

Scaling Human-AI Collaboration: Translating Multi-Source Assessment Data into Formative Learner Profiles

Hongwen Guo * Matthew S. Johnson Luis Saldivia Michelle Worthington
Jeremy Lee and Kadriye Ercikan ¹

¹ETS Research Institute, 660 Rosedale Rd, Princeton, NJ 08541

Abstract

We present a scalable dual-agent architecture translating multi-source assessment data – integrating performance with process logs – into formative data insights. Decoupling classification from text generation, embedding expert rubrics, and optimizing latency enables rapid first-draft generation. These insights reveal underlying learning behaviors, facilitating targeted intervention without increasing teachers' cognitive burden.

Keywords: Multi-source data, Dual-agent framework, Human-centered AI, Scalable feedback system

1 Introduction

As a key feedback mechanism, assessments play a fundamental role in the education system (Shute, 2008). With the paradigm shift of AI technologies from being mere tools for efficiency (e.g., automatic grading or predicting at-risk students) to becoming active collaborators in the classroom (Carnegie Learning, 2025), it is increasingly critical that AI-generated feedback is data-driven, contextually sound, and strictly free of hallucinations (Google DeepMind, 2024; Sahlman et al., 2025).

Training AI systems to effectively assist teaching requires human-in-the-loop protocols alongside large-volume, high-quality assessment data (Khalil et al., 2025; Hoffmann et al., 2022; Holstein and Alevan, 2022). Digitally-based large-scale assessments (LSAs), such as the National Assessment of Educational Progress (NAEP), uniquely satisfy this need by offering the rich empirical logs necessary to ground AI agents in psychometric reality. Specifically, integrating traditional response data with process data extracted from click streams—referred to collectively in the study as *multi-source data*—has been shown to reveal students' underlying problem-solving strategies

(Kroehne and Goldhammer, 2018; Winne, 2020), validate performance (Ercikan et al., 2023), provide evidence of engagement that illuminates holistic learning states (Guo et al., 2024), and evaluate how digital tools functioned (Guo et al., 2024b).

To address the analytical and scalability challenges identified in prior exploratory research, the current study investigates an optimized, dual-agent system built with teachers and for teachers. Specifically, we investigate the following research questions. *RQ-1*: How do architectural optimizations, specifically the implementation of Selective Few-Shot Learning (SFSL), reduce inference latency to enable the scaling of this formative feedback generation from multi-source assessment data? *RQ-2*: To what extent does a decoupled, dual-agent architecture generate highly accurate, formative learner narratives without introducing hallucinations? *RQ-3*: How do AI-generated narratives and visualizations of multi-source data function as an effective "first draft" to help educators uncover students' underlying problem-solving behaviors—such as time management, regulated navigation behavior, and productive/unproductive struggle—enabling them to better understand students' need to facilitate targeted interventions?

To address these questions, we utilized NAEP multi-source data to develop a novel dual-agent framework designed to assist educators in interpreting complex assessment logs. This architecture decouples cognitive tasks: a Classification Agent evaluates empirical student data using rigorously defined features, while a Teaching Agent synthesizes these findings into growth-oriented, formative feedback. To maintain both technical rigor and educational relevance, we leveraged the In-Context Learning (ICL) paradigm (Dong et al., 2024), embedding expert-validated rubrics and examples directly into the system workflow. Furthermore, architectural optimizations—specifically Selective Few-Shot Learning (SFSL) and natural

*hguo@ets.org; Corresponding author

language feature engineering—were implemented to enhance scalability and drastically reduce inference latency, ensuring the reliable processing of large-scale datasets.

Ultimately, this framework establishes a scalable blueprint for transforming raw assessment data into actionable formative assets. By employing SFSL and natural language feature translation, technical profiling reveals an 18-fold reduction in execution time. This efficiency gain contributes to moving AI from a theoretical possibility to a practical, deployable reality for large-scale assessment systems.

The subsequent sections detail the background literature, system architecture, and evaluation methodology, followed by a discussion of the framework’s implications.

2 Related Work

To contextualize the development of the dual-agent architecture, it is necessary to trace the evolution of how multi-source assessment data has been processed and utilized for educational feedback.

Multi-Source Data and Learner Profiles. Prior research has increasingly relied on human-centered AI (HAI) approaches to discover student engagement profiles from multi-source data. By integrating performance scores with behavioral process data collected from LSAs (e.g., using pseudonymized, restricted-use NAEP data), researchers can contextualize a student’s performance beyond a unidimensional score: not only we know what scores student received on a test, but also we can “see” how they engaged with and solved the problems on the test. Earlier HAI frameworks utilized human expertise to engineer explainable features, subsequently applying clustering and active machine learning techniques to identify typical engagement patterns (Guo et al., 2024).

LLM Agents and Scalability Bottlenecks. To bridge the gap between statistical clustering and actionable classroom feedback, a recent exploratory study tested multi-agent AI systems on a small, controlled sample of validated student profiles (Guo et al., 2025). This early work confirmed that agentic AI could successfully enrich feedback by generating individualized profiles, identifying strengths, and suggesting intervention strategies.

However, transitioning from traditional machine learning to generative Large Language Models (LLMs) in assessment contexts exposed critical bottlenecks. 1. *Formative Grounding:* Ensuring AI narratives remain strictly grounded in empirical

Data Type	
Item score	0,1, 1, 0, 2, 0, ...
Item response time	41, 112, 150, 292, ...
Item visit sequence	1,2,3,4, 3,4, 5, 6, 7, 14, ...

Table 1: Multi-source data for one student on 15-item block with a 30-minute time limit.

classroom practice rather than generating plausible but hallucinated pedagogical advice. 2. *Inference Latency:* There are concerns associated with the prohibitive computational cost and processing time, when feeding massive prompt payloads (including few-shot examples and numerical feature mappings) into the model context window.

The current paper departs significantly from the above prior exploratory efforts by directly addressing these critical bottlenecks. Rather than focusing purely on algorithmic exploration, this work operationalizes AI models into a deployable, scalable infrastructure designed with educators in the loop, transforming theoretical LLM capabilities into a high-fidelity pipeline capable of generating safe and formative feedback at scale.

3 Methodology

To realize the scalable, educator-in-the-loop infrastructure proposed above, we utilized a multi-phase, human-centered AI research design. The methodology traces the translation of raw multi-source data through an initial teacher co-design pilot, subsequent architectural optimizations, and a final large-scale system evaluation.

3.1 Data Source and Explainable Features

NAEP multi-source data used in this study are the NAEP restricted-use data collected from 2017 Grade Eight Math assessment (NCES, 2021), which include, for each student, item response scores, item response times (in seconds), and item visit sequences (refer to Table 1) on a released item block.

A dozen explainable features (refer to Table 2) were created from the multi-source data, which were also visualized in the navigation and position plots (refer to Figure 1 for details). These explainable features enable us to communicate with educators who participated in the study, as well as AI agents.

3.2 Phase I: Teacher Training Module and Co-Design Pilot

NAEP is well-known as the national report. However, participation in these NAEP assessments is

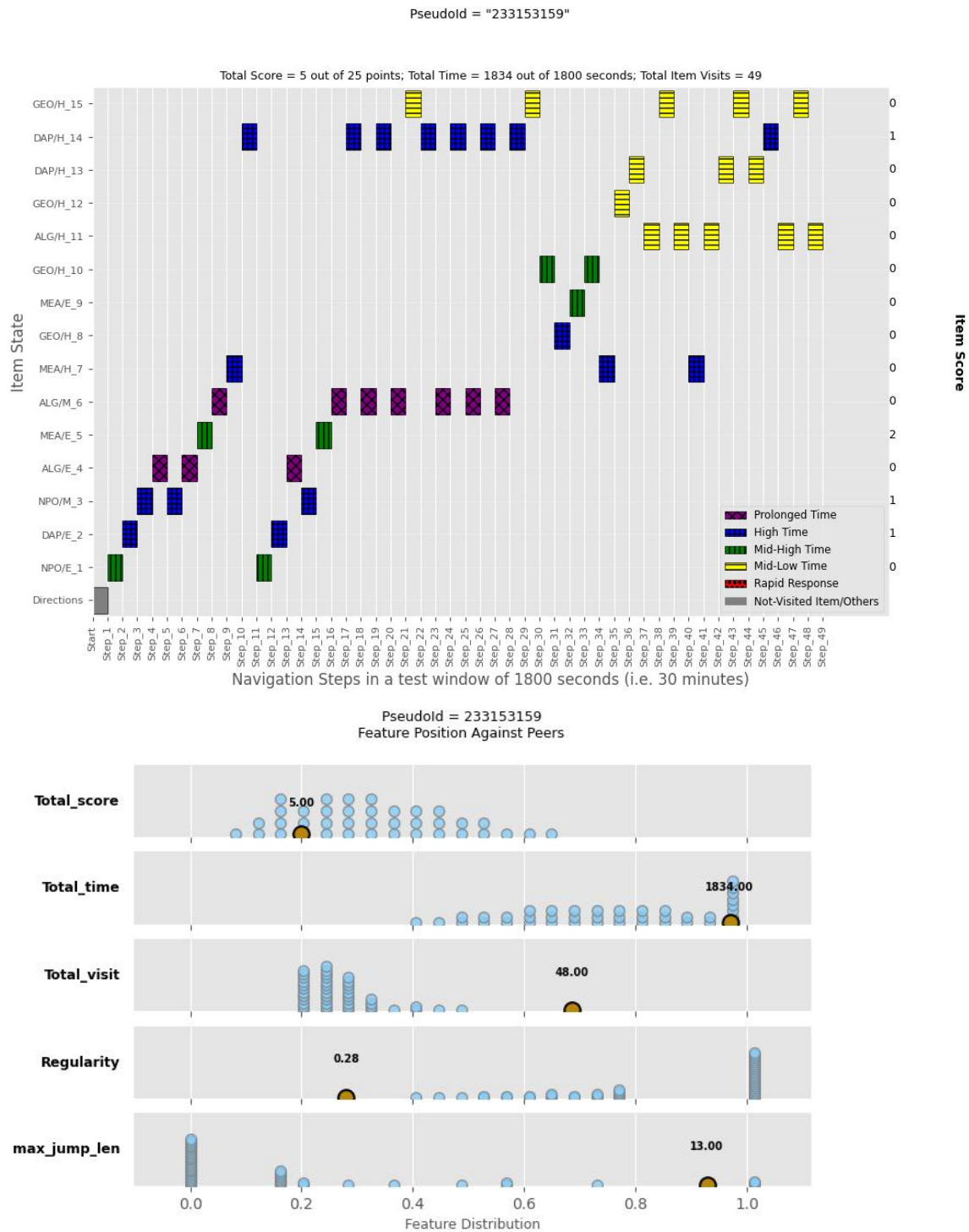


Figure 1: Visualization of one student's data. The navigation plot (upper panel) traces how a student worked through the math problems on the studied item block (15 items within a 30-minute time limit). The text above the plot shows this student's overall score (5 out of 25), time spent (1815 over 1800 seconds), and total number of item visits (49). The x-axis shows the step sequence the student took from one item to another. The Y-axis on the right shows the item sequence (1 - 15), associated content area (algebra-ALG, data analysis and probability and statistics-DAP, geometry-GEO, measurement-MEA, and number property and operation-NPO)s, and item difficulties (Easy, Medium hard, or Hard). In the legend, mid-low time, mid-high time, and high time are defined as: between RR and 50%tile, 50%tile to 75%tile, and 75% to PL, respectively. The position plot (lower panel) reflects where the student's data features are positioned compared to their peers. This plot contains a few key features, standardized within [0,1]. The background distribution depicts that of the reference group, where each circle stands for 50 students. Any value less than 50 are not shown. Because of the explainability of the features, the position plot provides contexture information to understand a student's performance and behaviors with their peers as a reference.

Name	Description	Interpretation
Total score	Sum of item scores	Value range (discrete): [0, 25]. High value: good performance; low value: low performance. Most important feature, affecting interpretation of others.
Total time	Sum of item response times	Value range (continuous): (0, 1800]. Low value: less engaged; high value: issues in time management. Important feature, affecting interpretation of others
Total visit	Sum of item visits	Value range (discrete) [1, 73]. A student visiting all items just once has a value of 14. Low value: few item visits/less engaged; high value: issues in behavior regulation. Important feature, affecting interpretation of others
Not-reached (NR)	Number of not-reached items (no response time)	Value range (discrete): [0, 14]. High value: worked on few items; low value: worked on many items. Speeded: if non-zero NR & high time.
Rapid response (RR)	Sum of rapid-responded items	Value range (discrete): [0, 14]. Low value: more engaged; high value: issues with effort. RR: likely not spent adequate time to understand/work the item and associated with low effort and guessing an answer on a multiple-choice question.
Prolonged time(PL)	Sum of items with prolonged times (over 95 percentile)	Value range (discrete): [0, 6]. High value: likely struggling on a high number of items. Non-zero value may indicate struggling, mostly due to lack of knowledge and skills to solve the problem(s), and subsequently led to NR items (i.e., test was speeded).
Navigation Regularity (Reg)	A measure to show whether a student mostly followed the order of item presentation on the test	Value range (continuous): [0.23, 1]. High value: orderly navigation through items (value of 1 indicates always moving forward or backward one item a time; no skipping around); low value: irregularly navigated through items. Also refer to item visit and jump related features for context.
Max jump length	Largest jump from one item to another	Value range (discrete): [0, 14]. High value may indicate global review; low value local review (moving forward or backward one item is counted as no jump with a value of zero).
Max jump location	Indicator to show from which item the max jump occurred.	Value range (discrete): [0, 15]. High value may indicate global review; low value pre-review.

Table 2: Main features created and their interpretation. Revised and reproduced from (Guo et al., 2025) with the studied data. Due to space limitations, detailed definitions for Reg and RR, for example, can be found in Guo et al. (2025) and references therein.

limited to a representative sample, meaning most teachers and students may never encounter the assessments, let alone have the opportunity to engage with NAEP’s granular process data (NCES, 2021).

The initial pilot involved a focused cohort of nine educators with extensive (20-years or more) teaching experiences. To assist participants to understand and interpret the multi-source data, we conducted extensive training sessions to introduce NAEP math assessments and student data, explain the meaning of the features, and illustrate how to read the plots. Participants were also instructed on how the data may provide insights to contextualize students’ performance. For example, Figure 1 shows that the student got a total score of 5 out of 25 points, used maximum testing time, and visited the items 49 times. Combined with the granular information, such as spending prolonged time on Item 4 (with three visits) and Item 5 (with seven visits), and frequent jumping between Item 6 and Item 14, the data may indicate the student lacked the knowledge and skills necessary to solve the problems, had issues with regulating navigation be-

haviors and time management, but they persistently tried hard on the assessment.

A dedicated teacher module (refer to Figure 2) was developed to assist participants. On the module, educators participated in three activities. In *Activity 1 – Student Experience and How students Performed*, teachers experience the NAEP Math assessment themselves to know digitally-based NAEP platform and understand under what circumstances the multi-source data were collected. They also have opportunities to explore how students performed on those items.

A crew of agents (three teachers, one professional coach, and a researcher; similar to that used in (Guo et al., 2025)) was experimented to create structured narratives (student profiles, strengths, areas for growth, and possible interventions). In *Activity 2 and Activity 3 – Evaluate AI Suggested Narratives*, student data and AI narratives are presented for teachers to elicit feedback, so that data visualizations and AI narratives can be improved for Phase II.

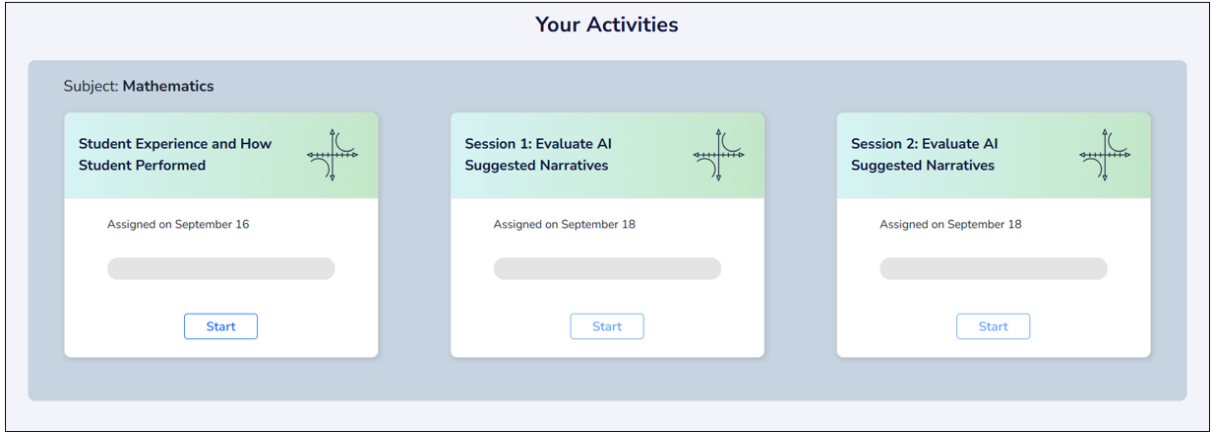


Figure 2: The teacher module.

3.3 Phase II: Model Iteration and Architectural Refinement

The experiment and formative feedback elicited from the Phase I pilot cohort revealed major gaps in the initial system design. First, we noticed that the average inference time was unacceptably long, taking three to five minutes per student, mainly due to the crew of five agents. Second, the early classification logic was deemed too generic to facilitate targeted interventions. Third, while educators found the visualizations of the multi-source data highly valuable, they noted that the AI crew generating narratives frequently suffered from task confusion—blending raw statistical classification with formative feedback, which occasionally led to hallucination without data evidence.

Dual-Agent Architecture. To address these gaps, we fundamentally restructured the system into a decoupled architecture, explicitly separating cognitive workloads between a Classification Agent (Dr. Metric) and a Teaching Agent (Dr. Coach). This decision stems from the established limitations of monolithic LLMs, which frequently suffer from task interference and elevated hallucination rates when simultaneously handling rigorous analytical classification and generative tasks (Ji et al., 2023). In educational contexts, attempting to process complex empirical logs and synthesize formative feedback within a single prompt often yields ungrounded recommendations (Macina et al., 2023). Because separating cognitive responsibilities is a standard method for enforcing model factuality (Du et al., 2023), this dual-agent design is an architectural prerequisite for systemic safety.

This dual-agent system strategically decouples cognitive responsibilities. Dr. Metric analyzes stu-

dent multi-source data and suggests a primary and a secondary profile. Dr. Coach synthesizes data and Dr. Metric’s recommendation into growth-oriented formative feedback in two sets of narratives (i.e., student strength, areas for growth, and potential intervention based on learning science). The two sets of narratives, as well as the opportunities to enter text, provide more flexibility for teachers to provide feedback based on their extensive knowledge and teaching experiences (refer to Figure 4 for one student example).

Engineering optimization. To eliminate the hallucinations flagged in Phase I, we embedded the expert language based on the Phase I participants’ feedback into profiling rubrics and a few-shot examples (consisting of data, profiles, and narratives) for each profile. The structured rubrics helped Dr. Metric to recommend two profiles aligned with data evidence; these two recommended profiles, in turn, enforced guardrails via Selective Few-Shot Learning (SFSL), to reduce latency and for Dr. Coach to ground narratives in teaching practice.

Figure 3 details the design of the dual-agent system with SFSL in three stages. *Stage I - Initialization.* The system prepares the environment by loading the necessary resources. This includes the Profile Rubric (guidelines for classification improved upon Phase I study (refer to Table 4); Few-shot Examples (demonstrations/context created from Phase I study). It also instantiates the two AI agents. *Stage II: Agentic Reasoning.* This is the core processing stage where a student’s multi-source data (input CSV) is ingested. Agent-1 performs a zero-shot classification based on profile rubrics and suggests primary and secondary student profiles. The suggested two profiles guide Agent-2

Stage I: Initialization		Stage II: Agentic Reasoning		Stage III: Final Results	
	Profile Rubric		Student Data (Input)	\$	Calculate Cost & Latency
	Few-shot Examples	Analyst Agent:	Classification (Zero-shot)		State Change: add strength, growth, and intervention
		Fetch	Primary Profile		Write to CSV (line-by-line)
Agent-1	Analyst (Dr. Metric)		Secondary Profile		
Agent-2	Teacher (Dr. Coach)	Teacher Agent:	Dynamic (few-shot)		Error Log

Figure 3: The dual-agent system

to select the two sets of relevant few-shot examples to generate data-driven AI narratives. *Stage III: Final Results.* The system finalizes the output and tracks performance. It calculates operational metrics (such as running time) and records state change (adding details on profile, strength, growth, and intervention). Finally, the AI narratives are written to a CSV file line-by-line for each student, and any issues are captured in an Error Log.

3.4 Phase III: Large-Scale System Evaluation

To rigorously validate the finalized dual-agent infrastructure and answer our core research questions, Phase III transitioned from iterative co-design to a large-scale deployment simulation. The evaluation was structured across two distinct dimensions: teaching practice alignment (assessing system efficacy and safety) and operational scalability (assessing computational deployability).

Participants and Task Design. The expanded evaluation phase involving 47 educators and a corpus of over 2,000 students’ multi-source data. As in Phase I, educators participated in the training sessions, and then logged onto the teacher training module (refer to Figure 2) to participate in the three activities, presented with the dual-agent generated growth-oriented narratives and the accompanying multi-source visualizations (refer to Figure 4 for the student example); this helps to simulate a real-world deployment scenario. They also took an exit survey to provide general feedback to the study, particularly on how the AI-generated narratives influenced their understanding of the students.

Data Collection and Metrics: Operational Scalability. Because the primary bottleneck in deploying agentic architectures for large-scale assessments is processing speed and compute cost, data collection focused on deterministic computational metrics rather than stochastic model variance to answer

RQ-1 (Latency & Scalability).

We systematically logged the system execution time across the pipeline’s optimization iterations when applying engineering techniques against a baseline model (T-0) that uses numerical features, feature mapping to natural language, and full few-shot examples in the dual-agent framework (refer to the first two columns in Table 3. These engineering techniques are intended to reduce latency and added to the baseline model one by one. *T-1:* The selective few-shot learning, instead of injecting the full few-shot examples (Dong et al., 2024). Dr. Coach only reads those two sets of examples associated with the two profiles recommend by Dr. Metric. *T-2:* Use natural language as input, instead of the numerical features with mapping. That is, we prepared the numerical features into natural language as input, so that the pipeline eliminates the feature mapping data. *T-3:* Application-level cache, which reduces the redundancy of generating narratives for the same data patterns.

Data Collection and Metrics: Teaching Practice Alignment. To directly address RQ-2 (Formative Validity & Safety) and RQ-3 (Human-AI Augmentation), data collection focused on both quantitative human-AI agreement and qualitative user feedback to measure the system’s utility as a collaborative tool.

Each educator independently reviewed AI generated narratives. For each student, they first evaluated Dr. Metric’s profile classification against the expert-validated rubrics; and then they evaluated Dr. Coach’s suggested narratives on the student’s strength, growth, and possible intervention using their own teaching expertise and practical knowledge. Human-AI Agreement rates on the feedback quantify systemic safety and the elimination of hallucinations. Qualitative data analysis on the exit survey is used to evaluate whether

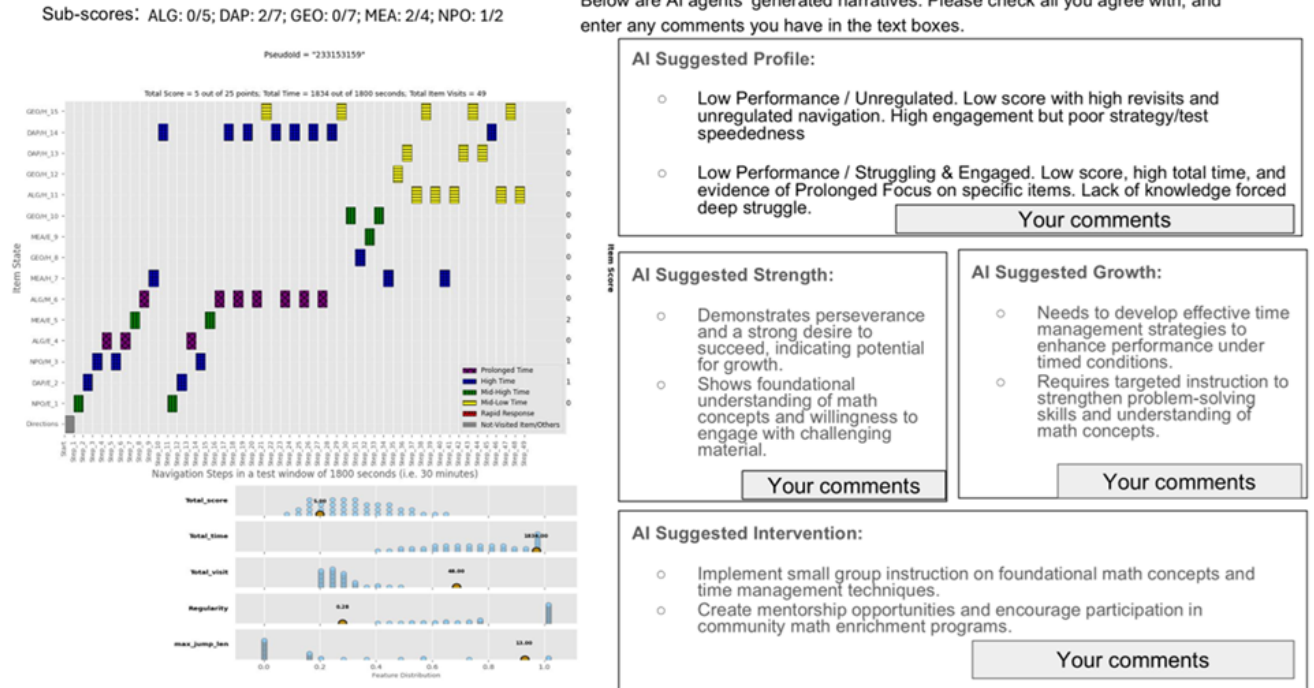


Figure 4: Data and visualization (left) and AI narratives (right) for one student. On the screen, AI suggested narratives for this student are presented (two profiles, two strength, two growth, and two interventions) for teachers to evaluate. In addition, space is dedicated for teachers to enter comments. For this student, AI-suggested profile highlighted low performance, high time spent on the test, but also acknowledged their strength in high engagement, persistent effort, and different strategies in solving problems.

these "first drafts" successfully revealed underlying problem-solving behaviors—such as time management, persistence, or rapid guessing—and whether this multi-source perspective would help them to design targeted pedagogical interventions without increasing their overall cognitive burden.

4 Results

4.1 Operational Scalability

Table 3 details the execution time comparisons across three optimization iterations (T-1 to T-3). The baseline dual-agent system (T-0) exposed severe In-Context Learning (ICL) token bottlenecks. Processing a massive prompt payload—which included exhaustive few-shot examples, expert rubrics, and the complex mapping of numerical features into natural language—resulted in prolonged latency (averaging 90 seconds to generate formative narratives per student) and frequent API rate-limit errors.

Implementation of the Selective Few-Shot Learning (T-1) shows that execution time dropped to 60 seconds (a 1.5x speed increase), mainly due to reduction of injected few-shot examples. Further

leverage of the explainability of the features via converting students' multi-source data directly into descriptive text prior to inference (T-2), it drastically reduced the generation time to 15 seconds (an additional 4x speed increase), due to the reduction of the heavy feature mapping. Finally, implementing application-level caching via LangChain (T-3) eliminated redundant queries for identical profile combinations, lowering the average processing time to just 5 seconds (an additional 3x speed increase).

Working in concert, these three engineering optimizations achieved an 18-fold reduction in overall executing time and inference latency. The broader trade-offs of these techniques—evaluating their impact on narrative richness, formative accuracy, consistency, data privacy, and latency—are synthesized in Table 3 in details.

4.2 Formative Validity

To assess system safety and formative validity, we quantified the alignment between the AI-generated outputs and human expert evaluations. For each participating educator, a composite human-AI

Table 3: Comparison of Execution Time by Engineering Optimization

	Techniques	Pros	Cons	Latency
Base (T-0)	Numerical features, mapping, full demo	Nuanced info, individualized feedback	High latency, risk of rate limits	90s (1x)
T-1	Selective few-shot demonstration	Focused dynamic context, faster	Fewer examples exposed, risk of less accurate interpretation	60s (1.5x)
T-2	Natural language features, no mapping	Privacy-friendly, fast compute	Less nuance, risk of less accurate interpretation	15s (6x)
T-3	Application-level cache	Consistent narratives, ultra-fast	Reduced language variation	5s (18x)

agreement score (Yes_Rate_Total) was computed by calculating the proportion of affirmative response ("Yes") across the two sets of narrative dimensions (i.e., profile, strength, growth, intervention). Figure 5 displays these agreement rates, categorized by the primary set (left panel) and secondary set (right panel) generated by dual agents. The results demonstrate a robust baseline consensus, with agreement rates stabilizing at approximately 75% across the majority of the profiles. This level of expert validation confirms that the decoupled architecture effectively mitigates hallucinations, keeping the formative feedback grounded in the empirical data. Notably, minor deviations from this 75% average occurred within the "Unengaged/Minimal Effort" (right panel) and "Very Low Performance / Low Engagement" (left panel) profiles; however, these variances may be quantitatively attributable to the limited sample sizes within these specific classifications ($n < 5$ and $n < 45$, respectively) and contextually attribute to the challenges in identifying strength when students did not engage.

4.3 Human-AI Augmentation

Evaluating the system's true potential for human-AI collaboration requires understanding its perceived utility from the educator's perspective. Drawing on data from the post-evaluation exit surveys, we assessed whether the participating educators felt these "first drafts" successfully revealed underlying problem-solving behaviors and reduced their analytical burden.

The exit survey data indicates that educators strongly valued the multi-source data that integrate performance with processes. Rather than relying solely on a unidimensional performance score, teachers reported that the multi-source data visualizations were informative (which 80% of participants rated as "good" or "excellent") and the

Teaching Agent's narrative helped them conceptualize the students' underlying behaviors (with 60% rating this generative capability as "good" or "excellent"). The qualitative feedback underscored these values: several educators noted that the system helps "differentiate students' needs" and provides "an idea for moving forward." One teacher particularly appreciated the strength-based narratives for lower-performing students, commenting that "students would love to know that their effort was acknowledged." Another educator emphatically stated that these data insights "should be the starting point for any intervention."

Crucially, the system was not designed to replace human judgment, but to augment it. Survey responses indicated a high perceived readiness to translate these narratives into action. Educators reported that having the raw data synthesized into a structured "first draft" significantly lowered the cognitive barrier to interpreting complex click stream logs. By offloading the heavy computational task of feature mapping and initial data synthesis to the dual-agent architecture, educators found the system "easy to navigate" and noted that it provided "productive feedback" that "could save time." In other words, such a system enables educators to more rapidly extract data formative insights to guide targeted interventions, confirming the framework's value as a collaborative asset.

5 Discussion and Conclusion

Phase III evaluation involving about fifty educators and over 2,000 student profiles confirmed that the proposed dual-agent architecture can effectively translate complex, multi-source assessment data into safe and actionable formative narratives. By explicitly decoupling empirical classification from narrative generation and introducing targeted architectural optimizations, this work demonstrates how AI can move beyond general exploratory research

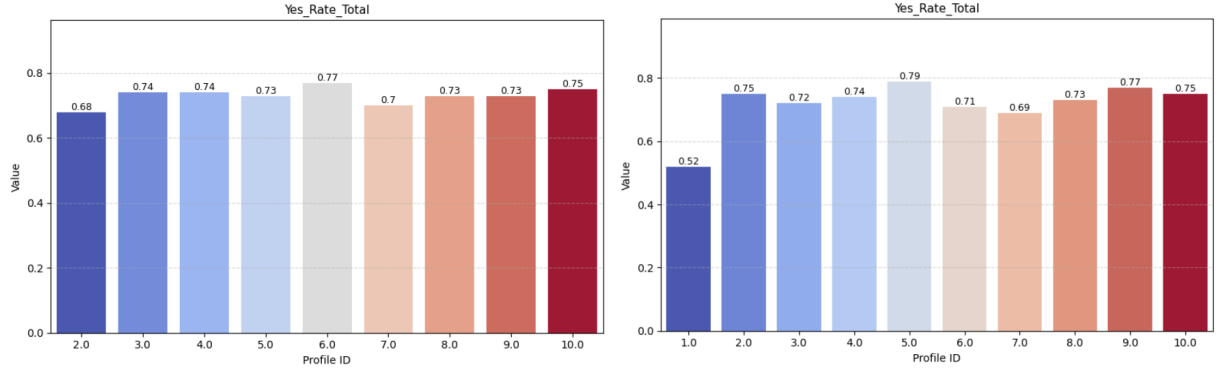


Figure 5: The overall human-AI agreement rates by the primary profile categories (left panel) or the secondary profile categories (right panel) by the dual-agents.

Table 4: Student Holistic Profiles and Corresponding Numeric Label

Profile ID	Behavior Profile Description
1	Unengaged/Minimal Effort. Attempted few items or took minimal time
2	Very Low Performance / Low Engagement. Very low score despite reaching most items. Low effort (low time/regular visit)
3	Low Performance / Basic Engagement. Low score, but completed most items with a consistent, low time commitment. Low effort level
4	Low Performance / Persistent Effort. Low score, but used full/regular time. Engaged and navigated most items with mixed, persistent strategies
5	Low Performance / Unregulated. Low score with high revisits and unregulated navigation. High engagement but poor strategy/test speededness
6	Low Performance / Struggling & Engaged. Low score, high total time, and evidence of Prolonged Focus on specific items. Lack of knowledge forced deep struggle
7	Medium Performance / Balanced. Consistent medium score, time, and visit behavior. Meets expected performance and strategic norms
8	Medium-High Performance / Strategic Effort. Medium-to-high score. Good time use with some Prolonged Focus (strategic engagement/struggle)
9	High Performance / Expected Strategies. High score with balanced time and visit behavior. Consistent, efficient performance
10	Very High Performance / Mastery. Highest score. Expected time uses and navigation. Represents the highest level of efficiency and content mastery

into a robust, deployable infrastructure to assist educators.

By answering the three core research questions, this study contributes to the advancement of human-AI synergy in educational assessment in three key areas: *A. Engineering for Scalability*. While many agentic systems remain theoretical due to high computational costs and inference latency, this study contributes a validated blueprint for scalable deployment. By integrating the selective few-shot learning technique and natural language feature translation, we achieved an 18-fold reduction in inference latency. This technical innovation breaks the primary bottleneck of large-scale assessments, paving the way to transform agentic systems from slow, offline tools into responsive computational "teammates", and to potentially enhance the use of assessment data. *B. Operationalizing Human*

Knowledge through Agentic Design. The proposed dual-agent architecture is explicitly grounded in human expertise. By leveraging formative findings from our pilot study, we constructed expert-validated profile rubrics and a repository of targeted demonstration narratives (few-shot examples). Constraining the teaching agent with these human-designed guardrails can effectively mitigate AI hallucinations, proving that generative models are most reliable—and most educationally valid—when strict empirical indicators dictate their outputs. *C. From Data Processing to Teacher Partnership*. Ultimately, this work contributes to advances of the assessment field by redefining the role of AI in processing multi-source data. We show that by offloading the heavy cognitive burden of interpreting click stream logs the AI, educators can rapidly access the "invisible" underlying problem-

solving behaviors of their students. Providing this structured "first draft" of formative insights empowers teachers to focus on their unique, irreplaceable strengths: relationship-building and designing targeted, empathetic interventions.

However, given the scope of the study, several key limitations remain. First, although the decoupled system significantly reduces hallucinations via human guardrails, the feedback generated remains synthetic—it is based on retrieving and matching data patterns rather than possessing a genuine cognitive understanding of the student's holistic context. Second, while the evaluation utilized offline and historical NAEP data to validate the architecture, this approach does not sufficiently account for live deployment challenges, specifically concerning real-time data streaming infrastructure, data privacy, and ethical compliance in active classrooms. Third, strict data privacy regulations and the scarcity of high-quality, multi-source classroom datasets necessitated the use of NAEP data to validate this architecture. However, because NAEP is fundamentally a summative, large-scale assessment, translating its empirical logs into formative classroom insights introduces an unavoidable inferential gap. While our framework demonstrates the mechanical feasibility of this translation, there remains an inherent construct mismatch between the macro-level policy design of NAEP and the micro-level pedagogical needs of formative assessments.

Future work would focus on transitioning from offline simulation to live classroom environments. For example, the system can be deployed as a 'wellness check' for students to work through the NAEP assessment, and for teachers to view their students' performance and processes and prepare a formative feedback based on AI narratives. When the system is deployed in a human-in-the-loop pilot study in classrooms, we could measure how teachers modify the AI-generated drafts before sharing them or acting upon them. This textual "edit distance" will serve as a quantitative metric for trust and practical utility. Furthermore, through the accumulation of this teacher-edited feedback, future work could compare the current dual-agent in-context learning performance against a fine-tuning approach, which theoretically represents the highest form of efficiency for inference.

Finally, while the immediate application of this framework is to generate formative feedback, its potential as a professional development tool is equally significant. By exposing pre-service teacher and

educators with limited resource to these rich multi-source data and AI-generated "first drafts," the system can function as an educational simulator to train educators to recognize complex engagement patterns they might otherwise miss, effectively preparing them for the next generation of digital assessment systems.

Acknowledgements

The work is supported, in part, by the Gates Foundation. Any opinions expressed here are those of the authors and do not necessarily reflect those of ETS or the Gates Foundation.

References

- Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, 78(1), 153–189.
- Carnegie Learning. (2025). *The State of AI in Education 2025*. Technical Report.
- Google DeepMind & Google Research. (2024). *Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach*. Technical Report.
- Sahlman, W. A., Ciechanover, A. M., & Grandjean, E. (2025). *Khanmigo: Revolutionizing Learning with GenAI*. Harvard Business School Case 824-059 (Revised).
- Khalil, M., Liu, Q., & Jovanovic, J. (2025). AI for data generation in education: Towards learning and teaching support at scale. *British Journal of Educational Technology*, 56(3), 993–998.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training Compute-Optimal Large Language Models. *Advances in Neural Information Processing Systems*, 35, 30016–30030.
- Holstein, K., & Aleven, V. (2022). Designing for Human-AI Complementarity in K-12 Classrooms. In *Proceedings of the 23rd International Conference on Artificial Intelligence in Education (AIED 2022)*, 142–155. Springer.
- Ercikan, K., Guo, H., & Por, H. H. (2023). Uses of process data in advancing the practice and science of technology-rich assessments. In K. Ercikan & J. W. Pellegrino (Eds.), *Innovating assessments to measure and support complex skills* (p. 211). Routledge.
- Guo, H., Johnson, M., Ercikan, K., Saldivia, L., & Worthington, M. (2024). Large-scale assessments for learning: A human-centred AI approach to contextualizing test performance. *Journal of Learning Analytics*, 11(2), 229–245.

- Guo, H., Johnson, M., Saldivia, L., Worthington, M., & Ercikan, K. (2024). Human-Centered AI for Discovering Student Engagement Profiles on Large-Scale Educational Assessments. *Journal of Measurement and Evaluation in Education and Psychology*, 15(1), 1–15.
- Guo, H., Johnson, M. S., Saldivia, L., Worthington, M., & Ercikan, K. (2025). Leveraging Multi-AI Agents for Teacher Co-Design. In *Proceedings of the 1st Artificial Intelligence in Measurement and Education Conference (AIME-Con)*, 25–34. Association for Computational Linguistics.
- Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., & Li, L. (2024). A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1107–1128. Association for Computational Linguistics.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). *Survey of hallucination in natural language generation*. *ACM Computing Surveys*, 55(12), 1–38.
- Macina, J., Kochmar, E., Bellet, C., & Ryan, R. (2023). *Opportunities and challenges in using large language models for educational feedback*. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 1–12).
- Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). *Improving factuality and reasoning in language models through multiagent debate*. *arXiv preprint arXiv:2305.14325*
- Kroehne, U., & Goldhammer, F. (2018). How to Conceptualize, Represent, and Summarize Indicator Variables Extracted from Computer-Based Assessment Log Data. *Frontiers in Psychology*, 9, 2297. <https://doi.org/10.3389/fpsyg.2018.02297>
- Winne, P. H. (2020). Construct and Consequential Validity for Learning Analytics Based on Trace Data. *Computers in Human Behavior*, 112, 106457. <https://doi.org/10.1016/j.chb.2020.106457>
- National Center for Education Statistics. (2021). *Process Data From the 2017 NAEP Grade 4 and Grade 8 Mathematics Assessment*. U.S. Department of Education, Institute of Education Sciences. Retrieved from https://www.nationsreportcard.gov/process_data/
- Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving Factuality and Reasoning in Language Models through Multi-agent Debate. *arXiv preprint arXiv:2305.14325*. <https://doi.org/10.48550/arXiv.2305.14325>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 1–38. <https://doi.org/10.1145/3571730>
- Macina, J., Dong, N., Huang, K., & Kapur, M. (2023). MathDial: A Dialogue Dataset for Mathematics Tutoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 3118–3134. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.176>
- Moore, S., Nguyen, H. A., Bier, N., & Stamper, J. (2024). Assessing the Quality of LLM-Generated Feedback for Open-Response Tasks. In *Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK '24)*, 1–10. Association for Computing Machinery. <https://doi.org/10.1145/3636555.3636892>
- Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023). AutoGen: Enabling Next-Gen LLM Applications via Multiple Agent Conversation. *arXiv preprint arXiv:2308.08155*. <https://doi.org/10.48550/arXiv.2308.08155>