INTISARI

Latar Belakang: Penggunaan kecerdasan buatan dalam pendidikan kedokteran terus berkembang dan membuka peluang untuk mendukung proses penilaian serta pemberian umpan balik. Dalam tugas esai yang bersifat formatif, umpan balik yang tepat waktu, konsisten, dan berkualitas merupakan hal yang sangat penting. Namun, penilaian dan pemberian umpan balik oleh dosen seringkali menyita waktu yang tidak sedikit. Bukti mengenai reliabilitas penilaian dan persepsi terhadap kualitas umpan balik menggunakan GPT-5 dalam pendidikan kedokteran masih sangat terbatas.

Tujuan: Mengevaluasi reliabilitas penilaian berbasis GPT-5 dibandingkan penilaian dosen, serta membandingkan persepsi kualitas umpan balik tertulis dari GPT-5 dan dosen pada esai formatif mahasiswa kedokteran tahun kedua.

Metode: Penelitian kuantitatif ini melibatkan 61 mahasiswa kedokteran preklinik tahun kedua dalam blok metodologi penelitian. Mahasiswa mengerjakan empat soal tugas esai formatif. GPT-5 dan dosen secara independen menilai esai serta memberikan umpan balik tertulis menggunakan rubrik yang sama. Reliabilitas penilaian dianalisis menggunakan Intraclass Correlation Coefficient (ICC), sedangkan kualitas umpan balik dinilai oleh mahasiswa dan pakar menggunakan rubrik kualitas umpan balik naratif. Perbandingan berpasangan dilakukan menggunakan uji Wilcoxon signed-rank.

Hasil: Kesesuaian penilaian antara GPT-5 dan dosen tergolong baik (ICC[3,1] = 0,808; 95% CI, 0,699–0,880). Baik mahasiswa maupun pakar menilai kualitas umpan balik GPT-5 lebih tinggi dibandingkan umpan balik dosen ($p < 0,001$).

Kesimpulan: GPT-5 menunjukkan reliabilitas penilaian yang baik, tetapi cenderung memberikan skor yang lebih tinggi sehingga memerlukan kalibrasi dan pengawasan oleh dosen. GPT-5 berpotensi menjadi pelengkap peran dosen dalam penilaian formatif, bukan sebagai pengganti utama.

Kata kunci: Kecerdasan buatan, Penilaian, Reliabilitas, Esai, Umpan balik

ABSTRACT

Introduction: The use of artificial intelligence in medical education is rapidly advancing, creating new opportunities to support assessment and feedback processes. In formative essay assignments, timely, consistent, and high-quality feedback is essential. However, grading and providing feedback by faculty members is often time-consuming. Evidence regarding the reliability of AI-based assessment and perceptions of feedback quality using GPT-5 in medical education remains limited.

Purpose: To evaluate the reliability of GPT-5-based assessment compared to faculty grading, and to compare perceptions of written feedback quality between GPT-5 and faculty on formative essays submitted by second-year medical students.

Methods: This quantitative study involved 61 preclinical second-year medical students enrolled in a research methodology block. Students completed four formative essay assignments. GPT-5 and faculty independently graded the essays and provided written feedback using an identical rubric. Assessment reliability was analyzed using the Intraclass Correlation Coefficient (ICC), while feedback quality was evaluated by both students and experts using a narrative feedback quality rubric. Pairwise comparisons were conducted using the Wilcoxon signed-rank test.

Results: Agreement between GPT-5 and faculty assessments was good (ICC[3,1] = 0.808; 95% CI, 0.699–0.880). Both students and experts rated GPT-5 feedback quality significantly higher than faculty feedback ($p < 0.001$).

Conclusion: GPT-5 demonstrated good assessment reliability, although it showed a tendency to assign higher scores, indicating the need for calibration and human oversight. GPT-5 has the potential to serve as a complement to faculty in formative assessment, rather than as a primary replacement.

Keywords: Artificial intelligence, Assessment, Reliability, Essay, Feedback