

Deteksi Jawaban Buatan AI pada Tugas Pemrograman Berbasis Sistem Penilaian Otomatis

¹Ahmad Muyassar Ibrahim

¹Program Studi Teknik Informatika, UIN Alauddin Makassar, Indonesia
Email: ahmad.muyassar@uin-alauddin.ac.id^{1*}

Received : 01 Januari 2026
Accepted : 27 Februari 2026
Published : 30 Maret 2026

ABSTRAK

Pemanfaatan kecerdasan buatan (AI) oleh mahasiswa dalam menyelesaikan tugas pemrograman terus mengalami peningkatan dan menimbulkan tantangan baru dalam menjaga integritas akademik. Penelitian ini mengkaji pola serta tingkat kemunculan jawaban yang dihasilkan oleh AI pada tugas pemrograman yang dikumpulkan melalui Learning Management System (LMS) dengan fitur penilaian otomatis. LMS yang digunakan dikembangkan dengan model ADDIE dan dilengkapi mekanisme deteksi AI terintegrasi sebagai media pengumpulan data. Penelitian ini melibatkan 109 mahasiswa dari dua mata kuliah, yaitu Pemrograman Web Dasar dan Pemrograman Perangkat Bergerak, di UIN Alauddin Makassar pada Semester Genap 2026/2027. Dari 186 submission tugas yang terkumpul, sistem mengidentifikasi 49 submission atau 26,3% sebagai jawaban yang diduga dibuat menggunakan AI. Analisis dilakukan terhadap distribusi flagging pada setiap mata kuliah, perbandingan nilai sebelum dan setelah penalti, serta karakteristik umum jawaban yang terindikasi AI. Hasil penelitian menunjukkan bahwa tingkat penggunaan AI pada tugas pemrograman berbahasa Indonesia tergolong cukup tinggi, dengan TGS-642 pada mata kuliah Pemrograman Web mencatat persentase flagging tertinggi sebesar 41,5%. Penelitian ini memberikan kontribusi dalam memahami fenomena penggunaan AI generatif pada pendidikan pemrograman berbahasa Indonesia serta menawarkan rekomendasi praktis untuk pengelolaan integritas akademik.

Kata Kunci: Deteksi AI, Integritas Akademik, Kecerdasan Buatan, Pemrograman, Tugas Otomatis

ABSTRACT

The use of artificial intelligence (AI) by students in completing programming assignments continues to increase and has created new challenges in maintaining academic integrity. This study examines the patterns and prevalence of AI-generated answers in programming assignments submitted through a Learning Management System (LMS) equipped with automatic assessment features. The LMS used in this study was developed using the ADDIE model and integrated with an AI detection mechanism as the data collection platform. This study involved 109 students from two courses, namely Basic Web Programming and Mobile Programming, at UIN Alauddin Makassar during the Even Semester of the 2026/2027 academic year. From a total of 186 assignment submissions, the system identified 49 submissions, or 26.3%, as answers suspected to have been generated using AI. The analysis covered the distribution of flagging across courses, score comparisons before and after penalties, and the general characteristics of answers indicated as AI-generated. The results show that the level of AI use in Indonesian-language programming assignments is relatively high, with TGS-642 in the Web Programming course recording the highest flagging percentage at 41.5%. This study contributes to understanding the phenomenon of generative AI use in Indonesian programming education and offers practical recommendations for managing academic integrity.

Keywords: AI Detection, Academic Integrity, Artificial Intelligence, Programming, Automated Assessment

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license



1. PENDAHULUAN

Kemajuan pesat kecerdasan buatan generatif, khususnya *Large Language Model* (LLM) seperti ChatGPT, Gemini, dan Claude, telah mengubah cara mahasiswa berinteraksi dengan tugas akademik secara fundamental (Cotton et al., 2024). Kemampuan LLM dalam menghasilkan kode program yang fungsional dan terstruktur dengan baik dari sebuah deskripsi sederhana menjadikannya alat yang sangat menarik bagi mahasiswa pemrograman (Denny et al., 2024). Di satu sisi, ini berpotensi mendukung pembelajaran; di sisi lain, penggunaan yang tidak diungkapkan menimbulkan permasalahan serius terhadap integritas akademik dan validitas evaluasi pembelajaran.

Permasalahan ini semakin relevan di Indonesia seiring meningkatnya penetrasi penggunaan AI di kalangan mahasiswa perguruan tinggi. Mata kuliah pemrograman menjadi salah satu yang paling rentan karena output-nya berupa kode program dan dapat dengan mudah dihasilkan oleh AI tanpa mahasiswa memahami logika dibaliknya (Perkins et al., 2024). Kondisi ini menciptakan ilusi kompetensi yang berbahaya bagi perkembangan kemampuan teknis mahasiswa dan melemahkan fungsi evaluasi sebagai alat ukur capaian pembelajaran.

Secara global, adopsi kecerdasan buatan dalam pendidikan tinggi telah berkembang pesat pada berbagai fungsi, mulai dari sistem tutor cerdas, analitik pembelajaran, hingga penilaian otomatis (Zawacki-Richter et al., 2019). *Learning Management System* (LMS) menjadi infrastruktur utama yang menghubungkan mahasiswa dengan proses evaluasi digital, sehingga menjadikannya titik strategis untuk memantau perilaku pengumpulan tugas secara sistematis (Turnbull et al., 2021). Ketika mekanisme pengumpulan dan penilaian tugas telah terdigitalisasi penuh, peluang untuk mengintegrasikan deteksi penggunaan AI secara otomatis menjadi terbuka, sesuatu yang sulit diwujudkan pada alur pengumpulan tugas konvensional berbasis kertas atau surel.

Pada saat yang sama, LLM juga semakin banyak dimanfaatkan pada sisi pengajar sebagai mesin penilai otomatis. Sejumlah studi menunjukkan bahwa LLM mampu menilai tugas pemrograman dengan akurasi yang mendekati penilai manusia (Jukiewicz, 2025; Mohamed et al., 2025). Namun demikian, penerapan LLM di ruang kelas nyata juga menyingkap persoalan integritas: Chiang et al. (2024) melaporkan bahwa dalam perkuliahan dengan lebih dari seribu mahasiswa, sebagian mahasiswa justru memanipulasi evaluator berbasis LLM melalui teknik *prompt hacking* untuk memperoleh skor tinggi tanpa memenuhi rubrik. Temuan tersebut menegaskan bahwa relasi antara mahasiswa dan AI dalam konteks evaluasi bersifat dua arah, yaitu AI dapat menjadi alat bantu penilaian sekaligus alat untuk mengelabui penilaian, sehingga mekanisme deteksi menjadi komponen yang tidak terpisahkan dari sistem penilaian otomatis.

Beberapa penelitian telah mengidentifikasi tingginya prevalensi penggunaan AI dalam tugas akademik secara global. Cotton et al. (2024) mendokumentasikan normalisasi cepat penggunaan ChatGPT dalam pengerjaan tugas mahasiswa. Perkins et al. (2024) menunjukkan bahwa deteksi teks buatan AI menggunakan kombinasi penilaian akademik dan perangkat lunak memberikan hasil lebih baik dibanding menggunakan salah satunya saja. Namun, hampir seluruh studi dilakukan pada konteks berbahasa Inggris. Sejauh ini belum ada penelitian yang secara empiris mengukur prevalensi dan pola penggunaan AI generatif pada tugas pemrograman berbahasa Indonesia dalam konteks nyata perkuliahan.

Dari perspektif teoretis, fenomena ini menyentuh dua konsep fundamental dalam evaluasi pendidikan, yaitu integritas akademik dan validitas asesmen. Integritas akademik menghendaki bahwa karya yang dikumpulkan mahasiswa benar-benar merepresentasikan usaha dan pemahamannya sendiri, sedangkan validitas asesmen menghendaki bahwa skor yang diberikan benar-benar mengukur kompetensi yang dituju. Jawaban buatan AI yang tidak diungkapkan melanggar keduanya sekaligus: karya bukan milik mahasiswa dan skor tidak lagi mencerminkan kemampuan sesungguhnya. Oleh karena itu, pengukuran empiris terhadap prevalensi jawaban buatan AI merupakan langkah awal yang diperlukan sebelum institusi dapat merumuskan kebijakan pengelolaan AI yang proporsional, tidak terlalu longgar sehingga evaluasi kehilangan makna, tetapi juga tidak terlalu represif sehingga menutup peluang pemanfaatan AI yang etis.

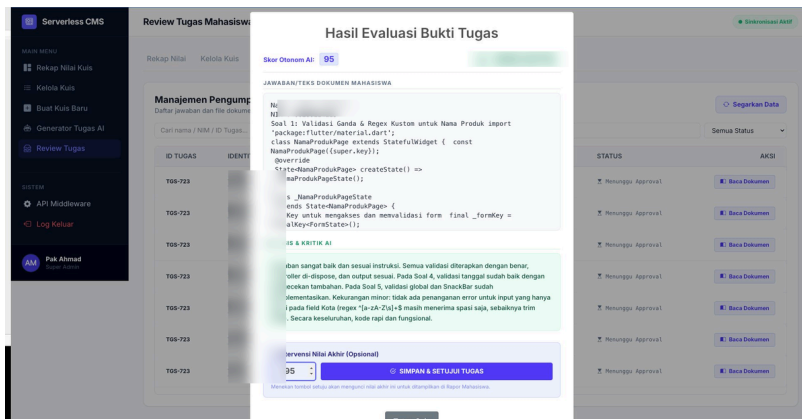
Penelitian ini mengisi celah tersebut dengan menganalisis 186 submission tugas nyata dari dua mata kuliah pemrograman di UIN Alauddin Makassar yang dikumpulkan melalui sistem LMS terintegrasi dengan mekanisme deteksi AI otomatis. Tujuan penelitian ini adalah: (1) mengukur prevalensi jawaban berindikasi AI pada tugas pemrograman berbahasa Indonesia, (2) menganalisis pola distribusi flagging per mata kuliah dan per tugas, dan (3) mengidentifikasi karakteristik umum submission yang terdeteksi sebagai buatan AI.

2. METODE PENELITIAN

Penelitian ini merupakan penelitian deskriptif kuantitatif yang menganalisis data faktual dari sistem LMS yang aktif beroperasi. Data dikumpulkan secara *longitudinal* selama Semester Genap 2026/2027 di Program Studi Teknik Informatika, UIN Alauddin Makassar.

A. Platform dan Sistem Deteksi

Data bersumber dari sistem LMS berbasis web yang dikembangkan menggunakan model ADDIE (Adeoye, 2024; Branch, 2009) dan dideploy pada domain materiku.dosengpt.com. Sistem mengintegrasikan mekanisme deteksi AI yang bekerja secara otomatis saat mahasiswa mengumpulkan tugas. Mekanisme deteksi menganalisis karakteristik teks dan kode yang dikumpulkan menggunakan Anthropic Claude API dengan *prompt* terstruktur yang dirancang untuk mengidentifikasi pola khas jawaban buatan AI, seperti struktur yang terlalu sempurna, ketiadaan komentar kontekstual, konsistensi gaya penulisan yang tidak wajar, dan kurangnya jejak proses berpikir. *Submission* yang terdeteksi otomatis menerima penalti 50% dari skor asli sebelum dikunci ke rapor mahasiswa. Gambar 1 menampilkan antarmuka *review* tugas dengan mekanisme deteksi dan intervensi nilai.



Gambar 1. Antarmuka review tugas menampilkan skor otonom AI dan kolom deteksi buatan AI

B. Populasi dan Data

Populasi penelitian adalah seluruh *submission* tugas dari 109 mahasiswa terdaftar pada dua mata kuliah: Pemrograman Web Dasar (kelas A, B, E sebanyak 77 mahasiswa) dan Pemrograman Perangkat Bergerak (kelas B, C sebanyak 59 mahasiswa), dengan 28 mahasiswa mengambil keduanya. Total tiga tugas dianalisis: TGS-642 (Web, Pertemuan 3), TGS-655 (PPB, Pertemuan 6), dan TGS-723 (PPB, Pertemuan 9). Seluruh 186 *submission* yang masuk diikutsertakan dalam analisis (total sampling).

C. Teknik Analisis Data

Data dianalisis menggunakan statistik deskriptif. Variabel yang dianalisis meliputi: (1) jumlah dan persentase *submission* terdeteksi AI per tugas dan per mata kuliah, (2) distribusi skor sebelum dan sesudah penalti, dan (3) karakteristik umum *submission* berindikasi AI. Perbandingan skor menggunakan selisih rata-rata dan distribusi frekuensi.

3. HASIL DAN PEMBAHASAN

A. Prevalensi Deteksi Jawaban Buatan AI

Dari 186 total *submission* tugas yang diterima sistem, sebanyak 49 *submission* (26,3%) terdeteksi sebagai jawaban yang diduga dihasilkan oleh AI generatif. Tabel 1 menyajikan rekapitulasi lengkap per tugas.

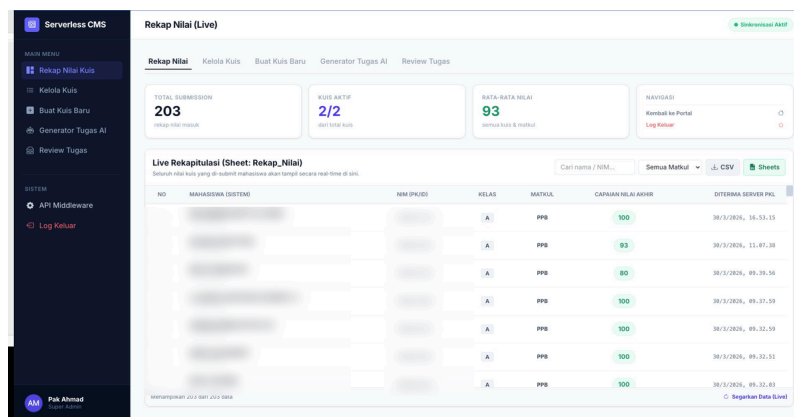
Tabel 1. Rekapitulasi Deteksi Jawaban Buatan AI per Tugas

ID Tugas	Pertemuan	Mata Kuliah	Total	Valid	AI Flag	% Flagging
TGS-642	Pertemuan 3	Pemrograman Web Dasar	41	24	17	41,5%
TGS-655	Pertemuan 6	Pemrograman Perangkat Bergerak	96	74	22	22,9%
TGS-723	Pertemuan 9	Pemrograman Perangkat Bergerak	49	39	10	20,4%
Total			186	137	49	26,3%

Angka 26,3% ini menunjukkan bahwa lebih dari seperempat mahasiswa yang mengumpulkan tugas menggunakan AI generatif untuk menghasilkan jawaban mereka. Angka ini konsisten dengan temuan Cotton et al. (2024) yang mendokumentasikan prevalensi tinggi penggunaan AI dalam tugas akademik, sekaligus menjadi validasi empiris pertama untuk konteks pemrograman berbahasa Indonesia.

B. Distribusi Flagging per Mata Kuliah

TGS-642 (Pemrograman Web, Pertemuan 3) mencatat persentase *flagging* tertinggi sebesar 41,5%, hampir dua kali lipat rata-rata keseluruhan. Ini dapat dikaitkan dengan topik pertemuan 3 yang relatif lebih dasar (HTML/CSS/JavaScript fundamental), sehingga AI lebih mudah menghasilkan kode yang benar secara sintaksis dan terlihat seperti pekerjaan manusia. Sebaliknya, TGS-723 (PPB, Pertemuan 9) yang membahas *Flutter state management* yang lebih kompleks hanya mencatat 20,4% *flagging*. Gambar 2 menampilkan dasbor rekap nilai *live* dengan 203 *submission* kuis yang menjadi konteks sistem keseluruhan.



Gambar 2. Dasbor rekap nilai *live* konteks sistem LMS terintegrasi

C. Perbandingan Skor Sebelum dan Sesudah Penalti

Seluruh *submission* yang terdeteksi AI menerima penalti 50% otomatis dari skor awal yang diberikan sistem penilaian AI. Tabel 2 menyajikan perbandingan rata-rata skor per tugas.

Tabel 2. Perbandingan Rata-rata Skor Sebelum dan Sesudah Penalti AI

ID Tugas	n Flagged	Rata Skor Awal	Rata Setelah Penalti	Selisih Rata-rata
TGS-642	17	71,2	35,6	-35,6 poin
TGS-655	22	78,4	39,2	-39,2 poin
TGS-723	10	80,1	40,1	-40,1 poin
Gabungan	49	76,6	38,3	-38,3 poin

Rata-rata skor awal *submission* berindikasi AI (76,6) justru lebih tinggi dari rata-rata skor valid keseluruhan (72,9), yang mengindikasikan bahwa AI menghasilkan kode yang terlihat lebih baik secara teknis namun tidak mencerminkan pemahaman mahasiswa yang sesungguhnya. Ini memperkuat argumen perlunya mekanisme deteksi: tanpa penalti, mahasiswa yang menggunakan AI justru mendapat nilai lebih tinggi.

D. Karakteristik Umum Submission Berindikasi AI

Berdasarkan analisis pola oleh mekanisme deteksi, *submission* yang terdeteksi AI umumnya menunjukkan karakteristik: (1) struktur kode yang sangat terorganisir dengan komentar dalam bahasa Inggris meskipun instruksi tugas dalam bahasa Indonesia, (2) implementasi fitur-fitur tambahan yang tidak diminta dalam soal namun umum dihasilkan AI, (3) ketiadaan keragaman penulisan yang biasanya muncul dari proses *trial-and-error* mahasiswa, dan (4) konsistensi gaya yang berlebihan dalam seluruh bagian kode. Temuan ini konsisten dengan karakteristik AI-generated text yang diidentifikasi Perkins et al. (2024) pada konteks esai akademik.

Temuan ini juga dapat dibandingkan dengan pengalaman internasional dalam skala yang jauh lebih besar. Chiang et al. (2024) menemukan bahwa ketika evaluator berbasis LLM digunakan pada perkuliahan lebih dari seribu mahasiswa, muncul perilaku manipulatif berupa *prompt hacking*, yaitu mahasiswa menyisipkan instruksi tersembunyi agar evaluator memberikan skor maksimal. Pola pada penelitian ini berbeda tetapi masih satu

rumpun: mahasiswa tidak memanipulasi evaluator, melainkan mendelegasikan pengerjaan tugas kepada AI. Keduanya menunjukkan bahwa setiap komponen otomatis dalam rantai evaluasi berpotensi dieksploitasi, sehingga desain sistem penilaian otomatis perlu mengantisipasi kedua arah penyalahgunaan tersebut secara bersamaan.

Dari sisi mata kuliah, Pemrograman Web Dasar menunjukkan prevalensi lebih tinggi dibanding Pemrograman Perangkat Bergerak. Ini sejalan dengan ketersediaan lebih banyak contoh kode HTML/CSS/JavaScript di training data LLM dibanding Flutter/Dart yang lebih spesifik. Denny et al. (2024) mencatat bahwa LLM lebih andal untuk bahasa pemrograman yang memiliki representasi lebih besar dalam data pelatihan mereka, sehingga output-nya lebih meyakinkan dan lebih sulit dibedakan dari kode buatan manusia.

Tingginya rata-rata skor awal submission berindikasi AI juga selaras dengan temuan literatur penilaian otomatis. Mohamed et al. (2025) dan Jukiewicz (2025) menunjukkan bahwa kode yang dihasilkan LLM cenderung unggul pada dimensi yang mudah diukur mesin, seperti kebenaran sintaksis, kerapian struktur, dan kelengkapan fitur, yang merupakan dimensi yang sama dengan rubrik penilaian otomatis. Akibatnya, terjadi bias sistematis yang menguntungkan jawaban buatan AI apabila deteksi tidak diterapkan. Hal ini menyiratkan bahwa sistem penilaian otomatis dan sistem deteksi AI perlu dipandang sebagai satu kesatuan desain, bukan dua komponen terpisah, karena akurasi penilaian tanpa jaminan keaslian justru dapat memperbesar ketidakadilan.

Fenomena ini memiliki implikasi pedagogis yang signifikan. Jika dibiarkan tanpa pendeteksian, mahasiswa yang menggunakan AI generatif akan lulus evaluasi tanpa benar-benar menguasai kompetensi yang diharapkan. Mekanisme deteksi terintegrasi dalam sistem seperti yang dikembangkan dalam penelitian ini menawarkan pendekatan proaktif yang lebih efektif dibanding mengandalkan pengawasan manual dosen, terutama di kelas besar dengan puluhan submission per sesi.

Dari sisi kebijakan institusi, temuan penelitian ini mendukung rekomendasi Perkins et al. (2024) bahwa penanganan penggunaan AI tidak dapat mengandalkan perangkat deteksi semata, melainkan perlu dikombinasikan dengan pertimbangan akademik dosen. Mekanisme yang diterapkan pada sistem ini mengikuti prinsip tersebut: deteksi bekerja otomatis, tetapi dosen tetap memegang kendali akhir melalui fitur intervensi nilai sebelum skor dikunci. Selain itu, tingginya flagging pada tugas bertopik dasar mengindikasikan perlunya desain soal yang lebih kontekstual dan personal, misalnya tugas yang menuntut penjelasan proses berpikir, refleksi atas kesalahan, atau keterkaitan dengan data spesifik kelas, sehingga jawaban tidak dapat dihasilkan utuh oleh AI tanpa keterlibatan kognitif mahasiswa.

E. Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan dalam menafsirkan hasil. Pertama, penandaan submission sebagai buatan AI didasarkan pada mekanisme deteksi otomatis yang belum divalidasi silang dengan konfirmasi langsung dari mahasiswa, sehingga kemungkinan positif palsu maupun negatif palsu tidak dapat sepenuhnya dihilangkan. Kedua, penelitian dilakukan pada satu institusi dengan tiga tugas dari dua mata kuliah, sehingga generalisasi temuan ke konteks yang lebih luas memerlukan kehati-hatian. Ketiga, penelitian ini tidak menelusuri motif mahasiswa dalam menggunakan AI, padahal pemahaman terhadap motif tersebut penting untuk merancang intervensi pedagogis yang tepat. Keterbatasan-keterbatasan ini sekaligus menjadi arah pengembangan bagi penelitian selanjutnya.

4. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengukur dan menganalisis prevalensi penggunaan AI generatif pada tugas pemrograman berbahasa Indonesia di UIN Alauddin Makassar. Dari 186 submission tugas nyata, 49 (26,3%) terdeteksi sebagai jawaban berindikasi buatan AI, dengan persentase tertinggi pada mata kuliah Pemrograman Web Dasar (41,5% pada TGS-642) dan terendah pada tugas Flutter yang lebih kompleks (20,4% pada TGS-723). Rata-rata skor awal *submission* berindikasi AI (76,6) lebih tinggi dari skor rata-rata keseluruhan (72,9), mengkonfirmasi bahwa tanpa mekanisme deteksi, mahasiswa pengguna AI justru mendapat keuntungan nilai yang tidak adil. Sistem deteksi terintegrasi dengan penalti otomatis 50% terbukti menjadi mekanisme penjaga integritas akademik yang efektif dan skalabel. Penelitian ini merupakan validasi empiris pertama tentang prevalensi dan pola penggunaan AI generatif pada tugas pemrograman berbahasa Indonesia. Untuk penelitian selanjutnya, disarankan untuk mengembangkan rubrik deteksi yang lebih granular berdasarkan kompleksitas topik, memperluas analisis ke perguruan tinggi lain di Indonesia, serta mengeksplorasi pendekatan pedagogis yang dapat menjaga integritas sekaligus memanfaatkan AI sebagai alat bantu belajar yang etis.

REFERENSI

- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239. <https://doi.org/10.1080/14703297.2023.2190148>
- Denny, P., Leinonen, J., Prather, J., Luxton-Reilly, A., Amarouche, T., Becker, B. A., & Reeves, B. N. (2024). Prompt problems: A new programming exercise for the generative AI era. *Proceedings of the 55th ACM Technical Symposium on Computer Science Education*, 296–302. <https://doi.org/10.1145/3626252.3630909>
- Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2024). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics*, 22(1), 89–113. <https://doi.org/10.1007/s10805-023-09492-6>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), Article 39. <https://doi.org/10.1186/s41239-019-0171-0>
- Chiang, C.-H., Chen, W.-C., Kuan, C.-Y., Yang, C., & Lee, H.-Y. (2024). Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, 2489–2513. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.146>
- Mohamed, K., Yousef, M., Medhat, W., Mohamed, E. H., Khoriba, G., & Arafa, T. (2025). Hands-on analysis of using large language models for the auto evaluation of programming assignments. *Information Systems*, 130, Article 102523. <https://doi.org/10.1016/j.is.2024.102523>
- Jukiewicz, M. (2025). A comparative study of large language models in programming education: Accuracy, efficiency, and feedback in student assignment grading. *Applied Sciences*, 15(18), Article 10055. <https://doi.org/10.3390/app151810055>
- Turnbull, D., Chugh, R., & Luck, J. (2021). Learning management systems, an overview. *Encyclopedia of Education and Information Technologies*. Springer. https://doi.org/10.1007/978-3-030-10576-1_82
- Branch, R. M. (2009). *Instructional design: The ADDIE approach*. Springer. <https://doi.org/10.1007/978-0-387-09506-6>
- Adeoye, M. A. (2024). Revolutionizing education: Unleashing the power of the ADDIE model for effective teaching and learning. *Jurnal Pendidikan Indonesia*, 13(1), 202–209. <https://doi.org/10.23887/jpiundiksha.v13i1.68624>