



Teacher Decision-Making and Assessment Consistency in Process-Oriented Arts Education

Sarah J. McKenzie¹, James R. Wallace¹, Emily K. Thompson^{1*}

¹Faculty of Fine Arts and Music, The University of Melbourne, Melbourne, VIC 3010, Australia

Corresponding Author*: Sarah J. McKenzie E-mail: sarah.mckenzie@unimelb.edu.au

ARTICLE INFO

Keywords

teacher decision-making
assessment consistency
arts education
inter-rater reliability
rubric design
process-oriented evaluation
assessment fairness

Published

01 September 2026

ABSTRACT

This study examines teacher decision-making and assessment consistency in process-oriented arts education by analyzing how instructors evaluate creative learning processes under different assessment conditions. The study is designed to involve approximately 60 arts teachers and 900 student portfolios from visual arts, design, creative practice, and interdisciplinary arts courses. Each portfolio is evaluated by at least three teachers, generating around 2,700 independent assessment records. Evaluation indicators include originality, technical improvement, experimentation, reflective depth, process documentation, final presentation quality, and overall achievement. The study compares open-ended teacher judgment with structured rubric-based assessment to examine changes in scoring variation, inter-rater reliability, and assessment consistency. Quantitative analysis applies intraclass correlation coefficient, generalizability theory, ANOVA, mixed-effects modeling, and variance decomposition to identify whether teacher background, teaching experience, course type, and rubric use significantly affect assessment results. Teacher interviews are further used to explain how instructors balance artistic freedom, institutional standards, and fairness in grading. The innovation of this study lies in treating arts assessment as both a pedagogical and measurement issue, providing empirical evidence on whether structured assessment tools can improve fairness and consistency without weakening professional judgment in creative education.

1. Introduction

Assessment in arts education requires teachers to make complex judgments about both creative processes and learning outcomes. In process-oriented courses, student achievement cannot be adequately represented by the quality of the final artwork alone. Teachers are also expected to consider idea development, technical progress, experimentation, revision, reflective thinking, process documentation, and presentation quality. These dimensions provide important evidence of creative learning, but they are often difficult to evaluate consistently because artistic performance is open-ended and highly dependent on context. Previous studies have shown that assessment outcomes may be influenced by rater experience, task characteristics, scoring criteria, and different interpretations of creativity and artistic quality (Abdellatif & El-Wakeel, 2025; J. Huang, Xu, et al., 2026). Assessment quality is also closely related to broader concerns about teaching quality and

Citation: McKenzie, S. J., Wallace, J. R., & Thompson, E. K. (2026). Teacher decision-making and assessment consistency in process-oriented arts education. *Journal of Adaptive Learning and Psychological Growth*, 1(2), 14-22.

3154-7915/© The Authors. Published by J&L Academic Group PLT. This is an open access article under the CC BY 4.0 license.
<https://doi.org/10.64744/jalpg.2026.278>.

educational equity. Evidence from public art education suggests that systematic assessment can support the monitoring of instructional quality and help determine whether students are evaluated under sufficiently fair and comparable conditions (Lin et al., 2026). This perspective extends the function of assessment beyond grading and positions it as an important mechanism for examining the quality, transparency, and equity of arts education. If teachers rely entirely on unrestricted professional judgment, comparable portfolios may receive noticeably different scores; however, excessively rigid criteria may constrain the flexibility needed to recognize diverse forms of creative achievement. Arts assessment must therefore balance professional judgment, fairness, reliability, and sensitivity to the individual nature of creative work (Seow, 2026; Zheng & Makar, 2022).

Rubrics are widely used to improve transparency and consistency in classroom assessment. By defining relevant dimensions of performance and describing different levels of achievement, rubrics can make assessment expectations more visible, support formative feedback, and help both teachers and students understand the qualities associated with successful work (Jonsson et al., 2025; Z. Zhang et al., 2025). In arts education, structured criteria may be particularly useful because they allow broad concepts such as creativity or artistic achievement to be divided into more interpretable components, including originality, technical development, experimentation, reflection, process engagement, and presentation. Rubrics may also provide a common reference point when several teachers assess the same work, thereby reducing the influence of purely personal scoring preferences (Shumaker et al., 2025; Wu et al., 2026). Nevertheless, their effectiveness depends heavily on how the criteria are designed and interpreted. When rubrics are treated as fixed checklists rather than flexible assessment guides, they may encourage superficial compliance and limit teachers' ability to recognize unexpected or unconventional forms of creativity (Qi & Qiao, 2026; Zhang et al., 2026). This limitation is especially relevant in visual arts, design, creative practice, and interdisciplinary arts courses, where successful learning often emerges through repeated experimentation, material exploration, revision, failure, and reflection rather than through a single predetermined route (Sattar et al., 2025). Teacher decision-making therefore remains central even when structured assessment tools are introduced. Instructors may rely on different forms of evidence when judging the same student work. Some may emphasize originality and conceptual development, while others may place greater weight on technical improvement, effort, process documentation, fulfillment of course objectives, or the quality of the final presentation. These decisions can also be shaped by disciplinary background, teaching experience, familiarity with particular artistic practices, and institutional expectations (Hong et al., 2026; Xiong & Guo, 2026). Such variation does not necessarily indicate poor assessment practice, because professional interpretation is an important part of evaluating complex creative work. However, excessive variation becomes problematic when students receive substantially different scores because of who happens to assess their portfolios rather than because of meaningful differences in performance. This raises an important distinction between legitimate professional judgment and avoidable rater-related inconsistency (Boichuk, 2025; Li et al., 2026). The relationship between reliability and professional autonomy is therefore particularly important in process-oriented arts assessment. A highly standardized scoring system may improve agreement while failing to capture the richness of creative work, whereas completely open judgment may preserve artistic sensitivity but weaken comparability across teachers and courses (Call et al., 2026; Y. Zhang et al., 2025). Effective assessment requires a structure that can support shared standards without eliminating professional interpretation. Existing research has examined rubric design, creativity assessment, teacher judgment, and classroom evaluation, but these areas are frequently treated as separate research problems. Fewer studies have investigated how structured criteria and teacher decision-making

interact when several instructors evaluate the same portfolio. As a result, it remains unclear whether rubric-based assessment primarily reduces random rater variation, changes the dimensions that teachers emphasize, or simply shifts disagreement from overall scores to the interpretation of individual criteria (R. Huang, 2026; Shao, 2026). Methodological limitations further restrict current understanding of assessment consistency in arts education. Many studies rely on relatively small samples, single-course settings, individual institutions, or descriptive analyses. Such designs provide useful contextual insights but offer limited evidence about whether observed scoring patterns can be generalized across teachers and course types (Gao et al., 2026; Lee et al., 2025). Large-scale portfolio studies involving multiple independent raters remain relatively uncommon. In addition, assessment reliability is often summarized using mean scores, percentage agreement, or simple correlations. These measures can describe broad similarity between ratings but cannot fully explain why scores differ. Two teachers, for example, may show a strong correlation while consistently assigning different score levels. More advanced approaches, including intraclass correlation coefficients, generalizability theory, mixed-effects modeling, and variance decomposition, provide a stronger basis for distinguishing variation associated with students, teachers, courses, criteria, and interactions among these factors (Köroğlu, 2025; Xu et al., 2026). Such methods are particularly valuable in arts assessment because several potential sources of variation operate simultaneously. Course context may also influence how assessment criteria function. Visual arts courses may emphasize material handling, visual expression, and experimentation, while design courses may place greater weight on problem solving, iteration, functionality, and communication. Creative practice and interdisciplinary arts courses may combine several of these expectations and allow even broader interpretations of successful performance. Consequently, a rubric that produces stable scores in one course may not necessarily function in the same way in another. Similarly, teachers with different disciplinary backgrounds may interpret terms such as originality, progress, experimentation, or reflective depth differently even when they use the same scoring descriptors. Examining assessment consistency across multiple course types and teacher groups can therefore provide a more realistic understanding of how structured assessment operates in authentic arts education settings.

The present study addresses these conceptual and methodological gaps by examining teacher decision-making and assessment consistency in process-oriented arts education using a large multi-rater portfolio dataset. The study includes approximately 60 arts teachers and 900 student portfolios from visual arts, design, creative practice, and interdisciplinary arts courses, generating approximately 2,700 independent assessment records. Each portfolio is evaluated by multiple teachers, allowing direct analysis of the extent to which scores vary across raters and assessment conditions. The study compares open-ended professional judgment with structured rubric-based assessment to determine whether clearer scoring criteria can improve consistency and fairness without unnecessarily restricting professional interpretation. Intraclass correlation coefficients and generalizability theory are used to evaluate inter-rater reliability and identify major sources of measurement variation, while mixed-effects models and variance decomposition are employed to examine the effects of teacher characteristics, course context, and assessment format. The study also considers how different assessment dimensions contribute to agreement and disagreement, with particular attention to originality, technical development, experimentation, reflection, process documentation, presentation, and overall achievement. By distinguishing variation attributable to students from variation associated with teachers and assessment conditions, the research seeks to clarify when disagreement reflects legitimate professional interpretation and when it represents avoidable inconsistency. The findings are expected to provide empirical evidence for the design of

more reliable and flexible arts-assessment rubrics, strengthen moderation and assessor-training practices, and support more transparent portfolio evaluation. More broadly, the study contributes to the development of assessment systems that can protect the openness and diversity of creative learning while improving comparability, instructional quality, and educational fairness in process-oriented arts education.

2. Materials and Methods

2.1 Sample and Study Setting

This study examined portfolio assessment in process-oriented arts education. The sample included about 60 arts teachers and 900 student portfolios from visual arts, design, creative practice, and interdisciplinary arts courses. Each portfolio was scored by at least three teachers, resulting in about 2,700 independent assessment records. The portfolios contained final works and process evidence, including sketches, drafts, revision notes, experiment records, reflective statements, and presentation files. The teachers differed in teaching experience and course background. They mainly worked in studio teaching, design practice, creative production, and interdisciplinary arts teaching. Portfolios with missing process records, incomplete final works, or unclear student information were removed before scoring. All materials were anonymized to reduce bias related to student identity, class section, or teacher familiarity.

2.2 Experimental Design and Control Setting

A comparative assessment design was adopted to test whether structured rubrics could improve scoring consistency. The open-judgment condition served as the control setting. In this condition, teachers scored portfolios based on their usual professional judgment. The rubric-guided condition served as the experimental setting. In this condition, teachers assessed portfolio materials with a shared scoring rubric. The rubric covered seven indicators: originality, technical improvement, experimentation, reflective depth, process documentation, final presentation quality, and overall achievement. This design was chosen because arts assessment depends strongly on teacher expertise, but unclear criteria may increase score variation. By comparing the two conditions, the study examined whether clear criteria could reduce scoring differences while still allowing professional judgment.

2.3 Measurement Methods and Quality Control

Teachers scored the assigned portfolios independently under the required assessment condition. Scores were recorded for originality, technical improvement, experimentation, reflective depth, process documentation, final presentation quality, and overall achievement. In the open-judgment condition, teachers gave scores and short comments according to their own assessment practice. In the rubric-guided condition, teachers used the same scoring scale and written descriptors for each indicator. Before formal scoring, sample portfolios were used for scorer training and rule checking. Portfolio order was randomized to reduce order effects. Teachers did not receive student names or previous course grades. All scoring sheets were checked for missing values, repeated entries, and abnormal score patterns. Records with large score gaps among raters were reviewed, but the original scores were retained for reliability analysis.

2.4 Data Processing and Model Formulas

The raw scores were checked, cleaned, and coded by portfolio, teacher, course type, assessment condition, and scoring indicator. Mean scores and standard deviations were calculated for each indicator under the two assessment conditions. Score variation was examined by

comparing the spread of teacher ratings across portfolios. Inter-rater reliability was measured using the intraclass correlation coefficient:

$$ICC = \sigma^2_p / (\sigma^2_p + \sigma^2_r + \sigma^2_e)$$

where σ^2_p is portfolio variance, σ^2_r is rater variance, and σ^2_e is residual variance. A mixed-effects model was also applied to test the effects of assessment condition, teacher experience, and course type on scoring results:

$$Y_{ijk} = \beta_0 + \beta_1 Condition_i + \beta_2 Experience_j + \beta_3 CourseType_k + u_j + v_k + \varepsilon_{ijk}$$

where Y_{ijk} is the score, $Condition_i$ is the assessment condition, $Experience_j$ is teacher experience, $CourseType_k$ is the course category, u_j and v_k are random effects, and ε_{ijk} is the error term.

2.5 Statistical Analysis and Interview Coding

Intraclass correlation coefficient analysis was used to compare inter-rater reliability between the open-judgment and rubric-guided conditions. Generalizability theory was used to estimate how much portfolios, raters, course types, and assessment conditions contributed to total score variation. ANOVA was used to compare mean score differences across course types and scoring indicators. Mixed-effects modeling was applied to test whether teacher background, teaching experience, course type, and rubric use affected assessment results. Variance decomposition was then used to identify the main sources of scoring inconsistency. Teacher interviews were coded to explain the statistical results. The coding focused on originality, progress, technical quality, process evidence, artistic freedom, institutional standards, and fairness. The interview findings helped explain why rubrics reduced some scoring differences while still preserving room for teacher judgment.

3. Results and Discussion

3.1 Score variation under open and rubric-guided assessment

The results indicated clear differences between open judgment and rubric-guided assessment. In the open-judgment condition, teacher scores varied more across the seven indicators. The largest differences were found in originality, experimentation, and reflective depth. These indicators depend on teachers' interpretation of creative intent, learning process, and personal expression. Therefore, they were more easily affected by teaching background and course experience. In the rubric-guided condition, score variation was reduced, especially for technical improvement, process documentation, and final presentation quality. These indicators were supported by visible portfolio evidence, such as sketches, drafts, revision records, and final works. However, originality remained difficult to score consistently. This result suggests that rubrics can improve scoring stability, but they cannot fully remove differences in artistic judgment. This finding is consistent with recent research on performance-based assessment, which reported that rater interpretation and rubric clarity can strongly affect scoring reliability (Du et al., 2026; Wu et al., 2026). The present study extends this issue to process-oriented arts education and shows that assessment consistency depends on both the rubric and the type of creative evidence being judged (Fig. 1).

Gage R&R Study - XBar/R Method			
Source	VarComp	%Contribution (of VarComp)	
Total Gage R&R	6.23878	63.93	
Repeatability	2.06825	21.19	
Reproducibility	4.17053	42.73	
Part-To-Part	3.52060	36.07	
Total Variation	9.75938	100.00	

Source	StdDev (SD)	Study Var (6 * SD)	%Study Var (%SV)
Total Gage R&R	2.49776	14.9865	79.95
Repeatability	1.43814	8.6288	46.04
Reproducibility	2.04219	12.2531	65.37
Part-To-Part	1.87633	11.2580	60.06
Total Variation	3.12400	18.7440	100.00

Number of Distinct Categories = 1

Fig. 1. Score differences across assessment indicators in the open-judgment and rubric-guided groups.

3.2 Inter-rater reliability and structured criteria

Inter-rater reliability was higher in the rubric-guided condition than in the open-judgment condition. This result suggests that shared criteria helped teachers apply assessment standards more consistently. The strongest improvement appeared in process documentation and technical improvement. These two indicators were easier to judge because the evidence was clear and directly observable. By contrast, originality and reflective depth showed smaller reliability gains. These indicators required teachers to judge risk-taking, personal voice, and conceptual quality. Different teachers may value these elements in different ways. This result shows that rubrics are useful, but they should not be treated as tools that remove all professional differences (J. Huang, Yin, et al., 2026; Phuong et al., 2025). The finding is close to recent studies on rubric-guided assessment, which found that clear criteria can reduce grading error and support more stable evaluation. In this study, rubrics worked best when portfolio evidence could be matched with written descriptors.

3.3 Sources of inconsistency in portfolio scoring

The mixed-effects model and variance decomposition showed that score differences came from several sources. Portfolio quality explained a large part of the total variance. This means that teachers could identify real differences in student performance. However, rater effects were still important, especially in the open-judgment condition. Teacher experience, course type, and disciplinary background affected scoring decisions. Teachers from design courses tended to place more weight on technical refinement and final presentation. Teachers from creative practice courses paid more attention to experimentation and process evidence. These differences do not mean that one group was more accurate than another. They show that arts assessment is shaped by different professional standards. After rubric use, rater-related variance decreased, but it did not disappear. Therefore, consistency in arts assessment should not mean full agreement among teachers. A more suitable goal is controlled variation, in which teachers share basic standards while keeping space for professional judgment (Fig. 2).

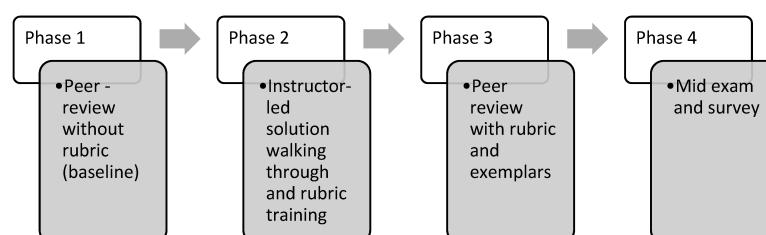


Fig. 2. Relationship among rubric use, teacher judgment, rater effects, and assessment consistency.

3.4 Teacher interviews and educational meaning

The teacher interviews helped explain the statistical results. Many teachers stated that the rubric made assessment standards clearer and helped them explain scores to students. They also reported that process evidence became easier to discuss when the criteria were linked to drafts, revisions, and reflective notes. However, some teachers were concerned that a strict rubric might reduce attention to unusual creative choices. This concern was mainly related to originality and experimentation. These responses suggest that rubrics should be used as flexible guides rather than fixed checklists. In process-oriented arts education, fairness does not mean judging all portfolios in the same way. It means using clear standards, giving reasons for scores, and treating different forms of creative growth with care (Baracco et al., 2025; Lin et al., 2020). Compared with studies that focus mainly on general classroom assessment, this study provides portfolio-based evidence from arts education. It shows that well-designed rubrics can improve fairness and reliability while preserving teacher judgment in creative learning.

4. Conclusion

This study explored teacher decision-making and scoring consistency in process-oriented arts education. The results showed that rubric-guided assessment reduced score differences among teachers, mainly in technical improvement, process documentation, and final presentation quality. Originality, experimentation, and reflective depth were still harder to score consistently because these areas depend more on teacher judgment. The model results showed that portfolio quality, teacher experience, course type, and assessment condition all affected scoring outcomes. Interview data further showed that rubrics helped teachers clarify standards and give fairer feedback. However, very rigid rubric use may reduce attention to creative choice and individual development. The main value of this study is that it views arts assessment as both a teaching issue and a measurement issue. The findings show that well-designed rubrics can improve fairness and reliability while keeping space for professional judgment. These results may support portfolio assessment design, teacher training, and course evaluation in visual arts, design, and interdisciplinary arts education. This study has some limits. The sample was drawn from selected arts courses, and the assessment focused on portfolio-based learning. Future research should include more institutions, longer observation periods, and different creative tasks to test the wider use of these findings.

Acknowledgments

We are grateful to all respondents who participated in this study.

References

1. Abdellatif, R., & El-Wakeel, H. (2025). Assessing creative outcomes in studio-based learning: a comparative assessment of analytical rubrics. *International Journal of Design Creativity and Innovation*, 13(1), 41-66.
2. Baracco, B., Escribano, N., Da Silva, D., Belliard, V., Ceballos, L., & Fuentes, V. (2025). Factors influencing the agreement between teachers and students in the assessment of preclinical endodontics using a rubric. *BMC Medical Education*, 25(1), 626.
3. Boichuk, I. V. (2025). Psychological and pedagogical aspects of objectivity in assessing productive speech activities in large-scale standardized testing. *Global Prosperity*, 5(4).

4. Call, M. L., Jiang, K., & Idso, C. (2026). Star advantage: Employee value creation and capture in the age of artificial intelligence. *Human Resource Management*, 65(1), 151-167.
5. Du, Y., Chen, Z., Liu, Q., & Dong, Y. (2026). Co-identification of Vibration, Structural Looseness, and Surface Damage in Industrial Equipment Operation Videos.
6. Gao, G., Gao, R., Lu, C., Gao, R., & Kuang, Y. (2026, March). Security Governance Methods and Quantitative Evaluation for Enterprise SMS and Verification Code Systems. In 2026 International Conference on Generative Artificial Intelligence and Information Security (GAIIS) (pp. 455-458). IEEE.
7. Hong, Z., Chen, H., & Xu, T. (2026). Long-Horizon Load Stability Testing and Degradation Forecasting for High-Concurrency Cloud Storage. Available at SSRN 7115859.
8. Huang, J., Xu, T., Yin, J., & Yang, J. (2026). Data Quality Monitoring and Intelligent Scheduling Optimization in Financial Data Workflow Automation. Available at SSRN 7179338.
9. Huang, J., Yin, J., Yang, J., & Xu, T. (2026). Construction of Audit Evidence Chain and SOX Control Verification Mechanism for Transaction-level Financial Data in High Volume E-commerce Platform. Available at SSRN 7179259.
10. Huang, R. (2026). Less Is More: When Data Augmentation Hurts Small CNN Generalization. *Journal of Computer and Communications*, 14(7), 1-11.
11. Jonsson, A., Panadero, E., Pinedo, L., & Fernández-Castilla, B. (2025). Using rubrics for formative purposes: identifying factors that may affect the success of rubric implementations. *Assessment in education: principles, policy & practice*, 32(2), 192-211.
12. Köroğlu, M. (2025). Inter-Rater Reliability Analysis in Performance-Based Assessment: A Comparison of Generalizability Coefficients and Rater Consistency. *Ahmet Keleşoğlu Eğitim Fakültesi Dergisi*, 7(2), 218-234.
13. Lee, S., Bustamante-Barreto, A., & Yoo, J. E. (2025). Toward a more discriminative and culturally responsive TALIS teacher self-efficacy in multicultural classrooms scale: cross-national evidence from Japan and Colombia using the generalized partial credit model. *Large-scale Assessments in Education*, 13(1), 35.
14. Li, J., Ma, R., & Gu, Y. (2026). From Audit Requirements to Computable Software Architecture: A Rule-Engine and Evidence-Indexing Method for Trusted Digital Infrastructure.
15. Lin, N., Xu, Y., Yang, H., Zhang, G., Zhang, M., Wang, S., ... & Li, X. (2020). Dissociating the neural correlates of the sociality and plausibility effects in simple conceptual combination. *Brain Structure and Function*, 225(3), 995-1008.
16. Lin, T., Spivack, Z., & Chen, A. (2026). Monitoring Teaching Quality and Validating Educational Equity in Public Art Education. Available at SSRN 6512778.
17. Phuong, H. Y., Le, T. T., & Pham, T. T. (2025). Assessment for learning or assessment for scoring? Washback of analytic rubrics in Vietnamese EFL classrooms. *Language Testing in Asia*, 15(1), 32.
18. Qi, C., & Qiao, X. (2026, May). From Traditional Machine Learning to Large Language Models: The Evolution of Infrastructure from an Engineering Perspective. In 2026 6th International Symposium on Computer Technology and Information Science (ISCTIS) (pp. 503-506). IEEE.

19. Sattar, G., Rustemova, S., & Nebessayeva, Z. (2025). Socially Interactive Approaches and Digital Technologies in Art Education: Developing Creative Thinking in Students During Art Classes. *Open Education Studies*, 7(1), 20250096.
20. Seow, E. (2026). When formal criteria fail: Judgment and validity in arts assessment. *Journal of Arts and Humanities*, 15(03), 01-05.
21. Shao, W. (2026). Design and Implementation of an Open-Source Security Framework for Cloud Infrastructure. *arXiv preprint arXiv:2604.03331*.
22. Shumaker, M. D., Rivers, M. L., & Tauber, S. K. (2025). Point values on scoring rubrics influence self-regulated learning for STEM material. *Behavioral Sciences*, 15(4), 532.
23. Wu, J., Zhang, J., Wu, D., & Peng, Y. (2026). Visual Transformation Mechanisms for Integrating Digital Cultural Heritage Resources into Junior High School Art Curricula.
24. Wu, J., Zhang, J., Wu, D., & Peng, Y. (2026). Visual Transformation Mechanisms for Integrating Digital Cultural Heritage Resources into Junior High School Art Curricula.
25. Xiong, W., & Guo, Y. (2026). A Study on the Phenomenon of Abnormal Load Concentration and Its Cross-system Transmission Characteristics in High-density Building Operations. Available at SSRN 7120661.
26. Xu, T., Zhang, J., & Zhu, W. (2026). Fairness-Accuracy Trade-offs and Calibration Mechanisms in Machine Learning-Based Subprime Credit Scoring. Available at SSRN 6893759.
27. Zhang, Y., Gu, W., & Wang, J. (2025, December). Construction of Wind Farm Asset Health Index Based on Multi-Dimensional Indicators and Analytic Hierarchy Process and Its Correlation with Operational Performance. In *Proceedings of the 2025 International Conference on Digital Transformation and Management* (pp. 317-323).
28. Zhang, Z., Gao, Y., & Tong, Y. (2026). CausalML4CX: A Causal Inference and Machine Learning Framework for Customer Experience Optimization with Application in Banking.
29. Zhang, Z., Wang, J., Li, Z., Wang, Y., & Zheng, J. (2025). Annocoder: A mti-agent-based code generation and optimization model. *Symmetry*, 17(7), 1087.
30. Zheng, J., & Makar, M. (2022). Causally motivated multi-shortcut identification and removal. *Advances in Neural Information Processing Systems*, 35, 12800-12812.