# Comparison of Generative AI and Peer Assessment in Essay Evaluation: Preliminary Study

**Yasuhisa TAMURA**[a*]
[a]*Faculty of Science and Technology, Sophia University, Japan*
*ytamura@sophia.ac.jp

**Abstract:** This paper addresses the timely and critical issue of evaluating students' essays using generative AI, particularly by leveraging student peer assessment as a gold standard, departing from traditional teacher-centric evaluations. While essay submission is vital for knowledge consolidation and logical thinking, evaluation remains time-consuming and subjective. Recent advancements in generative AI offer potential solutions for efficient and objective assessment. Prior research has shown moderate to high agreement between generative AI and human (teacher) graders across various contexts, often with detailed rubrics and prompts. However, this study uniquely focuses on student peer evaluations, recognizing their debated but increasingly accepted reliability. This research aims to answer two key questions: (RQ1) Can generative AI, when guided by rubrics, produce evaluation results similar to student peer assessments for natural language outputs? (RQ2) How do evaluations differ across various generative AI models? To address these, we will quantitatively analyze the characteristics of both student peer evaluations and multiple generative AI models. We will also investigate the impact of different rubric description formats (conceptual vs. example-inclusive), focusing on "Potemkin understanding" in AI. A preliminary experiment involving 117 undergraduate research plans, evaluated by five peers and five generative AIs (Gemini Flash, Gemini Pro, ChatGPT 4o, ChatGPT o3, Claude Sonnet 4), showed correlation coefficients (r=0.677-0.698, excluding ChatGPT 4o), which are below the reliability threshold of r > 0.8, indicating a need for improved rubric descriptions. This study proposes a shift from the "automated grading with teacher grades as true value" paradigm to a social constructivist view, recognizing assessment as part of the learning activity. Academically, it will pioneer a new field of multi-stance learning by utilizing diverse assessment language from learners as "weak teachers" for AI. Practically, it promises to improve immediate feedback quality in large lectures and MOOCs, foster assessment literacy, and address regulatory concerns regarding AI opacity and the absence of human judgment by adopting a hybrid structure of student peer evaluation and generative AI. This approach will also facilitate the development of localized evaluation models reflecting cultural and linguistic diversity.

**Keywords:** Generative AI, Automated Essay Scoring, Peer Assessment, Rubrics

## 1. Introduction

In the educational field, requiring students to submit essays and reports is an effective method for solidifying knowledge and improving logical thinking skills. It has been used for a long time. However, the task of evaluating these deliverables requires a great deal of time and effort, making it difficult to maintain the objectivity and fairness of the evaluation. To solve this problem, efforts have been made to use rubrics to perform objective, efficient, and fair assessments (Brookhart, 2013).

Meanwhile, with the rapid development of generative AI, including ChatGPT, its use has begun in various fields of education. Samala (2025) surveyed 453 previous studies and classified the use of generative AI into the following six areas:

- Content generation (syllabus, teaching materials, question creation)
- Dialogue and personalized learning support (AI tutors, Q&A bots)
- Evaluation and feedback (automated grading, formative assessment)
- Educational management (grade management, data analysis, and administrative efficiency)
- Research and creative support (paper drafting, idea generation)
- Ethics and governance (fairness, academic integrity, policy)

In this study, we focused on summative assessment of essays and reports. There is a wealth of prior research on summative assessment using generative AI. Most of this work has shown that generative AI can perform summative assessment of student essays with moderate to high agreement with human graders. e.g., undergraduate psychology (Wetzler, 2025), jurisprudence (Alimardani, 2024), and English (Escalante, 2023). Consistent agreement has been achieved across a variety of contexts, using detailed and analytical rubrics and carefully designed prompts. Mizumoto (2023) reported a 54.33% accuracy rate with human evaluation in automated essay evaluation using GPT-3. Quah (2024) found a strong correlation between ChatGPT-4–based evaluations and instructor grading in dental school examinations (r = 0.752–0.848). Many of these previous studies compared teacher evaluations with AI evaluations.

In this study, we focus on student peer evaluations of work, rather than teacher evaluations, which are used as indicators in the previous studies. There has been much debate regarding the reliability and validity of this peer evaluation. Related research is divided into reports that "student peer evaluations are comparable (reliable) with teacher evaluations" and reports that "the reliability and validity are questionable." As examples of the former, Fukazawa (2010) cites three studies that analyzed the correlation between teacher evaluations and peer evaluations and verified their validity, finding a high correlation. The correlation coefficients between teacher evaluations and peer evaluations in each study were r=0.68 - .80 (Miller, 1996), r=0.83 (Hughes, 1993), and r=0.89 (Stefani, 1994).

In contrast, examples of reports that question the reliability and validity of student peer assessment include Stefani (1994), who found that "the mean values of peer assessments tend to be higher than those of faculty assessments, and the standard deviations of peer assessments tend to be smaller than those of faculty assessments, " and Freeman (1995), who reported that "the mean values of peer assessments tend to be higher than those of faculty assessments, and the standard deviations of peer assessments tend to be smaller than those of faculty assessments." This study will proceed based on the former position, that "student peer assessments are reliable." Based on the above, this study will pose the following two research questions (RQs).

- RQ1: When summarizing and evaluating natural language outputs such as essays and reports, is it possible to use rubrics generated by generative AI? Will the results of the student evaluation be similar to those of peer evaluation?
- RQ2: When summarizing and evaluating natural language artifacts such as essays and reports, what is the effectiveness of different generative AI models? Isn't the evaluation different?

To clarify the RQs above, this study sets the following objectives:

- Quantitatively analyze the characteristics of student-to-student rubric evaluation and the evaluation characteristics of multiple types of generation AI. Analyze and clarify the differences in natural language understanding and evaluation characteristics of generative AI.
- Prepare multiple rubric descriptions to minimize discrepancies between students' peer evaluation results and the evaluations generated by the AI.

In particular, we will focus on the "Potemkin understanding" (generative AI can explain concepts, but has difficulty using them to reason, apply, and practice) pointed out by Mancoridis (2025), and clarify the difference in evaluation between rubrics that only include concepts and rubrics that include examples.

This research shifts the conventional paradigm of "automated grading with teacher grades as the true value". It embodies a social constructivist view of assessment that considers assessment "part of the learning activity" by adopting student-to-student peer

assessment as the gold standard. Academically, it pioneers a new field of multi-stance learning, utilizing the diverse assessment language generated by learners as a "weak teacher" for generative AI, with ripple effects on the theoretical development of both educational technology and natural language processing. In terms of implementation, it has the direct benefit of improving the quality of immediate feedback in large-scale lectures and MOOCs without increasing the burden on instructors, and also fostering assessment literacy and metacognitive abilities in the process of students critically examining AI output, which is a direct benefit of educational digital transformation. Ofqual (2024) (The Office of Qualifications and Examinations Regulation) announced that scoring solely by AI does not meet regulatory requirements and should be limited to auxiliary use. By contrast, mainstream approaches still treat teacher-assigned grades as the sole ground truth for training. This research, by adopting a hybrid structure of student peer evaluation and generative AI, is differentiated by being one of the first attempts to simultaneously resolve the "opacity of AI" and "absence of human judgment" concerns raised by regulatory agencies. Furthermore, by incorporating peer evaluation data into AI training, it is possible to build a localized evaluation model that reflects cultural and linguistic diversity, establishing a competitive advantage for international
expansion as a Japanese educational evaluation solution.

Previous research related to the above RQ1, "When summatively evaluating natural language outputs such as essays and reports, will rubric evaluation using generative AI produce results similar to those of student peer evaluation?" is Banihashem (2024). Generated peer feedback and feedback using ChatGPT-3.5 for argumentative essays written by 74 graduate students using the same prompts, and compared the feedback content using qualitative coding and MANOVA. ChatGPT provided ample insight into aspects of explanatory and sentence structure, while students excelled at problem identification, and the two were complementary. There was no significant correlation between essay quality and feedback quality. Usher (2025). The group projects of 76 undergraduate students were evaluated by (1) an AI chatbot (GPT-4), (2) peers, and (3) instructors, and scores and feedback quality were compared using a mixed-methods study. The AI consistently awarded higher scores than human evaluators, and the feedback was detailed but occasionally off-topic. Peer and instructor ratings were similar, revealing a high degree of individuality. The previous studies primarily focused on the quality of feedback, not summative evaluation.

Regarding RQ1, this study and previous studies share the commonality of comparing students' peer assessments with the generative AI's assessments and dealing with written output such as reports. However, this study differs in that it focuses on the degree of agreement of rubric scores. In contrast, Banihashem (2024) focuses on the content of feedback, and Usher (2025) focuses on both score differences and feedback. Furthermore, this study differs in that it focuses on the grades and agreement rate for each rubric item. At the same time, Banihashem (2024) uses three-dimensional (cognitive, affective, and constructive) coding, and Usher (2025) focuses on score distribution and feedback detail.

Regarding RQ2, "Do evaluations differ between different models of generative AI?", previous research has compared these models (Seßler, 2025; Yavuz, 2025). The results show that commercial and customized systems outperform initial and open-source systems. While generative AI performs best in the areas of language and mechanics, it has been noted to be less consistent in judging content quality and critical analysis. Several papers have reported biases toward more lenient evaluations of low-performing essays and more strict evaluations of high-performing essays. Regarding RQ2, this study and previous studies share the point that "student work is graded by multiple generative AIs and compared with each other." However, while previous studies evaluate generative AI based on teacher evaluation, this study differs in that it uses students' mutual evaluation as the standard.

While the primary outcome of [1 Research Objectives, Research Methods, etc. (continued)] is to support teachers by "reducing the burden on teachers using AI," this research differs in that it aims to present operational guidelines and rubric templates as assignments that complement/expand student-led peer assessment.

The first objective of this study is to empirically clarify the extent to which summative assessments by generative AI can coincide with student peer assessments. Specifically, we will compare the rubric assessment results of generative AIs and students on approximately 300 essays and reports assigned in lecture courses, and clarify the distribution of assessment agreement and its determining factors by analyzing correlation coefficients and cluster structures.

The second objective is to clarify an effective rubric description format for increasing the aforementioned agreement. We set two conditions: a rubric that lists only conceptual terms, and a rubric that also lists specific examples, and estimate the discrepancy between the generative AI and student assessments using a Bayesian hierarchical model. This will enable us to quantitatively identify which items exhibit the "Potemkin understanding" proposed by Mancoridis (2025). Through these analyses, this research will:

(1) define the differences in the natural language comprehension and evaluation characteristics of each generative AI model,
(2) extract the rubric structure (vocabulary level, example position, number of scales) that maximizes evaluation agreement, and
(3) propose a "standard rubric template" that can be immediately used in educational settings.

Finally, we will pilot the application of this rubric to not only humanities and social science assignments but also STEM reports to verify cross-sectional validity, thereby demonstrating the academic and practical scope of the feasibility of summative evaluation using generative AI.

## 2. Preliminary Experiment

### *2.1 Experiment Settings*

We conducted a preliminary experiment to verify the feasibility of this research. In the Spring 2025 semester, a specialized course for third-year undergraduate students, "Introduction to Human Informatics Research," 117 undergraduate students of Sophia University, Japan, wrote research plans (two pages of A4) based on their interests. Five other students evaluated the plans using a rubric. The rubric consisted of nine items and assessed the style, expression, logical consistency, and other aspects of the research plans. The concrete rubrics are shown in Appendix. Five generative AIs (Gemini Flash, Gemini Pro, ChatGPT 4o, ChatGPT o3, and Claude Sonnet 4) were also given the same rubric to evaluate the 117 students' research plans. Files of the research plan were input manually (not via API) to reduce the workload of developing APIs.

### *2.2 Results*

The four correlation coefficients, as shown in Figure 1, were r = 0.677-0.698, except for ChatGPT 4o. These were below the r > 0.8 required for reliability, and further improvements to the rubric description are necessary. We also tried changing the instructions (prompts) given to the generating AI, such as "evaluate strictly," but this did not result in a significant change in the correlation coefficient.

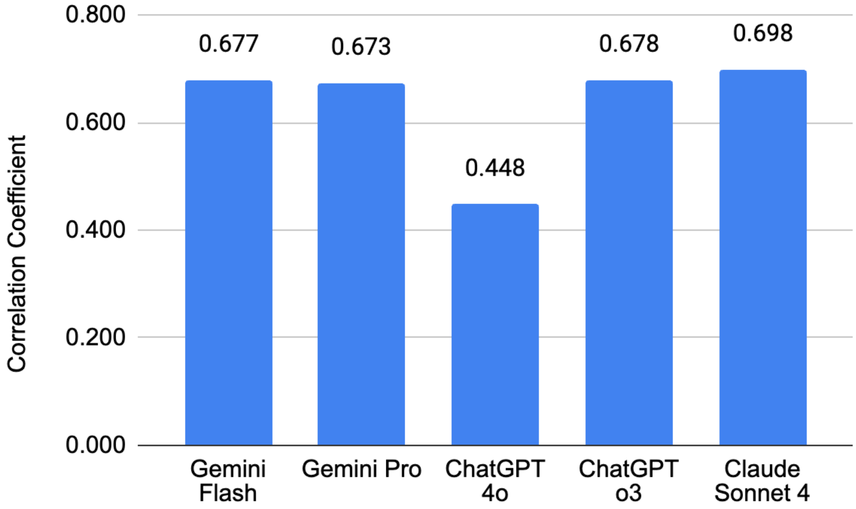

*Figure 1.* Preliminary Experiment Result

Furthermore, we examined the variance of the evaluations from five types of generative AI and the variance of mutual evaluations by students for each subject's evaluation. Figure 2 shows a scatter plot of these two types of variance values. Incidentally, the variance of these values was 114.48 for generative AI and 226.56 for mutual evaluation. This indicates that the evaluation results of generative AI are more concentrated around the average value than the results of mutual evaluation, meaning there is less difference in evaluations among the generative AIs.

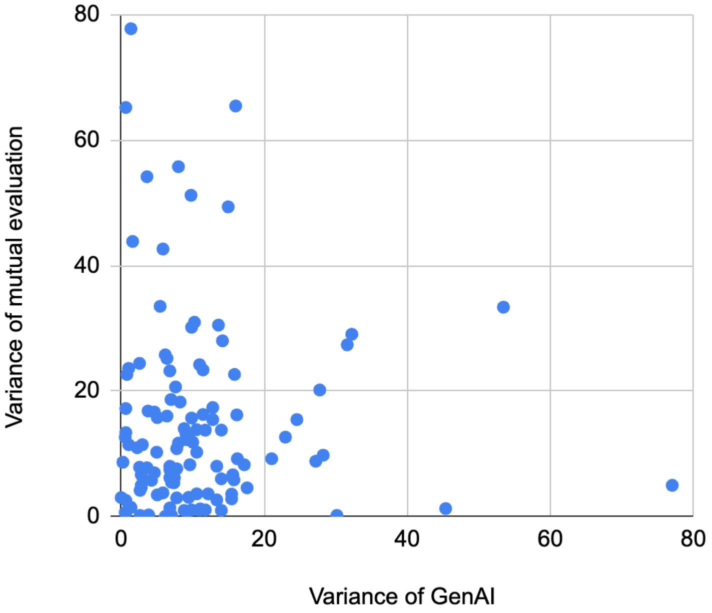

*Figure 2.* Scatter plot of variances between mutual evaluation and generative AI

## 3. Conclusion and Future Works

The preliminary experiment, which revealed correlation coefficients between human peer assessment and generative AI evaluations ranging from r=0.677-0.698 (excluding ChatGPT 4o), falling below the desired reliability threshold of r > 0.8, underscores the critical need for further refinement in our methodology. This initial finding directly informs and strengthens the objectives of our future work, which aims to address these discrepancies and develop more robust assessment tools.

Our immediate future work will focus on two primary objectives, directly building upon the insights gained from the preliminary study:

- Empirically Clarifying Agreement between Generative AI and Student Peer Assessments: We will expand our empirical investigation to approximately 300 essays and reports from lecture courses. This larger dataset will allow for a more comprehensive analysis of the correlation coefficients and cluster structures of assessment agreement. By comparing the rubric assessment results from a central generative AI with those from students, we aim to identify the distribution of agreement and its determining factors precisely. This will directly address the lower-than-expected correlations observed in the preliminary experiment, seeking to understand the nuances of agreement and disagreement.
- Identifying Effective Rubric Description Formats to Enhance Agreement: A key lesson from the preliminary experiment is the necessity for improved rubric descriptions. Therefore, our second objective is to clarify an effective rubric description format that maximizes agreement between generative AI and student assessments. We will specifically compare two conditions: rubrics that list only conceptual terms versus rubrics that also include specific examples. Utilizing a Bayesian hierarchical model, we will quantitatively estimate the discrepancies between generative AI and student assessments. This detailed analysis will enable us to pinpoint which rubric items are susceptible to "Potemkin understanding" by generative AI, as highlighted by Mancoridis (2025). The goal is to extract the optimal rubric structure (considering vocabulary level, example positioning, and scale numbers) that maximizes evaluation agreement.

Through these analyses, this research will:

- Define the differences in natural language comprehension and evaluation characteristics of each generative AI model, contributing to a more nuanced understanding of their capabilities and limitations in assessment.
- Propose a "standard rubric template" that is immediately usable in diverse educational settings, addressing the practical need for reliable AI-supported assessment tools.

Finally, we will pilot the application of this refined rubric not only to humanities and social science assignments but also to STEM reports. This cross-sectional validation will demonstrate the broader academic and practical applicability of our findings regarding summative evaluation using generative AI. By iteratively refining the rubric and analyzing a larger dataset, we expect to significantly improve the agreement between AI and human peer evaluations, moving closer to a truly synergistic assessment system that leverages the strengths of both.

As described above, the proposed approach aimed for summative assessment. However, this approach will also apply to formative assessment: to give feedback comments to learners for their draft essays. We will try to enhance the function for the formative assessment in the proposed system.

## References

Alimardani, A. (2024). Generative artificial intelligence vs. law students: An empirical study on criminal law exam performance. Law, Innovation and Technology, 16(2), 777–819. https://doi.org/10.1080/17579961.2024.2392932

Banihashem, S. K., Noroozi, O., Taghizadeh Kerman, N., Karami, M., & Biemans, H. J. A. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? International Journal of Educational Technology in Higher Education, 21(1), 1–23. https://doi.org/10.1186/s41239-024-00455-4

Brookhart, S. M. (2013). How to create and use rubrics for formative and summative assessment. ASCD. ISBN: 978-1-4166-1507-1

Escalante, J. L., Lim, J. C., Datu, J. P., & Ocon, V. L. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. International Journal of Educational Technology in Higher Education, 20(1), 1–17. https://doi.org/10.1186/s41239-023-00425-2

Freeman, M. (1995). Peer assessment by groups of group work. Assessment & Evaluation in Higher Education, 20(3), 289–300. https://doi.org/10.1080/0260293950200305

Fukazawa, M. (2010). Validity of Peer Assessment of Speech Performance, Annual Review of English Language Education in Japan (ARELE), 21 (0), 181-190. (in Japanese)

Hughes, I. E., & Large, B. J. (1993). Staff and peer-group assessment of oral communication skills. Studies in Higher Education, 18(4), 379–385. https://doi.org/10.1080/03075079312331382281

Mancoridis, M., Weeks, B., Vafa, K., & Mullainathan, S. (2025). Potemkin understanding in large language models. arXiv. https://doi.org/10.48550/arXiv.2506.21521

Miller, L., & Ng, R. (1996). Autonomy in the classroom: Peer assessment. In R. Pemberton, E. S. L. Li, W. W. F. Or, & H. D. Pierson (Eds.), Taking control: Autonomy in language learning (pp. 133–146). Hong Kong University Press.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring, Research Methods in Applied Linguistics, 2(2), 100050, https://doi.org/10.1016/j.rmal.2023.100050

Ofqual (2024). Policy paper: Ofqual's approach to regulating the use of artificial intelligence in the qualifications sector, https://www.gov.uk/government/publications/ofquals-approach-to-regulating-the-use-of-artificial-intelligence-in-the-qualifications-sector/ofquals-approach-to-regulating-the-use-of-artificial-intelligence-in-the-qualifications-sector (retrieved 2025-8-12)

Quah, W. F., Zheng, Y., Sng, T. H. G., Yong, Y. S. W., & Islam, S. S. (2024). Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. BMC Medical Education, 24(1), 796. https://doi.org/10.1186/s12909-024-05881-6

Samala, A. D., Rawas, S., Wang, T., Reed, J. M., Kim, J., Howard, N. J., & Ertz, M. (2025). Unveiling the landscape of generative artificial intelligence in education: a comprehensive taxonomy of applications, challenges, and future prospects. Education and Information Technologies, 30(3), 3239-3278.

Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In Proceedings of the 15th International Learning Analytics and Knowledge Conference (pp. 462-472).https://doi.org/10.1145/3706468.3706527

Stefani, L. A. (1994). Peer, self and tutor assessment: Relative reliabilities. Studies in Higher Education, 19(1), 69–75.

Usher, M. (2025). Generative AI vs. instructor vs. peer assessments: a comparison of grading and feedback in higher education. Assessment & Evaluation in Higher Education, 1–16.

Wetzler, E. L., Cassidy, K. S., Jones, M. J., Frazier, C. R., Korbut, N. A., Sims, C. M., ... & Wood, M. (2025). Grading the graders: Comparing generative AI and human assessment in essay evaluation. Teaching of Psychology, 52(3), 298-304. https://doi.org/10.1177/00986283241282696

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. British Journal of Educational Technology, 56(1), 150-166. https://doi.org/10.1111/bjet.13494

## Appendix: Research Proposal Evaluation Rubric

#1: Clarity of Problem Statement and Background
0: The research topic and background are unclear, and no context is provided.
1: The topic is stated, but the explanation of the background remains ambiguous.
2: The background and context are understandable but lack sufficient persuasiveness.
3: The research topic is clearly defined, and the background is explained logically.

#2: Literature-Based Explanation of Background and Theory
0: No literature is cited, and the theoretical basis is unclear or not presented.
1: Literature is cited, but the theoretical explanation is superficial or inaccurate, or the reliability of sources is questionable.
2: Literature is cited, but the relationship to the hypothesis or the handling of theory is somewhat unclear.
3: The theory and underlying assumptions are clearly explained based on reliable literature, and their connection to the hypothesis is logically established.

#3: Logical Consistency of Research Objectives and Hypotheses

0: The objectives or hypotheses are unclear or lack consistency.
1: The objectives or hypotheses are stated, but the connection between them is weak.
2: The objectives and hypotheses are generally consistent.
3: The objectives and hypotheses are clearly stated and logically connected.

#4: Originality Based on Comparison with Previous Research
0: Little or no reference is made to previous research, and no claim of originality is presented.
1: Previous research is introduced, but the differences from the present study or its originality are unclear, resembling a mere list.
2: Differences from previous research are presented to some extent, but clarity and strength of claim are somewhat lacking.
3: Multiple relevant prior studies are introduced, and the differences and originality are logically and structurally explained. The use of figures or comparison tables further enhances the evaluation.

#5: Validity and Consistency of Research Methods
0: The methods are unclear or disconnected from the hypothesis.
1: The methods are stated but lack consistency.
2: The proposed methods correspond to the objectives and hypotheses.
3: The methods are clearly stated and fully consistent with the objectives.

#6: Description of Expected Results and Analytical Methods
0: Expected results are unclear or absent, and analytical methods are not presented.
1: The description of expected results and analytical methods is ambiguous, with leaps in logic and questionable validity.
2: Either the expected results or analytical methods are somewhat unclear, but the proposal is generally valid overall.
3: The expected trends of results based on the hypotheses are clearly presented, and the analytical methods (e.g., statistical techniques, visualization) are appropriate and explicitly described.

#7: Social Significance and Potential Impact
0: No mention is made of social significance or potential impact.
1: The potential impact of the research results is unclear; although mentioned, the statement is weak.
2: Social significance is mentioned but remains abstract or limited to generalities; further elaboration is desirable.
3: The potential impact on society, industry, education, environment, etc., in the event of research success is concretely described, with a clear outlook presented.

#8: Clarity of Structure, Figures/Tables, and Terminology
0: The structure is disorganized, figures and tables are absent or hinder understanding, and terminology is frequently misused.
1: The structure and figures/tables are often unclear, and some terms are used inappropriately.
2: The proposal is generally readable, and figures and terminology are reasonably clear, though several areas for improvement remain.
3: Headings and paragraph structure are clear, technical terms are used accurately, and figures, tables, and schematic diagrams are effectively placed and explained.

#9: Accuracy of References and Compliance with APA Style
0: The reference list is absent or does not follow APA style at all, and the proposal text does not cite references.
1: Problems exist in reference consistency; formatting is inconsistent, errors are frequent, and many listed references are not cited in the text.
2: The reference list generally follows APA style, but minor errors or omissions remain (e.g., missing volume/issue numbers), and there are slight inconsistencies between the list and in-text citations.
3: The reference list fully complies with APA style, with complete and accurate inclusion of author names, year, journal title, volume, issue, page numbers, DOI, etc., and is consistent with in-text citations.