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

Zhengxiang Wang, Veronika Makarova, Zhi Li,

Jordan Kodner, Owen Rambow

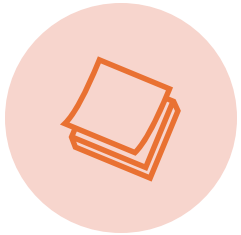
Acknowledgements

- Zhengxiang Wang, Veronika Makarova, and Zhi Li would like to thank **Social Sciences and Humanities Research Council of Canada (SSHRC)** for funding the writing project ("Collaborative development of written academic genre awareness by international graduate students") under the **Insight Development Grants (430-2020-00179).**
- They also appreciate three graduate students, i.e., **Leslee G. Mann, Abdelrahman Alqudah, and Hanh Pham** who expertly assessed the participants' submitted writings, and the participants who participated in the project.

- Zhengxiang Wang, Jordan Kodner, and Owen Rambow are grateful for the supports from the **Institute for Advanced Computational Science** (IACS) at Stony Brook University, in particular the free GPT access it provides.
- Zhengxiang Wang is supported by **IACS's Junior Researcher Award** since Fall 2024.

- We thank **Hannah Stortz** for providing manual annotations for our study, and **Tatiana Luchkina** and **Yongjun Zhang** for their helpful feedback.

Outline



INTRODUCTION

CORPUS

EVALUATION

RESULTS

CONCLUSION

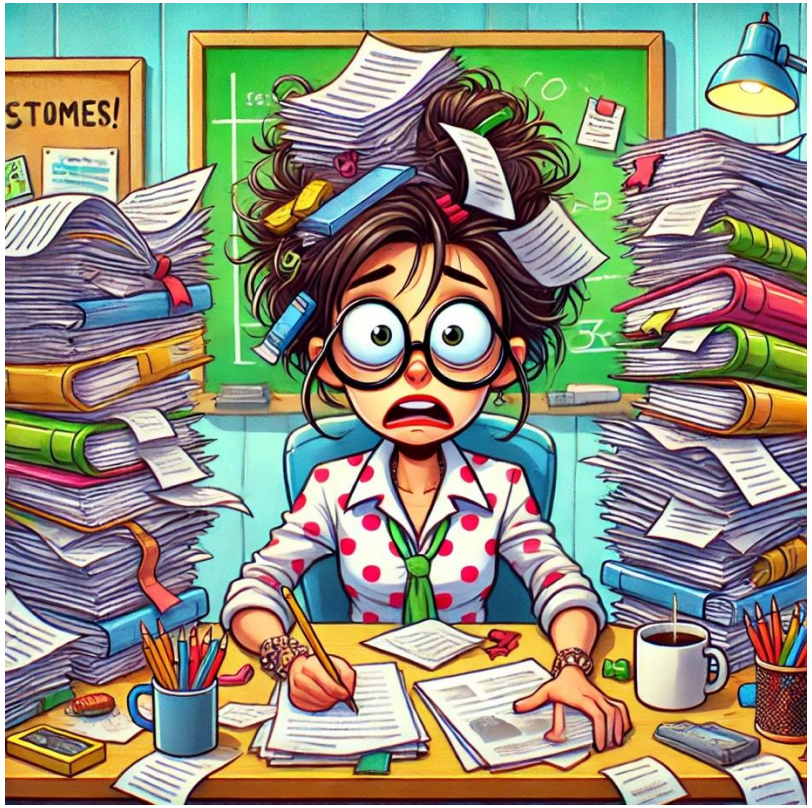
Introduction

Manual Writing Assessments

- Both time-consuming and labor-intensive

- Even more demanding and challenging in the case of **multi-dimensional analytic assessments**
  - Assigning scores and providing comments based on multi-dimensional analytic criteria

- Other drawbacks:
  - ➢ Easily affected by fatigue, mood, and biases etc.
  - ➢ Consistency and reliability issues with analytic assessments (see reviews by Banno et al., 2024)
  - ➢ Often not provided due to the significant time, cost, and expertise required to produce them



Credit: GPT-4o

Existing Non-LLM AWE Systems

- AWE: Automated Essay Evaluation, including

1) Automated Essay Scoring

- Mostly holistic scoring (Ke and Ng, 2019)
- Unidimensional analytic scoring (Jong et al., 2023; Banno et al., 2024)
  - Organization (Persing et al., 2010), thesis clarity (Persing and Ng, 2013), prompt adherence (Persing and Ng, 2014), argument strength (Persing and Ng, 2015), stance (Persing and Ng, 2016)

2) Feedback Comment Generation

- Mostly corrective, e.g., grammar error correction (Nagata, 2019; Han et al., 2019; Babakov et al., 2023)
- Mostly sentence-level rather than essay-level (Behzad et al., 2024b)

LLMs as AWE Systems: Existing Research

1) Automated Essay Scoring

- Holistic scoring (Mizumoto and Eguchi, 2023; Yancey et al., 2023; Wang and Gayed, 2024)
- Unidimensional analytic scoring, e.g., discourse coherence (Naismith et al., 2023)
- Multi-dimensional analytic scoring (Yavuz et al., 2024; Banno et al., 2024)

2) Feedback Comment Generation

- Holistic feedback (Behzad et al., 2024a,b)
- Corrective feedback (Mizumoto et al., 2024; Song et al., 2024)
- Multi-dimensional analytic feedback (Guo and Wang, 2024; Behzad et al., 2024a; Han et al., 2024)

3) Joint Essay Scoring and Feedback Generation

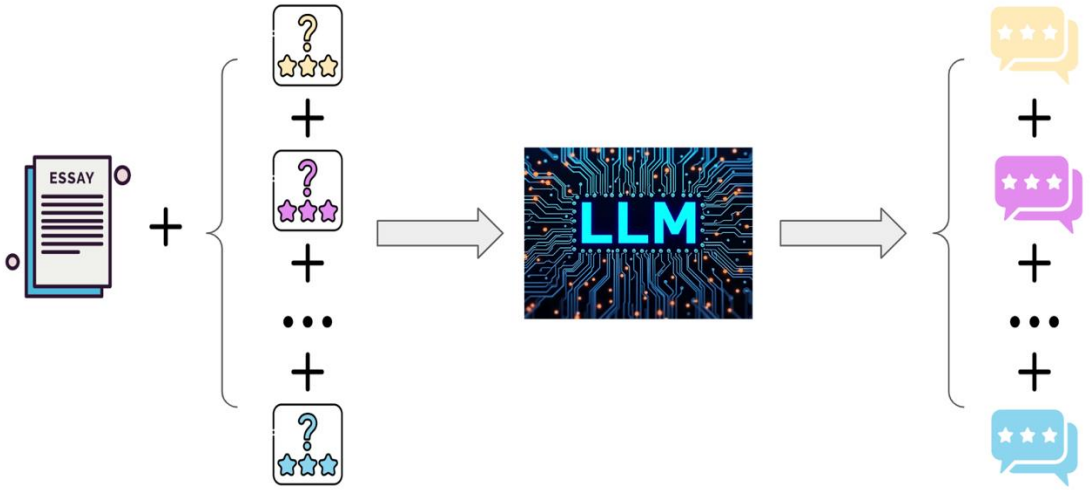
- Holistic score and comment (Stahl et al., 2024) on short essays by native speakers with no human reference comments

LLMs for Multi-Dimensional Analytic Assessments: Motivations

- **Research Gap**
  - Understudied, evidenced by a lack of a related corpus (Banno et al., 2024)

- **Possibility**
  - Strong instruction-following capabilities
  - Promising results from previous studies

- **Impact**
  - Increasing popularity and pedagogical significance



- **Promise**
  - Accessibility: free or highly affordable, real-time responses
  - Inclusiveness: multi-lingual/modal, personalized responses
  - Knowledge: Internet-scale linguistic and world knowledge

Addressing the Research Gap

Corpus

- Literature reviews written by L2 graduate students

- Assessed by independent human experts according to 9 analytic assessment criteria

Evaluation

- Various LLMs prompted to assess the corpus using the same criteria under various conditions

- A novel LLM-based feedback comment quality evaluation framework.

Research Question

- Can LLMs provide *reasonably good* multi-dimensional analytic writing assessments?

Corpus

Overview

- **Background**: A research project at the University of Saskatchewan (Canada) in 2021
  - Three rounds of a 5-unit online tutorial series, 13 weeks each round
  - Voluntary participation: 51 authors contributed, but only 31 completed everything

- **Basic Statistics**: 141 literature reviews by 51 L2 graduate students
  - Average word count: 930 words (w/o references), or 1321 words (w/ references)

- **Five Broad Topics**: drawn from the humanities and social sciences
  - (1) social consequences of legalized cannabis, (2) Canadian linguistic landscape,
  - (3) online learning, (4) lessons from the COVID-19 pandemic, and (5) pacifism

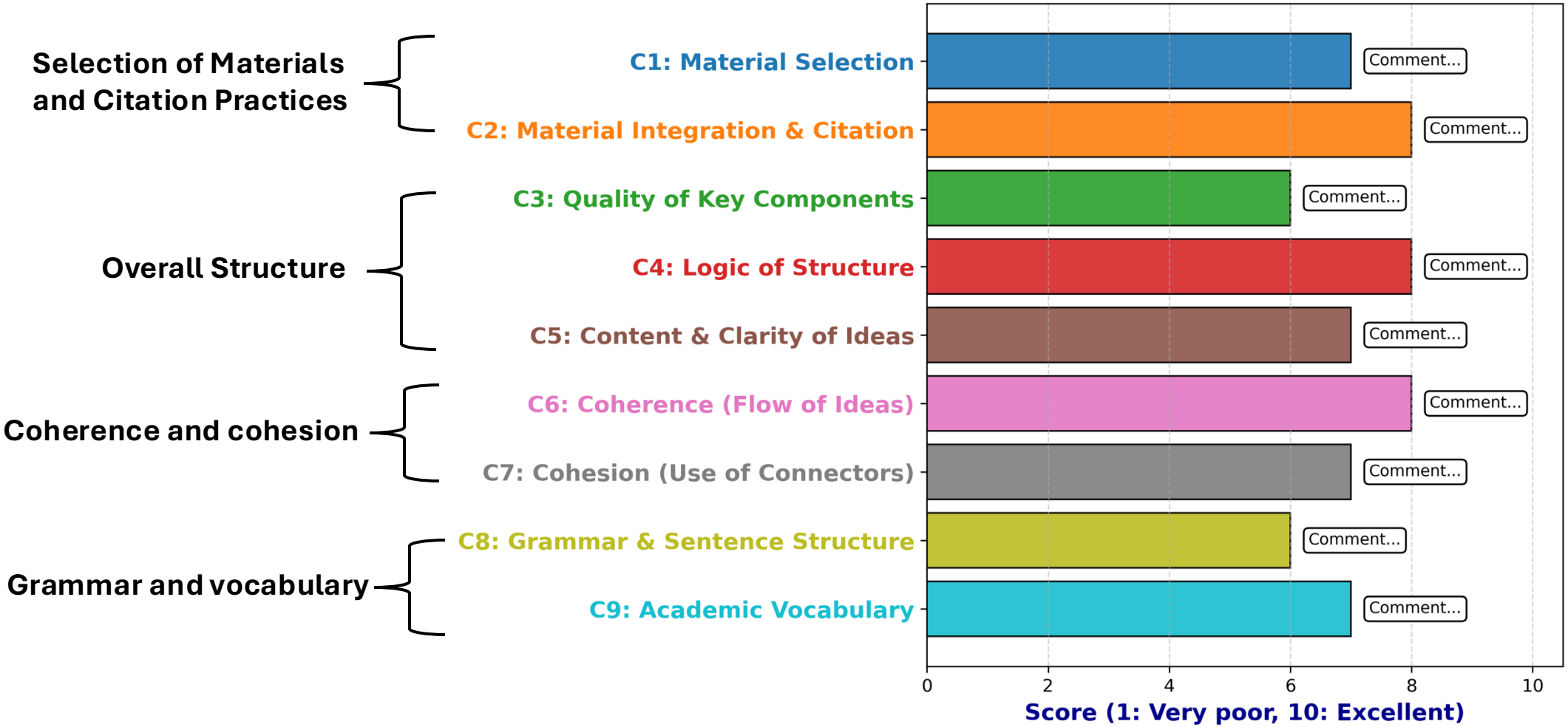
- **No Data Contamination Concerns**:
  - (1) Created prior to the release of ChatGPT;
  - (2) Never been made available to the public

Assessors

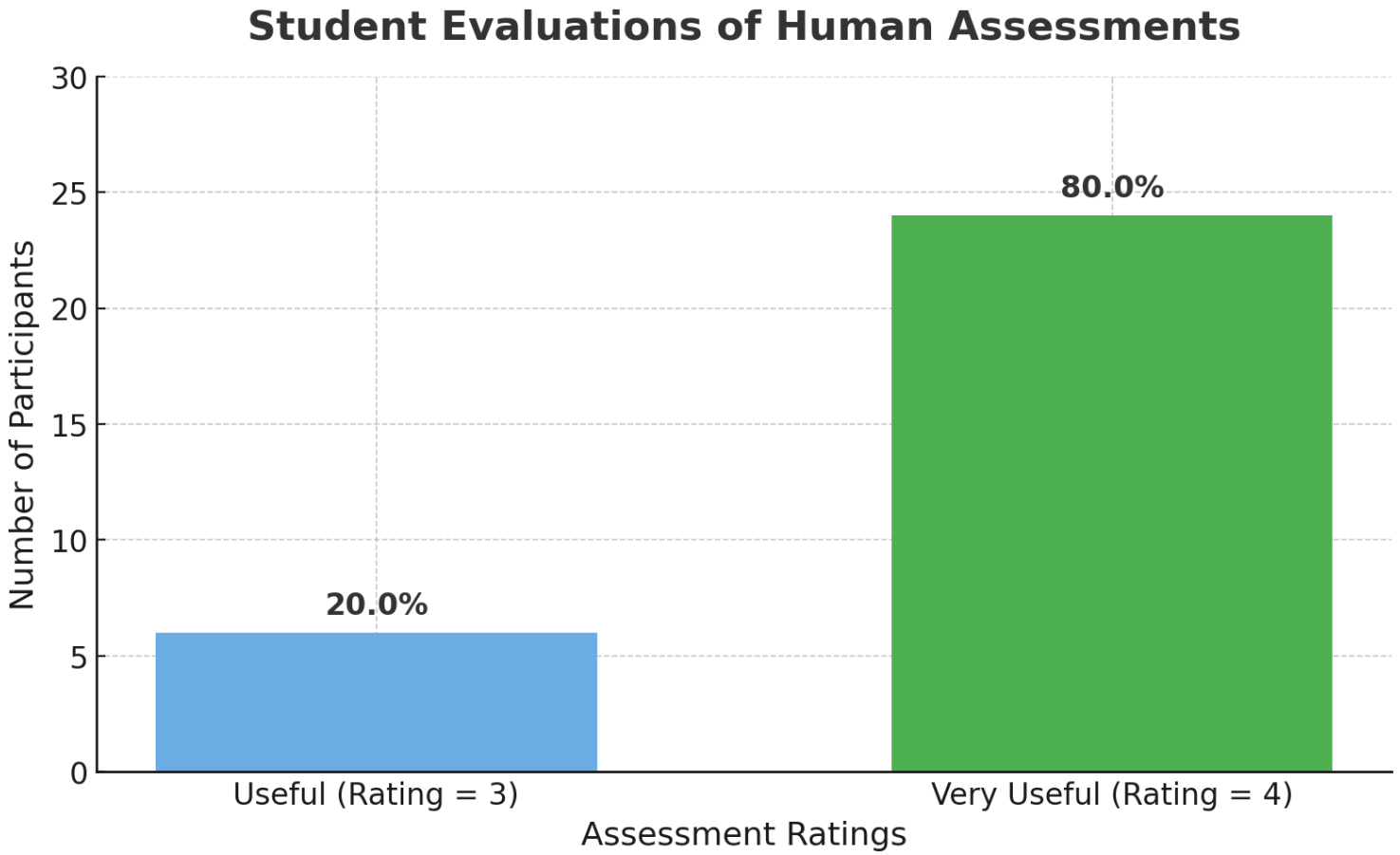
Code	Role	Rounds	Topics	# Essays
A	Graduate RA	1	1-5	27
B	Graduate RA	1-3	1-5	141
C	Faculty Member	1-3	1, 2, 5	93
D	Faculty Member	1	2	4
E	Faculty Member	1-3	3, 4	43
F	Graduate RA	2, 3	1-5	106

- Majority of the essays assessed by three (94.3%) or two (5.0%) independent human experts.

Assessments



Acceptable Assessment Quality



Thirty of 31 participants who completed all writing tasks evaluated the quality of human assessments on a **4-point scale** (1-4) in an anonymous final project survey.

Corpus Availability

Evaluation

Evaluation of Scores

- **Quadratic Weighted Kappa (QWK)**: A metric for rating inter-rater agreement
  - Ranging from 0 (random agreement) to 1 (perfect agreement)
  - Can be negative when agreement is worse than chance
  - 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)

An assigned score

$$\mathbf{W}_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

Number of possible scores

$$\kappa = 1 - \frac{\sum_{i,j} \mathbf{W}_{i,j} \mathbf{O}_{i,j}}{\sum_{i,j} \mathbf{W}_{i,j} \mathbf{E}_{i,j}}$$

Observed frequencies of score pairs *i* and *j*

Expected frequencies of score pairs *i* and *j*

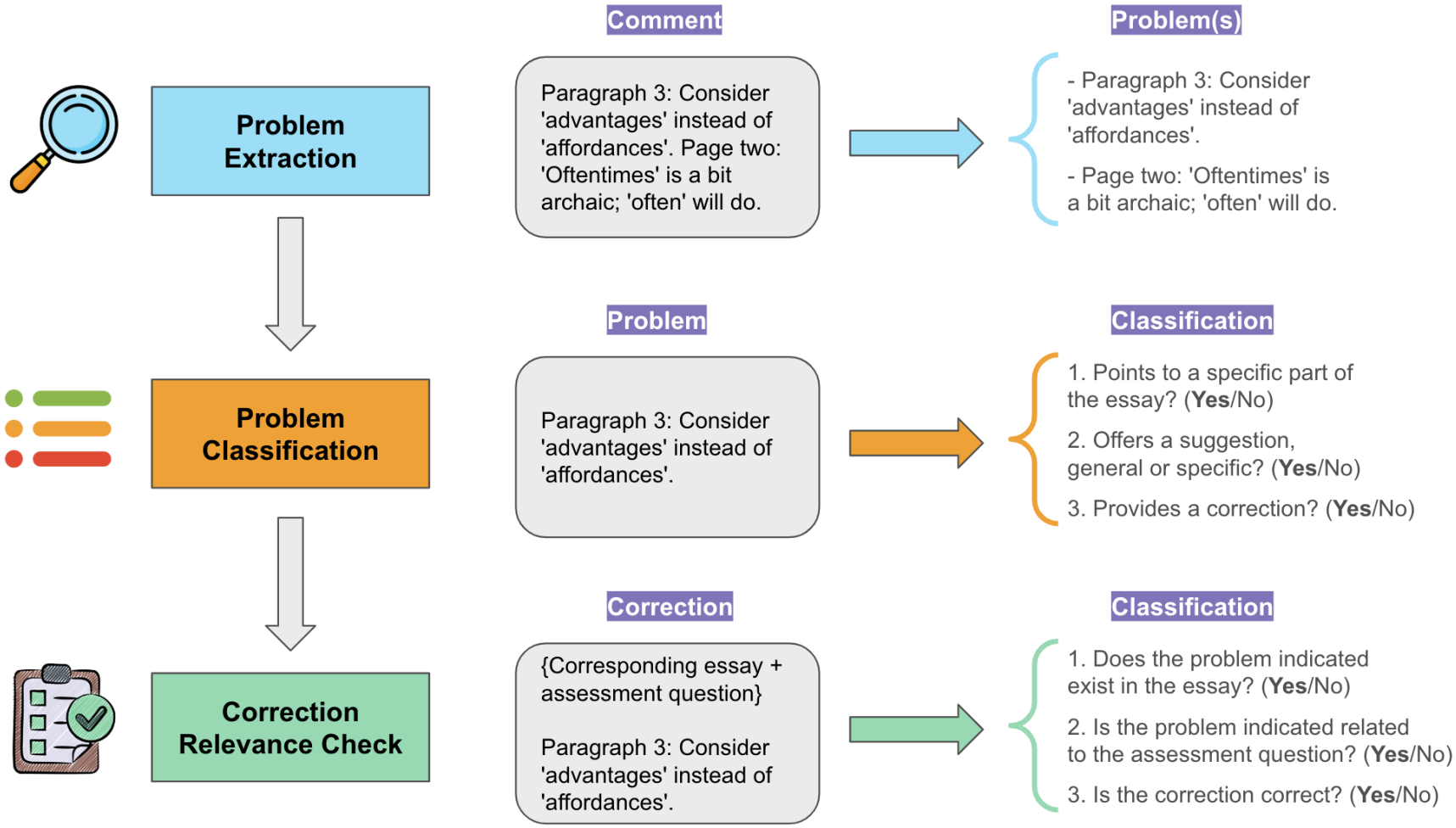
- **Adjacent Agreement Rate (AAR)**: A metric for measuring practical agreement
  - Percentage of scores (from two raters) that lie within a specified threshold *k* of one another
  - This study considers two scores within 1 point, i.e., minimum ordinal difference, as equivalent (AAR1)
  - Addressing the limitation of QWK's chance correction & observed scoring biases/inconsistency issues

Evaluation of Comments: Existing Methods

- Mostly reliant on manual judgments (Chiang and Lee, 2023; Han et al., 2023; Stahl et al., 2024; Behzad et al., 2024a,b)
  - E.g., employing assessment questions to guide human annotators to assess on a Likert scale
- **Drawbacks**: Expensive, time-consuming, not scalable, and may not always be reproducible

- For L2-related comments, common criteria for assessing comment quality include
  - **Specificity** (Han et al., 2023; Stahl et al., 2024; Behzad et al., 2024a,b)
  - **Relevance** (Han et al., 2023; Stahl et al., 2024; Behzad et al., 2024a,b)
  - **Helpfulness** (Han et al., 2023; 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*

Evaluation of Comments: Our Proposal



An Example Implementation of the Pipeline

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.

Input Comment

An Example Implementation of the Pipeline

Problem Extraction

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.

Input Comment

Problem Classification

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

For Extracted Problem 1

Correction Relevance Check

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

For Extracted Problem 1

An Example Implementation of the Pipeline

Problem Extraction

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.

Input Comment

Problem Classification

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

For Extracted Problem 2

Correction Relevance Check

~~N/A since the extraction problem contains no corrections.~~

For Extracted Problem 2

An Example Implementation of the Pipeline

Problem Extraction

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.

Input Comment

Problem Classification

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

For Extracted Problem 3

Correction Relevance Check

~~N/A since the extraction problem contains no corrections.~~

For Extracted Problem 3

Validations of the Proposed Pipeline

Validated against manual annotations from two annotators

Automatic validation

		TP: Num of Correctly Extracted Problems	FP: Num of Incorrectly Extracted Problems	FN: Num of Problems not Extracted
Problem Extraction	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**
Problem Classification	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**
Correction Relevance Check	"Yes" Rate (Positive Samples)	0.92	0.85	0.89
	"No" Rate (Negative Samples)	0.93	0.92	0.92

Results

Highlights

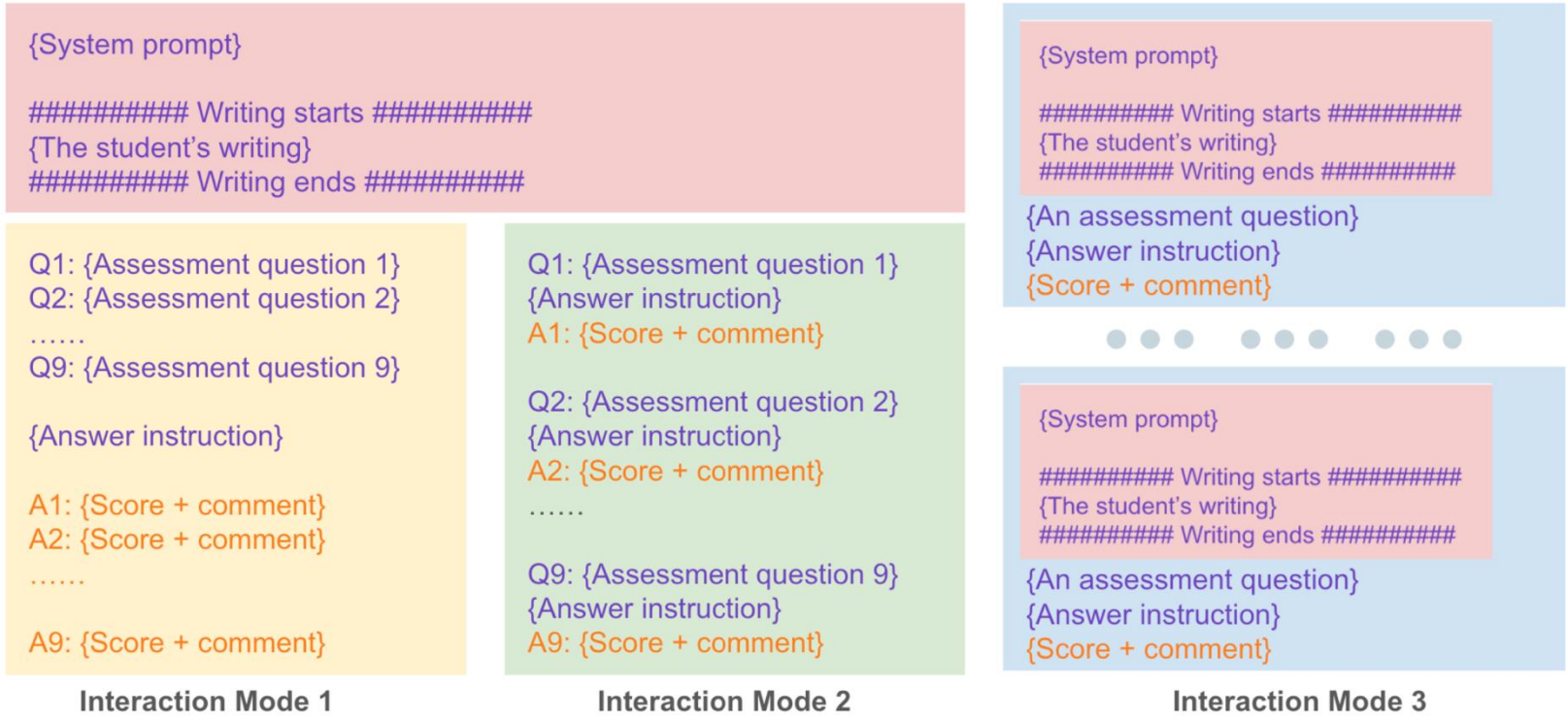
- **LLMs can generate reasonably good and generally reliable assessments**
  - Scores: Can approximate human-assigned scores, typically within 1 point
  - Comments: Can identify more relevant, specific writing problems than human assessors
  - Score-Comment Correlation: Expected negative correlations observed in human/LLM-generated assessments
  - Reliability: highly stable scores + decently similar comments

- **Our proposed feedback comment quality evaluation framework is further validated**
  - Provides a more effective and fine-grained metric in measuring score-comment correlation
  - Can measure specificity and helpfulness levels of comments in a more fine-grained and interpretable way

Main Experiments

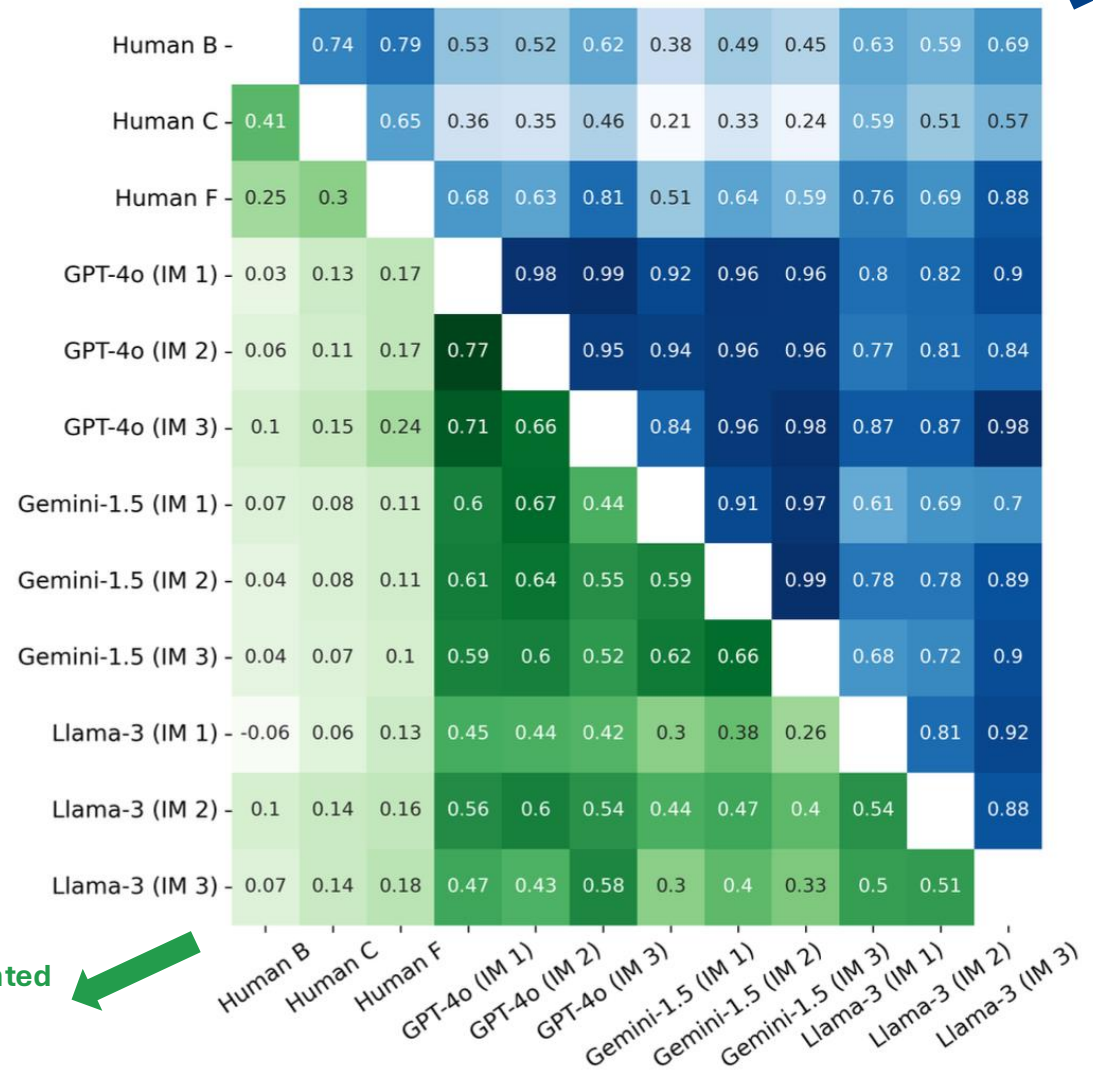
LLM Prompting

- **LLMs**: Three popular models, i.e., (1) GPT-4o; (2) Gemini-1.5; and (3) Llama-3
- **Prompt design**: a system prompt + an input essay + an assessment instruction



Scores: Overall Agreement

Adjacent Agreement Rate with k=1 (AAR1)

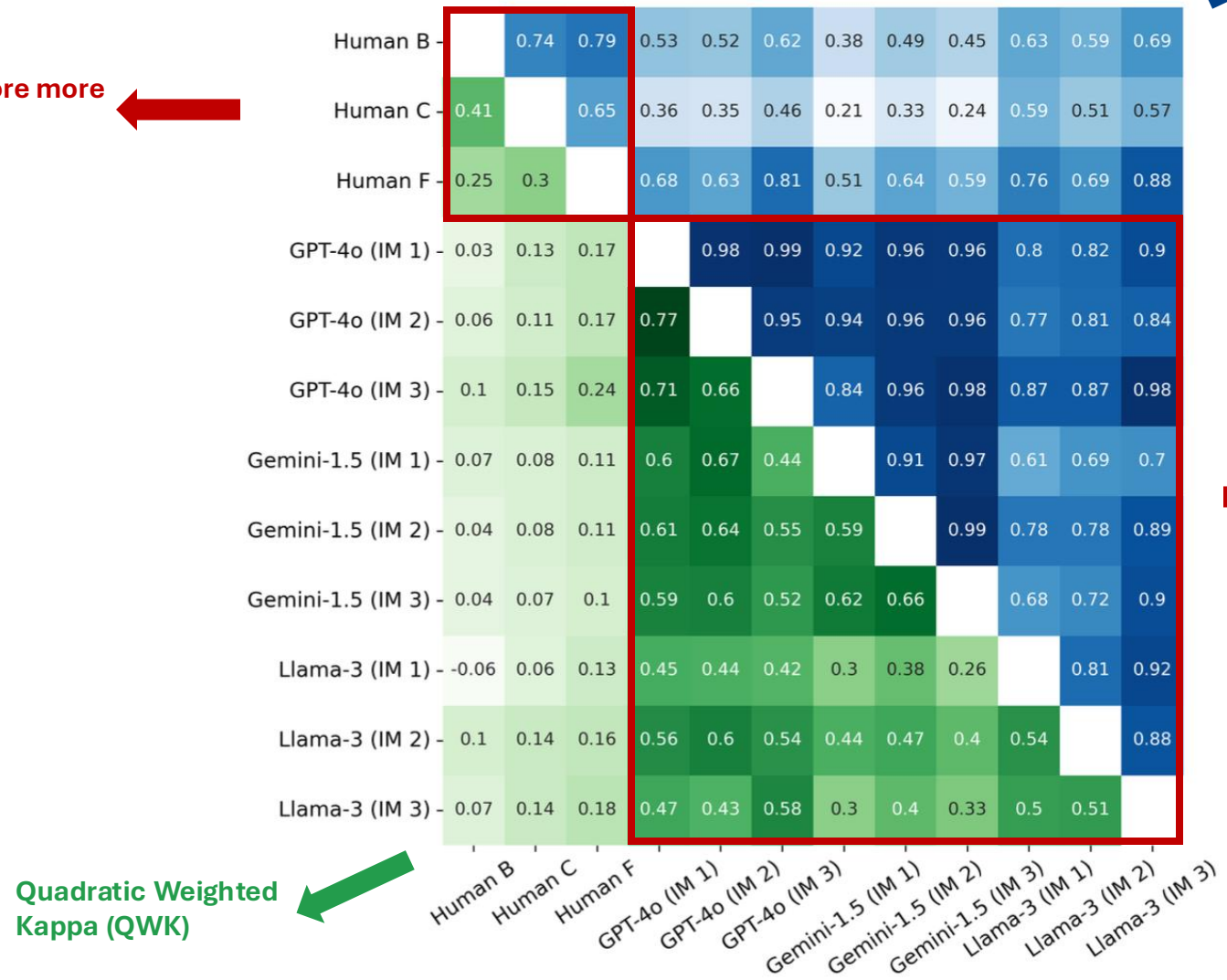


Quadratic Weighted Kappa (QWK)

Scores: Overall Agreement

Adjacent Agreement Rate with k=1 (AAR1)

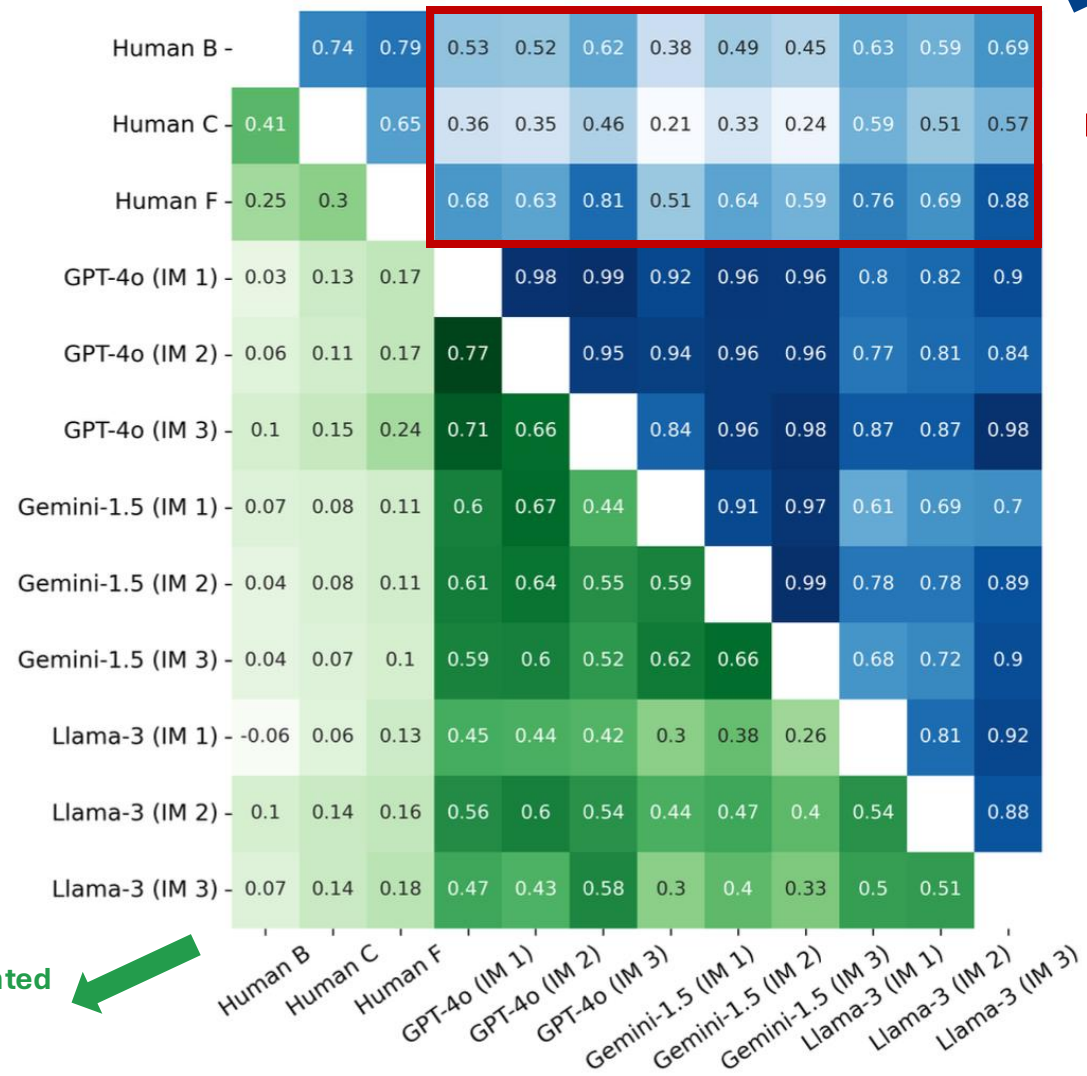
Humans score more like humans



LLMs score more like LLMs

Quadratic Weighted Kappa (QWK)

Scores: Overall Agreement



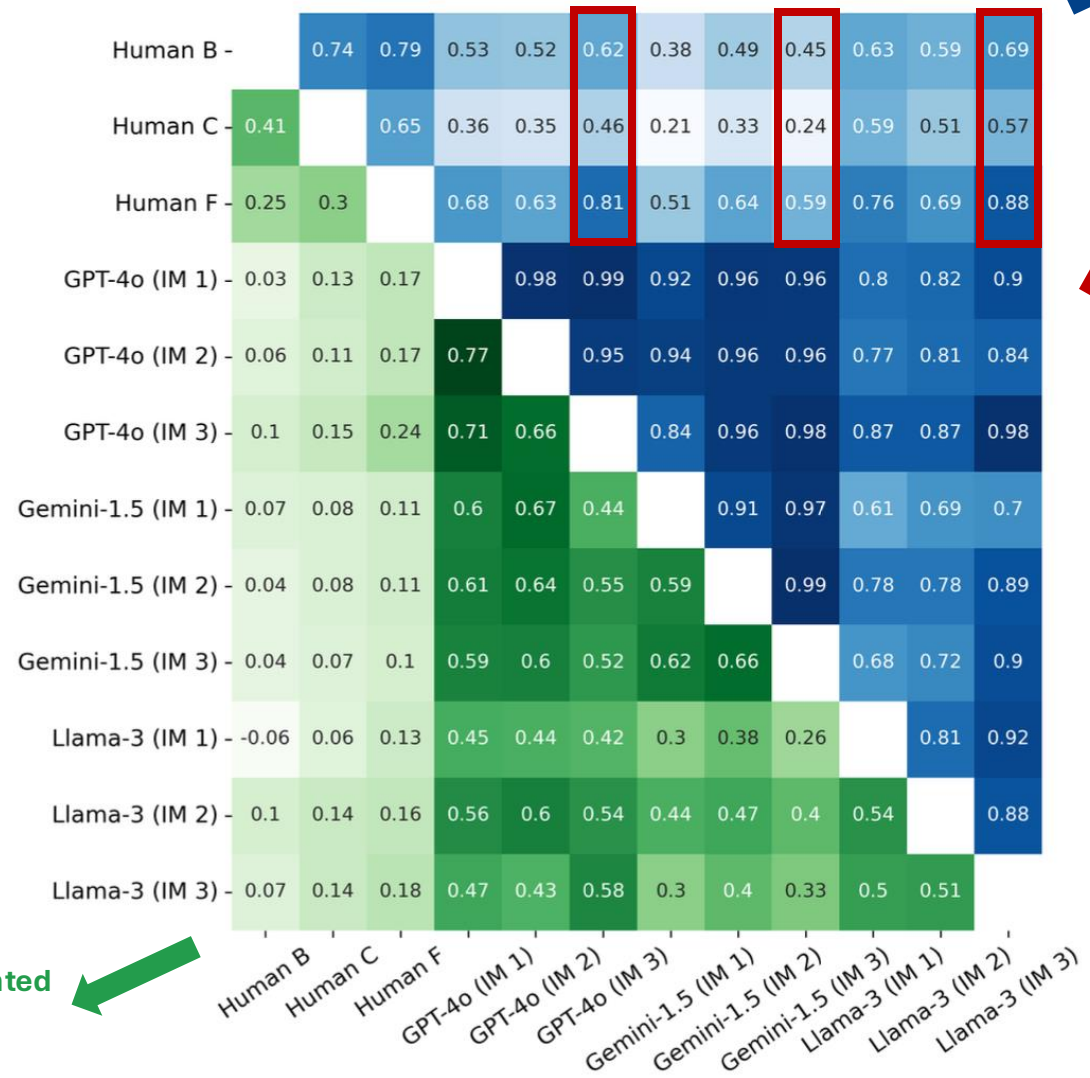
Adjacent Agreement Rate with k=1 (AAR1)

LLMs can score approximately like humans, typically within 1 point.

Quadratic Weighted Kappa (QWK)

Scores: Overall Agreement

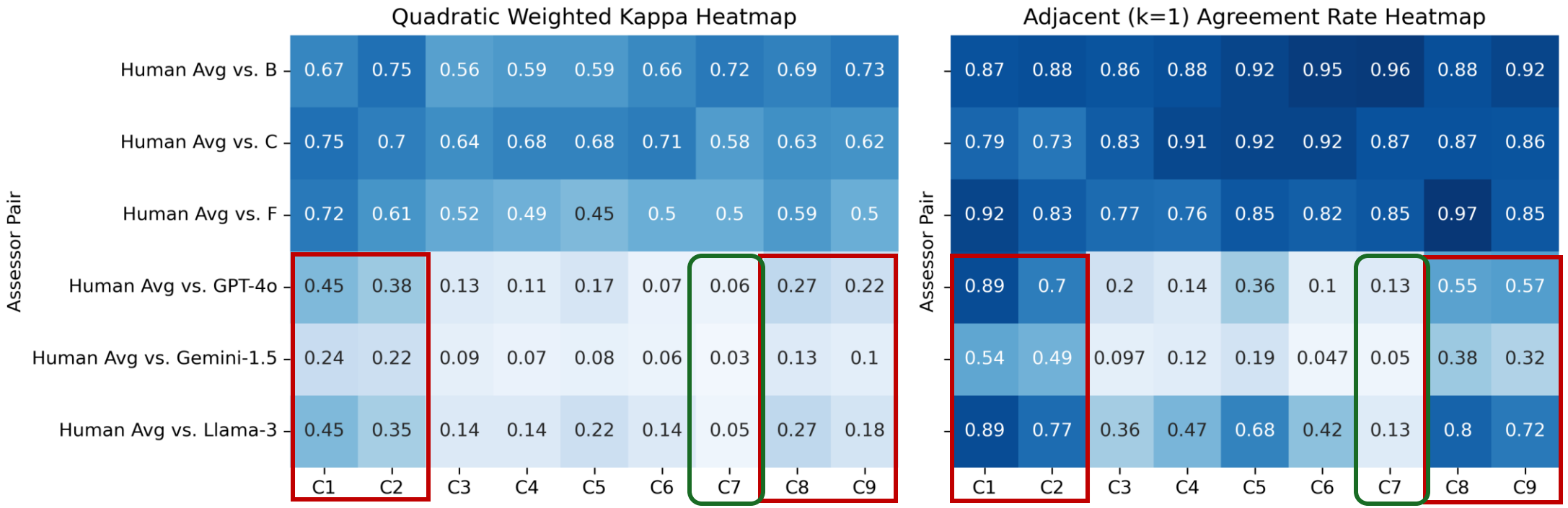
Adjacent Agreement Rate with k=1 (AAR1)



Human-LLM agreement tends to be higher when LLMs respond to each assessment criterion separately under IM 3.

Quadratic Weighted Kappa (QWK)

Scores: Criterion-Level Agreement



- **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.

Scores: Main Observations

- **Overall Observations**
  - Humans score more like humans and LLMs score more like LLMs
  - LLM-assigned scores can approximate human-assigned ones, typically within 1 point
  - Interaction Mode 3 produces most-aligned scores → more independent scoring decisions

- **Criterion-Level Observations**
  - Relatively higher human-LLM agreement on
    1. C1: Material selection
    2. C2: Material integration and citation
    3. C8: Grammar and sentence structure
    4. C9: Academic vocabulary

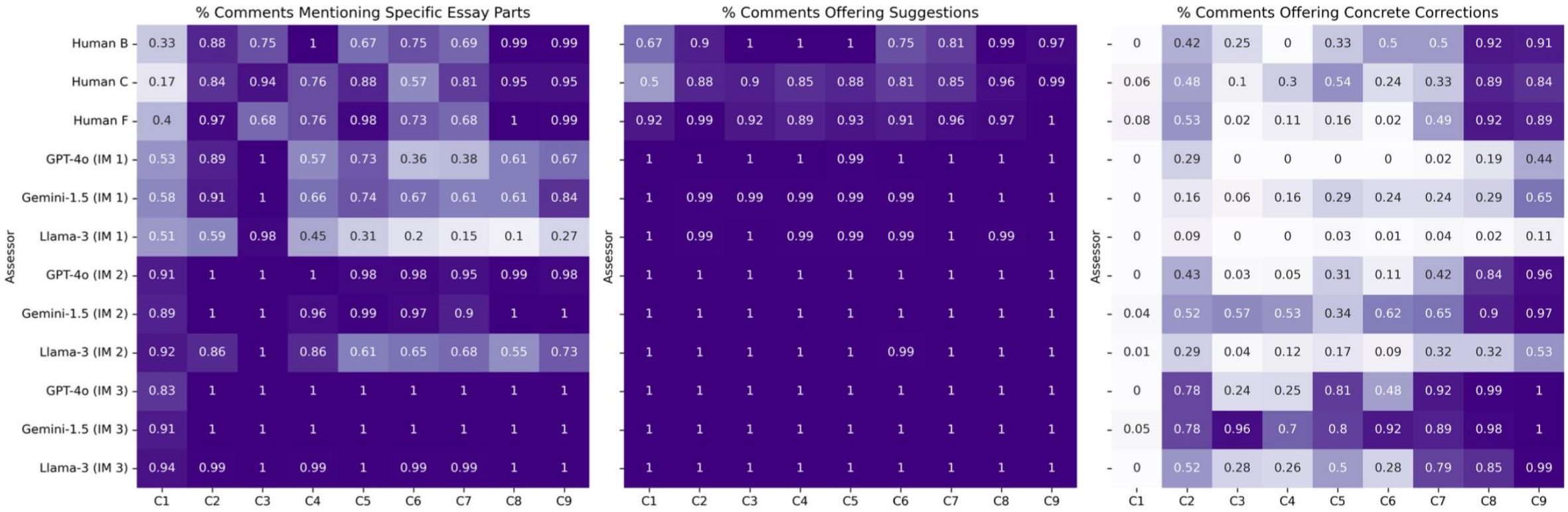
  - Rather poor human-LLM agreement on (LLMs assigning lower scores)
    - C7 (cohesion or use of connectors)

Comments: Overall Statistics

Assessor	Avg Comment Rate	Avg Comment Len	Avg Problem Rate	Avg Problem Num
Human B	0.24	104±85	0.97	3.8±3.5
Human C	1.00	62±85	0.56	1.3±1.8
Human F	0.90	47±58	0.63	1.3±1.6
GPT-4o (IM 1)	1.00	65±14	1.00	2.1±0.9
Gemini-1.5 (IM 1)	1.00	97±33	1.00	2.4±1.00
Llama-3 (IM 1)	1.00	68±20	1.00	2.2±0.8
GPT-4o (IM 2)	1.00	347±46	1.00	5.0±1.2
Gemini-1.5 (IM 2)	1.00	477±698	1.00	5.9±2.7
Llama-3 (IM 2)	1.00	370±112	1.00	6.6±2.8
GPT-4o (IM 3)	1.00	381±65	1.00	6.1±2.0
Gemini-1.5 (IM 3)	1.00	571±182	1.00	8.2±3.3
Llama-3 (IM 3)	1.00	399±67	1.00	6.4±2.3

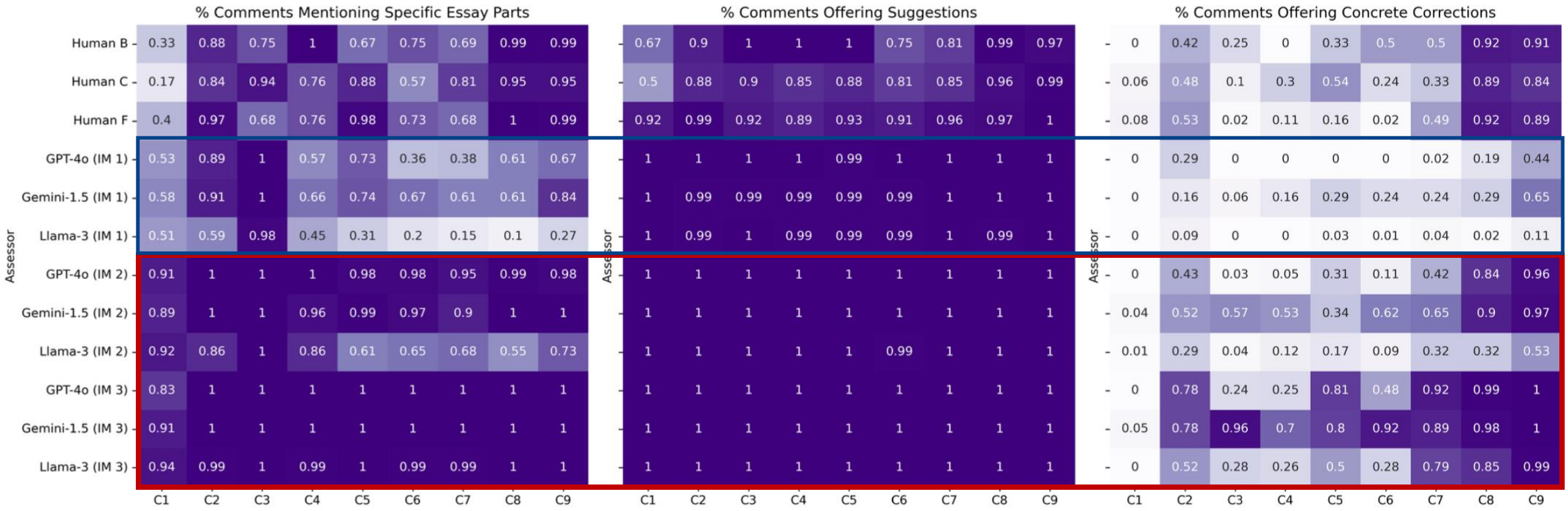
The table shows the percentage of time an assessor provided a comment, and when they did, the average length of these comments, the percentage of comments identifying a problem, and the average number of problems identified in each comment.

Comments: Criterion-level Characterizations



- **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.

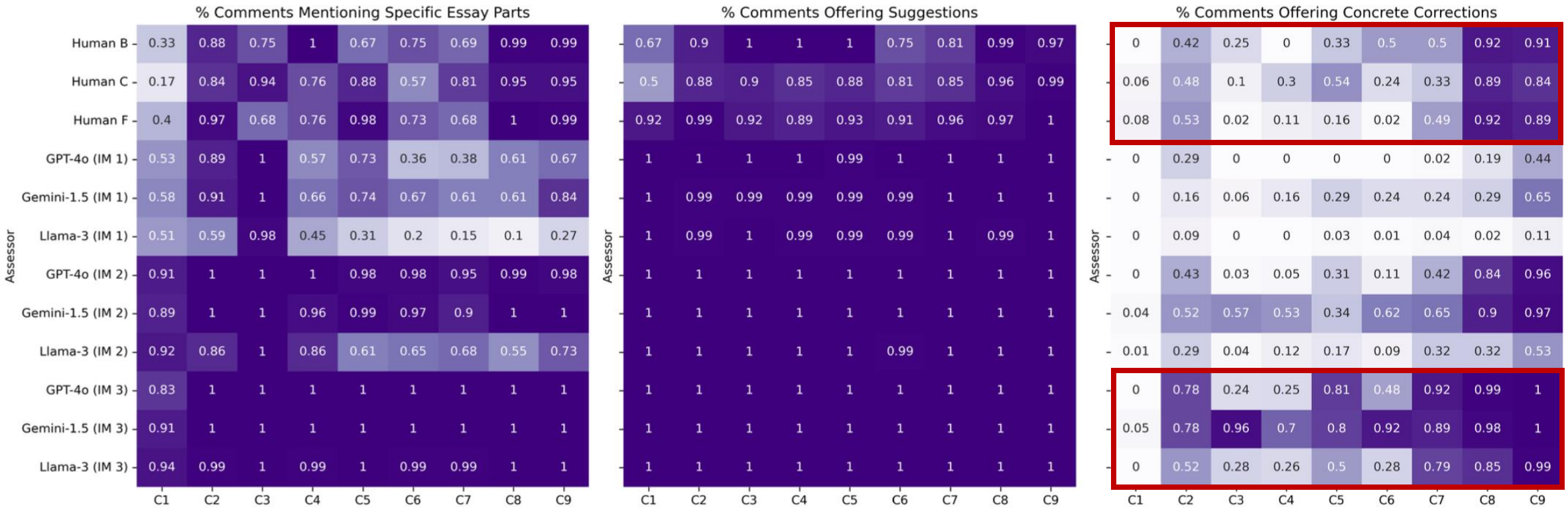
Comments: Criterion-level Characterizations



Interacting with LLMs one question at a time leads to more elaborate, specific, and helpful comments.

- **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.

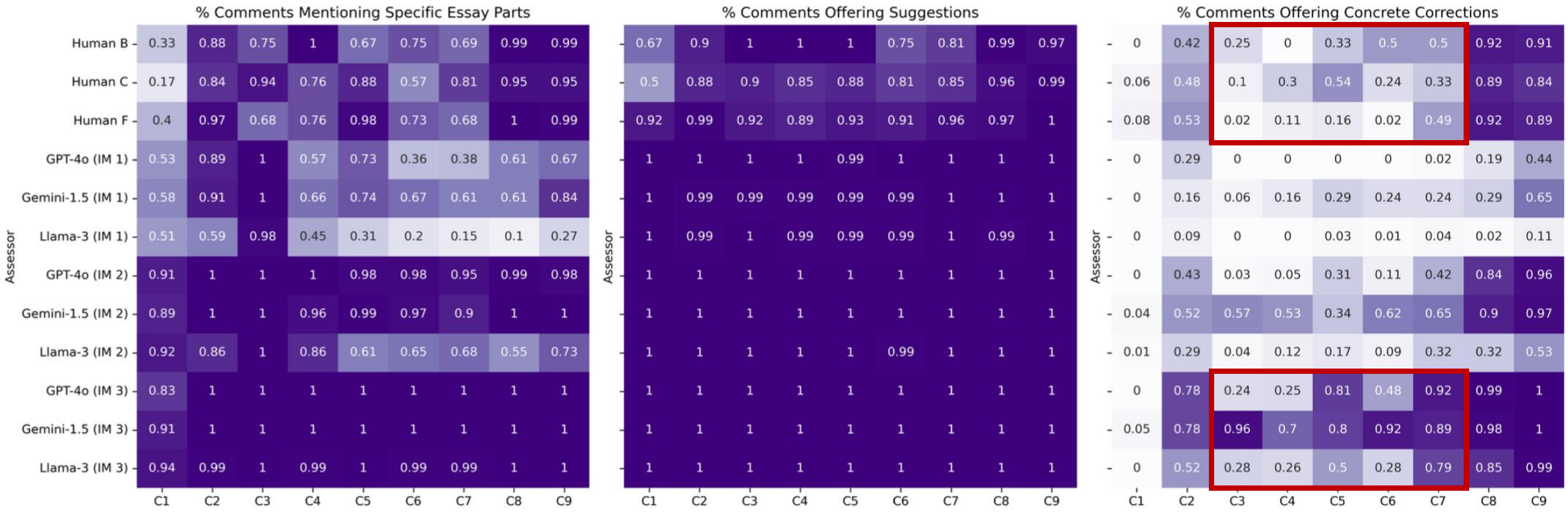
Comments: Criterion-level Characterizations



Interacting with LLMs one question at a time leads to more elaborate, specific, and helpful comments, particularly in IM 3.

- **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.

Comments: Criterion-level Characterizations



LLMs can be more specific than humans on assessing subjective criteria.

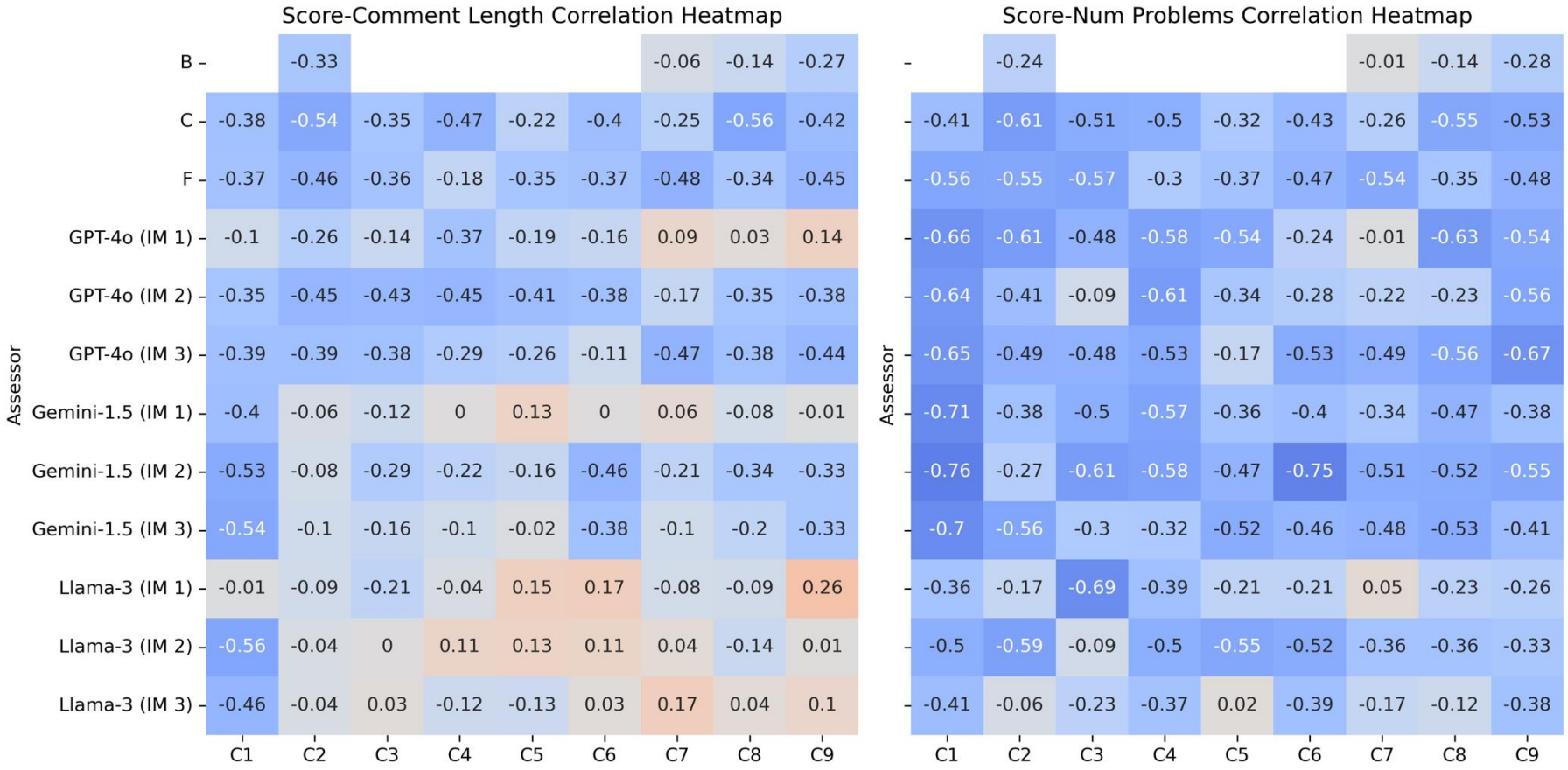
- **C1**: Material selection. **C2**: Material integration and citation; **C3**: Quality of key components.
- **C4**: Logic of structure. **C5**: Content and clarity of ideas. **C6**: Coherence (flow of ideas).
- **C7**: Cohesion (use of connectors). **C8**: Grammar and sentence structure. **C9**: Academic vocabulary.

Comments: Main Observations

- **Overall Observations**
  - LLMs always provide comments and identify problems, but humans do not
  - Interacting with LLMs one question at a time leads to more elaborate, specific, and helpful comments
  - LLMs can be more specific than humans on assessing subjective criteria

- **Criterion-Level Observations**
  - Human assessors tend to be relatively more specific on technical criteria
    - (1) C2: Material integration and citation; (2) C8: Grammar and sentence structure;
    - (3) C9: Academic vocabulary
  - Under Interaction Mode 3, LLMs are more specific than humans on subjective criteria
    - (1) C3: Quality of key components; (2) C4: Logic of structure; (3) C5: Content and clarity of ideas;
    - (4) C6: Coherence (flow of ideas); (5) C7: Cohesion or use of connectors
  - Both humans and LLMs are less specific on C1 (material selection)

Score-Comment Interaction: Expected



- **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.

Summary

- **LLMs can generate reasonably good multi-dimensional analytic assessments**
  - Scores: Can approximate human-assigned scores, typically within 1 point
  - Comments: Can identify more relevant, specific writing problems than human assessors, particularly on subjective criteria
  - Score-Comment Correlation: Expected negative correlations observed in human/LLM-generated assessments

- This is particularly true when LLMs are prompted in IM 3 where each assessment question is asked independently of each other.

Further Analyses

Re-examining Our Assumption about Feedback Comment Quality

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

Spearman Rank correlations between the specificity and helpfulness scores assigned by o1-mini and the number of different types of problems identified by our framework under different conditions.

Results

Scoring reliability metrics

Comment reliability metrics

	Condition	QWK	AAR1	BLEU	ROUGE-L	BERTScore
GPT-4o-Aug, IM1, Default	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-Aug, IM1, Default	GPT-4o-May-IM2	0.81	0.99	0.15	0.29	0.70
	GPT-4o-May-IM3	0.83	1.00	0.20	0.31	0.71
Llama-3, IM1, Default	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

Conclusion

Conclusion

- **LLMs can generate reasonably good and generally reliable multi-dimensional analytic assessments**
  - Promising tools for assessing academic English writing
  - Pedagogical potential: self-regulated learning for L2 learners; teaching assistance for instructors

- **Our feedback comment quality evaluation framework is effective and interpretable and can potentially serve as an alternative to human/LLM direct judgments.**

Contributions

Finding

- We provide comprehensive empirical evidence that LLMs can generate reasonably good and generally reliable multi-dimensional analytic writing assessments.

Corpus

- We release a corpus of L2 English graduate- level literature reviews, annotated with multi-dimensional analytic assessments.

Framework

- Proposes and validates a novel framework for evaluating feedback comment quality, more interpretable, cost-efficient, scalable, and reproducible than manual evaluation methods.

Limitations

Indirectness

- Our feedback comment quality evaluation is indirect
- A large-scale manual evaluation remains necessary

Insufficiency

- Insufficient validations for the proposed feedback comment evaluation framework
- Insufficient experimentations for the reliability evaluation

Comprehensiveness

- No qualitive analysis given the complexity and data-driven nature of the study

Thank You!

@ZhengXian9_Wang
zhengxiang.wang@stonybrook.edu