Transparent Algorithmic Integration in the *Systematic Review* (IATSR): A Methodological Proposal for Educational Research in the Age of Artificial Intelligence

Integrazione Algoritmica Trasparente nella *Systematic Review* (IATSR): Una proposta metodologica per la ricerca pedagogica nell'epoca dell'intelligenza artificiale

Maurizio Gentile
Dipartimento di Scienze Umane; Università LUMSA, Roma (Italy); m.gentile@lumsa.it – https://orcid.org/0000-0002-2410-8353
Piergiuseppe Ellerani
Dipartimento di Scienze dell'Educazione; Alma Mater Studiorum Università di Bologna (Italy); piergiusepp.ellerani@unibo.it
https://orcid.org/0000-0002-4730-3928

OPEN ACCESS

DOUBLE BLIND PEER REVIEW

ABSTRACT

This article proposes, describes and argues for a methodological structure (IATSR, Transparent Algorithmic Integration in the Systematic Review) for conducting systematic reviews assisted by algorithmic tools in the non-bibliometric fields of educational research. The proposal stems from a problem now felt in editorial practice and documented in the international literature: the spread of undeclared use of so-called generative artificial intelligence in scholarly production, with effects on the quality, verifiability and equity of educational research. We argue that algorithmic integration, if made constitutive and declared within a protocol built on PRISMA 2020, can be carried out rigorously, transparently and verifiably. The workflow unfolds in seven phases, adopts the items already present in PRISMA-trAIce, and treats the human-in-the-loop principle as unavoidable. The contribution delivers five theoretical-methodological products – the formalisation of the workflow in seven phases, the R1/R2 classification of sources, the epistemological risk matrix, the mapping onto the PRISMA 2020 checklist items and the adapted flow diagram – which remain usable independently of the full adoption of the workflow. The proposal is accompanied by explicit qualitative validity judgements stated as expert *a priori* estimates rather than the outcome of a pilot study or inter-rates validation: the empirical testing of the protocol constitutes the main line of future research. The contribution thus positions itself as a contribution to the methodology of educational research, to research training and to the debate on the transparent assessment of scholarly production.

Il contributo propone, descrive e argomenta una struttura metodologica (IATSR, Integrazione Algoritmica Trasparente nella *Systematic Review*) per condurre revisioni sistematiche assistite da strumenti algoritmici nei campi non-bibliometrici afferenti ai settori PAED-01 e PAED-02. La proposta muove da un problema ormai avvertito nella prassi editoriale e documentato nella letteratura internazionale: la diffusione di un uso non dichiarato della cosiddetta intelligenza artificiale generativa nella produzione scientifica, con effetti sulla qualità, sulla verificabilità e sull'equità della ricerca educativa. Si argomenta che l'integrazione algoritmica, se resa costitutiva e dichiarata del protocollo entro il modello PRISMA 2020, possa essere condotta in modo rigoroso, trasparente e verificabile. Il workflow si articola in sette fasi, assume gli elementi già presenti in PRISMA-trAIce (Holst et al., 2025) e considera l'ineludibilità del principio *human-in-the-loop*. Il contributo consegna cinque prodotti teorico-metodologici – la formalizzazione del workflow in sette fasi, la classificazione R1/R2 delle fonti, la matrice di rischio epistemologico, la mappatura sulle voci PRISMA 2020 e il diagramma di flusso adattato – riutilizzabili anche indipendentemente dall'adozione integrale del workflow. L'ipotesi è accompagnata dall'esplicitazione di giudizi qualitativi di validità formulati *a priori* e come stime esperte, e non l'esito di una prova pilota o di una validazione della concordanza tra valutatori: la messa alla prova empirica del protocollo costituisce la principale linea di ricerca futura. Il contributo si colloca quindi come apporto alla metodologia della ricerca pedagogica, alla formazione alla ricerca e al dibattito sulla valutazione trasparente della produzione scientifica in PAED.

KEYWORDS
Systematic review, PRISMA 2020, PRISMA-trAIce, Artificial Intelligence, Human-in-the-loop, Transparency, Reproducibility, Educational research, Research training
Revisione sistematica, PRISMA 2020, PRISMA-trAIce, Intelligenza Artificiale, *Human-in-the-loop*, Trasparenza, Riproducibilità, Ricerca pedagogica, Formazione alla ricerca

Citation: Gentile, M., & Ellerani, P. (2026). Transparent Algorithmic Integration in the *Systematic Review* (IATSR): A Methodological Proposal for Educational Research in the Age of Artificial Intelligence. *Formazione & insegnamento, 24*(2), 23-35. https://doi.org/10.7346/-fei-XXIV-02-26_03

Conflicts of interest: The Authors declare no conflicts of interest.
Acknowledgments: The authors received no specific funding for this work.
Authorship: Both Authors have contributed equally to the manuscript. Conceptualization (M. Gentile, P. Ellerani); Methodology (M. Gentile, P. Ellerani); Writing – first draft (M. Gentile, P. Ellerani); Writing – review and editing (M. Gentile, P. Ellerani). Both authors read and approved the final version.

DOI: https://doi.org/10.7346/-fei-XXIV-02-26_03

Submitted: May 21, 2026 • **Accepted**: July 14, 2026 • **Published on-line**: July 20, 2026

Pensa MultiMedia: ISSN 2279-7505 (online)

1. Introduzione

1.1. Il problema: l'uso non dichiarato dell'IA nella produzione scientifica

L'ipotesi di fondo della proposta muove da una condizione definibile come "potenziale problema" ormai avvertita nelle pubblicazioni a carattere scientifico e documentata nella letteratura internazionale.

Una prima considerazione annota come nel panorama della ricerca accademica in ambito pedagogico e delle scienze dell'educazione si registri, a partire dal 2023, una proliferazione significativa di contributi nei quali l'apporto di sistemi di intelligenza artificiale generativa rimane sistematicamente non dichiarato. Questo fenomeno produce distorsioni epistemologiche: da un lato introduce nelle produzioni scientifiche bias computazionali e allucinazioni fattuali (tra cui riferimenti bibliografici inesistenti, dati empirici non verificabili, argomentazioni circolari non rilevate); dall'altro genera una forma di competizione asimmetrica tra ricercatori, laddove chi utilizza algoritmi senza trasparenza accumula vantaggi in termini di produttività senza assumere i corrispondenti oneri metodologici. Nelle revisioni di riviste nazionali di Classe A si segnalano, con crescente frequenza, ricezioni caratterizzate da bibliografie contenenti riferimenti non verificabili o privi di DOI reale, sezioni di rassegna "*state of the art*" con una struttura argomentativa interna incoerente, utilizzo di terminologia specifica decontestualizzata e una certa uniformità stilistica presumibilmente incompatibile con la complessità tematica spesso dichiarata. Questi "indicatori", pur non potendo essere verificati con certezza dai revisori, sono fortemente descrittivi di un utilizzo algoritmico non supervisionato e non dichiarato.

Una seconda considerazione introduce elementi di evidenza a supporto delle esperienze emerse nella prima considerazione. Kobak et al. (2025) hanno stimato – da un'analisi su più di quindici milioni di abstract biomedici indicizzati in PubMed, attraverso l'uso di marcatori lessicali comparsi dopo la diffusione di ChatGPT – che almeno il 13,5% degli articoli pubblicati nel 2024 sia stato prodotto con un qualche grado di intervento di un modello linguistico di grandi dimensioni, con punte attorno al 40% in alcuni sottocorpora disciplinari. La ricerca di Gray (2024) converge, pur utilizzando metodi di analisi differenti, sulla definizione di un fenomeno in rapida crescita e fortemente variabile per disciplina, area geografica e sede editoriale.

Il punto critico non è l'uso in sé dei sistemi algoritmici, bensì la *non-dichiarazione*. Iniziative di documentazione come Academ-AI (Glynn, 2024) hanno raccolto centinaia di casi di testo verosimilmente generato da algoritmi e non segnalato nei lavori scientifici pubblicati, riconoscibili da "impronte" caratteristiche o da frammenti di interfaccia rimasti nel testo. Al riguardo, Glynn (2024) annota come vi sia crescente consapevolezza che i casi documentati rappresentino probabilmente solo una frazione del fenomeno realmente esistente, gran parte del quale resta non rilevabile. Ne conseguono due ordini di possibili distorsioni:

– L'introduzione, nelle produzioni scientifiche, di *allucinazioni fattuali*. Il caso più documentato è quello dei riferimenti bibliografici inesistenti: un'analisi condotta su 2,5 milioni di articoli biomedici ha rilevato una progressione netta nella frequenza di citazioni "fabbricate" dagli algoritmi, da circa 1 articolo su 2.828 nel 2023 a 1 articolo su 458 nel 2025, fino a 1 articolo su 277 nelle prime settimane del 2026 (Topaz et al., 2026). Alcuni studi sperimentali condotti da Linardon et al. (2025) confermano la portata del rischio a monte: in sei rassegne generate con GPT-4o, il 19,9% delle citazioni è risultato interamente "fabbricato" e il 45,4% di quelle reali conteneva errori bibliografici, spesso DOI errati o non validi, con tassi più elevati sui temi meno coperti dalla letteratura;
– L'evoluzione più carsica, ma potenzialmente esponenziale e con effetti reali sulla politica accademica, della natura competitiva nella produzione quantitativa a scapito dell'equità: chi impiega strumenti algoritmici senza trasparenza accumula vantaggi di produttività – bibliometrica e non-bibliometrica – senza attribuzione, introducendo un'asimmetria metodologica a danno di chi rinuncia a tali strumenti per non comprometterne la verificabilità. È una distorsione che incide direttamente sul sistema di valutazione della ricerca, e quindi sulla comunità pedagogica nel suo complesso.

1.2. Perché il problema riguarda PAED-01 e PAED-02

La ricerca educativa, seppur di ambito non-bibliometrico – sia nelle pubblicazioni in Classe A sia nelle ricerche dottorali – assume sempre più frequentemente modelli che possano dimostrare la ripetibilità e la rigorosità dell'impianto metodologico, come risposta al dibattito, anch'esso sempre più stringente, sul presunto maggiore valore attribuito alla ricerca dei settori bibliometrici. In tal senso, il modello PRISMA è uno tra i più noti e utilizzati: eppure la ricerca educativa, per complessità e spesso ampiezza dei fenomeni osservati e indagati, non è un semplice ambito in cui "mettere alla prova" un workflow nato in settori bibliometrici come quello medico. Possiamo assumere che anche la ricerca educativa, pur essendo un campo non-bibliometrico, possa essere un luogo in cui il problema di ricerca assume tratti specifici e dove un protocollo trasparente potrebbe divenire particolarmente necessario, come qualificazione del proprio operato. Annotiamo alcune ipotesi che sostengono questo assunto:

– Gran parte della ricerca in PAED è *non-bibliometrica e tematicamente densa*: costrutti come "apprendimento cooperativo", "valutazione formativa" o "professionalità docente" non hanno la standardizzazione terminologica della ricerca clinica, medica o delle cosiddette *hard sciences*, il che amplifica sia il rischio di generazione su temi meno familiari (Linardon et al., 2025) sia la difficoltà di una ricerca bibliografica esaustiva e definita;
– La comunità educativa lavora spesso con *risorse limitate* – singoli ricercatori, piccoli gruppi, accesso parziale alle banche dati a pagamento o dipendente dagli accordi delle Università di appartenenza – e con tempi lunghi di "messa alla prova" delle ipotesi, condizione nella quale l'efficienza promessa dagli algoritmi è attraente, con il rischio di un uso opaco come "funzione" riparativa della subalternità alla quale è posta la ricerca educativa;
– Le scienze della formazione hanno una responsabilità formativa propria nel tempo dell'IA: ovvero "insegnano" *come si fa ricerca*. Di conseguenza, la condivisione di uno o più protocolli dichiarati e verificabili non diviene solo uno strumento di lavoro, bensì un oggetto di *formazione* e un presidio di *accountability* disciplinare.

Possiamo aggiungere una ragione riflessiva. Le scienze della formazione non sono soltanto potenziali utenti di strumenti di IA, ma ne costituiscono da tempo un campo di studio: la ricerca educativa ha prodotto un numero consistente di analisi sulle applicazioni della cosiddetta intelligenza artificiale, sull'apprendimento e sull'insegnamento (Luckin et al.,

2016; Holmes et al., 2019), al punto che le revisioni sistematiche sull'IA in educazione si sono a loro volta moltiplicate (Zawacki-Richter et al., 2019). Produrre sintesi delle evidenze *sull'*IA ricorrendo in modo opaco a strumenti *di* IA sarebbe un'incoerenza difficilmente sostenibile: in PAED, dove l'IA è insieme oggetto e strumento, un protocollo dichiarato e verificabile potrebbe divenire non un accessorio, bensì una condizione di (nuova) identità disciplinare.

L'integrazione dell'IA prometterebbe, in effetti, vantaggi rilevanti di efficienza – le rassegne sull'automazione dello screening riportano riduzioni del carico manuale tra il 30% e il 70% (O'Mara-Eves et al., 2015) e, in studi più recenti, riduzioni di tempo superiori al 50% nella maggioranza dei casi e oltre il 75% in alcuni disegni a doppio screening – e una possibile democratizzazione del metodo, che renderebbe la revisione sistematica accessibile anche a gruppi con risorse limitate. Tali vantaggi sono legittimi solo a condizione che l'uso degli strumenti sia documentato in modo trasparente e sottoposto a verifica umana, come richiesto dalle proposte di estensione del modello PRISMA (Holst et al., 2025) e dalle posizioni congiunte degli enti di *evidence synthesis* (Flemyng et al., 2025).

1.3. Obiettivo e natura della proposta

L'obiettivo del contributo è rendere l'integrazione algoritmica non solo dichiarata, ma *costitutiva* del protocollo metodologico. La scelta si fonda su tre premesse:

- La *trasparenza* come requisito epistemologico, poiché la replicabilità impone di documentare tutti gli strumenti impiegati, inclusi gli algoritmi;
- Il principio *human-in-the-loop* come garanzia di validità, per cui l'algoritmo accelera operazioni descrittive e bibliografiche ma la valutazione critica della pertinenza, la sintesi interpretativa e l'argomentazione restano prerogativa del ricercatore;
- L'*integrazione metodologica* come contributo scientifico in sé, in un momento in cui gli standard disciplinari per PAED non sono ancora codificati.

L'ipotesi contenuta in questo lavoro non presenta la validazione definitiva di un metodo, ma è una proposta argomentata di un workflow trasparente e passibile di validazione. I giudizi di validità attribuiti alle singole fasi sono valutazioni esperte degli autori, esplicitate come tali nel par. 5, e non l'esito di una valutazione esterna, di un confronto sistematico tra versioni del workflow o di un'analisi della concordanza tra valutatori su dati reali.

1.4. I "prodotti" metodologici del contributo

Pur essendo il workflow una proposta (ancora da validare) non significa, conseguentemente, che la proposta stessa (il workflow che presentiamo) non produca risultati. Di fatto, nella successiva Tabella 1, sono resi espliciti e dichiarati cinque prodotti, ciascuno autonomamente riutilizzabile anche se non venisse adottata IATSR nel suo insieme. (a) Il primo è la formalizzazione del workflow in sette fasi sequenziali, con l'assegnazione esplicita a un sistema di controllo (umano esclusivo, assistito, logistico) e la definizione delle verifiche/responsabilità *human-in-the-loop*. (b) Il secondo è la classificazione R1/R2 delle fonti, che distingue la riproducibilità esecutiva da quella procedurale e vorrebbe fornire una regola operativa per trattare fonti non deterministiche senza rinunciare alla verificabilità: è, a nostro avviso, il prodotto teoricamente più incisivo (in altre parole il punto dal quale si sono originate curiosità, riflessioni, ipotesi, sperimentazioni) poiché confuterebbe l'obiezione che l'uso di strumenti generativi rende, per definizione, non riproducibile una revisione. (c) Il terzo è la matrice di rischio epistemologico, che associa a ciascun rischio noto dell'integrazione algoritmica una misura di mitigazione ancorata a una specifica fase del processo. (d) Il quarto è la mappatura del workflow sulle voci della checklist PRISMA 2020, che documenterebbe la compatibilità con lo standard. (e) Il quinto è l'adattamento del diagramma di flusso PRISMA, che rende visibili i punti di innesto dell'IA, i conteggi per fonte e la quota di studi inclusi provenienti unicamente da fonti non bibliometriche (Tabella 1).

Prodotto	Tipologia	Cosa rende evidente	Rif. testo
P1. Formalizzazione del workflow in sette fasi	Modello procedurale	Rende esplicito e ispezionabile dove l'IA entra e dove è esclusa	par. 4
P2. Classificazione R1/R2 delle fonti	Contributo concettuale	Distingue riproducibilità esecutiva e procedurale; consente di usare fonti non deterministiche senza perdere verificabilità	par. 5.3
P3. Matrice di rischio epistemologico	Strumento di controllo	Associa a ogni rischio noto una mitigazione ancorata a una fase	par. 6.3
P4. Mappatura sulle voci PRISMA 2020	Strumento di reporting	Documenta la compatibilità con lo standard senza pretendere lo statuto di estensione	App. A
P5. Diagramma di flusso PRISMA adattato	Strumento di reporting	Rende contabile il contributo marginale delle fonti R2 e visibili i *checkpoint human-in-the-loop*	App. B

Tabella 1. I cinque prodotti metodologici del contributo, con la loro natura e la sezione in cui sono sviluppati

2. Quadro di riferimento

Dalle considerazioni precedenti proponiamo gli elementi su cui si basa l'ipotesi, che considera il modello PRISMA e le sue estensioni, la specificità di PRISMA-trAIce, l'uso dell'IA nelle *literature review*, il principio *human-in-the-loop*, i concetti di riproducibilità e trasparenza e, infine, le ragioni per cui le scienze della formazione costituirebbero il campo elettivo della proposta.

2.1. PRISMA 2020 e le sue estensioni

Il modello PRISMA (Page et al., 2021) è lo standard internazionale per la conduzione e la rendicontazione di revisioni sistematiche e meta-analisi. Nato in ambito medico-clinico (Moher et al., 2009), è stato progressivamente esteso alle scienze sociali, psicologiche e dell'educazione (Grant & Booth, 2009; Xiao & Watson, 2019). La sua famiglia comprende estensioni per le *scoping review* (PRISMA-ScR; Tricco

et al., 2018) e per gli studi di accuratezza diagnostica (PRISMA-DTA; McInnes et al., 2018). Quattro proprietà lo rendono adatto a sostenere l'ipotesi: il riconoscimento internazionale; la struttura a flusso documentato, che impone la tracciatura quantitativa di ogni fase di inclusione/esclusione; la modularità, che consente l'innesto di strumenti algoritmici in fasi delimitate senza compromettere l'integrità del processo; l'esistenza di un'estensione dedicata proprio all'IA come strumento.

2.2. PRISMA-trAIce: l'IA come strumento, non come oggetto

Le linee guida PRISMA-AI riguardano l'IA come *oggetto* di ricerca (revisioni di studi sull'IA), mentre PRISMA-trAIce (Holst et al., 2025) è un'estensione *discipline-agnostic* del PRISMA 2020 dedicata alla reportistica trasparente dell'IA quando essa è usata come *strumento* metodologico nella sintesi delle evidenze. Di fatto essa rappresenta la base del workflow ipotizzato. PRISMA-trAIce introduce item di reportistica su dichiarazione e versione degli strumenti, misure di accordo tra esito umano e algoritmico, discussione dei limiti, e un adattamento del diagramma di flusso che esplicita i punti di innesto dell'IA. La proposta IATSR recepisce questi item e intende adattarli al contesto PAED.

2.3. L'IA nelle literature review: efficienza e rischi

L'evidenza sull'automazione dello screening è ben documentata sul versante dell'efficienza (O'Mara-Eves et al., 2015) (cfr. par. 1.2). È tuttavia un'efficienza che riguarda la *logistica* del processo (prioritizzazione, deduplicazione, ordinamento), non la decisione di pertinenza. Sul versante generativo, i rischi già richiamati – invenzione di citazioni, errori bibliografici, appiattimento interpretativo – impongono che ogni output sia verificato. La proposta tiene insieme i due versanti: usa l'IA dove l'evidenza ne mostra il vantaggio (recupero e logistica) e la confina dove l'evidenza ne mostra il rischio (decisione e interpretazione).

2.4. Il principio human-in-the-loop

Il principio per cui l'IA assiste ma non sostituisce il giudizio del ricercatore è la posizione condivisa dei principali enti di *evidence synthesis* che hanno codificato standard metodologici per la conduzione e il reporting delle revisioni sistematiche – il *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins et al., 2024), i Campbell Standards/MECCIR (Aloe et al., 2024), il *JBI Manual for Evidence Synthesis* (Aromataris et al., 2024) e le *Guidelines and Standards for Evidence Synthesis in Environmental Management* della Collaboration for Environmental Evidence (2022) – e, in una dichiarazione congiunta, hanno stabilito che gli esseri umani restano responsabili dei giudizi e dei risultati delle revisioni e che l'uso dell'IA va dichiarato e verificato (Flemyng et al., 2025). IATSR assume questo principio come vincolo non derogabile e ne fa il criterio di assegnazione dei ruoli a ogni fase.

2.5. Riproducibilità e trasparenza

L'integrazione di strumenti non deterministici impone di distinguere due accezioni di riproducibilità. La *riproducibilità esecutiva* consiste nel riottenere lo stesso insieme di record rieseguendo la medesima query sulla stessa fonte: la garantiscono le banche dati booleane, non gli strumenti generativi, i cui modelli e corpora variano nel tempo e restituiscono un ranking anziché un insieme stabile. La *riproducibilità procedurale* consiste invece nella possibilità, per un terzo, di ricostruire e ispezionare esattamente ciò che è stato fatto, anche quando una riesecuzione produrrebbe risultati lievemente diversi. La posizione metodologica della proposta è che l'affidabilità si basi sul nucleo deterministico e sul protocollo preregistrato, mentre il contributo dell'IA è reso verificabile convertendo la riproducibilità esecutiva mancante in riproducibilità procedurale documentata (Holst et al., 2025).

2.6. Specificità per le scienze della formazione

Un punto tecnico chiarisce perché, in PAED, il nucleo bibliografico debba restare ancorato a banche dati con vocabolario controllato. Gli strumenti semantici *AI-powered* (per esempio Consensus, Elicit) attingono prevalentemente a Semantic Scholar e OpenAlex e non indicizzano né ERIC né PsycINFO. Pur offrendo una copertura ampia, queste fonti presentano, proprio nelle aree umanistiche e sociali, una qualità dei metadati e una copertura interna delle citazioni più deboli: l'analisi di Culbert et al. (2025) mostra una copertura di referenze nettamente più bassa nelle Arti e Discipline umanistiche e intermedia nelle Scienze sociali, mentre *Semantic Scholar* presenta una bassa completezza dei metadati di tipologia documentale. Ne segue che il loro valore per PAED è l'*apporto tematico* (recupero semantico di contributi pertinenti), non la copertura settoriale né la sintassi booleana riproducibile, che restano affidate al nucleo ERIC + PsycINFO. Non a caso, le revisioni sistematiche sull'IA in ambito educativo già condotte (per esempio Zawacki-Richter et al., 2019) hanno dovuto comporre fonti eterogenee a cavallo tra banche dati educative e informatiche: un'esperienza che conferma la necessità, in PAED, di ancorare il recupero a un nucleo a vocabolario controllato e di trattare gli strumenti semantici come apporto tematico complementare.

3. L'oggetto della proposta: il workflow IATSR

3.1. Definizione e statuto

Assumiamo *IATSR (Integrazione Algoritmica Trasparente nella Systematic Review)* come una proposta di workflow operativo per la conduzione di revisioni sistematiche in PAED che innesta strumenti algoritmici in fasi delimitate di un processo strutturato secondo PRISMA 2020, sotto vincolo *human-in-the-loop* e con reporting conforme a PRISMA-trAIce.

Quanto allo statuto, IATSR *non è* un'estensione formale di PRISMA – le estensioni PRISMA seguono procedure di consenso e validazione che qui non sono state condotte – bensì un workflow PRISMA-compatibile che adotta integralmente PRISMA 2020 come framework e recepisce gli item di reporting di PRISMA-trAIce. Non è inoltre un nuovo modello di reporting in concorrenza con i precedenti: è una declinazione operativa, calibrata su PAED, di standard già esistenti. Come tale è offerta alla riflessione e alla discussione della comunità scientifica.

3.2. Architettura delle fonti a tre livelli

Il recupero documentale è organizzato in tre livelli, distinti per funzione e per grado di riproducibilità. La *spina dorsale documentale* combina un nucleo riproducibile – ERIC (Insti-

tute of Education Sciences - IES, 2024) e PsycINFO (American Psychological Association - APA, 2024), interrogati con ricerca booleana e vocabolario controllato e con conteggio certo dei record – e un apporto tematico *AI-powered*, ossia un motore semantico di cui Consensus (Consensus NLP, 2024) è l'esempio adottato, documentato per query, versione, data e numero di record. A questa si affiancano, in funzione complementare e opzionale, uno *strato AI-nativo* operante sul corpus del database madre – per esempio Scopus AI (Elsevier, 2024) o Web of Science Research Assistant, ove disponibile l'accesso istituzionale – e un *ulteriore strato semantico*, per esempio Elicit (Kung, 2023). Il conteggio dei record è tenuto distinto per fonte, e i record inclusi provenienti unicamente da fonti non deterministiche sono marcati esplicitamente. L'architettura completa è descritta nella Figura 1.

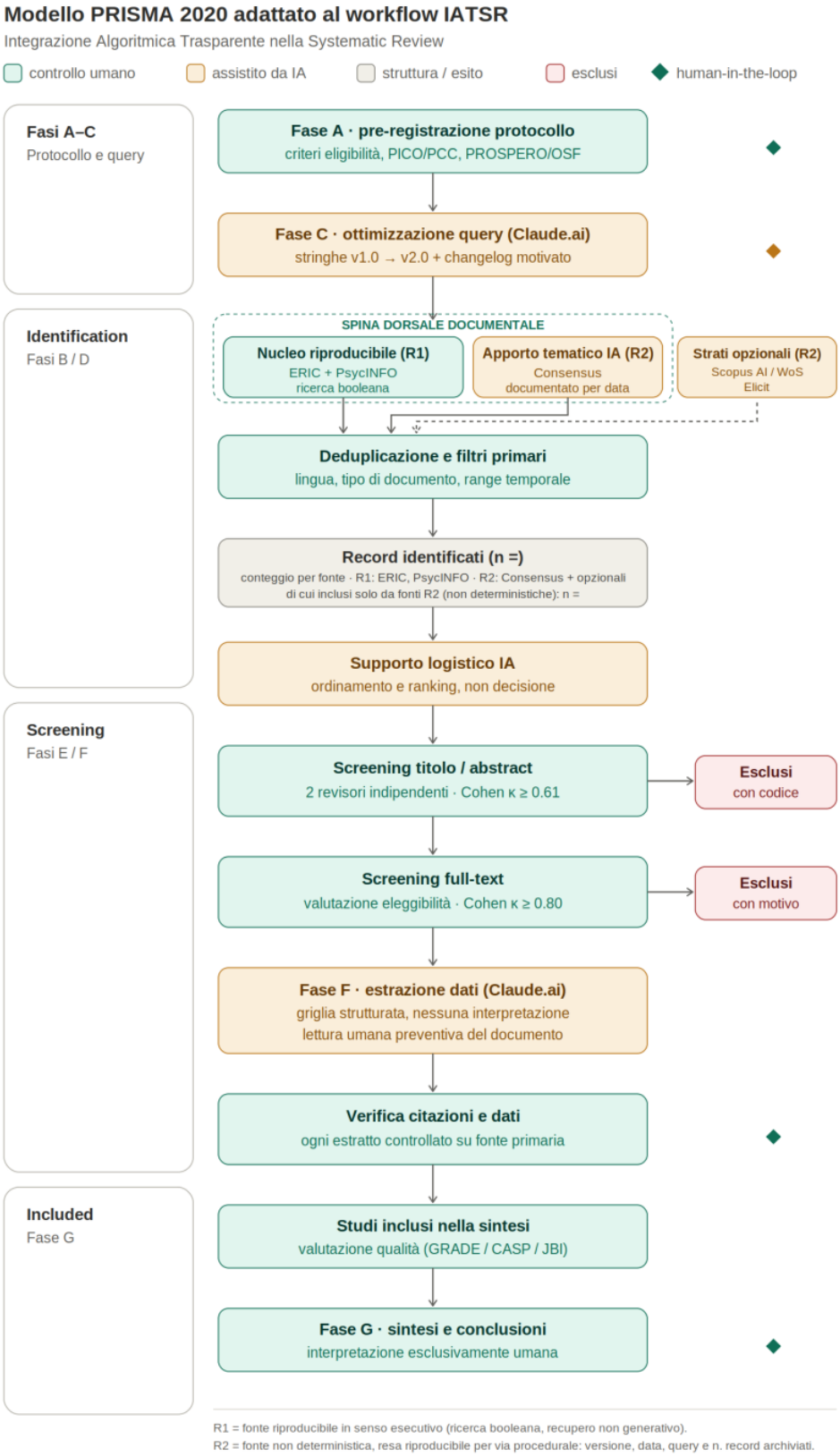

Figura 1. Architettura di IATSR. Spina dorsale documentale (nucleo R1 ERIC + PsycINFO e apporto tematico R2), strato AI-nativo opzionale e ulteriore strato semantico

3.3. Strumenti come categorie funzionali: l'indipendenza dal brand

Un'avvertenza metodologica essenziale, sollevata in sede di *pre-review*, riguarda l'indipendenza dal *brand*. Il protocollo è definito su *funzioni*, non su prodotti: ogni strumento nominato è un esempio declinato di una categoria funzionale, sostituibile con un equivalente di pari capacità. Ciò vale in particolare per il modello linguistico impiegato nell'ottimizzazione delle query e nell'estrazione strutturata: nel presente lavoro l'esempio adottato è un LLM generalista come Claude.ai, ma il protocollo non dipende da quel prodotto e resta valido con qualunque LLM di capacità comparabile. Questa scelta protegge il workflow dall'obsolescenza: un aggiornamento di versione o un cambio di fornitore non ne invalida l'impianto, ma incide solo sui parametri da documentare (versione, data, prompt). La Tabella 2 esplicita, per ciascuna categoria, i quattro statuti d'uso attribuiti: essi vanno tenuti distinti, laddove solo il primo vincola la qualità della proposta (funzione del workflow) come di seguito descritto:

- *prescrittivo*: la funzione è richiesta dal workflow; ometterla significa non applicare IATSR. La prescrizione riguarda sempre una funzione, mai un prodotto;
- *raccomandato*: la funzione è fortemente consigliata e la sua omissione va motivata e dichiarata, ma non invalida l'applicazione del protocollo;
- *opzionale*: la funzione è impiegabile in via complementare; la sua presenza o assenza non incide sulla conformità, purché sia documentata;
- *esemplificativo*: lo strumento nominato è uno fra i molti che istanziano la categoria; nessuno strumento commerciale è mai prescritto da IATSR.

Funzione nel workflow	*Esempi di strumenti*	*Statuto d'uso nel workflow*	*Fase*
Framework di reporting	PRISMA 2020 e relative estensioni (ScR, DTA, trAIce)	Prescrittivo	A–G
Pre-registrazione del protocollo	PROSPERO; OSF	Raccomandato (omissione da motivare)	A
Nucleo bibliografico riproducibile (booleano, vocabolario controllato)	ERIC; PsycINFO equivalenti in caso di accesso limitato	Prescrittivo (categoria); DB elettivi per PAED	B, D
Motore semantico *AI-powered* (apporto tematico)	Consensus; Elicit	Opzionale / esemplificativo	B, D
Strato AI-nativo sul corpus madre	Scopus AI/AI Discovery; WoS Research Assistant	Opzionale (ove disponibile)	D
Ottimizzazione query ed estrazione strutturata	LLM generalista equivalente	Opzionale (se usata, vincolata da Fasi C e F)	C, F
Piattaforme di screening collaborativo	Rayyan; Covidence	Opzionale	E
Calcolo della concordanza tra valutatori e analisi qualitativa	R; SPSS; NVivo; Atlas.ti	Prescrittivo (la funzione: misurare e riportare la concordanza)	E, F

Tabella 2. Categorie funzionali, strumenti esemplificativi e statuto d'uso. Il protocollo è definito sulle funzioni; gli strumenti sono sostituibili

Possiamo quindi rendere maggiormente esplicito come il workflow sia prescrittivo della presenza di un nucleo bibliografico riproducibile – interrogabile con sintassi booleana, ancorato a un vocabolario controllato e con conteggio certo dei record – non di una specifica coppia di banche dati. ERIC e PsycINFO sono, per PAED, l'istanza elettiva: la prima è ad accesso libero, la seconda richiede un abbonamento istituzionale di cui non tutti i gruppi dispongono. Dove PsycINFO non sia accessibile, la funzione resta prescritta e va soddisfatta con una fonte equivalente per vocabolario controllato e riproducibilità della query (per esempio Scopus o Web of Science ove disponibili, o il solo ERIC integrato da una ricerca strutturata su una fonte a copertura psicologica accessibile), dichiarando la sostituzione e le sue implicazioni di copertura nella sezione dedicata alla strategia di ricerca. Le condizioni locali di accesso sono dunque un parametro da documentare, non una condizione di validità del protocollo: un gruppo con accesso ridotto applica IATSR in modo pienamente conforme purché dichiari il nucleo effettivamente impiegato e ne discuta i limiti di copertura.

4. Descrizione del workflow: le sette fasi

Le sette fasi sono sequenziali e interdipendenti. Per ciascuna si fornisce un breve paragrafo discorsivo che ne chiarisce funzione e logica di strutturazione, seguito da una tabella di sintesi a supporto (non sostitutiva del testo). Il dettaglio operativo completo – stringhe, griglie, protocolli di interazione – è rinviato ai materiali supplementari, per non appesantire la linea argomentativa. Il modello PRISMA 2020 adattato al workflow, reso in forma tabellare nell'App. B, distingue i passaggi a giudizio umano esclusivo da quelli assistiti dall'IA e marca i *checkpoint human-in-the-loop*; le etichette adottano le categorie funzionali della Tabella 2, non i nomi commerciali.

4.1. Fase A – Adozione di PRISMA e pre-registrazione

La fase fissa, *prima* di avviare la ricerca, il framework e i criteri decisionali. La pre-adozione di PRISMA e la pre-registrazione del protocollo (PROSPERO o OSF) sono la garanzia primaria contro i bias di conferma e la selezione *post-hoc* dei criteri: fissando *ex ante* domanda, criteri di eligibilità e regole decisionali, rendono riproducibile il giudizio anche quando la superficie di ricerca si modifica.

Dimensione	*Contenuto*	*Ruolo*
Input	Domanda di ricerca; ipotesi; *background PAED-01-02*	Umano
Processo	Mappatura delle 27 voci PRISMA 2020; definizione *a priori* dei criteri; pre-registrazione	Umano
Output	Protocollo formalizzato; criteri di eligibilità; domanda PICO/PCC	Umano

Tabella 3. Fase A – Adozione di PRISMA e pre-registrazione

4.2. Fase B – Stringhe booleane iniziali e prima ricerca sulla spina dorsale

La domanda di ricerca è tradotta in operatori booleani e in termini del vocabolario controllato (Thesaurus ERIC, APA Thesaurus). La prima ricerca esplorativa interroga la spina dorsale: nucleo riproducibile (ERIC, PsycINFO) e apporto tematico *AI-powered*. Per le fonti non deterministiche si registrano fin da subito query, versione, data e numero di record, data la natura variabile dei risultati.

Dimensione	*Contenuto*	*Ruolo*
Input	Protocollo approvato; costrutti teorici	Umano
Processo	Traduzione booleana; vocabolario controllato; prima ricerca; export per fonte	Umano (+ recupero assistito sulla fonte *AI-powered*)
Output	Stringhe v1.0 per banca dati; record per fonte; analisi di copertura	Umano

Tabella 4. Fase B – Stringhe booleane iniziali e prima ricerca sulla spina dorsale

4.3. Fase C – Affinamento delle stringhe con LLM

È la fase metodologicamente più innovativa e più delicata. Un LLM è impiegato per identificare sinonimi disciplinari mancanti, verificare la coerenza logica degli operatori e proporre formulazioni alternative (*sensitivity analysis* sulle stringhe). Il vincolo non derogabile è che ogni suggerimento sia valutato dal team con voto esplicito (accettato / modificato / rifiutato) e documentato in un *changelog* motivato; l'LLM non genera mai interpretazioni sostantive dei risultati.

Dimensione	*Contenuto*	*Ruolo*
Input	Stringhe v1.0; risultati preliminari; criteri di eligibilità	Umano
Processo	Sottomissione con prompt strutturato; sinonimi, prossimità, coerenza, alternative	Assistito da IA
Output	Stringhe v2.0; log completo delle interazioni; *changelog* motivato	Umano (validazione)

Tabella 5. Fase C – Affinamento delle stringhe con LLM

4.4. Fase D – Seconda ricerca con le stringhe ottimizzate

La ricerca è eseguita con le stringhe v2.0 secondo l'architettura a tre livelli (par. 3.2). Il confronto quantitativo tra v1.0 e v2.0 funge da *sensitivity analysis* integrata; il confronto tra i record del nucleo riproducibile e quelli aggiunti dalle fonti *AI-powered* consente di quantificare il contributo marginale di ciascuna fonte e di documentarlo in modo trasparente. Successive sono la de-duplicazione e filtri primari (lingua, tipo di documento, range temporale).

Dimensione	*Contenuto*	*Ruolo*
Input	Stringhe v2.0 approvate; criteri confermati	Umano
Processo	Ricerca a tre livelli; export; de-duplicazione; filtri primari	Umano (+ recupero assistito)
Output	Dataset candidati; diagramma PRISMA aggiornato; tabella di tracciamento	Umano

Tabella 6. Fase D – Seconda ricerca con le stringhe ottimizzate

4.5. Fase E – Screening e selezione delle evidenze

Fase interamente *human-driven*. Lo screening avviene in due stadi (titolo/abstract; *full-text*) con due revisori indipendenti, risoluzione documentata dei disaccordi e misurazione della concordanza tra valutatori. Ciò che il workflow prescrive è la funzione – misurare la concordanza, riportarla e discuterla – non un valore-soglia universale. Le soglie convenzionali di Landis & Koch (1977), che collocano l'accordo sostanziale a $\kappa \geq 0{,}61$ e quello quasi perfetto a $\kappa \geq 0{,}81$, sono assunte come riferimento orientativo e non come criterio normativo. L'IA può al massimo supportare la logistica – e qui il termine va precisato: deduplicazione, ordinamento e prioritizzazione dei record tramite *active learning* – senza mai assumere la decisione di inclusione. Per i record

da fonti non bibliometriche si verifica inoltre manualmente lo stato di peer review e di eventuale ritrattazione. Restano non derogabili l'indipendenza dei due revisori e la tracciabilità delle decisioni; è la soglia, non la procedura, a essere contestuale.

Dimensione	*Contenuto*	*Ruolo*
Input	Dataset candidati; criteri di eligibilità	Umano
Processo	Screening in 2 stadi; Kappa; verifica peer-review/ritrattazione per fonti non bibliometriche	Umano (logistica assistita)
Output	Lista finale degli inclusi; diagramma PRISMA; *evidence mapping*	Umano

Tabella 7. Fase E – Screening e selezione delle evidenze

4.6. Fase F – Lettura e analisi con supporto algoritmico

Punto di massimo rischio per l'introduzione di errori (allucinazioni nelle citazioni, perdita di sfumature, appiattimento di posizioni teoriche). Le regole non derogabili: il ricercatore legge il documento *prima* di sottometterlo all'LLM; l'LLM esegue solo estrazione strutturata secondo griglia, non interpretazione teorica; ogni citazione estratta è verificata sul testo originale (pagina e paragrafo); le sintesi algoritmiche sono materiale di lavoro, non testo riproducibile nel paper; i disaccordi umano/algoritmo sono documentati.

Dimensione	*Contenuto*	*Ruolo*
Input	Lista degli inclusi; griglia di estrazione *a priori*	Umano
Processo	Lettura umana integrale; estrazione strutturata assistita; verifica di ogni estratto	Assistito da IA (validazione umana)
Output	Griglia compilata; sintesi per documento; matrice di evidenze	Umano

Tabella 8. Fase F – Lettura e analisi con supporto algoritmico

4.7. Fase G – Sintesi e validazione finale

Fase interamente prodotto dell'expertise del team: sintesi narrativa o meta-analitica, verifica della risposta alle domande di ricerca, valutazione della qualità metodologica degli studi inclusi (GRADE, CASP, JBI), discussione collegiale. Gli algoritmi non partecipano alla sintesi interpretativa, alla discussione teorica né alla formulazione delle conclusioni. Il diagramma PRISMA finale documenta numericamente ogni passaggio.

Dimensione	*Contenuto*	*Ruolo*
Input	Griglie completate; matrice di evidenze; domande di ricerca	Umano
Processo	Sintesi; valutazione della qualità; discussione collegiale	Umano esclusivo
Output	Corpo argomentativo; diagramma PRISMA finale; sezione Discussione e Conclusioni	Umano esclusivo

Tabella 9. Fase G – Sintesi e validazione finale

5. Criteri di controllo e validazione

Questa sezione chiarisce la natura dei giudizi di validità e la distinzione tra ciò che il workflow prescrive, ciò che gli autori ritengono plausibile, ciò che è già stato verificato e ciò che dovrà essere validato.

5.1. Quattro statuti epistemici:

- *Prescritto dal workflow*: regole vincolanti del protocollo (es. lettura umana prima dell'estrazione; verifica di ogni citazione; doppio screening con Kappa).
- *Ritenuto plausibile dagli autori*: giudizi esperti *a priori* sulla robustezza metodologica delle scelte, fondati su letteratura e standard, ma non ancora messi alla prova su dati propri.
- *Già verificato*: al momento, nulla del workflow è stato oggetto di prova pilota o validazione empirica da parte degli autori. La conformità dichiarata agli standard è una mappatura, non un audit esterno.
- *Da validare in studi futuri*: efficacia reale del guadagno di *recall* in Fase C/D; stabilità delle valutazioni peer review su corpora PAED; tassi effettivi di errore in Fase F; trasferibilità ad altri team; comportamento degli indici di concordanza su corpora PAED a bassa prevalenza di inclusi.

5.2. La natura dei giudizi "alta / critica / conforme"

I giudizi qualitativi assegnati alle fasi nelle versioni precedenti del documento (*alta*, *critica*) e alle voci di conformità (*conforme*, *compatibile*) sono *stime esperte degli autori*, formulate *a priori* e per confronto con checklist e standard, non l'esito di una valutazione esterna, di un confronto sistematico tra versioni del workflow o di un'analisi della concordanza tra valutatori su dati reali. La Tabella 10 li riporta come tali, esplicitandone la base e lo stato di validazione.

Fase	*Stima a priori (autoriale)*	*Base del giudizio*	*Stato di validazione*
A	Plausibilità elevata	Pre-registrazione = presidio riconosciuto contro i bias	Da validare empiricamente
B	Plausibilità elevata	Vocabolario controllato + record certi nel nucleo	Da validare empiricamente
C	Plausibilità elevata se *human-in-the-loop*	Capacità multi-dominio dell'LLM; rischio se non vincolato	Da validare (guadagno di *recall* reale)
D	Plausibilità elevata	Sensitivity analysis integrata v1.0/v2.0 e per fonte	Da validare empiricamente
E	Fase critica, *human-driven*	Nessun algoritmo sostituisce il giudizio di pertinenza	Da validare (stabilità della concordanza tra valutatori su corpora a prevalenza variabile)
F	Plausibilità elevata con controlli rigorosi	Punto di massimo rischio di errore algoritmico	Da validare (tassi di errore reali)
G	Plausibilità elevata	Sintesi interamente umana; tracciamento PRISMA	Da validare empiricamente

Tabella 10. I giudizi di validità come stime esperte *a priori*, con base e stato di validazione. Nessuna fase è al momento validata empiricamente

5.3. Riproducibilità: classificazione R1/R2, tracce e sensitivity analysis

Le fonti sono classificate in R1 (riproducibili in senso esecutivo: ERIC, PsycINFO; query booleana, conteggio stabile) e R2 (riproducibili in senso procedurale: motori semantici *AI-powered* ed eventuali strati AI-nativi; output non deterministici). Per ogni fonte R2 si archiviano versione del motore, data di accesso, stringa o prompt esatti e numero di record restituiti a quella data, depositando come materiale supplementare (OSF) l'export effettivo "congelato", così che la verifica avvenga sull'insieme realmente recuperato e non dipenda dalla possibilità di riottenere gli stessi record. Una *sensitivity analysis di riproducibilità* verifica se le conclusioni reggono usando le sole fonti R1: se reggono, lo strato IA risulta additivo; se una conclusione dipende da studi recuperati esclusivamente da fonti R2, tali studi sono segnalati esplicitamente. Nel diagramma di flusso, la casella dei record identificati distingue il conteggio per fonte e marca quanti degli inclusi provengano unicamente da R2.

Si riconosce, infine, che una quota di non-riproducibilità esecutiva dello strato IA è ineliminabile: la posizione sostenuta non è che il workflow sia integralmente riproducibile, bensì che l'affidabilità è garantita dal nucleo deterministico e dal protocollo, mentre il contributo dell'IA è reso verificabile per via documentale e quantificato nella sua incidenza.

5.4. Conformità agli standard internazionali (mappatura)

La Tabella 11 è una *mappatura autoriale* del workflow rispetto agli standard, non una certificazione esterna. È questo lo statuto da attribuire alle voci "*conforme*" e "*compatibile*".

Standard / Framework	*Esito della mappatura (autoriale)*
PRISMA 2020 (Page et al., 2021)	Allineato: le 27 voci sono mappate; il diagramma è prodotto obbligatoriamente.
PRISMA-trAIce (Holst et al., 2025)	Coerente (standard emergente): recepiti gli item di reporting trasparente, classificazione delle fonti per coinvolgimento di IA e per riproducibilità, adattamento del *flow diagram*.
COPE (2023)	Allineato: dichiarazione dell'uso dell'IA coerente con le posizioni COPE.
ICMJE (2023)	Allineato: strumenti IA dichiarati e non attribuiti come autori; responsabilità del contenuto in capo al team.
Cochrane/Campbell/JBI/CEE (Flemyng et al., 2025)	Coerente: *human-in-the-loop* come vincolo non derogabile.
OSF – *Preregistration*	Raccomandato: pre-registrazione in PROSPERO o OSF.
FAIR Principles	Allineato sul versante *Reusable*: log e tracce come materiale supplementare.
Linee guida ANVUR (VQR)	Compatibile: la trasparenza dichiarativa è coerente con i criteri di originalità e rigore.

Tabella 11. Mappatura del workflow rispetto agli standard. Si tratta di un confronto condotto dagli autori, non di una valutazione esterna

5.5. Dichiarazione AI, prompt log e appendici

Il paper esito del workflow includerà una sezione metodologica esplicita (*AI Tools Declaration*) che documenti, per ciascuna fonte: stringhe/prompt esatti, piattaforma e versione, data di ricerca, numero di record, filtri; per le fonti R2, in aggiunta, versione del motore e numero di record a quella data. Le appendici obbligatorie sono: (A) stringhe v1.0 e v2.0 con *changelog*; (B) diagramma PRISMA completo; (C) tabella di accordo tra valutatori; (D) log selezionati delle interazioni con l'LLM; (E) griglia di estrazione con marcatura human/AI per ogni voce; (F) mappatura delle voci PRISMA 2020 sul workflow; (G) diagramma di flusso PRISMA adattato a IATSR in forma tabellare. Le appendici A e B sono riportate in calce al presente documento. La collocazione – nel corpo del paper o in appendice – segue il criterio di alleggerimento argomentativo: nel testo la logica e i giudizi; in appendice il dettaglio operativo riproducibile.

6. Discussione

Alcune considerazioni sulla possibile "*stabilità motivazionale*" del workflow per l'utilizzo all'interno della comunità PAED.

6.1. Contributo alla metodologia della ricerca pedagogica

In un panorama in cui l'uso degli algoritmi è pervasivo ma raramente formalizzato, proporre un protocollo dichiarato e verificabile costituirebbe un'innovazione metodologica. Il ricercatore non si limita a usare gli strumenti, ma li inserisce in un framework epistemologico esplicito che ne definisce perimetro, limiti e procedure di controllo e che, in quanto tale, è anche oggetto di formazione alla ricerca.

6.2. Contributo a una produzione scientifica più equa

La formalizzazione risponde alle asimmetrie competitive descritte nell'introduzione: adottare e pubblicare un protocollo trasparente crea uno standard di riferimento che riduce il divario tra chi usa gli algoritmi in modo opaco e chi vi rinuncia per non comprometterne la verificabilità, restituendo a questi ultimi l'accesso ai vantaggi legittimi degli strumenti.

6.3. Contributo alla valutazione della ricerca

La dichiarazione metodologica strutturata offre ai revisori uno strumento per distinguere un uso trasparente e fondato degli algoritmi da un uso opaco finalizzato alla mera ottimizzazione bibliometrica. È un contributo al dibattito sulla ridefinizione dei criteri di qualità della ricerca nell'era dell'IA. La matrice di rischio epistemologico (Tabella 12) sistematizza i principali rischi e le misure di mitigazione previste.

Rischio	*Probabilità*	*Impatto*	*Mitigazione nel workflow*
Allucinazioni bibliografiche	Media	Critico	Verifica manuale di ogni citazione su fonte primaria (Fase F)
Bias di conferma algoritmico	Alta	Alto	Confronto con banche dati tradizionali su campione (Fase D)
Perdita di sfumature interpretative	Alta	Medio	Sintesi algoritmiche come materiale di lavoro; sintesi finale umana (Fase G)
Inconsistenza tra sessioni dell'LLM	Media	Medio	Documentazione integrale dei prompt; *versioning* delle stringhe
Over-reliance sul ranking semantico	Media	Alto	Doppio screening indipendente con Kappa (Fase E)
Mancata identificazione di letteratura grigia	Alta	Medio	Ricerca complementare in repository; ERIC; strato semantico
Inclusione di preprint non peer-reviewed (fonti R2)	Media	Alto	Verifica umana dello stato di peer review prima dell'inclusione
Inclusione di articoli ritirati	Bassa	Alto	Controllo ritrattazioni (Retraction Watch, Crossref) in *full-text*

Tabella 12. Matrice di rischio epistemologico e misure di mitigazione

7. Limiti

La proposta presenta ovviamente dei limiti che ne circoscrivono, allo stato, gli assunti in termini di ipotesi riflessiva e di condivisione di un problema, e che intendono orientare la discussione e la ricerca futura verso la messa alla prova sistematica:

– *Assenza di validazione empirica.* Il limite principale: il workflow è argomentato, non ancora messo alla prova su dati reali (né pilota, né peer-review, né confronto tra versioni);
– *Dipendenza dagli strumenti e instabilità delle piattaforme.* Pur essendo il protocollo *brand-independent*, la sua esecuzione dipende da strumenti che evolvono o possono essere dismessi; mutano la traccia da documentare e, potenzialmente, la copertura. In modo particolare, le pre-analisi sono state realizzate con l'utilizzo di strumenti *AI-powered* (motore semantico e LLM generalista);
– *Copertura incompleta nelle scienze sociali e umanistiche.* Le fonti semantiche *AI-powered* sono più deboli proprio in PAED (Culbert et al., 2025), il che confina il loro ruolo all'apporto tematico e non elimina il rischio di lacune;
– *Differenza tra systematic e scoping review.* Il workflow va

calibrato sul tipo di revisione (PICO vs PCC; PRISMA-ScR), con implicazioni su criteri e sintesi non ancora dettagliate;
– *Difficoltà per ricercatori isolati.* Il doppio screening e la discussione collegiale presuppongono un team; per il singolo ricercatore vanno previste soluzioni equivalenti (es. screening differito, terzo revisore esterno);
– *Rischio di falsa sicurezza metodologica.* La presenza di un protocollo dichiarato non garantisce di per sé la qualità: senza un'applicazione rigorosa e l'accortezza documentale, la trasparenza può diventare una rassicurazione formale.

8. Conclusioni

Innanzi al fattore emergente di un uso dell'IA nella ricerca tanto diffuso quanto raramente dichiarato, la proposta IATSR sosterrebbe la seguente tesi: l'integrazione algoritmica può essere resa rigorosa rendendola costitutiva e trasparente del protocollo, entro PRISMA 2020 e sotto vincolo *human-in-the-loop*. Il contributo vorrebbe offrire alla comunità pedagogica una proposta di workflow argomentato, verificabile e da "mettere alla prova della validazione", insieme agli strumenti di reporting per documentarne l'eventuale funzionalità. La fase successiva – ovvero una possibile sperimentazione, idealmente concordata con riviste di settore e ancorata a una domanda di ricerca – potrebbe divenire un'opportunità per trasformare la proposta in una fase di riflessione condivisa. Il passaggio a una prassi condivisa – ammesso che la comunità disciplinare lo ritenga desiderabile – richiederebbe, inoltre, almeno tre condizioni, ad oggi tutte non soddisfatte: prove su corpora reali; applicazioni indipendenti da parte di gruppi diversi dagli autori; verifiche della stabilità delle decisioni, in particolare della concordanza tra valutatori su corpora a diversa prevalenza di inclusi. Fino ad allora, ciò che offriamo alla comunità pedagogica PAED non è un metodo validato, ma un oggetto di discussione.

9. Appendici

9.1. Appendice A – Mappatura delle voci PRISMA 2020 sul workflow IATSR

La Tabella 13 riporta le voci della checklist PRISMA 2020 più critiche ai fini della validità metodologica del workflow, con la rispettiva applicazione. Costituisce l'evidenza puntuale a sostegno dell'affermazione, ripresa nel par. 5.4 e nella Tabella 11, secondo cui le 27 voci sono mappate nel processo.

Voce PRISMA 2020	*Applicazione nel workflow IATSR*
Voce 5 – *Eligibility criteria*	I criteri di inclusione/esclusione sono definiti *a priori* dal ricercatore e non sono modificabili dagli algoritmi.
Voce 6 – *Information sources*	ERIC e PsycINFO come nucleo booleano riproducibile della spina dorsale; un motore semantico *AI-powered* (es. Consensus) come apporto tematico, documentato per data; Scopus AI/AI Discovery o WoS Research Assistant ed Elicit come strati complementari opzionali; un LLM come strumento di affinamento delle query.
Voce 7 – *Search strategy*	Le stringhe booleane sono documentate in entrambe le versioni (pre e post affinamento algoritmico), con *changelog* motivato.
Voce 8 – *Selection process*	La selezione finale dei documenti è operata esclusivamente dal ricercatore umano (doppio screening con calcolo del Kappa).
Voce 9 – *Data collection process*	L'estrazione dei dati è condotta con supporto algoritmico ma validata dal ricercatore umano su ogni estratto.
Voce 16 – *Study selection*	Il diagramma di flusso PRISMA documenta i numeri di ogni fase, inclusi i passaggi algoritmici (cfr. Appendice B).
Voce 24 – *Limitations*	Il paper finale discute esplicitamente i limiti introdotti dall'uso degli algoritmi (cfr. par. 7).

Tabella 13. Voci critiche della checklist PRISMA 2020 e loro applicazione nel workflow. Riproduzione, in forma de-brandizzata, della mappatura presente nella prima versione del documento

9.2. Appendice B – Diagramma di flusso PRISMA 2020 adattato a IATSR (forma tabellare)

La Tabella 14 rende in forma tabellare il flusso PRISMA 2020 adattato al workflow, con l'architettura delle fonti a tre livelli, la distinzione tra passaggi a controllo umano e passaggi assistiti dall'IA e i *checkpoint human-in-the-loop* (marcati con ⍰). I conteggi (n) vanno compilati con i valori effettivi della revisione. R1 = fonte riproducibile in senso esecutivo; R2 = fonte non deterministica, resa tracciabile per via procedurale (cfr. Par. 5.3).

Stadio	*Passaggio del flusso*	*n*	*Controllo / checkpoint*
Identificazione	Record identificati – nucleo R1: ERIC (ricerca booleana)	n =	Umano
Identificazione	Record identificati – nucleo R1: PsycINFO (ricerca booleana)	n =	Umano
Identificazione	Record identificati – apporto tematico R2: motore semantico (es. Consensus), per data	n =	Assistito da IA
Identificazione	Record identificati – strati opzionali R2: Scopus AI/WoS; Elicit	n =	Assistito da IA
Identificazione	Record rimossi prima dello screening (duplicati; filtri lingua, tipo, range temporale)	n =	Umano

Screening	Record sottoposti a screening titolo/abstract (2 revisori indipendenti)	n =	Umano concordanza misurata e riportata (soglia pre-registrata) ◆
Screening	Record esclusi con codice motivazionale	n =	Umano
Screening	Supporto IA: deduplicazione, ordinamento, prioritizzazione (non decisione)	—	Assistito da IA
Screening	Report valutati a *full-text* per eleggibilità	n =	Umano · concordanza misurata e riportata (soglia pre-registrata) ◆
Screening	Report esclusi con motivo (incl. verifica peer review/ritrattazione per fonti R2)	n =	Umano ◆
Inclusi	Studi inclusi nella sintesi	n =	Umano ◆
Inclusi	di cui inclusi unicamente da fonti R2 (marcati esplicitamente)	n =	Umano

Tabella 14. Flusso PRISMA 2020 adattato a IATSR. ◆ = *checkpoint human-in-the-loop*. Adattamento del *PRISMA 2020 Statement* (Page et al., 2021) e del *flow diagram* PRISMA-trAIce (Holst et al., 2025)

9.3. Dichiarazione di utilizzo algoritmico (Methodological Disclaimer)

"Il presente documento costituisce una proposta argomentata di protocollo metodologico, non una validazione conclusa. Gli strumenti algoritmici menzionati sono impiegati in modo strumentale, dichiarato e documentato, con processo di revisione *human-in-the-loop* in ogni fase. Nessun contenuto del paper finale è prodotto o attribuito ad algoritmi senza validazione critica del team. I nomi commerciali compaiono come esempi declinabili in categorie funzionali e non come componenti strutturali del protocollo.

Per la stesura della presente proposta, gli algoritmi sono stati utilizzati:

- Nella generazione della Figura 1 con Claude.ai PRO su informazioni fornite dagli autori;
- Nel controllo delle tabelle di affidabilità, effettuato su prompt alimentati dai ricercatori con Claude.ai PRO;
- Per il controllo del testo e dell'affidabilità in base al modello di PRISMA-trAIce (Holst et al., JMIR AI 2025 con Claude.ai PRO);
- Per il controllo della coerenza interna del testo finale con Claude.ai PRO."

Riferimenti bibliografici

Aloe, A. M., Dewidar, O., Hennessy, E. A., Pigott, T., Stewart, G., Welch, V., Wilson, D. B., & Campbell MECCIR Working Group. (2024). Campbell standards: Modernizing Campbell's methodologic expectations for Campbell Collaboration intervention reviews (MECCIR). *Campbell Systematic Reviews, 20*, Article e1445. https://doi.org/10.1002/cl2.1445

American Psychological Association. (2024). *APA PsycInfo* [Research database]. APA PsycNet. https://www.apa.org/pubs/databases/psycinfo

Aromataris, E., Lockwood, C., Porritt, K., Pilla, B., & Jordan, Z. (Eds.). (2024). *JBI manual for evidence synthesis*. JBI. https://doi.org/10.46658/JBIMES-24-01

Pullin, A. S., Frampton, G. K., Livoreil, B., & Petrokofsky, G. (Eds.). (2022). *Guidelines and standards for evidence synthesis in environmental management* (Version 5.1). Collaboration for Environmental Evidence. https://www.environmentalevidence.org/information-for-authors

Consensus NLP, Inc. (2024). *Consensus* [*AI-powered* academic search engine]. https://consensus.app

COPE Council. (2023). *COPE position statement on AI tools in scholarly publishing*. https://doi.org/10.24318/cCVRZBms

Culbert, J. H., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2025). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics, 130*, 2475–2492. https://doi.org/10.1007/s11192-025-05293-3

Elsevier. (2024). *Scopus AI (AI Discovery)* [Generative AI research tool]. https://www.elsevier.com/products/scopus/scopus-ai

Flemyng, E., Noel-Storr, A., Macura, B., Gartlehner, G., Thomas, J., Meerpohl, J. J., Jordan, Z., Minx, J., Eisele-Metzger, A., Hamel, C., Jemioło, P., Porritt, K., & Grainger, M. (2025). Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI and the Collaboration for Environmental Evidence 2025. *Environmental Evidence, 14*, Article 20. https://doi.org/10.1186/s13750-025-00374-5

Glynn, A. (2024). *Suspected undeclared use of artificial intelligence in the academic literature: An analysis of the Academ-AI dataset* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2411.15218

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal, 26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Gray, A. (2024). *ChatGPT "contamination": Estimating the prevalence of LLMs in the scholarly literature* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2403.16887

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2024). *Cochrane handbook for systematic reviews of interventions* (Version 6.5, updated August 2024). Cochrane. https://training.cochrane.org/handbook

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Holst, D., Moenck, K., Koch, J., Schmedemann, O., & Schüppstuhl, T. (2025). Transparent reporting of AI in systematic *literature reviews*: Development of the PRISMA-trAIce checklist. *JMIR AI, 4*, Article e80247. https://doi.org/10.2196/80247

Incorrect terminology in Table 2 [Correction]. (2019). *JAMA, 322*(20), 2026. https://doi.org/10.1001/jama.2019.18307

Institute of Education Sciences. (2024). *Education Resources Information Center (ERIC)* [Bibliographic database]. U.S. Department of Education. https://eric.ed.gov

International Committee of Medical Journal Editors. (2023). *Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals*. https://www.icmje.org/recommendations/

Kobak, D., González-Márquez, R., Horvát, E.-Á., & Lause, J. (2025). Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances, 11*(27), Article eadt3813. https://doi.org/10.1126/sciadv.adt3813

Kung, J. (2023). Elicit (product review). *Journal of the Canadian Health Libraries Association / Journal de l'Association des bibliothèques de la santé du Canada, 44*(1), 15–18. https://doi.org/10.29173/jchla29657

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Linardon, J., Jarman, H. K., McClure, Z., Anderson, C., Liu, C., & Messer, M. (2025). Influence of topic familiarity and prompt specificity on citation fabrication in mental health research using large language models: Experimental study. *JMIR AI, 4*, Article e80371. https://doi.org/10.2196/80371

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson Education.

McInnes, M. D. F., Moher, D., Thombs, B. D., McGrath, T. A., Bossuyt, P. M., & the PRISMA-DTA Group. (2018). Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: The PRISMA-DTA statement. *JAMA, 319*(4), 388–396. https://doi.org/10.1001/jama.2017.19163

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine, 6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.-1000097

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews, 4*, Article 5. https://doi.org/10.1186/2046-4053-4-5

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Topaz, M., Roguin, N., Gupta, P., Zhang, Z., & Peltonen, L.-M. (2026). Fabricated citations: An audit across 2·5 million biomedical papers. *The Lancet, 407*, 1779–1781. https://doi.org/10.1016/-S0140-6736(26)00603-3

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*, *169*(7), 467–473. https://doi.org/10.7326/M18-0850

Xiao, Y., & Watson, M. (2019). Guidance on conducting a systematic *literature review*. *Journal of Planning Education and Research, 39*(1), 93–112. https://doi.org/10.1177/0739456X17723971

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39. https://doi.org/10.1186/s41239-019-0171-0