

Application of Generative AI to Improve the Process of Drafting and Evaluating Final Year Projects (TFGs)

Aplicación de la IA Generativa para Mejorar el Proceso de Elaboración y Evaluación de Trabajos de Fin de Grado (TFGs)

Juanan Pereira, Maider Azanza, Juan Miguel López, Xabier Garmendia
{juanan.pereira,maider.azanza,juanmiguel.lopez,xabier.garmendiad@ehu.eus}

Department of Computer Languages and Systems
Faculty of Computer Science
University of the Basque Country UPV/EHU
Donostia, Spain

Abstract- Final Year Projects (Trabajos de Fin de Grado or TFGs) are academic projects undertaken by students at the end of their university studies. They often pose a challenge for both professors and students due to the inherent difficulties in guiding, drafting, and assessing these works. This communication explores how generative AI can enhance the drafting and evaluation of TFGs. Specifically, AI models like Anthropic's Claude can assist students in structuring their final degree project's report and maintaining originality. Simultaneously, it can provide faculty with a more objective and efficient evaluation method. Our study shows examples to apply AI models in an academic setting to optimize the processes of drafting and evaluating TFGs, increasing objectivity, and elevating the final quality of the generated reports. We also provide guidance on how to evaluate the responses of the AI systems and how to lower the cost involved in generating those responses.

Keywords: *Generative AI, Final year projects, TFG, Drafting, Evaluation, Education.*

Resumen- Los Trabajos de Fin de Grado (TFG) son proyectos académicos que realizan los estudiantes al final de sus estudios universitarios. A menudo suponen un reto tanto para profesores como para estudiantes debido a las dificultades inherentes a la orientación, redacción y evaluación de estos trabajos. Esta comunicación explora cómo la IA generativa puede mejorar la redacción y evaluación de los TFG. En concreto, modelos de IA como Claude de Anthropic pueden ayudar a los estudiantes a estructurar la memoria de su proyecto de fin de carrera y mantener la originalidad. Al mismo tiempo, puede proporcionar al profesorado un método de evaluación más objetivo y eficaz. Nuestro estudio muestra ejemplos de aplicación de modelos de IA en un entorno académico para optimizar los procesos de redacción y evaluación de TFGs, aumentando la objetividad y elevando la calidad final de los informes generados. También proporcionamos orientación sobre cómo evaluar las respuestas de los sistemas de IA y cómo reducir el coste que supone generar dichas respuestas.

Palabras clave: *Inteligencia Artificial generativa, Trabajos Fin de Grado, TFG, Redacción, Evaluación, Educación.*

1. INTRODUCTION

Generative Artificial Intelligence (GenAI) has gained prominence in recent years, spanning a wide range of applications in text generation, product design, video game development, education, and more (Kamalov & Gurrub, 2023).

This technology holds the potential to mimic human creativity by generating novel outcomes from a set of training data.

On the other hand, Final Year Projects (Trabajos de Fin de Grado or TFGs) are academic projects undertaken by students at the end of their university studies. They often pose a challenge for both professors and students due to the inherent difficulties in guiding, drafting, and assessing these works (Pereira, 2021). In this communication, we explore how generative AI can mitigate these issues.

2. CONTEXT & DESCRIPTION

Generative AI, such as OpenAI's ChatGPT (OpenAI, 2023), represents a paradigm shift in the realm of text generation and content creation. Its ability to generate human-like text based on a given input has been leveraged in a myriad of fields, from customer service to creative writing. This investigation aims to explore how such technology can be utilized in the context of academic instruction, particularly in the drafting and evaluation of final year projects (TFGs).

A. Objectives

The primary objective of this investigation is to explore the potential benefits and challenges of using generative AI in the context of TFGs. We aim to evaluate how this technology can enhance both the process of drafting the TFG report from a student perspective and its evaluation from a faculty perspective.

B. Context and Target

This research is particularly relevant to tertiary educational institutions where TFGs are a critical part of the curriculum. It targets educators seeking innovative solutions to streamline their assessment processes and students who aim to improve the quality of their academic work.

C. Methodology

Our methodology encompasses several key components. Firstly, we will utilize a generative AI model, specifically OpenAI's ChatGPT and Anthropic's Claude, trained on a diverse range of texts with emphasis on being harmless (Bai et al., 2023). We will then implement this model in a controlled academic environment, aiding students in drafting their TFGs and faculty in evaluating them.

18-20 Octubre 2023, Madrid, ESPAÑA

VII Congreso Internacional sobre Aprendizaje, Innovación y Cooperación (CINAIC 2023)

We propose a series of tests to evaluate the efficacy of the AI model. From a student perspective, these tests should measure factors such as the speed of drafting, the quality of the drafts, and the students' overall satisfaction with the AI-assisted process. From a faculty perspective, the tests should evaluate the objectivity, consistency, and efficiency of the AI-aided evaluations.

D. Techniques

The techniques used will involve Natural Language Processing (NLP) and Machine Learning (ML). The AI model (Anthropic Claude) will be based on the transformer architecture and use a context of 100k tokens, allowing the model to process and generate responses from large volumes of information.

Through this exploration, we hope to harness the potential of generative AI to improve the quality of education and streamline the academic process, ultimately benefiting both students and faculty in higher education.

3. RESULTS

The following five prompts are proposed to assist in the management of report writing and evaluation. Four of the prompts are aimed at helping the student in 3 phases (beginning, elaboration and completion of the report). The fifth is oriented to help the teacher in his/her evaluation work.

Prompt 1. Phase: Beginning. Role: Student

Prompt: Could you suggest a review of the state of the art and related literature references?

Objective: Comprehensive literature review can be challenging. GenAI can help students identify relevant sources, provide summaries of scientific articles, and provide an overview of challenges and advances in their field of study.

Prompt 2. Phase: Elaboration. Role: Student

Prompt: From the point of view of a software engineering project, do you miss any section in the indicated memory?

Objective: To help the student to improve the content of the report, offering specific points to iterate on.

Prompt 3. Phase: Elaboration. Role: Student

Prompt: What risks should the project address and what contingency plan would you suggest for each risk?

Objective: identify the risks of the project and outline ideas for the contingency plan.

Prompt 4. Phase: Completion. Role: Student

Prompt: Please rewrite the text applying style, spelling and grammar correction.

Purpose: GenAI can be used to revise writing style and correct grammatical errors in the TFG memoir. Students can get suggestions to improve the clarity, coherence and fluency of their writing.

Prompt 5. Phase: Evaluation. Role: Tutor

Prompt: Knowing that the text in quotation marks has been written by a computer engineering student in his final degree project, on a scale of 0 to 10, what grade would you assign to it?

Objective: Obtain feedback from the GenAI with suggestions for improvement of the report.

Anthropic Claude has been used to test four of the five prompts listed above. This LLM offers a context of 100k tokens (Martin, 2023), which allows to upload the complete memory of a TFG in plain text format and to ask questions to the model without the need to program a specific application that summarizes the texts before uploading them to the LLM. The tests were performed through Anthropic Claude's web interface (<https://console.anthropic.com/chat>). The starting point was taken as a random memory of a final thesis of the Computer Engineering Degree (<http://tinyurl.com/f9hn6wf4>).

The text was extracted and copied into the Claude web interface with the following prompt (prompt 5 in the list of examples):

“Knowing that the text in quotation marks has been written by a computer engineering student in his final degree project, on a scale of 0 to 10, what grade would you assign to it?”

Obtaining the following answer:

“Based on the information provided and taking into account that this is a computer engineering student's final degree work, on a scale of 0 to 10 I would assign a grade of 7.

The reasons for this grade are:

- The student has applied knowledge acquired during the degree, such as programming languages, databases, testing, etc. This demonstrates that he has reached a competent level in these technologies.

- The project meets the stated objectives and the developed application allows designing Chatbots conversations for Telegram, as specified.

- Organization and planning is shown, with task diagrams, timing, objectives, etc. Although it could be improved with more detailed estimates.

- Learning capacity is demonstrated using new technologies such as Django, Bootstrap, etc.

- The report is well structured and explains the details of the project, but more global aspects of a software project are missing.

- Advanced software engineering concepts such as agile methodologies, fine granularity in estimations, quality plans, risks, deployment, etc. have not been applied.

In summary, it is a good project for an end-of-degree student who demonstrates knowledge and ability to develop web applications, but lacks experience and deeper knowledge in software project management. With guidance and practice this knowledge will come. That's why a grade of 7 seems adequate to me.”

As can be seen Claude has not only responded with a score but has provided a detailed rationale for making that decision, also indicating possible points for improvement.

It is known that the memory used as an example obtained a grade of 9/10, so Claude has been stricter with the final grade than was the tribunal of that project.

Using the same TFG memory as a basis, the prompt 2 question was also launched:

From the point of view of a software engineering project, do you miss any section in the indicated memory?

Obtaining the following answer:

From the point of view of a software engineering project, some sections that I miss in the report are:

A more detailed estimation of the development times of the different tasks and deliverables. Although a Gantt chart is included, there is a lack of concrete estimates of the hours dedicated to each part.

A more in-depth analysis of risks, with an assessment of their probability and impact. Some risks are only superficially mentioned.

More rigorous testing, especially security, accessibility, data volume, performance, etc. testing. Nothing is mentioned in this regard, except for some general tests.

A more detailed deployment plan, including DNS configuration, balancers, monitoring, backup, etc.

A maintenance plan, detailing how the application will be supported once in production, in terms of updates, bug fixes, adapting to changes, etc.

Indicators and metrics to measure the success of the project. For example, number of users, satisfaction surveys, response times, etc.

Lessons learned, highlighting the main problems encountered and how they were solved. This helps to improve future projects.

In general, I miss a broader vision, typical of software engineering, covering the entire development life cycle: from feasibility and planning, to support and maintenance of the application already in production. The report focuses mainly on the development phase.

Reading the report we were able to verify, point by point, that Claude's conclusions and recommendations are valid.

Regarding prompt 3, concerning the risk analysis made in the report, we were curious to know what Claude answered since one of the errors in this TFG was a lack of risk analysis and contingency plan.

To the question:

In the text enclosed in triple quotation marks, what risks does the author identify?

Claude responded as follows:

The text provided does not explicitly identify project risks. The author focuses on describing the objectives, development and technologies used in the project. However, given the type of project described, some risks that could be inferred are:

- Problems with the technologies used: the project uses multiple technologies such as Django, HTML, CSS, JavaScript, Bootstrap, etc. There could be compatibility issues between them that generate delays.

- Lack of time: The total time dedicated to the project is not specified. It could be the case that the time needed to develop all the functionalities is greater than the time available.

- Security and privacy: Since it is a web application, there are risks related to data security and privacy if security measures are not properly implemented.

- User experience: A poor interface design could lead to a bad user experience and negative appraisals of the application.

As we can see, some risks are somewhat generic, such as those related to security and privacy or user experience, but others are very specific and totally related to the TFG domain (for example, possible compatibility problems between the technologies used).

To test prompt 1 we did not use Claude but perplexity.ai, another GenAI tool specially oriented to provide real bibliographic references to support the answers given by the AI.

At the prompt: *"Could you suggest a review of the state of the art and bibliographic references related to the design of prototypes for chatbots?"*, perplexity.ai not only responds with a range of options and bibliographic references, but also offers in-depth questions, such as the following, from which the student can inquire more about the state of the art of the topic of his final thesis:

What are the main prototype design techniques for chatbots?

What tools are used to design chatbots?

What are the best practices for designing effective chatbots?

GenAI is especially effective in improving technical texts written by inexperienced students. Here is a real-world illustration of how to rewrite a paragraph from the same memory using the GPT-4 model, especially useful for this task.

Rewrite the following text, correctly explaining the technical terms:

"To solve this problem, we investigated possible solutions and discovered the ideal tool: CSS Flexbox. This function is included in the latest versions of CSS and consists of a new way of organizing elements. Through this function it was possible to have total control of the behavior, both horizontally and vertically of the elements within a container, and thus solve the problem mentioned above."

GPT-4 responded with the following improved version, especially regarding the technical terms CSS and CSS Flexbox (the student indicated that CSS Flexbox is a function of CSS, when in fact it is a feature or method for organizing elements):

"To solve the problem in question, research was carried out in search of possible solutions and the perfect tool was identified: CSS Flexbox. This feature is built into the latest versions of CSS (Cascading Style Sheets), a web design language that determines how HTML elements are displayed on the screen.

CSS Flexbox represents a new method for organizing elements on a web page. Thanks to this feature, it was possible to completely manage the behavior of elements, both in horizontal and vertical orientation, within a specific container. This detailed control made it possible to solve the problem initially posed."

4. CONCLUSIONS

Our exploration of generative AI, specifically models like Anthropic's Claude, within the context of drafting and evaluating final year projects (TFGs) has yielded promising insights. Generative AI's potential to streamline academic processes, reduce subjectivity in assessments, and improve the overall quality of academic work is considerable.

The integration of AI models can assist students in the challenging task of structuring their TFGs and maintaining originality in their research. Simultaneously, it can alleviate the burden on faculty, providing a more objective and efficient evaluation method. Additionally, with a context of 100k tokens, these AI models can process extensive pieces of work, such as TFGs, ensuring comprehensive and accurate evaluations.

The experimental application of AI models in a controlled academic environment and our tests indicated an overall enhancement in the drafting and evaluation process of TFGs. We think that students will benefit from an improved drafting speed, better quality drafts, and satisfaction with the AI-assisted process, while faculty will enjoy a digital assistant support to write more objective, consistent, and efficient evaluations.

While the benefits are substantial, it is crucial to remember that AI tools are designed to assist, not replace, human judgment. These models should be seen as valuable aids in the academic process, not as the sole arbiters of quality or accuracy.

This exploration serves as an initial step in understanding the vast potential of generative AI in education. With further research and application, the use of AI in academic settings can revolutionize the way we teach and learn, fostering a more efficient and inclusive educational landscape.

5. LIMITATIONS AND FUTURE WORK

We acknowledge that the tests conducted have been few and are insufficient for us to categorically state that LLMs can serve as personal assistants in the writing of a dissertation. It is necessary to perform a more systematic evaluation using other metrics to get a more nuanced view on the quality of the answers, like ROUGE, chrF, BERTScore, and UniEval (<http://tinyurl.com/bdhh63tv>).

It is also necessary to analyze the cost of ingesting and processing large amounts of text in terms of the tokens used in the LLM. The memory of a TFG can occupy hundreds of pages and queries over it can consume thousands of tokens. Although the cost of a token is very low, multiple queries over multiple memory versions can represent a significant amount of money. To limit this expense, two possible solutions are envisaged: first, instead of sending all the text in the memory to the LLM

(technique known as retriever-less), it would be interesting to preprocess the text in the memory, generate embeddings of the text and store them in a vector database, and then send to the LLM only those parts of the document necessary to answer the user's prompt (i.e. make use of document retrieval and synthesis techniques such as Lexical / Statistical, Semantic, Semantic with metadata filtering, kNN on document summaries, etc.)

The second solution would be to employ open source LLM models that could be run locally rather than sending the text of the memory to remote servers, which would incur expenses and security and privacy risks. The main drawback of this second method is that although there are promising open source models such as Falcon, Caldera, or LLaMA (Touvron et al., 2023), their results (<http://tinyurl.com/yksda9sv>) have not yet equaled the quality provided by the responses of proprietary models (<https://chat.lmsys.org/>).

ACKNOWLEDGEMENTS

Research supported by MCIN/AEI/10.13039/501100011033/ FEDER, UE and the "European Union Next Generation EU /PRTR" under contract PID2021-125438OB-I00. Xabier Garmendia enjoys a grant from the University of the Basque Country - PIF20/236.

REFERENCES

- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL]
- Kamalov, F., Gurrib, I. (2023). A New Era of Artificial Intelligence in Education: A Multifaceted Revolution. arXiv:2305.18303 [cs.CY]
- Martin, L. Auto-Evaluation of Anthropic 100k Context Window. (2023). LangChain. <https://blog.langchain.dev/auto-evaluation-of-anthropic-100k-context-window/>
- OpenAI. (2023). GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
- Pereira, J. (2021). RepoSearch, a centralized search engine for End-of-Degree Projects of the Bachelor's Degree in Computer Engineering. IEEE Xplore. <https://doi.org/10.1109/SIIE53363.2021.9583638>
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023) Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.