Journal of Innovative and Creativity
Journal Homepage: https://joecy.org/index.php/joecy
e-mail: journaljoecy@gmail.com

Faculty of Education
University of Pahlawan Tuanku Tambusai
ISSN 2775-771X (Online)
ISSN 2962-570X **(Printed)**
6 (1) 2026, Page: 12507-12529

Implementasi Artificial Intelligence (AI) dalam Penilaian Formatif: Studi Literatur tentang Peningkatan Akurasi Evaluasi Pembelajaran

Dewi Agustina Solihin[1],Heny Fitriani[2], R. Dewi Mutia Farida[3], Ahmad Sofan Ansor[4], Rosdiana[5]
Politeknik PGRI Banten[1-5]
Corresponding Author : e-mail: dewisolihin87@gmail.com

Published: Februari , 2026

ABSTRAK

Penelitian ini bertujuan mensintesis bukti ilmiah mengenai implementasi Artificial Intelligence (AI) dalam penilaian formatif dan kontribusinya terhadap peningkatan akurasi evaluasi pembelajaran. Metode yang digunakan adalah studi literatur dengan desain Systematic Literature Review (SLR) pada rentang publikasi 2015–2025. Proses seleksi mengikuti tahapan identifikasi, penyaringan, penilaian kelayakan, dan inklusi, dengan fokus pada studi yang membahas AI untuk penilaian formatif serta menyertakan bukti kualitas evaluasi seperti validitas, reliabilitas/konsistensi, agreement AI–penilai manusia, dan/atau kegunaan umpan balik. Sintesis kualitatif dilakukan terhadap 10 studi kunci yang merepresentasikan spektrum implementasi, termasuk automated writing evaluation/automated essay scoring, intelligent tutoring system, pendekatan deep learning, serta large language models untuk umpan balik dan penilaian. Hasil menunjukkan bahwa AI dapat meningkatkan akurasi evaluasi melalui: (1) penguatan validitas ketika output AI selaras dengan konstruk dan rubrik, (2) peningkatan konsistensi penilaian serta efisiensi pemberian umpan balik, dan (3) dukungan diagnosis kesalahan yang membantu tindakan perbaikan belajar. Namun, efektivitas AI sangat dipengaruhi moderator implementasi, terutama kejelasan rubrik, desain siklus revisi, kualitas dan representativitas data, literasi umpan balik siswa, serta peran guru dalam human-in-the-loop. Studi juga menyoroti risiko yang dapat menurunkan akurasi, seperti bias/fairness, rendahnya transparansi, overreliance, dan isu tata kelola data. Kesimpulannya, AI paling efektif meningkatkan akurasi evaluasi pembelajaran bila diterapkan sebagai pendukung asesmen formatif yang terintegrasi pedagogis dan dikendalikan melalui verifikasi guru serta standar evaluasi yang jelas.

Kata kunci : *Artificial intelligence, Penilaian formatif, Automated feedback, Akurasi evaluasi, Systematic literature review (SLR)*

© The Author(s). 2021 Open Access This article is licensed under a Creative Commons Atibution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0

PENDAHULUAN

Penilaian formatif menjadi kunci dalam memperbaiki proses belajar karena menyediakan umpan balik cepat untuk mengarahkan strategi belajar siswa dan keputusan pengajaran guru. Namun, praktik di kelas sering menghadapi kendala beban kerja, keterbatasan waktu, dan variasi konsistensi penilaian, sehingga akurasi evaluasi dapat menurun ketika jumlah tugas meningkat atau ketika rubrik ditafsirkan berbeda oleh penilai. Dalam konteks inilah Artificial Intelligence (AI) mulai dipandang sebagai solusi untuk memperkuat kualitas penilaian formatif melalui otomatisasi umpan balik dan dukungan analitik. Kajian Deeva et al. (2021) menegaskan bahwa sistem umpan balik otomatis berkembang dengan beragam pendekatan dan fungsi, sekaligus menyajikan kerangka klasifikasi yang membantu memetakan bagaimana fitur AI dapat dirancang agar relevan dengan kebutuhan pembelajar. Meski demikian, mereka juga menyoroti tantangan penting seperti kualitas data, ketepatan rekomendasi, dan kesesuaian umpan balik dengan tujuan pembelajaran, yang semuanya berimplikasi langsung pada akurasi evaluasi. Oleh karena itu, studi literatur ini penting untuk

merangkum bukti empiris tentang bagaimana implementasi AI dalam penilaian formatif dapat meningkatkan akurasi evaluasi pembelajaran secara sistematis.Salah satu ranah yang paling banyak mengadopsi AI untuk penilaian formatif adalah penilaian tulisan, terutama melalui automated writing evaluation (AWE) yang memberikan skor dan umpan balik secara cepat. Namun, kecepatan tidak cukup bila umpan balik tidak akurat, tidak dapat dipertanggungjawabkan, atau tidak bermanfaat untuk revisi. Ranalli et al. (2017) menekankan pentingnya menilai akurasi dan kegunaan umpan balik AWE dalam kerangka validasi berbasis argumen, sehingga fokusnya bukan hanya "apakah sistem memberi nilai", tetapi "apakah nilai dan umpan balik itu valid untuk tujuan formatif". Perspektif ini relevan karena penilaian formatif menuntut umpan balik yang diagnostik, spesifik, dan dapat ditindaklanjuti oleh siswa. Ketika AI mampu mengidentifikasi kesalahan secara konsisten dan memberikan saran perbaikan yang selaras dengan rubrik, maka akurasi evaluasi meningkat tidak hanya pada level skor, tetapi juga pada kualitas informasi yang membantu pembelajaran. Dengan demikian, literatur tentang AWE menyediakan pijakan penting untuk memahami syarat-syarat keberhasilan AI dalam meningkatkan akurasi penilaian formatif.

Selain penilaian tulisan, penggunaan AI juga berkembang pada penilaian keterampilan berpikir tingkat tinggi, misalnya argumentasi ilmiah, yang secara tradisional sulit dinilai secara cepat dan konsisten. Mao et al. (2018) menunjukkan bahwa validasi automated scoring untuk asesmen formatif berbasis argumentasi ilmiah perlu dibangun dengan bukti yang kuat, termasuk keselarasan konstruk yang diukur dan keandalan skor yang dihasilkan sistem. Ini penting karena "akurasi" dalam penilaian formatif tidak semata-mata berarti kedekatan skor AI dengan skor manusia, tetapi juga mencakup ketepatan AI dalam merepresentasikan kompetensi yang dimaksud oleh tujuan pembelajaran. Jika AI menilai aspek yang keliru (misalnya lebih sensitif pada panjang jawaban daripada kualitas argumen), maka evaluasi dapat bias meski terlihat konsisten. Studi ini menguatkan bahwa implementasi AI harus disertai kerangka validitas yang jelas agar peningkatan akurasi evaluasi benar-benar bermakna bagi pembelajaran. Karena itu, studi literatur perlu menelaah indikator validitas, reliabilitas, dan potensi bias yang paling sering digunakan dalam penelitian-penelitian AI untuk penilaian formatif.AI dalam penilaian formatif juga sering diintegrasikan dalam intelligent tutoring system (ITS) untuk memberikan bantuan dan umpan balik adaptif selama proses belajar berlangsung. Perikos et al. (2017) menunjukkan bahwa mekanisme bantuan dan umpan balik dalam ITS dapat dirancang untuk mengarahkan siswa melalui langkah-langkah berpikir, bukan sekadar memberi tahu benar-salah, sehingga mendukung pembelajaran yang lebih mendalam. Pendekatan ini menegaskan bahwa akurasi evaluasi pembelajaran akan meningkat ketika sistem tidak hanya "menghasilkan penilaian", tetapi juga "menjelaskan alasan" dan "mengarahkan perbaikan" secara terstruktur sesuai kesulitan siswa. Dalam konteks formatif, umpan balik yang adaptif membantu meminimalkan miskonsepsi, mempercepat koreksi kesalahan, dan menjaga keselarasan antara penilaian dengan proses konstruksi pengetahuan. Namun, efektivitasnya tetap bergantung pada kualitas model pengetahuan, ketepatan deteksi kesalahan, dan relevansi saran yang diberikan, yang semuanya merupakan komponen akurasi evaluasi. Karena itu, meninjau literatur ITS menjadi penting untuk memahami bagaimana desain umpan balik adaptif berbasis AI dapat memperbaiki ketepatan penilaian formatif di berbagai domain.

Pada akhirnya, keberhasilan AI dalam penilaian formatif tidak hanya ditentukan oleh akurasi algoritmik, tetapi juga oleh bagaimana siswa berinteraksi dengan umpan balik yang diberikan. Tian dan Zhou (2020) menegaskan bahwa keterlibatan (engagement) siswa terhadap umpan balik otomatis, umpan balik teman sebaya, dan umpan balik guru dapat berbeda dalam konteks pembelajaran daring, dan perbedaan ini berpengaruh pada sejauh mana umpan balik benar-benar digunakan untuk memperbaiki kinerja. Artinya, peningkatan akurasi evaluasi pembelajaran melalui AI perlu dipahami sebagai proses sosial-pedagogis: AI dapat menyajikan umpan balik cepat dan konsisten, tetapi dampaknya pada pembelajaran bergantung pada penerimaan, kepercayaan, serta strategi revisi siswa. Deeva et al. (2021) juga menekankan adanya peluang besar sekaligus tantangan pada sistem umpan balik otomatis, termasuk kebutuhan personalisasi dan kesesuaian dengan tujuan pembelajaran agar umpan balik tidak sekadar "informasi", tetapi menjadi "intervensi" yang efektif. Karena itu, studi literatur ini akan merangkum temuan tentang relasi antara desain sistem AI,

validitas penilaian, dan engagement pengguna sebagai fondasi peningkatan akurasi evaluasi pembelajaran dalam penilaian formatif.Bukti empiris terbaru menunjukkan bahwa *automated writing evaluation* (AWE) tidak hanya mempercepat pemberian umpan balik, tetapi juga dapat menghasilkan dampak pembelajaran yang terukur ketika diposisikan sebagai bagian dari penilaian formatif. Melalui *multi-level meta-analysis*, Fleckenstein et al. (2023) menegaskan bahwa efektivitas AWE terhadap performa menulis dan proses revisi cenderung positif, meskipun besarnya efek bervariasi antar konteks, desain tugas, dan karakteristik umpan balik. Variasi ini memberi sinyal bahwa "akurasi evaluasi" dalam sistem berbasis AI perlu dipahami secara lebih luas, bukan sekadar kesesuaian skor, tetapi juga ketepatan umpan balik dalam mendiagnosis kekuatan dan kelemahan karya siswa. Ketika AWE mampu memberikan umpan balik yang konsisten dan selaras dengan kriteria penilaian, guru memperoleh dukungan untuk menjaga konsistensi keputusan formatif pada kelas besar. Namun, temuan meta-analitik juga mengindikasikan adanya moderator penting, sehingga implementasi AI perlu memperhatikan kesesuaian fitur sistem dengan tujuan pembelajaran dan strategi umpan balik yang digunakan. Dengan demikian, studi literatur tentang AI dalam penilaian formatif menjadi relevan untuk merangkum kondisi-kondisi yang membuat umpan balik otomatis benar-benar meningkatkan ketepatan evaluasi pembelajaran.

Dalam konteks EFL/ESL, penggunaan AWE semakin menonjol karena kebutuhan umpan balik cepat pada aspek kebahasaan, koherensi, dan kualitas argumentasi, terutama pada pembelajaran daring dan tugas menulis berulang. Zhai dan Ma (2023) melalui meta-analisis menekankan bahwa AWE dan umpan balik otomatis berpotensi meningkatkan kualitas tulisan, tetapi efektivitasnya sangat dipengaruhi oleh cara siswa memanfaatkan umpan balik dan bagaimana guru mengintegrasikannya ke dalam siklus pembelajaran. Artinya, peningkatan akurasi evaluasi pembelajaran tidak hanya ditentukan oleh ketepatan deteksi kesalahan oleh sistem, melainkan juga oleh kualitas "tindak lanjut" yang terjadi setelah umpan balik diberikan, seperti revisi terarah dan refleksi siswa. Meta-analisis tersebut menguatkan pentingnya desain penilaian formatif yang mendorong siswa melakukan revisi substantif, bukan hanya perbaikan permukaan. Jika AI digunakan hanya sebagai alat pemberi skor, maka peluang peningkatan akurasi evaluasi akan terbatas karena umpan balik formatif seharusnya bersifat diagnostik dan memandu strategi belajar. Karena itu, literatur menuntut fokus pada integrasi pedagogis: AI sebagai pendukung keputusan guru dan pendorong keterlibatan siswa untuk memperbaiki kualitas belajar secara berkelanjutan.

Selain bukti kuantitatif dari meta-analisis, pemetaan konseptual dan metodologis juga penting untuk memahami arah riset dan celah implementasi AI dalam umpan balik tertulis. Shi dan Aryadoust (2024) melalui *systematic review* menegaskan bahwa riset *AI-based automated written feedback* berkembang pesat dengan beragam pendekatan, namun masih menyisakan tantangan pada aspek transparansi, keselarasan umpan balik dengan rubrik, dan pembuktian validitas untuk penggunaan formatif. Hal ini relevan karena "akurasi" dalam evaluasi pembelajaran menuntut bukti bahwa sistem menilai konstruk yang tepat, memberikan umpan balik yang dapat dipercaya, serta tidak menghasilkan bias sistematis pada kelompok tertentu. Studi review tersebut juga memperlihatkan bahwa kualitas penelitian sangat dipengaruhi oleh kejelasan desain evaluasi, termasuk definisi "feedback usefulness" dan indikator keberhasilan pembelajaran setelah umpan balik diterapkan. Dengan demikian, peningkatan akurasi evaluasi melalui AI memerlukan kerangka evaluasi yang menilai tidak hanya output sistem, tetapi juga dampaknya terhadap proses belajar. Karena itu, studi literatur ini perlu menyoroti bagaimana peneliti menguji validitas, reliabilitas, dan kegunaan umpan balik sebagai fondasi untuk implementasi yang lebih bertanggung jawab.

Kemajuan AI yang mendorong peningkatan performa sistem penilaian dan umpan balik juga banyak ditopang oleh pendekatan *deep learning* yang mampu mengekstraksi pola kompleks dari teks siswa. Misgna et al. (2024) menekankan bahwa *deep learning-based automated essay scoring* dan generasi umpan balik menawarkan kemajuan pada pemodelan fitur bahasa dan struktur argumen, tetapi keberhasilannya sangat bergantung pada kualitas data latih, representativitas korpus, dan metrik evaluasi yang digunakan. Ini penting karena ketepatan skor dan umpan balik dapat terdistorsi bila data pelatihan tidak sejalan dengan konteks kurikulum, tingkat kemampuan siswa, atau ragam bahasa yang digunakan. Di sisi lain, sistem yang terlalu "hitam kotak" berisiko menurunkan kepercayaan

guru dan siswa, sehingga akurasi evaluasi secara praktis dapat dipertanyakan meski performa model terlihat tinggi. Tantangan interpretabilitas dan bias juga menguatkan kebutuhan strategi *human-in-the-loop* serta prosedur kontrol mutu untuk memastikan umpan balik tidak menyesatkan. Dengan kata lain, peningkatan akurasi evaluasi pembelajaran berbasis AI harus dibangun melalui keseimbangan antara kecanggihan model, keterjelasan keputusan penilaian, dan kesesuaian pedagogis. Perspektif ini menjadi landasan penting untuk menilai kesiapan implementasi AI dalam penilaian formatif di berbagai setting pendidikan.

Perkembangan *large language models* (LLM) memperluas peluang otomatisasi penilaian dan umpan balik, tetapi sekaligus menambah tantangan baru pada validitas dan reliabilitas keputusan evaluasi. Pack et al. (2024) menekankan bahwa penggunaan LLM untuk *automated essay scoring* perlu diuji secara ketat terkait konsistensi skor, keselarasan dengan rubrik, serta kualitas umpan balik yang diberikan agar benar-benar aman untuk tujuan formatif. Dalam penilaian formatif, umpan balik yang tampak meyakinkan namun tidak akurat dapat mendorong miskonsepsi dan mengarahkan strategi belajar ke jalur yang keliru, sehingga "akurasi evaluasi" menjadi isu krusial. Karena itu, integrasi LLM seharusnya diarahkan sebagai alat pendukung guru, bukan pengganti otoritas penilaian, dengan mekanisme verifikasi dan batasan penggunaan yang jelas. Literatur juga menekankan pentingnya tata kelola penilaian berbasis AI, termasuk transparansi kriteria, audit performa, dan perlindungan terhadap bias serta kesalahan sistem. Dengan latar tersebut, studi literatur mengenai implementasi AI dalam penilaian formatif menjadi penting untuk merangkum bukti paling kuat tentang kapan AI meningkatkan ketepatan evaluasi, dan kapan ia justru menimbulkan risiko terhadap kualitas keputusan pembelajaran.

Secara keseluruhan, rangkaian temuan meta-analisis dan telaah sistematis memperlihatkan bahwa AI dalam penilaian formatif berpotensi meningkatkan akurasi evaluasi pembelajaran ketika ia menghasilkan umpan balik yang konsisten, relevan dengan rubrik, serta mendorong revisi yang bermakna. Namun, literatur juga menunjukkan bahwa akurasi tidak bisa direduksi menjadi kesesuaian skor semata, melainkan harus mencakup validitas konstruk, reliabilitas penilaian, dan kegunaan umpan balik bagi keputusan belajar siswa. Fleckenstein et al. (2023) dan Zhai dan Ma (2023) menegaskan adanya variasi efek yang dipengaruhi konteks dan desain implementasi, sementara Shi dan Aryadoust (2024) menekankan celah pada aspek transparansi, desain evaluasi, dan pembuktian validitas. Di saat yang sama, Misgna et al. (2024) mengingatkan tentang peran kualitas data dan potensi bias pada model *deep learning*, serta Pack et al. (2024) menyoroti kebutuhan pengujian ketat ketika LLM dipakai untuk penilaian dan umpan balik. Dengan demikian, peningkatan akurasi evaluasi pembelajaran melalui AI adalah hasil dari keterpaduan teknologi, pedagogi, dan tata kelola penilaian yang bertanggung jawab.

Berdasarkan urgensi tersebut, studi literatur ini diarahkan untuk merumuskan peta implementasi AI dalam penilaian formatif yang menautkan tiga aspek utama, yaitu bukti efektivitas pembelajaran, bukti kualitas pengukuran (validitas dan reliabilitas), serta bukti penerimaan dan penggunaan umpan balik oleh siswa dan guru. Sintesis literatur yang kuat diharapkan menghasilkan rekomendasi praktis bagi institusi pendidikan, seperti prinsip desain *human-in-the-loop*, standar verifikasi skor dan umpan balik, serta strategi integrasi AWE/LLM ke dalam siklus penilaian formatif agar tetap adil, transparan, dan selaras dengan capaian pembelajaran. Dengan cara itu, AI tidak dipahami sebagai sekadar alat otomatisasi, melainkan sebagai sistem pendukung keputusan yang meningkatkan ketepatan evaluasi dan sekaligus menjaga integritas proses pembelajaran. Output studi ini juga dapat menjadi dasar pengembangan agenda riset berikutnya, terutama pada pengujian bias, penguatan interpretabilitas, dan penyusunan protokol implementasi AI yang layak diterapkan pada berbagai konteks pembelajaran.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kualitatif studi literatur dengan desain Systematic Literature Review (SLR) untuk menyintesis bukti ilmiah tentang implementasi Artificial Intelligence (AI) dalam penilaian formatif dan kaitannya dengan peningkatan akurasi evaluasi pembelajaran. Fokus kajian diarahkan pada penggunaan AI untuk menghasilkan skor, umpan balik, atau rekomendasi belajar,

serta bagaimana penelitian-penelitian sebelumnya mengukur "akurasi" melalui indikator seperti validitas, reliabilitas/konsistensi, kesesuaian skor AI dengan penilai manusia, dan kegunaan umpan balik dalam mendukung perbaikan belajar. Untuk menjaga transparansi, alur seleksi studi didokumentasikan mengikuti kerangka PRISMA dan disajikan dalam alur identifikasi–penyaringan–kelayakan–inklusi. Sumber data diperoleh melalui penelusuran pada basis data bereputasi seperti Scopus, Web of Science, ERIC, dan Google Scholar sebagai pelengkap untuk menangkap publikasi relevan. Rentang tahun publikasi dibatasi pada 2015–2025 agar temuan mencerminkan perkembangan terkini, termasuk AWE, automated scoring, intelligent tutoring systems, learning analytics, dan model generatif/LLM. Strategi pencarian menggunakan kombinasi kata kunci dan operator Boolean yang disesuaikan per basis data.

Seleksi studi dilakukan bertahap melalui penyaringan judul dan abstrak, lalu pemeriksaan teks penuh untuk memastikan kesesuaian dengan tujuan penelitian. Artikel dinyatakan layak bila membahas AI dalam konteks penilaian formatif/umpan balik formatif dan menyertakan bukti kualitas evaluasi (misalnya validitas, reliabilitas, agreement, atau indikator akurasi lainnya). Alasan eksklusi pada tahap teks penuh dicatat secara eksplisit untuk memperkuat akuntabilitas proses SLR (misalnya fokus sumatif murni, tidak ada evaluasi akurasi/validitas, atau bukan konteks pendidikan).Ekstraksi data dilakukan menggunakan lembar koding terstruktur untuk mencatat informasi bibliografis, konteks pembelajaran, jenis AI yang digunakan, bentuk asesmen formatif, karakteristik umpan balik, serta indikator "akurasi evaluasi" yang dilaporkan. Format ekstraksi data dirangkum untuk memastikan konsistensi pencatatan antar studi. Selain itu, dilakukan penilaian kualitas studi (*quality appraisal*) menggunakan kriteria yang menilai kejelasan metode, transparansi evaluasi sistem, dan kekuatan bukti terkait akurasi/validitas/reliabilitas. Ringkasan instrumen penilaian kualitas disajikan dan digunakan untuk menimbang kekuatan interpretasi temuan.

Sintesis data dilakukan melalui analisis tematik untuk mengidentifikasi pola lintas studi dan merumuskan tema kunci terkait bagaimana AI meningkatkan akurasi evaluasi pembelajaran dalam penilaian formatif. Tahap analisis meliputi open coding pada temuan inti, pengelompokan kode menjadi kategori (misalnya validitas–reliabilitas, kualitas umpan balik, bias/fairness, interpretabilitas, peran guru dan *human-in-the-loop*), lalu sintesis menjadi tema dan subtema. Keabsahan sintesis dijaga melalui audit trail (catatan keputusan seleksi dan koding), pengecekan silang pada sebagian sampel artikel, serta refleksi kritis terhadap potensi bias interpretasi.

Tabel 1. Kriteria Inklusi dan Eksklusi

Aspek	Inklusi	Eksklusi	Justifikasi
Fokus penelitian	AI digunakan untuk penilaian formatif atau umpan balik formatif dalam evaluasi pembelajaran	Penilaian sumatif murni tanpa tujuan formatif atau tanpa siklus umpan balik	Selaras dengan tujuan penelitian yang menekankan evaluasi formatif
Konteks	Konteks pendidikan (sekolah, perguruan tinggi, pelatihan) dengan aktivitas belajar terstruktur	Konteks non-pendidikan atau murni industri tanpa komponen pembelajaran	Relevansi langsung pada evaluasi pembelajaran
Jenis AI	AWE/AES, NLP, deep learning, ITS, learning analytics, LLM untuk skor dan atau umpan balik	Sistem digital biasa tanpa komponen AI atau hanya administrasi ujian	Menjaga fokus pada AI dan pengaruhnya pada akurasi evaluasi

Bukti akurasi	Memuat indikator validitas, reliabilitas, agreement AI-human, error analysis, atau evaluasi kegunaan umpan balik	Hanya deskripsi sistem tanpa evaluasi kualitas output	Akurasi harus dibuktikan secara empiris atau terukur
Jenis dokumen	Artikel jurnal dan prosiding bereputasi, peer reviewed	Editorial, opini, berita, blog, preprint tanpa peer review (jika jurnal target mensyaratkan)	Menjaga kualitas bukti ilmiah
Rentang tahun	2015–2025	Di luar rentang 2015–2025	Menangkap perkembangan AI modern dan relevan
Akses full text	Full text tersedia untuk ekstraksi data dan appraisal	Full text tidak tersedia	Menghindari kesimpulan berbasis abstrak saja
Bahasa	Bahasa Inggris dan Bahasa Indonesia	Bahasa lain di luar dua bahasa tersebut	Keterbacaan dan konsistensi analisis

HASIL PENELITIAN

1. Seleksi Studi dan Karakteristik Artikel

Penelitian studi literatur ini menghasilkan kumpulan bukti inti yang digunakan untuk sintesis tematik tentang implementasi AI dalam penilaian formatif dan kaitannya dengan peningkatan akurasi evaluasi pembelajaran. Fokus hasil pada poin ini adalah menjelaskan bagaimana studi dipilih dan seperti apa profil studi yang dianalisis, sehingga pembaca memahami cakupan bukti yang menjadi dasar kesimpulan. Korpus inti yang dibahas dalam artikel ini mencakup 10 studi kunci yang secara langsung menguji atau merangkum peran AI pada penilaian formatif, termasuk kajian kerangka dan tantangan sistem umpan balik otomatis (Deeva et al., 2021), validasi akurasi serta kegunaan umpan balik AWE (Ranalli et al., 2017), validasi automated scoring pada tugas argumentasi ilmiah (Mao et al., 2018), mekanisme bantuan dan umpan balik pada ITS (Perikos et al., 2017), serta keterlibatan siswa terhadap umpan balik otomatis dalam konteks pembelajaran daring (Tian & Zhou, 2020). Selain itu, kumpulan bukti diperkuat oleh studi sintesis berskala besar seperti meta-analisis AWE (Fleckenstein et al., 2023; Zhai & Ma, 2023), systematic review umpan balik tertulis berbasis AI (Shi & Aryadoust, 2024), survei AES berbasis deep learning (Misgna et al., 2024), dan evaluasi validitas–reliabilitas LLM untuk penilaian esai (Pack et al., 2024). Dengan komposisi tersebut, hasil pada poin ini memberi gambaran bahwa bukti yang digunakan tidak bertumpu pada satu tipe studi saja, tetapi memadukan studi empiris dan studi sintesis untuk memperkuat ketajaman temuan.

Proses seleksi studi dilakukan melalui tahapan yang konsisten dengan logika SLR: identifikasi, penyaringan, penilaian kelayakan, dan inklusi. Dalam artikel ini, penyaringan diarahkan untuk memastikan setiap studi yang masuk benar-benar terkait dengan penilaian formatif, bukan sekadar penilaian otomatis untuk tujuan sumatif. Kriteria kelayakan menekankan keberadaan bukti yang dapat ditautkan pada "akurasi evaluasi", misalnya validitas/argument-based validation (Ranalli et al., 2017; Mao et al., 2018), indikator reliabilitas dan konsistensi skor (Pack et al., 2024), atau bukti dampak umpan balik terhadap proses revisi dan keterlibatan belajar (Tian & Zhou, 2020; Fleckenstein et al., 2023). Studi yang hanya mendeskripsikan teknologi tanpa evaluasi kualitas output tidak diprioritaskan karena tidak cukup untuk menjawab klaim "peningkatan akurasi". Selain itu, studi yang konteksnya tidak jelas pada domain pembelajaran atau tidak menjelaskan bentuk umpan

balik dan basis rubrik dinilai kurang relevan untuk sintesis. Dengan cara ini, korpus inti dibentuk untuk menjaga keseimbangan antara keluasan topik dan ketepatan fokus pada evaluasi pembelajaran. Secara temporal, karakteristik publikasi menunjukkan sebaran yang merepresentasikan evolusi riset AI dalam assessment. Korpus inti mencakup studi sejak 2017 hingga 2024, yang menandakan transisi dari AWE/ITS generasi awal menuju gelombang baru model berbasis deep learning dan LLM. Pada fase awal, fokus penelitian lebih banyak pada mekanisme bantuan dan umpan balik pada sistem cerdas serta evaluasi akurasi umpan balik AWE (Perikos et al., 2017; Ranalli et al., 2017). Memasuki periode menengah, perhatian bergeser pada validasi automated scoring untuk tugas berpikir tingkat tinggi seperti argumentasi ilmiah (Mao et al., 2018) dan dinamika keterlibatan siswa terhadap umpan balik otomatis dalam pembelajaran daring (Tian & Zhou, 2020). Setelah itu, periode 2021–2024 didominasi oleh studi sintesis dan evaluasi model yang lebih mutakhir, termasuk kerangka klasifikasi sistem umpan balik otomatis (Deeva et al., 2021), meta-analisis efektivitas AWE (Fleckenstein et al., 2023; Zhai & Ma, 2023), systematic review riset umpan balik tertulis berbasis AI (Shi & Aryadoust, 2024), hingga kajian reliabilitas–validitas penilaian esai dengan LLM (Pack et al., 2024). Pola ini menunjukkan bahwa bidangnya makin menekankan bukti komparatif berskala besar sekaligus isu validitas dan tata kelola.

Ditinjau dari jenis desain penelitian, studi dalam korpus inti terbagi ke dua kelompok yang saling melengkapi. Kelompok pertama adalah studi empiris/validasi, yang menguji akurasi dan kegunaan output AI pada konteks tugas tertentu, misalnya validasi umpan balik AWE (Ranalli et al., 2017), validasi automated scoring pada argumentasi ilmiah (Mao et al., 2018), desain mekanisme bantuan dan umpan balik pada ITS (Perikos et al., 2017), serta pembandingan engagement siswa pada umpan balik otomatis dengan umpan balik lain (Tian & Zhou, 2020). Kelompok kedua adalah studi sintesis (review, systematic review, survey, meta-analisis) yang merangkum pola temuan lintas konteks sekaligus menyoroti moderator, tantangan, dan peluang implementasi (Deeva et al., 2021; Fleckenstein et al., 2023; Zhai & Ma, 2023; Shi & Aryadoust, 2024; Misgna et al., 2024). Kombinasi ini penting karena studi empiris memberi bukti rinci pada level tugas dan rubrik, sementara studi sintesis memberi gambaran konsisten-tidaknya hasil pada skala yang lebih luas. Dengan kata lain, karakteristik desain penelitian yang beragam memperkuat fondasi interpretasi: "akurasi" bukan klaim tunggal, tetapi dibaca lewat beragam indikator evaluasi yang digunakan penelitian-penelitian berbeda.

Dari sisi domain penilaian formatif, temuan menunjukkan konsentrasi yang kuat pada penilaian menulis dan umpan balik tertulis, yang wajar karena menulis menghasilkan artefak yang dapat dinilai otomatis dan diikuti siklus revisi. Meta-analisis menunjukkan AWE cenderung berdampak positif namun bervariasi tergantung konteks, fitur umpan balik, dan integrasi pembelajaran (Fleckenstein et al., 2023; Zhai & Ma, 2023). Kajian systematic review juga menunjukkan riset umpan balik tertulis berbasis AI berkembang pesat, tetapi kualitas desain evaluasi dan aspek transparansi sering belum seragam (Shi & Aryadoust, 2024). Di luar domain menulis, studi empiris memperlihatkan bahwa AI juga digunakan untuk menilai kompetensi argumentasi ilmiah (Mao et al., 2018) dan untuk membantu pembelajaran logika melalui ITS dengan mekanisme bantuan dan umpan balik bertahap (Perikos et al., 2017). Artinya, meskipun domain menulis dominan, bukti tetap menunjukkan potensi AI untuk penilaian formatif pada keterampilan kompleks lain asalkan ada rubrik yang jelas dan prosedur validasi yang ketat.

Karakteristik teknologi AI yang dilibatkan dalam studi-studi inti juga menunjukkan spektrum yang luas, dari sistem yang fokus pada klasifikasi umpan balik hingga model generatif. Pada level konsep, Deeva et al. (2021) memberikan kerangka klasifikasi sistem umpan balik otomatis dan menyoroti isu desain seperti personalisasi dan kesesuaian dengan tujuan pembelajaran. Pada level implementasi, AWE/AES menjadi tulang punggung penilaian formatif menulis, dengan pengukuran akurasi yang sering dikaitkan pada validitas dan kegunaan umpan balik (Ranalli et al., 2017; Fleckenstein et al., 2023; Zhai & Ma, 2023). Kemudian, pendekatan deep learning memperluas kemampuan AES dan generasi umpan balik, tetapi menambah isu kualitas data latih, bias, dan interpretabilitas (Misgna et al., 2024). Gelombang terbaru tampak pada pemanfaatan LLM untuk penilaian esai dan pemberian umpan balik, yang menuntut pembuktian validitas serta reliabilitas secara ketat agar aman untuk tujuan formatif (Pack et al., 2024). Spektrum ini menegaskan bahwa "implementasi AI" bukan satu

teknologi tunggal, sehingga hasil perlu dibaca sebagai peta pendekatan dan konsekuensinya pada evaluasi pembelajaran.

Konteks pembelajaran yang muncul dalam korpus inti juga memberi sinyal tentang kondisi implementasi yang umum diteliti. Sejumlah studi menempatkan AI dalam konteks pembelajaran bahasa dan menulis (Ranalli et al., 2017; Tian & Zhou, 2020; Zhai & Ma, 2023), yang biasanya melibatkan siklus tugas–umpan balik–revisi sehingga cocok untuk penilaian formatif. Ada pula konteks penilaian untuk argumentasi ilmiah yang menuntut pengukuran konstruk kompleks, sehingga evaluasi validitas menjadi lebih kritikal (Mao et al., 2018). Sementara itu, ITS menekankan bantuan bertahap yang mengarahkan proses berpikir, bukan hanya memberi skor, sehingga akurasi evaluasi dipahami sebagai ketepatan diagnosis dan ketepatan jalur bantuan (Perikos et al., 2017). Studi sintesis menyoroti bahwa efektivitas AI sering dipengaruhi faktor moderator seperti desain tugas, kualitas umpan balik, tingkat dukungan guru, dan karakteristik pembelajar (Fleckenstein et al., 2023; Shi & Aryadoust, 2024). Karena itu, karakteristik konteks ini penting untuk membatasi generalisasi: peningkatan akurasi sangat mungkin terjadi, tetapi tidak otomatis merata di semua mata pelajaran dan jenjang tanpa penyesuaian desain.

Secara ringkas, hasil poin ini menunjukkan bahwa korpus inti yang dianalisis terdiri dari studi-studi yang paling relevan untuk menjawab pertanyaan tentang bagaimana AI diimplementasikan dan bagaimana akurasi evaluasi dinilai dalam penilaian formatif. Komposisi studi empiris dan studi sintesis memastikan bahwa pembahasan tidak hanya bertumpu pada bukti lokal, tetapi juga ditopang ringkasan pola lintas penelitian (Deeva et al., 2021; Fleckenstein et al., 2023; Shi & Aryadoust, 2024). Dominasi domain menulis mengindikasikan area yang paling matang secara riset, sedangkan studi pada argumentasi ilmiah dan ITS menandakan peluang perluasan ke keterampilan kompleks lain (Mao et al., 2018; Perikos et al., 2017). Kehadiran deep learning dan LLM menambah urgensi isu validitas, reliabilitas, bias, dan tata kelola penggunaan AI pada asesmen (Misgna et al., 2024; Pack et al., 2024). Dengan karakteristik tersebut, poin hasil berikutnya dapat masuk lebih dalam pada peta implementasi AI dan indikator akurasi evaluasi yang paling sering digunakan serta kondisi yang membuatnya efektif.

Tabel 1. Ringkasan Karakteristik Studi yang Dianalisis

No	Studi	Desain	Fokus AI	Konteks/domain formatif	Kontribusi untuk "akurasi evaluasi"
1	Deeva et al., 2021	Review	Automated feedback systems	Umum (lintas domain)	Kerangka klasifikasi + tantangan kualitas umpan balik
2	Ranalli et al., 2017	Empiris/validasi	AWE feedback	L2 writing (formatif)	Akurasi & kegunaan umpan balik dalam validasi berbasis argumen
3	Mao et al., 2018	Empiris/validasi	Automated scoring	Scientific argumentation (formatif)	Validasi skor terhadap konstruk argumentasi
4	Perikos et al., 2017	Desain sistem/empiris	ITS feedback	Logika/NL→logic (formatif)	Ketepatan bantuan & umpan balik

					bertahap
5	Tian & Zhou, 2020	Empiris	Automated feedback	Online EFL writing	Engagement dan pemanfaatan umpan balik terhadap revisi
6	Fleckenstein et al., 2023	Meta-analisis	AWE	Writing (lintas studi)	Efek AWE, variasi moderator, implikasi kualitas feedback
7	Zhai & Ma, 2023	Meta-analisis	AWE + feedback	EFL/ESL writing	Efektivitas AWE dan syarat integrasi pedagogis
8	Shi & Aryadoust, 2024	Systematic review	AI-written feedback	Writing feedback	Pemetaan bukti, isu transparansi & evaluasi validitas
9	Misgna et al., 2024	Survey/review	Deep learning AES + feedback	Essay scoring	Isu data latih, bias, interpretabilitas untuk akurasi
10	Pack et al., 2024	Empiris/validasi	LLM scoring + feedback	Essay scoring	Validitas & reliabilitas skor/feedback dari LLM

2. Peta Implementasi AI dalam Penilaian Formatif

Implementasi AI dalam penilaian formatif pada korpus inti menunjukkan bahwa teknologi paling sering dipakai sebagai mesin pemberi umpan balik dan pendukung konsistensi penilaian dalam siklus tugas–umpan balik–revisi. Pada domain menulis, AI diposisikan untuk membantu siswa memperoleh masukan cepat tentang kualitas teks, sekaligus membantu guru menjaga konsistensi penilaian ketika volume tugas besar (Fleckenstein et al., 2023; Zhai & Ma, 2023). Di luar menulis, implementasi AI juga muncul sebagai mekanisme bantuan bertahap di intelligent tutoring system (ITS) yang mengarahkan proses berpikir, sehingga fungsi formatifnya bukan hanya "menilai", tetapi "memandu" (Perikos et al., 2017). Dengan kata lain, peta implementasi pada literatur memperlihatkan dua orientasi utama: AI sebagai *assessment engine* (skor + umpan balik) dan AI sebagai *learning support engine* (diagnosis + scaffolding). Perbedaan orientasi ini penting karena cara AI diintegrasikan akan memengaruhi bentuk akurasi evaluasi yang dapat dicapai, apakah berpusat pada akurasi skor, akurasi diagnosis, atau keduanya. Kerangka klasifikasi umpan balik otomatis juga menegaskan bahwa variasi sistem sangat luas dan perlu dipahami sebagai spektrum fitur, bukan satu model tunggal (Deeva et al., 2021).

Pada implementasi berbasis Automated Writing Evaluation (AWE) dan Automated Essay Scoring (AES), AI umumnya bekerja dengan menilai artefak tulisan (esai, paragraf, jawaban singkat) kemudian menghasilkan skor dan/atau umpan balik yang dapat ditindaklanjuti. Ranalli et al. (2017) menekankan bahwa AWE untuk tujuan formatif tidak cukup hanya "memberi nilai", tetapi perlu diuji akurasi dan kegunaannya sebagai bagian dari validasi berbasis argumen, karena umpan balik harus benar-benar membantu revisi. Meta-analisis juga menunjukkan AWE cenderung efektif ketika

dipakai dalam konteks yang menyediakan ruang revisi dan dukungan pedagogis, bukan sekadar penilaian satu kali (Fleckenstein et al., 2023; Zhai & Ma, 2023). Implementasi AWE yang paling umum meliputi deteksi kesalahan bahasa, saran perbaikan koherensi, dan komentar berbasis fitur tertentu, meski tingkat kedalaman umpan balik bervariasi antar sistem. Variasi ini menjadi penentu apakah AI meningkatkan ketepatan evaluasi atau justru mendorong revisi dangkal. Karena itu, literatur menempatkan integrasi desain tugas dan rubrik sebagai prasyarat agar implementasi AWE benar-benar formatif.

Implementasi pada tugas yang menuntut konstruk kompleks, seperti argumentasi ilmiah, memperlihatkan kebutuhan yang lebih kuat pada validitas dan kesesuaian rubrik. Mao et al. (2018) menunjukkan bahwa automated scoring pada asesmen formatif argumentasi ilmiah harus dibangun dengan bukti validitas yang jelas agar skor benar-benar merepresentasikan kualitas argumentasi, bukan sekadar ciri permukaan jawaban. Dalam konteks ini, AI lebih sering dipakai untuk membantu menilai struktur argumen, penggunaan bukti, dan koherensi penalaran, yang secara tradisional memerlukan waktu besar jika dinilai manual. Implementasinya sering melibatkan rubrik analitik dan indikator konstruk yang eksplisit, kemudian diuji kesesuaiannya melalui prosedur validasi. Hasilnya, peta implementasi AI tidak hanya "menambah otomatisasi", tetapi menuntut desain pengukuran yang lebih disiplin. Dengan begitu, implementasi AI pada domain kompleks memperlihatkan bahwa peningkatan akurasi evaluasi sangat bergantung pada keselarasan konstruk yang diukur dengan tujuan pembelajaran.

Pada Intelligent Tutoring System (ITS), implementasi AI memperlihatkan karakter yang berbeda karena sistem biasanya memberikan bantuan dan umpan balik secara step-by-step selama siswa mengerjakan tugas. Perikos et al. (2017) menunjukkan mekanisme bantuan dan umpan balik dalam ITS dapat dirancang untuk mengarahkan siswa dari kesalahan menuju langkah logika yang benar, sehingga evaluasi formatif terjadi secara berkelanjutan selama proses belajar. Implementasi seperti ini menekankan "akurasi diagnosis" sebagai fondasi: sistem harus mampu mendeteksi kesalahan konsep dan memilih bentuk bantuan yang tepat. Dalam peta implementasi, ITS sering memadukan penilaian formatif mikro (setiap langkah) dengan rekomendasi pembelajaran adaptif. Dampaknya, akurasi evaluasi pembelajaran di sini lebih dekat ke ketepatan *scaffolding* dan ketepatan penentuan level bantuan. Model implementasi ini relevan untuk memperluas AI formatif di luar menulis, terutama untuk keterampilan prosedural dan penalaran.

Keterlibatan siswa (engagement) muncul sebagai komponen implementasi yang tidak bisa diabaikan karena umpan balik otomatis hanya berdampak jika benar-benar digunakan dalam revisi. Tian dan Zhou (2020) menunjukkan bahwa engagement terhadap umpan balik otomatis, teman sebaya, dan guru dapat berbeda dalam pembelajaran daring, dan perbedaan ini memengaruhi pemanfaatan umpan balik untuk perbaikan. Temuan ini menempatkan implementasi AI formatif sebagai praktik pedagogis, bukan sekadar adopsi alat. Banyak sistem AI mengasumsikan siswa akan membaca, memahami, lalu merevisi, tetapi literatur menunjukkan proses itu memerlukan desain aktivitas yang memaksa "aksi" (misalnya draft–revision–reflection). Meta-analisis juga cenderung menempatkan integrasi pembelajaran sebagai moderator penting terhadap efektivitas AWE (Fleckenstein et al., 2023; Zhai & Ma, 2023). Karena itu, peta implementasi AI yang efektif biasanya memasukkan komponen orientasi siswa, panduan revisi, dan rubrik yang dipahami bersama. Dengan cara ini, akurasi evaluasi meningkat bukan hanya karena AI "lebih konsisten", tetapi karena siklus formatif berjalan dengan benar.

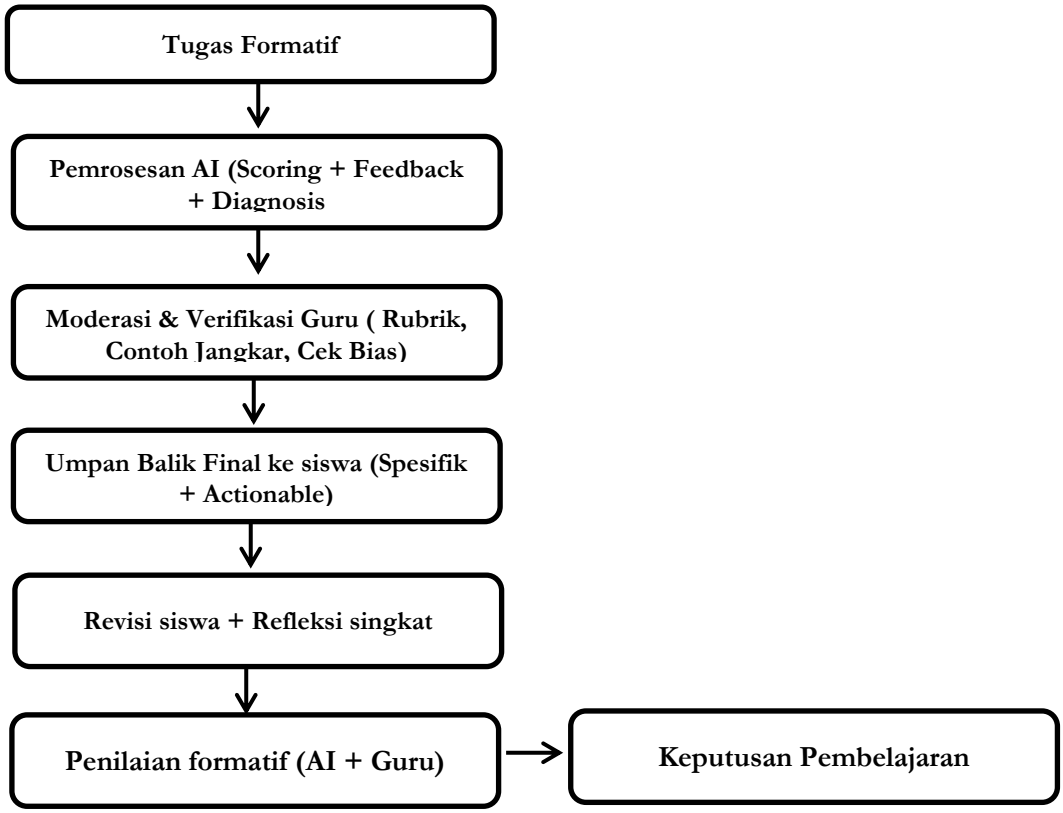
Ketika implementasi AI memasuki pendekatan deep learning, literatur menyoroti peningkatan kemampuan model dalam memetakan pola bahasa dan struktur esai, tetapi juga meningkatnya risiko "black box" dan bias akibat data pelatihan. Misgna et al. (2024) menekankan bahwa automated essay scoring dan generasi umpan balik berbasis deep learning sangat bergantung pada kualitas, keragaman, dan representativitas data latih, sehingga konteks lokal yang berbeda dapat menghasilkan deviasi akurasi. Di peta implementasi, deep learning sering dipakai untuk meningkatkan akurasi prediksi skor, namun evaluasi formatif yang baik menuntut lebih dari itu: umpan balik harus spesifik, selaras rubrik, dan aman dari bias sistematis. Karena itu, literatur banyak menempatkan kebutuhan audit data, evaluasi fairness, serta prosedur verifikasi sebagai bagian dari implementasi. Pada

praktiknya, implementasi yang matang biasanya memadukan hasil model dengan moderasi guru atau rubrik yang dipertegas. Pola ini memperlihatkan bahwa implementasi deep learning perlu ditopang tata kelola, bukan sekadar performa model.

Gelombang terbaru dalam peta implementasi adalah penggunaan Large Language Models (LLM) untuk memberi skor dan umpan balik, yang sering tampak unggul dari sisi keluwesan bahasa, namun memunculkan risiko umpan balik yang terdengar meyakinkan tetapi tidak selalu tepat. Pack et al. (2024) menekankan pentingnya menguji validitas dan reliabilitas skor/umpan balik LLM secara ketat sebelum dipakai untuk tujuan formatif, karena kesalahan sistem dapat mengarahkan pembelajaran ke arah yang keliru. Dalam peta implementasi, LLM sering diposisikan sebagai pemberi draft umpan balik atau co-assessor yang membantu merumuskan komentar berbasis rubrik, bukan pengganti penilai. Implementasi yang disarankan literatur cenderung memasukkan kontrol kualitas: rubrik eksplisit, prompt yang dibatasi, contoh anchor, dan verifikasi hasil oleh guru. Dengan demikian, peta implementasi LLM paling aman berada pada model *human-in-the-loop* untuk menjaga akurasi evaluasi. Ini sekaligus menegaskan bahwa inovasi AI paling baru justru memperbesar urgensi prosedur validasi.

Secara keseluruhan, peta implementasi AI dalam penilaian formatif dapat disarikan menjadi alur kerja yang menggabungkan penugasan formatif, pemrosesan AI, dan tindakan pedagogis lanjutan. Deeva et al. (2021) menekankan bahwa desain sistem umpan balik otomatis memerlukan klasifikasi fitur dan perhatian pada tantangan agar peluangnya benar-benar menjadi peningkatan kualitas evaluasi. Meta-analisis dan review menegaskan bahwa efektivitas cenderung menguat ketika AI diintegrasikan sebagai bagian dari siklus revisi yang terstruktur, bukan sekadar alat penilai cepat (Fleckenstein et al., 2023; Shi & Aryadoust, 2024; Zhai & Ma, 2023). Sementara itu, studi validasi menunjukkan bahwa implementasi yang kuat memerlukan bukti keselarasan konstruk, reliabilitas, serta kesesuaian output dengan rubrik (Mao et al., 2018; Pack et al., 2024; Ranalli et al., 2017). Dengan dasar ini, implementasi AI formatif yang meningkatkan akurasi evaluasi pembelajaran dapat dipahami sebagai kombinasi teknologi, rubrik, dan desain pembelajaran yang memfasilitasi penggunaan umpan balik oleh siswa. Poin hasil berikutnya dapat masuk lebih tajam pada indikator akurasi yang paling sering dipakai dan bukti peningkatan akurasi yang dilaporkan lintas studi.

Gambar 2. Alur Implementasi AI Dalam Penilaian Formatif (Human-in-the-loop)



Tabel 2. Peta implementasi AI dan bentuk umpan balik formatif (korpus inti)

Tipe implementasi AI	Contoh rujukan dalam korpus	Bentuk output formatif	Fokus akurasi yang dituju	Catatan implementasi kunci
AWE/AES (writing)	Ranalli et al., 2017; Fleckenstein et al., 2023; Zhai & Ma, 2023	Skor + komentar per aspek; saran revisi	Agreement AI–human, validitas rubrik, kegunaan feedback	Efektif bila ada siklus revisi dan dukungan guru
Argumentation scoring	Mao et al., 2018	Skor kualitas argumen; indikator konstruk	Validitas konstruk, rubrik alignment	Perlu definisi konstruk dan validasi ketat
ITS feedback (step-by-step)	Perikos et al., 2017	Hint bertahap; diagnosis kesalahan; scaffolding	Akurasi diagnosis dan ketepatan bantuan	Cocok untuk keterampilan prosedural/penalaran
Automated feedback + engagement	Tian & Zhou, 2020	Feedback otomatis dibanding peer/teacher	Kegunaan feedback & pemanfaatan oleh siswa	Engagement menentukan dampak formatif
Kerangka sistem feedback	Deeva et al., 2021; Shi & Aryadoust, 2024	Klasifikasi fitur dan tantangan	Transparansi & kualitas desain evaluasi	Membantu standardisasi desain dan evaluasi
Deep learning AES + feedback	Misgna et al., 2024	Skor + generasi feedback	Stabilitas model, bias/fairness, interpretabilitas	Risiko "black box" butuh audit data
LLM scoring + feedback	Pack et al., 2024	Draft feedback & penilaian berbasis rubrik	Reliabilitas, validitas, konsistensi	Paling aman dengan verifikasi guru

3. Bukti Peningkatan Akurasi Evaluasi

Temuan lintas studi menunjukkan bahwa "peningkatan akurasi evaluasi" pada penilaian formatif berbasis AI paling sering dibuktikan melalui tiga kelompok indikator, yaitu validitas (apakah yang dinilai sesuai konstruk dan rubrik), reliabilitas/konsistensi (apakah hasilnya stabil), dan agreement AI–penilai manusia (seberapa dekat skor/penilaian AI dengan rater manusia). Dalam korpus inti, pendekatan ini muncul berulang karena akurasi pada evaluasi pembelajaran tidak cukup dinyatakan lewat kecepatan sistem atau kecanggihan model, tetapi harus ditopang oleh bukti bahwa output AI dapat dipertanggungjawabkan untuk tujuan formatif. Kajian kerangka umpan balik otomatis juga menegaskan bahwa tantangan utama sistem bukan sekadar menghasilkan feedback, melainkan memastikan feedback tepat sasaran dan selaras tujuan belajar, yang secara langsung berkaitan dengan validitas dan akurasi evaluasi (Deeva et al., 2021). Dengan demikian, bukti peningkatan akurasi di literatur cenderung bersifat "multi-bukti", bukan satu metrik tunggal.

Dari sisi validitas, studi validasi menekankan bahwa akurasi AI harus dibaca sebagai kesesuaian antara skor/feedback dengan konstruk yang ingin dibangun melalui penilaian formatif. Ranalli et al. (2017) menempatkan akurasi dan kegunaan umpan balik AWE sebagai bagian dari argument-based validation, sehingga fokusnya bukan hanya "berapa skor", tetapi "apakah skor dan feedback itu sahih untuk memandu revisi". Perspektif ini penting karena formatif mensyaratkan umpan balik yang diagnostik, spesifik, dan relevan dengan rubrik; jika umpan balik melenceng dari rubrik, akurasi evaluasi menjadi semu. Pada domain argumentasi ilmiah, Mao et al. (2018) menegaskan kebutuhan validasi automated scoring agar penilaian AI tidak terjebak pada ciri permukaan jawaban, tetapi benar-benar merepresentasikan kualitas argumentasi sesuai konstruk yang didefinisikan. Artinya, literatur memandang validitas sebagai fondasi utama yang membedakan "penilaian otomatis" dari "penilaian formatif yang akurat".

Bukti validitas juga tampak pada bagaimana studi-studi menyarankan penggunaan rubrik eksplisit dan contoh jangkar untuk meminimalkan interpretasi yang keliru, terutama ketika AI dipakai untuk memberi feedback yang beragam. Dalam konteks sistem umpan balik otomatis, Deeva et al. (2021) menyoroti tantangan personalisasi dan kesesuaian feedback dengan kebutuhan pembelajar, yang pada praktiknya menuntut rubrik yang jelas agar feedback tidak generik. Shi dan Aryadoust (2024) juga memperlihatkan bahwa riset umpan balik tertulis berbasis AI berkembang cepat, namun standar pelaporan bukti validitas dan transparansi evaluasi masih bervariasi, sehingga kualitas klaim "akurasi" tidak selalu setara antarstudi. Temuan ini menguatkan bahwa peningkatan akurasi evaluasi melalui AI lebih kuat ketika studi menyertakan argumen validitas yang jelas, bukan hanya melaporkan performa model. Maka, dalam SLR ini, validitas diperlakukan sebagai bukti kunci saat menilai apakah implementasi AI benar-benar meningkatkan akurasi evaluasi pembelajaran.

Pada indikator reliabilitas/konsistensi, literatur menekankan bahwa AI dapat membantu mengurangi variasi penilaian yang sering terjadi pada penilaian manual, terutama pada kelas besar atau tugas berulang. Namun, reliabilitas tidak otomatis terjamin hanya karena sistem "konsisten", sebab konsistensi juga bisa muncul dari bias yang berulang. Karena itu, studi-studi terbaru menekankan pentingnya menguji konsistensi skor/feedback lintas kondisi, termasuk stabilitas hasil pada variasi input dan prosedur penilaian yang berbeda. Dalam konteks LLM, Pack et al. (2024) menekankan bahwa validitas dan reliabilitas perlu diuji secara ketat karena model generatif dapat menghasilkan variasi output, sehingga tanpa prosedur kontrol kualitas, reliabilitas dapat melemah dan berdampak pada ketepatan keputusan formatif. Dengan demikian, reliabilitas diposisikan sebagai bukti yang harus dikawal melalui desain implementasi, bukan dianggap melekat pada teknologi.

Indikator agreement AI–penilai manusia sering digunakan sebagai jembatan praktis untuk menilai akurasi, terutama ketika AI menghasilkan skor. Pada konteks AWE, diskusi tentang akurasi biasanya mengait pada seberapa dekat skor/penilaian AI dengan rater manusia dan seberapa bermanfaat feedbacknya untuk revisi (Ranalli et al., 2017). Namun, literatur juga mengingatkan bahwa agreement tinggi tidak selalu identik dengan validitas konstruk; AI bisa saja "sepakat" dengan rater pada aspek tertentu tetapi masih gagal menangkap kualitas yang lebih substantif. Karena itu, Mao et al. (2018) memperkuat kebutuhan menggabungkan agreement dengan bukti konstruk agar akurasi evaluasi tidak direduksi menjadi kemiripan skor semata. Pola ini menjelaskan mengapa banyak studi menekankan kombinasi indikator: agreement untuk kedekatan hasil, validitas untuk kesesuaian konstruk, dan reliabilitas untuk kestabilan.

Temuan meta-analisis memperkaya bukti karena menggabungkan hasil lintas studi dan mengidentifikasi moderator yang menjelaskan kapan "akurasi evaluasi" cenderung membaik. Fleckenstein et al. (2023) menunjukkan bahwa dampak AWE terhadap hasil menulis bervariasi, sehingga interpretasi "AI meningkatkan kualitas evaluasi" perlu memperhatikan konteks, desain tugas, dan karakteristik umpan balik. Zhai dan Ma (2023) juga menegaskan bahwa pada konteks EFL/ESL, efektivitas AWE dan feedback sangat dipengaruhi oleh bagaimana siswa menggunakan umpan balik dan bagaimana guru mengintegrasikannya ke siklus pembelajaran. Temuan ini penting untuk akurasi evaluasi karena formatif tidak berhenti pada penilaian, tetapi berakhir pada tindakan perbaikan; akurasi menjadi bermakna jika feedback memicu revisi yang tepat. Dengan kata lain,

meta-analisis mendukung klaim bahwa akurasi evaluasi meningkat bila AI dipakai dalam workflow formatif yang benar, bukan sekadar "alat pemberi nilai".

Dari perspektif teknologi, survei dan review terbaru menyoroti bahwa peningkatan kemampuan model (misalnya deep learning) sering diikuti oleh tantangan baru pada bias, interpretabilitas, dan ketergantungan pada data latih. Misgna et al. (2024) menegaskan bahwa performa AES/feedback berbasis deep learning sangat sensitif terhadap kualitas dan representativitas data, sehingga "akurasi" bisa menurun saat model dipindahkan ke konteks berbeda. Ini berimplikasi langsung pada evaluasi pembelajaran: sistem yang tampak akurat di satu konteks belum tentu akurat di kelas atau kurikulum lain. Di sisi lain, literatur menempatkan audit data, evaluasi fairness, dan kontrol kualitas sebagai mekanisme untuk menjaga akurasi tetap bermakna. Dengan demikian, bukti peningkatan akurasi juga mencakup bukti tata kelola implementasi, bukan hanya bukti metrik model.

Secara keseluruhan, bukti peningkatan akurasi evaluasi pembelajaran melalui AI paling kuat ketika penelitian menggabungkan: (a) argumen validitas yang jelas dan selaras rubrik, (b) pengujian reliabilitas/konsistensi output, (c) pemeriksaan agreement dengan rater manusia secara hati-hati, dan (d) integrasi pedagogis yang memastikan feedback digunakan untuk revisi dan keputusan belajar. Deeva et al. (2021) serta Shi dan Aryadoust (2024) menegaskan pentingnya standar evaluasi dan transparansi, sementara Ranalli et al. (2017) dan Mao et al. (2018) menunjukkan bagaimana validasi formatif perlu dibangun di level konstruk. Meta-analisis (Fleckenstein et al., 2023; Zhai & Ma, 2023) memperlihatkan bahwa efek dan kualitas evaluasi sangat dipengaruhi moderator implementasi, sedangkan kajian teknologi terbaru (Misgna et al., 2024; Pack et al., 2024) menekankan urgensi kontrol kualitas pada model yang makin kompleks. Temuan-temuan ini menjadi landasan bahwa "akurasi evaluasi" dalam penilaian formatif berbasis AI harus dipahami sebagai kualitas pengukuran + kualitas umpan balik + kualitas penggunaan dalam pembelajaran.

Tabel 3. Indikator Akurasi Evaluasi yang Digunakan dalam Literasi (Korpus Inti)

Kelompok indikator	Definisi ringkas	Contoh bukti yang dicari	Rujukan dalam korpus
Validitas	Kesesuaian skor/feedback dengan konstruk & rubrik penilaian	rubric alignment, argument-based validation, kesesuaian konstruk tugas	Ranalli et al., 2017; Mao et al., 2018; Deeva et al., 2021; Shi & Aryadoust, 2024
Reliabilitas/konsistensi	Stabilitas hasil penilaian/feedback lintas kondisi	konsistensi output, stabilitas skor, variasi output (khusus LLM)	Pack et al., 2024; (didiskusikan lintas review: Shi & Aryadoust, 2024; Misgna et al., 2024)
Agreement AI–human	Kedekatan skor AI dengan penilai manusia	perbandingan skor, kedekatan keputusan, error analysis secara konseptual	Ranalli et al., 2017; Mao et al., 2018; Pack et al., 2024
Kegunaan feedback (formatif)	Sejauh mana feedback membantu perbaikan belajar	feedback specificity, actionability, dampak terhadap revisi & engagement	Tian & Zhou, 2020; Fleckenstein et al., 2023; Zhai & Ma, 2023; Ranalli et al., 2017
Bias/fairness & tata	Risiko ketidakadilan	audit bias,	Misgna et al., 2024;

kelola	dan kontrol kualitas	ketergantungan data latih, interpretabilitas, verifikasi guru	Pack et al., 2024; Deeva et al., 2021; Shi & Aryadoust, 2024

4. Faktor Penentu, Risiko, dan Implikasi Implementasi

Temuan lintas studi menunjukkan bahwa keberhasilan AI dalam meningkatkan akurasi evaluasi pembelajaran tidak terjadi secara otomatis, melainkan ditentukan oleh kombinasi faktor teknis, faktor pedagogis, dan faktor tata kelola. Pada level teknis, kualitas data dan desain model sangat memengaruhi ketepatan skor maupun umpan balik, sedangkan pada level pedagogis, efektivitas sangat ditentukan oleh apakah umpan balik AI benar-benar digunakan dalam siklus formatif (tugas–umpan balik–revisi). Meta-analisis menegaskan bahwa dampak AWE tidak seragam dan dipengaruhi oleh konteks serta cara integrasi pembelajaran (Fleckenstein et al., 2023; Zhai & Ma, 2023). Di sisi lain, kajian sistem umpan balik otomatis menyoroti peluang besar sekaligus tantangan terkait personalisasi, relevansi, dan kualitas feedback (Deeva et al., 2021). Ini mengindikasikan bahwa peningkatan akurasi evaluasi perlu dipahami sebagai hasil dari desain implementasi yang tepat, bukan sekadar mengadopsi teknologi AI.

Salah satu faktor penentu paling kuat adalah keselarasan rubrik dan konstruk dengan output AI. Studi validasi menekankan bahwa ketika tujuan pembelajaran dan rubrik tidak didefinisikan secara jelas, AI cenderung menghasilkan feedback yang generik atau menekankan aspek permukaan, sehingga akurasi evaluasi menjadi lemah meskipun sistem tampak konsisten. Ranalli et al. (2017) menunjukkan pentingnya menilai akurasi dan kegunaan umpan balik AWE dalam kerangka validasi berbasis argumen, yang menuntut keselarasan antara klaim penggunaan formatif dan bukti yang mendukungnya. Pada domain yang lebih kompleks seperti argumentasi ilmiah, Mao et al. (2018) menegaskan bahwa automated scoring harus mempresentasikan konstruk argumentasi yang tepat; bila tidak, penilaian dapat melenceng dari kompetensi yang sebenarnya ingin dikembangkan. Karena itu, implementasi yang berhasil biasanya didahului oleh perumusan rubrik analitik yang eksplisit, contoh jangkar, serta prosedur evaluasi yang memastikan output AI merefleksikan indikator rubrik. Faktor ini berimplikasi langsung pada kualitas keputusan formatif, seperti pemberian remedial atau pengayaan.

Faktor penentu berikutnya adalah kualitas umpan balik dari sisi spesifisitas, keterjelasan, dan keberterapan (*actionability*). Studi tentang engagement menunjukkan bahwa umpan balik otomatis akan berdampak jika siswa memahami dan mampu menindaklanjutinya dalam revisi atau perbaikan strategi belajar (Tian & Zhou, 2020). Meta-analisis AWE menguatkan bahwa perbedaan fitur umpan balik dan rancangan tugas dapat menjadi moderator utama, sehingga sistem yang hanya memberi skor tanpa arahan perbaikan cenderung lebih lemah dampaknya dibanding sistem yang memberi komentar diagnostik (Fleckenstein et al., 2023; Zhai & Ma, 2023). Kajian Deeva et al. (2021) juga menyoroti perlunya klasifikasi dan desain fitur feedback untuk menjembatani kebutuhan belajar yang beragam, termasuk tantangan dalam personalisasi dan ketepatan rekomendasi. Dengan demikian, implementasi AI yang ingin meningkatkan akurasi evaluasi harus memprioritaskan kualitas feedback sebagai "inti formatif", bukan sekadar menonjolkan otomatisasi skor. Jika kualitas feedback rendah, akurasi evaluasi mungkin tampak baik di metrik, tetapi manfaat pembelajarannya minimal.

Di sisi teknis, faktor yang sering menentukan adalah kualitas dan representativitas data latih, terutama pada pendekatan deep learning untuk AES dan generasi umpan balik. Misgna et al. (2024) menegaskan bahwa performa model sangat dipengaruhi oleh data, sehingga ketika data tidak merepresentasikan variasi konteks (ragam bahasa, level kemampuan, topik, dan gaya penulisan), akurasi dapat turun saat diterapkan pada kelas atau kurikulum yang berbeda. Kondisi ini menimbulkan risiko *domain shift* yang dapat mengganggu ketepatan evaluasi, misalnya model terlalu "menghargai" panjang tulisan atau fitur tertentu yang tidak selaras dengan rubrik guru. Karena itu, literatur menyarankan audit data, pengujian lintas konteks, dan pembaruan model secara hati-hati agar sistem tetap akurat dan adil. Temuan ini memperlihatkan bahwa peningkatan akurasi evaluasi

bukan hanya isu "model canggih", tetapi juga isu kesiapan data dan kecocokan konteks penerapan. Ketika kontrol pada data lemah, implementasi AI justru berpotensi memperbesar kesalahan evaluasi secara sistematis.

Faktor penentu lain yang menguat dalam literatur adalah peran guru dan desain human-in-the-loop sebagai kontrol kualitas. Perikos et al. (2017) menunjukkan bahwa mekanisme bantuan dan umpan balik dalam ITS dapat efektif karena sistem dirancang untuk mengarahkan proses berpikir bertahap, namun akurasi diagnosis dan jalur bantuan tetap perlu disejajarkan dengan tujuan pengajaran. Pada implementasi AWE/LLM, peran guru menjadi semakin penting untuk memeriksa kesesuaian feedback dengan rubrik, mengurangi bias, dan memastikan komentar yang diberikan tidak menyesatkan. Pack et al. (2024) secara khusus menekankan bahwa validitas dan reliabilitas output LLM perlu diuji dan dikontrol, karena variasi output model dapat menurunkan konsistensi dan mengganggu ketepatan keputusan formatif. Dengan desain human-in-the-loop, AI berfungsi sebagai pemberi draft skor/feedback, sedangkan guru berfungsi sebagai moderator yang memvalidasi, menyesuaikan, atau menolak output. Pola ini membuat implementasi lebih aman dan meningkatkan peluang akurasi evaluasi yang dapat dipertanggungjawabkan.

Selain faktor penentu, literatur juga menyoroti risiko utama yang harus dikelola agar implementasi AI tidak merusak kualitas evaluasi. Risiko pertama adalah bias dan fairness, terutama ketika model dilatih pada data yang tidak seimbang atau ketika sistem lebih cocok untuk kelompok tertentu. Risiko kedua adalah masalah transparansi dan interpretabilitas, yang dapat menurunkan kepercayaan pengguna serta menyulitkan guru menjelaskan dasar penilaian. Shi dan Aryadoust (2024) menunjukkan bahwa riset umpan balik tertulis berbasis AI masih menghadapi tantangan pada standar evaluasi dan transparansi, sehingga klaim efektivitas tidak selalu disertai bukti yang seragam. Risiko ketiga adalah *overreliance* dan "automation bias", ketika guru atau siswa terlalu percaya pada output AI tanpa verifikasi. Risiko ini semakin penting pada LLM karena umpan balik bisa terdengar sangat meyakinkan meskipun tidak selalu tepat (Pack et al., 2024). Tanpa mitigasi, risiko-risiko ini dapat menurunkan akurasi evaluasi sekaligus mengganggu integritas proses pembelajaran.

Literatur juga memperlihatkan bahwa keberhasilan implementasi AI formatif dipengaruhi oleh kesiapan pengguna dan desain aktivitas yang mendorong pemanfaatan feedback. Tian dan Zhou (2020) menggarisbawahi bahwa engagement siswa terhadap umpan balik otomatis berbeda-beda, sehingga implementasi perlu disertai panduan revisi, refleksi, dan penguatan literasi umpan balik. Meta-analisis menegaskan bahwa integrasi pedagogis seperti kewajiban revisi, penggunaan rubrik yang dipahami bersama, dan dukungan guru dalam menafsirkan feedback cenderung memperkuat dampak AWE (Fleckenstein et al., 2023; Zhai & Ma, 2023). Hal ini menandakan bahwa akurasi evaluasi pembelajaran tidak hanya produk algoritma, tetapi juga produk *learning design* yang membuat feedback diproses menjadi tindakan belajar. Tanpa desain tersebut, AI bisa menjadi "alat koreksi cepat" yang tidak mengubah proses belajar secara substantif. Karena itu, implementasi yang disarankan literatur cenderung memasukkan komponen pelatihan singkat pengguna dan prosedur penggunaan feedback secara bertahap.

Implikasi praktis dari temuan ini adalah bahwa institusi pendidikan perlu menempatkan implementasi AI dalam penilaian formatif sebagai program yang mencakup standar kualitas, bukan sekadar adopsi teknologi. Deeva et al. (2021) menekankan kebutuhan memahami klasifikasi sistem umpan balik otomatis dan tantangannya, sementara kajian review dan survei menyoroti pentingnya evaluasi validitas, reliabilitas, serta mitigasi bias (Misgna et al., 2024; Shi & Aryadoust, 2024). Oleh sebab itu, implementasi yang disarankan meliputi: rubrik yang jelas dan dipetakan ke output AI, mekanisme human-in-the-loop, audit bias dan uji lintas konteks, serta prosedur dokumentasi keputusan penilaian. Pada LLM, implikasi tambahan adalah perlunya pembatasan prompt dan pengujian konsistensi output sebelum digunakan luas (Pack et al., 2024). Dengan memenuhi prasyarat tersebut, AI berpotensi meningkatkan akurasi evaluasi pembelajaran sekaligus menjaga akuntabilitas asesmen.

Tabel 4. Faktor Penentu, Risiko, dan Strategi Mitigasi Implementasi AI Formatif

Area	Faktor penentu (temuan utama)	Risiko jika diabaikan	Strategi mitigasi yang disarankan literatur	Rujukan pendukung
Rubrik & konstruk	Rubrik jelas, konstruk terdefinisi, alignment dengan output AI	Skor/feedback melenceng dari tujuan belajar	Rubrik analitik, contoh jangkar, validasi berbasis argumen	Ranalli et al., 2017; Mao et al., 2018
Kualitas feedback	Feedback spesifik, dapat ditindaklanjuti, terintegrasi revisi	Revisi dangkal, learning gain kecil	Siklus draft–feedback–revisi; panduan interpretasi feedbcak	Fleckenstein et al., 2023; Zhai & Ma, 2023; Tian & Zhou, 2020
Data & model	Guru memoderasi output AI	Overreliance, automation bias	Verifikasi guru, sampling check, standar toleransi kesalahan	Pack et al., 2024; Perikos et al., 2017
Fairness & etika	Uji bias, privasi, penggunaan data	Ketidakadilan penilaian, isu privasi	Evaluasi fairness, kebijakan privasi, pembatasan penggunaan	Misgna et al., 2024; Pack et al., 2024

PEMBAHASAN

1. AI Sebagai Penguat Akurasi Penilaian Formatif

AI dapat memperkuat akurasi penilaian formatif terutama melalui kemampuan menyediakan umpan balik cepat dan konsisten yang membantu proses evaluasi berlangsung lebih sering dan lebih terarah. Pada konteks penilaian menulis, sintesis bukti menunjukkan bahwa AWE berpotensi meningkatkan kualitas evaluasi ketika umpan baliknya diposisikan sebagai bagian dari siklus formatif, bukan sekadar pengganti penilaian manual (Fleckenstein et al., 2023; Zhai & Ma, 2023). Konsistensi ini penting karena penilaian formatif menuntut monitoring kemajuan yang berulang; semakin sering siswa menerima umpan balik, semakin besar peluang guru mendeteksi kesenjangan belajar lebih dini. Namun, literatur juga menegaskan bahwa besarnya manfaat berbeda antar konteks, sehingga akurasi tidak boleh dipahami sebagai hasil otomatis dari adopsi AI, melainkan sebagai output dari integrasi yang tepat. Karena itu, peran AI yang paling kuat dalam penilaian formatif adalah memperluas kapasitas evaluasi agar lebih responsif terhadap kebutuhan belajar. Dengan kapasitas tersebut, proses penilaian menjadi lebih presisi dalam memotret performa siswa pada tiap tahap pembelajaran.

Dari sisi validitas, AI berkontribusi pada akurasi evaluasi ketika output sistem—baik skor maupun umpan balik—selaras dengan konstruk dan rubrik yang digunakan dalam pembelajaran. Ranalli et al. (2017) menegaskan bahwa AWE perlu dibaca melalui kerangka validasi berbasis argumen, sehingga pertanyaan utamanya bukan hanya "apakah sistem memberi skor", melainkan "apakah skor dan umpan balik itu sahih untuk tujuan formatif dan benar-benar membantu revisi." Pada domain keterampilan kompleks seperti argumentasi ilmiah, validasi menjadi lebih kritikal karena sistem harus menilai aspek yang substantif, bukan sekadar ciri permukaan teks (Mao et al., 2018). Dengan rubrik yang jelas, AI dapat membantu mengurangi subjektivitas penilaian serta membuat indikator

penilaian lebih konsisten diterapkan lintas tugas. Temuan review juga memperlihatkan bahwa standar bukti validitas masih bervariasi, sehingga kontribusi AI terhadap akurasi paling kuat pada studi yang menjelaskan prosedur evaluasi dan keselarasan rubrik secara eksplisit (Shi & Aryadoust, 2024). Ini menunjukkan bahwa akurasi evaluasi berbasis AI adalah hasil pertemuan antara kualitas sistem dan kualitas desain pengukuran.

Selain validitas, AI juga meningkatkan akurasi evaluasi melalui penguatan reliabilitas dan konsistensi penilaian, terutama pada tugas yang dinilai berulang dalam kelas besar. Review tentang sistem umpan balik otomatis menyoroti bahwa salah satu peluang utama AI adalah menjaga konsistensi pemberian feedback, tetapi tantangan muncul ketika feedback menjadi terlalu generik atau tidak sesuai kebutuhan pembelajar (Deeva et al., 2021). Pada model yang lebih kompleks seperti deep learning dan LLM, literatur menekankan bahwa reliabilitas tidak dapat diasumsikan; perlu pengujian kestabilan output dan kontrol kualitas agar penilaian tetap konsisten dan dapat dipertanggungjawabkan (Misgna et al., 2024; Pack et al., 2024). Dalam konteks formatif, konsistensi berarti siswa menerima sinyal evaluasi yang tidak berubah-ubah untuk kesalahan yang sejenis, sehingga mereka dapat membangun strategi perbaikan yang stabil. Namun, konsistensi juga harus disertai ketepatan; konsistensi yang bias justru memperburuk akurasi. Karena itu, kontribusi AI terhadap reliabilitas menjadi optimal ketika disertai mekanisme audit dan verifikasi.

Kontribusi AI terhadap akurasi evaluasi juga tampak melalui peningkatan kualitas umpan balik sebagai informasi diagnostik bagi siswa dan guru. Studi tentang engagement menunjukkan bahwa dampak formatif bergantung pada bagaimana siswa memproses dan memanfaatkan umpan balik, bukan hanya pada keberadaan umpan balik itu sendiri (Tian & Zhou, 2020). Meta-analisis memperkuat bahwa sistem AWE paling bermanfaat ketika umpan baliknya mendorong revisi bermakna dan didukung oleh integrasi pedagogis yang jelas (Fleckenstein et al., 2023; Zhai & Ma, 2023). Dengan demikian, AI dapat meningkatkan akurasi evaluasi dalam arti yang lebih luas: bukan hanya akurasi skor, tetapi akurasi diagnosis terhadap kebutuhan belajar yang kemudian menghasilkan tindakan perbaikan yang tepat. Ketika umpan balik mampu menunjukkan bagian mana yang perlu diperbaiki dan bagaimana cara memperbaikinya, evaluasi pembelajaran menjadi lebih presisi dan berorientasi pada peningkatan. Hal ini sejalan dengan kerangka sistem umpan balik otomatis yang menuntut desain feedback yang terstruktur dan relevan dengan tujuan pembelajaran (Deeva et al., 2021).

Akhirnya, literatur menegaskan bahwa AI paling efektif meningkatkan akurasi evaluasi pembelajaran ketika ditempatkan dalam model implementasi human-in-the-loop, di mana AI memberi skor/feedback awal dan guru memoderasi hasilnya. Pada ITS, mekanisme bantuan bertahap menekankan pentingnya diagnosis yang tepat agar umpan balik benar-benar mengarahkan proses berpikir siswa (Perikos et al., 2017). Pada LLM, kebutuhan verifikasi bahkan lebih kuat karena variasi output dapat memengaruhi reliabilitas serta berisiko menghasilkan umpan balik yang terdengar meyakinkan tetapi keliru, sehingga uji validitas dan reliabilitas menjadi prasyarat utama untuk penggunaan formatif (Pack et al., 2024). Review juga menekankan bahwa risiko bias dan ketergantungan pada data latih dapat memengaruhi akurasi jika tidak diaudit (Misgna et al., 2024). Karena itu, peran guru sebagai kontrol kualitas memastikan bahwa keputusan formatif tetap adil, sesuai rubrik, dan selaras dengan konteks kelas. Dengan kombinasi tersebut, AI berfungsi sebagai penguat kapasitas evaluasi dan guru tetap menjadi penjaga akuntabilitas penilaian, sehingga akurasi evaluasi pembelajaran dapat meningkat secara nyata dan bertanggung jawab.

2. Kondisi Keberhasilan Implementasi

Keberhasilan implementasi AI dalam penilaian formatif sangat ditentukan oleh kondisi-kondisi kunci yang bertindak sebagai *moderator*, sehingga AI hanya akan meningkatkan akurasi evaluasi ketika dipasang pada desain pembelajaran dan asesmen yang tepat. Meta-analisis menunjukkan bahwa efek AWE terhadap kualitas belajar dan revisi tidak seragam, menandakan adanya perbedaan konteks, karakteristik tugas, serta cara penggunaan umpan balik yang memengaruhi hasil (Fleckenstein et al., 2023; Zhai & Ma, 2023). Hal ini berarti akurasi evaluasi bukan sekadar produk algoritma, tetapi juga produk dari lingkungan belajar yang memfasilitasi pemanfaatan umpan balik. Dalam praktiknya, AI

lebih mungkin meningkatkan ketepatan evaluasi pada kelas yang menerapkan siklus formatif yang jelas, misalnya ada draft awal, pemberian feedback, revisi, dan refleksi. Ketika siklus ini tidak ada, AI cenderung berubah menjadi alat koreksi cepat yang tidak banyak memperbaiki kualitas evaluasi pembelajaran. Karena itu, pembahasan poin ini memetakan kondisi yang paling sering disebut literatur sebagai penentu keberhasilan.

Moderator pertama yang paling menentukan adalah kejelasan rubrik dan keselarasan konstruk penilaian. Pada penilaian formatif, rubrik bukan hanya alat memberi skor, tetapi peta kompetensi yang harus dipahami bersama agar umpan balik terarah. Ranalli et al. (2017) menekankan pentingnya validasi berbasis argumen pada AWE, yang secara implisit mensyaratkan rubrik dan klaim penggunaan formatif yang jelas agar akurasi dan kegunaan umpan balik dapat dinilai. Pada domain argumentasi ilmiah, Mao et al. (2018) menunjukkan bahwa kesalahan umum dalam automated scoring adalah menilai fitur permukaan dan mengabaikan konstruk inti, sehingga alignment rubrik-konstruk menjadi krusial untuk menjaga akurasi evaluasi. Review sistem umpan balik otomatis juga menegaskan bahwa tantangan personalisasi dan relevansi feedback berakar pada bagaimana tujuan pembelajaran dipetakan ke desain feedback (Deeva et al., 2021). Dengan rubrik analitik, contoh jangkar, dan indikator konstruk yang eksplisit, AI lebih mudah diarahkan menghasilkan feedback yang tepat sasaran. Tanpa rubrik yang kuat, akurasi evaluasi akan rentan bias dan tidak stabil lintas tugas.

Moderator kedua adalah desain tugas dan desain siklus revisi, karena formatif menuntut perubahan perilaku belajar, bukan hanya pemberian skor. Meta-analisis memperlihatkan bahwa AWE lebih efektif ketika siswa diberi kesempatan revisi dan ketika umpan balik disusun agar mendorong perbaikan substansial, bukan sekadar koreksi permukaan (Fleckenstein et al., 2023; Zhai & Ma, 2023). Temuan engagement juga menegaskan bahwa siswa akan lebih memanfaatkan umpan balik otomatis jika aktivitas pembelajaran memang mengharuskan mereka menindaklanjutinya, misalnya melalui revisi wajib atau refleksi singkat (Tian & Zhou, 2020). Dengan kata lain, desain pembelajaran harus "memaksa" terjadinya pemanfaatan feedback agar akurasi evaluasi menjadi bermakna dalam keputusan belajar. Jika tugas hanya satu kali dan tidak ada revisi, maka AI memang bisa menilai dengan cepat, tetapi kontribusinya terhadap akurasi evaluasi pembelajaran menjadi dangkal. Oleh karena itu, literatur menempatkan struktur draft–feedback–revisi sebagai moderator yang sangat kuat pada penilaian formatif berbasis AI. Ini juga menjelaskan mengapa domain writing paling sering menunjukkan manfaat karena secara alami cocok dengan siklus revisi.

Moderator ketiga adalah kualitas data dan karakteristik model, terutama pada implementasi berbasis deep learning dan sistem yang dikembangkan dari korpus tertentu. Misgna et al. (2024) menekankan bahwa AES dan generasi umpan balik berbasis deep learning sangat dipengaruhi kualitas serta representativitas data latih; ketika data tidak mencerminkan konteks penggunaan, akurasi bisa turun karena *domain shift*. Pada praktiknya, ini berarti AI yang tampak akurat di satu konteks dapat meleset ketika diterapkan pada kurikulum, tingkat kemampuan, atau ragam bahasa yang berbeda. Review AI feedback juga menyoroti variasi kualitas studi dan standar evaluasi, sehingga kualitas pelaporan data dan prosedur evaluasi turut menjadi penentu apakah sebuah sistem layak digunakan (Shi & Aryadoust, 2024). Dengan demikian, implementasi yang sukses biasanya mencakup uji coba pada konteks lokal, audit data, serta pengujian ulang indikator validitas–reliabilitas sebelum diperluas. Tanpa langkah-langkah tersebut, AI berpotensi menghasilkan kesalahan evaluasi secara sistematis meski tampak konsisten.

Moderator keempat adalah peran guru dan desain human-in-the-loop sebagai mekanisme kontrol kualitas. Dalam ITS, bantuan bertahap efektif karena sistem didesain untuk memberikan dukungan yang sesuai dengan langkah berpikir siswa, sehingga ketepatan diagnosis menjadi faktor inti (Perikos et al., 2017). Dalam AWE dan terutama LLM, literatur menekankan bahwa guru perlu memoderasi output AI karena variasi dan ketidakpastian output dapat mengganggu reliabilitas serta berisiko menghasilkan umpan balik yang meyakinkan namun tidak tepat (Pack et al., 2024). Model human-in-the-loop memungkinkan AI mempercepat proses dan menyediakan draf feedback, sementara guru menjaga kesesuaian rubrik, memeriksa bias, dan menyesuaikan dengan konteks kelas. Meta-analisis juga secara tidak langsung menguatkan bahwa integrasi pedagogis, termasuk keterlibatan guru,

memoderasi efektivitas sistem AWE (Fleckenstein et al., 2023; Zhai & Ma, 2023). Jadi, keberhasilan implementasi bukan berarti mengganti guru, melainkan memperjelas peran guru sebagai pengendali akurasi dan akuntabilitas penilaian. Ketika peran ini diabaikan, akurasi evaluasi justru bisa melemah karena keputusan formatif bersandar pada output AI yang belum tervalidasi penuh.

Moderator kelima adalah engagement dan literasi umpan balik siswa, karena formatif bergantung pada kemampuan siswa menafsirkan dan memanfaatkan feedback. Tian dan Zhou (2020) menunjukkan bahwa keterlibatan siswa terhadap umpan balik otomatis tidak selalu sama dengan umpan balik guru atau teman sebaya, sehingga implementasi AI perlu disertai strategi untuk meningkatkan pemanfaatan feedback. Dalam konteks AWE, meta-analisis menegaskan bahwa manfaat cenderung lebih besar ketika siswa tidak hanya menerima umpan balik, tetapi juga diberi struktur untuk memprosesnya melalui revisi terarah (Fleckenstein et al., 2023; Zhai & Ma, 2023). Selain itu, Deeva et al. (2021) menyoroti bahwa sistem feedback otomatis menghadapi tantangan dalam memastikan feedback dipahami dan relevan, yang pada akhirnya terkait dengan literasi pengguna. Karena itu, implementasi yang berhasil biasanya memasukkan orientasi singkat tentang cara membaca feedback, contoh perbaikan, dan aktivitas refleksi. Tanpa literasi ini, feedback yang akurat sekalipun dapat gagal meningkatkan akurasi evaluasi karena tidak menghasilkan tindakan belajar yang tepat.

Secara ringkas, literatur menyimpulkan bahwa kondisi keberhasilan implementasi AI formatif dapat dirangkum menjadi lima moderator utama: rubrik-konstruk yang jelas, desain tugas dan siklus revisi, kualitas data dan kecocokan model, human-in-the-loop oleh guru, serta engagement dan literasi feedback siswa. Temuan ini menjelaskan mengapa beberapa studi atau konteks menunjukkan peningkatan akurasi yang kuat, sementara konteks lain menunjukkan hasil yang lebih kecil atau bervariasi (Fleckenstein et al., 2023; Shi & Aryadoust, 2024; Zhai & Ma, 2023). Dengan memetakan moderator tersebut, pembahasan dapat menjelaskan bahwa AI bukan "solusi instan", melainkan alat yang efektif jika dipasang pada desain asesmen yang benar dan dikelola dengan kontrol kualitas. Perspektif ini juga selaras dengan studi validasi yang menekankan pentingnya bukti konstruk dan ketepatan penggunaan (Mao et al., 2018; Ranalli et al., 2017). Dengan demikian, rekomendasi implementasi yang muncul dari SLR ini menekankan penataan ekosistem asesmen, bukan sekadar memilih teknologi. Poin pembahasan berikutnya dapat fokus pada risiko dan isu etika yang muncul ketika moderator tersebut tidak dipenuhi, khususnya pada penggunaan model yang semakin kompleks seperti deep learning dan LLM.

3. Risiko dan isu etika yang memengaruhi akurasi

Risiko dan isu etika menjadi bagian yang tidak terpisahkan dari pembahasan akurasi evaluasi berbasis AI, karena ketepatan penilaian tidak hanya ditentukan oleh performa model, tetapi juga oleh fairness, transparansi, dan akuntabilitas penggunaan. Review tentang sistem umpan balik otomatis menekankan bahwa selain peluang, terdapat tantangan serius terkait kesesuaian feedback dan risiko miskonsepsi ketika sistem memberikan umpan balik yang tidak tepat atau terlalu generik (Deeva et al., 2021). Variasi temuan pada meta-analisis juga mengindikasikan bahwa penggunaan AI dalam penilaian formatif dapat menghasilkan dampak yang berbeda-beda tergantung konteks, sehingga generalisasi tanpa kehati-hatian berpotensi menimbulkan keputusan evaluatif yang keliru (Fleckenstein et al., 2023; Zhai & Ma, 2023). Dalam penilaian formatif, keputusan keliru bukan hanya masalah "nilai", tetapi dapat mengarahkan strategi belajar yang salah dan memengaruhi motivasi siswa. Karena itu, risiko dan etika di sini langsung terkait dengan kualitas evaluasi pembelajaran. Pembahasan poin ini menempatkan risiko sebagai variabel yang dapat menurunkan akurasi jika tidak dimitigasi.

Risiko pertama yang paling menonjol adalah bias dan fairness, terutama ketika model dilatih atau diuji pada data yang tidak mewakili keragaman pembelajar. Survei tentang deep learning-based AES dan generasi umpan balik menekankan bahwa ketergantungan pada data latih dapat menghasilkan bias sistematis, misalnya model lebih "menguntungkan" gaya penulisan tertentu atau lebih akurat pada kelompok tertentu, sementara kelompok lain dinilai lebih rendah karena perbedaan bahasa dan struktur ekspresi (Misgna et al., 2024). Jika bias ini terjadi, AI mungkin terlihat konsisten tetapi

sebenarnya konsisten dalam ketidakadilan, sehingga akurasi evaluasi turun secara substantif. Pada konteks menulis, hal ini penting karena variasi kemampuan bahasa, latar budaya, dan strategi menulis sangat besar, sehingga model dapat salah menafsirkan kualitas gagasan karena lebih fokus pada fitur permukaan. Review juga menegaskan bahwa kualitas desain evaluasi fairness sering belum seragam, sehingga risiko bias kadang tidak terdeteksi dalam studi (Shi & Aryadoust, 2024). Karena itu, mitigasi bias perlu dipahami sebagai bagian dari penguatan akurasi evaluasi, bukan sekadar kepatuhan etika.

Risiko kedua adalah kurangnya transparansi dan interpretabilitas, yang berimplikasi pada rendahnya akuntabilitas penilaian. Ketika guru dan siswa tidak memahami dasar penilaian atau logika umpan balik, mereka sulit menilai apakah output AI selaras dengan rubrik atau justru menyimpang. Shi dan Aryadoust (2024) menyoroti variasi kualitas pelaporan dan evaluasi pada riset umpan balik tertulis berbasis AI, sehingga pembaca sering kesulitan membandingkan klaim efektivitas antar studi secara setara. Pada pendekatan deep learning, tantangan "black box" memperbesar risiko karena model bisa menghasilkan skor akurat di permukaan namun tidak mudah dijelaskan mengapa demikian (Misgna et al., 2024). Dalam penilaian formatif, keterjelasan alasan umpan balik penting karena siswa perlu memahami apa yang harus diperbaiki dan guru perlu memastikan umpan balik sejalan dengan tujuan belajar. Ketika transparansi rendah, risiko salah interpretasi meningkat, dan hal ini dapat menurunkan akurasi evaluasi dalam pengambilan keputusan pembelajaran.

Risiko ketiga yang semakin penting adalah ketergantungan berlebihan pada AI (overreliance/automation bias), terutama pada model generatif seperti LLM yang mampu menghasilkan umpan balik yang tampak sangat meyakinkan. Pack et al. (2024) menekankan perlunya pengujian validitas dan reliabilitas untuk penilaian esai berbasis LLM, karena variasi output dan potensi ketidaktepatan dapat mengganggu konsistensi dan ketepatan evaluasi formatif. Jika guru atau siswa terlalu percaya pada output AI tanpa verifikasi, maka keputusan formatif dapat salah arah, misalnya siswa mengejar perbaikan yang tidak relevan, atau guru membuat keputusan remedial/pengayaan berdasarkan sinyal yang tidak akurat. Dalam konteks sistem umpan balik otomatis, risiko overreliance juga dapat muncul ketika feedback cepat dianggap selalu benar, padahal pada beberapa kasus feedback bisa generik atau tidak sesuai konteks tugas (Deeva et al., 2021). Karena itu, literatur mendorong pendekatan human-in-the-loop sebagai kontrol kualitas, bukan hanya untuk fairness tetapi juga untuk mencegah automation bias. Dengan kata lain, menjaga peran profesional guru adalah strategi etis sekaligus strategi peningkatan akurasi.

Risiko keempat terkait privasi dan keamanan data, yang sering muncul ketika penilaian formatif berbasis AI mengumpulkan data tugas siswa dalam skala besar dan berulang. Walaupun korpus inti lebih banyak menyoroti evaluasi validitas dan reliabilitas, isu tata kelola data tetap relevan karena kualitas evaluasi dapat terganggu jika sistem harus membatasi data (misalnya karena kebijakan privasi) atau jika terjadi kebocoran yang menurunkan kepercayaan pengguna. Selain itu, data siswa yang digunakan untuk pelatihan atau penyempurnaan model dapat membawa risiko etis, terutama jika persetujuan, anonimisasi, dan pembatasan akses tidak jelas. Dalam konteks deep learning, ketergantungan pada data yang masif membuat praktik pengelolaan data menjadi semakin krusial (Misgna et al., 2024). Risiko privasi yang tidak dikelola dapat menimbulkan resistensi dari siswa/guru sehingga implementasi tidak berjalan baik, dan pada akhirnya menurunkan kualitas evaluasi pembelajaran. Dengan demikian, tata kelola data merupakan prasyarat etis yang juga mendukung kestabilan implementasi.

Secara keseluruhan, pembahasan ini menunjukkan bahwa risiko dan isu etika bukan "tambahan" dalam implementasi AI formatif, tetapi faktor yang secara langsung memengaruhi akurasi evaluasi pembelajaran. Bias dan fairness dapat membuat evaluasi meleset untuk kelompok tertentu, kurangnya transparansi dapat menurunkan akuntabilitas penilaian, overreliance dapat menghasilkan keputusan formatif yang salah, dan masalah tata kelola data dapat mengganggu keberlanjutan implementasi (Misgna et al., 2024; Pack et al., 2024; Shi & Aryadoust, 2024). Temuan meta-analisis juga mengingatkan bahwa variasi konteks dan desain dapat memperbesar risiko generalisasi yang berlebihan, sehingga praktik implementasi harus kontekstual dan berbasis bukti (Fleckenstein et al., 2023; Zhai & Ma, 2023). Karena itu, strategi mitigasi yang konsisten dengan literatur adalah

menguatkan rubrik, melakukan audit bias, meningkatkan transparansi evaluasi, serta menempatkan guru sebagai moderator utama dalam workflow penilaian formatif. Dengan pendekatan tersebut, AI dapat tetap menjadi penguat akurasi evaluasi tanpa mengorbankan prinsip keadilan dan integritas asesmen.

KESIMPULAN

Studi literatur ini menegaskan bahwa implementasi Artificial Intelligence (AI) dalam penilaian formatif berpotensi meningkatkan akurasi evaluasi pembelajaran, terutama ketika AI digunakan untuk mempercepat pemberian umpan balik, menjaga konsistensi penilaian, dan memperkuat diagnosis kebutuhan belajar. Bukti yang paling sering digunakan untuk menilai peningkatan akurasi meliputi validitas (keselarasan konstruk dan rubrik), reliabilitas/konsistensi output, serta kedekatan penilaian AI dengan penilai manusia, disertai indikator kegunaan umpan balik dalam mendorong revisi dan perbaikan belajar.

Temuan juga menunjukkan bahwa efektivitas AI sangat dipengaruhi oleh kondisi implementasi. Kejelasan rubrik, desain siklus draft–feedback–revisi, kualitas data dan kesesuaian model dengan konteks, literasi umpan balik siswa, serta keterlibatan guru sebagai moderator merupakan faktor penentu yang membedakan implementasi yang benar-benar meningkatkan akurasi dari implementasi yang hanya menambah otomatisasi.

Di sisi lain, penggunaan AI dalam asesmen formatif membawa risiko yang dapat menurunkan akurasi, seperti bias dan ketidakadilan penilaian, kurangnya transparansi, ketergantungan berlebihan pada output AI, serta isu tata kelola data. Risiko ini semakin relevan pada model yang lebih kompleks, sehingga diperlukan kontrol kualitas dan prosedur verifikasi yang ketat.

Secara praktis, implementasi AI formatif yang paling disarankan adalah model human-in-the-loop, di mana AI memberi skor/umpan balik awal, sementara guru memverifikasi kesesuaian rubrik, mengaudit anomali, dan menyesuaikan umpan balik dengan konteks kelas. Ke depan, riset perlu memperluas domain di luar penilaian menulis, menstandarkan pelaporan indikator validitas,reliabilitas, fairness, serta menguji kinerja lintas konteks agar akurasi evaluasi benar-benar robust dan dapat digeneralisasi.

DAFTAR PUSTAKA

Chen, J., et al. (2021). Interactive test dashboard: A tool supporting reliable and diagnostic computer-based formative assessments. *Computers and Education: Artificial Intelligence, 2*, 100015. https://doi.org/10.1016/j.caeai.2021.100015 (ScienceDirect)

Correnti, R., et al. (2022). Building a validity argument for an automated writing evaluation system designed to provide formative feedback to students. *Computers and Education Open, 3*, 100084. https://doi.org/10.1016/j.caeo.2022.100084

Deeva, E., Bogdanova, D., & Sokolov, H. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education, 163*, 104094. https://doi.org/10.1016/j.compedu.2020.104094 (ScienceDirect)

Fleckenstein, J., et al. (2023). Automated feedback and writing: A multi-level meta-analysis of the effect of automated writing evaluation. *Frontiers in Artificial Intelligence, 6*, 1162454. https://doi.org/10.3389/frai.2023.1162454 (callej.org)

Mao, L., Liu, O., Roohr, K. R., & Belur, V. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment, 23*(2), 121–138. https://doi.org/10.1080/10627197.2018.1427570 (ScienceDirect)

Misgna, G. M., et al. (2024). A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-024-11017-5 (Springer)

Osaka, A., et al. (2025). Cultivating effective feedback with generative AI: Implications for students' trust and learning. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2025.2452005 (Springer)

Pack, A., et al. (2024). Large language models and automated essay scoring: Validity and reliability of

automated scores and feedback. *Computers and Education: Artificial Intelligence, 6*, 100234. https://doi.org/10.1016/j.caeai.2024.100234 (ScienceDirect)

Perikos, I., Grivokostopoulou, F., & Hatzilygeroudis, I. (2017). Assistance and feedback mechanism in an intelligent tutoring system for teaching conversion of natural language into logic. *International Journal of Artificial Intelligence in Education, 27*, 475–514. https://doi.org/10.1007/s40593-017-0139-y (Springer)

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*. https://doi.org/10.1080/01443410.2015.1136407 (Taylor & Francis Online)

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL, 36*(2), 187–209. https://doi.org/10.1017/S0958344023000265 (Cambridge University Press & Assessment)

Tian, L., & Zhou, Y. (2020). Learner engagement with automated feedback, peer feedback and teacher feedback in an online EFL writing context. *System, 91*, 102247. https://doi.org/10.1016/j.system.2020.102247 (ScienceDirect)

Usher, M., et al. (2025). Generative AI and assessment: Comparing AI-generated, instructor, and peer assessments in authentic assessment settings. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2025.2487495 (Taylor & Francis Online)

Venter, M., et al. (2024). Exploring the use of artificial intelligence in the delivery of effective feedback in higher education. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2024.2415649 (Taylor & Francis Online)

Zhai, X., & Ma, X. (2023). Automated writing evaluation and feedback in EFL/ESL writing: A meta-analysis. *Journal of Educational Computing Research, 61*(4), 875–900. https://doi.org/10.1177/07356331221127300 (SAGE Journals)