Teacher-Involved Automatic Characteristics Classification in Handwritten Math Answer Process

Shunsuke TONOSAKI[a*], Taito KANO[a], Chia-Yu HSU[b] & Hiroaki OGATA[b]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*tonosaki.shunsuke.75i@st.kyoto-u.ac.jp

Abstract: In Japanese junior high schools, while digital logs of students' handwritten math answers offer utilization opportunities for classroom sharing, manually reviewing them is burdensome for teachers, and existing automatic classification methods often fail to meet their pedagogical needs. This study, co-designed with teachers, defined four pedagogically classification labels, identified their associated handwriting features, and subsequently evaluated classification models. Among the models tested, XGBoost was most effective, notably meeting our success criterion (precision > 0.5) for teacher support on the diagram problem type (0.532), thus demonstrating the practical feasibility of this approach. Our findings highlight the importance of a teacher-involved approach to designing learning analytics systems for practical classroom use.

Keywords: learning analytics, learning logs, handwriting process

1. Introduction

In Japanese elementary and junior high schools, students are provided with a digital device such as PC, tablets and pen devices for educational purposes through the GIGA School Program, and these devices are now used instead of paper for math problem-solving (MEXT, 2020). This enables the accumulation of students' handwriting processes as learning logs (Ogata et al., 2018), and visualization through the conversion and replay of these logs as videos has been achieved (Yoshitake et al., 2020). Teachers can use the replay function to observe students' thought processes regarding their difficulties and diverse ways of thinking, share these with the entire class, and create opportunities for mutual learning.

However, checking each student's log requires significant time and effort, placing a heavy burden on teachers. To reduce the burden on teachers and enable them to quickly access answers that match their purpose for sharing answers, it is important to provide support for summarizing and classifying the characteristics of each answer.

While various studies have classified handwritten answers, their criteria are often defined by researchers. These have included rule-based models focused on student stumbles (Asai et al., 2012), machine-learning for clustering (Yoshitake et al., 2020) or predicting performance (Stahovich & Lin, 2016), and LLMs for content recognition (Caraeni et al., 2025). However, these approaches did not consider labels required by teachers, even though incorporating user perspectives is crucial in Learning Analytics (Mavrikis et al., 2019).

This study proposed a teacher-involved co-design process for defining classification labels and criteria for handwritten math answers. To verify the feasibility of this co-design approach, several basic classification models were developed and evaluated. To guide this process, we address the following research questions:

RQ1: Which categories of handwriting process characteristics can be derived from teachers' observations?

RQ2: Which features of the handwriting process affect the teacher-focused characteristics?

RQ3: Is it feasible to classify handwritten answers based on teacher-defined labels for real-world practice?

2. Context

This study analyzed the handwriting problem-solving processes of 226 first- and second-year junior high school students in Japan using the Learning and Evidence Analytics Framework (LEAF) system (Ogata et al., 2018). The problems were answered using a pen device on the eBook tool (Book Roll) within the LEAF system, and two teachers—one proficient in using its analysis tool (Log Palette) for student assessment, and the other an experienced math teacher skilled in interpreting answer characteristics—reviewed replay videos of each student's answer using Log Palette (Yoshitake et al., 2020). The dataset comprised 292 description-type (law of exponents) and 222 diagram-type (inscribed angle theorem) answers, derived from three problems of each type presented one problem per page (Figure 1, left). We used two problem types because prior research suggests that both student handwriting behavior and the characteristics teachers focus on can differ by problem type (Tonosaki et al., 2024).

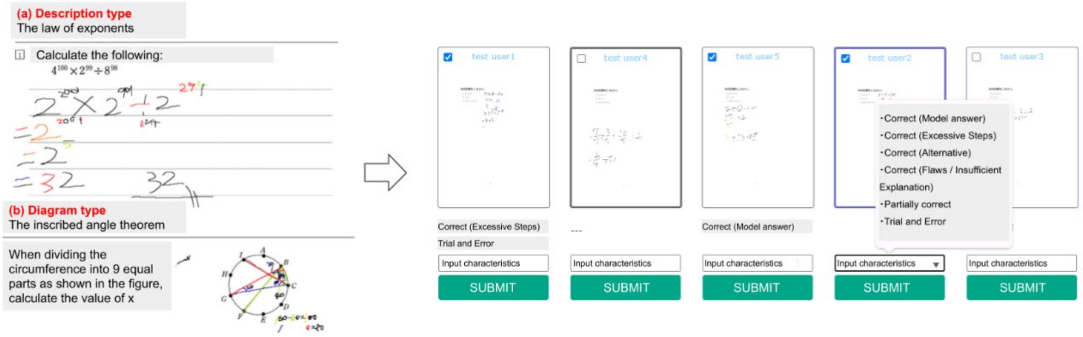

Figure 1. Examples and the Labeling Function of Handwriting Answering Process

To collect data for analysis, we first asked two teachers to label answers suitable for sharing in class, using a Log Palette function that allowed predefined or custom labels to each answer (Figure1, right). For the subsequent automatic classification, we used two data types from the logs: the performance score (*PFM*) and 13 handwriting process features (Table 1), all of which were standardized per problem type.

Table 1. *Handwriting Process Features*

Features	Description
TAT	Total answering time (second)
TNS	Total number of strokes
TNE	Total number of erasers
Speed	Whole answering speed (*TNS* / *TAT*)
AD	Average of duration
TSD	Total number of short durations (< *AD*) from each previous stroke
TLD	Total number of long durations (≥ *AD*) from each previous stroke
VD	Variance of duration
AST	Average of stroke time (second)
TSL	Total number of stroke lengths
ASL	Average of stroke lengths
ASLS	Average of each stroke speed (stroke length / stroke time)
VSLS	Variance of each stroke speed

3. Analysis & Results

3.1 Categorized Labels (RQ1)

We unified and categorized teacher labels through semi-structured interviews and the Card Sorting method (Upchurch et al., 2001). In this process, teachers first discussed the meaning and criteria for each label and then grouped them based on their practical classroom use. We refined these provisional categories, and the final labels and their definitions were confirmed by the teachers via email, at which point data saturation was reached.

The initial 12 labels provided by teachers, reflecting criteria such as performance, content, and stroke movement, were grouped into the four final classification labels. These labels, whose definitions are shown in Table 2, had the following distribution: "*Standard Model Answer*" (n=33 Description, n=14 Diagram), "*Creative Alternative Answer*" (n=15, n=12), "*Trial-and-Error or Mistake Answer*" (n=6, n=68), and "*Unclear Process Answer*" (n=1, n=37).

Table 2. *Categorized Labels and Their Definitions by Teachers (translated)*

Categorized Labels	Definition
Standard Model Answer	Uses appropriate mathematical notation and terminology, proceeds smoothly with sufficient detail, and reaches an expected answer.
Creative Alternative Answer	An answer reached through an original process, different from the model answer.
Trial-and-Error or Mistake Answer	Regardless of correctness, the answer shows visible trial and error or mistakes recognizable by others.
Unclear Process Answer	Regardless of correctness, the answer includes unclear steps that require additional explanation from the student.

3.2 Feature Selection & Analysis (RQ2)

To identify characteristic features, we examined the relationship between the features in Table 1 and categorized labels using two approaches: a teacher-driven questionnaire and a data-driven analysis. First, in the questionnaire, two teachers rated the relationship between each feature and label as 'Positive,' 'Negative,' or 'No Relationship.' Second, for the data-driven analysis, we used SHAP to quantify the explanatory power of each feature independently of the model (Lundberg & Lee, 2017). We trained an XGBoost model and considered the top five features ranked by their SHAP values as highly explanatory for each label.

Comparing the questionnaire and data analysis (Table 3), several features were commonly identified as significant, such as *PFM* for "*Standard Model Answer*" (description-type), *TAT* for the same label (diagram-type). In contrast, feature selection for “*Unclear Process Answer*” was challenging, as teachers reported 'No Relationship' for the diagram-type, and insufficient data prevented data-driven analysis for the description-type.

Table 3. *Results of each Feature Selection Method (Positive: +, Negative: -)*

Labels	Problem type	Questionnaire	Data Analysis (SHAP)
Standard Model Answer	*Description*	*PFM+*, *TAT+*, *TNS+*, *Speed+*	*TNE+*, *ASLS-*, *PFM+*, *VSLS-*, *Speed-*
	Diagram	*TAT+*, *TNS-*, *TNE-*, *Speed-*, *ASLS+*, *AD-*, *TSD-*	*PFM+*, *TAT+*, *TSD+*, *AST+*, *VSLS+*
Creative Alternative Answer	*Description*	*TAT-*, *TSL+*	*ASLS-*, *PFM+*, *AD-*, *TAT+*, *Speed+*
	Diagram	*TAT+*, *TNS+*, *TNE+*, *Speed+*, *ASLS+*, *AD+*, *TSD-*	*TNS+*, *VD-*, *AST+*, *PFM+*, *VSLS+*
Trial-and-Error or Mistake Answer	*Description*	*TAT+*, *TNE+*, *AD+*	*Speed+*, *AST+*, *AD-*, *PFM-*, *ASLS-*
	Diagram	*TNE+*, *Speed+*, *ASLS-*, *VSLS+*, *TSL+*, *ASL+*, *AST-*	*Speed-*, *VD+*, *TLD+*, *TSD+*, *TNS-*
Unclear Process Answer	*Description*	*TAT-*, *TNS-*, *Speed-*, *AD-*	(Nothing)
	Diagram	(Nothing)	*VSLS+*, *TNS-*, *ASL+*, *TSD+*, *VD-*

3.3 Multi-Label Classification (RQ3)

To verify the feasibility of automatically classifying answers using the teacher-defined labels, we evaluated three multi-label classification models, each offering different practical benefits: a transparent teacher-driven (rule-based) model, a high-performance data-driven (machine-learning) model, and a content-aware LLM-driven model. For evaluation, we prioritized macro-average precision to ensure the reliability of labels suggested to teachers. To define the effectiveness criterion for RQ3, we set a success threshold of a precision of at least 0.5. The threshold was chosen because it implies that if a teacher reviews two answers flagged by the models with a certain label at least one will be correctly classified, which we considered a sufficient level for supporting teachers. We also calculated Recall, F1 score, and Accuracy as the overall evaluations of the models. All models used the 13 handwriting features (Table 1) and performance score (*PFM*) as input.

First, for the teacher-driven classification, we implemented a rule-based model. We tested four feature selection strategies based on two sources—teacher questionnaires and data analysis (SHAP): (1) using only features from the questionnaire, (2) using only features from data analysis, (3) using features common to both (AND), and (4) using features from either source (OR). The classification rule was based on whether a feature's value exceeded its median, with the threshold adjusted to maximize precision.

Second, we implement machine-learning using handwriting process features for data-driven classification. The XGBoost, Support Vector Machine (SVM), and Random Forest were used in this study. For the three models, we conducted classification tasks using the features selected by SHAP, followed by parameter tuning using Grid Search. The models' performances were evaluated using 5-fold cross-validation, and the differences between the machine-learning models were tested for statistical significance using paired t-tests.

Third, for the LLM-driven classification, we used OpenAI's GPT-4o for multi-label image recognition. The prompt provided the model with the problem context, the student's answer image, label definitions, the desired output format, and few-shot examples, and instructed it as follows: "Look at the answer below and, based on the definitions, assign all applicable labels (all independent) accurately and return them in an array." We then tested three prompt variations: one with no features (None), one with all 14 features (All), and one with a pre-selected subset of features (Selected). For the 'Selected' condition, we identified the top five features that most improved precision when added to the prompt individually in a preliminary test. All features provided to the LLM were expressed in natural language on a five-point scale (e.g., "Answering Time: Top").

As shown in Table 4, the machine-learning models significantly outperformed the other approaches. XGBoost achieved the highest precision for the description-type (0.496) and the diagram-type (0.532), with the latter meeting our success criterion. Paired t-tests confirmed that XGBoost's precision was statistically significantly higher than the other models for both types ($p < .05$). In contrast, both the rule-based and LLM-based models resulted in low precision.

Table 4. *Evaluations of Each Multi-Label Classification Model*

	Method	Precision	Recall	F1 Score	Accuracy
(a) Rule-Based (comparison by feature selections)					
Description	Only Questionnaire	0.106	0.438	0.161	0.716
Diagram	OR	0.200	0.551	0.268	0.650
(b) Machine-Learning (comparison by models)					
Description	XGBoost	0.496	0.501	0.496	0.880
Diagram	XGBoost	0.532	0.525	0.515	0.747
(c) LLM (comparison by feature selections)					
Description	Selected	0.128	0.348	0.182	0.107
Diagram	None	0.149	0.283	0.187	0.107

4. Discussion

4.1 Co-Designing Pedagogical Labels (RQ1)

Through our labeling process with teachers, we identified four types of labels intended for classroom use. Based on the interview responses and our interpretation, each label has a distinct usage scenario. For example, “*Standard Model Answer*” can be used as exemplary and "*Creative Alternative Answer*" to spark student interest. "*Trial-and-Error or Mistake Answer*” is seen as opportunities to share common mistakes, while "*Unclear Process Answer”* serves as a reminder for students to show their work.

Unlike labels from prior work aimed at detecting when students stumble (Asai et al., 2012), our labels were defined for the pedagogical purpose of sharing answers in class. This highlights the need to design classification labels flexibly according to their educational use.

4.2 Analysis of Teacher-Focused Features (RQ2)

Our analysis revealed a difference in focus between the teacher questionnaires and the data analysis (Table 3). The teachers tended to select visually distinguishable, time-level features like total answering time (*TAT*), while the data-driven approach identified fine-grained, stroke-level features (e.g., *ASLS*, *AD*). These features reflected different student states depending on the problem type. For description-type problems, stroke timing features were prominent, suggesting that a student's calm or rushed state was a key indicator. For instance, a "*Standard Model Answer*" was associated with calm, consistent work. For diagram-type problems, features related to process accumulation like the number of strokes (*TNS*) were more significant, reflecting statuses such as having a clear goal despite many steps ("*Creative Alternative Answer*") or making repeated corrections ("*Trial-and-Error or Mistake Answer*").

Notably, performance (*PFM*) was rarely a top feature, distinguishing our study from those focused on performance prediction (Caraeni et al., 2025; Stahovich & Lin, 2016) and highlighting the value of behavioral data in capturing teacher-defined characteristics.

4.3 Effectiveness of Classification for Teacher Support (RQ3)

Our results show that the XGBoost machine-learning model performed significantly better than the other models. Notably, it achieved effective classification for the diagram-type (0.532) by meeting our success criterion (precision > 0.5), though it fell just short for the description-type (0.496). This demonstrates the practical feasibility of building a teacher-support tool, at least for diagram-type problems. Such a tool can effectively surface relevant student answers for review, thereby reducing teachers' search effort. In contrast, the rule-based and LLM-based approaches showed limited effectiveness.

The differing results among the models can be attributed to the unique complexity of the handwriting process. Machine-learning models likely succeeded by identifying complex patterns within this behavioral data. In contrast, the rule-based model's low precision suggests that simple rules are insufficient. This difficulty implies that the process's complexity may be a factor in the burden teachers face when identifying answer characteristics. Similarly, the LLM likely struggled as it analyzed only the final static image, lacking access to the crucial time-series data of the writing process.

4.4 Limitation & Future Work

This study has several limitations that suggest directions for future work, including: (1) Data Scale and Imbalance: The data was labeled by only two teachers, and the inter-rater reliability was not calculated, resulting in a small, imbalanced dataset that could lead to overfitting. Future work should involve more teachers and problems to create a more robust data set. (2) Feature Selection: As our use of SHAP may not be optimal for all models (e.g., SVM). Employing model-specific feature selection techniques could further improve performance. (3)

LLM Application: The LLM approach showed low precision and high computational costs. Performance could be enhanced through prompt optimization and fine-tuning, while costs could be reduced by using open-source models. (4) Hybrid Model Development: A key future direction is to combine the teacher-driven, data-driven, and LLM approaches. Integrating rule-based pedagogical intent with machine-learning's behavioral pattern recognition and the LLM's content understanding could create a more accurate and educationally relevant hybrid system. (5) Practical System Implementation: To ensure any resulting tool effectively reduces teacher burden, it is crucial to visualize the classification results and continue the co-design process by including teacher feedback from the prototype stage (Mavrikis et al., 2019).

5. Conclusion

This study, co-designed with teachers, established four pedagogical labels for classifying handwritten math answers intended for classroom sharing. We identified the process features influencing these labels and found that among the models tested, the XGBoost machine-learning model was the most effective at automatic classification. While not for full automation, these models offer practical support by helping teachers efficiently select relevant answers for classroom discussion. This study's primary contribution is its teacher-involved methodology, demonstrating the importance of co-designing learning analytics systems to reflect classroom needs, especially when dealing with abstract data like handwriting processes.

Acknowledgements

This work was partly supported by the Council for Science, 3rd SIP JPJ012347, and JSPS KAKENHI, Grant Number 23H00505.

References

Asai, H., Nozawa, A., Sonoda, S., & Hayato, Y. (2012). *Onrain tegaki dēta o mochiita gakushū-sha no tsumazuki kenshutsu [Student's stumble detection using online handwriting data]*. DEIM Forum A8-4. http://db-event.jpn.org/deim2012/proceedings/final-pdf/a8-4.pdf

Caraeni, A., Scarlatos, A., & Lan, A. (2025). Evaluating GPT-4 at grading handwritten solutions in math exams. *15th International Conference on Learning Analytics & Knowledge (LAK25)*, 126–128. https://www.solaresearch.org/wp-content/uploads/2025/02/LAK25_CompanionProceedings-Final.pdf

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1705.07874

Mavrikis, M., Geraniou, E., Gutierrez Santos, S., & Poulovassilis, A. (2019). Intelligent analysis and data visualisation for teacher assistance tools: The case of exploratory learning. *British Journal of Educational Technology: Journal of the Council for Educational Technology*, *50*(6), 2920–2942. https://doi.org/10.1111/bjet.12876

MEXT. (2020). *The image of the transformation of learning brought by "1 device for 1 student with a high-speed network."* https://www.mext.go.jp/en/content/20200716-mxt_kokusai-000005414_04.pdf

Ogata, H., Majumdar, R., Akçapınar, G., Hasnine, M., & Flanagan, B. (2018). *Beyond learning analytics: Framework for technology-enhanced evidence-based Education and learning*. http://hdl.handle.net/2433/237322

Stahovich, T. F., & Lin, H. (2016). Enabling data mining of handwritten coursework. *Computers & Graphics*, *57*, 31–45. https://doi.org/10.1016/j.cag.2016.01.002

Tonosaki, S., Kano, T., Hamada, S., Horikoshi, I., & Ogata, H. (2024). Toward contextualized handwriting process analysis: Comparison between problem types in math. *The 32nd International Conference on Computers in Education*. https://doi.org/10.58459/icce.2024.5063

Upchurch, L., Rugg, G., & Kitchenham, B. (2001). Using card sorts to elicit Web page quality attributes. *IEEE Software*, *18*(4), 84–89. https://doi.org/10.1109/MS.2001.936222

Yoshitake, D., Flanagan, B., & Ogata, H. (2020). Supporting group learning using pen stroke data analytics. *28th International Conference on Computers in Education Conference Proceedings*, *1*, 634–639. https://repository.kulib.kyoto-u.ac.jp/dspace/handle/2433/259798