# Explain-from-Stroke: Capturing Invisible Learning Processes Through Handwriting Dynamics Analysis

**Ryosuke Nakamoto[1], Brendan Flanagan[1], Kohei Nakamura[2], Hiroaki Ogata[1]**

[1]Kyoto University, Kyoto, Japan
[2]Osaka Kyoiku University, Osaka, Japan
s0527225@gmail.com

## Abstract

Educational assessment requires understanding student problem-solving processes, not just final answers. Current AI-driven analytics focus on static outcomes, missing valuable insights from temporal dynamics. We present Explain-from-Stroke, a practical framework that captures invisible learning processes by integrating handwriting dynamics with vision-language models. Our approach extracts temporal features—writing speed, pauses, and revisions—providing supplementary context for generating meaningful insights into hidden aspects of student reasoning. Deployed with real classroom data from a Japanese secondary school, our system demonstrates 18.2% improvement in cognitive depth analysis over static approaches. This work provides educators with accessible process-oriented analysis that reveals invisible learning processes using standard tablet technology.

## 1 Introduction

Educational assessment traditionally emphasizes correctness over process understanding. Two students might reach identical answers through vastly different reasoning—one via deep conceptual understanding, another through memorization. This product-focused approach limits educators' ability to provide targeted instructional support. Digital writing technologies offer new opportunities for process-oriented assessment. While sophisticated cognitive tools exist, they remain impractical for classroom deployment (Oviatt, Lin, and Sriramulu 2018). Pen stroke data from standard tablets provides accessible insights into student thinking without additional infrastructure. Recent vision-language models show educational promise but remain limited to static analysis (Kasneci et al. 2023). The DrawEduMath benchmark reveals significant improvement opportunities in educational content analysis (Baral et al. 2025).

**Cognitive basis of pauses.** Our approach is grounded in evidence that pauses in writing often index higher-level planning and cognitive restructuring rather than mere motor delay. Process-tracing studies show that writing alternates between execution bursts and pauses associated with planning and increased cognitive effort (Alves, Olive, and Castro 2008; Limpo and Alves 2018). These findings motivate treating kinematic pauses as probabilistic indicators of cognitive load rather than deterministic labels.



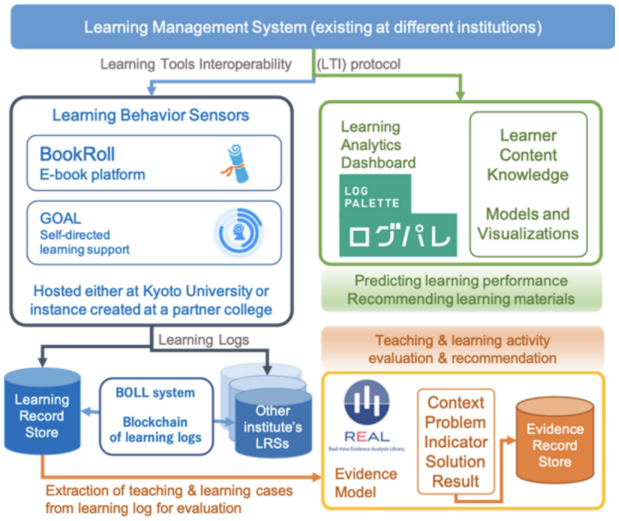
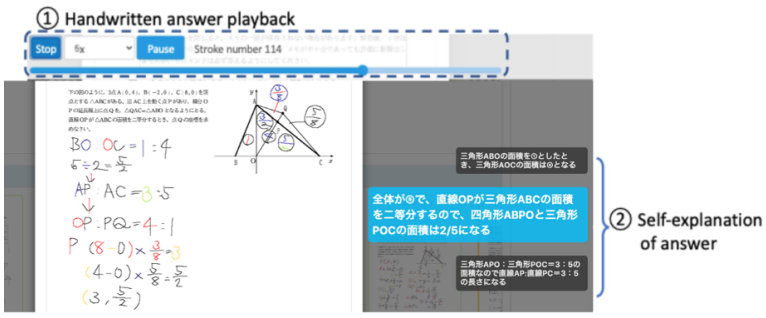


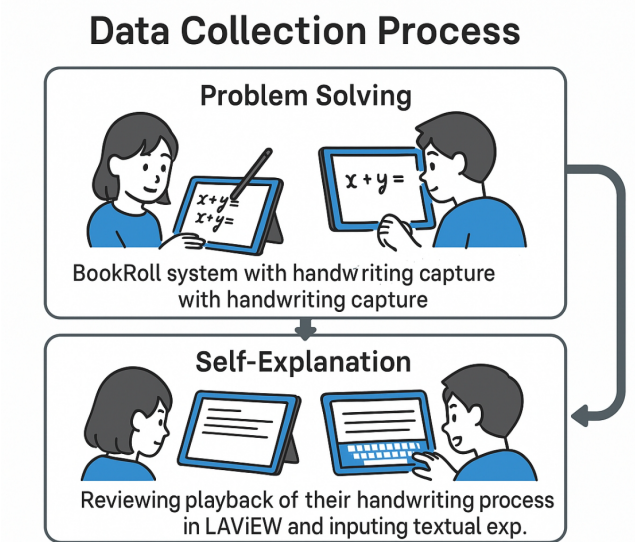

Figure 1: LEAF Architecture (adapted from Ogata et al. (Ogata et al. 2022)).

This work addresses: **Can temporal handwriting data capture invisible learning processes and generate explanations that reveal hidden aspects of student reasoning?** We focus on practical educational applications that make invisible cognitive processes visible.

**Key Contributions:** (1) **Invisible Process Capture**: A practical framework revealing hidden learning processes by supplementing static analysis with temporal behavioral insights. (2) **Accessible Deployment**: Implementation using standard tablet technology without specialized infrastructure. (3) **Real-World Validation**: Evidence from classroom deployment showing meaningful improvements in capturing invisible reasoning processes. (4) **Hidden Cognitive Insights**: Demonstration of enhanced understanding of student strategies invisible in traditional analysis.




(a) Student UI for answer playback and self-explanation (from Nakamoto et al. (2024)).




(b) Data Collection Process.

Figure 2: Data collection interface (a) and workflow (b).

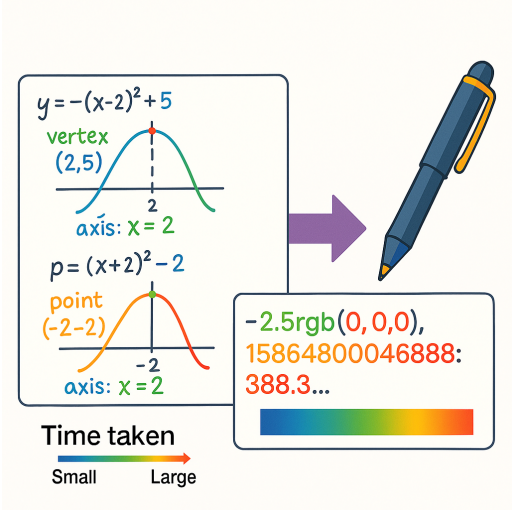



Figure 3: Pen Stroke Data sample (synthetically generated for illustration).

## 2 Related Work

### 2.1 Self-Explanation in Education

Self-explanation enhances learning by encouraging reasoning articulation (Chi et al. 1994; Rittle-Johnson, Loehr, and Durkin 2017). Students generating explanations achieve deeper understanding through inference generation, knowledge integration, and misconception repair (Chi 2000; Durkin 2011). VanLehn's work identifies learning impasses as critical understanding catalysts (VanLehn et al. 2003). Recent research shows combining textual explanations with handwriting features improves impasse detection from 74.0% to 80.06% (Nakamoto et al. 2025), suggesting complementary value in temporal behavioral data.

### 2.2 Vision-Language Models in Education

Large language models create new educational technology possibilities (Kocon et al. 2023). However, current VLMs remain limited to static analysis, unable to capture temporal dynamics revealing cognitive processes (Bubeck et al. 2023). Educational applications face consistency challenges across diverse contexts (Baral et al. 2025).

### 2.3 Handwriting Dynamics and Cognition

Research establishes strong connections between handwriting behavior and mental processes. EEG studies reveal greater and more widespread brain connectivity patterns during handwriting compared to typing (Van der Weel and Van der Meer 2024). The distinction between "thinking time" and "writing time" correlates with higher-order versus motor functions (Andersen et al. 2021). Digital pen features successfully model cognitive performance in medical assessments (Barz et al. 2020), suggesting handwriting dynamics provide valuable cognitive insights complementing traditional methods.

### 2.4 Data Structure

Pen stroke data captures timestamped coordinates with metadata:

```
<2.5rgb(0,0,0),1586480064688:
388.300:192.637>
```

This encodes stroke width, color, timing, and spatial coordinates, as illustrated in Figure 3. Temporal analysis reveals pause patterns indicating reflection; spatial analysis captures writing speed and complexity. These multimodal signals enable inference of cognitive states invisible in final products.

## 3 Methodology

### 3.1 Educational Context and Data Collection

**Deployment Environment.** Data were collected at a Japanese secondary school using the Learning and Evidence Analytics Framework (LEAF) (Flanagan and Ogata 2018), which supports tablet-based handwriting capture and learning analytics. LEAF integrates stroke logging, playback interfaces, and explanation input, as illustrated in Figures 1, 2a, and 2b.

**Data Collection Protocol.** From January 2021 to June 2023, students solved mathematics problems using digital

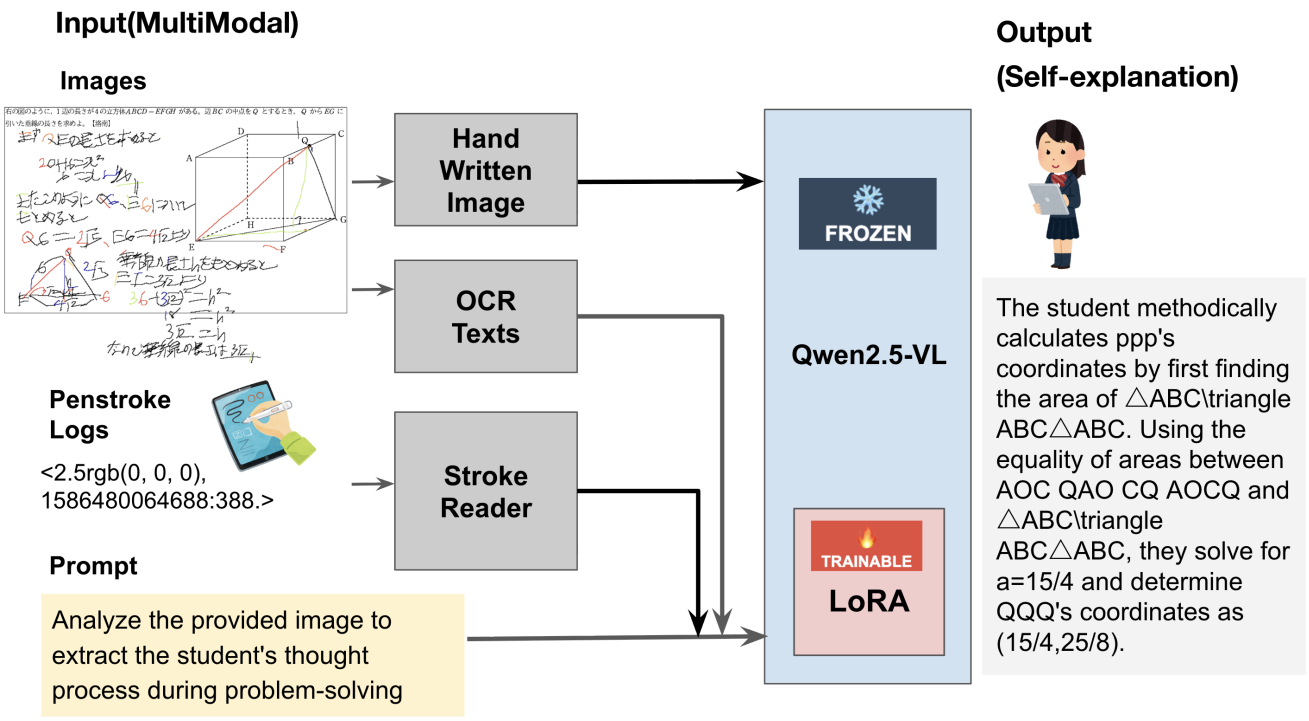



Figure 4: Explain-from-Stroke Model Architecture

pens on tablets, with full stroke-level logs recorded. After solving, students reviewed their answer playbacks and provided written self-explanations. Institutional ethics procedures and student consent were obtained.

A total of 2,300 authentic classroom sessions were collected. To increase coverage of reasoning patterns while preserving educational authenticity, GPT-4o generated additional thoughtful variants, expanding the dataset to 8,160 samples. After filtering low-quality or incomplete traces, the final dataset comprised 6,301 training, 1,340 validation, and 96 test samples, with 423 excluded.

## 3.2 Process Analysis Framework

Handwriting data consist of timestamped coordinate sequences:

$$S = \{(t_i, x_i, y_i)\}_{i=1}^{n}.$$

**Temporal Features.** For each stroke point, we compute the time delta

$$\Delta t_i = \frac{t_i - t_{i-1}}{1000},$$

and instantaneous velocity

$$v_i = \frac{\|\mathbf{p}_i - \mathbf{p}_{i-1}\|}{\Delta t_i}, \quad \mathbf{p}_i = (x_i, y_i).$$

We extract summary statistics such as point count $n$, total writing time $T_{total}$, mean and median velocity, and maximum velocity.

**Behavioral Features.** Pauses are identified as

$$P = \{i : \Delta t_i > 0.6\ \text{s}\},$$

indicating potential planning or cognitive restructuring episodes.

**Stroke Classification.** Strokes are categorized by length and speed into {short, medium, long} × {slow, moderate, fast} based on $(n)$, $(T_{total})$, and $s = n/T_{total}$.

**Educational Feature Processing.**

1. **Data Parsing**: Extract stroke width, timestamp, and coordinates.
2. **Temporal Segmentation**: Segment strokes using 600ms gaps as cognitive pause indicators.
3. **Feature Extraction**: Compute writing speed, pause duration, spatial complexity, and variance.
4. **Educational Mapping**: Interpret temporal and kinematic features as behavioral proxies of invisible learning processes.

## 3.3 Model Architecture and Training

**Model Variants.** To isolate the effect of temporal features, we evaluate three configurations sharing the same VLM backbone:

- **Base Model**: Qwen2.5-VL-3B-Instruct used without fine-tuning. Inputs: problem image + OCR text.
- **No Stroke**: Fine-tuned model receiving image + OCR tokens only. Controls for the effect of task-specific fine-tuning without temporal data.
- **With Stroke**: Identical to No Stroke but additionally receives serialized temporal descriptors. Inputs: image + OCR text + temporal stroke features.

| Model | Fine-Tuned | Temporal | LoRA |
|---|---|---|---|
| Base | No | No | No |
| No Stroke | Yes | No | Yes |
| With Stroke | Yes | Yes | Yes |

Table 1: Summary of the three model variants.

| Metric | Base | No Stroke | With Stroke | % Change |
|---|---|---|---|---|
| Overall Score | 6.12 | 6.17 | **6.31** | +2.27% |
| Content Accuracy | 9.65 | 9.70 | **10.00** | +3.09% |
| **Cognitive Depth** | 1.80 | 1.65 | **1.95** | **+18.18%** |
| Clarity | 7.60 | **8.00** | 7.60 | -5.00% |
| **Prof. Quality** | 2.40 | 2.20 | **2.60** | **+18.18%** |
| Completeness | 9.12 | 9.29 | **9.41** | +1.29% |

Table 2: Model performance comparison. % Change is With Stroke vs. No Stroke. The most significant improvements are highlighted.

**Architecture.** Figure 4 illustrates the system pipeline. Images are encoded by the visual encoder and text is processed by the language encoder. For the **With Stroke** model, temporal features are serialized as lightweight key–value tokens (e.g., `pause_duration:0.8s; avg_speed:0.45px/ms`) and appended to the OCR token sequence. This design preserves the base architecture while allowing the model to jointly attend to visual, textual, and temporal signals.

**LoRA Fine-Tuning.** Both the No Stroke and With Stroke models are fine-tuned using Low-Rank Adaptation (LoRA) applied to all projection layers. We set `r = 16` and `alpha = 32`. Training uses `max_steps = 4000`, learning rate `2e-5`, batch size `2` with gradient accumulation `4`, warm-up of `400` steps, and cosine scheduling. The Base model is not fine-tuned.

### 3.4 Educational Evaluation

Model outputs were evaluated using an LLM judge (GPT-4o-mini) across five educational criteria: *Content Accuracy*, *Cognitive Depth*, *Clarity*, *Professional Quality*, and *Completeness*. Evaluation prioritized insight into learners' reasoning processes rather than surface-level answer descriptions, aligning with formative assessment principles.

Temporal descriptors were serialized and appended to the OCR text using key–value tokens (e.g., `pause_duration:0.8s; burst_length:12`), enabling the language module to learn temporal–semantic correspondences without modifying the visual encoder.

## 4 Results

### 4.1 Educational Impact Assessment

Evaluation on the 96 test samples demonstrates that incorporating temporal stroke features leads to measurable improvements across multiple educational dimensions (Table 2). The stroke-aware model achieves the highest overall score among the three systems, with especially notable gains in **Cognitive Depth** (+18.18%) and **Professional Quality** (+18.18%) compared with the No Stroke baseline. These metrics capture the richness and structure of the generated explanations, indicating that temporal descriptors influence how the model characterizes student reasoning.

Content Accuracy also improves slightly, suggesting that temporal cues contribute to clearer identification of relevant information from handwritten work. Meanwhile, Clarity shows a minor decrease relative to the No Stroke model, reflecting a tendency toward longer or more detailed explanations when temporal features are present.

To assess statistical reliability, a Wilcoxon signed-rank test was applied to paired outputs from the No Stroke and With Stroke conditions. Differences in Cognitive Depth and Professional Quality were statistically significant ($p < 0.05$), supporting the robustness of the observed improvements.

### 4.2 Qualitative Output Characteristics

Representative examples from the evaluation set (Table 3) highlight how temporal features influence model output. In Samples 16 and 46, the stroke-aware model produces more detailed descriptions of student actions, including inferred ordering of steps or shifts in attention. By contrast, in Sample 82, irregular or ambiguous temporal cues result in less coherent explanations.

These examples illustrate characteristic patterns in model behavior but do not by themselves determine their educational interpretation. We further examine broader implications in the Discussion section.

## 5 Discussion

### 5.1 From Kinematics to Cognition

Our results demonstrate that pen stroke dynamics provide a unique window into cognitive processes. When students hesitate, the pen stops moving—this pause represents more than a null signal; it indicates cognitive load, planning, or memory retrieval (Andersen et al. 2021). The With Stroke model processes these kinematic signals as proxies for real-time cognitive states, moving beyond a surface-level "multi-structural" understanding to a "relational" understanding. As seen in Samples 16 and 46, this enables the model to identify the sequence of physical actions, moving beyond surface-level analysis to deeper process understanding. However, this reliance on kinematic data is not without risks. The failure case in Sample 82, where the With Stroke model hallucinates a nonsensical numerical procedure, suggests that noisy or ambiguous stroke patterns can be misinterpreted. This highlights a critical trade-off: while stroke data unlocks deeper process analysis, it can also introduce a new vector for error if not robustly interpreted.

### 5.2 Educational Applications

This approach supports a shift from summative, product-focused assessment toward formative, process-aware feedback. Building on Nakamoto et al. (2025), potential applications include: **Real-time Support**: AI tutors can respond to extended pauses with scaffolds or strategy prompts. **Classroom Analytics**: Teachers can view dashboards that surface

| Sample ID | Educational Significance | Generated Educational Insights |
|---|---|---|
| 16 | **Strategic Approach Detection**: Process analysis suggests systematic construction approach through temporal pattern inference. | **Base Model:** "The task here is first to understand what information the problem itself holds. This demonstrates the basic cognitive process of understanding the problem details."<br>**No Stroke:** "To solve this problem, it is first necessary to understand the problem's requirements. Here, they are trying to calculate the length of AE using similar triangles."<br>**With Stroke:** "**Based on temporal patterns in the handwriting data, the analysis suggests systematic geometric construction, with inferred line drawing from point A and methodical connection to points B and C.**" |
| 46 | **Process Organization**: Enhanced model infers systematic problem organization through temporal behavioral patterns. | **Base Model:** "The student first identifies 'a certain point D and E' by visually storing this part. This is a visual cognitive process."<br>**No Stroke:** "The student first confirms the shape of triangle ABC and its angles/sides. They also reconstruct the properties of triangles and proof methods."<br>**With Stroke:** "**Temporal analysis suggests the student began with systematic triangle construction ($\Delta$ABC), then organized given conditions ($\angle$A=35°, $\angle$B=65°) based on inferred writing patterns.**" |
| 82 | **System Limitation**: Demonstrates potential for misinterpretation when temporal patterns are ambiguous or noisy. | **Base Model:** "The student's action of first organizing the expression is important. Here, the task of eliminating 'x-2' and 'x' from the denominator requires mathematical understanding."<br>**No Stroke:** "To solve this problem, one must first organize the expression. The goal is to eliminate the denominator."<br>**With Stroke:** "**The temporal analysis attempts to infer procedural steps but generates unclear sequences, highlighting limitations in pattern interpretation when behavioral cues are ambiguous.**" |

Table 3: Qualitative comparison of generated educational insights.

learners showing prolonged difficulty, following approaches such as Kishi and Miura (Kishi and Miura 2018).

### 5.3 Model Capabilities and Limitations

The 18.2% improvement in cognitive depth reflects the ability of modern LLMs to synthesize abstract temporal information into coherent educational explanations. Contemporary vision-language models can infer systematic approaches from basic temporal features such as pause duration, writing speed, and stroke complexity. However, Sample 82 illustrates critical limitations when temporal patterns are ambiguous. The model produces nonsensical procedural sequences when presented with irregular pause patterns and inconsistent writing speeds. This failure mode suggests that while modern LLMs excel at pattern recognition and narrative generation, they may generate overconfident interpretations when underlying data is noisy or insufficient. **Individual Variation Challenges**: Handwriting patterns vary due to non-cognitive factors, meaning a "systematic" pattern may just be personal preference, not a deliberate cognitive strategy. Consequently, temporal cues must be interpreted as probabilistic indicators, not deterministic mappings, underscoring the need for context and uncertainty displays.

### 5.4 Validation Requirements and Future Research

Since LLM interpretative capabilities vary significantly, future research must employ multiple model architectures (GPT-4o, Claude, Gemini) to distinguish between robust pattern recognition and model-specific artifacts. **Methodological Recommendations**: Future work should focus on: (1) developing robust validation frameworks that distinguish genuine pattern recognition from model-specific artifacts, (2) establishing uncertainty quantification for temporal interpretations, (3) validating findings with think-aloud protocols to understand actual cognitive processes, and (4) creating individual baseline patterns to distinguish cognitive from motor behaviors.

## 6 Conclusion

Temporal handwriting cues enhance process-aware reasoning in VLMs, yielding meaningful gains in cognitive depth and interpretive quality. By integrating lightweight temporal descriptors with visual and textual inputs, Explain-from-Stroke reveals aspects of student reasoning that remain invisible in static assessments. These results highlight the value of behavioral traces for supporting more nuanced educational analytics and point toward future work on robustness, fairness, and multi-modal process modeling.

## Acknowledgments

This work was supported by JSPS Grant-in-Aid for Scientific Research (B) JP23H01001 and (A) JP23H00505. We also acknowledge assistance from Anthropic (Claude 4) and AI tools used for generating Figs. 2, 3; all content was reviewed by the authors.

## References


Alves, R. A.; Olive, T.; and Castro, S. L. 2008. Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill. *International Journal of Psychology*, 43(6): 969–979.

Andersen, S. L.; Sweigart, B.; Glynn, N. W.; et al. 2021. Digital Technology Differentiates Graphomotor and Information Processing Speed Patterns of Behavior. *Journal of Alzheimer's Disease*, 82(1): 17–32.

Baral, S.; Lucy, L.; Knight, R.; Ng, A.; Soldaini, L.; Heffernan, N.; and Lo, K. 2025. DrawEduMath: Evaluating Vision Language Models with Expert-Annotated Students' Hand-Drawn Math Images. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Barz, M.; Altmeyer, K.; Malone, S.; Lauer, L.; and Sonntag, D. 2020. Digital Pen Features Predict Task Difficulty and User Performance of Cognitive Tests. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, 23–32. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368612.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Chi, M.; de Leeuw, N.; Chiu, M.; and Lavancher, C. 1994. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3): 439–477.

Chi, M. T. H. 2000. Self-explaining: The dual processes of generating inference and repairing mental models. In *Advances in Instructional Psychology: Educational Design and Cognitive Science*, volume 5, 161–238. Erlbaum.

Durkin, K. 2011. The Self-Explanation Effect when Learning Mathematics: A Meta-Analysis. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, DC.

Flanagan, B.; and Ogata, H. 2018. Learning analytics platform in higher education in Japan. *Knowledge Management & E-Learning*, 10(4): 469–484.

Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103: 102274.

Kishi, K.; and Miura, M. 2018. Detecting Learners' Weak Points Utilizing Time Intervals of Pen Strokes. *International Journal of Learning Technologies and Learning Environments*, 1(1): 61–77.

Kocon, J.; Cichecki, I.; Kaszyca, O.; Kochanek, M.; Szydlo, D.; Baran, J.; Bielaniewicz, J.; Gruza, M.; Janz, A.; Kanclerz, K.; et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99: 101861.

Limpo, T.; and Alves, R. A. 2018. Effects of planning strategies on writing dynamics and final texts. *Acta Psychologica*, 188: 97–109.

Nakamoto, R.; Flanagan, B.; Dai, Y.; Takami, K.; and Ogata, H. 2024. Unsupervised techniques for generating a standard sample self-explanation answer with knowledge components in a math quiz. *Res. Pract. Technol. Enhanc. Learn.*, 19: 016.

Nakamoto, R.; Flanagan, B.; Dai, Y.; Yamauchi, T.; Takami, K.; and Ogata, H. 2025. Integrating self-explanation and operational data for impasse detection in mathematical learning. *Research and Practice in Technology Enhanced Learning*, 20: 019.

Ogata, H.; Majumdar, R.; Yang, S. J. H.; and Warriem, J. M. 2022. Learning and Evidence Analytics Framework (LEAF): Research and Practice in International Collaboration. *Information and Technology in Education and Learning*, 2(1): Inv–p001–Inv–p001.

Oviatt, S.; Lin, J.; and Sriramulu, A. 2018. Dynamic handwriting signal features predict domain expertise. *ACM Transactions on Interactive Intelligent Systems*, 8(3): 1–26.

Rittle-Johnson, B.; Loehr, A.; and Durkin, K. 2017. Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM*, 49(4): 599–611.

Van der Weel, F.; and Van der Meer, A. 2024. Handwriting but not typewriting leads to widespread brain connectivity: a high-density EEG study with implications for the classroom. *Frontiers in Psychology*, 14: 1219945.

VanLehn, K.; Siler, S.; Murray, C.; Yamauchi, T.; and Baggett, W. 2003. Why Do Only Some Events Cause Learning During Human Tutoring? *Cognition and Instruction*, 21(3): 209–249.