MASTER'S THESIS

EXPLORATORY STUDIES LAIA@UIB: PRE-TRAINED NATURAL LANGUAGE PROCESSING MODELS TO RESPOND TO DOMAIN-SPECIFIC INPUTS.

Lunara Nurgaliyeva

Master's Degree in Intelligent Systems (MUSI)

Specialisations: Computer Vision, Mobile Robotics

Centre for Postgraduate Studies

Academic Year 2022-23

EXPLORATORY STUDIES LAIA@UIB: PRE-TRAINED NATURAL LANGUAGE PROCESSING MODELS TO RESPOND TO DOMAIN-SPECIFIC INPUTS

Lunara Nurgaliyeva

Master's Thesis
Centre for Postgraduate Studies
University of the Balearic Islands

Academic Year 2022-23

Key words:
LLM, BLOOM, fine-tuning, question answering, FAQ, PEFT

Thesis Supervisor's Name: Antoni Jaume
Thesis Supervisor's Name: Isaac Lera

Exploratory Studies LAIA@UIB: Pre-trained Natural Language Processing models to respond to domain-specific inputs

Lunara Nurgaliyeva
Tutors: Antoni Jaume Capó, Isaac Lera Castro
Trabajo de fin de Máster Universitario en Sistemas Inteligentes (MUSI)
Universitat de les Illes Balears
07122 Palma, Illes Balears, Espanya
lunara.nurgaliyeva1@estudiant.uib.cat

ABSTRACT

With the rapid development of conversational agents, there is a need to create models that can understand human requests and respond effectively to them. This thesis explores the process of fine-tuning a pre-trained BLOOM language model with 3 billion parameters for use in a chatbot question-answering system. The study was conducted using both quantitative and qualitative assessment methods, including analysis of keyword relevance, text readability, and sentiment analysis of model responses. Methods for optimizing memory use for loading and configuring the model were also discussed and were found to be effective in saving resources. The results of fine-tuning the BLOOM model are promising in terms of relevance, readability and tone of responses. In summary, this study demonstrates the potential of fine-tuning large pre-trained models such as BLOOM to create effective question-answering systems in chatbots, but more research is needed to further adapt to different contexts and refine aspects of model behavior.

***Index Terms*—LLM, BLOOM, fine-tuning, question answering, FAQ, PEFT**

I. INTRODUCTION

Natural Language Processing (NLP) has become a key technology in the current rapidly changing digital era revolutionizing how computers can understand and interact with human language. Pre-trained language models now can provide remarkable abilities in a wide range of language comprehension tasks. In particular, large amount of attention and recognition were gained by nowadays popular GPT-3 (Generative Pre-trained Transformer), BERT (Bidirectional Encoder Representations from Transformers), and DistilBERT (Distilled BERT), for their capabilities to learn from extensive amount of text data and capture complicated contextual information, which led to an increase in research exploring how to utilize these pre-trained models for applications in specific domains.

The main objective of this thesis is to fine-tune smaller version of pre-trained BigScience Large Open-science Open-access Multilingual Language Model (BLOOM) for one of the downstream tasks, specifically question answering, which in the future can be efficiently utilized as a chatbot for University of Balearic Islands (UIB) website to answer Frequently Asked Questions (FAQs). The motivation behind this research arises from the need to enhance the efficiency and automation of UIB website's FAQ section to some extent. As the university's online platform is one of the primary information sources, it is crucial to ensure that users are able to quickly obtain answers to their questions. However, given that there can be a high volume of questions from prospective and current students, parents, and visitors, it might be difficult to handle them manually. This can be enhanced by an automated question-answering system, based on advanced NLP technology that is capable of providing information while maintaining high level accuracy and consequently providing a great user experience along with promotion of accessibility and support for the university community. In addition to that, it might be quite challenging to deploy pre-trained models as fully functionable chatbot considering significant computational power and memory, resources that may not be readily available in many practical applications, including running the FAQ section on a university server. Therefore, the research also covers various techniques, specifically Parameter-Efficient Fine-Tuning (PEFT) and QLoRA, that are implemented to significantly reduce the computational load of fine-tuning and deploying a pre-trained model while maintaining its efficacy.

To achieve the stated objective, several crucial tasks will be undertaken. Firstly, a comprehensive dataset of different questions and their corresponding answers and relative context will be preprocessed to serve as the foundation for fine-tuning a smaller version of the BLOOM model. Subsequently, the pre-trained BLOOM model will be fine-tuned using above-mentioned approaches after which it will be rigorously evaluated using qualitative analysis and various quantitative metrics, specifically accuracy, readability score and sentiment analysis.

In conclusion, this research aims to demonstrate the effective way of using pre-trained NLP models for specialized question answering tasks, specifically in the context of enhancing the FAQ section on the university website in terms of intelligence and responsiveness.

II. BACKGROUND

The transformer architecture presented in paper "Attention is All You Need" represents a significant breakthrough in the NLP field. Unlike previous approaches based on recurrent or convolutional neural networks, the transformer uses an "attention" mechanism to establish global connections between inputs and outputs [21]. This allows to process information in parallel, which leads to a significant increase in speed and performance. The essence of the transformer model is the presence of two key components: an encoder that processes the input data, and a decoder responsible for generating the output data [21]. This architecture facilitates a variety of NLP tasks, including machine translation and text summarization, and is the foundation for many current state-of-the-art models in this field.

Large language models (LLMs) are a class of machine learning models that specialize in understanding and generating human-readable text. These models use the transformer architecture and are trained on large corpus, massive dataset, having up to hundreds of billions of parameters, hence making them incredibly versatile and capable. One of such autoregressive LLMs similar to GPT3 is BLOOM, the world's largest open multilingual language model with a 176 billion parameters trained on 46 natural languages and 13 programming languages that was developed and released by a collaboration of hundreds of researchers using industrial-scale computational resources, including Jean Zay supercomputer [12]. Since in the scope of this thesis it is nearly impossible and infeasible to fine-tune a model that was trained using extensive amounts of resources, one of the several smaller versions of the models that have been trained on the same dataset was used, specifically BLOOM-3B, a BLOOM model with 3 billion parameters.

BLOOM-3B is a Transformer-based Language Model, consisting of a decoder-only structure, architecture of which includes an input embeddings layer, transformer blocks, and an output language-modeling layer [6]. Even though BLOOM-3B is one of the small versions, it still comes with a set of challenges similar to bigger pre-trained models. It requires a considerable amount of computational resources, for example, it is needed 20GB RAM to load the BLOOM-3B model to use it for inference. not mentioning the fine-tuning, storing and deploying the model, since the fine-tuned models are the same size as the original pre-trained model [6].

To address and overcome these challenges, the researchers have come up with an approach known as Parameter-Efficient Fine-tuning (PEFT). PEFT allows fine-tuning of pre-trained models in a way that maintains their efficacy comparable to full fine-tuning while only having a small number of trainable parameters thereby significantly reducing the computational and storage costs [9]. Moreover, PEFT also addresses the problem of "catastrophic forgetting", where a model trained on one task, and then trained on a second task, "forgets" how to perform the first task [3].

Among the several methods of PEFT currently supported by HuggingFace, Low-Rank Adaptation (LoRA), proposed by Microsoft researchers, has shown remarkable promise. According to the results of their empirical experiments, in comparison with GPT-3 175B fine-tuned with Adam, LoRA is capable of decreasing the number of trainable parameters by 10,000 times and the requirement of GPU memory by 3 times, which makes the performance of LoRA on-par or better than fine-tuning in model quality on RoBERTa or GPT-3, despite the fact of having fewer number of trainable parameters, a higher training throughput, and no additional inference latency [5]. Main idea of the LoRA is to fine-tune high-parameter models like GPT-3 by introducing low-rank matrices that substantially reduce the overall number of parameters that need to be adjusted during the fine-tuning process. In greater detail, during fine-tuning, changes in weights are stored in two decomposed matrices instead of in one large matrix, for example, 1000×1000 matrix can be decomposed into 1000×n and n×1000 matrices, where n is the rank of a matrix, number of linearly independent rows/columns in a matrix and lower rank can be chosen based on the downstream task [5]. The reason behind this is that LLMs in general capture a wide range of features to be versatile enough for different tasks with decent accuracy, however, when these models are tailored to a specific task, such as question answering, only fewer numbers of features are needed to be considered.

Further advancements led to the recent development of Quantized Low-Ranking Adaptation (QLoRA) technique, which extends the benefits of LoRA and helps to avoid the problems regarding time complexity and memory during the model loading as well, by adding an extra layer of efficiency: the quantization of weight values to compress a pre-trained model. In particular, this is done by converting high-resolution data types, 32 bits (FP32), in which elements of the weight matrices are stored, to lower-resolution ones, 4 bits or 8 bits [2]. It is important to mention that during fine-tuning, QLoRA employs backpropagation only through the LoRa adapters, while keeping, for instance, the 4-bit quantized pre-trained language model parameters unchanged, which means that only during training of LoRa adapters the model weights are returned to FP32 [2]. This contributes to a great decrease of the memory requirements and speed up of calculations. According to the results obtained by HuggingFace, they could achieve a gain of 1.65x and 1.96x memory footprint for the BLOOM model with 3 billion and 176 billion parameters respectively [1, 11].

A. Analysis of the State-Of-The-Art

For LLM to perform a specific task, such as question answering, it must be fine-tuned to this particular task. According to HuggingFace, there are different question answering variants based on the inputs and outputs [8] :

- **Extractive**: The model extracts the answer from a context.
- **Open Generative**: The model generates free text directly based on the context.
- **Closed Generative**: There is no context and the answer is completely generated by a model.

The HuggingFace's transformer library provides AutoModelForQuestionAnswering class, which allows to use and fine-tune any LLM, including BLOOM for extractive question answering. The problem with using this class or ready pipeline is that the answers must be an explicit part of the provided context. While in the case of building an automated FAQ system, it should be able to generate answers from the context, that is, to be open generative question answering. Majority of the current state-of-the-art are focused on the optimization of extractive question-answering systems. In the paper [22], authors train the BioBERT model for question answering for specific domains of biomedical questions, which is limited to factoid, list, and yes/no type questions that require concise answers of one-two words phrases. Moreover, in a recent paper [13], authors are concentrating on building a question-answering system for a low-resource language, Vietnamese using Intent Classification model and the Machine Reading Comprehension model. Despite the encouraging results, the system is based on extractive question answering, providing only short answers instead of full sentences. This thesis aims to address building of question-answering system that can provide comprehensive and informative answers in the form of full sentences while also extracting information from the context.

III. EXPERIMENT

A. Dataset

Considering severe limitations in the availability and small size of FAQ data from the UIB website and the fact that a reliable and extensive dataset is needed to effectively fine-tune a model, it was decided to fine-tune BLOOM model on SQuAD version 2.0 dataset provided by Stanford University. This dataset is large and diverse enough, containing over 100,000 questions, including FAQs from other universities, where each question has a corresponding piece of text that is the answer [4]. This richness and variety of question-answer pairs provide a solid foundation from which to fine-tune the model and enable it to better understand and respond to a wider range of queries.

B. Preprocessing

To prepare the dataset for use as the input to the model for specific task of question answering, each record containing context, question, and answer were preprocessed and converted into a structured string. In cases where there is no answer to a specific question in the dataset, "Can not find answer" placeholder was inserted. This strategy prepares the model for real-life scenarios where some questions may remain unanswered, which increases its adaptability and robustness. After the content has been formatted, a tokenization process converted the text data into a numeric format that is easily interpreted by the machine learning mode, which allows the model to work effectively with the data.

C. Modelling

In order to optimize the training step in terms of speed and memory efficiency, BLOOM-3B model was loaded in 4 bit using NF4 quantization with double quantization and with the compute dtype bfloat16. Specifically speaking, bitsandbytes library from transformers was leveraged to set above-mentioned parameters during loading of the model. According to theoretical considerations and empirical results from the paper, using NF4 quantization is recommended for better performance [2]. Moreover, using 16-bit compute dtype makes the matrix multiplication and training faster (default torch.float32) [2]. It is also essential to note that double quantization helps to enhance the memory efficiency without significantly sacrificing the model's performance.

Following these optimizations, the model was prepared for fine-tuning using the PEFT technique, LoRA. The LoRA configuration defines several important parameters, such as rank (r=8) and target modules in the model architecture that are fine-tuned (for example, "query_key_value"). Task-specific variables were also set, such as dropout rate and task type, in this case, Causal Language Modeling (CAUSAL_LM). Enabling LoRA with a selected rank further reduces the number of trainable parameters. This ensures that only a small portion of the overall model parameters are updated during fine-tuning, reducing computational overhead. This allows the model to specialize in answering frequently asked questions while retaining general knowledge from a pre-trained base. After applying the LoRA method, it resulted only in 13% of trainable parameters in the modified model.

Thus, these methodological decisions, including 4-bit and double quantization, as well as low-rank adaptation, made the fine-tuning process efficient in terms of the number of parameters and computations.

D. Training

The model was trained in small batches, since it prevents system memory overload, especially when working with larger models. However, to compensate for the limitations of small batch sizes, gradient accumulation was used. This method merges gradients in multiple forward and backward passes before updating, effectively simulating a larger batch size without additional memory overhead. The learning rate, that is, the rate at which the model adjusts its internal parameters, was set to a moderate level, which allows to carefully tune the model, avoiding rapid and unstable changes to its architecture. Likewise, a warm-up period was used at the beginning of the learning process to allow the learning rate to gradually increase, avoiding sudden jumps that could negatively impact the fine-tuning process. The training process used 16-bit numerical precision instead of the standard 32-bit, which reduced memory usage without significant loss in model training quality. The custom AdamW optimizer was also used to further reduce memory requirements. Consequently, less than 15GB of memory was required to fine-tune the model with 3 billion parameters on the dataset of size 100000.

Overall, the provided methodology aims to adapt a high-performing, pre-trained language model to a specific task where the selected training parameters and techniques reflect a carefully considered strategy to achieve balancing computational efficiency with model performance.

E. Inference

The saved fine-tuned model [10] was loaded using 8 bit quantization for inference [7]. To evaluate the model, 11 questions with answers were collected from the FAQs section of UIB website. Since the collected answers were large and informative enough, they were used as the context related to the question, and other corresponding answers, that were used as ground-truth answers, were manually generated. The inference included passing a pair of context and question to a fine-tuned model to generate answer with following preprocessing of generated answer.

F. Evaluation metrics

The answers generated by the fine-tuned model were evaluated against manually generated expected, human-like answers, using following quantitative metrics: keywords matching, readability score and sentiment analysis.

- **Keyword Matching** is used to check the relevance. The number of relevant keywords that appear in both the expected and model-generated answers are counted, where a higher match indicates better relevance. In addition, to check the relevance with the context the number of matching keywords are counted in both answer and context. This is done by tokenizing the texts and removing stopwords, then by counting the number of matching keywords between the two texts.
- **Readability Score** is calculated using the Flesch Reading Ease formula, where the higher the score means that it is easier to understand the text of the answer.
- **Sentiment Analysis** checks the tones (positive, neutral, negative) of answers. It is done by using TextBlob to give a polarity score. The score ranges from -1 to 1, where a score closer to 1 indicates a positive sentiment and closer to -1 indicates negative.

IV. RESULTS AND DISCUSSION

Part of the questions with corresponding context as well as the model-generated responses and expected human-like responses used for model evaluation are presented in table I with additional 4th query for further analysis. Remaining part of the of the questions with corresponding context as well as the model-generated responses and expected human-like responses used for model evaluation are presented in table III and IV. Additional queries used for qualitative analysis are depicted in the table II.

A. Quantitative analysis

From figure 1, it can be observed that keyword match scores between model-generated responses and expected human responses ranged from 2 to 23. Given the highest score of 23, the model can perform very well in some cases. Low scores can be explained by the presence of additional information in the responses generated by the model. For example, in the case of the answer in the 1st query in table I, it can be seen that the additional sentences from the context were provided. Moreover, even though in the case of the 2nd query in table

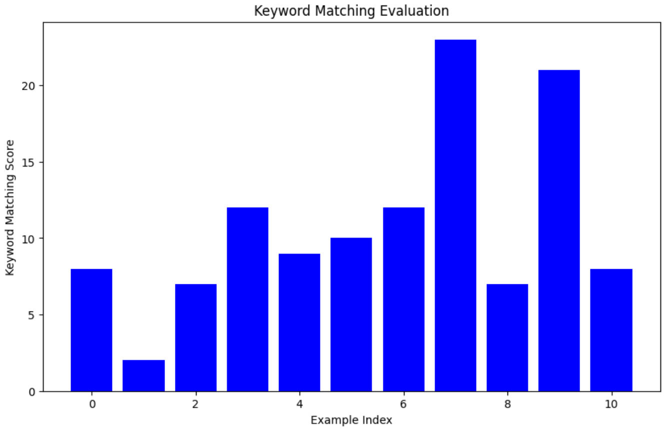

Figure 1. Keyword Matching scores of model-generated answers with human-like answers

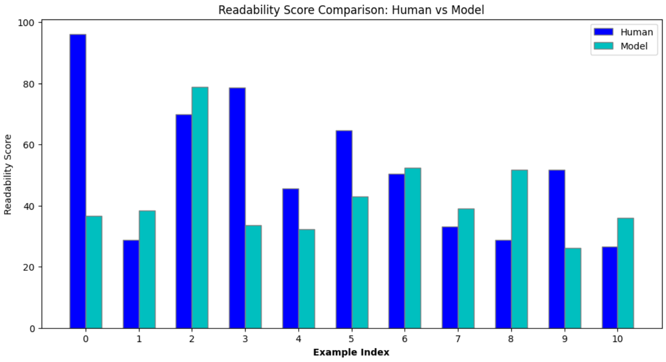

Figure 2. Readability Score comparison between model-generated and human-like answers

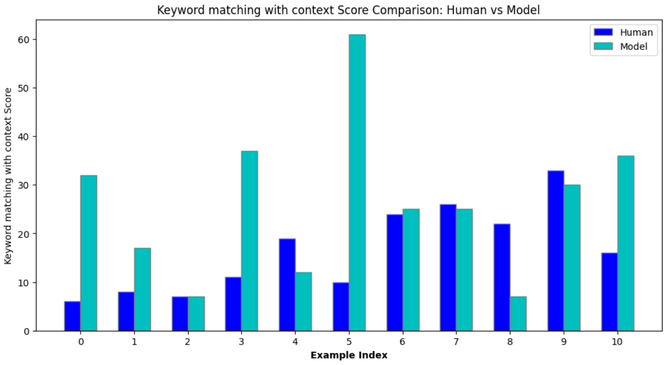

Figure 3. Keyword and context matching score comparison between model-generated and human-like answers

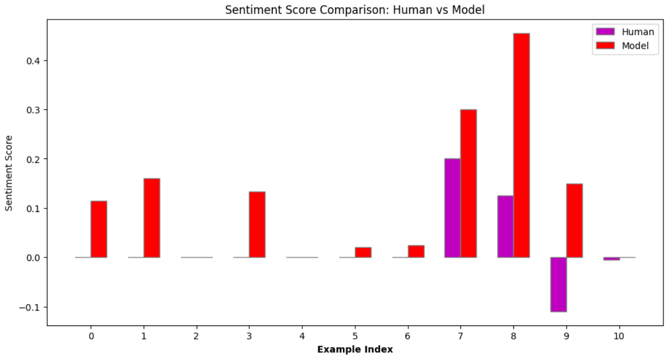

Figure 4. Sentiment analysis score comparison between model-generated and human-like answers

I the keyword match score was 2, the answers in both cases can be considered as the right answer.

Readability scores for responses generated by the model ranged from 26.14 to 78.75, while for human-like responses, from 26.51 to 96.18 (figure 2). On average, human-like responses were more likely to receive higher readability scores, however, in half of the cases model-generated answers were easier to read compared to human-like responses. Looking more specifically, in the 1st query in table I, there is a huge difference in readability score, 96.18 for human-like answer and 36.63 for model-generated answer, but as it was mentioned before, model-generated answer included extra information from context, being almost exact as the context. This means that readability of the answer can directly depend on the readability of the context in case of model-generated answer.

Considering the keyword match score between the context and the responses, the match scores of responses generated by the model ranged from about 7 to 61 and expected human responses ranged from 6 to 33, which can be observed from figure 3. This indicates that, on average, model responses included a higher proportion of relevant keywords compared to expected human responses. In the case of 5th query in figure 3 (3rd query in table III), the model response was highly similar to the context compared to expected human answer, where each of them have keyword matching with context score of 61 and 10 respectively, which means that model can generate very accurate results. The lower scores for model-generated answers can be explained by the fact that model is fine-tuned for open generative question answering meaning that it is possible to generate additional or more general information.

Figure 4 represents that sentiment analysis scores for model-generated responses ranged from 0 to 0.455, while all but four (two positive and two negative) expected human-like responses scored 0. This suggests that model-generated responses were generally more neutral and positive.

Overall, the general finding is that model-generated responses were generally more effective in terms of keyword matching and readability, which are important aspects of effective communication, which means that model-generated responses were not much inferior and in most cases even superior to human responses. Moreover, sentiment analysis showed that the responses generated by the model were generally more positive, which can be useful in customer service scenarios, for example, in terms of increased user satisfaction.

B. *Qualitative analysis*

In addition to the quantitative evaluation that provided valuable insights in terms of keyword matching, readability, and sentiment analysis, a qualitative examination of the model's behavior is essential for a comprehensive understanding of its performance.

It was found that the model can generate general answers, as in the case of the 3rd query in table I. Meanwhile, mentioning the specific information in both context and question for example, university name, results in accurate response which is the case in the 4th query in table I. However, this property can be very useful in terms of providing general answers which are not limited by the context. Taking as an example 1st query in table II, where model successfully provided information about the location of UIB even with non-related context, demonstrates that the model was able to answer questions accurately without having the right context.

Moreover, since the model is trained on multiple languages, it was also discovered that the model was capable of providing responses in English to questions asked in different languages (2nd query in table II). This is highly important in providing good user experience, where users can ask questions in different languages.

However, it was also found that sometimes model can generate not very accurate results, as in the case of 7th query in table IV, where instead of "40 percent of the credits" model interpreted this as "40 credits".

Summarizing qualitative analysis, model possesses various capabilities and is able to provide complete answers whether it is related to context or general information. But in some cases it can have problems with being precise and accurate regarding details mentioned in the context, which means that the model requires further improvements.

V. CONCLUSION

Taking all things into consideration, the study focuses on the effective fine-tuning of the BLOOM model, a pre-trained language model with 3 billion parameters, for question answering tasks applied in the context of chatbots. The fine-tuned model was evaluated through rigorous set of both qualitative and quantitative measures such as keyword relevance, readability, and sentiment analysis. The study successfully fine-tuned the BLOOM-3B model for specialized question answering tasks, demonstrating its adaptability and potential for specialized applications. This customization process tailored the model to the specific requirements of the question-answering chatbot. The thesis also introduced methods PEFT techniques, notably LoRA and QLoRA, for efficient use of memory to load and fine-tune the model, making it more accessible to various devices and applications. Despite the limitations regarding accuracy of specific details provided in the model-generated answers, the study revealed interesting aspects of the model's behavior, indicating its adaptability and potential for further

Table I
RESULTS OF MODEL INFERENCE

	Context	**Question**	**Model-generated answer**	**Expected human-like answer**
1	Current fees are around 260 euro per year, of which 210 euro are for academic tutoring and the remainder for admin fees and medical insurance in the event of a school-related accident. For more information, please see the following section: 'Doctoral School > Useful Information > Regulations > Decree on Public Fees' [17].	What's the total cost per year?	The total cost per year is around 260 euro per year, of which 210 euro are for academic tutoring and the remainder for admin fees and medical insurance in the event of a school-related accident. For more information, please see the following section: 'Doctoral School > Useful Information > Regulations > Decree on Public Fees'.	The total cost per year is around 260 euro.
2	Current fees are around 260 euro per year, of which 210 euro are for academic tutoring and the remainder for admin fees and medical insurance in the event of a school-related accident. For more information, please see the following section: 'Doctoral School > Useful Information > Regulations > Decree on Public Fees' [17].	What are admin fees for?	The administration fees for the doctoral school are set by the Decree on Public Fees. For more information, please see the following section: 'Doctoral School > Useful Information > Regulations > Decree on Public Fees'.	The admin fees are for administrative services and medical insurance in the event of a school-related accident.
3	Following information is regarding PhD program of University of Balearic Islands: Current fees are around 260 euro per year, of which 210 euro are for academic tutoring and the remainder for admin fees and medical insurance in the event of a school-related accident. For more information, please see the following section: 'Doctoral School > Useful Information > Regulations > Decree on Public Fees' [17].	How much does a PhD cost?	The cost of a PhD varies depending on the country and the university. The average cost of a PhD in the United States is around $60,000	The cost for the PhD program at the University of Balearic Islands is around 260 euro per year.
4	Following information is regarding PhD program of University of Balearic Islands: Current fees are around 260 euro per year, of which 210 euro are for academic tutoring and the remainder for admin fees and medical insurance in the event of a school-related accident. For more information, please see the following section: 'Doctoral School > Useful Information > Regulations > Decree on Public Fees' [17].	How much does a PhD cost at University of Balearic Islands?	The cost of a PhD program at University of Balearic Islands is around 260 euro per year, of which 210 euro are for academic tutoring and the remainder for admin fees and medical insurance in the event of a school-related accident. For more information, please see the following section: 'Doctoral School > Useful Information > Regulations > Decree on Public Fees'.	The cost for a PhD program at the University of Balearic Islands is around 260 euro per year.

research and tuning. As a further research, experimentation with different hyper-parameters during fine-tuning and inference, as well as different aspects of prompt engineering can be suggested. Moreover, various patterns or common issues can identified to address the issues with accuracy of specific information provided in the model-generated responses.

Overall, this thesis makes significant contributions to the field of fine-tuning large language models for specific applications, and aspires to set a precedent for how large, sophisticated NLP models can be adapted for domain-specific challenges in a computational and resource-efficient manner. This study lays the foundation for further research and practical implementation in the rapidly developing field of chatbot technology.

VI. APENDICES

Additional results of model inference used for evaluation can be observed in table III and IV.

REFERENCES

[1] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, 2022.

[2] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

[3] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

[4] S. N. Group et al. Squad 2.0: The stanford question answering dataset. *Retrievable at https://rajpurkar. github. io/SQuAD-explorer. Accessed on Jan 8th*, 2022.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[6] HuggingFace. Bigscience/bloom-3b, n.d. https://huggingface.co/bigscience/bloom-3b. Accessed: 15.06.2023.

[7] HuggingFace. Efficient inference on a single gpu, n.d. https://huggingface.co/docs/transformers/perf_infer_gpu_one. Accessed: 13.08.2023.

[8] HuggingFace. What is question answering?, n.d. https://huggingface.co/tasks/question-answering. Accessed: 20.06.2023.

[9] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[10] L. Nurgaliyeva. Fine-tuned bloom-3b, 2023. https://huggingface.co/aranulunara/outputs. Accessed: 27.08.2023.

[11] A. Pagnoni. Artidoro/qlora: Qlora: Efficient finetuning of quantized llms, May 23 2023. https://github.com/artidoro/qlora. Accessed: 22.08.2023.

[12] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[13] P. Tran, D. Nguyen, H.-A. Tran, T. Nguyen, and T. Tran. Building a closed-domain question answering system for a low-resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(3):1–14, 2023.

[14] University of the Balearic Islands. Application, access and admission, 2023. https://cep.uib.eu/en/Com_hi_puc_accedir/Preinscripcio/. Accessed: 17.06.2023.

[15] University of the Balearic Islands. Do i need to take face-to-face classes?, 2023. https://edoctorat.uib.eu/en/Gestio_Doctorands/FAQs/Detall/Do-I-need-to-take-face-to-face-classes.cid655900. Accessed: 17.06.2023.

[16] University of the Balearic Islands. How are thesis tutors and supervisors assigned?, 2023. https://edoctorat.uib.eu/en/Gestio_Doctorands/FAQs/Detall/How-are-thesis-tutors-and-supervisors-assigned.cid655901. Accessed: 17.06.2023.

[17] University of the Balearic Islands. How much does a phd cost?, 2023. https://edoctorat.uib.eu/en/Gestio_Doctorands/FAQs/Detall/How-much-does-a-PhD-cost.cid655824. Accessed: 17.06.2023.

[18] University of the Balearic Islands. What does a phd comprise?, 2023. https://edoctorat.uib.eu/en/Gestio_Doctorands/FAQs/Detall/What-does-a-PhD-comprise.cid655829. Accessed: 17.06.2023.

[19] University of the Balearic Islands. What is the difference between an official and a specific university master's degree?, 2023. https://cep.uib.eu/en/Te-podem-ajudar/Preguntes-frequents/Detall/What-is-the-difference-between-an-official-and.cid628398. Accessed: 17.06.2023.

[20] University of the Balearic Islands. What is the minimum number of credits i must pass?, 2023. https://cep.uib.eu/en/Te-podem-ajudar/Preguntes-frequents/Detall/What-is-the-minimum-number-of-credits-I-must-pass.cid628399. Accessed: 17.06.2023.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[22] W. Yoon, J. Lee, D. Kim, M. Jeong, and J. Kang. Pre-trained language model for biomedical question answering. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 727–740. Springer, 2019.

Table II
RESULTS OF MODEL INFERENCE WITH NON-RELATED QUESTIONS

	Context	**Question**	**Model-generated answer**
1	On a PhD programme at University of Balearic Islands, there are no classes per se, although depending on the programme where you are registered, there may be some face-to-face activities.There may be two types of activities: Cross-cutting: the academic commissions for each PhD programme decide on whether these activities are obligatory or elective for students. They are organised by the Doctoral School and are common to all PhD programmes. For more information, please see the Doctoral School Training Activities. Specific: the academic commissions for each PhD programme decide on whether these activities are obligatory or elective, and whether they are specific to their students. For more information about these activities, please see the PhD studies on offer section. You will need to select your PhD programme and go to the ’Training Activities’ section. If you have any academic questions, you may contact the programme coordinator via the contact form in the ’General Information’ section for the programme [15].	Where is University of Balearic Islands located?	University of Balearic Islands is located in Palma de Mallorca, Spain.
2	At the University of Balearic Islands, official master’s degrees: their clear and explicit designation is ’Master’s Degree in...’. These programmes aim for students to acquire advanced, specialised or multidisciplinary academic training geared towards an academic or professional specialisation, or to promote the start of research activity. These degrees are official and valid across Spain, with full academic recognition, and enable students to work in regulated professions in accordance with applicable regulations in each instance. Passing an official master’s programme grants you the right to obtain the degree certificate with the specific RUCT designation (University and Degree Registry at the Ministry of Education, Culture and Sport). Unofficial master’s degrees: these programmes may not use the designation ’Master’s Degree inâ€¦’. They are included in the postgraduate programme selection, are UIB-specific and represent an additional academic option. These master’s programmes and other UIB-specific courses (University Specialist/Expert) enable the university to provide an agile quality response to academic requirements, as well as refresher courses for students (for society in general and for employment opportunities). They are an essential element in our education system and provide top-flight and necessary training [19].	¿Cuál es la diferencia entre un máster oficial y no oficial en la Universidad de las Islas Baleares?	The difference between an official and an unofficial master’s degree is that the former is a master’s degree that is recognised by the RUCT (University and Degree Registry at the Ministry of Education, Culture and Sport). The latter is a master’s degree that is not recognised by the RUCT. These master’s programmes and other UIB-specific courses (University Specialist/Expert) enable the university to provide an agile quality response to academic requirements, as well as refresher courses for students (for society in general and for employment opportunities). They are an essential element in our education system and provide top-flight and necessary training.

Table III
RESULTS OF MODEL INFERENCE

	Context	**Question**	**Model-generated answer**	**Expected human-like answer**
1	Current fees are around 260 euro per year, of which 210 euro are for academic tutoring and the remainder for admin fees and medical insurance in the event of a school-related accident. For more information, please see the following section: ’Doctoral School > Useful Information > Regulations > Decree on Public Fees’ [17].	How much for academic tutoring?	The academic tutoring fee is 210 euro per year. The fee is paid in advance and is not refundable	The cost for academic tutoring is 210 euro per year.
2	Following information is regarding PhD program of University of Balearic Islands: Current fees are around 260 euro per year, of which 210 euro are for academic tutoring and the remainder for admin fees and medical insurance in the event of a school-related accident. For more information, please see the following section: ’Doctoral School > Useful Information > Regulations > Decree on Public Fees’ [17].	Any fees for accidents at PhD program of University of Balearic Islands?	The University of Balearic Islands has a doctoral school, which is a public institution. The doctoral school is a public institution, and therefore, it is not possible to provide information about the fees for accidents at the doctoral school. However, the doctoral school has a doctoral accident insurance, which covers the costs of medical treatment and rehabilitation of students who suffer an accident during their studies.	Yes, in the PhD program of the University of Balearic Islands, a portion of the fees is allocated for medical insurance in the event of a school-related accident. This is included in the remainder of the 260 euro annual fee that is not covered by the 210 euro for academic tutoring.
3	In this instance, the originals must be submitted between 1st November and 31st March of the current year. Until original documents are submitted, the CEP is unable to issue any academic certificate for the master’s programme, including the final degree certificate. The original and a photocopy must be submitted to the admin services in the Antoni Maria Alcover i Sureda building (Centre for Postgraduate Studies). You must request an appointment to do this step. Where documents are sent by post, a certified copy of each document must be included. The only stamps accepted are those issued by a registered Spanish notary public, a diplomatic service or the issuing body for the document, as long as payment for the certified stamp is accredited [14].	When do the originals need to be submitted?	The originals must be submitted between 1st November and 31st March of the current year. Until original documents are submitted, the CEP is unable to issue any academic certificate for the master’s programme, including the final degree certificate. The original and a photocopy must be submitted to the admin services in the Antoni Maria Alcover i Sureda building (Centre for Postgraduate Studies). You must request an appointment to do this step. Where documents are sent by post, a certified copy of each document must be included. The only stamps accepted are those issued by a registered Spanish notary public, a diplomatic service or the issuing body for the document, as long as payment for the certified stamp is accredited.	The originals must be submitted between 1st November and 31st March of the current year.
4	On a PhD programme at University of Balearic Islands, there are no classes per se, although depending on the programme where you are registered, there may be some face-to-face activities.There may be two types of activities: Cross-cutting: the academic commissions for each PhD programme decide on whether these activities are obligatory or elective for students. They are organised by the Doctoral School and are common to all PhD programmes. For more information, please see the Doctoral School Training Activities. Specific: the academic commissions for each PhD programme decide on whether these activities are obligatory or elective, and whether they are specific to their students. For more information about these activities, please see the PhD studies on offer section. You will need to select your PhD programme and go to the ’Training Activities’ section. If you have any academic questions, you may contact the programme coordinator via the contact form in the ’General Information’ section for the programme [15].	Do I need to take face-to-face classes on a PhD programme at University of Balearic Islands?	Yes, you will need to take face-to-face classes on a PhD programme at University of Balearic Islands. You will need to select your PhD programme and go to the ’Training Activities’ section. If you have any academic questions, you may contact the programme coordinator via the contact form in the ’General Information’ section for the programme.	No, on a PhD programme at the University of Balearic Islands, there are no classes per se. However, depending on the programme where you are registered, there may be some face-to-face activities. These activities could be either cross-cutting or specific to your programme, and the academic commissions for each PhD programme will decide whether these activities are obligatory or elective for students.

Table IV
CONTINUE OF RESULTS OF MODEL INFERENCE

	Context	**Question**	**Model-generated answer**	**Expected human-like answer**
5	At University of Balearic Islands Both the thesis tutor and supervisor are assigned by the academic commission. Your pre-registration application must include a summary of your research project and area. You will also need to make your own thesis supervisor proposal, which will be taken into account when assigning your thesis supervisor(s) [16].	How are thesis tutors and supervisors assigned at University of Balearic Islands?	The thesis tutor and supervisor are assigned by the academic commission. Your pre-registration application must include a summary of your research project and area. You will also need to make your own thesis supervisor proposal, which will be taken into account when assigning your thesis supervisor(s).	At the University of Balearic Islands, both the thesis tutor and supervisor are assigned by the academic commission. While making your pre-registration application, you are required to include a summary of your research project and area. You also have the option to make your own thesis supervisor proposal, which will be taken into account when the academic commission is assigning your thesis supervisor(s).
6	At the University of Balearic Islands PhD candidates undertake and produce original research in any field of knowledge: PhD thesis. Doctorands must pass annual assessments throughout their enrolment period (initial and subsequent assessments) that serve to monitor their research [18].	What does a PhD comprise at the University of Balearic Islands?	The PhD is a research degree awarded by the University of Balearic Islands. It is a four-year degree that is awarded to candidates who have successfully completed a master’s degree and have demonstrated a high level of research competence	At the University of Balearic Islands, a PhD comprises undertaking and producing original research in any field of knowledge, culminating in a PhD thesis. Additionally, PhD candidates must pass annual assessments throughout their enrolment period. These assessments, which include initial and subsequent assessments, serve to monitor the progress of their research.
7	The University of Balearic Islands has established a general academic progress plan on a master’s programme, depending on whether students are full- or part-time: Full-time master’s students must pass at least 40 percent of their enrolled credits (rounded up) for the academic year, Part-time master’s students must pass at least 10 percent of their enrolled credits (rounded up) for the academic year, Where students only enrol for one subject, the aforementioned rules do not apply. More information is available at the following link: ’Information on Academic Regulations, Academic Progress and Attendance Requirements and the Minimum Number of Initial Enrolled Credits’ [20].	What is the minimum number of credits I must pass on a master’s programme to continue studying at the University of Balearic Islands?	The minimum number of credits required to continue studying at the University of Balearic Islands is 40 credits (rounded up) for full-time students and 10 credits (rounded up) for part-time students. More information is available at the following link: ’Information on Academic Regulations, Academic Progress and Attendance Requirements and the Minimum Number of Initial Enrolled Credits’.	The minimum number of credits you must pass to continue studying a master’s programme at the University of Balearic Islands depends on whether you are a full-time or part-time student: Full-time master’s students must pass at least 40 percent of their enrolled credits (rounded up) for the academic year. Part-time master’s students must pass at least 10 percent of their enrolled credits (rounded up) for the academic year. Note that if you only enroll for one subject, these aforementioned rules do not apply.
8	Official master’s degrees: their clear and explicit designation is ’Master’s Degree in...’. These programmes aim for students to acquire advanced, specialised or multidisciplinary academic training geared towards an academic or professional specialisation, or to promote the start of research activity. These degrees are official and valid across Spain, with full academic recognition, and enable students to work in regulated professions in accordance with applicable regulations in each instance. Passing an official master’s programme grants you the right to obtain the degree certificate with the specific RUCT designation (University and Degree Registry at the Ministry of Education, Culture and Sport). Unofficial master’s degrees: these programmes may not use the designation ’Master’s Degree inâ€¦’. They are included in the postgraduate programme selection, are UIB-specific and represent an additional academic option. These master’s programmes and other UIB-specific courses (University Specialist/Expert) enable the university to provide an agile quality response to academic requirements, as well as refresher courses for students (for society in general and for employment opportunities). They are an essential element in our education system and provide top-flight and necessary training [19].	What is the difference between an official and unofficial master’s degree?	Official master’s degrees are those that are officially recognised by the Spanish Ministry of Education, Culture and Sport. They are recognised by the RUCT (University and Degree Registry at the Ministry of Education, Culture and Sport) and are valid across Spain. Unofficial master’s degrees are those that are not officially recognised by the Spanish Ministry of Education, Culture and Sport. They are included in the postgraduate programme selection, are UIB-specific and represent an additional academic option. They are an essential element in our education system and provide top-flight and necessary training.	An official master’s degree at the University of Balearic Islands is nationally recognized and enables you to work in regulated professions. It also allows you to obtain a specific degree certificate. On the other hand, an unofficial master’s degree is specific to UIB and doesn’t offer national recognition or the same professional eligibility as an official master’s degree.