

Development of Open Large Language Models for Artificial Intelligence Digital Textbooks

¹ Youngho Lee

¹ *Department of Computer Education, Daegu National University of Education, Daegu, Republic of Korea*

Abstract –Artificial Intelligence (AI) is being utilized in various fields, and research on generative AI, particularly within natural language processing (NLP) technology, is actively being conducted. Currently, research related to generative AI in the education sector utilizes closed large language models (LLMs) like GPT. However, these models have limitations, as they are difficult to fine-tune and incur high costs. This study aims to explore the potential educational applications of Open LLMs by fine-tuning and comparing the performance of Llama2 and Polyglot, which are built on a Korean-based model, with Llama3, which is not based on a Korean model. The experimental results, using a question-and-answer dataset from elementary school social studies and science subjects, showed that the Llama2 13B model exhibited the highest performance, followed by the Polyglot 12.8B model. The Llama3 8B model achieved approximately 93.08% of the performance of the Llama2 13B model and about 98.63% of the performance of the Polyglot 12.8B model. This indicates that even relatively small, non-Korean-based models can demonstrate high performance.

These results suggest that future development of LLMs base models may omit the process of converting them into language-specific base models. Additionally, fine-tuning Open LLMs for educational applications shows potential for providing personalized education.

Keywords – Open LLMs, fine-tuning, performance of LLMs, AI digital textbooks, Llama3, Llama2, Polyglot.

1. Introduction

The digital textbook project in Korea aims to introduce advanced technologies such as artificial intelligence into education to promote digital transformation in the field. Beginning in 2025, this plan will gradually introduce digital textbooks in subjects such as English, mathematics, and information technology [1]. The existing textbooks will be digitized from paper-based formats to e-books, and customized technologies will be developed to support students and teachers in learning and teaching efficiently. These technologies include intelligent tutoring systems using AI to provide human-like lessons, metaverse environments, and extended reality (XR). Additionally, the Ministry of Education plans to prepare training sessions for teachers to learn how to use digital textbooks and will operate pilot schools to facilitate digitalization [1]. Artificial intelligence (AI) plays a crucial role in digital textbooks and is utilized in various ways to enhance the educational experience. In particular, AI technology is implemented in Korea's digital textbook project in the following ways: First, there is the intelligent tutoring system. This system uses AI to enable computers to conduct lessons similar to human teachers. AI identifies students' learning levels and preferences, providing personalized learning paths and responding to students' questions in real-time. Next, AI is used for learning diagnostics and assessments. It evaluates students' learning progress and achievements to determine areas where additional support is needed.

DOI: 10.18421/TEM134-14

<https://doi.org/10.18421/TEM134-14>


Corresponding author: Youngho Lee,
Department of Computer Education, Daegu National University of Education, Daegu, Republic of Korea
Email: yhlee@dnu.ac.kr

Received: 19 June 2024.

Revised: 19 September 2024.

Accepted: 04 November 2024.

Published: 27 November 2024.

 © 2024 Youngho Lee; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDeriv 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

This allows teachers to plan their lessons more efficiently and provide tailored guidance to each student. In this way, AI plays a pivotal role in improving the quality of education in digital textbooks, offering students a more personalized and effective learning experience. Recently, generative AI technology has been having various impacts on education. Several researchers have conducted studies on the influence of generative AI on students and teachers. Grassini [2] explored the potential impacts and issues of AI and ChatGPT on education. ChatGPT has the ability to provide personalized learning experiences and real-time feedback [2]. Emphasizing the potential of AI tools to reshape educational norms, there is the need to reconsider educational standards and establish ethical guidelines [2]. Grassini highlighted the necessity for research on how AI technologies like ChatGPT are being implemented and utilized in educational settings. Additionally, Hakiki *et al.* [3] conducted a study on the impact of generative AI on learning outcomes. The data analysis of participants showed that the group using ChatGPT achieved higher performance levels. This study suggests that ChatGPT can have a positive effect on educational technology [3]. In addition, research targeting teachers has also been conducted. Ming *et al.* [4] explored the transformative impact of ChatGPT on teacher professional development. They found that ChatGPT can assist teachers in guiding students more effectively by providing personalized learning support and real-time feedback [4]. As shown above, most current research on generative AI in education utilizes closed LLMs like ChatGPT. Bommasani *et al.* [5] suggested that models trained on large-scale data, such as BERT, DALL-E, and GPT-3, are transforming the paradigm of AI. Recently, various Open LLMs have been developed to overcome the limitations of closed LLMs. To maximize the performance of generative AI, it is crucial to adjust models for specific tasks or industries. Although closed LLMs, such as GPT-4, exhibit excellent performance, fine-tuning them presents several challenges. The fine-tuning process requires extensive computational resources and powerful hardware, leading to significant costs. Additionally, a substantial amount of high-quality data is needed, and the process of collecting and preparing this data can incur additional expenses [6], [7]. Thus, while the fine-tuning of closed LLMs is necessary, it incurs substantial costs. Preparing high-quality datasets, understanding neural networks, and optimizing hyperparameters are complex tasks. However, the field of NLP is rapidly evolving, with new technologies and more advanced models continuously being developed. Meta's research team has introduced a new model called Llama.

Initially, Llama was intended for academic use only and consisted of a collection of base models with parameters ranging from 7 billion to 65 billion. With the release of Llama 2, it became available for commercial use, and the model expanded to 70 billion parameters. Llama demonstrated that it could be run on consumer hardware and that cutting-edge models could be created using only public datasets. Notably, the 13 billion parameter model outperformed GPT-3 in terms of performance [8]. Llama 2 is a pre-trained large language model optimized for conversational use cases. The Llama 2-Chat model outperforms other open-source conversational models on most benchmarks and could serve as a suitable alternative to closed models in terms of utility and safety [9]. The recently announced Llama3 model by Meta features configurations ranging from 800 million to 70 billion parameters, achieving state-of-the-art performance in various tasks. These models are pre-trained on over 15 trillion tokens, demonstrating impressive results in zero-shot and few-shot evaluations [10]. Additionally, the Polyglot model is a significant tool that offers cross-lingual transfer, improved performance in non-English languages, and high flexibility and performance in multilingual contexts. These models are playing increasingly important roles in various research and practical environments [11]. The CFA Institute (2023) emphasized that fine-tuning LLMs is cost-effective and reproducible, presenting models that can drive new advancements in downstream tasks. Researchers have explored fine-tuning techniques using smaller models like Llama 7B, promoting the growth of open LLM. Various studies have shown that fine-tuning significantly enhances performance across multiple fields such as education, healthcare, and manufacturing, particularly in solving complex problems. This demonstrates that LLMs possess innovative and practical applications in various sectors. Thus, Open LLMs allow the development of LLMs optimized for specific tasks. In this study, an Open LLM-based language model optimized for the educational context of students in the educational field is developed. Fine-tuned models based on the Open LLMs Llama2, Llama3, and Polyglot were developed to address the research questions. Datasets corresponding to elementary school social studies and science subjects were utilized for fine-tuning, and the accuracy of the fine-tuning results was compared. Among the models used in this study, Llama2 and Polyglot-Ko were based on models that were further developed into Korean-based models, while Llama3 was not. This research demonstrates the potential for using LLMs in the educational field and underscores the importance of developing optimized educational models.

2. Literature Review

This section provides an overview of various applications of large language models (LLMs) in the education sector. It explores how LLMs are used to enhance educational experiences through personalized learning, language support, automated assessments, and feedback analysis.

2.1. Review of LLM Application Cases in the Education Field

LLMs have various applications in the education field, including personalized learning, language learning support, automated assessment, and educational feedback analysis. These applications hold the potential to enhance the quality of education and improve the learning experience. Gan *et al.* [12] investigated how LLMs can be utilized in digital and smart education. They presented new methods and approaches to achieve personalized learning, intelligent tutoring, and educational assessment goals using LLMs [12]. The study emphasized the potential of LLMs to contribute to improving the quality of education. Cavojský *et al.* [13] also explored the educational applications of ChatGPT, highlighting its usefulness as a learning assistant and research support tool. Hicke *et al.* [14] evaluated the ability of LLMs to simulate the role of a teacher in educational dialogues and provide valuable insights. Their results showed that GPT-4 outperformed other models, demonstrating its effectiveness in educational applications. Caines *et al.* [15] explored the potential for integrating LLMs into AI-based language education and assessment systems. They evaluated the performance of LLMs in text generation, automated grading, and grammar error correction, while also presenting the challenges, including ethical considerations. Parker *et al.* [16] assessed the potential of using LLMs to analyze educational feedback surveys. They demonstrated methods for classifying, extracting, performing thematic analysis, and sentiment analysis on survey responses using LLMs. Their findings confirmed that the GPT-4 model achieved human-level performance.

Additionally, there are research findings indicating that LLMs can be effective tools for language learning. Peng *et al.* [17] evaluated the effectiveness of LLMs in language learning, particularly in spoken language learning. The study found that LLMs showed high performance in understanding and applying spoken language knowledge, and suggested that various prompting techniques could be used to improve their performance. Hamaniuk [18] explored how LLMs can be utilized in educational technology for tasks such as machine translation, prompt programming, and abstract text reasoning. The study highlighted the potential of LLMs as tools for creative writing and effective paraphrasing [18].

Lee *et al.* [19] designed the CoAuthor dataset to support creative and argumentative writing using LLMs, exploring how these models can enhance language, creativity, and collaboration skills. The study found that collaboration between humans and AI could increase the efficiency and creativity of writing tasks [19]. Additionally, the suggestions from AI models inspired human writers and helped overcome creative blocks during the writing process. Additionally, LLMs are being utilized as tools in programming courses to generate learning materials and assist students in generating code. Krüger and Gref [20] evaluated the effectiveness of LLMs as educational tools in undergraduate computer science programs. They presented the models with lecture materials, assignments, and past exam questions, finding that the models performed well overall but had limitations in mathematical calculations [20]. Sarsa *et al.* [21] presented a method for using LLMs, such as OpenAI Codex, to automatically generate programming assignments (including sample solutions and test cases) and code explanations. This study suggested that generating programming assignments and code explanations can reduce the time and effort educators spend developing learning materials [21]. Ross *et al.* [22] conducted a study on developing a "programmer's assistant" system using LLMs to aid developers in writing code through interactive dialogues. The study indicated that this system can provide developers with interactive assistance on various coding tasks and demonstrated the effectiveness of such a system. In addition, various studies have explored the educational applications of LLMs [22]. Kung *et al.* [23] demonstrated the potential of LLMs in medical education and clinical decision support by showing that models like ChatGPT can achieve near-passing or passing scores on the United States Medical Licensing Examination.

These examples illustrate that LLMs can be effectively utilized in the education sector in various ways, such as generating learning materials, supporting students, and assisting educators. They highlight the potential of this technology to enhance the quality of educational experiences.

2.2. Analysis of the Effectiveness of LLM Fine-Tuning Studies

This section examines the effectiveness of fine-tuning techniques applied to LLMs. It highlights key studies that demonstrate the improvements in LLM performance across different fields, including education, by using fine-tuning methods. In 2023, a research team at Stanford University developed the Alpaca model by fine-tuning the Llama 7B model with 52,000 instruction-following data points using text-davinci-003.

This process was completed with just \$100 and three hours of training time. Despite its small size and low budget, Alpaca demonstrated strong performance [24]. The CFA Institute (2023) considered this achievement significant in two aspects. First, it presented a cost-effective and reproducible model with performance comparable to much larger proprietary models. Second, it opened the door to new advancements in downstream tasks through the fine-tuning of open-source LLMs [8]. Shashidhar *et al.* [25] also emphasized the potential for advancements in downstream tasks through the fine-tuning of open-source LLMs. Through small models like Llama, researchers were able to explore innovative techniques in fine-tuning, leading to the explosive growth of open-source LLMs capable of competing with proprietary models.

Various studies have shown that fine-tuning LLMs can significantly improve performance in specific tasks. For example, LLMs have demonstrated excellent performance in the fields of education, program synthesis, medical data analysis, and aligning with diverse human preferences. Liu *et al.* [26] explored methods to improve the fine-tuning of LLMs for solving mathematical problems. The study employed various fine-tuning strategies to significantly enhance the LLMs' ability to solve mathematical problems [26]. As a result, these methods substantially improved the models' mathematical problem-solving capabilities, achieving high accuracy, particularly with complex problems. The CFA Institute (2023) explored methods for fine-tuning LLMs to process unstructured data, such as financial data analysis. This approach enabled more effective analysis of complex data patterns in the financial sector [8]. Xia *et al.* [27] studied how fine-tuning LLMs could be effectively utilized in the manufacturing sector. Their research demonstrated that fine-tuning LLMs can significantly enhance the accuracy and efficiency of manufacturing processes [27]. Wang *et al.* [28] presented the ClinicalGPT model, which was fine-tuned using various medical datasets and comprehensively evaluated [28]. This study demonstrated that fine-tuning LLMs for the medical field can significantly improve diagnostic accuracy and the reliability of treatment recommendations.

The fine-tuned model performed exceptionally well in various medical scenarios, providing superior results compared to existing medical AI systems. Bakker *et al.* [29] explored methods for fine-tuning LLMs to find consensus among people with diverse preferences. The fine-tuned model was shown to be effective in helping people coordinate their opinions and reach agreements, playing a crucial role in managing the complexities of social interactions.

Weysow *et al.* [30] demonstrated that LLMs could generate meaningful code snippets based on natural language descriptions [30]. The fine-tuned model exhibited excellent performance in solving various programming tasks, particularly in code generation and error correction. The study's results indicated that fine-tuning LLMs could significantly reduce human effort and time in programming tasks.

These studies demonstrate that fine-tuning LLMs has innovative and practical applications across various fields, supporting the need for developing fine-tuned LLMs in the educational sector as explored in this research.

3. Methodology

In this study, we developed LLMs for social studies and science subjects in Korean elementary schools and measured their performance. The specific steps carried out are as follows: First, the development of LLMs was conducted by fine-tuning Open LLMs. For fine-tuning, we used the dataset from Lee [31]. Following this, we measured the accuracy of the fine-tuned models using the GPT-4 API.

3.1. Dataset

The dataset used in this study consists of question-and-answer pairs presented in the teachers' guides for social studies and science subjects in Korean elementary schools, covering grades 3 to 6. The amplified dataset was generated based on the seed dataset, using the GPT-4 with a self-instruction method [32].

The amplified dataset was utilized as the training dataset, while the seed dataset was used as the test dataset for fine-tuning the LLMs. An overview of the datasets used is provided in Table 1.

Table 1. Overview of seed and amplified datasets

Subject	No. Seed Dataset	No. Amplified Dataset	Hugging Face URL
Science	1,061	10,665	JosephLee/science_textbook_elementary_kor_seed JosephLee/science_textbook_elementary_kor
Society	2,217	22,170	JosephLee/society_textbook_elementary_kor_seed JosephLee/society_textbook_elementary_kor

3.2. Base Model

The models used in this study, Llama-2-KoEn, Llama3, and Polyglot-Ko, are categorized into Korean-based models and general-based models. As described earlier, the Llama-2-KoEn and Polyglot-Ko models are existing base models that have been converted into Korean-based models. The Llama-2-KoEn model is an advanced version of Llama 2, enhanced by incorporating Korean and English corpus to improve its language processing capabilities. This model is configured with 13 billion parameters. It is an autoregressive language model based on an optimized transformer architecture. Polyglot-Ko is a large-scale Korean autoregressive language model series developed by the Polyglot team at EleutherAI. This model features 40 transformer layers, with a model dimension of 5,120 and a feed-forward dimension of 20,480. Each head has a dimension of 128, divided into a total of 40 heads, with rotary position embeddings (RoPE) applied to the 64 dimensions of each head. The Polyglot-Ko-12.8B model was trained on 863GB of Korean data curated by TUNiB, addressing the performance deficiencies that existing multilingual models often exhibit in non-English languages [11].

Next, the Llama3 model is an autoregressive language model that uses an optimized transformer architecture, incorporating supervised fine-tuning and reinforcement learning from human feedback. It has both a pre-trained version and an instruction-tuned version, with this study utilizing the instruct version. The instruction-tuned model is particularly optimized for conversational use cases and has demonstrated superior performance compared to existing open-source chat models on common industry benchmarks. An overview of the base models used can be seen in Table 2.

Table 2. Overview of base models

Model	Size	Korean-based model	Hugging Face URL
Polyglot-ko	12.8B	O	beomi/polyglot-ko-12.8b-safetensors
Llama2-koen	13B	O	beomi/llama-2-koen-13b
Llama3	8B	X	meta-llama/Meta-Llama-3-8B

3.3. Fine-Tuning Method

Active research is being conducted on various techniques for fine-tuning LLMs. In this study, we used the QLoRA (Quantized Low-Rank Adaptation) technique for model fine-tuning.

QLoRA involves fixing the weights of the pre-trained model and injecting trainable low-rank matrices into each layer, which significantly reduces the number of parameters. Hu *et al.* [33] reported that using this technique, it is possible to reduce the number of parameters by 10,000 times and the GPU memory requirements by 3 times in models like GPT-3, without any performance degradation. Additionally, Sun *et al.* [34] compared the full parameter tuning and LoRA-based tuning methods using the Llama model. The experimental results demonstrated that LoRA-based tuning offers significant advantages in terms of training costs and also showed superior performance in the model's effectiveness. Subsequently, Dettmers *et al.* [35] proposed QLoRA as a memory-efficient approach that allows fine-tuning of a 6.5 billion parameter model on a single 48GB GPU. In this study, they trained Low Rank Adapters (LoRA) by backpropagating gradients through a pre-trained language model quantized to 4 bits. He proposed the QLoRA technique, which utilizes the 4-bit NormalFloat (NF4) data type and double quantization of quantization constants. This method achieved 99.3% of ChatGPT's performance by fine-tuning on small, high-quality datasets and demonstrated that fine-tuning could be completed on a single GPU within 24 hours. Additionally, Huang *et al.* [36] explored the performance of the Llama3 model using various low-bit quantization methods. They evaluated the model across multiple zero-shot and few-shot benchmarks, demonstrating that low-bit quantization can significantly reduce model size and computational requirements while maintaining high performance. In this study, we applied the QLoRA method, a fine-tuning technique known for its performance and effectiveness, to fine-tune the base models of Open LLMs.

3.4. Model Accuracy Measurement Method

To measure the accuracy of the models developed in this study, we assessed the models' responses based on the seed dataset from the question-and-answer datasets. Specifically, the method for calculating the model's accuracy is as follows: First, we constructed a dataset by randomly extracting 20% of the test data from the question-and-answer datasets for science and social studies subjects.

The 'question' portion of the extracted dataset was input into the developed LLMs to generate responses. The correctness of the responses was then evaluated using the GPT-4 API, with a scoring scale ranging from 1 to 5. To minimize measurement errors, this process was repeated five times, and the average score was calculated.

Then, using the GPT-4 API, we entered the following prompt to evaluate the correctness of the answers on a scale of 5 to 1. To reduce errors in accuracy measurement, this process was repeated five times, and the average score was calculated.

4. Research Results

This section presents the outcomes of the model fine-tuning process and evaluates the performance of each fine-tuned LLM based on specific datasets.

4.1. Results of Model Fine-Tuning

This subsection details the steps taken to fine-tune three pre-trained LLMs—Polyglot, Llama2, and Llama3—using quantization and Low-Rank Adaptation techniques to optimize performance. This study used three pre-trained language models as base models: Polyglot (beomi/polyglot-ko-12.8b-safetensors), Llama2 (beomi/Llama-2-koen-13b), and Llama3 (meta-Llama/Meta-Llama-3-8B).

For each model, we loaded the tokenizer and the model itself, applying 4-bit quantization using BitsAndBytesConfig to enhance the model's efficiency. This quantization setup reduces computational memory requirements and increases model loading and computation speed. The configured quantization parameters include using torch.bfloat16 as the data type, specifying nf4 as the quantization type, and enabling double precision quantization.

We applied the LoRA (Low-Rank Adaptation) technique to the models, expanding only the essential parameters to increase adaptability for specific tasks without retraining the entire model. The LoRA settings include a rank $r=8$, a scaling factor $lora_alpha=32$, and a dropout rate of 0.05. The additional LoRA matrices $W1$ and $W2$ are used to fine-tune the model's responses.

During training, the update of model parameters is performed using the gradient of the loss function for the batch data, following the standard backpropagation formula. The formula used for updating the model parameters is:

$$\theta(t+1) = \theta(t) - \eta \cdot \nabla \theta L(\theta(t), D_{batch})$$

where $\theta(t)$ represents the model parameters at iteration t , η is the learning rate, and $\nabla \theta L$ is the gradient of the loss function based on the batch data D_{batch}

4.2. Results of Model Accuracy Measurement

This subsection outlines the accuracy measurements of the fine-tuned models based on their performance on science and social studies datasets. Comparative results between Polyglot, Llama2, and Llama3 are analyzed. The performance of each model was measured on five randomly extracted groups from the test dataset, with scores evaluated on a scale of 5 points.

The accuracy of models fine-tuned on the science question-and-answer dataset is as follows. The Llama2 13B model demonstrated the highest performance with an average score of 4.259, while the Polyglot 12.8B model showed the lowest performance with a score of 4.021. The Llama3 8B model exhibited intermediate performance with an average score of 3.966.

Table 3. Accuracy measurement results of the model fine-tuned with science question-and-answer dataset

Model	1	2	3	4	5	Average
Polyglot	3.989	4.213	4.042	3.892	3.968	4.021
Llama2	4.234	4.260	4.178	4.360	4.261	4.259
Llama3	4.090	3.979	3.907	3.899	3.955	3.966

The accuracy of models fine-tuned on the social studies question-and-answer dataset is as follows. The Llama2 13B model demonstrated the highest performance, achieving an average score of 4.432. Conversely, the Polyglot 12.8B model showed the lowest performance with a score of 4.058. The Llama3 8B model exhibited intermediate performance, with an average score of 4.158.

Table 4. Accuracy measurement results of the model fine-tuned with society question-and-answer dataset

Model	1	2	3	4	5	Average
Polyglot	4.010	4.204	4.005	3.971	4.101	4.058
Llama2	4.373	4.521	4.516	4.307	4.444	4.432
Llama3	4.082	4.208	4.039	4.231	4.229	4.158

The following graph compares the performance of three LLM models, Polyglot 12.8B, Llama2 13B, and Llama3 8B, on the science question-and-answer dataset. The X-axis represents each group, while the Y-axis represents the performance scores of the models measured in each group.

The Polyglot 12.8B model is represented by black bars, the Llama2 13B model by dark gray bars, and the Llama3 8B model by light gray bars.

The average performance scores for each model are indicated by dashed lines of the respective colors.

The black dashed line represents the average score of approximately 4.021 points for the Polyglot 12.8B model, the dark gray dashed line represents the average score of approximately 4.259 points for the Llama2 13B model, and the light gray dashed line represents the average score of approximately 3.966 points for the Llama3 8B model.

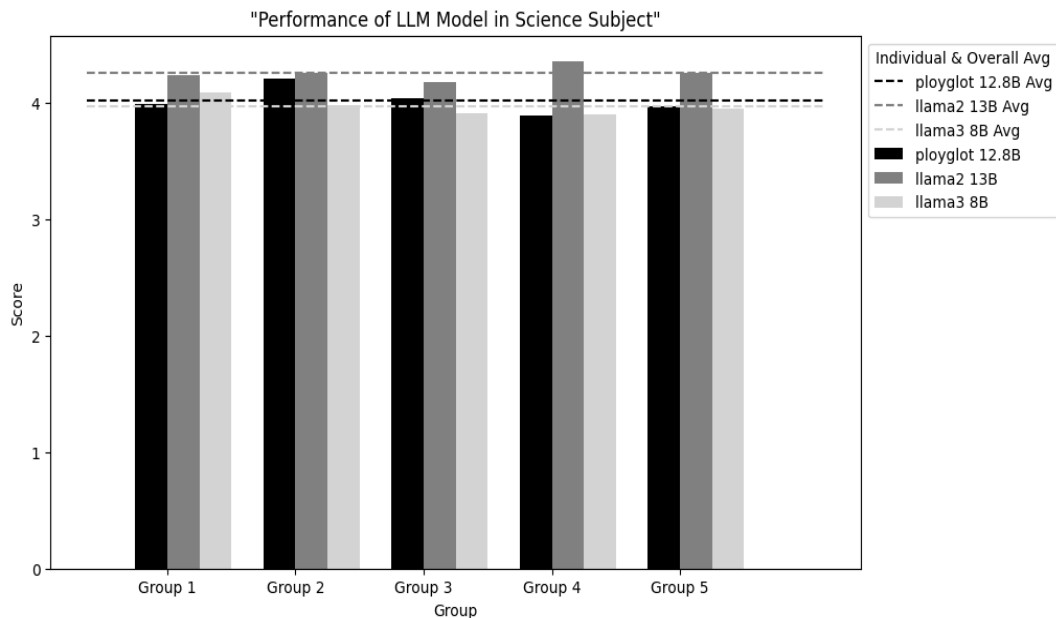


Figure 1. The performance of three LLM models, Polyglot 12.8B, Llama2 13B, and Llama3 8B, on the science textbooks dataset

Figure 1 compares the performance of three LLM models, Polyglot 12.8B, Llama2 13B, and Llama3 8B, on the science textbooks dataset. The X-axis represents each group, while the Y-axis represents the performance scores of the models measured in each group. The Polyglot 12.8B model is represented by black bars, the Llama2 13B model by dark gray bars, and the Llama3 8B model by light gray bars. The average performance scores for each model are indicated by dashed lines of the respective colors. The black dashed line represents the average score of approximately 4.021 points for the Polyglot 12.8B model, the dark gray dashed line represents the average score of approximately 4.259 points for the Llama2 13B model, and the light gray dashed line represents the average score of approximately 3.966 points for the Llama3 8B model.

Each model's performance exhibits similar patterns across the five groups, demonstrating consistent performance around the average score with minimal variation.

The Polyglot 12.8B model consistently showed performance in the early 4-point range across all groups, while the Llama2 13B model achieved the highest average score, indicating superior performance. In contrast, the Llama3 8B model demonstrated performance similar to the Polyglot 12.8B model but recorded slightly lower scores.

The following graph illustrates the performance comparison of three LLM models, Polyglot 12.8B, Llama2 13B, and Llama3 8B, on the social studies dataset. The Polyglot 12.8B model is represented by black bars, the Llama2 13B model by gray bars, and the Llama3 8B model by light gray bars. The black dashed line represents the average score of the Polyglot 12.8B model, approximately 4.058 points, the gray dashed line represents the average score of the Llama2 13B model, approximately 4.432 points, and the light gray dashed line represents the average score of the Llama3 8B model, approximately 4.158 points.

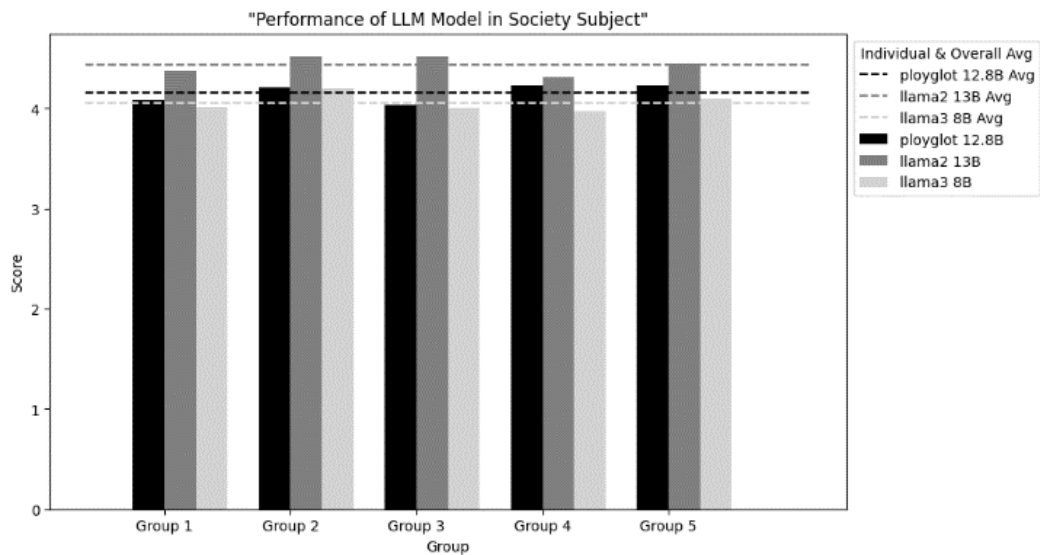


Figure 2. The performance of three LLM models, Polyglot 12.8B, Llama2 13B, and Llama3 8B, on the society textbooks dataset

Each model's performance exhibits similar patterns across the five groups, showing consistent performance near the average score with minimal variation. The Polyglot 12.8B model demonstrated performance in the early to mid-4-point range across all groups, while the Llama2 13B model recorded the highest average score, indicating excellent performance. On the other hand, although the Llama3 8B model scored lower than the Llama2 13B model, it outperformed the Polyglot 12.8B model.

When comparing the performance of each model, the Llama2 13B model exhibited the highest accuracy in both the science and social datasets, while the Polyglot 12.8B model showed intermediate accuracy in the science dataset and the lowest accuracy in the social dataset. The Llama3 8B model demonstrated intermediate accuracy in the social dataset and the lowest accuracy in the science dataset.

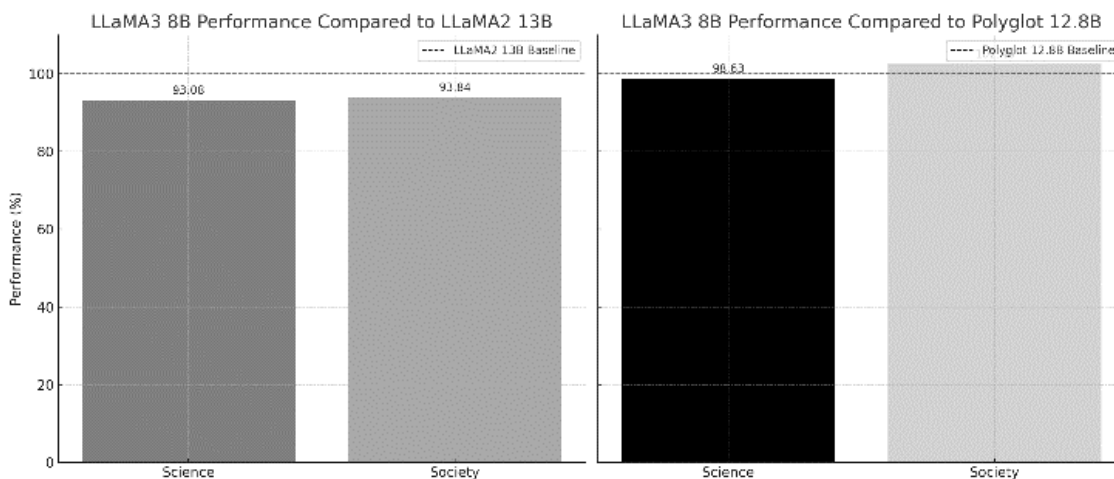


Figure 3. Comparison of the Llama 3 Model with the Llama 2 and Polyglot Models

When examining in detail the model fine-tuned based on the Llama3 8B base model, the following observations are made: Fine-tuned on subject-specific datasets, the Llama3 8B model demonstrated performance approximately 93.08% of Llama2 13B model in social studies and around 93.84% in science. Compared to the Polyglot 12.8B model, it achieved approximately 98.63% in social studies and 102.46% in science.

The findings align with previous research indicating that larger models tend to exhibit better performance. However, it is noteworthy that even the relatively smaller-sized Llama3 model shows high accuracy. Moreover, while Llama2 and Polyglot models were fine-tuned based on Korean base models, the Llama3 model underwent fine-tuning using a general model.

Despite this difference, the performance of the Llama3 model showed little deviation compared to the two Korean base models, underscoring its strong inherent performance.

The results of this experiment offer valuable insights into how fine-tuned LLM models generate responses accurately within specific domains. Particularly, the exceptional performance of the Llama2 13B model can serve as a significant criterion for setting future research directions. Additionally, despite its relatively smaller size, the Llama3 8B model's high performance suggests a need to reassess the relationship between model size and performance.

5. Conclusion

Artificial intelligence is being utilized in various aspects of education, with a particular emphasis on research related to generative AI in NLP technology. However, most current research relies on Closed LLMs like GPT. While Closed LLMs demonstrate excellent performance, they present challenges in fine-tuning and are costly. This poses practical difficulties in providing diverse personalized learning pathways in educational settings. Therefore, there is a need for a new approach using Open LLMs.

Thus, this study aimed to overcome these limitations and explore the educational potential of Open LLMs. For this purpose, three Open LLMs were selected, including two Korean-based models (Llama2, Polyglot) and one non-Korean-based model (Llama3). Fine-tuning was conducted based on question-and-answer datasets corresponding to social studies and science in Korean elementary schools. Through this process, the performance of each model could be evaluated and compared.

The performance comparison results revealed that Korean-based models outperformed non-Korean-based models relatively. The Llama2 13B model exhibited the highest average score, demonstrating excellent performance, followed by the Polyglot 12.8B model. The performance of the Llama2 13B model remained consistently high across all groups, particularly excelling in the social studies dataset. In contrast, the Llama3 8B model showed performance approximately 93.08% compared to the Llama2 13B model and approximately 98.63% compared to the Polyglot 12.8B model. This indicates that despite being a non-Korean-based model, it performed well on Korean datasets. It is noteworthy that the non-Korean-based Llama3 8B model performed comparably well to the Korean-based models.

This suggests the possibility of skipping the process of converting future LLM base models into Korean-based models. Such an approach could increase cost-effectiveness and reduce the complexity of the fine-tuning process.

Based on this, it is conceivable that by fine-tuning future Open LLM base models using quantization techniques, LLMs with performance similar to existing Closed LLMs could be developed. Furthermore, it is anticipated that this approach would enable personalized education by fine-tuning LLMs for specific grades or levels of students. Furthermore, this study, by demonstrating the potential of Open LLMs, can promote the adoption of digital textbooks and personalized education. For instance, AI-based educational tools like intelligent tutoring systems can offer personalized learning paths tailored to students' learning levels and preferences. Moreover, AI can respond to students' questions in real-time, assess their learning progress, and identify areas needing additional support. Through this, teachers can plan education more efficiently and provide tailored guidance to each student.

In conclusion, the educational potential of fine-tuning Open LLMs holds significant promise, paving the way for effective adoption of digital textbooks and personalized education. Future research should expand the application scope of these Open LLMs and explore their potential utility in various educational settings. Additionally, optimization of LLMs tailored to diverse educational demands and the establishment of supporting infrastructure are necessary. Through such efforts, we can establish a new paradigm in AI-based education.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF-2340007811) grant funded by the Korea government.

References:

- [1]. Ministry of Education. (2023). MOE to unlock personalized education for all with AI-embedded textbooks. English MOE. Retrieved from: https://english.moe.go.kr/boardCnts/viewRenew_al.do?boardID=265&boardSeq=95268&lev=0&searchType=null&statusYN=W&page=1&s=english&m=02_01&opType=N [accessed: 08 May 2024]
- [2]. Grassini, S. (2023). Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. *Education Sciences*. Doi: 10.3390/educsci13070692
- [3]. Hakiki, M., et al. (2023). Exploring the impact of using Chat-GPT on student learning outcomes in technology learning: The comprehensive experiment. *Advances in Mobile Learning Educational Research*. Doi: 10.25082/amlr.2023.02.013

- [4]. Ming, G. K., & Mansor, M. (2023). Exploring the Impact of Chat-GPT on Teacher Professional Development: Opportunities, Challenges, and Implications. *Asian Journal of Research in Education and Social Sciences*, 5(4), 54-67.
- [5]. Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. *ArXiv preprint, arXiv:2108.07258*.
- [6]. Patel, D., et al. (2023). The limits of prompt engineering in medical problem-solving: a comparative analysis with ChatGPT on calculation based USMLE medical questions. *MedRxiv*. Doi: 10.1101/2023.08.06.23293710
- [7]. Valdez, D., Bunnell, A., Lim, S. Y., Sadowski, P., & Shepherd, J. A. (2024). Performance of progressive generations of GPT on an exam designed for certifying physicians as Certified Clinical Densitometrists. *Journal of Clinical Densitometry*, 27(2), 101480. Doi: 10.1101/2023.07.25.23293171
- [8]. CFA Institute. (2024). *Unstructured data and AI: Fine-tuning LLMs to enhance the investment process*. Research & Policy Center. Retrieved from: <https://rpc.cfainstitute.org/en/research/reports/2024/unstructured-data-and-ai> [accessed: 25 May 2024].
- [9]. Touvron, H., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint, arXiv:2307.09288*.
- [10]. Touvron, H., et al. (2023). Llama: Open and Efficient Foundation Language Models. *ArXiv preprint, arXiv:2302.13971*. Doi: 10.48550/arXiv.2302.13971
- [11]. Ko, H., et al. (2023). A Technical Report for Polyglot-Ko: Open-Source Large-Scale Korean Language Models. *ArXiv, abs/2306.02254*. Doi: 10.48550/arXiv.2306.02254
- [12]. Gan, W., Qi, Z., Wu, J., & Lin, J. C. W. (2023). Large language models in education: Vision and opportunities. *2023 IEEE international conference on big data (BigData)*. IEEE. Doi: 10.48550/arXiv.2311.13160
- [13]. Čavojský, M., Bugár, G., Kormaník, T., & Hasin, M. (2023). Exploring the Capabilities and Possible Applications of Large Language Models for Education. *2023 21st International Conference on Emerging eLearning Technologies and Applications (ICETA)*, 91-98. IEEE.
- [14]. Hicke, Y., Masand, A., Guo, W., & Gangavarapu, T. (2023). Assessing the efficacy of large language models in generating accurate teacher responses. *ArXiv preprint, arXiv:2307.04274*. Doi: 10.48550/arXiv.2307.04274
- [15]. Caines, A., et al. (2023). On the application of large language models for language teaching and assessment technology. *ArXiv preprint, arXiv:2307.08393*. Doi: 10.48550/arXiv.2307.08393
- [16]. Parker, M. J., Anderson, C., Stone, C., & Oh, Y. (2024). A large language model approach to educational survey feedback analysis. *ArXiv preprint, arXiv:2309.17447*. Doi: 10.48550/arXiv.2309.17447
- [17]. Peng, L., Nuchged, B., & Gao, Y. (2023). Spoken Language Intelligence of Large Language Models for Language Learning. *ArXiv preprint, arXiv:2308.14536*. Doi: 10.48550/arXiv.2308.14536
- [18]. Hamaniuk, V. A. (2021). The potential of Large Language Models in language education. *Educational Dimension*. Doi: 10.31812/ed.650
- [19]. Lee, M., Liang, P., & Yang, Q. (2022). CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Doi: 10.1145/3491102.3502030
- [20]. Krüger, T., & Gref, M. (2023). Performance of Large Language Models in a Computer Science Degree Program. *European Conference on Artificial Intelligence*, 409-424. Cham: Springer Nature Switzerland.
- [21]. Sarsa, S., Denny, P., Hellas, A., & Leinonen, J. (2022). Automatic generation of programming exercises and code explanations using large language models. *Proceedings of the 2022 ACM Conference on International Computing Education Research*, 1. Doi: 10.1145/3501385.3543957
- [22]. Ross, S., Martinez, F., Houde, S., Muller, M., & Weisz, J. (2023). The Programmer's Assistant: Conversational Interaction with a Large Language Model for Software Development. *ArXiv preprint, arXiv:2302.07080*. Doi: 10.1145/3581641.3584037
- [23]. Kung, T., et al. (2022). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health*, 2. Doi: 10.1371/journal.pdig.0000198
- [24]. Taori, R., et al. (2023). Alpaca: A strong, replicable instruction-following model. *Stanford University Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>
- [25]. Shashidhar, S., Chinta, A., Sahai, V., Wang, Z., & Ji, H. (2023). Democratizing LLMs: An Exploration of Cost-Performance Trade-offs in Self-Refined Open-Source Models. *ArXiv preprint, arXiv:2310.07611*. Doi: 10.48550/arXiv.2310.07611
- [26]. Liu, Y., Singh, A., Freeman, C., Co-Reyes, J., & Liu, P. (2023). Improving Large Language Model Fine-tuning for Solving Math Problems. *ArXiv preprint, arXiv:2310.10047*. Doi: 10.48550/arXiv.2310.10047
- [27]. Xia, L., Li, C., Zhang, C., Liu, S., & Zheng, P. (2024). Leveraging error-assisted fine-tuning large language models for manufacturing excellence. *Robotics and Computer-Integrated Manufacturing*, 88, 102728. Doi: 10.1016/j.rcim.2024.102728
- [28]. Wang, G., Yang, G., Du, Z., Fan, L., & Li, X. (2023). ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation. *ArXiv preprint, arXiv:2306.09968*. Doi: 10.48550/arXiv.2306.09968
- [29]. Bakker, M., et al. (2022). Fine-tuning language models to find agreement among humans with diverse preferences. *ArXiv preprint, arXiv:2211.15006*. Doi: 10.48550/arXiv.2211.15006

- [30]. Weyssow, M., Zhou, X., Kim, K., Lo, D., & Sahraoui, H. (2023). Exploring parameter-efficient fine-tuning techniques for code generation with large language models. *ArXiv preprint, arXiv:2308.10462*. Doi; 10.48550/arXiv.2308.10462
- [31]. Lee, Y. (2023, December). Research on Dataset Generation in the Development of Large Language Models for Digital Textbooks. *2023 3rd International Conference on Robotics, Automation and Artificial Intelligence (RAAI)*, 1088-1093. IEEE.
- [32]. Peng, B., Li, C., He, P., Galley, M., & Gao, J. (2023). Instruction tuning with gpt-4. *ArXiv preprint, arXiv:2304.03277*.
- [33]. Hu, E. J., *et al.* (2021). Lora: Low-rank adaptation of large language models. *ArXiv preprint, arXiv:2106.09685*.
- [34]. Sun, X., Ji, Y., Ma, B., & Li, X. (2023). A comparative study between full-parameter and lora-based fine-tuning on chinese instruction data for instruction following large language model. *ArXiv preprint, arXiv:2304.08109*. Doi: 10.48550/arXiv.2304.08109
- [35]. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *ArXiv preprint, arXiv:2305.14314*. . Doi: 10.48550/arXiv.2305.14314
- [36]. Huang, W., *et al.* (2024). How Good Are Low-bit Quantized Llama3 Models? An Empirical Study. *ArXiv preprint, arXiv:2404.14047*.