Taiwan LLM Tutor: Revolutionizing Taiwanese Secondary Education with Large Language Model

Jia-Wei Liao **Ji-Jia Wu** **Kun-Hsiang Lin** **Kang-Yang Huang**
National Taiwan University, Taipei, Taiwan
{d11922016, r11922086, p12922005, r11944070}@csie.ntu.edu.tw

Abstract

In this project, we introduce an innovative AI-based approach to improve education in underserved regions, focusing on a customized academic dataset and LLM tuning methods. Centered on the General Scholastic Ability Test (GSAT) dataset, our approach aims to provide fair educational support in areas with limited resources. We curated a GSAT dataset and expanded it with a comprehensive social studies question bank. The research extensively explores both BERT-based and LLM-based architectures, with particular emphasis on the Taiwan LLM. It also incorporates advanced training techniques such as LoftQ and the Lion optimizer. Our results demonstrate significant performance improvements over existing models on GSAT benchmarks, with notable advancements in the integration of visual data through Vision BERT. This paper lays the groundwork for more interactive AI-driven educational tools and outlines future research directions to further enhance AI's capabilities in education. Our source code is available at https://github.com/jwliao1209/TWLLM-Tutor.git.

1 Introduction

In disadvantaged areas, educational resources are severely limited, posing significant challenges for many children who lack access to personalized academic support. The COVID-19 pandemic has led to a shift to online education, resulting in a higher prevalence of computing devices in households that can be used for educational purposes. Notably, marginalized communities experience a heightened lack of educational resources, emphasizing the need for improved one-on-one instructional approaches. This context forms the foundation of our research initiative. Our project is dedicated to developing an Artificial Intelligence model specifically designed to enhance the educational experience of these groups. This AI-based intervention is designed to address disparities in educational access by providing customized pedagogical support and creating equitable learning opportunities for students in resource-limited settings.

In the course of our research, we systematically collected a comprehensive dataset from the General Scholastic Ability Test (GSAT) administered by the College Entrance Examination Center (CEEC). This dataset has undergone a rigorous process of organization and cleansing to ensure its relevance and accuracy. After careful evaluation and refinement, around 1,500 questions were chosen to form the core dataset for our study. This carefully selected set of questions serves as the empirical foundation for our analysis and the development of our AI model. It ensures a strong and representative sample for our educational research efforts.
To enhance the robustness of our AI model, we augmented our original dataset by integrating a comprehensive *Social Studies* question bank, accompanied by detailed answer explanations. This question bank is categorized into three primary disciplines: *Geography*, with 322 questions; *History*, comprising a substantial 9,205 questions; and *Civics*, consisting of 2,035 questions.

The significantly smaller number of *Geography* questions in our dataset can be attributed to the specific nature of our AI model, which is based on Large Language Models (LLM). LLMs are inherently text-based and, as a result, have limitations in processing non-textual data inputs, such as images or tables. Consequently, during the process of compiling the dataset for *Geography*, a significant number of questions that heavily rely on visual aids such as maps, diagrams, or tabular data were excluded. This exclusion was necessary because these types of questions are incompatible with the current capabilities of our LLM-based model. Therefore, the dataset for *History* and *Civics* could be more comprehensively compiled due to their predominantly text-based question formats. In contrast, the *Geography* section was limited to questions that could be presented and interpreted solely in a textual format. This methodology ensures that our AI model operates within its functional capacity, focusing on text-based analysis and interpretation, which is crucial for maintaining the integrity and effectiveness of the model in educational applications.

This addition significantly broadens the scope and depth of our dataset, offering a more nuanced and extensive range of educational content. Such a diverse dataset is crucial for training our AI model to comprehend and tackle a wide range of questions, thereby enhancing its accuracy and effectiveness in educational applications. This strategic expansion of our dataset is crucial to ensuring that our AI model is not only well-rounded but also capable of providing comprehensive and contextually rich educational support.
Our initial approach involved using BERT (Devlin et al., 2018) to train a multiple-choice task as our baseline model. Recognizing that some questions included images, which cannot be processed as input by the standard language model, we introduced 'Vision BERT', which incorporates vision technology from CLIP (Radford et al., 2021). This enhancement enables our model to process visual input, thereby broadening its applicability to a wider variety of question types.

Focusing on the Taiwan LLM (Lin and Chen, 2023), we conducted a comprehensive investigation into Prompt Engineering, examining the impact of different prompts on model performance. Additionally, we explored advanced training methodologies. Notably, we implemented a novel fine-tuning architecture called LoftQ. (Li et al., 2023), specifically designed to optimize the performance of our model.

Due to budget constraints, we used the web-based version of GPT for manual question input, bypassing the need for OpenAI's API. This method facilitated ongoing research by allowing queries to generate accurate responses from the model. Our innovative approach highlights our dedication to resourcefulness in advancing AI capabilities. We conducted a performance benchmark of our 'Vision BERT', a model that integrates a visual encoder from CLIP, against GPT-3.5. The results showed that 'Vision BERT' outperformed BERT by 6.9%. Additionally, utilizing the Lion optimizer (Chen et al., 2023) and our proprietary Question Bank Dataset, our model outperformed Taiwan LLM by 45.7% on the self-collected GSAT dataset. Our contribution can be summarized as follows:

1. Create an academic dataset for AI applications in education.

2. We surveyed several LLM tuning methods in order to achieve competitive performance compared to GPT-3.5.

3. We have taken the initial step to integrate visual data with our model's input, enabling it to answer questions not only based on language data but also on visual information.

4. Our proposed method outperforms state-of-the-art methods on the self-collected GSAT benchmarks.

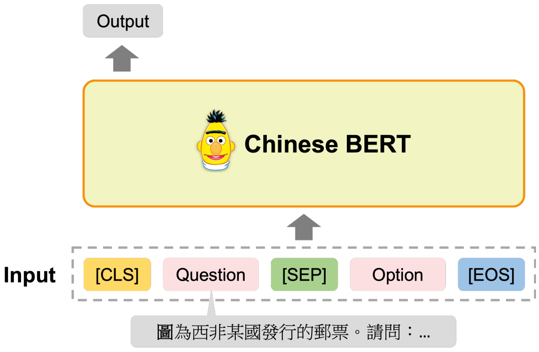
2 Method

In this project, we explore two main approaches to addressing the multiple choice problem: a BERT-based architecture and an LLM-based architecture. For each method, we thoroughly investigate various alternatives to improve performance, conducting extensive research by reviewing academic papers and engaging in technical surveys. Additionally, we leverage our expertise to apply deep learning techniques from the field of computer vision to the realm of text.

For BERT-based methods, we utilize the BERT model as a baseline and propose a Vision BERT approach to address questions that involve visual data. On the other hand, we conduct a comprehensive exploration of LLM methods. By leveraging multiple techniques, including LoRA (Hu et al., 2021), LoftQ (Li et al., 2023), and employing the novel optimizer Lion (Chen et al., 2023), we achieved a significant improvement in solving *Social Studies* questions in GSAT, resulting in a substantial performance gain.

2.1 Architecture (I): BERT for Multiple Choice

We utilize the `hfl/chinese-bert-wwm-ext` model. The detailed model architecture is illustrated in Figure 1.

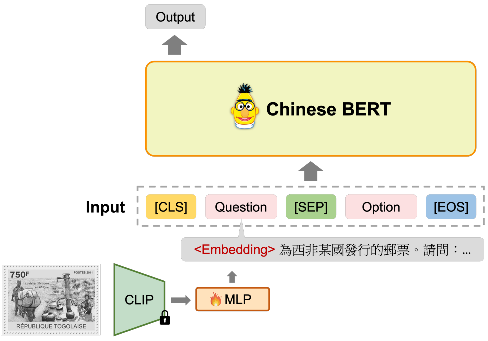


圖為西非某國發行的郵票。請問：...

Figure 1: Architecture of Chinese BERT for multiple choice.

2.2 Architecture (II): Vision BERT for Multiple Choice

In the GSAT, some questions may include image data, requiring students to consider both textual and visual information to answer correctly. However, the standard BERT model only supports textual input. To overcome this limitation and make use of visual input, we introduce the Vision BERT architecture. This architecture includes an additional pre-trained CLIP model and a learnable MLP in addition to the original BERT model. As depicted in Figure 2, the visual input is processed by passing the image through CLIP's visual encoder to obtain the visual embedding. Subsequently, we utilize the MLP to convert this visual embedding into a feature vector, which can then be considered as a word embedding and fed into the original BERT model. In each experiment, we freeze the weights of the pretrained CLIP model and maintain consistent hyperparameter settings. 2.1 remains the same.

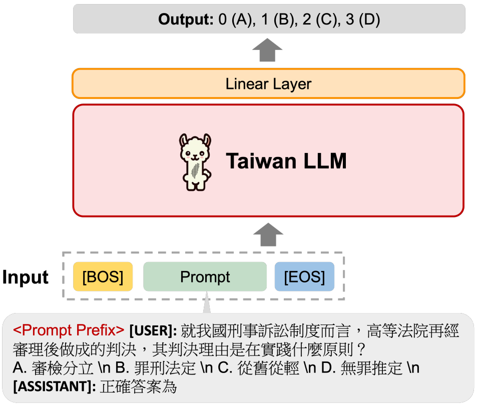


<Embedding> 為西非某國發行的郵票。請問：...

Figure 2: Architecture of Vision BERT for multimodal multiple choice.

2.3 Architecture (III): Taiwan LLM for Multiple Choice

As illustrated in Figure 3, we utilize the Taiwan LLM to construct our Taiwan LLM-based Multiple Choice Model. This architecture is built on top of the Taiwan LLM by adding an extra linear layer to its final layer. This extension adjusts the output dimension to four, with labels 0, 1, 2, and 3 corresponding to the correct answers (A), (B), (C), and (D), respectively. To fully leverage the valuable and powerful information acquired during Taiwan LLM's pretraining, we keep the pretrained weights of the Taiwan LLM fixed and exclusively fine-tune the additional final layer.

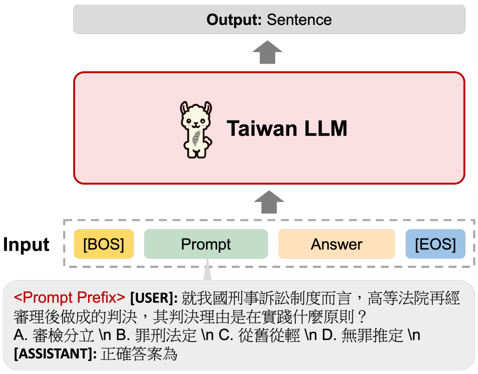


<Prompt Prefix> **[USER]:** 就我國刑事訴訟制度而言，高等法院再經審理後做成的判決，其判決理由是在實踐什麼原則？
A. 審檢分立 \n B. 罪刑法定 \n C. 從舊從輕 \n D. 無罪推定 \n
[ASSISTANT]: 正確答案為

Figure 3: Architecture of Taiwan LLM for multiple Choice.

2.4 Architecture (IV): Taiwan LLM for Instruction Tuning

For instruction tuning, we combine prompts and answers as the model's input and configure standardized outputs to facilitate accurate calculation. The model's architecture is illustrated in Figure 4.



<Prompt Prefix> **[USER]:** 就我國刑事訴訟制度而言，高等法院再經審理後做成的判決，其判決理由是在實踐什麼原則？
A. 審檢分立 \n B. 罪刑法定 \n C. 從舊從輕 \n D. 無罪推定 \n
[ASSISTANT]: 正確答案為

Figure 4: Architecture of Taiwan LLM for instruction tuning.

2.5 Low-Rank Adaptation

In the field of machine learning, low-rank approximation is a well-known technique. (Udell and Townsend, 2019) demonstrated that high-dimensional data can be approximated using low-rank matrices. In 2021, a team from Microsoft introduced LoRA (Low-rank Adaptation) (Hu et al., 2021; Aghajanyan et al., 2021) architecture, marking the first application of low-rank techniques to fine-tuning neural networks. Subsequently, researchers combined quantization techniques with LoRA to propose QLoRA (Dettmers et al., 2023), significantly reduces the computational resources required for fine-tuning large language models (LLMs). In this project, we introduce an improved version based on QLoRA, named LoftQ (LoRA Fine-Tuning Aware Quantization) (Li et al., 2023). Compared to QLoRA, LoftQ provides a more precise approximation of the original parameter matrix, leading to a significant improvement in benchmark performance. We present the LoftQ algorithm and offer our mathematical insights.

LoftQ employs an N-bit quantized weight Q and a low-rank matrix $L \in \mathbb{R}^{m\times n}$ to approximate the pre-trained weight W. The optimization problem is formulated as follows:

$$\min_{Q,L} \|W - Q - L\|_F \quad \text{s.t.} \quad \operatorname{rank}(L) \le r \qquad (1)$$

To reduce the hardware usage, the matrix L is decomposed into $AB^\top$, where $A \in \mathbb{R}^{m\times r}$ and $B \in \mathbb{R}^{r\times n}$. This reformulates the problem as follows:

$$\min_{Q,A,B} \|W - Q - AB^\top\|_F \qquad (2)$$

Since Problem 2 is NP-hard, we employ alternating optimization to approximate the solution of the reformulated problem:

$$Q_t = \operatorname*{argmin}_{Q} \|W - Q - A_{t-1}B_{t-1}^\top\|_F \qquad (3)$$

$$(A_t, B_t) \in \operatorname*{argmin}_{(A,B)} \|W - Q_t - AB^\top\|_F \qquad (4)$$

Finally, Equation 3 can be represented as $Q_t = q_N(W - A_{t-1}B_{t-1}^\top)$ with N-bits quantization function q_N. Additionally, the Eckart-Young theorem provides a closed-form solution for Equation 4, which can be obtained directly by solving the singular value decomposition (SVD). The detailed process is presented in Algorithm **??**.

2.6 Optimization

We primarily utilize AdamW and Lion as our optimizers. AdamW is a stochastic optimization technique that modifies the traditional approach to weight decay in Adam's method (Kingma and Ba, 2014) by decoupling the weight decay process from the gradient update. The algorithm 1 is described as follows:

Algorithm 1 AdamW Optimizer

```
Input: Objective function $f(\boldsymbol{\theta})$, learning rate $\gamma_t$, weight decay rate $\lambda$, $\beta_1$, $\beta_2$, and $\epsilon$
Initialization: Parameters $\boldsymbol{\theta}_0$, first moment $\mathbf{m}_0 \leftarrow \mathbf{0}$, and second moment $\mathbf{v}_0 \leftarrow \mathbf{0}$

while $\boldsymbol{\theta}_t$ not converged do
    $\mathbf{g}_t \leftarrow \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_{t-1})$
    $\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t$
    $\mathbf{v}_t \leftarrow \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\mathbf{g}_t \odot \mathbf{g}_t$
    $\hat{\mathbf{m}}_t \leftarrow \dfrac{\mathbf{m}_t}{1-\beta_1^t}$
    $\hat{\mathbf{v}}_t \leftarrow \dfrac{\mathbf{v}_t}{1-\beta_2^t}$
    $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \gamma_t \left( \dfrac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon} + \lambda \boldsymbol{\theta}_{t-1} \right)$
end while
return $\boldsymbol{\theta}_t$
```

To accelerate model convergence, we investigated the new optimizer Lion (EvoLved Sign Momentum), which introduced by the Google Brain team in 2023. The Lion algorithm was discovered through program search, as demonstrated in Algorithm 2. Distinctively, Lion uses the sign operator, which provides computational efficiency compared to AdamW's division and exponentiation operations. This efficiency reduces the computational time by 2% to 15%. Moreover, Lion demonstrates significantly improved convergence speed and overall performance compared to AdamW, especially when used with large batch sizes. In this project, we utilized Lion's rapid convergence property to significantly accelerate the completion of an extensive series of experiments.

Algorithm 2 Lion Optimizer

```
**Input:** Objective function $f(\boldsymbol{\theta})$, learning rate $\gamma_t$, weight decay rate $\lambda$, $\beta_1$, and $\beta_2$
**Initialization:** Parameters $\boldsymbol{\theta}_0$ and first moment $\mathbf{m}_0 \leftarrow \mathbf{0}$
**while** $\boldsymbol{\theta}_t$ not converged **do**
    $\mathbf{g}_t \leftarrow \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_{t-1})$
    $\mathbf{u}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1)\mathbf{g}_t$
    $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \gamma_t \cdot (\text{sign}(\mathbf{u}_t) + \lambda \boldsymbol{\theta}_{t-1})$
    $\mathbf{m}_t \leftarrow \beta_2 \mathbf{m}_{t-1} + (1 - \beta_2)\mathbf{g}_t$
**end while**
**return** $\boldsymbol{\theta}_t$
```

3 Experiments

The experiments were performed on a personal computer equipped with a single NVIDIA GeForce RTX 4090 GPU with 24 GB of VRAM, and a server configuration featuring a single RTX A6000 GPU with 49 GB of VRAM.

3.1 Datasets and Experimental Settings

For training and evaluating our model, we used the GSAT questions from the years 83 to 112, totaling approximately 2,300 questions, with at least 500 queries containing images. Furthermore, our self-compiled question bank (referred to as QB) contributed approximately 11,000 questions, distributed across History (9,000), Geography (300), and Civics (2,000). This diverse dataset provides a comprehensive foundation for robust model training and evaluation.

In the instruction tuning of Taiwan LLM, we utilized a specific template as depicted in Figure. 5, which is structured into two distinct segments. The first segment, labeled as *Question*, incorporates a *<Prompt Prefix>* and follows with directives combined with the *<Question and Options>* provided by the USER, concluding with the word ASSISTANT. The second segment, termed *Answer*, comprises the *<Correct Option>* along with the *<Answer Explanation>*. This structured approach enables Taiwan LLM to accurately respond to queries and generate comprehensive explanations, thereby enhancing its learning and explanatory capabilities.

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>**

Figure 5: Prompt template.

3.2 Quantitative Results for Different Models

In our experiment, the model training employed the entire QB along with GSAT questions from 83 to 107, specifically excluding those with images. For the evaluation, we utilized GSAT questions 108 to 112, excluding image-based questions, to assess the model's performance.

In our comparative analysis, we evaluated various models and methodologies, as detailed in Table 1. This includes Zero-shot ChatGPT, *Chinese-BERT* utilizing a Multiple Choice framework, *Taiwan-LLM-7B* in both Multiple Choice and Instruction Tuning configurations. The results are presented in Table 2, reveal a significant finding: the model *Taiwan-LLM-7B*, which was trained using instruction tuning with LoftQ exclusively on the GSAT dataset from the years 83 to 107, achieved the highest performance with a score of 0.4789. However, it is important to note that despite this achievement, the performance of Taiwan LLM still falls short when compared to the more advanced ChatGPT-3.5 model. This result offers valuable insights into the capabilities and limitations of current language models, especially in the context of specific training strategies and datasets.

In examining the experimental results presented in Table 3, a notable observation arises

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

Table 1: Hyper-parameters for models training.

regarding the effectiveness of various training methodologies. The model trained using the LoftQ framework demonstrates significantly better performance when evaluated against the GSAT dataset. The improved performance can be attributed to the specialized features and optimization strategies inherent in the LoftQ training approach, which appear to align well with the complexities and nuances of the GSAT questions. Such a finding not only emphasizes the potential of LoftQ as a robust training mechanism but also highlights the importance of selecting an appropriate training framework that resonates with the specific characteristics of the dataset in question.

Table 4 presents the outcomes of leveraging QB provides a compelling illustration of the impact of training data. The data clearly shows that using the QB as the primary source for model training significantly improves its overall performance. This improvement can be attributed to the diverse and comprehensive nature of the QB, which covers a wide range of topics and question types, providing a rich and varied training environment for the model. Such an approach ensures that the model is not only exposed to a wide range of information but also learns to handle various query formats and complexities. This finding is significant because it underscores the importance of a well-curated and extensive training dataset in developing robust and high-performing AI models, especially in applications requiring a deep understanding of diverse subject matter.

For case studies, we randomly selected three examples by comparing the answers and explanations from the instruction-tuned Taiwan LLM. In Figure 6, it is evident that Taiwan LLM, following instruction tuning, provides accurate and clearer explanations, while highlighting GPT-3.5's tendency for factual inaccuracies, as demonstrated in a particular historical example. Besides, Figure 7 demonstrates that GPT 3.5's inaccuracies, stemming from its hallucinations, result in unreliable explanations, whereas our trained model provides more accurate and trustworthy answers. Last but not least, Figure 9 indicates that GPT 3.5 often produces explanations with logical inconsistencies, while our trained model generates more coherent and logically sound explanations, highlighting the effectiveness of our training approach.

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.

3.3 Ablation Study

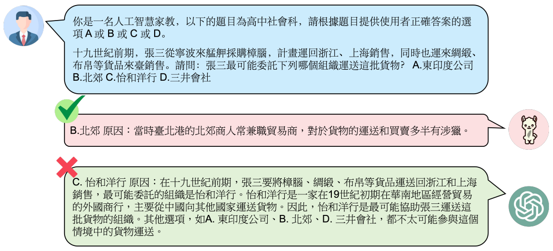
Taking one step further, the data presented in Table 5 offers a comprehensive overview of our model's performance metrics on various datasets, with a specific focus on QB and GSAT. The striking similarity in the model's performance on both of these datasets is noteworthy. This comparable level of effectiveness indicates a strong correlation between the model's training on QB and its capacity to effi-

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.

ciently process the GSAT queries. This observation has profound implications for the preparation strategies for GSAT, as it suggests that QB serves as a highly relevant and beneficial tool for such preparations. It demonstrates that the diversity and complexity of the questions in QB effectively mirror the structure and content of the GSAT, making it an invaluable resource for students and educators in their preparatory endeavors. Thus, this finding not only validates the comprehensiveness of QB but also highlights its practical utility in educational contexts, especially for high-stakes examinations such as the GSAT.

Focusing on evaluating the impact of integrating visual data on the performance of Vision BERT. This model, distinctively trained on datasets from both the GSAT and QB, was specifically designed to process and interpret questions that contain images or tables. The study involved a complex process of embedding visual elements into the linguistic framework of the model, essentially converting visual data into a format that is compatible with word embeddings. This integration was enhanced by the diverse range of textual data from the QB. As a result, Vision BERT, enhanced with visual data processing capabilities, exhibited a significant performance improvement, outperforming the baseline model (without visual data integration) by an impressive 24%. This finding not only confirms the effectiveness of incorporating visual data in language models but also provides critical insights into the improvements possible through the integration of multi-modal data integration, underscoring its potential in complex AI applications.



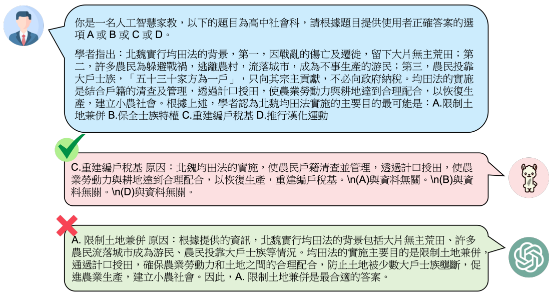
你是一名人工智慧家教，以下的題目為高中社會科，請根據題目提供使用者正確答案的選項 A 或 B 或 C 或 D。

十九世紀前期，張三從寧波來艋舺採購樟腦，計畫運回浙江、上海銷售，同時也運來綢緞、布帛等貨品來臺銷售。請問：張三最可能委託下列哪個組織運送這批貨物？ A.東印度公司 B.北郊 C.怡和洋行 D.三井會社

✓ B.北郊 原因：當時臺北港的北郊商人常兼職貿易商，對於貨物的運送和買賣多半有涉獵。

✗ C. 怡和洋行 原因：在十九世紀前期，張三要將樟腦、綢緞、布帛等貨品運送回浙江和上海銷售，最可能委託的組織是怡和洋行。怡和洋行是一家在19世紀初期在華南地區經營貿易的外國商行，主要從中國向其他國家運送貨物。因此，怡和洋行是最可能協助張三運送這批貨物的組織。其他選項，如A. 東印度公司、B. 北郊、D. 三井會社，都不太可能參與這個情境中的貨物運送。

Figure 6: An example of a history question and the corresponding responses generated by Taiwan LLM and ChatGPT (I).



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。

學者指出：北魏實行均田法的背景，第一，因戰亂的傷亡及遷徙，留下大片無主荒田；第二，許多農民為躲避戰禍，逃離農村，流落城市，成為不事生產的游民；第三，農民投靠大戶士族，「五十三十家方為一戶」，只向其宗主貢獻，不必向政府納稅。均田法的實施是結合戶籍的清查及管理，透過計口授田，使農業勞動力與耕地達到合理配合，以恢復生產，建立小農社會。根據上述，學者認為北魏均田法實施的主要目的最可能是：A.限制土地兼併 B.保全士族特權 C.重建編戶稅基 D.推行漢化運動

✓ C.重建編戶稅基 原因：北魏均田法的實施，使農民戶籍清查並管理，透過計口授田，使農業勞動力與耕地達到合理配合，以恢復生產，重建編戶稅基。\n(A)與資料無關。\n(B)與資料無關。\n(D)與資料無關。

✗ A. 限制土地兼併 原因：根據提供的資訊，北魏實行均田法的背景包括大片無主荒田、許多農民流落城市成為游民、農民投靠大戶士族等情況。均田法的實施主要目的是限制土地兼併，通過計口授田，確保農業勞動力和土地之間的合理配合，防止土地被少數大戶士族壟斷，促進農業生產，建立小農社會。因此，A. 限制土地兼併是最合適的答案。

Figure 7: An example of a history question and the corresponding responses generated by Taiwan LLM and ChatGPT (II).

3.3.1 Prompt engineering for Taiwan LLM

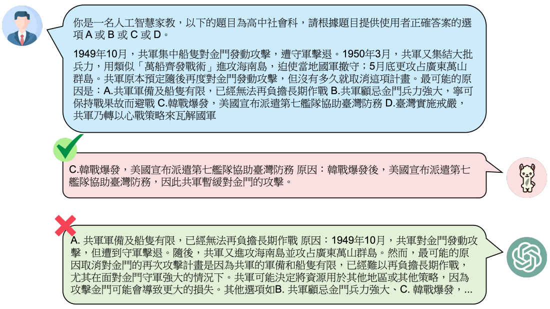
Drawing inspiration from recent advancements in prompt engineering, our study acknowledges the unique preferences of each LLM

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.

Testing Dataset	Model	Method	Explanation	Accuracy
108-112 History GSAT	Chinese BERT	MC		0.4742
History QB (205)	Chinese BERT	MC		0.4780
108-112 History GSAT	Taiwan LLM	MC		0.5773
History QB (205)	Taiwan LLM	MC		0.5463
108-112 History GSAT	Taiwan LLM	IT + QLoRA		0.5360
History QB (205)	Taiwan LLM	IT + QLoRA		0.3463
History QB (205)	Taiwan LLM	IT + QLoRA	✓	0.6000
108-112 History GSAT	Taiwan LLM	IT + LoftQ	✓	0.6082
History QB (205)	Taiwan LLM	IT + LoftQ	✓	0.6098

Table 5: Training with the History QB (9000 questions) results in comparable performance on both the GSAT and QB, indicating its effectiveness for GSAT preparation.



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。

1949年10月，共軍集中船隻對金門發動攻擊，遭守軍擊退。1950年3月，共軍又集結大批兵力，用類似「萬船齊發戰術」進攻海南島，迫使當地國軍撤守；5月底更攻占廣東萬山群島。共軍原本預定隨後再度對金門發動攻擊，但沒有多久就取消這項計畫。最可能的原因是：A.共軍軍備及船隻有限，已經無法再負擔長期作戰 B.共軍顧忌金門兵力強大，寧可保持戰果故而避戰 C.韓戰爆發，美國宣布派遣第七艦隊協助臺灣防務 D.臺灣實施戒嚴，共軍乃轉以心戰策略來瓦解國軍

C.韓戰爆發，美國宣布派遣第七艦隊協助臺灣防務 原因：韓戰爆發後，美國宣布派遣第七艦隊協助臺灣防務，因此共軍暫緩對金門的攻擊。

A. 共軍軍備及船隻有限，已經無法再負擔長期作戰 原因：1949年10月，共軍對金門發動攻擊，但遭到守軍擊退。隨後，共軍又進攻海南島並攻占廣東萬山群島。然而，最可能的原因取消對金門的再次攻擊計畫是因為共軍的軍備和船隻有限，已經難以再負擔長期作戰，尤其在面對金門守軍強大的情況下。共軍可能決定將資源用於其他地區或其他策略，因為攻擊金門可能會導致更大的損失。其他選項如B. 共軍顧忌金門兵力強大、C. 韓戰爆發，...

Figure 8: An example of a history question and the corresponding responses generated by Taiwan LLM and ChatGPT (III).

towards specific instructional prompts. To achieve this goal, we compiled and tested seven different prompt templates on the Taiwan LLM under zero-shot conditions. The baseline performance is represented by an orange bar chart with a score of 0.2936, reflecting the model's output without the application of any prompt engineering techniques. Notably, the prompt *Take a deep breath* significantly surpassed other variations in effectiveness. Conversely, Taiwan LLM showed a notable aversion to prompts related to *doggy treat* rewards, indicating a preference for specific types of instruction over others.

3.3.2 Experimental Results for Vision BERT

As demonstrated in Table 6, our proposed Vision BERT demonstrates superior performance compared to BERT by integrating both textual and visual information. This improvement is evident whether training on a limited dataset (GSAT) or a relatively larger dataset (GSAT+QB).

Training Dataset	Model	Accuracy
83-107 GSAT	BERT	0.3351
83-107 GSAT	Vision BERT	0.3514
83-107 GSAT + QB	BERT	0.3892
83-107 GSAT + QB	Vision BERT	**0.4162**

Table 6: Results of test performance for 108-112 GSAT using BERT and Vision BERT.

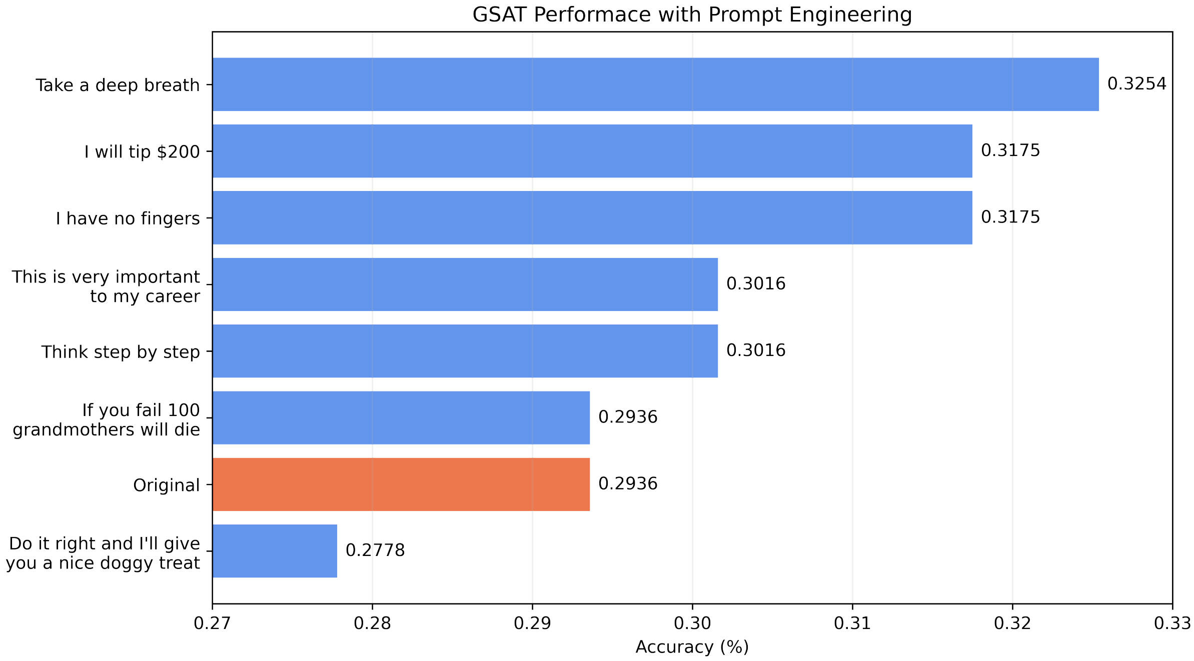


Figure 9: Prompt engineering on zero-shot Taiwan LLM.

4 Discussion

In expanding our research scope, it becomes imperative to explore beyond the confines of specialized QB and consider the integration of raw data from textbooks as a potential training source for our models. Textbooks, which represent the cornerstone of traditional educational systems, are rich repositories of structured knowledge that cover a wide spectrum of subjects in a format meticulously curated for educational purposes. The incorporation of textbook content into our training regimen promises to imbue the model with a more profound understanding of academic concepts. This approach could potentially bridge the gap between academic theory and the practical application of knowledge, enhancing the model's ability to comprehend and respond to complex educational queries. Moreover, the diversity and depth of information contained in textbooks could significantly augment the model's versatility, enabling it to cater to a broader range of academic disciplines and learning levels. Thus, the inclusion of textbook data, in conjunction with our existing question bank resources, could mark a pivotal step towards developing more sophisticated and educationally tuned AI models.

5 Conclusion and Future Work

In this project, we have innovatively improved AI applications in education by developing a customized academic dataset and conducting a comprehensive survey of Large Language Model (LLM) tuning methods. Our approach, benchmarked against advanced models such as ChatGPT, resulted in significant performance improvements. A significant advancement was the integration of visual data processing with LLMs, enabling the model to interpret both textual and visual inputs. This multi-modal capability expands the usefulness of AI in education, especially in subjects that rely heavily on visual content. Our work lays the groundwork for more interactive and comprehensive AI-driven educational tools, and it establishes a foundation for future research on integrating advanced AI technologies in educational settings.

For future research directions, there are several avenues to enhance the capabilities and performance of our model, while also addressing current limitations and expanding its potential applications.

1. **Investigate Lightweight Fine-Tuning Methods:** Explore additional lightweight fine-tuning methods, such as QLora (Dettmers et al., 2023), to

address computational resource limitations. These methods can improve model performance while preserving efficiency.

2. **Integrate Image-Related Data:** Investigate suitable methods for integrating incorporate image-related data, such as figures and tables, into the model input. This integration can improve the model's ability to answer questions related to visual content, which is prevalent in various domains.

3. **Enhance Explanations with Reinforcement Learning:** Investigate the application of reinforcement learning from human feedback (RLHF) to enhance the quality of the detailed explanations generated by the model. This can lead to more informative and contextually relevant explanations.

4. **Collaborate with Junyi Academy:** Consider establishing collaboration with Junyi Academy (均一), as they have the potential to advance this research. Collaboration can lead to valuable insights and resources for further development.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, et al. 2023. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yixiao Li, Yifan Yu, Chen Liang, Pengcheng He, Nikos Karampatziakis, Weizhu Chen, and Tuo Zhao. 2023. Loftq: Lora-fine-tuning-aware quantization for large language models. *arXiv preprint arXiv:2310.08659*.

Yen-Ting Lin and Yun-Nung Chen. 2023. Language models for taiwanese culture. Code and models available at https://github.com/MiuLab/Taiwan-LLaMa.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Madeleine Udell and Alex Townsend. 2019. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160.