Comparing Human and LLM Evaluations on AI-Generated Critical Thinking Items: Implications for Valid Applications of Automatic Item Generation

Euigyum Kim[1], Salah Khalil[2] and Hyo Jeong Shin[1,*]

[1]Sogang University, Seoul, South Korea
[2]MACAT, London, United Kingdom

Abstract

As a core 21st-century skill, critical thinking (CT) has garnered increasing attention in today's information society. Although growing interest has led to the development of various CT assessments and frameworks, research on leveraging large language models (LLMs) for the automatic generation and validation of CT items remains limited. To address this gap, this study examines AI-generated CT items developed based on MACAT's PACIER framework. We employed a human-in-the-loop evaluation approach, in which a human expert and four LLMs independently rated each item on a three-point quality scale and conducted qualitative reviews to identify item-level issues. The results demonstrate marked differences between the human and LLM evaluations. The human reviewer delivered more discerning and variable evaluations, whereas the LLMs exhibited greater uniformity and consistency, but tended to be permissive and generous in their judgments. Notably, the human expert identified subtle flaws that the LLMs failed to detect, such as imprecise terminology, overly suggestive answer choices, and culturally biased content, all of which pose threats to the validity of the assessment. These insights affirm the essential role of human engagement in validating and optimizing the automatic item generation (AIG) process for complex latent constructs such as CT.

Keywords

Automatic Item Generation, Critical Thinking, Item Validity, Human-in-the-Loop, Large Language Models

1. Introduction

In the rapid and continuous evolution of the information society, critical thinking (CT) has emerged as an essential skill to effectively navigate and interpret the vast amount of information encountered in daily life. Rather than passively consuming information, recent educational research underscores the importance for young adults to critically analyze and evaluate the knowledge through structured reflection and reasoning processes [1]. This emphasis has placed CT as a core competency of the 21st century [2], prompting extensive scholarly research into its effective cultivation [3, 4] and robust assessment methods and practices [5, 6].

CT skills involve awareness of underlying assumptions, an understanding that what appears intuitive may, in fact, be the opposite, and a recognition that conflicting assumptions require deeper reasoning based on logical principles [7]. These skills require reflective and analytical thinking, the evaluation of the credibility of information, and the ability to make sound inferential judgments [8]. In this regard, CT goes beyond rote memorization or simple reading comprehension and is essential for sound decision-making. However, CT is inherently difficult to evaluate and measure due to its nature as a latent competency. Consequently, the growing interest in CT has driven extensive educational research, along with concerted efforts to develop valid and reliable approaches to assess CT [9].

Recent advances in artificial intelligence (AI) and large language models (LLMs), offer promising avenues for automating the generation and development of assessment items [10], especially those that target complex cognitive competencies. In fact, a growing body of notable research has emerged that

EvalLAC'25: 2nd Workshop on Automatic Evaluation of Learning and Assessment Content, July 26, 2025, Palermo, Italy
*Corresponding author.
egkim@sogang.ac.kr (E. Kim); salah@macat.com (S. Khalil); hshinedu@sogang.ac.kr (H. J. Shin)
https://www.macat.com (S. Khalil)
0009-0007-3716-3945 (E. Kim); 0000-0003-2867-494X (H. J. Shin)

applies AI-driven methodologies within the domain of educational evaluation [11]. These include automated scoring systems [12, 13, 14], automatic short answer grading [15], assessment of the complexity and quality of automatically generated questions [16], automatic item generation (AIG) [10, 17], and automated translation of items for international large-scale assessments [18]. Collectively, these studies illustrate the expanding role of AI and LLMs in scaling and refining educational assessment, while also raising important considerations concerning the validity of their applications.

Traditionally, the validation of assessment items has depended heavily on expert judgment, guided by rigorous standards established by professional testing organizations. Although these expert-led procedures play a crucial role in ensuring the validity of the construct and the reliable assessment [19], they require substantial time and resources, especially in large-scale testing environments. In response, there has been growing interest in hybrid models that combine AIG with systematic human review as a way to improve efficiency while ensuring the validity of the assessment. A notable example is Duolingo's *Item Factory*, which demonstrates how such an approach can effectively achieve both scalability and psychometric rigor through human-AI collaboration [20].

Despite these advancements, ensuring the validity and psychometric soundness of AI-generated items remains a challenge [21], particularly when assessing complex and latent competencies such as CT. In addition, there remains a gap in research that explicitly examines whether LLMs can enable automated assessment by identifying items that capture the multifaceted nature of CT and emulate the rigor of human expert judgment to ensure a valid evaluation. Therefore, this study examines both quantitative and qualitative differences between human- and LLM-based review on AI-generated CT assessment items, in order to obtain insights that can inform the development of more valid and collaborative human–AI frameworks for AIG. Consequently, this study aims to comprehensively examine the following research questions.

1. What quantitative differences emerge between human and LLM evaluations of AI-generated items?
2. From a qualitative point of view, are there aspects of human feedback that LLMs tend to overlook or fail to capture on AI-generated items?

By addressing these research questions, this study aims to identify key factors that impede the validity of items that are generated by LLMs within the context of AIG, and to offer practical insights for enhancing the quality and validity of AIG applications.

2. Experimental Methodology

2.1. PACIER Framework

The CT items analyzed in this research are derived from MACAT's CT assessment solutions. MACAT specializes in developing CT learning materials and evaluating CT skills through sophisticated assessment frameworks and educational resources [22]. MACAT has developed the PACIER framework that encompasses six key sub-domains of CT: Problem solving, Analysis, Creative thinking, Interpretation, Evaluation, and Reasoning [23]. In the item development process, MACAT has employed LLM-based AIG to develop CT assessment items, and these items underwent rigorous human expert reviews prior to administration [24]. Recent research performed psychometric analyses of the data from the pilot study, demonstrating that these AIG items functioned properly and provided empirical support for their construct validity and measurement reliability [25].

2.2. Data

The items analyzed in this study consist of multiple-choice questions (MCQs) that were designed to assess CT competencies among 7th and 8th grade students. A total of 18 AI-generated MCQs were selected, comprising 3 items for each of the 6 domains defined by the PACIER framework. Each item consists of a passage and a corresponding question, followed by multiple-choice options designed to

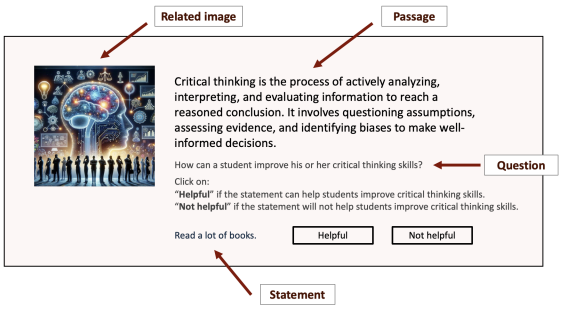

Figure 1: An example of a CT assessment item.

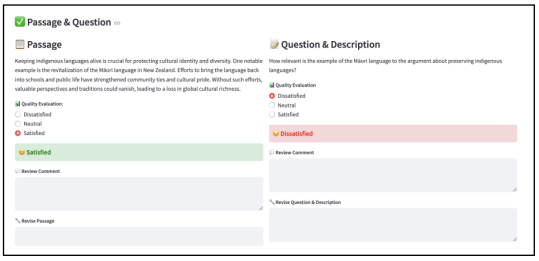

Figure 2: The item review tool interface.

evaluate the relevance of each statement. Although standard assessments typically employ 4 statement options per MCQ, the items in this study were deliberately generated with 10 statement options to facilitate more flexible and rigorous quality review and analysis. A representative example and structure of these CT assessment items is shown in Figure 1.

2.3. Review Method

A subject matter expert in CT conducted a structured evaluation of the 18 items. For this study, a custom-built item review platform, as illustrated in Figure 2, was developed to facilitate systematic expert review. The human reviewer used the platform to review each component of the items (passage, question, and statement options), using a three-point scale rating system. The items were assessed according to their suitability for effective CT evaluation, according to the following scale.

- ***Dissatisfied***: Items fundamentally flawed or inappropriate for use in CT assessments, necessitating immediate exclusion.
- ***Neutral***: Items that require modification to improve clarity, relevance, or appropriate difficulty.
- ***Satisfied***: Items suitable for immediate implementation or minor revisions only.

In addition to the ratings, the expert was asked to provide detailed qualitative feedback on the components of individual items. For items rated as *neutral* (score 2) or *satisfied* (score 3), the reviewer was additionally asked to offer concrete revision suggestions to guide subsequent improvement of the components. For items rated as *Dissatisfied* (score 1), reviewer was only asked to provide evaluative feedback, without specific suggestions for revision.

Concurrently, four distinct OpenAI's language models (GPT-4o, GPT-4.5-preview, o1-mini, and o3-mini) independently reviewed the same items leveraging a LLM-as-a-judge approach [26]. Although LLM as a judge has shown close alignment with human judgments in a variety of tasks, particularly with minimal tuning or refinement [27, 28], its effectiveness in evaluating complex cognitive processes or assessing the quality of AIG targeting higher-order skills remains limited. Furthermore, to ensure that both human and LLM reviewers perform independent evaluations based on identical criteria, all LLMs were provided with the same evaluation guidelines as those given to a human expert, implemented in the form of the standard prompt illustrated in Figure A.1 of Appendix A.

For the LLM review, we used the OpenAI API via the developer platform, setting the temperature parameter to 0 and fixing the seed to ensure the reproducibility and consistency of the results. Moreover, the prompts were designed to reflect the same context and information provided to human reviewers, thus multiple prompt variations were not explored. The purpose of this approach was to investigate how the LLM would generate reviews based solely on the context and information similar to that given to human experts. In other words, this study prioritized aligning the experimental conditions as closely as possible with those of the human reviewers.

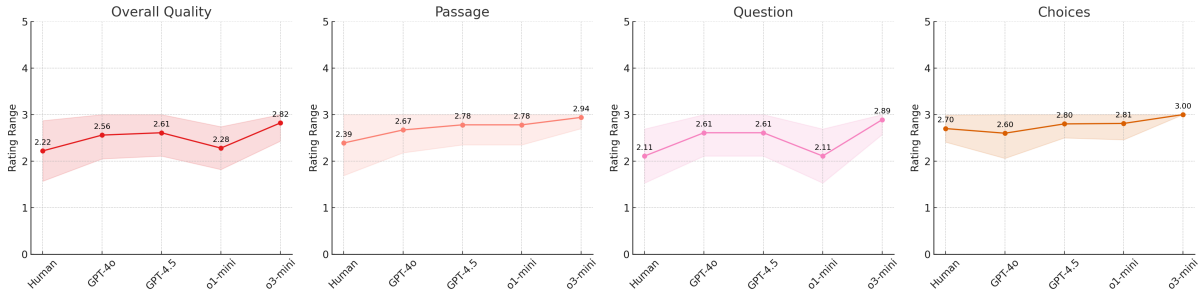

Figure 3: Comparison of rating results across item components. Error bands represent the standard deviation from the mean score.

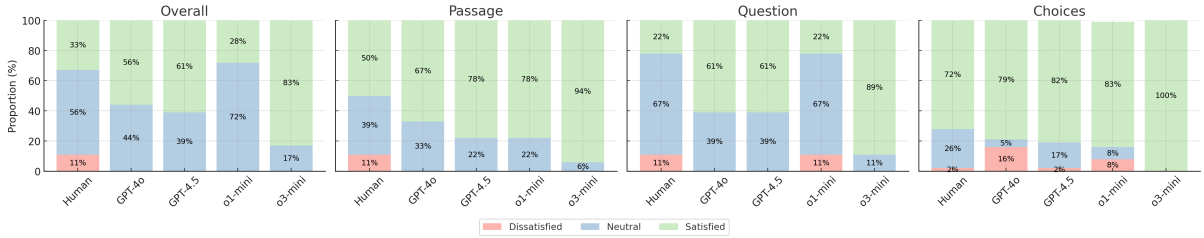

Figure 4: Distribution of rating percentages for MC item reviews across four item components. Each bar represents the proportion of responses in each rating category.

3. Results

3.1. Quantitative Comparison of Quality Ratings

An examination of review results across reviewers on 18 MCQs, as shown in Figure 3, highlights marked differences in rating consistency. The human reviewer exhibited the highest variability across all review areas. For *Overall Quality*, the human's standard deviation (SD = 0.65) exceeded that of GPT-4.5-preview (SD = 0.50) and o3-mini (SD = 0.39). A similar pattern is observed for *Passage* ratings, where the human's variability peaked (SD = 0.70), compared to GPT-4o (SD = 0.49) and GPT-4.5-preview (SD = 0.43). The human reviewer assigned scores more strictly and variably, whereas certain LLMs, particularly o3-mini, exhibited near-zero variance in specific components. For example, in the *Choices* component, o3-mini recorded a zero standard deviation (SD = 0).

In addition, the results indicate that LLMs tended to assign consistently favorable scores with relatively low dispersion. In contrast, the human rater applied the scale more flexibly and made near-exclusive use of the minimum score 1, which corresponds to *Dissatisfied*. This rating was rarely, if ever, assigned by any LLM, a pattern that is clearly evident in Figure 4. The human expert's scoring behavior suggests greater sensitivity to qualitative distinctions and a stronger willingness to penalize perceived deficiencies.

Figure 4 shows that rating distributions for MCQs varied markedly across evaluators. LLMs exhibited highly uniform scoring patterns across MCQs, with ratings predominantly clustered in the *Satisfied* category and infrequent use of lower categories such as *Dissatisfied*. From the perspective of human rater behavior, this pattern may reflect a form of leniency bias [29]. However, the human reviewer made fuller use of the rating scale, including multiple instances of the lowest score (11% in several components), indicating a more discriminating position that may better capture subtle variations in the quality of the items. Although systematic leniency and low dispersion observed in LLM evaluations may benefit consistency, these advantages come at the cost of reduced sensitivity. Therefore, integrating human expert review, alongside the examination of detailed qualitative comments, may facilitate the more nuanced and construct-aligned AIG system.

3.2. Analysis on Qualitative Feedback by Human

A qualitative analysis was conducted to gain deeper insight into the human expert's evaluation of the AIG items. Based on the expert's feedback, we identified three key thematic categories, summarized in Table 1. The expert provided detailed comments revealing factors that can undermine the validity of the assessment that are often overlooked by LLM reviews.

Table 1
Thematic categorization of human expert's feedback on AI-generated CT assessment items.

[1] Terminology that undermines test validity
– "(...) is a difficult formulation for not-so-strong readers."
– "(...) is unnecessarily vague scientific jargon."
– "The term (...) might be too technical for many students and may lead to incorrect interpretations."
[2] Item content that reduces discriminative validity
– "Too simple and easy to see the answer."
– "Answer appears verbatim in the passage."
– "Item is too simple; the answer is too easy to infer."
[3] Cultural bias that compromises item fairness
– "(...) portrayed in a one-sided positive light."
– "The concept of the (...) varies by culture and perspective."
– "(...) might be ideal in some contexts, while (...) may carry a clearer negative connotation."

These thematic categories reveal several qualitative limitations in AIG based on LLM. First, the expert identified instances of terminology that undermines test validity, noting that vague or overly technical language can obscure item clarity and compromise construct validity. LLMs frequently overlooked such subtleties and reduced their judgments to simplistic difficulty labels such as 'easy' or 'hard'.

In addition, the expert pointed to the content of the item that reduces discriminative validity, including unintentional clues and obvious answer choices. Such issues were considered to potentially weaken the clarity of the item and reduce the diagnostic value of CT evaluation.

Furthermore, cultural bias that compromises item fairness emerged as a pertinent concern. The human reviewer highlighted items that embedded ethnocentric assumptions or failed to reflect diverse sociocultural contexts, raising potential concerns about equity and cross-cultural validity.

Taken together, these findings highlight the continued importance of expert human review in detecting nuanced flaws that current LLMs may fail to recognize. Although the generous tendencies of LLMs have been previously reported in automated scoring [30] research and question evaluation [31] research, it is noteworthy that this study identifies specific threats to validity that are often overlooked as a result of such tendencies, as well as key considerations that should be addressed in the application of AIG. As AIG becomes more integrated into the development of various assessments, incorporating human-in-the-loop mechanisms will be critical to safeguarding item validity and test reliability [21]. In the end, in the context of LLM-powered AIG, it is essential that factors compromising validity be systematically incorporated into the item generation and self-review mechanisms of LLMs.

4. Conclusion & Discussion

4.1. Conclusion

This study compared the evaluation of AI-generated CT items by the human expert and four LLMs acting as reviewers. Through both quantitative and qualitative reviews, the findings indicated that, while LLMs demonstrated a high degree of internal consistency in their reviews, their critiques frequently lacked depth and were often overly generous, failing to detect subtle factors that could compromise the

validity of the items. In contrast, the human expert provided more discerning critiques, identifying pivotal issues that could compromise the validity of the item.

Consequently, discernible discrepancies observed between human and LLM reviewers underscore the necessity of systematic and ongoing human involvement to ensure the validity of items generated through AIG, especially when applied to measuring higher-order competencies such as CT. In this regard, human evaluators would play an indispensable role in proactively identifying and addressing potential threats to item validity throughout the AIG process. Accordingly, this study offers guidance by articulating essential considerations that should be provided to human reviewers or incorporated into AIG prompting, thereby informing future implementations and advancing research in AIG.

4.2. Discussion

One of the main implications of this study lies in identifying key considerations for the actual implementation of LLM-powered AIG. Although human involvement remains indispensable, strategically optimizing human participation can enable maximal validation with minimal involvement. In fact, leveraging the data collected in this study may offer a valuable pathway toward more efficient item review. In particular, quantitative evaluations and direct revision suggestions from the human expert provide feedback data for reinforcement learning from human feedback [32]. In addition, qualitative feedback on nuanced issues can guide prompt optimization [33]. These insights may support the refinement of LLM behavior and the development of more effective human-AI collaboration in educational assessment practices [21]. These findings would also be adopted beyond CT, suggesting a broader applicability of LLM-based AIG when evaluating various higher-order thinking skills. In this regard, this research sets the stage for future investigations aimed at refining interactions between humans and AI, thereby enhancing the validity of AI-generated items in various domains.

However, this research has several limitations that should be addressed in future studies. First, due to practical constraints, this study involved only one human reviewer, which limits the generalizability of the review process and raises concerns about potential bias resulting from the individual rater's severity. Future research should include multiple human reviewers to thoroughly examine inter-rater agreement and evaluation consistency.

Second, this study did not involve advanced fine-tuning of the LLMs to better detect validity-threatening factors, nor did it employ repeated measures to verify the consistency of review ratings. Consequently, future studies should systematically explore adjustments of LLMs' various parameters and diverse prompt design strategies to clearly define and refine the role of LLMs as item reviewers.

Lastly, the amount of review data utilized in this study is insufficient to fully support the research findings, which poses a limitation. Although the findings of this preliminary experiment are exploratory in nature, future studies should secure a larger volume of data and involve multiple experts who review a diverse set of items to improve the reliability of the results. Furthermore, it is necessary to verify whether the instructions provided to LLM reviewers can consistently produce reliable evaluation results for various types of items.

Despite these limitations, this study identified key considerations necessary for applying AIG to the assessment of higher-order skills. By addressing the aforementioned findings and implications, future research could further enhance the effectiveness of LLMs and contribute to the development of valid assessments in the field of automated evaluation.

Acknowledgments

We would like to express our deep gratitude to the anonymous reviewers. Their invaluable feedback and constructive suggestions were crucial in enhancing the quality of this paper and have greatly benefited our ongoing research. The improvements made based on their comments are incorporated specifically in the **4.2. Discussion** section.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and DeepL in order to: Grammar and spelling check, paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] R. T. Pithers, R. Soden, Critical thinking in education: A review, Educational Research 42 (2000) 237–249. doi:10.1080/001318800440579.

[2] World Economic Forum, New vision for education: Unlocking the potential of technology, https://widgets.weforum.org/nve-2015/chapter1.html, 2015. Accessed: 2025-05-19.

[3] C. P. Dwyer, A. Walsh, An exploratory quantitative case study of critical thinking development through adult distance learning, Educational Technology Research and Development 68 (2020) 17–35. doi:10.1007/s11423-019-09659-2.

[4] H. C. Yeh, S. H. Yang, J. S. Fu, Y. C. Shih, Developing college students' critical thinking through reflective writing, Higher Education Research & Development 42 (2022) 244–259. doi:10.1080/07294360.2022.2043247.

[5] O. L. Liu, L. Mao, L. Frankel, J. Xu, Assessing critical thinking in higher education: the heighten™ approach and preliminary validity evidence, Assessment and Evaluation in Higher Education 41 (2016) 677–694. doi:10.1080/02602938.2016.1168358.

[6] O. L. Liu, A. Shaw, L. Gu, G. Li, S. Hu, N. Yu, P. Loyalka, Assessing college critical thinking: Preliminary results from the chinese heighten® critical thinking assessment, Higher Education Research & Development 37 (2018) 999–1014. doi:10.1080/07294360.2018.1467381.

[7] H. Tunjungsari, B. Takwin, Understanding critical thinking practice in everyday life through stages of critical thinking skills and disposition, Mind, Brain, and Education 15 (2021) 225–231. doi:10.1111/mbe.12279.

[8] R. H. Ennis, Critical thinking dispositions: Their nature and assessability, Informal Logic 18 (1996) 165–182. doi:10.1353/jge.2007.0011.

[9] Council for Aid to Education (CAE), Critical thinking assessment for academic and career success, 2025. URL: https://cae.org/solutions/critical-thinking-assessment/, accessed: 2025-05-17.

[10] Y. Attali, A. Runge, G. T. LaFlair, K. Yancey, S. Goodwin, Y. Park, A. A. von Davier, The interactive reading task: Transformer-based automatic item generation, Frontiers in Artificial Intelligence 5 (2022). doi:10.3389/frai.2022.903077.

[11] J. Hao, A. A. von Davier, V. Yaneva, S. Lottridge, M. von Davier, D. J. Harris, Transforming assessment: The impacts and implications of large language models and generative ai, Educational Measurement: Issues and Practice (2024). doi:doi.org/10.1111/emip.12602.

[12] T. Shibata, M. Uto, Enhancing cross-prompt automated essay scoring by selecting training data based on reinforcement learning, in: Proceedings of the First Workshop on Automated Evaluation of Learning and Assessment Content (EvalLAC 2024), volume 3772 of *CEUR Workshop Proceedings*, CEUR-WS.org, Recife, Brazil, 2024.

[13] G. G. Lee, E. Latif, X. Wu, N. Liu, X. Zhai, Applying large language models and chain-of-thought for automatic scoring, Computers and Education: Artificial Intelligence 6 (2024) 100213. doi:10.1016/j.caeai.2024.100213.

[14] J. Y. Jung, L. Tyack, M. von Davier, Towards the implementation of automated scoring in international large-scale assessments: Scalability and quality control, Computers and Education: Artificial Intelligence 8 (2025) 100375. doi:10.1016/j.caeai.2025.100375.

[15] R. V. Ivanova, S. Handschuh, Evaluating llms' performance at automatic short-answer grading, in: Proceedings of the First Workshop on Automated Evaluation of Learning and Assessment Content (EvalLAC 2024), volume 3772 of *CEUR Workshop Proceedings*, CEUR-WS.org, Recife, Brazil, 2024.

[16] S. Alkhuzaey, F. Grasso, T. R. Payne, V. Tamma, Towards automatic evaluation of questions

generated from ontologies, in: Proceedings of the First Workshop on Automated Evaluation of Learning and Assessment Content (EvalLAC 2024), volume 3772 of *CEUR Workshop Proceedings*, CEUR-WS.org, Recife, Brazil, 2024.

[17] S. Khan, Y. Huang, S. Pu, V. Tarasov, A. Andrade, R. Meisner, D. Edwards, A. A. von Davier, Sphinx: an automated generation system for english reading comprehension assessment, in: International Conference on Learning Analytics and Knowledge, 2020.

[18] J. Y. Jung, L. Tyack, M. von Davier, Combining machine translation and automated scoring in international large-scale assessments, Large-scale Assessments in Education 12 (2024). doi:10.1186/s40536-024-00199-7.

[19] T. M. Haladyna, M. C. Rodriguez, Developing and Validating Test Items, Routledge, London, UK, 2013.

[20] A. A. von Davier, A. Runge, Y. Park, Y. Attali, J. Church, G. LaFlair, The item factory: Intelligent automation in support of test development at scale, in: Machine Learning, Natural Language Processing, and Psychometrics, Information Age Publishing, Charlotte, NC, 2024, pp. 1–25.

[21] A. A. von Davier, J. Burstein, Ai in the assessment ecosystem: A human–centered ai perspective, in: P. Ilic, I. Casebourne, R. Wegerif (Eds.), Artificial Intelligence in Education: The Intersection of Technology and Pedagogy, volume 261 of *Intelligent Systems Reference Library*, Springer, Cham, 2024. doi:10.1007/978-3-031-71232-6_6.

[22] MACAT, Critical thinking assessments, https://www.macat.com/critical-thinking-assessments, 2025. Accessed: 2025-05-14.

[23] MACAT, Critical thinking, https://www.macat.com/critical-thinking, 2025. Accessed: 2025-05-14.

[24] H. J. Shin, S. Li, J. H. Ryoo, A. A. von Davier, T. Lubart, S. Khalil, The nature and measure of critical thinking: The pacier framework and assessment, 2025. Manuscript submitted for publication.

[25] H. J. Shin, S. Li, J. H. Ryoo, A. A. von Davier, Harnessing artificial intelligence for generating items in critical thinking tests, 2024. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Philadelphia, PA, April 2024.

[26] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, arXiv preprint arXiv:2306.05685 (2023). doi:10.48550/arXiv.2306.05685.

[27] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, J. Guo, A survey on llm-as-a-judge, arXiv preprint arXiv:2411.15594 (2024). doi:10.48550/arXiv.2411.15594.

[28] A. S. Thakur, K. Choudhary, V. S. Ramayapally, S. Vaidyanathan, D. Hupkes, Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges, arXiv preprint arXiv:2406.12624 (2024). doi:10.48550/arXiv.2406.12624.

[29] E. W. Wolfe, Identifying rater effects using latent trait models, Psychology Science 46 (2004) 35–51.

[30] K. Seßler, M. Fürstenberg, B. Bühler, E. Kasneci, 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 (LAK 2025), LAK '25, Association for Computing Machinery, New York, NY, USA, 2025, pp. 462–472. URL: https://doi.org/10.1145/3706468.3706527. doi:10.1145/3706468.3706527.

[31] S. Bhat, H. Nguyen, S. Moore, J. Stamper, M. Sakr, E. Nyberg, Towards automated generation and evaluation of questions in educational domains, in: Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022), volume 3166 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3166/.

[32] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, in: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), Curran Associates Inc., 2017, pp. 4299–4307.

[33] X. Lin, Z. Dai, A. Verma, S.-K. Ng, P. Jaillet, B. K. H. Low, Prompt optimization with human feedback, arXiv preprint arXiv:2405.17346 (2024). doi:10.48550/arXiv.2405.17346.

A. Detailed Prompts for LLMs' Item Review

The prompt given to the LLM is shown in Figure A.1. In addition to the content of the prompt, the model was instructed to structure the review output in a CSV format.

Prompt for LLMs' Item Review

You are a Critical Thinking Assessment's **Item Review Expert** with extensive experience in educational evaluation and test design, specializing in critical thinking.

Your role is to systematically evaluate the quality of test items based on established frameworks, ensuring fairness, reliability, and alignment with learning objectives.

Review Context
- The exam items are designed for Grade 7~8 learners.
- Each item consists of a Passage, a Question, and 10 Answer Choices (each with an Explanation).
- Your task is to rigorously evaluate the quality of each component and provide structured feedback.

PACIER Framework (Cognitive Process Dimensions)
The PACIER categorizes cognitive processes into six distinct levels:
- **Problem-Solving (P):** (...)
- **Analysis (A): ** (...)
- **Creative Thinking (C):** (...)
- **Interpretation (I):** (...)
- **Evaluation (E):** (...)
- **Reasoning (R):** (...)

Each test item should align with at least one PACIER category, ensuring it assesses critical thinking skills effectively.

Evaluation Methodology
1. Assessment Criteria
  - Passage: Relevance, clarity, and cognitive demand.
  - Question: Alignment with passage, clarity, and ability to assess critical thinking.
  - Answer Choices: Plausibility of distractors, clarity, and correctness of explanations.

2. Comparative Judgment
  - Evaluate each component relative to high-quality reference items to ensure consistency.

3. Rating Scale
  - Dissatisfied: Fundamentally flawed or inappropriate for assessment and thus discarded without revision suggestions.
  - Neutral: Requires revisions to improve clarity, relevance, or difficulty. You should provide detailed feedback and specific revision recommendations.
  - Satisfied: Suitable for immediate use or required minimal edits. You could directly accept these items or suggest minor enhancements.

4. Actionable Feedback
  - Provide concise but specific feedback justifying each rating.

Additional Guidelines
- Ensure alignment with cognitive and linguistic proficiency standards.
- **Maintain consistency** across evaluations to avoid bias.
- Do not include markdown, bullet points, or additional explanations.

Figure A.1: Prompts provided with detailed instructions and context for LLMs' item review.