AI-Generated Assessments Project: Start-Up Report

Overview

This report outlines the initial approach and methodology for the AI-Generated Assessments Project, as contracted through ConCOVE (Centre of Vocational Excellence for Construction & Infrastructure). The intention here is to outline the rationale for the AI models that have been selected for this project as well as providing an update on onboarding and training for the experts being utilised for for this project.

AI Models and Approach

Since the inception of this research project in early February, there have been three significant step changes in frontier AI models, alongside the introduction of powerful open-source alternatives. These advancements have substantially improved our testing outputs from early February to the present day, offering exciting possibilities for our mahi.

Model Selection for Phase 1 Research

The AI model selection process was based on three key criteria:

- Performance: The model's ability to generate high-quality, relevant assessments
- Cost: Ensuring the model provides value for money and fits within project budgets
- Alignment with project objectives: The model's features and capabilities must closely match our specific assessment generation needs

It is important to note here that AI technology is changing at a rapid pace. What is being explored and discussed here is therefore a starting-off point for this piece of research, rather than an end-point. The intention therefore, is to commit to Claude for the whole research project and to act as a control. In phase 2, the personalisation, other models will be tested and explored to understand how close they get to or exceed Claude.

While promising, the new OpenAI model (GPT-4o1) has limited access and produces inconsistent results. Open-source models like Meta's LLaMA perform well but significantly trail behind Anthropic and OpenAI's offerings.

CONCOSE TŪHURA

Claude 3.5 Sonnet

Claude 3.5 Sonnet, developed by Anthropic, stood out as a superior choice for generative AI assessments due to its exceptional performance, advanced capabilities, and unique features:

Unparalleled Performance

Claude 3.5 Sonnet sets new industry benchmarks across various cognitive tasks:

- **Graduate-level reasoning**: Excels in complex problem-solving and analysis.
- **Undergraduate-level knowledge**: Demonstrates broad expertise across multiple disciplines.
- **Coding proficiency**: Solves 64% of coding problems, a significant improvement over its predecessor.
- **High scores on knowledge-based benchmarks**: Notably, it achieved 88.70% on the MMLU (Massive Multitask Language Understanding) benchmark. MMLU is based on multiple-choice assessment questions across many areas of study, making it particularly relevant for our assessment generation tasks.

Advanced Capabilities

1. **Natural Language Processing**: Exhibits remarkable understanding of nuance, context, and complex instructions.
2. **Content Generation**: Produces high-quality, engaging content with a natural, relatable tone.
3. **Visual Reasoning**: Demonstrates superior capabilities in interpreting charts, graphs, and images.
4. **Multitask Proficiency**: Excels in various domains, including legal, finance, and philosophy.

Speed and Efficiency

Claude 3.5 Sonnet operates at twice the speed of its predecessor, making it ideal for complex, time-sensitive assessment tasks.

Workbench and Evaluation Features

A significant advantage of Claude 3.5 Sonnet is its integrated workbench and evaluation features. These tools greatly assist in the development and refinement of AI-powered assessments:

- **Real-time Collaboration**: The Artifacts feature creates a dynamic workspace for integrating AI-generated content into projects and workflows.

- **Iterative Development**: Allows for rapid prototyping and testing of assessment items and rubrics.
- **Performance Analysis**: Enables detailed evaluation of the model's responses, facilitating continuous improvement of assessment quality.

Safety and Ethical Considerations

Claude 3.5 Sonnet maintains high standards of safety and privacy, rated at ASL-2, ensuring responsible and ethical use in educational contexts.

By leveraging Claude 3.5 Sonnet's advanced capabilities, speed, and integrated development features, educators and assessment designers can create more sophisticated, accurate, and engaging generative AI assessments. This model's combination of intelligence, efficiency, and user-friendly tools makes it an optimal choice for pushing the boundaries of AI-assisted educational evaluation.

Expert Onboarding and Evaluation

A critical component of our research methodology is the involvement of subject matter experts. We have carefully onboarded and trained these experts to provide effective feedback and guidance throughout the assessment generation process. This training ensures that our experts:

1. Understand the specific objectives of AI-generated assessments
2. Can evaluate the content for accuracy, relevance, and alignment with VET standards
3. Provide constructive feedback that can be used to refine and improve the AI's output
4. Consider cultural and ethical implications in their reviews

Their expertise and insights will be invaluable in refining our AI-generated assessments and ensuring they meet the high vocational education and training standards.

Quality and relevance of generated questions

Human Benchmark Creation

To objectively measure the quality of our AI-generated assessments, we will create a "human AI" assessment. This will serve as our control, treated identically to AI-generated content during evaluation, ensuring a fair and robust comparison. We have successfully onboarded a control assessment writer who will create this human-generated assessment, focusing specifically on the Trade Essential Micro-Credential.

CONCOR TŪHURA

Success Metrics

Our ultimate goal is to achieve AI-generated assessments that require no human intervention ("no human in the loop"). However, we recognise that due to the inherent creativity of Generative AI, some level of human oversight may be necessary, regardless of our best efforts to eliminate errors.

To quantify this, we've developed the following success metrics:

- Human-in-the-Loop Factor: This will be an estimated time based on:
  a) How long it would take to review the AI-generated assessment
  b) The estimated time to fix any identified issues

  This factor will be incorporated into the overall Assessment rating provided by our expert reviewers.

- Quality Comparison: We'll compare the AI-generated assessments against our human-created benchmark, evaluating factors including:

  - Accuracy of content
  - Alignment with learning outcomes
  - Clarity and coherence of questions
  - Appropriateness of difficulty level

- Iteration Efficiency: We'll measure how quickly we can adapt the AI prompts and methodology from feedback and improve its output over successive generations.

- Diversity and Inclusivity: We'll assess how well the AI-generated content caters to diverse learner needs and avoids bias, ensuring our assessments are inclusive and representative.

These metrics will provide a comprehensive view of the AI's performance and help us identify areas for improvement as we progress through the project phases.

Next Steps

As we progress through Phase 1, we are looking forward to producing:

1. Literature Review:

   - Detailed analysis of AI's performance in assessment generation
   - Insights gained from expert reviews and iterations

- Comparisons between AI-generated and human-created assessments
- Implications for the VET sector and potential for wider application

2. AI-generated assessment:
    - Develop and refine AI models and prompts for assessment generation.
    - Iteratively improve the generated assessment utilising expert panel feedback.
    - Produce an AI generated assessment for Trades Essentials (Micro-credential) which passes moderation standards and meets expert panel requirements.