Research Trends in Humanities RTH 11 (2024) – ISSN 2284-0184
Sezione *BEC Bio-Education & Cognition*

I Large Language Model sono capaci di valutare i compiti scritti degli studenti?
Uno studio pilota in Università

Daniele Agostini
Università di Trento

Introduzione
Fin dal pubblico rilascio di ChatGPT, il 30 novembre 2022, e, conseguentemente, di tutti i suoi concorrenti, i Large Language Model (LLM) si sono diffusi per un ampio utilizzo da parte del pubblico. Una delle aree nelle quali il loro impiego ha avuto dal primo momento un impatto maggiore è quella dell'Educazione e dell'Istruzione (Baytak, 2023; Elbanna, Armstrong, 2023; Extance, 2023; Roy et al., 2023; Saif et al., 2023; Tiwari et al., 2023). Di particolare interesse per l'oggetto di questo paper è l'utilizzo che ne è stato fatto fin dai primi momenti nell'ambito dell'istruzione superiore, sia da parte dei docenti che da parte degli studenti (Perkins, 2023; Roy et al., 2023; Sullivan et al., 2023).

La rapidità con cui i Large Language Models (LLM) si sono integrati nel tessuto dell'istruzione superiore, sia a livello dei docenti che degli studenti, solleva domande fondamentali sulla loro efficacia e affidabilità. In questo contesto, gli LLM promettono di rivoluzionare il modo in cui gli insegnanti interagiscono con gli studenti, gestiscono il carico di lavoro e personalizzano l'esperienza di apprendimento (Elbanna, Armstrong, 2023).

Mentre si rileva il potenziale per un miglioramento nell'accessibilità e nella personalizzazione dell'apprendimento, si pone la questione critica della loro capacità di valutare in modo oggettivo e imparziale le prestazioni degli studenti. Tale applicazione nella valutazione dell'apprendimento rimane un campo relativamente inesplorato, con implicazioni significative per la pratica educativa e la teoria pedagogica.

Questo studio esplora l'impiego dei principali LLM nel contesto specifico della valutazione degli elaborati degli studenti, con un focus sulla loro precisione capacità di valutare rispettando una rubrica di valutazione elaborata dal docente.

Contesto
Dalla diffusione di ChatGPT, avvenuta più di un anno fa, e a seguito del rilascio di modelli concorrenti, i Large Language Models (LLM) hanno iniziato a giocare un ruolo significativo nel panorama tecnologico. Sebbene gli LLM fossero già esistenti da tempo, il loro impatto è rimasto relativamente limitato fino all'introduzione di interfacce utente semplici e intuitive, come il livello di "chat", che hanno reso questi strumenti più accessibili al grande pubblico. Questa democratizzazione è stata catalizzatrice dell'uso commerciale e generale degli LLM, con un conseguente incremento degli investimenti in questo settore da parte di istituzioni, aziende e privati (Babina et al., 2023; Bloomberg, 2023; Hammond, 2023; Lee et al., 2022).

OpenAI ChatGPT, Anthropic Claude, Microsoft Copilot e Google Bard sono solo alcuni degli LLM più utilizzati, inoltre a questi si sono aggiunti i molto più numerosi modelli opensource ai quali LLAMA di Meta ha dato un grande impulso. Parallelamente, si è assistito a una crisi dei motori di ricerca, con gli LLM che offrono nuove modalità di interrogazione e analisi della conoscenza e dei dati, un'interazione più naturale e delle risposte abbastanza precise ed esaustive senza richiedere abilità di ricerca avanzata e soprattutto senza tutti passaggi che nei normali motori di ricerca separano il bisogno di una particolare informazione dal suo reperimento (ad esempio liste di siti web da selezionare, accettazione di cookies, banner pubblicitari).

Anche le istituzioni ed agenzie educative hanno risposto a questa tendenza, integrando LLM e AI generativa a vari livelli dei loro curricula e offrendo corsi specifici per sfruttare queste nuove tecnologie. C'è stata una presa di coscienza della necessità di una AI Literacy che permetta ai professionisti di vari settori, compreso quello dell'istruzione, di conoscere le caratteristiche principali dell'AI generativa, i vari strumenti a disposizione, le loro capacità e modalità di impiego nei rispettivi settori (Biagini et al., 2023; Cetindamar et al., 2024; Kong et al., 2023; Wang et al., 2023; Weber et al., 2023).

Tuttavia, questo rapido sviluppo ha anche sollevato questioni critiche relative al trattamento dell'informazione. Mentre gli LLM offrono potenzialità enormi in termini di analisi e generazione di dati, emergono preoccupazioni riguardo all'accuratezza, alla privacy e all'etica nella gestione delle informazioni e nella proprietà degli output. Queste sfide rappresentano un campo in continua evoluzione, richiedendo un'attenzione costante e una valutazione critica per garantire un impiego responsabile degli LLM (Gerdes, 2022; Jang, 2023; Majeed, Hwang, 2023; Samuelson, 2023).

La prima, breve, reazione degli enti di istruzione superiore è stata di difesa. Alcune università sono tornate agli esami scritti a mano e alle interrogazioni orali per contrastare un possibile utilizzo di LLM da parte degli studenti per lo svolgimento delle prove d'esame (Perkins, 2023; Yeo, 2023). Di pari passo il mercato ha iniziato ad offrire software per l'individuazione dei compiti scritti da LLM. Tali software si sono rivelati assolutamente inefficaci e hanno causato problematiche gestionali e legali agli istituti nel momento in cui degli studenti sono stati a torto accusati di usare aver fornito testi generati da una AI (van Oijen, 2023; J. Wang et al., 2023; Weber-Wulff et al., 2023).

Enti nazionali ed internazionali, così come gruppi di università, hanno fornito prontamente linee guida che, pur mantenendo alcune attenzioni, sono andate nella direzione di accettare l'utilizzo degli LLM in modo etico ed efficace per i compiti che possono portare vantaggio alle istituzioni, ai docenti e agli studenti. Alcuni esempi significativi sono UNESCO (Miao et al., 2023; Sabzalieva, Valentini, 2023) , JISC National Centre for AI (Webb, 2023), Russell Group (Russell Group, 2023), Ministero dell'Educazione Nazionale francese (GTnum, 2023), il Dipartimento per l'Educazione degli USA (Miguel A. Cardona et al., 2023) e University City London (UCL, 2023)

Un campo per il quale viene riconosciuto un potenziale grande beneficio, soprattutto in fatto di sostenibilità, è quello della valutazione. Allo stesso tempo si indica la necessità di molta cautela in quanto gli LLM generalisti non sembrano ancora in grado di gestire autonomamente la valutazione dei compiti degli studenti (Swiecki et al., 2022; Webb, 2023), mentre LLM aggiustati per tale compito sembrano poter dare buoni risultati (Martin et al., 2023). Anche in questo caso poi, come per gli studenti ci sono delle responsabilità e delle considerazioni etiche sulla produzione di elaborati con AI, anche i docenti hanno responsabilità e doveri etici riguardo la valutazione dei compiti, dalla quale può dipendere la carriera degli studenti (motivazione personale, media, borse di studio, accettazione a master o dottorati, ecc).

Quadro teorico

L'idea di utilizzare l'AI per aiutare il docente nei suoi compiti e ad arrivare a delle decisioni precise, prive di bias e informate è presente in molta letteratura a partire dagli anni '80 (Lepage, Roy, 2023).

L'opportunità di utilizzare gli LLM per la valutazione degli apprendimenti viene analizzata anche nell'era immediatamente precedente a ChatGPT, nella quale tuttavia i modelli transformer, compreso GPT3 di OpenAI, erano già ben presenti, da Tamkin et al. (Tamkin et al., 2021), sottolineando come fra i loro utilizzi in capo educativo ci fossero quelli di:

- Riepilogare: i LLM possono riassumere lunghe porzioni di testo. Ciò può aiutare a fornire riepiloghi concisi di lunghi invii da parte degli studenti. La sintesi può tenere conto di diversi

parametri presenti nel testo, fornendo informazioni esattamente sugli aspetti che il docente vuole valutare.

- Porre e rispondere a domande: i LLM possono comprendere un pezzo di testo e rispondere a domande al riguardo, nonché porre domande al riguardo, se richiesto. Questo può essere utilizzato per creare feedback interattivi ed esperienze di apprendimento.
- Classificare: gli LLM possono classificare il testo in categorie predefinite. Questo può essere utilizzato per la valutazione assistita o per classificare il feedback degli studenti.
- Rilevamento del plagio: confrontando la somiglianza tra diversi pezzi di testo, i LLM possono aiutare a rilevare potenziali casi di plagio sia tra studenti che tra studenti e materiale originale.
- Misurare la somiglianza semantica: gli LLM possono misurare la somiglianza semantica tra due parti di testo. Questo può essere utilizzato per abbinare le domande degli studenti con risposte o risorse pertinenti e aiutare l'insegnante nella valutazione del lavoro dello studente.
- Generare feedback: sulla base della valutazione del lavoro di uno studente, i LLM possono generare feedback personalizzato. Funzionerebbe ancora meglio se il LLM avesse degli appunti dell'insegnante sul compito su cui lavorare.
- Valutare la conoscenza: gli LLM possono essere utilizzati per valutare la comprensione di un argomento da parte di uno studente in base alle sue comunicazioni scritte, soprattutto se adeguatamente formati sui compiti corretti e avendo una rubrica di valutazione a cui fare riferimento.

Ognuna di queste sette applicazioni è fondamentale per l'impiego degli LLM nel campo della valutazione degli apprendimenti.

Infine, successivamente all'introduzione di ChatGPT e degli altri LLM accessibili a tutti, l'UNESCO ha diffuso le linee guida "AI and education: Guidance for policy-makers" (Miao et al., 2023) che suggeriscono le seguenti azioni per quanto riguarda la valutazione degli apprendimenti:

1. Testare e implementare le tecnologie di intelligenza artificiale per supportare la valutazione di varie dimensioni di competenze e risultati.
2. Cautela quando si adotta una valutazione automatizzata con risposte a domande chiuse basate su regole.
3. Utilizzare la valutazione formativa coadiuvata dall'intelligenza artificiale come funzione integrata dei sistemi di gestione dell'apprendimento (LMS) in modo da analizzare i dati di apprendimento degli studenti con maggiore precisione ed efficienza e ridurre i pregiudizi umani.
4. Valutazioni progressive basate sull'intelligenza artificiale per fornire aggiornamenti regolari a insegnanti, studenti e genitori.
5. Esaminare e valutare l'uso del riconoscimento facciale e di altre intelligenze artificiali per l'autenticazione e il monitoraggio degli utenti nelle valutazioni online remote.

Lavorando su questi approcci teorici, indicazioni e linee guida è stato sviluppato il modello AI-MAAS (AI-Mediated Assessment Academics and Students), al momento in fase di validazione, che prevede due possibili implementazioni degli LLM per la valutazione degli apprendimenti: l'implementazione di valutazione mediata per la valutazione formativa e quella per la valutazione sommativa (Agostini, Picasso, 2023).

Per entrambe queste implementazioni è importante che l'LLM scelto sia in grado di valutare seguendo una rubrica di valutazione fornita dal docente o degli studenti.

Le esperienze in questo senso non sono ancora molte. Martin et al. (Martin et al., 2023) hanno lavorato su questa possibilità partendo dall'esigenza di dare compiti di ragionamento, concettualizzazione ed elaborazione, e dal fatto che la correzione di grandi quantità di risposte e compiti aperti spesso si rivela essere non sostenibile. I ricercatori hanno quindi dimostrato in un compito di chimica che gli LLM possono essere usati a questo scopo. In questo caso, infatti, è stata raggiunta una corrispondenza quasi perfetta fra i punteggi dati dagli esseri umani e dall'LLM. È da sottolineare tuttavia che Martin et al. non si sono limitati all'uso di un LLM per raggiungere questo risultato. Essi, infatti, hanno seguito la seguente procedura:

1. Utilizzato la tecnica di apprendimento automatico non supervisionato HDBSCAN (Clustering spaziale basato sulla densità gerarchica di applicazioni con rumore) per raggruppare gli argomenti degli studenti in diversi gruppi. Ciò ha aiutato a scoprire schemi nelle discussioni.
2. Mappato i risultati del clustering su strutture guidate dalla teoria delle modalità di ragionamento e dei livelli di granularità per creare una rubrica di punteggio olistico con 20 categorie.
3. Assegnato un punteggio a tutti gli argomenti manualmente in base alla rubrica per creare un set di dati etichettato.
4. Comparato le prestazioni di diversi modelli linguistici preaddestrati (BERT, RoBERTa, SciBERT) sulla classificazione degli argomenti nelle 20 categorie di rubriche. BERT large uncased ha ottenuto i migliori risultati.
5. Addestrato un classificatore di *deep neural network* utilizzando il set di dati etichettato per automatizzare il punteggio di nuovi argomenti in base alla rubrica. Il modello ha raggiunto una precisione dell'87% su un set di test prolungato.
6. Convalidato il modello utilizzando tecniche come la generazione di argomenti artificiali, la conduzione di analisi della black box e il calcolo dei punteggi di importanza delle caratteristiche. Ciò ha contribuito a garantire che il modello si basasse su parole chiave simili a quelle dei valutatori umani.

L'ottimo risultato è quindi frutto di modelli allenati su un compito e una popolazione specifica e non ci si può aspettare che la procedura applicata venga utilizzata da un qualsiasi docente non specializzato in Machine Learning. L'idea di questo studio infatti era di presentare un modello operativo e dimostrarne la fattibilità.

Altri studi hanno impiegato LLM per la valutazione ma senza confrontare la valutazione data dalla AI con quella data da un docente. Questi studi hanno avuto risultati soddisfacenti nella valutazione di compiti di Inglese L2 (Koraishi, 2023) e nel supportare l'autovalutazione (Ali et al., 2023). Esistono altri esempi di utilizzo del Machine Learning nella valutazione di compiti legati alle STEM, ma senza l'utilizzo di LLM (Ouyang et al., 2023).

In questo articolo si presenta uno studio pilota sulle capacità dei maggiori LLM di valutare compiti autentici a partire da una rubrica di valutazione. L'individuazione di un LLM particolarmente capace in questo compito sarebbe un buon punto di partenza per dare seguito a sperimentazioni per la valutazione supportata dalla AI.

L'utilizzo di LLM per la valutazione nell'istruzione superiore potrebbe infatti permettere di utilizzare approcci didattici e valutativi che prima non potevano essere sostenibili e scalabili, garantendo un maggiore rispetto dell'allineamento costruttivo (Biggs, 1996) e dunque un miglioramento della qualità e dell'efficacia della didattica universitaria.

Metodologia e strumenti

Questo studio esplora l'impiego dei principali LLM nel contesto specifico della valutazione degli elaborati degli studenti, con un focus sulla loro precisione capacità di valutare rispettando una rubrica di valutazione elaborata dal docente. L'obiettivo è di comprendere se e quali modelli possano essere utilizzati da docenti, universitari e non, non esperti di Machine Learning, per valutare i prodotti scritti degli studenti, anche in presenza di compiti e domande aperte, grazie rubriche di valutazione.

Lo studio pilota si è svolto all'Università di Trento nel contesto di un corso universitario di specializzazione per le attività di sostegno didattico agli alunni con disabilità, nel modulo riguardante l'utilizzo delle nuove tecnologie per l'inclusione. Vi hanno preso parte anonimamente 88 studenti suddivisi in 21 gruppi e 2 docenti valutatori, esperti in pedagogia sperimentale e valutazione. Non è stato raccolto alcun dato riguardo l'anagrafica degli studenti. Ai gruppi è stato richiesto di svolgere un compito autentico, ovvero di progettare un intervento didattico mirato ad un a determinata classe (che poteva andare dalla 1° della scuola primaria alla 5° della scuola secondaria di secondo grado, a seconda della composizione del gruppo) che tenesse conto dell'integrazione delle tecnologie educative e dell'inclusione degli alunni con bisogni educativi speciali. Tale intervento didattico si sarebbe potuto svolgere in due o più incontri. Per portare a termine tale compito sono state date ai gruppi due ore e trenta minuti di tempo ed è stato fornito un template per la progettazione didattica costituito dalle seguenti sezioni: Discipline coinvolte, Classe e grado scolastico, Titolo intervento, Obiettivi e risvolti pratici (outcome), Contesto e ambiente (formale, informale, tipo di setting, ecc), Tecnologie previste con pro e contro e tipi di utilizzo, Valutazione e scansione con dettagli. Nella parte di scansione con i dettagli si chiedeva di esplicare la programmazione con descrizioni sintetiche delle varie attività didattiche, il compito del docente e quelli degli alunni. All'interno di questo quadro i gruppi avevano facoltà di proporre la propria programmazione originale. Il prodotto finale di ogni gruppo è quindi costituito da un file word contenente la programmazione dell'intervento didattico secondo il template descritto.

Per la valutazione dei prodotti è stata approntata la seguente rubrica di valutazione (Tab.1) consistente in cinque criteri di valutazione con cinque livelli per ogni criterio.

Tabella 1. Rubrica per la Valutazione dell'Intervento Didattico

Criteri di Valutazione	**Assente (0 punti)**	**Livello Basso (1 punto)**	**Livello Intermedio (2 punti)**	**Livello Alto (3 punti)**	**Livello Avanzato (4 punti)**
Obiettivi e Risvolti	Obiettivi assenti	Obiettivi poco chiari o incoerenti	Obiettivi chiari ma non pienamente integrati	Obiettivi chiari e ben integrati	Obiettivi eccellenti e perfettamente integrati con risvolti innovativi
Uso e Interazione con le Tecnologie e AI Generativa	Uso e interazione con tecnologie assenti	Uso e interazione limitati o inadeguati, assenza di AI generativa	Uso e interazione adeguati, AI generativa presente ma non sfruttata appieno	Buon uso e interazione, inclusione efficace di AI generativa	Uso e interazione eccellenti e innovativi, sfruttamento ottimale di AI generativa
Progettazione: Coerenza con obiettivi, Coerenza interna e Originalità	Progettazione assente	Progettazione incoerente, non in linea con gli obiettivi dichiarati o poco originale	Progettazione coerente ma con originalità e innovazione limitata	Progettazione sia coerente che originale/ innovativa	Progettazione estremamente coerente e altamente originale/ innovativa
Inclusività	Pensiero inclusivo assente	Inclusività didattica inadeguata o non efficace	Pensiero inclusivo presente ma senza particolari elementi distintivi o poco sviluppato	Inclusività proattiva e ben integrata con la progettazione	Spiccata inclusività, integrata in tutta la progettazione e le attività dell'intervento
Valutazione	Valutazione assente	Piani e metodi di valutazione non chiari o non adeguati	Piani e metodi di valutazione abbastanza adeguati, ma non chiarissimi	Buoni piani e metodi di valutazione, coerenti e con qualche elemento innovativo	Piani di valutazione eccellenti e altamente innovativi e coerenti con gli obiettivi didattici

Tutti i prodotti dei gruppi di studenti sono stati successivamente valutati da due valutatori umani esperti e cinque LLM. Gli LLM selezionati per questo studio sono stati quelli concorrenti più diffusi al momento, più un outsider:

1. **ChatGPT 3.5-turbo di OpenAI**: la versione migliorata del modello di rilascio iniziale di ChatGPT. Al momento è il modello disponibile per la versione libera di ChatGPT ed è ancora uno dei migliori modelli utilizzabili. Link: https://chat.openai.com/

2. **ChatGPT 4 di OpenAI**: È l'evoluzione di ChatGPT 3.5-turbo, con capacità di comprensione e generazione del testo notevolmente migliorate. Questo modello offre risposte più accurate e dettagliate, rendendolo adatto per un'ampia varietà di applicazioni. Questo modello è disponibile solamente per i clienti che pagano il piano premium. Link: https://chat.openai.com/

3. **Claude 2 Chat di Anthropic**: Una versione avanzata del modello di intelligenza artificiale di Anthropic. Claude 2 si distingue per la sua capacità di comprendere e rispondere a domande complesse con un alto livello di accuratezza e sensibilità contestuale. Infatti, la sua maggiore caratteristica è proprio quella di potergli inviare dei file per l'elaborazione e di avere un ampissimo contesto, ovvero la capacità di "leggere" e tenere conto di prompt e file testuali molto lunghi (fino 100.000 token, approssimabili a 75.000 parole). L'uso di questo modello è gratuito ma limitato 50 messaggi al giorno. La versione a pagamento non è disponibile in Italia. Link: https://claude.ai

4. **Bing Chat/Autopilot di Microsoft**: Un modello di intelligenza artificiale integrato col motore di ricerca Bing. Bing Chat, attivando l'opzione corretta, si basa su un modello GPT-4, lo stesso tipo di modello utilizzato da ChatGPT 4. È particolarmente utile per la sua capacità intrinseca di effettuare ricerche sul web. Inoltre, si può utilizzare gratuitamente. Link: https://www.bing.com/chat

5. **Bard di Google**: Il modello AI di Google, noto per la sua integrazione con l'enorme database di informazioni di Google. Bard è progettato per fornire risposte rapide e accurate a una vasta gamma di domande, sfruttando la vasta conoscenza disponibile su Internet. Si tratta di un modello gratuito, ma sembra mancare dell'accuratezza e delle capacità di ragionamento degli altri modelli presentati. Link: https://bard.google.com/chat

6. **OpenChat 3.5 di OpenChat/AlignmentLabs**: Un modello di intelligenza artificiale opensource di soli 7 miliardi di parametri (per referenza, ChatGPT 3.5 si stima abbia 175 miliardi di parametri) che può funzionare anche in locale sui computer desktop. In molti benchmark per LLM raggiunge i livelli di ChatGPT 3.5. Trattandosi un modello OpenSource è gratuito. Link: https://openchat.team

Tutti questi LLM possono "comprendere" e scrivere in italiano, anche se non è da escudere che le prestazioni in inglese possano essere diverse (presumibilmente migliori dato che la maggior parte dell'allenamento è fatto in quella lingua). Inoltre, in questo caso si trattava di compiti totalmente anonimi e privi di dati sensibili, tuttavia la maggior parte di questo modelli fornisce soluzioni dal punto di vista della privacy. Entrambi i modelli di OpenAI e anche Bing Chat danno la possibilità di non salvare le conversazioni e non utilizzarle per l'allenamento del modello. Claude 2 di Anthropic non effettua nessun allenamento sui dati delle chat, mentre Google Bard permette di non registrare l'attività e le conversazioni e gestire i dati salvati. OpenChat 3.5 non salva nulla nella sua versione online, inoltre, essendo opensource, è pensato per essere utilizzato in locale sul proprio computer: in tal modo non verrà trasmesso in Internet alcun dato.

Prompting

L'obiettivo di questo studio è comprendere quali modelli potrebbero essere utilizzati da docenti universitari (e non) per valutare i prodotti degli studenti. Per questo motivo non sono state utilizzate tecniche di prompting eccessivamente sofisticate, bensì quello che un docente potrebbe fare fornendo istruzioni chiare e dando i dati di contesto necessari alla valutazione.

Di seguito i due prompt che sono stati dati agli LLM per eseguire la valutazione dei prodotti:

Prompt 1:

Il seguente documento è un compito svolto dagli studenti. Dovrai correggerlo dopo avere ricevuto nel mio secondo prompt la rubrica di valutazione. Se hai capito, rispondi solo "capito" a questo primo prompt, senza aggiungere altro.

Nota che i seguenti testi fanno parte delle istruzioni date agli studenti per le varie sezioni del documento e non sono prodotti da essi:

<inizio testi già presenti come istruzioni nel documento>

[qui sono forniti tutti i testi già presenti nel template]

<fine testi già presenti nel documento come istruzioni>

<inizio documento prodotto dagli studenti>
[qui viene incollato il documento prodotto dai gruppi]

<fine documento prodotto dagli studenti>

Prompt 2:

Valuta l'intervento didattico che ho fornito nel primo prompt e che è stato creato da degli studenti del corso di specializzazione per il sostegno. La competenza chiave di questo compito risiedeva nel saper programmare un intervento didattico che fosse inclusivo per gli studenti e che utilizzasse in questo senso anche le tecnologie a disposizione (meglio se venivano usate tecnologie di AI generativa). Allo stesso tempo la progettazione didattica si doveva dimostrare efficace per arrivare agli obiettivi che loro stessi si sono posti. Gli studenti non hanno avuto molto tempo, quindi non veniva richiesto di programmare in grande dettaglio.

Utilizza la rubrica valutativa seguente per assegnare i punteggi e poi presenta la valutazione in forma di lista dei criteri della rubrica con i relativi punteggi.

<inizio rubrica di valutazione dell'Unità di Apprendimento>

Rubrica di valutazione UdA:

Criterio di Valutazione: Obiettivi e Risvolti (Objectives and Outcomes)

- *Assente (assegnare 0 punti): Obiettivi assenti*
- *Livello Basso (assegnare 1 punto): Obiettivi poco chiari o incoerenti*
- *Livello Intermedio (assegnare 2 punti): Obiettivi chiari ma non pienamente integrati*
- *Livello Alto (assegnare 3 punti): Obiettivi chiari e ben integrati*
- *Livello Avanzato (assegnare 4 punti): Obiettivi eccellenti e perfettamente integrati con risvolti innovativi*

[… continua con gli altri criteri…]

<fine rubrica di valutazione dell'Unità di Apprendimento>

Il primo prompt è stato diversificato per Chat GPT-4 e Claude 2 che hanno la possibilità di accettare documenti allegati, quindi per questi due modelli il primo prompt non conteneva il testo del prodotto dei gruppi bensì il relativo file allegato. Bing Chat non supporta documenti di testo allegati ai prompt, ma può leggere documenti aperti con il browser Microsoft Edge, ed è stato utilizzato in questo modo; anche in questo caso quindi nel primo prompt non è stato necessario copiare il testo del prodotto.

Infine, per tutti gli LLM si è utilizzata una procedura di prompting *zero-shot*, cioè non si è dato ai modelli nessun esempio di valutazione dei compiti da parte degli esseri umani. Un docente universitario potrebbe effettivamente dare questo tipo di esempi per migliorare la qualità delle valutazioni dell'LLM, ma in questo caso l'obiettivo era di selezionare i modelli più "portati" a questo tipo di valutazione, con cui poter iniziare a portare avanti una ricerca più approfondita, e solo successivamente studiare il modo di ottimizzarne i risultati.

Attenzione ai token e al contesto

Quando si utilizza un LLM, è essenziale comprendere alcune caratteristiche che definiscono le possibilità di utilizzo di un determinato modello.

Il primo concetto da tenere a mente è quello di token. Per semplificare, i token possono essere considerati in questo contesto come unità di testo. Un token può essere composto da una parola, un pezzo di parola o, al limite, anche da un singolo carattere. Le caratteristiche dei token possono variare da modello a modello ma, in generale, è abbastanza prudente assumere che, in media, per l'inglese ci vogliano da un token a un token e mezzo e per l'italiano da un token e mezzo a due per ogni parola.

Il secondo concetto da considerare è la finestra di contesto (context window). Essa rappresenta la quantità di token che un modello di linguaggio può considerare contemporaneamente per la generazione delle risposte. Il contesto può dipendere dal modello utilizzato e dalla memoria a disposizione. Se andassimo oltre al contesto di un determinato modello, questo provocherebbe un errore nel caso che accada in un unico prompt, oppure, nel corso di una conversazione più lunga, potrebbe semplicemente non tenere più in considerazione le prime parti del dialogo per fare spazio alle più recenti. È chiara dunque l'importanza che esso ha per la generazione di risposte coerenti e pertinenti.

Inoltre, bisogna tenere conto che non sono solo i prompt dell'utente ad occupare contesto, ma anche le risposte del modello.

Questo è il motivo per il quale nel primo prompt si è deciso di chiedere al modello di rispondere solamente "capito" nel caso fosse tutto chiaro. In questo modo si è cercato di preservare più contesto possibile per l'elaborazione della risposta finale sulla valutazione. In mancanza di questa gli LLM tendevano a scrivere delle lunghe risposte che andavano già ad analizzare il prodotto senza ancora avere la rubrica a disposizione.

Per preservare la finestra di contesto, alcuni LLM impongono un limite di caratteri ai prompt che si possono inviare e alla lunghezza delle risposte generate che sono inferiori all'ampiezza massima. Questo è il motivo per il quale si è optato per la suddivisione delle istruzioni in due prompt separati.

Di seguito una tabella che illustra l'ampiezza massima della finestra di contesto per ognuno dei modelli utilizzati:

Tabella 2. Finestre di contesto degli LLM utilizzati. Per tutti i modelli le finestre di contesto si intendono riguardo le versioni Chat e non riguardo le API. Attenzione, questa caratteristica può variare con gli aggiornamenti.

Large Language Model (versioni disponibili in Italia, novembre 2023)	**Finestra di contesto (in token)**
ChatGPT 3.5-turbo	8.192
ChatGPT 4	32.000
Claude 2	100.000
Bing Chat/Autopilot	Non dichiarata, c.a. 13.500 (prompt massimo 4000 caratteri)
Google Bard	Non dichiarata, c.a. 1024
OpenChat 3.5	8.192

Metodi di analisi

I dati di valutazione che sono che sono stati analizzati allo stato attuale sono i livelli assegnati da ogni valutatore (LLM e umani) ai vari criteri della rubrica, per ognuno dei 21 prodotti dei gruppi. Ognuno dei 7 valutatori ha assegnato ad ogni prodotto un livello per ognuno dei 5 criteri. Ogni valutatore ha quindi assegnato un livello a un totale di 105 criteri.

Per ottenere informazioni dai dati sono state utilizzate alcune tecniche statistiche quali una Principal Component Analysis (PCA), l'analisi della varianza e la creazione di un indice di disaccordo fra valutatori. Per le analisi statistiche si sono utilizzati i software Microsoft Excel e JASP (basato su R).

Risultati

Dopo aver analizzato a vista tutti i risultati, prima di procedere con ulteriori analisi, si è deciso di escludere Google Bard. Tale LLM ha infatti assegnato un punteggio alto (3) il 93,3% delle volte e intermedio (2) per la restante percentuale, dimostrando di non essere stato in grado di valutare con efficacia alcun prodotto, probabilmente anche a causa della sua limitata finestra di contesto.

Un altro controllo che è stato implementato durante le prove con i diversi LLM riguarda la consistenza della valutazione. Questa è stata effettuata chiedendo per tre volte la valutazione dello stesso prodotto ad ogni LLM, avviando ogni volta una nuova sessione. Da queste prove abbiamo accertato i seguenti comportamenti:

- ChatGPT-4 è sempre stato coerente. Nessuna variazione di valutazione nei tre tentativi.
- ChatGPT-3.5-turbo è stato incoerente. Varia la valutazione di due o più criteri, di uno o due punti.
- Bing Chat / Autopilot è sempre stato coerente. Nessuna variazione di valutazione nei tre tentativi.

- Claude 2 è quasi sempre coerente. Variazione di un criterio di un punto.
- OpenChat 3.5 è stato incoerente. Variazione di due criteri, di un punto.

Claude 2 si è dimostrato l'unico capace di valutare più di un prodotto allo stesso tempo mantenendo la coerenza, probabilmente grazie all'ampia finestra di contesto.

Bing Chat è stato l'unico che alla comunicazione che si trattava di prodotti di studenti avvisava l'utente che nessun dato della corrente conversazione sarebbe stato salvato a causa del materiale potenzialmente sensibile.

Principal Component Analysis

La prima analisi effettuata, oltre ai dati descrittivi, è stata la PCA, una tecnica di riduzione dimensionale che ci permette di individuare delle variabili latenti all'interno dei dati e che può rappresentare un modello generale dei dati .

Dalla PCA effettuata sui dati vengono individuati tre componenti principali (Tab. 3).

Tabella 3. Loadings dei Componenti della PCA

	RC1	RC2	RC3	Uniqueness
e3	0.687		-0.455	0.347
e4	0.654			0.481
e6	0.652			0.498
e2	0.462			0.684
e0		0.838		0.319
e1		0.698		0.393
e7			0.921	0.154

Nota. Il metodo di rotazione applicato è promax.

Il primo componente (RC1) è formato dai loading dei valutatori e2, e3, e4 ed e6 che corrispondono rispettivamente agli LLM ChatGPT-4, ChatGPT-3.5, Claude 2 e Bing Chat. Il secondo componente (RC2) invece è costituto da quelli di e0 ed e1, corrispondenti ai valutatori umani 1 e 2. Infine, il terzo componente (RC3) è costituto principalmente da e7, OpenChat 3.5, e dal loading negativo di Chat GPT-3.5.

Come si può apprezzare in Fig. 1, ChatGPT-4 contribuisce in misura minore degli altri al componente RC1. Cercando di dare un nome ai componenti individuati, RC1 si potrebbe chiamare "Valutazione LLM proprietari / ad alto numero di parametri", RC2 "Valutazione Umani", RC3 "Valutazione LLM OpenSource a basso numero di parametri".

Dal grafico si può anche apprezzare come ChatGPT-4 contribuisca positivamente anche a RC2 con un loading di circa 0,3.

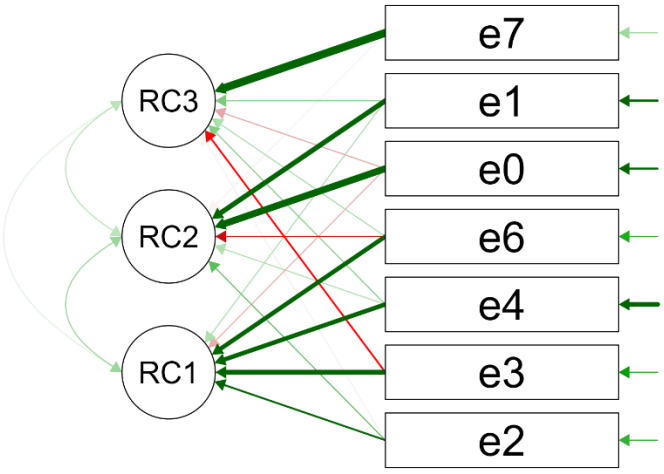


Figura 1: Diagramma dei loading della PCA

Analisi della varianza dei punteggi per prodotto e criterio di valutazione

Per comprendere come siano differite le valutazioni da criterio a criterio e da valutatore a valutatore si è proceduto ad analizzare la varianza delle diverse variabili dello studio.

I criteri, numerati o abbreviati in alcuni dei grafici, sono quelli riportati in Tab 1 e hanno la seguente corrispondenza: criterio 0: Obiettivi e Risvolti, 1: Uso e Interazione con le Tecnologie e AI Generativa, 2: Progettazione: Coerenza con obiettivi, Coerenza interna e Originalità, 3: Inclusività, 4: Valutazione.

In primo, luogo si è cercato di capire quali criteri della valutazione fossero quelli con la minore e maggiore varianza (Tab. 4) per comprendere quali fossero quelli valutati in modo più concorde da tutti i valutatori.

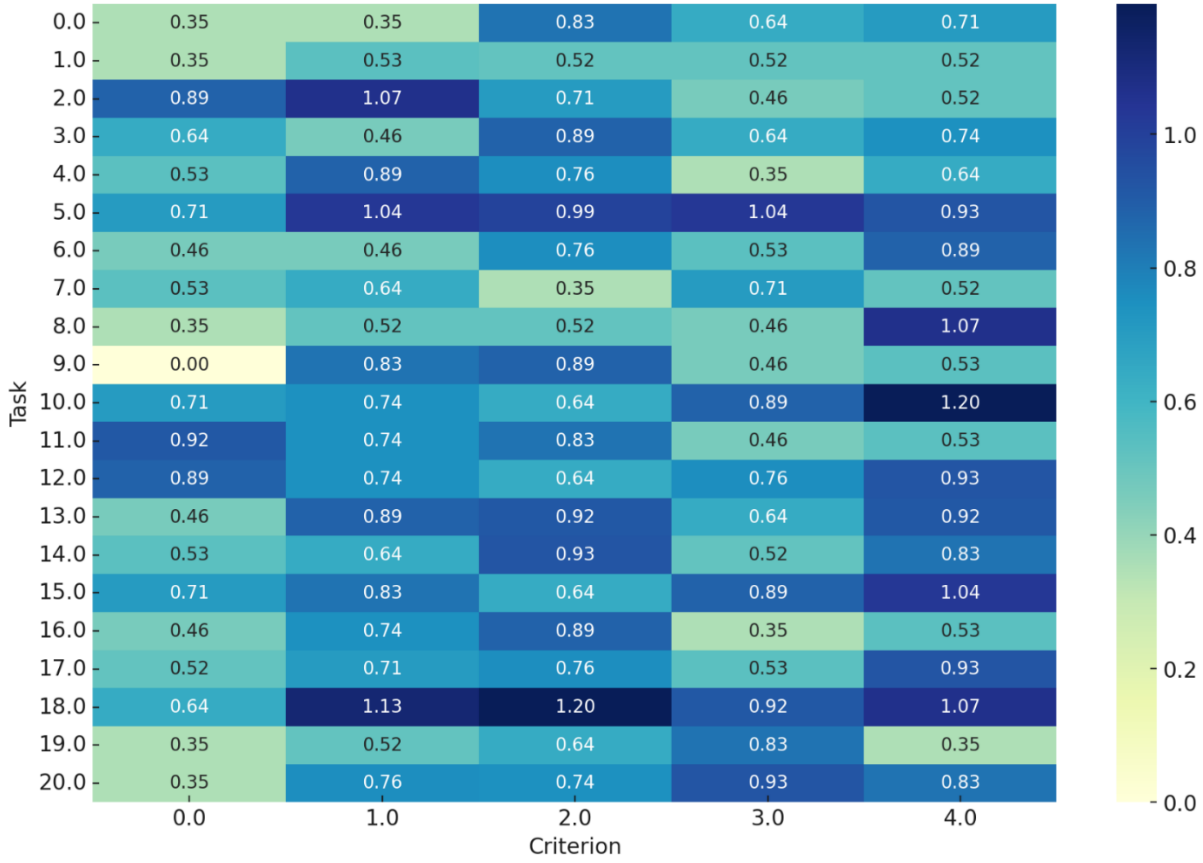
Il criterio con la minima varianza (deviazione standard) in tutti i prodotti è il Criterio 0 (Obiettivi), con una varianza media di circa 0,5. Ciò suggerisce che esista un livello elevato di accordo tra i valutatori nel valutare la bontà degli obiettivi e dei risvolti progettati. D'altra parte, il criterio con la varianza massima tra tutte le attività è il Criterio 4 (Valutazione), con una varianza media di circa 0,772. Ciò indica un maggiore livello di disaccordo o incoerenza nel modo in cui i valutatori hanno valutato la correttezza e l'opportunità delle valutazioni programmate.

Tabella 4. Varianza media delle valutazioni assegnate ai criteri

Rank	Criterio	Varianza media	Varianza media in % sul punteggio da 0 a 4
1	4 - Valutazione	0.772	19.31
2	2 - Progettazione	0.763	19.08
3	1 - Uso Tecnologie	0.726	18.14
4	3 - Inclusività	0.644	16.11
5	0 - Obiettivi	0.541	13.54

Nella Tabella 5 è possibile vedere come alcuni prodotti in particolare abbiano messo in disaccordo i valutatori, specialmente il prodotto numero 18. Altri prodotti come, ad esempio, il numero 0 e il numero 1 hanno ricevuto valutazioni più omogenee. In altri prodotti ancora, come nel prodotto n.8, è solo un singolo criterio che non viene valutato omogeneamente.

Tabella 5. Varianza dei punteggi per i diversi Prodotti (Task) e Criteri (Criterion).



Indice di Disaccordo

Inizialmente si è proceduto per ogni criterio a confrontare la differenza fra i punteggi assegnati dai valutatori umani (e0 ed e1) e a calcolarne la differenza con la media dei valutatori LLM. Dalle tabelle 6 e 7 si può apprezzare come gli LLM abbiano assegnato dei punteggi più simili al valutatore e1. In Tabella 8 sono mostrati anche i risultati del confronto fra i punteggi assegnati dai due valutatori umani. Fra essi sono presenti talvolta differenze anche più marcate rispetto a quelle fra umano ed LLM, raggiungendo in più di un'occasione i tre punti di differenza. Tuttavia, si può anche notare come, analizzando criterio per criterio, o anche prodotto per prodotto, dove c'è il maggiore disaccordo fra valutatori umani ed LLM si può notare una notevole concordanza dei valutatori umani.

Ad esempio, se si analizza il prodotto ("task" nella tabella) 18, si può notare come esso fosse problematico a livello di varianza (Tab. 5), e di disaccordo (Tab. 6 e 7) soprattutto per quanto riguarda i criteri 1 e 2 (Uso delle tecnologie e Progettazione). Se si guarda la Tabella 8 si può notare come siano invece concordanti le valutazione dei due valutatori umani.

Per ottenere una metrica più robusta e per meglio capire quali valutatori hanno assegnato punteggi più simili per i vari criteri, si è sviluppato un "Indice di Disaccordo" (IdD). Tale indice combina la Differenza media fra i punteggi assegnati ad un criterio e la variabilità di questa differenza. E' stato calcolato in modo da capire per ogni criterio quali siano i valutatori più simili a quelli umani (compresa l'opzione che il valutatore più simile sia l'altro valutatore umano). Esso è costruito come segue: Indice di Disaccordo = (Differenza media + Variabilità della differenza) / 2.

Tabella 6. Differenza di punteggi assegnati fra e0 (valutatore umano 0) e la media dei valutatori LLM.

Tabella 7. Differenza di punteggi assegnati fra e1 (valutatore umano 1) e la media dei valutatori LLM.

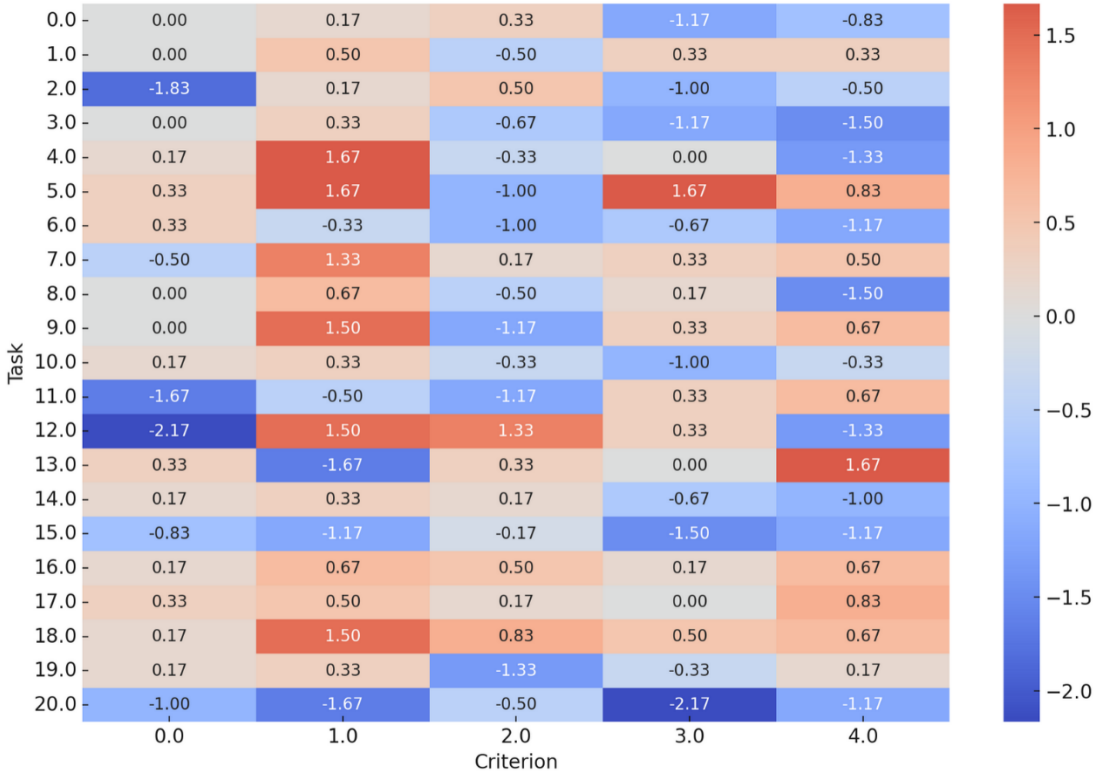
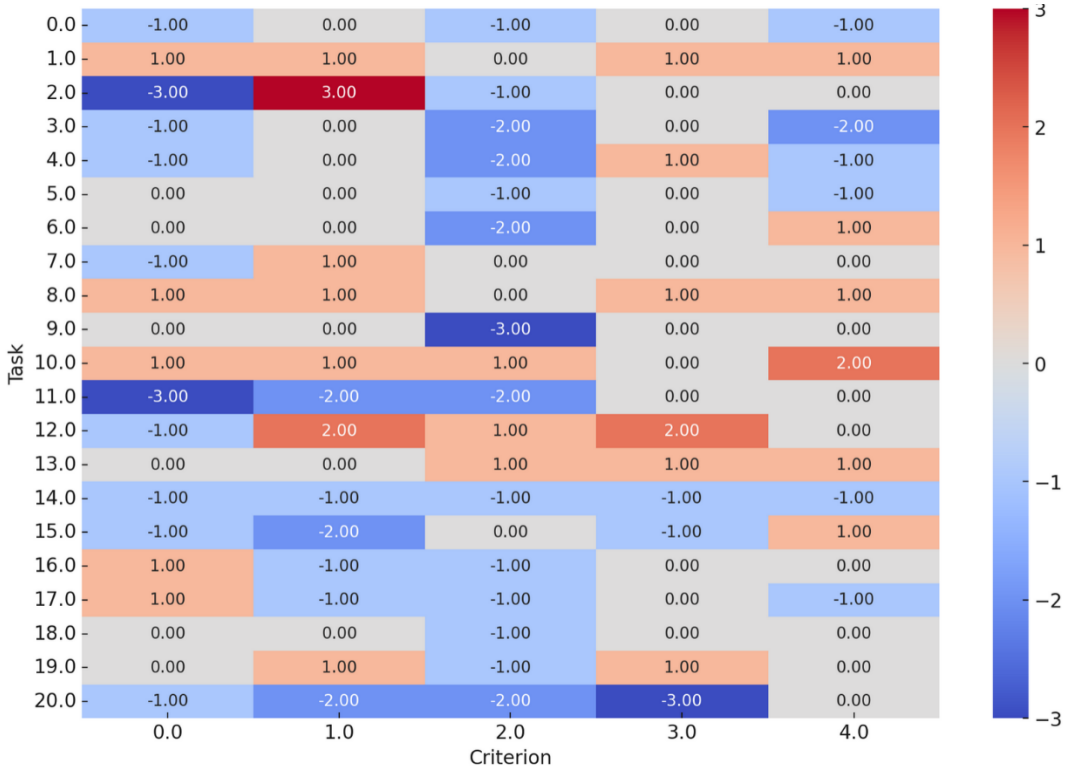


Tabella 8. Differenza di punteggi assegnati fra e0 (valutatore umano 0) ed e1 (valutatore umano 1).



Quindi:

• La “Differenza media” è la differenza media assoluta nei voti tra il valutatore in questione e il valutatore di riferimento (e0 o e1) in tutti i compiti e criteri.
• La “Variabilità della differenza” è la deviazione standard dei punteggi di differenza tra il valutatore e il valutatore di riferimento, che riflette quanto coerenti siano queste differenze tra compiti e criteri diversi.

L'Indice di Disaccordo viene calcolato individualmente per ciascun valutatore. Fornisce un'unica misura che incapsula sia l'entità media delle differenze di valutazione rispetto al valutatore di riferimento sia la consistenza di tali differenze. Un valore più alto indica un maggiore disaccordo complessiva nella valutazione rispetto al valutatore di riferimento.

Il valore più alto possibile per l'indice per un valutatore si raggiungerebbe se valutasse sempre alla massima differenza da e0 (4 punti) senza alcuna variazione. In questo caso, sia la dissomiglianza media che la deviazione standard sarebbero 4. Pertanto, anche l’indice composito, essendo la media di questi due valori, sarebbe 4.

Valutatori simili al valutatore umano e0

Applicando questa metrica (Tab. 9) abbiamo verificato come il valutatore che dato punteggi più simili al valutatore e0 è stato l’altro valutatore umano, e1. L’LLM (l’intelligenza artificiale) più in accordo con e0 è invece stato e2, ovvero ChatGPT-4, seguito a grande distanza da Claude 2 (e4) e da OpenChat 3.5.

Il valutatore che più si discosta dai punteggi assegnato da e0, è e6, ovvero Bing Chat.

Tabella 9. Indici di Disaccordo fra il valutatore umano e0 e gli altri valutatori.

Valutatori	Differenza Media	Variabilità della Differenza	Differenza Media Standardizzata (Z-score)	Indice di Disaccordo (IdD)
Umano e1 - Umano e0	0.8381	0.8101	-0.8343	-0.0121
ChatGPT-4 (e2) - Umano e0	0.8857	0.8004	-0.4230	0.1887
ChatGPT-3.5 (e3) - Umano e0	0.9714	0.7527	0.3173	0.5350
Claude 2 (e4) - Umano e0	0.9333	0.7754	-0.0118	0.3818
Bing Chat (e6) - Umano e0	1.1619	0.8562	1.9623	1.4093
OpenChat 3.5 (e7) - Umano e0	0.9429	0.8184	0.0705	0.4445

Concentrandosi ad osservare il singolo criterio (Tab. 10), si può notare come l'indice di accordo con gli altri valutatori vari di criterio in criterio, confermando tuttavia nel complesso quanto appreso dal calcolo generale. È interessante notare come nei criteri che abbiamo riconosciuto essere più critici per la valutazione, come Valutazione, Progettazione e Uso Tecnologie (Tab. 4) i valutatori più simili a e0 siano l'altro valutatore umano e OpenChat 3.5.

Tabella 10. Indici di disaccordo dei valutatori rispetto al valutatore umano e0 suddivisi per criterio:

Criteri	IdD e1-e0	IdD e2-e0	IdD e3-e0	IdD e4-e0	IdD e6-e0	IdD e7-e0	Valutatore più simile	Secondo Val. + Simile
0 - Obiettivi	0.8679	0.7311	0.5583	0.5749	0.6606	0.8107	ChatGPT-3.5	Claude 2
1 - Uso Tecnologie	0.8969	0.9841	0.9922	1.0105	1.1209	0.8412	OpenChat 3.5	Human e1
2 - Progettazione	0.9678	0.9183	0.9700	1.0345	1.1464	0.8107	OpenChat 3.5	ChatGPT-4
3 - Inclusività	0.6910	0.5749	0.8026	0.6374	0.9700	0.7796	ChatGPT-4	Claude 2
4 - Valutazione	0.6625	0.9629	0.9262	0.9183	1.0909	1.1342	Human e1	Claude 2

Valutatori simili al valutatore umano e1

Si era già potuto verificare da una veloce analisi della Tabella 7 che il valutatore umano e1 ha assegnato punteggi in generale più simili a quelli degli LLM rispetto al valutatore e0. Questo viene confermato dai dati in Tabella 11 che dimostrano come il valutatore in generale più in accordo con e1 sia stato un LLM, in particolare Claude 2, seguito a brevissima distanza da ChatGPT-4 e infine molto staccato ChatGPT-3.5.

È interessante notare come l'altro valutatore umano, e0, sia il secondo valutatore più in disaccordo con e1 mentre quello più in disaccordo è Bing Chat, come nel caso precedente.

Tabella 11. Indici di disaccordo fra il valutatore umano e1 e gli altri valutatori.

Valutatori	Differenza Media	Variabilità della Differenza	Differenza Media Standardizzata (Z-score)	Indice di Disaccordo (IdD)
Umano e0 - Umano e1	0.8381	0.8101	0.7328	0.7715
ChatGPT-4 (e2) - Umano e1	0.7333	0.6831	-0.5211	0.0810
ChatGPT-3.5 (e3) - Umano e1	0.7810	0.7336	0.0489	0.3912
Claude 2 (e4) - Umano e1	0.7048	0.7196	-0.8631	-0.0718
Bing Chat (e6) - Umano e1	0.8952	0.8077	1.4168	1.1123
OpenChat 3.5 (e7) - Umano e1	0.8286	0.7398	0.6188	0.6793

Anche nel caso del valutatore e1, andando ad osservare singolarmente i criteri (Tab. 12), si può notare come l'indice di accordo per la valutazione dei criteri più critici (Valutazione, Progettazione e Uso Tecnologie) i valutatori più simili a e1 siano l'altro valutatore umano e0 e ChatGPT-4. Anche in questo caso OpenChat 3.5 risulta il più concorde sul criterio dell'uso delle tecnologie e in generale abbastanza concorde su tutta la linea.

Tabella 12. Indici di disaccordo dei valutatori rispetto al valutatore umano e1 suddivisi per criterio:

Criteri	IdD e0-e1	IdD e2-e1	IdD e3-e1	IdD e4-e1	IdD e6-e1	IdD e7-e1	Valutatore Più Simile	Secondo Val. + Simile
0 - Obiettivi	0.8679	0.6440	0.6589	0.6367	0.5779	0.7797	Bing Chat	Claude 2
1 - Uso Tecnologie	0.8969	0.8873	0.8107	0.8366	1.0532	0.7652	OpenChat 3.5	ChatGPT-3.5
2 - Progettazione	0.9678	0.5583	0.6790	0.6238	0.8366	0.7559	ChatGPT-4	Claude 2
3 - Inclusività	0.6910	0.6985	0.7490	0.7157	0.8969	0.7796	Human e0	ChatGPT-4
4 - Valutazione	0.6625	0.7311	0.8786	0.7446	0.8186	0.8536	Human e0	ChatGPT-4

Discussione

Rispetto all'obiettivo di capire se gli LLM attuali possano essere utilizzati da docenti non esperti di Machine Learning per valutare i prodotti scritti degli studenti anche in presenza di compiti e domande aperte grazie a rubriche di valutazione; dalle analisi effettuate emergono alcuni elementi interessanti:

- L'ampiezza della finestra di contesto è estremamente importante per questo tipo di compiti.
- Due dei modelli più potenti provati hanno performato molto bene (ChatGPT-4 e Claude 2), ma un modello altrettanto potente come BingChat ha dato giudizi molto diversi da quelli dei valutatori umani.
- OpenChat 3.5, un modello OpenSource e molto piccolo da 7 miliardi di parametri, ha dato valutazioni più simili a quelle degli esseri umani rispetto a modelli sulla carta molto più potenti.
- Dalla PCA appare come i valutatori umani, in generale, abbiano una modalità di valutazione diversa da quella degli LLM, ma anche come ChatGPT-4 sia probabilmente l'LLM che, in generale, è più vicino alla modalità umana.

- L'Indice di Disaccordo indica la stessa direzione, ma dettagliando il punto di vista dei singoli valutatori umani emergono altri dettagli: mentre per il valutatore umano e0 il valutatore più concorde è il valutatore umano e1, seguito da ChatGPT-4, non è vero il contrario. Il valutatore più concorde con il valutatore umano e1 è stato Claude 2, seguito da ChatGPT-4.
- La concordanza dei valutatori umani si vede in casi particolari (come il compito 18) e mediamente nella valutazione dei criteri più complessi come la bontà della valutazione prevista e il corretto utilizzo e implementazione delle tecnologie.

In sintesi, secondo i dati a disposizione, sembra che ChatGPT-4 di OpenAI e Claude 2 di Anthropic siano i modelli che si presentano come i migliori candidati per un utilizzo come supporto alla valutazione di compiti scritti per i docenti, in presenza di una rubrica di valutazione perché sono quelli che sono stati più concordi alle valutazioni dei due esperti valutatori umani.

OpenChat 3.5 merita una menzione perché pur essendo un modello da 7 miliardi parametri, quindi che può funzionare in locale sui computer desktop della maggior parte delle persone, è stato più aderente alle valutazioni degli esseri umani rispetto a modelli molto più grandi come Bing Chat, ChatGPT-3.5 e Google Bard (che è stato escluso dall'analisi).

Un criterio da tenere a mente è anche l'impatto delle differenze di valutazione fra i vari valutatori. In una rubrica di valutazione come quella proposta, con valutazioni che si traducono in punteggi da 0 a 4, scattando di un punto alla volta, la varianza ha un impatto diverso a seconda del suo valore: per il Criterio 0 (Obiettivi), ad esempio, con una varianza di circa 0.54, l'effetto non è particolarmente importante. Questo perché la varianza è inferiore a 1, il che implica che la maggior parte delle valutazioni è raggruppata intorno al punteggio medio e raramente sposta il livello di un intero punto. Per il Criterio 4 (Valutazione), con una varianza di circa 0.77, l'effetto è più significativo. Questo livello di varianza indica che le valutazioni sono più distribuite e possono facilmente spostare il livello di apprendimento rilevato di un punto intero o più.

Se si tiene questo in considerazione, guardando le Tabelle 10 e 12 si può capire come, al momento, nessuno degli LLM possa essere utilizzato per una valutazione autonoma per tutti i criteri, soprattutto per quanto riguarda quelli più complessi (Webb, 2023). Tuttavia, sicuramente ChatGPT-4, Claude 2 e, con alcune cautele, anche OpenChat 3.5, potrebbero essere utilizzati come supporto alla valutazione sia per quanto riguarda il livello di valutazione sommativa che per quello di valutazione formativa (in quel caso interagendo soprattutto con gli studenti) come descritte nel modello AI-MAAS (Agostini, Picasso, 2023).

Conclusioni

La domanda alla base di questo studio era se e quali LLM attuali possano essere utilizzati da docenti, anche privi di esperienza tecnica, per valutare i prodotti scritti degli studenti in presenza di compiti e domande aperte grazie a rubriche di valutazione. L'impiego di queste tecnologie, infatti, potrebbe rendere la valutazione più sostenibile, scalabile e permettergli di mettere in atto una programmazione più coerente con gli obiettivi di apprendimento dichiarati.

La risposta derivante da questo studio sembra essere affermativa, a patto che tali LLM non vengano utilizzati senza una supervisione. Questi dovrebbero essere considerati come un supporto al docente e non come sostituti per il compito di valutazione perché dai dati a disposizione non risultano essere ancora abbastanza affidabili per eseguire questo compito senza supervisione. In questo, dunque, vengono confermate le ultime linee guida (Miao et al., 2023; Webb, 2023).

Tuttavia, non tutti i modelli esaminati sono consigliabili per coadiuvare il docente nella valutazione. Alcuni non hanno le "capacità" (o l'allenamento) per eseguire un compito così complesso (è il caso di Bing Chat); altri, come Google Bard, non hanno una finestra di contesto abbastanza ampia per correggere dei testi elaborati. D'altra parte, da un punto di vista pratico, è positiva la possibilità data da Claude 2 di correggerne anche più alla volta.

Lo studio ci ha permesso di fare una cernita rispetto agli LLM esaminati e selezionarne due coi quali fare delle analisi più estese e approfondite: ChatGPT-4 e Claude 2. Per selezionare un terzo modello e, possibilmente, un quarto oltre ai due menzionati, potrebbe essere interessante fare un esperimento simile solamente con modelli Opensource data la buona prestazione di OpenChat 3.5 rispetto a modelli molto più grandi. La comunità dei LLM Opensource è molto attiva e durante il 2023 è stata capace di alzare esponenzialmente il livello qualitativo dei modelli, anche grazie ad aziende ed istituzioni come Meta (che ha rilasciato LLAMA e LLAMA2), TII (il Technology Innovation Institute di Abu Dhabi che ha rilasciato i modelli Falcon), Mistral (col modello Mistral) e HuggingFace (una piattaforma per lo sviluppo e la condivisione di LLM) che hanno creato un ecosistema a supporto di questi sforzi che si sta rivelando molto efficace.

Una volta selezionati pochi modelli si potrà procedere ad un prossimo studio, utilizzando un tipo di prompting multi-shot, quindi portando agli LLM degli esempi di valutazioni eseguite in modo ottimale prima di passare al compito di valutazione. I feedback testuali ai compiti sono una potenzialità enorme degli LLM, e quelli che potrebbero essere forniti durante la valutazione seguendo i criteri forniti nella rubrica meriterebbero un approfondimento particolare (Agostini, Picasso, 2023; Sabzalieva, Valentini, 2023; Sullivan et al., 2023; Tamkin et al., 2021).

Riguardo le limitazioni del presente studio, risiedono nel campione di prodotti degli studenti che deve essere aumentato in modo consistente, così come il numero di esperti valutatori umani e le discipline coinvolte nelle prove. Anche la rubrica di valutazione può essere ottimizzata rendendola più chiara, tuttavia non si ritiene di dover semplificare troppo i criteri, in modo da non cadere nell'errore di valutare gli LLM su rubriche apposite, che non sarebbero quelle che un docente creerebbe nella quotidianità del suo lavoro. Questo studio non ha inoltre considerato tutte le implicazioni e le problematiche etiche, di protezione dei dati e legislative. Anche in questo caso il panorama è in veloce evoluzione.

Si conclude sottolineando il rapidissimo ritmo di aggiornamento degli LLM all'interno delle loro piattaforme: velocità del modello, ampiezza della finestra di contesto e capacità possono quindi variare in poco tempo e alcune conclusioni di questo studio potrebbero presto necessitare, anch'esse, un aggiornamento.

Bibliografia

Agostini, D., Picasso, F. (2023, November 6). Large Language Models for Sustainable Assessment and Feedback in Higher Education: Towards a Pedagogical and Technological Framework. *Proceedings of the First International Workshop on High-Performance Artificial Intelligence Systems in Education Co-Located with 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023)*.

AIxEDU 2023 High-performance Artificial Intelligence Systems in Education, Aachen. https://ceur-ws.org/Vol-3605/

Ali, F., Choy, D., Divaharan, S., Tay, H. Y., Chen, W. (2023). Supporting Self-Directed Learning and Self-Assessment Using TeacherGAIA, a Generative AI Chatbot Application: Learning Approaches and Prompt

Engineering. *Learning: Research and Practice*, *9*(2), 135–147. https://doi.org/10.1080/23735082.2023.2258886

Babina, T., Fedyk, A., He, A. X., Hodson, J. (2023). *Firm Investments in Artificial Intelligence Technologies and Changes in Workforce Composition* (Working Paper 31325). National Bureau of Economic Research. https://doi.org/10.3386/w31325

Baytak, A. (2023). The Acceptance and Diffusion of Generative Artificial Intelligence in Education: A Literature Review. *Current Perspectives in Educational Research*, *6*(1), Article 1. https://doi.org/10.46303/cuper.2023.2

Biagini, G., Cuomo, S., Ranieri, M. (2023, November 6). Developing and Validating a Multidimensional AI Literacy Questionnaire: Operationalizing AI Literacy for Higher Education. *Proceedings of the First International Workshop on High-Performance Artificial Intelligence Systems in Education Co-Located with 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023)*.

AIxEDU 2023 High-performance Artificial Intelligence Systems in Education, Aachen. https://ceur-ws.org/Vol-3605/

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347–364. https://doi.org/10.1007/BF00138871

Bloomberg, L.P. (2023). *Generative AI to Become a $1.3 Trillion Market by 2032*, Research Finds. https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/

Cetindamar, D., Kitto, K., Wu, M., Zhang, Y., Abedin, B., Knight, S. (2024). Explicating AI Literacy of Employees at Digital Workplaces. *IEEE Transactions on Engineering Management*, *71*, 810–823. https://doi.org/10.1109/TEM.2021.3138503

Elbanna, S., Armstrong, L. (2023). Exploring the integration of ChatGPT in education: Adapting for the future. *Management & Sustainability: An Arab Review*, *3*(1), 16–29. https://doi.org/10.1108/MSAR-03-2023-0016

Extance, A. (2023). ChatGPT has entered the classroom: How LLMs could transform education. *Nature*, *623*(7987), 474–477. https://doi.org/10.1038/d41586-023-03507-3

Gerdes, A. (2022). A participatory data-centric approach to AI Ethics by Design. *Applied Artificial Intelligence*, *36*(1). https://doi.org/10.1080/08839514.2021.2009222

GTnum. (2023, April). Intelligence artificielle et éducation: Apports de la recherche et enjeux pour les politiques publiques. *Carnet Hypothèses 'Éducation, numérique et recherche'*. https://edunumrech.hypotheses.org/8726

Hammond, G. (2023, December 27). *Big Tech outspends venture capital firms in AI investment frenzy*. https://www.ft.com/content/c6b47d24-b435-4f41-b197-2d826cce9532

Jang, C. (2023). Coping with vulnerability: The effect of trust in ai and privacy-protective behaviour on the use of ai-based services. *Behaviour & Information Technology*. https://doi.org/10.1080/0144929X.2023.2246590

Kong, S.-C., Cheung, W. M.-Y., Zhang, G. (2023). Evaluating an Artificial Intelligence Literacy Programme for Developing University Students' Conceptual Understanding, Literacy, Empowerment and Ethical Awareness. *Educational Technology & Society*, *26*(1), 16–30.

Koraishi, O. (2023). Teaching English in the Age of AI: Embracing ChatGPT to Optimize EFL Materials and Assessment. *Language Education and Technology*, *3*(1), Article 1. https://langedutech.com/letjournal/index.php/let/article/view/48

Lee, Y. S., Kim, T., Choi, S., Kim, W. (2022). When does AI pay off? AI-adoption intensity, complementary investments, and R&D strategy. *Technovation*, *118*, 102590. https://doi.org/10.1016/j.technovation.2022.102590

Lepage, A., Roy, N. (2023). A review of the literature from 1970 to 2022 on the roles of teachers and artificial intelligence in the field of AI in education. *Médiations et Médiatisations*, *16*, 30–50. https://doi.org/10.52358/mm.vi16.304

Majeed, A., Hwang, S. O. (2023). When AI Meets Information Privacy: The Adversarial Role of AI in Data Sharing Scenario | *IEEE Journals & Magazine* | IEEE Xplore. *IEEE Access*, *11*, 76177–76195. https://doi.org/10.1109/ACCESS.2023.3297646

Martin, P. P., Kranz, D., Wulff, P., Graulich, N. (2023). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*, *n/a*(n/a). https://doi.org/10.1002/tea.21903

Miao, F., Holmes, W., Ronghuai, H., Hui, Z. (2023). *AI and education: Guidance for policy-makers*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000376709?posInSet=1&queryId=dc9add31-5176-42f3-9537-4819566551e9

Miguel A. Cardona, E. D., Rodríguez, R. J., Ishmael, K. (2023). *Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations*. https://policycommons.net/artifacts/3854312/ai-report/4660267/

Ouyang, F., Dinh, T. A., Xu, W. (2023). A Systematic Review of AI-Driven Educational Assessment in STEM Education. *Journal for STEM Education Research*, *6*(3), 408–426. https://doi.org/10.1007/s41979-023-00112-x

Perkins, M. (2023). Academic Integrity Considerations of AI Large Language Models in the Post-Pandemic Era: ChatGPT and Beyond. *Journal of University Teaching and Learning Practice*, *20*(2). https://eric.ed.gov/?id=EJ1382355

Roy, S., Gupta, V., Ray, S. (2023). Adoption of AI ChatBot like Chat GPT in Higher Education in India: A SEM Analysis Approach. *Economic Environment*, *4*(46), 130–149. https://doi.org/10.36683/2306-1758/2023-4-46/130-149

Russell Group. (2023). *New principles on use of AI in education*. Russell Group. https://russellgroup.ac.uk/news/new-principles-on-use-of-ai-in-education/

Sabzalieva, E., Valentini, A. (2023). *ChatGPT and artificial intelligence in higher education: Quick start guide*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000385146

Saif, N., Khan, S. U., Shaheen, I., Alotaibi, A., Alnfiai, M. M., Arif, M. (2023). Chat-GPT; validating Technology Acceptance Model (TAM) in education sector via ubiquitous learning mechanism. *Computers in Human Behavior*, 108097. https://doi.org/10.1016/j.chb.2023.108097

Samuelson, P. (2023). Generative AI meets copyright. *Science (New York, N.Y.)*, *381*(6654), 158–161. https://doi.org/10.1126/science.adi0656

Sullivan, M., Kelly, A., Mclaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching*. https://doi.org/10.37074/jalt.2023.6.1.17

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075

Tamkin, A., Brundage, M., Clark, J., Ganguli, D. (2021). *Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models* (arXiv:2102.02503). arXiv. http://arxiv.org/abs/2102.02503

Tiwari, C. K., Bhat, Mohd. A., Khan, S. T., Subramaniam, R., Khan, M. A. I. (2023). What drives students toward ChatGPT? An investigation of the factors influencing adoption and usage of ChatGPT. *Interactive Technology and Smart Education*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/ITSE-04-2023-0061

UCL. (2023, September 12). Using generative AI (GenAI) in learning and teaching. *Teaching & Learning*. https://www.ucl.ac.uk/teaching-learning/publications/2023/sep/using-generative-ai-genai-learning-and-teaching

van Oijen, V. (2023, March 31). AI-generated text detectors: Do they work? | SURF Communities. *Surf Communities*. https://communities.surf.nl/en/ai-in-education/article/ai-generated-text-detectors-do-they-work

Wang, B., Rau, P.-L. P., Yuan, T. (2023). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, *42*(9), 1324–1337. https://doi.org/10.1080/0144929X.2022.2072768

Wang, J., Liu, S., Xie, X., Li, Y. (2023). *Evaluating AIGC Detectors on Code Content* (arXiv:2304.05193). arXiv. https://doi.org/10.48550/arXiv.2304.05193

Webb, M. (2023). A Generative AI Primer. *National Centre for AI*. https://nationalcentreforai.jiscinvolve.org/wp/2024/01/02/generative-ai-primer/

Weber, P., Pinski, M., Baum, L. (2023). Toward an Objective Measurement of AI Literacy. *PACIS 2023 Proceedings*. https://aisel.aisnet.org/pacis2023/60

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1), 26. https://doi.org/10.1007/s40979-023-00146-z

Yeo, M. A. (2023). Academic integrity in the age of Artificial Intelligence (AI) authoring apps. *TESOL Journal*, *14*(3), e716. https://doi.org/10.1002/tesj.716