



Melany Macias

Finetuning and Improving Prediction Results of LLMs Using Synthetic Data

Metropolia University of Applied Sciences

Bachelor of Engineering

Information Technology

Bachelor's Thesis

22 May 2024

Abstract

| | |
|---------------------|--|
| Author: | Melany Macias |
| Title: | Finetuning and Improving Prediction Results of LLMs Using Synthetic Data |
| Number of Pages: | 42 pages |
| Date: | 22 May 2024 |
| Degree: | Bachelor of Engineering |
| Degree Programme: | Information Technology |
| Professional Major: | Smart Systems |
| Supervisors: | Sakari Lukkarinen, Senior Lecturer Mika Hämäläinen, Project Manager |

The Metropolia University of Applied Sciences initiated a project intended to enhance its Moodle platform with an AI-powered plugin, designed to improve the educational process for educators. Central to this initiative is the development of a chatbot designed to engage with users in conversations about teaching material and sustainability, specifically about Sustainable Development Goals.

This thesis evaluates several open-source large language models — Llama 3 (8B), Gemma (2B and 7B), and Phi 2 (2.7B) — implementing a methodology that includes training dataset generation, automated evaluation, comparative analysis, and error analysis. Training data was created by collecting sustainability-related documents and using Mistral (7B) to convert plain text into Q&A pairs. Then, these base models were finetuned with the generated sustainability data as well as general datasets designed for dialogue and summarizations.

The model's performances were measured using the BLEU, ROUGE, AND METEOR scores to assess the quality of text generation, while comparative analysis focused on evaluating the model efficiency relative to the resources consumed and the parameters size, and an error analysis was done to classify the error types. The study shows that finetuning always improved the performance; the best performing model being finetuned was Gemma (7B) with a METEOR score of 0.25, and the maximum time taken during finetuning was 8 hours and 30 minutes.

Keywords: AI, LLMs, SDGs, Sustainability, Python, Chatbot, Moodle

Table of Contents

| | | |
|-------|--|----|
| 1 | Introduction | 1 |
| 2 | Theoretical Foundations | 3 |
| 2.1 | Fundamentals of AI and Deep Learning | 3 |
| 2.2 | Large Language Models | 4 |
| 2.3 | Open-Source Software | 7 |
| 2.3.1 | Moodle's Infrastructure | 8 |
| 2.3.2 | Hugging Face Platform | 9 |
| 2.4 | Sustainable Development Goals | 10 |
| 3 | AI Plugin Overview | 12 |
| 3.1 | Current Status | 12 |
| 3.2 | Technical Architecture | 13 |
| 3.3 | Justification with Educational Goals and Contributions | 15 |
| 4 | Methodological Approach | 17 |
| 4.1 | Criteria and Selection of the LLM models | 17 |
| 4.2 | Evaluation Methodology | 19 |
| 4.3 | Datasets Generation | 20 |
| 5 | Finetuning | 25 |
| 6 | Results and Evaluation | 28 |
| 6.1 | Automation Evaluation | 28 |
| 6.2 | Comparative Analysis | 31 |
| 6.3 | Error Analysis | 36 |
| 7 | Discussions | 40 |
| 8 | Conclusion | 42 |
| 9 | References | 43 |

List of Abbreviations

| | |
|--------|---|
| AI | Artificial Intelligence |
| API | Application Programming Interface |
| BLEU | Bilingual Evaluation Understudy |
| CPU | Central Processing Unit |
| CSS | Cascading Style Sheets |
| CUDA | Compute Unified Device Architecture |
| DL | Deep Learning |
| GPT | Generative Pre-trained Transformer |
| GPU | Graphics Processing Unit |
| HTML | Hyper Text Markup Language |
| IIS | Internet Information Services |
| LAMP | Linux, Apache, MySQL, PHP |
| LLM | Large Language Model |
| LLaMA | Large Language Model Meta AI |
| LoRA | Low-Rank Adaptation of Large Language Models |
| METEOR | Metric for Evaluation of Translation with Explicit Ordering |
| ML | Machine Learning |
| NER | Named Entity Recognition |
| NLP | Natural Language Processing |
| OSS | Open-Source Software |
| PDF | Portable Document Format |
| PEFT | Parameter-Efficient Fine-Tuning |
| PHP | Hypertext Preprocessor |
| QLoRA | Quantized Low-Rank Adaptation |

| | |
|-------|---|
| Q&A | Question and Answer |
| RNN | Recurrent Neural Networks |
| ROUGE | Recall-Oriented Understudy for Gisting Evaluation |
| SDG | Sustainable Development Goals |
| SQL | Structured Query Language |
| UAS | University of Applied Sciences |
| VLMS | Virtual Learning Management Systems |

1 Introduction

The rapid advancement of artificial intelligence (AI) is drastically influencing global industries, drawing parallels with transformative technologies such as the internet and electricity in their historical impact (Ng, 2017). This technological revolution has significantly transformed society across various domains, and particularly, education (Zawacki-Richter et al., 2019), as a vital mechanism for societal progress and individual growth, has notably benefited from the advancements in AI (León & Viña, 2017). Artificial intelligence is not just a theoretical advancement but a practical tool that is reshaping education through the automation of administrative tasks, enabling personalized learning experiences, identifying areas where students require additional support, and facilitating student guidance outside the classroom (Garrido, 2012). These capabilities underscore the potential of AI to not only complement traditional teaching methods but to revolutionize how educational content is delivered, and consequently, it has led to the personalization of learning experiences, which research indicates can boost motivation and reduce dropout rates among students (Sáiz et al., 2020).

After delineating the general impact of AI in education, it becomes pertinent to examine how these advancements are being applied within educational platforms. One of the most relevant educational digital systems is Moodle, an established and widely utilized Learning Management System (LMS), which serves as a cornerstone in the digital education landscape since its launch in 2002. Moodle presents a steady user base growth, specifically as of February 2024, there are 408,648,444 users across 242 countries (Moodle, 2024a). Additionally, Moodle is known for its compatibility with third-party applications that enables seamless integration of diverse content types, therefore is an ideal platform for innovative AI integrations. Consequently, the largest applied sciences institution in Finland, Metropolia University of Applied Sciences, uses Moodle to link instructors and students in a dynamic learning environment.

This thesis oversees the development and integration of an AI-powered plugin into Metropolia's Moodle platform. Conceived as part of a broader Metropolia UAS initiative, the plugin performs multiple functionalities: summarizing educational content, aligning teaching materials with Sustainable Development Goals (SDGs), suggesting assignments that correlate with the lectures content and a specific SDG, as well as a conversational chatbot that allows users to interact with the teaching material. These features aim to improve the platform's educational resources' usability and relevance, supporting both Metropolia's mission and the global movement to elevate the SDGs in higher education while deepening the meaning of the content through interactive interaction. Moreover, the plugin seeks to enrich educational engagement by enabling interactive dialogue between users and content.

Chatbots, through personalized and adaptive responses, can significantly enhance the learning experience by providing immediate feedback, clarifying complex concepts, and encouraging active learning. The potential for these technologies in educational settings is further demonstrated by the projection that the worldwide chatbot market would surpass \$994 million by 2024 (Dooley, 2019). This research focuses on the development of a conversational chatbot within the plugin, aiming to optimize interactions between users and educational materials. The process contains the finetuning of several Large Language Models (LLMs), assesses their performance and compares the outcomes during real-time interactions with lectures contents. Finetuning LLMs can be beneficial as it plays a crucial role in enhancing chatbots by enabling them to understand and generate human language more effectively, and these can be used to suit specific enterprise needs or tasks, offering a more engaging and customized experience. Consequently, this thesis not only illustrates the potential of LLMs to make educational tools more dynamic and interactive but also showcases their wide applicability in meeting diverse educational needs.

2 Theoretical Foundations

This section delineates the core principles and technologies pivotal to the AI plugin's development for Moodle. It elucidates the interplay between Moodle's open-source framework, the advancements in LLMs due to key platforms such as Hugging Face, and the integration of Sustainable Development Goals (SDGs) into educational content.

2.1 Fundamentals of AI and Deep Learning

Artificial Intelligence (AI) has become a reality and is now used in almost every field from education to healthcare to finance. AI strives to develop systems that would be able to perform tasks that would require intelligence if done by a human; this is achieved through learning algorithms that try to mimic the learning process. Emerging from the broader domain of AI, machine learning equips machines with human-like intelligence through experience rather than through direct programming (Das et al., 2015).

Nevertheless, because of how poorly conventional machine-learning methods handle raw natural data, an innovative approach to pattern detection labelled as representation learning was established (Bengio et al., 2013). This approach, through deep learning, empowers computational models built with several layers of processing to master data representations across various abstraction layers (LeCun et al., 2015). Using the concept of neurons in the human brain as a basis, these are computational models made of the algorithms known as artificial neural networks, though they operate on different principles than the biological structures. These models process data through layers of abstractions, allowing machines to make sense of vast amounts of information with unprecedented accuracy (Schmidhuber, 2015).

The cornerstone of deep learning's success lies in its neural networks' ability to automatically extract and learn features from data, which is achieved through training processes involving backpropagation (Rumelhart et al., 1986) and gradient descent (Lecun et al., 1998), where the model iteratively adjusts its

parameters to minimize the difference between its predictions and the actual outcomes. This capability not only enhances the model's performance over time but also adapts to new data, making it versatile and scalable across domains (Schmidhuber, 2015). As a result, deep learning has transformed areas such as natural language processing, speech recognition and computer vision. These advancements have allowed machines to undertake activities previously thought to be the sole domain of human intellect (Goodfellow et al., 2018).

In educational contexts, deep learning facilitates the development of systems that can, for instance, assess student essays or provide real-time language translation services, thereby enabling more inclusive and accessible learning environments. Deep learning's capacity for personalization and adaptability has profound implications for eLearning platforms, manifesting in intelligent tools that cater to the unique needs of each learner. Deep learning's contribution ranges from creating responsive eLearning environments, from chatbots that act as virtual assistants to performance indicators that adapt course content based on learner engagement (Muniasamy & Alasiry, 2020).

Complex neural networks can perform exceptionally well at activities that have historically been attributed to human intellect, such understanding and producing natural language (Arana et al., 2023), and an example of this capacity are LLMs, described on the following section, which process large amounts of textual data to extract context and generate human-like content. (Pallagani et al., 2024).

2.2 Large Language Models

LLMs are predicated on deep neural networks that understand, generate, and interpret human-like text, therefore these models can accurately predict the subsequent word in a phrase since they have been trained on a large amount of data. LLMs such as GPT-3 (Brown et al., 2020), GPT-4 (OpenAI et al., 2024), and LLaMA (Touvron et al., 2023) have shown outstanding performance across a variety of natural language processing (NLP) applications (Zhao et al., 2024). Referred to as “complex black-box systems” (Zhao et al., 2024), they rely on a sophisticated process involving unsupervised pre-training followed by finetuning,

which molds them for various applications, including conversational AI in educational platforms.

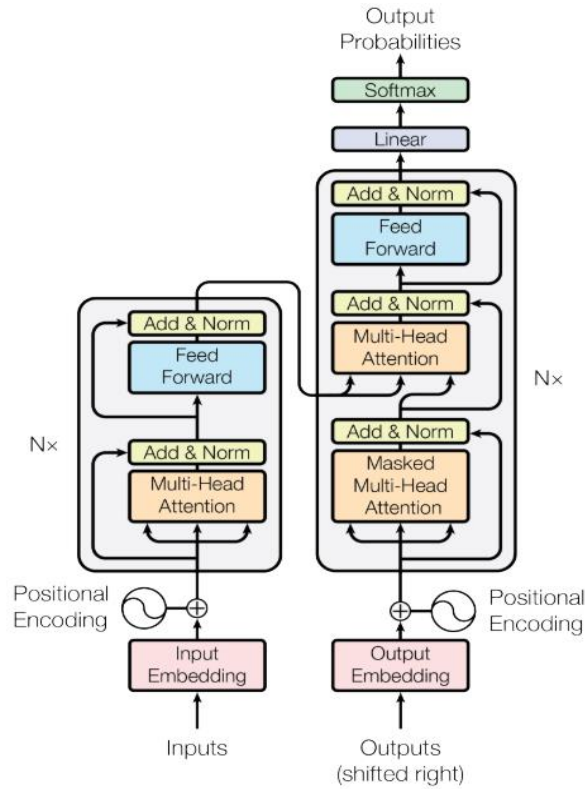


Figure 1: The Transformer - model architecture (Vaswani et al., 2017)

The Transformer architecture (Vaswani et al., 2017) shown in Figure 1, is the foundation upon which LLMs are built. The unique aspect of this architecture is its utilization of attention mechanisms that enable the model to assess the significance of various segments of the input data in parallel. This method brings efficient and effective processing of linguistic information, providing a basis for the pre-training of LLMs. During this phase, models absorb diverse linguistic patterns from extensive text datasets, establishing a broad linguistic competence.

Structurally, the Transformer is composed of two principal segments: the encoder and the decoder. The encoder's role is to convert input sequences into a series of higher-dimensional continuous representations. The decoder then utilizes these representations to reconstruct the target output sequence. A distinctive feature of both the encoder and the decoder is the integration of a self-attention mechanism that facilitates the evaluation of the sequence in its entirety, rather

than in isolation. Positional encodings are incorporated alongside the input to compensate for the model's non-recurrent nature, thereby ensuring the retention of sequential order information.

The complexity and the large number of parameters in LLMs significantly enhance their ability to learn and capture diverse linguistic features and dependencies from large-scale datasets. The increasing number of parameters in models like the ones from Llama 2 (Touvron et al., 2023), which series of models ranges between 7 to 70 billion parameters. These parameters enable models to maintain multiple layers of abstraction, allowing them to perform complex tasks such as translation, summarization, and conversation with high accuracy. However, the increase in parameters also lead to higher computational demands and memory consumption, making these models resource intensive (Kachris, 2024). Consequently, there is a need for a balance between the benefits of increased parameters and the practical challenges of deploying these computationally intensive models in real-world applications, therefore there are multiple studies that aim to compress models (Chavan et al., 2024).

Parameter-Efficient Fine-Tuning (PEFT) techniques such as Low Rank Adaptation (Hu et al., 2021) and its variant Quantized Low Rank Adaptation (Dettmers et al., 2023), which address the computational challenges posed by LLMs which usually worked by reducing the number of trainable parameters in a neural network. LoRA is an adapted based approach that works by breaking down the weight update matrix into smaller matrices which are used to update the model with backpropagation, and these new parameters are updated in the smaller matrices instead of directly on the model, therefore preserving the model's pre-trained capabilities. It also applies a model quantization technique which is the process of representing a number using fewer bits than the original number's data type, such as using 16-bit integers instead of 32-bit floating-point numbers.

On the other hand, QLoRA extends these benefits by incorporating double dequantization, 4-bit NormalFloat and paged optimizers, thus reducing memory requirements further and facilitating deployment on less capable hardware.

These resource saving techniques are highly beneficial during the finetuning process, which is helpful for customizing these pre-trained base LLMs to perform specific tasks. Finetuning involves supervised learning on task-focused datasets and reinforcement learning techniques to refine the models' outputs. This process is instrumental in concentrating the models' extensive linguistic knowledge on the precise requirements of targeted applications, such as conversational AI interfaces for educational platforms (Zhao et al., 2024).

Illustrative of the potential of finetuning is the innovative application in the field of modern poetry generation, where two pretrained neural models, RoBERTa (Liu et al., 2019) and GPT-2 (Radford et al., 2019), are finetuned for generating French poetry. The poetry produced was evaluated favorably by human judges, particularly in terms of understandability with its best score of 3.79 on a 5-point scale (Hämäläinen et al., 2022). Similarly, businesses frequently fine-tune LLMs to meet unique needs, leading to improvements in customer service.

Therefore, it is highly beneficial to personalize LLMs, and in the context of this thesis, customizing an LLM for the Moodle plugin symbolizes a step towards enriching the educational experience while also contributing to the academic community's understanding of LLMs' role in education.

2.3 Open-Source Software

Open-source software (OSS) represents a paradigm shift in how technology is developed, distributed, and maintained, offering significant benefits across various sectors, including education. Open source allows for the inspection, modification, and enhancement of source code, promoting a culture of open collaboration and decentralized production that drives technological advancements (GitHub, 2008; Red Hat, 2019). The Linux operating system exemplifies OSS's success, becoming the world's largest open-source project and showcasing the model's scalability and adaptability.

The value of OSS in education is multifaceted, encompassing cost savings, increased security, and enhanced flexibility. Consequently, educational

institutions often turn to OSS to avoid the licensing costs associated with proprietary software, allocating resources more efficiently towards teaching and research (Apereo Foundation Board, 2019). This economic advantage coupled with OSS's adaptability allows institutions to tailor software to their needs, fostering innovation within the educational landscape.

2.3.1 Moodle's Infrastructure

Moodle stands as a significant advancement in the educational sector, offering a dynamic platform for Virtual Learning Management Systems (VLMS) which is now the industry standard for many schools and universities worldwide (Campo et al., 2021). At its core, Moodle's architecture is designed for scalability and customization, which is demonstrated by foundational LAMP framework stack: Apache, Linux, PHP, and MySQL.

In Figure 2 below, Moodle's software architecture according to MoodleDocs (Moodle, 2020) is shown. PHP is an open-source scripting language extensively utilized for web development purposes (PHP, 2024) and serves as the programming language for Moodle, along with the integration of HTML, JavaScript, and CSS. Moodle also uses MySQL as the database management system, and the web server Apache hosts the Moodle application and handles requests from users' browsers.

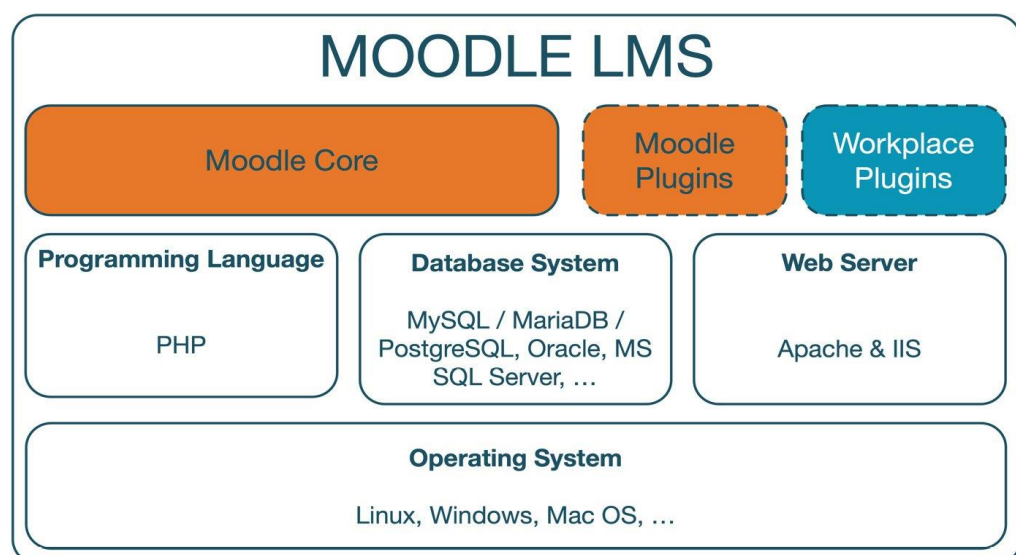


Figure 2: Moodle's overall architecture (Moodle, 2024a)

Additionally, Moodle Plugins are enhancements that enable educational institutions to customize and extend the core functions of Moodle VLMS. With a community-contributed library of over 2000 plugins, users can greatly enhance and personalize the learning experience (Moodle, 2024b).

Developed under an open-source philosophy, Moodle allows for the addition of new functionalities through the described plugin mechanism. This architecture not only facilitates the seamless integration of AI tools but also significantly enhances Moodle's pedagogical capabilities, therefore supporting the incorporation of innovative functionalities without disrupting the platform's core structure.

The 'Intelligent' plugin (Nalli et al., 2021) for the automatic formation of heterogeneous student groups, exemplifies Moodle's capacity for leveraging ML to enhance the educational experience. This plugin applies clustering algorithms to educational instructions' data stored in Moodle to form groups which maximize student diversity, consequently aiming to improve collaborative learning outcomes. Similarly, the 'eOrientation' plugin (Sáiz et al., 2020), which emphasizes Moodle's potential for prompt detection of potential drop-out students through machine learning approaches.

These examples illustrate Moodle's capacity for incorporating AI functionalities, which corresponds with the thesis's objective of finetuning a conversational model.

2.3.2 Hugging Face Platform

The emergence of the Transformers architecture (Vaswani et al., 2017) marked a significant advancement in the field of NLP. Transformers have largely superseded previous NLP architectures, particularly those based on recurrent neural networks (RNNs), in many tasks due to their efficiency and performance. However, it is important to note that RNN-based models still excel in certain specific areas. Furthermore, the emergence of model pre-training methods,

which allow training models on vast general corpora has become crucial and this shift in NLP showed the need for a platform that could make such complex models usable and feasible for everyday applications. Hugging Face arose to meet this demand, offering a central hub that offers access to these complex models for everyone (Wolf et al., 2020).

Huggingface.co (Hugging Face, 2024a) has become a leading platform within the ML community, recognized for its Transformers library. This open-source library is devoted to supporting Transformer-based architectures and facilitating the sharing of pretrained models. It is constructed around a unified API, featuring a collection of pretrained models created by and for the community (Wolf et al., 2020). It is maintained by Hugging Face's team of engineers and researchers, along with the support of a community comprising over 400 external contributors.

Within the Hugging Face platform, the integration of tokenizers, transformer models, and task-specific heads is crucial for the development of LLMs suitable for educational contexts. Tokenizers play a pivotal role in preparing the text data by converting the raw textual input into a format that transformer models can process. These models, based on the Transformer architecture, utilize self-attention mechanisms to generate contextually rich word embeddings. Finally, task-specific heads are tailored to NLP tasks, such as sentiment analysis or question-answering (Tunstall et al., 2022), which are integral to developing responsive and intelligent conversational agents in Moodle.

2.4 Sustainable Development Goals

The 2030 Agenda for Sustainable Development is a shared vision of humanity and a plan for action for people and the planet, for the present and for future generations. At the center of this agenda are the 17 SDGs which are a powerful appeal to action for all countries, both developed and developing, in a global partnership.

Embedded in these 17 goals is the recognition that overcoming poverty and other deprivations needs strategies that simultaneously tackle climate change,

safeguard our oceans and forests, improve health and education, and foster economic growth (United Nations, 2024).



Figure 3: Sustainable Development Goals (United Nations, 2024)

Within the context of achieving sustainable growth, SDGs, illustrated in Figure 3 above, serve as a crucial framework for organizations and educational institutions. For Metropolia UAS, the SDGs are more than a set of objectives; they are a directive that shapes the institution's commitments to integrating sustainability into its developmental strategies. The university's roadmap towards sustainability by 2030 emphasizes that acquiring sustainability know-how is imperative for the entire Metropolia community, ensuring that the principles underlying the SDGs are actively implemented in both professional and societal contexts in the hopes to create graduates that take sustainability into consideration during their professional lives. (Metropolia, 2023).

Given this, the use of technologies like LLMs presents a great opportunity to achieve these goals within the educational setting. The previous study showed that it is possible to fine-tune LLMs to predict the relevant SDGs from the description of the university courses, in specific the efficacy of various smaller language models in aligning academic curricula with SDGs (Kharlashkin et al., 2024). This intersection of sustainability and cutting-edge AI technology presents a major opportunity for development and research focus.

3 AI Plugin Overview

3.1 Current Status

The AI plugin is an additional customization layer that can be used to Moodle to facilitate teacher's work by analyzing their teaching content and allowing real-time interaction with a chatbot. By the time of writing the thesis, The AI plugin for Moodle is currently on the testing phase, with full deployment scheduled for the autumn term of 2024 at Metropolia UAS.

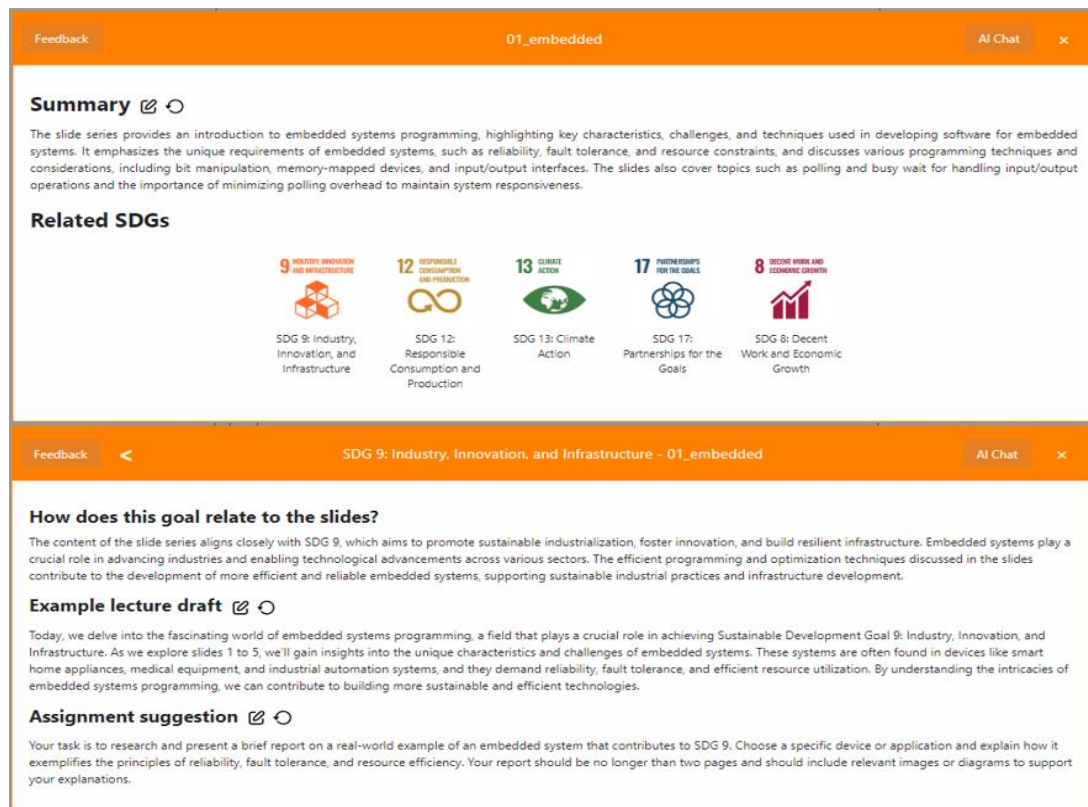


Figure 4: AI Plugin interface.

Figure 4 above shows two different windows, the upper one showing the user interface of the AI plugin which contains the summarization feature as well as the related goals available to be selected, and the bottom window containing more information about a specific goal that was selected from the first image. This interface aims to showcase the plugin's capacity to present educational materials in a manner that is both approachable and acknowledgeable to global sustainability targets.

Key features of the plugin include the following:

- **Summarization Feature:** Encapsulates the educational content uploaded to Moodle into a concise summary.
- **SDG Alignments:** Aligns teaching materials with relevant SDGs, making it easier for educators to integrate them into classroom discussions.
- **Chatbot:** Offers real-time, personalized interaction through chatbots based on global course content or specific document material.
- **Course Overview:** Contains a summary of the course, key skills, and keywords to improve accuracy in curriculum development.

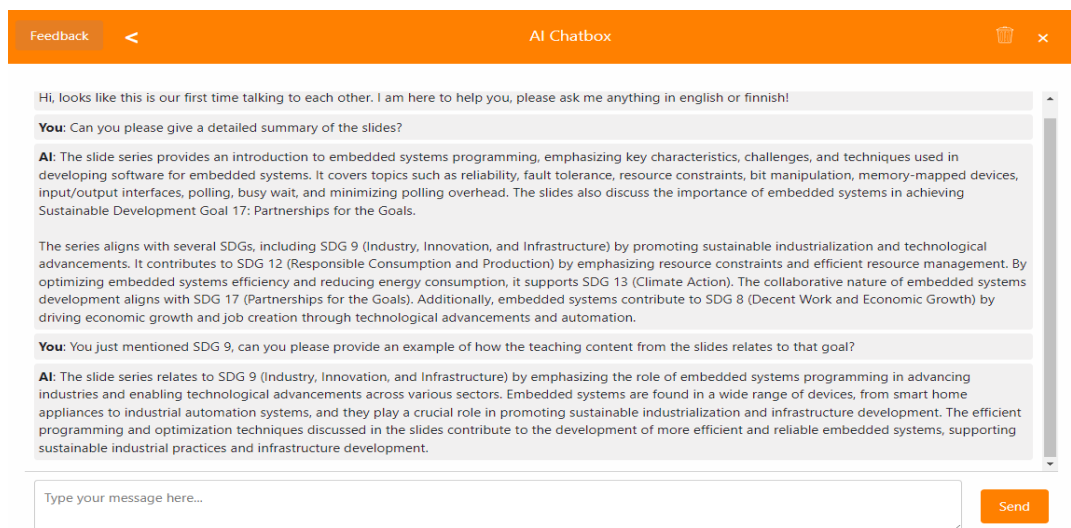


Figure 5: Example of a chatbot conversation for the LLM retrieval.

Figure 5 presented above illustrates a typical interaction between the LLM and the end user, showing a practice example of the chatbot feature.

Thus, these capabilities complement the sustainability goals of the educational institution and are a reaction to the need for improving the accessibility of teaching resources within the Moodle platform.

3.2 Technical Architecture

The AI plugin's application is designed to integrate into Moodle's infrastructure as an injected plugin. The stack includes the programming languages JavaScript,

PHP and Python, databases including MySQL, the vector database Chroma (Chroma, 2023), and Google Cloud Platform's Vertex AI (Google Cloud, 2024).

Figure 6 below depicts the current technical framework and the planned enhancement with an in-house LLM which will replace the Vertex AI highlighted in red. It aims to show the flow of information and the interaction between different system components.

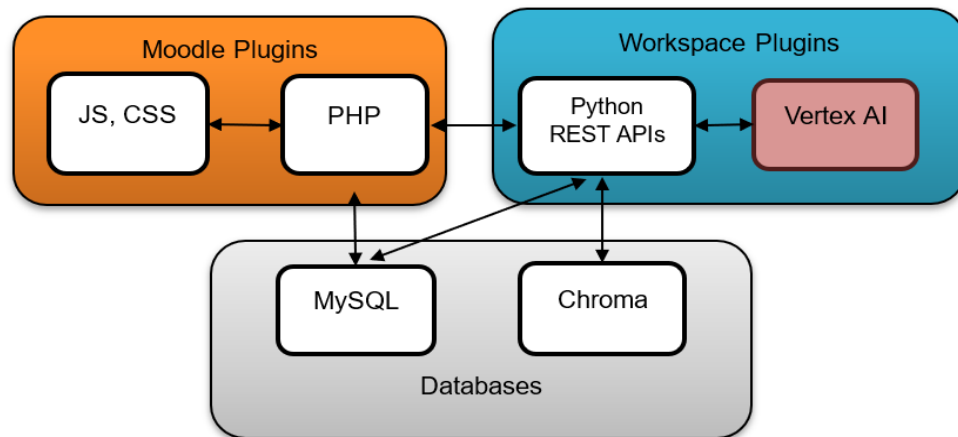


Figure 6: AI Plugin Software Design.

As shown in the picture, JavaScript and CSS are employed for frontend development, as these injected files are key in enabling user interactions with the plugin's features directly within the Moodle environment. The server-side PHP backend of Moodle serves as an intermediary, handling AJAX requests that are triggered by client-side user actions. This component manages the logic for permission and authentication, and it also uses data caching to maximize system efficiency by determining when requests to the Python server APIs are required.

The APIs on the Python side serve as a gateway to AI operations, database transactions, and Moodle's responses. It contains the AI functionality which lies in the interaction with Vertex AI and the use of LangChain library (LangChain, 2024), as well as the material's data cleaning procedure, the summarization and SDG suggestions process and the chatbot's conversational flow.

The plugin is designed to use two databases: MySQL, which is used for persistent data storage, and Chroma, an open-source database which makes simple the process of building LLM applications (Chroma, 2023). The capacity of the chatbot to extract contextually relevant data from the educational material's vector embeddings is performed by Chroma and LangChain.

3.3 Justification with Educational Goals and Contributions

The decision Metropolia is making to put the principles of the SDG into the syllabus is an important part of its vision for a sustainable society in 2030 (Metropolia, 2023). The plugin contributes to the integration of sustainability in the educational process through the provision of a plugin for summarizing the educational content, connecting the course material with related SDG goals, and chatbot interactions, which is aligned with Metropolia's commitment.

The goal of this thesis is to provide a transition for the integrated AI plugin's chatbot from an API based on external servers, in this case specifically, Vertex AI to an internally developed and trained LLM. This shift is motivated by several key considerations:

- **Privacy and Data Security:** Establishing an in-house LLM guarantees that all interactions maintained are constricted to the environment owned by Metropolia, which guarantees the security of academic data.
- **Cost Efficiency:** Via operating independently from the external APIs, the plugin becomes more affordable in the long run where the avoidance of high-volume API calls becomes expense associated to the user.
- **Customization and Relevance:** Many benefits come from a finetuned LLM. Firstly, it can be tailored specifically to the university curriculum, and making sure the language of the chatbot is relevant and supportive to our and our students' educational objectives.

The study tackles several critical research questions that guide the finetuning of these models to align with Metropolia's specific educational needs:

- What is the optimal model size for LLMs to efficiently support educational applications while maintaining computational sustainability?
- Can smaller, finetuned models deliver performance on par with or better than larger models within the context of educational chatbots?
- How significantly does finetuning on targeted educational content enhance LLM performance compared to generic models?

Consequently, this thesis will not only consider a technical need within the plugin's development but also ensure that it fits into the strategic framework of Metropolia UAS that calls for progressive and environmentally friendly learning. This research details the process of selecting, generating training data, finetuning and discussing the implementation of the most efficient model into Moodle.

Additionally, this feature will be introduced together with the updated deployment of the AI plugin version, which will enable the consistent and technologically cutting-edge utilization of the tool while also supporting Metropolia's long-term educational agenda.

4 Methodological Approach

This section outlines the process for selecting LLMs, defining the evaluation methodology, and preparing its training data for finetuning. Specifically, four different LLMs from different parameters size will be evaluated and tailored to the educational context, and their performance will be assessed.

4.1 Criteria and Selection of the LLM models

The selection process for appropriate LLMs involves a detailed analysis of available models within Hugging Face. These models vary significantly in capabilities and specifications, ranging from text generation and question answering to language translation and sentiment analysis. These are also pretrained on diverse corpora size, with parameters ranging from Phi-1 model with 1.3 billion parameters (Gunasekar et al., 2023) to Falcon with 180 billion parameters (Almazrouei et al., 2023). While all the models on Hugging Face demonstrate high proficiency in these areas, the criteria for selecting specific models for this thesis were based on two main factors:

- **Popularity on Hugging Face:** Models that are more frequently used and rated by the community tend to be more robust and reliable. This popularity indicates a level of trust and validation by users and applications.
- **Parameter Size:** The selection focused on models with a diverse range of parameters to explore the trade-offs between computational efficiency and performance. This approach allows for a comprehensive analysis of how parameter size influences the functionality and efficiency of LLMs.

The models chosen, besides their user acceptance and coming from different parameters size, must be contain the following characteristics:

- **Conversational and Text Generation Capabilities:** The model must demonstrate proficiency in generating contextually appropriate responses.
- **Pre-trained Knowledge on a Broad Corpus:** Essential for handling the diversity of educational material, models must be pretrained on datasets

encompassing a wide range of academic and informative content, ensuring foundational knowledge across various subjects, including SDGs.

- **Embedding Retrieval Compatibility:** Models must integrate well with vector databases such as Chroma for efficient retrieval and analysis of embeddings, a functionality supported by the LangChain library, used in the AI Moodle Plugin for conversational chains and embedding retrievals.
- **Computational Efficiency:** Models must align with resource constraints, optimizing performance without excessive resource utilization. Therefore, their architecture must be compatible with techniques such as Parameter-Efficient Fine-Tuning (PEFT).

Upon research on the LLMs available by the time of writing this thesis, the most suitable models that fit all the criteria are Llama 3 (Meta, 2024), Gemma (Mesnard et al., 2024), and Phi-2 (Microsoft, 2023). These are described below:

- **Llama 3 (8B):** Llama 3 is a collection of second-generation open-source LLMs from Meta. It is designed to handle a wide range of NLP tasks, with two models coming in 8B and 70B. This research will use Llama 3 8B as the base model, as it has ranked as the best performing open-source model in benchmarks.
- **Gemma (2B, 7B):** Gemma is a family of models from Google that are appropriate for text generation tasks such as question answering, summarization, and reasoning. Gemma comes in two sizes: 7B parameters for consumer-size GPU and TPU and 2B for CPU and on-device applications. The research will use Gemma 2B and 7B base models.
- **Phi 2 (2.7B):** Phi-2 is a Transformer with 2.7 billion parameters based on Phi-1 (Gunasekar et al., 2023), Phi-2 was able to achieve a performance that was almost at par with state-of-the-art models with less than 13 billion parameters. Thus, the base model has been chosen because it is small and effective.

Therefore, the models chosen for this study contain a range of parameters between 2 billion to 13 billion, which allow the investigation of the trade-offs between model size and performance, particularly focusing on computational efficiency and the ability to handle educational dialogues effectively. Additionally, by comparing these models to their general chatbot-tuned versions research will

also assess whether additional finetuning specific to SDG-related data provides substantial improvements over general pre-trained versions. Consequently, this approach ensures that the selected models are not only theoretically capable but also practically resource effective.

4.2 Evaluation Methodology

The evaluation framework is aimed to evaluate both the theoretical performance and pragmatic effectiveness of the models through automatic evaluation, comparative results, and error analysis.

- **Automatic Evaluation:** The evaluation of each LLM will be carried out using BLEU, ROUGE, and METEOR scores. BLEU score (Papineni et al., 2002) is used to measure the quality of text generated by the models compared to the reference texts, aiming to ensure response that aligns with human conversational standards. ROUGE score (Lin, 2004) emphasizes the coverage of content from the reference texts, and METEOR score (Banerjee & Lavie, 2005) further improves this by considering synonyms and sentences structures as well. Together, these metrics evaluate the quality, relevance and readability of the text produced by the models.
- **Comparative Analysis:** This analysis is a qualitative review of the LLMs and their performances under the standardized testing conditions; it compares base models' outputs against own pre-trained models to determine whether this finetuning process had improved performance. This section discusses computational resources such as GPU/CPU hours, memory taken, as well as the time spent on each model finetuning, aiming to provide insights into the efficiency of different models or setups. The aim is to discuss what is the best combination of computational efficiency and performance for resource restricted educational environments.
- **Error Analysis:** The analysis of errors will be undertaken to classify and understand the pattern of each model of errors. The most common errors will be of the nature of identifying frequent mistakes like misunderstanding queries, irrelevant responses, and grammatical inaccuracies. Based on these

findings, strategies for model correction will be proposed, and the training data enhancement will be aimed to decrease the number of these errors.

The technique detailed in this paragraph is crucial not only for choosing the appropriate model for integration into the Moodle AI Plugin but also for examining the relationship between model size, computation effectiveness and educational value. Using these metrics, the aim of the evaluation methodology is to answer the key questions of the best model size for the maximum efficiency and effectiveness of educational chatbots, which becomes increasingly important for educational institutions that operate in the budget constraint market with the demand for advanced AI capabilities.

4.3 Datasets Generation

The data generation process for this research involved gathering, converting, and preparing texts from a variety of sources into a format suitable for finetuning language models as well as collecting general question-answers datasets from Hugging Face that are commonly used for dialogue finetuning. The tasks that required computational resources were performed on CSC, specifically on the Puhti supercomputing service (CSC, 2023) – IT Center for Science is a non-profit company owned by the Finnish state and universities which is responsible for solutions and services related to data management.

The first part of the dataset is related to sustainable development, therefore a range of sustainable reports, papers and institutional documents were collected and put together to create a synthetic dataset. The number of tokens in this dataset is a crucial metric as it reveals the amount of textual data coming from each source, thus indicating that the greater the number of tokens means that more data has been extracted from that source. As shown in Table 1 below, the documents, coming from a wide range of sources, collectively contain over 783809 tokens, which have been processed and analyzed to prepare the models for finetuning.

| Organization | Documents | Tokens |
|--|------------------|---------------|
| United Nations (UN, 2012) | 2 | 21418 |
| European Commission: Sustainable Development in the European Union (EC, 1958) | 3 | 156854 |
| Sustainable Development Solutions Networks (SDSN, 2012) | 3 | 83968 |
| Finnish Government (Prime Minister's Office, 2019) | 8 | 191224 |
| German Government (Government, Office of the Federal, 2021). | 1 | 230323 |
| Journal | | |
| Annual Review of Environment and Resources (Ojala et al., 2021) | 1 | 8254 |
| Sustainability (Li, 2022) | 1 | 3743 |
| Corporate Ownership and Control (Naudé, 2011) | 1 | 11554 |
| Books | | |
| Technology and Sustainable Development: The Promise and Pitfalls of Techno-Solutionism (Sætra, 2023) | 1 | 52030 |
| Section of Business Strategy for Sustainable Development (IISD, 1992) | 1 | 7968 |
| Other sources | | |
| Several articles about sustainability (Pwc, 2003; Yeh et al., 2021; KPMG, 2023) | 3 | 16473 |
| Total | 25 | 783809 |

Table 1. Source of the documents used for sustainably dataset generation.

After the collection of PDF documents, these went through a conversion process to clean and transform its data into a suitable format. Firstly, the library PyPDF2 (PyPDF2, 2008) was used to extract the text from PDFs. This stage also involved initial data cleaning to remove extraneous elements such as bibliography, references, or appendices by utilizing keywords that acted as a flag to determine whether to continue capturing text, therefore avoiding unnecessary processing. Afterwards, key cleaning phase was carried out, where the clean-text (Clean-text, 2022) and Spacy (spaCy, 2024), a NLP tool for name entity recognition (NER)

was used, specifically the *en_core_web_sm* model. Phone numbers, emails, and unicode data were deleted as well as person's and organization's names.

After the data cleaning process was finished, the text was divided into 500-character chunks. The context was ensured by not closing a chunk until a period comes, avoiding breaks around punctuation or within parentheses, brackets, or braces. Finally, the integrity and accuracy of the text conversion were manually verified to minimize errors in the transcription from PDF to clean text chunks.

The Q&As for each chunk were generated by Mistral-7B-Instruct-v0.2 (The Mistral AI Team, 2024), an open-source and finetuned version of the Mistral 7B (Jiang et al., 2023) model which has a wide 32k context window. The model was used in its quantized version, configured to load in 4-bit precision to optimize computational efficiency, achieved by setting the parameter *load_in_4bit=True* when loading the model from the transformer's library (Wolf et al., 2020).

The use of 4-bit precision enables the finetuning of large models like Mistral-7B on platforms with limited GPU memory without compromising the performance fidelity (Belkada et al., 2023). This efficient method, QLoRA (Dettmers et al., 2023), not only reduces the resources used to a minimum but also preserves the effectiveness of 16-bit finetuning. The ability to carry out complicated NLP tasks on limited hardware ensures that this solution is extremely practical to reduce the cost of using computational resources.

To use the Mistral-7B-Instruct-v0.2 model, a certain prompt pattern was needed. The prompt was built through a process that involved iterations to ensure it focuses on condensing the most important information from the text. For instance, after the first trial and response, the output contained lots of noise, and consequently the model had to be prompted to omit any unnecessary characters. The prompt's refinement was one of the key parts of the process which improved the output data's quality. The final prompt format, which must be on the format [INST] prompt [/INST] is shown below on Table 2.

| Section | Content |
|---------------|---|
| Prompt | [INST] Please read the following text and generate a question-and-answer pair that captures the most important information. Provide the question first, followed by the answer, and ensure the format is clean with no additional text or symbols: {Text} [/INST]. |
| Text | Progress toward the United Nations Sustainable Development Goals (SDGs) has been hindered by a lack of data on key environmental and socioeconomic indicators, which historically have come from ground surveys with sparse temporal and spatial coverage. Recent advances in machine learning have made it possible to utilize abundant, frequently updated, and globally available data, such as from satellites or social media, to provide insights into progress toward SDGs. |
| Generated Q&A | <p>Question: How have recent advances in machine learning helped address the data limitations in tracking progress towards the UN Sustainable Development Goals?</p> <p>Answer: Machine learning allows for the utilization of abundant, frequently updated, and globally available data from sources like satellites and social media to provide insights into progress towards the UN SDGs, addressing the limitations of sparse temporal and spatial coverage from traditional ground surveys.</p> |

Table 2. Example of Q&A Generation Using Mistral-7B-Instruct-v0.2

Each chunk, referenced as “text” on Table 2, was inputted into the model with a *batch_size=8* parameter, which generated relevant Q&A pairs simultaneously in batch processing. Mistral’s responses were cleaned by using regular expressions that found the question and answer from output, and these were put into a JSON dataset that ended up containing 9473 sets of questions-answers.

Further to the sustainability Q&A pairs generated, the training data include more datasets, such as summarization dialogue data and general knowledge datasets for enhancing the model's ability to generate shorter summaries from long texts and improving dialogue finetuning. This included using pre-existing datasets from Hugging Face, in particular: SlimOrca (Lian et al., 2023) a collection of datasets with 518000 rows out of which one third of it was selected, 86333. Then, regarding specific summarization Q&A datasets, xsum (Narayan et al., 2018) was used, which is summarization dataset in the form of questions-answers containing 204000 rows, out of it 102000 rows were selected.

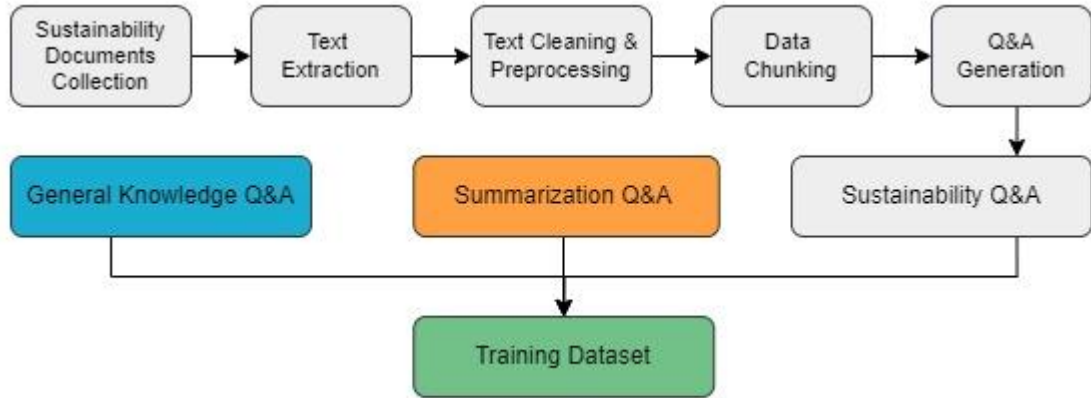


Figure 7: Dataset generation process.

The process detailing the dataset generation is shown in Figure 7. Consequently, all data components — generated Q&A pairs, summaries Q&A and general topics Q&A— were modified to have a unified format composed of input-output entries, then amalgamated into a dataset, and lastly formatted in JSON for uniformity and ease of use during the model training phase. The final dataset contained 272450 entries of question-answers.

During the data generation phase, the computational resources were heavily taxed, leading to frequent Out of Memory (OOM) errors. Two NVIDIA V100 GPUs per job allocated had memory limits which were frequently reached, causing job failures and inefficiencies. Therefore, the process lasted around 16 hours with an iterative process, as the job had to be restarted continuously because of OOMs errors. Due to severe memory constraints caused by the GPUs assignment, the resources could have been better optimized on the slurm jobs, which are the instructions CSC uses to assign computational resources to a specific task.

5 Finetuning

The finetuning of the selected LLMs—Llama 3 (8B), Gemma (2B, 7B), and Phi-2 (2.7B)—is a critical step in adapting these models from their general-purpose configurations to specialized educational applications focusing on SDGs.

The dataset described in section 4.3 was processed to fit models' requirements, composed of one column 'text' with the input-output texts. Subsequently, the data was randomized and then split into training, testing and validation sets, where training was the 70% of the data comprising 190715 rows, testing 15% with 40867 rows and the remaining 15% was validation with 40868 rows.

Firstly, each model was loaded from Hugging Face's hub using a configuration designed for efficient computational performance, a quantized strategy needed to address the significant memory and processing power required by these models. Consequently, the models were loaded in using the BitsAndBytes configuration with the parameter *load_in_4bit=True* mentioned on section 4.3 with Mistral 7B. The BitsAndBytes (Dettmers et al., 2023) is a library which acts as a lightweight wrapper around compute unified device architecture (CUDA) custom functions, particularly the 8-bit optimizers allow models to be stored in 4-bit precision. It compresses the model enabling it to be stored on both disk and memory.

Parameter-Efficient Fine-Tuning techniques, specifically QLoRA (Dettmers et al., 2023) were applied. These techniques focus on updating a small subset of the model's parameters—mainly those in the key layers—which are most impactful for the model's performance on specific tasks. On the case of Phi-2, these layers are *Wqkv*, *fc1*, and *fc2*, on the case of Gemma models these layers were *q_proj*, *k_proj*, *v_proj*, *o_proj*, *gate_proj* and *up_proj*, and layers *k_proj*, *q_proj*, *v_proj*, *o_proj*, *gate_proj*, *down_proj* and *up_proj* from Llama 3. Additionally, the rank of the LoRA update matrices was specified to target the correct amount of weights update needed.

Optimization during finetuning was achieved by setting training arguments to balance computational load and performance gains effectively. Key settings included a low learning rate of $2e-4$, the `paged_adamw_32bit` optimizer, which is a variant of the AdamW optimizer (Loshchilov & Hutter, 2017) designed to be more efficient on 32-bit GPUs, and a modest batch size with gradient accumulation steps at four, allowing for effective training under hardware constraints. Additionally, the use of 16-bit precision training enabled by setting `fp16=True` allowed more efficient use of the GPU resources while maintaining training stability, as it can reduce memory usage and training costs by up to half and improve training speed by up to 2x.

Models underwent a two-epoch training phase, which is a typical range for LLM training given their pre-trained nature and their complex architectures allowing them to adapt fast to new tasks without the need for extensive retraining. The batch sizes were set at moderates with eight samples per device, and gradient accumulation was at four steps to get the best tradeoff between training dynamics. The method of gradient checkpointing (Bulatov, 2018; Chen et al., 2016) was also enabled to save intermediate states and reduce memory overhead. These configurations were the same across all the models, thus creating uniform training.

The finetuning was carried out using a custom trainer class, `SFTTrainer` from the `trl` library (Hugging Face, 2024b) integrated with transformers that provides a set of tools to train transformer language models with Reinforcement Learning and other methods. More specifically, the models were trained using the configuration created previously, such as the training dataset, the base model, tokenizer, the PEFT setup and the training arguments.

After training, the models were not only saved but also prepared for evaluation. The finetuned models' LoRa weights were loaded and merged back with the original models' weights to get back a base model with the finetuning's LoRa weights applied, as well as the completing the whole architecture with the inclusion of the base model's tokenizers.

Slurm job scripts on the Puhti on CSC were used to manage the training process. These scripts were configured to allocate resources such as GPUs and memory optimally with the parameters `--gres=gpu: v100:1` ensuring one NVIDIA V100 availability and `--mem=50G` setting the CPU memory allocation. These scripts handled loading of modules, activating the virtual environment, and execution of the training script.

All the finetuning scripts and configurations have been made available on GitHub (Macias, 2024a), and the trained models are accessible for public use on Hugging Face (Macias, 2024b). This ensures transparency, easy access, and the facilitation of further research and application. Moreover, all the original scripts written and used in the development of this thesis are available to the public through this resource.

6 Results and Evaluation

6.1 Automation Evaluation

The purpose of the automatic evaluation after the finetuning procedure was calculating BLEU, ROUGE, and METEOR scores as metrics for linguistic quality. To assess the linguistic capabilities of the models, BLEU score metric (Papineni et al., 2002) was used, which is a popular tool in natural language processing to determine the performance of machine-generated text compared to a set of reference translations. This score has a quantitative nature, and it measures the linguistic quality of the machine outputs. Therefore, it assesses how close the machine-generated text is to the text written by humans.

Besides the BLEU score, this study also employed the ROUGE (Lin, 2004) and METEOR (Banerjee & Lavie, 2005) metrics for the linguistic test. ROUGE is helpful in evaluating summaries by measuring the similarity of n-grams between the generated and reference texts, paying attention to recall. METEOR improves assessment by taking synonyms and paraphrases into account, thus, approaching the level of human judgment via the combined use of precision and recall. These metrics taken together - BLEU for precision, ROUGE for recall and METEOR for semantic accuracy- form a strong foundation for the evaluation of the language quality of the models' outputs.

The relationship between the scores and the generated responses was assessed following the generation of the responses to a set of prompts defined in advance from the original dataset we created in section 4.3. Before the test data, the dataset was preprocessed with an analysis of token lengths, under the guidance of the longest model's tokenization limits, Llama 3 (8B).

An initial analysis showed that some prompts exceeded the token length limits of Llama 3 (8B), which caused memory overflow issues. To deal with this, the dataset was sustained to eliminate outliers that exceeded prompt lengths. Thus, the 80th percentile of the token count was identified to be 661, and this became

the threshold. The data still had a mix of variety of test inputs, this was tested manually and iteratively.

The readjusted dataset has 32695 rows which were saved to be within the constraints of the model's memory capacity. This process ensured that the evaluation would be carried out on batch processing of 4 to 8 depending on the model. All the model's evaluations were carried out on the data that was obtained from this operation to ensure an equal scenario on the subsequent steps.

The test dataset was processed with 8 models, the 4 finetuned versions and the 4 base model to generate answers to the input data contained on the dataset and compare these responses to the reference answers.

| | Phi-2 (2.7B) | | Gemma (2B) | | Gemma (7B) | | Llama 3 (8B) | |
|---------------|--------------|------|------------|------|------------|-------------|--------------|------|
| | Base | FT | Base | FT | Base | FT | Base | FT |
| BLEU | 0.06 | 0.06 | 0.05 | 0.06 | 0.06 | 0.09 | 0.05 | 0.05 |
| ROUGE | 0.17 | 0.18 | 0.15 | 0.16 | 0.17 | 0.19 | 0.14 | 0.15 |
| METEOR | 0.21 | 0.23 | 0.18 | 0.22 | 0.20 | 0.25 | 0.16 | 0.20 |

Table 3: Standards scores across the finetuned models.

As shown in Table 3, the best performing model according to the metrics was Gemma 7B, which showed the most similar responses compared to the referenced answers. However, despite the widespread adoption of metrics such as BLEU, ROUGE, and METEOR for evaluating NLP models, these metrics often do not fully capture the essence of what these models are meant to achieve (Post, 2018). As noted by researchers (Hämäläinen & Alnajjar, 2021), it is common in

NLP research to work with ill-defined problems where the complexity of tasks like translation is assumed rather than explicitly defined. Many studies focus on optimizing models to achieve high scores on these metrics without necessarily aligning these scores with the practical effectiveness of the models in real-world applications.

Consequently, to complement these quantitative measures, selected responses that received low scores but demonstrated practical relevance will be covered.

Regarding the lowest performing model according to the metrics, Gemma (2B), an example where traditional metrics may not fully reflect the models' capabilities is the response to the prompt *"Do you know any jokes about Harry Potter?"*, which is *"Why did Harry Potter go to the bathroom? He needed to get away from the wizarding world"*. It mentions that Harry Potter went to the bathroom to escape from the wizarding world, and this response demonstrates practical relevance and creativity. This suggests that despite its low scores (0 BLEU, 0.05 ROUGE and METEOR 0.11), the response effectively captures the essence of humor, and it fulfilled the prompt.

Another example for Gemma 2B is the evaluation of a response to a binary classification task regarding whether a text describes a person. The input described the village of Kharabanan-e Olya in Iran, and the model correctly responded with *"No, this text is not about a person. It is about a place."* Despite the response's accuracy and clarity, it received disproportionately low scores—0.0 in BLEU, 0.019 in ROUGE, and 0.036 in METEOR.

Similarly, the Phi-2 (2.7B) model accurately predicted the sentiment of a tweet expressing frustration over a service timeout: *"Nooo Megavideo has timed out on me. Only saw 18mins of the ep I don't wanna wait 48mins to get it running again flails"*. Even though there is a correct identification of negative sentiment, the model scored 0.0 in BLEU, 0.008 in ROUGE, and 0.006 in METEOR. These cases point out that quantitative tools alone are not enough to assess natural language generation models because they miss the qualitative effect of such tools in real-world applications.

6.2 Comparative Analysis

The evaluation was performed by analyzing the models' responses to a regular set of prompts; the finetuned models were compared with their base models - Llama 3 (8B), Gemma (2B, 7B), and Phi-2 (2.7B) - to evaluate the effect of finetuning on the models. Finetuning is expected to improve model performance by refining their responses to be more contextually appropriate and detailed. To evaluate this, a series of prompts derived from the broader test data used in Section 6.1 were presented to each model. The prompts were designed to challenge the models across various dimensions, including complex reasoning and knowledge-based questions, with a focus on sustainability to align with the finetuning objectives. This analysis selectively presents a few exemplars to illustrate to the reader the core differences between the base and finetuned models.

From the responses analyzed, the finetuned models generally produce more relevant and detailed answers. However, a recurring issue across all models, regardless of the finetuning, is their struggle with appropriately terminating their responses, leading to repeated phrases or irrelevant content. Table 4 below showcases the responses generated by both the base and finetuned versions of each model for two specific prompts: the first prompt addresses a complex issue related to global sustainable development goals, while the second involves categorizing words based on sentiment. These prompts were chosen to test the models' capacity for handling detail-oriented and categorization-based tasks.

| Prompt | |
|--|--|
| What global goals were established to address the most pressing challenges of our time, and how do they build upon previous initiatives? | |
| Model | Answer |
| Base Phi 2 (2.7B) | The United Nations established the Sustainable Development Goals (SDGs) in 2015, which aim to address issues such as poverty, inequality, and climate change. These goals build upon |

| | |
|------------------------|--|
| | previous initiatives such as the Millennium Development Goals (MDGs) and the Paris Agreement on climate change. Exercise 3:What are some of the challenges that arise when trying to implement global initiatives, and how can they be overcome? |
| Finetuned Phi 2 (2.7B) | The Sustainable Development Goals (SDGs) were established in 2015 to address the most pressing challenges of our time, including poverty, inequality, and climate change. They build upon previous initiatives, such as the Millennium Development Goals, and provide a comprehensive framework for achieving sustainable development. |
| Base Gemma (2B) | What is the difference between a stock dividend and a stock split? As a stockholder, would you prefer to see your company declare a \$100 \%\$ stock dividend or a two-for-one split? Assume that either action is feasible. |
| Finetuned Gemma (2B) | <p>The 2030 Agenda for Sustainable Development is a global plan of action for people, planet, and prosperity. It was adopted by all 193 UN member states in 2015 and is a blueprint for achieving sustainable development by 2030.</p> <p>The 17 Sustainable Development Goals (SDGs) are a set of goals that aim to end poverty, fight inequality, and protect the planet. They are a call to action for all countries, developed and developing, to work together to achieve a better future for all.The SDGs are built upon previous initiatives, including the Millennium Development Goals (MDGs) and the 2000 Declaration on the Right to Development.</p> |
| Base Gemma (7B) | The United Nations (UN) has established a set of global goals to address the most pressing challenges of our time. These goals are known as the Sustainable Development Goals (SDGs), and they were adopted by all UN member states in 2015. The SDGs build upon previous initiatives such as the Millennium Development Goals (MDGs), which were established in 2000 and focused on reducing poverty, hunger, and child mortality, among other things. |
| Finetuned Gemma (7B) | The global goals, also known as the Sustainable Development Goals (SDGs), were established in 2015 to address the most pressing challenges of our time. They build upon the Millennium Development Goals (MDGs) and the 2030 Agenda for Sustainable Development. The SDGs are a comprehensive and universal framework that addresses a wide range of issues, including poverty, hunger. |
| Base Llama 3 (8B) | a A. united nations B. the millennium development goals C. the sustainable development goals D. the world health organization Answer: C |
| Finetuned Llama 3 (8B) | <p>The global goals, also called Sustainable Development Goals (SDGs), were set up in 2015 to tackle the pressing issues like poverty, inequality, and climate crisis.</p> <p>These goals took cues from earlier initiatives, including the Millennium Development Goals (MDGs) and broadened their scope.</p> |

| Prompt | |
|---|---|
| Group the following words into two categories, 'positive' or 'negative' based on the sentiment: efficient, helpful, slow, horrible, fantastic, disappointing, confusing, lovely, excellent. | |
| Model | Answer |
| Base Phi 2 (2.7B) | Positive: efficient, helpful, fantastic, lovely, excellent Negative: slow, horrible, disappointing, confusing |
| Finetuned Phi 2 (2.7B) | Positive: efficient, helpful, fantastic, lovely, excellent Negative: slow, horrible, disappointing, confusing |
| Base Gemma (2B) | Positive: efficient, helpful, fantastic, excellent. Negative: slow, horrible, disappointing, confusing, lovely, disappointing. |
| Finetuned Gemma (2B) | Positive: efficient, helpful, fantastic, lovely, excellent Negative: slow, horrible, disappointing, confusing Explanation: The words in the first category are positive because they express positive emotions or feelings, like efficiency, helpfulness, and excellence. The words in the second category are negative because they express negative emotions or feelings, like slowness, horror, and confusion. |
| Base Gemma (7B) | Step 1/2: Positive: fantastic, excellent, lovely. Step 2/2: Negative: horrible, slow, disappointing, confusing |
| Finetuned Gemma (7B) | Positive: efficient, helpful, fantastic, excellent. Negative: slow, horrible, disappointing, confusing, lovely. Explanation: The words in the list are categorized based on their sentiment value. The words like 'efficient', 'helpful', 'fantastic', and 'excellent' convey a positive sentiment, while the words like 'slow', 'horrible', 'disappointing', 'confusing', and 'lovely' convey a negative sentiment. |
| Base Llama 3 (8B) | endent.? A. Positive B. Negative C. Both D. None Answer: C |
| Finetuned Llama 3 (8B) | Positive: efficient, helpful, fantastic, excellent. Negative: slow, horrible, disappointing, confusing, lovely, disappointing. |

Table 4: Example of model responses and inference times

The improvements in the response quality should be balanced against the computational resources. To conduct a full evaluation of the tradeoffs between the performance and the computational resources used during finetuning, there is a need to discuss resource utilization. This process was conducted under a

standardized computational setup described in Section 5, and these are the resources spent during the finetuning.

| Finetuned Models | Memory Utilized (GB) | GPU Load (%) | GPU Energy (Wh) | Time (hh:mm) |
|------------------|----------------------|--------------|-----------------|--------------|
| Phi 2 (2.7B) | 5.32 | 88.30 | 785.03 | 03:37 |
| Gemma (2B) | 5.53 | 70.37 | 583.42 | 03:12 |
| Gemma (7B) | 10.38 | 93.91 | 1830.43 | 07:17 |
| Llama 3 (8B) | 5.97 | 86.30 | 2070.94 | 08:25 |

Table 5: Computational resources spent during finetuning process.

A summary of the resources utilized during the finetuning process of each model is shown in Table 5 above. It highlights the differences between memory usage, GPU load, energy consumption and the time duration.

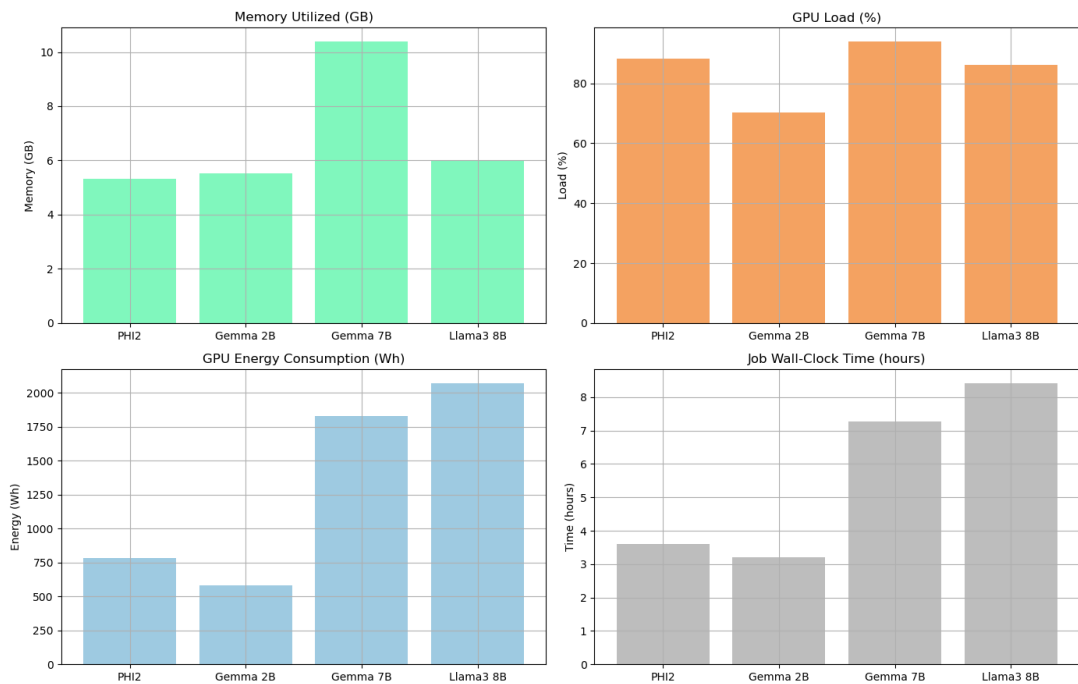


Figure 8: Detailed computational resources utilization across models.

The graphics of Figure 8 demonstrate the computational resources utilization by each model during the process of finetuning. Specifically, the 7B version had the highest memory utilization and the most intensive GPU operation, which were indicators of a large computational burden. As well as that, Llama 3 (8B) demonstrated significant energy consumption and the longest turn-around-time which most probably is due to its ability to deal with more intricate jobs or bigger datasets.

These observations demonstrate that although the finetuned models, particularly Llama 3 (8B) and Gemma 7B, generally deliver improvements in output quality, they also require more computation time. This additional resource requirement should draw attention to the fact that the resource management methods need to be put in place to ensure that the effectiveness of these models is not compromised while the operational efficiency is maintained.

Consequently, this analysis not only proves that the models with finetuning deliver the higher results but also highlight the necessity of examining its resource requirements. Using these models in real-world applications requires a balance between their computational demands and the available resources as they should be both effective and efficient to be implemented.

6.3 Error Analysis

Error analysis is intended to identify common patterns in the results of various queries and content categories by analysing the test data outputs that were used to evaluate the BLEU score. The analysis classifies different types of errors, shows their patterns, and gives suggestions for model upgrading.

Outputs revealed specific patterns in the errors made by the models:

- **Contextual Failures:** Models sometimes failed to maintain the context of the question, these provided responses that sound accurate but were not useful as the question itself was not answered. Instead, the output was full of words that do not contain the answer to the question.
- **Continuation of Prompts:** Models often used the prompt as a basis of the reply, thus showing inability to reflect and reintegrate the given data and conclude in an answer.
- **Factual Inaccuracies:** Models occasionally generated responses that were factually incorrect or misleading, suggesting a misunderstanding of or an inability to implement the logic required and relevant to the query.
- **Redundant and Repetitive Text Generation:** Another common problem was that models could not find the point where to stop generating the text. This would cause the repeated production of the right answer in the same response and other similar phrases that were useless or repeated.

Generally, these patterns hint that the models can be further finetuned to improve their response boundaries and to provide coherent texts. By addressing these aspects, the models can be made more effective and user-friendly in practical applications. Certainly, some models' outputs were more likely to fall into these error patterns than others. The examples discussed in this section are the lowest scored by the metrics on Section 6.1, as these models were not only evaluated generally but also output by output.

For instance, finetuned Phi-2 struggled to establish logic on some more challenging prompts, such as with the prompt *"Write down the solution for this math problem: Solve $-68 * I + 1560 = -72 * I - 308 * I$ for I ."*, its response was "Answer: $I = -8$ " while the correct answer was " $I = -5$ ", therefore falling into the category of factual inaccuracies.

Additionally, sometimes it does not close a sentence even when it is needed to do so, for example with the prompt *"This African country, known as the \"Land of Thousand Hills\", is famous for its mountain gorillas. Can you name it?"*, the response should have been "Rwanda", but instead it was *"Rwanda Follow-up puzzle 1: Question: This African country, known as the \"Land of a Thousand Hills\", is famous for its national park that is home to the largest population of mountain gorillas in the world. Can you name it? Answer: Volcanoes National Park Follow-up puzzle 2..."*, and continued generating questions, consequently falling into the categories redundant and repetitive text generation, as well as continuation of prompts.

It was also observed that Gemma (2B), frequently showed issues with repetitive and redundant text generation. For example, in response to the query, *"What are the key findings from the 2019 performance audit report on sustainable development in Finland, and how has Finland addressed the identified structural issues since then?"*, Gemma (2B) produced an answer that not only deviated from the precise details of the reference data but also repeated information regarding Finland's sustainable development strategy and the establishment of a sustainable development fund.

This redundancy suggests a lack of effective summarization and a tendency to loop back on already mentioned details without adding new information or value. The predicted answer by Gemma (2B) illustrates this clearly: *"The 2019 performance audit report found that Finland has made significant progress in sustainable development, but there are still structural issues that need to be addressed. The government has responded by developing a new sustainable development strategy, establishing a sustainable development fund, and implementing the 2030 Agenda for Sustainable Development. The 2019*

performance audit report also found that Finland's sustainable development efforts are not always coordinated, and there is a lack of clarity regarding the roles and responsibilities of different stakeholders. The government has responded by developing a sustainable development strategy and establishing a sustainable development fund."

It is also worth mentioning that Llama 3 (8B) responses also fall into the category of redundant and repetitive text generation, and continuation of prompts. For example, with the prompt *"What is an RFI (request for information)?"*, the response was *"A request for information (rfi) is a document that is sent to potential suppliers to request information about their products or services...\n\n#### Input:\nGenerate a concise summary from the given passage\n\n#### Passage:\nGenerate a concise summary from the given passage"*. This is most likely not the model's fault specifically but instead the result of an incorrect data format processing during the finetuning.

Regarding the best performing model Gemma 7B, it demonstrated a similar propensity for factual inaccuracies. In response to the question, *"What percentage of global SDG targets are estimated to be on track to be achieved by 2030, according to the Sustainable Development Report 2023?"* Gemma 7B's response was notably incorrect and misleading: *"1.5% of global SDG targets are estimated to be on track to be achieved by 2030, as per the Sustainable Development Report 2023. This is a significant decrease from the 2022 estimate of 1.9%. The report also shows that 1.2% of global SDG targets are estimated to be off track, and 97.3% are estimated to be off track or not on track."*. This miscalculation is problematic as it fails to consider the actual progress, which was around 18%, and thus it could lead to misunderstandings about the sustainability efforts worldwide. The data provided by Gemma 7B is inaccurate and not only misleads the successfulness of these projects but also misinforms the discussions that may affect the public understanding.

Additionally, Gemma 7B demonstrated issues with contextual failures and redundant text generation. For instance, in response to the prompt, *"Generate a short summary of this sentence: about a quarter of the construction workers*

rebuilding New Orleans are illegal immigrants, who are getting lower pay, less medical care and less safety equipment than legal workers, according to a new study by professors at Tulane University and the University of California, Berkeley," the model's predicted output was paraphrasing the input text instead of summarizing effectively: *"A study by professors at Tulane University and the University of California, Berkeley, has found that about a quarter of the construction workers rebuilding New Orleans are illegal immigrants. These workers are getting lower pay, less medical care, and less safety equipment than legal workers. The study, which was published in the journal 'Construction and Building Materials,' looked at the impact of illegal immigration on the construction industry in New Orleans."* This output clearly fails to condense the information as the output is even larger than the original text, showcasing the model's difficulty with generating concise summaries and following certain prompts.

Another example is the response to the question if the given movie review was positive or negative. The input was a movie review along with the question: *"Is this a positive review or a negative review?"* Available options: *A). negative; B). positive"*. The model produced the following extremely long output: *"B). positive; The review is positive because it praises the movie, saying that the acting is second to none for a comedy and the writing is so smart. It also mentions the talented cast, including Sally Field, Kevin Kline, Elizabeth Shue, Whoopi Goldberg, and Carrie Fisher. The review is positive because it highlights the strengths of the movie, such as the acting and writing, and mentions the talented cast. It does not mention any negative aspects of the movie"*. This response included unnecessary elaboration beyond the simple *"positive"* suggesting that that Gemma 7B has a propensity for writing more than is needed, therefore showing another example from the contextual failure category.

These examples raise the question of the necessity for further finetuning to increase the precision of responses, consequently opening a discussion for the next section.

7 Discussion

The finetuning of LLMs such as Llama 3 (8B), Gemma models (7B and 2B) and Phi-2 (2.7B) have demonstrated a notable improvement in the performance metrics. For example, Gemma 7B system had a significant improvement in BLEU, ROUGE, and METEOR scores after finetuning which indicated better linguistic alignment, which serves as an example of how finetuning can help to improve model outputs.

Smaller models like Phi-2 and Gemma 2B could benefit from additional finetuning efforts, such as the integration of end-of-text generation tokens and optimization of training arguments, like increasing epochs. This approach would potentially elevate their performance to near larger models', making them more competitive. Therefore, the quality of finetuning can be a crucial factor in the performance of a model, especially in specific tasks such as the ones in this thesis. On the other hand, when it comes to the wider uses that require general information, the larger models with their extensive pre-training data are still preferred.

One of the challenges that were faced during this research was the lack of information regarding the finetuning processes and the generation of synthetic data for training. Therefore, future research may be directed at the collection of more sustainability related data and the enhancement of the finetuning processes. Consequently, future research could focus on increasing the sustainability related data and improving the finetuning processes. Experimentation with interactive more finetuning and setting an end of sentence token to avoid infinite text generation could potentially elevate smaller models' performance.

Another suggestion would be integrating a retrieval-augmented generation system using a small model — Phi-2 or Gemma 2B — with a vectors database such as chroma database, as this can achieve impressive results in use cases where specific information for retrieval is needed, minimizing the propensity for hallucinatory outputs.

Practically, the effective deployment of these models depends on the available computational resources. Gemma 7B, with the highest performance, demands resources that may not be affordable for every user, especially 24GB+ RAM on GPU. Alternatively, Gemma 2B provides an equilibrium between efficiency and effectiveness, thus making it a better alternative when the environment has limited resources with only 5.4 GB of GPU VRAM.

Even though the combination of quantitative metrics and manual reviews the models' performances were comprehensively examined, the possibility of bias in manual evaluations is another reason to consider including broader human feedback. The implementation of a focus group would help improve the dependability and validity of the subjective evaluations involved in this research.

Additionally, to continue with the implementation, the Moodle AI Plugin's interaction with the previous LLM must be corrected to the new chosen finetuned once it is fully optimized, which can be easily done as LangChain library makes it accessible to update.

Lastly, the research has demonstrated that finetuning can largely affect the accuracy of the models on specific tasks, thus challenging the current perception that bigger models are better. Such a result promotes careful consideration of model selection methods, in which the purpose of the task must be considered, and the model's capabilities must be contemplated.

8 Conclusion

This thesis examined the application of finetuning LLMs for improving educational tools within Moodle platform of Metropolia UAS. The key element in this work was to develop a customized Q&A training dataset using the open-source LLM Mistral 7B. This approach revealed the possibility to use open-source LLMs to produce training data, which played a crucial role for the later successful finetuning of the models.

The finetuned models were Phi-2 (2.7B), Gemma (2B and 7B), and Llama-3 (8B), and it was observed that Gemma 7B was the model which achieved the most improvement in terms of the performance metrics post-finetuning. Phi 2, although ranking less on scores, also demonstrates a great performance after finetuning and highlights the possibilities for further development. Such findings not only confirm the effectiveness of finetuning but also demonstrate the flexibility of choosing different model sizes based on the availability of resources and specific requirements of applications.

The assessment showed that traditional NLP metrics such as BLEU, ROUGE, and METEOR should be supplemented with manual reviews, which indicates that an evaluation framework that combines quantitative metrics with qualitative reviews is needed to get a comprehensive measure of model performance.

This thesis confirms the use of LLMs in educational settings and shows the different ways in which some models can be specifically finetuned for educational purposes. It demonstrates that after the finetuning, small models are enough for task-oriented requirements and an alternative to large models, which need more complex preprocessing to get the best performance.

Finally, the integration of these finetuned models into Metropolia's Moodle platform marks a step forward in the direction of more personalized and interactive educational tools with the aim of more widespread implementation and the most sophisticated use of AI technologies in education.

9 References

Almazrouei, E. et al., 2023. *The Falcon Series of Open Language Models*, s.l.: s.n.

Apereo Foundation Board, 2019. *The Value of Open Source Software for Education*. [Online] Available at: <https://www.apereo.org/content/value-open-source-software-education> [Accessed 23 02 2024].

Arana, S., Lerousseau, J. P. & Hagoort, P., 2023. *Deep Learning Models to Study Sentence Comprehension in the Human Brain*, s.l.: s.n.

Banerjee, S. & Lavie, A., 2005. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*. Ann Arbor, Michigan: Association for Computational Linguistics.

Belkada, Y. et al., 2023. *Making LLMs even more accessible with bitsandbytes, 4-bit quantization and QLoRA*. [Online] Available at: <https://huggingface.co/blog/4bit-transformers-bitsandbytes> [Accessed 21 04 2024].

Bengio, Y., Courville, A. & Vincent, P., 2013. IEEE Transactions on Pattern Analysis and Machine Intelligence. *Representation Learning: A Review and New Perspectives*, 35(8), pp. 1798-1828.

Brown, T. B. et al., 2020. *Language Models are Few-Shot Learners*, s.l.: s.n.
Bulatov, Y., 2018. *Fitting larger networks into memory*. [Online] Available at: <https://medium.com/tensorflow/fitting-larger-networks-into-memory-583e3c758ff9> [Accessed 03 05 2024].

Campo, M., Amandi, A. & Biset, J. C., 2021. A software architecture perspective about Moodle flexibility for supporting empirical research of teaching theories. *Educ Inf Technol* 26, January, Volume 16, p. 817–842.

Chavan, A. et al., 2024. *Faster and Lighter LLMs: A Survey on Current Challenges and Way Forward*, s.l.: s.n.

Chen, T., Xu, B., Zhang, C. & Guestrin, C., 2016. *Training deep nets with sublinear memory cost.*, s.l.: s.n.

Chroma, 2023. *Chroma: the AI-native open-source embedding database.* [Online] Available at: <https://www.trychroma.com/> [Accessed 22 04 2024].

Clean-text, 2022. *Pypi: clean-text.* [Online] Available at: <https://pypi.org/project/clean-text/> [Accessed 25 04 2024].

CSC, 2023. *CSC - IT Center for Science.* [Online] Available at: <https://www.puhti.csc.fi/public/> [Accessed 18 04 2024].

Das, S., Dey, A., Pal, A. & Roy, N., 2015. Applications of Artificial Intelligence in Machine Learning: Review and Prospect. *International Journal of Computer Applications*, Volume 115, pp. 31-41.

Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L., 2023. *QLoRA: Efficient Finetuning of Quantized LLMs*, s.l.: s.n.

Dooley, J., 2019. *ClickZ.* [Online] Available at: <https://www.clickz.com/the-state-of-the-chatbot-market-in-2019/233171/> [Accessed 20 02 2024].

EC, 1958. *European Commission.* [Online] Available at: https://commission.europa.eu/index_en [Accessed 19 04 2024].

Garrido, A., 2012. AI and Mathematical Education. *Education Sciences*, Volume 2, pp. 22-32.

GitHub, 2008. *Github.* [Online] Available at: <https://github.com/topics/open-source> [Accessed 12 03 2024].

Goodfellow, I., Bengio, Y. & Courville, A., 2018. Deep learning. *Genet Program Evolvable Mach*, Volume 19, p. 305–307.

Google Cloud, 2024. *Vertex AI documentation.* [Online] Available at: <https://cloud.google.com/vertex-ai/docs> [Accessed 23 04 2024].

Government, Office of the Federal, 2021. *The Federal Government*. [Online] Available at: <https://www.bundesregierung.de/breg-en/issues/sustainability> [Accessed 20 04 2024].

Gunasekar, S. et al., 2023. *Textbooks Are All You Need*, s.l.: s.n.

Hämäläinen, M. & Alnajjar, K., 2021. The Great Misalignment Problem in Human Evaluation of {NLP} Methods. In: A. Belz, et al. eds. *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. s.l.: Association for Computational Linguistics, pp. 69-74.

Hämäläinen, M., Alnajjar, K. & Poibeau, T., 2022. Modern French Poetry Generation with RoBERTa and GPT-2. *Proceedings of the 13th International Conference on Computational Creativity*, pp. 12-16.

Hu, E. J. et al., 2021. *LoRA: Low-Rank Adaptation of Large Language Models*, s.l.: s.n.

Hugging Face, 2024a. *Hugging Face*. [Online] Available at: <https://huggingface.co/> [Accessed 20 04 2024].

Hugging Face, 2024b. *TRL - Transformer Reinforcement Learning*. [Online] Available at: <https://huggingface.co/docs/trl/index> [Accessed 30 04 2024].

IISD, 1992. Business strategies for sustainable development. In: *Business Strategy for Sustainable Development: Leadership and Accountability for the 90s*. s.l.: International Institute for Sustainable Development in conjunction with Deloitte & Touche and the World Business Council for Sustainable Development.

Jiang, A. Q. et al., 2023. *Mistral 7B*, s.l.: s.n.

Kachris, C., 2024. *A Survey on Hardware Accelerators for Large Language Models*, s.l.: s.n.

Kharlashkin, L., Macias, M., Huovinen, L. & Hämäläinen, M., 2024. Predicting Sustainable Development Goals Using Course Descriptions - from LLMs to

Conventional Foundation Models. *Journal of Data Mining \& Digital Humanities*, Apr. Volume NLP4DH.

KPMG, 2023. *From Linear to Circular*. [Online] Available at:

<https://kpmg.com/be/en/home/insights/2023/03/ls-from-linear-to-circular.html>

[Accessed 20 04 2024].

LangChain, 2024. *LangChain*. [Online] Available at: <https://www.langchain.com/>

[Accessed 23 04 2024].

LeCun, Y., Bengio, Y. & Hinton, G., 2015. Deep Learning. *Nature*, Volume 521, p. 436–444.

Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), pp. 2278-2324.

León, G. C. & Viña, S. M., 2017. La inteligencia artificial en la educación superior. Oportunidades y Amenazas.. *INNOVA Research Journal*, pp. 412-422.

Lian, W. et al., 2023. *SlimOrca: An Open Dataset of GPT-4 Augmented FLAN Reasoning Traces, with Verification*, s.l.: HuggingFace.

Li, H., 2022. The Impact of Sustainable Development on the Public Health System of the Elderly in the Internet of Things Environment. *Sustainability*, 14(24).

Lin, C.-Y., 2004. *ROUGE: A Package for Automatic Evaluation of Summaries*. Barcelona: Association for Computational Linguistics.

Liu, Y. et al., 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, s.l.: s.n.

Loshchilov, I. & Hutter, F., 2017. *Decoupled Weight Decay Regularization*, s.l.: s.n.

Macias, M., 2024a. *Fine-Tuning Conversational Large Languages Models for Moodle*. [Online] Available at: <https://github.com/nessa01macias/Fine-Tuning-Conversational-Large-Languages-Models-for-Moodle> [Accessed 30 04 2024].

Macias, M., 2024b. *Models & Datasets - Melany Macias, Hugging Face*. [Online] Available at: <https://huggingface.co/nessa01macias> [Accessed 30 04 2024].

Mesnard, T. et al., 2024. *Gemma: Open Models Based on Gemini Research and Technology*, s.l.: s.n.

Meta, 2024. *Introducing Meta Llama 3: The most capable openly available LLM to date*. [Online] Available at: <https://ai.meta.com/blog/meta-llama-3> [Accessed 02 05 2024].

Metropolia, 2023. *Sustainable future is built on competence and innovative solutions*. [Online] Available at: <https://www.metropolia.fi/kestavakehitys/en/Home> [Accessed 18 03 2024].

Microsoft, 2023. *Phi-2: The surprising power of small language models*. [Online] Available at: <https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/> [Accessed 29 04 2024].

Moodle, 2020. *Moodle - Docs*. [Online] Available at: https://docs.moodle.org/dev/Moodle_architecture [Accessed 20 04 2024].

Moodle, 2024a. *Moodle*. [Online] Available at: <https://stats.moodle.org/> [Accessed 20 03 2024].

Moodle, 2024b. *Plugins*. [Online] Available at: <https://moodle.org/plugins/?q=> [Accessed 20 03 2024].

Muniasamy, A. & Alasiry, A. D. L., 2020. Deep learning: The impact on future eLearning. *International Journal of Emerging Technologies in Learning (Online)*, 15(1), p. 188.

Nalli, G., Amendola, D., Perali, A. & Mostarda, L., 2021. Comparative Analysis of Clustering Algorithms and Moodle Plugin for Creation of Student

Heterogeneous Groups in Online University Courses. *Applied Sciences*, Volume 11, p. 5800.

Narayan, S., Cohen, S. B. & Lapata, M., 2018. *Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization*, s.l.: ArXiv.

Naudé, M., 2011. Sustainable development in companies: Theoretical dream or implementable reality?. *Corporate Ownership and Control*, 01, Volume 8, pp. 352-364.

Ng, A., 2017. *Why AI Is the New Electricity*. Stanford Graduate School of Business. [Online] Available at: <http://stanford.io/2mwODQU> [Accessed 05 03 2024].

Ojala, M., Cunsolo, A., Ogunbode, C. & Middleton, J., 2021. Anxiety, Worry, and Grief in a Time of Environmental and Climate Crisis: A Narrative Review. *Annual Review of Environment and Resources*, Volume 46, pp. 45-58.

OpenAI et al., 2024. *GPT-4 Technical Report*, s.l.: s.n.

Pallagani, V. et al., 2024. *On the Prospects of Incorporating Large Language Models (LLMs) in Automated Planning and Scheduling (APS)*, s.l.: s.n.

Papineni, K., Roukos, S., Ward, T. & Zhu, W. J., 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 07.p. 311–318.

PHP, 2024. *What is PHP?*. [Online] Available at: <https://www.php.net/manual/en/intro-what-is.php> [Accessed 12 03 2024].

Post, M., 2018. A Call for Clarity in Reporting {BLEU} Scores. In: O. Bojar, et al. eds. *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels: Association for Computational Linguistics, pp. 186-191.

Prime Minister's Office, 2019. *Valtioneuvosto*. [Online] Available at: <https://julkaisut.valtioneuvosto.fi/handle/10024/161601> [Accessed 19 04 2024].

Pwc, 2003. *Integral Business*. [Online] Available at: www.pwc.com/sustainability [Accessed 10 04 2024].

PyPDF2, 2008. *PyPDF2*. [Online] Available at: <https://pypdf2.readthedocs.io> [Accessed 15 04 2024].

Radford, A. et al., 2019. *Language Models are Unsupervised Multitask Learners*, s.l.: s.n.

Red Hat, 2019. *Red Hat*. [Online] Available at: <https://www.redhat.com/en/topics/open-source/what-is-open-source> [Accessed 12 03 2024].

Rumelhart, D., H. G. & Williams, R., 1986. Learning representations by back-propagating errors. *Nature*, Volume 323, p. 533–536.

Sætra, H., 2023. *Technology and Sustainable Development: The Promise and Pitfalls of Techno-Solutionism*. s.l.:Østfold University College.

Sáiz, M. C., Marticorena, R. & García, C. I., 2020. Monitoring Students at the University: Design and Application of a Moodle Plugin. *Applied Sciences*, Volume 10, p. 3469.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. *Neural Networks*, Volume 61, pp. 85-117.

SDSN, 2012. *Sustainable Development Solutions Network: A Global Initiative for the United Nations*. [Online] Available at: <https://www.unsdsn.org/> [Accessed 20 04 2024].

spaCy, 2024. *spaCy*. [Online] Available at: <https://spacy.io/> [Accessed 20 04 2024].

The Mistral AI Team, 2024. *Hugging Face: mistralai/Mistral-7B-Instruct-v0.2*. [Online] Available at: <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2> [Accessed 20 04 2024].

Touvron, H. et al., 2023. *LLaMA 2: Open and Efficient Foundation Language Models*, s.l.: s.n.

Tunstall, L., Werra, L. & Wolf, T., 2022. *Natural language processing with transformers*. Revised Edition ed. s.l.:O'Reilly Media, Inc.

UN, 2012. *United Nations Department of Economic and Social Affairs: Sustainable Development*. [Online] Available at: <https://sdgs.un.org/> [Accessed 19 04 2024].

United Nations, 2024. *Department of Economic and Social Affairs: Sustainable Development*. [Online] Available at: <https://sdgs.un.org/goals> [Accessed 01 04 2024].

Vaswani, A. et al., 2017. *Attention Is All You Need*, s.l.: s.n.

Wolf, T. et al., 2020. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*, s.l.: s.n.

Yeh, C. et al., 2021. *SustainBench: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning*, s.l.: s.n.

Zawacki-Richter, O., Marín, V., Bond, M. & Gouverneur, F., 2019. Systematic review of research on artificial intelligence applications in higher education – where are the educators?. *Int J Educ Technol High Educ*, Volume 16, p. 39.

Zhao, H. et al., 2024. Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.*, April.15(2157-6904).