Quaderni di Comunità

Persone, Educazione e Welfare nella società 5.0

Community Notebook

People, Education, and Welfare in society 5.0

n. 1/2026

POLICIES, PRACTICES AND QUALITY ACROSS EDUCATION, TRAINING AND LABOUR

Edited by
Concetta Fonzo, Laura Evangelista

eurilink
University Press

Iscrizione presso il Registro Stampa del Tribunale di Roma
al n. 172/2021 del 20 ottobre 2021

Eurilink University Press rl
Via Gregorio VII, 601 - 00165 Roma
www.eurilink.it - ufficiostampa@eurilink.it
ISBN: 979 12 82274 12 8
ISSN: 2785-7697 (Print)
ISSN: 3035-2525 (Online)

Prima edizione, giugno 2026
Progetto grafico di Eurilink

Si ringrazia Eleonora Zecca per il contributo all'editing

INDICE

EDITORIALE
Concetta Fonzo, Laura Evangelista 13

RUBRICA *EDUCATION* 21

1. The Involvement of Student Associations in Quality Assurance Mechanisms of Educational Reforms in Italy
Astrid Favella, Emiliane Rubat du Mérac 23
2. Le competenze emergenti in enologia: qualità e coerenza nei percorsi di istruzione e formazione
Paolo Brogioni 33

RUBRICA EMPOWERMENT 43

1. Intelligenza Artificiale: un approccio antropo-centrico, etico, inclusivo
Alessandro Barca, Mariella Tripaldi 45

SAGGI 55

1. Verso un sistema di apprendistato di qualità: standard europei, lavoro dignitoso e governance multilivello. Il caso della Regione Toscana
Miriana Bucalossi 57
2. Valutare la qualità della formazione professionale in Italia: evidenze empiriche e prospettive di policy del quadro EQAVET
Massimiliano Mazzanti, Nicolò Barbieri, Alessandro Montanaro, Laura Evangelista, Concetta Fonzo 85

3. Regulatory Fragmentation and Quality in Training: The Case of the Mediterranean Yachting Sector
Fabio Croci 115
4. Validazione digitalizzata delle competenze nell'apprendimento non formale europeo
Giuseppe Palomba, Enrico Elefante 143
5. The Evolution of Microcredentials within Italy's Continuing Vocational Training System: Regulatory Advances and Social Implications
Alessandra Pedone 171
6. Digital Transformation: Processes, Organisational Models and Osh Training
Sara Stabile, Rosina Bentivenga, Emma Pietrafesa, Edvige Sorrentino, Margherita Bernabei, Silvia Colabianchi, Francesco Costantino 203
7. Il valore euristico di Data, Digital e AI Literacy per la valutazione delle scuole nel Sistema Nazionale di Valutazione
Michela Freddano, Miriam Mariani 239
8. The AI Turn in Higher Education: From Labour Market to Employment Challenges
Danilo Boriati, Mariangela D'Ambrosio 277
9. Ripensare la valutazione con l'Intelligenza Artificiale: qualità, equità e sostenibilità pedagogica nell'istruzione superiore
Francesco Pio Sarcina, Michele Baldassarre 305
10. Tra trasformazioni digitali e capitale relazionale: una lettura sociologica dell'esperienza universitaria per ripensare le politiche del diritto allo studio
Giuseppe Monteduro, Daria Panebianco, Sara Nanetti 337
11. Un approccio basato sui diritti umani per la formazione del servizio sociale. L'esperienza del progetto europeo Fundamental Rights in Daily Actions of Social Workers (FRIDAS) nella coproduzione di strumenti partecipativi
Cecilia de Baggis, Vittoria Grillo, Andrea Bilotti 371

12. Coil In Engineering Educational Activities: Challenges and Opportunities
Néstor Mora Núñez, Juan Carlos Calabria Sarmiento 399

APPROFONDIMENTO 427

Costruire futuro. Un modello di didattica trasformativa per l'orientamento professionale
Domenico Barricelli 429

9. RIPENSARE LA VALUTAZIONE CON L'INTELLIGENZA ARTIFICIALE: QUALITÀ, EQUITÀ E SOSTENIBILITÀ PEDAGOGICA NELL'ISTRUZIONE SUPERIORE

di Francesco Pio Sarcina*, Michele Baldassarre**

Abstract: L'Intelligenza Artificiale sta trasformando la valutazione nell'istruzione superiore, rendendo possibili nuove forme di *scoring*, feedback e *assessment* adattivo, ma riaprendo questioni di validità, equità e responsabilità. Questo studio propone una sintesi narrativa della letteratura internazionale (2020–2025) sull'uso operativo dell'IA nei processi di *assessment* ed *evaluation*. L'analisi, articolata su opportunità, criticità e implicazioni socio-pedagogiche, mostra che i risultati più solidi emergono quando l'IA è vincolata da rubriche, esempi ancora e procedure di confronto, con supervisione umana. Parallelamente, segnala limiti di trasparenza e fairness, rischi per l'integrità e *trade-off* nei sistemi di *proctoring*. Si propone un orientamento che integri in modo responsabile l'IA al fine di rafforzare qualità, equità e sostenibilità della valutazione.

Parole chiave: Valutazione formativa, Intelligenza Artificiale, Istruzione superiore, Qualità dell'istruzione, Equità algoritmica.

* Dipartimento di Scienze della Formazione, Psicologia, Comunicazione, Università degli Studi di Bari "Aldo Moro", francesco.sarcina@uniba.it.
** Dipartimento di Scienze della Formazione, Psicologia, Comunicazione, Università degli Studi di Bari "Aldo Moro", michele.baldassarre@uniba.it.

Sebbene gli autori abbiano condiviso l'intera conduzione della ricerca *ivi* presentata e l'impostazione dell'articolo, si attribuisce a Francesco Pio Sarcina la scrittura dei paragrafi: 1; 1.1; 1.2; 1.3; 2; 3; 4; a Michele Baldassarre il paragrafo: Introduzione.

Abstract: Artificial intelligence is reshaping assessment in higher education, enabling new forms of scoring, feedback, and adaptive assessment while reopening questions of validity, equity, and accountability. This study offers a narrative synthesis of the international literature (2020–2025) on the operational use of AI in assessment and evaluation processes. Organised around opportunities, challenges, and socio-pedagogical implications, the analysis shows that the most robust outcomes emerge when AI is constrained by rubrics, anchored examples, and comparative procedures, under human oversight. At the same time, it highlights transparency and fairness limitations, risks to academic integrity, and trade-offs in AI-based proctoring systems. The paper advances a direction for responsibly integrating AI to strengthen the quality, equity, and sustainability of assessment.

Keywords: Formative assessment, Artificial Intelligence, Higher education, Quality of education, Algorithmic fairness.

Introduzione

La valutazione dell'apprendimento è uno dei momenti della pratica didattica in cui l'Intelligenza Artificiale (IA) trova sempre più spazio. L'utilizzo di questo strumento nei processi di valutazione non è limitato al semplice supporto, ma può influire concretamente su feedback, punteggi, decisioni e criteri. Si tratta quindi di un cambiamento che interroga le fondamenta stesse del giudizio valutativo, richiedendo una riflessione sulla natura dell'evidenza e sulla distribuzione della responsabilità tra attori umani e sistemi artificiali (Crompton e Burke, 2023; Gardner *et al.*, 2021).

L'arrivo dell'IA generativa e dei *Large Language Models* (LLM) ha accelerato due dinamiche che, insieme, mettono sotto pressione i modelli valutativi oggi consolidati. La prima è legata alla produzione degli elaborati: testi e riassunti possono essere generati

con facilità e in forme sempre più convincenti, rendendo fragile l'idea che il prodotto finale sia sempre una traccia affidabile del lavoro dello studente (Scarfe *et al.*, 2024; Lodge *et al.*, 2023). La seconda dinamica, complementare, riguarda l'uso degli stessi modelli per automatizzare parti del giudizio: *scoring*[1] di risposte aperte, generazione di feedback, costruzione di item e supporto alla personalizzazione dell'*assessment*[2] sono applicazioni sempre più esplorate nella letteratura recente (Dai *et al.*, 2024; Latif e Zhai, 2024; Lee *et al.*, 2024).

In questo scenario, la questione non è stabilire se l'IA 'funzioni' in astratto, ma in quali condizioni produca una valutazione di qualità. La letteratura mostra risultati promettenti, soprattutto quando l'uso dei modelli è vincolato da rubriche, àncore, procedure di calibrazione e controlli di coerenza; allo stesso tempo evidenzia che prestazioni e affidabilità variano in modo sensibile a seconda del compito, della disciplina, della lingua e delle scelte di design valutativo (Arslan *et al.*, 2024; Tate *et al.*, 2024). In altre parole, l'IA non sostituisce la progettazione, ma la rende più esigente.

Accanto alla qualità, il tema dell'equità diventa immediatamente concreto. Se la valutazione è mediata da sistemi algoritmici, aumentano i rischi di trattamenti diseguali legati a *bias* linguistici e culturali, ma anche a differenze di accesso agli strumenti e competenze d'uso. Inoltre, le risposte istituzionali orientate alla sorveglianza tecnologica introducono tensioni tra la tutela dell'integrità e la salvaguardia della privacy e del benessere degli

[1] Nel contesto docimologico, si riferisce al processo di assegnazione di un punteggio numerico o di un livello di merito a una prestazione (come un test o un saggio) sulla base di criteri predefiniti.

[2] In questo lavoro, con il termine *assessment* si intende indicare il processo complessivo di raccolta di evidenze sull'apprendimento degli studenti, finalizzato a monitorare i progressi (valutazione formativa) o a certificarne i risultati (valutazione sommativa).

studenti. Le ricerche sulla *fairness*[3] ricordano che l'equità non si riduce a una correzione del modello, ma dipende dall'intero sistema decisionale, nonché da criteri, dati, procedure di revisione e *accountability* (Ferrara, 2024). Anche sul piano delle rappresentazioni sociali, non c'è un consenso automatico: l'accettazione dell'IA in processi di *assessment* dipende da fiducia, rischio percepito, trasparenza e chiarezza delle regole (Kizilcec *et al.*, 2024; Lai *et al.*, 2024).

C'è poi un terzo elemento, spesso meno visibile ma decisivo per i contesti di istruzione superiore: la sostenibilità pedagogica. La valutazione non è solo certificazione, orienta lo studio, regola la motivazione e costruisce senso attraverso feedback e interazione. In questa prospettiva, i benefici dell'IA non coincidono con l'automazione pura, ma con la possibilità di fornire riscontri più frequenti e mirati, senza impoverire la dimensione dialogica e la responsabilità professionale del docente (Cavalcanti *et al.*, 2021; Dai *et al.*, 2024). Allo stesso tempo, i modelli generativi portano rischi specifici, come la variabilità delle risposte, gli errori plausibili, le allucinazioni, che possono compromettere la fiducia nella valutazione se non sono gestiti con pratiche di verifica e supervisione (Huang *et al.*, 2024; Kalai *et al.*, 2025).

Alla luce di queste tensioni pedagogiche, il presente contributo propone una sintesi narrativa della letteratura internazionale nel periodo compreso tra il 2020 e il 2025 per analizzare come l'IA venga integrata operativamente nei processi di *assessment* ed *evaluation*[4] nell'*Higher Education* e quali implicazioni

[3] Termine che indica l'equità e l'imparzialità di un sistema valutativo. Un processo è *fair* quando garantisce che ogni studente sia valutato senza pregiudizi e che i risultati non siano influenzati da variabili esterne al merito (come l'origine culturale o sociale).

[4] A differenza dell'*assessment*, che si concentra sullo studente, l'*evaluation* riguarda solitamente il giudizio di valore su un intero programma formativo, un corso o un'istituzione, per verificarne l'efficacia e la qualità.

ne derivino in termini di qualità, equità e sostenibilità pedagogica. L'analisi esplora le opportunità tecno-pedagogiche dei sistemi adattivi, le criticità metodologiche legate alla validità dei modelli e le implicazioni etiche relative al ruolo del docente. In chiusura, si discute un orientamento critico in cui l'IA viene integrata come strumento di potenziamento sotto vincoli e responsabilità esplicitamente pedagogici (Cardona *et al.*, 2023; UNESCO, 2025).

1. Le tre lenti per leggere l'IA nella valutazione

In ambito universitario valutare non significa semplicemente attribuire un punteggio, ma costruire un'inferenza ragionata sull'apprendimento a partire da evidenze che sono sempre, inevitabilmente, parziali. La valutazione è una sequenza di passaggi: si definisce un obiettivo (che cosa si intende osservare), si progetta un compito (che cosa si chiede), si raccoglie una performance (che cosa lo studente produce), si interpreta quella performance secondo criteri (come si decide che cosa vale) e infine si prende una decisione (che cosa certifico o che cosa restituisco come feedback). In questa prospettiva, tipica della docimologia e della *assesment literacy*, la qualità non dipende solo dalla precisione del voto, ma dalla coerenza dell'intero ragionamento valutativo: se i passaggi intermedi sono deboli o incoerenti, anche uno *scoring* tecnicamente accurato rischia di poggiare su un'interpretazione fragile (Race, 2007; Gardner *et al.*, 2021).

Spostare la discussione su quale anello della catena sia presidiato dall'IA permette di distinguere la fluidità formale del risultato dalla sua adeguatezza docimologica. Per il docente, questo significa riconoscere che l'IA può velocizzare i passaggi operativi ma non assicura automaticamente il significato dell'inferenza, richiedendo il mantenimento del controllo epistemico sulla valutazione.

L'IA può alleggerire il carico operativo, ad esempio nella produzione di feedback preliminari o nella classificazione di risposte, ma richiede al docente di esplicitare meglio ciò che spesso rimane implicito: il costrutto, le soglie, gli indicatori, gli esempi àncora, le eccezioni. In pratica, integrare l'IA non elimina la competenza valutativa: la rende più visibile e, se ben gestita, può persino rafforzarla perché costringe a rendere tracciabile il ragionamento.

Per gli studenti, la stessa cornice porta un altro significato: la valutazione non è solo un esito, ma un messaggio sul valore e una guida implicita a "come si studia" e "che cosa conta". Se l'IA entra nella catena senza trasparenza, lo studente rischia di trovarsi di fronte a criteri percepiti come opachi o instabili, con un impatto diretto su fiducia e *agency*. Se invece i criteri sono esplicitati e l'uso dell'IA è dichiarato (ad esempio, dove produce feedback, dove suggerisce miglioramenti, dove supporta la coerenza della correzione), la valutazione può diventare più leggibile e più negoziabile in senso educativo: lo studente capisce quali evidenze contano, può interpretare il feedback, può migliorare in modo più mirato (González-Calatayud *et al.*, 2021).

Infine, considerare la valutazione come catena aiuta a inquadrare anche un nodo pratico che, con l'IA generativa, diventa più delicato: la natura dell'evidenza. Se una parte della *performance* è co-prodotta con strumenti esterni, l'inferenza sull'apprendimento non può basarsi solo sul prodotto finale; deve spesso spostarsi sul processo, sulla giustificazione, sulla capacità di discutere scelte e limiti, sull'uso critico degli strumenti. Questo non significa rinunciare alla valutazione, ma riposizionarla: chiarire quali forme di supporto sono ammesse e che cosa, in quell'attività, costituisce prova credibile di apprendimento. Da qui discende la logica di fondo che guiderà le sezioni successive: l'integrazione dell'IA può essere sostenibile e persino migliorativa, ma solo se viene collocata consapevolmente dentro i passaggi della valutazione, con criteri e

responsabilità espliciti (Race, 2007).

L'analisi che segue approfondisce questa dinamica attraverso le tre dimensioni complementari della qualità (validità e affidabilità), dell'equità (trasparenza e *accountability*) e della sostenibilità pedagogica delle pratiche.

1.1 La qualità

Quando si discute di qualità della valutazione assistita (Trajkovski e Hayes, 2025) dall'IA, il punto critico è che "qualità" non è un'etichetta unica. In letteratura, almeno tre dimensioni si intrecciano in questa parola, ma non coincidono con essa: validità, affidabilità e utilità didattica. La validità riguarda la forza dell'interpretazione, cioè se l'evidenza raccolta supporta davvero l'inferenza sull'apprendimento che intendiamo fare. L'affidabilità riguarda la coerenza del giudizio, in particolare se i criteri e le procedure producono risultati stabili al variare del valutatore, del momento o di compiti comparabili. L'utilità didattica, infine, riguarda ciò che la valutazione produce nell'ecosistema formativo e, quindi, se genera informazioni che orientano lo studio, rendono visibili gli standard e attivano miglioramento, soprattutto attraverso feedback praticabile (Gardner *et al.*, 2021).

L'IA tende a performare in modo asimmetrico su queste dimensioni: i LLM, ad esempio, possono mostrare un'elevata coerenza formale nello *scoring* (affidabilità) pur restando distanti dal costrutto specifico del corso o della disciplina (validità). Per questo, gli studi che valutano l'impiego dei modelli diventano informativi solo quando esplicitano a quale dimensione di qualità stanno rispondendo e con quali procedure di verifica: rubriche, esempi àncora o analisi di coerenza intra-modello.

Una parte consistente delle ricerche recenti insiste proprio su questi dispositivi di verifica incrociata. I lavori sull'*automatic*

scoring con i LLM mostrano che vincolare il modello con rubriche e criteri espliciti, e testarne l'allineamento rispetto a valutatori umani, può migliorare in modo significativo la coerenza dello *scoring* e la sua spendibilità operativa (Latif e Zhai, 2024; Lee *et al.*, 2024; Tate *et al.*, 2024). Tuttavia, gli stessi studi segnalano che la qualità non è un attributo garantito: cambia con disciplina, tipo di compito, lingua, livello di competenza e sensibilità del modello a formulazioni e contesto. In altri termini, la domanda "può l'IA correggere?" ha poco valore se non viene tradotta in una domanda più precisa: "quali condizioni rendono lo *scoring* sufficientemente valido, affidabile e utile per quel corso e per quegli studenti?" (Arslan *et al.*, 2024).

Parlare di qualità in chiave valutativa implica considerare anche l'impatto sul lavoro docente e sull'esperienza dello studente (Galliani, 2015). Se l'IA viene usata per rendere più frequente il feedback, la qualità non si misura solo nell'accuratezza tecnica del messaggio, ma nella sua coerenza con gli obiettivi del compito e nella capacità di sostenere l'autoregolazione dello studente. E se l'IA viene usata nello *scoring*, la qualità include la possibilità di spiegare il giudizio e renderlo contestabile in modo ragionevole, evitando che la standardizzazione diventi opacità. In questa prospettiva, validità, affidabilità e utilità didattica non sono tre slogan: sono tre requisiti che aiutano a distinguere l'innovazione che migliora la valutazione da quella che la rende solo più rapida.

Se la lente della qualità ci obbliga a distinguere validità, affidabilità e utilità, il feedback rappresenta il mezzo che rende visibili queste dimensioni di qualità nella pratica. L'IA entra in questa fase con la promessa di aumentare frequenza e tempestività del riscontro in contesti ad alta numerosità o in attività iterative, ma alcuni autori fanno notare che l'automazione produce valore solo se il messaggio resta comprensibile, mirato e legato a criteri espliciti (Cavalcanti *et al.*, 2021; Lo *et al.*, 2026). Con i LLM, questa possibilità si estende: non si tratta più soltanto di feedback

standardizzati o su errori tipici, ma di risposte linguisticamente ricche, adattabili e, in teoria, personalizzabili. Proprio per questo, però, diventa necessario chiarire un punto: un feedback ben scritto non è automaticamente un feedback didatticamente efficace.

Le ricerche più recenti che valutano la capacità dei LLM di generare feedback mettono in evidenza questa ambivalenza. Da un lato, i modelli possono produrre suggerimenti articolati e pertinenti, anche su compiti complessi; dall'altro, la qualità varia sensibilmente e dipende dal modo in cui il compito è descritto, dai criteri forniti e dalla presenza di ancoraggi (esempi, rubriche, standard). Quando mancano questi vincoli, il feedback tende a diventare generico, eccessivamente accomodante o, in alcuni casi, fattualmente scorretto: un rischio che, in valutazione, non è neutro perché può indirizzare lo studente verso revisioni inutili o sbagliate. Per questo gli studi di valutazione insistono su procedure di controllo e comparazione sistematica, mostrando che funziona soprattutto ciò che è incastonato in un *design* che rende il feedback verificabile e allineato agli obiettivi (Dai *et al.*, 2024).

Un secondo aspetto, spesso sottovalutato, riguarda l'effetto del feedback sulla responsabilità e sull'autoregolazione. Un feedback formativo non serve soltanto a correggere un errore: serve a sostenere lo studente nel comprendere il criterio e nel ri-orientare le proprie strategie. Se l'IA produce feedback come sostituto totale dell'interazione didattica, il rischio è che lo studente impari a negoziare con il sistema artificiale senza sviluppare davvero consapevolezza del perché certi standard contino. Se invece l'IA è usata come primo livello di riscontro (da discutere, validare e integrare) può favorire un apprendimento più riflessivo, perché rende più frequente l'occasione di confronto con criteri e revisioni, lasciando al docente il ruolo di regia e di garanzia epistemica.

In quest'ottica, l'integrazione dell'IA nel feedback è più promettente quando è pensata come "architettura a livelli": un

riscontro rapido e standardizzato che aumenta la frequenza del dialogo valutativo e una supervisione umana che garantisca validità, equità e pertinenza, nei passaggi che richiedono interpretazioni sottili o alta responsabilità.

1.2 L'equità e la trasparenza

Quando l'IA entra nella valutazione accademica, l'equità non può essere trattata come un requisito aggiuntivo da verificare alla fine. Diventa, piuttosto, una proprietà del sistema socio-tecnico che prende decisioni: non solo il modello, ma anche dati, regole, procedure, contesto d'uso e possibilità di revisione. Questo spostamento è importante perché evita due semplificazioni opposte: da un lato, pensare che basti correggere il *bias* per rendere giusta una valutazione; dall'altro, concludere che ogni automazione sia inevitabilmente iniqua. La letteratura sulla *fairness* mostra che *bias* e disparità possono emergere a più livelli (definizione del costrutto, raccolta dei dati, soglie decisionali, interpretazione dei risultati) e che la mitigazione richiede interventi combinati, non un'unica soluzione tecnica (Ferrara, 2024; Hort *et al.*, 2023; Kheya *et al.*, 2024).

In ambito valutativo, questa complessità assume una forma molto concreta: l'IA può cambiare chi è avvantaggiato e chi è penalizzato da un certo formato di *assessment*. Differenze linguistiche, culturali e disciplinari possono riflettersi nel modo in cui uno studente "viene letto" da un sistema di *scoring* o nel tipo di feedback che riceve. In questo senso, integrare l'IA non è mai una scelta neutra: è una scelta con effetti distributivi che vanno esplicitati e governati.

Un secondo punto riguarda la trasparenza, intesa non come dettaglio tecnico, ma come condizione di legittimità. In valutazione, la trasparenza significa che studenti e docenti possono capire quali

criteri contano, come vengono applicati e in che modo l'IA contribuisce al giudizio o al feedback. Qui la posta in gioco è la fiducia: se i criteri sono percepiti come opachi o automatici, aumenta la distanza tra studente e processo valutativo; se invece l'uso dell'IA è dichiarato, circoscritto e ancorato a rubriche, la valutazione tende a risultare più leggibile e, soprattutto, più contestabile in modo ragionevole. Le ricerche sulle percezioni di docenti e studenti confermano che accettazione e disponibilità ad adottare l'IA dipendono in modo sensibile da fiducia, rischio percepito e chiarezza delle regole d'uso (Kizilcec *et al.*, 2024; Lai *et al.*, 2024).

Una parte della discussione internazionale sull'*assessment reform* suggerisce quindi un cambio di baricentro: non inseguire la tecnologia con nuove barriere, ma ripensare compiti, criteri e modalità di verifica. Ciò significa valorizzare forme di valutazione che rendono più visibile il processo e che chiedono allo studente non solo di produrre un elaborato, ma di argomentare, fare collegamenti, applicarsi e riflettere sui limiti delle proprie decisioni. Spostare il baricentro verso un'integrità costruita nel *design* della prova (Lodge *et al.*, 2023) rende l'*accountability* un tema centrale: se l'IA è ammessa come supporto (ad esempio per *brainstorming*, revisione linguistica o feedback preliminare), serve una cornice che distingua l'uso legittimo dall'alterazione dell'evidenza, definendo *policy* esplicite su cosa resti attribuibile allo studente e cosa all'automazione.

Allo stesso tempo, quando l'IA viene impiegata dall'istituzione o dal docente per supportare lo *scoring*, l'*accountability* riguarda il versante opposto: chi risponde dell'esito, quali controlli sono previsti, quali margini di revisione esistono. Con l'IA l'integrità non viene letta esclusivamente nella dimensione dell'*anti-cheating*, ma è una questione di trasparenza, responsabilità e progettazione valutativa che tenga insieme credibilità del giudizio e sostenibilità

delle pratiche.

Nei processi valutativi l'errore non è solo un dettaglio tecnico: può incidere su esiti, progressioni e percezioni di giustizia. Con i modelli generativi, la criticità più insidiosa non è l'errore evidente, ma l'errore plausibile: risposte fluenti e apparentemente coerenti che contengono imprecisioni, salti logici o informazioni inventate. Le rassegne sulle allucinazioni dei LLM mostrano che si tratta di un fenomeno strutturale, con cause multiple e strategie di mitigazione non risolutive in modo isolato (Huang *et al.*, 2024).

Per la valutazione universitaria questo significa che quando l'IA contribuisce alla generazione di feedback o allo *scoring*, deve essere trattata come una fonte fallibile di evidenza, non come un'autorità. In breve, l'IA può essere utile, ma solo entro un disegno che renda l'incertezza gestibile e l'errore correggibile.

1.3 La sostenibilità pedagogica

La sostenibilità pedagogica è spesso il criterio che decide se l'IA resta un'esperienza interessante "da provare" o se può davvero diventare parte della *routine* valutativa. In ambito accademico, la sostenibilità pedagogica riguarda la capacità di mantenere nel tempo pratiche valutative che, pur integrando l'automazione, non ne restino schiacciate. Se l'IA può attenuare la tensione strutturale tra numeri elevati e tempi di correzione, il rischio è lo spostamento dell'onere didattico: ridurre il costo di produzione dei riscontri potrebbe aumentare la quantità richiesta, e quindi di indebolire il tempo dedicato a criteri, discussione degli standard e cura delle eccezioni.

La sostenibilità, allora, non dipende tanto dall'adozione dello strumento, quanto dalla costruzione di un flusso di lavoro realistico: rubriche esplicite, esempi àncora, punti in cui l'IA può operare in autonomia limitata e punti in cui è prevista una supervisione umana

sistematica, soprattutto per i casi ambigui. Su questi temi la letteratura che si concentra sul ruolo e le funzioni del feedback automatico è utile perché mostra un principio ricorrente: l'automazione produce valore quando è parte di un *design* didattico e non un semplice aggiunta alla correzione (Cavalcanti *et al.*, 2021).

In questa prospettiva, la sostenibilità non dipende solo dall'abilità del singolo docente nel governare gli *output*, ma dalla costruzione di flussi di lavoro realistici e supportati da una chiara *governance* istituzionale. Per gli studenti, ciò significa garantire che il feedback rimanga coerente con gli obiettivi del corso, evitando che l'abbondanza di commenti generati dall'IA si trasformi in rumore valutativo senza reale orientamento. A livello di sistema, le università sono dunque chiamate a gestire la coerenza tra i corsi e a definire linee guida sull'uso legittimo degli strumenti, assicurando la possibilità di spiegare e difendere le decisioni valutative.

2. Una sintesi narrativa della letteratura sull'IA e la valutazione

La scelta di una sintesi narrativa quale strategia di ricerca risponde alla natura eterogenea del campo d'indagine. Gli studi analizzati (2020-2025) variano sensibilmente per tecnologie, contesti disciplinari e disegni metodologici, richiedendo una lettura comparativa capace di integrare evidenze empiriche e indicazioni di *policy*.

La ricerca è stata condotta attraverso interrogazioni mirate su principali database e archivi di riferimento per l'educazione e le scienze sociali, precisamente con Scopus ed ERIC. Le stringhe di ricerca hanno combinato termini relativi a *artificial intelligence*, *generative AI*, *large language models*, *automated scoring*, *learning*

analytics, *proctoring*[5] con termini relativi a *assessment*, *evaluation*, *grading*, *feedback*, *academic integrity*, *higher education*. La selezione ha privilegiato criteri di pertinenza qualitativa e densità informativa rispetto alla mera estensione campionaria, al fine di garantire intercettare contributi direttamente pertinenti ai processi valutativi universitari rispetto agli assi analitici del lavoro.

Sono stati inclusi studi che soddisfacevano criteri di aderenza stringenti: contesto di *Higher Education*; presenza di un impiego di IA connesso in modo diretto alla valutazione dell'apprendimento; impatto operativo su un esito valutativo oppure supporto valutativo immediato e tracciabile. Sono stati esclusi i lavori che trattavano l'IA in educazione in senso generico, gli studi centrati su didattica o *AI literacy* senza un nesso operativo con *assessment* ed *evaluation*, e i contributi focalizzati su strumenti o contesti non universitari. Per mantenere elevata la coerenza con lo scopo del presente studio, sono stati inoltre esclusi gli studi in cui l'IA compariva solo come sfondo teorico o come tecnologia potenziale, senza evidenze d'uso o senza un collegamento chiaro a esiti o pratiche valutative.

Per ciascun contributo incluso è stata effettuata un'estrazione strutturata delle informazioni chiave: tecnologia (es. LLM, ML predittivo, sistemi di *scoring*, *proctoring*), dominio disciplinare e contesto (corso, livello, modalità online o in presenza), tipo di prova (risposta aperta, *essay*, *quiz*, prove autentiche, esami online), ruolo dell'IA nel processo (supporto, co-valutazione, automazione di componenti), *outcome* e metriche riportate (accuratezza e allineamento con umani, qualità del feedback, impatti su apprendimento e percezioni), oltre a rischi dichiarati e strategie di mitigazione (rubriche, controlli di *bias*, supervisione umana, *policy* di

[5] Si tratta di sistemi di monitoraggio automatizzato o assistito dalla tecnologia per la sorveglianza degli esami a distanza, volti a garantire l'integrità della prova e prevenire comportamenti scorretti.

trasparenza).

L'analisi è stata condotta tramite una *thematic analysis* di tipo comparativo, organizzata lungo tre assi interconnessi: opportunità tecno-pedagogiche (automazione e personalizzazione di scoring e feedback, adattività, supporto decisionale); criticità strutturali e metodologiche (validità e affidabilità, *bias*, trasparenza, integrità accademica, *accountability*); implicazioni etiche e socio-pedagogiche (dimensione relazionale della valutazione, ruolo docente, condizioni di equità e partecipazione). Questa struttura ha permesso di leggere in modo coerente studi tra loro diversi, evitando sia una rassegna meramente descrittiva, sia una discussione puramente normativa: le evidenze sono state interpretate alla luce delle tre lenti e messe in dialogo per evidenziare convergenze, divergenze e condizioni di applicabilità nei contesti reali di istruzione superiore.

3. Risultati e discussione per ogni asse

Nel primo asse, le opportunità tecno-pedagogiche diventano credibili soprattutto quando gli studi non si limitano a far correggere delle prove a un LLM, ma lo inseriscono in un dispositivo valutativo con criteri espliciti e confronti controllati (Tabella 1). Ad esempio, Kim (2025) testa ChatGPT-4 in un contesto realistico di *placement test* di inglese accademico, confrontando diverse strategie di *prompting* e verificando sia la coerenza interna (*intra-rater*) sia l'allineamento con punteggi e *placement* umani: il messaggio non è che GPT-4 funzioni in sé, ma che l'affidabilità dipende dalla progettazione docimologica (rubriche, esempi di *scoring*, informazioni linguistiche) e dalla cura del *prompting*. In modo ancora più procedurale, Lan *et al.* (2025) costruiscono un *chatbot* generativo che imita i passaggi che i docenti seguono prima di valutare: lo

applicano a 254 report tecnici di studenti di ingegneria e verificano la relazione tra punteggi del sistema e punteggi dei docenti, anche su dimensioni analitiche (*task fulfillment*, *language*, *organisation*, *formatting*); le correlazioni non risultano uniformi tra dimensioni, e questo è informativo perché indica che l'IA può performare meglio su alcuni aspetti del compito che su altri. Un'ulteriore evidenza operativa è portata dallo studio di Impey *et al.* (2025), che sperimentano GPT-4 per valutare brevi produzioni scritte su temi scientifici in tre MOOC (astronomia, astrobiologia, storia e filosofia dell'astronomia): il sistema riceve rubriche, risposte modello e voti totali forniti dall'istruttore e l'analisi mira a capire se la valutazione automatizzata possa avvicinarsi alla valutazione docente in classi numerose. Qui l'innovazione non è la retorica sull'IA, ma l'idea che un LLM possa sostenere la valutazione della scrittura scientifica dove tradizionalmente si ripiega su quiz perché correggere testi porta via troppo tempo. In modo simile, Gao *et al.* (2024) mettono GPT-4o alla prova su domande concettuali in un corso universitario di ingegneria meccanica (con coorti di circa 225 studenti), confrontandolo con i *teaching assistant* e vincolando entrambi alla stessa rubrica: valutano l'accordo tramite correlazioni di Spearman in impostazioni *zero-shot* e *few-shot*[6], mostrando un potenziale di scalabilità quando la rubrica guida la correzione. Infine, Anghel *et al.* (2025) propongono *CourseEvalAI* per rendere più trasparente e consistente la valutazione di risposte e spiegazioni generate da LLM su contenuti autentici universitari: combinano *fine-tuning* supervisionato[7] (LoRA) con rubriche distinte per tipi di risposta e misure di affidabilità *inter-rater*. Anche contributi più architetturali, come Dimari *et al.* (2024), insistono sul fatto che l'automazione del

[6] Modalità di utilizzo dell'IA senza esempi precedenti (*zero-shot*) o fornendo alcuni esempi guida (*few-shot*) per istruire il modello sul compito richiesto.
[7] Si riferisce al processo più ampio di valutazione e classificazione delle prestazioni degli studenti, che sfocia solitamente nell'attribuzione di un voto o di un giudizio di merito.

grading in prove aperte richiede un *framework* che tenga insieme accuratezza, integrità accademica e implicazioni pedagogiche: qui l'evidenza è meno sperimentale, ma segnala che l'adozione funzionante di sistemi di IA è un progetto socio-tecnico, non un *plug-in*.

Tabella 1: Sintesi delle evidenze per l'asse Opportunità tecno-pedagogiche

Asse 1 - Opportunità tecno-pedagogiche			
Studio	*Contesto e Target*	*Intervento Tecnologico*	*Risultati / Implicazioni*
Anghel et al. (2025)	Esami universitari (Informatica): risposte aperte e spiegazioni.	Ottimizzazione (*Fine-tuning*) di modelli linguistici su rubriche specifiche.	L'uso di dati autentici e procedure di calibrazione aumenta la consistenza e la trasparenza del giudizio.
Kim (2025)	Test di posizionamento (*Placement test*) di inglese accademico.	Uso di GPT-4 per la valutazione automatizzata dei saggi.	L'efficacia dipende dal *design*: rubriche dettagliate e istruzioni (*prompt*) strutturate riducono la variabilità.
Lan et al. (2025)	Ingegneria: valutazione di report tecnici e scrittura disciplinare.	Assistente virtuale progettato per imitare la procedura di correzione dei docenti.	Risultati più solidi su aspetti formali (lingua e struttura) rispetto a criteri meno formalizzabili (contenuto).

Impey et al. (2025)	Corsi online su larga scala (MOOC) in ambito scientifico.	Modelli alimentati con risposte modello e criteri del docente.	Rende possibile fornire feedback su testi complessi in contesti dove solitamente si usano solo test a risposta chiusa.
Gao et al. (2024)	Prove d'esame in Informatica (multimodali e a risposta libera).	Confronto tra esperti umani e intelligenza artificiale come valutatore.	L'automazione richiede protocolli di validazione rigorosi; i compiti d'esame reali restano una sfida per i modelli.
Dimari et al. (2024)	Istruzione Superiore (contesti vari).	Analisi dei sistemi di attribuzione automatica dei voti (*automated grading*).	L'integrazione va letta come un progetto che unisce qualità, integrità e gestione delle regole (*governance*).

Fonte: Ricerca a cura degli autori

Nel secondo asse, le criticità emergono con forza proprio negli studi che misurano validità e affidabilità e nei lavori sull'integrità, perché mostrano dove l'IA rischia di trasformarsi da supporto a vulnerabilità del sistema (Tabella 2). Pecuchova *et al.* (2025), ad esempio, testano diversi modelli (tra cui GPT-4o, Claude3, PaLM2 e SBERT) nel *grading* di risposte aperte in un corso universitario di informatica: raccolgono risposte da 110 studenti su 25 domande e confrontano i modelli con due esperti umani, usando metriche esplicite (*precision*, *recall*, F1, falsi positivi/negativi). Il

risultato più interessante del lavoro evidenzia che gli approcci basati su una risposta di riferimento possono penalizzare risposte che contraddicono il riferimento pur essendo semanticamente corrette; quindi, il problema non è che il modello sbaglia, ma il criterio di valutazione automatizzata può essere troppo rigido rispetto alla varietà epistemicamente legittima delle risposte. In ambito linguistico, Yavuz *et al.* (2025) affrontano direttamente affidabilità e validità nel *grading* di *essay* EFL in *Higher Education* con rubrica analitica: coinvolgono 15 docenti esperti che valutano tre saggi di qualità diversa e riportano livelli alti di accordo per un modello *fine-tuned*; tuttavia, la prospettiva degli autori non è trionfalistica, perché gli stessi insistono sulla necessità di ulteriore *fine-tuning* e di cautela nel passaggio dalla dimostrazione di affidabilità a una validità robusta in contesti diversi. Ghapanchi *et al.* (2023) evidenziano come la fiducia acritica negli strumenti di *detection* sia insufficiente, confermando la necessità di modelli più sofisticati come la *linguistic fingerprint*[8] proposta da Kutbi *et al.* (2024) per rilevare il *contract cheating*[9]. Nella *detection* comportamentale, Trabelsi *et al.* (2023) e gli studi di Tzeng *et al.* (2023; 2024) descrivono sistemi basati su *deep learning* e tracciamento dello sguardo (webcam, *eye-tracking*) per identificare tendenze anomale. Sebbene tali sistemi riportino elevate prestazioni nel riconoscimento di *pattern*, l'implicazione non è demonizzare lo strumento, quanto riconoscere il *trade-off*[10] strutturale tra controllo e diritti (*privacy*,

[8] Letteralmente "impronta linguistica", si riferisce al profilo stilistico unico derivato dall'analisi computazionale della scrittura di un individuo (scelte lessicali, sintassi, punteggiatura). Viene utilizzata come evidenza biometrica testuale per verificare se un elaborato è stato effettivamente scritto dallo studente a cui è attribuito.

[9] Pratica illecita in cui uno studente incarica una terza parte (un'altra persona o un *tool* di IA) di produrre un lavoro accademico al proprio posto. A differenza del plagio classico (copia-incolla), il testo può risultare formalmente originale ai *software* tradizionali, rendendo necessaria l'analisi dell'impronta linguistica per rilevarlo.

[10] In ambito valutativo e tecnologico, indica una situazione di compromesso tra due obiettivi o requisiti contrastanti, in cui il miglioramento di un aspetto comporta

falsi positivi, clima di fiducia) e la necessità di *governance* e contestabilità. In questa scia si colloca anche du Plessis (2025), che discute l'impatto potenziale di ChatGPT sull'*online assessment* evidenziando che gli esempi d'uso reale in prove sono ancora limitati: un richiamo utile per evitare affermazioni assolute e per motivare la tua proposta di riprogettazione come risposta proporzionata, non reattiva.

Tabella 2: Sintesi delle evidenze per l'asse Criticità strutturali e metodologiche

Asse 2 - Criticità strutturali e metodologiche			
Studio	*Contesto e Target*	*Intervento Tecnologico*	*Risultati / Implicazioni*
Pecuchova et al. (2025)	Correzione di risposte aperte in corsi di Informatica.	Confronto tra 11 modelli diversi e valutazione di esperti umani.	I sistemi basati su risposte di riferimento possono penalizzare risposte corrette ma originali; serve flessibilità interpretativa.
Yavuz et al. (2025)	Scrittura in lingua straniera (EFL) nell'istruzione superiore.	Valutazione della coerenza del giudizio tramite rubriche analitiche.	La stabilità del voto non garantisce la validità: è necessaria la supervisione umana sui criteri interpretativi.
Ghapanchi et al. (2023)	Integrità accademica e originalità degli elaborati.	Test di saggi generati dall'intelligenza artificiale tramite *software* antiplagio (Turnitin).	Gli strumenti di controllo non sono sufficienti: serve ripensare il design dei compiti e le politiche di trasparenza.

necessariamente il peggioramento di un altro.

Kutbi et al. (2024)	Identificazione di testi prodotti da terzi (*contract cheating*).	Analisi dell'impronta linguistica tramite apprendimento automatico.	Lo strumento può supportare l'indagine ma presenta rischi di falsi positivi; l'uso non deve essere mai totalmente automatico.
Trabelsi et al. (2023)	Sistemi di monitoraggio per esami online (*proctoring*).	Utilizzo di algoritmi di riconoscimento immagini per rilevare comportamenti anomali.	Forte tensione tra controllo e *privacy*; rischio di sanzioni ingiuste senza una supervisione umana e regole chiare.
Tzeng et al. (2023/24)	Esami a distanza nell'istruzione superiore.	Monitoraggio dello sguardo e dei movimenti tramite *webcam* e intelligenza artificiale.	Richiede estrema trasparenza e gestione dei falsi positivi per non compromettere il clima di fiducia.
du Plessis (2025)	Valutazione online in contesti universitari.	Discussione teorica su opportunità e rischi (equità e originalità).	Necessità di passare da una logica difensiva a una riprogettazione che valorizzi il processo di apprendimento.

Fonte: Ricerca a cura degli autori

Nel terzo asse, le implicazioni etiche e socio-pedagogiche si chiariscono quando la letteratura porta l'attenzione su come cambia la relazione valutativa e su quali competenze servono per non perdere la dimensione educativa (Tabella 3). Mithi *et al.* (2024), con interviste a docenti e studenti in un'università sudafricana, non misurano le *performance* del modello, ma ricostruiscono esperienze e percezioni sull'uso dell'IA generativa in compiti di valutazione formativa: l'integrazione non è solo questione tecnica, ma si intreccia

con aspettative, contesto e pratiche reali. Eager e Brunton (2023) adottano un registro più propositivo: discutono *affordance* e criticità dei LLM e offrono indicazioni per scrivere *prompt* istruzionali, includendo anche un caso per illustrare l'uso dell'IA nella progettazione dell'*assessment*; l'IA viene presentata come uno strumento tra altri, e l'accento cade sullo sviluppo professionale docente. Fiedler (2023), infine, porta una cornice dialogica: se l'IA rende più facile produrre *output* indistinguibili, la valutazione deve spostarsi progressivamente dal "saper rispondere" al "saper chiedere, argomentare e valutare risposte", recuperando pratiche di insegnamento dialogico e rendendo la valutazione più legata al ragionamento e meno al prodotto isolato.

Tabella 3: Sintesi delle evidenze per l'asse Implicazioni etiche e socio-pedagogiche

Asse 3 - Implicazioni etiche e socio-pedagogiche			
Studio	*Contesto e Target*	*Intervento Tecnologico*	*Risultati / Implicazioni*
Lee et al. (2024)	Studio su feedback incentrato sull'uomo (*human-centric*).	Feedback formativo in corsi universitari numericamente numerosi.	L'IA agisce come infrastruttura che rende il feedback sostenibile mantenendo il controllo in mano al docente.
Mithi et al. (2024)	Studio qualitativo tramite interviste a docenti e studenti.	Esperienze e percezioni sull'uso dell'IA nella valutazione formativa.	L'integrazione è un fatto sociale: servono nuove strategie di progettazione per preservare l'autonomia dello studente.
Eager e Brunton (2023)	Commentario e analisi delle potenzialità dei modelli linguistici.	Sviluppo delle competenze dei docenti e design della valutazione.	L'IA è uno strumento tra molti; il suo valore dipende dall'integrazione con competenze

			pedagogiche e politiche chiare.
Fiedler (2023)	Proposta di una cornice pedagogica dialogica.	Ripensamento del paradigma valutativo nell'era dell'IA generativa.	Spostare il focus dal "saper rispondere" al "saper argomentare e valutare"; rafforza il ruolo del docente come garante di senso.

Fonte: Ricerca a cura degli autori

Conclusioni e prospettive future

La sintesi narrativa converge su un punto netto: l'IA può rafforzare la valutazione universitaria non perché decide meglio, ma perché, se ben progettata, può rendere più praticabili coerenza, tempestività e personalizzazione in passaggi che oggi sono spesso fragili o insostenibili. Le evidenze su *automated scoring* e *grading* mostrano che i risultati più promettenti emergono quando l'IA è vincolata da rubriche esplicite, esempi àncora e procedure di confronto con valutatori umani, e quando viene usata per supportare (non sostituire) il processo valutativo: in questi casi aumenta la consistenza e diventa più difendibile sul piano docimologico (Gao *et al.*, 2024; Anghel *et al.*, 2025; Impey *et al.*, 2025; Kim, 2025; Lan *et al.*, 2025). In parallelo, l'uso dei LLM per il feedback può rendere più continua la regolazione formativa, ma solo se il riscontro resta ancorato a criteri e obiettivi; altrimenti la maggiore produzione rischia di trasformarsi in rumore valutativo, senza reale guadagno per l'apprendimento.

Allo stesso tempo, la letteratura segnala che l'integrazione non è mai neutra: la qualità tecnica non elimina le questioni di equità, trasparenza e responsabilità. Studi su *open-ended grading* mettono in evidenza limiti strutturali di approcci *reference-based* e

la necessità di controllare falsi positivi/negativi e rigidità interpretative (Pecuchova *et al.*, 2025), mentre i lavori sull'integrità accademica e sulle tecnologie di *detection* o *proctoring* mostrano un *trade-off* reale tra controllo, *privacy* e fiducia, con il rischio di soluzioni sproporzionate se non accompagnate da *governance* e contestabilità (Ghapanchi *et al.*, 2023; Trabelsi *et al.*, 2023; Tzeng *et al.*, 2023; Kutbi *et al.*, 2024; Tzeng *et al.*, 2024). Da qui discende una conclusione operativa: il criterio di adozione non può essere solo "funziona", ma "funziona in modo giustificabile", con criteri espliciti, comunicazione agli studenti, tracciabilità delle decisioni, possibilità di revisione e supervisione umana nei passaggi più delicati della valutazione. Senza una *governance* consapevole, il rischio è quello di un'adozione "a macchia di leopardo" che produce profonde disomogeneità: studenti valutati con standard percepiti come diversi non per scelta pedagogica, ma per differenze di *tool* e pratiche tra i singoli corsi. In questo scenario, la rapidità offerta dai nuovi strumenti andrebbe a discapito della significatività, offrendo una valutazione sicuramente più veloce, ma meno capace di sostenere l'apprendimento (Dai *et al.*, 2024).

Al fine di tradurre tali evidenze in un orientamento operativo per decisori e docenti, è necessario sistematizzare l'integrazione dell'IA attraverso principi guida che ne presidino la sostenibilità pedagogica e l'equità algoritmica. In primo luogo, le istituzioni accademiche dovrebbero promuovere protocolli di trasparenza che garantiscano il "diritto alla spiegazione" e alla contestabilità del giudizio, assicurando che la valutazione automatizzata rimanga un supporto analitico e mai un'autorità decisionale priva di supervisione umana. Parallelamente, il *design* valutativo deve evolvere verso modelli orientati al processo, in cui l'oggetto della valutazione non sia il prodotto isolato (facilmente delegabile alla macchina), bensì la capacità dello studente di interagire criticamente con l'IA, giustificando le scelte compiute e riflettendo

sui limiti degli *output* generati. Tale approccio richiede una *governance* che valorizzi il principio del *human-in-the-loop*, dove l'automazione viene impiegata per scalare la frequenza del feedback formativo senza tuttavia esautorare il docente dalla propria responsabilità epistemica e deontologica nella sintesi valutativa finale. Infine, la sostenibilità di questa prospettiva dipende in buona parte da un investimento sistemico nella formazione continua e nella creazione di comunità di pratica, volte a ridurre le asimmetrie nelle competenze d'uso e a garantire che l'innovazione tecnologica sia sempre vincolata da finalità pedagogiche esplicite e condivise.

Per il futuro, l'orientamento più auspicabile è quello di usare l'IA per aumentare qualità del processo valutativo, mantenendo al docente il controllo sui nuclei ad alta responsabilità (definizione del costrutto, standard, interpretazione dei casi ambigui, decisione finale) e costruendo per gli studenti condizioni di comprensibilità e partecipazione. Le traiettorie di ricerca futura dovrebbero dunque concentrarsi sulla validazione longitudinale di questi sistemi, indagando come la mediazione algoritmica modifichi nel tempo l'*agency* dello studente e la professionalità del docente, al fine di garantire che l'università resti un luogo di valutazione equa, rigida nel metodo ma flessibile nel dialogo educativo, pur operando in un ecosistema inevitabilmente ibrido.

Bibliografia

Anghel, C., Crăciun, M. V., Pecheanu, E., Cocu, A., Anghel, A. A., Iacobescu, P., Maier, C., Andrei, C. A., Scheau, C., & Dragosloveanu, S. (2025). CourseEvalAI: Rubric-Guided Framework for Transparent and Consistent Evaluation of Large Language Models. *Computers*, 14(10). https://doi.org/10.3390/computers14100431.

Arslan, B., Lehman, B., Tenison, C., Sparks, J. R., López, A. A., Gu, L., & Zapata-Rivera, D. (2024). Opportunities and challenges of using generative AI to personalize educational assessment. *Frontiers in Artificial Intelligence*, 7. https://doi.org/10.3389/frai.2024.1460651.

Cardona, M. A., Rodríguez, R. J., & Ishmael, K. (2023). *Artificial Intelligence and the Future of Teaching and Learning. Insights and Recommendations Artificial Intelligence and the Future of Teaching and Learning*. https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf.

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y. S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2. Elsevier B.V. https://doi.org/10.1016/j.caeai.2021.100027.

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, 20(22). https://doi.org/10.1186/s41239-023-00392-8.

Dai, W., Tsai, Y. S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, 7. https://doi.org/10.1016/j.caeai.2024.100299.

Dimari, A., Tyagi, N., Davanageri, M., Kukreti, R., Yadav, R., & Dimari, H. (2024). AI-Based Automated Grading Systems for open book examination system: Implications for Assessment in Higher

Education. *International Conference on Knowledge Engineering and Communication Systems Proceedings (ICKECS)*. https://doi.org/10.1109/ICKECS61492.2024.10616490.

Eager, B., & Brunton, R. (2023). Prompting Higher Education Towards AI-Augmented Teaching and Learning Practice. *Journal of University Teaching and Learning Practice*, 20(5). https://doi.org/10.53761/1.20.5.02.

Ferrara, E. (2024). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, 6(3). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/sci6010003.

Fiedler, K. D. (2023). Teaching Students to Question: Dialogic Teaching in the Age of Artificial Intelligence. *29th Annual Americas Conference on Information Systems*. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85192887242&partnerID=40&md5=6188f3ae405ee9d6a085766089894d0c.

Galliani, L. (a cura di). (2015). L'agire valutativo. Manuale per docenti e formatori. Editrice La Scuola.

Gao, R., Guo, X., Li, X., Lekshmi-Narayanan, A. B. L., Thomas, N., & Srinivasa, A. R. (2024). Towards Scalable Automated Grading: Leveraging Large Language Models for Conceptual Question Evaluation in Engineering. *Proceedings of Machine Learning Research*, 264, 186–206. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85219516480&partnerID=40&md5=751d57872403a6ea7f4170293e19d44b.

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in

educational assessment: 'Breakthrough? Or buncombe and ballyhoo?' *Journal of Computer Assisted Learning*, 37(5). John Wiley and Sons Inc. https://doi.org/10.1111/jcal.12577.

Ghapanchi, A. H., Ghanbarzadeh, R., & Purarjomandlangrudi, A. (2023). An Initial Investigation on Originality of Text Generated by Generative AIs Like ChatGPT. *Proceedings of the Information Systems Education Conference, ISECON*. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85173038292&partnerID=40&md5=5354611b58d8683eb961a163e99c3bce.

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Review Artificial Intelligence for Student Assessment: A Systematic Review. *Applied Sciences*, 11(5467). https://doi.org/10.3390/app11125467.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2023). Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *arXiv*. http://arxiv.org/abs/2207.07068.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2024). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 1(1). https://doi.org/10.1145/3703155.

Impey, C., Wenger, M., Garuda, N., Golchin, S., & Stamer, S. (2025). Using Large Language Models for Automated Grading of Student Writing about Science. *International Journal of Artificial Intelligence in Education*, 35(4), 1825–1859. https://doi.org/10.1007/s40593-024-00453-7.

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why Language Models Hallucinate. *arXiv*. http://arxiv.org/abs/2509.04664.

Kheya, T. A., Bouadjenek, M. R., & Aryal, S. (2024). The Pursuit of Fairness in Artificial Intelligence Models: A Survey. *arXiv*. http://arxiv.org/abs/2403.17333.

Kim, Y. (2025). Automated Essay Scoring With GPT-4 for a Local Placement Test: Investigating Prompting Strategies, Intra-Rater Reliability, and Alignment with Human Scores. *TESOL Quarterly*, 59(S1), S318–S329. https://doi.org/10.1002/tesq.3405.

Kizilcec, R. F., Huber, E., Papanastasiou, E. C., Cram, A., Makridis, C. A., Smolansky, A., Zeivots, S., & Raduescu, C. (2024). Perceived impact of generative AI on assessments: Comparing educator and student perspectives in Australia, Cyprus, and the United States. *Computers and Education: Artificial Intelligence*, 7. https://doi.org/10.1016/j.caeai.2024.100269.

Kooli, C., Yusuf, N., & Sarhan, M. Y. (2026). Colloquial engagement theory with AI awareness (CET-AIA): A new creative pedagogical framework for ethical assessment in the age of artificial intelligence. *Thinking Skills and Creativity*, 60. https://doi.org/10.1016/j.tsc.2025.102051.

Kutbi, M., Al-Hoorie, A. H., & Al-Shammari, A. H. (2024). Detecting contract cheating through linguistic fingerprint. *Humanities and Social Sciences Communications*, 11(1). https://doi.org/10.1057/s41599-024-03160-9.

Lai, C. Y., Cheung, K. Y., Chan, C. S., & Law, K. K. (2024).

Integrating the adapted UTAUT model with moral obligation, trust and perceived risk to predict ChatGPT adoption for assessment support: A survey with students. *Computers and Education: Artificial Intelligence*, 6. https://doi.org/10.1016/j.caeai.2024.100246.

Lan, G., Li, Y., Yang, J., & He, X. (2025). Investigating a customized generative AI chatbot for automated essay scoring in a disciplinary writing task. *Assessing Writing*, 66. https://doi.org/10.1016/j.asw.2025.100959.

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, 6. https://doi.org/10.1016/j.caeai.2024.100210.

Lee, G. G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6. https://doi.org/10.1016/j.caeai.2024.100213.

Lo, J., Wong, C., Ng, A., Wong, P., Cheung, D., & Lai, P. (2026). Stretching AI's reach: Assessing an AI-driven feedback system for extended academic writing. *Computers and Education: Artificial Intelligence*, 10. https://doi.org/10.1016/j.caeai.2025.100511.

Lodge, J. M., Howard, S., & Bearman, M. (2023). *Assessment reform for the age of artificial intelligence*. Australian Government Tertiary Education Quality and Standards Agency. https://www.teqsa.gov.au/sites/default/files/2023-09/assessment-reform-age-artificial-intelligence-discussion-paper.pdf.

Mills, K., Ruiz, P., Lee, K.-W., Coenraad, M., Fusco, J., Roschelle, J., & Weisgrau, J. (2024). *AI Literacy: A Framework to Understand,*

Evaluate and Use Emerging Technology. Digital Promise. https://files.eric.ed.gov/fulltext/ED671235.pdf.

Mithi, J., Madzvamuse, S., Mbanje, S., & Lomahoza, S. (2024). Generative Artificial Intelligence and Formative Assessment: Perspectives from Higher Education in South Africa. *Proceedings of the International Conference on Education Research*, 449–458. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85216081927&partnerID=40&md5=0a4e87aa3a16b0648ab378aaea47eac9.

Pecuchova, J., Benko, Ľ., & Drlík, M. (2025). Automated Grading of Open-Ended Questions in Higher Education Using GenAI Models. *International Journal of Artificial Intelligence in Education*, 35(6), 3813–3846. https://doi.org/10.1007/s40593-025-00517-2.

Race, P. (2007). *The Lecturer's Toolkit. A Practical Guide to Assessment, Learning and Teaching*. Routledge.

Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024). A real-world test of artificial intelligence infiltration of a university examinations system: A "Turing Test" case study. *PLoS ONE*, 19(6). https://doi.org/10.1371/journal.pone.0305354.

Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence*, 7. https://doi.org/10.1016/j.caeai.2024.100255.

Trabelsi, Z., Ambali Parambil, M. M. A., Alnajjar, F., & Ali, L. (2023). Behavioral-based Real-Time Cheating Detection in Academic Exams using Deep Learning Techniques. *ETLTC-ICETM2023 International*

Conference Proceedings, 2909(1). https://doi.org/10.1063/5.0181921.

Trajkovski, G., & Hayes, H. (2025). *AI-Assisted Assessment in Education Transforming Assessment and Measuring Learnin*. Palgrave MacMillan.

Tzeng, J.-W., & Zhuang, Z.-X. (2024). Exploring the Relationship Between Learning Achievement of Remote Exam Student-Problem Chart Integrated with Convolutional Neural Network and Webcam Eye-tracking Trajectories. *5th International Conference on Control, Robotics, and Intelligent System Proceedings*, 13404. https://doi.org/10.1117/12.3050116.

Tzeng, J.-W., Hsueh, C.-Y., Lee, C.-A., & Shih, W.-Y. (2023). Identifying the Correlation Between Online Exam Answer Trajectory and Test Behavior Based on Artificial Intelligence and Eye Movement Detection Technology. *2023 International Conference on Consumer Electronics Proceedings*, 503–504. https://doi.org/10.1109/ICCE-Taiwan58799.2023.10226745.

UNESCO (2025). *AI and the future of education: Disruptions, dilemmas and directions*. https://doi.org/10.54675/keck1261.