ADL HW3 Report B09611048

1. How much training data do you use?
   I used 3000 and 4000 training data in this training. On the other hand, I used 750 and 1000 data for validation respectively.

   The result is listed below.

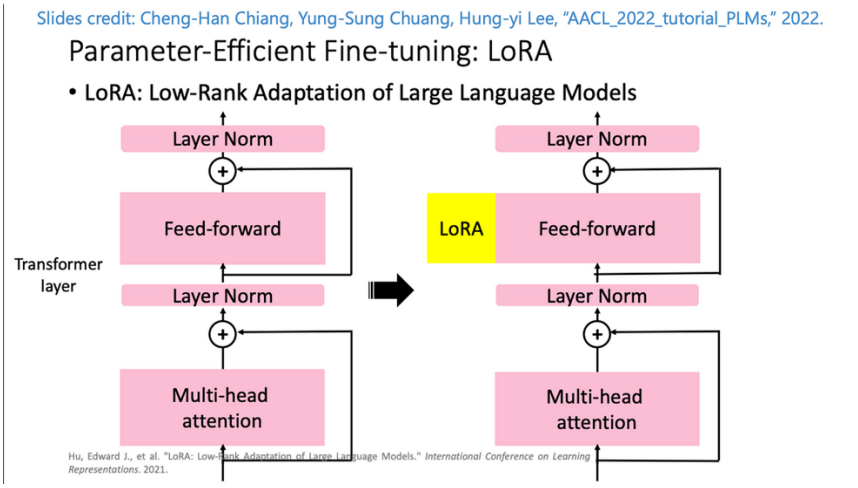
Data size	steps	eval_loss	ppl
3000	500	1.1567	4.748
4000	500	1.3138	3.646

   I chose to train 4000 data in the end, the numbers of data will strongly effect the performance of the model.

2. How did you tune your model?
   In this computational generation, finetuning a whole model cost a lot computational power, which is rather inefficient when you are dealing with a new task.

   To encounter the problem, in this homework, we uses another fine-tuning method that is relatively efficient, which is called PEFT (Parameter-Efficient Fine-Tuning). In this assignment, we use Low-Rank Adaptation (Lora) as the method.

   The concept of Lora is that freeze the weights of the original model (Let's say it is Taiwan Llama checkpoint in this assignment). And it will train a little model, acting as some patches or plus-ins. The adapt little model will do little adjust on the whole LLM while generating. And this method will than have similar actions comparing to finetuning the whole model. By doing so, we came decrease the parameters and the computational power while finetuning the model.



Concept of Lora

   As for Qlora we used in this assignment, it added the feature Quantile quantization in the model which will change the float numbers into a special datatype NF4, which can

reduce memory usage. (Image you use approximated integers to store float numbers, which can reduce the memory from 32 bytes to 8 bytes, quantization is doing the same concept.)

3. What hyper-parameters did you use?

lora rank	4
lora_alpha	16
lora_dropout	0.0
Learning rate	2e-4
Training steps	500
max_training_dataset	4000
max_eval_dataset	1000
per_device_train_batch_size	1
gradient_accumulation_steps	16
lr_scheduler_type	'constant'

Bnb-config:

```
BitsAndBytesConfig(
    load_in_4bit=True,
    llm_int8_threshold=6.0,
    llm_int8_has_fp16_weight=False,
    bnb_4bit_compute_dtype=torch.float32,
    bnb_4bit_use_double_quant=True,
    bnb_4bit_quant_type='nf4',
)
```

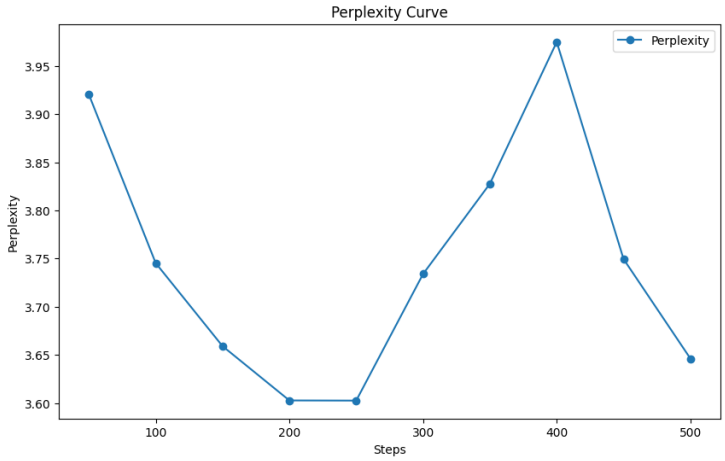
4. What is the final performance of your model on the public testing set? (2%)
I used two kinds of prompts for training:

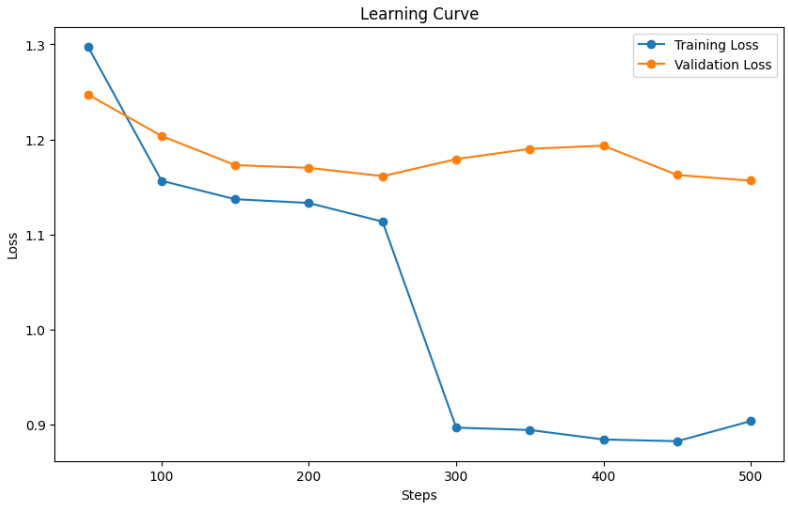
Prompts	ppl (500 steps)
"你是人工智慧助理，以下是用戶和人工智能助理之間的對話。你要對用戶的問題提供有用、安全、詳細和禮貌的回答。USER: {instruction} ASSISTANT:"	3.60
"你是人工智慧國文大師，以下是用戶和人工智能助理之間的對話。你要對用戶的問題提供有用、安全、詳細和禮貌的回答。USER: {instruction} ASSISTANT:"	3.59

In my test, different prompts don't really affect the performance of the PEFT model. The second prompt is only a little higher than the second prompt.

5. Plot the learning curve on the public testing set (2%)

steps	perplexity	Training loss	Validation loss
50	3.9203932566642763	1.297800	1.247280
100	3.7452657055854797	1.156600	1.203718
150	3.6590152835845946	1.137200	1.173057
200	3.6026388711929322	1.133200	1.170200
250	3.602478358745575	1.113600	1.161433
300	3.7339287343025207	0.896900	1.179391
350	3.8277029271125795	0.894400	1.190202
400	3.9748926978111268	0.884400	1.193541
450	3.749453022003174	0.882600	1.162635
500	3.6461181592941285	0.903800	1.156736





When the step number increases, the perplexity and the eval loss has slightly increase as well, which cause by having the problems like over fitting. I choose the checkpoint with the lowest perplexity, which also has the lowest eval loss as well.

6. Zero shot

Prompts:

Prompt 1	"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: {instruction} ASSISTANT:"
Prompt 2	你是人工智慧國文大師，以下是用戶和人工智能國文大師之間的對話。請根據用戶的問題輸出答案，將文言文翻譯成現代文或將現代文翻譯成文言文。USER: {instruction} ASSISTANT:

Perplexity:

Prompt	Perplexity
Prompt 1	5.453014215946197
Prompt 2	5.161542772769928

In the observation, calling the LLM a "**master**" can help the performance of the model.

7. Few shot

I use Prompt 2 in the zero shot section as my base prompt.

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

I use three examples for each prompt, for the first prompt. I will give it rather short contexts for each examples. And in the second prompt, I will choose those with the structure that is more like a complete story, which will be rather long.

And for prompt 3, I add one more examples to prompt 1.

Prompt 1	"你是人工智慧國文大師，以下是用戶和人工智能國文大師之間的對話。請根據用戶的問題輸出答案，將文言文翻譯成現代文或將現代文翻譯成文言文。以下是三個例子，1. 古文：是歲，京師及州鎮十三水旱傷稼。 現代文：當年，京都及各州鎮十三處發生水旱災害，損傷莊稼。2. 古文：自今令在必行，毋有所遏。 現代文：今天令在必行，不得有任何阻攔。3. 古文：在那時候，秦、晉是強國 現代文：當是之時，秦晉為彊國。USER: {instruction} ASSISTANT:"
Prompt 2	"你是人工智慧國文大師，以下是用戶和人工智能國文大師之間的對話。請根據用戶的問題輸出答案，將文言文翻譯成現代文或將現代文翻譯成文言文。以下是三個例子，1． 古文：鞦九月，親祀南郊，加尊號天冊金輪聖神皇帝，大赦天下，改元為天冊萬歲，大闢罪已下及犯十惡常赦所不原者，鹹赦除之，大酺九日。現代文：鞦九月，皇上親自在鼕至日到南郊祭天，加尊號曰天丹金輪聖神皇帝，大赦天下，改元為天冊萬歲，死罪以下以及犯瞭十大罪惡、常規赦免而不能饒恕的人，都加以赦免，特許天下百姓大宴飲九天。2． 古文：棄疾為憲時，嘗攝帥，每嘆曰： 福州前枕大海，為賊之淵，上四郡民頑獷易亂，帥臣空竭，急緩奈何！ 至是務為鎮靜，未期歲，積錢至五十萬緡，榜

	曰： 備安庫。 現代文：辛棄疾為提點刑獄時，曾主持軍事，每每嘆道 ：福州前麵是海，為盜賊藏身之所，上四郡的百姓頑獷易亂，而帥府空虛，一旦有事，怎麼辦？ 於是他以安撫為務，積極儲備，不到一年，積錢五十萬緡，稱 備安庫。3. 古文：後鴉仁兄子海珍知之，掘曷父伯道並祖及所生母閤五喪，各分其半骨，共棺焚之，半骨雜他骨，作五袋盛之，銘袋上曰 荀曷祖父母某之骨。現代文：後來，羊鴉仁哥哥的兒子羊海珍得知情況，便挖開荀曷的父親荀伯道、他祖父母以及生母共五個死者的墳墓，各取齣一半骸骨，放在一起和棺材同時燒掉，又用剩餘的一半屍骨加上彆的亂七八糟的骨頭混在一起，分彆裝入五隻袋子，在袋子的錶麵上寫著 荀曷祖父母某人之骨 。USER: {instruction} ASSISTANT”
Promp 3	“你是人工智慧國文大師，以下是用戶和人工智能國文大師之間的對話。請根據用戶的問題輸出答案，將文言文翻譯成現代文或將現代文翻譯成文言文。以下是三個例子，1. 古文：是歲，京師及州鎮十三水旱傷稼。 現代文：當年，京都及各州鎮十三處發生水旱災害，損傷莊稼。2. 古文：自今令在必行，毋有所遏。 現代文：今天令在必行，不得有任何阻攔。3. 古文：在那時候，秦、晉是強國現代文：當是之時，秦晉為彊國。 4. 古文：辛未，命吳堅為左丞相兼樞密使，常楙參知政事。現代文：初五，命令吳堅為左承相兼樞密使，常增為參知政事。USER: {instruction} ASSISTANT:”

Perplexity:

Prompt	Perplexity
Prompt 1	4.92259840965271
Prompt 2	4.920815424442291
Prompt 3	4.873627830505371

From the results, we can have the conclusion that the length of the examples does no effect the performance of the model. But the increase of the number of examples may help.

8. Difference between Zero shot, few shot and Qlora.

Perplexity:

Method	
Zero-shot basic prompt	5.453014215946197
Zero-shot master prompt	5.161542772769928
Few-shot 3-example prompt	4.92259840965271
Few-shot 4-example prompt	4.873627830505371
Qlora (checkpoint-250 steps)	3.602478358745575

Comparing the outputs for the same instructions:

Instruction: 我當時在三司，訪求太祖、仁宗的手書敕令沒有見到，然而人人能傳誦那些話，禁止私鹽的建議也最終被擱置。

Method	Result
Zero-shot master prompt	我當時在三司，訪求太祖、仁宗的手書敕令沒有見到，然而人人能傳誦那些話，禁止私鹽的建議也最終被擱置。
Few-shot 4 example promt	我當時在三司，訪求太祖、仁宗的手書敕令沒有見到，然而人人能傳誦那些話，禁止私鹽的建議也最終被擱置。
Qlora	時在三司，敕不見，人誦之，禁止不為。

From our observation, some of the models will encounter time that it didn't do the task. And Qlora has the most specific generation and performance