



Transformasi Evaluasi Pembelajaran di Era Kecerdasan Buatan Generatif: Kajian Pustaka Atas Redesain Asesmen Otentik, Validitas-Reliabilitas Instrumen, dan Pergeseran Evaluasi Formatif di Pendidikan Tinggi

Saidatun Nisa Nasution^{1*}

¹ Sekolah Tinggi Ilmu Tarbiyah Serdang Bedagai, Indonesia

DETAIL ARTIKEL	ABSTRAK
Riwayat: Diterima : 15 Juli 2026 Disetujui : 21 Juli 2026 Dipublikasikan : 23 Juli 2026	Proliferasi perangkat kecerdasan buatan generatif (AI generatif) telah mengguncang asumsi fundamental evaluasi pembelajaran di pendidikan tinggi, khususnya premis bahwa sebuah produk tulisan yang dikumpulkan mahasiswa secara andal mencerminkan kompetensi individu yang bersangkutan. Penelitian ini menerapkan desain kualitatif dengan pendekatan studi pustaka (library research) untuk mengkaji bagaimana bidang evaluasi pembelajaran merespons disrupsi tersebut, dengan tiga fokus, yaitu: (1) redesain instrumen asesmen menuju tugas otentik yang tahan terhadap AI; (2) validitas dan reliabilitas instrumen deteksi AI yang digunakan untuk menjaga integritas evaluasi; serta (3) pergeseran paradigma evaluasi sumatif menuju formatif yang dimungkinkan oleh umpan balik berbasis AI generatif. Data dihimpun dari jurnal terindeks, satu studi empiris multi-alat berskala internasional, dan literatur teori evaluasi klasik yang terbit dalam rentang 2011-2025, kemudian dianalisis secara tematik. Hasil kajian menunjukkan bahwa literatur redesain asesmen mengerucut pada tugas berorientasi proses dan berpikir tingkat tinggi yang secara struktural sulit diotomatisasi penuh, selaras dengan teori penyelarasan konstruktif (constructive alignment). Namun demikian, pengujian empiris terhadap alat deteksi AI mengungkap persoalan reliabilitas yang serius, dengan tingkat akurasi di bawah 80% pada mayoritas alat yang diuji serta risiko tuduhan keliru (false accusation) yang terdokumentasi, sehingga alat tersebut tidak layak dijadikan satu-satunya bukti dalam keputusan evaluasi yang berisiko tinggi. Sementara itu, umpan balik otomatis berbasis AI generatif menunjukkan potensi untuk menskalakan umpan balik formatif yang cepat, meski kualitas pedagogisnya masih bervariasi dan tetap memerlukan pengawasan manusia. Penelitian ini menyimpulkan bahwa sistem evaluasi yang dapat dipertanggungjawabkan di era AI generatif harus secara simultan meredesain objek yang dinilai, memperlakukan teknologi deteksi sebagai bukti pendukung dan bukan bukti final, serta mengarahkan ulang budaya evaluasi dari penjagaan gerbang sumatif menuju praktik formatif yang berkelanjutan dan dialogis.
Kata Kunci: <i>Evaluasi Pembelajaran, Kecerdasan Buatan Generatif, Asesmen Otentik.</i>	
*Korespondensi: Saidatun Nisa Nasution Sekolah Tinggi Ilmu Tarbiyah Serdang Bedagai saydanasty08@gmail.com	
Konflik Kepentingan: Penulis menyatakan tidak ada konflik kepentingan dalam penelitian ini.	
Cara Mengutip: Nasution, S. N. (2026). Transformasi Evaluasi Pembelajaran di Era Kecerdasan Buatan Generatif: Kajian Pustaka Atas Redesain Asesmen Otentik, Validitas-Reliabilitas Instrumen, dan Pergeseran Evaluasi Formatif di Pendidikan Tinggi. <i>Tarbiyah : Jurnal Ilmu Pendidikan dan Pengajaran</i> , 5(2), 63-69. https://doi.org/10.64464/tarbiyah.v5i2.269	
Lisensi:  © 2026 Penulis. Dilisensikan di bawah Lisensi Creative Commons Atribusi 4.0 Internasional (CC BY 4.0)	

ABSTRACT

The proliferation of generative artificial intelligence (generative AI) tools has shaken the fundamental assumptions of learning evaluation in higher education, particularly the premise that a piece of writing submitted by a student reliably reflects that individual's own competence. This study employs a qualitative design with a library research approach to examine how the field of learning evaluation is responding to this disruption, focusing on three areas: (1) the redesign of assessment instruments toward AI-resistant authentic tasks; (2) the validity and reliability of AI detection instruments used to safeguard evaluation integrity; and (3) the paradigm shift from summative toward formative evaluation enabled by generative AI-based feedback. Data were collected from indexed journals, one international-scale multi-tool empirical study, and classical evaluation theory literature published between 2011 and 2025, and were subsequently analyzed thematically. The findings indicate that the assessment redesign literature converges on process-oriented and higher-order thinking tasks that are structurally difficult to fully automate, in line with constructive alignment theory. However, empirical testing of AI detection tools reveals serious reliability issues, with accuracy rates below 80% for the majority of tools tested, as well as documented risks of false accusation, rendering such tools unsuitable as the sole evidence in high-stakes evaluation decisions. Meanwhile, automated feedback based on generative AI shows potential for scaling rapid formative feedback, although its pedagogical quality remains variable and continues to require human oversight. This study concludes that an accountable evaluation system in the era of generative AI must simultaneously redesign the object being assessed, treat detection technology as supporting rather than definitive evidence, and redirect the culture of evaluation from summative gatekeeping toward sustainable and dialogic formative practice.

A. PENDAHULUAN

Evaluasi pembelajaran senantiasa berdiri di atas satu asumsi dasar yang jarang dipersoalkan secara eksplisit: bahwa sebuah karya yang diserahkan mahasiswa, baik berupa esai, laporan, maupun jawaban ujian tertulis, merupakan representasi yang sah dari proses berpikir dan kompetensi individu yang menghasilkannya. Sepanjang sejarah pendidikan tinggi modern, berbagai kerangka teori evaluasi, mulai dari taksonomi ranah kognitif hingga model penyelarasan konstruktif (*constructive alignment*) yang menghubungkan capaian pembelajaran, aktivitas belajar, dan tugas asesmen dalam satu kesatuan yang koheren, dibangun di atas asumsi tersebut (Biggs & Tang, 2011, hlm. xxiii). Ketika asumsi dasar ini goyah, seluruh bangunan evaluasi yang berdiri di atasnya turut terguncang.

Guncangan itulah yang sesungguhnya sedang terjadi sejak kemunculan kecerdasan buatan generatif (*Generative Artificial Intelligence/GenAI*)

secara masif pada akhir 2022. Model bahasa besar seperti ChatGPT kini mampu menghasilkan esai, jawaban analitis, bahkan laporan penelitian yang secara kualitas sulit dibedakan dari karya manusia dalam sekali perintah (Mishra, Warr, & Islam, 2023). Bagi bidang keilmuan evaluasi pembelajaran, implikasinya bersifat mendasar: jika sebuah tugas tertulis dapat diselesaikan sepenuhnya oleh mesin tanpa keterlibatan kognitif mahasiswa, maka nilai yang dihasilkan dari tugas tersebut kehilangan validitasnya sebagai indikator capaian belajar, betapapun rapinya instrumen penilaian yang digunakan untuk mensukornya. UNESCO, melalui panduan globalnya, secara eksplisit menyerukan agar institusi pendidikan meninjau ulang metode asesmen mereka sebagai salah satu langkah mendesak dalam merespons GenAI (UNESCO, Miao, & Holmes, 2023).

Respons dunia akademik terhadap persoalan ini bergerak, secara garis besar, pada tiga arah. Arah pertama bersifat ofensif-adaptif: alih-alih mempertahankan

bentuk asesmen lama, sejumlah pendidik mulai merancang ulang instrumen evaluasi agar berfokus pada proses, konteks nyata, dan keterampilan berpikir tingkat tinggi yang secara struktural lebih sulit diotomatisasi penuh oleh AI (Khlaif, dkk: 2025). Arah kedua bersifat defensif-teknis: institusi berupaya menjaga integritas evaluasi dengan mengandalkan perangkat lunak pendeteksi teks hasil AI, meski keandalan perangkat ini sendiri menjadi bahan perdebatan serius di kalangan peneliti (Weber-Wulff dkk., 2023). Arah ketiga bersifat transformatif-paradigmatik: alih-alih memandang AI generatif semata sebagai ancaman, sebagian pendidik mulai mengintegrasikannya sebagai mesin pemberi umpan balik formatif yang dapat menyediakan respons cepat dan berkelanjutan bagi mahasiswa (Lee & Moore, 2024).

Persoalannya, ketiga arah respons tersebut jarang dibahas secara terintegrasi dalam satu kajian. Literatur mengenai redesain asesmen umumnya berdiri sendiri dari literatur mengenai validitas alat deteksi, dan keduanya jarang disandingkan dengan kajian pergeseran ke evaluasi formatif berbasis AI, padahal ketiganya sesungguhnya merupakan tiga respons terhadap satu akar persoalan yang sama: krisis validitas evaluasi di tengah otomatisasi produksi teks. Di sinilah gap research yang hendak diisi oleh penelitian ini. Konteks Indonesia menambah urgensi tersendiri, mengingat kajian kualitatif terhadap mahasiswa di Kota Kupang menunjukkan bahwa AI generatif telah terintegrasi secara mendalam dalam praktik akademik sehari-hari mahasiswa, jauh mendahului kesiapan sistem evaluasi institusi untuk meresponsnya secara memadai (Ataupah, dkk: 2025), sementara Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi baru mulai menyusun panduan penggunaannya di ranah pembelajaran pada 2024 (Direktorat Pembelajaran dan Kemahasiswaan, 2024).

Kebaruan (*novelty*) yang ditawarkan penelitian ini terletak pada penyatuan tiga respons evaluasi pembelajaran tersebut, redesain instrumen, validitas-reliabilitas alat deteksi, dan pergeseran ke evaluasi formatif, ke dalam satu kerangka analisis yang koheren dan secara khusus ditinjau dari perspektif keilmuan evaluasi pembelajaran, bukan sekadar teknologi pendidikan secara umum. Berdasarkan uraian di atas, penelitian ini mengarahkan

pembahasannya pada tiga fokus, yaitu: (1) redesain instrumen dan desain asesmen otentik sebagai respons terhadap AI generatif; (2) validitas dan reliabilitas instrumen evaluasi, khususnya problematika alat deteksi teks hasil AI; serta (3) pergeseran paradigma evaluasi dari sumatif menuju formatif melalui pemanfaatan umpan balik berbasis AI generatif. Ketiga fokus ini akan didalami secara komprehensif pada bagian Hasil dan Pembahasan.

B. METODE

Penelitian ini menggunakan pendekatan kualitatif dengan desain studi pustaka (*library research*), yang mengandalkan sumber data sekunder berupa dokumen ilmiah tanpa memerlukan pengumpulan data lapangan secara langsung. Pendekatan ini relevan dengan tujuan penelitian yang bersifat sintesis-konseptual terhadap teori dan temuan empiris yang telah dipublikasikan mengenai evaluasi pembelajaran di era AI generatif, sebagaimana lazim digunakan dalam kajian tinjauan literatur pada bidang asesmen pendidikan tinggi (Gonsalves, 2025).

Sumber data terdiri atas artikel jurnal terindeks bereputasi, satu laporan studi empiris multi-alat berskala internasional mengenai keandalan perangkat deteksi AI, dokumen panduan kelembagaan dari UNESCO dan Kemendikbudristek, serta buku teori evaluasi pembelajaran klasik, dengan rentang waktu publikasi 2011-2025. Rentang ini sengaja mencakup literatur teori evaluasi sebelum era GenAI (seperti teori penyelarasan konstruktif) untuk memastikan bahwa analisis tidak semata reaktif terhadap teknologi terbaru, melainkan tetap berpijak pada fondasi keilmuan evaluasi pembelajaran yang telah mapan. Pengumpulan data dilakukan melalui penelusuran basis data ilmiah seperti Google Scholar, ScienceDirect, Springer, dan MDPI, menggunakan kata kunci “assessment redesign generative AI”, “AI detection tool validity reliability”, dan “formative feedback generative AI higher education”. Kriteria inklusi meliputi: (a) relevansi tematik langsung dengan tiga fokus penelitian; (b) diterbitkan oleh penerbit atau jurnal yang kredibel dan terindeks; serta (c) dapat diverifikasi melalui tautan DOI atau akses daring resmi.

Data yang terkumpul dianalisis menggunakan teknik analisis tematik, yaitu proses identifikasi dan pengelompokan pola gagasan yang berulang pada berbagai sumber, yang kemudian disintesis ke dalam tiga tema besar sesuai fokus penelitian. Untuk fokus kedua, yaitu validitas dan reliabilitas instrumen evaluasi, penelitian ini secara khusus menyintesis data kuantitatif yang telah dipublikasikan dari studi pengujian empiris terhadap empat belas alat deteksi teks hasil AI (Weber-Wulff dkk., 2023), sebagai bentuk data sekunder yang diolah ulang secara deskriptif, bukan data primer yang dikumpulkan sendiri oleh peneliti. Alur tahapan penelitian selengkapnya divisualisasikan pada diagram di atas. Karena sepenuhnya berbasis dokumen, penelitian ini tidak memerlukan prosedur persetujuan etik partisipan, namun tetap menjunjung etika akademik melalui parafrasa dan atribusi sumber yang konsisten pada setiap kutipan.

C. HASIL DAN PEMBAHASAN

Sintesis atas literatur yang dikaji menghasilkan tiga tema besar yang saling berkelindan, sebagaimana digambarkan pada kerangka konseptual di atas. Ketiganya merupakan tiga lapis respons keilmuan evaluasi pembelajaran terhadap satu krisis validitas yang sama.

1. Redesain Instrumen dan Desain

Asesmen Otentik di Era AI Generatif

Respons pertama yang paling banyak dibahas dalam literatur adalah upaya merancang ulang instrumen evaluasi itu sendiri. Khlaif, Al-Abed, Salama, dan Abu Eideh (2025) mendefinisikan apa yang mereka sebut sebagai tugas “tahan-AI” (*AI-resistant assessment*), yaitu tugas evaluasi yang secara sengaja meminimalkan ketergantungan pada AI generatif dengan menekankan keterampilan yang sulit ditiru mesin, seperti berpikir kritis, kreativitas, dan pengambilan keputusan etis. Temuan penelitian tersebut, yang melibatkan sejumlah pendidik sebagai partisipan, menunjukkan bahwa motivasi utama di balik redesain asesmen mencakup upaya menjaga integritas akademik, mempersiapkan mahasiswa untuk dunia kerja, dan menyesuaikan diri dengan kebijakan institusional, meskipun proses ini juga menghadapi tantangan berupa kebutuhan pengembangan profesional

dosen serta kekhawatiran mengenai kesetaraan akses.

Gonsalves (2025) memperluas argumen ini melalui kerangka teoretis yang disebut desain asesmen kontekstual, dengan menegaskan bahwa kehadiran AI generatif tidak sekadar menantang metode asesmen konvensional, tetapi telah mengungkap kerapuhan yang sesungguhnya sudah lama ada dalam konsep validitas dan keaslian karya pada asesmen tradisional. Argumen ini penting secara teoretis karena menggeser cara pandang: AI generatif bukanlah penyebab tunggal krisis validitas evaluasi, melainkan katalisator yang mempercepat terbukanya kelemahan yang sebelumnya tersembunyi. Pandangan ini selaras dengan prinsip penyelarasan konstruktif yang telah lama dikemukakan Biggs dan Tang (2011), yang menegaskan bahwa asesmen yang baik harus dirancang selaras dengan capaian pembelajaran yang dituju sejak awal, bukan sekadar menjadi alat pengecekan di akhir proses (Biggs & Tang, 2011, hlm. xxiii). Ketika prinsip ini diterapkan secara konsisten, tugas evaluasi secara alami akan mengarah pada penilaian proses dan penerapan pengetahuan pada konteks nyata, bentuk yang jauh lebih sulit diselesaikan penuh oleh AI generatif dibandingkan pertanyaan berbasis hafalan.

Beberapa jenis redesain asesmen otentik yang teridentifikasi secara konsisten dalam literatur meliputi tugas berbasis analisis studi kasus nyata, presentasi lisan dengan tanya jawab langsung, penilaian bertahap yang mendokumentasikan proses penyusunan draf, serta refleksi tertulis yang menghubungkan materi dengan pengalaman personal mahasiswa. Mishra, Warr, dan Islam (2023) turut menegaskan bahwa institusi pendidikan tinggi perlu mengembangkan kerangka kebijakan yang jelas mengenai kapan dan bagaimana AI generatif boleh digunakan dalam suatu tugas, sebab ambiguitas kebijakan justru menjadi sumber ketidakadilan tersendiri dalam proses evaluasi. Meskipun demikian, redesain instrumen semata tidak cukup untuk menjamin integritas evaluasi apabila tidak disertai mekanisme verifikasi yang andal, sebuah persoalan yang menjadi fokus tema berikutnya.

2. Validitas dan Reliabilitas Instrumen Evaluasi: Problematika Alat Deteksi AI

Ketika institusi pendidikan tinggi mengandalkan perangkat lunak pendeteksi teks hasil AI sebagai instrumen tambahan dalam proses evaluasi, pertanyaan mendasar keilmuan evaluasi pembelajaran pun muncul: seberapa valid dan reliabel sesungguhnya alat tersebut? Studi empiris paling komprehensif mengenai hal ini dilakukan oleh Weber-Wulff dan tujuh rekan peneliti (2023), yang menguji empat belas alat deteksi teks hasil AI secara sistematis menggunakan lima puluh empat dokumen uji yang mencakup teks tulisan manusia asli, teks hasil AI murni, teks hasil AI yang diedit manusia, dan teks hasil parafrasa. Studi multinasional ini merupakan salah satu pengujian paling ketat yang pernah dipublikasikan mengenai topik tersebut dan menjadi rujukan penting bagi keilmuan evaluasi

pembelajaran dalam menilai kelayakan teknologi deteksi sebagai instrumen asesmen.

Data yang dihasilkan studi tersebut memperlihatkan persoalan reliabilitas yang serius. Seluruh alat yang diuji mencatatkan tingkat akurasi di bawah 80%, dan hanya lima dari empat belas alat yang mencapai akurasi di atas 70%. Lebih mengkhawatirkan lagi, enam dari empat belas alat yang diuji menghasilkan positif keliru (*false positive*), yaitu mengklasifikasikan tulisan asli karya manusia sebagai hasil AI generatif, dengan salah satu alat bahkan mencatatkan bahwa separuh dari seluruh klasifikasi positifnya merupakan tuduhan keliru terhadap mahasiswa yang sesungguhnya menulis sendiri karyanya. Ringkasan temuan tersebut dan implikasinya bagi praktik evaluasi pembelajaran disajikan pada tabel berikut.

Tabel 1. Ringkasan temuan pengujian reliabilitas alat deteksi teks hasil AI

Aspek yang Diuji	Temuan Empiris	Implikasi bagi Evaluasi Pembelajaran
Akurasi keseluruhan (14 alat diuji)	Seluruh alat mencatatkan akurasi di bawah 80%; hanya 5 alat di atas 70%	Alat deteksi belum layak dijadikan bukti tunggal yang menentukan
Tingkat positif keliru (<i>false positive</i>)	6 dari 14 alat menghasilkan tuduhan keliru terhadap tulisan asli manusia	Berisiko merugikan mahasiswa jujur secara tidak adil
Kasus ekstrem (GPT Zero)	Sepuluh dari klasifikasi positif bersifat tuduhan keliru	Tidak layak digunakan sebagai dasar sanksi akademik
Konsistensi terhadap teks hasil parafrasa	Tingkat kegagalan mendeteksi meningkat signifikan pada teks yang telah diparafrasa	Mahasiswa dapat menghindari deteksi tanpa mengubah esensi tugas

Sumber: disintesis dari Weber-Wulff, Anohina-Naumeca, Bjelobaba, Foltýnek, Guerrero-Dib, Popoola, Sigut, dan Waddington (2023).

Temuan ini memiliki konsekuensi metodologis yang signifikan bagi keilmuan evaluasi pembelajaran. Dalam kerangka teori pengukuran pendidikan, sebuah instrumen yang digunakan untuk mengambil keputusan berisiko tinggi, seperti penjatuhan sanksi akademik atas tuduhan plagiarisme berbasis AI, semestinya memenuhi standar validitas dan reliabilitas yang ketat sebelum dapat dipercaya sebagai bukti yang berdiri sendiri. Ketika enam dari empat belas alat yang diuji justru berisiko menuduh mahasiswa jujur secara keliru, penggunaannya sebagai satu-satunya dasar keputusan evaluasi jelas melanggar

prinsip keadilan penilaian (*assessment fairness*) yang menjadi salah satu pilar utama keilmuan evaluasi pembelajaran. Weber-Wulff dkk. (2023) sendiri secara eksplisit merekomendasikan agar pendekatan preventif, yaitu redesain tugas sebagaimana dibahas pada subbagian 3.1, diprioritaskan ketimbang pendekatan deteksi semata, sebuah rekomendasi yang secara langsung menjembatani kedua fokus penelitian ini. Persoalan validitas alat deteksi ini pada gilirannya mendorong sebagian pendidik untuk mengalihkan perhatian dari fungsi evaluasi sebagai alat pengawasan menuju fungsi evaluasi sebagai alat pembelajaran itu sendiri,

sebagaimana dibahas pada tema berikutnya.

3. Pergeseran ke Evaluasi Formatif

Tema ketiga menunjukkan sisi yang secara paradigmatis berbeda dari dua tema sebelumnya. Alih-alih memandang AI generatif semata sebagai ancaman yang harus dideteksi dan dibendung, sejumlah pendidik mulai mengintegrasikannya sebagai instrumen aktif dalam proses evaluasi formatif, yaitu evaluasi yang bertujuan memandu proses belajar secara berkelanjutan, bukan sekadar menghakimi hasil akhir. Lee dan Moore (2024), melalui kajian sistematis terhadap literatur mengenai umpan balik otomatis berbasis GenAI di pendidikan tinggi, menemukan bahwa teknologi ini menunjukkan potensi nyata dalam menyediakan umpan balik yang cepat dan terskala, sebuah kebutuhan yang selama ini sulit dipenuhi secara manual oleh dosen pada kelas berjumlah besar.

Potensi ini sangat relevan dengan prinsip klasik evaluasi formatif yang menekankan bahwa umpan balik akan paling efektif ketika diberikan segera setelah performa mahasiswa diamati, bukan berminggu-minggu setelahnya. Dalam konteks ini, AI generatif berfungsi mengatasi salah satu keterbatasan struktural evaluasi formatif konvensional, yaitu keterbatasan waktu dan tenaga pendidik dalam memberikan umpan balik individual secara konsisten kepada seluruh mahasiswa. Lee dan Moore (2024) turut mencatat bahwa efektivitas umpan balik otomatis ini sebanding dengan umpan balik manual yang diberikan dosen dalam meningkatkan capaian belajar mahasiswa pada sejumlah studi yang mereka tinjau, meskipun kualitasnya tetap bervariasi tergantung pada desain prompt dan tingkat pengawasan pendidik terhadap keluarannya.

Meskipun demikian, pergeseran ini tidak lepas dari kehati-hatian metodologis. Lee dan Moore (2024) turut mengingatkan adanya kekhawatiran etis terkait potensi bias dalam sistem umpan balik berbasis AI serta risiko terhadap otonomi belajar mahasiswa apabila panduan institusional mengenai penggunaannya tidak jelas. Kekhawatiran ini bergema dengan peringatan Selwyn (2019) bahwa otomatisasi peran pendidik, termasuk dalam fungsi memberikan penilaian dan umpan balik, tidak pernah bersifat netral

secara nilai dan senantiasa membawa serta pertanyaan mengenai siapa yang sesungguhnya berwenang menentukan kualitas belajar seorang mahasiswa. Dengan kata lain, pemanfaatan AI generatif untuk evaluasi formatif idealnya diposisikan sebagai pelengkap yang memperluas kapasitas dosen memberikan umpan balik, bukan sebagai pengganti penuh peran profesional dosen dalam memaknai dan mengontekstualisasikan capaian belajar mahasiswa.

Ketiga tema tersebut, redesain instrumen, validitas alat deteksi, dan pergeseran ke evaluasi formatif, pada akhirnya membentuk satu rangkaian logis yang saling menguatkan: redesain asesmen mengurangi ketergantungan pada deteksi represif, kesadaran akan keterbatasan validitas alat deteksi mendorong urgensi redesain tersebut, dan pergeseran ke evaluasi formatif menyediakan jalan keluar paradigmatis yang mengubah AI generatif dari ancaman menjadi mitra dalam proses evaluasi yang lebih berorientasi pada pertumbuhan belajar mahasiswa.

D. KESIMPULAN

Kajian pustaka ini menegaskan bahwa kehadiran kecerdasan buatan generatif telah menimbulkan krisis validitas yang mendasar dalam keilmuan evaluasi pembelajaran di pendidikan tinggi, sekaligus mendorong tiga jalur respons yang saling melengkapi. Pada aspek redesain instrumen, kajian ini menyimpulkan bahwa asesmen otentik yang berorientasi proses, konteks nyata, dan berpikir tingkat tinggi merupakan strategi paling mendasar dan berkelanjutan dalam menjaga validitas evaluasi, karena secara struktural mengurangi kemungkinan tugas diselesaikan penuh oleh mesin, sejalan dengan prinsip penyelarasan konstruktif yang telah lama menjadi fondasi keilmuan asesmen. Pada aspek validitas dan reliabilitas instrumen evaluasi, kajian ini menemukan bukti empiris yang kuat bahwa alat deteksi teks hasil AI belum memenuhi standar keandalan yang memadai untuk dijadikan bukti tunggal dalam pengambilan keputusan evaluasi berisiko tinggi, mengingat tingkat akurasi yang rendah dan risiko tuduhan keliru yang nyata terhadap mahasiswa yang jujur. Pada aspek pergeseran paradigma, kajian ini menyimpulkan bahwa AI generatif menyimpan potensi transformatif untuk

memperkuat evaluasi formatif melalui umpan balik yang cepat dan terskala, asalkan tetap ditempatkan sebagai instrumen pendukung yang berada di bawah pengawasan dan pemaknaan profesional dosen, bukan pengganti peran evaluatif dosen itu sendiri.

Berdasarkan simpulan tersebut, penelitian ini merekomendasikan agar program studi dan institusi pendidikan tinggi, khususnya di Indonesia, segera menyusun pedoman evaluasi pembelajaran yang secara eksplisit mengintegrasikan ketiga aspek tersebut: pelatihan perancangan asesmen otentik bagi dosen, kebijakan yang membatasi penggunaan alat deteksi AI semata sebagai indikator awal dan bukan bukti final, serta kerangka pemanfaatan AI generatif untuk umpan balik formatif yang tetap menempatkan penilaian akhir di tangan pendidik. Bagi peneliti selanjutnya, disarankan untuk melengkapi kajian pustaka ini dengan penelitian empiris lapangan, baik berupa pengembangan dan uji validitas instrumen asesmen otentik baru maupun studi eksperimental mengenai efektivitas umpan balik berbasis AI generatif terhadap capaian belajar mahasiswa pada konteks program studi tertentu di Indonesia, sehingga temuan konseptual dalam kajian ini dapat diuji dan diperkaya secara empiris.

E. REFERENSI

- Ataupah, N., Lulu, M. C. U., Dianita, H., Folrati, M. P., Leo, C., & Wijaya, R. P. C. (2025). Integrasi AI generatif dalam kehidupan akademik mahasiswa di Kota Kupang. *Jurnal Inovasi Penelitian Ilmu Pendidikan Indonesia*, 2(6), 420-435. <https://jipipi.org/index.php/jipipi/article/view/114>
- Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university (4th ed.)*. Open University Press/McGraw-Hill. hlm. xxiii.
- Direktorat Pembelajaran dan Kemahasiswaan, Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi. (2024). *Buku panduan penggunaan generative artificial intelligence (GenAI) pada pembelajaran di perguruan tinggi*. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi.
- Gonsalves, C. (2025). Contextual assessment design in the age of generative AI. *Journal of Learning Development in Higher Education*, (34). <https://doi.org/10.47408/jldhe.vi34.1307>
- Khlaif, Z. N., Al-Abed, W. A., Salama, N., & Abu Eideh, B. (2025). Redesigning assessments for AI-enhanced learning: A framework for educators in the generative AI era. *Education Sciences*, 15(2), 174. <https://doi.org/10.3390/educsci15020174>
- Lee, S. S., & Moore, R. L. (2024). Harnessing generative AI (GenAI) for automated feedback in higher education: A systematic review. *Online Learning*, 28(3), 82-104. <https://olj.onlinelearningconsortium.org/index.php/olj/article/view/4593>
- Mishra, P., Warr, M., & Islam, R. (2023). Challenges and opportunities of generative AI for higher education as explained by ChatGPT. *Education Sciences*, 13(9), 856. <https://doi.org/10.3390/educsci13090856>
- Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education*. Polity Press.
- UNESCO, Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000386693>
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19, Article 26. <https://doi.org/10.1007/s40979-023-00146-z>
- Yusuf, A., Pervin, N., & Román-González, M. (2024). Generative AI and the future of higher education: A threat to academic integrity or reformation? Evidence from multicultural perspectives. *International Journal of Educational Technology in Higher Education*, 21, Article 21. <https://doi.org/10.1186/s41239-024-00453-6>