

# Instruct, Not Assist: LLM-based Multi-Turn Planning and Hierarchical Questioning for Socratic Code Debugging

Priyanka Kargupta\*, Ishika Agarwal\*, Dilek Hakkani-Tur, Jiawei Han

<sup>1</sup>Department of Computer Science, University of Illinois at Urbana-Champaign

{pk36, ishikaa2, dilek, hanj}@illinois.edu

## Abstract

Socratic questioning is an effective teaching strategy, encouraging critical thinking and problem-solving. The conversational capabilities of large language models (LLMs) show great potential for providing scalable, real-time student guidance. However, current LLMs often give away solutions directly, making them ineffective instructors. We tackle this issue in the code debugging domain with **TreeInstruct**, an Instructor agent guided by a novel state space-based planning algorithm. TreeInstruct asks probing questions to help students independently identify and resolve errors. It estimates a student’s conceptual and syntactical knowledge to dynamically construct a question tree based on their responses and current knowledge state, effectively addressing both independent and dependent mistakes concurrently in a multi-turn interaction setting. In addition to using an existing single-bug debugging benchmark, we construct a more challenging multi-bug dataset of 150 coding problems, incorrect solutions, and bug fixes— all carefully constructed and annotated by experts. Extensive evaluation shows TreeInstruct’s state-of-the-art performance on both datasets, proving it to be a more effective instructor than baselines. Furthermore, a real-world case study with five students of varying skill levels further demonstrates TreeInstruct’s ability to guide students to debug their code efficiently with minimal turns and highly Socratic questioning.

## 1 Introduction

With the rapidly expanding conversational and reasoning abilities of large language models (LLMs), there has been a substantial rise in demand for exploiting their capabilities within a multitude of educational applications (Kasneci et al., 2023) in order to widen accessibility via personalized feedback. Specifically, several recent works explore

\*These authors contributed equally to this work.

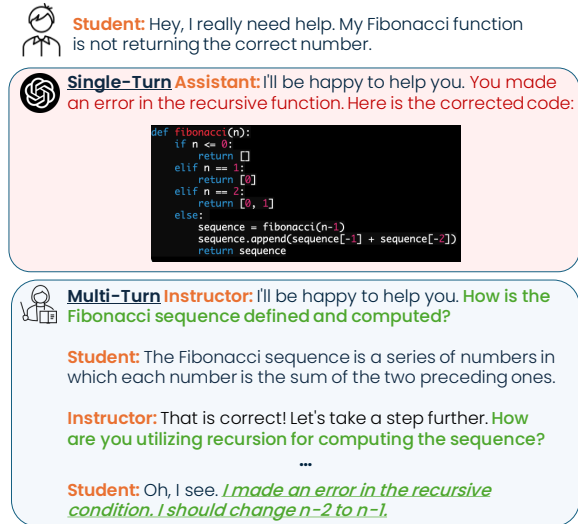


Figure 1: The Instructor’s goal is to generate multi-turn Socratic questions while guiding the Student towards the correct solution.

the use of LLMs for providing feedback and guidance to students (Wang et al., 2023; Kazemitabaar et al., 2024; Sheese et al., 2024; Lyu et al., 2024). However, LLMs are typically optimized to generate customer-serving, assistant-like responses, which also translates into the types of questions asked. Especially for educational domains, this style of questioning can be suboptimal (Cotton, 1988; Sahamid, 2016; Yang et al., 2005; Wilson, 1987). For instance, if a student is seeking help from an instructor for correcting their mistakes (e.g., debugging their buggy code), we consider two forms of potential responses: **assistant-like** and **instructor-like**. As shown in Figure 1, an assistant-like response would not be a successful educational interaction, as it leads to the Assistant directly providing an answer. On the other hand, an Instructor-like response reflects the educational philosophy of *Socratic questioning*.

Socratic questioning is a teaching strategy where the Student independently solves their problem by answering *guiding* questions, instead of being

given the *solution directly* (Wilson, 1987). This is a more effective learning strategy because the weight of learning falls on the Student as they must put in effort to answer a question as opposed to solely relying on the model (Cotton, 1988; Kasneci et al., 2023). Therefore, we aim to re-orient an LLM to be an Instructor, not an assistant, by asking Socratic questions that (1) help the Student understand their mistakes, and (2) do not directly provide the answer. To tackle these challenges, we propose **TreeInstruct** based on the following principles:

1. **State space estimation:** An Instructor plans its conversation with a Student based on the “distance” between their initial answer and the optimal, correct answer within the estimated state space. In other words, it tracks the knowledge state of the Student within this space throughout the Instructor-Student interactions.
2. **Tree-based Socratic questioning:** An Instructor generates turn-level Socratic questions conditioned on both the Student’s current knowledge state *and* misunderstanding(s), the latter derived from their responses to the Instructor’s questions. This step dynamically constructs a Socratic question tree.
3. **Adaptive conversation restructuring:** An Instructor updates their initial conversation plan based on how the Student is progressing in the conversation, as reflected by updates (or lack thereof) to the Student’s knowledge state. This planning can include both *questioning* and *teaching* actions.

While these principles can apply to many educational domains, this paper focuses on code debugging, which presents unique challenges. Real-world code debugging often involves multiple, potentially interdependent conceptual and syntactical bugs. For instance, Figure 1 shows that first resolving the Student’s conceptual misunderstanding of recursion in Fibonacci helps them identify their recursive syntactical bug (Figure 1). However, existing work fails to account for such nuances and assumes single-turn feedback (Kazemitabaar et al., 2024; Wang et al., 2023; Lyu et al., 2024). This ignores the sub-steps required for the Student to understand each bug.

In contrast, TreeInstruct constructs a multi-turn debugging plan (*state representation*), defined as the set of Student misunderstandings and mistakes (*state variables*) to be resolved in order to comprehend and correct their bug(s). We define all

potential paths to complete these tasks as the *state space*. We traverse the space using Socratic questions and trace which variables have been resolved, grounded based on the Student’s responses.

While existing LLM-based tutors are effective in fixing the Student’s code with high success, they are either prone to directly revealing code answers or cannot be adapted to new Student responses. For example, CodeAid (Kazemitabaar et al., 2024) (specifically, the "Help Fix Code" and "Question from Code" modules, as these are most similar to our setting) directly provides code or pseudocode 57% of the time, and achieves a mere 55% rate of helpfulness. On the other hand, TreeInstruct exploits the state space to dynamically construct a tree of questions based on (1) incorrect Student responses, or (2) gaps in the Student’s knowledge. The sibling and parent-child relationships between questions reflect the manner in which they traverse the state space. Finally, it exploits both the Student’s knowledge state and any proposed bug fixes to serve as the dynamic stopping condition. Overall, TreeInstruct takes a more structured approach to multi-turn conversational feedback, as (1) grounding the conversation on the state space representation ensures that all bugs are sufficiently addressed, and (2) constructing a tree based on the Student’s current level of understanding allows for more relevant and personalized question generation.

We summarize our contributions below:

- To the best of our knowledge, TreeInstruct is the first work to explore state space estimation and dynamic tree-based questioning for multi-turn Socratic instruction.
- We construct a novel multi-bug debugging dataset with 150 expert-annotated, challenging conceptual and syntactical bugs and their fixes.
- Extensive experiments on an existing benchmark and our constructed dataset demonstrate that TreeInstruct can be universally applied to both open and closed source-settings. We also showcase that TreeInstruct’s strong Socratic questioning abilities widely outperform all baselines through both (1) rigorous quantitative and qualitative expert evaluation (on average, *preferred* 78.43% of the time; Student fixes code 24.55% *more*) and (2) real-world interactions with students of varying coding abilities.

**Reproducibility:** We release our data and source code<sup>1</sup> to facilitate further studies.

<sup>1</sup><https://github.com/agarwalishika/TreeInstruct>

---

**Algorithm 1** TreeInstruct

---

**Require:**  $P$  (Problem Description),  $B$  (Buggy Code, Bug Descriptions),  $C$  (Corrected  $B$  Code, Bug Fixes)

- 1:  $S = \{\tau_1, \tau_2, \dots, \tau_k\} \leftarrow \text{GenerateState}(P, B, C)$  ▷ Section 3.2: State representation: (resolved?, task)
- 2:  $l \leftarrow 0, Q \leftarrow \{l : \square\}, H \leftarrow \square, F \leftarrow \{\}$  ▷ Tree level, question list/level, conv. history, Student bug fixes
- 3:  $q \leftarrow \text{GenerateQuestion}(P, B, C, \tau_1)$  ▷ Section 3.3: Generate initial question
  
- 4: **while**  $\exists \tau \in S$  s.t.  $\neg \text{isResolved}(S, F, C)$  **do** ▷ Section 3.4: Process while tasks or bugs remain
- 5:      $r \leftarrow \text{StudentResponse}(q)$
- 6:      $v, w \leftarrow \text{VerifyResponse}(q, r)$  ▷ Section 3.3: is  $r$  to  $q$  correct ( $v$ ); why or why not ( $w$ )?
- 7:      $H.add(q, r)$
- 8:      $Q[l].add(q)$
- 9:     **if**  $v = \text{false}$  **then** ▷ **Incorrect student response**
- 10:          $q \leftarrow \text{GenerateSiblingQuestion}(\tau, Q[l], H, w)$  ▷ Section 3.3: factor in *why* the student was incorrect
- 11:     **else** ▷ **Correct student response**
- 12:          $S, w \leftarrow \text{UpdateUnderstanding}(S, q, r)$  ▷ Section 3.3: tasks  $\tau_i \dots \tau_k$  resolved? If  $\neg S[\tau]$ , why ( $w$ )?
- 13:         **if**  $\neg S[\tau]$  **then**
- 14:              $q \leftarrow \text{GenerateChildQuestion}(\tau, Q[l], H, w)$  ▷ Section 3.3: factor in *why*  $\tau$  was unresolved
- 15:              $l \leftarrow l + 1$  ▷ Advance to next tree level
- 16:         **else** ▷ **Task  $\tau$  resolved**
- 17:              $F \leftarrow \text{GetStudentBugFixes}(H)$  ▷ Section 3.4: ask Student for bug fixes (if any)
- 18:              $l \leftarrow 0, Q \leftarrow \{l : \square\}$  ▷  $\tau$  resolved  $\rightarrow$  create new tree

---

## 2 Related Works

### 2.1 Knowledge Tracing

Knowledge tracing tracks student knowledge to personalize their learning experience, including understanding specific concepts, behavior, and recall ability. There are two primary methods: probabilistic and deep learning-based. Probabilistic knowledge tracing, as it was first introduced, uses a Hidden Markov Model (HMM) to maintain binary states, learned and unlearned, for each skill as learners engage with exercises. This approach, from which we draw inspiration, updates the likelihood of these states based on performance (Corbett and Anderson, 1994; Yudelson et al., 2013). Some models use open-ended paths to states (Rafferty et al., 2016), while others use deep learning-based, long-term memory capabilities essential for learning (Piech et al., 2015). These methods are performative, but such state spaces hinder effectiveness and require large amounts of annotated training data.

Our methodology addresses the challenge of limited annotated data by dynamically generating states during interactions between instructors and students. We monitor these evolving states through a component we refer to as the Verifier. Using these dynamically generated states, we tailor the educational experience by personalizing the sequence and type of questions posed to learners.

### 2.2 Socratic Reasoning in Educational AI

There have been several works exploring Socratic reasoning in education (Herbel-Eisenmann and Breyfogle, 2005; Wang and Demszky, 2024; Alic

et al., 2022; Demszky and Hill, 2022). More recently, prior work (Al-Hossami et al., 2023b,a) has highlighted the poor performance of prompting-based methods in performing Socratic Reasoning for the education domain (Achiam et al., 2023), even with Chain-of-Thought (CoT) (Wei et al., 2022), as they often give away answers without asking clarifying questions, or the questions are unrelated to the student’s response or original bug (Achiam et al., 2023). In contrast, TreeInstruct mitigates this issue by explicitly grounding the question generation step on both a target state variable  $\tau$  and any Student *misunderstanding* gauged from their previous response.

### 2.3 LLMs for Interactive Education

Recent generative approaches within the AI tutoring space have attempted to generate responses which cater to the student’s type of mistake or request, but only in single-turn settings. CodeAid (Kazemitabaar et al., 2024) is an *assistive* tool that helps students debug their code. In their "Help Fix Code" and "Question from Code" modules, the Instructor provides single-turn responses to the Student for answering questions, explaining concepts, and helping to write code. However, these modules direct the Student towards where their mistake is and uses natural language to describe the bug fixes. In contrast, TreeInstruct aims to *instruct* the Student socratically through questions, such that even in natural language, the bug fixes are not provided. BRIDGE (Wang et al., 2023) is an Instructor-like framework that aims to help students with math mistakes. The LM estimates the type of error, the

strategy of error remediation, and the instructor intention behind the remediation (all are chosen from a predetermined set). However, our methodology makes use of a more structured planning approach that accounts for the inherent multi-turn nature of educational guidance.

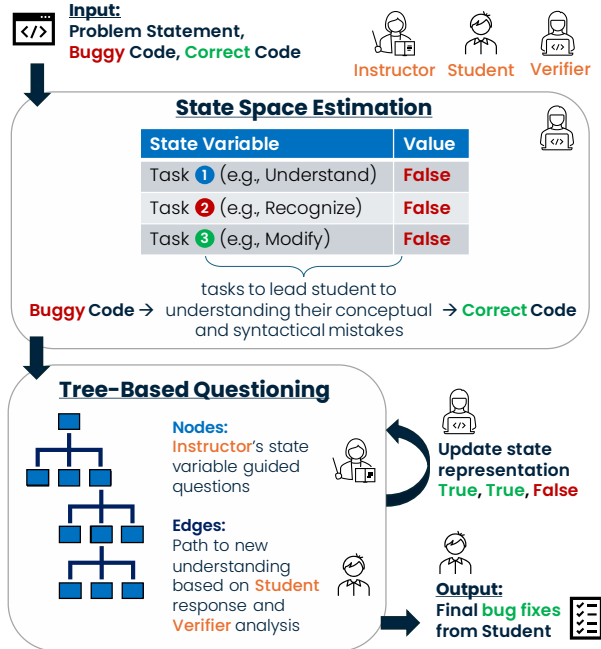


Figure 2: We propose TreeInstruct, a novel tree-guided instructional questioning framework for meaningful educational debugging guidance.

### 3 Methodology

As shown in Figure 2, **TreeInstruct** aims to dynamically guide the multi-turn conversation based on its estimated state space. Section 3.1.2 provides an overview of the three different agents we use and their respective roles during the state space generation/update and tree construction processes (outlined in Figure 3). This allows TreeInstruct to respond to the Student’s current level of understanding adequately. Algorithm 1 contains the pseudocode for all components in our method.

#### 3.1 Preliminaries

##### 3.1.1 Problem Description

As input, the Instructor is given the Student’s buggy code that contains  $e$  errors, a problem statement, bug descriptions, and their respective fixes. The Instructor guides the Student to generate a list of all bug fixes based on their interactions with the Instructor. The overall goal is for the Student to resolve their own conceptual and syntactical errors in a Socratic fashion to reach the correct code. Note

that we assume bug fixes are provided, a common scenario in educational settings (e.g., assignments, exams) where ground-truth solutions are available from human instructors or platforms like LeetCode. As our focus is steering an LLM towards Socratic guidance, generating these solutions for real-world tutoring applications is left for future work.

##### 3.1.2 Agents

In a real-world setting, a Socratic educator (e.g., an instructor, a teaching assistant) executes two tasks when interacting with a Student: (1) ask relevant questions to the Student, and (2) assess the Student’s understanding based on their responses. Following this cyclical pattern, we break down our educator into two roles: an Instructor and a Verifier, with persona prompts specified in Tables 12 and 14 in Appendix I, respectively. The Instructor and Verifier perform their respective tasks specified in Algorithm 1 via zero and one-shot prompting. The *Instructor* agent’s job is to generate questions to ask the Student (*GenerateQuestion*, *GenerateSiblingQuestion*, and *GenerateChildQuestion* in Alg. 1; details provided in Section 3.3). The *Verifier* agent has a significantly more involved role:

1. *State Space Estimation* (Section 3.2): The Verifier determines a set of tasks which will lead a Student to understanding and correcting their problem and buggy code. This is *GenerateState* in Alg. 1.
2. *Assess Student Response* (Section 3.3): Once the Student answers the Instructor’s question, the Verifier must judge the response’s accuracy, given the question-answer pair interaction. This is *VerifyResponse* in Alg. 1.
3. *Assess Student Understanding of Target State Variable* (Section 3.3): To update the Student’s state space representation, the Verifier must determine whether the Student would have needed a sufficient understanding of the target state variable in order to generate their response. This is *UpdateUnderstanding* in Alg. 1.
4. *Verify Student Bug Fixes* (Section 3.4): Each time the Student understands a target state variable, they are asked to provide, if any, recommended bug fixes based on the conversation history. This serves as an early stopping condition. This is *isResolved* in Alg. 1.





ated question intended therefore, the question can be rephrased or made more specific. To ensure this, we ground the question generation based on two things: (1) the previous questions from level  $l$ , and (2) the Verifier’s explanation for why the Student got the question wrong.

**Conditional generation of child questions.** The Instructor *conditionally generates* child questions at level  $l + 1$  if and only if the *Student correctly answers the Instructor question* (addresses the question and has no mistakes in their answer), but still does not understand the target state variable  $\tau_i$  (line 14 in Alg. 1). As shown in the fourth question of Figure 3, these questions aim to guide the Student to a more complete understanding of the target state. To ensure this, we ground the question generation on two things: (1) the previous questions from level  $l - 1$ , and (2) the Verifier’s explanation of the gaps in the Student’s target state understanding.

### 3.4 Adaptive Conversation Restructuring

Once the Verifier agent determines that the target state/task has been resolved, we exploit the same process to update all remaining tasks  $\tau \in S$ , as multiple dependent bugs may have been concurrently resolved within the same tree. After at least the target state variable has been resolved (line 13 in Alg. 1), we create a new tree for any remaining tasks, as shown in the first interaction of Figure 3. This step is crucial to the multi-bug setting, as independent bugs would benefit from having separate, distinct trees of questioning.

For further adaptiveness to the conversation, we additionally provide (1) an early stopping condition based on the Student’s intermediate bug fixes, and (2) a maximum tree width and depth threshold, after which TreeInstruct chooses to teach the Student their remaining gap in knowledge.

- *Bug fixes:* After a task  $\tau$  has been resolved, the Student is prompted to provide a list of natural language bug fixes (e.g. “Replace  $i$  with  $i+1$  on line 6.”) based on their entire conversation history with the Instructor. The Verifier will determine if all the ground-truth bug fixes have an *isomorphic* counterpart within the set of suggested Student bug fixes. Isomorphism can be defined as (1) having the same conclusion or output, (2) sharing the same underlying logical structure or pattern, and/or (3) being convertible to each other through a series of logical transfor-

mations<sup>2</sup>. If all ground-truth bug fixes have been resolved, then we may stop early.

- *Teaching:* After generating a maximum number of sibling questions  $q$  or depth  $l$ , the Instructor appends the correct answer to  $Q[l][0]$  and re-ask  $Q[l][-1]$  to the Student. This ensures that the conversation flows in case the Student gets stuck.

## 4 Experiments

### 4.1 Experimental Setup

In order to evaluate TreeInstruct, we utilize a proxy *Student agent* based on the Mistral-7B-Instruct model (Jiang et al., 2023) to mimic the abilities of a student while responding to the Instructor. The prompt we use to define the Student persona is outlined in Table 13 of Appendix I. We additionally provide GPT-4 API experimental setup details in Appendix G.

### 4.2 Datasets




We evaluate our method on two datasets. First, we use the Socratic Debugging Benchmark dataset from (Al-Hossami et al., 2023b), which consists of 149 problems—each with a problem statement, student buggy code, bug fixes and descriptions in English, and correct code. However, these problems lack sufficient difficulty, often requiring small fixes and minimal problem comprehension. To evaluate TreeInstruct on more challenging problems, we craft a novel dataset, **MULTI-DEBUG**, based on 50 popular programming problems<sup>3</sup> (16 easy, 29 medium, and 5 hard according to their corresponding LeetCode labels). It features longer problems and more involved concepts, requiring TreeInstruct to have more extensive reasoning capabilities for guiding the Student.

For each problem, expert annotators (Appendix D) injected 1, 2, and 3 syntactical or conceptual bug(s) that a typical student would make (a total of 150 different samples for MULTI-DEBUG). Conceptual bugs usually cause runtime errors or result in incorrect output. Examples include misunderstanding the problem statement, encountering an infinite loop, or incorrectly using a library or mathematical operator ( $/$  vs  $//$  in Python). Syntactical bugs cause compilation errors due to incorrect

<sup>2</sup>For example, the bug fixes “Replace `if i <= 0` with `if i < 0`” and “Check if `i` is strictly less than 0” are isomorphic because they have the same conclusion and share the same underlying pattern of the mathematical operator  $<$ .



<sup>3</sup><https://github.com/Garvit244/LeetCode>

Table 1: Results on Socratic Debugging Benchmark (Single Bug). **Bolded** and † values denote the top 2 methods.

Methods	#T	Syntactical (42 samples)				Conceptual (107 samples)			
		Success	Relevant	Indirect	Logic	Success	Relevant	Indirect	Logic
Vanilla 	3.23	<b>80.95</b>	83.72†	76.19	78.70†	76.64†	87.35†	80.32†	78.79†
Bridge 	6.00	78.57†	76.50	82.24†	41.72	62.14	78.12	79.86	34.38
<b>TreeInstruct</b> 	5.41	77.27	<b>92.01</b>	<b>96.48</b>	<b>88.95</b>	<b>80.26</b>	<b>95.63</b>	<b>89.10</b>	<b>94.63</b>

Python syntax (e.g., missing a colon). For each bug, we keep track of its fix and description.

### 4.3 Baselines

We compare TreeInstruct to several baselines and their variants. Given that LLM-based Socratic instruction is a new evolving area with few existing work and no multi-turn methods, we adapt a prompting and an existing single-turn method to our task. The first baseline, **Vanilla**, is given the same input as TreeInstruct’s Instructor. The base model is prompted to ask Socratic questions to the Student—it does not use any explicit conversational structure or estimate the Student’s knowledge. We use both Meta-Llama-3-8B-Instruct  (Touvron et al., 2023) and GPT-4  (Achiam et al., 2023) as base models for the Vanilla baseline.

Second, we use **BRIDGE** (Wang et al., 2023). Since we are adapting this for Socratic code debugging, we use their predetermined error types, remediation strategies, and remediation intentions to guide the question generation, along with the problem-specific input given to TreeInstruct’s Instructor. For both baselines, we limit the conversations to 20 turns per number of bugs.

### 4.4 Evaluation Metrics:

We perform qualitative and quantitative evaluation of our methods. The scores are averaged across all turns and then averaged across all problems. In the result tables, we scale the scores by 100.

**Qualitative:** We develop a binary scale to assess the Socratic quality of questions. (Al-Hossami et al., 2023b) identifies multiple dimensions of Socratic questioning, including relevance to specific bugs, implicitness of the answer, and structural coherence. For each question, we measure the below attributes of the conversation manually (giving a score of 1 if the attribute is met, and 0 otherwise):

- **Relevance (Relevant):** The instructor’s question was pertinent to the errors in the student’s code.

- **Indirectness (Indirect):** The instructor’s question refrained from directly revealing solutions to the bugs.
- **Logical Flow (Logic):** The instructor’s question promoted a coherent conversation, facilitating the student’s problem-solving process.

**Quantitative:** We apply quantitative metrics to objectively evaluate the effectiveness and efficiency of our framework.

- **Overall Success Rate (Success):** We check whether the final list of bug fixes generated by the Student,  $B_S$ , and the ground truth set of big fixes,  $B_{GT}$ , are isomorphic (Section 3.4). The success rate is calculated as  $|B_{GT} \cup B_S|/|B_{GT}|$ .
- **Average Number of Turns (#T):** We compute the average number of turns required by the method to reach the goal state. This metric provides insight into the efficiency and depth of the interaction process.

### 4.5 Overall Results





In Tables 1 and 2, we see that with **more structured representations of student knowledge and conversation state**, TreeInstruct demonstrates **significant improvements beyond the baselines**. Across all multi-bug settings, we see an overall improvement of 16.6% and 11.59% in the success rates for syntactical and conceptual bugs, respectively. We also see an improvement of 13.47% and 14.89% for syntactical and conceptual bugs, respectively, across the three conversation metrics. For 1-bug, we see that the Vanilla  baseline has the highest success for conceptual bugs, *and* the lowest Indirectness score, indicating that questions were very direct, and gave hints towards the bug fixes, which evidently increased the success rate. We see the same trend in the syntactical, single bug on Vanilla  setting in Table 1. Overall, TreeInstruct demonstrates **strong performance despite drastically different base models**,  and . TreeInstruct’s runtime performance is detailed in Appendix G.3.

Table 2: Results on the **MULTI-DEBUG** dataset. In total, 1-bug has 29 syntactical and 21 conceptual bugs, 2-bug has 50 syntactical and 50 conceptual bugs, and 3-bug has 78 syntactical and 72 conceptual bugs. **Bolded** and † values denote the top 2 methods, respectively. The bottom two rows are ablation studies performed on 3-Bug setting.


















Bugs	Methods	#T	Syntactical				Conceptual			
			Success	Relevant	Indirect	Logic	Success	Relevant	Indirect	Logic
1	Vanilla 	2.36	71.43	92.16	55.12	84.15	<b>78.57</b>	94.58	59.17	84.17
	BRIDGE 	16.60	50.00	93.93	<b>98.04</b>	24.23	68.00	97.27	<b>96.67</b>	35.38
	TreeInstruct 	7.24	<b>76.19</b>	93.98†	94.08	85.28†	71.43	97.57†	93.02†	86.02†
	TreeInstruct 	3.94	75.00†	<b>100.00</b>	95.59†	<b>96.63</b>	76.92†	<b>100.00</b>	88.01	<b>94.76</b>
2	Vanilla 	8.32	53.26	83.45	74.41	60.82	62.50	86.96	74.13	59.90
	BRIDGE 	15.28	34.88	89.47	89.33	52.40	42.71	89.67	88.06	46.64
	TreeInstruct 	9.04	66.67†	93.00†	92.17†	84.59†	72.62†	94.15†	92.58†	81.46†
	TreeInstruct 	6.14	<b>69.32</b>	<b>97.96</b>	<b>98.47</b>	<b>90.14</b>	<b>73.91</b>	<b>99.58</b>	<b>98.47</b>	<b>94.45</b>
3	Vanilla 	17.48	44.00†	69.88	64.31	52.38	67.00	84.68	84.68	41.51
	BRIDGE 	8.44	19.00	87.78	83.95	64.95	43.00	90.09	85.78	44.65
	TreeInstruct 	10.46	43.00	95.68†	88.88	80.94†	72.00†	96.76†	<b>97.95</b>	83.28†
	TreeInstruct 	10.46	<b>73.00</b>	<b>100.00</b>	<b>99.27</b>	<b>95.57</b>	<b>92.00</b>	<b>98.40</b>	95.89†	<b>93.63</b>
	 No State	16.34	25.51	51.61	97.21	41.09	53.00	47.57	94.70	20.36
	 No Teaching	9.69	30.61	90.75	97.61†	72.84	50.00	94.62	95.17	68.78

Table 3: Results on the side-by-side evaluation. **Bolded** and † values denote the top 2 comparisons, respectively. Note: S-bug refers to the Socratic Debugging Benchmark. We abbreviate TreeInstruct as TI.


Comparison	3-bug	2-bug	1-bug	S-bug
TI vs BRIDGE	71.43†	68.00†	<b>83.67</b>	<b>94.63</b>
TI vs Vanilla	<b>100.00</b>	<b>90.00</b>	69.39†	50.33†
BRIDGE vs Vanilla	57.14	62.00	40.82	24.83

**Side-by-Side Evaluation:** Using the same evaluation metrics (Section 4.4), we performed a side-by-side evaluation that measures how often a user prefers TreeInstruct over the baselines. The results are shown in Table 3, in which we see that on average, TreeInstruct is given a **higher ranking than BRIDGE 79.43%** of the time, and a **higher ranking than Vanilla 77.43%** of the time. The details of this computation are in Appendix C. When specifically comparing TreeInstruct against BRIDGE, the key differences were that BRIDGE began with specific questions which revealed bug fixes and ended with general questions, lacking a logical flow in the conversation. In addition to being limited in a multi-bug setting, it posed less effective follow-up questions. In contrast, TreeInstruct prioritized high-level conceptual questions early on in order to build the foundation for later


code-specific questions. Additionally, it asked for new information each turn and handled multiple bugs effectively. This is explicitly demonstrated in the example shown in Appendix B, Table 4.



**Human Student Interaction:** We conducted a separate case study where human students directly interacted with TreeInstruct. We gathered five human volunteers of varying levels of programming backgrounds and knowledge, and presented each volunteer with three Single Bug problems and three 3-bug (MULTI-DEBUG) problems. The study shows that TreeInstruct can adapt to various levels of students effectively, as the scores are mostly comparable between our volunteers. The volunteers mentioned that **TreeInstruct helped them learn programming concepts by forcing them to critically think about their mistakes**, instead of *trying random solutions*. Also, our volunteers had better interactions with TreeInstruct  than TreeInstruct  because the  Verifier is of higher quality. Appendix F includes further details of this study, with quantitative results in Table 10.



#### 4.6 Ablation Studies

We perform two ablations on all 50 3-bug problems (bottom two rows of Table 2). For  **No State**, we remove the state space representation, basing the question generation on the conversation history,



Verifier feedback on the Student’s answer, and the Student-proposed bug fixes. For  **No Teaching**, we remove the teaching functionality that kicks in after three consecutive incorrect Student answers. The conversation is still guided by the state space representation, tree-based questioning, its updates, and the Student-proposed bug fixes.


When compared to TreeInstruct , **No State success rates drop by 18.25%, Relevance by 46.63%, and Logic by 51.39%**. Without the grounding provided by the state space representation, the conversations deviate from the real bugs and contain many repeated questions that the Student already answered. Additionally, when compared to TreeInstruct , **No Teaching success rates and Logic scores for drop by 17.20% and 11.32%, respectively**. When the Student does not know some base-level of foundational knowledge, it is better to break the unnecessary cycle of asking questions and allow the Instructor to teach these concepts. Overall, these performance differences demonstrate the impact of each component.

**Verifier Subtask Performance:** To understand the Verifier’s abilities to perform its designated subtasks (Section 3.1.2), given their complexity, we evaluate the Verifier on each subtask for the 3-bug setting. We demonstrate that for both  and , on the most difficult setting, both Verifier models feature high accuracy rates across all subtasks (86.3% average performance across all subtasks), most notably `generateState`, which shows that the base model does not have significant impact. Overall, this shows that TreeInstruct can reliably depend on the Verifier to complete such tasks. The detailed results are included in Appendix A.

#### 4.6.1 Analysis

**Fine-grained and dependency-aware state spaces improve conversation quality.** In our evaluation, we saw that state spaces typically contain 4–5 state variables, evenly distributed across all bugs. In other words, the number of bugs is inversely proportional to the number of variables allocated to each bug. Thus, 1-bug uses all state variables—each featuring a more fine-grained subtask to solve the bug, whereas 2- or 3-bug would use fewer variables for each bug (example provided in Appendix E). 1-bug’s fine-grained state variables lead to increased Success and Relevance scores as the conversation delves deep into the Student’s root misunderstanding of the bug. On the other hand,

3-bug features more coarse-grained state variables with a higher *inter-bug dependency*. Consider the (recursive) Fibonacci problem in Figure 1. Suppose the student is missing a base case and incorrectly calling the recursive function. Solving one bug requires adequate understanding of the other, thereby making it easier to solve. Another example can be found in Appendix B, Table 4. The 2-bug setting does not *fully* experience the benefits of fine-grained state variables or high bug-interdependency, resulting in a slight dip in scores: on average, a 7.62%, 0.83%, and 2.85% drop in Success, Relevant, and Logic scores, respectively—nonetheless, better than the baselines.

**TreeInstruct tackles challenging problems more effectively.** Given the increase in problem difficulty between (1) the Socratic Debugging Benchmark (SDB) and MULTI-DEBUG (MD) datasets, and (2) from the 1 to 3-bug settings within MD, TreeInstruct is able to tackle challenging problems more effectively than baselines. The state space estimation breaks down problems into simple subtasks which the Instructor and Verifier can target using tree-guided questioning, even in complex conversations. Concretely, TreeInstruct’s performance in the SDB setting (Table 1) is comparable to that of the MD setting (Table 2), with only a 2.44% drop in performance from SDB to MD 1-bug and 1.53% drop from MD 1-bug to MD 3-bug. Comparatively, Vanilla  drops by 14.55% from 1-bug to 3-bug. Overall, TreeInstruct’s Relevant, Indirect, and Logic scores remain high regardless of difficulty.

## 5 Conclusion

This paper proposes a novel method, **TreeInstruct**, for state space estimation and dynamic tree-based questioning for multi-turn Socratic instruction. We construct a novel multi-bug debugging dataset, **MULTI-DEBUG**, with 150 expert-annotated conceptual and syntactical problems and buggy solutions/fixes. Extensive experiments on an existing benchmark and MULTI-DEBUG demonstrate that TreeInstruct can be universally applied to both open and closed source models. We also showcase that TreeInstruct’s strong Socratic questioning abilities widely outperform all baselines through both (1) rigorous quantitative and qualitative expert evaluation (preferred over 77.94% of the time), and (2) real-world interactions with students of varying coding abilities.

## 6 Limitations & Future Work

While TreeInstruct provides an intuitive framework which demonstrates promising results for effective multi-turn Socratic instruction, it contains a few limitations that form the foundation for future, impactful research areas.

Firstly, Tables 1 and 2 shows high qualitative scores for the questions asked by TreeInstruct. While these are encouraging, the success rates still have large room for improvement—the highest success rate is 77.27%. This indicates that Socratic questions alone are not sufficient for teaching a student to debug their code. We judge the efficacy of questions locally, whereas the next step would be to judge them globally across the conversation. We leave it to future work to devise an effective global questioning scheme and evaluation metric.

Additionally, our method is dependent on the base model’s reasoning capabilities, specifically for the Verifier agent. In our results, with a stronger model, we see higher scores for Logic and Success. Although our method shows comparable results between GPT-4 and Llama-3-8b, this may be a bottleneck, as stronger and bigger models require a higher deployment cost.

Next, in the few failure cases, we see some adverse effects of our method’s reliance on the reasoning capabilities of the base model. First, our method can get stuck into a cyclical conversation with the Student if they are particularly weak in an area and cannot understand the target state even after multiple rounds of direct questioning and teaching. In these cases, the number of turns rises to 20-30.

Moreover, syntactical bugs might be “harder to see” for the language model, as it goes against the generation process to generate syntactically incorrect code. Breaking it down, a language model is trained to generate code with a colon at the end of for loops, if-statements, and method signatures, so if buggy code has a missing colon, the language model might ignore it. This results in syntactical bugs being harder to solve.

These limitations give way to exciting future work. Firstly, we can make use of vision language models to provide students with multi-modal teaching strategies, instead of relying solely on language. Additionally, we can enhance the framework, so it will explore new instruction methods when the questioning becomes cyclical. This can also help make the Instructor more reliable to gen-

erate consistent output across multiple runs on the same problem. Furthermore, we can utilize a structured fine-tuning approach to help the model better leverage the Verifier feedback and tree-based question generation process to make hierarchical Socratic planning and questioning inherent to a model. Overall, TreeInstruct can also be extended to automatically generalize to different teaching domains (e.g., quantitative reasoning).

## 7 Ethics Statement

We are committed to the transparency and reproducibility of our research. We encourage our research community to make use of our open-source code and dataset to further improve our methodology. Our research involves the evolving intersection of large language models (LLMs) and education, where the deployment of language model instructors and their interactions with students have been relatively unexplored. The role of technology and language models is being widely discussed with respect to its impact on student dependence and lack of critical thinking. Given the rapid and wide-scale deployment of LMs to the public, we emphasize the importance of designing Socratic dialogue systems in the hopes of bettering educational support for all students and educators.

## 8 Acknowledgements

This research project has benefited from the Microsoft Accelerate Foundation Models Research (AFMR) grant program, through which leading foundation models hosted by Microsoft Azure and access to Azure credits were provided to conduct the research. Furthermore, we would like to thank Mihir Kavishwar, Krish Agarwal, Sonia Agarwal, Nirav Diwan, and Shradha Sehgal for their help and feedback on our work.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Erfan Al-Hossami, Razvan Bunescu, Justin Smith, and Ryan Teehan. 2023a. Can language models employ the socratic method? experiments with code debugging. *arXiv preprint arXiv:2310.03210*.
- Erfan Al-Hossami, Razvan Bunescu, Ryan Teehan, Laurel Powell, Khyati Mahajan, and Mohsen Dorodchi.

- 2023b. Socratic questioning of novice debuggers: A benchmark dataset and preliminary evaluations. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 709–726.
- Sterling Alic, Dorottya Demszky, Zid Mancenido, Jing Liu, Heather Hill, and Dan Jurafsky. 2022. Computationally identifying funneling and focusing questions in classroom discourse. *arXiv preprint arXiv:2208.04715*.
- Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4:253–278.
- Kathleen Cotton. 1988. Classroom questioning. *School improvement research series*, 5:1–22.
- Dorottya Demszky and Heather Hill. 2022. The ncte transcripts: A dataset of elementary math classroom transcripts. *arXiv preprint arXiv:2211.11772*.
- Beth A Herbel-Eisenmann and M Lynn Breyfogle. 2005. Questioning our patterns of questioning. *Mathematics teaching in the middle school*, 10(9):484–489.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Z 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. *arXiv preprint arXiv:2401.11314*.
- Wenhan Lyu, Yimeng Wang, Tingting Rachel Chung, Yifan Sun, and Yixuan Zhang. 2024. Evaluating the effectiveness of llms in introductory computer science education: A semester-long field study. *arXiv preprint arXiv:2404.13414*.
- Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. *Advances in neural information processing systems*, 28.
- Anna N Rafferty, Emma Brunskill, Thomas L Griffiths, and Patrick Shafto. 2016. Faster teaching via pomdp planning. *Cognitive science*, 40(6):1290–1332.
- Husniah Sahamid. 2016. Developing critical thinking through socratic questioning: An action research study. *International Journal of Education and Literacy Studies*, 4(3):62–72.
- Brad Sheese, Mark Liffiton, Jaromir Savelka, and Paul Denny. 2024. Patterns of student help-seeking when using a large language model-powered programming assistant. In *Proceedings of the 26th Australasian Computing Education Conference*, pages 49–57.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Rose E Wang and Dorottya Demszky. 2024. Educonvokit: An open-source library for education conversation data. *arXiv preprint arXiv:2402.05111*.
- Rose E. Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2023. [Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes.](#)
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Judith D Wilson. 1987. A socratic approach to helping novice programmers debug programs. *ACM SIGCSE Bulletin*, 19(1):179–182.
- Ya-Ting C Yang, Timothy J Newby, and Robert L Bill. 2005. Using socratic questioning to promote critical thinking skills through asynchronous discussion forums in distance learning environments. *The american journal of distance education*, 19(3):163–181.
- Michael V. Yudelson, Kenneth R. Koedinger, and Geoffrey J. Gordon. 2013. Individualized bayesian knowledge tracing models. In *Artificial Intelligence in Education*, pages 171–180, Berlin, Heidelberg. Springer Berlin Heidelberg.

## A Breakdown of the Verifier Agent’s Subtask Performance

We perform ablation studies for each module on our most difficult setting (MULTI-DEBUG 3-bug), with both GPT-4 and Llama-3-8B Verifier Agents. For the sake of time, we used 20% of the dataset. For each subtask (generateState, verifyResponse, updateUnderstanding, isResolved), we define the following metrics of success:

Table 4: A comparison of questions asked by BRIDGE and TreeInstruct - between Q1 and Q2, and Q2 and Q3, the Student gave a correct answer.

	BRIDGE	TreeInstruct
Q1	What is the difference between a list and a set in Python, and how might this affect the performance of your code when checking if an item is in 'temp'?	Can you explain the difference between a list and a set in Python, and how their properties might affect the functionality of your code?
Q2	Can you explain why you chose to use a list to store the characters of the string and what might be the implications of this choice?	That is correct! Let's take a step further. Given that a set only contains unique items, how might this property of a set be useful in solving this problem?
Q3	N/A	Can you explain how you are currently adding elements to your 'temp' collection in your code?

- **generateState**: the proportion of bugs that would be sufficiently resolved if the tasks (state space variables) were followed by the Student.
- **verifyResponse**: a binary 0/1 for correct Verifier judgment and 0/1 for the correct explanation behind the Verifier's judgment:  $\frac{\text{verification\_score} + \text{explanation\_score}}{2}$ .
- **updateUnderstanding**: the proportion of state variables that were correctly validated (whether to update the state variable) and 0/1 for the correct explanation behind the Verifier's judgment:  $\frac{\text{validation\_score} + \text{explanation\_score}}{2}$ .
- **isResolved**: the proportion of Student proposed bug fixes which the Verifier correctly identifies as isomorphic/non-isomorphic to the corresponding ground truth bug fixes.

The results are in Table 5. Despite this being our most difficult setting, both Verifier models feature high accuracy rates across all subtasks, most notably `generateState`, which shows that the choice of base model does not have significant impact. Note: `isResolved` is primarily used for early stopping (e.g., if a student proposes all correct bugs during the first turn, then TreeInstruct stops early), so this subtask is not necessary. This subtask simply reduces computational costs.

Table 5: Performance of each Verifier subtask.

Subtasks	Llama-3-8B	GPT-4
<code>generateState</code>	90.0	100.0
<code>verifyResponse</code>	85.7	86.7
<code>updateUnderstanding</code>	88.6	85.9
<code>isResolved</code>	72.2	81.4

## B Code Bug Interdependency

Table 4 shows an example of the effect of interdependent bugs, taken from real conversations that BRIDGE and TreeInstruct had with our student simulator `Mistral-7B-Instruct`. The questions aim to solve: (1) using a set instead of a list, and (2) using `.add()` instead of `.append()` to add items. As shown, Questions 1 and 2 tackle Bug 1, and Question 3 tackles Bug 2. The interdependency of changing the data structure first makes it easier to resolve the second bug, yielding higher success rates for the 3-bug setting.

## C Side by Side Evaluation

As mentioned in the main text, we perform a side-by-side evaluation to measure the percentage of times a user prefers our method TreeInstruct over the baselines. Preference was measured as the average of all conversation metrics across syntactical and conceptual bugs. Based on the metrics, we assign each method a ranking (1, 2, or 3). Table 3 shows that TreeInstruct was preferred 68-94.6% of the time over the baselines. On average, TreeInstruct was preferred over BRIDGE 79.43% of the time, and over Vanilla 77.43% of the time.

**Interpretation.** When we say TreeInstruct was preferred 79.43% more over BRIDGE, this means that across all 50 3-bug problems and ranking configurations, TI was given a higher ranking than BRIDGE (TI is ranked #1 while BRIDGE is ranked #2, TI is ranked #1 while BRIDGE is ranked #3, TI is ranked #2 while BRIDGE is ranked #3) 79.44% of the time. Each of the 50 problems can have multiple preferences (TI over BRIDGE, TI over Vanilla, Bridge over Vanilla, etc.) which is why they will not necessarily add up to 100.

Table 4 shows a comparison between the questions asked by BRIDGE and TreeInstruct. Ap-



Table 6: Inter-annotator agreement on the relevance, indirectness, logic scores, and bug type determination. Conducted on a random subset of the Single-Bug (SB) and MULTI-DEBUG 3-Bug (3-Bug) datasets.

Setting	Cohen’s $\kappa$				% Agreement			
	Success	Relevant	Indirect	Logic	Success	Relevant	Indirect	Logic
<b>SB</b>	61.97	71.88	76.62	1.0	91.67	94.44	88.89	1.0
<b>3-Bug</b>	1.0	78.90	78.90	1.0	1.0	97.83	97.83	1.0

pendix B specifies the details behind the targeted problem in the Table. On the left, we see that BRIDGE asks very similar questions regarding the design choice of using a list versus a set. TreeInstruct, on the right, asks one question regarding the design choice, one question about how to add elements to the correct data structure, and one intermediate question about why the design choice is suited towards the problem. Updating the state space, grounding questions towards the Student’s capabilities, and connecting concepts back to the problem are principles that rank TreeInstruct higher than BRIDGE.

## D Human Expert Evaluators

As mentioned before, the injected bugs into the dataset (Section 4.2) and evaluation in Tables 1 and 2 were obtained using volunteer human expert evaluators: two computer science teaching assistants with at least four years of high-school, undergraduate, and graduate-level teaching experience, with proficiency in Python and located in the USA. Their experiences equipped them with the first-hand experience of mistakes made by beginner programmers, leading to more realistic injected bugs. Furthermore, for evaluation, they were given the same set of instructions and the following set of guidelines:

- Assign a score of 1 for Relevance if the question will eventually lead the Student to understand their bug(s).
- Assign a score of 0 for Indirect if a question explicitly or implicitly states a solution.
- Assign a score of 0 for Logic if the current question does not naturally flow from the Student’s previous answer.

They were given a CLI-based annotation system. In our GitHub repository, we have provided the script ([question\\_human\\_evaluation.py](#)) and an example of the evaluation interaction ([eval-](#)

[uation\\_example.pdf](#)). Below are some special cases/considerations the evaluators were also given:

- If the Verifier is wrong and asks the same question despite the Student getting the question correct, give a score of 0 for Relevance.
- If a question seems out of order, give a score of 0 for Logic.
- If a question deep into the conversation is vague, gives a score of 0 for Relevance and Logic.
- If the answer is provided in a hint after 2 rounds of similar questions, and the Student still does not understand, do not penalize the Instructor for Indirect.
- For determining Success, do not penalize the Student if the bug fix is in natural language rather than code.

On a random subset of the Single-Bug and MULTI-DEBUG 3-Bug datasets, we computed the inter-annotator agreement on the relevance, indirectness, and logic scores (Section 4.4) and bug type determination. We show these agreement results in Table 6.

## E Comparing State Space Representations in Multi-Bug Settings

Here, we compare the state space representations of the 1-bug, 2-bug, and 3-bug settings for the two sum problem. In the two sum problem, given is an array of integers and a target value. The goal is to return the indices of two numbers that add up to the target value. Below is the correct code.

```

1. def twoSum(self, nums, target):
2.     d = {}
3.     for i in range(len(nums)):
4.         difference = target - nums[i]
5.         if difference in d:
6.             return [d[difference], i]
```

```

7.     d[nums[i]] = i
8.     return d

```

In the 1-bug setting, the Student mistakenly writes `nums[i]-target` instead of `target-nums[i]` on line 4. In the 2-bug setting, along with the previous bug, the Student also initializes `d` as a list (`d=[]`) instead of a dictionary on line 2. Finally, in the 3-bug setting, the Student forgets to add a colon at the end of the if-statement on line 5.

Tables 7, 8, and 9 outline the state space representations for the 1-bug, 2-bug, and 3-bug settings. As shown, 1-bug uses 3 states (states 1, 2, and 3) to solve the same but that 3-bug uses 1 state (state 1) to solve. This means the 1-bug state representation is much less compact than that for 3-bug.

1. Understand the problem statement and the requirement to find two numbers that add up to a specific target.
2. Understand the logic behind calculating the difference as `target - nums[i]`.
3. Correctly implement the difference calculation in the code.

Table 7: State space representation for 1-bug on the two-sum problem.

1. Understand how to correctly calculate the difference between the target and the current number in the array.
2. Understand the difference between lists and dictionaries in Python.
3. Correctly initialize a dictionary in Python.
4. Understand how to use a dictionary to store and retrieve values in Python.



Table 8: State space representation for 2-bug on the two-sum problem.

## F Interactive Evaluation with Human Students

For our main evaluation, we used `Mistral-7B-Instruct` to represent a Student. We noticed that Mistral is an overconfident model that (1) suggests incorrect bug fixes in between the conversations and (2) jumps to fix bugs that do not exist in the code. Therefore, we worked with human students to test our method on the following two settings: Socratic Debugging

1. Understand how to correctly calculate the difference as `'target-nums[i]'`.
2. Understand how to initialize a dictionary using `''` instead of `'[]'`.
3. Understand how to use a dictionary to store and retrieve values.
4. Understand the correct syntax for an if-condition, including the necessary colon at the end.

Table 9: State space representation for 3-bug on the two-sum problem.



on `TreeInstruct`  and 3-bug on `TreeInstruct` . We gathered 5 human volunteers of varying levels of programming backgrounds and knowledge (ensuring to anonymize their identities):

- **Level 1:** Student knows how `TreeInstruct` works; they act as an adversary to intentionally provide bad inputs that will try to make the method fail.
- **Level 2:** Student is new to `TreeInstruct`; they are a basic programmer who has been learning to code in Python for a few months.
- **Level 3:** Student is new to `TreeInstruct`; they are a non-computer science major who does not use Python often, but knows the basic high level concepts of data structures and syntax.
- **Level 4:** Student is new to `TreeInstruct`; they have been using Python for 2 years and are in their final year of undergraduate education in computer science.
- **Level 5:** Student knows how `TreeInstruct` works; they act as an ally to intentionally provide good inputs, so the method can resolve the bugs in as little turns as possible.

When conducting the study, we adhered to the following experimental process:

1. We presented the student with the problem statement and gave them as much time as they needed to fully understand it.
2. The students were given two minutes to review the buggy code. We noted down how many bugs each of the students were able to identify before their conversation.



Table 10: Results of human student evaluation across S(ingle)-bug (Socratic Debugging benchmark) and 3-bug (MULTI-DEBUG dataset) settings, broken down by the student level (Lvl).

Bugs	Lvl	Avg. Turns	Syntactical				Conceptual			
			Success	Relevant	Indirect	Logic	Success	Relevant	Indirect	Logic
S-bug 	1	6.0	<b>100.00</b>	66.67	66.67	<b>100.00</b>	<b>100.00</b>	91.67 <sup>†</sup>	<b>100.00</b>	50.79
	2	12.0	<b>100.00</b>	66.67	83.33 <sup>†</sup>	75.00 <sup>†</sup>	50.00 <sup>†</sup>	<b>100.00</b>	<b>100.00</b>	50.00
	3	8.0	0.00 <sup>†</sup>	87.50 <sup>†</sup>	<b>100.00</b>	50.00	<b>100.00</b>	67.50	90.00 <sup>†</sup>	42.50
	4	1.0	<b>100.00</b>	<b>100.00</b>	0.00	<b>100.00</b>	<b>100.00</b>	57.14	<b>100.00</b>	64.29 <sup>†</sup>
	5	1.0	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>75.00</b>
3-bug 	1	19.0	83.33 <sup>†</sup>	75.93	97.92 <sup>†</sup>	74.77	<b>100.00</b>	<b>100.00</b>	88.89 <sup>†</sup>	79.49
	2	11.7	83.33 <sup>†</sup>	<b>100.00</b>	<b>100.00</b>	78.57	<b>100.00</b>	<b>100.00</b>	86.67	82.50
	3	6.67	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	85.71 <sup>†</sup>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>
	4	4.7	<b>100.00</b>	93.33 <sup>†</sup>	<b>100.00</b>	76.67	<b>100.00</b>	<b>100.00</b>	83.33	88.89 <sup>†</sup>
	5	3.0	<b>100.00</b>	<b>100.00</b>	83.33	<b>100.00</b>	<b>100.00</b>	83.33 <sup>†</sup>	<b>100.00</b>	83.33

- The students conversed with TreeInstruct until they were able to identify all the bugs present in the code.

We provide the results of this interactive study in Table 10. We used the same three single and 3-bug questions for all students, leading to 30 human student interactions in total. We also conducted a post-interaction interview with each of the students and provide an overview of their feedback below:

**Socratic questioning helped students learn programming concepts.** The Level 3 student stated that, "If there was no conversation, I would be put off from attempting to fix and just try a bunch of different things based on the errors." Overall, students of Levels 2-4 (students with no knowledge of the system) were not able to identify all the bugs before their interactions, but ended up solving them independently under the Socratic guidance of TreeInstruct.

**Underlying model had a significant impact on user experience.** Students had a significantly better experience with TreeInstruct  compared to TreeInstruct . Specifically, the quality of the Verifier determined whether the questions posed by the Instructor would be overly repetitive or not.

## F.1 Analysis

Table 10 contains the results. We see that from Level 1 to Level 5, the conversation have fewer turns, especially in the 3-bug setting. Additionally, we see that syntactical bugs are harder to solve for weaker students (on average, a success rate of 86.67%), which is intuitive as these students do not

have a strong foundation in Python syntax. On the other hand, conceptual bugs are easier to solve (on average, a 95% success rate). Overall, the results show that our method can adapt to various levels of students effectively.

## G Model Inference Experimental Setup

### G.1 GPT-4 API

For GPT-4, we made use of OpenAI’s GPT-4 API. Overall, we use temperature sensitivity  $t = 0$  for all generation tasks, except for  $t = 0.1$  for state space estimation and  $t = 0.3$  for instructor question generation. Using \$30 / 1M input tokens and \$60 / 1M output tokens, we break down the cost for each method. TreeInstruct uses an average of 35,000 input tokens and 4,000 output tokens, which adds up to **\$1.29 per conversation**. BRIDGE uses an average of 18,000 input and 5,500 output tokens, which adds up to **\$0.87 per conversation**. Vanilla uses an average of 31,000 output and 2,200 output tokens, which adds up to **\$1.06 per conversation**.

### G.2 Mistral and Llama

We run the Mistral-7B-Instruct-0.2 and Llama models locally on 2 NVIDIA-RTX A6000 GPUs. For one pass on a dataset (i.e., 150 problems/conversations), TreeInstruct takes approximately 4 hours.

### G.3 Runtime Comparison

We present the per-turn runtime for TreeInstruct (TI) across all multi-bug settings in Table 11. Llama was run on 1 NVIDIA Tesla V100 32GB, and OpenAI’s GPT-4 API was used for serverless inference. While the table indicates that Llama takes approximately 1 min per turn, this is due to

a hardware bottleneck. With better hardware, TI-Llama should be able to run much faster to provide close to real-time tutoring. This is indicated by TI-GPT4 taking approximately 7 seconds per turn.

Table 11: Average runtime (in seconds) TreeInstruct takes for each turn, for each setting.

Setting	Llama-3-8B	GPT-4
1-bug	56.27	7.09
2-bug	65.69	7.77
3-bug	61.22	7.65

## H License

All the datasets used in this work, including our own, are under the Apache 2.0 License. Our use of existing artifact(s) is consistent with their intended use, specifically for the Socratic Debugging benchmark and in general, programming practice and feedback for the problems used in the MULTI-DEBUG dataset.

## I Prompts

A few of the prompts use one-shot learning, and the fields are prefixed with "example". These examples are hand chosen, with no criteria in mind. The example problem relates to a solution that outputs the Fibonacci sequence of length  $n$ , where  $n$  is the input. We provide the specific prompts starting from the next page.



You are an Instructor helping a Student debug their code to solve the following problem statement (after tag 'problem'). You have access to their buggy code (after tag 'bug\_code'). Do not ask questions that explicitly or implicitly mention the following:

Table 12: Instructor agent persona prompt

You are a Student writing code to solve the above problem statement (after tag 'problem'), and you have written the below buggy code (after tag 'buggy\_code'). You are seeking help from your Instructor help solve your 'buggy\_code'. Your role is to answer the questions that the Instructor asks you as if you were an introductory programmer with a beginner's level of coding knowledge.

Table 13: Student agent persona prompt

You are an assistant to the Instructor helping a Student debug their code to solve the following problem statement (after tag 'problem'). Your role is to determine the Student's understanding (or lack thereof) within the Instructor-Student interactions. You have access to the correct code (after tag 'correct\_code'). Assume the Student is a introductory programmer with a beginner's level of coding knowledge.

Table 14: Verifier agent persona prompt

Given the student's buggy code (after tag 'buggy\_code'), bug description (after tag 'bug\_description'), bug fixes (after tag 'bug\_fixes'), and the correct code (after tag 'correct\_code') for solving the problem statement (after tag 'problem'), we define the state representation of a set of Instructor-Student interactions as a series of necessary tasks which lead the Student from their 'buggy\_code', with bugs described in 'bug\_description', to understanding and correcting their conceptual and syntactical mistakes to reach 'correct\_code' with the 'bug\_fixes'.

We define a state representation as a list of state attributes, where each attribute denotes a specific task that is NECESSARY for the student to successfully understand and implement the given problem. A NECESSARY task directly addresses at least one of the 'bug\_description's and thus, is NOT ALREADY ADDRESSED in 'buggy\_code'. In other words, if a task is not successfully completed, the Student will never be able to correct their 'buggy\_code' to 'correct\_code'.

If the student's 'buggy\_code' shows that they have already understood and implemented a specific task, DO NOT INCLUDE that task as a state attribute since it is REDUNDANT.

The list should be ordered, with earlier attributes/tasks given priority over later ones (e.g., conceptual understanding tasks are a pre-requisite and thus more important than syntactical tasks). The following is an example of the state representation for the given example problem statement: example problem: Implement a Fibonacci sequence using recursion.

—

In-context example from Table 16

—

In the above example format ('example\_explanation' and 'example\_state\_representation'), output an 'explanation' for your plan on which state\_attributes to output and the 'state\_representation' with all NECESSARY 'state\_attribute's for the following 'problem' statement based on 'correct\_code', the Student's 'buggy\_code', the 'bug\_description's, and its respective 'bug\_fixes':

Table 15: Internal Verifier prompt to estimate the state space representation; corresponds to the GenerateState() method in line 1 of Alg 1.

Input:

example\_problem:

Implement a fibonacci sequence using recursion.

example\_buggy\_code:

```
def Fibonacci(n): if n <= 0: print("Incorrect input") elif n == 1: return 1 else: return Fibonacci(n-1)
```

example\_bug\_description:

On line 7, the function only recursively calls 'Fibonacci(n-1)', which will then only return '1' from the edge case on line 5. Instead, the function should consider that the nth term of the Fibonacci sequence is computed as the sum of the preceding n-1 and n-2 values.

example\_bug\_fixes:

Replace 'return Fibonacci(n-1)' with 'return Fibonacci(n-1) + Fibonacci(n-2)' on line 7.

example\_correct\_code:

```
def Fibonacci(n): if n <= 0: print("Incorrect input") elif n == 1: return 1 else: return Fibonacci(n-1) + Fibonacci(n-2)
```

Output:

example\_explanation: Based on the 'bug\_description', the bug involves an incorrect recursive call, which indicates that the Student does not understand that the Fibonacci sequence requires taking the sum of the preceding two terms for getting the current value. An example where 'n' is equal to '2' points this mistake out. Finally, based on the 'bug\_fixes', the Student must modify the recursive call to add 'Fibonacci(n-2)' in order to reach the 'correct\_code' state.

example\_state\_representation:

state\_attribute\_1: Understand the definition of the Fibonacci Sequence.

state\_attribute\_2: Consider the preceding two items in the sequence for computing the current value in the Fibonacci Sequence.

state\_attribute\_3: Consider the example when 'n' is equal to '2'.

state\_attribute\_4: Correctly recursively call 'Fibonacci(n-2)' and add it to the existing 'Fibonacci(n-1)'.

Table 16: In context example for Table 15

The Student has written code (after tag 'student\_code') to solve the problem (after tag 'problem') and is answering a question (after tag 'Student') from the Instructor (after tag 'Instructor') based on their understanding of the 'problem' and their 'student\_code'. IF the Student suggests a solution to a bug they identify, also consider the following:

Ensure that the Student's suggestion is isomorphic to any one of the bug fixes mentioned in the provided 'bug\_fixes'; if not, then 'answer\_has\_no\_mistakes' should be "False". A Student's suggestion is isomorphic to a bug fix if they (1) have the same conclusion or output, (2) share the same underlying logical structure or pattern, and (3) are convertible to each other through a series of logical transformations.

Answer the following questions and within your reasoning, think about how you would answer the "instructor\_question" yourself and include this in your "explanation".: answer\_addresses\_question: <Does the Student's response (after tag 'Student') directly answer the Instructor's question (after tag 'Instructor')? Output "True or "False"> answer\_has\_no\_mistakes: <Is the Student's response (after tag 'Student') to the Instructor's question (after tag 'Instructor') logical (no logical errors or mistakes)? Output "True or "False">

Instructor: {Instructor question}  
Student: {Student response}  
bug\_fixes: {bug fixes}  
student\_code: {student code}

Table 17: Internal Verifier prompt to assess the accuracy of the Student response with respect to the Instructor's question; corresponds to the VerifyResponse() method in line 6 of Alg 1.

A Student has sufficient understanding of a certain topic (specified at tag "target\_understanding") when the responses that they provide to the Instructor (specified in the "conversation\_history") would REQUIRE them to comprehend "target\_understanding". This can either be demonstrated (1) explicitly, where the Student directly mentions "target\_understanding", OR (2) implicitly, where their reasoning is isomorphic to completing the task in "target\_understanding". A Student's reasoning is isomorphic to the "target\_understanding" if they (1) have the same conclusion or output, (2) share the same underlying logical structure or pattern, and (3) are convertible to each other through a series of logical transformations.

Based on the Student's response (after tag 'student\_response') to the Instructor's question (after tag 'instructor\_question') and the conversation history (after tag 'conversation\_history'), do you believe that the Student needed to sufficiently comprehend the "target\_understanding" in order to provide their responses (after tag 'Student' in 'conversation\_history') to the Instructor's questions (after tag 'Instructor' in 'conversation\_history') throughout the conversation history? Include specific quotes from the "conversation\_history" in your "explanation". Within your reasoning, think about how you would answer the "instructor\_question" yourself and include this in your "explanation".

Instructor: {Instructor question}  
Student: {Student response}  
target\_understanding: {target understanding}

Table 18: Internal Verifier prompt to update the state space with respect to a single-turn Instructor-Student interaction; corresponds to the UpdateUnderstanding() method in line 12 of Alg 1.

Are any bug fixes mentioned in the conversation that you have had with the Instructor (under tag "conversation\_history")? If no, return "None". If yes, then follow the format below:

First, based on your current understanding of the problem (tag "problem") and your conversation with the Instructor, summarize (after tag "bug\_summarization") the bugs in the code explicitly mentioned within the "conversation\_history" that you believe will revise your buggy code (after tag "buggy\_code") to a correct implementation of the "problem" statement. Then, based on this summary, output a list of the explicitly mentioned bug fixes (from "bug\_fix\_1" to "bug\_fix\_n", where  $n$  is the number of bug fixes to make), each described briefly.

An example format/wording of a brief bug fix would be: "Replace 'i' with 'i+1' on line 6."

conversation history: { convo history }

Table 19: Instructor to Student prompt that asks the Student to generate a list of bug fixes; corresponds to the GetStudentBugFixes() method in line 17 of Alg 1.

For the problem description given above (after tag 'problem'), you are given two sets of bug fixes (under tags 'suggested\_bug\_fixes' and 'correct\_bug\_fixes'). For each bug fix in 'correct\_bug\_fixes', is there at least one bug fix in 'suggested\_bug\_fixes' that is isomorphic? Two bug fixes are isomorphic if they (1) have the same conclusion or output, (2) share the same underlying logical structure or pattern, and (3) are convertible to each other through a series of logical transformations. Output "True" or "False" as your answer with an explanation.

suggested bug fixes: { student\_bf }

correct bug fixes: { correct\_bf }

Table 20: Internal Verifier prompt check if the Student has suggested all the correct bug fixes that are present in the ground truth set of bug fixes, corresponds to isResolved() in line 4 of Alg. 1.

Based on the student's current level of understanding, as demonstrated through their conversation history (tag "conversation\_history"), what is 1 follow-up question with the same level of depth and difficulty RELATIVE to the 'previous\_questions' that you could ask based on the Student's explanation that would help them reach the "target\_understanding"? Make sure that the question addresses the reasons why the Student got the previous question(s) wrong, as detailed in tag "misunderstanding", such that the Student is more likely to resolve these misunderstandings. You must generate a question such that any correct answer to your question should automatically reflect the "target\_understanding" and resolve the "misunderstanding".

target\_understanding: { target }

conversation\_history: { conversation history }

previous\_questions: { previous questions }

previous\_misunderstanding: { explanations }

These questions should help the Student arrive at the answer themselves; do NOT give any direct hints towards the solution (under tag "bug\_fixes" and tag "bug\_description").

bug\_fixes: { bug fixes }

bug\_descriptions: { bug descriptions }

Table 21: Internal Instructor prompt to generate a sibling question; corresponds to the GenerateSiblingQuestion() method in line 10 of Alg 1.



Based on the student's current level of understanding, as demonstrated through their conversation history (tag "conversation\_history"), what is 1 follow-up question with increasing depth and difficulty RELATIVE to the 'previous\_questions' that you could ask based on the Student's explanation that would help them reach the "target\_understanding"? Make sure that the question addresses the reasons why the Student has not reached the "target\_understanding", as detailed in tag "misunderstanding", such that the Student is more likely to resolve these "misunderstanding"s by answering your question.

target\_understanding: {target}

conversation\_history: {conversation history}

previous\_questions: {previous questions}

previous\_misunderstanding: {explanations}

These questions should help the Student arrive at the answer themselves; do NOT give any direct hints towards the solution (under tag "bug\_fixes" and tag "bug\_description").

bug\_fixes: {bug fixes}

bug\_descriptions: {bug descriptions}

Table 22: Internal Instructor prompt to generate a child question; corresponds to the GenerateChildQuestion() method in line 14 of Alg 1.

Based on the buggy code and the target understanding state (under tag "target\_understanding"), what is one question (k=1) that you could ask that would help the Student reach the "target\_understanding"? These questions should help the Student arrive at the answer themselves; do NOT give any direct hints towards the solution (after tag 'bug\_fixes').

These questions should help the Student arrive at the answer themselves; do NOT give any direct hints towards the solution (under tag "bug\_fixes" and tag "bug\_description").

target\_understanding: {target}

bug\_fixes: {bug fixes}

bug\_descriptions: {bug descriptions}

Table 23: Internal Instructor prompt to generate the initial question; corresponds to the GenerateQuestion() method in line 3 of Alg 1.