Online-Journals.org

iJAC International Journal of Advanced Corporate Learning

iJAC | eISSN: 1867-5565 | Vol. 19 No. 1 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijac.v19i1.58935

TLIC PAPER

Low-Resource Strategies for IELTS Writing Preparation using Free Generative AI

Rohib Adrianto Sangia (✉)

University of Aberdeen, Aberdeen, United Kingdom

r.sangia.22@abdn.ac.uk

ABSTRACT

The study assesses whether free, web accessible generative AI can produce IELTS Academic Writing band scores and feedback credible for formative use in low-resource contexts. A corpus of 110 official Task 1 and Task 2 responses was scored by four systems accessed at no cost: ChatGPT with GPT 4, Gemini 1.5 Flash, Claude 3.5 Sonnet, and DeepSeek. Agreement with official scores was quantified using Lin's concordance correlation coefficient and the Bland-Altman method. Feedback was coded for structure, depth, specificity, and tone, with intercoder reliability estimated by Krippendorff's alpha. Two systems showed minimal central bias with limits of agreement, while one system systematically underestimated bands. Agreement within a one-band tolerance was high. Feedback was criterion-linked and actionable but often lacked precise phrasing and had uneven clarity. Qualitative coding reliability was acceptable. Findings suggest that free AI can offer indicative banding and revision cues when combined with rubric checking and human oversight, especially in situations where bandwidth is limited. The study logs access conditions to support replication and sets priorities for self-directed learning.

KEYWORDS

IELTS Academic Writing, generative AI, scoring agreement, feedback quality, low resource settings

1 INTRODUCTION

Inequities in IELTS Academic Writing preparation are particularly pronounced in remote and rural, low-income settings, where students face limited teacher availability, high travel costs, weak internet connectivity, and scarce study materials. Hattie and Timperley [1] argue that what learners need most is actionable feedback that is specific, timely, and linked to clear next steps. When feedback arrives weeks after a practice essay, the learning effect diminishes, and delays of more than about one to two days have been shown to reduce gains [2]. In many crowded classrooms,

Sangia, R. A. (2026). Low-Resource Strategies for IELTS Writing Preparation using Free Generative AI. *International Journal of Advanced Corporate Learning (iJAC)*, 19(1), pp. 93–104. https://doi.org/10.3991/ijac.v19i1.58935

This article is an expanded version of a paper presented at The Learning Ideas Conference, held in New York, NY, USA, June 11–13, 2025. Article submitted 2025-09-29. Revision uploaded 2025-10-29. Final acceptance 2025-12-03.

reliability also suffers because teachers have little time and uneven assessment training, which leaves self-directed learners without consistent guidance [3]. For example, a village student who writes a Task 2 response on public transport may wait until the next school visit to hear anything useful, by which time the draft and the motivation have both gone cold.

Free, web-based generative AI (GenAI) offers a pragmatic solution by providing instant feedback and rough band estimates, which encourage more frequent practice [3], [4]. Hattie and Timperley [1] explain why such feedback can serve as a scaffold for planning, monitoring, and evaluation. Huang and Mizumoto [5] report that many students felt that immediate comments helped them identify specific weaknesses. At the same time, scoring agreement with human raters is not guaranteed, and risks such as bias, hallucinated advice, privacy concerns, and uneven access temper the promise [3], [6]. This study treats these tools as support for learning rather than as systems for certification or proctoring, and any band estimate is used only to guide revision, not to replace human judgment.

The alignment construct examines how close scores and feedback from free AI tools align with human judgments and official IELTS rubrics. In self-directed learning, this alignment supports sound self-assessment and targeted practice, as noted by Younas, et al. [3] and Song and Song [7]. Agreement rather than simple association is the relevant construct for learner decision making, because high correlation can coexist with significant absolute errors [7]. Song and Song [7] warn that high correlations may mask persistent discrepancies that mis-calibrate learners.

Alongside scoring, feedback acts as instructional support by pointing out the exact next steps that help learners plan, monitor, and evaluate their writing in practical ways [3], [7]. Zimmerman [8] describes self-regulated learning as a cycle of planning, monitoring, and evaluation, and Jin, et al. [4] show that AI assistance can scaffold this cycle. Hattie and Timperley [1] and Shute [2] emphasize that quality feedback shows structure, provides depth beyond surface errors, targets specific points, and maintains a motivational tone that suits the learner. For IELTS, comments should align with the following criteria: Task Achievement or Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy, providing explicit criterion references and concrete next steps that students can apply in the next draft [6], [7]. A practical example is a Task 1 note that advises adding a summary sentence of key trends, a Task 2 tip to vary linking devices, and a Lexical Resource hint to replace ‘very’ with more precise modifiers. Evidence suggests gains in confidence and planning when such targeted feedback is paired with reflection and tracking; however, reports also warn of hallucinations, bias, and privacy risks that can erode trust or fairness [4], [6]. Operationalization can use a rubric that checks for criterion references, task-linked goals, concrete actions, and a tone matched to the level, extending the guidance set out by Hattie and Timperley [1] and by Shute [2].

The IELTS Writing band is the target outcome for each sample answer, meaning that each response to Task 1 or Task 2 is evaluated on a 0 to 9 scale in 0.5 increments [9]. The bands are anchored to descriptors that cover Task Achievement or Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy [9]. Some might argue that a range better reflects uncertainty in borderline cases, yet IELTS intentionally removes that flexibility to test alignment cleanly [9]. In practice, removing ranges also prevents later disputes about which end of the interval should drive feedback priorities. The decision aligns with self-directed learning because students can more easily compare their progress with that of their peers across study sessions [6]. However, Kim, et al. [6] also note that students often appreciate nuance in feedback, which means a single score must be paired with comments to feel fair.

We therefore maintain a single-valued scoring system while encouraging qualitative notes in the data collection forms, as the rubrics themselves are descriptive rather than purely numeric [9]. This establishes whether free AI tools supply scores that align with human judgment without confusing band labels.

The primary aim of this study is to assess the degree of alignment between free GenAI models and official IELTS Writing band scores for Tasks 1 and 2. The focus is on ChatGPT, Claude, DeepSeek, and Gemini, all of which can be accessed online at no cost to learners. Agreement is treated as the closeness and consistency of AI scores to human rater scores for the same sample answer. To make the scores useful for learner decision-making, the study will examine agreement, bias, and reliability together. Our first research question asks to what extent the four GenAI models produce band scores that agree with official scores across GenAI models. Timely, indicative scores can support planning and revision in self-directed learning when they are accurate and precise [1], [2]. Hattie and Timperley [1] argue that feedback and goals are most effective when linked to specific criteria that learners can see and use. Jin, et al. [4] report that misplaced confidence can alter strategy choices and slow real progress in writing practice. For low-resource contexts, even small biases matter because learners may not have fast access to expert correction [2], [6].

The secondary aim is to assess the usefulness of AI feedback for self-directed improvement in IELTS Writing. Actionable feedback is defined as specific, timely, and directly tied to strategies that a learner can apply. Hattie and Timperley argue that practical feedback answers the questions of where to go next, how to close the gap, and what quality looks like [1]. Shute [2] notes that shorter delays boost motivation and help knowledge stick when feedback is clear and targeted. The study will code feedback on structure, depth, specificity, and motivational tone, aligned to the IELTS descriptors for Task Achievement or Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. Our second research question asks how the feedback qualities from the four models align with human marker expectations and the band descriptors. Kim and colleagues report that students value encouragement but become wary when comments feel generic or detached from the text [6]. Ahmed, et al. [10] caution that GenAI may hallucinate details or gloss over errors, and they also raise concerns about privacy in educational settings. A concrete example is a Task 1 report that lists data without an overview, where helpful feedback prompts the inclusion of a clear summary sentence and a comparison of the main trends. For Task 2, actionable guidance might require a direct thesis in the introduction and topic sentences that clearly link back to the question. Such moves are linked to better planning, monitoring, and evaluation in self-regulated learning [4]. The motivational tone will be checked for warmth and balance to ensure that praise does not overshadow the precise next steps. There is a benefit to a friendly tone for self-efficacy; however, empty praise can inflate confidence without improving writing quality [4], [6]. Timeliness and clarity will also be rated because both are known to affect engagement and persistence in practice [2]. The resulting map of strengths and gaps will guide the staged use of AI in low-resource contexts, including decisions about when to trust AI advice and when to seek human review.

The study contributes to self-directed learning in low-resource settings by demonstrating how free, web-accessible GenAI tools, such as ChatGPT, Claude, DeepSeek, and Gemini, can support IELTS Academic Writing practice. Self-directed learning is defined as a learner taking initiative to set goals, choose resources, and assess progress, especially in contexts where teachers or broadband access is scarce [6], [11]. Lin [12] argues that verification steps, such as rubric checks,

plagiarism screening, and triangulating outputs across tools, reduce these risks and related protocols. For example, a student could draft a Task 2 essay, request criterion-based scoring, compare each point to the official descriptors, revise, and then sample a second model before committing changes. Learners can also use offline-first habits by drafting locally and batching feedback requests to save data in areas with weak connectivity [11]. When model outputs conflict, appear vague, or contradict rubric language, learners should seek a second AI opinion or ask a qualified peer or teacher if one is available. Findings apply solely to the four free models available in the study window and do not generalize to paid tiers, plug-ins, or other AI systems that may behave differently over time. Risks such as bias, hallucinations, privacy concerns, and inequitable access are ongoing limitations. The use of guardrails, including rubric verification and cross-model comparison, is considered essential for the responsible adoption and implementation of new models.

2 MATERIALS AND METHODS

This study employed a mixed-methods explanatory design, first quantifying score agreement and then examining the quality of feedback. In the quantitative phase, we compared AI band scores with official human scores across two IELTS writing tasks. The qualitative phase analyzed the written feedback that the models attach to those scores to judge the usefulness for learning. Integration occurred through joint displays and triangulation, which established numeric agreement alongside feedback profiles. Lin [12] is followed to maintain reproducible and straightforward prompts across tools. Evidence in educational measurement emphasizes that agreement is not equivalent to association; therefore, we selected methods that assess both accuracy and precision [13], [14]. Scholars of AI feedback argue that it can enhance motivation for self-directed learning; however, others caution that overuse can distort judgment and create a false sense of certainty [7], [15]. The design, therefore, treats the numbers and the words as connected pieces of the same puzzle without assuming that high agreement implies high educational value.

The unit of analysis was the official sample answer, which was a written response to either Task 1 or Task 2, as defined by IELTS. Task 1 involves describing visual data, and Task 2 requires an argumentative essay. We analyzed them separately while also producing pooled estimates to show the overall pattern across the corpus. The corpus comprised 110 sample answers, sourced from Cambridge IELTS Student Book Editions and the official IELTS website, with the edition, year, and page range or URL recorded for each item. The study employed both quantitative and qualitative methodologies to assess the effectiveness of GenAI technologies in comparison to human evaluators. A dataset of 110 authentic essays was selected from official preparation materials to ensure validity and relevance. Four leading GenAI models were selected for their multimodal input capabilities, specifically ChatGPT-4 (released in June 2024), Gemini 2.0 Flash (released in July 2024), Claude 3.5 Sonnet (released in October 2024), and DeepSeek (released in June 2024). Only sample answers with clear provenance, complete text, and permission for educational use were included following fair dealing guidance for research, and the final set lists 110 entries as the analytic base [16]. Each sample answer had an official human band serving as the criterion against which GenAI scores were judged, and when only per-criterion bands were available, we computed the composite as the mean of the four criteria and rounded to the nearest half band using IELTS guidance with ties such as 0.25 or 0.75 rounded up in line with standard practice [9], [17]. The GenAI

systems under test were accessed via their public free interfaces, and for each, we recorded the date, time, version cues, and any rate limits or refusals, and following Lin's recommendation for stable prompts we used fixed templates with default settings, entered no personal or sensitive data, and removed any residual metadata prior to storage, and complied with platform terms of service [12]. To examine variability, we repeated runs on a subset and stored the raw outputs locally for audit, while logging session interruptions and replacing items from the same domain as needed to maintain balanced topic coverage. A single prompt protocol was used across systems with the exact wording requesting one precise IELTS band so that inputs contained only the official prompt and full answer without rubrics or chain of thought, with temperature and randomness left at defaults, seed controls noted if absent, and non-conforming outputs were post-processed using a pre-registered rule to extract the first valid band or mark the case as missing.

Agreement between GenAI and human scores was quantified using Lin's Concordance Correlation Coefficient, Bland–Altman analysis, and Krippendorff's alpha. [14], [18] defines CCC as the product of Pearson correlation, which captures precision, and a bias correction term that captures accuracy. We computed the CCC for each model against the official score and reported task-specific and pooled values [19]. We report model-specific bias and limits of agreement to accompany the outside limits of agreement proportions [20]. Krippendorff [21] provided a reliability metric for multiple raters over interval data, using squared distance, along with bootstrap confidence intervals. Because bands are discrete, we note that CCC and alpha can be muted by restricted range, a point that encourages cautious interpretation rather than triumphal claims [17]. Sensitivity checks were performed by removing sample answers at the floor (band four or below) and at the ceiling (band eight or above) to see whether mid-range agreement behaved differently. Analyses were reported separately for Task 1 and Task 2 to reflect their distinct rhetorical demands and to investigate whether topic and structure influence the error profile [7].

The qualitative strand analyzed the models' written feedback that accompanied each score, capturing it verbatim and unedited for every sample answer and model. A codebook was developed a priori to define the structure, depth, specificity, and motivational tone, with alignment anchors drawn from the IELTS descriptors for Task Achievement, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. Two independent coders were trained on the scheme, piloted on fifteen sample answers, and then coded the corpus. For qualitative coding, we calculated Krippendorff's alpha per dimension across coders. For band score agreement, we did not use alpha per model because alpha is a multi-rater index. The use of public sample answers and official scores eliminated the need for personal data, and storage adhered to encrypted-at-rest and in-transit practices with role-based access [16]. Learner guidance emphasized responsible use, with encouragement to triangulate AI comments with teachers and peers, a stance praised by some for its balance but critiqued by others who fear it may still normalize over-reliance on tools [15], [22]. Equity notes addressed low-bandwidth realities through offline drafting, batching queries, and simple interfaces, ensuring that rural learners are not left out, in line with sector recommendations for inclusive design [23].

3 RESULTS

We define differences as the model minus the official answer across 110 Task 1 and Task 2 sample answers. Therefore, negative values indicate underscoring relative to

the reference, and the across-model means sit at approximately minus 0.38 bands. ChatGPT shows a strong underscoring tendency with a mean of about minus 0.92 and a median of minus 1.00, and it records 90 under, three overs, and 17 exact matches (15.5%) while roughly 40.9% fall within half a band and 72.7% within one band, and an illustrative case is Task 2 – Sample Text 35 where official is 6.0 and ChatGPT gives 3.0. Claude is gentler with a mean near minus 0.37 and a median of 0, and it hits 34 exact matches (30.9%) with 64.5% within half a band and 84.5% within one band while ranging from minus 3.0 to plus 1.5, and its tilt remains low with 53 under and 23 overs, including 33 cases more than a band below official. DeepSeek is closest at the center with a mean around minus 0.17 and a median of 0, and it reaches 36 exact matches (32.7%) with 73.6% within half a band and 94.5% within one band while ranging from minus 2.5 to plus 1.5, and it still shows 47 under and 27 overs including a notable overshoot in Task 2 – Sample Text 10 where official is 4.0 and DeepSeek gives 5.5. Gemini is nearly unbiased, with a mean of minus 0.05 and a median of 0. It delivers 30 exact matches (27.3%), with 62.7% within half a band and 93.6% within one band, spanning a range of –2.0 to +2.5. Overestimation pockets exist, such as Task 2 – Sample Text 6, where the official is 3.5 and Gemini gives 6.0. In aggregate, this includes four Gemini and two DeepSeek cases that exceed a full band. Tails matter for decision-making because about 27.3% of ChatGPT's outputs stray beyond one band, whereas only about 6 to 7% of Gemini and roughly 5% of DeepSeek do so, and the within-one-band counts sketch a practical envelope with 80 for ChatGPT, 93 for Claude, 104 for DeepSeek, and 103 for Gemini. Precision at the half-band level diverges, as around 41% of ChatGPT cases meet this tighter tolerance, and roughly 74% of DeepSeek cases do likewise. This spread signals different risk profiles for fine-grained calibration. Overestimation is relatively rare for ChatGPT, with only 3 cases, but more common for Gemini, at 35, and for DeepSeek, at 27, while Claude sits at 23. Underestimation dominates for ChatGPT at 90, remains prevalent for Claude at 53, moderates for DeepSeek at 47, and is notable for Gemini at 45, so three models have medians at zero. In contrast, ChatGPT scores minus one, indicating that DeepSeek and Gemini cluster most closely with the official data, while Claude and, especially, ChatGPT warrant more caution, as all figures are derived from the dataset.

Table 1. Quantitative findings

Model	CCC	Bias	SD of Differences	Outside LOA	Krippendorff's α (overall)
ChatGPT	0.514	–0.923	0.733	0.054	0.636
Claude	0.722	–0.373	0.825	0.064	0.636
DeepSeek	0.747	–0.168	0.692	0.046	0.636
Gemini	0.622	–0.046	0.783	0.054	0.636

The Table 1 summarizes quantitative descriptors for each model against the Official Score as the reference, and the concordance correlation coefficient (CCC) values are 0.514 for ChatGPT, 0.722 for Claude, 0.747 for DeepSeek, and 0.622 for Gemini. Bias is defined as Model minus Official, and, using the Official Score as the center, all four biases are negative, with values of –0.923 for ChatGPT, –0.373 for Claude, –0.168 for DeepSeek, and –0.046 for Gemini. Furthermore, for example, a bias of –0.923 for ChatGPT corresponds to an average model to official difference of 0.923 bands in the negative direction. The spread of pairwise differences, captured by

the standard deviation (SD), is 0.733 for ChatGPT, 0.825 for Claude, 0.692 for DeepSeek, and 0.783 for Gemini, and across models, the SD ranges from 0.692 to 0.825. The proportion of observations outside the Bland–Altman limits of agreement is 0.054 for ChatGPT, 0.064 for Claude, 0.046 for DeepSeek, and 0.054 for Gemini, and these outside-LOA proportions fall between 0.046 and 0.064 as tabulated. Krippendorff's alpha is reported as an overall index across models, with a value of 0.636 for all four entries. All figures use one decimal place for bands and three decimal places for indices in the table. Ordered by the reported CCC, the sequence is ChatGPT 0.514, Gemini 0.622, Claude 0.722, and DeepSeek 0.747, and ordered by the magnitude of the recorded bias, the sequence is ChatGPT –0.923, Claude –0.373, DeepSeek –0.168, and Gemini –0.046, while CCC values overall occupy the interval 0.514 to 0.747 under this dataset description. Considering dispersion and outlier share, the SD ordering is Claude 0.825, Gemini 0.783, ChatGPT 0.733, and DeepSeek 0.692, and the outside-LOA ordering is Claude 0.064, ChatGPT 0.054, Gemini 0.054, and DeepSeek 0.046. A compact example shows Official versus DeepSeek with bias –0.168 and SD 0.692, while Official versus Gemini shows bias –0.046 and SD 0.783, and proponents emphasize CCC as a single concordance summary. At the same time, critics argue that different plots and the sharing of outside limits better describe pairwise behavior. In summary, these numerals indicate the numerical position of the four model score series in relation to the Official Score reference within this dataset.

Across the corpus, all four systems explicitly structure feedback according to the IELTS criteria, routinely pairing strengths with weaknesses, such as inaccurate data reporting, limited lexical range, and gaps in organization. ChatGPT's Task 1 entries tend to focus on actionable, itemized coaching on form, and they often provide rewrites for connectors and phrasing without lengthy commentary. Gemini typically presents a concise overview with a Band 6 rationale, striking a balance between strengths, such as the effective use of specific figures, and weaknesses, including grammatical errors and imprecise language. Claude often opens with a direct statement of intent to analyze the response, enumerating Task Achievement, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy, while noting minor imprecisions in wording. DeepSeek's Task 1 feedback often clusters around Band 5, with causes such as chart misinterpretation and basic cohesion. However, the set also contains very low ratings for task mismatch and higher-band summaries when a clear overview is present. For illustration, one DeepSeek pass assigns a very low rating to an off-topic floor plan description, while other passes present mid-band summaries of coherent trend reporting. In Task 2, ChatGPT again frames feedback under the four criteria and assigns a single band, mixing supportive notes about overall progression with cautions about repetition and awkward constructions. Gemini's Task 2 comments typically begin with a band statement and a criterion-by-criterion breakdown, praising clear structure while observing that transitions can be smoother and vocabulary remains adequate rather than sophisticated. Claude's Task 2 feedback often begins with systematic framing, listing brief observations about development and cohesion while also recording occasional agreement errors. DeepSeek structures Task 2 feedback into sections such as key issues and reasoning, and notes concrete shortcomings, including the failure to state a preferred medium, alongside comments about basic paragraph organization and simple linking. Across models, the feedback language remains anchored to Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy, and it repeatedly pairs encouragement about paragraphing with counterpoints regarding transitions. Examples are sometimes embedded as paraphrased fragments rather than verbatim wording, and these appear next to brief rationales that point to

specific sentences or clauses. Across tasks, the four systems employ a consistent descriptive framework, alternating between concise overviews and itemized notes that reference both positive features and limiting factors.

4 DISCUSSION

Across 110 scripts, the four generative models exhibit varying degrees of agreement with the Official Score, as defined by the model minus the official score. Table 1 reports concordance correlation coefficients that anchor overall agreement, with DeepSeek at 0.747, Claude at 0.722, Gemini at 0.622, and ChatGPT at 0.514. Interpreting the CCC as a joint index of accuracy and precision places DeepSeek and Claude closer to concordance, whereas ChatGPT lags. Bias profiles are uniformly negative relative to the Official Score, with averages of minus 0.168 for DeepSeek, minus 0.373 for Claude, minus 0.046 for Gemini, and minus 0.923 for ChatGPT. The analysis shows ChatGPT to be distinctly conservative, with a median of minus one and 90 underestimates against only three overestimates. By contrast, the dataset indicates that Gemini and DeepSeek center on a median of zero and exhibit minor mean shifts, suggesting a near-unbiased central tendency. Precision varies, as the standard deviation of differences spans 0.692 to 0.825, with DeepSeek the tightest and Claude the widest. Agreement within one band is frequent for DeepSeek and Gemini, at 94.5% and 93.6%, respectively. It is lower for Claude at 84.5% and notably weaker for ChatGPT at 72.7%. At the stricter half-band tolerance, performance diverges sharply, from roughly 41% for ChatGPT to about 74% for DeepSeek, with Claude at 64.5% and Gemini at 62.7%. Extremes matter, and the share beyond one band spans about 27.3% for ChatGPT, versus roughly 5% for DeepSeek and about 6 to 7% for Gemini. Bland–Altman outliers remain limited, with outside limits proportions of 0.054 for ChatGPT, 0.064 for Claude, 0.046 for DeepSeek, and 0.054 for Gemini, although tail behavior still differentiates risk. Concrete instances illustrate the pattern, such as Task 2 Sample Text 35, where the official is 6.0 and ChatGPT assigns 3.0, Task 2 Sample Text 10, where DeepSeek gives 5.5 against an Official 4.0, and Task 2 Sample Text 6, where Gemini yields 6.0 for an Official 3.5. Directionality also varies, as overestimation is rare for ChatGPT in three cases, yet more common for Gemini (35) and DeepSeek (27), while Claude posts 23. Conversely, underestimation dominates for ChatGPT (90%), but it is moderated for the other systems. Krippendorff's alpha remains constant at 0.636 across entries, stabilizing the overall reliability signal while masking cross-model differences in bias and dispersion. Viewed together, the evidence suggests the strongest practical agreement for DeepSeek and Gemini, moderate alignment for Claude, and comparatively weak concordance for ChatGPT, considering both central tendency and dispersion.

In terms of quality, the feedback from the four models aligns structurally with human markers by organizing feedback under the categories of Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy, which signals a baseline concordance with the band descriptors. I interpret the recurrent pairing of strengths and weaknesses as echoing human marks practice, particularly where accuracy of data, lexical range, and organization are treated as balancing forces. In Task 1, ChatGPT tends to provide itemized coaching maps to address cohesion and clarity concerns in the descriptors, yet the coaching flavor extends beyond typical exam commentary. Gemini's concise mid-band framing resembles the descriptor language of adequate control, although its steady center of gravity may compress the tails that human markers sometimes award. The systematic

enumeration used by Claude mirrors the rater grid and, in this reading, aligns well with boundary judgments across criteria, while its micro-level notations might overemphasize precision. Evidence from DeepSeek suggests alignment at Band 5, where basic cohesion and misinterpretation are foregrounded; however, the sharp penalties for task mismatch point to a severity pattern that not every examiner would express so strongly. Interpreted across Task 2, all models foreground progression, structure, and paragraphing in a way that supports human expectations of stance and logical development, although the weighting of lexical sophistication versus clarity varies. For example, a script with a clear structure but repetitive phrasing would be read by the models as solidly organized yet lexically constrained, which is broadly consistent with mid-band human judgments. On balance, the feedback language aligns with the band descriptors and highlights the same levers that influence scores, while distributional tendencies and register choices introduce subtle yet meaningful deviations from calibrated human standards.

Ethical considerations surrounding GenAI in language education center on learner dependency, equity of access, algorithmic bias, privacy, and transparency, with the cited sources identifying these as significant risks. As Abu-Rayyash [24] cautions, overreliance on automated support can blunt critical thinking and creativity, and this interpretation implies that classroom design should scaffold tool use rather than outsource cognition. Access is a fairness issue as well as a technical one, since Mohamed [25] highlights differential availability across contexts. While advocates claim that freely available tools broaden participation, detractors counter that the costs of computing and connectivity entrench existing divides. Human oversight appears as a stabilizing mechanism that preserves independent skill formation and provides context-sensitive judgment, a position reinforced by Zainuddin, et al. [26] and echoed in the passage's emphasis on balancing advantages with learner development.

Regarding bias, Mohamed [25] argues that models can reproduce social prejudices unless they are monitored and recalibrated, and a plausible failure mode is genre or dialect preferences that consistently lower scores for specific groups. Privacy risks extend from training corpora to operational data trails, and Akhtar [27] signals that poorly governed collection and retention can expose personal information, for example, when drafts or metadata are retained beyond necessity. Transparency and accountability emerge as procedural counterweights, with Mohamed [25] and related guidance emphasizing auditability, explainable rationales, and accessible appeal processes. Rather than a settled equilibrium, the promise of GenAI is accompanied by continuous obligations for monitoring, adjustment, and human validation, and the practical balance between innovation and safeguards remains open to scrutiny.

5 CONCLUSION

Against the Official Score, the models show uneven agreement: DeepSeek and Gemini cluster tightly around the human bands, while Claude sits in a respectable midfield, and ChatGPT tends to undershoot. Read as concordance rather than mere association, these patterns indicate that only DeepSeek and Gemini deliver broadly actionable band estimates for unsupervised practice, with Claude serviceable and ChatGPT requiring caution. It answers RQ1 in favor of near-concordance for DeepSeek and Gemini, moderate alignment for Claude, and conservative bias for ChatGPT, consistent with an agreement-first framing that avoids the correlation trap [7], [14]. On feedback, all four organize comments by the IELTS criteria, a structural

echo of human marking that supports self-regulated cycles of planning, monitoring, and evaluation [8], [9]. Hattie and Timperley [1] emphasize specificity and next steps, and in this corpus, the models typically meet that bar, for instance, by asking a Task 1 writer to add a one-sentence overview or a Task 2 writer to state a thesis and tighten topic sentences. Still, Kim, et al. [6] and Ahmed, et al. [10] caution that generic phrasing, hallucination, and privacy risks can erode trust. Our reading notes occasionally reveal superficial praise or over-penalization for task mismatch, which an experienced examiner might temper. Practically, the ethical balance is to treat free GenAI as formative scaffolding rather than certification, to triangulate outputs across tools, and to verify against rubrics, a protocol advocated by Lin [12] and compatible with equitable, offline-first habits in low-bandwidth contexts [11]. Scholars argue that immediacy raises motivation and helps knowledge stick, while detractors warn that overuse may distort judgment and entrench inequity, a live tension in self-directed learning [2], [6], [15]. Taken together, the evidence supports a cautious but constructive adoption strategy in which DeepSeek and Gemini provide the most dependable indicative bands, all four supply rubric-shaped feedback that can accelerate revision, and governance measures deal squarely with bias, privacy, transparency, and uneven access so that rural learners benefit without surrendering judgment [4], [15].

These findings imply a cautious adoption pathway in which DeepSeek and Gemini serve as first-line sources of indicative bands for formative practice, with Claude as a secondary option and ChatGPT used chiefly for criterion-linked feedback, while its conservative bias is explicitly anticipated; for example, teachers can require rubric cross-checks and a second model pass before learners' revisions. To protect equity and privacy in low-bandwidth contexts, programs should lean on offline-first routines, local drafting with batched queries, minimal data retention, and transparent consent practices. For classroom reliability, simple local calibration should be trialed, such as fitting a minor intercept-only correction on a held-out set of twenty scripts to align each model to the center, alongside routine bias audits that inspect errors by task type, topic, and band. Future studies should conduct multi-site randomized trials that compare learning gains from AI-supported drafting with business-as-usual across Task 1 and Task 2, track long-term retention and self-efficacy, and include multilingual cohorts that highlight dialect sensitivities. Methodologically, researchers should extend agreement analysis to the criterion level, report band-conditioned limits of agreement with proportional bias tests, quantify tail risk, and replicate across updates, paid tiers, and prompt variants while conducting fairness audits for demographic and topical subgroups. Systems research should investigate teacher-in-the-loop workflows, privacy-preserving logging, lightweight confidence flags, and rationale quality checks, so that feedback remains rapid and specific without encouraging overconfidence or eroding human judgment.

6 REFERENCES

[1] J. Hattie and H. Timperley, "The power of feedback," *Review of Educational Research*, vol. 77, no. 1, pp. 81–112, 2007. https://doi.org/10.3102/003465430298487

[2] V. J. Shute, "Focus on formative feedback," *Review of Educational Research*, vol. 78, no. 1, pp. 153–189, 2008. https://doi.org/10.3102/0034654307313795

[3] M. Younas, D. Abdel Salam El-Dakhs, and Y. Jiang, "A comprehensive systematic review of AI-driven approaches to self-directed learning," *IEEE Access*, vol. 13, pp. 38387–38403, 2025. https://doi.org/10.1109/ACCESS.2025.3546319

[4] F. Jin, C.-H. Lin, and C. Lai, "Modeling AI-assisted writing: How self-regulated learning influences writing outcomes," *Computers in Human Behavior*, vol. 165, p. 108538, 2025. https://doi.org/10.1016/j.chb.2024.108538

[5] J. Huang and A. Mizumoto, "The effects of generative AI usage in EFL classrooms on the L2 motivational self system," *Education and Information Technologies*, vol. 30, no. 5, pp. 6435–6454, 2025. https://doi.org/10.1007/s10639-024-13071-6

[6] J. Kim, S. Yu, R. Detrick, and N. Li, "Exploring students' perspectives on generative AI-assisted academic writing," *Education and Information Technologies*, vol. 30, no. 1, pp. 1265–1300, 2025. https://doi.org/10.1007/s10639-024-12878-7

[7] C. Song and Y. Song, "Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students," *Frontiers in Psychology*, vol. 14, p. 1260843, 2023. https://doi.org/10.3389/fpsyg.2023.1260843

[8] B. J. Zimmerman, "Becoming a self-regulated learner: An overview," *Theory into Practice*, vol. 41, no. 2, pp. 64–70, 2002. https://doi.org/10.1207/s15430421tip4102_2

[9] J. Read, "Test review: The international english language testing system (IELTS)," *Language Testing*, vol. 39, no. 4, pp. 679–694, 2022. https://doi.org/10.1177/02655322221086211

[10] Z. Ahmed *et al.*, "The generative AI landscape in education: Mapping the terrain of opportunities, challenges, and student perception," *IEEE Access*, vol. 12, pp. 147023–147050, 2024. https://doi.org/10.1109/ACCESS.2024.3461874

[11] C. Rose *et al.*, "A conference (Missingness in Action) to address missingness in data and AI in health care: Qualitative thematic analysis," *Journal of Medical Internet Research*, vol. 25, p. e49314, 2023. https://doi.org/10.2196/49314

[12] Z. Lin, "Techniques for supercharging academic writing with generative AI," *Nature Biomedical Engineering*, vol. 9, no. 4, pp. 426–431, 2025. https://doi.org/10.1038/s41551-024-01185-8

[13] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016. https://doi.org/10.1016/j.jcm.2016.02.012

[14] I. K. L. Lawrence, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989. https://doi.org/10.2307/2532051

[15] L. Yan, S. Greiff, Z. Teuber, and D. Gašević, "Promises and challenges of generative artificial intelligence for human learning," *Nature Human Behaviour*, vol. 8, no. 10, pp. 1839–1850, 2024. https://doi.org/10.1038/s41562-024-02004-5

[16] M. Al-Kfairy, D. Mustafa, N. Kshetri, M. Insiew, and O. Alfandi, "Ethical challenges and solutions of generative AI: An interdisciplinary perspective," *Informatics*, vol. 11, no. 3, p. 58, 2024. https://doi.org/10.3390/informatics11030058

[17] S. Herbold, A. Hautli-Janisz, U. Heuer, Z. Kikteva, and A. Trautsch, "A large-scale comparison of human-written versus ChatGPT-generated essays," *Scientific Reports*, vol. 13, no. 1, p. 18617, 2023. https://doi.org/10.1038/s41598-023-45644-9

[18] I. K. L. Lawrence, "Corrections (A concordance correlation coefficient to evaluate reproducibility)," *Biometrics*, vol. 56, no. 1, pp. 324–325, 2000. https://doi.org/10.1111/j.0006-341X.2000.00324.x

[19] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY: Chapman and Hall/CRC, 1994. https://doi.org/10.1201/9780429246593

[20] J. M. Bland and D. G. Altman, "Agreement between methods of measurement with multiple observations per individual," *Journal of Biopharmaceutical Statistics*, vol. 17, no. 4, pp. 571–582, 2007. https://doi.org/10.1080/10543400701329422

[21] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 4th ed. Thousand Oaks, CA: SAGE Publications, Inc., 2019. [Online]. Available: https://doi.org/10.4135/9781071878781. [Accessed: Sep. 29, 2025].

[22] L. Li, Z. Ma, L. Fan, S. Lee, H. Yu, and L. Hemphill, "ChatGPT in education: A discourse analysis of worries and concerns on social media," *Education and Information Technologies*, vol. 29, no. 9, pp. 10729–10762, 2024. https://doi.org/10.1007/s10639-023-12256-9

[23] A. D. Samala *et al.*, "Unveiling the landscape of generative artificial intelligence in education: A comprehensive taxonomy of applications, challenges, and future prospects," *Education and Information Technologies*, vol. 30, no. 3, pp. 3239–3278, 2025. https://doi.org/10.1007/s10639-024-12936-0

[24] H. Abu-Rayyash, "Revolutionizing translator training through human-AI collaboration: Insights and implications from integrating GPT-4," *Current Trends in Translation Teaching & Learning E*, vol. 10, pp. 259–301, 2023. https://doi.org/10.51287/cttl20239

[25] M. S. P. Mohamed, "Exploring ethical dimensions of AI-enhanced language education: A literature perspective," *Technology in Language Teaching & Learning*, vol. 6, no. 3, p. 1813, 2024. https://doi.org/10.29140/tltl.v6n3.1813

[26] N. Zainuddin *et al.*, "Responsible and ethical use of artificial intelligence in language education: A systematic review," *Forum for Linguistic Studies*, vol. 6, no. 5, pp. 316–325, 2024. https://doi.org/10.30564/fls.v6i5.7092

[27] Z. B. Akhtar, "Generative artificial intelligence (GAI): From large language models (LLMs) to multimodal applications towards fine tuning of models, implications, investigations," *Computing and Artificial Intelligence*, vol. 3, no. 1, p. 1498, 2024. https://doi.org/10.59400/cai.v3i1.1498

7 AUTHOR

Rohib Adrianto Sangia is a final year of PhD Student at School of LLMVC, University of Aberdeen, Scotland, United Kingdom (E-mail: r.sangia.22@abdn.ac.uk).