Efficient AI: Optimizing Performance, Speed, and Sustainability

Dhruv Dixit, Dhavan Antala, Harsh Gautam
Department of Electrical & Computer Engineering, Dr. Hao Wang

Problem Statement

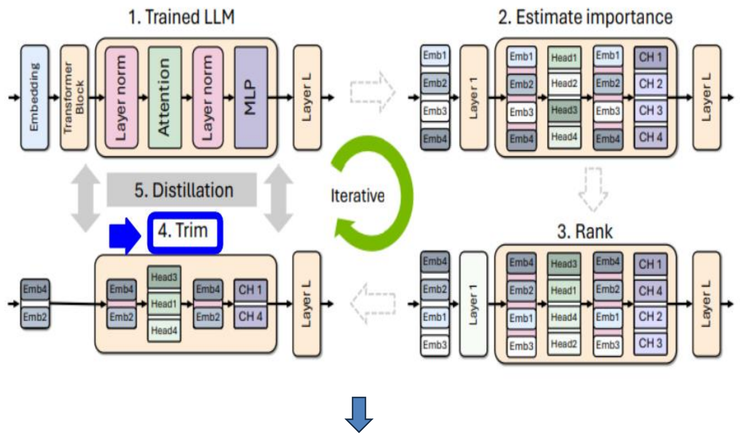
How can we **retain the power** of large language models **while reducing** their computational and memory requirements?

Introduction

The Goal of this project is optimizing model inference speed for LLaMA 3 8B by applying **pruning**, **quantization**, and **distillation**, resulting in faster inference with minimal GPU usage.

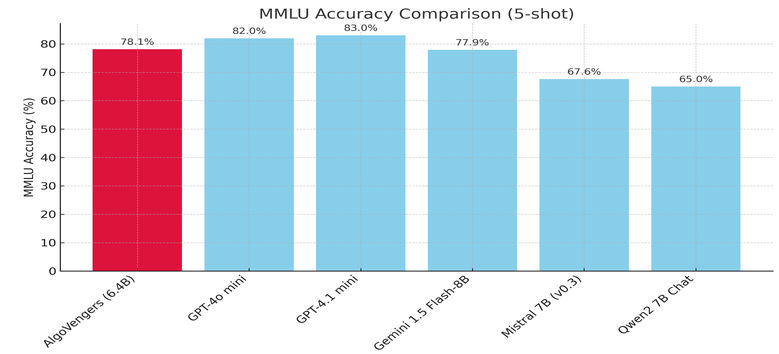
Methodology

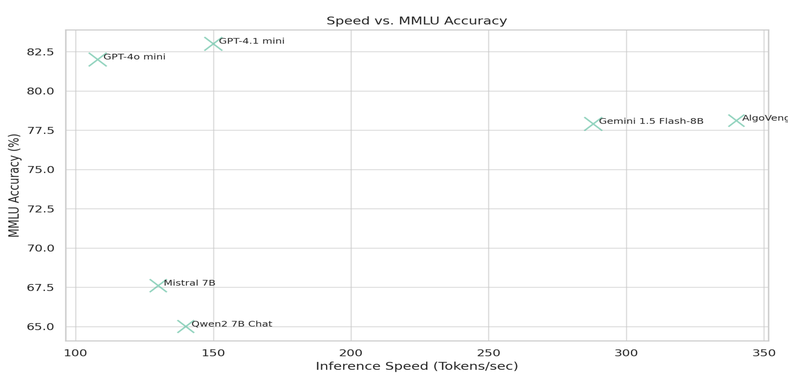
- **Model:** LLaMA 3 8B
- **Pipeline:** Pruning → GPTQ/AWQ Quantization (8-bit → 4-bit) → Distillation
- **Optimization:** NVIDIA NeMo / Megatron
- **Techniques:**
  - Pruned from 8B to 6.4B parameters
  - Distillation with 70B LLaMA-3 teacher (long-form QA)
- **Tools:** Hugging Face, bitsandbytes, Flash Attention 2, PyTorch
- **Evaluation:**
  - Metrics: EM, F1, latency, memory



Results

- **Inference Speed:**
  - 2.1× faster than baseline LLaMA 3 8B
- **Memory Usage:**
  - Reduced peak GPU memory by ~25%
- **Accuracy Retention:**
  - Within ±1.2% of original on HotpotQA/NarrativeQA
- **Distilled Model Performance:**
  - Outperformed Mistral 7B on long-text QA (F1: +2.3%)





STEVENS INSTITUTE OF TECHNOLOGY

DuckAI