Un framework basato sull'intelligenza artificiale a supporto della didattica della programmazione

Vanessa Barbaro, Alessandro Pagano, Veronica Rossano[1]
[1] Università degli Studi di Bari "Aldo Moro" – Dipartimento di Informatica
vanessa.barbaro@studenti.uniba.it, alessandro.pagano@uniba.it,
veronica.rossano@uniba.it

Abstract

Nel contesto del rapido avanzamento e della crescente integrazione dell'intelligenza artificiale (IA) in molteplici settori, il suo impiego nell'ambito educativo si configura come una forza profondamente trasformativa. In particolare, l'IA generativa e i modelli linguistici di grandi dimensioni (LLM) offrono prospettive promettenti per il potenziamento dell'insegnamento della programmazione, automatizzando il processo di creazione delle domande d'esame.

Il presente studio propone un framework innovativo per la generazione automatica di domande d'esame basate sull'intelligenza artificiale, con un focus specifico sulla progettazione, valutazione e ottimizzazione dei prompt. Tale approccio è finalizzato a garantire la produzione di esercizi di programmazione di elevata qualità e consolidati fondamenti pedagogici.

Lo studio si avvale di modelli di intelligenza artificiale come ChatGPT, Gemini, Copilot e Claude, per valutare la loro efficacia nella formulazione di domande d'esame strutturate. Mediante un'analisi comparativa, le domande vengono esaminate in funzione della chiarezza, completezza, livello di difficoltà e allineamento con gli obiettivi formativi. I risultati evidenziano come l'utilizzo di prompt basati su esempi consenta di ottenere domande maggiormente coerenti e contestualmente appropriate, mentre i prompt aperti tendono a generare output caratterizzati da ambiguità o eccessiva generalità.

Attraverso il perfezionamento delle strategie di prompt engineering, la ricerca dimostra che l'intelligenza artificiale può efficacemente supportare i docenti nell'automatizzazione della progettazione delle valutazioni, contribuendo a una significativa riduzione del carico di lavoro senza compromettere il rigore accademico. Tuttavia, rimangono sfide rilevanti, quali i pregiudizi presenti nei contenuti generati dall'IA e la imprescindibile necessità di una supervisione umana. Il modello proposto si inserisce pertanto nel dibattito sull'integrazione responsabile dell'IA in ambito educativo e apre la strada a futuri sviluppi nei sistemi di apprendimento adattivo basati su tale tecnologia.

1 Introduzione

L'integrazione dell'Intelligenza Artificiale (IA) nel settore educativo sta ridefinendo rapidamente le metodologie tradizionali di insegnamento e apprendimento (Meyer et al., 2024), introducendo strumenti innovativi per ottimizzare l'esperienza formativa degli studenti. Tra le applicazioni più rilevanti figurano l'apprendimento personalizzato, la valutazione automatizzata e i sistemi di tutoraggio adattivo (Chen et al., 2023; Nguyen & Allan, 2024). In particolare, l'Intelligenza Artificiale Generativa (GenAI) e i modelli linguistici di grandi dimensioni (LLM) emergono come risorse fondamentali nella produzione automatizzata di contenuti educativi (Latif & Zhai, 2024), offrendo nuove opportunità soprattutto nella progettazione delle verifiche e valutazioni per l'insegnamento della programmazione.

La progettazione e la valutazione degli esami di programmazione rappresentano attività complesse, richiedendo competenze specialistiche e tempi elevati per garantire domande pedagogicamente valide

e coerenti con gli obiettivi formativi (Pozdniakov et al., 2024). In tale scenario, modelli di IA quali ChatGPT, Gemini, Copilot e Claude possono contribuire ad automatizzare la generazione di domande d'esame, riducendo significativamente il carico di lavoro dei docenti (GoodfellowIan et al., 2020; Vahlois et al., 2024). Tuttavia, restano aperti interrogativi sull'efficacia di tali strumenti nel contesto educativo, che richiedono ulteriori approfondimenti empirici.

Il presente studio si concentra sulla generazione automatizzata di domande d'esame per corsi di programmazione tramite modelli generativi basati sull'IA. Attraverso l'analisi comparativa di diverse strategie di prompt engineering, la ricerca mira a sviluppare un framework metodologico efficace per produrre domande chiare, complete e coerenti con i risultati di apprendimento attesi. Questo approccio si pone come alternativa al processo tradizionale, spesso oneroso e soggetto a disomogeneità, proponendo una metodologia più adattiva e scalabile. Lo studio esamina limiti e criticità di tale approccio, inclusi potenziali pregiudizi, imprecisioni e la imprescindibile necessità di supervisione umana.

Il documento è strutturato come segue: nella sezione di background è illustrata la letteratura a supporto dell'uso dell'IA in ambito educativo e l'evoluzione dei modelli LLM. La sezione metodologia descrive il framework sviluppato, le strategie di prompt adottate e le metriche per valutare le domande generate. La sezione dedicata ai risultati discute l'efficacia dei diversi modelli IA. Infine, le conclusioni sintetizzano i risultati chiave e suggeriscono sviluppi futuri per ottimizzare ulteriormente l'integrazione responsabile dell'IA nell'istruzione.

2 Background

I recenti progressi nell'intelligenza artificiale (IA) ne hanno favorito l'adozione crescente nell'istruzione (Kingma & Welling, 2014), dove strumenti come i modelli linguistici di grandi dimensioni (LLM) sono utilizzati per generare testo, fornire spiegazioni dettagliate, valutare le risposte degli studenti e creare contenuti educativi (Almasri, 2024; Wang et al., 2023). L'IA permette una personalizzazione significativa dei percorsi formativi, migliorando l'accessibilità e il coinvolgimento, soprattutto per studenti con esigenze diverse (Brown et al., 2020). Tuttavia, la presenza di pregiudizi nei modelli, le questioni etiche e la privacy dei dati rappresentano sfide importanti che richiedono un approccio responsabile e trasparente per l'adozione efficace dell'IA nell'educazione.

I recenti sviluppi nell'apprendimento automatico e deep learning hanno favorito lo sviluppo dell'Intelligenza Artificiale Generativa (GenAI), permettendo la creazione di modelli capaci di generare testi coerenti e rilevanti (Vaswani et al., 2017). In particolare, i modelli linguistici come il Generative Pre-trained Transformer (GPT), hanno rivoluzionato l'elaborazione del linguaggio naturale grazie alla loro capacità di fornire risposte contestualmente pertinenti (Molina et al., 2024). Tra i modelli più utilizzati vi sono GPT, Pathways Language Model (PaLM) e Claude, quest'ultimo con un particolare focus sulla trasparenza e sull'etica (Bender et al., 2021). Questi strumenti sono sempre più applicati nell'educazione, soprattutto per generare domande d'esame e automatizzare feedback. Tuttavia, persistono sfide legate ai pregiudizi nei modelli e all'affidabilità delle informazioni generate, che richiedono un uso attento e responsabile.

Disporre di un feedback immediato e personalizzato è fondamentale per un apprendimento efficace. I modelli di intelligenza artificiale (IA) mostrano un grande potenziale in questo ambito, riducendo il carico di lavoro degli insegnanti e migliorando l'apprendimento degli studenti.

Uno studio condotto su 459 studenti ha evidenziato che il feedback generato dall'IA ha aumentato il coinvolgimento e migliorato le competenze di scrittura rispetto all'assenza di feedback (Meyer et al., 2024). L'accuratezza del feedback IA risulta significativamente maggiore (87%) quando vengono utilizzati prompt strutturati con esempi, rispetto al caso in cui il feedback non contenga esempi (67%), sottolineando l'importanza di una progettazione accurata dei prompt. In questo contesto diversi studi

confermano che, nonostante il potenziale dell'IA, la supervisione umana rimane essenziale per garantirne l'affidabilità e la validità pedagogica (Nguyen & Allan, 2024).

Il fine-tuning dei modelli linguistici di grandi dimensioni (LLM) con contenuti educativi specifici può significativamente migliorare le prestazioni nella generazione di materiali didattici e valutazioni personalizzate. Tuttavia, permangono criticità come la scarsità di dati di qualità, distorsioni nei dati di addestramento e il rischio di generare contenuti fuorvianti (AI hallucinations), che limitano la loro piena adozione in ambito educativo. Uno studio comparativo tra ChatGPT-3.5 e BERT nelle valutazioni di un corso di chimica ha mostrato che ChatGPT-3.5, dopo fine-tuning, ha superato BERT del 9,1% in accuratezza (Latif & Zhai, 2024). Questo risultato sottolinea l'importanza del fine-tuning per migliorare l'appropriatezza contestuale dei modelli IA, sebbene rimanga aperta la questione relativa all'impatto di tali strumenti sullo sviluppo delle capacità critiche degli studenti.

L'integrazione dell'Intelligenza Artificiale (IA) nell'educazione, nonostante i notevoli benefici, presenta diverse sfide etiche e pratiche (Chen et al., 2023). Tra queste emergono preoccupazioni riguardo bias nei dati che possono generare valutazioni distorte e svantaggiare alcuni gruppi di studenti (Nakanishi, 2023). Inoltre, l'uso intensivo dell'IA rischia di limitare le interazioni dirette tra docenti e studenti, potenzialmente riducendo lo sviluppo del pensiero critico (Pozdniakov et al., 2024). Un'altra criticità riguarda l'integrità accademica, poiché l'IA potrebbe incoraggiare il plagio e compromettere l'autenticità delle valutazioni (Vahlois et al., 2024). Queste problematiche sottolineano la necessità di linee guida rigorose e di supervisione umana per un impiego equo e responsabile dell'IA (GoodfellowIan et al., 2020). Le ricerche future dovranno quindi indirizzarsi verso la definizione di quadri normativi ed etici per garantire l'efficacia educativa e l'integrità accademica.

3 Metodologia

Questo studio analizza l'uso dell'intelligenza artificiale (IA) per l'automazione della generazione di domande d'esame nella programmazione. La metodologia adottata è iterativa e combina tecniche di prompt engineering, valutazione dei modelli IA e raccolta di feedback qualitativi. La ricerca si articola in quattro fasi: definizione degli obiettivi pedagogici, progettazione e ottimizzazione dei prompt per migliorare chiarezza e rilevanza, valutazione della qualità delle domande generate, e sviluppo di un framework per la generazione standardizzata delle valutazioni.

L'analisi valuta l'efficacia di quattro chatbot generativi—ChatGPT, Gemini, Copilot e Claude—nella creazione automatica di domande di programmazione. I risultati ottenuti evidenziano opportunità e sfide legate all'integrazione dell'IA nella didattica universitaria, contribuendo così al dibattito sulla sua adozione nell'istruzione superiore.

3.1 Sviluppo del framework basato sull'intelligenza artificiale

Il framework proposto standardizza la generazione automatizzata di domande d'esame tramite intelligenza artificiale (IA), garantendone l'allineamento agli obiettivi didattici. Il processo metodologico comprende quattro fasi: definizione degli obiettivi di apprendimento mirati a competenze chiave della programmazione (ad esempio, progettazione algoritmica e scrittura di codice modulare nel linguaggio C), progettazione e ottimizzazione iterativa dei prompt, test comparativi dei modelli IA, e valutazione della qualità pedagogica delle domande prodotte.

La fase di ottimizzazione dei prompt prevede tre strategie: prompt basati su esempi, istruzioni dettagliate senza esempi e prompt aperti basati su parole chiave. Studi precedenti mostrano che i prompt basati su esempi producono domande di maggiore qualità (accuratezza 87%) rispetto ai prompt generici (accuratezza 67%) (Nguyen & Allan, 2024).

Nella terza fase, sono stati valutati quattro chatbot generativi — ChatGPT (OpenAI, 2022), Gemini (Google, 2022), Copilot (Microsoft, 2021) e Claude (Anthropic, 2023) — attraverso diversi tipi di prompt per analizzarne coerenza e accuratezza.

Infine, le domande generate dall'IA sono state valutate secondo chiarezza, completezza, difficoltà e correttezza, assicurando il loro appropriato livello pedagogico.

3.2 Strategia di progettazione rapida e metriche di valutazione

La progettazione rapida è stata realizzata tramite tre cicli iterativi con diverse strategie di prompt engineering (Wisniewski et al., 2020). La prima, basata su esempi, prevedeva domande d'esame predefinite come modelli di riferimento per l'intelligenza artificiale, ottenendo domande coerenti e strutturate. La seconda strategia, fondata su istruzioni dettagliate senza esempi, ha prodotto risultati meno consistenti, con frequenti ambiguità o dettagli eccessivi che potevano limitare l'autonomia degli studenti (Raffel et al., 2020). La terza strategia, basata su prompt aperti, ha generato domande poco specifiche e con suggerimenti impliciti, risultando inefficace come metodo di valutazione.

I risultati confermano che i prompt basati su esempi producono domande di maggiore qualità rispetto agli altri approcci (Raffel et al., 2020). Studi futuri dovrebbero concentrarsi sull'ottimizzazione dei modelli IA, sulle tecniche di valutazione adattiva e su meccanismi automatizzati di validazione per migliorare ulteriormente l'efficacia educativa dei contenuti generati dall'IA.

La qualità delle domande d'esame generate dall'intelligenza artificiale è stata valutata secondo chiarezza, completezza, difficoltà e correttezza (Tabella 1). La chiarezza riguarda l'assenza di ambiguità, la completezza verifica la copertura degli obiettivi didattici, la difficoltà valuta l'adeguatezza rispetto al livello degli studenti, mentre la correttezza esamina coerenza logica e sintattica.

L'analisi ha mostrato che ChatGPT e Copilot producono domande più coerenti e prive di ambiguità rispetto a Gemini e Claude, che spesso includono suggerimenti impliciti riducendo l'efficacia pedagogica delle valutazioni. Tali risultati hanno consentito di migliorare ulteriormente la progettazione dei prompt per assicurare una migliore autonomia e integrità pedagogica nelle valutazioni generate (Brown et al., 2020; Luckin, 2018).

Tabella 1 - Metriche per la valutazione dei prompt

Descrizione delle metriche	Descrizione delle metriche
Chiarezza del prompt	Il problema è chiaramente definito e non ci sono ambiguità. Le funzionalità sono specifiche e dettagliate, senza lasciare spazio a interpretazioni incerte.
Completezza	Il problema riguarda tutti gli obiettivi di apprendimento previsti. Ogni requisito è pertinente e necessario per risolvere il problema.
Difficoltà del prompt	La difficoltà del prompt è coerente con le competenze attese dagli studenti.
Requisiti di codifica chiari	I requisiti di codifica sono definiti in dettaglio e non presentano ambiguità.

La fase conclusiva dello studio ha valutato l'efficacia delle domande d'esame generate dall'intelligenza artificiale attraverso feedback di educatori e studenti in ambito universitario. Gli educatori hanno evidenziato che alcuni modelli di IA fornivano dettagli eccessivamente specifici, limitando la capacità degli studenti di sviluppare soluzioni autonome.

Gli studenti hanno percepito come più chiare e pedagogicamente efficaci le domande basate su prompt strutturati con esempi, mentre quelle generate con prompt aperti risultavano spesso ambigue. Il

framework proposto è stato perfezionato integrando questi feedback per garantire migliore chiarezza e maggiore allineamento con gli obiettivi didattici, minimizzando le soluzioni predefinite.

I risultati confermano che i prompt strutturati migliorano significativamente la qualità pedagogica delle domande generate dall'IA, supportando la progettazione automatizzata di valutazioni efficaci e rigorose.

4 Risultati

La valutazione delle domande d'esame di programmazione generate dall'intelligenza artificiale è stata condotta utilizzando quattro modelli distinti: ChatGPT, Gemini, Copilot e Claude. L'analisi si è concentrata sulla capacità di ciascun modello di produrre domande coerenti, complete e pedagogicamente valide, adottando diverse strategie di prompt engineering. I risultati hanno evidenziato variazioni significative nella qualità delle domande generate, a seconda dell'approccio di prompt impiegato e del modello di intelligenza artificiale utilizzato.

L'approccio basato su prompt strutturati con esempi ha prodotto le domande d'esame più coerenti e pedagogicamente efficaci. In questo scenario, l'intelligenza artificiale è stata guidata da un modello di domanda predefinito, il che ha permesso di ottenere output chiari e ben strutturati. ChatGPT e Copilot si sono distinti per la capacità di mantenere la coerenza formale delle domande, preservandone la chiarezza e la pertinenza con gli obiettivi di apprendimento. Al contrario, Gemini e Claude hanno mostrato una tendenza a generare domande parziali o eccessivamente generiche, includendo talvolta suggerimenti di implementazione non necessari. Questo aspetto rappresenta una criticità, in quanto suggerimenti eccessivi possono limitare l'autonomia degli studenti nella risoluzione dei problemi. I risultati ottenuti sono coerenti con le ricerche precedenti sul feedback generato dall'IA, che evidenziano come prompt ben strutturati contribuiscano a migliorare l'accuratezza e la coerenza delle risposte prodotte.

L'uso di prompt basati esclusivamente su istruzioni, senza fornire esempi espliciti, ha portato a risultati più eterogenei. Sebbene questo approccio abbia garantito una maggiore flessibilità nella generazione delle domande, la mancanza di una struttura di riferimento ha determinato livelli di dettaglio e chiarezza variabili tra i diversi modelli di IA. ChatGPT e Copilot hanno dimostrato una maggiore capacità di produrre domande ben allineate agli obiettivi di apprendimento, mentre Gemini e Claude hanno frequentemente introdotto dettagli tecnici superflui o omesso vincoli essenziali del problema. Questa variabilità suggerisce che i modelli di IA interpretano le istruzioni in modi diversi, sottolineando la necessità di una progettazione precisa dei prompt per garantire la qualità delle valutazioni generate.

L'approccio basato su prompt aperti, in cui venivano forniti solo argomenti generali senza ulteriori vincoli, ha generato le domande meno strutturate e meno efficaci dal punto di vista valutativo. In numerosi casi, le domande prodotte mancavano della specificità necessaria, risultando poco adatte per una valutazione rigorosa delle competenze degli studenti. Inoltre, mentre Gemini e Claude hanno talvolta generato contenuti con informazioni irrilevanti o imprecise (hallucinations), ChatGPT e Copilot tendevano a formulare domande vaghe o incomplete. Questi risultati evidenziano il ruolo cruciale di una guida strutturata nella progettazione dei prompt, poiché l'assenza di vincoli specifici può portare a output altamente variabili e imprevedibili, compromettendo la validità delle valutazioni prodotte dall'intelligenza artificiale.

4.1 Valutazione dei contenuti generati dall'intelligenza artificiale

Per valutare sistematicamente la qualità degli schemi d'esame generati dall'intelligenza artificiale, ogni output è stato valutato in base a quattro criteri pedagogici: chiarezza, completezza, livello di

difficoltà e correttezza (Tabella 1). I risultati rivelano che ChatGPT e Copilot hanno costantemente superato Gemini e Claude nel mantenere la struttura, la chiarezza e la pertinenza dei contorni.

I feedback degli educatori hanno evidenziato che le risposte di ChatGPT rispecchiavano da vicino le valutazioni progettate dall'uomo, in particolare quando sono stati utilizzati suggerimenti basati su esempi. Tuttavia, alcune risposte richiedevano ancora piccoli perfezionamenti per migliorare la specificità. Copilot ha dimostrato un'efficacia simile, ma occasionalmente ha introdotto ambiguità sintattiche nelle dichiarazioni dei problemi.

Gemini, pur producendo risposte linguisticamente valide, spesso mancava di precisione nella definizione dei vincoli del problema, rendendolo meno affidabile per la generazione di esercizi di programmazione di alta qualità. Claude, nonostante sia stato progettato per l'intelligenza artificiale etica e la trasparenza, ha mostrato prestazioni incoerenti, introducendo spesso dettagli non necessari o omettendo requisiti di problemi critici.

Un problema comune a tutti i modelli di intelligenza artificiale era l'eccessiva specificazione dei dettagli di implementazione. In diversi casi, le domande d'esame generate includevano suggerimenti espliciti o soluzioni di codifica, che potevano minare lo scopo della valutazione. Questa tendenza era più pronunciata in Gemini e Claude, rafforzando la necessità di suggerimenti accuratamente elaborati per evitare che l'IA fornisse una guida eccessiva (Wang et al., 2023).

4.2 Feedback di educatori e studenti

Lo studio ha previsto una fase di raccolta di feedback, coinvolgendo docenti universitari e studenti nella valutazione dell'usabilità e della rilevanza educativa degli schemi d'esame generati dall'intelligenza artificiale. I risultati indicano che le domande basate su prompt strutturati con esempi erano generalmente ben organizzate e coerenti con gli obiettivi del corso, mentre quelle generate attraverso suggerimenti aperti risultavano meno efficaci per la valutazione delle competenze.

Dal punto di vista dei docenti, gli schemi generati dall'intelligenza artificiale sono stati considerati un utile punto di partenza per lo sviluppo degli esami, ma hanno evidenziato la necessità di un intervento umano per garantirne l'allineamento con gli obiettivi curriculari specifici. È emersa, inoltre, una preoccupazione riguardo ai potenziali pregiudizi nei contenuti prodotti dall'IA, in particolare per quanto riguarda la selezione dei problemi e la distribuzione del livello di complessità. Alcuni docenti hanno osservato che le domande generate dall'intelligenza artificiale non sempre riflettevano la diversità cognitiva tipica delle valutazioni degli studenti in contesti reali, sottolineando la necessità di un ulteriore affinamento prima della loro piena integrazione nei sistemi di valutazione accademici.

Dal lato degli studenti, è stata rilevata una significativa variabilità nei livelli di difficoltà delle domande d'esame generate dall'IA. Alcune risultavano eccessivamente semplicistiche, mentre altre erano percepite come troppo complesse e prive di un adeguato equilibrio nella progressione della difficoltà. Le domande generate utilizzando prompt basati su esempi sono state considerate le più chiare e strutturate, mentre quelle ottenute attraverso suggerimenti aperti spesso mancavano di una formulazione precisa del problema o di criteri di valutazione espliciti. Questi risultati evidenziano l'importanza di una progettazione attenta dei prompt per garantire che le domande prodotte dall'intelligenza artificiale siano pedagogicamente valide e adeguate ai livelli di competenza degli studenti.

4.3 Punti di forza e limiti degli schemi d'esame generati dall'intelligenza artificiale

I risultati dello studio evidenziano il potenziale degli schemi d'esame generati dall'intelligenza artificiale nel semplificare la progettazione delle valutazioni, offrendo agli educatori soluzioni rapide e scalabili. La capacità di produrre varianti multiple di domande in pochi secondi rappresenta un

vantaggio significativo, soprattutto nei contesti accademici che richiedono valutazioni su larga scala, contribuendo a ridurre il tempo necessario per la loro elaborazione.

Nonostante questi benefici, persistono diverse limitazioni che devono essere affrontate prima di un'adozione diffusa delle valutazioni generate dall'intelligenza artificiale. Una delle principali preoccupazioni riguarda l'affidabilità dei contenuti. Sebbene i modelli di IA siano in grado di generare domande sintatticamente corrette e contestualmente pertinenti, possono talvolta introdurre errori fattuali, ambiguità o dettagli irrilevanti. La tendenza dell'IA a generare hallucinations, particolarmente evidente nelle formulazioni di problemi complessi, rappresenta una sfida critica che richiede ulteriori perfezionamenti nei metodi di generazione.

Un'altra limitazione riguarda la necessità di supervisione umana nella validazione delle domande d'esame prodotte. Anche se l'impiego di prompt basati su esempi migliora la chiarezza e l'accuratezza delle domande, non elimina del tutto la necessità di una revisione manuale da parte degli educatori. La verifica dell'accuratezza dei contenuti, l'adeguamento dei livelli di difficoltà e il perfezionamento dei suggerimenti rimangono attività fondamentali per garantire che le domande generate siano pienamente allineate agli obiettivi didattici.

Le implicazioni etiche dell'uso dell'intelligenza artificiale nella generazione di valutazioni rappresentano un ulteriore ambito di attenzione. Il rischio che i modelli di IA perpetuino i pregiudizi presenti nei dati di addestramento potrebbe portare alla formulazione di domande che favoriscono determinati approcci alla risoluzione dei problemi, svantaggiando così studenti con stili di apprendimento differenti. Questo aspetto evidenzia la necessità di sviluppare quadri di valutazione trasparenti e meccanismi di controllo volti a monitorare e mitigare eventuali distorsioni involontarie nelle valutazioni automatizzate.

4.4 Implicazioni per la progettazione della valutazione basata sull'intelligenza artificiale

I risultati di questo studio si inseriscono nel dibattito in continua evoluzione sull'educazione assistita dall'intelligenza artificiale, offrendo un'analisi approfondita delle migliori pratiche per la progettazione di valutazioni automatizzate. Uno degli aspetti più rilevanti emersi riguarda il ruolo cruciale dell'ingegneria dei prompt nella qualità dei contenuti generati dall'intelligenza artificiale. L'utilizzo di suggerimenti strutturati basati su esempi si è rivelato determinante nel migliorare l'affidabilità e la coerenza delle domande d'esame, consentendo un allineamento più efficace con gli obiettivi di apprendimento stabiliti dai docenti.

Un altro elemento chiave evidenziato dalla ricerca è la necessità di sviluppare una maggiore alfabetizzazione in ambito IA tra gli educatori. Con l'integrazione crescente dell'intelligenza artificiale nei contesti didattici, diventa essenziale per gli insegnanti acquisire competenze specifiche nell'ingegneria dei prompt, nella selezione dei modelli di IA e nella convalida dei contenuti generati. Il rafforzamento di queste competenze non solo consente di sfruttare appieno il potenziale dell'IA nella progettazione delle valutazioni, ma garantisce anche il mantenimento di standard qualitativi elevati nel processo di apprendimento e valutazione.

Le ricerche future dovrebbero approfondire gli effetti a lungo termine delle valutazioni generate dall'intelligenza artificiale sui risultati di apprendimento degli studenti. È necessario indagare se tali valutazioni favoriscano una comprensione concettuale più approfondita o, al contrario, incoraggino un approccio mnemonico e superficiale allo studio. Inoltre, lo sviluppo di strumenti avanzati di valutazione adattiva basati sull'IA potrebbe aprire la strada a nuovi modelli di apprendimento personalizzato, in cui l'intelligenza artificiale sia in grado di modulare dinamicamente la complessità delle domande in base alle prestazioni degli studenti, contribuendo così a un'educazione più mirata ed efficace.

5 Conclusioni e sviluppi futuri

L'integrazione dell'Intelligenza Artificiale (AI) nell'istruzione sta trasformando gli approcci pedagogici tradizionali, in particolare nel campo della generazione automatizzata di valutazioni. Questo studio ha esaminato l'applicazione di modelli di intelligenza artificiale generativa, tra cui ChatGPT, Gemini, Copilot e Claude, per generare domande d'esame di programmazione, valutandone l'efficacia attraverso tecniche strutturate di prompt engineering. I risultati indicano che gli strumenti basati sull'intelligenza artificiale hanno il potenziale per supportare gli educatori automatizzando la creazione di materiali di valutazione strutturati e pedagogicamente validi, a condizione che i suggerimenti siano accuratamente realizzati e perfezionati.

I risultati dimostrano che il suggerimento basato su esempi produce le domande d'esame più coerenti e ben strutturate, molto simili a quelle create dagli educatori umani. Al contrario, i suggerimenti aperti hanno spesso portato a domande ambigue o eccessivamente ampie, prive della necessaria specificità richiesta per valutazioni efficaci. Sebbene i contenuti generati dall'intelligenza artificiale fossero generalmente sintatticamente e contestualmente accurati, lo studio ha identificato incongruenze tra i diversi modelli, con alcuni che generavano suggerimenti di implementazione eccessivi o mancavano di vincoli di problema. Questi risultati sono in linea con ricerche precedenti che evidenziano l'importanza di un'ottimizzazione tempestiva per massimizzare l'affidabilità degli output generati dall'intelligenza artificiale.

Nonostante i suoi vantaggi, la generazione di domande d'esame basata sull'intelligenza artificiale non è priva di limitazioni. Lo studio ha identificato diverse sfide, tra cui l'affidabilità dei contenuti, l'allineamento pedagogico e le preoccupazioni etiche relative ai pregiudizi e all'equità dell'IA. Le domande generate dall'intelligenza artificiale spesso richiedono una convalida e un perfezionamento umano per garantire che siano in linea con gli standard accademici e gli obiettivi di apprendimento. Inoltre, la questione dei pregiudizi nei contenuti generati dall'intelligenza artificiale richiede lo sviluppo di linee guida etiche e quadri di valutazione per evitare distorsioni involontarie delle valutazioni.

I risultati sottolineano che l'intelligenza artificiale non dovrebbe sostituire gli educatori umani, ma piuttosto fungere da strumento di assistenza per semplificare la creazione di valutazioni, migliorare l'efficienza e fornire opportunità di apprendimento adattivo. Il framework basato sull'intelligenza artificiale proposto per la generazione di domande d'esame offre una metodologia strutturata per la progettazione, la valutazione e il perfezionamento rapidi, contribuendo al discorso più ampio sull'integrazione responsabile dell'intelligenza artificiale nell'istruzione.

Sebbene questo studio fornisca informazioni significative sull’efficacia delle valutazioni generate dall’intelligenza artificiale, ulteriori ricerche sono necessarie per indagare il suo impatto a lungo termine sui risultati di apprendimento, sul carico di lavoro degli insegnanti e sul coinvolgimento degli studenti. Una delle principali direzioni per le future indagini riguarda lo sviluppo di modelli di intelligenza artificiale ottimizzati specificamente per le applicazioni educative. Attualmente, i modelli linguistici di grandi dimensioni (LLM), come ChatGPT e Gemini, vengono addestrati su set di dati generici, che potrebbero non essere sempre in linea con i requisiti specifici della valutazione educativa. La loro ottimizzazione mediante fine-tuning su corpora educativi specializzati potrebbe migliorare l’accuratezza e la pertinenza dei contenuti generati, rendendo le valutazioni più coerenti con gli obiettivi didattici.

Un ulteriore ambito di esplorazione riguarda la personalizzazione adattiva delle valutazioni prodotte dall’intelligenza artificiale. Le ricerche future dovrebbero analizzare in che modo l’IA possa regolare dinamicamente la complessità delle domande in base alle prestazioni degli studenti, contribuendo a un’esperienza di apprendimento più personalizzata e reattiva. L’implementazione di sistemi di apprendimento adattivo basati sull’intelligenza artificiale potrebbe consentire la modifica in tempo reale delle valutazioni, offrendo agli studenti esercizi su misura che rispondano alle loro esigenze specifiche e ai loro progressi nell’acquisizione delle competenze.

È altresì fondamentale approfondire le implicazioni etiche legate all'uso dell'intelligenza artificiale nella valutazione educativa. Considerati i potenziali rischi legati ai pregiudizi nei modelli e alla diffusione di informazioni errate, la ricerca futura dovrebbe concentrarsi sullo sviluppo di protocolli di trasparenza, metodi per il monitoraggio dell'equità e quadri di spiegabilità, al fine di garantire che i contenuti generati dall'IA rimangano imparziali e privi di distorsioni. Inoltre, l'integrazione di meccanismi di supervisione guidati dagli insegnanti potrebbe fornire un livello di controllo umano necessario per mitigare eventuali errori e migliorare l'affidabilità delle valutazioni.

La collaborazione interdisciplinare tra educatori, ricercatori di intelligenza artificiale e responsabili politici si rivela essenziale per guidare l'evoluzione dell'istruzione basata sull'IA. La ricerca futura dovrebbe esplorare la scalabilità dell'integrazione delle valutazioni generate dall'intelligenza artificiale, valutando come possano essere efficacemente implementate su larga scala all'interno delle istituzioni educative e delle piattaforme di apprendimento online. Studi longitudinali che analizzino l'efficacia delle valutazioni automatizzate nel corso di più cicli accademici potrebbero offrire approfondimenti sulle tendenze delle prestazioni degli studenti, sui livelli di coinvolgimento e sull'impatto educativo complessivo.

Affrontando queste sfide e ampliando il quadro metodologico proposto, la generazione di valutazioni basate sull'intelligenza artificiale potrebbe affermarsi come uno strumento strategico nell'istruzione moderna. Questa evoluzione offrirebbe esperienze di apprendimento più scalabili, efficienti e personalizzate, garantendo al contempo il mantenimento dell'integrità accademica e del rigore pedagogico come principi guida dell'innovazione tecnologica.

Bibliografia

Almasri, F. (2024). Exploring the Impact of Artificial Intelligence in Teaching and Learning of Science: A Systematic Review of Empirical Research. *Research in Science Education*, *54*(5), 977–997. https://doi.org/10.1007/s11165-024-10176-3

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv*. https://www.semanticscholar.org/paper/Language-Models-are-Few-Shot-Learners-Brown-Mann/90abbc2cf38462b954ae1b772fac9532e2ccd8b0

Chen, Y., Chen, H., & Su, S. (2023). *Fine-Tuning Large Language Models in Education*. 718–723. https://doi.org/10.1109/ITME60234.2023.00148

GoodfellowIan, Pouget-AbadieJean, MirzaMehdi, XuBing, Warde-FarleyDavid, OzairSherjil, CourvilleAaron, & BengioYoshua. (2020). Generative adversarial networks. *Communications of the ACM*. https://doi.org/10.1145/3422622

Kingma, D. P., & Welling, M. (2014). *Auto-Encoding Variational Bayes*. https://dare.uva.nl/search?identifier=cf65ba0f-d88f-4a49-8ebd-3a7fce86edd7

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210

Lodovico Molina, I., Švábenský, V., Minematsu, T., Chen, L., Okubo, F., & Shimada, A. (2024). Comparison of Large Language Models for Generating Contextually Relevant Questions. In

R. Ferreira Mello, N. Rummel, I. Jivet, G. Pishtari, & J. A. Ruipérez Valiente (A c. Di), *Technology Enhanced Learning for Inclusive and Equitable Quality Education* (pp. 137–143). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72312-4_18

Luckin, R. (2018). Machine Learning and Human Intelligence: The Future of Education for the 21st Century. In *UCL IOE Press*. UCL IOE Press.

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, *6*, 100199. https://doi.org/10.1016/j.caeai.2023.100199

Nakanishi, T. (2023). An Inquirer-Responder Architecture Using LLMs: Emulating Virtual Hearing Q&A in Education. *2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 526–533. https://doi.org/10.1109/WI-IAT59888.2023.00088

Nguyen, H., & Allan, V. (2024). *Using GPT-4 to Provide Tiered, Formative Code Feedback* (p. 964). https://doi.org/10.1145/3626252.3630960

Pozdniakov, S., Brazil, J., Abdi, S., Bakharia, A., Sadiq, S., Gašević, D., Denny, P., & Khosravi, H. (2024). Large language models meet user interfaces: The case of provisioning feedback. *Computers and Education: Artificial Intelligence*, *7*, 100289. https://doi.org/10.1016/j.caeai.2024.100289

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Vahlois, I. B. D., Ong, M. A. S., Llagas, B. F. T., Dela Cruz, R. G. L., & Chu, S. B. (2024). ChatGPT as a Learning Aid for an Introductory Programming Course. In C. Herodotou, S. Papavlasopoulou, C. Santos, M. Milrad, N. Otero, P. Vittorini, R. Gennari, T. Di Mascio, M. Temperini, & F. De la Prieta (A c. Di), *Methodologies and Intelligent Systems for Technology Enhanced Learning, 14th International Conference* (pp. 153–165). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-73538-7_14

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, T., Lund, B. D., Marengo, A., Pagano, A., Mannuru, N. R., Teel, Z. A., & Pange, J. (2023). Exploring the Potential Impact of Artificial Intelligence (AI) on International Students in Higher Education: Generative AI, Chatbots, Analytics, and International Student Success. *Applied Sciences (Switzerland)*, *13*(11). https://doi.org/10.3390/app13116716

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.03087