**Architectural Decision Record (ADR)**

**1. Title**

**Adoption of Claude 3.5 Sonnet for Cost-Effective and Scalable Short-Answer Grading**

**2. Context**

Our organization requires an AI model capable of efficiently grading large batches of short-answer responses. The key requirements include:

* **High Token Limit:** Ability to process extensive context in a single pass.
* **Cost-Effectiveness:** Affordable pricing to maintain budget constraints.
* **Scalability:** Capability to handle increasing volumes without performance degradation.
* **High Accuracy:** Deliver precise and reliable grading outcomes.

After evaluating various models, **Claude 3.5 Sonnet** by Anthropic emerges as a strong candidate.

**3. Decision**

We have decided to integrate **Claude 3.5 Sonnet** into our short-answer grading system.

**Rationale**

* **High Token Limit:** Supports a context window of up to **200,000 tokens**, enabling the processing of large batches in a single API call. citeturn0search5
* **Cost-Effectiveness:** Competitive pricing at **$3 per million input tokens** and **$15 per million output tokens**, balancing performance with affordability. citeturn0search5
* **Scalability:** Demonstrates superior performance in complex tasks, including coding and multi-step workflows, ensuring scalability for large-scale grading operations. citeturn0search5
* **High Accuracy:** Outperforms previous models and competitors in various benchmarks, ensuring precise grading results. citeturn0search5

**4. Alternatives Considered**

**4.1. DeepSeek-R1**

* **Token Limit:** Offers a context window of **128,000 tokens**, which is substantial but less than Claude 3.5 Sonnet's capacity. citeturn0search0
* **Cost-Effectiveness:** More affordable at **$0.55 per million input tokens** and **$2.19 per million output tokens**, making it approximately 6.6 times cheaper than Claude 3.5 Sonnet. citeturn0search0
* **Scalability:** Utilizes a "mixture of experts" approach, enhancing efficiency and scalability. citeturn0search0
* **Accuracy:** Excels in tasks requiring deep reasoning, coding, and mathematical problem-solving. citeturn0search4

**Reason for Not Choosing:** Despite its cost advantages and strong performance in specific areas, DeepSeek-R1's lower token limit and focus on specialized tasks make it less suitable for our broad short-answer grading requirements.

**4.2. Llama 3**

* **Token Limit:** Supports a context window of up to **128,000 tokens**, similar to DeepSeek-R1.
* **Cost-Effectiveness:** As an open-source model, it allows for self-hosting, potentially reducing operational costs.
* **Scalability:** Designed to handle complex tasks efficiently, suitable for large-scale applications.
* **Accuracy:** Trained on a diverse dataset, offering improved performance over its predecessors.

**Reason for Not Choosing:** While Llama 3 offers flexibility and cost benefits due to its open-source nature, the infrastructure and maintenance required for self-hosting, along with its lower token limit compared to Claude 3.5 Sonnet, present challenges for our immediate deployment needs. It also would require an ML/AI team to maintain the model which Certifiable Inc. currently does not have.

**4.3. OpenAI's o3-mini**

* **Token Limit:** Supports a context window of up to **64,000 tokens**, which is significantly lower than Claude 3.5 Sonnet's capacity.
* **Cost-Effectiveness:** More affordable than Claude 3.5 Sonnet, making it a cost-effective option for many users. citeturn0search8
* **Scalability:** Suitable for large-scale applications requiring efficient processing of mathematical queries or basic text generation.
* **Accuracy:** While capable, it may not match the advanced reasoning capabilities of Claude 3.5 Sonnet, especially for complex coding tasks.

**Reason for Not Choosing:** The lower token limit and potentially reduced accuracy in complex tasks make o3-mini less aligned with our requirements for processing large batches of short-answer responses with high precision.

**5. Architecture Impact**

* **Integration:** Incorporate Claude 3.5 Sonnet via the Anthropic API into the existing grading infrastructure.
* **Batch Processing:** Leverage the 200K token context window to process multiple responses simultaneously, reducing API calls and latency.
* **Cost Monitoring:** Implement tracking mechanisms to monitor token usage and manage expenses effectively.

**6. Risks & Mitigation**

* **Risk:** Potential for higher-than-expected operational costs due to output token generation.
  + **Mitigation:** Optimize prompts to minimize unnecessary output tokens and regularly review usage patterns.
* **Risk:** Integration challenges with existing systems.
  + **Mitigation:** Conduct thorough testing in a staging environment before full deployment.

**7. Acceptance Criteria**

* **Cost Efficiency:** Maintain operational costs at or below **$0.018 per API call**, based on the pricing of $3 per million input tokens and $15 per million output tokens.
* **Processing Capacity:** Ability to handle up to **200,000 tokens** per API call, facilitating the grading of large batches of responses efficiently.
* **Accuracy:** Achieve a high correlation between AI-generated grades and human evaluations, targeting an agreement rate of **90% or higher**.

**8. Implementation Plan**

* **Phase 1:** Set up and configure access to the Claude 3.5 Sonnet API.
* **Phase 2:** Develop and test integration modules for batch processing.
* **Phase 3:** Deploy the application