

Valutazione delle informazioni generate dall'IA

Sviluppo di competenze per valutare criticamente l'accuratezza, l'affidabilità e i pregiudizi dei contenuti generati dall'IA

Numero di riferimento del progetto:
2024-1-BG01-KA220-HED-000249185



Disclaimer: Finanziato dall'Unione Europea. Le opinioni espresse sono però solo quelle dell'autore o degli autori e non riflettono necessariamente quelle dell'Unione Europea o dell'Agenzia Nazionale (NA). Né l'Unione Europea né l'NA possono essere ritenute responsabili per queste.



Co-funded by
the European Union



Obiettivi di apprendimento

In poche parole



- Identifica i tre pilastri fondamentali della valutazione dei contenuti dell'IA: Accuratezza, Affidabilità e Bias (ARB).
Applica framework di verifica strutturati (come CRAAP o ROBOT) specificamente adattati per output di IA generativa.
Riconosci specifiche modalità di fallimento dell'IA, incluse allucinazioni, fabbricazione di citazioni e pregiudizi culturali.
Sviluppa attività pratiche in classe che trasformino gli errori dell'IA in esercizi di pensiero critico.
Implementare protocolli di trasparenza che richiedano agli studenti di documentare e criticare l'uso degli strumenti di IA.

L'urgenza di

Alfabetizzazione critica dell'IA – concetti chiave e definizioni

- **Cambiamento nel consumo di informazioni:** Gli studenti usano l'IA come motore di ricerca principale e strumento di redazione, spesso bypassando la verifica tradizionale.
- **La trappola della 'plausibilità':** Gli LLM sono progettati per sembrare convincenti, non necessariamente per essere sinceri. Danno priorità alla probabilità linguistica rispetto all'accuratezza fattuale.
- **Impatto sulle discipline umanistiche:** L'IA spesso fatica con sfumature, contesto culturale e profondità interpretativa – componenti fondamentali della ricerca sulle discipline umanistiche.
- **Integrità Accademica 2.0:** L'integrità non riguarda più solo il 'non barare'; si tratta di verificare le affermazioni di verità degli strumenti che usiamo.
- **Empowerment:** La valutazione dell'insegnamento sposta la narrazione da "controllare gli studenti" a "dare potere agli studiosi" affinché usino gli strumenti in modo responsabile.



La necessità critica di Competenze di valutazione dell'IA

LA SFIDA:

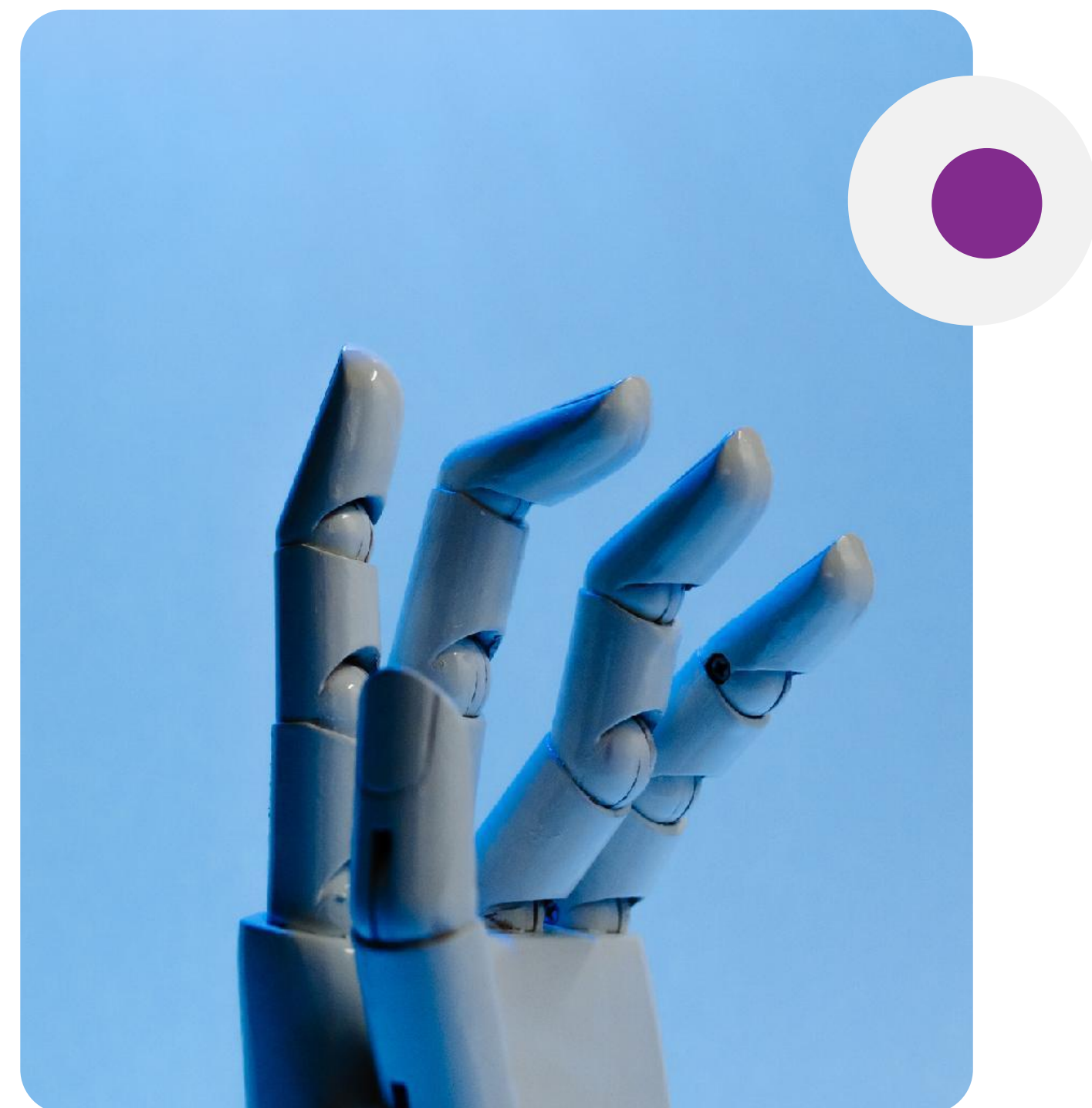
Gli strumenti di IA generano testi convincenti e fluenti che possono mascherare imprecisioni e pregiudizi. Gli studenti si affidano sempre più all'IA per ricerca, scrittura e analisi senza verifiche. Il tradizionale rilevamento del plagio è insufficiente per l'era dell'IA.

L'OPPORTUNITÀ:

La valutazione critica trasforma l'IA da minaccia a potente strumento pedagogico. Gli educatori umanistici sono in una posizione unica per insegnare una valutazione sfumata. Sviluppare le competenze di valutazione prepara gli studenti a un mondo professionale integrato dall'IA.

L'IMPERATIVO:

L'integrità accademica dipende da informazioni verificabili e affidabili. L'uso etico dell'IA richiede trasparenza e un coinvolgimento critico.





I tre pilastri:

Accuratezza, affidabilità, bias

ACCURATEZZA (Il Fact Check)

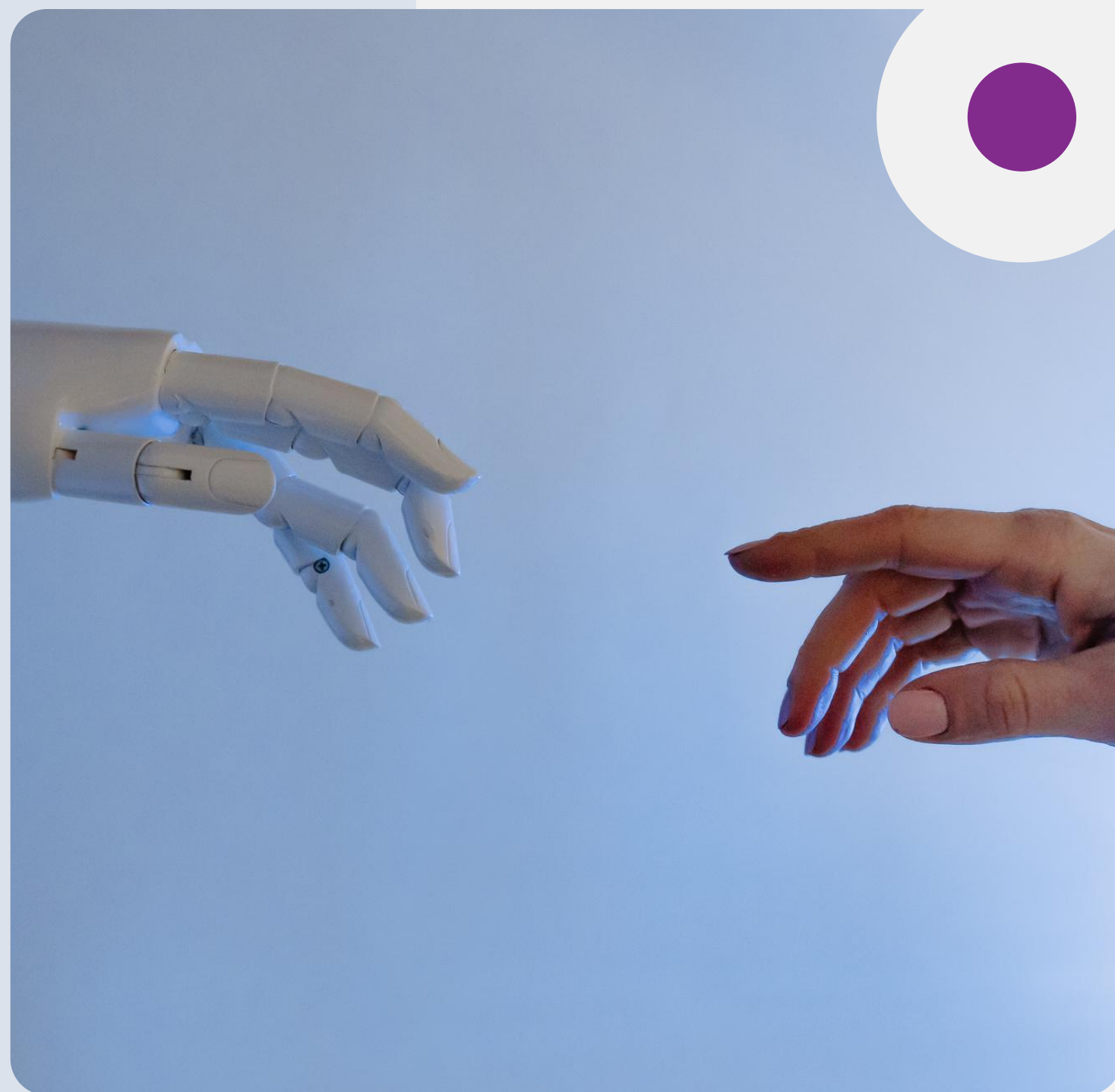
- Affermazioni specifiche possono essere incrociate con fonti accademiche consolidate? Il contenuto contiene incongruenze logiche? L'IA inventa dati, date o eventi (allucinazioni)?

AFFIDABILITÀ (Il controllo sorgente)

- L'IA può fornire citazioni reali e tracciabili? (Spesso non può). L'output è coerente su più prompt o su modelli diversi? L'IA fornisce una fonte di fonte trasparente?

BIAS (Il controllo della prospettiva)

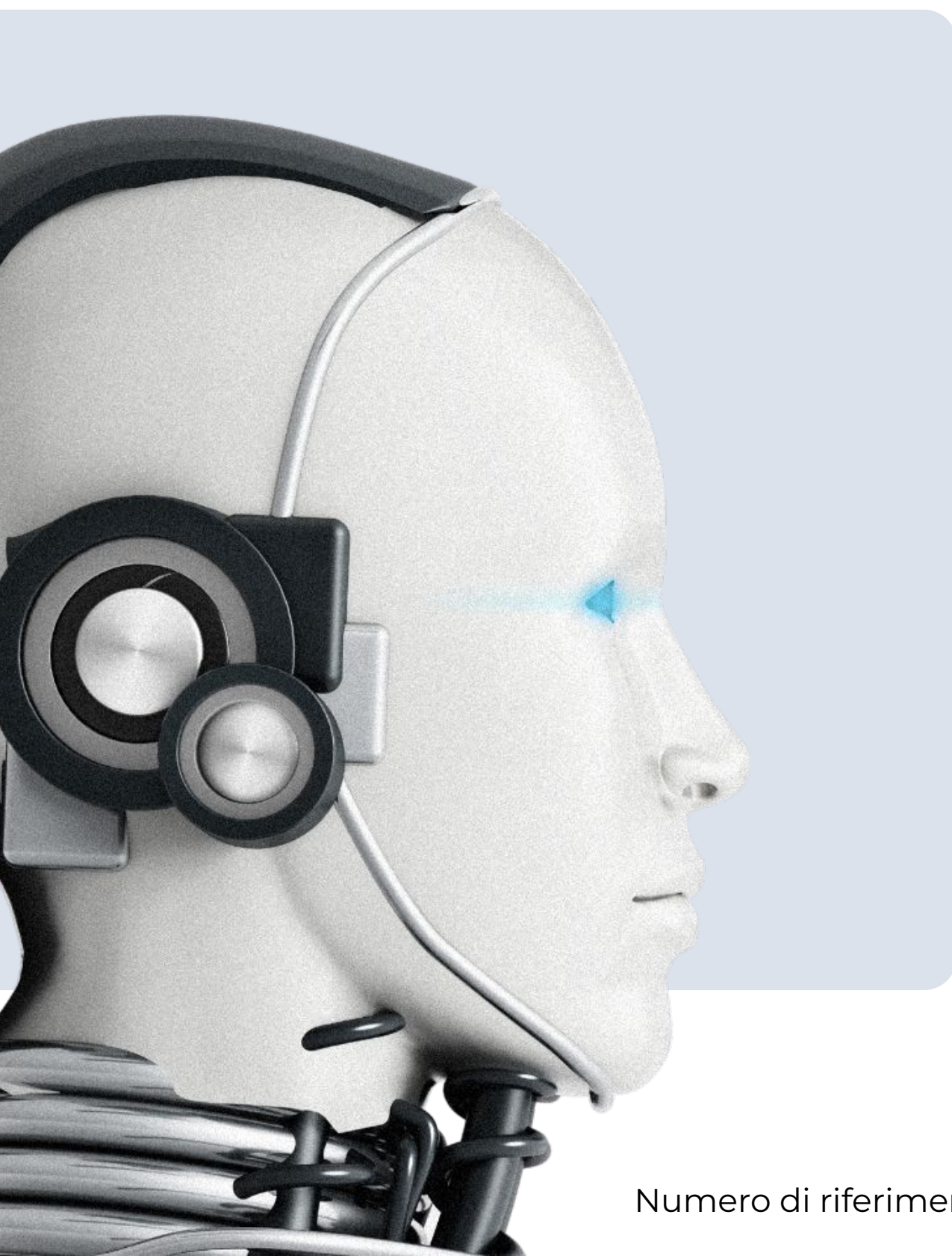
- Di chi è la voce dominante? (Il pregiudizio occidentale/anglocentrico è comune). Di chi manca la voce? (Gruppi emarginati, studi non in inglese). Il tono è oggettivo, o porta con sé stereotipi radicati?



Anatomia dell'IA

Limitazioni

- **Allucinazioni:** La generazione sicura di informazioni false. L'IA prevede la parola successiva, non 'conosce' i fatti.
Fabbricazione delle citazioni: L'IA può generare titoli di libri che sembrano reali, di autori reali, ma che in realtà non esistono.
Limiti di conoscenza: I modelli potrebbero non conoscere gli eventi degli ultimi 6-12 mesi.
Collasso del contesto: L'IA spesso appiattisce i dibattiti complessi in riassunti generici, perdendo la 'texture' del discorso umanistico.
Adulazione: L'IA tende ad essere d'accordo con la premessa dell'utente, anche se questa è errata o sbagliata.



Attività interattiva:

Individua l'allucinazione

SCENARIO: Uno studente presenta un elaborato sulla letteratura polacca citando:

'Kowalski, J. (2019). La metafora quantistica nelle prime opere di Tokarczuk. Pressa dell'Università di Varsavia.'

DISCUSSIONE:

Il titolo: Sembra accademico e plausibile.

L'autore: 'Kowalski' è un nome comune, il che lo rende più difficile da verificare rapidamente.

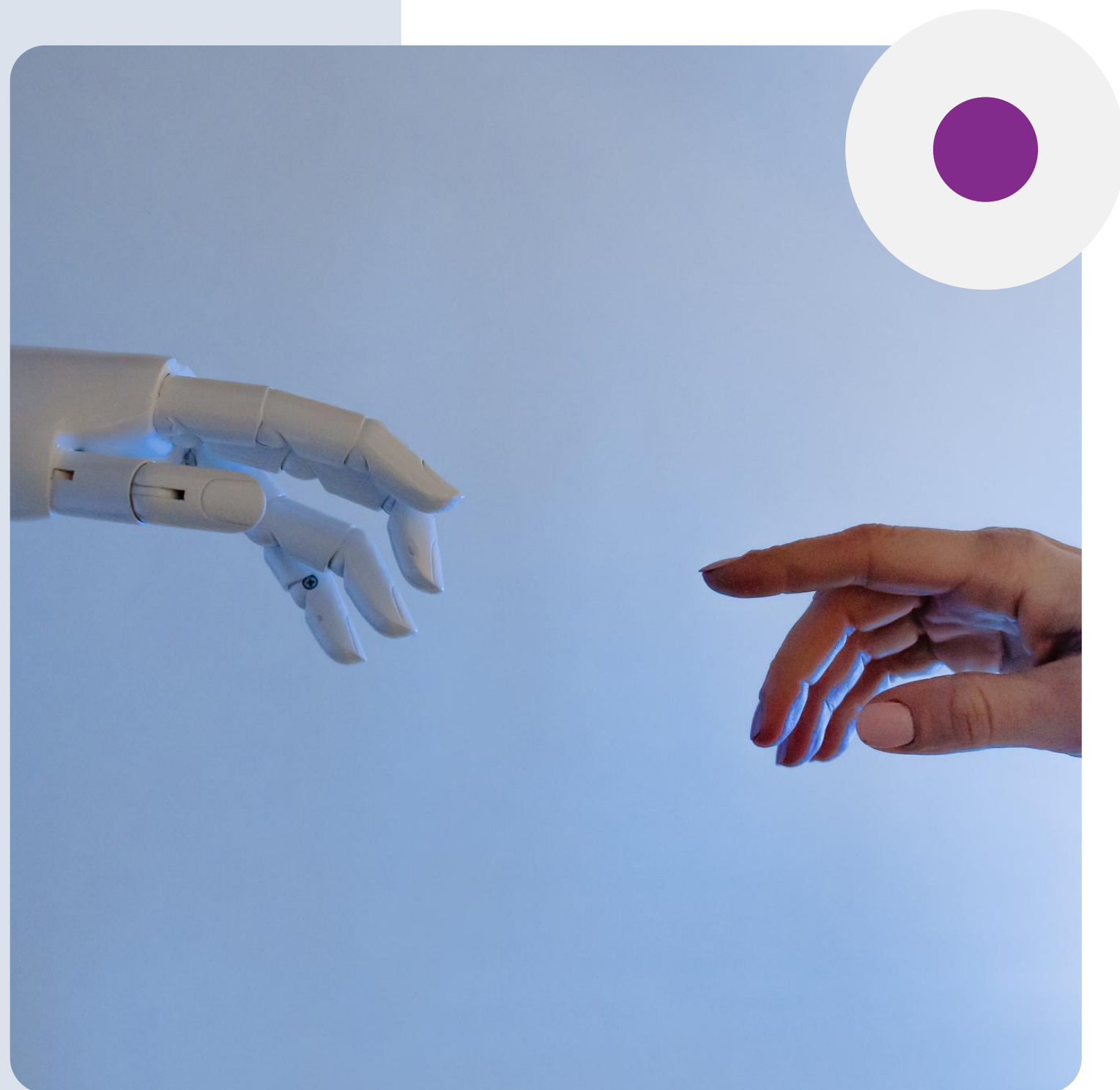
L'Editore: Una vera istituzione.

LA REALTÀ: Questo libro non esiste. L'IA ha combinato parole chiave reali per inventare una fonte.

AZIONE:

Come verificherei questo? (Google Scholar, Catalogo della Biblioteca, ricerca DOI).

Che feedback dai allo studente? (Concentrati sul processo di verifica, non solo sull'errore).



Framework: adattamento

CRAAP per l'IA

Test CRAAP tradizionale vs. esigenze dell'IA:

Valuta: i dati sull'IA sono spesso obsoleti. Controlla la data di 'data di scadenza delle conoscenze' del modello.

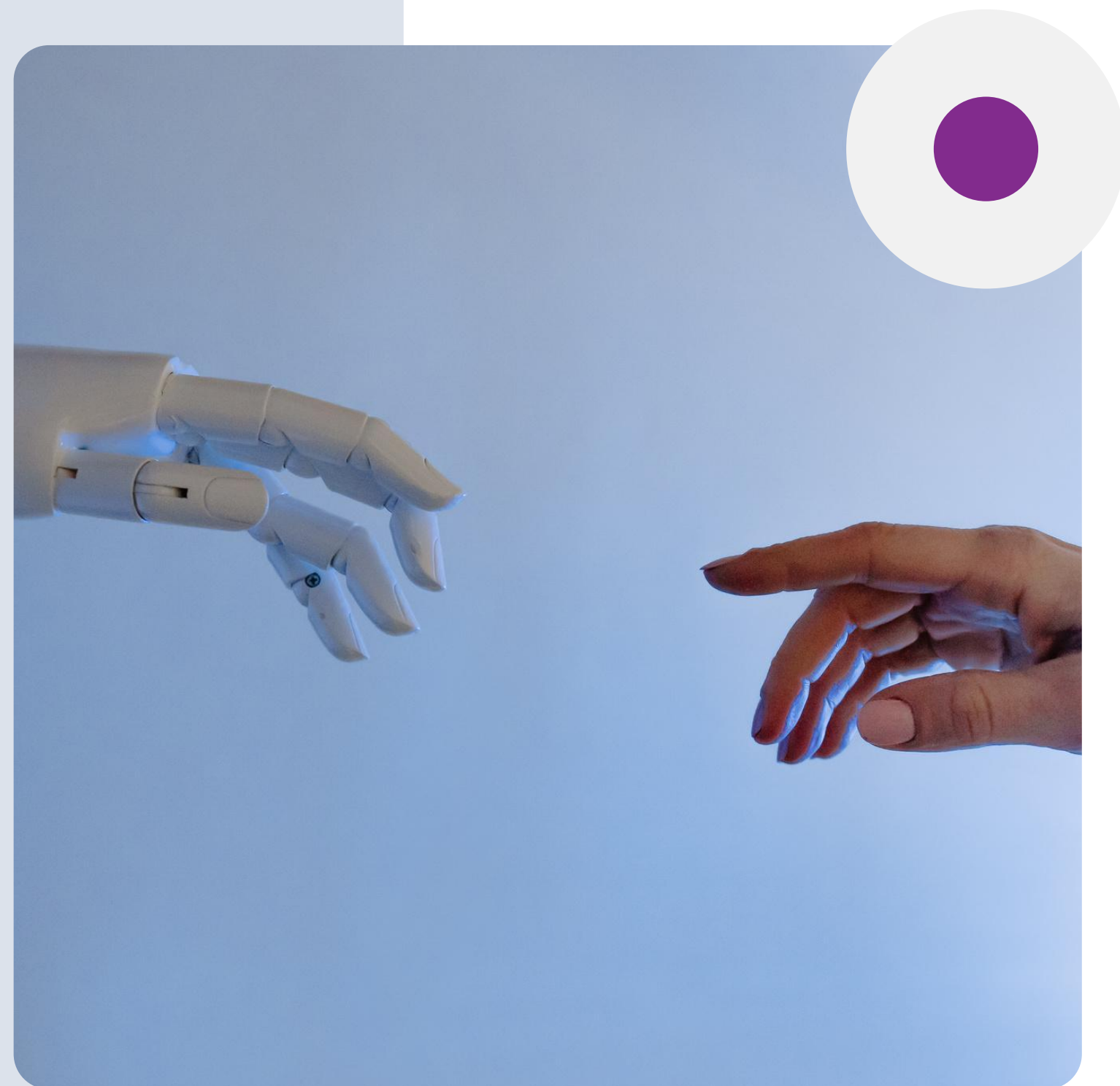
Rilevanza: l'IA risponde al prompt, ma risponde alla domanda di ricerca? (Adagiamento tempestivo vs. profondità intellettuale).

Autorità: l'IA NON ha autorità. È un aggregatore. Il controllo di 'Autorità' deve passare alla verifica delle fonti *originali* che l'IA afferma di utilizzare.

Accuratezza: Richiede una verifica esterna (Triangolazione con testi affidabili).

Scopo: Lo scopo dell'IA è soddisfare l'utente, non cercare la verità. Sii consapevole della sua tendenza a compiacere.

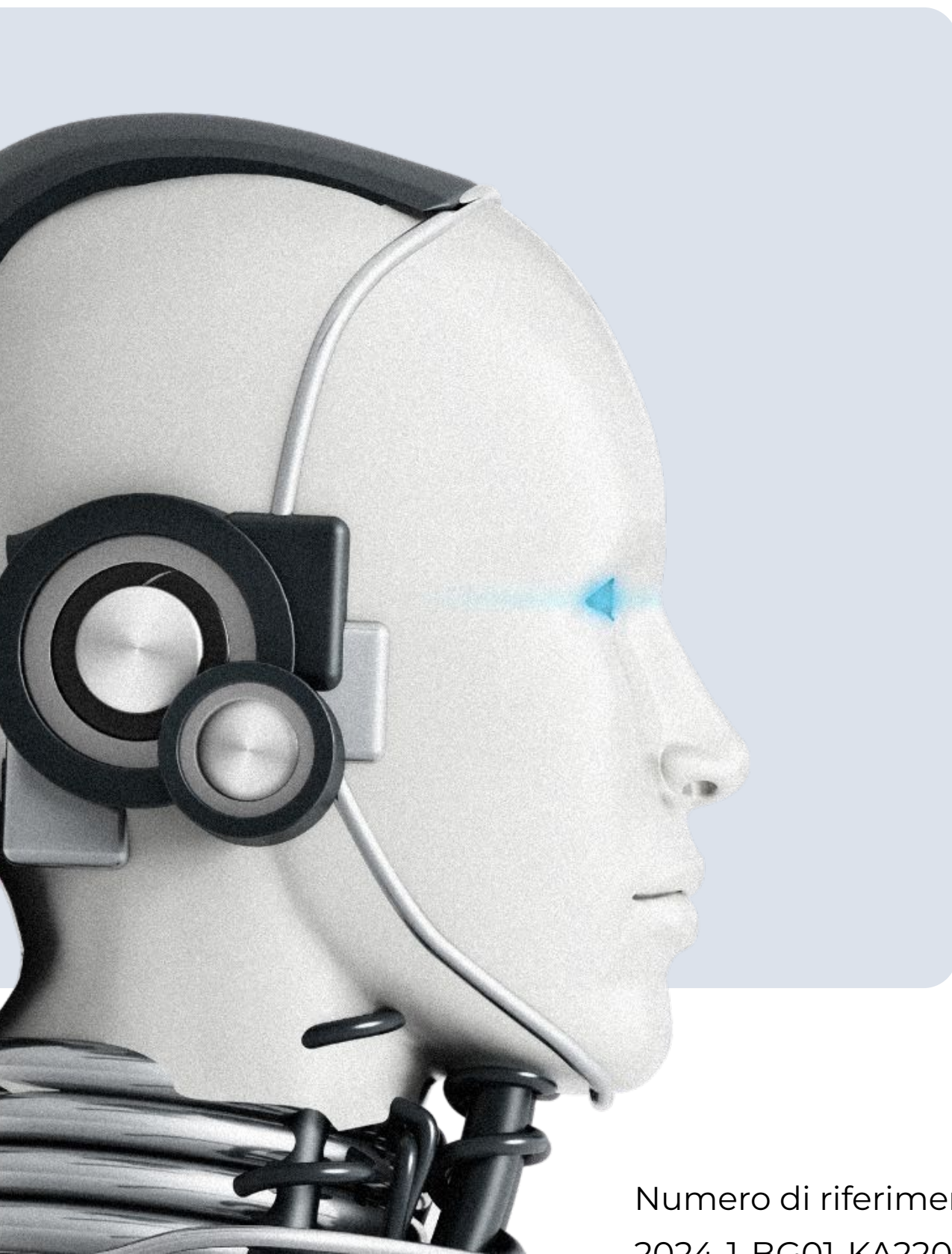
NUOVA METRICA: Riproducibilità. Si può ottenere di nuovo lo stesso risultato? Se no, non è uno strumento analitico affidabile.



Strategie di verifica

per l'aula

1. **Lettura laterale:** Non leggere la pagina; Apri nuove schede per verificare immediatamente i fatti.
Triangolazione: Verifica qualsiasi affermazione dell'IA con almeno due fonti distinte e indipendenti (ad esempio, un testo primario e un articolo peer-reviewed).
La politica di citazione 'Zero-Trust': Assumere che ogni citazione generata dall'IA sia falsa fino a quando non viene dimostrata vera tramite un DOI o un link alla biblioteca.
Ricerca inversa di immagini: Per immagini o grafici generati dall'IA, verifica se i dati esistono altrove.
Audit esperto: Incoraggia gli studenti a usare l'IA su un argomento che *già* conoscono bene per individuare gli errori sottili (Il 'Metodo Sandwich': Conoscenza umana -> bozza AI -> critica umana).



Attività interattiva:

Rilevamento del bias

SCENARIO: Chiedi all'IA 'I 10 filosofi più influenti del XX secolo'.

RISULTATO: La lista contiene 9 uomini e 1 donna. Tutti sono europei o americani.

ANALISI:

Bias di rappresentazione: Perché i pensatori non occidentali sono esclusi?

Bias linguistico: Il modello dà priorità ai dati di addestramento in lingua inglese.

Bias storico: Riproduce la marginalizzazione storica presente nel dataset.

SUGGERIMENTO PER LA BONIFICA:

Spunto negativo: 'Elenca filosofi influenti.'

Buon prompt: 'Elenca filosofi influenti del XX secolo, garantendo un equilibrio di genere e includendo prospettive non occidentali dall'Asia, dall'Africa e dall'America Latina.'



Strategie pedagogiche:

Insegnare l'uso critico dell'IA

- **Il compito 'AI Audit':** Gli studenti generano un saggio AI su un argomento, poi scrivono una critica identificando ogni errore, bias e lacuna.
Documentazione del processo: Richiedi agli studenti di inviare i loro chat log. Valuta la *qualità dei loro prompt* e la loro *verifica* dell'output.
Dibattiti in classe: Umano vs. IA. Una squadra dibatte un argomento; l'altra squadra usa l'IA. Confronta la profondità e la sfumatura degli argomenti.
Riprogettazione della valutazione: Spostati dal 'riassumere questo testo' (l'IA lo fa bene) a 'criticare questo caso di studio locale specifico usando questa teoria specifica' (l'IA fatica qui).
Rubriche trasparenti: Aggiungi una riga di rubrica per 'Valutazione critica delle fonti', premiando esplicitamente la correzione degli errori dell'IA.

L'IA della fiducia

Elenco di controllo per la valutazione



Prima di usare contenuti AI, chiedi:

VERIFICABILITÀ: Ci sono fatti o date specifiche che posso verificare?

FONTING: Ho trovato la fonte primaria per ogni citazione?

PROSPETTIVA: Ho chiesto esplicitamente punti di vista diversi?

TEMPESTIVITÀ: Questo argomento è recente? (Se sì, sii particolarmente scettico).

LOGICA: L'argomento regge o è solo una retorica fluida?

ETHICS: Ho rivelato di aver usato l'IA per questo lavoro?

CONTATTO UMANO: Richiede la mia interpretazione personale o conoscenza locale?

Punti chiave & Prossimi passi

1. L'IA è uno strumento, non un oracolo: predice il testo; Non accede alla verità.
La verifica è non negoziabile: dobbiamo passare da consumatori di informazioni a revisori delle informazioni.
La trasparenza costruisce fiducia: le discussioni aperte sulle limitazioni dell'IA riducono ansia e cattiva condotta.
Le competenze umanistiche sono competenze di IA: contesto, critica e ragionamento etico sono esattamente ciò che serve per gestire l'IA.



Coinvolgimento o Interattivo

Avvia l'attività





Riferimenti



1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). Sui pericoli dei pappagalli stocastici: i modelli linguistici possono essere troppo grandi? 🦜 Atti della Conferenza ACM 2021 su Equità, Responsabilità e Trasparenza, 610-623.
Alkaissi, H., & McFarlane, S. I. (2023). Allucinazioni artificiali su ChatGPT: implicazioni nella scrittura scientifica. Cureus, 15(2), e35179.
Caulfield, M. (2017). Alfabetizzazione web per i fact-checker degli studenti. Recuperato da <https://webliteracy.pressbooks.com/>.
Borji, A. (2023). Un archivio categorico dei fallimenti di ChatGPT. preprint arXiv arXiv:2302.03494.
Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). I rilevatori GPT sono di parte contro gli scrittori non madrelingua inglese. Patterns, 4(7), 100779..
Insegnare. International Journal of Teaching and Learning in Higher Education, 32(1), 39-48.



"Il futuro dell'istruzione non riguarda la sostituzione della mente umana riguarda l'espansione responsabile. Continua a mettere in discussione, continua a insegnare, continua a imparare."

Il consorzio:



<https://trustai.unibit.bg/>

 Trust AI project

Disclaimer: Finanziato dall'Unione Europea. Le opinioni espresse sono però solo quelle dell'autore o degli autori e non riflettono necessariamente quelle dell'Unione Europea o dell'Agenzia Nazionale (NA). Né l'Unione Europea né l'NA possono essere ritenute responsabili per queste.
Numero di riferimento del progetto: 2024-1-BG01-KA220-HED-000249185



Co-funded by
the European Union

