**[Final Year Project Proposal]**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sr# | Student Name | Roll Number | Credit Completed | Signature |
| 1 | Zainab Khalid | 22P-9061 | 100 |  |
| 2 | Faryal Hanif | 22P-9173 | 97 |  |
| 3 | Ume Taqadus | 22P-9393 | 100 |  |

**Suggested Supervisor**:

Faculty Member’s Name: Mr.Shahzeb Khan Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date (28th September, 2025)

**Project Details**

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Title** | Towards Lightweight Routing: Adaptive and Explainable Orchestration of Models and Agents | | |
| **Project Area of Specialization** | Artificial Intelligence | | |
| **List Related Core Subjects** | 1) Artificial Intelligence  2) Database Systems  3) Software Design And Analysis | | |
| **List Related Elective Subjects** | 1) Natural language processing  2) Data Science | | |
| **Project Start Date** | 2025-9-25 | **Project End Date** | 2026-5-10 |
| **Project Summary (less than 2500 characters)** | Our project aims to develop a lightweight, adaptive, and explainable routing framework for handling user queries across multiple locally hosted language models and agents. Instead of relying on a single large model or cloud-based APIs, the system will dynamically decide whether a query should be processed by a smaller local model, a specialized agent, or a more capable model.  The framework is designed to be computationally efficient on local hardware, optimizing CPU/GPU usage, memory consumption, and response time. It also adapts over time by learning from feedback and provides transparency by explaining why a particular route was chosen. The final outcome will be a working prototype that demonstrates how locally hosted models can handle queries efficiently, adaptively, and transparently, even in low-resource environments. | | |
| **Project Objectives (less than 2500 characters)** | * **Design a lightweight routing system** that directs user queries to either agents or the most suitable language model, reducing reliance on large models for simple tasks. * **Incorporate adaptive learning** so the router can improve decisions over time and escalate queries to stronger models when needed. * **Provide explainability and transparency** by showing users why a particular model or agent was chosen, and highlighting the trade off between performance and cost. * **Develop and demonstrate a prototype** that showcases efficient, transparent, and adaptive query routing across multiple models and agents. | | |
| **Project Implementation Method (less than 2500 characters)** | * Review existing routing approaches and define requirements for a lightweight system. * Prepare a dataset by aggregating model preference data and generating labels for training. * Develop a lightweight router to classify queries and direct them to agents or models. * Add adaptive features so the router improves with feedback and escalates when needed. * Implement explainability and cost awareness through a dashboard for transparent routing. * Test and refine the prototype to demonstrate efficiency, adaptability, and trustworthiness. | | |
| **Benefits of the Project (less than 2500 characters)** | This project promotes resource-efficient AI deployment by enabling multiple open-source LLMs to operate locally without relying on paid cloud APIs. It reduces hardware strain by routing simpler queries to smaller models while reserving heavier models for complex tasks. The system’s adaptive behavior improves decision quality over time, and its explainability features enhance user trust by making routing transparent. Ultimately, it demonstrates how AI systems can be sustainable, efficient, and trustworthy even in limited computational environments. | | |
| **Technical Details of Final Deliverable (less than 2500 characters)** | The final deliverable will be a working prototype of a lightweight query routing system that can direct user queries either to agents or to the most suitable language model. It will include a simple router for decision making, a small set of integrated agents for specialized tasks, and multiple models for handling queries of different complexity levels. The system will also feature adaptive mechanisms to improve over time and an explainability layer with a dashboard that shows the reasoning behind routing choices as well as the cost-performance trade offs. This prototype will demonstrate efficient, transparent, and resource friendly query handling in a practical and user focused way. | | |
| **Final Deliverable of the Project** | A functional prototype of a lightweight, adaptive, and explainable query routing system that directs user queries to agents or the most suitable language model, with a dashboard for transparency and cost-performance insights. | | |
| **Type of Industry** | Natural Language Processing (NLP) Applications  Generative AI Solutions | | |
| **Technologies** | Open-source LLMs, Simple Agents, | | |
| **Sustainable Development Goals** | Industry, Innovation, and Infrastructure  Decent Work and Economic Growth  Responsible Consumption and Production | | |

**Project Key Milestones**

|  |  |  |
| --- | --- | --- |
| **Elapsed time in (days or weeks or month or quarter) since start of the project** | **Milestone** | **Deliverable** |
| **Semester 1** | **FYP 1** | 1. Dataset Collection and Preprocessing 2. Router Design (lightweight model setup) 3. System Architecture / Workflow Draft 4. Initial Prototype Development (basic query routing) |
| **Semester 2** | **FYP 2** | 1. Adaptive Features Implementation 2. Explainability & Cost-Awareness Layer 3. Dashboard Development 4. Final Prototype Integration & Testing 5. Deployment and Demonstration |

**Project Equipment Details**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Item Name** | **Type** | **No. of Units** | **Per Unit Cost (in Rs)** | **Total (in Rs)** |
| **Cloud GPU Services** | **Service** | **3** | **5000** | 15000 |
| GSM module | Equipment | 5 | 10000 | 5000 |
| Others | Miscellaneous | 1 | 10000 | 10000 |
|  |  |  | **Total in (Rs)** | **30000** |