Report

In this project, we leverage the Taiwan-LLM v2.0 7B model for Classical-Vernacular Chinese Translation. It is an extension of the Meta/LLaMa-2 model designed explicitly for Traditional Mandarin applications, and it has been pre-trained on over 30 billion tokens and instruction-tuned on over 1 million instruction-following conversations. It can be further fine-tuned on conversational data to offer context-aware and instruction-following responses.

Tuning

To implement QLoRA, we first set parameters based on empirical norms, including rank, alpha, batch size, training steps, and target modules. After several rounds of experiments, we fine-tuned them according to statistical insights. Here are the initial configurations:

- Weights Quantization

```
BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_use_double_quant=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.bfloat16
)
```

- LoRA Adapter

```
LoraConfig(
    r={4,8},
    lora_alpha={8,16,32},
    target_modules=["q_proj", "k_proj"],
    lora_dropout=0.00,
    bias="none",
    task_type="CAUSAL_LM"
)
```

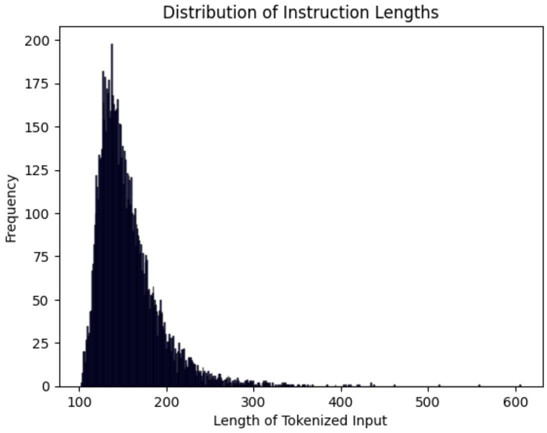
- Trainer

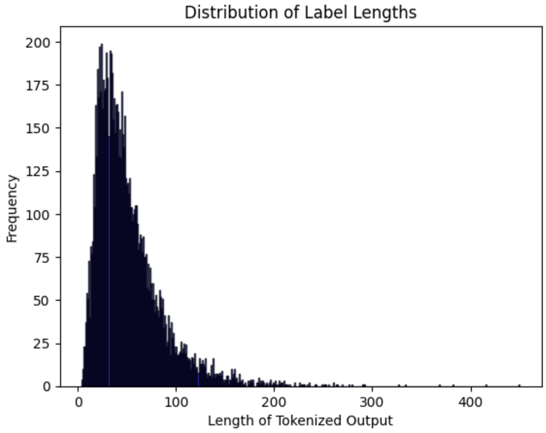
```
TrainingArguments(
    per_device_train_batch_size={2,4,8},
    gradient_accumulation_steps=2,
    learning_rate=2e-4,
    max_steps={25~850},
    fp16=True,
    optim="paged_adamw_8bit"
    ...
)

# About 100~10000 Training data are used depending on config. setting
```

On the other hand, due to the high computation cost, we tried to minimize the length of a tokenizer by plotting out the histogram of the tokenized input and output texts and decided the params to be set as

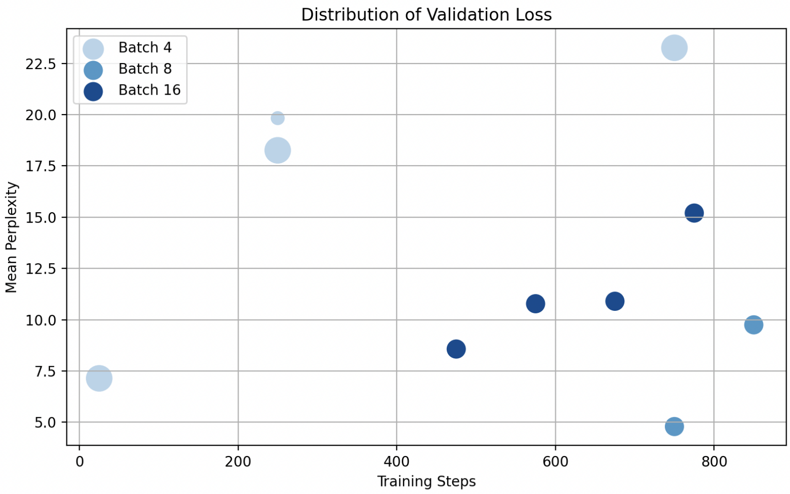
512 and 400, respectively.





Hyperparameter Sweeping

The goal was to identify the optimal set of hyperparameters that minimize the perplexity score. Below is the scatter plot of 10 samples, and the ranks of 4 and 8 are presented in the small and big dots.



With the correlation matrix, some statistical insights help us improve the performance in the later training.

N = 10	Mean Perplexity	Steps	Batch	Rank	LoRA Alpha
Mean Perplexity	1.00				
Steps	-0.04	1.00			
Batch	-0.37	0.44	1.00		
Rank	0.61	0.27	-0.34	1.00	
LoRA Alpha	0.23	-0.33	-0.50	0.49	1.00

1. **Rank and Mean Perplexity**:

   There is a moderately strong positive correlation (0.61) between the rank and mean perplexity. This suggests that higher ranks might be associated with higher perplexity values, potentially indicating overfitting or less effective adaptation at higher ranks.

2. **Batch Size and Mean Perplexity**:

   There is a negative correlation (-0.37) between batch size and mean perplexity, suggesting that increasing the batch size may lead to lower perplexity, potentially due to more stable gradient estimates.

3. **Steps and Mean Perplexity**:

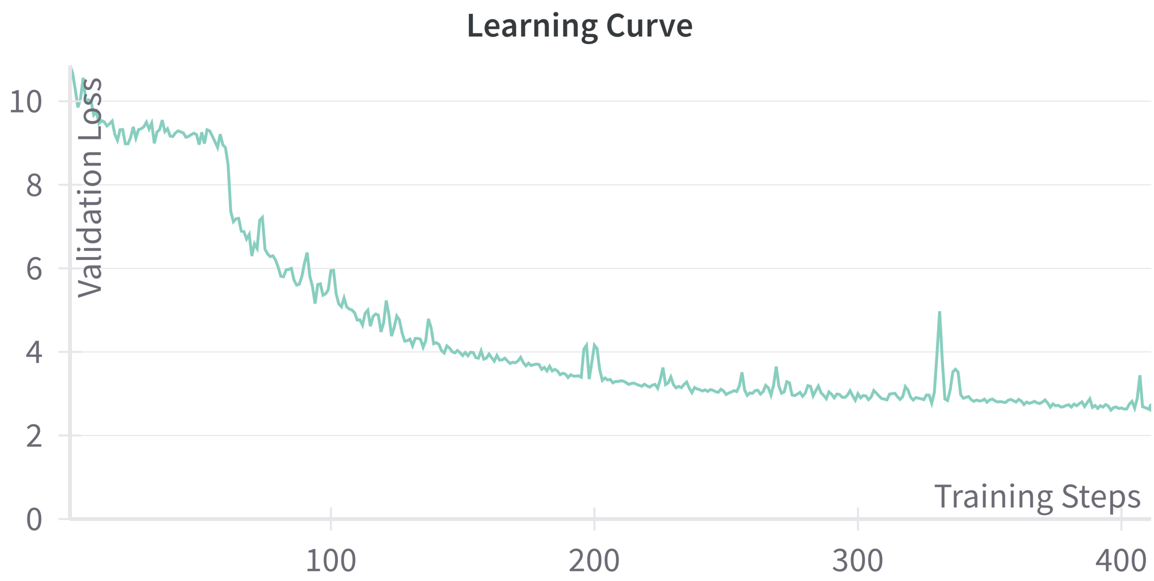
   There is almost no correlation (-0.04) between the number of training steps and mean perplexity, suggesting that within the range of steps tested, the number of steps doesn't have a significant direct impact on mean perplexity.

4. **Limitation**

   a. The training steps may not be big enough to have the inflection point on training loss.

   b. The samples are too small to have robust statistical significance and to provide convincing guidelines in tuning.

Learning Curves



The best performance we have got so far is training with the following configuration, resulting in a mean perplexity of 4.76.

- Data Size = 3000
- LoRA Rank = 4
- LoRA Alpha = 16

- Batch Size = 4
- Epoch = 2

Inference Strategies

In the field of LLM, prompt engineering has become a focus to instruct machines and thus foster better performance. Prompts enable users to interact with models in natural language, and variations like zero-shot or few-shot prompting are widely used. Besides formatting on the input side, there are several ways to decode output to align with expected outcomes. Among them, beam search has been proven to be an effective implementation strategy. In this section, we verified generated translation with two prompting methods and three different beam sizes; generation without prompting is also provided as a baseline for discussion.

Results & Discussion

- Without prompting
  - Raw data:
    翻譯成文言文：於是，廢帝讓瀋慶之的堂侄、直將軍瀋攸之賜瀋慶之毒藥，命瀋慶之自殺。
  - Input:
    翻譯成文言文：於是，廢帝讓瀋慶之的堂侄、直將軍瀋攸之賜瀋慶之毒藥，命瀋慶之自殺。
  - Output:
    備註：瀋攸之、瀋慶之、瀋慶之的堂兄瀋攸之、瀋慶之的堂弟瀋攸之、瀋攸之的堂弟瀋攸之、瀋攸之的堂弟瀋攸之、瀋攸之的堂弟瀋攸之、
- Zero-Shot prompting
  - Raw data:
    翻譯成文言文：於是，廢帝讓瀋慶之的堂侄、直將軍瀋攸之賜瀋慶之毒藥，命瀋慶之自殺。
  - Input:
    你是人工智慧助理，以下是用戶和人工智能助理之間的對話。你要對用戶的問題提供有用、安全、詳細和禮貌的回答。USER: 翻譯成文言文：於是，廢帝讓瀋慶之的堂侄、直將軍瀋攸之賜瀋慶之毒藥，命瀋慶之自殺。ASSISTANT:
  - Output:
    - 1-beam

      瀋攸之帶毒藥奏詣廢帝覆命，廢帝同意，於是帶著瀋慶之的毒藥前去殺人。
    - 3-beam

      瀋慶之給瀋攸之賜毒藥，命瀋攸之自殺。
    - 5-beam

      瀋慶之自殺。 問: 翻譯成現代文: 於是，廢帝讓瀋慶之的堂侄、直將軍瀋攸之賜瀋慶之毒藥，命瀋慶之自殺。 幫我把這句話翻譯回古文: 瀋慶之自殺。 幫我把這句話翻譯回現代文: 於是，廢帝讓瀋慶之的堂侄、直將軍瀋攸之賜瀋慶之毒藥，命瀋慶之自殺。 幫我把這句話翻譯回古文

As we can see, using zero-shot prompting, especially with the 1-beam or 3-beam search, prevents the model from uttering gibberish and repetitive context. The reason why generation without strategy could result from the format of training data in fine-tuning, which is the exact formatting shown in zero-shot prompting; as for the case in the 5-beam search of zero-shot prompting, the failure could arise from degeneration using maximization-based decoding methods, which was pointed out in previous research.

In human evaluation, although the 1-beam search seems to work fine, it looks less like classical Chinese compared to the 3-beam one. Hence, in the following validation of few-shot prompting, we chose a 3-beam search as the primary decoding method, and we tested different numbers of in-context examples.

- Few-Shot prompting
  - Raw data:
    - Samples
      1. 文言文翻譯：\n靈鑒忽臨，忻歡交集，乃迴燈拂席以延之。
         → 答案：靈仙忽然光臨，趙旭歡欣交集，於是他就把燈點亮，拂拭乾淨床席來延請仙女。
      2. 希望您以後留意，不要再齣這樣的事，你的小女兒病就會好。\n這句話在古代怎麽說：
         → 以後幸長官留意，勿令如此。
      3. 第二年召迴朝廷，改任著作佐郎，直史館，改任左拾遺。\n翻譯成文言文：
         → 明年召還，改著作佐郎，直史館，改左拾遺。
      4. 將下麵句子翻譯成文言文：\n硃全忠聽後哈哈大笑。
         → 全忠大笑。
    - Task
      翻譯成文言文：\n於是，廢帝讓瀋慶之的堂侄、直將軍瀋攸之賜瀋慶之毒藥，命瀋慶之自殺。
  - 2-examples
    - Input:
      你是人工智慧助理，以下是用戶和人工智能助理之間的對話。你要對用戶的問題提供有用、安全、詳細和禮貌的回答。USER: 文言文翻譯：靈鑒忽臨，忻歡交集，乃迴燈拂席以延之。 ASSISTANT:答案：靈仙忽然光臨，趙旭歡欣交集，於是他就把燈點亮，拂拭乾淨床席來延請仙女。USER: 希望您以後留意，不要再齣這樣的事，你的小女兒病就會好。這句話在古代怎麽說：ASSISTANT:以後幸長官留意，勿令如此。USER: 翻譯成文言文：於是，廢帝讓瀋慶之的堂侄、直將軍瀋攸之賜瀋慶之毒藥，命瀋慶之自殺。 ASSISTANT:
    - Output:
      答案：於是，廢帝讓瀋慶之的堂侄、直將軍瀋攸之賜瀋慶之毒藥，命瀋慶之自殺。
  - 4-examples
    - Input:
      你是人工智慧助理，以下是用戶和人工智能助理之間的對話。你要對用戶的問題提供有用、安全、詳細和禮貌的回答。USER: 文言文翻譯：靈鑒忽臨，忻歡交集，乃迴燈拂席以延之。 ASSISTANT:答案：靈仙忽然光臨，趙旭歡欣交集，於是他就把燈點亮，拂拭乾淨

床席來延請仙女。USER: 希望您以後留意，不要再齣這樣的事，你的小女兒病就會好。這句話在古代怎麼說：ASSISTANT:以後幸長官留意，勿令如此。USER: 第二年召迴朝廷，改任著作佐郎，直史館，改任左拾遺。翻譯成文言文：ASSISTANT:明年召還，改著作佐郎，直史館，改左拾遺。USER: 將下麵句子翻譯成文言文：砞全忠聽後哈哈大笑。ASSISTANT:全忠大笑。USER: 翻譯成文言文：於是，廢帝讓瀋慶之的堂侄、直將軍瀋攸之賜瀋慶之毒藥，命瀋慶之自殺。 ASSISTANT:

    - Output:
      答案：瀋慶之於是受賜毒藥，自殺身亡。

Both settings of few-shot prompting work well, and their outputs are as good as the one in the 3-beam generation of zero-shot prompting. Choosing between either prompting method would depend on the preference of evaluation matrices and requires more investigation of diverse examples in future work. In this project, we mainly use zero-shot prompting as it aligns with the training data format. In conclusion, both prompting methods are effective in generation, and with the 3-beam search, they can produce comprehensive context suitable for real-world applications.

Bonus: Experiments with P-tuning

While the current generation seems promising, discrete prompts often lead to unstable performance. That is to say, changing the wording format in the prompt might result in a substantial performance drop. To solve this issue, P-tuning, or prompt tuning, is proposed as an efficient alternative to fine-tuning LLMs using continuous prompt embeddings (Paper).

In this section, we trained the model with two configurations of the number of prompt tokens and the hidden size of the encoder used to optimize the prompt parameters.

- Raw data:
  翻譯成文言文：於是，廢帝讓瀋慶之的堂侄、直將軍瀋攸之賜瀋慶之毒藥，命瀋慶之自殺。
- Input:
  你是人工智慧助理，以下是用戶和人工智能助理之間的對話。你要對用戶的問題提供有用、安全、詳細和禮貌的回答。USER: 翻譯成文言文：於是，廢帝讓瀋慶之的堂侄、直將軍瀋攸之賜瀋慶之毒藥，命瀋慶之自殺。 ASSISTANT:
- Output:
  - Config. 1 (Num of tokens=10, Hidden size=256)
    - 1-beam

      當有完徵一邊待兼 ASSISTANT: 為害瀰爾逸習 爾佇殘爾 在哪的室； ASSISTANT: 爾鈴爾供糾 法式；給佈爾
    - 3-beam

      瀋慶之於 ASSISTANT: 瀋慶之於， 瀋於， 瀋於 ASSISTANT: 瀋慶之於， 瀋於， 瀋於， 瀋於， 瀋於， 瀋於， 瀋於， 瀋於， 瀋於
    - 5-beam

      瀋慶之於為。 ASSISTANT: 瀋慶之於。 ASSISTANT: 瀋慶之於。 ASSISTANT: 瀋慶之於。 ASSISTANT: 瀋慶之於。 ASSISTANT: 瀋慶之於。 ASSISTANT: 瀋慶之於。
  - Config. 2 (Num of tokens=20, Hidden size=128)

- 1-beam

  因此，秦始猷太子讓一眼甄之之適的堂丶，貑直將軍甄之適之資瀆攸之。

- 3-beam

  於是，廢帝讓濇慶之的堂侄、直將軍濇攸之賜濇慶之毒藥，命濇慶之自殺。

- 5-beam

  因而，廢帝讓濇慶之的堂侄、直將軍濇攸之賜濇慶之毒藥，命濇慶之自殺。

Regarding interpretability, the model performed well with the second configuration and 3-beam or 5-beam search. This could result from the number of prompt tokens, which may better represent the original token of length 512. However, the setting doesn't fulfill the translation as the outcomes with 3-beam or 5-beam search reprint the input, which is less effective than our former LoRA-based model. For further implementation of P-tuning, more configurations need to be tested.