Human and LLM Ratings in L2 Speaking Assessment: Task-Conditioned Divergences in KISTEC

Taishi Chika, Yusuke Tanaka, Katsunori Kanzawa, Ayaka Setoguchi
University of Toyama, Kumamoto Gakuen University, Kyoto Institute of Technology, Meiwa Elementary School
3190 Gofuku, Toyama 930-8555, Japan
tchika@las.u-toyama.ac.jp, yu-tanaka@kumagaku.ac.jp, kanzawa@kit.ac.jp, guchi.a.chan7@gmail.com

Abstract

This study investigates divergences between human and Large Language Model (LLM) ratings in second language (L2) speaking assessment, focusing on task-dependent scoring patterns. Although recent studies have demonstrated the potential of LLMs for automated scoring, relatively little attention has been paid to how agreement varies across speaking task types or to the direction of score divergence. Using the KIT Speaking Test Corpus (KISTEC), comprising 5,166 transcribed responses from 574 university students in Japan, we compared human Task Achievement (TA) ratings, which assess how well a response fulfils the task, with ratings generated by GPT-4o mini across nine speaking tasks. We examined agreement, directional divergence, and representative examples. Overall agreement was consistent with previous findings (mean r = .725), although substantial variation emerged across tasks (r = .516–.849). The LLM tended to assign higher scores in Imagine tasks but lower scores in Compare and plan-and-organize tasks. Qualitative analyses suggest that these divergence patterns reflect differences in evaluative focus: human raters emphasized overall task accomplishment, whereas the LLM placed greater weight on the explicit realization of task requirements. These findings indicate that human–LLM disagreement in TA scoring is task-dependent and should be examined using task-specific and directional analyses.

Keywords: LLM Evaluation, L2 Speaking Test, Task Achievement (TA), The KIT Speaking Test Corpus (KISTEC)

1. Introduction

This study investigates divergences between human and Large Language Model (LLM) ratings in second language (L2) speaking assessment, with a particular focus on task-dependent scoring patterns. Recent advances in LLMs have opened new possibilities for automated assessment by enabling more direct evaluation of response content than conventional feature-based approaches, with recent studies demonstrating their potential for assessing content-related aspects of L2 production (Uchida, 2024).

Against this background, the present study addresses two research questions; (i) whether the degree and direction of human–LLM divergence vary across different speaking task types; and (ii) whether such divergence can be explained by differences in task characteristics.

To address these questions, we compared human and LLM ratings using transcribed responses from the KIT Speaking Test Corpus (KISTEC; Kanzawa et al., n.d.). The analysis focused on Task Achievement (TA), which evaluates how well the response in question accomplished the task. We examined the extent and direction of score divergence across task types and conducted qualitative analyses to explore differences in evaluative focus between agents.

The results reveal systematic task-dependent divergence between human and LLM ratings, suggesting differences in how TA is evaluated across task types.

The remainder of this paper is organized as follows. Section 2 provides the background, with particular attention to recent LLM-based approaches to speaking assessment, Section 3 describes the data and methodology, Section 4 presents the quantitative results, and Section 5 discusses the divergence patterns through qualitative analyses of representative examples. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Background

Automated assessment of L2 speaking has traditionally relied on statistical models based on linguistic features such as lexical sophistication, grammatical complexity, and fluency features (e.g. filler, silent pause). Rather than evaluating response content directly, these approaches estimate human ratings from observable linguistic characteristics (cf. Zechner and Evanini, 2019). Although such models have achieved reasonable predictive performance especially in the aspects of delivery, they remain limited in assessing TA, which requires evaluating how adequately a speaker fulfils the communicative requirements of a task.

Recent advances in LLMs have introduced a different approach. Instead of predicting scores from surface linguistic features, LLMs can evaluate response content directly through natural-language prompts, making automated assessment of semantic aspects partially feasible. Among recent studies, Uchida (2024) investigated the validity of the LLM (GPT-4) for assessing learner production in the ICNALE GRA corpus (Ishikawa, 2023). The study reported a high correlation with human ratings in writing (up to r = .888), whereas agreement for speaking remained substantially lower (r = .561). The study further demonstrated that explicitly specifying the scoring rubric within the prompt improved agreement to r = .641 in speaking, suggesting that prompt engineering substantially influences LLM scoring performance.

Despite these promising results, several issues remain unresolved. First, previous studies have primarily evaluated LLM performance using overall agreement measures such as correlation coefficients. While these metrics quantify the degree of correspondence between human and LLM ratings, they do not characterize the extent or direction of disagreement. Consequently, systematic tendencies for an LLM to over- or under-score particular responses remain difficult to identify.

Second, relatively little attention has been paid to variation across speaking task types. Speaking tests typically consist of tasks with substantially different discourse structures, including Imagine, Summarization, Compare, and Problem Solving. Because these tasks require different kinds of information organization and communicative strategies, they may also involve different evaluative priorities. It therefore remains unclear whether human–LLM agreement is consistent across task types.

Finally, disagreement between human and LLM ratings should not be interpreted solely as evidence of model failure. In speaking assessment, observed divergences may instead reflect differences in evaluative focus between human raters and LLMs. For example, task-specific requirements may be weighted differently even when both rating agents assign scores according to the same rubric. Distinguishing these possibilities is essential for understanding the role of LLMs in speaking assessment.

Motivated by these gaps, the present study investigates human–LLM divergence in TA scoring from both quantitative and qualitative perspectives. Specifically, we examine the extent and direction of rating divergence across speaking tasks and explore whether the observed patterns can be interpreted in terms of differences in evaluative focus.

3. Data & Method

3.1 The KIT Speaking Test Corpus (KISTEC)

The present study used the KIT Speaking Test Corpus (KISTEC; Kanzawa et al., n.d.), a learner corpus consisting of spoken English responses produced by university students in Japan. The corpus contains 5,166 transcribed monologue responses collected from 574 first-year university students across three test versions (191 for Ver. 1, 191 for Ver. 2, and 192 for Ver. 3), each of whom was allocated 7 minutes and 45 seconds of response time across the nine tasks. In total, the corpus contains approximately 290,000 words. The KISTEC speaking test comprises nine instructions assigning a range of communicative speaking tasks. Responses are evaluated in terms of two dimensions: Task Delivery (TD), which primarily concerns fluency and delivery, and Task Achievement (TA), which concerns the extent to which the task is accomplished.

Table 1, which presents Task Instruction for each Task Type, summarizes the task structure adopted in KISTEC.

It should be noted that all responses were rated holistically on the official 0–5 KISTEC TA scale (Table 2), rather than by scoring individual task components separately. After training with example responses and accuracy screening, the original audio responses for each task were independently rated by a different pair consisting of one native and one non-native speaker of English (18 raters in total). Ratings differing by no more than one point were averaged, yielding scores in 0.5-point increments; discrepancies of two points or more were adjudicated by an experienced senior rater, whose score was used as the final TA score.

Q	Task Type	Task Instruction
1	Imagine	Imagine why the object is left here.
2	Imagine	Imagine what the person is thinking.
3	Compare	Compare two alternatives and justify your choice.
4	Summarization	How are the two speakers' opinions different?
5	Take Position	Which opinion do you support? Give reasons.
6	Summarization	What problem is the speaker facing?
7	Problem Solving	How would you solve the problem?
8	Plan and Organize	Explain how you would organize it.
9	Persuade	Explain why the listener should choose it.

Table 1: Task instructions for the nine speaking tasks[1]

Score	Task Achievement (TA)
5	The task is achieved, being developed with a satisfactory level of detail.
4	The task is mostly achieved, with some supporting detail in places.
3	The task is minimally or partially achieved, being supported with some basic detail.
2	The task is addressed, but there is no or very little supporting detail.
1	The task remains essentially unachieved, though there may be some relevant words.
0	There is no relevant contribution (e.g., the content is entirely unconnected to the topic).

Table 2: TA rubric adopted in KISTEC

3.2 Method

Human ratings were compared with scores generated by GPT-4o mini using a zero-shot prompt. For each response, the model received the task instruction, the TA rubric (Table 2), the student's transcribed response, and, for Imagine tasks, the corresponding stimulus image. The human raters and the LLM evaluated the same task content in general, although it was presented through different modalities[2].

To improve scoring consistency, four constraints were incorporated into the prompt. First, the model was instructed to act as a strict rater for an English-speaking test. Second, it was instructed to evaluate TA only, ignoring TD features such as pronunciation, fillers, pauses, and other

[1] The official task labels for Q4 and Q6 are I*dentify Different Values* and *Identify Problem*, respectively. In the present study, these tasks are jointly classified as *Summarization* because both require participants to summarize information presented in a dialogue. Full task descriptions are available on the official KISTEC website (https://kitstcorpus.jp).

[2] Human raters evaluated the original audio responses (publicly unavailable), whereas the LLM received transcribed responses as input. This difference in input modality should be considered when interpreting the results; however, the present analyses focus exclusively on TA rather than TD.

fluency-related phenomena. Third, the output was restricted to a single score from 0 to 5 in 0.5-point increments, without explanation[3]. Finally, the model was instructed to incorporate visual information when available. The complete prompt is reproduced below[4]:

> You are a strict rater for an English-speaking test. Score only Task Achievement on a scale of 0 to 5, in increments of 0.5. You're given: (i) Task Instruction, (ii) Rubric, and (iii) Response.
>
> CAUTION:
> - Evaluate only Task Achievement (content). Ignore delivery & fluency.
> - Output only a single score {0, 0.5, 1.0, ..., 5.0}. No explanation.
> - If an image is provided, include the information from the picture when evaluating the Task Achievement. If no image is provided, evaluate based only on the question and student response.

The scoring system was implemented using Google Apps Script (GAS) and the OpenAI API. To avoid potential context effects from previous interactions, each rating was processed in an independent session. The temperature was fixed at 0 to maximize scoring consistency.

3.3 Evaluation Metrics

Table 3 summarizes the primary evaluation metrics used in the present study.

Metrics	Description
Pearson's *r*	Score correlation
±1 agreement	Responses within ±1 point
RMSE	Root mean squared error
UPPER	≥1-point overscoring by the LLM
LOWER	≥1-point underscoring by the LLM

Table 3: Evaluation metrics used in the present study

First, Pearson's correlation coefficient (*r*) was calculated for each speaking task to examine the overall association between human and LLM ratings. Second, the ±1 agreement rate was computed as the proportion of responses for which the human and LLM scores differed by no more than one point. Third, root mean squared error (RMSE) was used to quantify the average extent of score differences.

In addition to these conventional measures, the present study introduced two directional indices to characterize systematic scoring tendencies. UPPER represents the proportion of responses for which the LLM score exceeded the human score by at least one point, whereas LOWER denotes the proportion of responses for which the LLM score was at least one point lower than the human score. Unlike correlation coefficients, these indices capture both the extent and the direction of disagreement, allowing systematic over- and under-scoring tendencies to be examined separately across task types.

Finally, responses were grouped according to their human TA scores to examine whether the extent and direction of human–LLM divergence varied across proficiency groups.

4. Results

4.1 Overall Results

Table 4 summarizes human–LLM agreement across the nine speaking tasks. Pearson's correlation coefficients (*r*) ranged from 0.516 to 0.849 (mean $r = 0.725$), broadly consistent with previous findings (Uchida, 2024). Likewise, the ±1 agreement rate remained high for most tasks, although noticeable variation was observed across task types, particularly in Compare and Plan-and-Organize tasks.

4.2 Task-dependent Divergence

While the overall metrics summarized in Table 4 indicate moderate agreement between human and LLM ratings, they do not fully capture the direction of score divergence. The rightmost columns of Table 4 report the directional indices UPPER and LOWER for each speaking task.

Imagine tasks (Q1–Q2) exhibited relatively high proportions of UPPER cases, indicating that the LLM more frequently assigned higher scores than human raters. In contrast, Compare (Q3) and Plan-and-Organize (Q8) tasks showed markedly higher LOWER values, indicating a tendency for the LLM to assign lower scores. The remaining tasks exhibited comparatively balanced distributions of UPPER and LOWER.

4.3 Proficiency Effects

Figure 1 illustrates the distribution of human–LLM divergence across three proficiency groups based on human TA scores (Low: 0–2.0, Mid: 2.5–3.0, High: 3.5–5.0). Agreement was generally high for lower-scoring responses but decreased especially among higher-scoring responses. This tendency was particularly pronounced for Compare and Plan-and-Organize tasks, whereas Imagine tasks showed relatively stable distributions across proficiency groups.

[3] For actual evaluation, each response was independently scored by two trained raters in 1-point increments. Because the final human score was calculated as the average of the two ratings, the resulting scores were represented in 0.5-point increments. Accordingly, the LLM was instructed to output scores on the same 0.5-point scale.

[4] The placeholders (i)–(iii) were automatically replaced at runtime with the corresponding task instruction, TA rubric, and transcribed student response. For Imagine tasks (Q1–Q2), the corresponding stimulus image was additionally supplied to the model; no image was provided for the remaining task types.

			Mean Scores		Association	Agreement		Divergence		
Q	**Task Type**	**Ver**	**Human**	**LLM**	***r***	**Exact**	**±1**	**RMSE**	**UPPER**	**LOWER**
1	Imagine	1	2.88	3.07	0.828	31%	96%	0.71	22%	7%
		2	2.62	3.15	0.802	20%	82%	0.93	41%	6%
		3	2.82	2.75	0.849	36%	92%	0.76	15%	21%
2	Imagine	1	3.14	3.16	0.772	29%	95%	0.77	21%	19%
		2	2.38	3.08	0.516	19%	72%	1.17	42%	5%
		3	3.20	3.58	0.808	21%	86%	0.84	31%	6%
3	Compare	1	3.22	2.01	0.700	4%	57%	1.36	1%	83%
		2	3.18	2.02	0.684	3%	56%	1.29	1%	81%
		3	3.41	2.03	0.765	2%	42%	1.49	0%	89%
4	Summarization	1	2.66	2.71	0.771	36%	96%	0.67	15%	11%
		2	2.17	1.80	0.827	34%	95%	0.68	3%	27%
		3	2.41	2.86	0.737	28%	88%	0.83	33%	5%
5	Take Position	1	2.81	2.12	0.702	12%	85%	0.91	2%	44%
		2	2.70	1.78	0.802	9%	70%	1.10	1%	63%
		3	2.70	2.20	0.771	24%	92%	0.71	0%	28%
6	Summarization	1	2.69	2.30	0.795	20%	93%	0.79	4%	28%
		2	3.02	2.96	0.758	34%	96%	0.66	10%	13%
		3	2.57	2.43	0.794	31%	94%	0.66	6%	18%
7	Problem Solving	1	3.02	1.88	0.786	6%	58%	1.30	0%	73%
		2	2.97	2.20	0.815	19%	85%	0.99	2%	58%
		3	2.65	1.99	0.822	23%	84%	0.90	1%	47%
8	Plan-and-Organize	1	3.29	1.83	0.527	8%	37%	1.79	6%	80%
		2	3.32	1.85	0.525	5%	38%	1.66	1%	87%
		3	3.10	2.13	0.603	10%	65%	1.21	2%	60%
9	Persuade	1	2.93	2.71	0.572	29%	86%	0.81	9%	22%
		2	3.23	2.72	0.616	29%	90%	0.79	2%	37%
		3	3.03	2.69	0.625	29%	96%	0.69	4%	25%

Table 4: The results of the rating

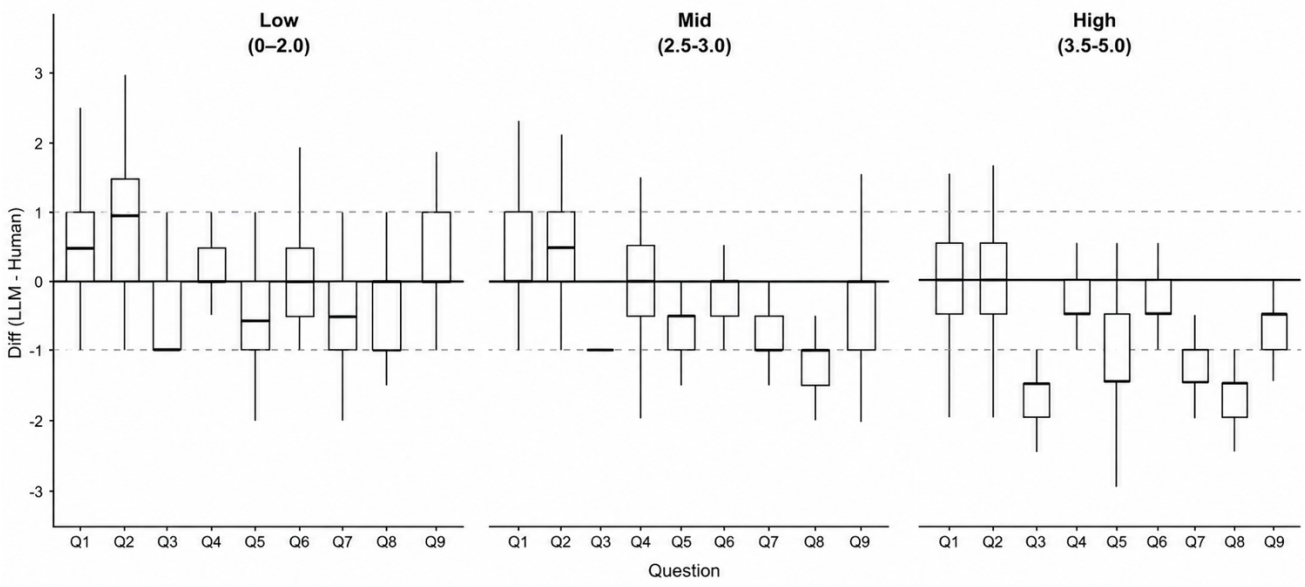

Figure 1: Distribution of human–LLM rating divergence across human TA score groups

5. Discussion

5.1 Interpreting Human-LLM Divergence

In Section 4, the quantitative analyses revealed systematic variation in human–LLM divergence across task types. Although such disagreement can be interpreted as evidence of model limitations, it is not the only plausible interpretation in the context of speaking assessment. Since TA requires evaluating the overall task accomplishment of responses rather than relying solely on linguistic features, divergence between agents may also reflect differences in how human raters and LLMs operationalize the criterion of TA. The following sections examine representative examples from three divergence patterns summarized in Table 5.

Pattern	Representative Task Types
Convergence	Predefined content to be summarized: Summarization (Q4, Q6)
UPPER	Open-ended inference from visual information: Imagine (Q1, 2)
LOWER	Integration of multiple task requirements: Compare (Q3), Plan-and-Organize (Q8)

Table 5: Representative task categories for qualitative analysis

These case studies are intended to illustrate possible interpretations rather than provide definitive explanations.

5.2 Convergence

The summarization tasks (Q4 and Q6) exhibited relatively high agreement between human and LLM ratings. Before examining tasks characterized by systematic UPPER and LOWER tendencies, we first consider this high-agreement pattern as a baseline for interpreting human–LLM divergence. According to the results shown in Table 4 and Figure 1, these tasks showed both high ±1 agreement rates and relatively small score differences across proficiency levels, which provide a useful reference for identifying conditions under which the two rating agents apply similar evaluative criteria.

Because the essential content to be evaluated is largely specified in the task itself, both human raters and the LLM appear to evaluate similar aspects of response quality. Consequently, differences in evaluative focus are less likely to emerge than in tasks involving more flexible discourse organization. The example below presents a representative response from the summarization task (Q4), for which the human and LLM ratings differed by only 0.5 points.

> Mariam did not submit her assignment on time. And, the professor criticized her. Uh, Mariam thinks that is very, uh, correct for the professor to be angry at her because it is her fault and, uh, the professor just cares about her if she does not study hard. But Bill thinks that the professor is just treating them like children, and, um, they are all adults and they know what they're doing.

(Ver. 2, Q4, 18123046.txt; Human = 4.5; LLM = 4.0)

In this task, the communicative requirements are largely determined by the dialogue itself. A successful response is expected to summarize predefined information, including the central event and the viewpoints expressed by the two speakers. Because the content to be evaluated is explicitly specified by the task, both human raters and the LLM appear to attend to largely the same communicative elements. As a result, differences in evaluative focus are less likely to emerge than in tasks requiring speakers to generate or integrate their own ideas.

This example suggests that convergence between human and LLM ratings is more likely when the information structure is constrained by the task itself, leaving less room for systematic differences in TA evaluation.

5.3 UPPER

In contrast, the Imagine tasks (Q1 and Q2) exhibited relatively frequent UPPER cases, particularly among lower-scoring responses (0─2.0 in TA).

Figure 2. Imagine what the man is thinking. (Ver. 2; Q2)

> The elderly man is uh waiting for a calling from his family because mm he has not meet his family for a long time. And this day, uh, today, eh the family want to call him, and also he want call family. But then, the calling was not come here. He the man even waiting and not calling.

(Ver. 2, Q2; 18123017.txt; Human = 1.0, LLM = 4.0)

For comparison, the next example received similar scores from both rating agents.

> I think that the man is thinking about his business. There's some reason for this. First, ah I think the man is in his office because there is I think telephone and some some shelve and he wearing suits tie and he is sitting on in front of desk, so I think he's in his office, so he's thinking about his work or business. I think he has some trouble because he is not happy I think. So he has a problem or his business and thinking about it.

(Ver 2. Q2; 18151164.txt; Human = 4.0, LLM = 4.0)

One possible interpretation is that human raters evaluated not only the plausibility of the imagined scenario but also the extent to which the inference was supported by the visual stimulus. The LLM, in contrast, appeared to assign relatively high scores when the inferred scenario was internally coherent, even when the supporting visual evidence was comparatively limited. Importantly, this does not necessarily imply that the LLM ignored the image itself. Rather, the two rating agents may have differed in how much inferential elaboration they considered acceptable for fulfilling the task requirement of imagining the speaker's thoughts.

These observations suggest that the UPPER tendency in Imagine tasks may reflect different interpretations of what constitutes successful task completion. Whereas human raters appeared to emphasize well-grounded inferences supported by the visual stimulus, the LLM appeared to tolerate broader but coherent elaborations, particularly for lower-scoring responses.

5.4 LOWER

Compare (Q3) and Plan-and-Organize (Q8) tasks exhibited the highest proportions of LOWER cases. As shown in Figure 1, this tendency became increasingly pronounced among responses receiving relatively high human ratings. Unlike the tasks discussed in the last section, these tasks require speakers to integrate multiple content elements into a coherent response. The divergence may therefore reflect differences in how human raters and the LLM evaluate these requirements.

The next example illustrates a representative response from the comparison task (Ver. 3; Q3) *Where would you take your guests from abroad, to the countryside or to a big city? Explain the reasons for your choice, comparing the advantages and disadvantages of both*.

> If I were to take my guests from abroad, I would take them to I would rather take them to the countryside than to a big city. Umm, I would I chose the countryside because it is I think that Japan is all about nature. And in the big city, it does include a lot of subcultures in Japan, however, I think the countryside is portrays Japan's potential beauty and its nature has been left from a long time ago and it is really pretty.

(Ver 3. Q3; 18161017.txt; Human = 5.0, LLM = 2.0)

Although the response clearly expresses a preference for the countryside and contrasts it with the attractions of the city, it does not explicitly address several task components specified in the instructions, such as the disadvantages of each option. Human raters nevertheless appeared to evaluate the response favourably because it presented a coherent argument supporting the speaker's choice. The LLM, by contrast, appeared to assign greater weight to whether the individual task components were explicitly realized, resulting in a substantially lower score.

A similar pattern was observed in the Plan-and-Organize task (Ver.1; Q8): *You have been asked to make a promotional video of your university. Explain how you would organize it*.

> I want to get, at first, I I have two thing I want to do to make promotional video. At first, I want to get scenery of my university. Scenery is important. If it is and not good, many people don't want to go. So I want to get beautiful scenery on in this cam campus. And second thing is I want to get I want to do interview to many people. And they're they have they have different major and so I want I want to know about their ai idea to to their major and it is it gives many information to some some people to do to see it. So I want to organize it to.

(Ver 1. Q8; 18141051.txt; Human = 5.0, LLM = 2.0)

For comparison, the following example presents a more explicit plan for organizing a promotional video.

> First, I ask to my friend to organize it together. Next, I talk to some teacher and some student to do things and ask to interview about the Kyoto Institute of Technology university and I take a video around the university. And I want to especially take a movie about a club activity, for example, the tennis club, track and field club, volleyball club, music club, um and so on. The club activity is featured the color of the university, I think.

(Ver 1. Q8; 18141003.txt; Human = 5, LLM = 4)

Both responses identify appropriate content for a promotional video. However, the former example provides relatively limited information about how the proposed video would be organized, whereas the latter explicitly presents a sequence of actions together with concrete filming targets. Human raters appeared to regard both responses as communicatively successful, whereas the LLM assigned a considerably lower score, possibly because several task components remained only partially specified.

Taken together, these examples suggest that the observed LOWER tendency may not simply reflect poorer LLM performance. Rather, it may arise from differences in the evaluative focus adopted by the two rating agents. Human raters appeared to assess the overall task accomplishment of a response, whereas the LLM placed greater emphasis on the explicit realization of individual task components. As the number of required components increases, differences in how incomplete or partially expressed information is evaluated become more likely to influence the final score. This interpretation also explains why LOWER cases were concentrated among higher-scoring responses. In such responses, the overall communicative goal had largely been achieved, although some task requirements remained only implicitly or incompletely expressed.

6. Conclusion

This study investigated divergences between human and LLM ratings of TA in L2 speaking assessment. Using KISTEC, we examined not only the overall association between human and LLM ratings but also the extent and direction of score divergence across different speaking task types. Although overall agreement was comparable to that reported by Uchida (2024), substantial task-dependent divergence was observed.

In particular, the LLM tended to assign higher scores for Imagine tasks but lower scores for Compare and Plan-and-Organize tasks, especially among responses receiving relatively high human ratings.

The qualitative analyses further suggested that these divergence patterns may reflect differences in evaluative focus between human raters and the LLM. Rather than interpreting all disagreement as evidence of model limitations, the present study hinted that task structure may influence how TA is interpreted by different rating agents. These findings highlight the importance of examining task-specific and directional divergence, rather than relying solely on overall agreement measures, when evaluating LLM-based speaking assessment.

One limitation of the present study is its use of a single LLM. Consequently, it remains unclear whether the observed divergence patterns reflect model-specific behaviour or more general characteristics of LLM-based assessment. Future research should therefore examine whether similar patterns emerge across different LLMs and model generations[5].

In addition, the proposed interpretations should be tested more systematically through controlled interventions. One possible approach is to manipulate individual components of a response and examine how such modifications affect scores assigned by the same model. Such analyses may help identify which aspects of task accomplishment contribute to human–LLM divergence and inform task-sensitive approaches to LLM-based scoring.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22K00736. We would like to thank an anonymous reviewer for constructive comments.

8. Bibliographical References

Ishikawa, S. (2023). *The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English*. Routledge.

Koizumi, R. and In'nami, Y. (2022). Assessing functional adequacy using picture description tasks in classroom-based L2 speaking assessment. *JLTA Journal*, 25:60–79.

Kuiken, F. and Vedder, I. (2022). The assessment of functional adequacy in language performance. *Journal on Task-Based Language Teaching and Learning*, 2(1):1–7.

Uchida, S. (2024). Evaluating the accuracy of ChatGPT in assessing writing and speaking: A verification study using ICNALE GRA. *Learner Corpus Studies in Asia and the World*, 6:1–12.

Zechner, K. and Evanini, K., (Eds.), (2019). *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous* Speech. Routledge.

9. Language Resource References

Kanzawa, K., Kobayashi, Y., Lee, J., Mitsunaga, H., Mori, M., Tanaka, Y., and Chika, T. (n.d.). *The KIT Speaking Test Corpus (KISTEC)*. https://kitstcorpus.jp.

[5] A preliminary analysis using GPT-5 (with minimal reasoning effort) followed the same experimental procedure, except that the temperature parameter was not configurable for GPT-5. Higher correlations with human ratings were observed, while the pronounced LOWER tendency in Compare and Plan-and-Organize tasks persisted. These results are not included in the main analysis and should be interpreted cautiously.