COMP4901B: Large Language Models

Assignment 1 Report

HE, Wenqian
Student ID: 20860896

October 3, 2025

1 Part 1: Data Preprocessing

1.1 Cleaning Logic

For each paragraph in the text, I do the following cleaning steps:

- if there are more than 1 whitespaces between words, I replace them with a single whitespace
- strip the paragraph of extra whitespace
- remove paragraphs that have too few words (less than 2 words)
- remove paragraphs that are mostly non-alphanumeric, such as ?,/,-, etc. (less than 50% alphanumeric characters)
- remove paragraphs that have too many repeated characters in one word, such as "aaaaaa", "bbbbbb", etc. (more than 5 consecutive characters)

1.2 Heuristic Quality Filter Logic

I count the number of bad words in the text, if there are more than the number of bad words threshold (I use 1 as the threshold), I reject the text.

1.3 English Text Detection Logic

First, I filter out the alphabetic characters, such as "a", "B", "我", etc., excluding the symbols like "?", "/", "-", etc.
Then, from the remaining alphabetic characters, I count the number of English alphabetic characters (i.e. "a-zA-Z").
If the ratio of English alphabetic characters to the total number of alphabetic characters is greater than the English character ratio threshold (I use 0.9 as the threshold), I accept the text.

1.4 Results

Here is the record numbers of WARC records processed and the number that pass all filters:

```
9 passed out of 30 records processed.
Cleaned documents saved to: cleaned_test.txt
121 deduplicated out of 219 records processed.
```

2 Part 2: Pretraining

The model is trained on Mac book with mps device enabled, which is achieved by modifying the `run_llama.py` file. This training takes about 9 hours.
Here is the training command:

```
python run_llama.py \
  --run_name run6-fix-loss \
  --option pretrain \
  --data_path train_100M \
  --block_size  256 \
  --batch_size 512 \
  --micro_batch_size 32 \
  --epochs 1 \
  --tokenized_dir train_100M/tokenized \
  --use_gpu  \
  --val_path dev \
  --val_tokenized_dir dev/tokenized \
  --val_per_steps 200 \
  --test_path  test \
  --test_tokenized_dir test/tokenized \
  --auto_resume \
  --warmup_ratio 0.1 \
  --lr 1e-3
```

The training loss declines quickly at the beginning, and then climb up until step 300, finally keep declining stably.
Here are some important curves:

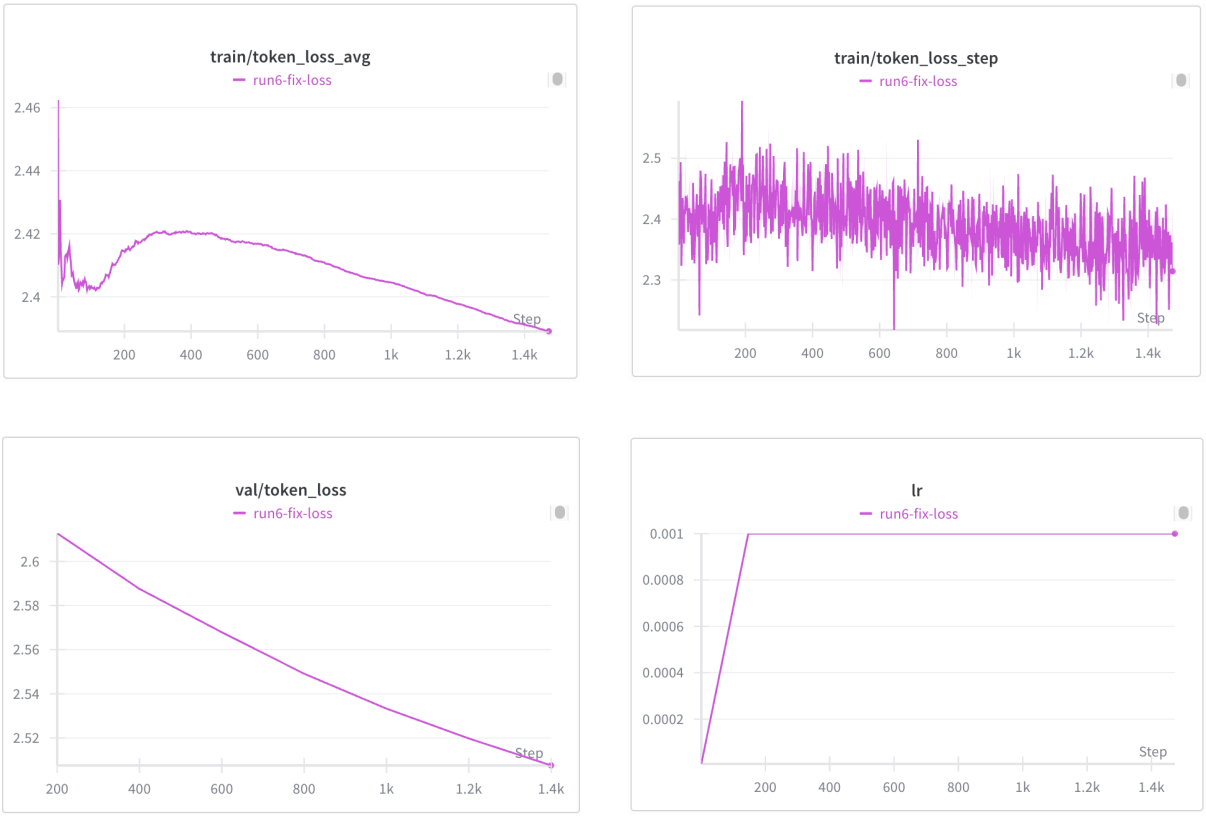

Figure 1: Training metrics visualization

3 Part 3: Generation

Provided model:

```
python run_llama.py --pretrained-model-path llama2-42M-babylm.pt --option generate

// temperature = 0.0
White Bird is a 2023 American war drama movie starring Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt,

// temperature = 0.5
White Bird is a 2023 American war drama movie starring James Dee, Eddie Quinn, Larry Dee, James McDonald, George J. L. Smith, David Dee, John Duck, James McDonald, John H. W. Bush. It was distributed by 20th Century Fox. = = = PG4021 = = = A LITTLE
```

My pretrained model:

```
python run_llama.py --pretrained-model-path run6-fix-loss-pretrain-1-0.001.pt --option generate

// temperature = 0.0
White Bird is a 2023 American war drama movie starring John Wayne, and is the second movie of the same name by John Wayne. It was directed by John Wayne. = = = 2023-2023 season = = = The 2023-2023 season was the 10th season of the National Hockey League (NHL). It was the

// temperature = 0.5
White Bird is a 2023 American war drama movie starring Gary Leigh. It was directed by John Wilder. = = = Cade = = = Cade is a municipality in the province of Limburg in the canton of Limburg in Switzerland. = = = Cade County, Pennsylvania = = = Cade County is a county in the U.S. state of Pennsylvania.
```

From the generation results, although using temperature = 0.0 is more coherent, such as the generated sentence from my pretrained model, which echos "John Wilder" again, I think the temperature = 0.5 is better. Using greedy method will lead to very repetitive and boring generation, as shown in the generated sentence from the provided model. Using temperature sampling will lead to more diverse and interesting generation.