ADL HW3 Report

Student ID: m11203404
Name: 陳旭霖

Q1: LLM Tuning

- Describe

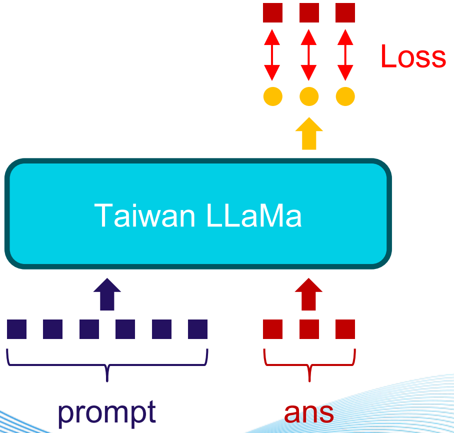
Q: How much training data did you use?

A:I utilized a total of 4096 training samples, with the aim of having it evenly divisible by a batch size of 16.

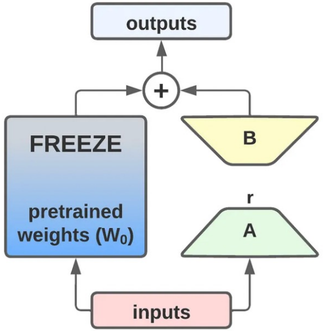
Q: How did you tune your model?

A:I employ LoRA and 4-bit Quantization techniques for Instruction Tuning on Taiwan-LLaMa. Instruction Tuning involves using detailed prompts during training to help the language model understand the concepts of the described problem. The model's output is then compared to the actual answer, and the loss is calculated accordingly.

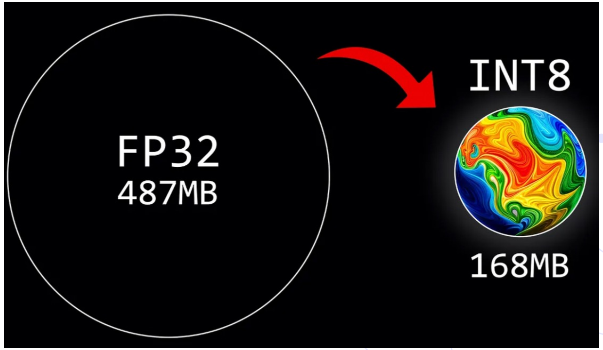
- **Instruction tuning**



- **LoRA**



- **Quantization**



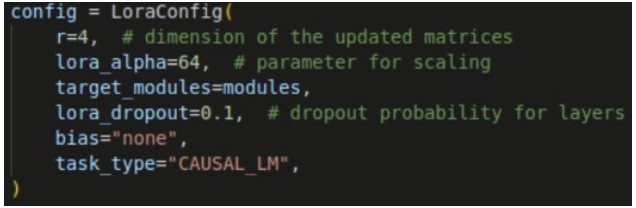
Q1: LLM Tuning

- Describe

Q: What hyper-parameters did you use?

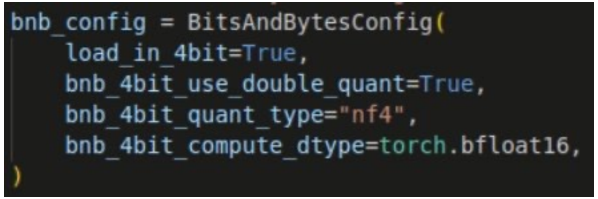
A:

- **Peft config**



```
config = LoraConfig(
    r=4,  # dimension of the updated matrices
    lora_alpha=64,  # parameter for scaling
    target_modules=modules,
    lora_dropout=0.1,  # dropout probability for layers
    bias="none",
    task_type="CAUSAL_LM",
)
```

- **bnb config**



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

- **Training parameters**

1. Steps: 256
2. Batch_size: 16
3. Warmup_steps: 5
4. Learning_rate: 2e-4
5. Optimizer: paged_adamw_8bit

Q1: LLM Tuning

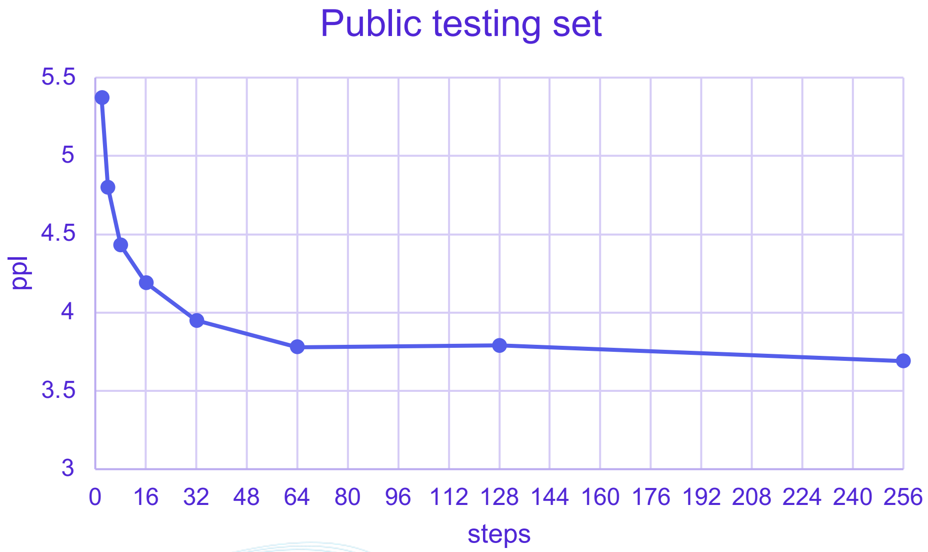
- Show your performance:

Q: What is the final performance of your model on the public testing set?

A: ppl → 3.69

Q: Plot the learning curve on the public testing set

A:



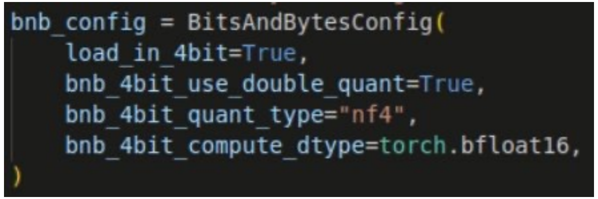
Q2: LLM Inference Strategies

- Zero-Shot:

Q: What is your setting? How did you design your prompt?

A:

- **bnb config**



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

- **Inference parameters**

1. Model: Taiwan-LLM-7B-v2.0-chat
2. Tokenizer: Taiwan-LLM-7B-v2.0-chat
3. Data type: Bfloat16
4. Max_new_tokens: 512

- **Prompt**

文言文主要指以秦漢書面語為標準，脫離口語而寫成的文字。 語言學上，先秦及西漢時使用的語言被稱為上古漢語。 此時傳世的一些文獻，如《詩經》《論語》《左傳》《韓非子》《史記》 ，被視為文言文的範本。USER: {instruction} ASSISTANT:

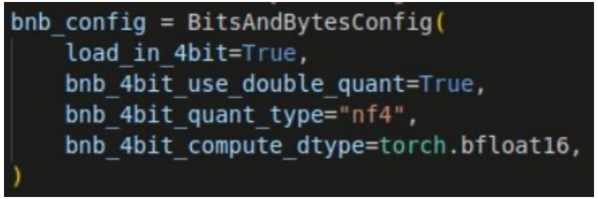
Q2: LLM Inference Strategies

- Few-Shot:

Q: What is your setting? How did you design your prompt?

A:

- **bnb config**



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

- **Inference parameters**

1. Model: Taiwan-LLM-7B-v2.0-chat
2. Tokenizer: Taiwan-LLM-7B-v2.0-chat
3. Data type: Bfloat16
4. Max_new_tokens: 512
5. Shot: 3

- **Prompt**

以下會給你幾段文言文與白話文翻譯的範例，請你學習其中規則，並翻譯。

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: 文言文翻譯： 中宗與庶人嘗因正月十五日夜幸其第，賜賚不可勝數。ASSISTANT: 答案：唐中宗與韋庶人曾經在正月十五日夜到韋安石的宅第，並賜賞給他不可勝數的財物。

USER: {instruction} ASSISTANT:

Q2: LLM Inference Strategies

- Few-Shot:

Q: How many in-context examples are utilized? How you select them?

A: I utilized a total of 3 in-context examples. This decision was made after testing, as I observed that increasing the number to 10 did not yield better results. Therefore, I opted for 3 in-context examples.

Q2: LLM Inference Strategies

- Comparison:

Q: What’s the difference between the results of zero-shot, few-shot, and LoRA?

A: Regardless of whether it's zero-shot or few-shot generation, the results have not been very satisfactory. It's possible that the Classical Chinese translation task is too complex, and it's challenging to address such issues without adjusting the weights. However, when employing LoRA technology to adjust the weights, a significant improvement in translation performance can be observed.

Command: "文言文翻譯：\n靈鑒忽臨，忻歡交集，乃迴燈拂席以延之。"

Zero-shot: "靈鑒忽臨，忻歡交集，乃迴燈拂席以延之。"

Few-shot: "答案：靈鑒忽臨，忻歡交集，乃迴燈拂席以延之。"

LoRA: "答案：靈鑒忽然現身，忻歡招待他們交朋友，靈鑒就把燈籠拿著走，想迴去營席。"

GT: "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。"