Intelligenza Artificiale per la valutazione linguistica nell'ambito degli esami di certificazione Cambridge: Caso studio del B1 *Preliminary*

Luca Di Stefano
Università degli Studi di Bari, Aldo Moro
L.distefano5@studenti.uniba.it

Abstract
Cambridge Assessment English sets global standards in language tests, strictly tying exams to the CEFR framework, yet the rapid rise of AI poses new paths for test grading. While existing literature covers proprietary automated assessment systems, there is a gap regarding the interactional evaluation limits of off-the-shelf generative AIs. This study aims to fill this gap by exploring how commercial AIs, specifically NotebookLM and Gemini Pro, assess productive skills, comparing their outputs to human examiners' comments in a blind study of ten B1-level Italian learners. Data suggest AI limitations in the tested models when grading candidates: in the Writing test, they demand higher-level structures, failing to stick to strict CEFR frameworks. Challenges emerge in the Speaking test, where the selected AIs relied on text transcriptions, struggling to process prosodic features like pitch contour or speech rate, missing key facts of human interaction, and prioritizing pure grammar over it. All factors considered, AI is a highly efficient tool to track class progress regularly, but high-stakes formal exams must adhere to a human-in-the-loop methodology, for only real experts can currently grasp the deepest pragmatic traits of speech and fix AI's "blindness", ensuring a fair and valid test.

Keywords: Artificial Intelligence; Decoding and Language Assessment; Cambridge Assessment English; Productive Skills; Human-in-the-loop.

1. Introduzione

Il panorama educativo attuale vive una transizione dettata dall'integrazione dell'Intelligenza Artificiale nei processi di misurazione delle competenze linguistiche (Yu, Xu 2024: 356). Come analizzano studi analoghi, L'IA sta ridefinendo alcune pratiche educative, con un impatto nell'educazione linguistica che richiede strumenti operativi per un'integrazione consapevole (Cinganotto, Montanucci 2025).

A tal proposito, la letteratura recente evidenzia la necessità di un "ricorso consapevole e guidato che promuova le glottotecnologie, valorizzando i processi cognitivi e metalinguistici degli apprendenti" (Nitti 2026), un principio fondamentale per la presente indagine sul paradigma *human-in-the-loop*. Per questo motivo, il *language testing*, tradizionalmente basato sul Quadro Comune Europeo di Riferimento per la conoscenza delle lingue (QCER) (North *et al*. 2021: 25) e, ancor prima, sui principi fondanti di *validity* e *reliability* teorizzati dai principali studiosi (Bachman 1990; Alderson *et al*. 1995; Bachman, Palmer 1996; McNamara 2000; Barni

2005; Weir 2005; Fulcher 2010), sta ridefinendo i propri confini in relazione ai nuovi paradigmi dell'automazione. Enti certificatori come Cambridge Assessment English offrono standard valutativi consolidati nel bilanciare validità di costrutto e affidabilità statistica (Khalifa, Weir 2009: 3). La diffusione di modelli generativi commerciali e proprietari, unita all'impiego dell'*Automated Item Generation* (AIG), solleva dubbi sulla capacità dell'IA di replicare l'affidabilità del *testing* linguistico istituzionale (Kim *et al*. 2025: 920).

Attualmente l'attenzione si è spostata verso modelli di *Natural Language Generation*, i quali dimostrano ottime capacità nel *parsing* strutturale e nel riconoscimento di *pattern* sintattici e fonetici (McKnight *et al.* 2023: 99). I recenti modelli *audio-native* tentano di elaborare direttamente i fonemi per superare i limiti della sola trascrizione testuale. Tuttavia, proprio entro questi confini, limitatamente ai modelli testati, si palesano alcune inefficienze valutative, come il fenomeno della dominanza lessicale (Chen *et al.* 2026: 5848-5856). Si pensi a un *input* audio in cui il contenuto lessicale è semanticamente neutro ma i tratti acustici veicolano un'intenzione ironica o sarcastica. In questi casi, alcuni sistemi privilegiano il dato testuale, faticando a interpretare alcune sfumature prosodiche.

Nonostante la recente integrazione dei *thinking tokens* abbia migliorato la comprensione semantica rispetto alla *next-token prediction*, permangono criticità algoritmiche in aree valutative decisive. Difatti, nei modelli testati, la mancanza di una reale intelligenza sociale impedisce di valutare l'appropriatezza pragmatica in contesti sociali complessi, rendendo l'*output* dipendente da un *prompting* esplicito e guidato, privo della naturalezza intuitiva tipica dell'umano (Tsvilodub *et al*. 2026: 1). Questo deficit pragmatico emerge particolarmente negli impliciti comunicativi e nelle presupposizioni. Pur mappando dati acustici come il *pitch contour*, le IA adoperate nello studio faticano a interpretare le violazioni delle massime conversazionali, le formulazioni idiosincratiche e gli attivatori di presupposizione, i quali richiedono l'accesso alle intenzioni dell'interlocutore e a un *common ground* condiviso, (Tsvilodub *et al*. 2026: 1; Chen *et al*. 2026: 5848).

Queste limitazioni assumono maggiore rilevanza nelle abilità produttive, che costituiscono la dimostrazione pratica di competenza comunicativa. Difatti, per il *Natural Language Understanding*, il riconoscimento del non detto e della comunicazione interattiva rimane una sfida primaria (Tsvilodub *et al*. 2026: 1).

In questo contesto si inserisce la presente indagine. Sebbene la ricerca accademica abbia indagato i sistemi di *Automated Writing Evaluation* (Weigle 2013) e di *Automated Speaking Assessment* (Zechner *et al.* 2009), basati su metriche di accuratezza morfosintattica e fluidità acustica, si assiste a un limitato focus sulla dimensione relazionale e sull'affidabilità di modelli generativi commerciali nei contesti certificatori. In questo studio, dunque, ci si è soffermati sul deficit interpretativo delle IA adoperate dinanzi a questi elementi.

Come evidenziano studi passati, l'automazione dello *scoring* per il parlato spontaneo ha avuto la necessità di appoggiarsi a sistemi di riconoscimento vocale che faticavano a garantire un'analisi dettagliata delle componenti linguistiche più complesse, rendendo così indispensabile il ricorso alla valutazione umana per la validazione di contesti ad alto impatto (Zechner *et al.* 2009: 883-884).

Questa indagine empirica mira a colmare questa mancanza valutando l'efficacia di IA commerciali nelle prove di *Writing* e *Speaking* del B1 *Preliminary*. Confrontando l'*assessment* algoritmico con gli standard degli esaminatori umani, lo studio mira a dimostrare come i deficit pragmatici delle IA testate rendano necessario un paradigma *human-in-the-loop* per garantire una valutazione linguistica olistica e affidabile. Difatti, la competenza comunicativa non si limita alle conoscenze grammaticali, ma richiede la valutazione di modelli multicomponenziali che includano l'appropriatezza contestuale e pragmatica (Bachman 1990: 102; McNamara 2000).

2. Metodologie di Ricerca

Lo studio adotta metodologie comparative per misurare l'efficacia e l'affidabilità dei LLM nella generazione e valutazione di *Sample Test* Cambridge B1 *Preliminary*. La sperimentazione ha coinvolto dieci studenti di scuola secondaria, frequentanti un corso extracurriculare B1. In aggiunta, il campione è stato bilanciato per genere, includendo cinque soggetti di sesso maschile e cinque di sesso femminile, per facilitare i sistemi di riconoscimento vocale dell'IA e mitigare eventuali errori nell'attribuzione vocale per la prova di *Speaking*.

La generazione del *Sample Test* è stata affidata a NotebookLM[1], sfruttando la sua chiusura documentale per vincolare l'*output* a documenti istituzionali controllati. Al modello, dunque, è stata fornita documentazione di natura strettamente istituzionale, come il B1 *Preliminary Handbook for Teachers* (CUP&A 2024), i descrittori del Quadro Comune Europeo di Riferimento (CoE 2001), le *Research Notes* ufficiali della Cambridge University e, per il mantenimento dei confini lessicali, la *Vocabulary List* del livello in esame (CUP&A 2025).

Tuttavia, al fine di sopperire ai limiti di NotebookLM, è stato integrato un approccio ibrido, demandando al modello Gemini Pro l'attività di generazione del materiale visivo, come il diagramma a ragno tipico della prova di *Speaking* del livello B1 *Preliminary* (CUP&A 2024: 60, 64).

Il protocollo di somministrazione ha replicato il test *computer-based*, e, per prevenire la *test fatigue*, le prove ufficiali e quelle generate da IA sono state alternate su più giornate e somministrate tramite piattaforma Google Forms, con interfacce grafiche rese volutamente indistinguibili per i candidati, strutturate in moduli identici per layout, font e impostazioni di visualizzazione sia per le prove ufficiali che per quelle generate da IA. Questo accorgimento ha permesso di annullare i *bias* di aspettativa dei candidati sulla natura del test.

Per garantire trasparenza in fase di correzione, gli elaborati scritti e le tracce audio sono stati resi anonimi prima della valutazione asincrona: gli esaminatori umani hanno valutato le performance ignorando sia l'identità dei candidati, sia la natura della traccia.

Di seguito, vengono riportati i criteri di valutazione della prova di *Writing* (Immagine 1) e della prova di *Speaking* (Immagine 2), contenuti nell'*Handbook for Teachers*, ai quali è stato chiesto di fare specifico riferimento in fase di valutazione delle prove.

The following assessment scale, extracted from the one on the previous page, is used for marking candidate responses.

B1	CONTENT	COMMUNICATIVE ACHIEVEMENT	ORGANISATION	LANGUAGE
5	All content is relevant to the task. Target reader is fully informed	Uses the conventions of the communicative task to hold the target reader's attention and communicate straightforward ideas.	Text is generally well organised and coherent, using a variety of linking words and cohesive devices.	Uses a range of everyday vocabulary appropriately, with occasional inappropriate use of less common lexis. Uses a range of simple and some complex grammatical forms with a good degree of control. Errors do not impede communication.
4	*Performance shares features of Bands 3 and 5.*			
3	Minor irrelevances and/or omissions may be present. Target reader is on the whole informed.	Uses the conventions of the communicative task in generally appropriate ways to communicate straightforward ideas.	Text is connected and coherent, using basic linking words and a limited number of cohesive devices.	Uses everyday vocabulary generally appropriately, while occasionally overusing certain lexis. Uses simple grammatical forms with a good degree of control. While errors are noticeable, meaning can still be determined.
2	*Performance shares features of Bands 1 and 3.*			
1	Irrelevances and misinterpretation of task may be present. Target reader is minimally informed.	Produces text that communicates simple ideas in simple ways.	Text is connected using basic, high-frequency linking words.	Uses basic vocabulary reasonably appropriately. Uses simple grammatical forms with some degree of control. Errors may impede meaning at times.
0	Content is totally irrelevant. Target reader is not informed.	*Performance below Band 1.*		

B1 Preliminary Speaking Examiners use a more detailed version of the following assessment scales, extracted from the overall Speaking scales on page 64.

B1	GRAMMAR AND VOCABULARY	DISCOURSE MANAGEMENT	PRONUNCIATION	INTERACTIVE COMMUNICATION
5	Shows a good degree of control of simple grammatical forms, and attempts some complex grammatical forms. Uses a range of appropriate vocabulary to give and exchange views on familiar topics.	Produces extended stretches of language despite some hesitation. Contributions are relevant despite some repetition. Uses a range of cohesive devices.	Is intelligible. Intonation is generally appropriate. Sentence and word stress is generally accurately placed. Individual sounds are generally articulated clearly.	Initiates and responds appropriately. Maintains and develops the interaction and negotiates towards an outcome with very little support.
4	*Performance shares features of Bands 3 and 5.*			
3	Shows a good degree of control of simple grammatical forms. Uses a range of appropriate vocabulary when talking about familiar topics.	Produces responses which are extended beyond short phrases, despite hesitation. Contributions are mostly relevant, but there may be some repetition. Uses basic cohesive devices.	Is mostly intelligible, and has some control of phonological features at both utterance and word levels.	Initiates and responds appropriately. Keeps the interaction going with very little prompting and support.
2	*Performance shares features of Bands 1 and 3.*			
1	Shows sufficient control of simple grammatical forms. Uses a limited range of appropriate vocabulary to talk about familiar topics.	Produces responses which are characterised by short phrases and frequent hesitation. Repeats information or digresses from the topic.	Is mostly intelligible, despite limited control of phonological features.	Maintains simple exchanges, despite some difficulty. Requires prompting and support.
0	*Performance below Band 1.*			

[1] Chiamato "Gemini Notebook" a partire dal mese di luglio 2026.

Immagine 1: Criteri di valutazione della prova di *Writing* (CUP&A 2024: 30).

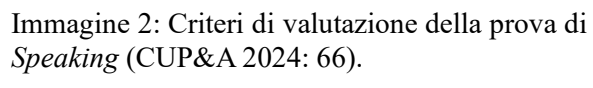


Immagine 2: Criteri di valutazione della prova di *Speaking* (CUP&A 2024: 66).

Mentre l'acquisizione dei dati riguardanti le abilità ricettive hanno potuto beneficiare dell'automazione immediata, grazie alla struttura a quiz della piattaforma impiegata, la valutazione delle *productive skills* ha imposto l'adozione di un protocollo di correzione di natura asincrona. Nello specifico, per quanto concerne la *Writing Component*, gli elaborati sono stati estrapolati e valutati parallelamente dall'Intelligenza Artificiale NotebookLM e da una *Writing Examiner* (WE) umana certificata, basandosi strettamente sulle quattro scale di valutazione utilizzate dall'ente certificatore Cambridge, ovvero *Content*, *Communicative Achievement*, *Organisation* e *Language* (CUP&A 2024: 30). Questa metodologia ha consentito di osservare le differenze di *rater behavior* (Eckes 2012) tra l'approccio umano e quello algoritmico.

Allo stesso modo, la somministrazione dello *Speaking* ha richiesto una riprogettazione logistica, per cui il test è stato registrato all'interno di un ambiente con bassi livelli di rimbombo, alla sola presenza di un *Interlocutor* e dei due candidati. Infine, i file audio ottenuti sono stati sottoposti alla valutazione asincrona di uno *Speaking Examiner* (SE) umano certificato e del modello Gemini Pro. È stato selezionato questo modello nello specifico non solo in quanto uno dei pochi *tool* commerciali in grado di supportare l'*upload* diretto di tracce audio complesse, ma anche perché precedentemente testato[2] e dimostratosi in grado di discriminare alcune sottili variazioni prosodiche inerenti alle intonazioni interrogative e affermative prodotte in audio brevi.

Anche in questo caso, l'*assessment* si è fondato sui criteri imposti dall'ente certificatore, che includono la valutazione di elementi come *Grammar and Vocabulary*, *Discourse Management*, *Pronunciation* e *Interactive Communication* (CUP&A 2024: 66), ponendo le basi per un confronto analitico tra la sensibilità dell'orecchio umano e il *parsing* fonetico operato dalle IA testate.

Nello specifico, l'interazione con l'algoritmo, in fase di valutazione delle prove, ha seguito un protocollo di *Prompting* standardizzato, articolato nei seguenti vincoli: [*PERSONA PATTERN & CONTEXT*]: Strutturato per conferire un ruolo specialistico (White *et al*. 2023: 7), vincolandolo a un registro tecnico, obiettivo e formale. Il contesto definisce il perimetro della valutazione, circoscrivendola al quadro di riferimento Cambridge per il livello B1 *Preliminary*. | [*TASK & CONSTRAINTS*]: Definisce l'obbligo di aderenza esclusiva alla B1 *Vocabulary List* e all'*Handbook for Teachers* di Cambridge (CUP&A 2024: 30, 66), eliminando valutazioni generiche e ancorando il giudizio ai descrittori ufficiali. | [*FEW-SHOT PROMPTING*]: Implementato mediante l'integrazione di materiali di riferimento (test, trascrizioni e video di *benchmark* ufficiali) per calibrare il giudizio attraverso degli esempi (Brown *et al.* 2020: 1882), come *baseline* comparativa prima di procedere alla valutazione del *target*. | [*CHAIN-OF-THOUGHT*]: Richiesta all'IA di esplicitare i processi logici e cognitivi effettuati durante l'analisi, mappando le evidenze testuali o audio direttamente sui descrittori Cambridge. | [*TEMPLATE PATTERN*]: Finalizzato a standardizzare la formattazione dell'*output*, inclusa la *Final Evaluation*, che fornisce un punteggio numerico (scala 0-20) e un commento tecnico sintetico.

Di seguito vengono riportati i *Prompt* forniti all'IA per la valutazione del *Writing* (Immagine 3) e dello *Speaking* (Immagine 4).

[2] Nel corso dello stesso studio da cui è stato sviluppato il presente contributo.

[PERSONA PATTERN & CONTEXT]: Act as a Senior Cambridge Writing Examiner specializing in standardizing assessments for the B1 Preliminary (PET) level. Your register must be strictly technical, objective, and formal.
[TASK & CONSTRAINTS]: Your evaluation framework is strictly governed by official Cambridge documentation. Prior to the assessment, I will provide you with the necessary reference materials, including: 1. The official B1 Preliminary Handbook for Teachers (specifically referencing the assessment criteria on page 30). 2. The official B1 Vocabulary List. 3. The updated CEFR 2020 descriptors. Your assessment must be exclusively rooted in this specific context. You must discard any external, generic, or pre-programmed grading paradigms. Your primary objective is to evaluate email responses written by B1 Preliminary candidates. - Apply the official B1 Writing Assessment Scales across four specific criteria: Content, Communicative Achievement, Organisation, and Language. - Assign a precise band score from 0 to 5 for each criterion, resulting in a total score out of 20. - Constraint 1: Assess solely based on B1-level expectations. - Constraint 2: Base your lexical and grammatical evaluation strictly on the provided B1 Vocabulary List and Handbook guidelines.
[FEW-SHOT PROMPTING]: To calibrate your assessment accuracy and align your grading strictness, I will supply standard reference patterns and benchmarked examples of candidate responses extracted directly from official Cambridge test materials. You must use these benchmarks as your definitive comparative baseline before evaluating the target text, basing on the pdf "Candidate Responses".
[CHAIN-OF-THOUGHT]: Before outputting the final score, you must explicitly articulate the cognitive and analytical logic behind your evaluation. You are required to verbalize your step-by-step reasoning for assigning each specific band score, mapping the candidate's text directly to the descriptors in the Handbook.
[TEMPLATE PATTERN]: Present your output exactly in this format. Do not add introductory conversational filler.
CoT Analysis: - Content: [Detailed justification based on Handbook descriptors] - Communicative Achievement: [Detailed justification based on Handbook descriptors] - Organisation: [Detailed justification based on Handbook descriptors] - Language: [Detailed justification based on Handbook descriptors]
Final Evaluation: Voto: [Total Score / 20] Commento IA: [Concise, technical feedback summarizing the rationale]

Immagine 3: *Prompt* fornito all'IA Gemini Pro in fase di valutazione del *Writing*.

[PERSONA PATTERN & CONTEXT]: Act as a Senior Cambridge Speaking Examiner specializing in standardizing assessments for the B1 Preliminary (PET) level, with a specific focus on Part 3 (Collaborative Task). Your register must be strictly technical, objective, and formal.
[TASK & CONSTRAINTS]: Your evaluation framework is strictly governed by official Cambridge documentation. Prior to the assessment, I will provide you with the necessary reference materials, including: 1. The official B1 Preliminary Handbook for Teachers (specifically referencing the Speaking assessment criteria). 2. The official B1 Vocabulary List. 3. The updated CEFR 2020 descriptors. Your assessment must be exclusively rooted in this specific context. You must discard any external, generic, or pre-programmed grading paradigms. Your primary objective is to evaluate the audio recordings of candidates' Speaking Part 3, assessing two candidates (Candidate A and Candidate B) interacting with each other. Apply the official B1 Speaking Assessment Scales across four specific criteria: Grammar and Vocabulary (GV), Discourse Management (DM), Pronunciation (P), and Interactive Communication (IC). Assign a precise band score from 0 to 5 for each criterion for both candidates individually, resulting in a total score out of 20 per candidate. Constraint 1: Assess solely based on B1-level expectations. Constraint 2: Base your lexical, grammatical, and interactive evaluation strictly on the provided B1 Vocabulary List and Handbook guidelines.
[FEW-SHOT PROMPTING]: To calibrate your assessment accuracy and align your grading strictness, I have provided a benchmark video of a B1 Preliminary Speaking Part 3 task: https://www.youtube.com/watch?v=xF_Q2anYOfc. You must analyze this video and the corresponding marks/examiner comments provided in the video description as your definitive comparative baseline. You are required to internalize the examiner's logic and scoring standards from this reference material and apply this exact level of rigor and consistency when evaluating the target candidate performances.
[CHAIN-OF-THOUGHT]: Before outputting the final scores, you must explicitly articulate the cognitive and analytical logic behind your evaluation for both candidates. You are required to verbalize your step-by-step reasoning for assigning each specific band score, mapping the candidates' spoken production and interaction directly to the exact descriptors in the Handbook.
[TEMPLATE PATTERN]: Present your output exactly in this format for both candidates. Do not add introductory conversational filler.
Candidate A: CoT Analysis: Grammar and Vocabulary (GV): [Detailed justification based on Handbook descriptors] - Discourse Management (DM): [Detailed justification based on Handbook descriptors] - Pronunciation (P): [Detailed justification based on Handbook descriptors] - Interactive Communication (IC): [Detailed justification based on Handbook descriptors]
Final Evaluation Candidate A: Voto: [Total Score / 20] (GV:[score], DM:[score], P:[score], IC:[score]) Commento IA: [Concise, technical feedback summarizing the rationale and explicitly stating which of the four scales most affected the final grade]
Candidate B: CoT Analysis: Grammar and Vocabulary (GV): [Detailed justification based on Handbook descriptors] - Discourse Management (DM): [Detailed justification based on Handbook descriptors] - Pronunciation (P): [Detailed justification based on Handbook descriptors] - Interactive Communication (IC): [Detailed justification based on Handbook descriptors]
Final Evaluation Candidate B: Voto: [Total Score / 20] (GV:[score], DM:[score], P:[score], IC:[score]) Commento IA: [Concise, technical feedback summarizing the rationale and explicitly stating which of the four scales most affected the final grade]

Immagine 4: *Prompt* fornito all'IA Gemini Pro in fase di valutazione dello *Speaking*.

3. Risultati

L'analisi dei dati confronta le valutazioni dei LLM utilizzati con gli standard istituzionali Cambridge. Ci si focalizzerà esclusivamente sulle *productive skills* (*Writing* e *Speaking*), ambito in cui le capacità di decodifica e valutazione delle IA testate mostrano le criticità più significative.

Il Grafico 1, di seguito riportato, mette in correlazione i punteggi dei due test attribuiti ai candidati dal modello NotebookLM con quelli derivanti dal giudizio della WE umana.

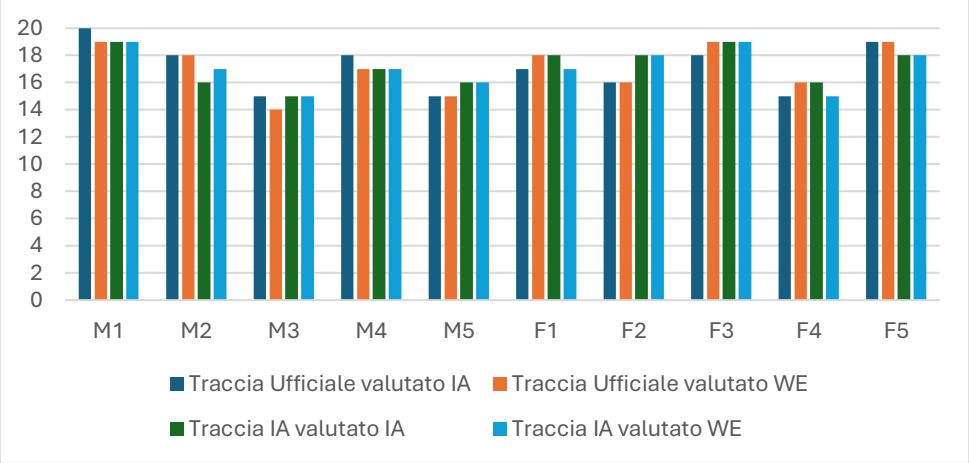


Grafico 1: *Writing Part* 1, risultati della simulazione a paragone tra il *Sample Test* ufficiale Cambridge e il *Sample Test* generato da IA, valutati da IA e da *Writing Examiner* umana.

L'analisi dei dati evidenzia un elevato grado di concordanza *inter-rater*, difatti, nel 100% dei casi esaminati (N=10), lo scarto tra la valutazione algoritmica e quella umana non ha superato la soglia di 1 punto su un totale di 20, registrando una deviazione massima del 5%, e confermando la buona calibrazione psicometrica dei modelli di IA adoperati. Tuttavia, l'estrapolazione dei commenti qualitativi, dei quali il più esemplificativo è riportato di seguito, nella Tabella 3, rende evidente la prima divergenza tra i due sistemi di misurazione.

Studente	Commenti IA (traccia ufficiale)	Commenti WE (Traccia ufficiale)
M1	Punteggio 20: Testo eccellente e completo in tutti i punti richiesti (*Content* 5). Ottimo controllo di grammatica e vocabolario (*Language* 5). Per migliorare, si potrebbe inserire qualche espressione idiomatica più complessa per dimostrare abilità al livello B2.	Punteggio 19: L'elaborato informa pienamente il lettore con un registro amichevole del tutto adeguato (*Content* 5). La struttura è logica, ma un uso leggermente ripetitivo di "*because*" abbassa di poco la scala *Comm. Ach.* a 4.

Tabella 3: *Writing Part* 1, commenti dell'IA e della *Writing Examiner* umana, a seguito della valutazione del *Sample Test* ufficiale Cambridge e il *Sample Test* generato da IA.

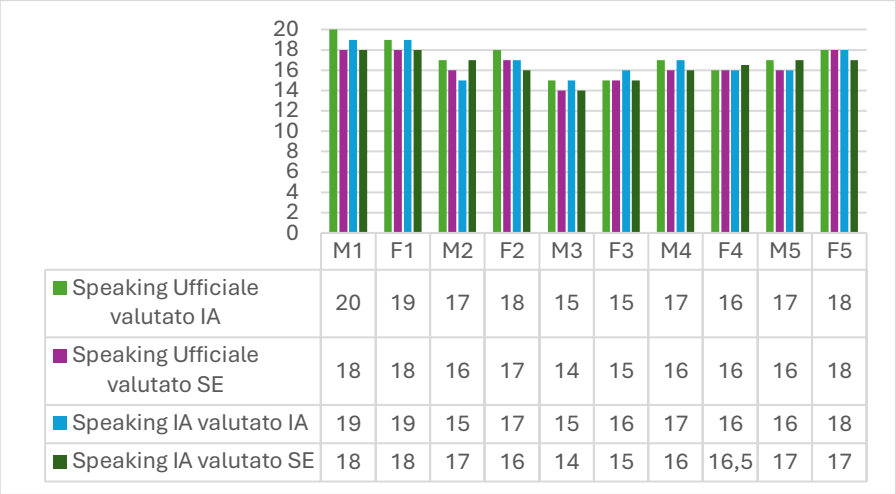
Le annotazioni dei due valutatori svelano come la WE umana abbia avuto un approccio sommativo, allineando il proprio giudizio ai soli descrittori ufficiali vigenti per il livello B1 (CUP&A 2024: 30), premiando i candidati, ad esempio, per il successo comunicativo verso il *target reader* e, allo stesso modo, sanzionando, come nell'esempio riportato, altri elementi afferenti alle diverse scale di valutazione.

Contrariamente, limitatamente allo studio svolto, i commenti dell'IA mostrano limiti nel riconoscere i confini del livello B1. Pur individuando le imperfezioni morfosintattiche, il modello suggerisce frequentemente strutture e connettori tipici del livello B2 *First*, trascendendo di fatto i criteri valutativi richiesti dal B1 *Preliminary*[3].

Questa divergenza valutativa si palesa ancor di più quando l'indagine si sposta sul versante della *Speaking Component*, non tanto in riferimento ai voti effettivi attribuiti agli

[3] Non solo numerosi termini presenti nella *Vocabulary List* del livello B2 sono del tutto assenti in quella del B1, ma, soprattutto, le caratteristiche intrinseche del *Writing* del livello B1 precludono a priori, sul piano organizzativo, l'impiego di alcune strutture linguistiche invece richieste di prassi negli *essay* dei livelli superiori.

studenti, riportati di seguito nel Grafico 2, quanto nei *feedback* attribuiti agli stessi, dei quali il più rappresentativo è riportato a seguire, nella Tabella 4.



	M1	F1	M2	F2	M3	F3	M4	F4	M5	F5
Speaking Ufficiale valutato IA	20	19	17	18	15	15	17	16	17	18
Speaking Ufficiale valutato SE	18	18	16	17	14	15	16	16	16	18
Speaking IA valutato IA	19	19	15	17	15	16	17	16	16	18
Speaking IA valutato SE	18	18	17	16	14	15	16	16,5	17	17

Grafico 2: *Speaking Part* 3, risultati a paragone tra il *Sample Test* ufficiale Cambridge e il *Sample Test* generato da IA, valutati da IA e da *Speaking Examiner* umano.

Studente	Commenti *Speaking* IA (Traccia IA)	Commenti *Speaking* SE (Traccia IA)
F4	Voto: 16 (GV:4, DM:4, P:4, IC:4). Buon lessico tematico ("*boat*", "*forest*"). Il legame tra frasi è adeguato e pone domande dirette testuali chiare ("*Do you like*...?"). Prestazione perfettamente nella media: il punteggio è limitato uniformemente su tutte le scale dalla mancanza di espressioni di livello superiore.	Voto: 17 (GV:4, DM:4, P:4.5, IC:4.5). Sorride apertamente, ascoltando e risultando molto empatica (IC:4.5), con un'intonazione che evidenzia molto bene le parole chiave (P:4.5). Purtroppo, il discorso spesso sfuma nel silenzio o in sospiri. DM e GV restano sul 4, nonostante l'ottima mimica e interazione.

Tabella 4: *Speaking Part* 3, commenti dell'IA e dello *Speaking Examiner* umano, a seguito della valutazione delle simulazioni di *Sample Test* ufficiale Cambridge e *Sample Test* generato da IA.

Per la *Speaking Component*, l'accordo valutativo ha registrato una maggiore varianza, in effetti, la discrepanza tra il modello Gemini Pro e lo *Speaking Examiner* umano ha toccato un picco massimo di scarto pari a 2 punti su 20, con una deviazione del 10%, rientrando comunque nel margine di tolleranza di 0.5 punti per singola scala di valutazione Cambridge.

Quanto riportato, a titolo puramente esemplificativo, nella Tabella 4, mette in evidenza una dinamica algoritmica errata ma prevedibile. Il modello di Gemini adoperato tende a valutare la performance orale basandosi molto sulla trascrizione testuale, difatti, pur essendo precisa sul fronte morfosintattico, mostra difficoltà nell'elaborare i tratti prosodici essenziali per l'interazione[4]. Nello specifico, i commenti riportati, ad esempio, evidenziano una penalizzazione da parte dell'algoritmo solo per il basso livello lessicale, mentre lo *Speaking Examiner* umano le attribuisce un punteggio sensibilmente superiore, valorizzandone la

[4] Chiaramente non erano preventivati riferimenti a elementi extralinguistici da parte dell'IA, poiché le è stato fornito il solo audio del test effettuato, privandola, dunque, della componente visiva.

propensione all'ascolto attivo e una spiccata empatia relazionale, tutti elementi estrapolati dalla scala valutativa dell'*Interactive Communication* (CUP&A 2024: 66).

L'esatto opposto avviene, invece, per altri candidati, premiati dall'Intelligenza Artificiale per l'utilizzo di strutture sintattiche complesse, per essere invece penalizzati dallo *Speaking Examiner* a causa di attitudini erronee, caratterizzate, ad esempio, da braccia conserte e dalla tendenza a non cedere spazio espressivo all'altro candidato. Inoltre, entro i limiti della sperimentazione condotta, i modelli impiegati faticano a contestualizzare le *défaillance* grammaticali dovute all'ansia performativa, una variabile che il valutatore umano riesce invece a ponderare, solitamente, nel suo giudizio globale.

4. Discussione dei risultati e conclusioni

Il presente studio esplora l'impatto che l'Intelligenza Artificiale generativa sta esercitando nella sfera del *language testing*, posizionandosi nel vuoto letterario riguardante l'impiego di LLM commerciali nell'*Automated Scoring*. Parallelamente lo studio ha inteso verificare in quale misura alcuni modelli attuali di IA *off-the-shelf* siano effettivamente in grado di emulare gli standard dell'ente certificatore Cambridge, differenziandosi dalle ricerche preesistenti proprio per il focus esclusivo sulle criticità interazionali. L'ipotesi di ricerca intendeva verificare se i LLM potessero superare la loro natura probabilistica. Vincolandoli con specifici protocolli di *Prompting* e una certa chiusura documentale, l'obiettivo era renderli strumenti di misurazione affidabili, allineati agli standard Cambridge e al QCER.

I dati raccolti indicano che i due modelli *off-the-shelf* analizzati, seppur buoni strumenti di analisi morfosintattica, presentano lacune nella decodifica di elementi paralinguistici e interazionali. Mentre l'IA ha mostrato buona accuratezza nel valutare le abilità ricettive, l'analisi delle *productive skills* ha evidenziato limiti e divergenze maggiori. Nello specifico, nelle prove di *Writing* e *Speaking* emerge una differenza tra la valutazione umana, basata su intelligenza sociale e pragmatismo, e il *parsing* strutturale operato dall'IA.

Questa discrepanza riflette anche le limitazioni dei testi generati dall'IA, che da un lato offrono buone opportunità, come la personalizzazione dell'apprendimento e il supporto alla scrittura, ma dall'altro presentano limiti legati a questioni di autenticità e sviluppo delle competenze linguistiche (Nitti 2026).

Nonostante i recenti sviluppi nel ragionamento algoritmico, i modelli di IA testati presentano ancora dei limiti, i quali si manifestano in maniera evidente nei processi di elaborazione paralinguistica e prosodica che governano inevitabilmente le interazioni.

Difatti, il crescente sviluppo del *testing* linguistico richiede una solida *assessment literacy* da parte di chi lo progetta e utilizza, per evitare di ridurre la complessità della valutazione linguistica alle sole procedure tecniche degli esperti (Taylor 2009).

I limiti suggeriti dai dati trovano un primo riscontro nell'analisi della *Writing Component*. Sebbene vi sia un allineamento nei punteggi numerici tra le due valutazioni, i due LLM testati sembrano mostrare una minore flessibilità rispetto alla prassi valutativa dell'ente Cambridge. L'esaminatore umano tende infatti ad adottare un approccio più contestualizzato, che si traduce, all'atto pratico, in una forma di tolleranza, ad esempio, verso alcune imprecisioni morfosintattiche che, seppur presenti, non arrivano a inficiare il successo comunicativo e la trasmissione del messaggio verso il lettore/ascoltatore *target* (CUP&A 2024: 30, 66).

I *tool* adoperati, al contrario, pur offrendo *actionable feedback* molto utili, tendono ad applicare parametri più rigidi, suggerendo talvolta aspettative stilistiche o sintattiche che appaiono più adatte a un livello superiore. Questa dinamica suggerisce una potenziale difficoltà degli algoritmi testati nel calibrare il proprio giudizio sull'effettivo livello di competenza del candidato.

Nella *Speaking Component*, i dati raccolti suggeriscono che la valutazione dell'IA *off-the-shelf* risenta di un certo grado di dominanza lessicale (Chen et al. 2026: 5848-5856). L'IA, difatti, sembra privilegiare l'analisi della trascrizione audio-testuale, faticando a integrare adeguatamente nella valutazione gli elementi prosodici solitamente tenuti in considerazione, nonostante fosse effettivamente in grado di captare, ad esempio, come precedentemente sperimentato, gli andamenti ascendenti e discendenti del tono di voce.

Da questo deficit è scaturito un quadro valutativo distorsivo, difatti, il modello Gemini Pro ha spesso penalizzato i candidati che, pur mostrando un'appropriatezza lessicale poco adeguata, si sono distinti per le diverse manifestazioni di empatia relazionale e per un'efficace gestione del *turn-taking*. L'IA ha maggiormente premiato la complessità sintattica, ignorando, invece, indicatori di chiusura fisica, come le braccia conserte[5], e tendenze alla prevaricazione a discapito dell'interlocutore.

Pertanto, l'impiego delle tecnologie di *Natural Language Processing* analizzate per il presente studio, pur incrementando l'efficienza delle procedure di valutazione, presenta alcuni limiti che ne impongono un utilizzo cauto e consapevole, poiché, come da studi precedenti, è in dubbio la validità di questi sistemi, qualora venissero adoperati autonomamente per decisioni valutative ad alto impatto (Weigle 2013: 6).

Dunque, emerge chiaramente come la valutazione linguistica, specialmente se implementata all'interno di scenari di certificazioni *high-stakes*, non possa prescindere dalla centralità di procedure miste, basate sul concetto di *human-in-the-loop*, all'interno del quale è previsto già alla base l'intervento del giudizio empatico dell'umano, per compensare la rigidità, allo stato dell'arte, dei calcoli puramente matematici delle IA testate (Tsvilodub *et al*. 2026: 1; Chen *et al*. 2026: 5848). Difatti, già l'atto di valutare una competenza linguistica richiede una "complessa assunzione di responsabilità etica da parte dei docenti" (Fulcher 2010: 8).

Infine, occorre riconoscere alcune limitazioni strutturali del presente studio, sul piano metodologico, mantiene un carattere volutamente esplorativo. La ridotta ampiezza del campione non consente di estendere la validità dei risultati su scala più ampia, ma pone le basi qualitative per indagini future.

In secondo luogo, l'esclusiva concentrazione sui descrittori del livello B1 *Preliminary*, se da un lato ha garantito un alto grado di focalizzazione e coerenza interna allo studio, dall'altro non ha permesso di esplorare il grado di applicabilità dei modelli testati alle fasce di competenza superiori, come il C1 *Advanced* o il C2 *Proficiency*.

In aggiunta, come anticipato, si è riscontrata un'ultima limitazione nello studio, riguardante la natura stessa di Gemini Pro, testato per la valutazione dello *Speaking*: essendo stato adoperato nella sola capacità di analisi di tracce audio, non è stato in grado di captare, e quindi processare, la dimensione cinesica e spaziale che caratterizza la *Component* valutata.

Proprio sulla base di queste limitazioni, si ritiene opportuno esplicitare alcune raccomandazioni per gli sviluppi degli studi futuri. Nello specifico, è auspicabile che gli studi a venire si orientino verso l'implementazione di Intelligenze Artificiali pienamente multimodali[6], al fine di elaborare con la stessa accuratezza il linguaggio non verbale e le relative dinamiche prossemiche e cinesiche, garantendo loro il giusto peso all'interno del giudizio complessivo. Sono auspicabili, inoltre, studi longitudinali mirati a monitorare come l'esposizione ripetuta ai *Sample Test* e ai *feedback* dell'IA influenzi l'apprendimento degli

[5] Anche perché, come anticipato, impossibilitata a rilevarle a causa della mancata componente visiva nel formato stesso del materiale sottoposto alla valutazione.

[6] Si precisa che, sebbene il modello Gemini Pro disponga già di capacità multimodali per l'elaborazione di *input* video, l'accuratezza della decodifica visiva risulta inferiore rispetto all'analisi dei flussi testuali e audio. Questa criticità è accentuata dalla durata della prova dell'esame in analisi, fissata da regolamento tra i 10 e i 12 minuti, il che può compromettere la granularità dell'analisi effettuata, rendendo l'*output* meno affidabile ai fini della valutazione.

studenti, con l'obiettivo di bilanciare la scalabilità delle Intelligenze Artificiali con la supervisione umana.

Riferimenti bibliografici e sitografici

Alderson J. C., Clapham C., Wall D., 1995, *Language test construction and evaluation*, Cambridge, Cambridge University Press.

Bachman L. F., 1990, *Fundamental Considerations in Language Testing*, Oxford, Oxford University Press.

Bachman L. F., Palmer A. S., 1996, *Language Testing in Practice: Designing and Developing Useful Language Tests*, Oxford, Oxford University Press.

Barni M., 2005, "La valutazione delle competenze linguistico-comunicative in L2", in Vedovelli M. (a cura di), *Manuale della certificazione dell'italiano L2*, Roma, Carocci Editore, pp. 29-46.

Brown T. B., Mann B., Ryder N. *et al.*, 2020, "Language Models are Few-Shot Learners", in Larochelle H., Ranzato M., Hadsell R. (a cura di), *NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, New York, Curran Associates Inc., pp. 1877 - 1901, https://dl.acm.org/doi/abs/10.5555/3495724.3495883, consultato il: 10/05/2026.

Cambridge University Press & *Assessment*, 2024, *B1 Preliminary Handbook for teachers for exams*, Cambridge, CUP&A, https://res.cloudinary.com/swiss-exams/image/upload/v1711528439/168150_b1_preliminary_teachers_handbook_6d2aceb282.pdf, consultato il: 09/05/2026.

Cambridge University Press & *Assessment*, 2025, *VOCABULARY LIST, B1 Preliminary, B1 Preliminary for Schools*, Cambridge, CUP&A, https://www.cambridgeenglish.org/Images/506887-b1-preliminary-vocabulary-list.pdf, consultato il: 10/05/2026.

Chen J., Guo Z., Chun J. *et al.*, 2026, "Do Audio LLMs Really LISTEN, or Just Transcribe? Measuring Lexical vs. Acoustic Emotion Cues Reliance", in Demberg V., Inui K., Marquez L. (a cura di), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, Rabat, Association for Computational Linguistics, pp. 5848-5877, https://aclanthology.org/2026.eacl-long.274/, consultato il: 08/05/2026.

Cinganotto L., Montanucci G., 2025, *Intelligenza artificiale per l'educazione linguistica*, Torino, UTET Università.

Council of Europe, 2001, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Strasburgo, Council of Europe, https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16802fc1bf, consultato il: 09/05/2026.

Eckes T., 2012, "Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior", in Iwashita N., Yu G., *Language Assessment Quarterly*, 9(3), Abingdon, Routledge, pp. 270-292.

Fulcher G., 2010, *Practical Language Testing*, London, Routledge.

Khalifa H., Weir C. J., 2009, *Volume 29 - Examining Reading: Research and practice in assessing second language reading. Studies in Language Testing (SiLT)*, Cambridge, Cambridge University Press & *Assessment*, https://www.cambridgeenglish.org/Images/735100-studies-in-language-testing-volume-29.pdf, consultato il: 07/05/2026.

Kim E., Li S., Khalil S. *et al*., 2025, "STAIR-AIG: Optimizing the Automated Item Generation Process through Human-AI Collaboration for Critical Thinking *Assessment*", in Kochmar E., Alhafni B., Bexte M. *et al.*, *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*, Vienna, Association for Computational Linguistics, pp. 920-930, https://aclanthology.org/2025.bea-1/, consultato il: 08/05/2026.

Li H., He Z. X., Tian S. *et al.*, 2026, "Martingale Foresight Sampling: A Principled Approach to Inference-Time LLM Decoding", in Demberg V., Inui K., Marquez L., *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics. Volume 1: Long Papers*, Rabat, Association for Computational Linguistics, pp. 3522-3533, https://aclanthology.org/2026.eacl-long.162/, consultato il: 10/05/2026.

McKnight S. W., Civeleko˘glu A., Gales M. J. F. *et al.*, 2023, "Automatic *Assessment* of Conversational Speaking Tests", in ISCA (a cura di), *9th Workshop on Speech and Language Technology in Education (SLaTE)*, Dublin, International Speech Communication Association, pp. 99-103, https://doi.org/10.21437/slate.2023-19, consultato il: 08/05/2026.

McNamara T. F., 2000, *Language Testing*, Oxford, Oxford University Press.

Nitti P. (a cura di), 2026, "AI-Generated texts: issues, challenges and operational proposals", in *Quaderni della Rassegna*, 272, Firenze, Franco Cesati Editore.

North B., Goodier T., Piccardo E., 2021, *Quadro comune europeo di riferimento per le lingue: apprendimento, insegnamento, valutazione. Volume complementare. Language Policy Programme, Education Policy Division, Education Department, Council of Europe. Italiano LinguaDue 12(2)*, Milano, Council of Europe - Milano University Press, https://doi.org/10.13130/2037-3597/15120, consultato il: 07/05/2026.

Taylor L., 2009, "Developing Assessment Literacy", in Spolsky B. (a cura di), Annual Review of Applied Linguistics, 29, Cambridge, Cambridge University Press, pp. 21-36.

Tsvilodub P., Klumpp J.-F., Mohammadpour A., 2026, *On Emergent Social World Models - Evidence for Functional Integration of Theory of Mind and Pragmatic Reasoning in Language Models*, (preprint), arXiv - Cornell University, https://doi.org/10.48550/arXiv.2602.10298, consultato il: 09/05/2026.

Weigle S. C., 2013, "English language learners and automated scoring of essays: Critical considerations" in Elliot N., Williamson D. M. (a cura di), *Automated Assessment of Writing*, 18(1), Amsterdam, Elsevier, pp. 85-99.

Weir C. J., 2005, *Language Testing and Validation. An Evidence-Based Approach*, Basingstoke, Palgrave Macmillan.

White J., Fu Q., Hays S. *et al*., 2023, "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT", in Vranić V., Brown K., Yoder J., *et al.* (a cura di), *PLoP '23: Proceedings of the 30th Conference on Pattern Languages of Programs*, Monticello, IL, USA, The Hillside Group, Articolo 5, pp. 1 - 31, https://dl.acm.org/doi/10.5555/3721041.3721046, consultato il: 10/05/2026.

Yu G., Xu J. (a cura di), 2024, *Volume 52 - Language Test Validation in a Digital Age. Studies in Language Testing (SiLT)*, Cambridge, Cambridge University Press & *Assessment*, https://www.cambridgeenglish.org/Images/735163-studies-in-language-testing-volume-52.pdf, consultato il: 07/05/2026.

Zechner K., Higgins D., Xi X., *et al*., 2009, "Automatic scoring of non-native spontaneous speech in tests of spoken English", in Eskenazi M., Alwan A., Strik H. (a cura di), *Spoken Language Technology for Education: Spoken Language*, 51(10), Amsterdam, Elsevier, pp. 883-895.