**Architectural Decision Record (ADR)**

**1. Title**

AI-Assisted Grading for Aptitude Test Using RAG and LLM

**2. Context**

Certifiable, Inc. requires a scalable solution to grade short-answer questions in the aptitude test. The existing system relies entirely on manual expert review, which is time-consuming (~3 hours per candidate) and does not scale with increasing demand. The goal is to automate grading while maintaining accuracy, fairness, and explainability.

Current grading challenges include:

* Slow processing times due to full manual review.
* High expert workload, limiting scalability.
* Inconsistencies in grading across different reviewers.
* Need for detailed feedback for candidates.

Available options:

1. Rule-based grading system
2. Fine-tuned LLM for direct grading
3. Hybrid approach using Retrieval-Augmented Generation (RAG) and LLM (Selected)

Key constraints:

* AI must achieve high accuracy (>80% agreement with human grading).
* Latency should be low (<2 seconds per response).
* Cost of inference must be optimized.
* Bias detection and fairness monitoring are required.

**3. Decision**

We will implement a **Hybrid AI-Assisted Grading System** using **Retrieval-Augmented Generation (RAG) with a fine-tuned LLM** for short-answer grading.

* **Chosen Approach:**
  + LLM pre-grades responses, categorizing them as **Clearly Correct, Clearly Incorrect, or Requires Review**.
  + Expert software architects review only flagged responses, reducing manual workload by 60-70%.
  + AI-generated feedback is attached to results for automated candidate feedback.
* **Why This Approach?**
  + Ensures consistency and explainability in grading.
  + Reduces expert grading workload while maintaining human oversight.
  + RAG component ensures contextual accuracy by referencing a knowledge base of expert-graded past answers.
* **Rejected Alternatives & Trade-offs:**
  + Rule-based grading: Too rigid and unable to handle nuanced answers.
  + Pure LLM grading: Lacks factual grounding, leading to possible hallucinations.
  + Hybrid RAG-LLM was selected for **accuracy, explainability, and scalability**.

**4. Architecture Impact**

* **New Components:**
  + AI Grading Engine with LLM and RAG
  + Candidate Answer Pre-Processing Service
  + Expert Review Module for flagged cases
  + AI Feedback Generator
  + AI Performance Monitoring Dashboard
* **Integration Points:**
  + Connects with Candidate Testing UI to receive responses.
  + Updates Candidate Database with graded responses.
  + Integrates with the Certification System for final grading decisions.
* **Performance Considerations:**
  + AI grading must maintain **<2 seconds per response**.
  + System should handle **10x increase in candidates** without failure.

**5. Risks & Mitigation**

|  |  |  |
| --- | --- | --- |
| **Risk** | **Impact** | **Mitigation Strategy** |
| AI may grade incorrectly | Medium | Keep expert review for flagged cases |
| High inference cost | High | Use lightweight models (Flan-T5) for simple cases |
| Bias in AI grading | High | Regular fairness audits and bias detection tools |
| Explainability concerns | Medium | Ensure AI provides justifications for grading decisions |

**6. Acceptance Requirements**

|  |  |  |
| --- | --- | --- |
| Requirement | Description | Acceptance Threshold |
| Accuracy | AI grading must align with human expert grading | ≥ 80% agreement |
| Speed | AI must process each response quickly | ≤ 2 seconds per response |
| Expert Workload Reduction | Reduce manual grading workload | ≤ 30% of cases require review |
| Bias Mitigation | No demographic-based grading discrepancies | ≤ 5% grading variation across groups |
| Explainability | AI must provide clear feedback on grading decisions | 100% of cases must include feedback |
| System Stability | AI must handle increased load | 99.9% system uptime |

**7. Monitoring & Validation**

* **Success Metrics:**
  + AI-human grading agreement of at least 80%.
  + Expert review workload reduced by 60%.
  + AI grading response time under 2 seconds.
  + Bias detection metrics to monitor fairness.
* **Monitoring Tools:**
  + AI Performance Dashboard (track grading accuracy and latency).
  + Model Drift Detection to ensure AI remains consistent over time.
  + Fairness & Bias Audits conducted quarterly.
* **Iteration Plan:**
  + Initial deployment with human oversight for validation.
  + Gradual increase in AI autonomy as confidence improves.
  + Continuous fine-tuning based on expert feedback.

**8. Decision Status**

* 🟡 **In Progress** (Pilot phase with expert oversight)

**9. Related Documents**

* AI Grading System Architecture Diagram
* Research Papers on RAG for Short-Answer Evaluation
* Certification System API Documentation