Capstone Project Title: LLM Compression Trade-off Evaluator

Sponsor:
Robins AFB – AFSC/EN

https://www.robins.af.mil/

Project Overview:
In this project, students will evaluate the inference quality on code-generation queries (metrics will be decided by the student team) of at least one open-source LLM undergoing various compression techniques (such as quantization, knowledge-distillation, or anything else.) This will be the first step in understanding the trade-off between LLM parameter-count and inference quality for code-generation tasks, potentially reducing the average size of LLMs needed for this use-case. (See sponsor provided slides for full technical details.)

Project Objectives:

- Analyze model compression techniques and their impact on query accuracy and speed.
- Build a dynamic query allocation prototype between full and compressed models.
- Design test cases focusing on tasks such as code generation and debugging.
- Evaluate performance using code-specific metrics like Code-BLEU.
- Visualize performance trends and resource savings via reporting dashboards.
- Document best practices for LLM deployment efficiency in cloud environments.

Scope of Work:

- Set up baseline LLM inference environment with multiple compression methods.
- Implement dynamic query routing logic based on query complexity.
- Conduct experiments on code-related queries, using BLEU/Code-BLEU metrics.
- Build a reporting dashboard summarizing accuracy, performance, and resource use.
- Provide end-to-end documentation, including data analysis scripts and system architecture.
- See sponsor provided slides for detailed technical guidance.

Expected Deliverables:

- LLM compression benchmarking scripts.
- Dynamic query routing proof-of-concept application.
- Code-BLEU metric evaluation reports.
- Interactive dashboard with performance visualizations.
- Final project report with deployment recommendations.

Milestones:

1. System Setup & Dataset Analysis (Weeks 1-4):

- **Deliverable**: initial environment setup and data preparation
- **Evaluated on**: completeness and accuracy of baseline configuration.

2. Model Compression & Query Routing Development (Weeks 5-10):

- **Deliverable**: working compression implementations and dynamic routing logic
- **Evaluated on:** functional correctness and routing efficiency.

3: Evaluation, Visualization & Final Report (Weeks 11-15):

- **Deliverable:** comparative evaluation results, interactive dashboard, and final project report
- **Evaluated on:** insightfulness, clarity, and operational recommendations.

Technology Stack:

- **Python**, **PyTorch**/**Transformers** libraries for LLM interaction
- **Pruning**, **Quantization**, and **Distillation** techniques for compression
- **Docker** for containerized deployment
- **Streamlit** or **Dash** for dashboard visualization
- **GitHub**/**GitLab** for version control
- **Cloud environment (AWS/GCP/Azure)** for optional deployment and testing

Student Learning Outcome / Resume Highlights:

- Hands-on experience with LLM inference and optimization.
- Practical knowledge of compression techniques in AI systems.
- Development of full-stack solutions including backend services and dashboards.

- Skills in Python-based experimentation and performance evaluation.
- Exposure to cloud services and cost-effective AI deployment strategies.

Sponsor Commitment:

The AFSC/EN team will provide guidance on expected evaluation metrics and access to example datasets or LLM environments where applicable. Regular feedback will be provided throughout the project lifecycle, with milestone reviews and mentoring support from technical leads.