ISSN 3063-7112
AEJ (Advances in Education Journal)

Peran Dosen Sebagai Korektor dalam Model Human-in-the-Loop (HITL) untuk Meningkatkan Akurasi Evaluasi Pembelajaran Berbasis Artificial Intelligence

Abdul Rahman[1]*
Universitas Lambung Mangkurat
email: abdulrahman@ulm.ac.id

Brezto Asagi Dewantara[2]
Universitas Lambung Mangkurat
email: brezto.tp@ulm.ac.id

*Korespondensi: abdulrahman@ulm.ac.id

History Artikel:
Diterima 20 Agustus 2025
Direvisi 22 Agustus 2025
Diterima 24 Agustus 2025
Tersedia online 26 Agustus 2025

ABSTRAK

Artificial Intelligence (AI)-based learning evaluation is efficient but lacks nuance and risks bias, and the integration of lecturer assessments into the system remains unclear. This study aims to systematically review Human-in-the-Loop (HITL) models to map lecturer roles and measure the impact of their interventions. The study used a literature review of Google Scholar, IEEE, ACM Digital Library, Scopus, dan ERIC databases (2016–2025) with empirical inclusion criteria; 15 studies were analyzed. The results show that while AI improves evaluation efficiency, three lecturer roles initiator, supervisor, and facilitator generally do not directly improve model accuracy. Conversely, the corrector role, which utilizes lecturer feedback for retraining, has the greatest potential for accuracy improvement, but empirical evidence remains limited. Therefore, a shift from simply "Human-in-the-Loop" to a structured feedback mechanism based on Intelligence Augmentation that enables lecturers to contribute to the continuous improvement of Artificial Intelligence models is needed.

Kata kunci:
AI-based learning evaluation, Human-in-the-Loop (HITL), the role of lecturers

Pendahuluan

Evaluasi dalam pembelajaran saat ini sedang mengalami perubahan besar berkat kemajuan teknologi seperti *Artificial Intelligence* (AI). Sistem seperti *Automated Essay Scoring* (AES) dan *Automated Writing Evaluation* (AWE) memungkinkan penilaian tulisan dilakukan secara otomatis. Hal ini membuat proses evaluasi pembelajaran lebih efisien, objektif, dan mampu memberikan umpan balik yang cepat dalam skala besar (Bai et al., 2022; Dikli, 2006). Meski efisiensi meningkat, kualitas evaluasi tidak otomatis terjamin. Agar manfaat AES dan AWE benar-benar bermakna instruksional, *output* sistem perlu disejajarkan dengan konstruk yang diukur, misalnya argumentasi, koherensi, kedalaman isi, dan orisinalitas, melalui kalibrasi dan pengawasan pendidik. Pendekatan *Human-in-the-Loop* dengan rubrik analitik, peninjauan sampel, serta elemen keterjelasan model (*explainability*) penting untuk menjaga validitas, reliabilitas, dan keadilan sekaligus meminimalkan kesalahan sistemik. Sistem ini sering dikritik karena cenderung hanya menilai aspek-aspek dangkal dari sebuah tulisan seperti tata bahasa atau panjang kalimat dan kurang memperhatikan kualitas argumen

dan kreativitas penulis. Selain itu, sistem tersebut juga berisiko mengandung bias algoritmik yang dapat merugikan kelompok siswa tertentu (Correnti et al., 2022; Litman et al., 2021).

Keterbatasan ini telah mendorong perubahan paradigma dari otomatisasi penuh menuju pendekatan *Intelligence Augmentation* (IA). Pendekatan IA, sebagaimana ditegaskan oleh Chris Dede, menekankan pentingnya kemitraan sinergis antara manusia dan mesin. Dalam kerangka ini, teknologi tidak dimaksudkan untuk menggantikan peran manusia, melainkan memperkuat kemampuan penalaran, interpretasi, dan pengambilan keputusan. Dengan kata lain, AI bertugas mengolah informasi secara cepat dan efisien, sementara dosen tetap memegang kendali untuk memastikan penilaian bersifat kontekstual, adil, dan bermakna bagi pembelajaran. Pendekatan ini menjadi fondasi penting bagi integrasi evaluasi berbasis AI yang tidak hanya efisien, tetapi juga sejalan dengan nilai-nilai pedagogis yang berpusat pada siswa (Liang, 2025).

Kerangka kerja ini menegaskan adanya perbedaan mendasar antara dua ranah kemampuan. *Reckoning*, yaitu serangkaian tugas komputasional yang berfokus pada pengolahan data, pengenalan pola, dan perhitungan matematis. Bidang ini merupakan kekuatan utama AI karena sistem mampu mengolah data dalam jumlah besar dengan cepat, presisi, dan konsisten. Di sisi lain terdapat *judgment*, yakni kapasitas khas manusia yang tidak hanya bergantung pada logika, tetapi juga melibatkan pertimbangan etis, pemahaman konteks sosial budaya, serta kepekaan terhadap nuansa dan kompleksitas situasi nyata. *Judgment* menuntut intuisi, pengalaman, serta nilai-nilai kemanusiaan yang tidak dapat direduksi menjadi sekadar angka atau pola. Dengan pembedaan tersebut, dapat dipahami bahwa AI berfungsi sebagai mitra yang unggul dalam aspek *reckoning*, yaitu pengolahan data dan pola secara cepat dan efisien. Sementara itu, *judgment* tetap menjadi ranah utama dosen karena menuntut pertimbangan etis, pemahaman konteks, serta kepekaan terhadap nuansa yang kompleks. Dengan demikian, keberadaan AI bukanlah untuk menggantikan peran dosen, melainkan memperkuat kapasitas mereka dalam menjaga kualitas, keadilan, dan makna proses evaluasi pembelajaran (Dede et al., 2021).

Dalam proses evaluasi, AI berperan mengerjakan pemrosesan awal, sedangkan dosen bertugas menilai kualitas dan konteks. Namun, cara mengintegrasikan penilaian dosen ke dalam sistem AI masih belum terdefinisi dengan jelas. Untuk menganalisis hal ini, peneliti mengidentifikasi dua peran utama dosen dalam model *Human-in-the-Loop* (HITL). Peran pertama adalah sebagai supervisor, yaitu peran pasif di mana dosen hanya memvalidasi atau menolak hasil keluaran AI guna memastikan mutu serta kepatuhan terhadap regulasi, misalnya yang diatur dalam *General Data Protection Regulation* (GDPR) (Colonna, 2024). Intervensi supervisor bersifat lokal, hanya memengaruhi keputusan pada saat itu dan tidak mengubah model AI yang mendasarinya. Sebaliknya, peran kedua adalah korektor, yaitu peran aktif ketika dosen memberikan koreksi dan umpan balik secara sistematis. Informasi dari koreksi tersebut kemudian dimasukkan kembali ke dalam sistem untuk melatih ulang dan menyempurnakan model AI, sehingga akurasi dan kualitas penilaian dapat terus ditingkatkan.

Meskipun peran supervisor sering dibahas dalam berbagai kajian kebijakan, peran korektor justru memiliki potensi besar dan dapat diwujudkan melalui pendekatan seperti *Interactive Machine Learning*. Pendekatan ini memungkinkan AI terus belajar dari masukan manusia, tetapi penerapannya dalam evaluasi pendidikan masih jarang diteliti secara mendalam (Z. Liu et al., 2021; Maadi et al., 2021). Masih terdapat kesenjangan yang cukup besar dalam penelitian empiris terkait hal ini. Hanya sedikit studi dalam bidang evaluasi pendidikan yang secara jelas merancang arsitektur peran korektor, apalagi yang mengukur dampaknya terhadap peningkatan akurasi sistem AI. Oleh karena itu, penelitian ini bertujuan menyajikan tinjauan sistematis tentang model *Human-in-the-Loop* (HITL) dalam evaluasi pembelajaran berbasis *Artificial Intelligence* (AI), dengan fokus pada identifikasi peran dosen serta pengukuran dampak intervensi mereka.

Metode

Penelitian ini menggunakan *literature review*. Proses pencarian literatur dilakukan melalui sejumlah basis data akademik utama, termasuk Google Scholar, IEEE, ACM Digital Library, Scopus, dan ERIC. Kata kunci yang digunakan baik secara terpisah maupun dalam kombinasi meliputi: *'Human-in-the-Loop' AND 'automated evaluation' AND 'education', 'AI-assisted grading' AND 'lecturer role', 'feedback loop' AND 'automated assessment', and 'expert oversight' AND 'automated scoring'*. Pencarian dibatasi pada artikel yang diterbitkan antara tahun 2016 dan 2025, untuk menangkap perkembangan kontemporer di bidang *Artificial Intelligence* dan pendidikan.

Penelitian dimasukkan dalam tinjauan ini jika memenuhi kriteria inklusi berikut: (1) diterbitkan dalam jurnal *peer-review* atau prosiding konferensi; (2) secara eksplisit membahas model atau sistem yang melibatkan intervensi manusia dalam proses evaluasi pembelajaran otomatis; (3) secara spesifik mendeskripsikan peran manusia seperti dosen atau pakar dalam model *human-in-the-loop*; dan (4) ditulis dalam bahasa Inggris. Kriteria eksklusi meliputi: (1) artikel yang hanya membahas AI dalam pendidikan secara umum tanpa berfokus pada aspek evaluasi; (2) artikel yang hanya mengusulkan arsitektur teoretis tanpa implementasi atau evaluasi empiris; dan (3) artikel yang intervensi manusianya terbatas pada penyediaan data pelatihan awal tanpa mekanisme umpan balik yang berkelanjutan.

Peneliti secara independen melakukan penyaringan awal terhadap judul dan abstrak artikel yang teridentifikasi. Artikel yang dianggap berpotensi relevan kemudian ditinjau lebih lanjut melalui tinjauan teks lengkap, berdasarkan kriteria inklusi dan eksklusi yang telah ditentukan sebelumnya. Pada tahap analisis data, peneliti mengidentifikasi dan menganalisis studi yang secara eksplisit mengembangkan dan/atau mengevaluasi model *Human-in-the-Loop* (HITL) dalam konteks mengevaluasi pembelajaran berbasis AI dalam pendidikan.

Hasil

Analisis dari 15 studi yang dipilih mengungkapkan empat tipologi utama model intervensi dosen, dengan perbedaan yang jelas dalam tujuan dan pengukuran dampak.

Tabel 1. Perbandingan Empat Peran Dosen dalam Model *Human-in-the-Loop* (HITL**)**

Fitur Utama	Inisiator	Supervisor	Fasilitator	Korektor
Tugas utama	Berikan data data awal atau *seed* (contoh esai berlabel) pada tahap awal.	Bertindak sebagai peninjau akhir; menyetujui, merevisi, atau menolak keluaran AI.	Membantu mahasiswa memahami dan memanfaatkan umpan balik dari AI.	Menganalisis kelemahan model AI dan merevisi strukturnya dengan cepat.
Arah Interaksi	Satu Arah: Keterlibatan hanya di awal, tanpa umpan balik.	Lingkaran Buntu: Koreksi bersifat individual dan tidak digunakan	Tidak Langsung ke AI: Interaksi difokuskan pada mahasiswa, bukan pada	Loop Tertutup: Umpan balik dari dosen secara eksplisit digunakan untuk menyempurnakan model AI.

		untuk melatih ulang AI.	peningkatan sistem AI.	
Dampak pada Model AI	Terbatas: Hanya memulai model. Peningkatan akurasi setelahnya tidak terukur.	Tidak ada: Tidak berkontribusi pada peningkatan akurasi atau kemampuan sistem AI secara keseluruhan.	Tidak ada: Intervensi dosen tidak digunakan untuk meningkatkan model AI.	Signifikan & Terukur: Meningkatkan akurasi prediktif model (diukur melalui AIC, BIC, RMSE).
Fokus Utama	Inisiasi proses otomatis.	Kontrol kualitas, akuntabilitas, dan kepatuhan hukum/etika.	Aspek pedagogis: meningkatkan kualitas tulisan dan hasil belajar siswa.	Peningkatan teknis berkelanjutan dan akurasi model AI.
Contoh Kerangka Kerja	TGOD (*Transudative Graph-based Ordinal Distillation*)	Model HITL umum untuk pertimbangan hukum dan etika.	Sistem eRevise	*"Menutup Lingkaran"* dalam sistem bimbingan belajar cerdas.

Dosen sebagai Inisiator: Umpan Balik Satu Arah

Dalam model ini, keterlibatan dosen hanya terjadi pada tahap awal proses evaluasi otomatis. Contoh yang jelas dapat dilihat pada kerangka kerja TGOD untuk penilaian esai otomatis / *Automated Essay Scoring* (AES) sekali jalan (Jiang et al., 2021). Pada model ini, dosen berperan sebagai inisiator dengan memberikan satu contoh esai berlabel untuk setiap tingkat skor. Contoh-contoh tersebut berfungsi sebagai data awal atau *seed* yang digunakan untuk memulai pelatihan model. Namun, setelah tahap inisiasi, tidak tersedia mekanisme bagi dosen untuk meninjau, mengawasi, atau mengoreksi proses evaluasi berikutnya. Dengan demikian, peran dosen terbatas hanya sebagai pemberi contoh, bukan sebagai supervisor atau korektor aktif. Karena tidak ada siklus umpan balik berkelanjutan, dampak keterlibatan dosen terhadap peningkatan akurasi model setelah tahap awal tidak dapat diukur maupun divalidasi secara empiris.

Dosen sebagai Supervisor: Menjaga Gerbang Kualitas

Dalam literatur, model *Human-in-the-Loop* (HITL) yang paling sering dianjurkan adalah peran supervisor, terutama karena alasan hukum dan etika (Colonna, 2024). Pada model ini, dosen bertugas sebagai peninjau akhir yang dapat menyetujui, merevisi, atau menolak hasil keluaran AI sebelum disampaikan kepada mahasiswa. Tujuan utama peran ini adalah memastikan akuntabilitas, transparansi, dan kepatuhan terhadap peraturan yang berlaku.

Namun, hasil analisis penelitian menunjukkan bahwa peran supervisor sering kali menjadi lingkaran buntu. Intervensi dosen hanya bersifat satu arah dan terbatas pada kasus tertentu. Misalnya, ketika dosen mengubah skor dari B menjadi A, koreksi tersebut tidak digunakan

kembali untuk melatih ulang atau memperbaiki model AI. Akibatnya, kontribusi manusia dalam model ini tidak berdampak pada peningkatan akurasi sistem secara keseluruhan. Fokus peran ini lebih pada menjaga kualitas keluaran semata, bukan pada peningkatan berkelanjutan dari proses evaluasi itu sendiri.

Dosen sebagai Fasilitator: Menjembatani AI dan Mahasiswa

Dalam model ini, fokus utamanya bukan pada penyempurnaan AI, melainkan pada pendampingan mahasiswa agar dapat memanfaatkan umpan balik AI secara efektif. Salah satu contoh yang menonjol adalah sistem eRevise, yang dirancang untuk memberikan umpan balik formatif mengenai penggunaan bukti dalam tulisan argumentatif siswa sekolah dasar (Correnti et al., 2022; Matsumura et al., 2022).

Pada pendekatan ini, peran dosen secara tegas didefinisikan sebagai fasilitator yang membantu mahasiswa memahami, merefleksikan, dan menindaklanjuti umpan balik yang dihasilkan AI. Studi mengenai eRevise memang mengukur dampak keterlibatan guru secara kuantitatif, tetapi fokus metriknya lebih pada peningkatan kualitas tulisan argumentatif siswa sekolah dasar, bukan pada akurasi atau kemampuan teknis model AI itu sendiri. Tidak ditemukan bukti bahwa intervensi guru digunakan sebagai masukan untuk memperbaiki kinerja AI. Dengan demikian, peran fasilitator terutama berkontribusi pada aspek pedagogis, yaitu mendukung proses belajar mahasiswa, dan tidak secara langsung terkait dengan pengembangan teknologi dari sistem evaluasi otomatis.

Dosen sebagai Korektor: Menutup Lingkaran untuk Peningkatan AI

Model korektor dalam *Human-in-the-Loop* (HITL) yang menjadi inti pertanyaan penelitian ini masih jarang ditemukan dalam literatur evaluasi pendidikan. Analogi yang paling kuat justru datang dari luar ranah penilaian esai, yaitu kerangka kerja "Closing the Loop" untuk meningkatkan model kognitif dalam sistem tutor cerdas (R. Liu & Koedinger, 2017). Dalam kerangka kerja tersebut, para ahli berperan sebagai korektor. Mereka menganalisis keluaran model statistik, menemukan kelemahan dalam model kognitif yang mendasarinya, lalu merevisi strukturnya secara langsung. Dampak intervensi diukur secara eksplisit melalui dua dimensi utama: (1) Peningkatan akurasi prediktif model, diukur dengan metrik statistik seperti AIC, BIC, dan RMSE. Hasilnya menunjukkan bahwa model yang telah direvisi lebih akurat dalam memprediksi kinerja mahasiswa. (2) Peningkatan hasil belajar mahasiswa, divalidasi melalui eksperimen A/B di kelas, yang membuktikan bahwa mahasiswa yang menggunakan tutor dengan model yang ditingkatkan memperoleh capaian belajar lebih baik. Sejauh ini, studi tersebut merupakan satu-satunya contoh implementasi loop umpan balik yang lengkap, di mana koreksi ahli tidak hanya diterapkan, tetapi juga diukur secara ketat dampaknya terhadap akurasi model dan hasil belajar.

Pembahasan

Temuan penelitian ini secara keseluruhan mendukung gagasan utama bahwa terdapat kesenjangan besar antara teori tentang kemitraan antara manusia dengan AI dan kenyataan praktik evaluasi pembelajaran. Peran korektor aktif, yaitu ketika penilaian manusia digunakan secara sistematis untuk meningkatkan perhitungan AI, hingga kini sebagian besar masih bersifat ideal dan belum banyak diterapkan dalam praktik nyata. Kesenjangan ini dapat dijelaskan oleh beberapa faktor utama. Pertama, terdapat tantangan teknis yang signifikan dalam membangun alur pembelajaran berkelanjutan untuk model AI yang kompleks. Kedua, terdapat perbedaan prioritas pedagogis dalam komunitas penelitian. Studi terdahulu seperti sistem eRevise secara sadar lebih berfokus pada dampak formatif pada siswa daripada penyempurnaan teknis model AI. Pendekatan ini berangkat dari pandangan bahwa tujuan

utama sistem *Automated Writing Evaluation* (AWE) adalah untuk melengkapi penilaian manusia, bukan untuk menggantikannya (Matsumura et al., 2022).

Sebagai perbandingan, pada domain berisiko tinggi seperti diagnostik medis, model korektor sudah diterapkan dengan lebih matang. Dalam patologi digital, umpan balik dari pakar secara rutin digunakan untuk menyempurnakan model klasifikasi citra, dan keberhasilannya diukur dengan metrik kinerja yang jelas, seperti akurasi, sensitivitas, dan spesifisitas (Amann et al., 2020; Ribeiro et al., 2016). Salah satu faktor pendorongnya adalah sifat kebenaran dasar yang lebih objektif, misalnya dalam menentukan apakah suatu lesi bersifat ganas atau tidak. Selain itu, tingginya risiko klinis juga mendorong adanya investasi besar dalam pengembangan sistem umpan balik yang ketat dan terukur.

Temuan ini membawa implikasi penting bagi pengembangan teknologi pendidikan. Pendekatan yang digunakan perlu bergeser dari sekadar *Human-in-the-Loop* (HITL) menuju desain yang lebih cermat tentang bagaimana lingkaran tersebut bekerja dan memberi nilai nyata. Dengan demikian, dosen tidak lagi dipandang hanya sebagai penjaga kualitas, tetapi juga sebagai pakar yang secara aktif berkontribusi dalam penyempurnaan model AI. Peran mereka sebagai agen pengetahuan dan evaluasi perlu diintegrasikan secara strategis dan berkelanjutan ke dalam siklus pembelajaran mesin.

Penelitian ini memiliki beberapa keterbatasan yang perlu dicatat. Pertama, jumlah studi yang dianalisis relatif terbatas, hanya 15 artikel, sehingga belum sepenuhnya merepresentasikan variasi implementasi *Human-in-the-Loop* (HITL) di bidang pendidikan. Kedua, fokus penelitian lebih banyak pada evaluasi berbasis tulisan seperti *Automated Essay Scoring* (AES) dan *Automated Writing Evaluation* (AWE), sehingga konteks lain seperti penilaian keterampilan praktik atau proyek masih kurang terwakili. Ketiga, meskipun secara teoritis peran dosen sebagai korektor dinilai paling potensial, bukti empiris yang mendukung dampaknya terhadap akurasi model AI dan hasil belajar mahasiswa masih sangat terbatas, sehingga kesimpulan penelitian lebih bersifat indikatif. Keempat, detail teknis mengenai mekanisme integrasi umpan balik dosen ke dalam pelatihan ulang AI belum banyak dijelaskan dalam literatur, sehingga menyulitkan analisis kelayakan implementasi. Kelima, pencarian literatur yang terbatas pada basis data tertentu berpotensi menimbulkan bias sumber. Terakhir, penelitian ini belum mendalami faktor non-teknis seperti literasi AI dosen, kesiapan institusi, dan aspek kebijakan maupun etika, padahal faktor tersebut sangat berpengaruh terhadap keberhasilan penerapan model HITL dalam evaluasi pembelajaran.

Kesimpulan

Berdasarkan hasil tinjauan sistematis terhadap 15 studi, penelitian ini menyimpulkan bahwa meskipun kecerdasan buatan (AI) terbukti mampu meningkatkan efisiensi evaluasi pembelajaran, kualitas dan akurasi evaluasi tetap sangat bergantung pada peran dosen dalam kerangka *Human-in-the-Loop* (HITL). Empat tipologi peran dosen berhasil diidentifikasi, yaitu inisiator, supervisor, fasilitator, dan korektor. Peran inisiator terbatas pada penyediaan data awal tanpa kontribusi lanjutan terhadap akurasi model. Supervisor berfungsi menjaga mutu dan kepatuhan regulasi, namun intervensinya bersifat lokal dan tidak berdampak pada peningkatan kinerja sistem AI. Fasilitator membantu mahasiswa memahami serta memanfaatkan umpan balik AI, dengan fokus utama pada aspek pedagogis, bukan teknis. Sebaliknya, peran korektor menawarkan potensi terbesar karena memungkinkan umpan balik dosen digunakan secara sistematis untuk melatih ulang dan menyempurnakan model AI, sehingga meningkatkan akurasi prediktif sekaligus berpotensi memperkuat hasil belajar. Sayangnya, bukti empiris mengenai implementasi peran Korektor dalam evaluasi pendidikan masih sangat terbatas, sehingga menimbulkan kesenjangan antara gagasan teoretis tentang kemitraan antara manusia dengan AI dan praktik nyata di lapangan.

Konsekuensi logis dari temuan ini adalah perlunya pengembangan pengetahuan yang lebih berorientasi pada desain arsitektur HITL yang memungkinkan peran Korektor dijalankan

secara efektif. Penelitian lanjutan perlu menguji dampak kuantitatif intervensi dosen dengan metrik yang terukur, sekaligus mengintegrasikan pendekatan Explainable AI (XAI) agar dosen dapat memahami dan mengoreksi model secara lebih bermakna. Dari sisi praksis pendidikan, dosen tidak boleh hanya diposisikan sebagai penjaga kualitas, melainkan sebagai pakar yang secara aktif berkontribusi pada penyempurnaan sistem. Untuk itu, diperlukan mekanisme umpan balik yang terstruktur, berkelanjutan, dan strategis dalam siklus pembelajaran mesin. Selain itu, literasi data dan AI di kalangan dosen harus ditingkatkan agar mereka mampu menjalankan peran ini dengan efektif. Dengan demikian, keseimbangan antara efisiensi teknis AI dan makna pedagogis dapat tercapai, sehingga sistem evaluasi di masa depan benar-benar mendukung peningkatan mutu pendidikan yang adil, kontekstual, dan berkelanjutan.

Referensi

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, VI (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20 (1), 1–9. https://doi.org/10.1186/s12911-020-01332-6

Bai, JYH, Zawacki-Richter, O., Bozkurt, A., Lee, K., Fanguy, M., Cefa Sari, B., & Marin, VI (2022). Automated Essay Scoring (AES) Systems: Opportunities and Challenges for Open and Distance Education. Conference: Tenth Pan-Commonwealth Forum on Open Learning (PCF10), September 1–10. https://doi.org/10.56059/pcf10.8339

Colonna, L. (2024). Teachers in the loop? An analysis of automatic assessment systems under Article 22 GDPR. International Data Privacy Law, 14 (1), 3–18. https://doi.org/10.1093/idpl/ipad024

Correnti, R., Matsumura, L.C., Wang, E.L., Litman, D., & Zhang, H. (2022). Building a validity argument for an automated writing evaluation system (eRevise) as a formative assessment. Computers and Education Open, 3 (February), 1–15. https://doi.org/10.1016/j.caeo.2022.100084

Dede, C., Etemadi, A., & Forshaw, T. (2021). Intelligence augmentation: Upskilling humans to complement AI (The Next Level Lab at the Harvard Graduate School of Education).

Dikli, S. (2006). An overview of automated scoring of essays. Journal of Technology, Learning, and Assessment, 5 (1), 1–35.

Jiang, Z., Liu, M., Yin, Y., Yu, H., Cheng, Z., & Gu, Q. (2021). Learning from graph propagation via ordinal distillation for one-shot automated essay scoring. IW3C2 (International World Wide Web Conference Committee), 2347–2356. https://doi.org/10.1145/3442381.3450017

Kumar, V., & Boulanger, D. (2020). Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value. Frontiers in Education, 5 (October), 1–22. https://doi.org/10.3389/feduc.2020.572367

Liang, L. (2025). Bridging Human Intelligence Augmentation (IA) and Classroom Practices via GenAI in Learning Engineering: A Response to Dr. Dede's Keynote Speeches on IA 2022-2025 Li (Lee) Liang, University of Sydney.

Litman, D., Zhang, H., Correnti, R., Matsumura, L. C., & Wang, E. (2021). A Fairness Evaluation of Automated Methods for Scoring Text Evidence Usage in Writing. In the International Conference on Artificial Intelligence in Education (AIED). Springer International Publishing. https://doi.org/10.1007/978-3-030-78292-4_21

Liu, R., & Koedinger, K. R. (2017). Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains. Journal of Educational Data Mining, 9 (1), 25–41.

Liu, Z., Guo, Y., & Mahmud, J. (2021). When and Why Does a Model Fail? A Human-in-the-Loop Error Detection Framework for Sentiment Analysis. Human Language Technologies, Industry Papers, 170–177. https://doi.org/10.18653/v1/2021.naacl-industry.22

Maadi, M., Khorshidi, H. A., & Aickelin, U. (2021). A review on human–ai interaction in machine learning and insights for medical applications. International Journal of Environmental Research and Public Health, 18 (4), 1–27. https://doi.org/10.3390/ijerph18042121

Matsumura, L.C., Wang, E.L., Correnti, R., & Litman, D. (2022). Designing Automated Writing Evaluation Systems for Ambitious Instruction and Classroom Integration. In AHA Fan Ouyang, Pengcheng Jiao, Bruce M. McLaren (Ed.), Artificial Intelligence in STEM Education: The Paradigmatic Shifts in Research, Education, and Technology (1st Edition, pp. 195–208). CRC Press. https://doi.org/10.1201/9781003181187

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session, 97–101. https://doi.org/10.18653/v1/n16-3020