



A Multimodal AI Framework for Automated Grading of Digital and Text-Based Handwritten Exams

M.Y. Virajani¹, K.D.S. Udayanthika¹, W.A.S. Wahalathanthri¹, K.N. Poornima¹, G.G.N. Sandamali^{1}*

¹ Department of Electrical and Information Engineering, Faculty of Engineering, University of Ruhuna

**Corresponding author: nadeesha@eie.ruh.ac.lk*

Manual examination grading is time-consuming, inconsistent, and difficult to scale. This study presents a multimodal Intelligent Exam Evaluation System for automated grading of digital and handwritten scripts across various question types, including short answers, lists, essays, equations, diagrams, graphs, and multiple-choice questions (MCQs). The framework integrates Transformer-based OCR (TrOCR), Retrieval-Augmented Generation, and multiple large language models in a context-grounded evaluation pipeline. Handwritten text is extracted using TrOCR, achieving 93.24% character-level and 76.59% word-level accuracy. Grading is based on rubric-aligned semantic evaluation, grounded in lecture materials and model answers stored as vector embeddings rather than exact word matching. Text-based grading was evaluated using OpenAI GPT, Google Gemini, Anthropic Claude, DeepSeek, and a fine-tuned DeepSeek-R1 model. Zero-shot and few-shot prompting were applied strictly as inference-time strategies to guide rubric-aligned evaluation. Fine-tuning was performed through continued pre-training on domain-specific academic text in structured JSON format to enhance subject-domain understanding. Evaluation included 500 text answers, 350 equation responses, 250 diagram answers, 250 graph answers, and 200 handwritten responses. Temperature sensitivity analysis at 0.0 and 0.2 showed that few-shot prompting at 0.0 provided the most stable and reproducible grading, suitable for high-stakes academic assessment. At this setting, the system achieved the highest matching accuracy of 72% for short answers using Gemini, 80% for list-type using OpenAI, 60% for essays using Claude, and 63.43% for equations using OpenAI. Multimodal evaluation achieved a strong correlation with human scores, with Pearson's correlation coefficient of 0.784 for diagrams using Claude and 0.787 for graphs using Gemini. Compared to prior automated grading systems focused primarily on MCQs or text-only responses, the proposed framework supports handwritten text, equations, and visual content within a single scalable pipeline.

Keywords: Automated Grading, Handwritten Text Recognition, Large Language Models, Multimodal Evaluation, Retrieval-Augmented Generation