**Project Proposal: GPU-Accelerated Mixture-of-Experts (MoE) Router with Quantized Routing and MLP Comparison**

**Summary**

This project aims to implement a GPU-accelerated inference kernel for the router component of a Mixture-of-Experts (MoE) transformer model, enhanced with quantized router logits and expert weights. Additionally, the project will compare the speed and accuracy of the MoE implementation against a standard dense MLP layer.

**Background Information**

Mixture-of-Experts models have become popular due to their ability to selectively activate parts of the model per input token, significantly reducing computation. Quantization techniques further enhance these efficiency benefits by reducing memory bandwidth and computation. This project explores the practical benefits and trade-offs of combining MoE architectures with quantization compared to traditional dense MLPs.

**Computation Details**

* **Inputs**: Token embeddings and router weight matrices.
* **Quantization**: Router logits (int8), expert weights (int8 or FP16).
* **Output**: Expert assignments per token and corresponding MLP outputs.

**Project Explanation**

This project includes three main components:

1. MoE router implementation for token assignment.
2. Expert MLP computation with quantized weights.
3. Comparison of MoE and standard dense MLP layers in terms of inference speed and accuracy.

**Questions to Address**

* How does quantization of router logits and expert weights impact inference speed and accuracy?
* What performance advantages does the MoE model provide over a dense MLP?

**Previous GPU Implementations**

Recent MoE models such as Switch Transformer and DeepSeek have successfully implemented GPU-accelerated quantized MoE systems, achieving substantial computational efficiency.

**Technical Challenges**

* Implementing efficient quantized matrix multiplication kernels in CUDA.
* Managing quantization scale factors and ensuring numerical stability.
* Optimizing token dispatching and batching logic to maximize GPU utilization.

**Problems to Solve**

* Designing GPU kernels for quantized routing and expert computation.
* Comparing MoE versus standard MLP layers comprehensively.

**Deliverables and Goals**

* CPU baseline implementation of MoE router and dense MLP.
* GPU (CUDA) kernels for MoE routing and expert computation with quantization.
* Benchmarking and detailed analysis comparing inference speed and accuracy.
* Comprehensive documentation and clear performance analysis.

**Week-by-Week Timeline**

* **Week 1:** CPU baseline implementations (MoE and dense MLP); initial correctness testing.
* **Week 2:** Implement initial CUDA kernels for MoE routing and experts; verify GPU correctness.
* **Week 3:** Integrate quantization; refine GPU kernels; begin benchmarking.
* **Week 4:** Conduct final benchmarks, document results, analyze performance trade-offs, and write final report.