Project 3

Gabriel Ferreira
gabferre@iastate.edu

1 Methods

This section outlines the methodology adopted to develop and evaluate a title generation model using the Llama3.2 3B architecture. The primary objective of this task is to generate relevant and coherent titles for research papers based solely on their abstracts. These papers pertain to animal QTL (Quantitative Trait Loci) studies. The methodology described here includes dataset, data pre-processing, Llama3.2 3B model, training configuration and procedures, and evaluation metrics.

1.1 Dataset

To support the title generation task, two labeled datasets were provided: (1) a training set for supervised learning, and (2) a held-out set for evaluation. Each instance consists of a research paper abstract paired with its corresponding title, and is represented as:

$$D = \{(a_i, t_i)\}_{i=1}^{N}$$

where a_i is the abstract of the i-th paper, serving as the input, and t_i is the target title to be generated.

The training dataset contains $N = 11,278$ entries, each with four fields: 'PMID', 'Title', 'Abstract', and 'Category'. For the purpose of this task, only the 'Abstract' and 'Title' fields were used.

The evaluation dataset contains $M = 1,097$ entries with the same structure, and was used to assess model performance. During evaluation, only the abstracts were provided to the model for inference, and the generated titles were compared against the ground truth titles to compute quality metrics.

1.2 Preprocessing for Llama3.2

To fine-tune Llama3.2 for the title generation task, the following preprocessing steps were applied to prepare the data in a format suitable for causal language modeling:

- **Prompt Formatting:** Each training instance was formatted into a natural language prompt with the structure: "`You are an expert at writing concise and informative scientific titles. \nAbstract: <abstract> \nTitle:`".

  The model was trained to generate the corresponding title following the prompt.

- **Tokenization with LLaMA Tokenizer:** All prompts and titles were tokenized using the Llama3.2 tokenizer, which converts the text into input IDs and target labels compatible with the model's vocabulary.

- **Sequence Concatenation:** Input (prompt) and target (title) sequences were concatenated to form a single token sequence. The model was trained to predict the tokens corresponding to the title, conditioned on the abstract.

- **Padding and Truncation:** Sequences were truncated or padded to a fixed maximum length to ensure consistent input size, with truncation applied from the left to retain the most relevant tokens near the target.

- **Train-Validation Splitting:** The dataset was randomly split into training (80%) and validation (20%) sets to facilitate model tuning and generalization assessment.

1.3 Model Architecture (Llama3.2-based)

For this title generation task, we fine-tuned the Llama3.2 3B model, a decoder-only transformer model designed for autoregressive text generation. Llama3.2 excels at generating coherent and contextually relevant sequences conditioned on prior input.

The model takes as input a formatted prompt, which is tokenized and fed into the transformer. The model is trained to autoregressively predict the next tokens that complete the title, using its learned internal representations of the abstract.

Specifically, we use the standard output sequence generated by Llama3.2 following the abstract prompt. The loss is computed over the title tokens only, using a causal language modeling

(CLM) objective with cross-entropy loss. No additional layers are stacked on top of Llama3.2; instead, the model learns to directly generate the title text by predicting each token in sequence, conditioned on the abstract.

This structure leverages Llama3.2's full transformer stack, where self-attention across the entire prompt and generated tokens allows it to maintain coherence and relevance throughout the generated title.

1.4 Training Parameters and Configuration

The Llama3.2 3B model was fine-tuned using the Hugging Face Transformers and PEFT (Parameter-Efficient Fine-Tuning) libraries with LoRA (Low-Rank Adaptation) to reduce memory usage and accelerate training. The training process was conducted on a single NVIDIA A100 GPU.

The main configuration details are as follows:

- **Batch Size:** 1 (with gradient accumulation steps set to 8 to simulate a batch size of 8)
- **Max Sequence Length:** 512 tokens
- **Learning Rate:** 2e-4 with cosine learning rate scheduler
- **Number of Epochs:** 1 (Only one epoch given a mix of early convergence and compute constraints)
- **Optimizer:** AdamW
- **Loss Function:** Cross-entropy loss computed over the title tokens only
- **Precision:** bfloat16 mixed precision enabled for faster training and reduced memory usage
- **LoRA Configuration:** Rank = 64, α = 16, Dropout = 0, applied to attention and projection layers

Evaluation on the validation set was performed after each epoch to monitor performance and avoid overfitting.

1.5 Evaluation Metrics

- **BLEU:** Measures n-gram precision between generated and reference titles, emphasizing exact word matches.
- **ROUGE-2:** Calculates bigram overlap to assess phrase-level similarity with a focus on recall.
- **ROUGE-L:** Evaluates the longest common subsequence, capturing fluency and structural similarity.

2 Results

The fine-tuned Llama3.2 7B model was evaluated using a held-out test set of research abstracts. The results from the validation phase and final evaluation on the test set are presented next.

2.1 Training and Validation Loss

The model was trained for a single epoch due to early convergence, as shown in Figure 1. Both training and validation losses steadily decreased and stabilized over time, indicating effective learning without signs of overfitting.

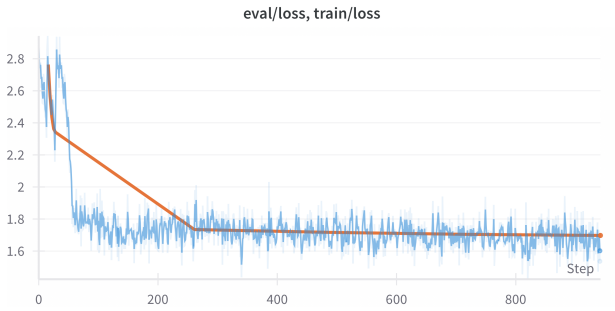

Figure 1: Training (blue line) and validation (orange line) loss throughout the fine-tuning process.

2.2 Generation Performance

After fine-tuning, the model was evaluated on a held-out test set of 1,097 abstracts. Inference was conducted using the fine-tuned Llama3.2 3B model with the custom prompt template defined earlier.

The generation quality was assessed using the metrics described above. The results are reported in Table 1.

Metric	Score
BLEU	0.1322
ROUGE-2	0.2575
ROUGE-L	0.4135

Table 1: Evaluation metrics on the test set.

The scores demonstrate that the fine-tuned model successfully learns to generate meaningful and partially overlapping titles. While the BLEU score indicates limited exact token overlap, the higher ROUGE-L suggests good structural similarity and relevance between predicted and reference titles.

References

[1] DataCamp. Fine-Tuning LLaMA 3 on Your Own Data. Available at: https://www.datacamp.com/tutorial/fine-tuning-llama-3-1.

[2] Hugging Face. Transformers Library. Available at: https://huggingface.co/transformers/.

[3] Hugging Face. PEFT: Parameter-Efficient Fine-Tuning. Available at: https://github.com/huggingface/peft.

[4] Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*. Available at: https://arxiv.org/abs/2106.09685.

Assistance from AI (ChatGPT) was used for refining the structure and wording of this report.