

Can AI Grade Like Humans? Agreement and Reliability in Exam Assessment

Ricardo Gouveia Rodrigues ¹, Andreia Ferreira ², Nathalia Suchek ³

¹²³ NECE, Faculty of Social and Human Sciences, Universidade da Beira Interior, Portugal

[https://doi.org/10.35609/gcbssproceeding.2026.1\(98\)](https://doi.org/10.35609/gcbssproceeding.2026.1(98))

ABSTRACT

Artificial intelligence (AI) is rapidly transforming higher education, with assessment emerging as one of its most impactful and contested applications. The automation of grading processes promises substantial improvements in efficiency, scalability, and standardisation. However, these potential benefits are accompanied by growing concerns about the reliability, consistency, fairness, and transparency of AI-based evaluation systems. These concerns are particularly critical in high-stakes academic assessment, where grading accuracy directly influences student outcomes and institutional credibility. Traditional grading relies on human evaluators who contribute contextual understanding, disciplinary expertise, and interpretative judgement. Nevertheless, human assessment is subject to limitations, including fatigue, variability, and potential bias. In contrast, AI systems are often assumed to provide more objective and consistent evaluations. Despite these assumptions, empirical evidence remains mixed. Recent studies suggest that while large language models (LLMs) can approximate human scoring patterns, they often exhibit weaker inter-rater agreement and inconsistent grading behaviour across different contexts. Moreover, existing research tends to focus on isolated performance indicators, such as accuracy or correlation, without simultaneously addressing both agreement between evaluators and internal consistency within evaluator groups. This gap limits the ability to fully evaluate the effectiveness and reliability of AI-based grading systems.

JEL Codes:

Keywords: artificial intelligence in education, automated assessment, inter-rater reliability, large language models, exam grading