



From Prompts to Scores: Generative AI vs. Human Marks in Writing for Standardized Language Testing

Rohib Adrianto Sangia

University of Aberdeen

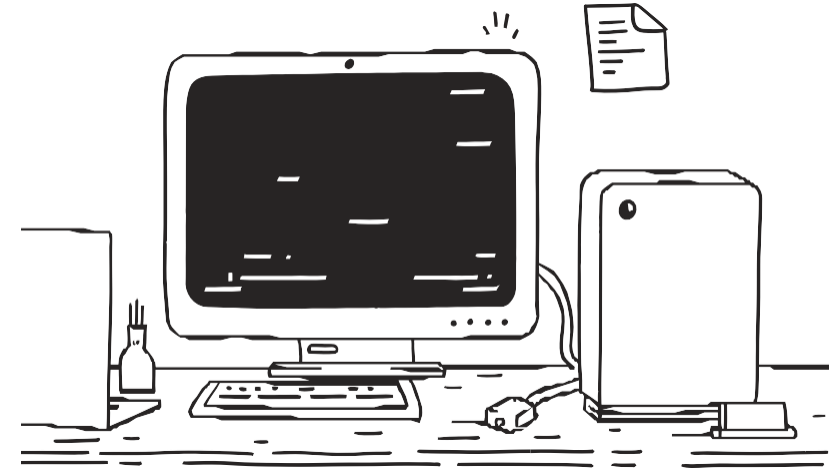
The Learning Ideas Conference 2025

Faculty House at Columbia University.

Introduction and Background

Many student in rural developing countries, dreams of studying abroad, but stuck in English

- No Access. No tutors. No feedback.
- All they have: Grit, Legacy PC and free AI tools.
- Can AI give him feedback as reliable as a human examiner?
- This question drives the heart of our inquiry.



Research Objectives



Primary Objective

Evaluate the statistical alignment of GenAI scores with human IELTS writing scores.



Secondary Objective

Assess the depth, structure, and usability of GenAI feedback for self-directed writing practice.

Methodology

1

Dataset

110 authentic IELTS Task 1 and Task 2 essays from official IELTS preparation materials.

2

GenAI Models Tested

ChatGPT-4o, Claude 3.5 Sonnet, Gemini 2.0, DeepSeek (June 2024 release). They are free versions and untrained.

3

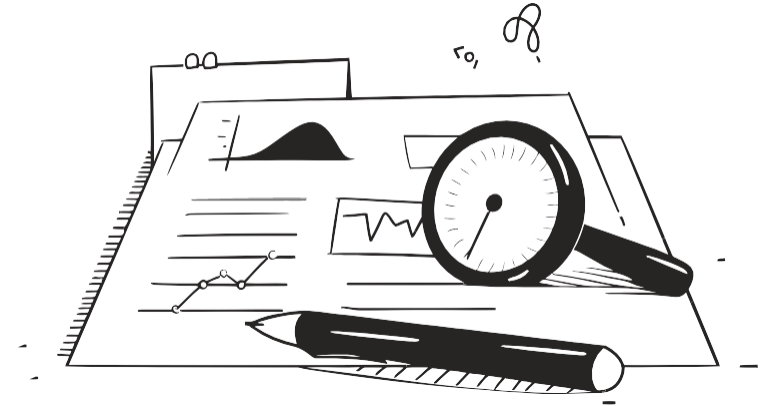
Evaluation Metrics

Pearson Correlation Coefficient (PCC), Intraclass Correlation Coefficient (ICC), Mean Absolute Error (MAE).

4

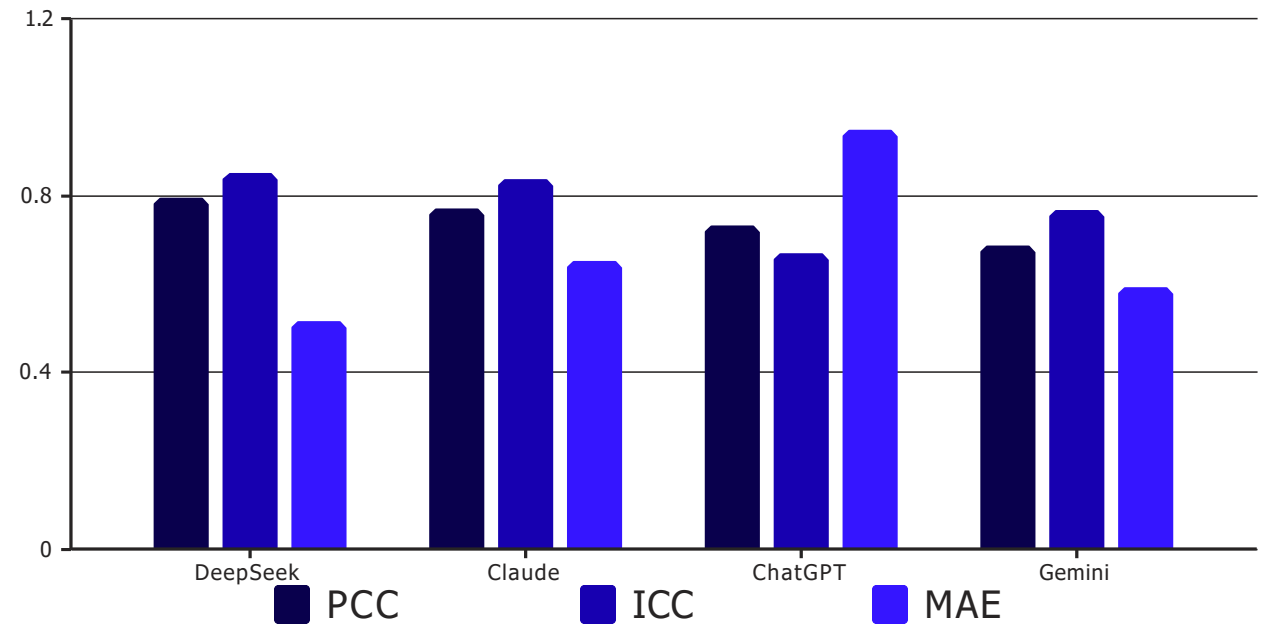
Feedback Analysis

Qualitative evaluation of feedback structure, depth, specificity, and motivational tone.



Results

Statistical Alignment:



DeepSeek showed the highest correlation with human scores, followed by Claude, ChatGPT, and Gemini.

Results

Feedback Analysis:

ChatGPT Claude

Provided comprehensive, structured, and actionable feedback.

DeepSeek

Offered concise feedback but lacked depth.

Gemini

Focused on error patterns but lacked text references.



Ethical Considerations

Critical Thinking

Potential reduction in critical thinking and creativity due to over-reliance on GenAI.



Algorithmic Bias

Data privacy concerns and accessibility issues.

Human Oversight

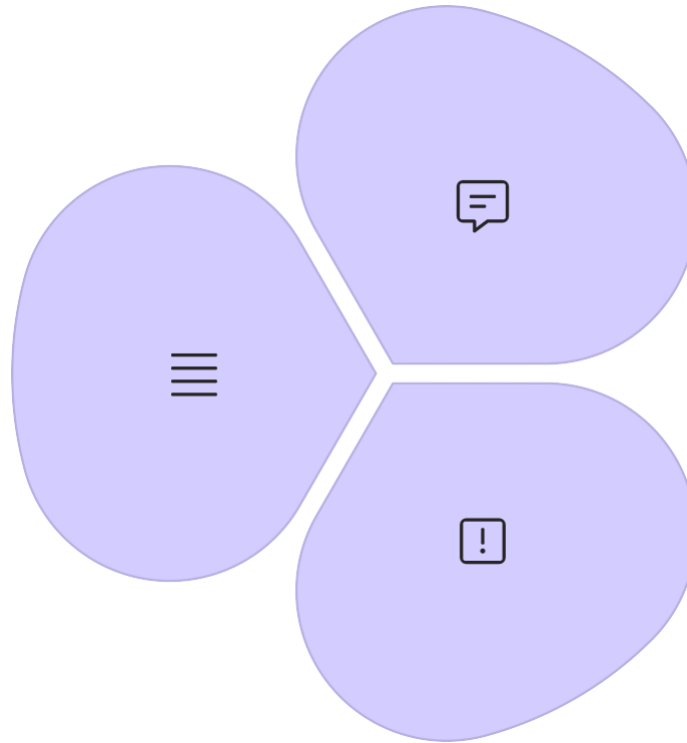
Need for human oversight to ensure fairness and equity.



Conclusion

Strong Alignment

GenAI models, particularly DeepSeek and Claude, demonstrate strong alignment with human evaluators.



Valuable Feedback

ChatGPT, along with Claude, provides valuable feedback for IELTS preparation in terms of self-directed learning.

Ethical Importance

Ethical considerations highlight the importance of human oversight in AI-driven assessments.

Acknowledgments

We would like to express our sincere gratitude to:



Indonesian Endowment Fund for Education (LPDP)

IELTS Test Format

(IELTS - Academic)

Listening

- 4 Recording
- **Questions:** 40 (worth 40 Marks)
- **Time:** 30 Minutes

✿ An extra 10 minutes to transfer your answers from your question booklet to the answer sheet.

Reading

- 3 Long Texts
- **Questions:** 40 (worth 40 Marks)
- **Time:** 60 Minutes

Note: There's **NO Extra Transfer Time** provided.

Writing

- 2 Writing Tasks:
 - **Task 1:** Describe (or summarize) a graph, table, chart or diagram (**20 minutes**)
 - **Task 2:** Write an Essay (**40 minutes**)

Speaking

You will talk to a certified examiner.

Time: 11-14 minutes

Part 1: General questions about home, family, work, studies and interests.

Part 2: Talk about a specific topic given in a task card (1-2 minutes). Then, one or two more questions on the topic.

Part 3: Further questions from the same topic.

Ref: <https://takeielts.britishcouncil.org/take-ielts/test-format> (as of: 05/Jan/2024)

visit: www.notesflux.com



Statistical Tools



1. The Pearson Correlation Coefficient (PCC) measures the linear relationship between two sets of scores. It indicates whether they move in the same direction; for example, if one score is high, the other tends to be high as well.
2. The Intraclass Correlation Coefficient (ICC) is a more rigorous measure. It assesses the consistency or agreement between scorers, considering whether the actual values are similar, not just following a similar pattern.
3. Finally, the Mean Absolute Error (MAE) calculates the average size of the difference between the paired scores. A smaller MAE signifies that the two sets of scores are, on average, closer together.

Results



Statistical Alignment:

	Task 1	Task 2	Average	Rank
Pearson Correlation Coefficient (PCC)				
Official vs. ChatGPT Score	0.707	0.757	0.732	3rd
Official vs. Claude AI Score	0.773	0.770	0.771	2nd
Official vs. DeepSeek Score	0.748	0.843	0.795	1st
Official vs. Gemini Score	0.690	0.683	0.686	4th
Intraclass Correlation Coefficient (ICC)				
Official vs. ChatGPT Score	0.639	0.696	0.668	4th
Official vs. Claude AI Score	0.820	0.854	0.837	2nd
Official vs. DeepSeek Score	0.806	0.897	0.852	1st
Official vs. Gemini Score	0.767	0.764	0.766	3rd
Mean Absolute Error (MAE)				
Official vs. ChatGPT Score	0.882	1.018	0.950	4th
Official vs. Claude AI Score	0.611	0.694	0.653	3rd
Official vs. DeepSeek Score	0.509	0.519	0.514	1st
Official vs. Gemini Score	0.500	0.683	0.591	2nd

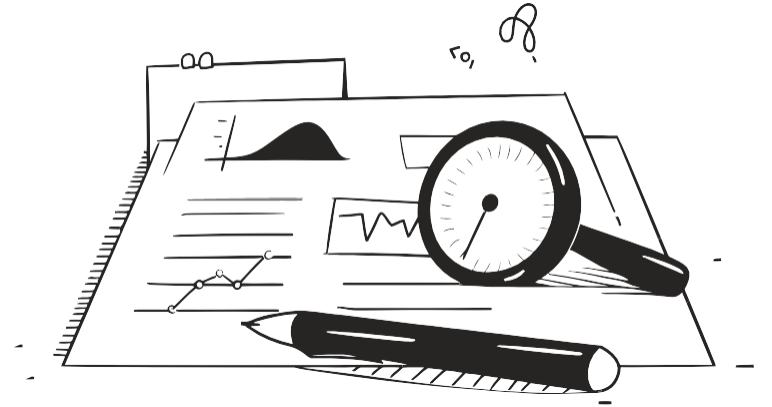
DeepSeek showed the highest correlation with human scores, followed by Claude, ChatGPT, and Gemini.

The given response for the IELTS Writing Task 1 describing the hydroelectric power generation process has several issues concerning coherence, grammar, and accuracy. Here is an evaluation based on the IELTS Writing Task 1 band descriptors:

1. Task Achievement: Band 5.0 • The response addresses the diagram and describes the key stages but lacks precision and completeness. • Some essential details are either missing or unclear, such as the role of the pump and how electricity flows through the system. • The explanation of the process is somewhat basic, and the description lacks a logical sequence in some parts. • The overview is present but could be clearer. The conclusion summarises the steps but does not sufficiently highlight the cyclical nature of the process.
2. Coherence and Cohesion: Band 5.0 • The structure of the response is somewhat logical but could be better organised. • There are some linking words (e.g., "first of all," "afterwards," "finally"), but their use is mechanical and lacks variety. • Some sentences are awkwardly structured, making it difficult to follow the flow of ideas. • The connection between the different stages of the process could be more clearly stated.
3. Lexical Resource: Band 5.0 • The vocabulary is limited and repetitive, with basic words such as "water," "electricity," and "process" being overused. • Some word choices are incorrect, such as "sun energy" instead of "solar energy" and "tourbin" instead of "turbine." • A wider range of precise technical terms (e.g., "hydroelectric dam," "power grid") would enhance clarity.
4. Grammatical Range and Accuracy: Band 4.5 - 5.0 • There are several grammatical errors, such as "as natural resources, like water, as well as special human-built equipment" (incorrect phrase structure). • Incorrect verb forms appear throughout, for example, "Evaporation forces clouds to form" (should be "causes clouds to form"). • Incorrect prepositions and articles are used, such as "the rain goes the water is collected" (should be "when it rains, the water is collected"). • Spelling mistakes ("stransformer" instead of "transformer," "tourbin" instead of "turbine") further reduce accuracy.

Overall IELTS Writing Band Score: 5.0 The response demonstrates an understanding of the process but lacks clarity, cohesion, and grammatical accuracy. To improve the score: • Use a wider range of vocabulary and more precise technical terms. • Improve grammatical accuracy, particularly with verb forms, prepositions, and article usage. • Organise ideas more logically, ensuring each stage of the process is clearly explained. • Avoid spelling errors and repetition by varying sentence structures.

ChatGPT Feedback



Claude Feedback

IELTS Task 1 Writing Response Analysis

Let me analyze this IELTS Task 1 writing response according to the four IELTS scoring criteria:

1. Task Achievement:

- Covers all main stages of the process
- Includes overview/summary
- Clear understanding of the process shown
- Meets word count requirement

2. Coherence and Cohesion:

- Clear progression using sequence markers ("First of all," "Afterwards," "Finally")
- Logical organization of information
- Good use of paragraphing
- Includes overview at the end

3. Lexical Resource:

- Good range of vocabulary related to the topic
- Some accurate technical terms ("reservoir," "transformer station," "high voltage cables")
- Few spelling errors ("tourbin" instead of "turbine", "stransformer")
- Appropriate word choices overall

4. Grammatical Range and Accuracy:

- Mix of simple and complex sentences
- Some minor errors ("how to use water in order to" should be "how water is used to")
- Generally good control of grammar
- A few awkward constructions ("involves as natural resources")

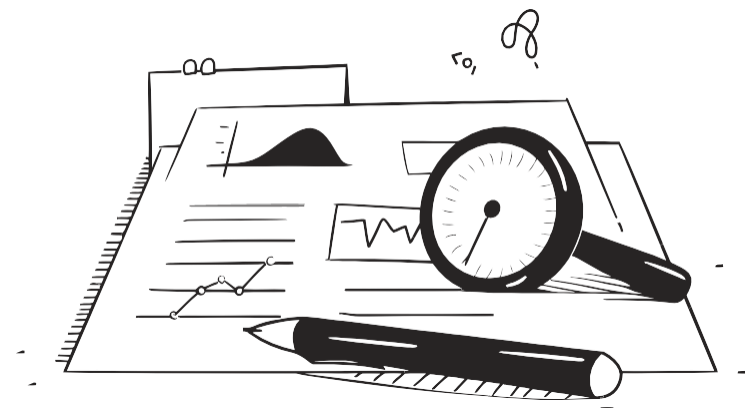
Overall Assessment

Based on these factors, this response would receive a **Band 6.5**.

The score of 6.5 reflects that the writer:

- Has presented and organized the information clearly
- Uses appropriate sequencing and linking words
- Shows good vocabulary control despite a few errors
- Demonstrates generally good grammatical control
- Provides a clear overview of the process
- Makes some errors but they don't impede communication

While the response shows good understanding and organization, the occasional language errors and awkward expressions prevent it from reaching a higher band score of 7.0 or above.



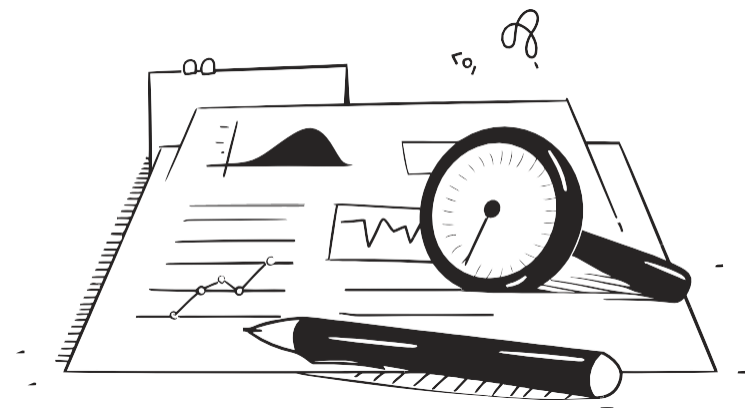
Appendix 6

Deepseek Feedback

The provided answer would likely receive a **Band 6** in the IELTS Writing Task 1. Here are the reasons:

1. **Task Achievement:** The response adequately summarizes the main features of the process and includes relevant details. However, there are some inaccuracies and omissions, such as not clearly explaining the role of the pump and the underground cables. The overview is present but could be more detailed.
2. **Coherence and Cohesion:** The response is generally coherent and logically organized. There is a clear progression of ideas, although some transitions could be smoother. The use of linking words is adequate but not always precise.
3. **Lexical Resource:** The vocabulary is sufficient to describe the process, but there is limited variety and some repetition. Some terms are used inaccurately (e.g., "sun energy" instead of "solar energy").
4. **Grammatical Range and Accuracy:** There are several grammatical errors, including incorrect word choices and awkward phrasing (e.g., "the rain goes the water is collected"). These errors affect the clarity and fluency of the response.

Overall, the response demonstrates a reasonable understanding of the task and provides a clear description of the process, but the inaccuracies and grammatical issues limit it to a Band 6.



The most accurate IELTS Writing score for the response provided is Band 6. Here's a breakdown based on the IELTS Writing Task 2 assessment criteria:

Task Response

The response addresses the prompt and presents a relevant opinion. It discusses reasons for insufficient respect towards older people and touches on the problems this might cause. However, the development of ideas is uneven. While some examples are provided (e.g., people having many roles, lack of pensions), they are not always fully explained or linked back to the core issue of respect. The discussion of solutions is also somewhat limited.

Coherence and Cohesion

The response is organized with a clear structure (introduction, body paragraphs, conclusion). However, the flow of ideas within paragraphs can be a bit disjointed. While some cohesive devices are used (e.g., "for example," "on the other hand"), they are not always effectively used to link ideas, leading to some choppiness.

Lexical Resource

The writer demonstrates an adequate range of vocabulary, but there are some noticeable inaccuracies (e.g., "cousins don't take care of them," "don't paying"). While the vocabulary is generally relevant, the errors and occasional awkward phrasing limit the overall impression.

Grammatical Range and Accuracy

The response contains a mix of simple and complex sentence structures, but there are some recurring grammatical errors that affect clarity. These include errors in articles, prepositions, verb tenses, and sentence structure. While the errors do not completely obscure meaning, they do hinder communication to some extent.

