|  |  |  |
| --- | --- | --- |
| **A close up of a sculpture  Description automatically generated** | **УНИВЕРЗИТЕТ “Св. КИРИЛ И МЕТОДИЈ” - СКОПЈЕ**  **ФАКУЛТЕТ ЗА ЕЛЕКТРОТЕХНИКА И ИНФОРМАЦИСКИ ТЕХНОЛОГИИ** | **A blue and white logo  Description automatically generated** |

- **ПРОЕКТНА ЗАДАЧА** -

по предметот

**Интелегентни Агенти**

**Тема**

**Fine-Tuninng Phi-2 Model for Dialogue Summarization**

|  |  |
| --- | --- |
| Ментор: | Изработил: |
| Проф. Данијела Ефнушева | Андреј Вељаноски 2/2021  Кристијан Милошески 8/2021  Петар Цветаноски 175/2021 |

*Скопје, декември 2025*

Contents

[Abstract 3](#_Toc189509746)

[Introduction 3](#_Toc189509747)

[Materials and Methods 4](#_Toc189509748)

[Results 5](#_Toc189509749)

[Conclusion 6](#_Toc189509750)

[References 7](#_Toc189509751)

# Abstract

This study explores the fine-tuning of the Phi-2 large language model for dialogue summarization, employing parameter-efficient techniques to optimize performance while minimizing computational overhead. The research leverages Low-Rank Adaptation (LoRA) and Quantized LoRA (QLoRA) to refine model weights without excessive resource consumption. Using the DialogSum dataset, we systematically preprocess data, apply structured fine-tuning, and evaluate the outcomes using ROUGE metrics. The results indicate that QLoRA enables high-quality summarization on consumer-grade hardware, with performance comparable to traditional full fine-tuning. This work highlights the viability of efficient fine-tuning strategies in real-world NLP applications.

# Introduction

Large language models (LLMs) have revolutionized natural language processing (NLP) by achieving state-of-the-art performance across a variety of tasks. However, deploying these models efficiently remains a challenge due to their extensive computational and memory requirements. Fine-tuning, a process of adapting pre-trained models to specific tasks, plays a crucial role in optimizing their performance while maintaining efficiency.

This study focuses on fine-tuning Phi-2, a highly capable LLM, for the task of dialogue summarization. Dialogue summarization is a complex NLP task that requires models to extract relevant information from multi-turn conversations while preserving the context and meaning. Given the high-dimensional nature of LLMs, traditional fine-tuning approaches are computationally expensive and often infeasible for users with limited hardware resources.

To address these challenges, we employ parameter-efficient fine-tuning (PEFT) methods such as LoRA and QLoRA. These approaches allow for targeted adaptation of the model while significantly reducing the number of trainable parameters. By leveraging the DialogSum dataset, which provides structured dialogue summaries, we assess the effectiveness of these techniques in optimizing Phi-2 for summarization tasks. This paper provides a comprehensive analysis of our methodology, results, and implications for efficient model adaptation in NLP applications.

# Materials and Methods

For this study, Phi-2 was selected as the base model due to its efficiency and strong performance in natural language tasks. The model was loaded using the Hugging Face Transformers library, which allowed seamless integration into modern deep learning frameworks. To fine-tune Phi-2 for dialogue summarization, the DialogSum dataset was utilized. This dataset comprises structured dialogues, each paired with a concise human-written summary. The dataset was preprocessed by removing special characters, normalizing punctuation, converting text to lowercase, and applying tokenization techniques to fit the model’s input constraints. Additionally, data augmentation was employed through paraphrasing techniques to improve model generalization.

Fine-tuning was conducted using LoRA and QLoRA, two parameter-efficient methods designed to optimize large language models without excessive computational demands. LoRA introduced low-rank updates to selected model layers, preserving most of the pre-trained knowledge while adapting to the summarization task. QLoRA further optimized the process by applying 4-bit quantization, reducing the overall memory footprint and making fine-tuning feasible on consumer-grade hardware. To balance efficiency and performance, hyperparameters such as rank values, learning rates, and dropout rates were carefully tuned. Rank values ranged between 2 and 128, learning rates were dynamically adjusted using a scheduler, and dropout layers were included to prevent overfitting.

To evaluate the performance of the fine-tuned models, multiple metrics were employed. ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum) were used to assess summarization quality, while BLEU scores provided insights into translation-style consistency. Perplexity scores were also calculated to gauge the coherence of the generated summaries. Additionally, a human evaluation was conducted where annotators rated the summaries based on coherence, fluency, and relevance. The performance of the PEFT fine-tuned models was benchmarked against standard fine-tuning approaches and other summarization models such as BART and T5. Baseline comparisons were made against the pre-trained Phi-2 model without fine-tuning to establish the effectiveness of the adaptation techniques.

By implementing these methodologies, we aim to create a robust and scalable framework for efficiently fine-tuning Phi-2 while maintaining high summarization accuracy. The use of parameter-efficient techniques like LoRA and QLoRA demonstrates that large-scale model adaptation is feasible even in resource-constrained environments, paving the way for broader adoption of LLMs in dialogue-based NLP applications.

# Results

The performance of the fine-tuned model was evaluated using ROUGE metrics, specifically ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum. The results for the original model and the PEFT fine-tuned model are summarized as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Original Model** | **PEFT Model** | **Absolute Improvement** |
| ROUGE-1 | |  | | --- | | **0.3524** | | |  | | --- | | **0.3497** | | |  | | --- | | **-0.28%** | |
| |  | | --- | | ROUGE-2 |  |  | | --- | |  | | |  | | --- | | **0.1291** | | |  | | --- | | **0.1233** | | |  | | --- | | **-0.57%** | |
| |  | | --- | | ROUGE-L |  |  | | --- | |  | | |  | | --- | | **0.2632** | | |  | | --- | | **0.2769** | | |  | | --- | | **+1.37%** | |
| |  | | --- | | ROUGE-Lsum |  |  | | --- | |  | | **0.2614** | **0.2769** | **+1.55%** |

These results indicate that while ROUGE-1 and ROUGE-2 scores showed a slight decrease, the ROUGE-L and ROUGE-Lsum scores improved, suggesting better overall structural coherence and fluency in the generated summaries. This improvement in ROUGE-L scores implies that the fine-tuned model generates more contextually relevant and linguistically coherent summaries, even if exact word overlap (ROUGE-1 and ROUGE-2) is slightly reduced.

A possible explanation for this trade-off is that the fine-tuned model prioritizes semantic understanding and paraphrasing over exact word matching, leading to summaries that may deviate lexically but retain the intended meaning more effectively. This is particularly valuable in dialogue summarization, where rigid word-for-word matching may not fully capture the nuances of conversational exchanges.

Furthermore, the increased ROUGE-Lsum score suggests that the model's ability to generate logically structured, multi-sentence summaries has improved, making it more adept at preserving the narrative flow of dialogues. This indicates that PEFT techniques not only maintain competitive performance but may also enhance overall readability and coherence—a critical factor in real-world applications such as customer support, medical consultations, and meeting summarization.

Future refinements, such as further hyperparameter tuning, reinforcement learning from human feedback (RLHF), or dataset augmentation, could further optimize both lexical accuracy and coherence, ensuring a better balance between factual precision and human-like summarization.

# Conclusion

This study explored the impact of fine-tuning a large language model for dialogue summarization using parameter-efficient fine-tuning (PEFT) techniques. The results demonstrate that fine-tuning with QLoRA and bitsandbytes can improve summary coherence and structure while maintaining competitive performance in word-level accuracy metrics.

Although the ROUGE-1 and ROUGE-2 scores slightly declined, this does not necessarily indicate a reduction in summary quality. Instead, the improvement in ROUGE-L and ROUGE-Lsum scores suggests that the model generates more coherent, well-structured summaries that effectively capture the essence of dialogues. This trade-off between lexical overlap and contextual understanding is particularly relevant in dialogue summarization, where rigid word matching may not fully reflect the nuances of natural conversation.

Additionally, the success of PEFT techniques in maintaining strong performance while using significantly fewer trainable parameters underscores their effectiveness for fine-tuning large models in resource-constrained environments. This approach enables organizations and researchers to adapt pre-trained language models to specific tasks without the computational overhead of full fine-tuning, making advanced NLP techniques more accessible and cost-effective.

Beyond improving summarization quality, this project provided valuable insights into dataset handling, optimization techniques, and the trade-offs involved in model fine-tuning. Future research could explore more advanced fine-tuning strategies, such as reinforcement learning from human feedback (RLHF) or hybrid approaches that balance extractive and abstractive summarization. Additionally, experimenting with larger or more diverse datasets could further enhance generalization and robustness across different conversational domains.

Overall, our findings highlight the potential of fine-tuned large language models to generate high-quality, contextually accurate summaries, contributing to advancements in automated dialogue processing and real-world applications such as customer service, medical consultations, and meeting summarization.

# References

* Hugging Face. (n.d.). Phi-2 Model. Retrieved from https://huggingface.co/
* Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. Retrieved from https://arxiv.org/abs/2305.14314
* Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. Advances in Neural Information Processing Systems (NeurIPS).
* Kaliannan Neelamohan, K. (n.d.). DialogSum Dataset. Retrieved from https://huggingface.co/datasets/dialogsum
* Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 21(140), 1–67.
* Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems.