Big Data CS 6220 : Project Proposal (Group 11)

Smart Tutor: Subject-Specific MoE based LLM System for K-12 Q&A.

Authors:

1. Kushal Ramaiya (kramaiya3)
2. Ramanathan Swaminathan (rswaminathan38)
3. Rutuja Kawade(rkawade3)
4. Venkata Sai Anirudh Kamaraj Vobbilisetty (vvobbilisetty6)

Objective:

The goal of this project is to build **MoE-Tutor**, a subject-aware question-answering system that routes each query to a specialized LoRA adapter on top of a single shared backbone (LLaMA-3.1-8B-Instruct). A router selects whether to activate the Math, Science, or General Knowledge adapter.

Students often ask a wide range of questions, and no single model performs equally well across all domains. By routing each question to a specialized expert trained on subject-specific data, we aim to improve both factual accuracy and explanation clarity.

Our ultimate goal is to build a **single unified model** that can generate responses tailored to the query style - for example, chain-of-thought reasoning for Math queries and concise factual answers for General Knowledge queries..

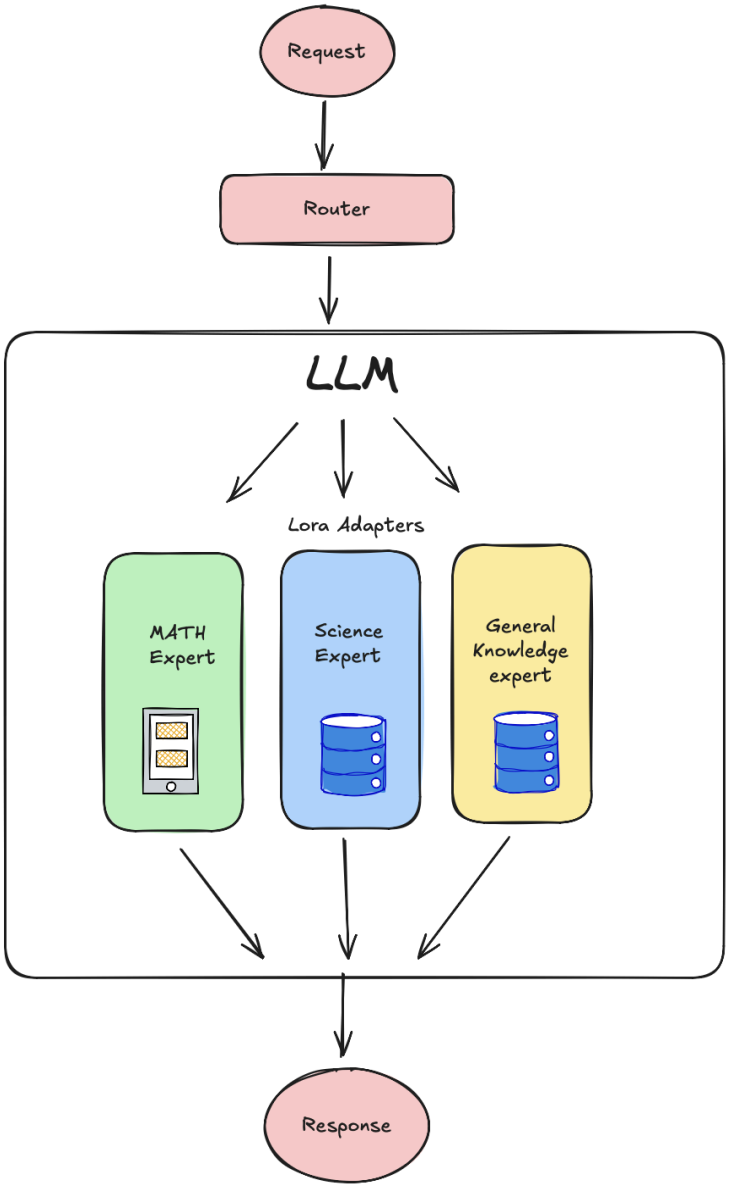
Dataset :

- Math — GSM8K
- Science : OpenBookQA , ARC (AI2 Reasoning Challenge)
- General Knowledge —TriviaQA , Natural Questions
- Retrieval corpus (for Science/GK)**:** Simple English Wikipedia

Three different domains are utilized here. The top datasets in these domains that are suitable for LLMs are collected. All these datasets will be concatenated together to form into a single dataset which will be trained on LLaMA-3.1-8B-Instruct which would be the baseline model.

The comparison between the baseline model and MoE model which uses the same shared LLaMA-3.1-8B-Instruct backbone with separate LoRA adapters fine-tuned for Math, Science, and GK .

Design:



drawn with https://excalidraw.com/

Our implementation consists primarily of 4 major building blocks Router, Maths,Science and General Knowledge mixture of experts.

Implementation:

Steps :

1. **Baseline (Joint FT)**:

   Supervised fine-tuning LLaMA-3.1-8B-Instruct with no adapters the concatenated dataset of Math (GSM8K), Science (ARC/OpenBookQA), and General Knowledge (TriviaQA+NQ). (no adapters, no routing)

2. **Expert LoRAs:**

   Each LoRA is trained on its subject-specific dataset while sharing the same LLaMA-3.1-8B backbone

   a. Math LoRA: trained on GSM8K to produce concise reasoning plus answer; calculator used only to evaluate simple expressions at inference.
   b. Science LoRA: trained on ARC/OpenBookQA with retrieved passages prepended; outputs short rationale + final choice/span.
   c. GK LoRA: trained on TriviaQA/NQ with retrieved context; outputs concise factual spans.

3. **Router** -

   The router checks each incoming query and activates the appropriate expert. We will use a lightweight DistilBERT classifier trained on our fine-tuning dataset, producing three output classes: Math, Science, and GK. To improve robustness, we also plan to explore training the router on a separate dataset, which will be investigated further during the project.

4. **Flow** :

   Input question -> Router decides domain ->Activates corresponding LoRA expert -> Output answer.

5. **Evaluation:**

   Math : Exact Match on final numeric answer (GSM8K).
   Science : Multiple-choice accuracy (ARC/OpenBookQA).
   GK : Exact Match/F1 (TriviaQA/NQ).
   System Metrics : Latency,Cost/Query

Observations:

1. Per-subject score metrics
2. Router confusion matrix
3. Ablation studies (tools on/off, top-1 vs top-2(routing logic)

Timeline:

- **Week 1–2**: Literature review, dataset selection, baseline agent setup.
- **Week 3–4**: Implement all the individual adapters with their datasets with LoRa fine tuning.
- **Week 5**: Router training to route queries.
- **Week 6**: Pipelines to send queries between models and router.
- **Week 7**: Compare results, prepare analytics on performance.
- **Week 8**: Write final report + prepare presentation/demo.

References:

1. *What is LoRA (Low-Rank Adaption)? | IBM*.https://www.ibm.com/think/topics/lora
2. Zilliz. (2024, November 15). LoRA Explained: Low-Rank Adaptation for Fine-Tuning LLMs. *Medium.*
3. *Mixture of Experts Explained* . https://huggingface.co/blog/moe
4. *What is mixture of experts?* https://www.ibm.com/think/topics/mixture-of-experts
5. *Chain-of-Thought Prompting – Nextra*. https://www.promptingguide.ai/techniques/cot

6. *Prompt Engineering Guide*. https://www.promptingguide.ai/techniques