ABSTRACT

Manual assessment of short-answer responses is prone to subjectivity and requires considerable time, motivating the development of Automatic Short Answer Grading (ASAG) systems capable of providing consistent and automated evaluation. Although various approaches have been proposed, studies directly comparing Semantic Embeddings based on Sentence-BERT (SBERT) and Generative Large Language Models (LLMs) across different question types while considering operational efficiency remain limited. This study compares six ASAG methods, namely SBERT zero-shot, Fine-tuned SBERT, GPT-5.4-nano zero-shot, GPT-5.4-nano few-shot, GPT-5.4 zero-shot, and GPT-5.4 few-shot, using the Mohler ASAG dataset, which is divided into Open-Ended (OE) and Close-Ended (CE) subsets. For the SBERT-based approach, the all-distilroberta-v1 model was fine-tuned using CosineSimilarityLoss on a training set balanced through LLM-based data augmentation. Meanwhile, the Generative LLM approach employed GPT API calls with zero-shot and few-shot prompting strategies. Performance was evaluated using Accuracy, Precision, Recall, and F1-score, while inference time and estimated API cost were used to assess operational efficiency. The results show that Fine-tuned SBERT achieved the highest F1-score on the OE subset (0.9241), whereas GPT-5.4 few-shot obtained the best performance on the CE subset with an F1-score of 0.9750. Furthermore, fine-tuning consistently improved SBERT performance, while the effectiveness of few-shot prompting depended on the model capacity. This study concludes that no single method consistently outperforms others across all ASAG question types. Fine-tuned SBERT is more suitable for Open-Ended questions, whereas GPT-5.4 few-shot performs better on Close-Ended questions. Therefore, the selection of an ASAG approach should be tailored to the characteristics of the questions and the operational efficiency requirements.

Keywords : Automated Short Answer Grading, Sentence-BERT, Generative Pre-trained Transformer, binary classification, fine-tuning, prompt engineering