Rating Short L2 Essays on the CEFR Scale with GPT-4

Kevin P. Yancey and **Geoffrey T. LaFlair** and **Anthony R. Verardi** and **Jill Burstein**
Duolingo
{kyancey,geoff,anthony.verardi,jill}@duolingo.com

Abstract

Essay scoring is a critical task used to evaluate second-language (L2) writing proficiency on high-stakes language assessments. While automated scoring approaches are mature and have been around for decades, human scoring is still considered the gold standard, despite its high costs and well-known issues such as human rater fatigue and bias. The recent introduction of large language models (LLMs) brings new opportunities for automated scoring. In this paper, we evaluate how well GPT-3.5 and GPT-4 can rate short essay responses written by L2 English learners on a high-stakes language assessment, computing inter-rater agreement with human ratings. Results show that when calibration examples are provided, GPT-4 can perform almost as well as modern Automatic Writing Evaluation (AWE) methods, but agreement with human ratings can vary depending on the test-taker's first language (L1).

1 Introduction

Automated writing evaluation (AWE) systems are commonly used to evaluate test-taker writing. AWE systems are deployed on large-scale, high-stakes writing assessments used for admissions to higher education institutions, and for lower-stakes US state writing assessments that provide information about K-12 students' academic writing performance. These systems typically use feature-engineering approaches that include rule-based and statistical natural language processing (NLP) methods. NLP is used to extract features from essay writing responses that are characteristic of writing quality. Features may include errors in grammar and spelling, discourse structure, discourse coherence, vocabulary usage, and sentence variety. Features may be rule-based or statistically derived. Statistical model methods, such as straightforward linear regression, are used to train (build) AWE scoring models for high-stakes scoring of writing assessments. Detailed descriptions of systems are available for major systems, including e-rater®, Intelligent Essay Assessor™, Intellimetric®, and PEG (Shermis and Burstein, 2013), and Cambium's automated essay scoring system (Lottridge, in press).

Recent advances in language modeling with neural transformer architectures (OpenAI, 2023; Brown et al., 2020) have the potential to revolutionize AWE. These large language models (LLMs) demonstrate an incredible potential to analyze and evaluate text which has implications for the future of AWE. In addition, GPT's intuitive, text-based interface lowers barriers for use, potentially increasing accessibility and adoption of these tools for AWE. The assumptions about how LLMs – specifically GPT-4 – can be used for AWE tasks, such as automated scoring and feedback need to be evaluated to determine how we can use them beneficially, and particularly to ensure that they can be used in a fair and ethical manner (Burstein, 2023).

Previous research evaluated GPT-3.5 for essay scoring tasks in an L2 context (Mizumoto and Eguchi, 2023). In this paper, we evaluate GPT-4 for a similar task, comparing it to GPT-3.5, human judgement, and a strong baseline using current AWE methods. We also explore various aspects that affect the accuracy of GPT's ratings, and its fairness across gender and L1.

2 Data

For our experiments, we used a human-rated dataset consisting of short essay responses collected as part of the Duolingo English Test, a high-stakes test of English for L2 learners. For this essay task, test-takers are given a short written prompt randomly selected from an item bank of about 700 items. Test-takers have 5 minutes to provide their essay response to the prompt. Two human raters used a scoring rubric aligned with the Common European Framework of Reference (CEFR) (Council of Europe, 2001).

We started by sampling 10,000 responses from

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 576–584

test sessions that took place over a 10-month period, controlling for L1 and gender. For L1, we limited responses to 7 of the most common L1 languages for the test, which also captures a broad range of language families: Arabic (ara), Mandarin Chinese (cmn), Telugu (tel), English (eng)[1], Spanish (spa), Gujarati (guj), and Bengali (ben). To ensure all CEFR levels were well represented in the final dataset[2], we used a simple CEFR classifier that uses logistic regression and NLP features to roughly estimate the CEFR level of each response. For the final dataset, we randomly sampled an equal number of responses for each combination of L1, gender, and estimated CEFR level from the 10,000 test sessions.

The scoring rubric was aligned to the CEFR scale and assessed each response based on its content, coherence, vocabulary, and grammar. The rubric instructed raters to assign each essay one of eight rating categories: six based on the CEFR scale, and two "unscorable" categories for minimal responses (e.g., provides no response or says they can't answer the question) and bad-faith responses (e.g., off-topic or nonsensical). The full rubric is provided in Appendix B.

Based on this rubric, two assessment researchers developed a set of calibration examples by collectively rating 676 essays, 180 of which were rated by both. The rubric and calibration examples were provided to two new human raters, who collectively rated 1,961 new essays, including a random sub-sample of 389 essays that were rated by both. Both new human raters were trained by one of the original assessment researchers and inter-rater agreement was routinely checked. Raters were provided feedback to help with calibration when necessary. The final Quadratic Weighted Kappa (QWK) between the two raters was 0.87. Ratings were roughly normally distributed (see Figure 1), with ∼53 % of essays receiving a rating of B1 or B2 and only ∼12 % getting a rating of A1 or C2.

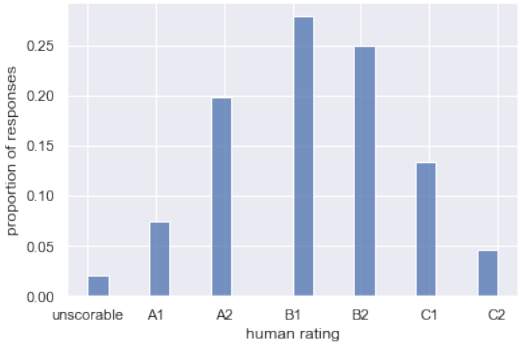

Figure 1: Distribution of Human Ratings by Raters 1 and 2

2.1 Methodology

In our experiments, we used the ChatGPT API to rate these short essay responses, comparing them to human judgements using the same rubrics.

In the system message, we instructed GPT to rate each provided essay in one of eight rating categories: one of the six CEFR levels or one of the two unscorable categories, `[No-Response]` and `[Nonsense/Off-Topic]`. In the default setting, we provided specific criteria the two unscorable categories, but not for CEFR levels[3]. See Appendix C for details.

In addition to the system message, we also provided GPT with varying numbers of calibration examples. These examples were randomly sampled from the set of 180 essays that were double-rated by assessment researchers where both researchers agreed on the same rating. The same number of examples were provided for each of the eight rating categories. We tested providing up to the maximum number of calibration examples that would fit into each model's token limit (generally two per category for GPT-3.5 and four per category for GPT-4)[4]. To avoid any possible interaction between essays, we used a fresh GPT conversation to rate each essay.

[1]Test-takers who identify their L1 as English may come from countries where English is an official language, such as India.These test-takers are required to take an English language proficiency test to attend an English-medium institution abroad.

[2]In particular, the DET test-taker population's proficiencies follow a unimodal distribution around the B1/B2 CEFR levels (Cardwell et al., 2022), and so uniform random sampling would have resulted in too few A1 and C2 essay responses being included in the dataset.

[3]Querying GPT-4 easily shows that it already has some built-in knowledge of CEFR, presumambly from its massive training corpora, and can even provide CEFR descriptors for various language skills verbatim, if prompted. So, it was reasonable to evaluate GPT's ability to apply CEFR rating categories accurately without a rubric. The same is not true for the unscorable rating categories, and preliminary experiments showed that GPT applied the unscorable labels much too broadly if their criteria weren't elaborated in the instructions to GPT.

[4]Note that this token limit applies to the entire GPT conversation, not just a single turn within the conversation, and thus this puts a hard limit on the number of calibration examples that can be provided.

Once all ratings were collected, we tabulated them on a scale of 0 – 6: assigning a 0 for both unscorable categories, and a score 1 – 6 for the CEFR levels. We then computed the inter-annotator agreement between GPT and rater 1 (n=1,175), computing 90% confidence intervals using bootstrapping and comparing this to the agreement between the two human raters. We also compared our results to two baselines: a machine learning (ML) classifier using only the response's character length, and a strong baseline representative of current AWE methods that use feature engineering and statistical modeling (Attali and Burstein, 2006; Foltz et al., 1999). The strong AWE baseline, which is used to score writing responses on the Duolingo English Test, uses XGBoost (Chen and Guestrin, 2016) and is trained on hundreds of thousands of short essay responses using 85 research-based linguistic features covering a wide range of writing sub-skills, including cohesion, grammatical complexity, lexical sophistication, grammatical and lexical accuracy, length, and relevance. A more detailed breakdown of these features are provided in Appendix A.

3 Experiments

We conducted three experiments. The first evaluates both GPT-3.5 and GPT-4 with a minimal rubric and up to the maximum number of calibration examples that fit within the GPT model's token limit. The second experiment evaluates various prompt engineering strategies for improving performance. The third experiment explores GPT-4's fairness properties across gender and L1.

3.1 Experiment 1: Calibration Only

In this first experiment, we evaluated GPT's ability to rate essay responses on the CEFR scale when provided only a minimal rubric (as described in Appendix C) and varying numbers of calibration examples.

Figure 2 shows the QWK between GPT and the first human rater, depending upon the model used and the number of calibration examples provided. When no calibration examples were provided, neither GPT-3.5 nor GPT-4 even outperform the baseline classifier using character length only. However, by providing just one calibration example for each rating category, GPT-4 almost matches the performance of the AWE baseline (QWK 0.81 vs 0.84, $p < 0.1$). Providing additional examples did not result in significant improvement. GPT-3.5, on the other hand, did not improve much when provided calibration examples, and only outperformed the length-only baseline when provided two calibration examples per rating category (i.e., the maximum possible with GPT-3.5's limit of 4,096 tokens).

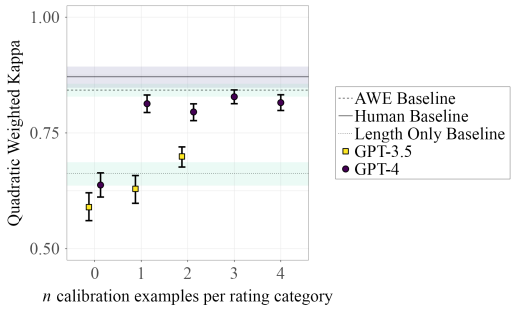

Figure 2: Human–GPT agreement when only calibration examples are provided (90% confidence intervals shown)

The confusion matrices in Figure 3 provide more insight. We see that when no examples were provided, both versions of GPT were generally able to identify unscorable responses, and did tend to assign slightly higher ratings to better essays, but mainly rated essays in the B1 – B2 range. When provided calibration examples, GPT-4 learned to use the full range of CEFR levels, but struggled to distinguish between adjacent CEFR levels compared to humans, especially for CEFR level B2. GPT-3.5, on the other hand, improves only slightly when provided calibration examples.

3.2 Experiment 2: Prompt Engineering

In our second experiment, we tested two strategies for improving the performance of GPT-4:

Detailed Rubric - In the system message, we replaced the minimal rubric used in the previous experiment with a detailed rubric that described the criteria for each CEFR level (see Appendix C).

Require Rationale - In the system message, we asked GPT to provide a rationale before providing its rating in order to elicit a chain of reasoning, which has been shown to improve the the ability of LLMs to perform complex tasks (Wei et al., 2022). This also meant providing rationales for the calibration examples, which could help GPT-4 better understand the reason for each example's rating.

Both of these techniques required significantly more token-space for the input prompt and thus lim-

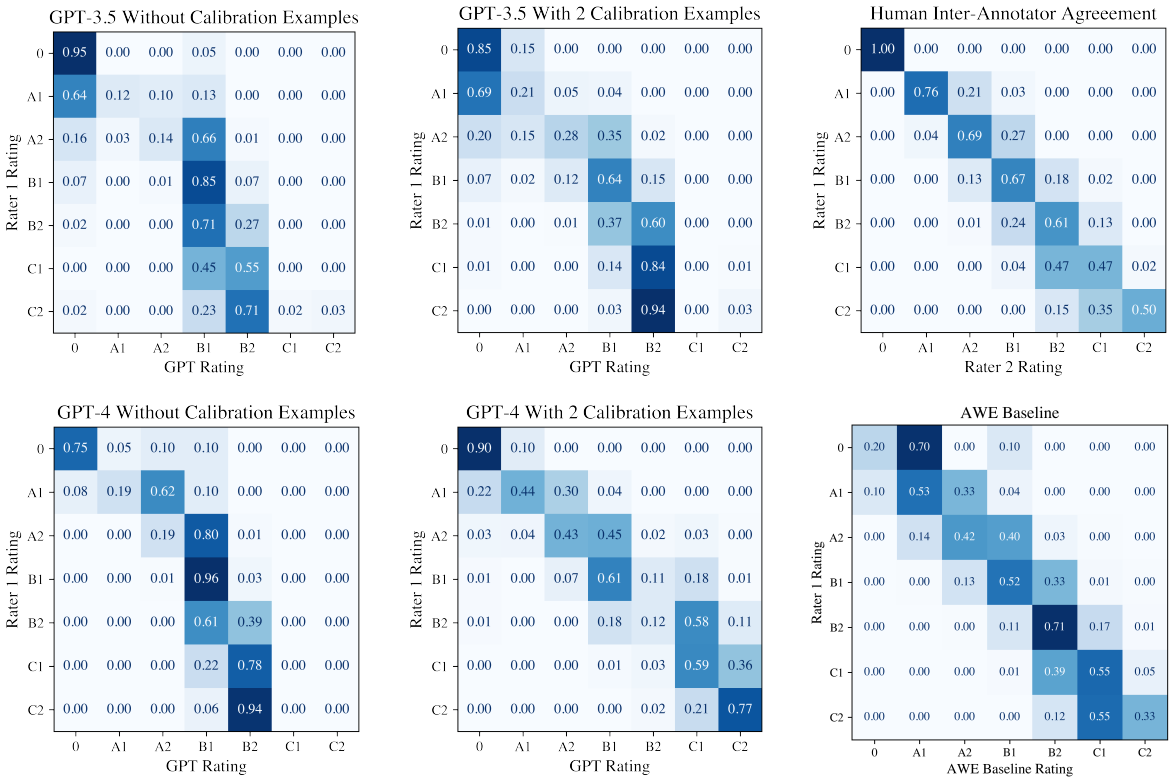

Figure 3: Confusion Matrices (Normalized by Rater 1's Rating)

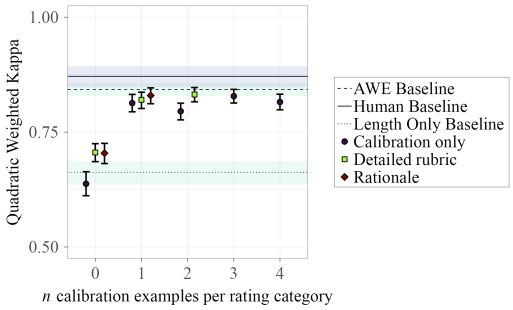

Figure 4: Human–GPT-4 agreement when various prompt engineering techniques are applied (90% confidence intervals shown)

ited the number of calibration examples that could be provided. Only up to two per rating category could be provided when using a detailed rubric, and only up to one per rating category when requiring rationales.

As seen in Figure 4, these strategies contributed substantial lift in performance when not providing calibration examples, but when at least one calibration example per rating category was provided, these techniques contributed negligible benefit.

3.3 Experiment 3: Fairness

Ensuring that raters do not show systematic bias that can affect scoring accuracy due to background characteristics of test-takers, such as gender or L1, is an important step in rater analysis with human raters (Jin and Eckes, 2022). This is also a needed step in developing AWE systems. To investigate the extent to which GPT-4's ratings are fair, we evaluated its performance for each gender and each of the L1 languages in the dataset.

To maximize statistical power and ensure that the analysis is not biased by a single human rater, we used all essays rated by any one of the raters or researchers in our dataset, except the 180 essays that were double-rated by the two researchers, which were reserved for calibration examples. The resulting dataset included 2,457 essays, roughly equally distributed among both genders and all L1s.

We found no significant differences in performance by gender, and while GPT-4's ratings were slightly positively biased compared to human ratings overall (by about +0.15 CEFR levels), this bias did not vary significantly by any gender or L1 ($p > 0.10$).

However, we did find that GPT-4 had less agreement with human ratings for essays written by L1

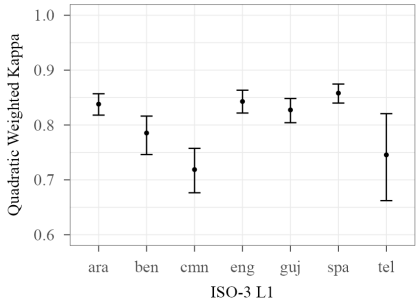

Figure 5: GPT-4 QWK by test-taker L1

speakers of some languages compared to others: QWK was lowest for L1 speakers of Telugu (tel) at 0.66 and highest for L1 speakers of Spanish (spa) at 0.89. A more detailed analysis showed that some of the differences in agreement by L1 was explained by differences in the distribution of human ratings for those L1s. The standard deviation of human ratings by L1 ranged from 1.04 for Telugu (tel) to 1.56 for Arabic (ara). Those L1s with narrower distributions of human ratings had a greater proportion of essays rated in categories for which GPT-4 had lower rates of agreement overall, such as B2, and thus brought down the QWK for those L1s.

We assume that the differences in the distribution of human ratings by L1 reflect systematic errors in the CEFR classifier used in sampling (see Section 2) and possibly differences in our underlying test-taker population. Thus we controlled for these distribution differences by recomputing QWK for each L1 using importance sampling so that all L1s would have the same effective distribution of human ratings. The results are shown in Figure 5. Even after the importance sampling correction is applied, GPT-4's ratings agreed less with human ratings for responses written by L1 speakers of Mandarin Chinese (cmn), Telugu (tel), and Bengali (ben) compared to those written by L1 speakers of Spanish (spa). It is possible that essays of some L1s are harder to distinguish and thus have less reliable human ratings, but our dataset does not consist of a sufficient number of double-rated essays to investigate this hypothesis, so we leave this for a future work.

4 Conclusion

We showed that unlike GPT-3.5, GPT-4 is able to attain performance similar to conventional Automated Writing Evaluation (AWE) models when rating short L2 essays. GPT-4 only required one calibration example per rating category to achieve near optimal performance, but other prompt engineering techniques we tried were not very helpful. Furthermore, when assessing fairness with respect to the test-taker's gender or L1, we found that while GPT-4 did not show bias in favor of any one group, it showed significantly less agreement with human ratings for some L1s. It is unclear whether this is due to the reliability of GPT-4 or that of the human ratings themselves. More research is needed to understand this discrepancy and its implications for fairness. Future research may also explore other prompt engineering strategies for improving GPT-4's performance at this task, or potentially fine-tuning GPT-3.5, enabling one to leverage dramatically more training data than what can be provided in a prompt. Perhaps most excitingly, future work may explore GPT-4's potential for providing feedback aligned to essay scoring: a task for which GPT-4 seems particularly well suited.

Acknowledgements

We thank the researchers and raters who contributed to building the dataset, and the reviewers who reviewed our paper and provided valuable feedback, particularly JR Lockwood, Ben Naismith, Klinton Blicknell, and Alina von Davier.

References

Yigal Attali. 2011. A differential word use measure for content analysis in automated essay scoring. *ETS Research Report Series*, 2011(2):i–19.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

CJ Bryant, Mariano Felice, and Edward Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. Association for Computational Linguistics.

Jill Burstein. 2023. Responsible ai standards.

Ramsey Cardwell, Geoffrey T LaFlair, and Burr Settles. 2022. Duolingo english test: Technical manual.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining, KDD '16, pages 785–794, New York, NY, USA. ACM.

Peter W Foltz, Darrell Laham, and Thomas K Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2):939–944.

Kuan-Yu Jin and Thomas Eckes. 2022. Detecting differential rater functioning in severity and centrality: The dual drf facets model. *Educational and Psychological Measurement*, 82(4):757–781.

S. Lottridge. in press. *Applications of transformer neural networks in processing examinee text*, MARCES Book Series. University of Maryland Press.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Danielle S McNamara and Arthur C Graesser. 2012. Coh-metrix: An automated tool for theoretical and applied natural language processing. In *Applied natural language processing: Identification, investigation and resolution*, pages 188–205. IGI Global.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press.

OpenAI. 2023. Gpt-4 technical report.

Marek Rei and Ronan Cummins. 2016. Sentence similarity measures for fine-grained estimation of topical relevance in learner essays. *arXiv preprint arXiv:1606.03144*.

Sarah E Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530.

Mark D Shermis and Jill Burstein. 2013. Handbook of automated essay evaluation. *NY: Routledge*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2019. Text readability assessment for second language learners. *arXiv preprint arXiv:1906.07580*.

A AWE Baseline Model Features

Here we provide a more detailed breakdown of the features used in our AWE baseline:

- 13 cohesion features, including overlap features and coreference counts (McNamara and Graesser, 2012)
- 3 grammatical complexity features, including max/mean dependency tree depth and mean sentence length (Schwarm and Ostendorf, 2005)
- 7 lexical sophistication features measuring the proportion of words at each CEFR level (including an out-of-vocabulary category for words that could not be found in the CEFR dictionary) (Xia et al., 2019)
- 51 lexical and grammatical accuracy features, measuring the error rates across a wide variety of error types (Bryant et al., 2017)
- 4 features using n-gram models over word-forms, lemmas, part-of-speech, and dependency tags to measure differential use of vocabulary and grammar across test-takers of different proficiency levels (Attali, 2011)
- 3 length features, including number of characters, words, and sentences
- 2 lexical diversity features derived from the Measure of Textual Diversity (MTLD) (McCarthy and Jarvis, 2010)
- 1 vocabulary control feature using n-gram models to measure idiomatic use of vocabulary
- 1 relevance feature, computed using IDF weighted word embeddings between the prompt and the response (Rei and Cummins, 2016)

B Scoring Rubric

Below are the criteria for each rating that were used in the rubric provided to human raters, and the system message prompts provided to ChatGPT (where applicable).

C2 The response fully achieves the task requirements: (1) the response is clear, relevant, fully developed, and is written in an appropriate

style (2) the response is smoothly-flowing, coherent, and cohesive throughout; (3) vocabulary (including collocations and idiomatic language) is accurate, appropriate, and precise; and (4) a wide range of grammatical structures are flexibly used, and there are no grammatical errors other than slips characteristic of expert speakers. Does the response have an excellent effect on the reader, such that the writer communicates their position/describes the image extremely effectively and in detail, there is no strain on the reader, and a very high level of language is used consistently throughout?

C1 The response achieves the task requirements: (1) the response is clear, relevant, appropriately developed, and is written in an appropriate style (2) the response is well-structured, coherent, and cohesive; (3) vocabulary (including collocations and idiomatic language) is accurate, appropriate, and demonstrates a broad range; and (4) a wide range of grammatical structures are used, and grammatical errors are rare. Does the response have a very good effect on the reader, such that the writer communicates their position/describes the image clearly and effectively at some length, with a high level of language used consistently throughout other than minor lapses which do not impact the communicative effect?

B2 The response mostly achieves the task requirements: (1) the response is mostly clear, relevant, developed, and written in an appropriate style (2) the response is generally well-structured, coherent, and cohesive despite occasional lapses; (3) vocabulary (including collocations and idiomatic language) is generally accurate and appropriate to the task; and (4) a range of grammatical structures are used, and grammatical errors usually do not impact communication. Does the response have a good effect on the reader, such that the writer communicates their position/describes the image fairly clearly and with some detail, with a level of language that allows them to successfully complete the task despite inaccuracies?

B1 The response partially achieves the task requirements: (1) the response is not always clear, relevant, developed, or written in an appropriate style (2) the response is somewhat organized but may lack coherence or cohesion at times; (3) vocabulary (including collocations and idiomatic language) is generally clear but limited; and (4) a limited range of grammatical structures are used with some errors which may impact communication. Does the response have a satisfactory effect on the reader, such that the writer communicates their position/describes the image despite lapses, with a level of language that allows them to generally complete the task despite errors?

A2 The response minimally achieves the task requirements and may be somewhat off-topic or underlength: (1) the response is limited to simple descriptions/personal opinions and topics and may be unclear, irrelevant, or written in an inappropriate style or format (2) the response uses some simple cohesive devices but may be repetitive or incoherent at times; (3) vocabulary is limited and often inaccurate or unclear; and (4) grammar structures are basic and there are frequent errors which may impact communication. Does the response have a poor effect on the reader, such that the writer communicates only basic impressions or opinions/a basic description, with a level of language that allows them to only minimally complete the task despite numerous errors?

A1 The response does not achieve the task requirements and may be off-topic or very underlength: (1) the response is limited to simple personal information and does not present a position/describe the image. Ideas are often unclear or irrelevant. (2) the response does not demonstrate organizational features and is composed of isolated phrases and sentences; (3) vocabulary is very limited, inaccurate, and is insufficient for the task; and (4) only basic grammatical structures are produced and errors predominate. Does the response have a very poor effect on the reader, such that the writer does not communicate a relevant position/adequately describe the image, with a level of language that does not allow them to successfully complete the task?

No-Response There is no response, it is very minimal, or the test-taker indicates that they cannot answer the question (e.g., "I don't understand", "Sorry my English is bad", etc.).

Nonsense/Off-Topic The test-taker does not respond to the prompt in good faith, repeats the prompt without responding to it, or intentionally goes off-task in an attempt to "trick" the system (e.g., by writing random words, writing in a non-English language, writing random strings of letters, or giving a memorized off-topic response).

C GPT Prompts

The wording and design of the prompts provided to GPT can affect its performance. In this appendix, we provide the exact details of each prompt we used.

For our purposes, there are two components to the GPT prompts: the system message and the conversation turns. The system message tells ChatGPT the role it is playing in the conversation, and helps set its behavior during the interaction. For the system messages, we used two different messages, depending on whether the rubric was provided or not.

When providing a minimal rubric to GPT without asking for a rationale, we used the following message:

```
You are a rater for writing responses on
    a high-stakes English language exam
    for second language learners. You
    will be provided with a prompt and
    the test-taker's response.

Ratings are based on the CEFR scale.
    Each rating should be one of the
    following: [A1], [A2], [B1], [B2], [
    C1], [C2], [Nonsense/Off-Topic], or
    [No-Response].

You should assign a [No-Response] rating
    if:
- There is no response to assess.
- There is no or very minimal response.
- The test-taker indicates they cannot
    answer the question (e.g., I don't
    understand, Sorry my English is bad,
    etc.).

You should assign a [Nonsense/Off-Topic]
    rating if:
- The test-taker is not responsive to
    the prompt in good faith:
- The test-taker repeats the prompt but
    does not respond to it.
- The test-taker intentionally goes off-
    task in some way to 'trick' the
    system, e.g., by writing random
    words, writing in a non-English
    language, writing random strings of
    letters, or giving a memorized off-
    topic response.

You should reply to each response with
    just your rating: do not explain or
    justify it.
```

When the rubric was provided to GPT, we used the message below, which adds the descriptions for each CEFR level. We used the same descriptions as defined in Appendix B, so we elide them here, replacing them with a comment between angled brackets <>, for brevity.

```
You are a rater for writing responses on
    a high-stakes English language exam
    for second language learners. You
    will be provided with a prompt and
    the test-taker's response.

Ratings are based on the CEFR scale.
    Each rating should be one of the
    following: [A1], [A2], [B1], [B2], [
    C1], [C2], [Nonsense/Off-Topic], or
    [No-Response].

Scoring Criteria:

For each CEFR rating, there is a
    description which addresses relevant
    aspects of language related Content,
    Discourse, Vocabulary, and Grammar.
    When assigning a score, the overall
    holistic impression should be
    considered  it is not necessary for
    a test=taker to achieve all of the
    positive characteristics of a grade
    as long as overall the descriptor is
    the best match.

Rating: [C2]
Description: <See description in
    Appendix A above>

<Repeated for ratings C1 - A1>

You should assign a [No-Response] rating
    if:
- There is no response to assess.
- There is no or very minimal response.
- The test-taker indicates they cannot
    answer the question (e.g., I don't
    understand, Sorry my English is bad,
    etc.).

You should assign a [Nonsense/Off-Topic]
    rating if:
- The test-taker is not responsive to
    the prompt in good faith:
- The test-taker repeats the prompt but
    does not respond to it.
- The test-taker intentionally goes off-
    task in some way to 'trick' the
    system, e.g., by writing random
    words, writing in a non-English
    language, writing random strings of
    letters, or giving a memorized off-
    topic response.

You should reply to each response with
    just your rating: do not explain or
    justify it.
```

In both cases, we explicitly instructed GPT not to explain or justify its responses, to ensure that a definitive rating that could be parsed and used in the evaluation would be provided. When we experimented with requesting rationales as described in Experiment 2, we replaced the last line with the following:

```
You should reply to each response with
    your rationale and rating in the
    following format:

Rationale: <<<Your rationale here.>>>

Rating: [<<<Your rating here.>>>]
```

The conversation turns were used to provide GPT with the essay to be rated, and to elicit a rating. It was also used to provide GPT with calibration examples, when applicable. In both cases, we used the same format.

The user message provides the essay prompt and the test-taker's response. As recommended by OpenAI, both are surrounded in triple-quotes.

```
Prompt: """
<Essay prompt placed here.>
"""

Response: """
<Essay response placed here.>
"""
```

The assistant response message following each user message would simply contain the rating in square brackets (e.g., `[B2]` or `[Nonsense/Off-Topic]`). In most cases, GPT would prefix its response with `Rating:`, which we simply dropped.