader community of researchers and developers to contribute to the field.

As we stated previously in 3, the educational domain is always looking for solutions that are attained to its particular conditions due to the importance of its functions and the confidential information related to the students, and the open source models that are available thanks to the effort of the tech community can help the HEIs in the road to integrate these solutions to the student's learning process.

This report has reviewed a series of literature that investigates and analyses the impact and the integration of Generative Artificial Intelligence in the educational domain and how this can solve the Academic Integrity crisis that, in part, this technology has created. Preliminary, the observer of this research could prejudice that from the selected LLMs to test, the most recognisable, the biggest and with more than a billion more parameters, from one of the most significant technological enterprises would be the one with the best performance. However, the results showed that a better-tuned and best-trained solution has better results than a bigger model, invalidating the idea that "the more data, the better" the quality of the training and the better the data quality implies a better performance in the end. However, the research does not state this by utilising one metric; the utilisation of a series of measurement tools to evaluate the performance of GenAI in this particular use case seems to amplify the ability of the researchers and analysts to

understand in a better way the performance of a GenAI model, and how much capability has to be trained in a particular topic and be as precise as the human capability.
Another significant conclusion of this research is the necessity to alter how students engage with LLMs. This change can significantly impact their understanding of and adherence to academic integrity policies. We propose an implementation plan for high educational institutions to leverage the findings of this research and promote a culture of academic integrity. Upon achieving the objectives outlined in this research, we will develop a proposal framework. This framework will guide the implementation of the proposed solution in the educational domain, ensuring a systematic and effective approach.

## 7.2 Future Work

The models have been evolving and improving; the models used and tested in this research will soon be obsolete, updated or replaced by more powerful ones. The hardware limitations presented by a single researcher and his access to a series of models that can be run on a laptop can be overpass by a better infrastructure in a bigger scale of investigation of these topics, that scaling will allow trying different hyperparameters settings and test without the time-consuming use of limited hardware.
The metrics to evaluate the text generation will be improved, and new ones will appear; the following studies can be performed with another tool or updated versions of the ones used in this report. Different, not only more prominent but also more efficient, models can be tested, and the various metrics presented in this report can be used to generate analysis and data-driven recommendations for the best tool to be implemented, especially in the academic domain.

# References


Alexander, K., Savvidou, C. and Alexander, C. (2023). Who wrote this essay? detecting ai-generated writing in second language education in higher education., *Teaching English with Technology JCI= 0.71 Cited=12* **23**(2): 25–43.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S. et al. (2024). Scaling instruction-finetuned language models, *Journal of Machine Learning Research* **25**(70): 1–53.

Crawford, J., Cowling, M. and Allen, K.-A. (2023). Leadership is needed for ethical chatgpt: Character, assessment, and learning using artificial intelligence (ai), *Journal of University Teaching & Learning Practice JCI = 0.87 Cited = 197* **20**(3): 02.

Darwish, S. M., Mhaimeed, I. A. and Elzoghabi, A. A. (2023). A quantum genetic algorithm for building a semantic textual similarity estimation framework for plagiarism detection applications, *Entropy* **25**(9): 1271.

Denny, P., Leinonen, J., Prather, J., Luxton-Reilly, A., Amarouche, T., Becker, B. A. and Reeves, B. N. (2024). Prompt problems: A new programming exercise for the generative ai era, *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, pp. 296–302.

Denny, P., Prather, J., Becker, B. A., Finnie-Ansley, J., Hellas, A., Leinonen, J., Luxton-Reilly, A., Reeves, B. N., Santos, E. A. and Sarsa, S. (2024). Computing education in the era of generative ai, *Communications of the ACM* **67**(2): 56–67.

Denny, P., Smith IV, D. H., Fowler, M., Prather, J., Becker, B. A. and Leinonen, J. (2024). Explaining code with a purpose: An integrated approach for developing code comprehension and prompting skills, *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, pp. 283–289.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .

Eaton, S. E. (2023). Postplagiarism: transdisciplinary ethics and integrity in the age of artificial intelligence and neurotechnology, *International Journal for Educational Integrity JCI= 2.21 Cited= 12* **19**(1): 23.

Farrelly, T. and Baker, N. (2023). Generative artificial intelligence: Implications and considerations for higher education practice, *Education Sciences JCI = 1.46 Cited = 19* **13**(11): 1109.

Gallent Torres, C., Zapata-González, A. and Ortego-Hernando, J. L. (2023). 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 JCI = 0.62* .

GeeksforGeeks (2024). Different techniques for sentence semantic similarity in nlp.
**URL:** *https://www.geeksforgeeks.org/different-techniques-for-sentence-semantic-similarity-in-nlp/*

Google (2024). Python rouge implementation.
**URL:** *https://github.com/google-research/google-research/tree/master/rouge*

Gupta, T. (2023). Research on the application of artificial intelligence in the education and teaching system, *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, IEEE, pp. 1168–1173.

HugginFace (2021). Sentence similarity.
**URL:** *https://huggingface.co/tasks/sentence-similarity*

Khan, U. F. and Gonzalez, C. (2023). Real world legal document summarization using llms and its efficiencies, National College of Ireland.

Kızılırmak, E. (2023). Text summarization: How to calculate rouge score.
**URL:** *https://medium.com/@eren9677/text-summarization-387836c9e178*

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries, *Text summarization branches out*, pp. 74–81.

Madiraju, P. (2022). Rouge your nlp results!
**URL:** *https://medium.com/@priyankads/rouge-your-nlp-results-b2feba61053a*

Makeleni, S., Mutongoza, B. H. and Linake, M. A. (2023). Language education and artificial intelligence: An exploration of challenges confronting academics in global south universities, *Journal of Culture and Values in Education JCI= 0.82 Cited=12* **6**(2): 158–171.

Maryamah, M., Irfani, M. M., Raharjo, E. B. T., Rahmi, N. A., Ghani, M. and Raharjana, I. K. (2024). Chatbots in academia: a retrieval-augmented generation approach for improved efficient information access, *2024 16th International Conference on Knowledge and Smart Technology (KST)*, IEEE, pp. 259–264.

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R. and Gerardou, F. S. (2023). Challenges and opportunities of generative ai for higher education as explained by chatgpt, *Education Sciences JCI = 1.46 Cited = 68* **13**(9): 856.

Moya, B. and Eaton, S. E. (2023). Examining recommendations for artificial intelligence use with integrity from a scholarship of teaching and learning lens, *RELIEVE-Revista Electrónica de Investigación y Evaluación Educativa JCI = 0.62* **29**(2).

NAIN (2021). Academic integrity: National principles and lexicon of common terms.
**URL:** *https://www.qqi.ie/sites/default/files/2021-11/academic-integrity-national-principles-and-lexicon-of-common-terms.pdf*

NewsCatcher (2022). Ultimate guide to text similarity with python.
**URL:** *https://www.newscatcherapi.com/blog/ultimate-guide-to-text-similarity-with-python*

Oravec, J. A. (2022). Ai, biometric analysis, and emerging cheating detection systems: The engineering of academic integrity?., *Education Policy Analysis Archives JCI= 0.31 Cited= 12* **30**(175): n175.

Patil, A., Han, K. and Jadon, A. (2024). A comparative analysis of text embedding models for bug report semantic similarity, *2024 11th International Conference on Signal Processing and Integrated Networks (SPIN)*, IEEE, pp. 262–267.

Perkins, M. (2023). Academic integrity considerations of ai large language models in the post-pandemic era: Chatgpt and beyond, *Journal of university teaching & learning practice JCI = 0.87 Cited = 250* **20**(2): 07.

Poulsen, S., Sarsa, S., Prather, J., Leinonen, J., Becker, B. A., Hellas, A., Denny, P. and Reeves, B. N. (2024). Solving proof block problems using large language models, *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, pp. 1063–1069.

Prather, J., Denny, P., Leinonen, J., Smith IV, D. H., Reeves, B. N., MacNeil, S., Becker, B. A., Luxton-Reilly, A., Amarouche, T. and Kimmel, B. (2024). Interactions with prompt problems: A new way to teach programming with large language models, *arXiv preprint arXiv:2401.10759* .

Prather, J., Reeves, B., Leinonen, J., MacNeil, S., Randrianasolo, A. S., Becker, B., Kimmel, B., Wright, J. and Briggs, B. (2024). The widening gap: The benefits and harms of generative ai for novice programmers, *arXiv preprint arXiv:2405.17739* .

PyPI (2019). semantic-text-similarity.
**URL:** *https://pypi.org/project/semantic-text-similarity/*

Quille, K., Gordon, D., Harte, M., Faherty, R., Hensman, S., Becker, B. A., Nolan, K., O'Leary, C., Hofmann, M., Alattyanyi, C. et al. (2024). Machine vs machine: Large language models (llms) in applied machine learning high-stakes open-book exams, *Revista de Educación a Distancia (RED)* **24**(78).

Singh, M. (2023). Maintaining the integrity of the south african university: the impact of chatgpt on plagiarism and scholarly writing, *South African Journal of Higher Education JCI= 0.29 Cited= 4* **37**(5): 203–220.

StackOverflow (2021a). How to compute the similarity between two text documents?
**URL:** *https://stackoverflow.com/questions/8897593/how-to-compute-the-similarity-between-two-text-documents*

StackOverFlow (2021b). Rouge score append a list.
**URL:** *https://stackoverflow.com/questions/67390427/rouge-score-append-a-list*

Zhao, Y., Xia, T., Jiang, Y. and Tian, Y. (2024). Enhancing inter-sentence attention for semantic textual similarity, *Information Processing & Management* **61**(1): 103535.

Zhelezniak, V., Savkov, A., Shen, A. and Hammerla, N. Y. (2019). Correlation coefficients and semantic textual similarity, *arXiv preprint arXiv:1905.07790* .