OpenCD: Empowering Diagnosis of Children's Mathematical Cognition through Open-ended Multimodal Tasks

Zhi Zheng
Department of Computer Science and Technology, BNRist, Tsinghua University
Beijing, China
zhengz22@mails.tsinghua.edu.cn

Chun Yu*†
Department of Computer Science and Technology, BNRist, College of AI, Tsinghua University
Beijing, China
chunyu@tsinghua.edu.cn

Weihao Chen
Department of Computer Science and Technology, Tsinghua University
Beijing, China
chenwh20@mails.tsinghua.edu.cn

Minzheng Song
Department of Computer Science, Univ. of North Carolina at Chapel Hill
Chapel Hill, NC, USA
minzheng@unc.edu

Binglin Liu
Department of Computer Science and Technology, Tsinghua University
Beijing, China
lbl23@mails.tsinghua.edu.cn

Jianyang Liu
Department of Computer Science and Technology, Tsinghua University
Beijing, China
liujiany22@mails.tsinghua.edu.cn

Shiyi Wang
Academy of Arts & Design, Tsinghua University
Beijing, China
shiyi-wa23@mails.tsinghua.edu.cn

Xutong Wang
Department of Computer Science and Technology, Tsinghua University
Beijing, China
wangxuto23@mails.tsinghua.edu.cn

Jie Cai
Department of Computer Science and Technology, Tsinghua University
Beijing, China
jie-cai@mail.tsinghua.edu.cn

Yuanchun Shi
Department of Computer Science and Technology, BNRist, Tsinghua University
Beijing, China
Qinghai University
Xining, China
shiyc@tsinghua.edu.cn

Abstract

Assessing children's cognitive development in early mathematics is vital for effective teaching. Compared to closed-ended questions, which may fail to capture nuanced developmental spectrum, open-ended elicitation tasks (e.g., asking students to manipulate objects or draw to represent numbers) serve as a promising approach to reveal deeper cognitive processes. However, their diverse and unstructured nature makes systematic analysis challenging for teachers. We present *OpenCD*, a teacher-facing system that automatically analyzes multimodal student responses to capture individualized insights. Based on Evidence-Centered Design, it combines Vision-Language Models (VLMs) and expert models to generate interactive diagnostic graphs and reports with traceability back to behavioral evidence. In our two-part evaluation, a validation study found 90.3% of the system's diagnoses "completely reasonable," and a user study showed that *OpenCD* reduced teachers' analysis burden and enhanced their insights into student thinking. Our work contributes to scalable process-based assessment for mathematical literacy.

CCS Concepts

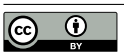
• **Applied computing** → **Computer-assisted instruction**; • **Human-centered computing** → **Information visualization**; • **Computing methodologies** → *Knowledge representation and reasoning*.

Keywords

Cognitive Diagnosis, Explainable AI, Formative Assessment, Mathematics Education, Vision-Language Model, Visual Analytics

ACM Reference Format:
Zhi Zheng, Chun Yu, Weihao Chen, Minzheng Song, Binglin Liu, Jianyang Liu, Shiyi Wang, Xutong Wang, Jie Cai, and Yuanchun Shi. 2026. OpenCD: Empowering Diagnosis of Children's Mathematical Cognition through Open-ended Multimodal Tasks. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 31 pages. https://doi.org/10.1145/3772318.3790318

*Corresponding author.
†Also with Key Laboratory of Pervasive Computing, Ministry of Education



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790318

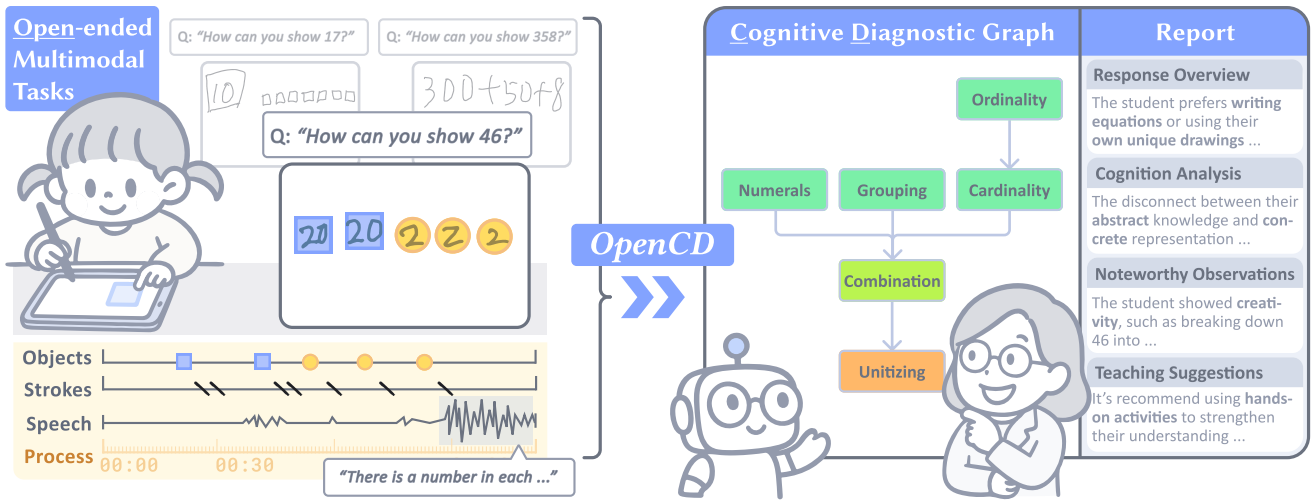


Figure 1: *OpenCD* analyzes children's multimodal temporal responses (e.g., objects, drawings, speech) to open-ended elicitation tasks (left), empowering teachers with an automatically generated interactive Cognitive Diagnostic Graph and a Comprehensive Report (right), revealing nuanced insights into each child's mathematical cognitive development.

1 Introduction

"How can you show 46?" Faced with this question, the 1st-grade student in Figure 1 pauses to think. After a moment, she drags a blue square onto the canvas and writes the number "20", and repeats. She then drags a yellow circle and writes "2", and repeats twice. When asked to explain, she says, "Each block has a number inside. That number represents how many sticks the block has." This vignette illustrates one of the diverse responses collected in our study.

In early mathematics education, it is crucial for children to ground their abstract understanding of concepts and operations in concrete experiences (e.g., learning about place value by bundling ten sticks into a single group) [4, 46]. Open-ended elicitation tasks—involving multimodal actions such as manipulating objects, drawing, writing equations, and verbalizing thoughts—serve as a powerful mechanism for teachers to observe these cognitive processes [77]. By analyzing these multimodal responses, teachers can look beyond correct answers to gauge students' developmental levels in concrete and abstract thinking, thereby enabling targeted scaffolding [80].

However, making sense of these rich assessments is prohibitively difficult in practice, requiring teachers to bridge the semantic gap between low-level behavioral signals and high-level pedagogical insights [16, 74]. The diversity and subtlety of open-ended responses demand time-consuming observation and expert-level analysis of the entire problem-solving process. Consequently, many schools revert to traditional computational tests that prioritize the accuracy of procedural skills over the depth of thinking. This narrow focus risks overlooking students' intrinsic conceptual understanding, potentially leaving foundational cognitive deficits unaddressed [76, 86]. While this dilemma calls for teacher-assistance technologies, "black-box" automation fails to foster the trust and agency required in such high-stakes contexts [35, 53]. Yet, research on how to effectively support teachers in making sense of such multimodal response process remains scarce.

To address this gap, we guide our research through three key questions:

- **RQ1:** What are the characteristics of children's multimodal open-ended responses, and what are teachers' specific sensemaking needs for such data?
- **RQ2:** How can we design a human-AI collaborative system that automates the diagnosis of unstructured process data while ensuring interpretability and teacher agency?
- **RQ3:** What is the perceived value of *OpenCD* in streamlining the analysis workflow and uncovering student thinking?

To answer these questions, we adopted a multi-phase research approach, moving from formative understanding to system design and rigorous evaluation, as presented in Figure 2.

We conducted a two-part formative study to inform the system design. An analysis of multimodal responses from 14 students, revealing the ambiguous characteristics of such data, informed our automated analysis pipeline. Subsequent interviews with five teachers clarified their sensemaking needs. We highlighted two design challenges: bridging the gap between the unstructured and unpredictable nature of student work and the structured requirements of assessment, while satisfying teachers' needs for both macro-level overview and micro-level transparency with evidence traceability.

Based on these insights, we present *OpenCD*, an AI system designed to empower teachers in *C*ognitive *D*iagnosis through *Open*-ended multimodal tasks. To ground the system with pedagogical knowledge, *OpenCD* adopted Evidence-Centered Design (ECD) [64], employing a hybrid architecture that synergizes the flexibility of VLMs with the reliability of rule-based expert models. This approach allows the system to interpret diverse, unstructured student responses while ensuring diagnostic consistency. For teachers, *OpenCD* visualizes individual learning progress through cognitive diagnostic graphs and aggregates class-wide data to reveal collective performance patterns. Crucially, the system ensures interpretability

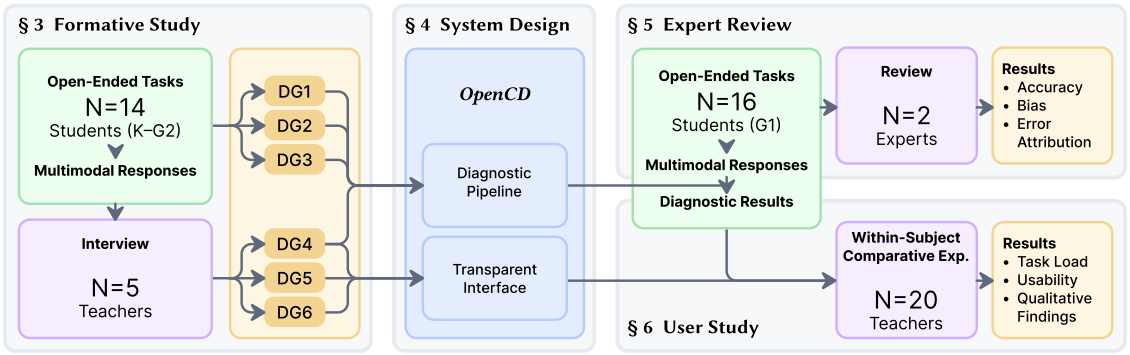


Figure 2: Overview of the research methodology and study roadmap. The project progressed through four phases corresponding to the paper sections: a formative study with students and teachers to define design goals (§ 3), the design and implementation of the *OpenCD* system (§ 4), an expert review to validate diagnostic accuracy (§ 5), and a user study to evaluate system utility with teachers (§ 6). The flowchart details the participants, data sources, and key outputs at each stage. Colors distinguish student- (Green) and teacher-related (Purple) studies, key findings (Yellow) and system design (Blue).

by enabling bi-directional traceability, allowing teachers to navigate seamlessly between high-level diagnostic conclusions and the specific behavioral evidence in student work.

We conducted a two-stage evaluation of our system. To quantitatively validate the system's diagnoses, two expert teachers reviewed the AI-generated diagnoses, based on a new collection of data from 16 first-grade students. The results showed that 90.3% of the diagnoses were deemed "completely reasonable". Secondly, to qualitatively evaluate the system's utility, we performed a within-subject user study with 20 primary school teachers. Interviews and survey data indicated that *OpenCD* significantly streamlined their analysis burden while deepening their insights into student thinking. Furthermore, the system's transparent presentation fostered appropriate trust, empowering teachers to verify AI judgments and identify potential discrepancies. We further discussed the design implications for explainable AI (XAI) in assessment and the tensions in teacher-AI collaboration.

In summary, this paper makes the following contributions:

- Empirical findings derived from a formative study that characterize students' multimodal open-ended responses and articulate teachers' specific sensemaking requirements for such unstructured data.
- The design and implementation of *OpenCD*, a human-AI collaborative system that automates the analysis of multimodal response processes and transparently presents diagnostic results with traceable evidence.
- A mixed-methods evaluation demonstrating the system's validity and its effectiveness in reducing teachers' analysis burden while enhancing pedagogical insights, and offering design implications for trustworthy XAI.

2 Related Work

2.1 Early Mathematics: Cognition, Teaching and Assessment

Mathematics learning is fundamentally a process of cognitive shaping. The theory of Embodied Cognition posits that mathematical thinking is not an isolated cerebral activity but is deeply rooted in sensorimotor experiences and physical interactions with the environment [5, 46, 68]. Consequently, effective learning relies on a multimodal interplay of actions, gestures, and language [67, 87]. For instance, Wu et al demonstrate that congruent physical movements can significantly reduce cognitive load and enhance estimation accuracy [103].

These theoretical insights align with Bruner's **Enactive-Iconic-Symbolic (EIS) model** [11, 12], and its practical application of the Concrete-Pictorial-Abstract (CPA) approach [49], which advocates for grounding abstract knowledge in concrete manipulations. Crucially, the ability to translate and connect these various modalities is known as Representational Fluency, a vital marker of deep conceptual understanding [18, 30, 50, 66]. Responding to calls for technologies that support such embodied learning [68], based on EIS model, *OpenCD* is designed to capture and analyze the full spectrum of student responses—from manipulating objects to drawing, writing, and verbal explanation.

Capturing such multimodal behaviors enables Formative Assessment—a process distinct from summative testing that aims to "close the gap" in learning through timely, diagnostic feedback [8, 82, 101]. Crucially, this requires looking beyond final answers to uncover students' underlying reasoning strategies and misconceptions [32]. However, such process-oriented diagnosis faces a scalability dilemma: traditional one-on-one observations [28] are prohibitively labor-intensive for daily practice. While emerging AI technologies offer potential solutions, current automated assessments largely focus on correctness in structured, closed-ended tasks [19], lacking the capability to interpret the nuanced demonstrations essential for identifying *why* a student struggles [91]. To bridge the gap, we employ **Evidence-Centered Design (ECD)** [64, 65] as our guiding framework. ECD provides a formal structure for designing assessments by explicitly linking student competencies, evidence, and tasks [90]. We translate these principles into *OpenCD* by defining the competency and evidence models as a computational knowledge base, thereby grounding the system's diagnostic inferences in pedagogical theory rather than black-box predictions.

2.2 AI System for Mathematics Education

HCI research has enriched mathematics education by exploring novel interaction modalities. By combining Extended Reality (XR) with tangible interactions, systems like ARMath [39] and MathBuilder [96] engage students in hands-on manipulation to support Enactive representations, while other works bridge the Iconic and Symbolic by visualizing complex notations [20, 26, 47] or translating sketches into equations [83]. Complementing these advancements in interaction modalities, empowering teachers with Generative AI (GenAI) has become a prominent research topic [69]. Current tools assist in various pedagogical aspects, ranging from lesson planning and resource generation [27, 78] to enhancing classroom management [41, 56, 89] and providing visual analytics [93]. However, these innovations primarily focus on pedagogy, leaving the assessment on students' diverse modes of representation underexplored.

To assess student mastery at scale, educational data mining often employs Cognitive Diagnosis Models (CDMs), ranging from foundational statistical models using expert-defined Q-matrices [23, 60, 94] to modern neural network-based approaches like NeuralCD [99] and knowledge graph-based systems [57, 100]. While efficient, these systems predominantly rely on dichotomous response data (correct/incorrect) from closed-ended questions, and consequently fail to capture the nuanced, unstructured cognitive processes essential for developing mathematical literacy.

This highlights a critical gap: the lack of scalable methods for analyzing the rich cognitive behaviors demonstrated in open-ended tasks. We aim to complement traditional assessments by conducting cognitive diagnosis based on students' process-oriented multimodal behaviors, thereby informing future pedagogy and Intelligent Tutoring System design.

2.3 Building Explainable AI for Educational Assessment

Analyzing multimodal behavioral data from assessment is fundamentally a sensemaking process [73, 74]. In HCI domains ranging from health to data science, researchers have developed visual analytics and human-in-the-loop systems to help experts bridge the semantic gap between low-level raw signals and high-level conceptual insights [45, 51]. In the educational context, this challenge is uniquely demanding: teachers must apply their pedagogical expertise to uncover the deep cognitive processes underlying observable behaviors [6]. This requires tools that go beyond simple data reduction to support a rigorous analysis of students' thinking [16].

However, in high-stakes educational contexts where teachers are ultimately accountable for decisions [35, 40], "black-box" automation is insufficient. HCI research advocates for XAI that fosters appropriate trust and collaboration [1, 59]. While general XAI methods like LIME and SHAP focus on post-hoc feature attribution [58, 79], such low-level metrics often fail to support deep causal inference. Effective explanations for teachers must instead align with domain knowledge and users' mental models [53, 98]. For teacher-facing systems, this means AI outputs must be interpretable in pedagogical terms, enabling teachers to verify algorithmic judgments against their expertise and maintain agency [29, 75].

The emergence of GenAI introduces new complexities. While GenAI excels at synthesizing intuitive explanation, its outputs are often "post-hoc rationalizations" rather than faithful reflections of reasoning [54, 84], potentially misleading teachers with plausible but ungrounded justifications. Despite established XAI principles, it remains underexplored how to design explainable GenAI systems to support teacher sensemaking in multimodal cognitive diagnosis.

3 Formative Studies

To address **RQ1** and guide the design of *OpenCD*, we conducted a two-part formative study. Part 1 investigated the characteristics of students' multimodal responses to open-ended tasks, informing the design of our automated analysis pipeline. Part 2 involved interviews with teachers to understand their sensemaking needs and preferences. We synthesized the findings from both parts into six design goals (DGs).

3.1 Study Rationale and Scope

To guide our study, we first established the target mathematical domain and response modalities based on a literature review.

For early mathematics, number sense is the foundational concept, encompassing children's initial understanding of numbers through basic operations in the early elementary grades [3, 72]. We therefore scoped our investigation to **number sense** at the first-grade level. To capture the full developmental spectrum of these core concepts, we recruited students ranging from late kindergarten through second grade.

Multimodal responses are essential for comprehensively reflecting students' mathematical cognition [68]. We based our design on the theoretical foundation of the well-established CPA approach [49]: the EIS model [12]. To align with this model and facilitate data collection on an iPad, we defined three primary interaction modalities: **dragging objects** (Enactive), **drawing** (Iconic), and **writing** (Symbolic). Furthermore, we included a **speech** modality, as extensive research highlights the importance of verbal explanations in revealing students' cognitive processes [50, 67].

3.2 Methods

Below, we describe the methods for Part 1 (Understanding Students) and Part 2 (Understanding Teachers) respectively.

3.2.1 *Part 1: Understanding Student Multimodal Responses to Open-ended Tasks.*

Participants. We recruited 14 students (6 boys, 8 girls; age M = 6.8, SD = 0.95) spanning late kindergarten through second grade (4 K, 6 G1, 4 G2). Recruitment was conducted via personal accounts on RedNote for in-person sessions in one city in China. Parents self-registered, and we filtered excess applicants solely to balance gender and grade representation, with no baseline assessments. The study was conducted at the start of the second semester, so students had completed one semester of prior instruction. Following IRB approval and parental written consent, students participated and received certificates as gifts upon completion. Participant details are provided in Appendix A.1.

Open-ended Tasks. Focusing on number sense, we designed multimodal tasks across 7 types, including number representation, quantity comparison, and basic operations [3, 62], supplemented by pre-designed follow-up questions (detailed in Appendix B.1).

Procedure. Sessions were conducted on-site (approx. 30–60 minutes per student). Participants interacted with a custom-developed web application on a 13-inch iPad with an Apple Pencil, which supported dragging, drawing, writing, and speaking modalities (interface details in Appendix C.1). The process followed a semi-structured format mimicking natural diagnostic dialogues: after each task, researchers may use follow-up questions to explore different response modalities. Throughout the session, we collected screen recordings, audio, and raw interaction data.

Data Analysis. Methodologically grounded in [10], we conducted a two-stage analysis (detailed in Appendix C.2). In the first stage, we analyzed student responses using a hybrid approach that combined a priori theoretical codes [31, 55] with open coding to identify 54 typical behaviors and infer cognitive states, while documenting our interpretative reasoning in memos. In the second stage, we performed a meta-analysis of these memos to extract the essential information needs and evidential requirements for diagnosis, synthesizing them into three core analytical challenges (F1-1 to F1-3) for system design.

3.2.2 *Part 2: Understanding Teachers' Sensemaking Needs from Interview.*

Participants. We recruited 5 experienced elementary school mathematics teachers. All had multiple years of teaching experience in elementary schools (8.6±3.2 years), including specific experience in first grade (3.8±2.4 years), and they came from different schools in 4 different cities in China. Recruitment was conducted via personal networks and private accounts on RedNote, through which we screened out applicants with insufficient teaching experience. Participants received 100 RMB (approx. $15 USD) for the one-hour session. See Appendix A.2 for details of teacher profiles.

Procedures. We conducted semi-structured interviews centered on three topics: (1) teachers' current practices and challenges in using open-ended tasks; (2) their perspectives on our proposed open-ended assessment method (we showed them examples of student responses); and (3) their expectations for the content and format of AI-generated diagnostic report (we showed some handcrafted examples).

Thematic Analysis. The interviews were audio-recorded, transcribed, and analyzed using thematic analysis [10]. Two researchers first independently performed open coding on two interview transcripts. They then discussed their codes and collaboratively developed a codebook. One researcher applied this codebook to the remaining three transcripts, and the second researcher reviewed the coded data. Finally, the researchers discussed and resolved any ambiguities or disagreements and jointly summarized the resulting themes.

3.3 Findings

Our two-part formative study yielded six key findings. We first report the three findings from Part 1, which highlight the core challenges in analyzing student responses and informed our algorithm design. We then present the three findings from Part 2, which articulated teachers' sensemaking needs and guided our system interface design.

3.3.1 *Challenges in Analyzing Student Responses.*

The open-ended tasks successfully elicited a rich variety of responses, and follow-up questions revealed an even wider range of student behaviors. However, the nature of these multimodal responses presents several analytical challenges.

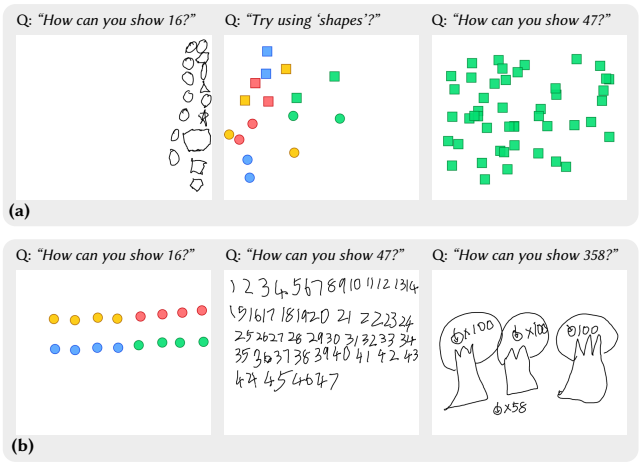


Figure 3: Examples of students showing consistent or varied strategies across tasks. (a) Student S5 consistently uses a one-to-one correspondence strategy. (b) In contrast, Student S2's strategies vary from dragging circles to writing numeral sequence and drawing semantic sketches.

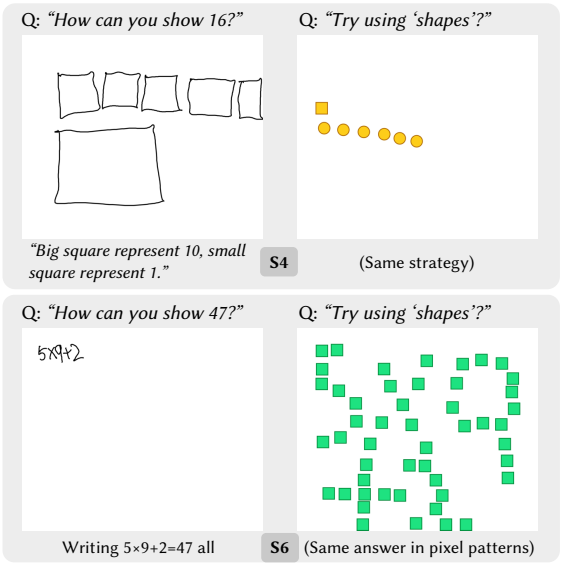


Figure 4: Examples of students showing "inertia of thinking" influenced by prior tasks. Left: S4 uses the same base-ten strategy (1 group of ten, 6 ones) with different items. Right: S6's object arrangement for 47 is a direct visual copy of their previous symbolic answer "5x9+2", making it difficult to understand on its own.

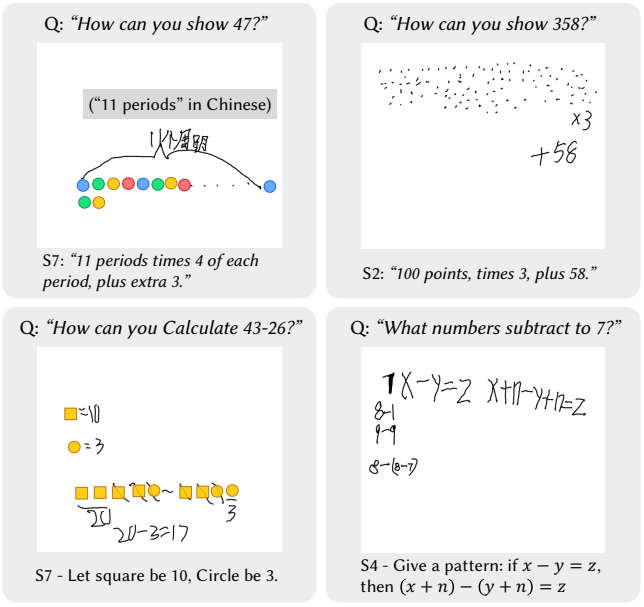


Figure 5: Examples of unique and idiosyncratic student responses. The students are shown inventing their own grouping rules, using mixed representations, defining custom units for calculation, and applying algebraic patterns.

Finding 1-1: Insufficiency of Single-Task Evidence. A student's performance on a single task is often incidental or incomplete, making it insufficient for a reliable cognitive diagnosis. As examples shown in Figure 3, some students consistently used the same method, while others demonstrated different strategies, switched methods when the task changed. This suggests that a holistic analysis across all of a student's responses is necessary for an accurate diagnosis.

Finding 1-2: Ambiguity and Information Loss in Final Products. The final state of a student's canvas is often insufficient to understand their intent and problem-solving method. We identified three crucial types of information required for interpretation: (1) *Process* is essential for deciphering cluttered final products and for understanding dynamic operations. (2) *Speech* is crucial for disambiguating visually similar behaviors and understanding a student's specific intent. (3) *Prior Responses*, which provides context, is necessary to interpret a student's current actions, especially to account for continuity in their reasoning (Figure 4).

Finding 1-3: Diversity and Unpredictability of Student Responses. The open-ended nature of the tasks leads to a vast diversity of student responses. While many responses aligned with established developmental frameworks, we also observed unique, idiosyncratic responses, such as the one described in the introduction (Figure 1) and others in Figure 5. Such responses cannot be anticipated or fully covered by a predefined rubric and require specific analysis. This highlights the limitation of purely rule-based systems, which cannot handle the "long tail" of student responses.

3.3.2 Teachers' Diverse Sensemaking Needs. All teachers highly valued the open-ended assessment approach, which helps uncover genuine student understanding, rather than relying on mechanical memorization. However, its practical application is limited by the time cost and the difficulty of observing the problem-solving process. All teachers were enthusiastic about our proposed system, but expressed diverse needs for the sensemaking of diagnostic results:

Finding 2-1: Teachers have divergent preferences on structured overviews and narrative reports. One teacher preferred a narrative summary, which she found familiar and easy to understand (T1). The others preferred a structured format (e.g., a knowledge graph or table) for an at-a-glance overview (T2, T3, T4, T5), but also wanted the option to access more detailed textual explanations and teaching suggestions (T3). This divergence resonates with challenges identified in teacher Visualization Literacy [75].

Finding 2-2: Teachers need access to the original student response and transparency of reasoning. All teachers stated that, in addition to the diagnostic results, they needed access to the students' original responses (T1-T5). They also wanted to see the reasons behind the system's judgments, such as how a student made a mistake (T1, T3), and expressed particular interest in reviewing typical or unique student responses (T1, T5). This is a need that aligns with research on maintaining teacher agency in AI-assisted systems [40].

Finding 2-3: Teachers required both individual and class-level views. Teachers emphasized that because they teach the class collectively, having a grasp of the class-level learning situation is a priority (T3, T4, T5). Specifically, T3 suggested using different colors to visualize the distribution of mastery levels, and T2 expressed a desire to see student responses grouped by behavioral category.

3.4 Design Goals

We synthesized six design goals (DGs) to guide the development of *OpenCD* based on the aforementioned six findings. The first three goals inform the algorithmic pipeline to address the challenges of multimodal analysis, while the latter three specify the interaction and visualization requirements for the teacher-facing interface.

- **F1-1 → DG1:** Accumulate evidence across multiple responses to form a holistic diagnosis.
- **F1-2 → DG2:** Incorporate multimodal process data (actions, speech) and prior context into the analysis pipeline.
- **F1-3 → DG3:** Utilize generative capabilities to interpret "long-tail" responses that fall outside predefined behavior sets.
- **F2-1 → DG4:** Provide both structured overviews for quick scanning and detailed textual reports for deep reading.
- **F2-2 → DG5:** Link high-level diagnostic conclusions back to specific behavioral evidence in the student's original response to ensure transparency.
- **F2-3 → DG6:** Integrate class-wide analytics for identifying common patterns with individual profiles for personalized intervention.

4 System Design of *OpenCD*

This section details the design of *OpenCD* (**RQ2**). We first introduce the core design rationale: the Evidence-Centered Design framework and the knowledge base derived from it (§ 4.1). We then describe how this framework support a VLM-rule hybrid diagnostic system

capable of processing multimodal responses from open-ended tasks (§ 4.2). Finally, we present the teacher-facing interface that visualizes the diagnostic results and behavioral evidence in a transparent and traceable way (§ 4.3), ensuring that *OpenCD*'s analysis is translated into understandable and trustworthy pedagogical insights.

4.1 Evidence-Centered Design for Automated Diagnosis

Traditional assessments often rely on closed-ended questions where behaviors are judged simply as right or wrong. In such cases, cognition is tightly linked with specific questions, such as in various Cognitive Diagnosis Models (CDMs) [23, 94]. In contrast, for open-ended tasks, it is the student behavior that serves as the crucial bridge between the tasks and cognition. Therefore, our approach involves analyzing multiple responses (**DG1**) to identify key behaviors that serve as evidence for cognitive diagnosis.

This analytical process relies on fine-grained pedagogical knowledge, including the typical behaviors, the key cognitive nodes, how behaviors map to cognition, and the developmental relationships between cognitive nodes. To structure this knowledge, we adopted the **ECD** framework [64]. We adapted it from a framework for assessment design [90] into a knowledge base and pipeline for automated diagnosis. Specifically, we used ECD's Evidence Model and Competency Model to form the knowledge base, thereby grounding *OpenCD* in pedagogical knowledge (Figure 6).

Typical Behaviors as Evidence. In ECD, the *Evidence Model* specifies the typical behaviors a student might exhibit for a given task type, and defines their mapping relationship to cognitive nodes. *OpenCD* will regard typical behaviors as evidence to diagnose the corresponding cognitive nodes.

Cognitive Graph. The *Competency Model* in ECD describes the space of student possible cognitive states. In our work, we aim to represent key cognitive nodes and their developmental relationships, for which we use a *Directed Acyclic Graph (DAG)*. It reflects the conceptual structures of cognitive development and the understanding of strategies across different representations. In the graph, each cognitive node serves as a unit for diagnosis. The edges between nodes represent *prerequisite relationships*. Traditionally, a prerequisite implies a constraint on mastery levels (i.e., mastery of a more advanced successor node ≤ mastery of a more foundational predecessor node) [100]. In *OpenCD*, we leverage this prerequisite structure for evidence propagation, as will be introduced in § 4.2.2.

Knowledge Base for Different Task Types. This knowledge base is specialized for different types of mathematical tasks. Drawing from relevant literature [17, 31] and previous analysis (§ 3.2.1), we developed this knowledge base for seven task types used in formative study (see Appendix G for details). The knowledge base from ECD serves a **dual purpose**: it supports the rule-based expert models to reliably map typical behaviors to cognition and to propagate evidence across the cognitive graph, and it acts as a source of domain knowledge to ground the VLM's interpretations.

4.2 VLM-Rule-Hybrid Diagnostic Pipeline

To accurately interpret students' diverse and often ambiguous multimodal response processes (**DG2**, **DG3**), we designed a VLM-rule hybrid system grounded in our ECD-derived knowledge base. To meet teachers' diverse needs, the system outputs both a diagnosis on the cognitive graph and a textual diagnostic report (**DG4**). The overall workflow is illustrated in Figure 6.

Our system integrates two types of large models: VLMs for processing multimodal input to analyze student response and collect evidence, and Large Language Models (LLMs) for text-based reasoning to synthesize evidence and generate diagnoses. In our implementation, both are using Gemini-2.5-Pro model [21]. Model configurations and prompts are detailed in the Appendix D.3 & H.

4.2.1 Understanding Multimodal Processes. (**DG2**) As analyzing only the final product is insufficient, our first challenge is to enable the AI to comprehend the entire dynamic, multimodal response process.

Converting Multimodal Temporal Data. Student responses are recorded by the iPad web app as multimodal temporal data, including pen strokes, object manipulations, and audio. Since current mainstream large models cannot directly process video or simultaneously handle images and audio streams [21], we convert this raw data into a text-and-image script. It preserves essential information—such as action descriptions and sequences, audio transcriptions, and keyframe renderings—to support process-based analysis. See Appendix D.2 for more details.

Response Comprehension. A VLM agent is assigned to interpret this text-and-image script with context of prior responses, to capture the student's core intent. We evaluated the agent's performance on the data collected from the formative study, achieving an accuracy of 93.6% (see details in Appendix D.1). The output of this agent serves three purposes: it is used iteratively to inform its own understanding of subsequent responses; it is passed, along with the full process script, to the subsequent behavior analysis agents; and it serves as a concise summary of the raw script for the final diagnosis agents, allowing them to incorporate multiple responses.

4.2.2 From Multimodal Behaviors to Pedagogical Evidence. (**DG3**) Student responses are diverse, including common, typical behaviors as well as many unexpected and unique methods, which are difficult for a purely rule-based system to cover. To balance flexibility and reliability, we designed a hybrid detection and mapping mechanism that combines VLMs with rule-based modules.

Behavior Detection. A VLM agent is assigned to identify key behaviors in the student's response. On one hand, the agent attempts to classify the student's actions with respect to "typical behaviors" from the knowledge base. On the other hand, if unclassified, the agent would provide an open-ended description for the behaviors.

Evidence Mapping. These identified behaviors are then mapped to nodes in the cognitive graph as direct evidence. For typical behaviors, this mapping is handled by a rule-based module to ensure accuracy. For unclassified behaviors, a VLM agent infers the corresponding cognitive node, guided by existed mappings provided in the knowledge base.

Evidence Propagation. This step is handled by a rule-based module. Using the *prerequisite* relations in the cognitive graph, behavioral evidence can inform the diagnosis of additional cognitive nodes, serving as indirect evidence. Evidence propagates along the edges: positive evidence propagates backward to support mastery

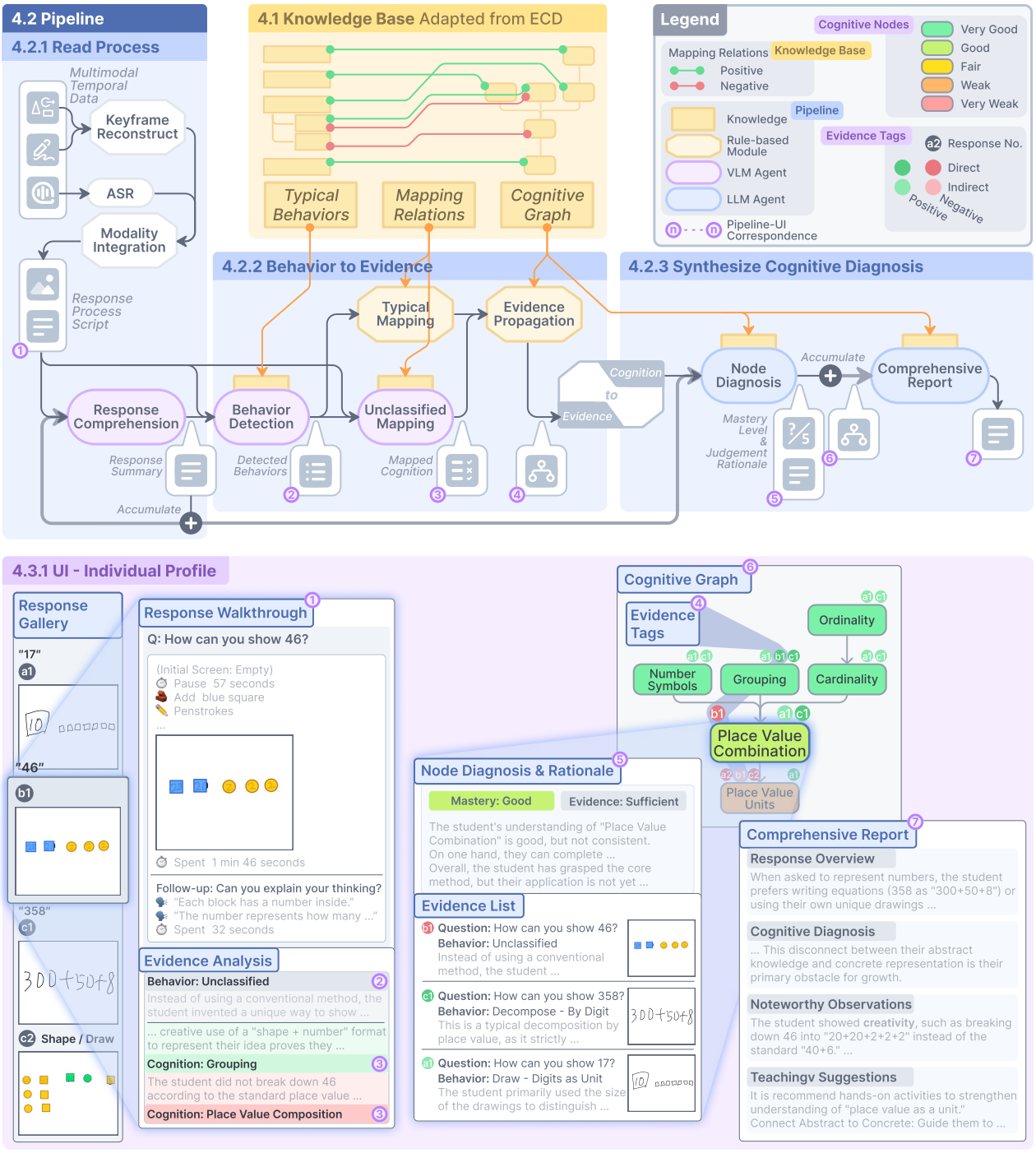


Figure 6: The architecture of *OpenCD*. (§ 4.1, top, yellow) The Knowledge Base, adapted from ECD, includes behaviors, mapping relations, and a cognitive graph. (§ 4.2, upper, blue) The Diagnostic Pipeline, uses a hybrid of rule-based modules and VLMs/LLMs, informed by the knowledge base. It consists of three stages: processing raw data, collecting behavioral evidence, and synthesizing final cognitive diagnosis. (§ 4.3, lower, purple) The Teacher-Facing Interface for an individual student, which visualizes these diagnoses, integrating the original Response Gallery, an interactive Cognitive Graph, and qualitative Reports. users can click the thumbnail of a response to view the Response Walkthrough and Evidence Analysis, or click a cognitive node to view its Rationale and Evidence List. The purple numbered tags indicate the transparent correspondence between the pipeline and UI.

of prerequisites, while negative evidence propagates forward to suggest non-mastery of more advanced nodes.

4.2.3 Synthesizing a Holistic Cognitive Diagnosis. After collecting behavioral evidence from all of a student's responses (**DG1**), the system performs a holistic cognitive diagnosis, providing both structured diagnostic results and a textual report (**DG4**).

Cognitive Node Diagnosis. An LLM agent reviews all the accumulated evidence (positive and negative, direct and indirect) for each cognitive node, and determines the sufficiency of the evidence (our measure of confidence [52]), assigns a mastery level (on a 5-point scale), and provides a textual rationale for its judgment.

Comprehensive Report. Based on all the preceding analysis, another LLM agent generates the final report including an overview of the student's responses, a cognitive analysis, noteworthy observations, and pedagogical suggestions.

4.3 Visualization of Diagnosis Results & Evidence

The application of AI diagnostics in education faces a core HCI challenge: How can we enable teachers to trust and effectively utilize the analytical results to support their pedagogical needs? Our interface design is guided by three goals: (1) integrating quantitative assessment with qualitative insights (**DG4**), (2) ensuring the transparency and traceability of the analysis process (**DG5**), and (3) supporting both individual and class-level perspectives (**DG6**).

4.3.1 Individual Diagnostic Profile. The interface for individual students is shown in Figure 6.

To satisfy teachers' need for both quantitative assessment and qualitative insights (**DG4**), the interface presents a textual report alongside the visualized cognitive graph. The graph uses **color-coding** to intuitively represent different mastery levels, allowing teachers to quickly identify weak cognitive nodes and developmental bottlenecks.

To build teacher trust in the AI's analysis (**DG5**), our interface features **bi-directional traceability**. This means teachers can explore the data from two directions:

- **From Response to Cognition:** When a teacher views a specific response, the interface highlights all the behavioral evidence detected and the corresponding cognitive nodes it informs.
- **From Cognition to Response:** When a teacher focuses on a cognitive node (e.g., by clicking a red node on the graph), the interface surfaces all the responses that support this judgment (i.e., *Evidence List*) along with the qualitative rationale.
- **Evidence as the Bridge:** All evidence is visually represented on the graph as intuitive *Evidence Tags*, attached to corresponding cognitive nodes. This serves a dual purpose: it can indicate the findings from each response (e.g., presence of problematic behaviors) while also implying the analytic status of each cognitive node.

Additionally, leveraging the metric defined in § 4.2.3, the interface presents evidence sufficiency as a proxy for diagnostic confidence. Nodes with insufficient evidence are visually highlighted (via opacity changes), proactively alerting teachers to potential uncertainty and mitigating the risk of misinterpretation.

4.3.2 Class Performance Dashboard. We designed the class-level dashboard as shown in Figure 7.

Since teachers manage the learning of the entire class, they hope to quickly grasp the overall learning landscape. This requires both statistical data on "how many students have mastered a concept," and qualitative insights into "how students are responding" (**DG6**). To this end, our class dashboard provides two complementary views:

- **Class Cognitive Graph:** This provides a cognitive mastery overview. It aggregates the diagnostic results (§ 4.2.3) for the entire class, using color to visualize the distribution of mastery levels. This allows teachers to quickly identify common areas of difficulty. Selecting a node can reveal student names at each mastery level.
- **Class Response Gallery:** This view categorizes student responses based on the "key behaviors" detected, providing an efficient, at-a-glance presentation of the different responding patterns. Teachers can immediately see which advanced strategies are prevalent, which typical novice approaches are common, and even specific types of errors.

5 Expert Evaluation of Diagnostic Accuracy

Before deploying *OpenCD* in a user study with teachers, we first needed to conduct a strict quantitative evaluation of its core diagnostic system's validity. Therefore, we designed an expert evaluation study guided by two primary questions: (1) How reasonable are the diagnoses generated by *OpenCD*? (2) What are the primary sources and patterns of error in the analysis?

To answer these questions, we recruited experienced elementary school mathematics teachers as experts. They carefully examined the reasonableness of the AI-generated diagnosis for each cognitive node and provided justifications for any results they deemed unreasonable. We then performed an in-depth error analysis based on their feedback.

5.1 Dataset

We recruited a new group of 16 first-grade students (gender balanced; age M = 7.3 years, SD = 0.29; 10 from public schools, 6 from private schools) to collect response data as the "test set" and then perform automated diagnosis using *OpenCD*. The study was conducted after the academic year, following the same protocol as formative study (§ 3.2.1), with three task types that are core to the first-grade curriculum and involve complex cognitive processes: Number Representation, Addition with Regrouping, and Subtraction with Regrouping. (See Appendix B.2 for details). This resulted in a total of 271 responses (16.9±2.0 for each student).

We processed the data collected using *OpenCD* with Knowledge Base established from formative study (See Appendix G for details). For each student at each cognitive node, the system generated (1) a **mastery level** (on a 5-point scale), (2) a judgement of **evidence sufficiency** (yes or no), and (3) a textual rationale explaining how the evidence was used to reach the conclusion. For the 3 task types, there are all together $6+7+7 = 20$ cognitive nodes, with 16 students, resulted $20 \times 16 = 320$ data points to review.

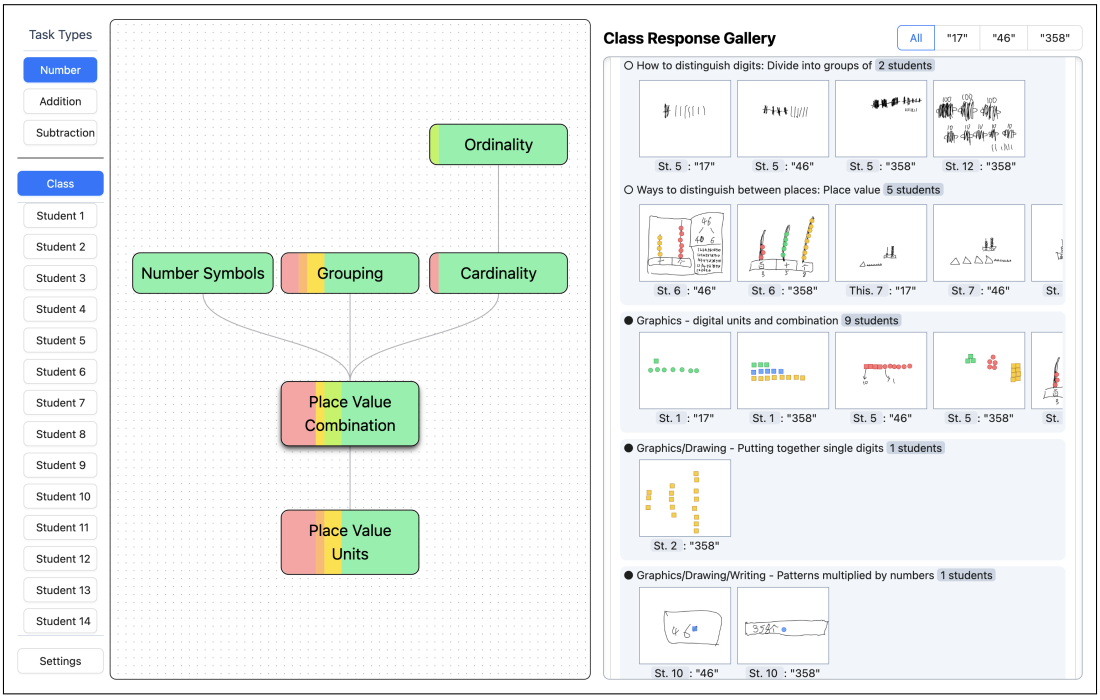


Figure 7: The Class Performance Dashboard, featuring two complementary views. Left: The *Class Cognitive Graph* visualizes the mastery distribution for each concept across the entire class, indicated by the colored segments in each node. Right: The *Class Response Gallery* groups student responses by their underlying strategies, allowing teachers to quickly identify common patterns and misconceptions.

5.2 Method

Expert Participants. We recruited two expert teachers (E1, E2) from different primary schools. E1 had 27 years of teaching experience, including 15 years teaching Grades 1–2. E2 had 28 years of teaching experience, including 5 years teaching Grades 1–2.

Procedure. Using a custom interface with *Scoring* and *Feedback* modes, the two experts followed a rigorous three-phase protocol (detailed in Appendix C.3): (1) Blind Rating: Experts first independently assessed the students' mastery and evidence sufficiency without AI; (2) Independent Review: Experts then evaluated the AI's diagnoses against their own baselines, labeling them as "Reasonable" or specifying error types (e.g., overestimation, missing evidence); (3) Consensus Meeting: Finally, experts met to discuss discrepancies and reach a final consensus on all evaluations.

5.3 Metrics

We used the following metrics:

- Diagnostic Reasonableness Rate: This was our primary outcome measure, defined as the percentage of diagnostic nodes rated "Completely Reasonable" based on the final consensus of the experts.
- Error Analysis: For diagnoses that were deemed unreasonable, we first statistically analyzed the error types. Then, guided by the experts' justifications, we performed a qualitative analysis to classify the cause of each error and traced it back to a root cause within the *OpenCD* pipeline.

5.4 Results

Diagnostic Accuracy. The results indicate a high degree of diagnostic accuracy. During the independent review phase, the inter-rater reliability (IRR) between the two experts was substantial (Cohen's $\kappa = 0.662$). After reaching consensus, experts rated **90.3% (289 out of 320) of the AI's diagnoses as "Completely Reasonable"**

The error types of the remaining 31 "Partially Unreasonable" cases are shown in Figure 8(a). Among this, the system underestimated the student's mastery level in 18 cases (including 1 substantial underestimation), significantly more than the 8 cases of overestimation. Furthermore, all 9 errors related to evidence sufficiency were misjudged as insufficient; there was 0 misjudged as sufficient.

Error Attribution. Our error attribution analysis traced the system's mistakes back to two main stages in the pipeline: final downstream errors in the Cognitive Node Diagnosis agent (§ 4.2.3) and upstream errors in the evidence analysis process (§ 4.2.2).

Downstream: Misestimating Evidence Strength. Downstream errors occurred in the Cognitive Node Diagnosis agent, which misestimated the strength or implication of correctly identified evidence. As shown in Figure 8 (b) (top row), this included: *Overestimating behavioral typicality* (e.g., giving undue credit for a simple instead of standard strategy). *Underestimating behavioral typicality* (e.g., dismissing a typical procedural error as simple carelessness). *Misreading strategy choice* (e.g., penalizing a student for using an appropriate alternative strategy).

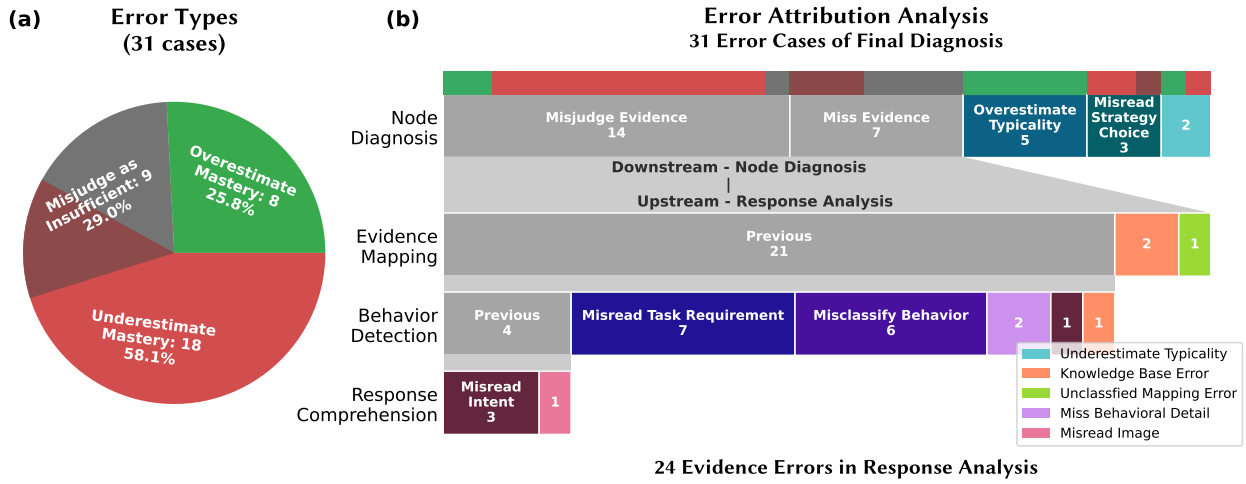


Figure 8: Analysis of the 31 cases rated as "Partially Unreasonable" by experts. (a) The distribution of final error types, showing that the system was more likely to underestimate (58.1%) than overestimate (25.8%) student mastery, and was likely to misjudge evidence as insufficient (29.0%). (b) An error attribution analysis that traces the root causes of these errors back to different stages of the diagnostic pipeline, from downstream errors in 'Node Diagnosis' to upstream errors in 'Response Comprehension.'

Upstream: Missing and Incorrect Evidence. Most errors (24 errors across 23 responses) originated upstream. These typically involved the system either missing behavioral evidence or identifying incorrect evidence (i.e., misjudging a behavior). A root cause analysis traced these 24 errors to the following agents:

- *Evidence Mapping.* Incomplete mapping rules in the Knowledge Base (e.g., failing to map a behavior as negative evidence for a higher-level cognition), failures in the Unclassified Behavior Mapping agent.
- *Behavior Detection (Primary Source).* This agent was the main source of errors, including misreading open-ended task requirements, misclassifying behaviors, missing fine-grained behavioral details, and misreading student intent. There was also error due to incomplete behavioral features in the Knowledge Base (e.g., lacking cues for process-based error detection).
- *Response Comprehension.* Errors included misreading ambiguous student intent or misreading a cluttered image.

Notably, the complexity of the Behavior Detection agent makes it particularly error-prone and a key focus for our future optimization.

5.5 Summary and Implications

Our expert evaluation demonstrates that *OpenCD* achieves a **high reasonableness rate exceeding 90%**. Crucially, its primary failure mode is shown as a clear **conservative bias**: a tendency to underestimate students' mastery and evidence sufficiency. Unlike general LLMs, which often exhibit overconfidence [63], our **Evidence-Centered Design promotes more cautious judgements**, thus avoiding misleading conclusions [22, 43].

In high-stakes applications like education, this conservative tendency of *OpenCD* can be interpreted as a **desirable safety feature**. When the system reports "insufficient evidence", it's in fact inviting the teacher to carry out a more in-depth review, fostering a more responsible human-AI collaboration model. In addition, as we will discuss in § 7.1.3, this conservative approach also reduces the risk of overlooking student problems and implicitly encourages students to express their thinking more clearly.

6 User Study with Teachers

Having established the diagnostic accuracy of *OpenCD*, we conducted a qualitative user study with 20 teachers to investigate its practical value and utility (**RQ3**). This study focused on how teachers perceive *OpenCD*'s potential to influence their analysis burden and diagnostic insights, and integrate into their existing pedagogical practices.

6.1 Methods

We conducted a within-subjects comparative experiment, having teachers perform student learning analysis—similar to their regular teaching practices—either with or without support from *OpenCD*'s diagnosis to demonstrate its value.

Participants. 20 elementary mathematics teachers were recruited via personal networks and snowball sampling, representing diverse professional profiles. Their total teaching experience ranged from 1 to 35 years (M = 9.3, SD = 9.25), with specific experience in Grades 1–2 from 0 to 7 years (M = 2.8 years, SD = 1.79). The participants were drawn from 15 different schools across 5 cities in China, adhering to the National Curriculum Standards. To ensure sample diversity and minimize institutional bias, we limited the number of participants from the same school. With a 7-point Likert scale, participants reported a high frequency of using educational technology (M = 6.45, SD = 0.76) and a generally positive attitude toward AI (M = 6.15, SD = 0.81). Profile details along with their educational context and technical background are presented in Appendix A.3.

Conditions & Tasks. We employed a within-subjects design with two conditions, using the dataset of 16 students from the expert evaluation study (§ 5.1). For the experiment, we selected two task types: *Number Representation* and *Addition with Regrouping* [1]. In each session, teachers analyzed a balanced set of 12 students (randomized order), while the remaining 4 students were reserved for the tutorial. In addition, video clips of every student's response were provided in both conditions via a cloud storage link.

- **Condition A (Manual):** Teachers accessed student response processes via the web app but without AI diagnoses (see Appendix C.4 for interface details).
- **Condition B (*OpenCD*-Assisted):** Teachers used the full *OpenCD* system.

Participants were framed in a scenario: *preparing for a new semester by analyzing students' open-ended pre-tests.* They were tasked with writing a report covering overall class performance, struggling students' situations, unexpected responses, and future teaching decisions.

Procedure. The study was conducted remotely using video conferencing. We used a counterbalanced 2×2 design to alternate the order of conditions and task types. The procedure consisted of three phases:

(1) *Onboarding (30 min):* Participants received training on the cognitive graphs and practiced with the system (Tutorial Mode) using the 4 training student datasets to ensure proficiency.
(2) *Experimental Sessions (2 × ∼50 min):* In each session, teachers analyzed the 12 students under the assigned condition and task type. They wrote an analysis report and then completed the NASA-TLX [33] and a self-reported performance scale. Breaks were allowed during the sessions.
(3) *Post-Study Interview (30–45 min):* We held semi-structured interviews to probe teachers' analytical processes and perceived value of *OpenCD*, followed by functional feedback surveys.

6.2 Data Collection and Analysis.

We collected following two types of data:

Quantitative data. NASA-TLX and self-reported performance (7-point Likert) were compared using the Wilcoxon signed-rank test. Responses to the interview surveys—including system feature ratings, multiple favorite selections, and Likert-scale scores for perceived influence and willingness to use—were analyzed using descriptive statistics.

Qualitative data. Semi-structured interviews were transcribed and analyzed using thematic analysis [10]. First, two researchers independently performed open coding on 4 transcripts to collaboratively develop an initial codebook. The remaining 16 transcripts were then divided between the researchers: each coded 8 transcripts and cross-checked the other 8. Finally, the researchers resolved all discrepancies through discussion to synthesize the key themes and insights.

[1]We didn't use "Subtraction" type because its underlying cognitive structure and student response patterns are highly similar to those of "Addition".

6.3 Findings

6.3.1 **Streamlining Teachers' Analysis Process**. We found that *OpenCD* effectively supported and facilitated teachers' needs for understanding student learning, significantly reducing their perceived cognitive burden (mental demand: W=25.5, p=.046*; effort: W=0, p=.0045**) and enhancing their confidence (W=4.5, p=.0081**), as shown in Figure 9. Teacher interviews further confirmed these findings.

Whole-Class Performance at a Glance. For class-wide instruction, a teacher's primary need is to quickly grasp the overall learning status of the class. Teachers found this process highly challenging when done manually. They had to analyze each student's work individually and then either rely on their often-unreliable memory (T2, T5, T12, T19) or invest significant effort in taking notes of their observations (T5, T8, T11).

In sharp contrast, *OpenCD*'s *Class Cognitive Graph* received the highest praise (Figure 10) and was considered "*the most useful for whole-class teaching*" (T16, T17). Teachers appreciated the cognitive-level statistics as intuitive, accurate, and clear at a glance (T2, T5, T9, T10, T13, T14, T15, T17, T19). They also noted that the behavioral-level clustering in *Class Response Gallery* preserved the diversity and common patterns in student responses, making the analysis more informative (T1, T9, T12, T18).

Optimizing Effort Allocation through AI-Guided Screening. Teachers' instructions are often problem-oriented, prioritizing their energy for struggling students who require more attention. Manual analysis forced them to review every student's response indiscriminately, a process widely described as tedious and exhausting (T1, T5, T10, T12, T15).

OpenCD acted as a "screener" to help teachers quickly pinpoint problems. Teachers would identify struggling students from *Class Cognitive Graph* (T10, T12, T14, T15, T17, T19, T20). Then in the *Individual Cognitive Graph*, they could navigate from red markers (cognitive nodes and *Evidence Tags*) to specific cognitive weaknesses (T2, T4, T8, T9, T10, T12, T15, T17) or corresponding negative evidence in the student's work (T1, T2, T6, T7, T8, T10, T12). This approach significantly optimized their effort allocation. As T12 stated: "*I can spend my time where it matters most. Typical students [with difficulties] are easier to find out with AI, so my attention is more focused and targeted.*"

6.3.2 **Enabling Deeper Insights into Student Thinking**. Qualitative feedback indicates that *OpenCD* empowered teachers to fully leverage the richness of process data, thereby helping them capture both deep cognitive structures and subtle behavioral cues.

Uncovering the Value of Response Process Information. Teachers broadly recognized the huge value of multimodal processes in *Response Walkthrough*, noting that they break through the limitations of paper to better reflect students' true thinking (T1, T2, T6, T11, T12, T13, T15, T16, T17). However, the high cost of manually analyzing process information often led them to overlook important details (T1, T2, T4, T9, T12, T13, T18).

***OpenCD* bridges the gap between the value of information and the cost of analyzing it**, by automating the detection of key behaviors, thus fully utilizing the significance of process data. T9

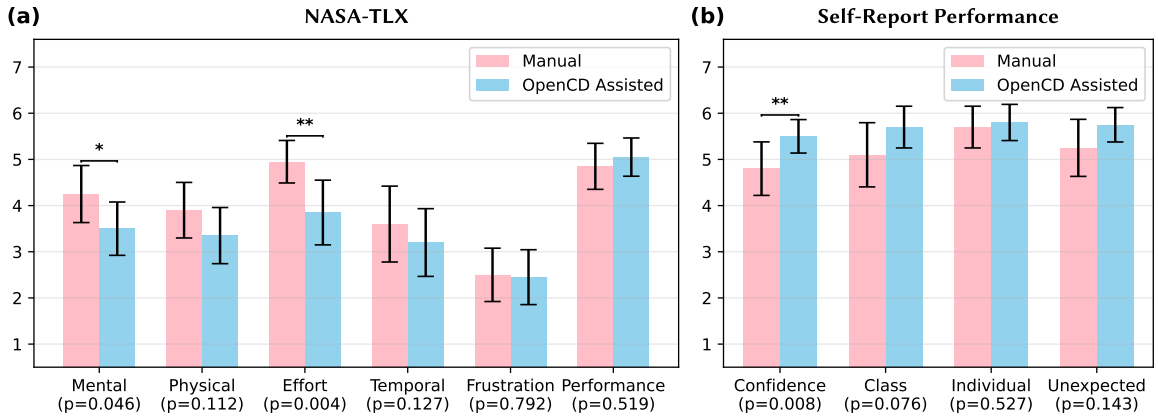


Figure 9: Comparison of teacher experience between the Manual condition and the *OpenCD*-Assisted condition (N=20). Error bars represent 95% confidence intervals. (a) NASA-TLX task load scores, where lower scores are better except for last Performance. *OpenCD* significantly reduced the perceived Mental Demand and Effort. (b) Self-reported performance scores, where higher scores are better. Using *OpenCD* significantly increased teachers' Confidence in their diagnoses. (p < 0.01, *p < 0.05)**

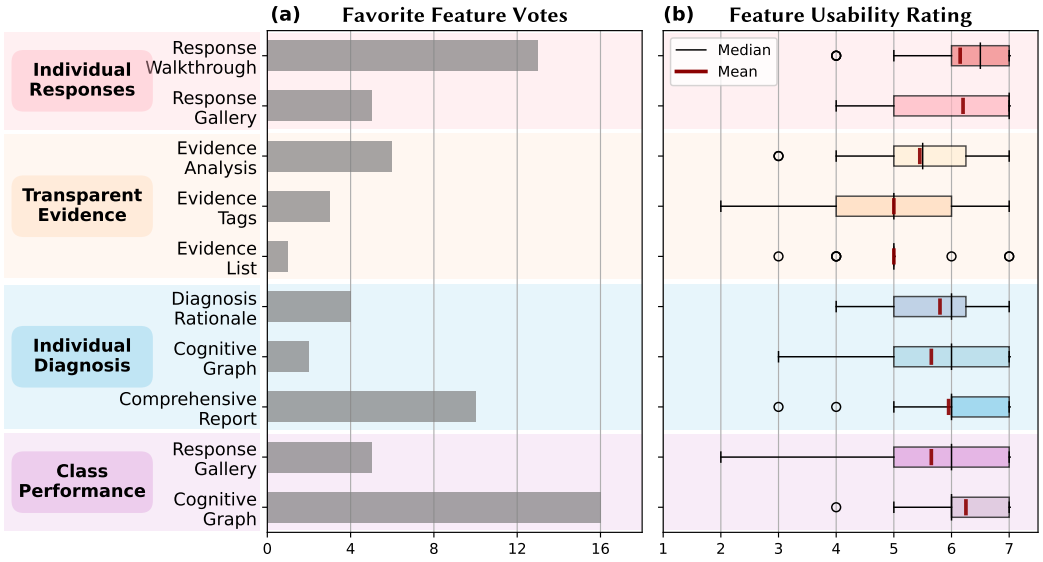


Figure 10: User study results (N=20) on feature usability ratings and preferences. This chart combines two visualizations for ten system features, which are grouped into four categories. (a) shows the number of votes for teachers' favorite feature(s). (b) shows the distribution of teachers' usability ratings for each feature on a 7-point Likert scale.

shared: "*When I saw the student's work was blank, I assumed they had no ideas. But the AI's different analysis reminded me to look at their process and what they said, and I found out they actually had right ideas [but they erased it in the end].*"

Enhancing the Granularity and Depth of Analysis. Guided by the knowledge base, *OpenCD* performs comprehensive diagnosis, which participants reported helped enhance the depth and granularity of their analysis. Several teachers admitted that their usual practices were relatively coarse or only focused on surface-level behaviors (T8, T9, T17, T19). The system's fine-grained cognitive diagnosis was seen as an effective supplement and extension to their own thinking (T4, T5, T9, T13, T16, T17, T19, T20).

T19 admitted: "*[To show a number,] a student arranged objects [as pixels] to form a pattern, and I just thought, 'okay, they just arranged them like that.' But AI told me it might be due to the child's insufficient understanding of number structure and suggested using more tools for practice. That helped me understand better.*" A participant with a principal's perspective (T1) also highly praised this: "*Our teachers' analysis usually stops at whether students can do it or not. But the AI can extract underlying mathematical competencies and mindset. This is something that average teacher cannot achieve.*"

Surfacing Subtle Behavioral Cues. Furthermore, *OpenCD* could capture nuanced yet significant behavioral details that even experienced teachers might neglect, thereby revealing profound differences in students' cognition (T2, T9, experience ≥ 8 years). T2 mentioned an example where *OpenCD* helped disclose hidden misconceptions behind procedural error: during a regrouping addition, "*the AI identified a student's error of erasing all items in the ones place instead of the required 10 items,*" which she had not noticed during a quick review. Another example shared by T9 highlighted identifying more advanced thinking strategies: when calculating 14+7, the student decomposed and added 14+6, "*I thought they tried to 'make ten.' But AI reminded me that he was 'making a round number,' which is essentially different. 'Making a round number' is to facilitate calculation which is more advanced, while 'making ten' is just place-value. AI actually broadened my perspective a little bit.*"

Nevertheless, a subset of teachers expressed concerns about "*excessive detail*" (T6, T8, T11, T12). They stressed the practical constraints of daily instruction, where managing large classes limits their capacity to attend to fine-grained individual differences (T8). We further discuss strategies to accommodate these varying preferences and mitigate potential friction in § 7.2.

6.3.3 **Potential for Supporting Professional Reflection**. Beyond assisting with in-the-moment diagnostic tasks, our qualitative data suggests that *OpenCD* may support teachers' reflection on their diagnostic practices.

Modeling Analytical Frameworks for Novices. For several less-experienced teachers (T5, T17, T19, experience ≤ 4 years), *OpenCD* acted as a pedagogical guide. The system's *Evidence Analysis* allowed them to observe how specific student actions could be mapped to underlying mathematical thinking. Teachers reported that it helped them internalize a more structured analytical framework (T5, T19). Veteran teachers (T2, T3, T8, T15, T16, experience ≥ 8 years) also commented that they would highly recommend this feature to novice colleagues for training purposes.

Prompting Reflection for Veteran Teachers. Even experienced teachers (T6, T9, T16, experience ≥ 8 years) reported that the system broadened their perspectives. T16 noted that "*after looking at the AI's analyses a lot, I noticed my own accuracy and granularity improved too. It kind of taught me along the way. Because people are naturally experience-driven (and it helped me move beyond).*" T6 mentioned that it inspired her to think more carefully and systematically during her subsequent manual analysis.

Inspiration from Personalized Suggestions. The personalized teaching suggestions generated by the system were highly praised for their specificity (T2, T4, T6, T9, T10, T16, T19). Even a veteran teacher (T14) admitted that while she was proficient at class lesson planning, she felt short at providing individualized guidance and planning for students' future development. Therefore, these detailed and targeted pedagogical suggestions were seen as a great help and had the potential for direct application in "home-school communication" (T1, T2, T3, T9, T10, T16).

6.3.4 **Trust and Reliance**. On these issues, teachers demonstrated complex and even divergent situations.

Transparency Fostered Trust and Enabled Error Detection. Quantitative results indicate high levels of teacher trust (Figure 11).

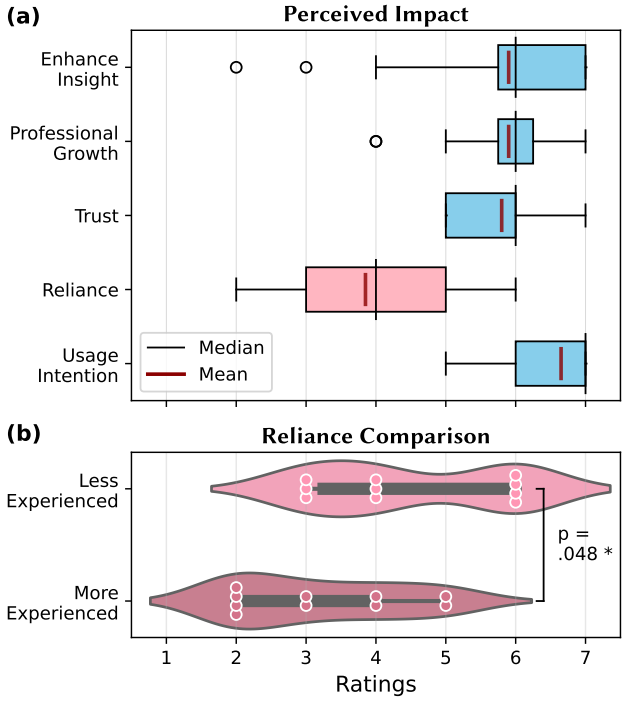


Figure 11: Teachers' perceived impact of *OpenCD* and an analysis of their reliance on the system (N=20). (a) Box plots showing high teacher ratings (on a 7-point scale) for the system's ability to Enhance Insight, foster Professional Growth, and their high Trust and Usage Intention. (b) A violin plot comparing self-reported reliance between 10 less experienced (≤ 6 years) and 10 more experienced (≥ 8 years) teachers. The less experienced group showed a significantly higher reliance on the system's diagnoses (*p < .05).

Qualitatively, participants attributed this to the system's transparency—specifically, the *Evidence Analysis* feature (T4, T9, T10, T11, T16, T18, T19) and the highlights of insufficient evidence (T18). These features allowed teachers to comprehend the AI's reasoning process (T10). Notably, high trust did not result in passive acceptance. Leveraging the source-level transparency provided by the *Response Gallery* and *Response Walkthrough*, the majority of teachers adopted a workflow of "**independent judgment followed by AI verification**" (T1–T6, T9, T11, T14, T19, T20). This approach enabled them to identify occasional AI errors, particularly cases where the system was overly conservative (T2, T7, T8, T9, T10, T12, T15).

The Reliance Gap. However, we observed a reliance gap based on teaching experience. As shown in Figure 11(b), teachers with less experience (≤ 6 years) reported significantly higher reliance on the system compared to their more experienced (≥ 8 years) counterparts (Mann-Whitney U test, U=24.0, p=.048*). Novice teacher T17 admitted, "*I would involuntarily follow the AI's line of thought, which makes it hard for me to think otherwise*", while others (e.g.,

T19) relied primarily on the AI-generated summaries rather than raw data.

7 Discussion

Our work contributes to the HCI literature on educational assessment tools by advancing beyond outcome-oriented evaluation and behavioral metrics [61, 95]—toward deep, process-oriented cognitive diagnosis. However, such deeper analysis introduces greater risks of automation, thereby imposing higher requirements for trustworthy XAI design. We address this through our grounded hybrid architecture and transparent interface design (§ 7.1). Furthermore, our evaluation reveals the nuanced tensions inherent in teacher-AI collaboration (§ 7.2). Ultimately, we demonstrate the potential for scalable process-oriented assessments in practice (§ 7.3).

7.1 Designing Trustworthy XAI for Educational Assessment

Our work demonstrates how the design of XAI shapes user trust and adoption in high-stakes educational contexts. We discuss three key design implications.

7.1.1 **Grounding GenAI via Hybrid Architecture**. While GenAI offers flexibility in interpreting unstructured multimodal data, its inherent stochasticity and tendency for hallucination pose significant risks in educational diagnosis [38]. Our study suggests that a hybrid architecture—synergizing the generative capabilities of VLMs with the deterministic reliability of rule-based systems—offers a reliable solution [85]. Crucially, we leveraged ECD as an educational framework for better interpretability. By grounding the VLM's outputs in a verified knowledge base, the system mitigates the unpredictability often associated with end-to-end models. As shown in § 5.4, this architectural choice was central to achieving the high diagnostic accuracy and effectively preventing model overconfidence, thereby establishing a solid foundation for trustworthy XAI.

7.1.2 **Fostering Trust Calibration through Process Transparency**. In the era of GenAI, systems are increasingly capable of generating persuasive *post-hoc rationalizations*, which can mislead users into unearned trust [37, 84]. To counter this, *OpenCD* prioritizes process transparency by faithfully exposing its underlying reasoning. This approach highlights a broader **XAI design opportunity for multi-agent systems**: intermediate agent outputs can be intentionally engineered as structured, intelligible artifacts for human visualization [102]. Accordingly, rather than simply explaining a final result, our pipeline visualizes meaningful intermediate states (e.g., *Evidence Tags*) that are inherently interactive. Unlike traditional feature-importance XAI methods (e.g., SHAP [58], LIME [79]) that offer mathematical approximations of logic, this process transparency provides semantic attribution, linking diagnostic conclusions directly back to student behaviors. This design fostered a "trust calibration" process observed in our user study: teachers first carefully reviewed these intermediate steps to verify the reasoning logic, and after establishing trust, shifted to utilizing conclusions directly. Thus, the system empowers teachers with the agency and the capability to identify discrepancies, serving as a vital safeguard for human-in-the-loop accountability.

7.1.3 **Aligning Algorithmic Conservatism with Formative Assessment Goals**. Formative assessment, unlike summative testing, aims to identify gaps in student learning to inform instruction, often prioritizing *recall* (detecting struggles) over precision [8]. Our evaluation reveals that the error pattern of *OpenCD* is a conservative bias—a tendency to underestimate students' mastery. Interestingly, teachers showed high tolerance because it aligns with their pedagogical goal of *screening*. They perceived such errors as beneficial strictness, which means student's expression was not clear enough, indicating room for improvement. This suggests that in educational tools, the key to trustworthiness is not necessarily eliminating "algorithmic bias", but aligning this bias with the specific pedagogical stakes [25].

7.2 Navigating the Tensions in Teacher-AI Collaboration

Beyond diagnostic accuracy, our findings reveal the nuanced tensions that emerge when AI tools interact with the diverse expertise and mental models of teachers.

7.2.1 **From Over-Reliance to Capacity Building**. We observed a divergence in usage patterns driven by professional expertise. Experienced teachers, possessing proficient domain skills, mainly treated the AI as a "verifier" for their independent judgments. In contrast, novice teachers showed a higher tendency towards Automation Bias [70], relying on AI outputs to guide their thinking [34]. This raises a critical concern regarding deskilling: long-term, passive reliance might hinder the professional development of novices and lead to skill decay in experienced teachers. Consequently, AI systems for professionals must be designed for capacity building, not just efficiency. Future designs can consider incorporating Cognitive Forcing Functions [13]—such as explicitly highlighting uncertain cases (e.g., via the insufficient evidence hints in *OpenCD*), or actively enabling and encouraging manual error correction [24]. Such mechanisms would encourage active engagement and independent judgment, thereby cultivating expertise rather than replacing it.

7.2.2 **Bridging the Gap between Algorithmic Frames and Mental Models**. Sensemaking is often described as the process of fitting Data into a Frame [44]. In *OpenCD*, the Cognitive Graph serves as a computational frame to structure the analysis [71]. However, friction arose when this frame mismatched a teacher's internal mental model. When this happened, teachers retreated to alternative formats—either identifying patterns directly from the "Data" level, or relying on the unstructured *Qualitative Reports* [15]. This mismatch manifested in two dimensions: granularity (some teachers accustomed to coarser-grained analysis) and structural logic (some accustomed to hierarchical "mind-map" instead of the prerequisite causal graphs).

This friction suggests that the "knowledge" underlying AI assessments should not be statically defined. Instead, this points towards the potential of **participatory knowledge base design**. Mechanisms that enable teachers to contribute to open-ended tasks, possible behaviors and underlying cognition, could resolve frame mismatches and enhance system acceptance [48]. More importantly, such an approach might offer opportunities for expert teachers to

encode and propagate their pedagogical wisdom, while serving as an evolving learning resource for novices.

Furthermore, our findings reveal a **tension between Micro-context and Macro-structure** [36]. For diagnosing individual student (Micro), the structured frame often conflicted with teachers' preference for rich, unreduced context. However, this same structure is necessary for class-level aggregation (Macro), which is highly valued by teachers (Figure 10). Therefore, effective educational AI must support multi-scale sensemaking: offering structures for aggregation while allowing fluid transitions to flexible narratives for individual inquiry. This duality should resonate with the two distinct pedagogical needs: the efficiency required for in-class collective instruction versus the depth required for post-class personalized tutoring.

7.3 Enabling Scalable Process-Oriented Assessment in Practice

Our research is not merely an evaluation of a specific tool, but an exploration of how AI can help overcome the practical barriers of implementing process-oriented assessment in real-world classrooms.

7.3.1 **Unlocking the Value of Multimodal Process Data for Math Literacy Assessment**. Teachers universally recognize that a student's problem-solving process contains far richer cognitive information than the static final product [14]. Similarly, regarding learning goals, there is a consensus on the need to shift focus from rote procedural skills to mathematical literacy [81]. However, in traditional practice, capturing the dynamic, multimodal evidence required for such assessment demands prohibitive effort [9]. Consequently, even open-ended tasks (e.g., drawing) are often treated as "closed-ended problems"—graded solely on whether the final outcome matches a standard template, fostering instructional practices centered on rote repetition and imitation.

The core contribution of *OpenCD* is that it significantly lowers the cost of analyzing this messy process data. By automating the interpretation of multimodal inputs, the system reclaims the value of open-endedness, making it feasible to incorporate process data into routine assessment at scale. This technological affordance facilitates a shift in assessment focus: from judging the simple "correctness of a behavior" to interpreting the "level of cognitive development." By surfacing students' intuitive expressions, the system can uncover profound understanding gaps—such as students who can perform algorithms but fail to represent numbers concretely—thereby providing a scalable solution for assessing and cultivating true mathematical literacy.

7.3.2 **Adapting Personalized Diagnosis to Collective Instruction Constraints**. Idealistic assessment philosophies must confront the constraints of real-world teaching environments, where collective instruction in large classes remains the norm [88]. As noted by participants, even with fine-grained individual diagnoses, the capacity for teachers to provide exclusive one-on-one guidance is limited. *OpenCD* addresses this by alleviating the cognitive burden of the diagnostic phase, enabling teachers to allocate their limited resources more effectively towards differentiated feedback within a collective setting [42].

Furthermore, this depth of diagnosis opens new possibilities for future AI-driven interventions. Traditional Intelligent Tutoring Systems (ITS) often limit interventions to recommending similar closed-ended problems [97]. In contrast, because *OpenCD* assesses cognitive development through multimodal processes, it lays the foundation for next-generation ITS that can suggest comprehensive interventions—such as hands-on manipulation tasks or drawing exercises—to holistically foster cognitive development in digital learning environments.

8 Limitations and Future Work

First, the generalizability of our findings is constrained by the scope of our participants and content. We focused exclusively on first-grade number sense concepts with students recruited from a single city. Additionally, the high EdTech familiarity of our teacher participants, while consistent with the ubiquity of digital infrastructure in Chinese schools, may have contributed to higher system acceptance compared to less digitized contexts. The applicability of our knowledge base and VLMs' performance on more complex, higher-grade topics remain to be validated. Moreover, as the interrelation of mathematical knowledge increases, our current method of manually constructing the cognitive graph becomes a scalability bottleneck. Future work aims to explore semi-automatic or fully automatic methods for constructing domain-specific cognitive graphs [2, 100] to efficiently expand the system's content coverage.

Second, our technical implementation involves trade-offs regarding processing and latency. Our current approach of converting multimodal data into a text-and-image script is a practical workaround; future advances in large multimodal models that can process video directly may offer more holistic analysis. Additionally, the multi-step analysis pipeline induces latency (approximately one minute per response), which currently precludes real-time adaptive questioning. Optimizing algorithmic efficiency to enable low-latency, adaptive assessment is a key technical direction. We also acknowledge that due to the resource-intensive nature of expert reviews, our validation focused on final diagnostic outcomes rather than the accuracy of intermediate steps like behavior detection. Furthermore, the current interface does not allow users to explicitly override AI errors as a feedback mechanism, which has the potential to refine the model over time.

Finally, regarding ecological validity, our user study involved teachers analyzing work from unfamiliar students recruited individually. This setup does not fully capture the dynamics of a real classroom, where teachers leverage prior knowledge of their students and the cohort exhibits common learning patterns. Furthermore, the benefits reported in this study rely on short-term teacher perceptions. While we have demonstrated the system's utility for educators, a critical remaining question is how these diagnostic insights translate into tangible benefits for learners. Consequently, the system's objective impact on teaching practices and student learning outcomes can only be ascertained through long-term, in-situ field studies. Future deployments should also address practical hardware constraints in schools, potentially exploring integration with existing classroom technologies like digital dot-matrix pens to improve ecological fit.

9 Conclusion

In this paper, we addressed the challenge of interpreting children's multimodal open-ended response processes to support teacher sensemaking. We presented *OpenCD*, a system that synergizes VLMs with expert models grounded in Evidence-Centered Design to perform fine-grained cognitive diagnosis. Through an interface designed for process transparency, the system enables teachers to trace AI judgments back to behavioral evidence, empowering them with deeper insights while effectively reducing analysis burden. Our work demonstrates the potential of scalable, process-based assessment and contributes to the design of trustworthy XAI for cognitive diagnosis.

Acknowledgments

This work was supported by the Natural Science Foundation of China under Grant No. 62132010. We also extend our special thanks to Chen Qi at Tsinghua University Primary School for his generous support and insightful guidance on the design and experiments of this research. Finally, we acknowledge the use of LLMs as writing assistants in the preparation of this manuscript. Their role was limited to improving the language, including grammar and clarity, and assisting with the translation of terms. The authors reviewed and edited all AI-generated suggestions and assume full responsibility for the final content of this paper.

References

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–18. doi:10.1145/3173574.3174156

[2] Erik Andersen, Sumit Gulwani, and Zoran Popovic. 2013. A Trace-Based Framework for Analyzing and Synthesizing Educational Progressions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 773–782. doi:10.1145/2470654.2470764

[3] Paul Andrews and Judy Sayers. 2015. Identifying Opportunities for Grade One Children to Acquire Foundational Number Sense: Developing a Framework for Cross Cultural Classroom Analyses. *Early Childhood Education Journal* 43, 4 (July 2015), 257–267. doi:10.1007/s10643-014-0653-6

[4] Arthur J Baroody. 2017. The Use of Concrete Experiences in Early Childhood Mathematics Instruction. In *Advances in Child Development and Behavior*. Vol. 53. Elsevier, 43–94.

[5] Lawrence W. Barsalou. 2008. Grounded Cognition. *Annual Review of Psychology* 59, 1 (Jan. 2008), 617–645. doi:10.1146/annurev.psych.59.103006.093639

[6] Melanie Bertrand and Julie A. Marsh. 2015. Teachers' Sensemaking of Data and Implications for Equity. *American Educational Research Journal* 52, 5 (Oct. 2015), 861–893. doi:10.3102/0002831215599251

[7] Camilla Björklund and Ulla Runesson Kempe. 2022. Strategies Informed by Various Ways of Experiencing Number Relations in Subtraction Tasks. *The Journal of Mathematical Behavior* 67 (Sept. 2022), 100994. doi:10.1016/j.jmathb.2022.100994

[8] Paul Black and Dylan Wiliam. 1998. Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice* 5, 1 (March 1998), 7–74. doi:10.1080/0969595980050102

[9] Paulo Blikstein and Marcelo Worsley. 2016. Multimodal Learning Analytics and Education Data Mining: Using Computational Technologies to Measure Complex Learning Tasks. *Journal of Learning Analytics* 3, 2 (Sept. 2016), 220–238. doi:10.18608/jla.2016.32.11

[10] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp063oa

[11] Jerome S. Bruner. 1964. The Course of Cognitive Growth. *American Psychologist* 19, 1 (Jan. 1964), 1–15. doi:10.1037/h0044160

[12] Jerome Seymour Bruner. 1974. *Toward a Theory of Instruction*. Harvard university press.

[13] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–21. doi:10.1145/3449287

[14] Hugh Burkhardt and Mathematical Sciences Research Institute. 2007. Mathematical Proficiency: What Is Important? How Can It Be Measured? In *Assessing Mathematical Proficiency*, Alan H.Editor Schoenfeld (Ed.). Cambridge University Press, 77–98.

[15] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–24. doi:10.1145/3359206

[16] Fabio C. Campos, June Ahn, Daniela K. DiGiacomo, Ha Nguyen, and Maria Hays. 2021. Making Sense of Sensemaking: Understanding How K–12 Teachers and Coaches React to Visual Analytics. *Journal of Learning Analytics* 8, 3 (July 2021), 60–80. doi:10.18608/jla.2021.7113

[17] Thomas P. Carpenter and James M. Moser. 1984. The Acquisition of Addition and Subtraction Concepts in Grades One through Three. *Journal for Research in Mathematics Education* 15, 3 (May 1984), 179. doi:10.2307/748348

[18] Stijn Ceuppens, Johan Deprez, Wim Dehaene, and Mieke De Cock. 2018. Design and Validation of a Test for Representational Fluency of 9th Grade Students in Physics and Mathematics: The Case of Linear Functions. *Physical Review Physics Education Research* 14, 2 (Aug. 2018), 020105. doi:10.1103/PhysRevPhysEducRes.14.020105

[19] Lijia Chen, Pingping Chen, and Zhijian Lin. 2020. Artificial Intelligence in Education: A Review. *IEEE access : practical innovations, open solutions* 8 (2020), 75264–75278. doi:10.1109/ACCESS.2020.2988510

[20] Neil Chulpongsatorn, Mille Skovhus Lunding, Nishan Soni, and Ryo Suzuki. 2023. Augmented Math: Authoring AR-Based Explorable Explanations by Augmenting Static Math Textbooks. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, San Francisco CA USA, 1–16. doi:10.1145/3586183.3606827

[21] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. doi:10.48550/ARXIV.2507.06261

[22] Valdemar Danry, Pat Pataranutaporn, Matthew Groh, and Ziv Epstein. 2025. Deceptive Explanations by Large Language Models Lead People to Change Their Beliefs About Misinformation More Often than Honest Explanations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–31. doi:10.1145/3706598.3713408

[23] Jimmy De La Torre. 2009. DINA Model and Parameter Estimation: A Didactic. *Journal of Educational and Behavioral Statistics* 34, 1 (March 2009), 115–130. doi:10.3102/1076998607309474

[24] Michael Desmond, Michael Muller, Zahra Ashktorab, Casey Dugan, Evelyn Duesterwald, Kristina Brimijoin, Catherine Finegan-Dollak, Michelle Brachman, Aabhas Sharma, Narendra Nath Joshi, and Qian Pan. 2021. Increasing the Speed and Accuracy of Data Labeling Through an AI Assisted Interface. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 392–401. doi:10.1145/3397481.3450698

[25] Shayan Doroudi and Emma Brunskill. 2019. Fairer but Not Fair Enough On the Equitability of Knowledge Tracing. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. ACM, Tempe AZ USA, 335–339. doi:10.1145/3303772.3303838

[26] Liam Franco Esparraguera, Kristoffer Selberg, Brian Lou, Jenny Sun, Beza Desta, Andrés Monroy-Hernández, and Parastoo Abtahi. 2024. Breaking the Plane: Exploring Real-Time Visualization of 3D Surfaces in Augmented Reality with Handwritten Input. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–9. doi:10.1145/3613905.3651032

[27] Haoxiang Fan, Guanzheng Chen, Xingbo Wang, and Zhenhui Peng. 2024. LessonPlanner: Assisting Novice Teachers to Prepare Pedagogy-Driven Lesson Plans with Large Language Models. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. ACM, Pittsburgh PA USA, 1–20. doi:10.1145/3654777.3676390

[28] Carmen Fariña, Phil Weinberg, Anna Commitante, and Linda Curtis-Bey. 2015. *Early Childhood Assessment in Mathematics Manual*. NYC Department of Education, 52 Chambers Street, New York, NY 10007.

[29] Yael Feldman-Maggor, Mutlu Cukurova, Carmel Kent, and Giora Alexandron. 2025. The Impact of Explainable AI on Teachers' Trust and Acceptance of AI EdTech Recommendations: The Power of Domain-specific Explanations. *International Journal of Artificial Intelligence in Education* 35, 5 (Dec. 2025), 2889–2922. doi:10.1007/s40593-025-00486-6

[30] Nicole L. Fonger. 2019. Meaningfulness in Representational Fluency: An Analytic Lens for Students' Creations, Interpretations, and Connections. *The Journal of Mathematical Behavior* 54 (June 2019), 100678. doi:10.1016/j.jmathb.2018.10.003

[31] Karen C. Fuson, Diana Wearne, James C. Hiebert, Hanlie G. Murray, Pieter G. Human, Alwyn I. Olivier, Thomas P. Carpenter, and Elizabeth Fennema. 1997.

Children's Conceptual Structures for Multidigit Numbers and Methods of Multidigit Addition and Subtraction. *Journal for Research in Mathematics Education* 28, 2 (March 1997), 130. doi:10.2307/749759

[32] Robert Glaser, Naomi Chudowsky, and James W Pellegrino. 2001. *Knowing What Students Know: The Science and Design of Educational Assessment.* National Academies Press.

[33] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. Vol. 52. Elsevier, 139–183.

[34] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 3 (May 2015), 407–434. doi:10.1177/0018720814547570

[35] Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven. 2019. Co-Designing a Real-Time Classroom Orchestration Tool to Support Teacher–AI Complementarity. *Journal of Learning Analytics* 6, 2 (July 2019), 27–52. doi:10.18608/jla.2019.62.3

[36] Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven. 2019. Designing for Complementarity: Teacher and Student Needs for Orchestration Support in AI-Enhanced Classrooms. In *Artificial Intelligence in Education*, Seiji Isotani, Eva Millán, Amy Ogan, Peter Hastings, Bruce McLaren, and Rose Luckin (Eds.). Vol. 11625. Springer International Publishing, Cham, 157–171. doi:10.1007/978-3-030-23204-7_14

[37] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4198–4205. doi:10.18653/v1/2020.acl-main.386

[38] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (Dec. 2023), 1–38. doi:10.1145/3571730

[39] Seokbin Kang, Ekta Shokeen, Virginia L. Byrne, Leyla Norooz, Elizabeth Bonsignore, Caro Williams-Pierce, and Jon E. Froehlich. 2020. ARMath: Augmenting Everyday Life with Math Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–15. doi:10.1145/3313831.3376252

[40] Alexander John Karran, Patrick Charland, Joé Trempe-Martineau, Ana Ortiz de Guinea Lopez de Arana, Anne-Marie Lesage, Sylvain Senecal, and Pierre-Majorique Leger. 2025. Multi-Stakeholder Perspective on Responsible Artificial Intelligence and Acceptability in Education. *npj Science of Learning* 10, 1 (2025), 44. doi:10.1038/s41539-025-00333-2

[41] Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. CodeAid: Evaluating a Classroom Deployment of an LLM-based Programming Assistant That Balances Student and Educator Needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–20. doi:10.1145/3613904.3642773

[42] Jinhee Kim. 2024. Types of Teacher-AI Collaboration in K-12 Classroom Instruction: Chinese Teachers' Perspective. *Education and Information Technologies* 29, 13 (Sept. 2024), 17433–17465. doi:10.1007/s10639-024-12523-3

[43] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *The 2024 ACM Conference on Fairness Accountability and Transparency*. ACM, Rio de Janeiro Brazil, 822–835. doi:10.1145/3630106.3658941

[44] Klein, Moon, and Hoffman. 2006. Making Sense of Sensemaking 2: A Macrocognitive Model. *IEEE Intelligent Systems* 21, 5 (Oct. 2006), 88–92. doi:10.1109/MIS.2006.100

[45] Laura Koesten, Kathleen Gregory, Paul Groth, and Elena Simperl. 2021. Talking Datasets – Understanding Data Sensemaking Behaviours. *International Journal of Human-Computer Studies* 146 (Feb. 2021), 102562. doi:10.1016/j.ijhcs.2020.102562

[46] George Lakoff and Rafael Núñez. 2000. *Where Mathematics Comes From.* Vol. 6. New York: Basic Books.

[47] Jimin Lee, Steven-Shine Chen, and Paul Pu Liang. 2025. Interactive Sketchpad: A Multimodal Tutoring System for Collaborative, Visual Problem-Solving. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. doi:10.1145/3706599.3719790

[48] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–35. doi:10.1145/3359283

[49] Yew Hoong Leong, Weng Kin Ho, and Lu Pien Cheng. 2015. Concrete-Pictorial-Abstract: Surveying Its Origins and Charting Its Future. *The Mathematics Educator* 16, 1 (2015), 1–18.

[50] Richard Lesh, Thomas R Post, and Merlyn Behr. 1987. Representations and Translations among Representations in Mathematics Learning and Problem Solving. In *Problems of Representations in the Teaching and Learning of Mathematics*. Lawrence Erlbaum, 33–40.

[51] Jiachen Li, Xiwen Li, Justin Steinberg, Akshat Choube, Bingsheng Yao, Xuhai Xu, Dakuo Wang, Elizabeth Mynatt, and Varun Mishra. 2025. Vital Insight: Assisting Experts' Context-Driven Sensemaking of Multi-modal Personal Tracking Data Using Visualization and Human-in-the-Loop LLM. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 3 (Sept. 2025), 1–37. doi:10.1145/3749508

[52] Jingshu Li, Yitian Yang, Q. Vera Liao, Junti Zhang, and Yi-Chieh Lee. 2025. As Confidence Aligns: Understanding the Effect of AI Confidence on Human Self-confidence in Human-AI Decision Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. doi:10.1145/3706598.3713336

[53] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–15. doi:10.1145/3313831.3376590

[54] Q. Vera Liao and Jennifer Wortman Vaughan. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review* Special Issue 5 (Feb. 2024), 1–53. doi:10.1162/99608f92.8036d03b

[55] Pi-Jen Lin and Wen-Huan Tsai. 2016. Enhancing Students' Mathematical Conjecturing and Justification in Third-Grade Classrooms: The Sum of Even/Odd Numbers. *Journal of Mathematics Education* 9, 1 (2016), 1–15.

[56] Ziyi Liu, Zhengzhe Zhu, Lijun Zhu, Enze Jiang, Xiyun Hu, Kylie A Peppler, and Karthik Ramani. 2024. ClassMeta: Designing Interactive Virtual Classmate to Promote VR Classroom Participation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–17. doi:10.1145/3613904.3642947

[57] Yu Lu, Yang Pian, Penghe Chen, Qinggang Meng, and Yunbo Cao. 2021. RadarMath: An Intelligent Tutoring System for Math Education. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 18 (May 2021), 16087–16090. doi:10.1609/aaai.v35i18.18020

[58] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in neural information processing systems* 30 (2017).

[59] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2025. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–23. doi:10.1145/3706598.3713423

[60] E. Maris. 1999. Estimating Multiple Classification Latent Class Models. *Psychometrika* 64, 2 (June 1999), 187–212. doi:10.1007/BF02294535

[61] Kelly McConvey, Shion Guha, and Anastasia Kuzminykh. 2023. A Human-Centered Review of Algorithms in Decision-Making in Higher Education. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15. doi:10.1145/3544548.3580658

[62] Alistair McIntosh, Barbara J Reys, and Robert E Reys. 2005. A Proposed Framework for Examining Basic Number Sense. *Subject Learning in the Primary Curriculum* (2005), 209–221.

[63] Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing Conversational Agents' Overconfidence through Linguistic Calibration. *Transactions of the Association for Computational Linguistics* 10 (2022), 857–872. doi:10.1162/tacl_a_00494

[64] Robert J. Mislevy, Russell G. Almond, and Janice F. Lukas. 2003. A BRIEF INTRODUCTION TO EVIDENCE-CENTERED DESIGN. *ETS Research Report Series* 2003, 1 (June 2003). doi:10.1002/j.2333-8504.2003.tb01908.x

[65] Robert J. Mislevy, Linda S. Steinberg, and Russell G. Almond. 2003. Focus Article: On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research & Perspective* 1, 1 (Jan. 2003), 3–62. doi:10.1207/S15366359MEA0101_02

[66] Mitchell J Nathan, Ana C Stephens, DK Masarik, Martha W Alibali, and Kenneth R Koedinger. 2002. Representational Fluency in Middle School: A Classroom Study. In *Proceedings of the Twenty-Fourth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, Vol. 1. ERIC Clearinghouse for Science, Mathematics and Environmental Education …, 462–472.

[67] Oi-Lam Ng. 2016. The Interplay between Language, Gestures, Dragging and Diagrams in Bilingual Learners' Mathematical Communications. *Educational Studies in Mathematics* 91, 3 (2016), 307–326. doi:10.1007/s10649-015-9652-9

[68] Erin R Ottmar, Candace Walkington, Dor Abrahamson, Mitchell J Nathan, Avery Harrison, and Carmen Smith. 2019. Embodied Mathematical Imagination and Cognition (EMIC) Working Group. *North American Chapter of the International Group for the Psychology of Mathematics Education* (2019).

[69] Rock Yuren Pang, Hope Schroeder, Kynnedy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–20. doi:10.1145/3706598.3713726

[70] Raja Parasuraman and Dietrich H. Manzey. 2010. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52, 3 (June 2010), 381–410. doi:10.1177/0018720810376055

[71] Samir Passi and Steven J Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–28. doi:10.1145/3274405

[72] Jean Piaget. 2013. *Child's Conception of Number: Selected Works Vol 2.* Routledge.

[73] Peter Pirolli and Stuart Card. 1999. Information Foraging. *Psychological Review* 106, 4 (Oct. 1999), 643–675. doi:10.1037/0033-295X.106.4.643

[74] Peter Pirolli and Stuart Card. 2005. The Sensemaking Process and Leverage Points for Analyst Technology as Identified through Cognitive Task Analysis. In *Proceedings of International Conference on Intelligence Analysis*, Vol. 5. McLean, VA, USA, 2–4.

[75] Stanislav Pozdniakov, Roberto Martinez-Maldonado, Yi-Shan Tsai, Vanessa Echeverria, Namrata Srivastava, and Dragan Gasevic. 2023. How Do Teachers Use Dashboards Enhanced with Data Storytelling Elements According to Their Data Visualisation Literacy Skills?. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. ACM, Arlington TX USA, 89–99. doi:10.1145/3576050.3576063

[76] Safâ Qetrani and Naceur Achtaich. 2022. Evaluation of Procedural and Conceptual Knowledge of Mathematical Functions: A Case Study from Morocco. *Journal on Mathematics Education* 13, 2 (May 2022), 211–238. doi:10.22342/jme.v13i2.pp211-238

[77] Sri RahayuniNgsiH, Sirajuddin SiRajuddiN, and Muhammad Ikram. 2021. Using Open-ended Problem-solving Tests to Identify Students' Mathematical Creative Thinking Ability. *Participatory Educational Research* 8, 3 (Aug. 2021), 285–299. doi:10.17275/per.21.66.8.3

[78] Prerna Ravi, John Masla, Gisella Kakoti, Grace C. Lin, Emma Anderson, Matt Taylor, Anastasia K. Ostrowski, Cynthia Breazeal, Eric Klopfer, and Hal Abelson. 2025. Co-Designing Large Language Model Tools for Project-Based Learning with K12 Educators. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–25. doi:10.1145/3706598.3713971

[79] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco California USA, 1135–1144. doi:10.1145/2939672.2939778

[80] Bethany Rittle-Johnson and Kenneth R. Koedinger. 2005. Designing Knowledge Scaffolds to Support Mathematical Problem Solving. *Cognition and Instruction* 23, 3 (Sept. 2005), 313–349. doi:10.1207/s1532690xci2303_1

[81] L M Rizki and N Priatna. 2019. Mathematical Literacy as the 21st Century Skill. *Journal of Physics: Conference Series* 1157, 4 (Feb. 2019), 042088. doi:10.1088/1742-6596/1157/4/042088

[82] D. Royce Sadler. 1989. Formative Assessment and the Design of Instructional Systems. *Instructional Science* 18, 2 (June 1989), 119–144. doi:10.1007/BF00117714

[83] Nazmus Saquib, Rubaiat Habib Kazi, Li-yi Wei, Gloria Mark, and Deb Roy. 2021. Constructing Embodied Algebra by Sketching. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. doi:10.1145/3411764.3445460

[84] Advait Sarkar. 2024. Large Language Models Cannot Explain Themselves. In *Workshop on Human-Centered Explainable AI (HCXAI) at the CHI Conference on Human Factors in Computing Systems (CHI '24)*. arXiv, Honolulu HI USA. doi:10.48550/ARXIV.2405.04382

[85] Md Kamruzzaman Sarker, Lu Zhou, Aaron Eberhart, and Pascal Hitzler. 2022. Neuro-Symbolic Artificial Intelligence: Current Trends. *Ai Communications* 34, 3 (2022), 197–209. doi:10.3233/AIC-210084

[86] Alan H. Schoenfeld and Mathematical Sciences Research Institute. 2007. What Is Mathematical Proficiency and How Can It Be Assessed? In *Assessing Mathematical Proficiency*, Alan H.Editor Schoenfeld (Ed.). Cambridge University Press, 59–74.

[87] Amanda Seccia and Susan Goldin-Meadow. 2024. Gestures Can Help Children Learn Mathematics: How Researchers Can Work with Teachers to Make Gesture Studies Applicable to Classrooms. *Philosophical Transactions B* 379, 1911 (2024), 20230156. doi:10.1098/rstb.2023.0156

[88] Yiming Shan. 2020. Enrolment Expansion in China: The Large Class Phenomenon. *Open Journal of Social Sciences* 8, 8 (2020), 1–13. doi:10.4236/jss.2020.88001

[89] Zekai Shao, Siyu Yuan, Lin Gao, Yixuan He, Deqing Yang, and Siming Chen. 2025. Unlocking Scientific Concepts: How Effective Are LLM-Generated Analogies for Student Understanding and Classroom Practice?. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–19. doi:10.1145/3706598.3714313

[90] Valerie Shute and Matthew Ventura. 2013. *Stealth Assessment: Measuring and Supporting Learning in Video Games.* The mit press.

[91] Valerie J Shute. 2008. Focus on Formative Feedback. *Review of educational research* 78, 1 (2008), 153–189. doi:10.3102/0034654307313795

[92] Robert S Siegler. 1984. Strategy Choices in Addition and Subtraction: How Do Children Know What to Do? *Origins of cognitive skills* (1984).

[93] Xiaohang Tang, Sam Wong, Kevin Pu, Xi Chen, Yalong Yang, and Yan Chen. 2024. VizGroup: An AI-assisted Event-driven System for Collaborative Programming Learning Analytics. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. ACM, Pittsburgh PA USA, 1–22. doi:10.1145/3654777.3676347

[94] Jonathan L Templin and Robert A Henson. 2006. Measurement of Psychological Disorders Using Cognitive Diagnosis Models. *Psychological methods* 11, 3 (2006), 287. doi:10.1037/1082-989X.11.3.287

[95] Giovanni M Troiano, Michael Cassidy, Daniel Escobar Morales, Guillermo Pons, Amir Abdollahi, Gregorio Robles, Gillian Puttick, and Casper Harteveld. 2025. CT4ALL: Towards Putting Teachers in the Loop to Advance Automated Computational Thinking Metric Assessments in Game-Based Learning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–23. doi:10.1145/3706598.3713368

[96] Almar Van Der Stappen, Yunjie Liu, Jiangxue Xu, Xiaoyu Yu, Jingya Li, and Erik D. Van Der Spek. 2019. MathBuilder: A Collaborative AR Math Game for Elementary School Students. In *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. ACM, Barcelona Spain, 731–738. doi:10.1145/3341215.3356295

[97] Kurt VanLehn. 2011. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational psychologist* 46, 4 (2011), 197–221. doi:10.1080/00461520.2011.611369

[98] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–15. doi:10.1145/3290605.3300831

[99] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su. 2022. NeuralCD: A General Framework for Cognitive Diagnosis. *IEEE Transactions on Knowledge and Data Engineering* 35, 8 (2022), 8312–8327. doi:10.1109/TKDE.2022.3201037

[100] Shuhan Wang, Fang He, and Erik Andersen. 2017. A Unified Framework for Knowledge Assessment and Progression Analysis and Design. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 937–948. doi:10.1145/3025453.3025841

[101] Dylan Wiliam. 2014. Formative Assessment and Contingency in the Regulation of Learning Processes. In *Annual Meeting of American Educational Research Association*. Philadelphia, PA, 1–13.

[102] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–22. doi:10.1145/3491102.3517582

[103] Yunyi Wu, Xinyun Cao, Mark Nielsen, Yichen Mao, and Fuxing Wang. 2024. Dragging but Not Tapping Promotes Preschoolers' Numerical Estimating with Touchscreens. *Journal of Experimental Child Psychology* 246 (2024), 105989. doi:10.1016/j.jecp.2024.105989

A Participants

A.1 Student Participants in Formative Study Part 1

We recruited 14 students for the first formative study. Their demographic details are provided in Table 1. Specific school contexts (e.g., public vs. private) were not recorded for these participants.

A.2 Teacher Participants in Formative Study Part 2

We recruited 5 teachers, all adhering to China's National Curriculum Standards. Their professional backgrounds are detailed in Table 2.

A.3 Teacher Participants in User Study

We recruited 20 teachers with diverse professional backgrounds, all of whom teach in accordance with China's National Curriculum Standards. Detailed profiles, including technical backgrounds, are provided in Table 3. For specific details regarding the scales used in the table, please refer to Appendix E.1.

Table 1: Demographics of Student Participants in the Formative Study.

ID	Grade	Gender	Age	ID	Grade	Gender	Age
S1	1	Female	6y8m	S8	2	Female	8y3m
S2	2	Female	7y8m	S9	K	Male	5y7m
S3	K	Male	4y8m	S10	2	Female	7y11m
S4	1	Male	7y4m	S11	K	Female	6y4m
S5	1	Male	6y7m	S12	K	Female	6y1m
S6	2	Male	7y6m	S13	1	Female	6y9m
S7	1	Male	6y10m	S14	1	Female	6y8m

Table 2: Profiles of Teacher Participants in the Formative Study.

ID	Teaching Exp. (years)	Grade 1 Exp. (years)	Gender	Remarks
T1	4	2	Female	
T2	8	1	Male	
T3	9	5	Female	
T4	9	4	Female	Private school
T5	13	7	Female	Private school

Table 3: Profiles of Teacher Participants in the User Study.

ID	Teaching Exp. (years)	Grades 1–2 Exp. (years)	Gender	Remarks	Open-ended Tasks	Cognitive Graph	Attitude toward AI	EdTech
T1	30	6	Male	Also Vice Principal	6	5	5	7
T2	13	7	Female	Private school A; Formative study T5	3	7	6	5
T3	35	5	Female	Retired	6	7	7	6
T4	3	3	Female		4	4	7	6
T5	4	2	Female	School B	7	6	7	7
T6	12	3	Female	School B	7	4	6	7
T7	4	4	Female		6	2	7	7
T8	20	2	Female	Private school	7	7	7	5
T9	8	3	Male	School C; Formative study T2	7	5	4	5
T10	6	2	Male	Tutoring institution	6	3	7	7
T11	1	1	Male	School C	5	5	6	6
T12	6	4	Male	Private school D	4	5	6	7
T13	5	2	Female		6	6	6	7
T14	9	4	Female	Private school D	3	5	6	6
T15	8	1	Female	Private school A	1	6	6	7
T16	9	3	Male	Tutoring institution	6	5	6	6
T17	1	1	Female		6	4	5	7
T18	2	0	Female	School B	2	5	6	7
T19	1	1	Female		3	4	7	7
T20	9	2	Female		5	4	7	7

B Open-ended Elicitation Tasks

B.1 Tasks in Formative Study

The following list presents details each task used in our formative study 1. For each task type, we specify the *Parameters* (e.g., specific numbers or quantities) used. Certain task types have initial settings, as shown in Figure 12.

Number Representation "How can you show the number N?"
Parameters (N): 9, 16, 24, 47, 358.

Quantity Comparison "Which are more, squares or circles?"
Parameters (pairs): 4 vs. 5, 7 vs. 7, 9 vs. 8, 12 vs. 13.

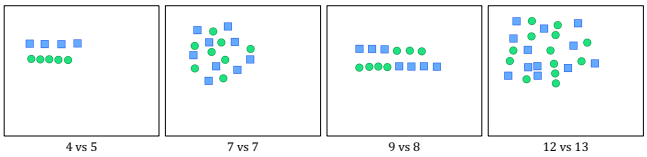
Odd/Even Judgement "Is the number of these objects odd or even?"
Parameters (count): 9, 14, 19, 16.

Addition "What is A+B? How do you calculate it?"
Parameters (A+B): 5+3, 6+8, 24+11, 38+17.

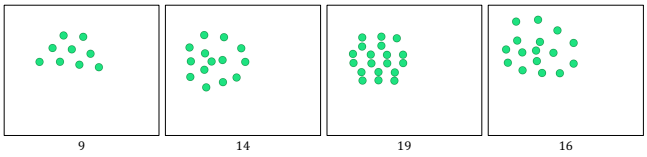
Subtraction "What is A-B? How do you calculate it?"
Parameters (A-B): 7-4, 13-8, 38-22, 43-26.

Addition Composition "What numbers add up to N?"
Parameters (N): 8, 14, 31.

Subtraction Decomposition "What numbers subtract to N?"
Parameters (N): 3, 7, 18.



(a) Tasks for Quantity Comparison



(b) Tasks for Odd/Even Judgement

Figure 12: Initial on-screen object arrangements for (a) the Quantity Comparison tasks and (b) the Odd/Even Judgement tasks.

To elicit deeper explanations or alternative strategies, we used a set of predefined follow-up questions based on the participant's previous response, including

- "Can you try using a different method?"
- "Can you try using objects?"
- "Can you try drawing?"
- "Can you try using objects or drawing?"
- "Can you write how you calculate?"
- "Can you explain your thinking?"
- "Can you find more answers?" (Primarily used for composition/decomposition tasks)

B.2 Tasks in Evaluation

The following list presents details each task used in our second round of student response data collection (for our expert review and user study). For each task type, we specify the parameters used.

Number Representation "How can you show the number N?"
Parameters (N): 17, 46, 358.

Addition with Regrouping "Can you show me how you calculate A+B?"
Parameters (A+B): 6+8, 15+7, 27+14.

Subtraction with Regrouping "Can you show me how you calculate A-B?"
Parameters (A-B): 13-7, 24-15, 45-28.

We used the following set of predefined follow-up questions:

- "Can you try using objects?"
- "Can you try drawing?"
- "Can you try using objects or drawing?"
- "Can you try writing?"
- "Can you explain your thinking?"
- "Can you explain your calculation?"

C Study Method Details

C.1 Student Response Platform

We developed a custom web-based data collection platform using a Vue 3 frontend and a Python backend, deployed on a 13-inch iPad with Apple Pencils. As shown in Figure 13, the interface supports multimodal interactions corresponding to three representational modes:

- **Enactive Mode:** Virtual manipulation of objects such as circles and squares from the side panel.
- **Iconic & Symbolic Modes:** Free-form drawing and writing using the Apple Pencil.

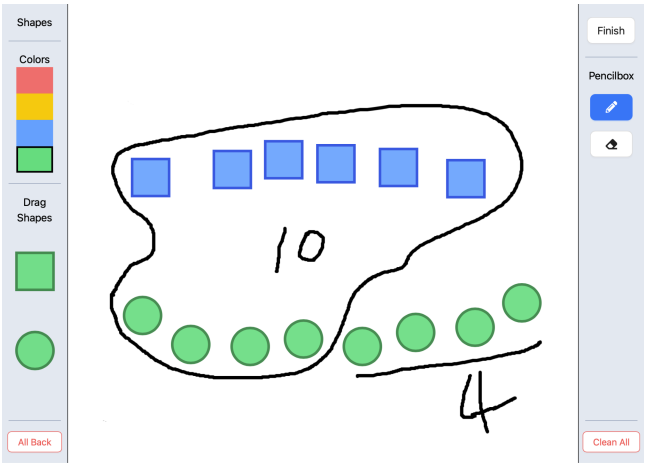


Figure 13: The student interface. The central area is the canvas for responses. Students can drag objects from the left panel or return them, and select a pen or eraser from the right panel.

The system captures synchronized audio and granular on-screen interaction logs:

- **Object Manipulation:** Events including adding, moving, clicking, and returning objects are logged with specific object IDs, coordinates, and timestamps.
- **Pen Strokes:** Drawing, erasing, and clear-all actions are recorded with stroke trajectories and timestamps.

C.2 Analysis in Formative Study Part 1

We conducted the following two-stage analysis, which is methodologically grounded in [10]:

First, we coded the student responses to identify key behaviors—that is, "what students did and how they did it." Examples of such codes include "student used one-to-one correspondence to represent a number" and "student used different objects to represent tens and ones." We began with an a priori set of behavioral codes informed by foundational research on the developmental levels of multi-digit numbers [31], odd/even numbers [55], and strategies of addition and subtraction [7, 92]. One researcher then employed open coding during the analysis to incorporate new behaviors not covered by the initial theoretical framework. A second researcher then reviewed and refined the codes. Through this process, we identified 54 typical behaviors across the tasks. Subsequently, based on established literature, we inferred students' cognitive states, such as their developmental level and patterns of strategy choice. Crucially, during this process, the researchers maintained analytical memos to document ambiguities, interpretation challenges, and the specific reasoning steps taken to resolve them.

In the second stage, we conducted a meta-analysis of our own analytical process to inform the design of an automated system. The analytical memos and documented challenges from Stage 1 served as the primary data for this stage. This meta-analysis was formally guided by two key questions: (1) "What information was essential for accurately understanding a student's intent?" and (2) "What behaviors served as evidence for diagnosing cognitive states?". Two researchers iteratively categorized the documented information needs and the types of evidence used for diagnosis. We then synthesized these categories into the core analytical challenges (F1-1 to F1-3) that the system must address.

C.3 Review Procedure of Expert Evaluation

Apparatus. As shown in Figure 14, we adapted the *OpenCD* interface for the review task into two modes:

- *Scoring Mode:* The results of the AI analysis are hidden, with access to students responses by *Response Walkthrough*. Experts can rate the student's mastery level and evidence sufficiency.
- *Feedback Mode:* The interface displays a side-by-side comparison of the expert's ratings and the AI's diagnosis. The experts can label the reasonableness of the AI diagnosis, annotate specific errors and provide explanations if unreasonable.

In addition, we provided experts with access to all video clips of every student's response (containing screen activity and audio) via a cloud storage link. We also provided detailed rating criteria, and examples of diagnostic logic and ratings for two sample students to familiarize the experts with the task.

Procedure. We used following three-phase protocal:

Onboarding & Rating. The goal of this phase was to ensure that the experts formed their own independent judgements of students before viewing the AI results. This phase lasted around 4–5 hours and involved the following steps:

- Introduction to criteria: We introduced the rating task and provided detailed rubrics for rating mastery level and evidence sufficiency.
- Data familiarization: The experts conducted an initial review of all student responses and took preliminary notes on whether each student performed well or poorly for each task type.
- Logic and rubric calibration: Using the system's Scoring Mode, experts practiced rating the responses from the first two students for all three task types. They iteratively rated students and viewed the examples to ensure a consistent understanding of the diagnostic logic.
- Independent Rating: The experts proceeded to independently rate the remaining 14 students on 3 task types and take notes of evidence.

Independent Review. The experts then entered the system's Feedback Mode, where their initial ratings were displayed alongside the AI-generated diagnoses for comparison. This phase lasted around 3 hours. Experts evaluated the AI's results with the rationale and evidence, following these steps:

- Reasonableness labeling: Experts classified each AI diagnosis as either "Completely Reasonable" or "Partially Unreasonable".
- Error Specification: If labeled "Partially Unreasonable", experts further specified the error type by selecting from predefined types: mastery level being substantially overestimate, slightly overestimate, slightly underestimate, or substantially underestimate, and evidence sufficiency misjudged as sufficient or insufficient.
- Qualitative Justification: Based on the rationale and evidence given by AI, experts provided written or verbal feedback of error causes, such as missing, misjudging or misusing evidence.

Consensus Meeting. Finally, the two experts met to discuss all discrepancies in their independent reviews, including differences of reasonableness and error types and reached a final consensus for all evaluated items. This phase lasted around 1 hour.

C.4 Apparatus for User Study

We adapted the *OpenCD* interface into two experimental modes:

- *Manual Mode:* This mode provided only the students responses and their processes, as shown in Figure 15. The cognitive graph and its manual scoring tools were available as optional aids.
- *OpenCD-Assisted Mode:* This mode offered the full functionality, with students responses plus *OpenCD*'s diagnosis.

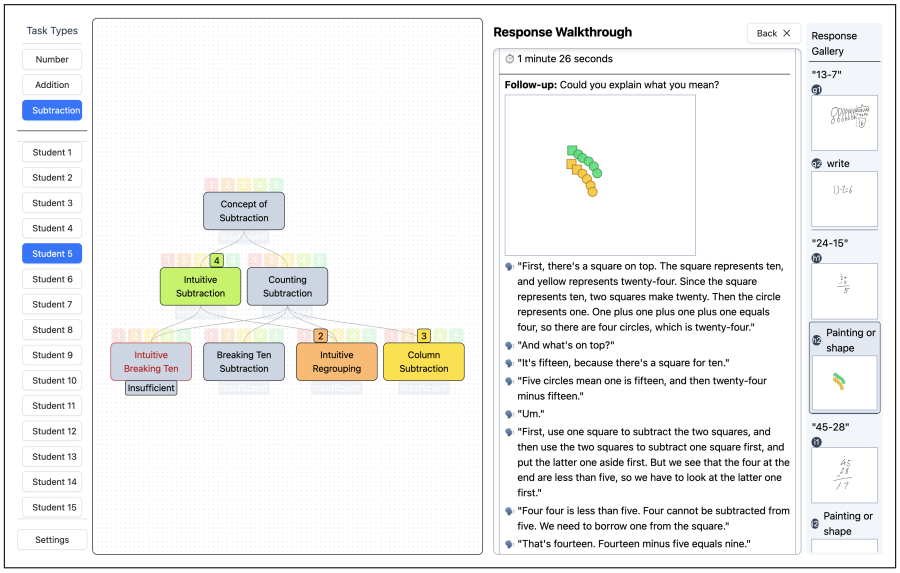
Each mode present data from the same 12 students. In the experiment, we randomized the student order and selected a balanced set of 12 students for each teacher.

Besides, we provided a Tutorial Mode, functionally identical to *OpenCD*-Assisted Mode but using data from the 4 remaining students. In addition, video clips of every student's response were accessible via a cloud storage link.

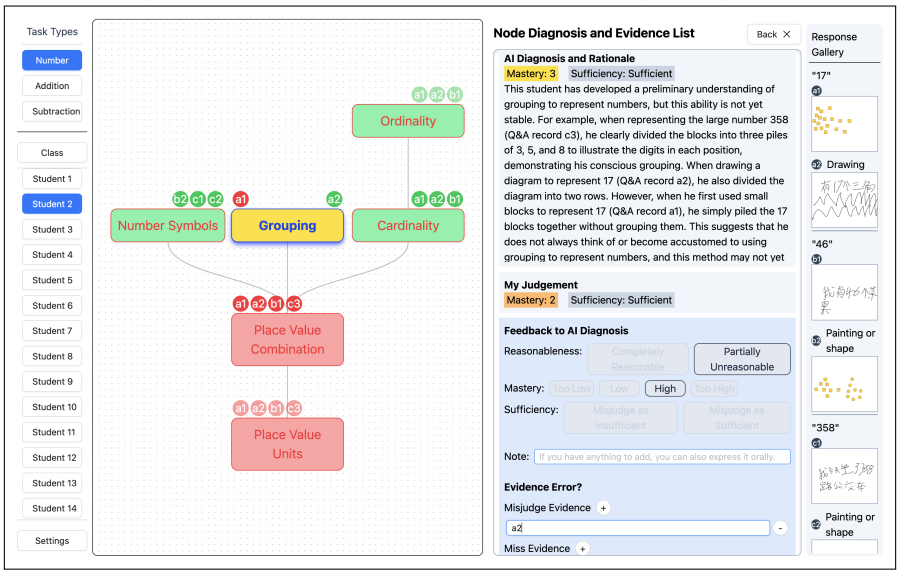
D System Implementation Details

D.1 Performance of VLM

To assess the VLM's performance, we examined the accuracy of the Response Comprehension agent using the student response data



(a) Scoring Mode for Expert Review (task Subtraction with Regrouping as example)



(b) Feedback Mode for Expert Review (task Number Representation as example)

Figure 14: The two modified user interface modes for the expert review study. (a) In Scoring Mode, experts rated student mastery based on the *Response Walkthrough*, with the AI's analysis hidden. (b) In Feedback Mode, experts saw the AI's diagnosis and provided feedback on its reasonableness.

collected during the formative study and identified several error types.

Out of the 299 responses collected in the formative study, the VLM accurately understood the students' visual outcomes, processes, and intentions in **93.6% (280/299) of cases**. Accuracy rates varied across different task types due to the diverse answering behaviors: 84.6% (44/52), 97.8% (45/46), 97.2% (35/36), 95% (57/60), 88.9% (40/45), 96.7% (29/30), and 100% (30/30) (following the order presented in Appendix B.1). Notably, the three types with relatively lower accuracy—*Number Representation*, *Addition*, and *Subtraction*—were associated with more diverse and complex student expressions. Therefore, these were the **three task types we selected for the subsequent Evaluation phase**.

Regarding the 19 cases of incorrect comprehension, we further identified the error types as follows:

- **Counting (9 cases):** The VLM miscounted the number of shapes or items in the image.

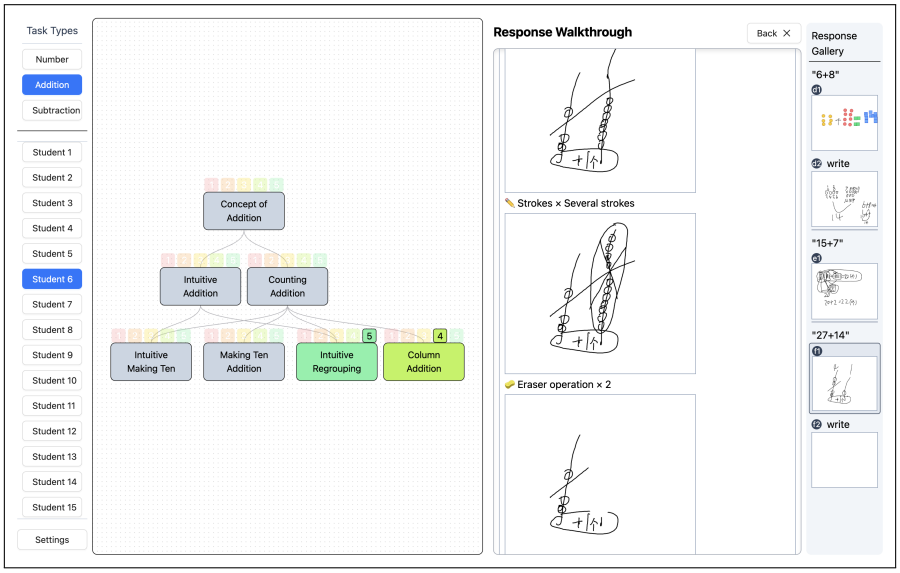


Figure 15: Manual Mode for User Study (Addition with Regrouping as example)

- **Intent Interpretation (5 cases):** The students' expressions were ambiguous and lacked verbal explanation. The VLM either failed to grasp the student's intent or hallucinated an intent.
- **Pattern Recognition (2 cases):** Students used shapes as pixels to construct a pattern, but the representation was unclear, and the VLM failed to recognize it.
- **Handwriting Recognition (1 case):** The VLM misidentified a "plus" sign as a Chinese character.
- **Layout Recognition (1 case):** The VLM incorrectly described the spatial layout of the shapes.
- **Process Hallucination (1 case):** The VLM hallucinated operational steps that the student did not actually perform.

Throughout our development process, consistent with Findings 1-2 from the formative study, we found that the process (descriptions of operations and intermediate canvas), the verbal explanations, and even prior responses significantly aided the VLM in understanding student intentions and visual outputs. (For example, while VLMs often struggle with counting, incorporating descriptions of the operation process—such as dragging to add shapes—effectively reduced counting errors.) Additionally, the context of the question itself helps the VLM understand intent and visuals; however, this can occasionally lead the VLM to over-rely on the question context. For instance, in cases where a student drew the wrong number of shapes, the VLM sometimes assumed the quantity matched the correct requirement of the question.

D.2 Encoding Multimodal Temporal Data

We use the Doubao API for ASR. Object manipulations are described textually, including quantities (e.g., "Add blue squares ×3"), to help the VLM count more accurately.

Screen images can be reconstructed from the interaction log. Due to the context window and attention limitations of VLMs, inputting the entire process frame-by-frame is infeasible. Therefore, we select a limited number of keyframes based on the following rules:

- The start and end frames of the response.
- Frames immediately before and after a major clear event (of either objects or pen strokes).
- Frames ever after a certain number of interactions have occurred.
- If a frame is blank, we use a textual note (e.g., "The canvas is clear") instead of an image.

D.3 Large Model Configuration

Both the VLM and the LLM Agent use the Gemini-2.5-Pro model. We set the temperature to 0 for deterministic outputs. The prompts are designed using a zero-shot Chain-of-Thought (CoT) approach, and the model is accessed via its API.

E Questionnaires

E.1 Teacher Background and Technical Proficiency

Ratings were collected on a 7-point Likert scale (1 = *Very Low / Negative*, 7 = *Very High / Positive*).

(1) **Familiarity with Open-ended Tasks:** Before this study, how familiar were you with assessment formats involving open-ended multimodal tasks (e.g., drawing, object manipulation)?
(2) **Familiarity with Cognitive Graphs:** Before this study, how familiar were you with the concept of Knowledge/Cognitive Graphs, including the developmental prerequisite rules (e.g., mastering advanced concepts requires mastering foundational ones)?
(3) **Attitude toward AI:** What is your general attitude toward the application of AI in education?

(4) **Frequency of EdTech Use:** How frequently do you use educational technology tools or digital teaching platforms in your daily instruction?

E.2 NASA-TLX

1 for *very low*, 7 for *very high*

(1) Mental Demand: What was the level of mental load for thinking, decision-making, or memory?
(2) Physical Demand: What was the level of physical demand experienced during the task?
(3) Effort: How much effort did you expend to achieve the current level of performance in the analysis report?
(4) Temporal Demand: What was the level of time pressure you experienced? (e.g., feeling rushed versus feeling unhurried with time for reflection).
(5) Frustration Level: How much frustration, irritability, or stress did you feel during the task?
(6) Performance: How successful and satisfied were you with your task performance?

E.3 Self-Report Performance

1 for *strongly disagree*, 7 for *strongly agree*

(1) **Confidence** in Analysis: I am very confident in the correctness of my conclusions from the learning analysis.
(2) **Class** Performance: I have comprehensively and deeply understood the overall class performance.
(3) **Individual** Learning Profiles: I have deeply understood the situations of representative students with stronger and weaker performance.
(4) **Unexpected** Answers: I have clearly understood the reasons behind some unexpected responses.

E.4 System Usability and Preferences

Feature Usefulness Rating: To what extent did the following system functionalities contribute to your understanding and analysis of student learning? (1 = *Not helpful at all*, 7 = *Very helpful*)

Feature Preference: Which system functionalities do you value most? (Select up to 4 out of the 10 features).

E.5 Perceived Impact and Usage Intention

1 for *strongly disagree*, 7 for *strongly agree*

(1) Enhance Insight: Without this AI system, I might miss important insights about student learning.
(2) Professional Growth: Sustained use of this system would enhance my own student learning analysis capabilities.
(3) Trust: I tend to trust the analytical conclusions provided by this AI system.
(4) Reliance: I rely on this AI system and would be largely influenced by its conclusions.
(5) Usage Intention: If available, I would use this AI system to assist in analyzing student learning profiles from open-ended, drawing-based tests.

F Interview Script

Comparing the Effects of Student Learning Analysis.

(1) What do you consider the most significant differences between the two modes of student learning analysis?
(2) In which aspects was the manual analysis less effective or satisfactory, and in which aspects did it perform well?
(3) Did the *OpenCD*-Assisted mode help you better understand students' learning progress? If yes, in what ways did it provide support?
(4) Were there aspects in which you felt the *OpenCD*-Assisted analysis was less effective, misleading, or even had negative side effects? If so, could you provide examples?
(5) How did the processes or workflows differ between the two modes?
(6) Did the *OpenCD*-Assisted mode help you identify details that might be overlooked in manual analysis? Were there any results you disagreed with or found misleading? If so, could you give specific examples?
(7) Did you find it necessary to revisit original video recordings? Based on the current presentation of students' responses (Response Walkthrough), is the provided information sufficient?

Evaluation of System Functionalities.

(1) How did the system features assist you? In what specific aspects do they play a role?
(2) What is your perspective on [system function]? Would its removal have a significant impact?

Impact, Applicability, and Future Prospects.

(1) Did you trust the diagnostic conclusions provided by the AI system, and why?
(2) Have you noticed any results from the AI analysis that you consider unreasonable or problematic?
(3) In what kinds of teaching activities or classroom scenarios do you see a demand for, and suitability of, such an AI analysis system? Can it be applied in other grade levels or knowledge domains of mathematics?
(4) Beyond teachers, do you think this system could be useful for other stakeholders?
(5) Did the AI analysis results support you in making instructional decisions or clarifying teaching directions?
(6) Has the system provided inspiration or insights that enhanced your teaching practice and analytical capacity?
(7) What limitations did you observe, and what suggestions would you propose for improvement?

G Knowledge Base of Expert Model

The expert model's Knowledge Base was developed to cover the 7 task types used in this study, informed by both academic literature and data from our first formative study. The full Knowledge Base includes detailed descriptions of each behavior and cognitive node. These descriptions are an essential part of the input provided to our VLM and LLM agents to guide their analysis. For brevity, the following content, which has been translated from the original Chinese, presents only the structure of the Knowledge Base and omits these detailed descriptions.

G.1 Typical Behaviors and Mapping Relations

```
Number Representation
Writing - Sequential Numbers (+)> Ordinality
Objects/Drawing - Unitary Quantity (+)> Cardinality
- Is Grouped: true/false (+/-)> Grouping
- Uses Grouping to Distinguish Place Values: true/false (+/-)> Place Value Combination
Drawing - Combination of Place Value Units (+)> Place Value Units
- Method of Distinguishing Place Values: "Grouping", "Size", "Shape", "Positional" (+)> Place Value Units
Objects - Combination of Place Value Units (+)> Place Value Units
Writing - Arabic Numeral (+)> Number Symbols
Objects - Numeral Shape Construction (+)> Number Symbols
Writing - Decomposition into Addition (+)> Grouping
- Is Decomposed by Place Value: true/false (+/-)> Place Value Combination
Objects/Drawing - Concatenating Digits (-)> Place Value Combination
Objects/Drawing/Writing - Object times Number (+)> Cardinality
Unable to Perform (Number Representation) (-)> Place Value Units

Quantity Comparison
Comparison by Counting (+)> Number Magnitude Comparison
- Compared Correctly: false (-)> Number Magnitude Comparison
Comparison by One-to-One Correspondence (+)> Intuitive Quantity Comparison
Unable to Perform (Quantity Comparison) (-)> Number Magnitude Comparison

Odd/Even Judgement
Judging Parity by Counting (+)> Odd/Even Number Sets
- Judged Correctly: false (-)> Odd/Even Number Sets
Judging Parity by Pairing (+)> Intuitive Parity
Judging Parity by Grouping and Combining (+)> Parity Operation Rules
- Parity Arithmetic Correct: false (-)> Parity Operation Rules
Unable to Perform (Judging Parity) (-)> Odd/Even Number Sets

Addition (with Regrouping)
Objects/Drawing - Expressing Addition as 'Adding To' (+)> Intuitive Addition
Objects/Drawing - Expressing Addition as 'Combining' (+)> Intuitive Addition
Objects/Drawing - Representing Addends and Sum Separately (+)> Intuitive Addition
Objects - Addition Equation Shape Construction (-)> Intuitive Addition
Writing - Addition Equation (+)> Concept of Addition
Objects/Drawing/Writing - Object times Number and Add (+)> Concept of Addition
Unable to Perform (Intuitive Representation of Addition) (-)> Intuitive Addition
Counting Addition (+)> Counting Addition
- Calculated Correctly: false (-)> Counting Addition
Memorized Addition (+)> Counting Addition
- Calculated Correctly: false (-)> Counting Addition
Making Ten Strategy (+)> Making Ten Addition
- Calculated Correctly: false (-)> Making Ten Addition
Place Value Decomposition Addition (+)> Column Addition
- Calculated Correctly: false (-)> Column Addition
Column Addition (+)> Column Addition
- Calculated Correctly: false (-)> Column Addition
Unable to Perform (Addition Calculation) (-)> Column Addition
Objects/Drawing - Composing a Ten (+)> Intuitive Making Ten
Objects/Drawing - Direct Addition, No Making Ten Shown (-)> Intuitive Making Ten
Objects/Drawing - Direct Addition, No Regrouping Shown (-)> Intuitive Regrouping
Objects/Drawing - Regrouping by Bundling (+)> Intuitive Regrouping
Objects/Drawing - Regrouping by Exchanging Ones for Ten (+)> Intuitive Regrouping

Subtraction (with Regrouping)
Objects/Drawing - Expressing Subtraction as 'Taking Away' (+)> Intuitive Subtraction
Objects/Drawing - Expressing Subtraction as 'Comparing' (+)> Intuitive Subtraction
Objects/Drawing - Representing Minuend, Subtrahend, and Difference Separately (+)> Intuitive Subtraction
Objects - Subtraction Equation Shape Construction (-)> Intuitive Subtraction
Writing - Subtraction Equation (+)> Concept of Subtraction
Objects/Drawing/Writing - Object times Number and Subtract (+)> Concept of Subtraction
Unable to Perform (Intuitive Representation of Subtraction) (-)> Intuitive Subtraction
Counting Subtraction (+)> Counting Subtraction
- Calculated Correctly: false (-)> Counting Subtraction
Memorized Subtraction (+)> Counting Subtraction
- Calculated Correctly: false (-)> Counting Subtraction
Breaking Ten Strategy (+)> Breaking Ten Subtraction
- Calculated Correctly: false (-)> Breaking Ten Subtraction
Subtracting to Ten Strategy (+)> Breaking Ten Subtraction
- Calculated Correctly: false (-)> Breaking Ten Subtraction
Column Subtraction (+)> Column Subtraction
- Calculated Correctly: false (-)> Column Subtraction
Think-Addition Strategy (+)> Counting Subtraction
Place Value Decomposition Subtraction (+)> Column Subtraction
- Calculated Correctly: false (-)> Column Subtraction
Unable to Perform (Subtraction Calculation) (-)> Column Subtraction
Objects/Drawing - Demonstrating Breaking Ten (+)> Intuitive Breaking Ten
Objects/Drawing - Demonstrating Subtracting to Ten (+)> Intuitive Breaking Ten
Objects/Drawing - Direct Subtraction, No Breaking Ten / Subtracting to Ten Shown (-)> Intuitive Breaking Ten
Objects/Drawing - Direct Subtraction, No Regrouping Shown (-)> Intuitive Regrouping (Subtraction)
Objects/Drawing - Expressing Regrouping by Unbundling (+)> Intuitive Regrouping (Subtraction)
Objects/Drawing - Expressing Regrouping by Trading Ten for Ones (+)> Intuitive Regrouping (Subtraction)

Additive Decomposition
Writing - Listing Additive Combinations (+)> Additive Decomposition
- Listed in Sequence: true/false (+/-)> Additive Compensation Principle
- Applied a Pattern to Generate More Results: true/false (+/-)> Additive Compensation Principle
Objects/Drawing - Representing Each Additive Decomposition Separately (-)> Intuitive Additive Compensation
Objects/Drawing - Showing Combinations by Regrouping Objects (+)> Intuitive Additive Compensation

Subtractive Decomposition
Writing - Listing Subtractive Combinations (+)> Subtractive Decomposition
- Listed in Sequence: true/false (+/-)> Constant Difference Principle
- Applied a Pattern to Generate More Results: true/false (+/-)> Constant Difference Principle
Objects/Drawing - Representing Each Subtractive Decomposition Separately (-)> Intuitive Constant Difference
Writing - Using Ellipsis to Show Infinity (+)> Infinity of Solutions
- Demonstrates Infinity: false (-)> Infinity of Solutions
Objects/Drawing - Showing Combinations by Adding/Removing Objects (+)> Intuitive Constant Difference
```

G.2 Cognitive Graph

```
Number Representation
Ordinality
Cardinality <- Ordinality
Grouping
Number Symbols
Place Value Combination <- Number Symbols, Grouping, Cardinality
Place Value Units <- Place Value Combination

Quantity Comparison
Concept of Quantity Comparison <- Cardinality
Number Magnitude Comparison <- Concept of Quantity Comparison
Intuitive Quantity Comparison <- Concept of Quantity Comparison

Odd / Even Judgement
Concept of Parity <- Cardinality
Odd/Even Number Sets <- Concept of Parity
Intuitive Parity <- Concept of Parity
Parity Operation Rules <- Grouping, Intuitive Parity

Addition (with Regrouping)
Concept of Addition
Intuitive Addition <- Concept of Addition
Counting Addition <- Concept of Addition
Making Ten Addition <- Counting Addition
Intuitive Making Ten <- Counting Addition, Intuitive Addition
Column Addition <- Counting Addition
Intuitive Regrouping <- Place Value Combination, Counting Addition, Intuitive Addition

Subtraction (with Regrouping)
Concept of Subtraction
Intuitive Subtraction <- Concept of Subtraction
Counting Subtraction <- Concept of Subtraction
Breaking Ten Subtraction <- Counting Subtraction
Intuitive Breaking Ten <- Counting Subtraction, Intuitive Subtraction
Column Subtraction <- Counting Subtraction
Intuitive Regrouping (Subtraction) <- Place Value Units, Counting Subtraction, Intuitive Subtraction

Additive Decomposition
Additive Decomposition <- Counting Addition
Additive Compensation Principle <- Additive Decomposition
Intuitive Additive Compensation <- Additive Compensation Principle

Subtractive Decomposition
Subtractive Decomposition <- Counting Subtraction
Constant Difference Principle <- Subtractive Decomposition
Intuitive Constant Difference <- Constant Difference Principle
Infinity of Solutions <- Constant Difference Principle
```

H Prompts

The prompts for our multi-agent system, presented below, were prompt-engineered and iteratively refined using the data collected in our first formative study.

Please note the following regarding these prompts:

- They have been translated from the original Chinese.
- Red text indicates placeholders for variables, abbreviations, or desired output.

H.1 Response Comprehension

```
You are given an open-ended math problem that can be answered using objects (pre-set squares and circles in four possible colors), strokes (for drawing or writing), and
       speech.
A student (around first grade) has answered this question. Your task is to understand the child's response process by analyzing the Response Process, which includes
      operations, speech recognition results (which may contain errors), and screenshots.

In the "Response Process":
    You will see the question, including the initial prompt and any follow-up questions.
    The log is formatted as follows:
        <emoji-pen> Stroke-related actions; can be 'writing' or 'drawing'.
        <emoji-sponge> Eraser-related actions; may indicate corrections or be used to demonstrate a dynamic process.
        <emoji-bricks> Object-related actions; includes adding, removing (returning), moving, clicking, etc.
        <emoji-speak> Speech content. Generally used to explain their actions or thoughts.
        Note that stroke and object actions may be followed by '×N' to indicate the number of operations. However, 'Stroke ×N' means N strokes were made, which does
              not necessarily equal the number of items drawn. You must use the screenshots to determine the actual images.
    The log also includes screenshots taken at the following times:
        At the beginning and end of the response;
        Before and after large-scale "remove object" or "eraser" operations;
        Periodically after a significant number of operations.

The current question type is: "%question_type%";

Additionally, identify the primary modalities the student used to answer (Objects, Drawing, Writing, Speech). You do not need to list every modality used, only the main
       ones. There can be more than one.

Finally, provide a detailed description of the student's response process. This description should summarize the "Response Process" and clearly explain how the student
      answered the question, such that a person who has not seen the log can understand the student's process from your description.
```

```
Notes:
    - There may be follow-up questions asking the student to clarify their meaning. In such cases, the student might use speech to explain their previous actions. Use
          their explanation to understand the response process, especially when the operational steps are long or complex and the details are not immediately clear from
           the actions alone. The student's spoken explanation is crucial here.
    - The student's use of the eraser could mean correction, but it could also be part of a dynamic demonstration. If a student erases objects or strokes, it might be
          to represent 'taking away' or to move an item, not necessarily to fix a mistake.
    - Although the Response Process is chronological, speech and actions may overlap in reality (i.e., the student speaks while performing an action). Therefore, the
          linear sequence in the log does not strictly correspond to the actual order, and a continuous series of actions might be interrupted by a speech entry. Do not
           over-interpret the sequential relationship between adjacent speech and action entries.
    - Screenshots are taken periodically after a certain number of actions and do not necessarily represent keyframes. Do not rely heavily on these intermediate
          screenshots to segment the student's response process. Focus on analyzing the sequence of operations and the speech content, using screenshots primarily to
          understand the content and layout of strokes and objects.
    - Do not over-speculate about the student's internal thoughts or mental state. Focus on describing the student's observable behaviors and response process: what
          they drew, wrote, said, and how they manipulated objects. Be reasonably detailed about the content shown in the screenshots.
    - You are not proficient at accurately counting the number of items in an image. When you need to output a quantity, you can refer to the problem description or in-
          progress object counts from the log. Since you may struggle to determine the exact number of hand-drawn items and identify counting errors, focus your
          analysis on the student's method and thinking process rather than on whether the quantity drawn is perfectly correct.
    - Do not hastily judge the student's response as incorrect. The provided screenshots are not a complete video record, and you cannot assume an action was performed
          incorrectly or not at all just because you didn't see it. Rely on the student's speech to infer their intent and actions. The speech may not be immediate; it
          might come after the response is complete.

Besides, some of the response or screenshots may not be easy to interpret, you can refer to prior Q&As to understand.
The prior questions and response summaries are:
%previous_QAs%

Response Process:
[text-and-image script of process]

When responding, use the following chain-of-thought steps. Please reply in the specified JSON format:
```json
{
    "Response Analysis": "First, review and analyze the student's operational steps one by one to clearly understand their response, but do not over-speculate about
          their internal thoughts or mental state. If there is clear speech, refer to the student's spoken explanation to understand their behavior and intent. (The
          student's process may not be perfectly neat, which could be due to the difficulty of using the tools rather than a lack of mathematical understanding. The
          screenshots you see are incomplete, and intermediate details may be missing; you should use speech to help understand the full process.)",
    "Primary Modalities": [...], // list[str], possible values are "Objects", "Drawing", "Writing", "Speech". List only the most central modalities used; there can be
          more than one, but try to list only the most essential ones.
    "Response Summary": "Describe and summarize the student's response process. Do not describe the question content; focus on the student's actions and how they
          answered the question. The final screenshot can be described in detail; intermediate screenshots are for your own understanding and do not need to be
          described deliberately."
}
```
```

H.2 Behavior Detection

```
You are given an open-ended math problem that can be answered using objects (pre-set squares and circles in four possible colors), strokes (for drawing or writing), and
       speech.
A student (around first grade) has answered this question. Your task is to detect the child's behavioral features by analyzing the response process, which includes
      operations, speech recognition results (which may contain errors), and screenshots.

In the "Response Process":
    ...Same as the prompt for Response Comprehension...

The "Behavioral Feature" format is as follows:
{
    "name": "Name of the behavioral feature",
    "description": "Description and criteria for the behavioral feature",
    "properties": [
        {
            "name": "Property name",
            "description": "Description and criteria for the property",
            "value_type": "The value type of the property, e.g., boolean, string, etc.",
        },
        ...
    ]
}
The name of a behavioral feature may follow the format 'Action Type - Specific Behavior,' where the action types include:
    Objects: Using pre-set squares or circles.
    Drawing: Using strokes to draw patterns.
    Writing: Using strokes to write numbers or symbols.
The action type can be a combination, such as 'Objects/Drawing,' indicating the behavior can occur with either action type.
'properties' may be an empty list or contain multiple properties.

"Behavioral features" will be provided to you as a dictionary of behavioral scopes, where each scope contains a list of features:
{
    "A Behavioral Scope": [
        {
            "name": "An Action Type - A Specific Behavior",
            ... // description, properties, etc.
        },
        ... // other behavioral features
    ],
    ... // other behavioral scopes and their features
}
Generally, at least one behavioral feature should be detected within each relevant behavioral scope.

When you detect a behavioral feature, you must:
    Provide a specific explanation of the behavior, such as what the child specifically did. This explanation is for the teacher to understand why this feature was
          identified.
    If the feature has properties, provide the value for each property, along with a specific explanation for that property.

Response Process:
[text-and-image script of process]

The understanding and summary of the response process above is as follows: <Response Summary>
%response_summary%
</Response Summary>

The dictionary of behavioral scopes (with their feature lists) is as follows: %behavior_list%
```

```
When responding, do not directly output the behavioral features. Instead, use the following chain-of-thought steps. Please reply in the specified JSON format:
```json
{
    "Behavioral Analysis": "Briefly analyze the student's response to determine which behavioral features exist within each behavioral scope. This includes identifying
           expected behaviors (based on the context, what was the student expected to do?) that were not performed. You should identify the most accurate behavioral
           feature. If a feature has properties, you must also analyze and determine their values.",
    "Behavioral Features": {
        // Iterate through each behavioral scope, including: %behavior_scopes%
        "A Behavioral Scope": [ // A list of behavioral features detected in this scope; can be one or more.
            {
                "name": "The name of the behavioral feature, which must exactly match a name from the provided 'Behavioral Features List'",
                "explanation": "A specific explanation of this behavior as performed by the child",
                "properties": [
                    {
                        "name": "Property name",
                        "explanation": "A specific explanation for this property of the child's behavior",
                        "value": boolean | string | ...
                    },
                    ...
                ]
            },
            ...
        ]
    }
}
```
```

H.3 Unclassified Mapping

```
You are given an open-ended math problem that can be answered using objects (pre-set squares and circles in four possible colors), strokes (for drawing or writing), and
       speech.
A student (around first grade) has answered this question. To analyze the student's response, a preliminary analysis has already been conducted, detecting several
      relevant behavioral features. However, some features did not match the standard list and were therefore labeled as "Other".
All standard behavioral features can be automatically mapped to a cognitive node as positive or negative evidence. However, "Other" features cannot be mapped directly.
      Your task is to further analyze these "Other" features and determine if they can serve as positive or negative evidence for any cognitive nodes.

In the "Response Process":
    ...Same as the prompt for Response Comprehension...

The understanding and summary of the response process above is as follows: <Response Summary>
%response summary%
</Response Summary>

The "Identified Behavioral Features" for this response are as follows:
%behavior_features%

All "Cognitive Nodes" and their associated "Standard Behavioral Features" are as follows:
%cognitions_connections%

You need to analyze the "Other" behavioral features from this response and determine if they can serve as positive or negative evidence for any cognitive nodes.
Try to understand the meaning of the behavior. A student using a special method could indicate several possibilities:
    The student used a non-standard, more advanced method, which could indicate mastery of a certain cognitive node.
    The student used a non-standard, more basic method, which could indicate mastery of a foundational cognitive node but not a more advanced one. (In this case, it
           might map to two cognitive nodes).
    The student has not mastered a standard method and therefore used a method incorrectly (making it erroneous, inefficient, or ineffective), which appears special.
           This could indicate a lack of mastery for the cognitive node associated with the standard method.

Note, you should output only the most relevant cognitive nodes, not all possible ones. Please try to limit the number of mapped cognitive nodes.
Regarding calculation methods, some are parallel alternatives. Therefore, using one calculation method does not necessarily imply a lack of mastery of another.
      Intuitively demonstrating one method does not necessarily imply a lack of intuitive understanding of another.

When responding, do not directly output the result. Instead, use the following chain-of-thought steps. Please reply in the specified JSON format:
```json
{
    "Feature Matching": "Carefully analyze the 'Other' behavioral feature to understand its meaning. Analyze which standard behavioral features it most closely
           resembles. (If a standard feature name is repeated or has specific properties, it may match several).",
    "Cognitive Mapping": "Based on the standard behavioral features, analyze the meaning of the related cognitive nodes and determine if the behavior can be mapped to
           mastery (or lack thereof) of a node. Unless the behavior is a very close match to a standard feature or node, do not map it to multiple cognitive nodes
           without strong justification.",
    "Behavior Feature": [ // Iterate through each 'Other' feature in the 'Identified Behavioral Features'
        {
            "outlier_behavior": "The name of the 'Other' behavioral feature, which must exactly match the name from the 'Identified Behavioral Features' input",
            "mappings": [ // This behavior may map to multiple cognitive nodes, but be conservative.
                {
                    "analysis": "Provide the analysis and justification for why this specific behavior indicates the mastery status of the cognitive node.",
                    "cognition": "The name of the cognitive node, which must exactly match a name from the provided 'Cognitive Nodes' list",
                    "positive": boolean // true for positive evidence (indicates mastery), false for negative evidence (indicates lack of mastery)
                }
                // Multiple mappings are possible, but should be used cautiously.
            ]
        }
    ]
}
```
```

H.4 Node Diagnosis

```
You are given an open-ended math assessment that includes multiple questions and interactions. Students can respond using objects (pre-set squares and circles in four
       possible colors), strokes (for drawing or writing), and speech. Through these tasks, we can conduct an in-depth analysis of a student's 'Mastery Level' of
       relevant cognitive nodes and the 'Evidence Sufficiency' (i.e., whether the student's actions provide enough evidence to support the mastery judgment).
A student (around first grade) has completed a series of these tasks.
Analyzing cognition for these open-ended responses requires inferring from behavioral features as evidence and making a judgment by considering the relationships
       between foundational and advanced cognitive nodes.
Each interaction in the assessment has already undergone a simple, individual analysis: behavioral features were identified and preliminarily mapped as positive or
       negative evidence to related cognitive nodes. However, a comprehensive analysis that synthesizes across multiple questions and interactions to determine the final
        Mastery Level and Evidence Sufficiency has not yet been performed. This is your task. Note: this assessment and scoring is intended for teachers, not students.
```

```
The current question type is: "%question_type%"
The cognitive node you need to analyze is: "%cognition_node%"
Its 'Meaning' is: %cognition_description%
All 'Related Behavioral Features' for this node are (positive indicates mastery, negative indicates lack of mastery):
%related_behaviors%
You must judge its mastery level. Your analysis should focus on the meaning of this node and its related behaviors, rather than speculating based on the node's name.

For this node, '%cognition_node%':
More foundational cognitive nodes are: %basic_cognitions%,
More advanced cognitive nodes are: %advanced_cognitions%;
You can use this to understand the developmental relationship and whether this node is foundational or advanced.
There are also other parallel nodes that have less influence on this one.

The 'Records of Questions, Responses, and Behavioral Feature Analyses' for the current assessment are:
%analysis_records%

Within this question type, tasks often have varying parameters and increasing difficulty. The student's chosen method may change; they may not always use the most
       advanced method. You need to analyze this carefully when judging mastery:
    If the student uses a basic method correctly in simple scenarios and an advanced method correctly in complex scenarios, it may simply mean the advanced method was
           unnecessary for the easier tasks. This should have a minor impact on the mastery judgment.
    However, if the reverse is true-the student fails to use an advanced method in complex scenarios or uses methods incorrectly/inconsistently-it suggests a lack of
           proficiency or true understanding, which should have a major impact on the mastery judgment.

The preliminary analysis provides positive and negative evidence mapped from behavioral features to cognitive nodes. Because nodes are not independent but exist in a
       developmental hierarchy (from foundational to advanced), evidence propagates along this hierarchy.
Evidence is categorized as direct or indirect based on whether it is propagated from other nodes:
- Direct Evidence: A behavioral feature from a response that is directly related to this cognitive node.
- Indirect Evidence: A behavioral feature that maps to a different node, from which an inference is propagated to the current node. This includes negative evidence for
       a foundational node or positive evidence for an advanced node. (If a node is mastered, its foundational nodes are also considered mastered; if a node is not
       mastered, its advanced nodes are also considered not mastered.)
The current 'Evidence' is as follows:
<Evidence List>
Positive Evidence (Total: %positive_count%): %positive_evidences%
Negative Evidence (Total: %negative_count%): %negative_evidences%
</Evidence List>
You must consider all evidence synthetically and not base your judgment on isolated instances.
When making a judgment, consider the quantity, sign (positive/negative), and strength of the evidence (i.e., how typical the behavior is and how strongly it relates to
       the node):
- Strong:
    (Positive) The student's response is exemplary, the process is demonstrated meticulously, or the verbal and physical expressions are very clear.
    (Negative) A classic misconception is observed, or the student is completely unable to perform the task.
    The behavioral feature is strongly correlated with this cognitive node.
- Weak:
    (Positive) Uses a similar but non-standard method (so mastery of the standard method is uncertain); the process is not meticulous, or verbal/physical expressions
           are unclear (so it's uncertain if the process and understanding are fully correct).
    (Negative) The error is atypical or appears to be a careless mistake.
    The behavioral feature is weakly correlated with this cognitive node (e.g., not using an advanced method on a simple problem; using a method with a similar but not
           identical underlying principle).
- **The strength of evidence is not affected by whether it is direct or indirect.** You should treat them equally:
    Evidence pointing to mastery of an advanced node is also evidence for mastery of its foundational nodes.
    Evidence pointing to a lack of mastery of a foundational node is also evidence for a lack of mastery of its advanced nodes.
    The strength depends on how typically the student's response exhibits the relevant behavioral features.
When the evidence is too sparse or weak, the evidence sufficiency is low. You can refer to the 'Records of Questions, Responses, and Behavioral Feature Analyses,' the '
       Node Meaning,' and 'Related Behaviors' to further infer the mastery level. **However, in this case, the Evidence Sufficiency score must be low**.

In summary, your analysis must consider:
- The evidence in the Evidence List, including both direct and indirect evidence. You also need to consider the student's response process to judge how prominently the
       behavior was displayed.
- **If evidence is sparse, "Evidence Sufficiency must be low."** You should then use the full interaction records to further infer the mastery level, but in such cases,
        you should avoid giving extremely high or low mastery scores due to the lack of sufficient evidence.

Finally, based on a synthesis of all direct and indirect evidence, provide a conclusion for the current cognitive node, including 'Mastery Level' and 'Evidence
       Sufficiency'.
[Mastery Level] The student's level of mastery for this node. You need to synthetically consider the quantity, strength, sign, and consistency of the evidence, as well
       as the node's developmental position (foundational nodes are easier to master; advanced nodes are harder).
    The mastery level is a 5-point scale:
    5: Excellent Mastery. Skillfully and correctly uses methods corresponding to this node with clear and thorough explanations; or skillfully uses more advanced
           methods.
    4: Good Mastery. Generally able to correctly use methods corresponding to this node, but the process may lack clarity or be purely mechanical; may occasionally use
           less efficient methods.
    3: Fair Mastery. Sometimes able to correctly use methods corresponding to this node, but other times uses them incorrectly or inadequately.
    2: Poor Mastery. Only understands the general form or procedure but frequently uses the corresponding methods incorrectly.
    1: Very Poor Mastery. Unable to use methods corresponding to this cognitive node.
    When evidence is very insufficient, the mastery score, while mostly a guess, should generally not be high (implying the student is not comfortable with the skill).
           Your specific inference should still be based on the student's responses.
[Evidence Sufficiency] Whether the student's actions provide sufficient evidence (both direct and indirect) to judge the mastery level. Sufficiency affects the
       certainty/confidence of your judgment.
    You need to synthetically consider the quantity and strength of the evidence. You should treat the strength of direct and indirect evidence equally. The specific
           strength depends on how typically the student's behavior exhibits the relevant features.
    The evidence sufficiency is a 4-point scale:
    4: Very Sufficient. The evidence is abundant or very strong, providing excellent support for the mastery judgment.
    3: Relatively Sufficient. The evidence is reasonably strong or numerous (e.g., three or more instances, or a very typical behavior), providing good support for the
           mastery judgment.
    2: Relatively Insufficient. The evidence is sparse or weak (e.g., only one instance, or an atypical behavior), not enough to confidently determine the student's
           mastery level; the diagnosis has high uncertainty.
    1: Very Insufficient. There is no or extremely weak evidence, and the judgment is almost entirely a guess (e.g., the student always used other methods, or the tasks
            did not involve this cognitive node).
Afterward, you must provide a "Conclusion Explanation" for other math teachers. Please use language that is accessible and easy for other teachers to understand.

<Judgment Focus>
Additionally, the 'Judgment Focus' varies slightly for different types of cognitive nodes:
For 'Calculation Method' cognitive nodes:
    Focus on the correctness of the calculation process and the final result.
    If evidence is sparse because the student skillfully used a parallel calculation method, it may indicate a preference for the other method. In this case, the
           evidence sufficiency for this node is low.
For 'Intuitive Representation of Calculation' cognitive nodes:
    Focus on whether the student understands the underlying principle and the manner of their representation. Minor quantitative errors in the process or result do not
           significantly impact the mastery judgment for this type of node.
    The student does not need to demonstrate the process completely every time; one reasonably complete demonstration is often sufficient to infer a high mastery level.
    If the student primarily uses other calculation methods, they may lack the opportunity to demonstrate their intuitive understanding of this method, making it
           difficult to judge this node. The evidence sufficiency will be low.
    If the student frequently uses this calculation method but struggles to represent it intuitively, their mastery level for this intuitive node is low.
```

```
For other types of cognitive nodes, there is no special judgment focus.
</Judgment Focus>

When responding, do not directly output the result. Instead, use the following chain-of-thought steps to analyze the mastery of '%cognition_node%'. Please reply in the
      specified JSON format:
```json
{
    "Evidence Analysis": "Review the 'Evidence List', analyze the direct and indirect evidence, and judge the strength of each piece of evidence. Determine if the
          evidence is sufficient to support a judgment of mastery.",
    "Conclusion Analysis": "Restate the 'Meaning' of the current cognitive node and any relevant 'Judgment Focus'. Based primarily on the evidence list and secondarily
          on the related behavioral features in the response records, determine the mastery level. Judge the evidence sufficiency based on the quantity and strength of
          the evidence.",
    "Mastery Level": 1|2|3|4|5, // int, 1-5 scale
    "Evidence Sufficiency": 1|2|3|4, // int, 1-4 scale; if the 'Evidence List' is sparse, this level should be low"
    </>"Rationale"</>: "Briefly explain the judgment for the mastery level and evidence sufficiency, referencing the evidence. If mastery is low, analyze the student's
          specific difficulties. (When explaining, describe the levels with words, not scores. When mentioning an interaction, cite the 'Interaction Record ID' for
          teachers to reference. Use teacher-friendly, accessible language; avoid overly mechanical language or the jargon used in the prompts above.)"
}
```
```

H.5 Comprehensive Report

```
You are given an open-ended math assessment that includes multiple questions and interactions. Students can respond using objects (pre-set squares and circles in four
      possible colors), strokes (for drawing or writing), and speech. Through these tasks, we can conduct an in-depth analys of student's cognition and mastery,
      specifically their 'Mastery Level' and 'Evidence Sufficiency' (i.e., whether the student's actions provide enough evidence to support the mastery judgment) for
      various cognitive nodes.
A student (end of first grade) has completed a series of these tasks. An AI has already performed a multi-level analysis and diagnosis, identifying behavioral features
      as evidence, mapping them to cognitive nodes, and providing 'Mastery Level' and 'Evidence Sufficiency' judgments for each node.
Your task is to synthesize all of the student's interactions and the diagnoses of all cognitive nodes to generate a comprehensive diagnostic report. The report should
      mainly include: a summary of the student's performance, an analysis of their cognitive state, any special observations, and pedagogical suggestions.

The current question type is: "%question_type%"

The relevant cognitive graph for the current problem is:
%cognitive_graph%
The edges in the graph represent the developmental relationships between cognitive nodes, from foundational to advanced. Generally, mastery of a foundational node is a
      prerequisite for mastery of an advanced node. Likewise, positive evidence for an advanced node also serves as positive evidence for its foundational nodes;
      conversely, negative evidence for a foundational node also serves as negative evidence for its advanced nodes.

The 'Records of Interactions and Behavioral Features' for the current assessment are:
%analysis_records%

The student's 'Cognitive Node Diagnosis' is as follows:
%cog_diagnosis%
Where:
    Mastery Level is a 5-point scale: 5: Excellent Mastery; 4: Good Mastery; 3: Fair Mastery; 2: Poor Mastery; 1: Very Poor Mastery.
    Evidence Sufficiency is a 4-point scale: 4: Very Sufficient; 3: Relatively Sufficient; 2: Relatively Insufficient; 1: Very Insufficient.
    When evidence is insufficient, the mastery level is largely an inference and has high uncertainty; further assessment is needed for an accurate judgment.

The diagnostic report you generate should be as concise as possible, focusing on the information most valuable to teachers. Teachers have large classes and limited time
      to focus on each student.
If the student performs well, be brief. If the student struggles, do not try to find irrelevant strengths for awkward praise. The report is for teachers, who need to
      understand the actual problems and receive practical advice, not empty compliments.
Some standard, high-quality performances (like using sticks in bundles or an abacus) may be the result of specific training from the teacher. Do not over-praise these
      common behaviors as if they were creatively invented by the student.
Summary of Performance:
    Based on the 'Records of Interactions and Behavioral Features,' concisely summarize their performance.
    Do not list every single interaction; instead, extract and generalize typical and important behaviors.
Cognitive Diagnosis:
    Based on the 'Cognitive Node Diagnosis,' the 'Cognitive Graph,' and typical behaviors, concisely analyze their cognitive state.
    Do not list the mastery status of every single node. Instead, focus on the nodes where mastery is weak. Use the structure of the cognitive graph and its
          developmental relationships to analyze the reasons and struggles for these weaknesses.
        Low mastery does not always mean a lack of understanding; it could be due to limitations in expression. Frame weaknesses as potential gaps in instruction or
              familiarity with the task format, rather than definitively stating the student 'does not understand' or 'has not mastered' the concept.
        When describing a node's status, use fluid language instead of restating the numerical scores for mastery level and evidence sufficiency.
        If the student performed poorly, do not offer awkward or forced praise.
    Furthermore, you can touch upon higher-order competencies like 'number sense,' 'symbolic awareness,' 'abstract thinking,' and 'mathematical intuition.' However, do
          not force these connections; they must be genuinely linked to the student's actual performance. If these competencies are not demonstrated, do not mention
          them.

Special Observations:
    This section is for noting any performance that is beyond their current grade level, is anomalous, requires special attention, is difficult to understand, or
          relates to metacognition. If there are none, do not write anything.
    You can also mention personality traits if they are particularly relevant, but do not force it.
    Do not assume a method was invented by the student, as you don't know if they have received similar training.
    If there is nothing particularly special to report, output an empty string to save the teacher's time.
Pedagogical Suggestions:
    Based on the cognitive diagnosis, provide concrete, actionable suggestions for the teacher that target the student's areas of weakness.
    Be as concise as possible. These are suggestions, not directives, as the teacher is the educational expert.
    Do not suggest learning ahead. If the student has mastered all current topics well, you can suggest activities for broader development or encourage free exploration
          . (You do not know the teacher's current instructional progress).

When responding, use the following chain-of-thought steps. No reasoning is required in the output. Please reply in the specified JSON format:
```json
{
    "Report Logic": "First, state the logic of this diagnostic report and how you will organize it, including any unique characteristics of this student. Decide if the
          'Special Observations' section is needed; if there's nothing special, omit it to save the teacher's time.",
    "Response Overview": "Based on the 'Records of Interactions and Behavioral Features,' concisely summarize their performance.",
    "Cognitive Analysis": "Based on the 'Cognitive Node Diagnosis,' the 'Cognitive Graph,' and typical behaviors, concisely analyze their cognitive state. You may even
          mention relevant core competencies.",
    "Noteworthy Observations": "Be as concise as possible. If there is nothing particularly noteworthy, output an empty string to save the teacher's time. (Note: some
          standard, excellent performances might just be what was taught in class, not necessarily a sign of the child's brilliance; a method was likely taught, not
          invented).",
    "Pedagogical Suggestions": "Based on the cognitive diagnosis, provide concrete, actionable suggestions for the teacher that target the student's areas of weakness."
}
```
```