Architecture Document: Essay Grading System (Fine-Tuned Qwen + LangGraph Framework)

---

1. Overview

The Essay Grading System is an **AI-driven pipeline** that evaluates student essays against multiple rubric dimensions. It combines a **fine-tuned Qwen language model** with **custom mapping layers and weighted scoring logic** to provide both **analytic feedback** and a **final aggregated score**.

---

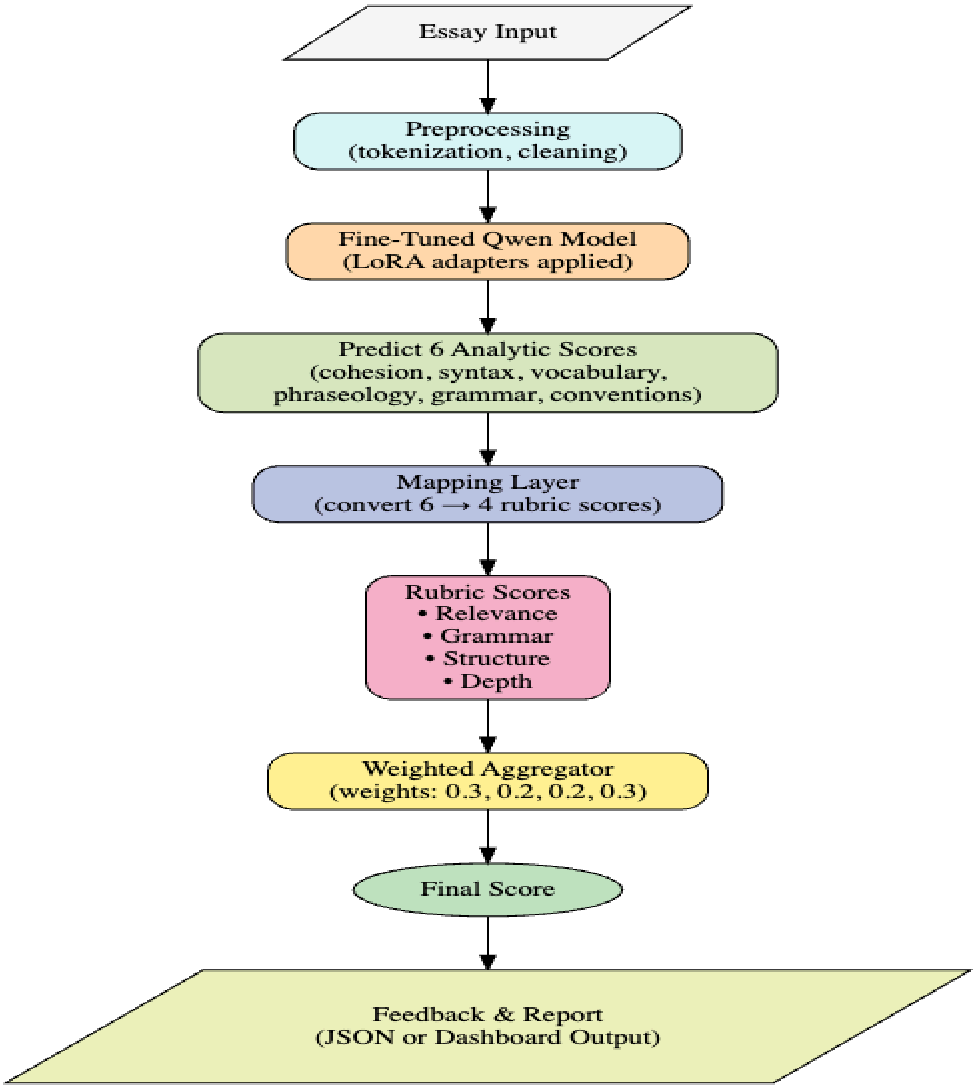
2. System Objectives

- Automate essay evaluation at scale.
- Provide **transparent, rubric-aligned scores** instead of a “black-box” grade.
- Leverage **fine-tuned open-source LLMs (Qwen)** for cost efficiency and flexibility.
- Allow easy integration with dashboards, APIs, or LMS platforms.

---

3. High-Level Architecture

Flow Diagram


Essay Input
Preprocessing (tokenization, cleaning)
Fine-Tuned Qwen Model (LoRA adapters applied)
Predict 6 Analytic Scores (cohesion, syntax, vocabulary, phraseology, grammar, conventions)
Mapping Layer (convert 6 ⟶ 4 rubric scores)
Rubric Scores
• Relevance
• Grammar
• Structure
• Depth
Weighted Aggregator (weights: 0.3, 0.2, 0.2, 0.3)
Final Score
Feedback & Report (JSON or Dashboard Output)

Steps:

1. **Essay Input**
   - Student submits an essay.
   - Input collected via API, web app, or batch upload.
2. **Preprocessing**
   - Tokenization, cleaning, trimming to model max length.
   - Ensures input fits LLM constraints and removes noise.
3. **Fine-Tuned Qwen Model (LoRA Adapters)**
   - Base: `Qwen/Qwen2.5-3B-Instruct`.
   - Fine-tuned using **Feedback Prize – ELL dataset** via **QLoRA**.
   - Outputs **six analytic scores**: cohesion, syntax, vocabulary, phraseology, grammar, conventions.
   - Chosen because:

        - Open-source (no API costs like GPT-4o).
        - Fine-tuned with domain data → improved grading consistency.
        - LoRA makes it feasible on consumer GPUs (memory-efficient).
4. **Mapping Layer (6 → 4 Rubric Scores)**
    - Converts ELL's six dimensions into four rubric categories:
        - **Relevance Score** ⟵ cohesion + vocabulary + phraseology.
        - **Grammar Score** ⟵ grammar (+ conventions weighted).
        - **Structure Score** ⟵ cohesion + syntax.
        - **Depth Score** ⟵ vocabulary + phraseology + syntax.
    - Justification: matches your original framework's design while leveraging analytic outputs.
5. **Rubric Scores Output**
    - Each score normalized to [0, 1].
    - Provides interpretable feedback across multiple dimensions.
6. **Weighted Aggregator**
    - Combines the four rubric scores into a **final score** using weights:
        - Relevance: **0.3**
        - Grammar: **0.2**
        - Structure: **0.2**
        - Depth: **0.3**
    - Justification: higher weight to **Relevance** and **Depth** as they reflect essay content/insight.
7. **Final Score**
    - A single numeric grade (0–1, scaled to 0–100 if needed).
    - Used for comparative ranking or direct grading.
8. **Feedback & Reporting**
    - JSON output for programmatic integration.
    - Optional dashboard visualizations:
        - Spider/Radar chart for rubric scores.
        - Histogram of final scores across a class.
    - Enables both **teachers** (analytics) and **students** (feedback).

---

4. Technology Stack

- **Model:** Qwen2.5-3B-Instruct (fine-tuned with LoRA).
- **Training Framework:** Hugging Face `transformers`, `trl`, `peft`, `bitsandbytes`.
- **Workflow Framework:** LangGraph for modular, DAG-based grading logic.
- **Storage/Interface:** JSON output; can be plugged into Flask, FastAPI, or LMS.
- **Visualization:** Graphviz for architecture diagrams, dashboards for analytics.

---

5. Why This Architecture

- **Fine-Tuned Qwen**: Free to Use SLM.
- **LoRA Adapters**: efficient fine-tuning within Kaggle/Colab GPU limits.
- **Mapping Layer**: bridges dataset labels → original rubric requirements.
- **Weighted Aggregation**: preserves grading philosophy, customizable for different institutions.
- **Modular Flow (LangGraph)**: easy to extend (e.g., add plagiarism check, coherence analysis).

---

6. Future Enhancements

- **Multi-task Training**: directly fine-tune Qwen to predict all four rubric scores instead of mapping.
- **Explainability**: add natural-language feedback per score.
- **Calibration**: align model outputs with human graders using linear regression or Platt scaling.
- **Deployment**: package as a REST API or microservice for LMS integration.
