SecretLLM Project Report

Aleksei Buvailik
TU Dresden
Matriculation number: 5271683
Codabench nickname: albu670g
email@domain.tld

Abstract

This report tracks my iterative development of a QA system built around LoRA adapters, structured prompting, and retrieval-augmented context. I detail how each change affected MCQ and SAQ accuracy, including parsing refinements, logprob scoring, and RAG variants. The final results summarize the best combined configuration and the limits of the current approach.

1 Introduction

I develop a QA pipeline that fine-tunes a base LLM with LoRA adapters and evaluates performance on MCQ and SAQ tasks. This section will motivate the task, define the evaluation setting, and summarize the main contributions.

2 System Overview

This section will present the high-level architecture and data flow. Key components include dataset preprocessing, LoRA-based training, task-specific inference, and optional retrieval augmentation.

3 Implementation

This section will detail how the system is implemented and configured.

3.1 LoRA Fine-Tuning

I will describe which transformer layers are adapted, how adapters are trained, and which hyperparameters are used for MCQ and SAQ tasks.

3.2 MCQ Inference via Logprob Scoring

I will summarize the log-probability scoring algorithm used to pick among answer options and the reranking mechanism based on country priors.

3.3 RAG for SAQ

I will explain how retrieval augmentation is integrated, including indexing, retrieval, and prompt construction with contextual snippets.

3.4 Parsing and Validation

I will describe how answer formats are parsed, validated, and retried for robustness in both MCQ and SAQ settings.

4 Experiments and Iterative Improvements

This section will report incremental changes and their impact, organized around the experimental records in `report/drafts/experiments.md` and linked to corresponding submissions and commits.

4.1 Baseline

I will document the default model performance for MCQ and SAQ as the baseline.

4.2 SAQ Iterations

I will outline prompt and parsing refinements, validation retries, and LoRA layer extensions for SAQ improvements.

4.3 MCQ Iterations

I will outline logprob scoring variants and reranking configurations tested for MCQ improvements.

4.4 RAG Variants

I will summarize retrieval settings (raw, stop-word removal, stemming) and the observed effects on SAQ performance.

5 Results

This section will consolidate the final MCQ and SAQ scores, with tables that match the reported experimental metrics and best combined configuration.

6 Discussion and Limitations

I will discuss observed trends, limitations of the current pipeline, and possible sources of error or variance.

7 Conclusion and Future Work

I will summarize the contributions and outline concrete next steps for further model and retrieval improvements.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.