Project Proposal: Microscaling Large Language Models for Edge Deployment

Team Members: Paapa Kusi, Anish Gaonkar, Param Gattupalli, Vinay Reddy Ratnam

---

1. Problem Statement and Motivation

Large Language Models (LLMs) achieve state-of-the-art performance in diverse NLP tasks, but their size makes them unsuitable for deployment on personal devices. This limits accessibility and raises privacy and sustainability concerns. Recent methods in pruning, quantization, and distillation offer paths toward microscaling LLMs by compressing them without fully sacrificing utility.

This project investigates:
“How far can we microscale LLMs (pruning + quantization) while preserving usefulness on edge devices?”

---

2. Objectives

- Compress open-source LLMs through **pruning and quantization**.
- Systematically evaluate **accuracy, latency, and memory trade-offs**.
- Benchmark under **edge constraints** (CPU/RAM limits).
- (Stretch goal) Deploy the best-performing compressed model on CoreML (iOS) or TensorFlow Lite (Android).

---

3. Proposed Methods

- **Baseline Models:** GPT-2 Medium (355M) and LLaMA-2-7B.

- **Compression Techniques:**
    - Structured pruning (layers, heads) and unstructured pruning (weights).
    - Quantization (INT8, INT4).
    - (Optional) Knowledge distillation.
- **Evaluation Metrics:**
    - **Intrinsic:** Perplexity on test corpus.
    - **Extrinsic:** Sentiment classification / QA benchmarks.
    - **Systems:** Model size, inference speed, memory footprint.
- **Hardware Setup:** HiPerGator GPUs (L4/B200, B200's if we can) with Docker-based resource caps to simulate phones.

---

4. Planned Experiments

1. **Baseline:** Evaluate uncompressed models.
2. **Pruning:** Test 10%, 50%, 90% pruning ratios.
3. **Quantization:** Apply INT8 and INT4 post-training quantization.
4. **Combined:** Pruning + quantization pipeline.
5. **Stretch:** Deploy best model on CoreML/TFLite, test on mobile hardware.

---

5. Division of Work

- **Compression Lead - Anish Gaonkar**
    - Implement pruning strategies (structured and unstructured).
    - Measure effects on perplexity, accuracy, and model stability.

    - Document pruning pipeline and ablation studies.
- **Quantization Lead - Param Gattupalli**
    - Apply post-training quantization (INT8, INT4).
    - Benchmark efficiency improvements (inference speed, memory footprint).
    - Compare with pruning results and analyze combined effects.
- **Evaluation & Benchmarking Lead - Vinay Reddy Ratnam**
    - Design evaluation framework (intrinsic metrics like perplexity, extrinsic tasks such as sentiment classification/QA).
      Run baseline experiments on uncompressed models.
    - Track system metrics (latency, RAM, storage) under resource constraints.
- **Deployment & Presentation Lead - Paapa Kusi**
    - Explore deployment of compressed models to edge environments (CoreML, TensorFlow Lite, Raspberry Pi/Jetson).
    - Develop demonstration scripts or mobile app prototype.
    - Organize project presentation materials and coordinate report visuals.

(All members will contribute to final report writing and presentation prep.)

---

6. Timeline

- **Week 1 (Oct):** Baseline runs, dataset prep, assign technical leads.
- **Week 2–3 (Oct):** Implement pruning & quantization experiments.
- **Week 4 (Nov):** Analyze results, prepare slides, rehearse presentation.
- **Week 5-6 (Nov):** Write and finalize NeurIPS-style report.

---

7. Expected Findings

- A set of **compression–performance trade-off benchmarks** for edge deployment.
- Analysis of the **“sweet spot”** where LLMs remain usable on constrained devices.
- Reproducible **codebase + experiments** for future UF researchers.
- (Optional) Live demo of microscaled LLM on a mobile device.