Conversational Learning Diagnosis via Reasoning Multi-Turn Interactive Learning

Fangzhou Yao[1], Sheng Chang[1], Weibo Gao[1], Qi Liu[1*]

[1] State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
{fangzhouyao, changsheng, weibogao}@mail.ustc.edu.cn, qiliuql@ustc.edu.cn

Abstract

Learning diagnosis is a critical task that monitors students' cognitive state during educational activities, with the goal of enhancing learning outcomes. With advancements in language models (LMs), many AI-driven educational studies have shifted towards conversational learning scenarios, where students engage in multi-turn interactive dialogues with tutors. However, conversational learning diagnosis remains underdeveloped, and most existing techniques acquire students' cognitive state through intuitive instructional prompts on LMs to analyze the dialogue text. This direct prompting approach lacks a solid psychological foundation and fails to ensure the reliability of the generated analytical text. In this study, we introduce ParLD, a preview-analyze-reason framework for conversational learning diagnosis, which leverages multi-agent collaboration to diagnose students' cognitive state over multiple dialogue turns. Specifically, ParLD comprises three main components: (1) Behavior Previewer, which generates a student behavior schema based on previous states and learning content; (2) State Analyzer, which diagnoses the tutor-student dialogue and behavior schema to update the cognitive state; and (3) Performance Reasoner, which predicts the student's future responses and provides verifiable feedback to support ParLD's self-reflection with the Chain Reflector. They operate sequentially and iteratively during each interaction turn to diagnose the student's cognitive state. We conduct experiments to evaluate both performance prediction and tutoring support, emphasizing the effectiveness of ParLD in providing reliable and insightful learning diagnosis.

Introduction

The rapid growth of educational technology has accelerated the adoption of online learning, prized for its flexibility and personalized experiences. Among various approaches, conversational learning (Long 2025; Thomas 1994; Jensen 2002) has emerged as a promising paradigm. It enables students to acquire knowledge through interactive dialogue with a human or AI-driven tutor, facilitating tailored guidance and adaptive feedback (Park et al. 2024; Lv et al. 2025). A typical scenario is illustrated in Figure 1(a), where teaching unfolds as a multi-turn dialogue centered on solving a given learning objective, often framed as solving a question. At each turn, the tutor adjusts hints, questions, and feedback based on the student's responses, gradually supporting the learner's progress by adapting to their current understanding. Effectively supporting this process requires learning diagnosis that continuously monitors the student's cognitive state (Clow 2013), such as their mastery level of key knowledge concepts relevant to the target question. For example, poor performance on an *algebra*-related question signals a need for targeted intervention in that domain.

Despite its promise, accurately assessing students' evolving cognitive states during multi-turn conversations remains challenging. Traditional methods for modeling students' cognitive states, such as Knowledge Tracing (KT) (Piech et al. 2015; Ghosh, Heffernan, and Lan 2020) and Cognitive Diagnosis Models (CDMs) (Lord 1980; Zhang et al. 2024), typically infer knowledge mastery from performance labels like the correctness of students' responses. While effective for discrete exercises, these approaches provide only coarse-grained estimates and fail to capture subtle, continuous cognitive changes occurring within a single problem-solving process. In contrast, learner responses in conversational learning are predominantly open-ended, context-sensitive texts. Cognitive information is distributed and dynamically evolves across multiple turns, making stable signals difficult to extract and limiting the applicability of label-based methods. This calls for fine-grained diagnosis methods to interpret rich textual interactions.

Recent advances in Large Language Models (LLMs) offer new opportunities for conversational learning diagnosis. LLMs possess strong language understanding, broad knowledge, and flexible reasoning (Wei et al. 2022; Yue et al. 2025), enabling them to interpret open-ended, context-rich responses and track evolving cognitive states (Laban et al. 2025; Scarlatos, Baker, and Lan 2025). However, existing work primarily applies LLMs to downstream tasks such as adaptive learning or scaffolding, often treating cognitive state analysis as a secondary objective (Liu et al. 2024a,b). These approaches typically perform coarse assessments, analyzing only the final response or the entire dialogue, thereby overlooking the fine-grained, turn-level cognitive dynamics, and potentially introducing bias (Shi, Liang, and Xu 2025; Echterhoff et al. 2024). Moreover, since cognitive states are latent constructs without observable ground-truth labels, validating LLM-generated outputs

*Corresponding Author.

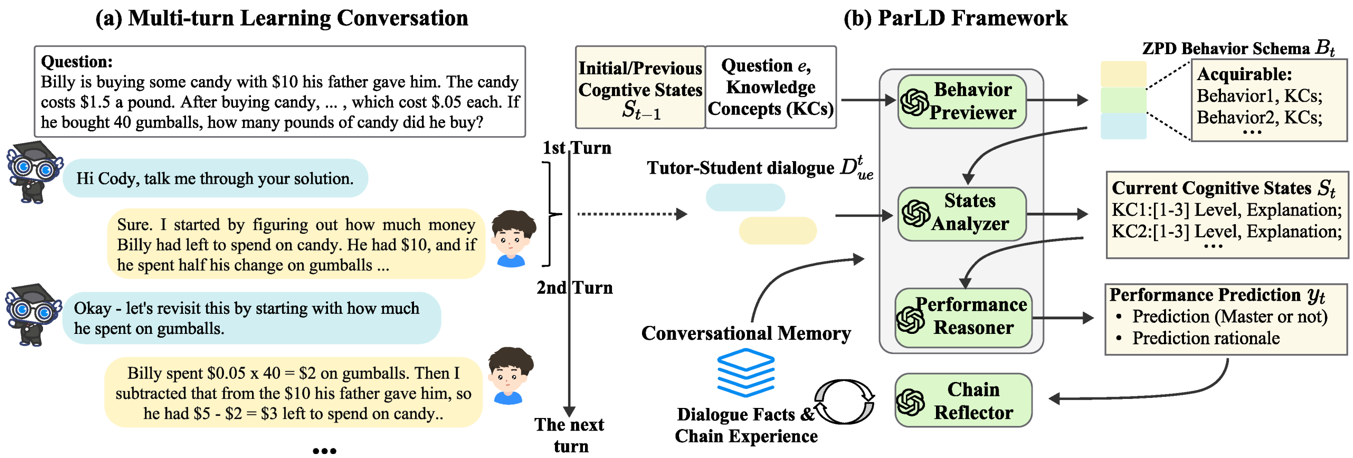

Figure 1: (a) A typical conversational learning scenario where the student interacts with a teacher or intelligent tutor in a turn-based manner. (b) The ParLD framework for diagnosing cognitive state through the preview-analyze-reason cycle, which can iteratively evolve with each interaction.

remains inherently difficult (Brown 2002; Bower 2014).

To address the aforementioned challenges, we introduce and formulate the Conversational Learning Diagnosis (CLD) task, aimed at analyzing a student's evolving cognitive state within multi-turn tutoring dialogues. To tackle this task, we propose **ParLD**, an agent-based framework built upon a ***P**review-**A**nalyze-**R**eason* chain. At a high level, ParLD consists of four core components that work collaboratively: a Behavior Previewer, a State Analyzer, and a Performance Reasoner, which together operate the preview-analyze-reason chain, and a Chain Reflector, which revisits the entire chain to refine the diagnosis results. This framework operates iteratively, with each turn unfolding in a structured sequence. First, inspired by Zone of Proximal Development (ZPD) theory (Shabani, Khatib, and Ebadi 2010), the Behavior Previewer projects a behavioral schema based on the student's prior cognitive state and the current learning objective; this schema outlines expected behaviors within their ZPD, providing concrete evidence for the subsequent phase. Next, the State Analyzer scrutinizes the live student-tutor interaction by mapping the dialogue against the projected schema, a process that allows it to update the student's mastery of relevant Knowledge Concepts (KCs). Then, the Performance Reasoner leverages this newly updated cognitive state to predict the student's likely performance on the learning question. Crucially, following each preview-analyze-reason chain, the Chain Reflector initiates a meta-cognitive loop. It reflects on the entire process, cross-references dialogue facts to refine its understanding, and updates an internal experience memory. This self-correction mechanism ensures that ParLD continuously adapts by allowing each component to draw upon the refined memory in subsequent turns, thereby progressively improving the diagnosis's fidelity over time. Our contributions are as follows:

- We first formulate the Conversational Learning Diagnosis (CLD) task, bridging the critical gap of fine-grained, dynamic tracking of student cognitive state within unstructured, multi-turn dialogues.
- We propose **ParLD**, a novel LLM-based agent framework designed to tackle the CLD task. ParLD's core innovation is its ***preview-analyze-reason*** chain, augmented by a self-correcting reflective mechanism. This allows the agent to learn from the conversational flow, continuously adapting and improving the fidelity of its diagnosis over time.
- We experimentally validate ParLD's effectiveness, demonstrating both high accuracy in performance prediction and the ability to generate impactful tutoring support from its reliable, insightful diagnosis.

Problem Definition

In this section, we begin by defining the core problem and its associated notations to formally address the challenges of conversational learning diagnosis.

Notations. We consider a learning session where a student u engages in a dialogue with a tutor t to solve a given question e. This question is associated with a set of knowledge concepts (KCs), denoted as K_e. The entire interaction is captured as a dialogue sequence $D_{ue} = \{d_1, d_2, \ldots, d_T\}$, composed of T turns. Each turn d_t consists of a pair of utterances from the tutor and the student. Finally, after the dialogue concludes, the student's overall performance on question e is recorded as r_{ue}, which can be a binary outcome (e.g., mastered, not mastered) or a qualitative score.

Definition 1 (Conversational Learning Diagnosis) *Given the dialogue history up to the t-th turn, $D_{ue}^t = \{d_1, \ldots, d_t\}$, the goal of Conversational Learning Diagnosis (CLD) is to infer the student's latent cognitive state S_t at that turn. This state, S_t, represents the student's evolving mastery level with respect to each relevant knowledge concept $k \in K_e$.*

The results of the learning diagnosis must accurately reflect the true cognitive state of students, ensuring that they are valuable for both students' self-assessment and the tutor's instructional decisions.

Methodology

In this study, we introduce ParLD, a novel multi-agent framework that operationalizes a preview-analyze-reason chain for conversational learning diagnosis. This architecture is specifically designed to infer and model student cognitive state within dynamic, multi-turn conversational learning environments.

Overview

As depicted in Figure 1(b), the ParLD framework is architected as a multi-agent system comprising four specialized modules: the ***Behavior Previewer***, the ***State Analyzer***, the ***Performance Reasoner***, and the ***Chain Reflector***. These agents operate in an iterative inference loop, processing turn-based dialogue to dynamically model the student's cognitive state. Within each conversational turn, the Behavior Previewer first projects a ZPD-Behavior schema. This projection is conditioned on the prior turn's cognitive state and question text, and the associated KCs. Subsequently, the State Analyzer scrutinizes the live tutor-student dialogue, mapping the interaction against the projected behavior schema to infer and update the student's cognitive mastery of the relevant KCs. Finally, the Performance Reasoner leverages this updated cognitive state to predict the probability of the student successfully answering the learning question. This predictive output is not merely an endpoint. It serves as a critical signal that initiates a reflective process by the Chain Reflector, allowing the ParLD framework to adaptively refine its internal memory and instructional strategy for subsequent interactions.

We provide detailed explanations of each component in the following subsections, and all prompts used in ParLD are available in the code repository.

Behavior Previewer

A primary challenge in diagnosing cognitive state from dialogue is that they are latent constructs, not directly observable. Attempting to map high-dimensional, unstructured text directly to a discrete diagnostic label is an ill-posed problem, often yielding unreliable results due to the significant semantic gap. To address this, our framework introduces the ***Behavior Previewer***: previewing a set of plausible and discriminative student behaviors before analyzing the t-th turn's dialogue. Inspired by the ZPD theory (Shabani, Khatib, and Ebadi 2010), which emphasizes the potential for cognitive growth through guided interaction, we formalize this preview as a ZPD-Behavior schema. This schema acts as a structured prior, constraining the subsequent diagnosis. It categorizes potential behaviors into three zones:

- **Mastered:** Behaviors demonstrated based on prior cognitive state.
- **Acquirable:** Behaviors that can be developed with teacher guidance.
- **Inaccessible:** Behaviors the student cannot perform even with guidance.

The zone schema is populated with specific behavioral descriptions and their associated KCs. At turn t, the Behavior Previewer agent generates this schema, B_t, by conditioning an LLM on the prior cognitive state S_{t-1}, the current question's features (question text e and KC set K_e), and a task-specific prompt, $\mathcal{P}_b$:

$$B_t = \text{LLM}(S_{t-1}, e, K_e, \mathcal{P}_b).$$

The prompt $\mathcal{P}_b$ is to instruct the LLM to generate ZPD-Behavior schema, thereby creating a bounded and interpretable hypothesis space for subsequent diagnosis phases.

State Analyzer

The ***State Analyzer*** serves as the core diagnostic engine within the ParLD framework. While other components are designed to support and refine its diagnosis, this module performs the primary function of inferring the student's mastery level for each relevant KC.

The key to its operation is the ZPD-Behavior schema (B_t), which provides a structured lens through which to interpret the raw dialogue. Instead of analyzing the dialogue in isolation, the State Analyzer maps the student's observed behaviors in the current turn's interaction, d_t, against the predicted behaviors outlined in B_t. For instance, if the student's utterances or problem-solving actions align with behavioral evidence described in the **Acquirable Zone** of the schema, the system can infer a positive shift in mastery for the associated KCs. This inference process is formally executed by prompting an LLM:

$$S_t = \text{LLM}(S_{t-1}, B_t, d_t, e, \mathcal{P}_a).$$

Here, $\mathcal{P}_a$ is a prompt engineered to instruct the model to perform this evidence-matching and state-updating task. The output, S_t, is a structured representation of the cognitive state. As illustrated in Figure 1(b), it contains key-value pairs for each KC, detailing not only the mastery level (e.g., `Good`, `Fair`, `Poor`) but also a textual explanation for the diagnosis (e.g., {`"KC1": {"level": "Poor", "explanation": "..."}`}). This structured, explainable output is crucial for enabling AI tutors to monitor learning trajectories and make informed instructional decisions.

Performance Reasoner

While the ZPD-Behavior schema provides strong priors for the ***State Analyzer***, the framework's reliability needs to be further enhanced by incorporating a verifiable feedback loop. This is the primary function of the ***Performance Reasoner***, which enables the Reflector to perform chain reflection on its diagnostic results.

Specifically, the Performance Reasoner takes the State Analyzer's output, the cognitive state S_t, and uses it to predict the student's **final performance** (r_{ue}) on the question e. This predictive task is formulated as:

$$y_t = \text{LLM}(S_t, e, \mathcal{P}_r).$$

Here, $\mathcal{P}_r$ represents a prompt specifically designed to elicit predictive reasoning from the LLM. The resulting output, y_t, is a structured tuple formatted as ($\hat{r}_t$, Rationale), where $\hat{r}_t$ is the predicted learning outcome for the current turn (e.g., $\in$ {mastered, not mastered}), and "Rationale" provides the textual justification for this prediction.

Crucially, this prediction is verifiable. Once the student's actual performance, r_{ue}, is observed at the end of the session, it can be compared against the prediction $\hat{r}_t$. This comparison provides a concrete error signal that is essential for the framework's self-reflection and refinement, which will be detailed in the next section.

Chain Reflector via Memory

To enable adaptation in a single learning turn, ParLD is equipped with a memory system and a reflective mechanism.

Conversation Memory. The **Conversation Memory**, $\mathcal{M}$, serves as an episodic buffer for the current learning session. It is designed to store a complete record of the operations within each turn, which we denote as a "turn trace", h_t.

At its core, this trace contains the dialogue from the current turn d_t, the generated ZPD-Behavior schema B_t, and the inferred cognitive state S_t. Crucially, if a reflection is triggered by the Chain Reflector during this turn, the resulting `R_trace` is also appended to h_t. After the operations of turn t are complete, its trace h_t is added to the memory. This update process is formally represented as:

$$\mathcal{M}_t = \mathcal{M}_{t-1} \cup \{h_t\}.$$

It is noted that conversational memory is ephemeral and is purged when a new learning conversation begins at a low storage cost.

Chain Reflector. The ***Chain Reflector*** is a critical component that drives ParLD's self-correction process. It is activated when a **discrepancy** occurs between the Performance Reasoner's prediction and the observed student performance at each interaction turn. Upon detection of such a discrepancy, the Reflector systematically revisits the entire preview-analyze-reason chain, querying the Conversation Memory to identify the root cause of the error.

For example, when auditing the preview-analyze subchain at turn t, the Reflector might ask: "Was the cognitive state S_t correctly inferred, given the dialogue d_t and the schema B_t?" This inquiry is guided by a specific prompt, $\mathcal{P}_{\text{reflect}}$, which generates a structured critique:

$$\mathcal{R}_{P \to A} = \text{LLM}(M_t, \mathcal{P}_{\text{reflect}}).$$

Here, $\mathcal{P}_{\text{reflect}}$ is the reflection prompt that justifies whether the ZPD-Behavior schema B_t accurately informs the cognitive state S_t. The $\mathcal{R}_{P \to A}$ is a structured output containing two key pieces of information: a `judgment` (e.g., accurate or not) and a `critique` (a textual explanation for the judgment). If the reflection indicates that the cognitive state was inaccurately inferred, the Chain Reflector triggers the State Analyzer to rerun the diagnosis process, utilizing $\mathcal{R}_{P \to A}$ and the internal memory associated with previous changes. The diagnosis continues iteratively until the performance prediction is accurate. Additionally, we introduce a *max_num* parameter to control the cost in the experimental setup, limiting the number of reflection iterations.

By leveraging the Chain Reflector via Conversation Memory, this mechanism ensures that the diagnostic model continuously adapts, thereby enhancing the accuracy and fidelity of its cognitive state diagnosis in subsequent interaction turns.

Experiments

This section details the empirical evaluation of our proposed framework, ParLD. The primary objective of ParLD is to generate accurate, turn-by-turn diagnoses of a student's cognitive state within a conversational learning context. However, evaluating the accuracy of these inferred states is inherently challenging, as the true cognitive state is a latent, unobservable construct. Therefore, to validate the efficacy of our framework, we assess its utility through two well-defined proxy tasks that are directly dependent on the quality of the generated diagnoses: (1) student performance prediction and (2) tutoring support. Our experimental design is guided by the following research questions:

- **RQ1:** How effective is ParLD at predicting student performance compared to baseline models?
- **RQ2:** What are the individual contributions of the Previewer and Reflector modules to ParLD's overall performance?
- **RQ3:** What is the utility of ParLD's diagnostic outputs for enhancing pedagogical decision-making in interactive learning scenarios?

Dataset

We use the MathDial and CoMTA datasets for experiments:

MathDial (Macina et al. 2023) is a large-scale dataset comprising 2,861 dialogues. These dialogues feature interactions between human teachers and a student agent simulated by InstructGPT (Ouyang et al. 2022). Each conversation is goal-oriented, designed to guide the student toward correctly solving a math problem they had previously answered incorrectly. For our experiments, we strictly adhere to the official train/test split provided with the dataset.

CoMTA (Miller and DiCerbo 2024) provides authentic interaction data, capturing real student dialogues with an intelligent tutoring system from Khan Academy. To ensure a focused evaluation on clear learning objectives, we curated a subset of this data by filtering 116 of 188 conversations that contained clear conversational goals, making them suitable for our turn-by-turn diagnostic analysis.

The conversations in both datasets are annotated with a final label indicating whether the student has fully mastered the question. This valuable label is used to validate the student's performance on the task.

Utilized LLMs

Considering the power, need for structured output, and cost-effectiveness of the model, we select **GPT-4.1** (OpenAI 2025) and **GPT-4o** (OpenAI 2024) via OpenAI's API service to construct the ParLD agent framework. All LLMs used in the experiments have their temperature set to **0** to ensure stable output. Additionally, the maximum reflection time in the Reflector is set to 2 and 1 in MathDial and CoMTa, respectively, to save costs. The code is available at https://github.com/fannazya/ParLD.

Models	MathDial		CoMTA	
	ACC↑	**F1 ↑**	**ACC↑**	**F1 ↑**
DKT	58.72	<u>65.26</u>	51.88	46.40
AKT	57.67	64.50	53.84	52.88
DKVMN	55.15	63.04	52.31	47.06
SAINT	58.25	63.98	51.76	47.06
SimpleKT	56.70	63.38	44.58	44.73
ParLD (GPT4o)	<u>65.08</u>	64.04	<u>57.02</u>	<u>56.84</u>
ParLD (GPT4.1)	**68.72**	**66.15**	**57.26**	**56.91**

Table 1: Comparison of model performance on the MathDial and CoMTA datasets, with the best and second-best models highlighted in **bold** and <u>underlined</u>, respectively. The arrow ↑ means the higher score, the better performance. These markers are also for the following results.

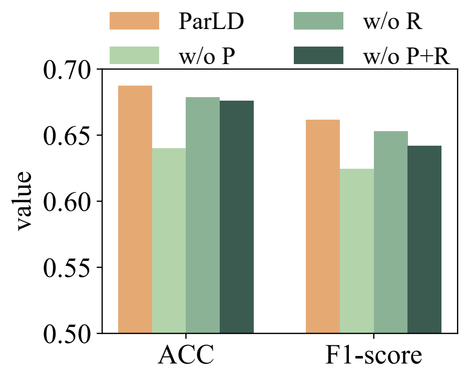

Figure 2: The ablation study of ParLD (GPT-4.1) on the MathDial dataset.

Performance Prediction Comparison (RQ1)

Motivation. In this section, we aim to evaluate whether the cognitive state generated by ParLD accurately reflects the students' learning progress. For precise evaluation, we use the annotated labels as the final learning outcome for each conversation, comparing them with the predicted outcomes based on the final cognitive state. Notably, this prediction is made on the final turn without triggering the chain reflection process.

Baselines and Evaluation Metrics. Given the limited work on the CLD task, we utilize several Knowledge Tracing (KT) models (Scarlatos, Baker, and Lan 2025) for comparisons: DKT (Piech et al. 2015), AKT (Ghosh, Heffernan, and Lan 2020), DKVMN (Zhang et al. 2017), SAINT (Choi et al. 2020), and SimpleKT (Liu et al. 2023), which predict students' future performance based on their dialogue learning history. To ensure a fair comparison, we only report the prediction at the final turn of the conversation. The effectiveness of the models is assessed using both classification and regression metrics, including accuracy (ACC) and F1-score.

Table 1 demonstrates that ParLD (GPT-4.1) achieves state-of-the-art performance on both the Mathdial and CoMTA datasets, outperforming all traditional KT models. GPT-4.0 also performs well, ranking second in most metrics, except for the F1 score on the Mathdial dataset. Specifically, ParLD (GPT-4.1) surpasses the best-performing DKT model on the Mathdial dataset by a margin of 10%. This result underscores the potential of LLMs to significantly improve learning diagnosis tasks, particularly in predicting students' cognitive states and the reasoning behind their final predictions. Notably, ParLD (GPT-4.1) outperforms ParLD (GPT-4.0), highlighting how enhanced LLM capabilities can lead to more accurate and insightful predictions of learning progress. This points to a promising direction for future educational technologies.

Ablation Study (RQ2)

To isolate the contributions of our proposed modules, we conducted an ablation study comparing the full ParLD framework against three variants in the performance prediction task. The first variant, **w/o P**, removes the Behavior **P**reviewer, thereby analyzing the dialogue directly without the constraining ZPD-Behavior schema. The second variant, **w/o R**, disables the **R**eflector, which means memory updates from performance feedback are not incorporated, allowing us to measure the impact of the reflection mechanism. Finally, the **w/o P+R** variant serves as a baseline without both the Behavior Previewer and Performance Reflector, relying solely on the State Analyzer for performance prediction.

The experimental results in Figure 2 demonstrate that, for both GPT-4.1 and GPT-4.0, ParLD outperforms all variants, confirming the contribution of each component to the reliability of its learning diagnosis. Notably, the "w/o P" variant, which relies solely on the State Analyzer, performs the worst in predicting whether a student has effectively learned the material. This highlights the critical role of the preview-analyze chain, supported by the psychological ZPD-Behavior schema, in providing the necessary context for the State Analyzer. In contrast, reflecting only on the analyze-reason chain can mislead the Analyzer, resulting in less accurate predictions.

Effects on Tutoring Support (RQ3)

In this section, we investigate whether the learning diagnosis results can assist the tutor system in enhancing instructional strategies to improve conversational learning outcomes. Evaluating tutoring support is inherently challenging, as tutoring interventions alter real students' learning states, making controlled comparisons impossible. To this end, we conduct simulation experiments and perform quality evaluations.

Tutoring Enhancement Simulation

To answer **RQ3**, we designed a simulation experiment to evaluate the pedagogical utility of ParLD's diagnostic outputs. The core motivation is that a more accurate cognitive diagnosis should lead to more effective tutoring instructions, which in turn should help students solve the problem more efficiently. We measure effectiveness by calculating the accuracy of simulated students' answers to the learned ques-

Model Setting		CR↑	Avg. T↓	Int. Avg. T↓
GPT-4.1	-ParLD	**72.22**	3.29	**2.83**
	-DA	62.96	3.28	3.04
	-DR	56.48	**3.25**	3.39
GPT-4o	-ParLD	**62.04**	**3.43**	**3.20**
	-DA	53.70	3.74	3.63
	-DR	49.07	3.81	3.63

Table 2: Comparison of GPT-4.1 and GPT-4o in tutoring support across different settings on the MathDial dataset. The arrow ↑(↓) means the higher (lower) score, the better (worse) performance.

tions when guided by instructions generated from ParLD's diagnostic results.

Simulation Setup. We use the same settings as the original dataset, simulating students with InstructGPT using the default profile information, and filtering 108 conversations where the simulated student did not answer correctly until the final turn. The simulation proceeds as follows: starting from the second turn, the teacher's utterance is modified by generating improved tutoring instructions based on the cognitive state produced by ParLD. At each turn, the student's response is verified for correctness; if it is correct, the simulation terminates. We set the maximum number of turns for each conversation to the original number of turns in the dataset. The model settings compared are Direct Respond (**DR**), which directly responds to the student's response, and Direct Analyze (**DA**), which simply analyzes the student's cognitive state in each turn and generates corresponding teaching instructions.

Evaluation Metrics. We use three metrics as follows:

- **Correct Rate (CR):** The percentage of simulated dialogues that end with a correct student answer.
- **Avg. Turns (Avg. T):** The average number of turns taken to reach a solution for those dialogues that were successfully solved.
- **Intersection Avg. Turns (Int. Avg. T):** The average number of turns for only those dialogues that were successfully solved by all compared models. This metric provides a fairer comparison of efficiency on a common set of problems.

Results Analysis. The results show that ParLD (GPT-4.1) and ParLD (GPT-4.0) significantly outperform both Direct Respond (DR) and Direct Analyze (DA), with DR performing the worst. Specifically, ParLD with GPT-4.1 helps 72.22% of students effectively learn the material, compared to just 56.48% with Direct Respond using GPT-4.1. This difference underscores the superior efficacy of ParLD in facilitating effective learning. It also highlights the reliability of the cognitive states diagnosed by ParLD, which are essential for adapting the tutoring process to each student's needs.

In terms of average dialogue turns (Avg. T), ParLD with GPT-4.0 reduces the number of interactions required for effective tutoring, while yielding results that are similar to those of DR and DA when used with GPT-4.1. This closeness is reasonable because more complex questions naturally require more turns for effective tutoring. The valid advantage of a lower Int. Avg. T on the common set of problems solved by all models using ParLD further supports this point. The reduced Int. Avg. T indicates that ParLD not only improves learning outcomes but also requires fewer dialogue turns to help students master the material efficiently.

In a word, ParLD offers a dual benefit: it helps students overcome learning challenges more effectively while requiring fewer interactions.

Quality Evaluations: To further evaluate the diagnostic quality of ParLD, we sampled 20 simulated conversations and assessed the reliability of the inferred cognitive states. Two mathematics-proficient students rated the diagnostic results in terms of Accuracy and Coherence using a 3-point scale (1 = worst, 3 = best). ParLD outperformed DA (2.475 vs. 2.1 in Accuracy; 2.7 vs. 2.425 in Coherence), indicating its advantage in conversational learning diagnosis.

Case Study (RQ3)

In this section, we present a specific case of multi-turn conversational learning diagnosis and tutoring from the MathDial dataset to demonstrate the strong learning diagnosis and tutoring enhancement capabilities of our proposed framework. By comparing DR (Direct Respond) and ParLD-supported tutoring (Figure 3(a) and (b)), we can observe that ParLD effectively mines evidence of areas where the student struggles, such as ***Subtraction***. In the following turn, the tutoring system generates targeted instructions focused on this ***Subtraction*** KC, helping students identify and fill knowledge gaps, ultimately enhancing learning outcomes. The DR cannot capture the student's confusion and merely describes the question without specifying the underlying knowledge concept. As a result, the student tutored by DR fails after 10 dialogue turns, whereas ParLD achieves the correct answer in just three turns.

Related Works

Learning Diagnosis

Learning Diagnosis aims to model and understand student learning processes by analyzing educational data (Clow 2013). Within this field, two prominent lines of research are Knowledge Tracing (KT) and Cognitive Diagnosis Models (CDMs). The KT research line (Liu et al. 2025; Gao et al. 2025a) focuses on predicting student performance by tracking knowledge evolution over time. This area has progressed from foundational models like Bayesian Knowledge Tracing (BKT) (Yudelson, Koedinger, and Gordon 2013) to modern deep learning approaches DKT (Piech et al. 2015) and its advanced variants (Ghosh, Heffernan, and Lan 2020; Shen et al. 2021; Liu et al. 2019). CDMs, conversely, aim to infer students' fine-grained proficiency on specific skills based on their question responses. This includes classic models like

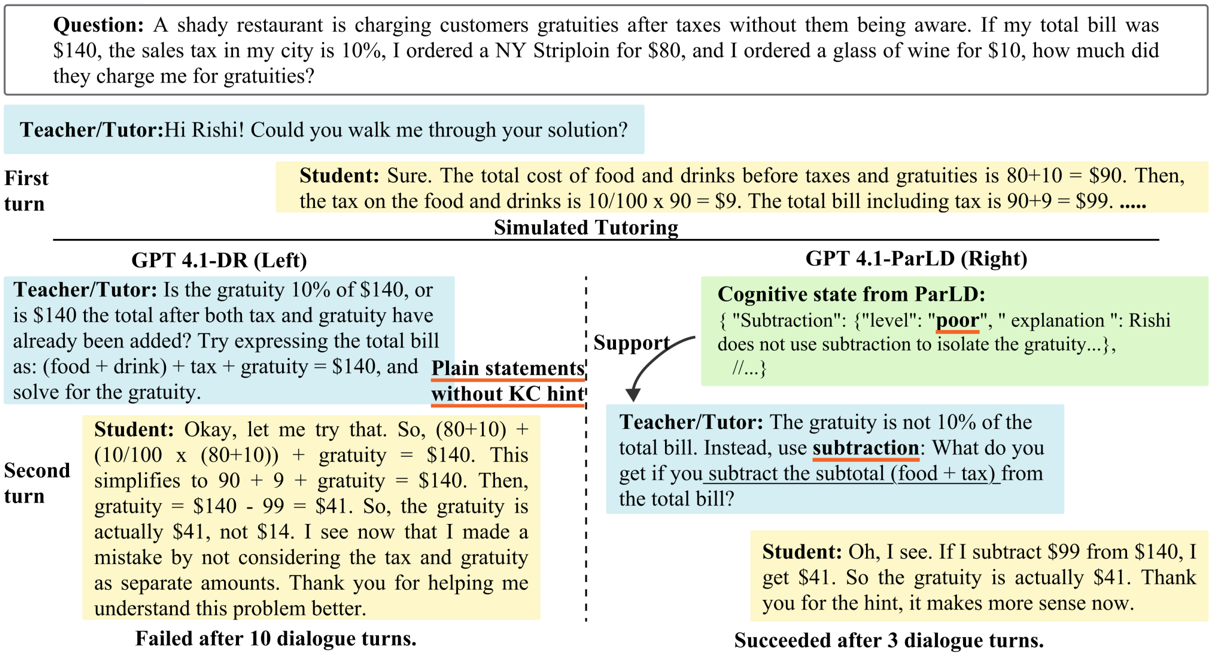

Figure 3: A specific case of multi-turn conversational learning diagnosis and tutoring from the MathDial dataset.

Item Response Theory (IRT) (Lord 1980) and more recent methods like NeuralCDM (Wang et al. 2020), which leverages neural networks to model complex student-question interactions. Additionally, recent deep learning-based CDMs have been developed to address practical challenges in online education, such as item scarcity (Chen et al. 2023; Yao et al. 2023), zero-shot diagnosis (Gao et al. 2024), and noisy interactions (Yao et al. 2024). A primary limitation of current methods is their fundamental reliance on structured, discrete data, such as right/wrong answers to questions. This makes them inherently ill-suited for the unstructured, open-ended nature of turn-by-turn conversational learning.

Conversational Learning

Conversational learning (Jensen 2002; Thomas 1994) leverages dialogue as a primary mechanism for knowledge acquisition and skill development. Early pioneering works in this area, often categorized as Intelligent Tutoring Systems (ITS), laid the foundational principles. Systems like AutoTutor (Graesser et al. 2004) and other tutorial dialogue systems (Olney, Graesser, and Person 2010) used techniques such as Latent Semantic Analysis (LSA) and scripted dialogue moves to guide students. However, these traditional systems often relied on heavily pre-authored content and rule-based dialogue managers, which made them labor-intensive to build and limited their flexibility. The advent of Large Language Models (LLMs) has catalyzed a paradigm shift in conversational learning (Milano, McGrane, and Leonelli 2023), overcoming many of these traditional limitations with their extraordinary knowledge and strong interactivity. Existing research has focused on leveraging LLMs as powerful tutors or educational tools for applications such as Socratic teaching (Liu et al. 2024a; Ding et al. 2024), generating dynamic learning scaffolds (Liu et al. 2024b), and recommending personalized learning paths (Lv et al. 2025; Cao and Wu 2025), behavior simulation (Zhang et al. 2025; Gao et al. 2025b). However, in many of them, learning diagnosis is treated as an intermediate step for a downstream task, rather than a primary research objective. This highlights a remaining gap in leveraging the full potential of LLMs for dynamic, reliable learning diagnosis within conversational settings.

Conclusion

In this paper, we addressed the critical challenge of unreliable and psychologically ungrounded learning diagnosis in conversational learning environments. To tackle this issue, we introduced ParLD, a novel preview-analyze-reason agent framework that leverages multi-agent collaboration for a more robust diagnosis of students' cognitive state. Unlike prevailing approaches that rely on direct and often brittle prompting of language models, ParLD implements a structured, iterative process. By first generating a Behaviour Preview schema, then using a State Analyzer to interpret the live dialogue in context, and finally leveraging a Performance Reasoner for self-reflection, ParLD establishes a more reliable analytical loop. Our experiments, focusing on the practical utility of the diagnosed states, demonstrated ParLD's effectiveness. The cognitive state produced by our framework significantly enhanced the accuracy of student performance prediction and proved valuable for informing appropriate tutoring support. In conclusion, ParLD represents a significant step towards more reliable, insightful, and actionable learning diagnosis, providing a foundational component for the next generation of truly adaptive conversational tutoring systems.

Acknowledgements

This research was partially supported by grants from the National Key Research and Development Program of China (Grant No. 2024YFC3308200), the National Natural Science Foundation of China (No. 62525606), the Key Technologies R & D Program of Anhui Province (No. 202423k09020039), and the Fundamental Research Funds for the Central Universities.

References

Bower, G. H. 2014. Cognitive psychology: An introduction. In *Handbook of Learning and Cognitive Processes (Volume 1)*, 25–80. Psychology Press.

Brown, R. J. 2002. The cognitive psychology of dissociative states. *Cognitive Neuropsychiatry*, 7(3): 221–235.

Cao, Y.; and Wu, Y. 2025. Dynamic Generation Model of AI-Personalized Learning Paths Based on Educational Psychology. In *Proceedings of the 2025 International Conference on Artificial Intelligence and Educational Systems*, ICAIES '25, 124–130. New York, NY, USA: Association for Computing Machinery. ISBN 9798400715068.

Chen, X.; Wu, L.; Liu, F.; Chen, L.; Zhang, K.; Hong, R.; and Wang, M. 2023. Disentangling cognitive diagnosis with limited exercise labels. *Advances in Neural Information Processing Systems*, 36: 18028–18045.

Choi, Y.; Lee, Y.; Cho, J.; Baek, J.; Kim, B.; Cha, Y.; Shin, D.; Bae, C.; and Heo, J. 2020. Towards an appropriate query, key, and value computation for knowledge tracing. In *Proceedings of the seventh ACM conference on learning@ scale*, 341–344.

Clow, D. 2013. An overview of learning analytics. *Teaching in Higher Education*, 18(6): 683–695.

Ding, Y.; Hu, H.; Zhou, J.; Chen, Q.; Jiang, B.; and He, L. 2024. Boosting large language models with socratic method for conversational mathematics teaching. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 3730–3735.

Echterhoff, J. M.; Liu, Y.; Alessa, A.; McAuley, J.; and He, Z. 2024. Cognitive Bias in Decision-Making with LLMs. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 12640–12653. Miami, Florida, USA: Association for Computational Linguistics.

Gao, W.; Liu, Q.; Li, R.; Zhao, Y.; Wang, H.; Yue, L.; Yao, F.; and Zhang, Z. 2025a. Denoising Programming Knowledge Tracing with a Code Graph-based Tuning Adaptor. KDD '25, 354–365. New York, NY, USA: Association for Computing Machinery. ISBN 9798400712456.

Gao, W.; Liu, Q.; Wang, H.; Yue, L.; Bi, H.; Gu, Y.; Yao, F.; Zhang, Z.; Li, X.; and He, Y. 2024. Zero-1-to-3: Domain-Level Zero-Shot Cognitive Diagnosis via One Batch of Early-Bird Students towards Three Diagnostic Objectives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8417–8426.

Gao, W.; Liu, Q.; Yue, L.; Yao, F.; Lv, R.; Zhang, Z.; Wang, H.; and Huang, Z. 2025b. Agent4edu: Generating learner response data by generative agents for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23923–23932.

Ghosh, A.; Heffernan, N.; and Lan, A. S. 2020. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2330–2339.

Graesser, A. C.; Lu, S.; Jackson, G. T.; Mitchell, H. H.; Ventura, M.; Olney, A.; and Louwerse, M. M. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36(2): 180–192.

Jensen, J. 2002. Conversational learning: An experiential approach to knowledge creation.

Laban, P.; Hayashi, H.; Zhou, Y.; and Neville, J. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

Liu, J.; Huang, Z.; Xiao, T.; Sha, J.; Wu, J.; Liu, Q.; Wang, S.; and Chen, E. 2024a. SocraticLM: Exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems*, 37: 85693–85721.

Liu, Q.; Huang, Z.; Yin, Y.; Chen, E.; Xiong, H.; Su, Y.; and Hu, G. 2019. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1): 100–115.

Liu, Z.; Guo, T.; Liang, Q.; Hou, M.; Zhan, B.; Tang, J.; Luo, W.; and Weng, J. 2025. Deep Learning Based Knowledge Tracing: A Review, A Tool and Empirical Studies. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, Z.; Liu, Q.; Chen, J.; Huang, S.; and Luo, W. 2023. simpleKT: a simple but tough-to-beat baseline for knowledge tracing.

Liu, Z.; Yin, S. X.; Lee, C.; and Chen, N. F. 2024b. Scaffolding language learning via multi-modal tutoring systems with pedagogical instructions. In *2024 IEEE conference on artificial intelligence (CAI)*, 1258–1265. IEEE.

Long, F. 2025. Conversational Learning in the Age of ChatGPT. *Studies in Philosophy and Education*, 44(3): 245–261.

Lord, F. M. 1980. *Applications of item response theory to practical testing problems*. Routledge.

Lv, R.; Liu, Q.; Gao, W.; Zhang, H.; Lu, J.; and Zhu, L. 2025. GenAL: Generative Agent for Adaptive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 577–585.

Macina, J.; Daheim, N.; Chowdhury, S.; Sinha, T.; Kapur, M.; Gurevych, I.; and Sachan, M. 2023. MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5602–5621. Singapore: Association for Computational Linguistics.

Milano, S.; McGrane, J. A.; and Leonelli, S. 2023. Large language models challenge the future of higher education. *Nature Machine Intelligence*, 5(4): 333–334.

Miller, P.; and DiCerbo, K. 2024. Llm based math tutoring: Challenges and dataset.

Olney, A. M.; Graesser, A. C.; and Person, N. K. 2010. Tutorial dialog in natural language. *Advances in intelligent tutoring systems*, 181–206.

OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276.

OpenAI. 2025. Introducing GPT-4.1 in the API.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Park, M.; Kim, S.; Lee, S.; Kwon, S.; and Kim, K. 2024. Empowering personalized learning through a conversation-based tutoring system with student modeling. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–10.

Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L. J.; and Sohl-Dickstein, J. 2015. Deep knowledge tracing. *Advances in neural information processing systems*, 28.

Scarlatos, A.; Baker, R. S.; and Lan, A. 2025. Exploring knowledge tracing in tutor-student dialogues using llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 249–259.

Shabani, K.; Khatib, M.; and Ebadi, S. 2010. Vygotsky's zone of proximal development: Instructional implications and teachers' professional development. *English language teaching*, 3(4): 237–248.

Shen, S.; Liu, Q.; Chen, E.; Huang, Z.; Huang, W.; Yin, Y.; Su, Y.; and Wang, S. 2021. Learning process-consistent knowledge tracing. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 1452–1460.

Shi, Y.; Liang, R.; and Xu, Y. 2025. EducationQ: Evaluating LLMs' Teaching Capabilities Through Multi-Agent Dialogue Framework. *arXiv e-prints*, arXiv–2504.

Thomas, A. 1994. Conversational learning. *Oxford Review of Education*, 20(1): 131–142.

Wang, F.; Liu, Q.; Chen, E.; Huang, Z.; Chen, Y.; Yin, Y.; Huang, Z.; and Wang, S. 2020. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6153–6161.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

Yao, F.; Liu, Q.; Hou, M.; Tong, S.; Huang, Z.; Chen, E.; Sha, J.; and Wang, S. 2023. Exploiting non-interactive exercises in cognitive diagnosis. *Interaction*, 100(200): 300.

Yao, F.; Liu, Q.; Yue, L.; Gao, W.; Li, J.; Li, X.; and He, Y. 2024. Adard: An adaptive response denoising framework for robust learner modeling. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3886–3895.

Yudelson, M. V.; Koedinger, K. R.; and Gordon, G. J. 2013. Individualized bayesian knowledge tracing models. In *International conference on artificial intelligence in education*, 171–180. Springer.

Yue, L.; Du, Y.; Wang, Y.; Gao, W.; Yao, F.; Wang, L.; Liu, Y.; Xu, Z.; Liu, Q.; Di, S.; et al. 2025. Don't Overthink It: A Survey of Efficient R1-style Large Reasoning Models. *arXiv preprint arXiv:2508.02120*.

Zhang, J.; Shi, X.; King, I.; and Yeung, D.-Y. 2017. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th international conference on World Wide Web*, 765–774.

Zhang, Z.; Wu, L.; Liu, Q.; Liu, J.; Huang, Z.; Yin, Y.; Zhuang, Y.; Gao, W.; and Chen, E. 2024. Understanding and improving fairness in cognitive diagnosis. *Science China Information Sciences*, 67(5): 152106.

Zhang, Z.; Zhang-Li, D.; Yu, J.; Gong, L.; Zhou, J.; Hao, Z.; Jiang, J.; Cao, J.; Liu, H.; Liu, Z.; et al. 2025. Simulating Classroom Education with LLM-Empowered Agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 10364–10379.