Taiwan-LLM Tutor: Large Language Models for Taiwanese Secondary Education

Jia-Wei Liao, Ji-Jia Wu, Kun-Hsiang Lin, Kang-Yang Huang
National Taiwan University, Taipei, Taiwan

GitHub

Abstract

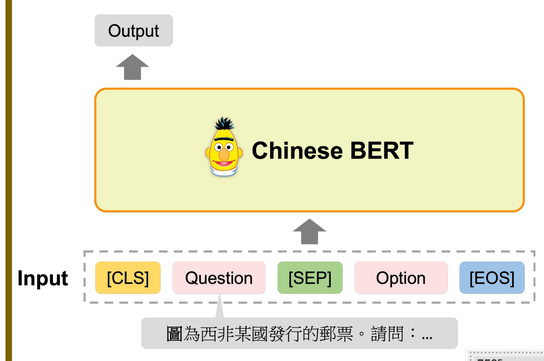
In this project, we utilize language models to develop an AI tutor aimed at enhancing the learning experiences of students in Taiwanese secondary education. We organized the General Scholastic Ability Test (GSAT) dataset and integrated it with a high school social studies question bank. Utilizing BERT and TWLLM as our foundational models, we designed four architectures to compare the results of fine-tuning. This initiative lays the groundwork for more interactive AI-driven educational tools and outlines future research directions to further enhance AI's role in education.

Method

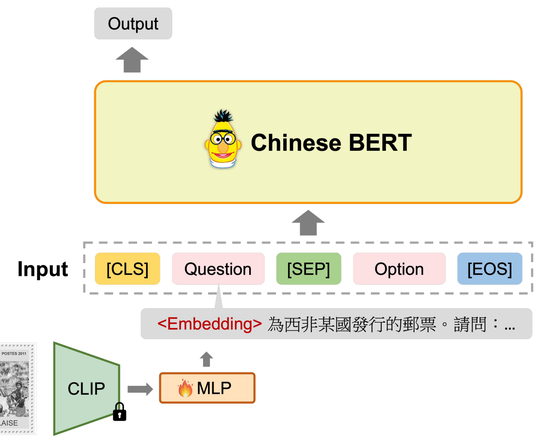
- Pre-trained model: Chinese BERT [1], TWLLM [2]
- Fine-tune method: QLoRA [3], LoftQ [4]
- Optimizer: AdamW, Lion

Architectures

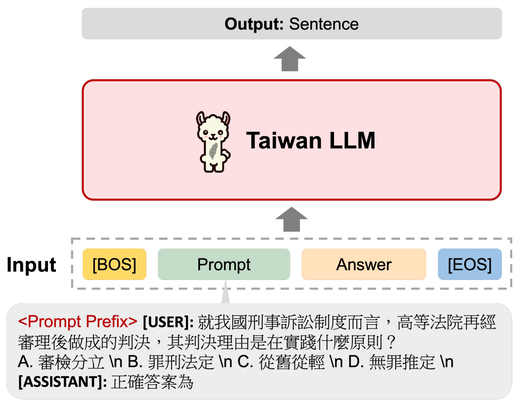
(a) BERT Multiple Choice



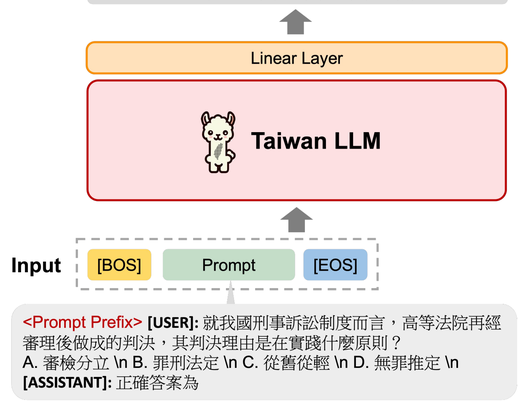
(b) Vision-BERT Multiple Choice



(c) TWLLM Instruction Tuning

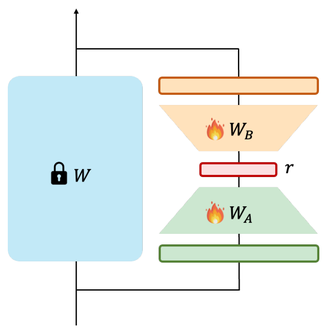


(d) TWLLM Multiple Choice



Fine-Tune Methods

(a) LoRA



(b) LoftQ

$$\min_{Q,A,B} \left\| W - Q - AB^{\top} \right\|_F$$

1: Initialize $A_0 \leftarrow 0, B_0 \leftarrow 0$
2: **for** t = 1 to T **do**
3: Obtain quantized weight $Q_t \leftarrow q_N(W - A_{t-1}B_{t-1}^{\top})$
4: Obtain low-rank approximation $A_t, B_t \leftarrow \text{SVD}(W - Q_t)$
5: **end for**

Conclusion

In this project, we utilize language models to develop an AI tutor aimed at improving the learning experiences of students in Taiwanese secondary education. We have compiled the General Scholastic Ability Test (GSAT) dataset and integrated it with a high school social studies question bank. Utilizing BERT and TWLLM as foundational models, we designed four architectures to compare the results of fine-tuning. This initiative establishes a foundation for more interactive AI-driven educational tools and outlines future research directions to enhance AI's role in education further.

For future endeavors, we aim to:
1. Enhance explanations by applying Reinforcement Learning from Human Feedback (RLHF).
2. Collaborate with Junyi Academy (均一教育) to improve educational outcomes.

Experiment Results

Question
<Prompt Prefix> 你是一名人工智慧家教，以下的題目為高中社會科，請根據題目提供使用者正確答案的選項 A 或 B 或 C 或 D。USER: <Question and Options> ASSISTANT：正確答案為

Answer
<Correct Option> 原因：<Answer Explanation>

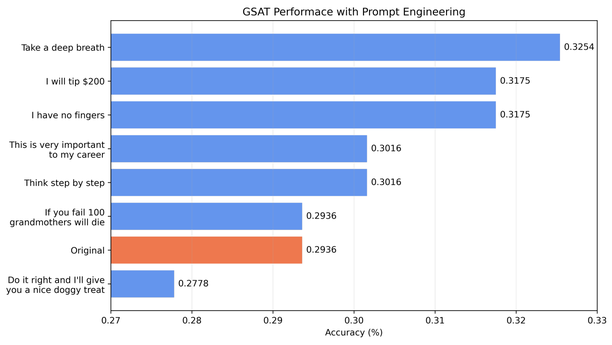
Model Setting

Model	BERT	Vision BERT	Taiwan LLM	Taiwan LLM	Taiwan LLM
Method	MC	MC	MC + QLoRA	IT + QLoRA	IT + LoftQ
Epochs	10	10	10	10	10
Batch size	8 × 16	8 × 16	16 × 1	16 × 1	4 × 4
Optimizer	AdamW	AdamW	AdamW or Lion	AdamW or Lion	AdamW
Learning rate	2e-5	2e-5	2e-4	2e-4	2e-4
Weight decay	1e-5	1e-5	1e-5	1e-5	1e-5
Scheduler	Linear	Linear	Constant	Constant	Constant
Warm up step	300	300	0	0	0

Quantitative Results

Model	Method	Accuracy
Chinese-BERT	MC	0.3568
Taiwan LLM	MC	0.3286
Taiwan LLM	IT+QLoRA	0.3380
Taiwan LLM	IT+LoftQ	**0.4789**
ChatGPT-3.5	Zero-shot	0.5000

Table 2: Result of test performance on 108-112 social GSAT.



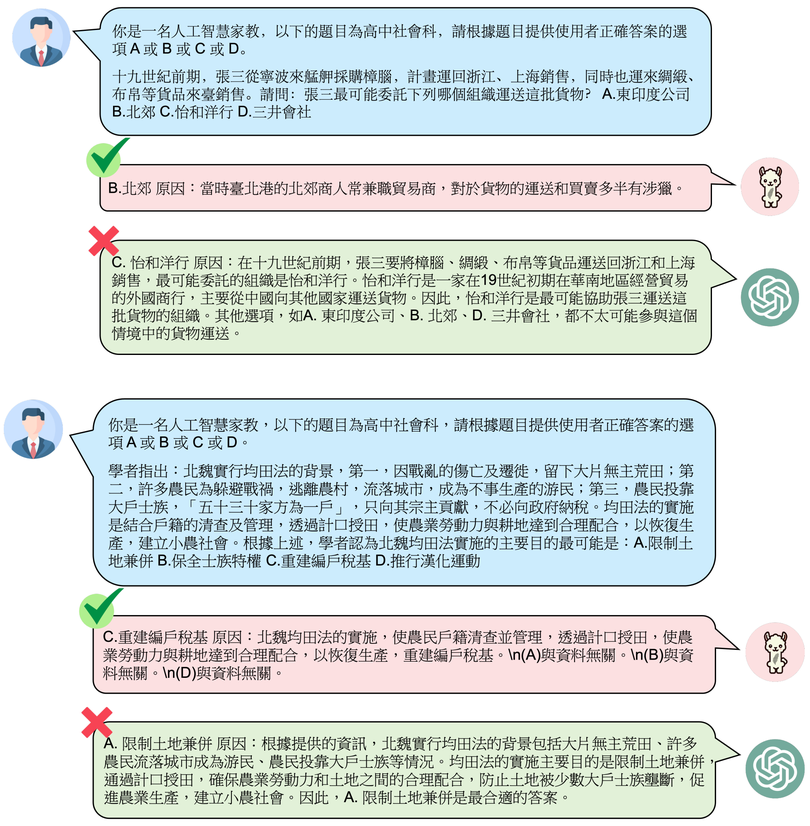
Training Dataset	Testing Dataset	Model	Method	Explanation	Accuracy
History QB (9000)	108-112 History GSAT	Chinese BERT	MC		0.4742
History QB (9000)	108-112 History GSAT	Taiwan LLM	MC		0.5773
History QB (9000)	108-112 History GSAT	Taiwan LLM	IT + QLoRA		0.5051
History QB (9000)	108-112 History GSAT	Taiwan LLM	IT + QLoRA	✓	0.5360
History QB (9000)	108-112 History GSAT	Taiwan LLM	IT + LoftQ	✓	**0.6082**
Civics QB (2035)	108-112 Civics GSAT	Chinese BERT	MC		0.4177
Civics QB (2035)	108-112 Civics GSAT	Taiwan LLM	MC		0.3418
Civics QB (2035)	108-112 Civics GSAT	Taiwan LLM	IT + QLoRA		0.4051
Civics QB (2035)	108-112 Civics GSAT	Taiwan LLM	IT + QLoRA	✓	0.4936
Civics QB (2035)	108-112 Civics GSAT	Taiwan LLM	IT + LoftQ	✓	**0.5443**

Table 3: Results of test performance on the 108-112 history GSAT and civics GSAT, respectively. The table illustrating experimental results demonstrates that the model trained using LoftQ exhibits superior performance on the GSAT dataset.

Training Dataset	Model	Method	Explanation	Accuracy
83-107 Social GSAT	Chinese BERT	MC		0.3568
Social QB (11347)	Chinese BERT	MC		**0.4507**
83-107 Social GSAT	Taiwan LLM	IT + QLoRA		0.3380
Social QB (11347)	Taiwan LLM	IT + QLoRA	✓	**0.5681**
83-107 Social GSAT	Taiwan LLM	IT + LoftQ		0.4789
Social QB (11347)	Taiwan LLM	IT + LoftQ	✓	**0.5446**

Table 4: The experimental results show that training the model with the QB notably improves its performance on the 108-112 social GSAT tests, compared to other training sets.

Demo Case



Reference

1. Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *Proceedings of naacL-HLT*. Vol. 1. 2019.
2. Lin, Yen-Ting, and Yun-Nung Chen. "Taiwan llm: Bridging the linguistic divide with a culturally aligned language model." *arXiv preprint arXiv:2311.17487* (2023).
3. Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." *arXiv preprint arXiv:2305.14314* (2023).
4. Li, Yixiao, et al. "Loftq: Lora-fine-tuning-aware quantization for large language models." *arXiv preprint arXiv:2310.08659* (2023).