AI-Driven Research Papers Classification: Fine-Tuning LLM with LoRA for Efficient Adaptation

Dipto Biswas
Dept. of Computer Software Engineering
Daegu Catholic University
Gyeongsan-si, South Korea
dipto.biswas94@gmail.com

Jaechoon Jo
Dept. of Computer Education
Jeju National University
Jeju-si, South Korea
jjo@jejunu.ac.kr

Joon-Min Gil*
School of Computer Engineering
Jeju National University
Jeju-si, South Korea
jmgil@jejunu.ac.kr

***Abstract*—Large Language Models (LLMs) have demonstrated remarkable capabilities across a range of Natural Language Processing (NLP) tasks, including text classification. However, effectively applying these models to specific tasks requires fine-tuning and high-quality labeled datasets, which are often scarce in specialized academic fields. This study proposes an innovative approach to efficient Research Paper Classification (RPC) by leveraging Low-Rank Adaptation (LoRA) to fine-tune the Large Language Model Meta AI (LlaMA-3 (3B)). LoRA enables adapting pretrained models with minimal parameter updates, significantly reducing the number of trainable parameters while maintaining high performance. We explore the application of this technique in RPC and evaluate its performance using the Top-n method in classification scenarios. We introduce a Top-n labels approach that leverages semantic relationships to provide precise, user-specific recommendations relevant to the context of search queries. Our experiments demonstrate that the LoRA-enhanced LlaMA-3 achieves competitive classification performance, surpassing baseline fine-tuning methods for RPC.**

***Index Terms*—Research Papers Classification (RPC), Large Language Models (LLM), Low-Rank Adaptation (LoRA), Large Language Model Meta AI (LlaMA)**

I. Introduction

In this study, we introduce the deep semantic capabilities of LlaMA-3, combined with LoRA's lightweight adaptability, for RPC tasks. We employ the Top-n labels method to improve RPC performance by capturing semantic relationships and providing precise, user-aligned suggestions based on contextual relevance in search queries. Our framework enables efficient, scalable, and accurate classification with minimal training overhead. It addresses the growing demand for intelligent research paper management in digital libraries, significantly enhancing the performance of research paper search and recommendation systems.

This research was supported by the Basic Science Research Program to the Research Institute for Basic Sciences (RIBS) of Jeju National University through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (RS-2019-NR040080).
*** Corresponding author: Joon-Min Gil**

II. System Model

The process for RPC using LLM starts with a structured pipeline. First, research paper datasets D serve as input, containing titles, abstracts, keywords, and optionally, full text or metadata. For each paper (abstract) $p \in D$, the goal is to prepare this data for downstream tasks. Let $P = (p_1, p_2, \ldots, p_n)$ represent the textual input, where each p_i is a Research Paper Abstract (RPA) to be standardized and preprocessed. In the data preprocessing step, raw text is cleaned and structured through multiple stages using the NLTK modules[1]. The preprocessing function P can be expressed as:

$$P(RPA) \rightarrow RPA' \tag{1}$$

where, RPA' denotes the preprocessed research paper abstract.

The next steps involve tokenizing the input RPA' with a SentencePiece tokenizer aligned with LlaMA-3, and truncating or padding each input sequence to a fixed maximum token length L (e.g., $L = 512$) [1] to ensure uniform input dimensions. The input sequence X is defined as:

$$X = \{x_1, x_2, \ldots, x_L\}, \quad x_i \in \mathbb{V} \tag{2}$$

where, $\mathbb{V}$ is the vocabulary space. The input sequence X is then embedded and passed to the LlaMA-3 model.

Given an input sequence X, the model generates contextualized token representations H by sequentially applying these components by LlaMA-3, expressed as:

$$H = \text{Transformer}(X) = \text{LN}(X + \text{MHSA}(X) + \text{FFN}(X)) \tag{3}$$

where, **MHSA** represents the multi-head self-attention mechanism, **FFN** signifies the position-wise feedforward network, and **LN** indicates layer normalization.

Specifically, LoRA injects trainable rank decomposition matrices $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times d}$ into the attention projections, typically applied to the query W_Q and value W_V matrices. The updated weight matrix $\tilde{W}$ is expressed as:

$$\tilde{W} = W + \Delta W = W + \alpha \cdot AB \tag{4}$$

[1]https://www.nltk.org/api/nltk.html

TABLE I: Comparative performance results for RPC using Top-5 labels.

Model	arXiv		DBLP		Elsevier		PubMed	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Gemini-pro(S)	0.4988	0.4780	0.4533	0.4632	0.4667	0.4546	0.4667	0.4546
GPT-4	0.5267	0.5059	0.5200	0.4066	0.4876	0.4820	0.4876	0.4820
Qwen(7B)	0.4900	0.4519	0.5133	0.4453	0.5133	0.4410	0.5133	0.4410
Vicuna(7B)	0.4700	0.4362	0.4676	0.4146	0.5467	0.5412	0.5471	0.5412
LlaMA-2(8B)	0.5363	0.5298	0.6333	0.5797	0.4718	0.4258	0.4718	0.4258
LlaMA-3 (LoRA)$^{Top\text{-}5}$	**0.6732**	**0.6529**	**0.6974**	**0.6584**	**0.7048**	**0.6786**	**0.6470**	**0.6240**

where W is the frozen pre-trained weight matrix (e.g., from LlaMA-3), $\Delta W(= \alpha \cdot AB)$ denotes the low-rank adaptation, and α is a scaling factor, which is commonly set to 1 or adjusted relative to the rank r.

Using a pooling strategy, we extract a global representation $h \in \mathbb{R}^d$ from the final token. This representation is passed through a linear classification layer, enhanced with LoRA [2], to produce raw logits $z \in \mathbb{R}^C$, defined as:

$$z = W_{\text{cls}}h + b + \alpha \cdot ABh \tag{5}$$

where, W_{cls} is the weight matrix that projects the d-dimensional embedding h into a C-dimensional space corresponding to the number of labels, while b is the bias vector associated with each label.

The raw logits are transformed into a categorical probability distribution over all classes using the softmax activation function, defined as:

$$\hat{y}_i = \text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}, \quad \forall i \in \{1, 2, ..., C\} \tag{6}$$

where z_i represents the logit associated with the i-th class, $\hat{y}_i \in [0, 1]$ denotes the predicted probability of class i, and C is the total number of labels. The final predicted label corresponds to the class with the highest predicted probability:

$$\hat{y} = \arg\max_i \hat{y}_i \tag{7}$$

This study optimizes the model using the Categorical Cross-Entropy (CCE) loss function. The loss for a single training example is defined as:

$$\mathcal{L}_{\text{CCE}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{8}$$

where $y_i \in \{0, 1\}$ is the one-hot encoded ground truth label (exactly one y_i is 1, others are 0), and $\hat{y}_i \in [0, 1]$ is the predicted probability for label i.

III. Experimental Design

A. Baseline Model

In this study, we evaluate leading Large Language Models such as Gemini-Pro, GPT-4, Qwen-7B, Vicuna-7B, and LLaMA-2-7B under identical settings to ensure a fair, apples-to-apples comparison.

B. Dataset Description

Experiments are conducted using four benchmark datasets: **(1) arXiv (2) DBLP**, **(3) Elsevier**, and **(4) PubMed**. These datasets are obtained using APIs from the "HuggingFace" platform. The statistical overview of each dataset is summarized in [3]

C. Model Hyperparameter

In this setup, we fine-tuned the `"open_LlaMA_3b"` model, which has 5,324,800 trainable parameters (0.16%), by using a LoRA configuration approach to enable efficient model adaptation with minimal additional parameters.

IV. Results and Discussion

Tables I illustrates the significant improvements achieved by LlaMA-3 (LoRA) in Top-5 labels across various datasets (arXiv, DBLP, Elsevier, and PubMed). Specifically, LlaMA-3 (LoRA) outperforms baseline models, achieving 10% improvements in accuracy (ACC) and F1-score (F1). This result is primarily attributed to LoRA-based fine-tuning, which enables effective model optimization with only a few parameter changes. Additionally, using the Top-n labeling method improves the model's ability to understand semantic relationships between research papers.

V. Conclusion

This study presents an innovative approach to RPC by employing LoRA to fine-tune the LlaMA-3 model. This model enhances the classification performance for user-aligned recommendations. The results indicate that LoRA-based fine-tuning considerably improves classification accuracy with the Top-n labeling method. This model effectively captures the semantic relationships among research papers, leading to more precise recommendations. The experimental findings of this study reveal that LlaMA-3 (LoRA) outperforms baseline models in accuracy and F1-score across various datasets for RPC.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

[2] 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," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2022.

[3] D. Biswas, T. Byun, J. Gil, and J. Jo, "A study on research paper classification using fine-tuning LLM based on lora," in *Proc. of the Korea Artificial Intelligence Conference (KAIC)*, 2025.