LLMs can Perform Multi-Dimensional Analytic Writing Assessments: A Case Study of L2 Graduate-Level Academic English Writing

Anonymous ACL submission

Abstract

The paper explores the performance of LLMs in the context of multi-dimensional analytic writing assessments, i.e. their ability to provide both scores and comments based on multiple assessment criteria. Using a corpus of literature reviews written by L2 graduate students and assessed by human experts against 9 analytic criteria, we prompt several popular LLMs to perform the same task under various conditions. To evaluate the quality of feedback comments, we apply a novel feedback comment quality evaluation framework. This framework is interpretable, cost-efficient, scalable, and reproducible, compared to existing methods that rely on manual judgments. We find that LLMs can generate reasonably good and generally reliable multi-dimensional analytic assessments. We release our corpus for reproducibility.

1 Introduction

Assessing the writing quality of essays manually is both time-consuming and labor-intensive. This task becomes even more demanding and challenging due to high cognitive load (Cai, 2015), when assessors have to assign scores and provide comments based on multi-dimensional analytic criteria, referred to here as *multi-dimensional analytic assessments* (see Fig. 1 for an illustration). For evaluation of non-native language (L2) learners' writing, such precise and multi-dimensional assessments are highly valuable and desirable, but they are often not provided, due to the significant time, cost, and expertise required to produce them. This is also evidenced by the dearth of publicly available L2 writing corpora annotated with multi-dimensional analytic assessments (Banno et al., 2024).

In recent years, large language models (LLMs) have emerged as promising tools for self-regulated writing assessments among L2 learners. A growing number of studies (Chiang and Lee, 2023; Mizumoto and Eguchi, 2023; Han et al., 2024; Yancey et al., 2023, *i.a.*) have indicated the general usefulness of LLMs for automated writing assessments. Given their increasing use for this task, the following question remains understudied: *can LLMs provide reasonably good multi-dimensional analytic writing assessments?* We use the phrase "**reasonably good**" intentionally, given the open-ended nature of the task, particularly generating essay-level feedback comments.

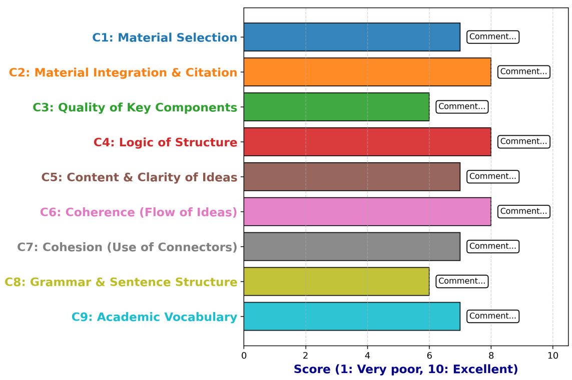


Figure 1: Multi-dimensional analytic assessments, where each assessment contains a score and a comment.

To address this question, we utilize an English-language corpus of literature reviews written by L2 graduate students and assessed by human experts on 9 analytic assessment criteria. We prompt various popular LLMs to assess the corpus using the same criteria under various conditions, and we examine the quality of their generated assessments compared to human-generated assessments.

Our study makes three primary contributions:

1. We provide comprehensive and reproducible evidence that LLMs can generate reasonably good and generally reliable multi-dimensional analytic writing assessments. This is the primary goal of this study; we do not argue in favor of a specific LLM, nor do we advocate replacing humans with LLMs for this task.

2. We release a corpus[1] of L2 English graduate-level literature reviews, annotated with multi-dimensional analytic assessments, which will facilitate future studies.

3. We propose and validate a novel LLM-based framework for evaluating the quality of feedback comments. This framework is time- and cost-efficient, scalable, and reproducible, compared to manual judgments. It is also interpretable, compared to direct quality ratings.

2 Related Work

Automated Writing Evaluation (AWE) We use AWE to include both automated essay scoring (AES)[2] and feedback comment generation (Shermis and Burstein, 2013). AWE systems have existed since the 1960s (Page, 1966) and have evolved over time with a predominant focus on AES (Ke and Ng, 2019; Hussein et al., 2019; Zhang and Zou, 2020; Uto, 2021; Lagakis and Demetriadis, 2021). Modern AWE systems use deep neural networks for scoring (Taghipour and Ng, 2016; Alikaniotis et al., 2016; Dong et al., 2017; Rodriguez et al., 2019; Yang et al., 2020; Xie et al., 2022) and feedback comment generation (Nagata, 2019; Han et al., 2019; Babakov et al., 2023). The latter task typically focuses on sentence-level grammatical error identification and correction (Behzad et al., 2024b). Existing non-LLM AWE systems mainly provide holistic assessment, with some specialized systems offering uni-dimensional analytic assessment based on a specific dimension of writing quality (Ke and Ng, 2019; Jong et al., 2023; Banno et al., 2024).

LLMs used for AWE Unlike prior AWE systems, LLMs can be prompted in natural language to jointly score and comment on a given essay. A growing body of research has explored the use of LLMs for assessing L2 writing. For AES, LLMs have been examined for holistic scoring (Mizumoto and Eguchi, 2023; Yancey et al., 2023; Wang and Gayed, 2024), discourse coherence scoring (Naismith et al., 2023), and multi-dimensional analytic scoring (Yavuz et al., 2024; Banno et al., 2024). For feedback comment generation, LLMs have been studied for generating corrective feedback (Mizumoto et al., 2024; Song et al., 2024), holistic feedback (Behzad et al., 2024a,b), and multi-dimensional analytic feedback (Guo and Wang, 2024; Behzad et al., 2024a; Han et al., 2024). Stahl et al. (2024) is the only study we know of which explores LLMs jointly performing scoring and feedback comment generation, but holistically. Moreover, their corpus contains short essays by native speakers and has no human reference comments.

Related Corpora Major L2 writing corpora include TOEFL11 (Blanchard et al., 2013), which contains scored essays from TOEFL tests, and CLC-FCE (Yannakoudakis et al., 2011), which includes error-annotated short texts in response to exam prompts. Other notable corpora are derived from online language learning platforms, such as EFCAMDAT (van Rooy, 2015), Write & Improve (Yannakoudakis et al., 2018), and LEAF (Behzad et al., 2024b), focusing on scoring, grammatical error correction, and personalized feedback, respectively. We are not aware of any publicly available corpora annotated with multi-dimensional analytic scores and comments jointly.

3 Corpus

Overview Our corpus[3] consists of 141 literature reviews written in English by 51 L2 graduate students, with an average word count of 1321 (930 excluding references). The reviews cover five broad topics from the humanities and social sciences, chosen to minimize the need for specialized disciplinary knowledge: (1) the social consequences of legalized cannabis, (2) the Canadian linguistic landscape, (3) online learning, (4) lessons from the COVID-19 pandemic, and (5) pacifism.

The corpus is a result of a large research project conducted at a Canadian university in 2021 with an aim to examine the developmental trajectory of literature review writing skills among L2 graduate students. The project involved three rounds of a 5-unit online tutorial series conducted over the course of 2021, with each round lasting 13 weeks (see Appendix A for details). Participation was voluntary, with 31 participants completing all five writing tasks across all rounds, and 20 further students completing at least one task before withdrawing.

Essay Authors The corpus authors comprise a diverse group of L2 learners, representing a wide range of first languages and enrolled in graduate programs across various disciplines at multiple

[1] Corpus will be released upon acceptance.

[2] AES is sometimes conflated with AWE in the literature (Hockly, 2019). We distinguish them.

[3] The corpus was used in our previous studies with a different focus and has never been made public. We omit citations for these studies for anonymity.

Code	Role	Rounds	Topics	# Essays
A	Graduate RA	1	1-5	27
B	Graduate RA	1-3	1-5	141
C	Faculty Member	1-3	1, 2, 5	93
D	Faculty Member	1	2	4
E	Faculty Member	1-3	3, 4	43
F	Graduate RA	2, 3	1-5	106

Table 1: Anonymized information for the six assessors (A–F). The columns "Rounds" and "Topics" indicate the specific rounds and writing topics they participated in. Assessors C and E never co-assessed together.

Canadian universities. Their English proficiency ranged from upper-intermediate to advanced, with an average score equivalent to IELTS Band 7 based on conversions from various standardized English language tests. Scores varied from IELTS 6.5 to 8.5, with a standard deviation of 0.55.

Human Assessments Most essays in the corpus were assessed by three (94.3%) or two (5.0%) independent human experts. As illustrated in Fig. 1, the assessments consist of scores on a 10-point scale and comments based on 9 analytic assessment criteria (see Appendix A.3 for details). While scores were required, comments were optional for the assessors. A total of six assessors with professional experience in English language teaching participated at different stages of the research project. Table 1 provides basic information about them.

Assessment Quality The 31 students who completed all writing tasks evaluated the quality of human assessments on a 4-point scale in an anonymous final project survey. Based on the 30 submitted survey responses, all participants agreed that the assessments were at least "useful" (rating = 3), with 24 participants (80%) rating them as "very useful" (rating = 4).

Data Contamination Since the corpus was created prior to the release of ChatGPT and has never been made public, it contains no LLM-generated contents and is free from the risk of data contamination (Jacovi et al., 2023; Sainz et al., 2023), making it an ideal resource for LLM evaluation.

4 A Novel Feedback Comment Quality Evaluation Framework

A common approach to evaluating feedback comment quality for an essay uses manual judgments (e.g., rating on a Likert scale), since generating essay-level feedback is an open-ended task. However, this approach is expensive, time-consuming, not scalable, and may not always be reproducible.

For L2-related feedback comments, common criteria for assessing comment quality include specificity, relevance, helpfulness (Han et al., 2024; Stahl et al., 2024; Behzad et al., 2024a,b), and the ability to identify writing problems (Stahl et al., 2024; Behzad et al., 2024a,b). These criteria reflect a common and practical need of L2 learners to be shown specific problems in their essays and how to correct them to improve their writing quality.

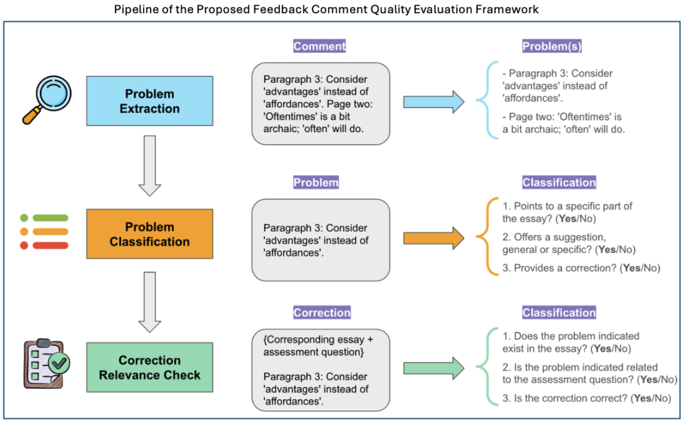
To address the issues of manual judgment, we propose an automatic evaluation framework that *evaluates the quality of a feedback comment in terms of its ability to effectively identify relevant writing problems within the assessed essay.* As illustrated in Fig. 2 (left), the framework utilizes LLMs to extract problems identified in assessment comments and to characterize their specificity and potential helpfulness. The framework consists of the following three steps.[4]

Problem Extraction We start out by extracting any writing problems stated or implied in assessment comments, along with any relevant contextual information for each problem, such as further explanations, suggestions for improvement, concrete corrections, or clarifying questions. We define a problem as any writing-related issue that affects the quality of the writing, such as citation errors, logical flaws, or grammatical mistakes.

Problem Classification The extracted problems are further characterized along three dimensions: whether an extracted problem **(1)** points to a specific part of the essay, **(2)** includes any form of suggestion (general or specific), and **(3)** provides a concrete correction that can be directly applied to fix an identified problem. These classifications offer a way to assess the *specificity* and *potential helpfulness* of related comments.

Correction Relevance Check We perform a sanity check to determine whether the proposed correction (and thus the comment) is in fact relevant to the original essay. The Correction Relevance Check also contains three binary classification questions for a more nuanced relevance analysis: **(1)** does the problem indicated in the correction exist in the essay? **(2)** is the indicated problem related to the given assessment question? and **(3)** is the correction correct?

[4] See Appendix B for additional details and explanations.



Validation Results for the Pipeline

	TP: Num of Correctly Extracted Problems	FP: Num of Incorrectly Extracted Problems	FN: Num of Problems not Extracted
IAA (Cohen's Kappa)	0.72	0.47	0.49
Exact Match Rate	0.78	0.78	0.98
LLM Task Performance	Precision: 0.86; Recall: 0.99; F1: 0.92		
	Specific Part	**Has a Suggestion**	**Has a Correction**
IAA (Cohen's Kappa)	0.90	0.30	0.89
Exact Match Rate	0.95	0.96	0.96
LLM Task Performance (Accuracy / Micro Average F1)	0.87 / 0.87	0.87 / 0.82	0.94 / 0.94
	In Essay	**In Question**	**Is Correct**
"Yes" Rate (Positive Samples)	0.92	0.85	0.89
"No" Rate (Negative Samples)	0.93	0.92	0.92

Figure 2: *Left*: Pipeline of the proposed feedback comment quality evaluation framework. The input and output for each step of the pipeline are illustrated using a human-generated comment on the use of academic vocabulary, with related tasks performed by an LLM. Answers to the 6 classification questions from the last two steps are highlighted in bold. *Right*: Validation results for the pipeline, where IAA (inter-annotator agreement) and exact match rate are measured between raw annotations by two annotators. See Appendix B for further details.

The results show that both human- and LLM-provided corrections are highly relevant, with answers to those three questions being "Yes" typically above 90% time (see Appendix B.3). We thus focus on the Problem Classification results when comparing human- and LLM-identified problems in the subsequent sections.

Validating the Proposed Framework The first author and a paid graduate student in Linguistics (native speaker) first annotated some held-out samples for training and developing the annotation guidelines. Each then independently annotated at least another 200 samples containing human- and LLM-generated comments or problems for Problem Extraction and Problem Classification. Afterward, they met to resolve disagreements before the inter-annotator agreement (IAA) was calculated.

We measure IAA using Cohen's Kappa. As is known (Feinstein and Cicchetti, 1990), Cohen's Kappa can provide misleading values with highly imbalanced class distributions. We therefore also provide exact match rates which have not been corrected for random agreement. Fig. 2 (right) shows that the IAA is typically high. When the Cohen's Kappa is low due to class imbalance (i.e., problems being incorrectly or not extracted is uncommon or rare and nearly all extracted problems contain a suggestion), the exact match rates are high. LLM task performance, evaluated based on the resolved annotations, is also notably high.

We automatically evaluate LLM performance on the Correction Relevance Check by assuming that human-identified corrections are generally relevant. Specifically, we assess whether the LLM classifies these corrections as mostly *relevant* when presented with their corresponding essays and assessment questions (positive samples), and as mostly *irrelevant* when paired with random essays and questions (negative samples). As shown in Fig. 2 (right), our results confirm this expectation.

5 Experiments

This sections describes and presents the main experiments conducted and the results obtained.

5.1 LLM Prompting

List of LLMs We evaluate variants of three popular LLMS: GPT-4O-2024-08-06 (GPT-4o, OpenAI et al., 2024a), GEMINI-1.5-FLASH (Gemini-1.5, Gemini Team et al., 2024), and LLAMA-3 70B-INSTRUCT (Llama-3, Grattafiori et al., 2024).

Default Prompt Setting All prompts contain a system prompt, an input essay, and an assessment instruction. There are four default conditions. **(1)** The system prompt contains not only essential background information, such as writing topic, but also helpful information regarding the L2 nature of the input essay, year of writing, the same general assessment guidance used by human assessors. **(2)** The input essay always includes references. **(3)** LLMs are instructed to produce a score before an optional comment for each assessment question **(4)** via greedy decoding, i.e., with temperature set to 0. Conditions 1-3 are used to maximize the alignment between human and LLM assessment conditions.

Interaction Modes We consider three possible user-LLM interaction modes, depending on how the 9 assessment questions are presented. In Interaction Mode 1 (IM 1), all questions are prompted at once in a single-turn conversation, where all LLM assessments are generated in a single response. In Interaction Mode 2 (IM 2), the questions are asked one at a time, with an LLM generating answers to each question in corresponding turns in a multi-turn conversation. In Interaction Mode 3 (IM 3), however, the assessment questions are provided independently of one another in 9 separate prompts to elicit 9 separate outputs from an LLM.

5.2 Baselines

Given the open-ended nature of the task, we compare raw assessments produced across individual assessors to understand the assessment patterns and behaviors of humans and LLMs. For a more robust statistical analysis, we only consider raw assessments made by assessors B, C, and F, since the essays they each assessed and co-assessed both cover at least half of the corpus.

5.3 Evaluation of Scores

Quadratic Weighted Kappa (QWK) This is a metric for rating inter-rater agreement. It ranges from 0 (random agreement) to 1 (perfect agreement), though it can be negative when agreement is worse than chance. QWK places higher penalties for larger score mismatches, but can yield misleadingly high or low values due to chance correction when the distribution of scores is highly skewed (Yannakoudakis and Cummins, 2015).

Adjacent Agreement Rate (AAR) AAR measures the percentage of scores (from two raters) that lie within a specified threshold k of one another. When $k = 0$, it assesses exact matches. For this study, we set $k = 1$ (AAR1), meaning raters' scores are treated as matching as long as they differ by no greater than 1.

We use AAR1 in addition to QWK to account for the limitation of QWK's chance correction, as we observe that both human- and LLM-assigned scores are highly biased toward the respective means. AAR1 also helps address observed scoring inconsistency issues (often by 1 point) by humans. See Appendix C.1 for more details and discussions.

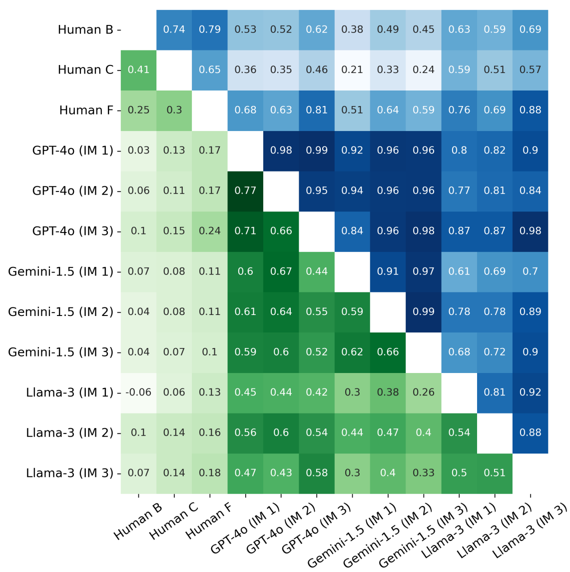


Figure 3: Heatmaps of overall QWK (bottom, green) and AAR1 (top, blue) among assessors. Darker shades indicate a higher degree of agreement.

5.4 Results

We compare human- and LLM-generated assessments in terms of scores, comments, and the interaction between scores and comments.

5.4.1 Scores

Figure 3 illustrates the overall scoring agreement between all pairs of assessors.

Humans score more like humans and LLMs score more like LLMs. More concretely, human-human QWK and AAR1 are almost always higher than the corresponding human-LLM agreement. Similarly, LLM-LLM agreement exceeds human-LLM agreement in virtually all cases, with a much larger margin, suggesting that LLMs may resemble each other in scoring more closely than humans resemble each other. Criterion-level agreement between human/LLM assessors shows similar patterns, as shown in Fig. 4.

LLMs can score approximately like humans. The best human-LLM AAR1 for the three LLMs ranges from 0.59 to 0.88, with all LLMs achieving an AAR1 above 0.5 with assessor F (Fig. 3). Moreover, the AAR1 scores between GPT-4o and assessor B and between Llama-3 and assessors B and C are always greater than 0.5. *Overall, it shows that LLMs generate sensible or reasonably good scores, often differing by no more than 1 point from the corresponding human-generated scores.*

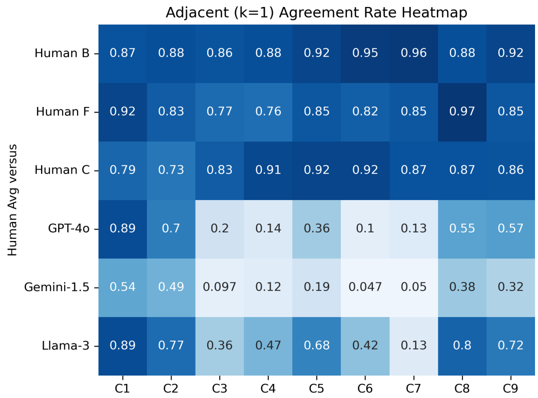


Figure 4: Criterion-level AAR1 between average human scores (“Human Avg”) and human or LLM assessors. See Appendix C.1 for full results for QWK and AAR1.

Assessor	Avg Comment Rate	Avg Comment Len	Avg Problem Rate	Avg Problem Num	Avg Corr Score - Cmt
Human B	0.24	104±85	0.97	3.8±3.5	**-0.20** / -0.17
Human C	1.00	62±85	0.56	1.3±1.8	-0.40 / **-0.46**
Human F	0.90	47±58	0.63	1.3±1.6	-0.37 / **-0.47**
GPT-4o (IM 1)	1.00	65±14	1.00	2.1±0.9	-0.11 / **-0.48**
Gemini-1.5 (IM 1)	1.00	97±33	1.00	2.4±1.00	-0.05 / **-0.46**
Llama-3 (IM 1)	1.00	68±20	1.00	2.2±0.8	0.01 / **-0.27**
GPT-4o (IM 2)	1.00	347±46	1.00	5.0±1.2	-0.37 / **-0.38**
Gemini-1.5 (IM 2)	1.00	477±698	1.00	5.9±2.7	-0.29 / **-0.56**
Llama-3 (IM 2)	1.00	370±112	1.00	6.6±2.8	-0.04 / **-0.42**
GPT-4o (IM 3)	1.00	381±65	1.00	6.1±2.0	-0.34 / **-0.51**
Gemini-1.5 (IM 3)	1.00	571±182	1.00	8.2±3.3	-0.21 / **-0.48**
Llama-3 (IM 3)	1.00	399±67	1.00	6.4±2.3	-0.04 / **-0.23**

Table 2: Overall statistics of feedback comments generated by human and LLM assessors. The last column shows the Spearman Rank correlations measured between scores and related comments (length / number of identified problems). Stronger negative correlations (smaller numbers) in each number pair are in bold.

Human-LLM agreement tends to be higher when LLMs respond to each assessment criterion separately under IM 3. This is particularly true compared to when LLMs respond to all criteria at once under IM 1, since IM 3 exhibits a generally higher agreement level (Fig. 3). This result may imply that, while human assessors score the 9 assessment criteria sequentially, they effectively make independent scoring decisions based on the specifics of each assessment question.

That said, the effect of interaction modes is overall limited, given the fairly close scores (i.e., high QWK/AAR1) assigned across them for each LLM. Therefore, we average human-LLM agreement for each LLM across the three interaction modes to obtain human-LLM agreement in Fig. 4.

The degree of human-LLM agreement varies across assessment criteria. For example, Fig. 4 shows that LLM-assigned scores are relatively closer to human-assigned scores on assessment criteria C1 (material selection), C2 (material integration and citation), C8 (grammar and sentence structure), and C9 (academic vocabulary) than the other criteria. Among criteria C3-C7, LLMs and humans agree rather poorly on C7 (use of connectors), with LLMs consistently assigning scores more than 1 point away from human-assigned ones.

5.4.2 Comments

Table 2 shows the percentage of time an assessor provided a comment, and when they did, the average length of these comments, the percentage of comments identifying a problem, and the average number of problems identified in each comment.

LLMs always provide comments and identify problems, but humans do not. This is an apparent advantage of LLMs since, unlike humans, they do not experience practical constraints like mental fatigue and limited time for writing comments. While humans show different tendencies in comment writing, they tend to write more comments and/or identify more problems (with longer comments) on criteria that are technical and objective, including C2, C8, and C9, also mentioned in the end of Section 5.4.1. See Appendix C.2 for details.

Interacting with LLMs one question at a time leads to more elaborate, specific, and helpful comments. LLM comments are much longer and identify more problems in IM 2 and IM 3 than in IM 1 (see Table 2). Additionally, Fig. 5 shows that comments generated in IM 1 are also less likely to refer to a specific essay part and offer a concrete correction than those generated in IM 2 and IM 3 or human-generated comments. This suggests that IM 2 and IM 3 provide higher levels of elaboration than IM 1. Furthermore, IM 3 produces more corrections than both IM 2 and humans across all assessment criteria, except C1, for which a correction is unlikely since it is about evaluating the relevance of cited references. In other words, *LLMs can be more elaborate, specific, and potentially helpful than humans in their comments.*

LLMs can be more specific than humans on assessing subjective criteria. While humans and LLMs (in IM 3) are comparably likely to include a correction in their comments for objective criteria C2, C8, and C9, LLMs’ comments (in IM 3)

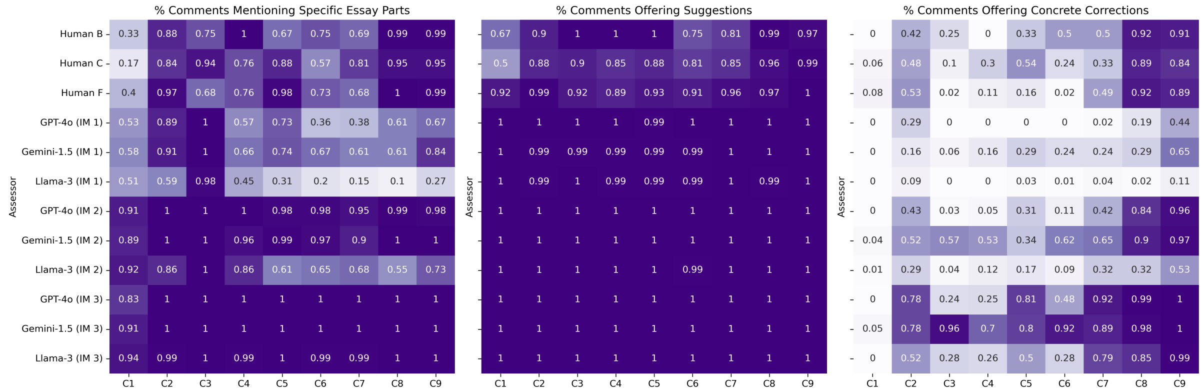


Figure 5: Percentage of comments identifying a problem that mentions a specific essay part (left), offers a comment (middle), and offers a concrete correction (right) across assessment criteria by different assessors.

tend to offer more corrections on other subjective criteria (e.g., C3: quality of key components, C4: logic of structure etc.), except for C1 (see above). This aligns with the observation that humans tend to comment more on objective criteria, since commenting on subjective criteria requires more explanations and can thus be more demanding to do.

5.4.3 Score-Comment Interaction

Since lower scores reflect a perception of more writing problems, an assessor typically needs to provide a more extensive feedback comment to both cover the identified problems and justify their low scores. We highlight this score-comment interaction by measuring the correlations between scores and the token counts of or the numbers of identified problems in the related comments.

As expected, the last column in Table 2 shows strongly negative score-comment correlations across both human- and LLM-generated assessments. The fact that these negative correlations are generally much stronger when measured with the number of identified problems suggests that it is a more fine-grained metric than comment length and also indicates the usefulness of our framework proposed in Section 4.

5.5 Summary

We show that LLMs can generate sensible scores, typically within 1 point of human-generated ones on a 10-point scale, and feedback comments that identify more writing problems than human assessors that are specific, and potentially helpful. This is particularly true when LLMs are prompted in IM 3 where each assessment question is asked independently of each other. Moreover, like humans, LLMs also generate assessments that exhibit an expected and negative score-comment correlation, justifying the validity of their assessments. *Overall, these results highlight that LLMs can generate reasonably good multi-dimensional analytic assessments.*

6 Further Analyses

6.1 Re-examining Our Assumption about Feedback Comment Quality

Our proposed feedback comment quality evaluation framework assumes that the quality of a feedback comment is related to how well it identifies relevant writing problems of an assessed essay. The framework extracts and characterizes problems of assessed essays identified in comments to evaluate the specificity and helpfulness of these comments.

To assess this assumption, we adopt an LLM-as-a-judge approach (Zheng et al., 2023), prompting OPENAI-O1-MINI-2024-09-12 (o1-mini, OpenAI et al., 2024b) to directly assess the specificity and helpfulness of a feedback comment, given the corresponding essay and assessment question on a 10-point scale. We do not define specificity and helpfulness to avoid injecting biases and choose all comments, generated by humans and LLMs, from one subjective criterion (C6: coherence or flow of ideas) and one objective criterion (C9: academic vocabulary) to balance our examination. We then calculate the corrections between these two scores produced by o1-mini and the number of different types of problems identified by our framework.

The results in Table 3 shows that the characteristics extractable from applying the framework correlate very well with the o1-mini-assigned specificity and helpfulness scores. In particular, the number of problems that mention specific essay parts and offer corrections appears to be overall stronger sig-

Condition		#Problems	#Specific	#Corrections
Humans	Specificity	0.57	0.66	0.63
	Helpfulness	0.65	0.70	0.62
LLMs	Specificity	0.62	0.80	0.61
	Helpfulness	0.64	0.77	0.58
C6	Specificity	0.68	0.78	0.51
	Helpfulness	0.72	0.74	0.48
C9	Specificity	0.59	0.79	0.77
	Helpfulness	0.64	0.76	0.74
IM 1	Specificity	-0.02	0.63	0.43
	Helpfulness	-0.03	0.50	0.44
IM 2	Specificity	-0.02	0.63	0.43
	Helpfulness	0.09	0.48	0.38
IM 3	Specificity	0.22	0.33	0.31
	Helpfulness	0.23	0.30	0.24

Table 3: Spearman Rank correlations between the specificity and helpfulness scores and the number of different types of problems identified by our framework under different conditions. Corrections with number of problems making a suggestion are omitted as they are nearly identical to those with "#Problems."

	Scores	Comments
GPT-4o-May	0.82 / 0.98	0.21 / 0.39 / 0.70
SP Simplification	0.78 / 0.98	0.24 / 0.43 / 0.72
Exclusion of References	0.69 / 0.95	0.26 / 0.44 / 0.73
Comment First	0.75 / 0.96	0.19 / 0.32 / 0.58
Temperature=1, run#1	0.73 / 0.96	0.10 / 0.30 / 0.67
Temperature=1, run#2	0.79 / 0.98	0.10 / 0.31 / 0.67
GPT-4o-May (IM 2)	0.81 / 0.99	0.15 / 0.29 / 0.70
GPT-4o-May (IM 3)	0.83 / 1.00	0.20 / 0.31 / 0.71
Llama3: SP Simplification	0.66 / 0.88	0.25 / 0.44 / 0.73
Llama3: Exclusion of Refs	0.71 / 0.90	0.25 / 0.44 / 0.74
Llama3: Comment First	0.51 / 0.81	0.24 / 0.44 / 0.72

Table 4: Reliability tests results. "QWK / AAR1" and "BLEU / ROUGE-L / BERTScore" are used to measure score stability and comment similarity, respectively.

nals of specificity and helpfulness than the mere number of problems, which shows negligible correlations for comments from IM 1 or IM 2. This shows the potential of our framework in providing a more fine-grained and interpretable measurement of specificity and helpfulness levels of comments.

6.2 Reliability of LLM-generated Assessments

We evaluate the reliability of LLM-generated assessments across different realistic conditions that mirror potential real-world use cases. To prevent experimental confounding, we change only one condition at a time for a given LLM in a specific interaction mode, assuming that users tend to interact with their chosen LLM in a consistent manner.

First, we consider GPT-4O-2024-08-06 (GPT-4o-Aug) in IM 1 with the default prompt setting from Section 5.1 as the baseline. To test the effect of model variant, we run the same experiment but with GPT-4O-2024-05-13 (GPT-4o-May). We also prompt GPT-4o-Aug while varying one of the four conditions in the default prompt setting (see Section 5.1) by (1) removing the helpful information from the system prompt, (2) excluding references in the input essays, (3) instructing LLMs to produce a comment before a score, or (4) setting temperature to 1 to increase output randomness.

To ensure the comprehensiveness of our experiments, we prompt GPT-4o-May in IM 2 and IM 3 under default prompt setting to study the effect of model variant under other interaction modes. We also prompt Llama-3 in IM 1 changing the first three conditions in the default prompt setting mentioned in the last paragraph. The baselines here are GPT-4o-Aug and Llama-3 prompted under respective interaction modes from Section 5.1.

We use QWK and AAR1 and three widely adopted machine translation metrics, i.e., BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BERTScore (Zhang et al., 2020), to evaluate the reliability of the generated scores and comments between contrastive condition pairs, respectively.

The results in Table 4 show that LLMs are capable of generating highly stable scores, with an AAR1 score at least 0.81 and mostly above 0.9 across all conditions. Their generated comments are also decently similar with BERTScore typically no lower than 0.67. A small-scale manual check and a correlation analysis performed in Appendix D further verify the validity of BERTScore.

7 Conclusion

This study provides evidence that LLMs can generate reasonably good and generally reliable multi-dimensional analytic assessments. Our findings highlight the promising role of LLMs in assessing academic English writing, especially for graduate-level literature reviews, which is a highly technical genre. In short, LLMs show strong pedagogical potential, benefiting both L2 learners and instructors for self-regulated learning or teaching assistance. We propose and validate a novel feedback comment quality framework to facilitate our analysis.

Looking ahead, future studies could further characterize and compare the writing problems identified by human- versus LLM-generated comments, offering deeper qualitative insights. Additionally, it would be valuable to develop a metric grounded in our proposed framework that can directly compare the relative quality of two sets of comments.

Limitations

Generality of Findings This study focuses on L2 graduate-level academic writing, specifically literature reviews in the humanities and social sciences. While this domain represents a significant subset of academic writing, the findings may not generalize to other genres (e.g., technical reports, creative writing) or proficiency levels (e.g., undergraduate or professional writers). Additionally, our study is limited to English, a high-resource language, which means our results may not be indicative of LLMs' capabilities in other languages, particularly low-resource ones. Future research should explore the applicability of our findings across diverse writing contexts and linguistic backgrounds.

Weakness of Our Assumption About Feedback Quality A key limitation of our approach is that it does not account for other factors that may influence the *perceived* quality of a feedback comment, such as politeness (e.g., rude comments may not be well received) or the logical coherence of the argument (e.g., illogical comments could be misleading). However, this concern is less pronounced for LLM-generated feedback comments, as LLMs are trained to align with human preferences and social norms (Ouyang et al., 2022). Moreover, these factors could potentially be incorporated into our framework by adding additional steps focused on politeness and argumentation etc.

Indirect Evaluation of Feedback Quality While our approach to measuring the general quality of LLM-generated assessments is intuitive and simple, it is inherently indirect. A large-scale manual evaluation remains necessary to more accurately assess and compare the quality of human- and LLM-generated multi-dimensional analytic assessments. Due to resource constraints, we leave this investigation to future studies.

Limited Validation and Reliability Testing Due to time and resource constraints, we were unable to comprehensively validate our proposed feedback comment quality evaluation framework. As a result, we may have overlooked some potential issues with the framework or the LLM outputs. Similarly, the reliability assessments we conducted are limited, with only one factor being changed at a time in each evaluation. More extensive experiments are needed to further validate our claim that LLM-generated assessments are generally reliable and to explore the conditions influencing this reliability.

Ethical Considerations

Corpus Creation The research project that led to the construction of the corpus was ethically reviewed and received approval from a Canadian institution for involving human participants. Participants provided informed consent to allow the use of their materials, with the option to withdraw at any time.

Human Annotations We compensated the hired annotator at a rate of approximately US$25 per hour, which exceeds the minimum wage in the region where the annotations took place.

Potential Biases in LLM Assessments LLMs are trained on large-scale datasets that may contain inherent biases, which can be reflected in their assessments. For example, they might systematically favor certain writing styles, linguistic structures, or cultural conventions, leading to biased evaluations. However, we argue that in contexts where human assessments are not readily accessible, the benefits of LLM-generated feedback – particularly for L2 learners – may outweigh potential biases. Furthermore, bias mitigation strategies, such as improved prompting techniques or advancements in LLM development, could help reduce these concerns.

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany. Association for Computational Linguistics.

Nikolay Babakov, Maria Lysyuk, Alexander Shvets, Lilya Kazakova, and Alexander Panchenko. 2023. Error syntax aware augmentation of feedback comment generation dataset. In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 37–44, Prague, Czechia. Association for Computational Linguistics.

Stefano Banno, Hari Krishna Vydana, Kate Knill, and Mark Gales. 2024. Can GPT-4 do L2 analytic assessment? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 149–164, Mexico City, Mexico. Association for Computational Linguistics.

Shabnam Behzad, Omid Kashefi, and Swapna Somasundaran. 2024a. Assessing online writing feedback resources: Generative AI vs. good samaritans. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources* *and Evaluation (LREC-COLING 2024)*, pages 1638–1644, Torino, Italia. ELRA and ICCL.

Shabnam Behzad, Omid Kashefi, and Swapna Somasundaran. 2024b. LEAF: Language learners' English essays and feedback corpus. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 433–442, Mexico City, Mexico. Association for Computational Linguistics.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Hongwen Cai. 2015. Weight-based classification of raters and rater cognition in an efl speaking test. *Language Assessment Quarterly*, 12(3):262–282.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kai Guo and Deliang Wang. 2024. To resist it or to embrace it? examining ChatGPT's potential to support teacher feedback in EFL writing. *Educ. Inf. Technol.*, 29(7):8435–8463.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. 2024. LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-LLM interaction. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 284–293, Miami, Florida, USA. Association for Computational Linguistics.

Wen-Bin Han, Jhih-Jie Chen, Chingyu Yang, and Jason Chang. 2019. Level-up: Learning to improve proficiency level of essays. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 207–212, Florence, Italy. Association for Computational Linguistics.

Nicky Hockly. 2019. Automated writing evaluation. *ELT Journal*, 73:82–88.

Mohamed Abdellatif Hussein, Hesham Hassan, and Mohammad Nassef. 2019. Automated language essay scoring systems: a literature review. *PeerJ Computer Science*, 5:e208.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

You-Jin Jong, Yong-Jin Kim, and Ok-Chol Ri. 2023. Review of feedback in automated essay scoring. *Preprint*, arXiv:2307.05553.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, IJCAI-2019, page 6300–6308. International Joint Conferences on Artificial Intelligence Organization.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Atsushi Mizumoto, Natsuko Shintani, Miyuki Sasaki, and Mark Feng Teng. 2024. Testing the viability of chatgpt as a companion in l2 writing accuracy assessment. *Research Methods in Applied Linguistics*, 3(2):100116.

Ryo Nagata. 2019. Toward a task of feedback comment generation for writing learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215, Hong Kong, China. Association for Computational Linguistics.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024b. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. 2019. Language models and automated essay scoring. *Preprint*, arXiv:1909.09482.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Mark D Shermis and Jill C Burstein, editors. 2013. *Handbook of automated essay evaluation*. Routledge, London, England.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2024. GEE! grammar error explanation with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 754–781, Mexico City, Mexico. Association for Computational Linguistics.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Masaki Uto. 2021. A review of deep-neural automated essay scoring models. *Behaviormetrika*, 48(2):459–484.

Bertus van Rooy. 2015. *Annotating learner corpora*, page 79–106. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Qiao Wang and John Maurice Gayed. 2024. Effectiveness of large language models in automated evaluation of argumentative essays: finetuning vs. zero-shot prompting. *Computer Assisted Language Learning*, page 1–29.

Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2724–2733, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584, Toronto, Canada. Association for Computational Linguistics.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569, Online. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

	T1	T2	T3	T4	T5
# Essays	50	16	31	13	31
Avg WC (w/o refs)	845	1169	926	1079	887
Avg WC (w/ refs)	1232	1583	1347	1666	1159

Table 5: Basic statistics of the corpus. "T" in each column stands for "Topic." "WC" means "word count."

Helen Yannakoudakis and Ronan Cummins. 2015. Evaluating the performance of automated text scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–223, Denver, Colorado. Association for Computational Linguistics.

Helen Yannakoudakis, Øistein E Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for esl learners. *Applied Measurement in Education*, 31(3):251–267.

Fatih Yavuz, Özgür undefinedelik, and Gamze Yavaş undefinedelik. 2024. Utilizing large language models for efl essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*.

Ruofei Zhang and Di Zou. 2020. Types, purposes, and effectiveness of state-of-the-art technologies for second and foreign language learning. *Computer Assisted Language Learning*, 35(4):696–742.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Corpus

A.1 Basic Corpus Statistics

Table 5 provides the basic statistics of the corpus. Note that throughout this study, we use the default word tokenizer of NLTK to compute word counts. See: `https://www.nltk.org/api/nltk.tokenize.html`.

A.2 Details of the 5-Unit Tutorial Series

Table 6 presents details of the 5-unit tutorial series, including the themes, notions, activities, duration, and writing task for each unit.

To support their writing, the authors were provided with a short, curated bibliography for each task, designed to help them focus on literature review writing while minimizing the effort required for bibliographic searches. Prior to submitting their final writing samples for expert assessments, the authors engaged in peer reviews (for topics 1, 3, and 5) or group collaboration (for topics 2 and 4).

A.3 Assessment Criteria

The 9 assessment criteria/questions provided to human assessors are detailed in Table 7.

B Feedback Comment Quality Evaluation Framework

B.1 Implementation

The framework is implemented using LLMs. More concretely, we used GPT-4O-2024-11-20 for Problem Extraction and Problem Classification, and GPT-4-TURBO-2024-04-09 for Correction Relevance Check. An example implementation of our framework can be found in Table 8.

Related prompts used for implementing our framework can be found in Appendix E.1.

B.2 Annotation

Guidelines Table 9 provides explanations and examples of what is considered as a problem for Problem Extraction, and the three characteristics relevant to Problem Classification: whether an extracted problem (1) refers to a specific part of the essay, (2) provides a suggestion (general or specific), and (3) offers a concrete correction.

Samples for Problem Extraction We employed stratified sampling to randomly select 100 human-generated feedback comments and 108 LLM-generated feedback comments. In total, there are 208 comments for manual annotations.

For LLM-generated comments, half of them were generated under Interaction Mode 1 and the other half under Interaction Modes 2 and 3. Comments from Interaction Modes 2 and 3 were sampled together to reduce manual annotation effort, as these comments tend to be lengthy. The sampling covered the 9 assessment criteria, with 2 comments from each of the 3 LLMs used, resulting in 9 * 3 * 2 = 54 comments from Interaction Mode 1 and another 54 comments from the combined Interaction Modes 2 and 3.

Unit	Key notions	Activities	Duration	Writing task
1. Genre of literature review	Components in literature review writing, material selection, citation practices	Interactive e-book, Peer-review, Discussion forum, quiz	3 weeks	Individual writing on the social consequences of legalized cannabis
2. Structure and logic in literature review	Types of logic structure, terms and abbreviations, Coherence, Cohesion	Interactive e-book, Discussion forum, quiz	2 weeks	Collaborative writing on Canadian linguistic landscape
3. Sentence structures	Sentence structures and variety, nominalization, Phrase bank and Swales' CARS (Creating a Research Space) model	Interactive e-book, Peer-review, Discussion forum, quiz	3 weeks	Individual writing on the pros and cons of online learning
4. Academic vocabulary	(academic) formulaic expressions and their functions	Interactive e-book, Discussion forum, quiz	2 weeks	Collaborative writing on lessons from the COVID- 19 pandemic
5. Grammar of reported speech	Direct vs. indirect speech, reporting verbs and expressions, verb tenses, modal verbs	Interactive e-book, Peer-review, Discussion forum, quiz	3 weeks	Individual writing on pacifism, peace-making, or just/justifiable war

Table 6: Details of the 5-unit online tutorial series.

Aspect	Criterion	Question
Selection of materials and citation practices	1. Material selection	On a scale of 10 (1: Very poor, 10: Excellent), how would you evaluate the author's selection of source materials in terms of relevance, quality, and quantity of the materials? Note: "If the draft has a noticeable issue regarding the number or the quality of the papers reviewed, please comment on the issue."
	2. Material integration and citation	On a scale of 10 (1: Very poor, 10: Excellent), how would you evaluate the writing for its integration of source materials (e.g., clarity of presenting information) and citation practices (e.g., use of APA or other style in both in-text citations and reference list)?
Overall structure	3. Quality of key components	On a scale of 10 (1: Very poor, 10: Excellent), how would you evaluate the writing for the quality or effectiveness of each component (i.e., Introduction, Body, and Conclusions)? Note: The introduction is expected to introduce a research area, identify issue(s), and/or state the significance of the issue(s). The body of literature review should present the relevant ideas or findings of the reviewed studies and/or identify a research gap. The conclusion(s) may identify research trends or controversies and highlight the contribution of this literature review.
	4. Logic of structure	On a scale of 10 (1: Very poor, 10: Excellent), how would you evaluate the logical structure of this literature review?
	5. Content and clarity of ideas	On a scale of 10 (1: Very poor, 10: Excellent), how would you evaluate the content and clarity of ideas expressed in this literature review?
Coherence and cohesion	6. Coherence	On a scale of 10 (1: Very poor, 10: Excellent), how would you evaluate the literature review for the quality of coherence (e.g., the connectivity and the naturalness of the flow of ideas in this draft)?
	7. Cohesion	On a scale of 10 (1: Very poor, 10: Excellent), how would you evaluate the literature review for the use of connectors (e.g., 'because,' 'therefore,' 'however,' 'likewise', and 'similarly') to link sentences in this draft?
Grammar and vocabulary	8. Grammatical and sentence structure	On a scale of 10 (1: Very poor, 10: Excellent), how would you evaluate the draft for grammatical accuracy, sentence length and sentence type variety?
	9. Academy vocabulary	On a scale of 10 (1: Very poor, 10: Excellent), how would you evaluate the draft for vocabulary quality (e.g., use of academic expressions, the correctness of word choice, the naturalness of collocations, the complexity of vocabulary, the use of stylistically acceptable vocabulary—not too colloquial, not excessively formal or not overusing terms)?

Table 7: The 9 assessment criteria/questions, reflecting 4 general aspects of writing quality.

Comment	Problem Extraction	Problem Classification	Correction Relevance Check
The author has generally done a good job of integrating the source materials into the text, with clear summaries and explanations of the findings. However, there are some areas where the citation practices could be improved. For example, some of the in-text citations are not formatted correctly (e.g., "Wilkinson ST, et al., 2015" should be "Wilkinson et al., 2015"), and there are some inconsistencies in the reference list (e.g., some sources have DOIs, while others do not). Additionally, the author could benefit from using more precise language when describing the findings of the studies, rather than relying on general statements.	There are some areas where the citation practices could be improved. For example, some of the in-text citations are not formatted correctly (e.g., "Wilkinson ST, et al., 2015" should be "Wilkinson et al., 2015").	1. Yes. The excerpt refers to a specific part of the essay by mentioning the in-text citation "Wilkinson ST, et al., 2015" and how it should be formatted. 2. Yes. The excerpt offers a suggestion by indicating that the in-text citation formatting should be corrected. 3. Yes. The excerpt provides a concrete correction by specifying the correct format for the citation as "Wilkinson et al., 2015." Final answers: Yes, Yes, Yes	1. Yes, the problem pointed out in the excerpt exists in the corresponding essay. The excerpt mentions that the citation "Wilkinson ST, et al., 2015" is incorrectly formatted and should be "Wilkinson et al., 2015". In the essay, the citation appears as "Wilkinson ST, et al., 2015", which is indeed not formatted correctly according to APA style guidelines that recommend omitting the first names and middle initials of authors in in-text citations. 2. Yes, the problem pointed out in the excerpt is relevant to the corresponding assessment question. The assessment question asks to evaluate the essay based on the integration of source materials and citation practices. The excerpt directly addresses an issue with citation practices, which is a key aspect of the assessment question. 3. Yes, the correction of the problem pointed out in the excerpt is correct. The suggested correction to change "Wilkinson ST, et al., 2015" to "Wilkinson et al., 2015" aligns with APA style guidelines, which recommend using only the surname of the first author followed by "et al." for works authored by multiple people, without including initials unless necessary to distinguish between authors with the same surname. Final answers: Yes, Yes, Yes
	There are some inconsistencies in the reference list (e.g., some sources have DOIs, while others do not).	1. No - The excerpt does not refer to a specific part of the essay, as it does not point to a particular location or provide a specific example from the reference list. 2. Yes - The excerpt suggests that the student should address the inconsistencies in the reference list, such as ensuring uniformity in the inclusion of DOIs. 3. No - The excerpt does not provide a concrete correction, as it does not specify exactly how to fix the inconsistencies or what the corrected references should look like. Final answers: No, Yes, No	
	The author could benefit from using more precise language when describing the findings of the studies, rather than relying on general statements.	1. No - The excerpt does not refer to a specific part of the essay. It does not mention a concrete location or provide any referencable details about where the issue occurs. 2. Yes - The excerpt offers a general suggestion for improvement by advising the author to use more precise language when describing the findings of the studies. 3. No - The excerpt does not provide a concrete correction or specific example of how to fix the issue. Final answers: No, Yes, No	

Table 8: An example of implementing our framework on a feedback comment on the second assessment criteria, i.e., material integration & citation, which is about citation practices. Our framework correctly identifies three problems identified in the comment and performs Problem Classification on each one of them, of which the first problem offers a concrete correction. The Correction Relevance Check is thus only performed on the first extracted problem.

Samples for Problem Classification We randomly sampled 100 problems extracted from both human- and LLM-generated comments, resulting in 200 problems for annotations.

Since the distribution of extracted problems across the nine assessment criteria are highly skewed, we ensured that there were at least 5 problems for each assessment criterion.

Problem Extraction For each feedback comment, the two annotators were provided with LLM-extracted problems and asked to identify the number of correctly extracted problems (true positives), the number of incorrectly extracted problems (false positives), and the number of problems not extracted (false negatives). The number of true negatives is always set to 0, as there is no negative prediction in problem extraction.

A problem is considered correctly extracted if the LLM output contains the exact or paraphrased problem stated or implied in the feedback comment. It is acceptable if additional information relevant to the problem, such as elaborations, suggestions, clarifying questions, or quoted text from the assessed essay, is not included in the LLM-identified problems, which appears to be uncommon based on our annotations. However, if the problem and relevant additional information are extracted as separate problems, only the stated or implied problem is counted as a true positive, and the relevant information is treated as a false positive. This over-segmentation is the primary source of errors in LLM-extracted problems.

Problem Classification For each extracted problem, the two annotators were asked to answer the three classification problems based on Table 9.

B.3 Correction Relevance Check

Table 10 demonstrates that comments generated by both humans and LLMs are overall highly relevant. However, human-generated comments tend to exhibit slightly lower relevance—either broadly or strictly—compared to those generated by LLMs.

We conducted a small-scale error analysis to investigate the reasons behind the 8%, 15%, and 9% of human-identified problems that GPT-4 incorrectly classified as not present in the essays, not adhering to the assessment criteria, and being incorrect, respectively.

Problems not Present in Essays We randomly selected 10 problems identified by GPT-4 as not present in the assessed essays. Upon reviewing each human-identified problem in the original essay, we found that 6 of these problems were indeed present, while 4 were not. Of the 4 problems that did not exist in the essays, 3 appeared to be misassigned comments (2 of these 3 were extracted from the same comment), while the remaining one seemed to be an assessor error. Among the 6 problems that GPT-4 misclassified, 4 were due to GPT-4 misunderstanding the identified problems, 1 was due to GPT-4 failing to locate a quoted word in the essay, and 1 was because GPT-4 mistakenly deemed the identified problem not to be a problem, despite its presence in the essay.

Problems not Adherent to the Assessment Criteria We randomly selected 10 problems identified by GPT-4 as not adhering to the assessment criteria. Of these, 9 were related to C8 (grammar & sentence structure), and 1 was related to C9 (academic vocabulary). Our manual validation showed that 7 of the problems were less related to grammar and sentence structure but more related to word choice or clarity of expression. The remaining 3 were misclassified by GPT-4, mostly due to its requirement that problems be explicitly related to both grammar and sentence structure in order to adhere to C8.

Correction being Incorrect We randomly selected 10 problems containing corrections identified by GPT-4 as incorrect. We found that 5 of these problems involved accurate corrections, all related to grammar. There were 2 corrections proposed to be suggestions and 3 corrections that require subjective judgments to determine their correctness.

Remarks Based on this error analysis, we can attributed the discrepancy in relevance to two primary reasons: (1) human comments often include (inconsistent use of) diacritics that complicate problem extraction and characterization, and (2) human assessors may occasionally deviate from instructions, providing corrections unrelated to the assessment question. These issues are less frequent in LLM-generated comments, which benefit from their strong adherence to instructions and the ability to handle extended context windows. That said, both human- and LLM-identified problems are highly relevant.

Characteristic	Explanation	Examples
If a problem is stated or implied in a comment	A problem is any writing-related issue that affects the quality of the writing, such as citation errors, logical flaws, coherence issues, grammatical mistakes, or inappropriate word choices, among others. The problem can be mentioned or implied in a given comment.	Positive Examples • Specify what the abbreviation "THC" stands for. (Implied problem: "THC" is unspecified) • There was a redundant use of "the legalization of cannabis". Negative Examples • Great grammatical skills, well done! • Final references are well formatted. In-text references are well integrated.
If a problem points to a specific part of the essay	A specific part refers to a part of the essay that is easily locatable. (1) It can be a specific word, phrase, sentence, paragraph, reference etc. used in the essay. (2) It can be a concrete location, such as "sentence 2 in paragraph 2," "in paragraph 6," "the first citation," or "the first sentence of the paper" and so on. (3) A less concrete location, such as "the introduction," or "the conclusion," is also considered a specific part if it is accompanied by some referenceable details.	Positive Examples • In Paragraph 2, the word "decay" is likely a mistake and should be replaced with "decade". • The sentence "This theory still is under debate even with many authors provide a justification for that" contains a grammatical error. The verb "provide" should be corrected to "providing." Negative Examples • Some of the sentences are a bit too long and fall apart a little. • Your paper would benefit from the use of expressions such as "as a result" or "the result" where cause and consequence are important.
If a problem offers some form of suggestions, general or specific	A suggestion indicates or implies ares of improvement. If the problem only contains a problem statement and it is unclear what direction the student should take to improve it, then there is no suggestion. A concrete correction is always considered a suggestion.	Positive Examples • Some sentences could be a bit shorter. • The use of a topic sentence for each paragraph in the main body could be improved. Negative Examples • The beginning of the literature review could be changed slightly. • The first sentence of the paper is confusing.
If a problem provides a concrete correction for an identified writing issue	A concrete correction is something that can be directly applied to an essay to fix a writing problem. Corrections should not require thinking to implement, i.e. text that can be copy-pasted, or actions that can be taken following an instruction (e.g., capitalize the first letter).	Positive Examples • The citation "(Toronto Star December 2016)" should be revised to "(Toronto Star 2016)" to align with proper citation practices. • "The advance of technologies" should be corrected to "the advancement of technologies". Negative Examples • The significance of South Australian policy is unclear, as it is the first citation and the only one in the Introduction. • The conclusion is a little too short.

Table 9: Explanations and illustrative examples of "problems" and their characterizations.

Assessor	In Essay	In Question	Is Correct	Broadly Relevant	Strictly Revelant
Human B	87.9	79.4	85.1	84.4	72.8
Human C	94.9	91.8	94.5	93.8	89.0
Human F	96.3	86.7	91.4	90.9	82.3
GPT-4o (IM 1)	100.0	100.0	100.0	100.0	100.0
Gemini-1.5 (IM 1)	95.6	99.6	98.0	95.6	95.2
Llama-3 (IM 1)	97.8	97.8	97.8	97.8	97.8
GPT-4o (IM 2)	99.6	100.0	100.0	99.6	99.6
Gemini-1.5 (IM 2)	98.3	98.8	97.5	97.1	96.6
Llama-3 (IM 2)	94.7	96.2	96.2	94.4	92.5
GPT-4o (IM 3)	100.0	99.5	99.8	99.8	99.2
Gemini-1.5 (IM 3)	98.8	97.8	99.0	98.8	96.8
Llama-3 (IM 3)	98.7	98.7	98.5	98.5	97.5

Table 10: Overall Correction Relevance Check results (%), representing the percentage of instances each attribute is true for corrections made by an assessor. "In Essay": whether the problem indicated in the correction exists in the essay. "In Question": whether the correction relates to the assessment question. "Is Correct": whether the correction is correct. "Broadly Relevant": applicable when both "In Essay" and "Is Correct" are true. "Strictly Revelant": applicable when both "Broadly Relevant" and "In Question" are true.

C Results

C.1 Scores

Scoring Ranges Table 11 summarizes the scoring ranges, in the form of means and standard deviations for each assessment criterion, as produced by three human assessors and the three LLMs under three interaction modes.

Full QWK/AAR1 Table 12 presents the full results for Quadratic Weighted Kappa (QWK) and Table 13 presents the full results for AAR1.

Inconsistencies in Scoring by Human Assessors First, there is an instance in the corpus, where assessor B accidentally assessed the same essay twice on separate days.[5] While assessor B provided identical scores for 5 out of the 9 assessment criteria, discrepancies of 1 point occurred for the remaining 4 criteria, with scores alternating between (8, 7), (8, 7), (4, 5), and (7, 8).

Second, we observe that human assessors assigned different scores to identical or similar comments, mostly within 1-point differences. For example, assessor F gave the same comment "Decent number of citations" three times but assigned three different scores: 6, 7, and 8. Similarly, assessor C assigned scores of 7 and 8 to the comment "Appropriate use of connectors." However, when the same comment is repeated, scores tend to be very close, typically within one point. For instance, assessor A assigned a score of 8 to the comment "Great use of academic words and formal tone" five times, with only one instance where the score was 9.

[5]Four days apart and assessor B had no access to their earlier assessments.

C.2 Comments

Table 14 presents the general statistics of feedback comments generated by human assessors and LLMs under the three interaction modes.

C.3 Score-Comment Interaction

Fig. 6 provides the full results of the correlations measured between scores and the token counts of or the numbers of identified problems in the related comments.

D Further Analyses

Table 15 provides five random example comment pairs sampled from GPT-4o-Aug and GPT-4o-May prompted under default prompt setting specified in Section 5.1. We find that when BERTScore is low (the last row), the comment pair is less similar compared to other pairs. While other two metrics (BLEU and ROUGE-L) are highly correlated with BERTScore (BLUE: 0.78, ROUGE-L: 0.88, Pearson), they consistently yield lower values than BERTScore. This indicates that these two lexical overlap-based metrics may be less effective at measuring comment reliability compared to the semantic similarity captured by BERTScore.

E Prompts

Note that, any word followed by a dollar sign "$" is a placeholder for all prompt templates included

Assessor	C1	C2	C3	C4	C5	C6	C7	C8	C9
Human B	6.7±0.9	6.5±1.2	7.5±1.2	7.7±1.1	7.7±1.1	7.6±1.1	7.3±1.1	7.2±1.1	7.5±1.1
Human C	7.8±1.3	7.6±1.3	7.9±1.0	7.8±1.3	7.8±1.1	7.9±1.1	8.1±0.9	7.7±1.1	8.2±0.9
Human F	7.0±1.0	6.6±1.0	6.9±0.9	7.0±0.8	7.1±0.8	7.1±0.8	7.2±0.8	7.3±0.7	7.0±0.8
GPT-4o (IM 1)	7.4±0.7	6.4±0.7	5.7±0.8	5.7±0.9	6.3±0.7	5.4±0.7	5.5±0.8	6.4±0.9	6.7±0.8
GPT-4o (IM 2)	6.9±0.7	6.0±0.8	6.0±0.8	5.6±1.1	6.2±0.8	5.4±0.9	4.9±0.8	6.2±0.9	6.8±0.9
GPT-4o (IM 3)	6.9±0.7	6.4±0.7	6.0±0.7	6.2±0.7	6.4±0.6	6.1±0.7	6.0±0.7	6.7±0.7	6.8±0.6
Gemini-1.5 (IM 1)	6.3±0.8	5.4±0.7	5.5±0.7	5.5±1.0	6.0±0.8	4.9±0.8	4.5±0.9	5.7±0.8	6.1±0.8
Gemini-1.5 (IM 2)	6.4±0.6	6.3±0.9	5.5±0.7	5.8±0.8	6.0±0.5	5.4±0.7	5.2±0.8	6.4±0.6	6.5±0.6
Gemini-1.5 (IM 3)	6.4±0.6	5.8±0.6	5.5±0.6	5.6±0.5	5.7±0.5	5.5±0.6	5.4±0.5	6.0±0.6	6.1±0.6
Llama-3 (IM 1)	7.5±0.5	7.4±0.7	6.4±0.9	6.4±1.2	7.1±0.7	6.2±0.8	5.2±0.7	7.8±0.5	7.1±0.7
Llama-3 (IM 2)	7.2±0.6	6.8±1.0	6.1±1.1	6.4±1.4	6.7±1.1	6.2±1.4	4.9±1.4	7.3±0.9	7.2±0.8
Llama-3 (IM 3)	7.2±0.5	6.9±0.5	6.4±0.6	6.7±0.6	6.8±0.4	6.7±0.6	5.9±0.6	6.8±0.4	6.8±0.5

Table 11: Means and standard deviations of scores assigned by three human assessors and three LLMs prompted under three interaction modes (IM), denoted by "IM" in parentheses. C1: Material selection. C2: Material integration and citation; C3: Quality of key components. C4: Logic of structure. C5: Content and clarity of ideas. C6: Coherence (flow of ideas). C7: Cohesion (use of connectors). C8: Grammar and sentence structure. C9: Academic vocabulary.

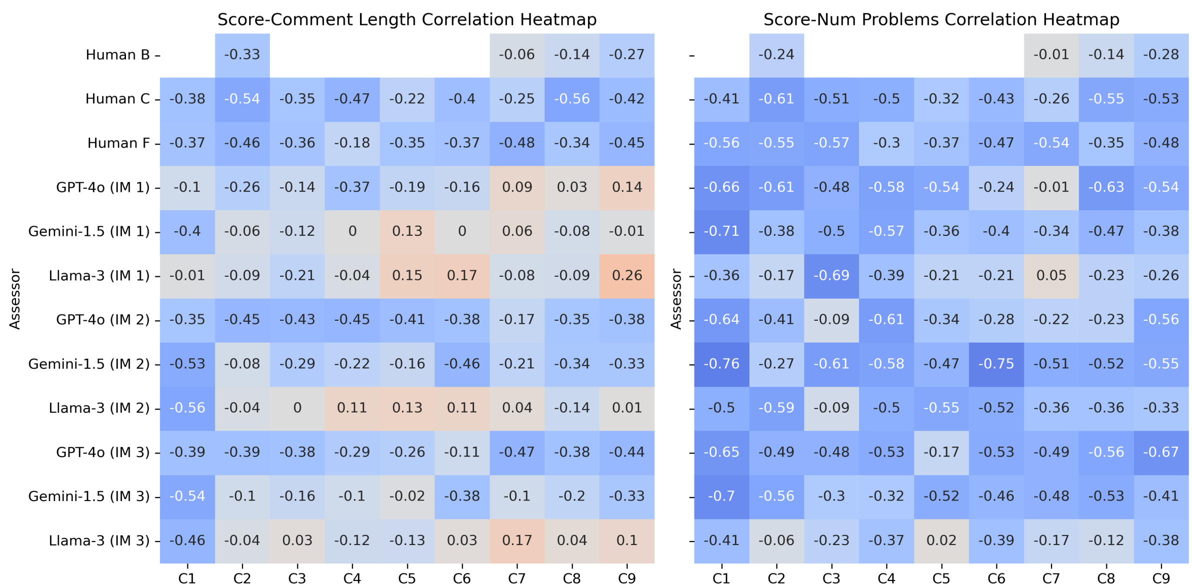


Figure 6: Heatmaps showing score-comment correlations between scores and the length of the related comments (left) and between scores and the number of problems identified in the related comments (right). Darker blue shades indicate a stronger negative correlation and darker orange shades a stronger positive correlation, with gray-ish colors indicating negligible correlations. To ensure meaningful analysis, correlations are calculated only when at least 10 score-comment pairs are available. C1: Material selection. C2: Material integration and citation; C3: Quality of key components. C4: Logic of structure. C5: Content and clarity of ideas. C6: Coherence (flow of ideas). C7: Cohesion (use of connectors). C8: Grammar and sentence structure. C9: Academic vocabulary.

Assessor	C1	C2	C3	C4	C5	C6	C7	C8	C9	Overall
Human B vs. Human F	0.36	0.32	0.18	0.12	0.11	0.09	0.20	0.24	0.26	0.25
Human B vs. Human C	0.41	0.39	0.34	0.36	0.43	0.51	0.40	0.36	0.43	0.41
Human F vs. Human C	0.52	0.29	0.29	0.24	0.23	0.24	0.17	0.28	0.20	0.30
Human B vs. GPT-4o (IM 1)	0.23	0.29	0.08	0.06	0.11	0.05	0.06	0.25	0.10	0.03
Human B vs. GPT-4o (IM 2)	0.33	0.20	0.08	0.05	0.12	0.05	0.06	0.20	0.15	0.06
Human B vs. GPT-4o (IM 3)	0.26	0.30	0.09	0.07	0.14	0.07	0.09	0.28	0.18	0.10
Human B vs. Gemini-1.5 (IM 1)	0.29	0.15	0.08	0.07	0.10	0.04	0.05	0.12	0.08	0.07
Human B vs. Gemini-1.5 (IM 2)	0.25	0.22	0.08	0.05	0.05	0.05	0.05	0.16	0.10	0.04
Human B vs. Gemini-1.5 (IM 3)	0.25	0.18	0.06	0.06	0.06	0.04	0.05	0.09	0.03	0.04
Human B vs. Llama-3 (IM 1)	0.10	0.04	0.07	-0.03	0.03	0.01	0.03	0.06	0.08	-0.06
Human B vs. Llama-3 (IM 2)	0.27	0.23	0.16	0.14	0.22	0.11	0.06	0.25	0.13	0.10
Human B vs. Llama-3 (IM 3)	0.26	0.18	0.07	0.13	0.14	0.14	0.07	0.09	0.06	0.07
Human C vs. GPT-4o (IM 1)	0.36	0.28	0.10	0.12	0.17	0.07	0.03	0.22	0.15	0.13
Human C vs. GPT-4o (IM 2)	0.27	0.21	0.14	0.07	0.15	0.05	0.04	0.20	0.16	0.11
Human C vs. GPT-4o (IM 3)	0.23	0.25	0.09	0.13	0.19	0.08	0.06	0.30	0.17	0.15
Human C vs. Gemini-1.5 (IM 1)	0.19	0.11	0.09	0.11	0.14	0.06	0.03	0.10	0.09	0.08
Human C vs. Gemini-1.5 (IM 2)	0.12	0.21	0.08	0.05	0.08	0.06	0.04	0.15	0.10	0.08
Human C vs. Gemini-1.5 (IM 3)	0.12	0.15	0.08	0.07	0.08	0.07	0.01	0.11	0.05	0.07
Human C vs. Llama-3 (IM 1)	0.24	0.16	0.09	0.08	0.21	0.08	0.02	0.22	0.10	0.06
Human C vs. Llama-3 (IM 2)	0.27	0.36	0.11	0.19	0.28	0.10	0.04	0.43	0.13	0.14
Human C vs. Llama-3 (IM 3)	0.27	0.30	0.10	0.15	0.17	0.16	0.06	0.20	0.12	0.14
Human F vs. GPT-4o (IM 1)	0.44	0.32	0.24	0.17	0.22	0.09	0.07	0.14	0.26	0.17
Human F vs. GPT-4o (IM 2)	0.51	0.30	0.36	0.17	0.25	0.12	0.06	0.10	0.27	0.17
Human F vs. GPT-4o (IM 3)	0.47	0.29	0.25	0.29	0.25	0.19	0.07	0.14	0.32	0.24
Human F vs. Gemini-1.5 (IM 1)	0.37	0.16	0.22	0.11	0.18	0.08	0.03	0.05	0.13	0.11
Human F vs. Gemini-1.5 (IM 2)	0.29	0.16	0.14	0.14	0.10	0.09	0.05	0.12	0.21	0.11
Human F vs. Gemini-1.5 (IM 3)	0.29	0.12	0.17	0.14	0.09	0.09	0.05	0.03	0.09	0.10
Human F vs. Llama-3 (IM 1)	0.32	0.07	0.28	0.24	0.18	0.19	0.04	0.10	0.23	0.13
Human F vs. Llama-3 (IM 2)	0.50	0.18	0.23	0.22	0.21	0.22	0.05	0.07	0.26	0.16
Human F vs. Llama-3 (IM 3)	0.50	0.21	0.27	0.35	0.19	0.25	0.04	0.05	0.12	0.18
GPT-4o (IM 1) vs. Llama-3 (IM 1)	0.59	0.01	0.35	0.30	0.28	0.20	0.36	0.07	0.41	0.45
GPT-4o (IM 1) vs. Gemini-1.5 (IM 1)	0.33	0.38	0.65	0.64	0.59	0.58	0.34	0.51	0.51	0.60
Llama-3 (IM 1) vs. Gemini-1.5 (IM 1)	0.23	0.01	0.27	0.23	0.19	0.14	0.28	0.04	0.26	0.30
GPT-4o (IM 2) vs. Gemini-1.5 (IM 2)	0.49	0.39	0.48	0.59	0.53	0.56	0.48	0.56	0.47	0.64
GPT-4o (IM 2) vs. Llama-3 (IM 2)	0.62	0.33	0.57	0.47	0.57	0.46	0.52	0.37	0.52	0.60
Gemini-1.5 (IM 2) vs. Llama-3 (IM 2)	0.33	0.32	0.36	0.35	0.27	0.36	0.30	0.30	0.23	0.47
Llama-3 (IM 3) vs. GPT-4o (IM 3)	0.56	0.40	0.48	0.53	0.36	0.46	0.55	0.56	0.58	0.58
Llama-3 (IM 3) vs. Gemini-1.5 (IM 3)	0.33	0.21	0.28	0.28	0.15	0.24	0.30	0.24	0.24	0.33
GPT-4o (IM 3) vs. Gemini-1.5 (IM 3)	0.49	0.50	0.50	0.44	0.34	0.49	0.38	0.41	0.35	0.52

Table 12: Full QWK (Quadratic Weighted Kappa) results between all assessor pairs, evaluated at the level of each assessment criterion and the whole essay (“Overall”). C1: Material selection. C2: Material integration and citation; C3: Quality of key components. C4: Logic of structure. C5: Content and clarity of ideas. C6: Coherence (flow of ideas). C7: Cohesion (use of connectors). C8: Grammar and sentence structure. C9: Academic vocabulary.

Assessor	C1	C2	C3	C4	C5	C6	C7	C8	C9	Overall
Human B vs. Human F	0.90	0.77	0.73	0.69	0.75	0.75	0.85	0.86	0.79	0.79
Human B vs. Human C	0.58	0.58	0.70	0.80	0.82	0.86	0.75	0.76	0.80	0.74
Human F vs. Human C	0.73	0.54	0.60	0.64	0.68	0.69	0.65	0.74	0.59	0.65
Human B vs. GPT-4o (IM 1)	0.89	0.84	0.33	0.27	0.46	0.25	0.36	0.74	0.60	0.53
Human B vs. GPT-4o (IM 2)	0.94	0.76	0.42	0.30	0.43	0.27	0.17	0.70	0.68	0.52
Human B vs. GPT-4o (IM 3)	0.94	0.85	0.37	0.41	0.50	0.40	0.54	0.84	0.74	0.62
Human B vs. Gemini-1.5 (IM 1)	0.88	0.59	0.27	0.24	0.33	0.13	0.12	0.46	0.42	0.38
Human B vs. Gemini-1.5 (IM 2)	0.94	0.75	0.28	0.32	0.30	0.23	0.29	0.71	0.62	0.49
Human B vs. Gemini-1.5 (IM 3)	0.94	0.72	0.29	0.25	0.23	0.24	0.32	0.60	0.45	0.45
Human B vs. Llama-3 (IM 1)	0.84	0.76	0.49	0.54	0.74	0.46	0.27	0.79	0.74	0.63
Human B vs. Llama-3 (IM 2)	0.88	0.73	0.38	0.50	0.65	0.45	0.25	0.77	0.71	0.59
Human B vs. Llama-3 (IM 3)	0.89	0.84	0.50	0.57	0.70	0.63	0.46	0.87	0.72	0.69
Human C vs. GPT-4o (IM 1)	0.74	0.54	0.20	0.22	0.46	0.10	0.13	0.45	0.42	0.36
Human C vs. GPT-4o (IM 2)	0.65	0.44	0.32	0.17	0.38	0.15	0.03	0.43	0.54	0.35
Human C vs. GPT-4o (IM 3)	0.62	0.57	0.27	0.39	0.48	0.33	0.24	0.66	0.57	0.46
Human C vs. Gemini-1.5 (IM 1)	0.42	0.21	0.17	0.15	0.30	0.05	0.03	0.29	0.23	0.21
Human C vs. Gemini-1.5 (IM 2)	0.47	0.51	0.16	0.26	0.35	0.17	0.08	0.53	0.45	0.33
Human C vs. Gemini-1.5 (IM 3)	0.47	0.32	0.14	0.18	0.22	0.15	0.09	0.38	0.24	0.24
Human C vs. Llama-3 (IM 1)	0.82	0.73	0.49	0.47	0.74	0.46	0.03	0.86	0.68	0.59
Human C vs. Llama-3 (IM 2)	0.71	0.67	0.24	0.42	0.57	0.31	0.13	0.84	0.72	0.51
Human C vs. Llama-3 (IM 3)	0.71	0.75	0.44	0.57	0.68	0.57	0.18	0.68	0.58	0.57
Human F vs. GPT-4o (IM 1)	0.94	0.82	0.61	0.56	0.80	0.37	0.41	0.71	0.87	0.68
Human F vs. GPT-4o (IM 2)	0.96	0.80	0.79	0.42	0.71	0.32	0.17	0.65	0.86	0.63
Human F vs. GPT-4o (IM 3)	0.94	0.85	0.71	0.83	0.87	0.70	0.60	0.86	0.93	0.81
Human F vs. Gemini-1.5 (IM 1)	0.77	0.63	0.56	0.50	0.64	0.19	0.10	0.45	0.72	0.51
Human F vs. Gemini-1.5 (IM 2)	0.80	0.76	0.54	0.65	0.62	0.43	0.27	0.76	0.89	0.64
Human F vs. Gemini-1.5 (IM 3)	0.80	0.75	0.53	0.57	0.53	0.44	0.35	0.57	0.75	0.59
Human F vs. Llama-3 (IM 1)	0.87	0.68	0.79	0.73	0.89	0.75	0.29	0.91	0.92	0.76
Human F vs. Llama-3 (IM 2)	0.97	0.75	0.58	0.65	0.76	0.56	0.28	0.81	0.89	0.69
Human F vs. Llama-3 (IM 3)	0.97	0.87	0.84	0.92	0.96	0.90	0.58	0.91	0.93	0.88
GPT-4o (IM 1) vs. Llama-3 (IM 1)	0.99	0.62	0.80	0.66	0.89	0.85	0.91	0.56	0.95	0.80
GPT-4o (IM 1) vs. Gemini-1.5 (IM 1)	0.82	0.89	1.00	0.96	0.99	0.99	0.77	0.89	0.93	0.92
Llama-3 (IM 1) vs. Gemini-1.5 (IM 1)	0.73	0.42	0.71	0.63	0.66	0.59	0.81	0.22	0.74	0.61
GPT-4o (IM 2) vs. Gemini-1.5 (IM 2)	1.00	0.90	0.95	0.94	0.99	0.97	0.93	0.97	0.96	0.96
GPT-4o (IM 2) vs. Llama-3 (IM 2)	0.99	0.72	0.87	0.67	0.90	0.71	0.84	0.71	0.91	0.81
Gemini-1.5 (IM 2) vs. Llama-3 (IM 2)	0.97	0.84	0.72	0.63	0.78	0.67	0.76	0.78	0.85	0.78
Llama-3 (IM 3) vs. GPT-4o (IM 3)	0.99	0.98	0.91	1.00	0.97	1.00	0.99	0.99	1.00	0.98
Llama-3 (IM 3) vs. Gemini-1.5 (IM 3)	0.97	0.85	0.81	0.91	0.85	0.81	0.99	0.94	0.97	0.90
GPT-4o (IM 3) vs. Gemini-1.5 (IM 3)	1.00	1.00	0.95	0.99	0.99	1.00	0.99	0.96	0.96	0.98

Table 13: Full AAR1 (adjacent agreement rate with $k = 1$) results between all assessor pairs, evaluated at the level of each assessment criterion and the whole essay (“Overall”). C1: Material selection. C2: Material integration and citation; C3: Quality of key components. C4: Logic of structure. C5: Content and clarity of ideas. C6: Coherence (flow of ideas). C7: Cohesion (use of connectors). C8: Grammar and sentence structure. C9: Academic vocabulary.

Attr	Assessor	C1	C2	C3	C4	C5	C6	C7	C8	C9
CR	Human B	5.6	28.2	2.8	0.7	2.1	3.5	13.4	91.5	66.2
	Human C	100	100	100	98.9	100	98.9	100	100	100
	Human F	99.1	97.2	96.2	80.2	87.7	83.0	89.6	90.6	82.1
	All LLMs	100	100	100	100	100	100	100	100	100
AL	Human B	$26_{\pm23}$	$43_{\pm32}$	$59_{\pm33}$	$45_{\pm0}$	$50_{\pm53}$	$34_{\pm26}$	$46_{\pm24}$	$147_{\pm83}$	$97_{\pm85}$
	Human C	$17_{\pm23}$	$104_{\pm122}$	$39_{\pm38}$	$56_{\pm77}$	$112_{\pm102}$	$38_{\pm69}$	$26_{\pm39}$	$103_{\pm88}$	$65_{\pm94}$
	Human F	$26_{\pm40}$	$77_{\pm75}$	$51_{\pm30}$	$30_{\pm39}$	$51_{\pm57}$	$31_{\pm31}$	$27_{\pm33}$	$52_{\pm67}$	$79_{\pm92}$
	GPT-4o (IM 1)	$79_{\pm10}$	$82_{\pm13}$	$72_{\pm8}$	$59_{\pm7}$	$61_{\pm10}$	$53_{\pm7}$	$55_{\pm9}$	$59_{\pm12}$	$67_{\pm10}$
	Gemini-1.5 (IM 1)	$98_{\pm22}$	$126_{\pm29}$	$120_{\pm33}$	$82_{\pm19}$	$91_{\pm27}$	$84_{\pm22}$	$85_{\pm24}$	$90_{\pm26}$	$99_{\pm49}$
	Llama-3 (IM 1)	$90_{\pm13}$	$91_{\pm18}$	$87_{\pm14}$	$60_{\pm10}$	$64_{\pm15}$	$54_{\pm12}$	$52_{\pm15}$	$56_{\pm13}$	$62_{\pm13}$
	GPT-4o (IM 2)	$291_{\pm44}$	$353_{\pm40}$	$332_{\pm30}$	$333_{\pm37}$	$374_{\pm39}$	$362_{\pm42}$	$347_{\pm36}$	$370_{\pm45}$	$357_{\pm42}$
	Gemini-1.5 (IM 2)	$378_{\pm88}$	$446_{\pm111}$	$512_{\pm106}$	$399_{\pm103}$	$425_{\pm121}$	$397_{\pm109}$	$867_{\pm2032}$	$468_{\pm148}$	$400_{\pm107}$
	Llama-3 (IM 2)	$331_{\pm35}$	$368_{\pm41}$	$438_{\pm111}$	$466_{\pm197}$	$357_{\pm68}$	$351_{\pm107}$	$317_{\pm88}$	$345_{\pm82}$	$355_{\pm97}$
	GPT-4o (IM 3)	$295_{\pm40}$	$437_{\pm40}$	$372_{\pm51}$	$380_{\pm40}$	$444_{\pm37}$	$402_{\pm39}$	$358_{\pm60}$	$422_{\pm50}$	$321_{\pm40}$
	Gemini-1.5 (IM 3)	$374_{\pm86}$	$654_{\pm107}$	$689_{\pm110}$	$592_{\pm94}$	$655_{\pm102}$	$559_{\pm125}$	$473_{\pm109}$	$642_{\pm131}$	$505_{\pm343}$
	Llama-3 (IM 3)	$333_{\pm35}$	$425_{\pm53}$	$409_{\pm49}$	$378_{\pm51}$	$481_{\pm56}$	$389_{\pm51}$	$366_{\pm58}$	$445_{\pm62}$	$362_{\pm44}$
PR	Human B	75	100	100	100	100	80	84	100	95
	Human C	19	72	54	50	81	40	29	79	80
	Human F	47	84	82	44	61	50	49	64	82
	All LLMs	100	100	100	100	100	100	100	100	100
AP	Human B	$1.1_{\pm1.0}$	$2.1_{\pm1.4}$	$2.0_{\pm1.2}$	$1.0_{\pm0}$	$1.3_{\pm0.6}$	$1.0_{\pm0.7}$	$1.2_{\pm0.9}$	$5.3_{\pm4.0}$	$3.5_{\pm3.0}$
	Human C	$0.2_{\pm0.6}$	$2.1_{\pm2.3}$	$0.9_{\pm1.1}$	$1.1_{\pm1.3}$	$2.1_{\pm1.8}$	$0.9_{\pm1.6}$	$0.4_{\pm0.6}$	$2.5_{\pm2.2}$	$1.9_{\pm1.8}$
	Human F	$0.7_{\pm1.0}$	$2.4_{\pm2.0}$	$1.4_{\pm1.0}$	$0.8_{\pm1.0}$	$1.2_{\pm1.4}$	$0.8_{\pm1.0}$	$0.7_{\pm0.9}$	$1.4_{\pm1.8}$	$2.3_{\pm2.2}$
	GPT-4o (IM 1)	$1.8_{\pm0.7}$	$2.3_{\pm0.8}$	$3.4_{\pm0.6}$	$2.3_{\pm0.8}$	$2.0_{\pm0.9}$	$1.8_{\pm0.7}$	$1.3_{\pm0.6}$	$1.9_{\pm0.7}$	$2.2_{\pm0.8}$
	Gemini-1.5 (IM 1)	$2.1_{\pm0.8}$	$2.6_{\pm0.9}$	$3.3_{\pm1.0}$	$1.9_{\pm0.7}$	$2.1_{\pm0.8}$	$2.5_{\pm0.8}$	$2.2_{\pm0.7}$	$2.4_{\pm0.8}$	$2.6_{\pm1.5}$
	Llama-3 (IM 1)	$2.2_{\pm0.5}$	$2.4_{\pm0.6}$	$3.1_{\pm0.9}$	$2.0_{\pm0.7}$	$2.3_{\pm0.8}$	$2.1_{\pm0.6}$	$1.5_{\pm0.7}$	$2.0_{\pm0.7}$	$2.3_{\pm0.5}$
	GPT-4o (IM 2)	$3.8_{\pm0.8}$	$4.8_{\pm1.0}$	$5.8_{\pm1.5}$	$4.6_{\pm1.1}$	$5.1_{\pm0.9}$	$5.5_{\pm1.1}$	$5.7_{\pm1.2}$	$5.0_{\pm0.9}$	$4.9_{\pm1.1}$
	Gemini-1.5 (IM 2)	$5.0_{\pm2.2}$	$5.7_{\pm2.5}$	$8.2_{\pm3.2}$	$5.7_{\pm2.6}$	$6.1_{\pm2.8}$	$5.9_{\pm2.7}$	$5.7_{\pm2.1}$	$5.0_{\pm2.2}$	$5.4_{\pm2.3}$
	Llama-3 (IM 2)	$5.0_{\pm1.7}$	$5.7_{\pm2.2}$	$8.4_{\pm3.0}$	$8.1_{\pm3.8}$	$6.7_{\pm2.7}$	$6.9_{\pm2.9}$	$6.2_{\pm1.9}$	$6.1_{\pm2.2}$	$6.6_{\pm2.4}$
	GPT-4o (IM 3)	$3.9_{\pm0.7}$	$6.5_{\pm1.7}$	$8.5_{\pm2.2}$	$5.6_{\pm1.1}$	$7.7_{\pm1.4}$	$5.8_{\pm1.0}$	$5.1_{\pm1.2}$	$6.7_{\pm1.8}$	$5.2_{\pm1.3}$
	Gemini-1.5 (IM 3)	$4.9_{\pm2.2}$	$9.0_{\pm2.9}$	$10.7_{\pm3.4}$	$9.3_{\pm2.6}$	$9.2_{\pm2.8}$	$8.2_{\pm2.9}$	$6.3_{\pm1.6}$	$9.1_{\pm3.2}$	$7.3_{\pm3.4}$
	Llama-3 (IM 3)	$5.1_{\pm1.6}$	$6.7_{\pm2.2}$	$8.9_{\pm2.1}$	$6.1_{\pm2.1}$	$7.6_{\pm1.9}$	$5.5_{\pm1.8}$	$5.4_{\pm1.7}$	$6.5_{\pm2.0}$	$5.9_{\pm2.0}$

Table 14: General statistics of feedback comments generated by human and LLM assessors. CR (%): comment rate, i.e., the percentage of time a comment is provided. AL: *average length* (measured in tokens) of the provided comments (excluding cases where comments are not given), along with their respective standard deviations. PR (%): problem rate, i.e., the percentage of time a problem is mentioned or implied in the *provided comments*. AP: *average number of problems* identified in the provided comments, along with their respective standard deviations. "All LLMs" means all three LLMs across the three interaction modes. C1: Material selection. C2: Material integration and citation; C3: Quality of key components. C4: Logic of structure. C5: Content and clarity of ideas. C6: Coherence (flow of ideas). C7: Cohesion (use of connectors). C8: Grammar and sentence structure. C9: Academic vocabulary.

Criterion	GPT-4o-Aug	GPT-4o-May	BLEU	ROUGE-L	BERTScore
Content & clarity of ideas	The content is generally relevant, but the clarity of ideas could be improved. Some sentences are complex and could be simplified for better understanding. For example, the sentence discussing Olson (2012) and Sanner (2008) could be broken down into shorter, more concise statements. Additionally, the review could benefit from more explicit explanations of key concepts, such as acculturation.	The content is relevant and covers important aspects of the topic. However, some ideas are not clearly articulated. For example, the sentence "This is made worse because 'the nursing profession faces increased numbers of linguistically and culturally diverse patients who speak little or no English'" could be rephrased for clarity. Additionally, some sentences are overly complex and could be simplified for better understanding.	0.20	0.30	0.63
Logic of structure	The logical structure of the literature review could be improved. The review jumps between different topics without clear transitions, making it difficult for the reader to follow the argument. A more organized structure with clear subheadings and a logical flow of ideas would enhance readability.	The logical structure of the literature review needs improvement. While the review covers important aspects of the psychological impact of COVID-19, the organization of the content is somewhat disjointed. Clearer subheadings and a more logical flow of ideas would help guide the reader through the review. For example, separating the discussion of different demographic groups or types of psychological impacts into distinct sections could enhance readability.	0.17	0.41	0.70
Quality of key components	The introduction provides a general overview of the topic but lacks a clear statement of the significance of the issue or specific research questions. The body of the literature review presents some relevant ideas but lacks a clear organization and focus on the pros and cons of online learning. The conclusion is missing, which is a significant omission as it should summarize key findings and highlight contributions or gaps in the literature.	The introduction lacks a clear thesis statement and does not adequately set up the significance of the topic. The body of the review presents relevant ideas but lacks a clear structure and flow. The conclusion is weak and does not effectively summarize the key points or identify research trends or gaps. A more structured approach with clear sections and subheadings would improve readability and coherence.	0.22	0.44	0.73
Grammar & sentence structure	The draft contains several grammatical errors and awkward sentence structures. For example, "The among of time is short considering the kind of assignments..." should be "The amount of time is short considering the type of assignments...". Varying sentence length and structure could also improve readability.	The draft has several grammatical errors and issues with sentence structure. For example, "The among of time is short considering the kind of assignments" should be "The amount of time is short considering the type of assignments." Additionally, varying sentence length and structure would make the writing more engaging. Proofreading for grammatical accuracy and sentence variety is recommended.	0.50	0.76	0.85
Grammar & sentence structure	The grammatical accuracy is generally acceptable, but there are instances of awkward phrasing and run-on sentences. Varying sentence length and structure could improve readability. Additionally, ensuring subject-verb agreement and correct punctuation would enhance grammatical accuracy.	- Positive: The grammar is generally accurate. - Improvement: Sentence length and variety need improvement. Some sentences are too long and complex, making them difficult to read. - Example: "The health arguments against the use of cannabis include its addictive nature (Hurd et al., 2014) . It has also been directly linked to a range of adverse outcomes in physical health, which include lung cancer (Aldington et al., 2008), impaired respiratory function, cardiovascular disease, elevated systolic blood pressure, stroke (Singh et al., 2012), mental disorders (Saban et al., 2014), which include schizophrenia, especially amongst young people (Casadio et al., 2011), undesirable cognitive changes (Crean et al., 2011)." This could be broken down into shorter sentences.	0.00	0.12	0.49

Table 15: Five random example comment pairs with their BLEU, ROUGE-L, and BERTScore scores provided.

in this section. For example, "$comment" is a placeholder for a comment.

E.1 Prompts for the Feedback Comment Quality Evaluation Framework Pipeline

The full prompt templates for the three steps in the pipeline of the feedback comment quality evaluation framework are given below. Among these three prompts, the prompt for Problem Extraction contains three in-context exemplars, whereas the prompts for the other two steps are zero-shot prompts.

E.1.1 Prompt for Problem Extraction

You will be given a feedback comment written for a student's essay. Your task is to identify and extract all the writing-related problems mentioned or implied in the comment, along with any explanations, suggestions, corrections, questions, quotations, or other relevant information provided in the comment for each extracted problem.

A writing-related problem is any issue that affects the quality of the writing, such as citation errors, logical flaws, coherence issues, grammatical mistakes, or inappropriate word choices, among others.

Extraction Instructions

- Each extracted problem must be clear and can be understood without the need to refer to the original comment.

- Each extracted problem must faithfully reflect the provided comment by including any relevant information. Relevant information includes a further explanation or an elaboration of the problem, a suggestion for improvement, a concrete correction, a clarifying question, an excerpt (possibly without quotation marks) from the student's essay, or any other relevant information that helps to understand the problem.

- Whenever possible, extract each problem and the relevant information as they are written in the comment.

Output Instructions

- Output each extracted problem along with their relevant information line by line headed by "-". - Output "None" if no writing-related problems are mentioned or implied in the comment.

Examples

Example 1 input:

The content is generally informative and relevant, but the clarity of ideas could be improved. Some sentences are overly complex and could be simplified for better understanding. For instance, the sentence "Gandhi's Satyagraha as an adequate substitute for violent methods of conducting social conflict in an early and thorough philosophical examination of Gandhi's attitude to violence in extreme group conflict" is difficult to parse and could be rephrased for clarity.

Example 1 output:

- The clarity of ideas could be improved. Some sentences are overly complex and could be simplified for better understanding. For instance, the sentence "Gandhi's Satyagraha as an adequate substitute for violent methods of conducting social conflict in an early and thorough philosophical examination of Gandhi's attitude to violence in extreme group conflict" is difficult to parse and could be rephrased for clarity.

Example 2 input:

The content and clarity of ideas are generally good, but there are some areas where the author could provide more depth or analysis. For example, the author could have explored the potential reasons why students in India may be more vulnerable to substance abuse, or discussed the implications of legalization for public health policy. To improve, the author could revisit the body of the literature review and provide more nuanced analysis of the findings.

Example 2 output:

- There are some areas where the author could provide more depth or analysis. For example, the author could have explored the potential reasons why students in India may be more vulnerable to substance abuse, or discussed the implications of legalization for public health policy. To improve, the author could revisit the body of the literature review and provide more nuanced analysis of the findings.

Example 3 input:

The author has generally done a good job of integrating source materials and presenting information clearly. However, there are some instances where the connections between ideas could be more explicitly stated, and the citation practices could be more consistent (e.g., some sources are cited with author names, while others are cited with only the year).

Example 3 output:

- There are some instances where the connections between ideas could be more explicitly stated.
- The citation practices could be more consistent (e.g., some sources are cited with author names,

while others are cited with only the year).

Input

$comment

Output

E.1.2 Prompt for Problem Classification

You will be given an excerpt of a feedback comment written for a student's essay. Your task is to answer the following questions:

1. Does the excerpt refer to a specific part of the essay? A specific part refers to a part of the essay that can be easily located by the student. For example, it can be a specific word, phrase, sentence, paragraph, reference etc. used in the essay. It can be a concrete location, such as "sentence 2 in paragraph 2," "in paragraph 6," "the first citation," or "the first sentence of the paper" and so on. A less concrete location, such as "the introduction," or "the conclusion," is also considered a specific part if it is accompanied by some referenceable details, such as "The significance of South Australian policy is unclear, as it is the first citation and the only one in the Introduction." Note that the excerpt may only contain a quoted text from the essay, in which case, the quoted text is considered a specific part.

2. Does the excerpt offer some form of suggestions, general or specific, for the student to improve the essay? If the excerpt only describes a problem and it is unclear what the student should do to fix it, then there is no suggestion. If the excerpt provides a concrete correction, it is considered a suggestion.

3. Does the excerpt provide a concrete correction for the student to apply? Note that when the excerpt only contains a quoted text from the essay and there are some notes indicating a correction (e.g., adding/removing a punctuation, correcting a spelling), this is considered a correction.

Answer each question with "Yes" or "No" based on the content of the excerpt and briefly justify your answer. After answering all the questions, produce your final answers in a newline separated by commas.

Excerpt: $excerpt

E.1.3 Prompt for Correction Relevancy Check

You will be given an excerpt of a feedback comment written for a student's essay according to an assessment question. Your task is to answer the following questions:

1. Does the problem pointed out in the excerpt exist in the corresponding essay? If the excerpt uses a quoted text to point out a problem, check if the quoted text is present in the essay. Please note that the quoted text may not be an exact match either due to misspellings, capitalization errors etc., or because the quoted already contains the correction in place.

2. Is the problem pointed out in the excerpt relevant to the corresponding assessment question? Check if the excerpt is broadly related to any aspect of the assessment question.

3. Is the correction of the problem pointed out in the excerpt correct? If the problem does exist in the essay, check if the correction fixes the problem or presents a plausible solution or improvement.

Here is the essay:

$essay

Here is the assessment question:

$question

Here is the excerpt:

$excerpt

Answer each question with "Yes" or "No" utilizing all the information provided and briefly justify your answer. After answering all the questions, produce your final answers in a newline separated by commas.

E.2 Prompts for the Main Experiments

Our prompts consist of three parts: (1) a system prompt part that provides general background information and specifies the writing topic and some general assessment guidance; (2) a writing part that includes an entire literature review (with references); (3) an assessment instruction part, where one or multiple assessment questions (see Table 7) are asked in various manners according to the interaction modes.

We keep the system prompt fixed across the three interaction modes. For the main experiments, the system prompt is as follows:

> You are an expert academic writing instructor specializing in graduate-level work, with particular experience supporting students who speak English as an additional language. You have been asked to evaluate a literature review submitted by a graduate student on the following topic: $Topic. The review was written in 2021, so references after this year are not expected.
>
> When assessing the student's writing, please strictly follow the instruction provided to you and make sure your score/feedback is carefully considered and constructive. Please provide your comments and/or suggestions with as much detail and specificity as possible. Please provide

specific examples of sentences, paragraphs or sections that you think could use improvement. If you write comments, please start them with something positive. Please proceed with things that could be improved, would make things clearer for the reader, would make the text flow better, etc.

For the writing part, we explicitly mark the beginning and end of the writing for clarity:

########## Writing starts ##########

$writing

########## Writing ends ##########

The specifics of how the assessment instruction part is constructed are detailed below.

E.2.1 Interaction Mode 1

In Interaction Mode 1, all assessment questions (see Table7) are asked at once:

Q1: {Assessment question 1}

Q2: {Assessment question 2}

...

Q9: {Assessment question 9}

After these assessment questions is an answer instruction:

For each of the 9 questions above, provide your comments or suggestions if any, followed by your score out of 10. Please indicate which question you are providing feedback for by starting your response with 'A1:', 'A2:', etc. Each response should use the following format:

Score: ...

Comments or suggestions: ...

Note that we use "if any" to denote the optionality of the comments and suggestions. We tried putting "(Optional)" after "Comments or suggestions," but that does not make a difference.

E.2.2 Interaction Mode 2

In Interaction Mode 2, the assessment questions are presented sequentially and one at a time. Below is the basic structure:

Q$_i$: {The ith assessment question.}

{Answer instruction}

A$_i$:

The answer instruction resembles the one used in the Interaction Mode 1.

Provide your score out of 10, followed by comments or suggestions if any. Your response should use the following format:

Score: ...

Comments or suggestions: ...

Note that, we append LLM's response to the ith assessment question to the original prompt to form a new prompt, to which the next assessment question is added. This way, the writing is only provided once (at the beginning), but the LLM will have access to previous assessment questions as well as its answers to those questions.

E.2.3 Interaction Mode 3

In Interaction Mode 3, each assessment question is asked independently, so there are 9 separate prompts for each essay.

The structure for the assessment part of the prompt is similar to that in Interaction Mode 2, but without indexation and prefix "Q/A":

{An assessment question.}

{Answer instruction}

The answer instruction works exactly the same as in Interaction Mode 2.

E.3 Prompts for the Follow-Up Experiments

E.3.1 System Prompt Simplification

Below is a simplified system prompt removing the helpful information from the default system prompt used in Section 5.

You are an expert academic writing instructor for graduate students. You have been asked to evaluate a literature review submitted by a student below. The writing is broadly related to the following topic: $Topic.

When assessing the student's writing, please strictly follow the instruction provided to you and make sure your score/feedback is carefully considered and constructive.

E.4 Prompts for Assessing Specificity and Helpfulness

You will be given a feedback comment written for a student's essay according to an assessment question. Your task is to rate the feedback comment on (1) specificity and (2) helpfulness, using a scale from 1 to 10, where 1 is the lowest and 10 is the highest. Conclude your response with the final ratings in this format: "Specificity: X, Helpfulness: X" (where X is a score from 1 to 10).

Here is the essay:

$essay

Here is the assessment question:

$question

Here is the feedback comment:

$feedback

Please rate the specificity and helpfulness of the feedback comment.