Project Proposal Submission (Week 4)

COMPSCI 792 – University of Auckland

Candidate's Name: Ruoyu Zhang

UoA ID Number: 984520034

Thesis Title:
Benchmarking and Fine-Tuning Open-Source Arabic Foundation Models for Language Understanding and Generation

Main Supervisor: Jiamou Liu

Background

Arabic is a morphologically rich language with diglossia between Modern Standard Arabic and regional dialects, which poses unique challenges in the large language models (LLMs). With only over 400 million people speak Arabic globally, most open-source LLMs have been trained with minimal Arabic data, often lacking dialectal and culturally nuanced representation. Even among models specifically trained for Arabic, the output often suffers from rigid or unnatural phrasing, particularly in generative tasks like summarization, dialogue, or open-ended question answering. This problem is more pronounced when models are prompted in dialects or informal settings.

However, several Arabic-centric foundation models have been developed in recent years, including Jais, Falcon Arabic, AceGPT, AraBERT, and Qwen-Arabic. These models differ significantly in architecture, training corpora, dialectal coverage, and fine-tuning strategies. While existing studies have explored their construction or individual performance, there is a lack of comprehensive, comparative evaluation, mainly focusing on naturalness, instruction-following, and cultural alignment in generative tasks.

This project aims to fill that gap by conducting a structured benchmarking study of Arabic LLMs. It will also explore resource-efficient fine-tuning strategies to enhance Arabic text generation, improving linguistic naturalness and fluency. The contribution lies in delivering a reproducible evaluation framework, identifying task-specific strengths and weaknesses, and potentially releasing enhanced checkpoints for real-world Arabic applications.

Aims and Objectives

- Benchmark the performance of multiple open-source Arabic foundation models across a range of NLP tasks.
- Analyse model behaviour in dialectal vs. Modern Standard Arabic (MSA) scenarios.
- Evaluate the effectiveness of instruction tuning on Arabic-centric educational datasets.
- Explore resource-efficient fine-tuning techniques (e.g., QLoRA).
- Provide reproducible tools, datasets, and results for future Arabic NLP research.

Research Design

Models: Jais, Falcon Arabic, AceGPT, Qwen-Arabic, Juhaina, ALLaM, and possibly Arabic-instructed variants of LLaMA and Gemma.
Tasks: Classification, summarisation, question answering, text generation.
Benchmarks: AraBench, AraT5Eval, MADAR, ArBench, CIDAR.
Tools: Hugging Face Transformers, PEFT, BitsandBytes, evaluate, lm-eval-harness.

Ethical Approvals

No ethical approval is required as all datasets and models are open-source and non-sensitive.

Resources

- - Computing: LIU AI Lab GPU servers.
  - Datasets: Translated instruction datasets: Alpaca, FLAN, SuperNI; NativeQA / SafetyQA: From Jais (authentic Arabic QA + safety);
  - Tools: Python, Hugging Face libraries, QLoRA setup.

Potential Problems and Mitigations

- Compute limitations: Running multiple large models locally may strain available GPU resources. We will mitigate this by selecting models under 13B, using quantified inference (e.g., 4-bit) and batching.
- Benchmark inconsistency: Different benchmarks may use incompatible formats or metrics. We will build a unified evaluation pipeline to standardize inputs and outputs for all tasks.
- Dialectal variation: Some benchmarks are MSA-heavy while others include dialects. We will categorize and tag results by language variant to ensure meaningful comparisons.
- Incomplete model documentation: Some open-source models lack clear documentation or tokenizer support. We will manually inspect and, if needed, adapt pre/post-processing for alignment.

Tentative Timeline (14-week Semester)

- **Week 4**: Finalize model list and complete literature review
- **Week 5**: Collect benchmark datasets and prepare evaluation scripts
- **Week 6**: Set up and validate unified benchmarking pipeline
- **Week 7**: Run zero-shot evaluations across selected tasks and models
- **Week 8**: Analyse and compare model performance (accuracy, fluency, coverage)
- **Week 9**: Extend evaluations to dialectal and generative benchmarks
- **Week 10**: Aggregate results and standardize evaluation metrics
- **Week 11**: Conduct qualitative evaluation (e.g., GPT judge, human sampling)
- **Week 12**: Summarize findings and prepare figures/tables
- **Week 13**: Draft methodology, results, and discussion sections
- **Week 14**: Finalize thesis draft and incorporate supervisor feedback

References

- Alkhowaiter et al. (2025). “Mind the Gap: Arabic Post-Training Datasets”
- Ismail et al. (2025). “Transformers for Arabic GEC”
- Alrefaie et al. (2024). “Arabic Tokenization Strategies”
- Qian et al. (2024). “CamelEval and Juhaina”
- Huang et al. (2024). “AceGPT: Arabic Adaptation of LLaMA”
- Arcee AI (2023). “Meraj: Arabic-Tuned Nova Model”

Supervisor Signature: