Automated Evaluation of Written Discourse Coherence Using GPT-4

Ben Naismith
Duolingo
ben.naismith@duolingo.com

Phoebe Mulcaire
Duolingo
phoebe@duolingo.com

Jill Burstein
Duolingo
jill@duolingo.com

Abstract

The popularization of large language models (LLMs) such as OpenAI's GPT-3 and GPT-4 have led to numerous innovations in the field of AI in education. With respect to automated writing evaluation (AWE), LLMs have reduced challenges associated with assessing writing quality characteristics that are difficult to identify automatically, such as discourse coherence. In addition, LLMs can provide rationales for their evaluations (ratings) which increases score interpretability and transparency. This paper investigates one approach to producing ratings by training GPT-4 to assess discourse coherence in a manner consistent with expert human raters. The findings of the study suggest that GPT-4 has strong potential to produce discourse coherence ratings that are comparable to human ratings, accompanied by clear rationales. Furthermore, the GPT-4 ratings outperform traditional NLP coherence metrics with respect to agreement with human ratings. These results have implications for advancing AWE technology for learning and assessment.

1 Introduction

Recent advances in large language models (LLMs; Brown et al., 2020), and in particular OpenAI's GPT-4 model (Eloundo et al., 2023; OpenAI, 2023), have led to a paradigm shift with regard to what machines can generate, such as coherent writing. We are now witnessing the potential power and exponential growth of AI in education, though the impact of LLMs used for educational purposes is still largely unexplored. For instance, applications not intended for educational purposes, such as ChatGPT, are being used in educational contexts – everyone with access to the internet can now ask ChatGPT to complete writing tasks, from generating outlines and ideas, to summarizing documents, to essay writing. With these novel capabilities, we can see immediate advantages, such as leveraging GPT-4 for instructional purposes (e.g., automatic item generation, see Attali et al., 2022), and disadvantages (e.g., increased plagiarism, see Eliot, 2022). In addition, we are learning about current potential shortcomings of LLMs (e.g., hallucinations or low-quality content generation) due to miscalibrated expectations of what LLMs can do or the pitfalls of non-optimized prompt engineering.

To further our understanding of one innovative application of AI in education, this paper presents an exploratory evaluation of LLMs for automated writing evaluation (AWE). Specifically, it is the first study to our knowledge to examine GPT-4's ability to provide a rating (score) and rationale for one aspect of writing quality – discourse coherence quality – in test-taker written responses to an online, high-stakes writing assessment item. Discourse coherence is notoriously challenging to satisfactorily assess using AWE, and as such, there is great value in determining whether state-of-the-art AI can be used to improve upon prior options. We believe that the method described in the paper should be generalizable to similar datasets that are publicly available. However, caution in the use of GPT-4 ratings is warranted due to limited reproducibility, the possibility of bias, and limited insight into the underlying processes that determine the ratings.

2 Background

In the field of AI in education, AWE is one of the most widely researched and mature areas. AWE systems evaluate written text quality (Shermis and Burstein, 2003, 2013; Attali and Burstein, 2006) and are widely used for high-stakes writing assessment and instruction. These systems are informed by theoretical writing subconstructs (i.e., factors contributing to writing quality) described in human scoring rubric criteria such as grammatical accuracy, lexical sophistication, relevance, and discourse coherence. These rubric criteria are developed and used by educational testing organizations for scoring purposes and are often informed by

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 394–403

education policy (e.g., Common Core Standards, 2010 and Council of Europe, 2020). AWE systems typically provide a holistic score that indicates the overall quality of writing, given a set of rubric criteria. The performance of these scores (accuracy) is then reported through human-system agreement, a well-studied evaluation measure that is typically quite high on modern systems (e.g., Bridgeman, 2013).

In recent years, large language models (and earlier models pretrained on unlabeled text) have been leveraged to good effect in various ways to improve AWE performance through the use of "transformers", a type of deep learning neural network. For example, Lagakis and Demetriadis (2021) found that the best AWE performance was achieved through a model incorporating linguistic features with the BERT language model (Devlin et al., 2019). More recently, Mizumoto and Eguchi (2023) explored the capabilities of GPT-3 to holistically rate test-taker essays in the TOEFL11 corpus (Blanchard et al., 2013). The researchers showed Human-GPT-3 agreement rates to be reasonable (exact agreement 54.33%, adjacent agreement 89.15%). The model's performance was then further improved by combining GPT ratings and a range of lexical, syntactic, and cohesion features, resulting in substantial Quadratic Weighted Kappa (QWK) of 0.61. Methodologically, it is important note that in their study, the same prompt was used in all conditions, and this prompt did not include examples or ask for rationales for the ratings. To our knowledge, there have been no similar studies with the newer GPT-4 or with comparing different prompt configurations to elicit ratings.

While AWE systems show strong performance for holistic scoring, scores for discourse coherence quality alone have been a challenging area of NLP research (Hearst, 1997; Barzilay and Lapata, 2008; Burstein et al., 2013; Somasundaran et al., 2014; Lai and Tetreault, 2018). Although some discourse features can be considered "surface-based," for example, pronoun referents and transition terms used in a text, operationalizing aspects of coherence such as the relationship between ideas is less straightforward and involves labor-intensive annotations or less easily interpretable LLM-derived features. In particular, it may be difficult to tell whether LLM-generated "analyses" of a text actually reflect the same aspects of writing that superficially similar human-written analyses describe.

Further complicating coherence assessment is the fact that different disciplines, from linguistics (Halliday and Hasan, 1976) to cognitive psychology (Graesser et al., 2004), to education research (Van den Broek et al., 2009), share slightly different views about how coherence is constructed by readers of a text. However, a common thread is that discourse coherence pertains to the textual continuity or flow of a text, that is, the overall sense of unity and meaning that is conveyed by a text. Within the construct of discourse coherence, assessment rubrics often directly or indirectly refer to subconstructs such as clarity (how easy to understand ideas and purpose; readability; and impact of lexis/grammar on coherence); flow (sequence/progression of ideas; use of linking words; and referencing); structure (appropriacy of paragraphing; introducing/concluding; and connection between topics); and effect on reader (naturalness of cohesion; appropriacy of cohesive features; repetitiveness; and helpfulness to reader for understanding the response).

3 Methods

In this section we describe the dataset of test-taker responses and the processes for evaluating them through human and automated means.

3.1 DET coherence (DET-Coh) dataset

The DET coherence (DET-Coh) dataset contains test-taker written responses from the operational Duolingo English Test (DET). The DET is a high-stakes English language test whose primary use is for higher-education admissions. One of the writing tasks, *Writing Sample*, is an independent writing task in which test takers respond to a prompt requiring them to produce a persuasive or narrative extended piece of writing in five minutes (see Cardwell et al., 2023, for further details). *Writing Sample* is scored using AWE; the scoring model includes features to assess the writing subconstructs of Content, Discourse coherence, Grammar, and Vocabulary.

In total, there are 500 written responses in the DET-Coh dataset, sampled from the operational DET during a 7-month span in 2022. DET-Coh was deliberately constructed and stratified so that it contains an equal distribution of males and females, as well as an equal distribution of the seven most common first-language groups in the DET test-taker population (Chinese, Arabic, Spanish, Telugu, En-

glish, Bengali, Gujarati). An approximately even distribution of proficiency levels was also ensured based on DET automated scoring models. These levels align with the levels of the Common European Framework of Reference (CEFR; Council of Europe, 2001, 2020), an international standard for describing language ability, ranging from level A1 (basic) to C2 (proficient) on a six-point ordinal scale.

3.2 Human scoring

Test-taker writing responses were scored by four expert raters, each with second language (L2) teaching qualifications, extensive L2 teaching experience, and L2 assessment experience with international proficiency exams. Of the original 500 responses, 20 were double rated collaboratively for standardization, and 80 were rated independently by pairs of raters to assess interrater agreement. The interrater agreement for these 80 items was 0.72 exact agreement and 0.93 QWK, indicating excellent agreement. Having established rater reliability, the remaining 420 responses were rated by a single rater each.

All ratings were based on writing coherence task rubrics created for this study (see Appendix A, Table 2, for full rubric text). The rubric was developed using a 6-point, holistic scale that was based on the six levels/descriptors from the CEFR, other coherence research studies, and publicly-available rubrics from testing organizations. A rating of 0 was also given to blank or bad-faith responses in which the test taker did not attempt to respond to the prompt. In addition, one rater produced paragraph-long rationales for 12 of the ratings (two at each scale point) for the purposes of few-shot prompting (6 responses) and qualitative analysis (6 responses).

3.3 GPT-4 ratings and rationales

To elicit GPT-4 coherence ratings and rationales, we used the OpenAI Python API. The full prompt given to GPT-4 for each student response consisted of the following ordered elements:

- Task – a short paragraph explaining the task of rating the coherence of a written text written by a language learner in response to a prompt
- Rubrics – see Section 3.2 for description
- Guidelines – bullet point guidelines relating to expected terminology and style
- Examples – six training items removed from the dataset (one from each scale point), accompanied by expert ratings and/or rationales (depending on the condition) for the ratings based on the rubrics
- Prompt – the prompt the test taker responded to
- Response – the test taker's response

Based on these elements, GPT-4 was called to complete three different conditions: 1) rating then rationale (rating-first), 2) rationale then rating (rationale-first), and 3) rating only (rating-only).

3.4 NLP coherence metrics

As a baseline, coherence ratings were predicted using a set of simple NLP features based on CohMetrix (Graesser et al., 2004):

- Binary overlap between sentence pairs: overlap of arguments, nouns, or word stems between two sentences
- Proportional overlap between sentence pairs: overlap of content words as a proportion of all content words in a sentence pair
- Coreference overlap: number of coreferent mentions between two sentences found using a neural coreference model (Lee et al., 2018)
- LSA similarity: measure of the similarity between two sentences calculated using an LSA model trained on a large sample of writing responses

Two versions of each feature were computed, one considering only adjacent sentence pairs ("local"), and one considering all pairs of sentences in a response ("global"). For each response, we fit a linear regression model using the features and human ratings for all other responses, then predicted the rating for the held-out response.

4 Results

4.1 Rating comparison

Ratings from GPT-4 and the baseline model are compared to the human ratings on all items not included in the prompt (Table 1); for double-rated items the second rating was used. The findings show that the baseline linear regression model is moderately predictive of the human ratings, reaching an adjacent agreement score of 0.82 and Spearman correlation (ρ) of 0.47 despite its simplicity.

Metric	Human-baseline model	Human-GPT-4 (rating-rationale)	Human-GPT-4 (rationale-rating)	Human-GPT-4 (rating-only)
Exact agreement	0.36 (0.31-0.40)	**0.56** (0.52-0.60)	0.53 (0.49-0.58)	0.51 (0.46-0.56)
Adjacent agreement	0.82 (0.78-0.85)	0.96 (0.95-0.98)	**0.97** (0.95-0.98)	0.95 (0.93-0.97)
Cohen's Kappa	0.13 (0.08-0.18)	**0.43** (0.38-0.48)	0.40 (0.36-0.46)	0.36 (0.31-0.42)
Quadratic Weighted Kappa	0.39 (0.33-0.45)	0.81 (0.79-0.84)	**0.82** (0.79-0.85)	0.78 (0.75-0.82)
Spearman's rho	0.47 (0.39-0.53)	**0.82** (0.79-0.85)	**0.82** (0.79-0.85)	0.79 (0.76-0.83)

Table 1: Coherence rating agreement rates, with bootstrapped 95% confidence intervals (percentile). Bold indicates the best performance for a metric. All GPT-4 conditions have significantly better agreement with human ratings than the baseline model across all metrics. The two GPT-4 conditions which produce a rationale have marginally (but not significantly) better agreement than the rating-only condition.

All GPT-4 conditions significantly outperform this baseline model, obtaining a correlation of 0.82 with the human rating in the rationale conditions.

Inspired by Mizumoto and Eguchi (2023), we also experimented with a linear regression model that includes the GPT-4 rating as an additional feature along with the baseline features, potentially combining the strengths of the two models. However, unlike that work, we found that the combined model performs almost identically to the GPT-4 ratings on their own and so do not analyze it further.

The rationale-first condition could be interpreted as a form of chain-of-thought (CoT) prompting (Wei et al., 2022) which has been shown to improve performance on reasoning tasks. That work also hypothesized that showing examples with the reasoning after the answer in the prompt could improve performance, by drawing attention to relevant aspects of the tasks, but found it performed similarly to the baseline and worse than CoT prompting. By contrast, we find that GPT-4's agreement is slightly improved by the use of rationales, regardless of their position. However, there are no significant differences between the agreement rates of any of the GPT-4 configurations, with all versions showing overlapping confidence intervals. These findings suggest that there is not a CoT effect for this task.

We focus on the rating-first condition for error analysis. GPT-4's ratings have less variance than human ratings (0.37 vs 0.42), especially producing fewer 1, 5, and 6 ratings (most samples rated 1 by humans are rated 2 by GPT-4). This behavior is actually in-line with a well-documented tendency of human raters, the central tendency effect, in which raters avoid the extremes of rating scales (McNamara et al., 2019). One hypothesis to account for this pattern is that GPT-4 is imitating trends found in its pre-training data. When GPT-4's ratings differ from human ratings (n=143), they are also slightly but significantly lower on average ($\mu = 3.17$ for GPT-4 in the rating-first condition vs $\mu = 3.41$ for the comparable human rating, p=0.04 with Welch's t-test). In the rating-first condition, GPT-4 mentions "spelling" in 43% of rationales where its rating differs from the human rating, versus only 30% of equally rated rationales. Speculatively, this may indicate an oversensitivity to spelling errors; human raters may be better able to discern the intended word while GPT-4's token-based representation may prevent such recognition.

4.2 Rationale comparison

The six human-generated rationales were compared to GPT-4 rationales in terms of their content and style. Figure 1 provides an example of a response with a 3 rating (CEFR B1; human and GPT-4 rating in agreement), answering a prompt about the advantages and disadvantages of using books, movies, and TV shows to learn about different cultures. Figure 2 shows the accompanying human and GPT-4 rationales. Of note, the trends exemplified in this set of examples hold true for all six pairs of human-

I think that we can learn different cultures from using books and watching movies and tv shows. We can be able to get to know how they live, eat, dance and party. There are many advantages and disadvantages learning about different cultures through books or movies. Firstly, the advantages of learning about the cultures are we can get to learn morals, codes and ethics of that certain culture and know their language, hardwork that they do in that characters and mostly the people mentality. Secondly, disadvantages are we acn also get encounter dru

Figure 1: Example test-taker response (3 rating / CEFR B1).

Human rationale	**GPT-4 rationale**
The response has variable coherence. A relevant position is presented and a point in favor of using media to learn about different cultures is given. However, the topic is not developed, with the second point about disadvantages ending abruptly so that the intended message is not conveyed. There is some appropriate use of basic discourse markers (e.g., "Firstly", "Secondly"). The language is sufficient to convey basic ideas on the topic, but lexis from the prompt is used repetitively and at times coherence is affected (e.g., "hardwork that they do in that characters"). Overall, the response produces a satisfactory effect on the reader, negatively impacted by language inaccuracies and lack of development.	The response has variable coherence. The writer begins by stating that it is possible to learn about different cultures through books, movies, and TV shows, and provides some examples of what can be learned. The writer then moves on to discuss the advantages and disadvantages of this approach. However, the response is unfinished and the disadvantages are not fully developed. The response demonstrates some organization, with the use of discourse markers such as "Firstly" and "Secondly" to link ideas. However, there are some inaccuracies in grammar and vocabulary that impact coherence, for example "we acn also get encounter dru". Overall, the response is somewhat effective in conveying the intended message, but the unfinished nature of the response and inaccuracies in language limit the rating to a 3.

Figure 2: Comparison of rationales for assessment of Figure 1.

GPT4 rationales we analyzed.

Comparing the content of the two rationales, there is a great deal of consistency, with both addressing the clarity, flow, structure, and effect on the reader. For example, both rationales describe how the writer's position is initially presented and provide a specific example. The two rationales also note the same main weakness relating to the lack of development of the second point. The two rationales then move on to describe how discourse markers are used to achieve local coherence, even highlighting the same two examples of *Firstly* and *Secondly*. Examples of coherence negatively affected by language inaccuracies are then given, though different examples are used to exemplify this point in the two rationales. Finally, both rationales summarize the reason for the overall satisfactory effect on the reader.

Likewise, in terms of style, the GPT-4 rationale has clearly adopted the examples and followed the guidelines from the prompt. The rationales use terminology such as *the writer* (rather than *the author/student/learner*), are written in the 3rd person, and are within the desired length range. The overall format of the rationale is also consistent, starting with an overall statement of coherence, moving to discuss each of the coherence subconstructs in turn, then closing with an overall description of the effect on the reader.

To further illustrate how GPT-4 rationales discuss and incorporate key concepts from the rubrics, we conducted a simple corpus analysis of key words. First, a frequency list was compiled of the most common words (tokens) in the rationales. We restricted this list to content words (nouns, verbs, adjectives, and adverbs) and only counted the first occurrence of each word in each rationale. Of interest, we noted commonly used words related to discourse coherence including *ideas* (n=509), *developed* (n=406), *impact* (n=297), *inaccuracies* (n=278), and [discourse] *markers* (n=264). Figure 3 presents a concordance of the first ten oc-

. The reader is able to discern some relevant	ideas	but the response is not well-organized or developed
a result the reader struggles to identify any relevant	ideas	. There is no evidence of discourse features such
the response contains a number of incomplete or incoherent	ideas	for example , the issue of scales to travel
lacks an overall structure appropriate for the task and	ideas	are not clearly presented or arranged . The discoursal
has minimal coherence . The writer expresses two basic	ideas	: that video conferencing applications are easy to learn
lacks an overall structure appropriate for the task and	ideas	are not clearly presented or arranged . Grammar and
coherence . It is possible to discern some relevant	ideas	, such as the writer's decision to date
lacks an overall structure appropriate for the task with	ideas	not clearly presented or arranged . As a result
is minimally coherent with the writer expressing two basic	ideas	: that taking notes with pen and paper takes
coherence . It is possible to discern some relevant	ideas	such as that travel can provide information and

Figure 3: Uses of the key term *ideas* in the GPT-generated rationales with local context.

currences of the most frequent of these key words, *ideas*, to provide the context in which this term is being used. Here we see that *ideas* are described in a number of ways, for example, *relevant*, *appropriate*, *basic*, and *incoherent*, all of which are descriptors used in the rubrics. As importantly, these *ideas* are discussed in terms of how they are presented and arranged in the response, and specific examples of test-taker ideas are listed, that is, there is a focus on content and meaning, not just mechanical use of linguistic features.

5 Discussion and conclusions

This study examined the effectiveness of using GPT-4 for assessing written discourse coherence of test-taker responses on a high-stakes English proficiency test. We found that GPT-4 is able to rate the coherence of writing samples with a good degree of accuracy in terms of agreement with the gold-standard human ratings; regardless of the exact order of the prompt (rating-first or rationale-first), the exact agreement rates were >0.5 and the QWK >0.8. Prompts eliciting rating-only performed slightly worse, though not significantly so. Importantly, all permutations of the GPT-4 prompt greatly outperformed a baseline NLP model composed of traditional coherence features. Human-GPT-4 agreement rates could likely be improved with further tailoring of the prompt; for example, based on the qualitative analysis, we might suggest additional guidelines to lower the weighting that GPT-4 assigns to spelling errors as it may be overvaluing their importance.

Studies such as this one have important implications for the field of AWE. There is often a tension between designing features that are easily interpretable but provide limited signal (e.g., the number of discourse markers) versus features which are less clearly aligned with human rubrics but which may provide more predictive power (e.g., perplexity of the response under a language model). The promise of ratings based on GPT-4 is that they may bridge this gap by providing quantitative features which seemingly are based on aspects of language of importance to the language assessment community. In the future we therefore expect to see research in a similar vein which looks at further optimizing prompts to elicit ratings and clear, interpretable rationales, especially for subconstructs of writing which have historically been a challenge to measure through automated means. In using LLMs in this manner, we could reduce the "epistemic opacity" of AWE processes (Ferrara and Qunbar, 2022), that is, modern automated assessment could become less of a black box, thereby improving stakeholder confidence in the results. Nevertheless, although these results are encouraging, it is important to recognize that the interpretability promised by generated rationales is limited: GPT-4's rationales may not accurately reflect the process used to assign the ratings. In particular, rationales may present rationalizations for decisions actually grounded in biasing features, as was found to be true of CoT explanations in Turpin et al. (2023). Rationales should therefore not be treated as offering insight into the *process of generating* ratings, even when they provide true and relevant information about the response.

The fact that rationales do not reflect a "thought process" by GPT does not, however, reduce their value in all contexts. As suggested in Mizumoto and Eguchi (2023), rationales can support language learning by providing instantaneous feedback. In the context of test takers of the DET, rationales such as the ones in this report are particularly useful because they are based on task- and construct-specific rubrics. For example, test takers completing a practice test would greatly benefit from feedback tailored to the writing subconstructs, such as discourse coherence, that will be assessed under

operational test conditions. GPT-4 could also then be further beneficially exploited by querying it to produce an improved version of the test taker's own response; in other words, a personalized model answer.

Figure 4 is an example of one such model answer, revising the response from Figure 1. The same prompt as before was used for generating this revision, with the following amendment:

> *Now, write a revised version of the following response with improved coherence according to the rubric. Stick closely to the original in content, and do not rewrite too extensively; simply improve the organization and complete unfinished ideas.*

In this revision, we see that the test taker's ideas are maintained, for example, the benefits of learning about how other cultures eat and dance. In addition, the appropriate use of some discourse markers from the original are left intact. In contrast, key coherence weaknesses from the original are addressed, most notably the lack of development of disadvantages and the language inaccuracies which impacted clarity. There remains some repetitiveness in the revision of language from the task prompt, but this issue did not prevent the revised response from being independently rated a 5 (CEFR C1) by both GPT-4 and a human rater. As such, this revision would seem a reasonable goal for this particular test taker.

On a broader level, the focus of our study, including the importance of transparency, is in line with the larger field of educational AI application development where responsible AI is a key focus (ATP, 2021; Dignum, 2021; ITC-ATP, 2021; Burstein, 2023; Department for Science, Technology & Innovation, 2023). As novel ideas, applications, and research questions emerge around the use of LLMs for educational purposes, it is essential that research communities investigating the use and impact of AI for education build a research agenda. In light of the need to ensure responsible use of AI in education, researchers need to anticipate and pressure test possible uses of AI for education to ensure fairness.

Acknowledgements

We thank the raters for their contribution to the DET-Coh dataset.

I believe that we can learn about different cultures by using books, watching movies, and TV shows. Through these mediums, we can gain insights into how people from various cultures live, eat, dance, and celebrate. There are both advantages and disadvantages to learning about different cultures through books, movies, and TV shows. Firstly, the advantages include being able to learn about the morals, codes, and ethics of a particular culture. Additionally, we can become familiar with their language and observe the hard work and mentality of the characters portrayed in these mediums.
On the other hand, there are also disadvantages to this approach. For example, we may encounter negative stereotypes or misrepresentations of a culture, which can lead to misunderstandings and misconceptions. Furthermore, relying solely on books, movies, and TV shows may not provide a comprehensive understanding of a culture, as these mediums often focus on specific aspects or situations.
In conclusion, while there are benefits to learning about different cultures through books, movies, and TV shows, it is important to be aware of the potential drawbacks and seek additional sources of information to gain a more accurate and well- rounded understanding of various cultures.

Figure 4: Response revised by GPT-4.

References

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

Yigal Attali, Andrew Runge, Geoff T. LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A. Von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Daniel Blanchard, Joel R. Tetreault, Derrick Higgins, A. Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native english. *ETS Research Report Series*, 2013:15.

Bren Bridgeman. 2013. *Human ratings and automated essay evaluation*, pages 243–254. Routledge.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, et al. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901.

Jill Burstein. 2023. Duolingo English Test Responsible AI Standards. Technical report, Duolingo.

Jill Burstein, Joel Tetreault, and Martin Chodorow. 2013. Holistic discourse coherence annotation for noisy essay writing. *Dialogue & Discourse*, 4(2):34–52.

Ramsey Cardwell, Ben Naismith, Geoffrey T. LaFlair, and Steven Nydick. 2023. Duolingo English Test: Technical Manual.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Council of Europe Publishing, Cambridge, UK.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing, Cambridge, UK.

Department for Science, Technology & Innovation. 2023. A pro-innovation approach to AI regulation. White paper, Crown.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Virginia Dignum. 2021. The role and challenges of education for responsible AI. *London Review of Education*, 19(1):1–11.

Lance Eliot. 2022. Enraged worries that generative AI ChatGPT spurs students to vastly cheat when writing essays, spawns spellbound attention for AI ethics and AI law. *Forbes*.

Tyna Eloundo, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint axXiv:2303.10130*.

Steve Ferrara and Saed Qunbar. 2022. Validity arguments for AI-based automated scores: Essay scoring as an illustration. *Journal of Educational Measurement*.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiquiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

Michael A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London, UK.

Marti A. Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. In *Computational Linguistics*, volume 23, pages 33–64.

International Test Commission and Association of Test Publishers (ITC-ATP). 2022. Guidelines for technology-based assessment. Technical report, International Test Commission and Association of Test Publishers, Washington, D.C.

Paraskevas Lagakis and Stavros Demetriadis. 2021. Automated essay scoring: A review of the field. In *Proceedings of the 2021 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–6. Institute of Electrical and Electronics Engineers.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: a dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of NAACL*, pages 687–692. Association for Computational Linguistics.

Tim F. McNamara, Ute Knoch, and Jason Fan. 2019. *Fairness, Justice and Language Assessment*. Oxford University Press, Oxford, UK.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Educational Technology Research and Development*.

National Governors Association Center for Best Practices and Council of Chief State School Officers. 2010. Common Core State Standards.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mark D. Shermis and Jill Burstein. 2003. *Automated essay scoring: A cross-disciplinary perspective*. Routledge, Mahwah, NJ.

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

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *The 25th International Conference on Computational Linguistics (COLING)*, pages 23–29.

The International Privacy Subcommittee of the ATP Security Committee (ATP). 2021. Artificial intelligence and the testing industry: A primer. Association of Test Publishers.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Paul Van den Broek, Charles R. Fletcher, and Kirsten Risden. 2009. Investigations of inferential processes in reading: A theoretical and methodological integration. *Discourse Processes*, 16:169–180.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Discourse coherence rubrics

Rating	Description
6 (C2)	The response is highly coherent: (1) the ideas and purpose of the response are completely clear, and lexical/grammatical choices effectively enhance coherence; (2) the response is smoothly-flowing, with a clear sequence of ideas which are cohesively linked using a range of discoursal features (including cohesive devices and referencing); (3) the response is logically and appropriately structured for the task, with topics effectively developed and expertly connected. *Do the discoursal features have an excellent effect on the reader, such that they are completely natural and do not attract any attention; they are appropriate for the text type; and they help the reader to understand the ideas in the response?*
5 (C1)	The response is coherent: (1) the ideas and purpose of the response are clear, and lexical/grammatical choices rarely impact coherence in any way; (2) the response has a clear progression and ideas are linked using a range of discoursal features (including cohesive devices and referencing), though there may be some under-/over-use; (3) the response is well-structured for the task, with topics appropriately introduced, developed, and concluded. *Do the discoursal features have a very good effect on the reader, such that they are mostly natural; they are appropriate for the text type; and they allow the reader to follow along easily?*
4 (B2)	The response is mostly coherent: (1) the ideas and purpose of the response are clear, and lexical/grammatical choices generally do not impact coherence though they may lead to some instances of confusion; (2) the response has a generally clear overall progression and ideas are generally linked effectively despite some inaccurate or unnatural use of cohesive devices and referencing; (3) the response is generally well-structured for the task, with topics usually developed in some detail though some arguments may lack clarity. *Do the discoursal features have a good effect on the reader, such that they are mostly appropriate despite some inaccuracies or repetitiveness, and they allow the reader to follow along?*
3 (B1)	The response has variable coherence: (1) the reader can generally follow the overall purpose and the main points made by the writer, though lexical/grammatical choices impact coherence at times; (2) the response demonstrates some organization, linking discrete elements in a linear sequence, though the use of referencing and cohesive devices may be inaccurate and the overall progression may be unclear; (3) the response contains evidence of some structure appropriate for the task, though topics are not always developed, clearly distinct, or clearly connected, and argumentation may lack coherence. *Do the discoursal features have a satisfactory effect on the reader, such that they are somewhat effective in conveying the intended message, despite inaccuracies or repetitiveness which impact coherence and cohesion?*
2 (A2)	The response has minimal coherence: (1) it is possible to discern some relevant ideas, though the overall purpose of the response may be incoherent and the lexical/grammatical choices lead to breakdowns in coherence other than for basic ideas; (2) there is limited evidence of organizational features including cohesive devices and referencing, and when used, such features may be inaccurate and lead to breakdowns in coherence; (3) the response lacks an overall structure appropriate for the task and ideas are not clearly presented or arranged. *Do the discoursal features have a poor effect on the reader, such that they are mostly not effective in conveying the intended message, with inaccuracies or repetitiveness often impacting coherence and cohesion?*
1 (A1)	The response mostly lacks coherence: (1) it is a strain on the reader to identify points the writer is trying to make, with lexical/grammatical choices greatly impacting coherence throughout; (2) there is no apparent logical organization of ideas other than simple isolated phrases, with no or minimal/inaccurate use of discoursal features such as linking and referencing; (3) there is no overall structure appropriate for the task and ideas are difficult to discern. *Do the discoursal features have a very poor effect on the reader, such that they are mostly not effective in conveying the intended message, with inaccuracies or repetitiveness often impacting coherence and cohesion?*
0	There is no response or the test-taker is not responsive to the prompt in good faith, e.g., the test taker repeats the prompt but does not respond to it, or the the test taker intentionally goes off-task in some way to "trick" the system, for example, by writing random words, writing in a non-English language, writing random strings of letters, or giving a memorized/plagiarized off-topic response.

Table 2: Discourse coherence rubrics used for human rating and GPT prompting