SimStep: Human-in-the-Loop Authoring of Interactive Educational Simulations Through Task-Level Abstractions

Zoe Kaputa
Stanford University
USA
kaputa@stanford.edu

Anika Rajaram
The Harker School
USA
26anikar@gmail.com

Vryan Almanon Feliciano
Stanford University
USA
vgfelica@stanford.edu

Zhuoyue Lyu
University of Cambridge
UK
zl536@cam.ac.uk

Maneesh Agrawala
Stanford University
USA
maneesh@cs.stanford.edu

Hari Subramonyam
Stanford University
USA
harihars@stanford.edu

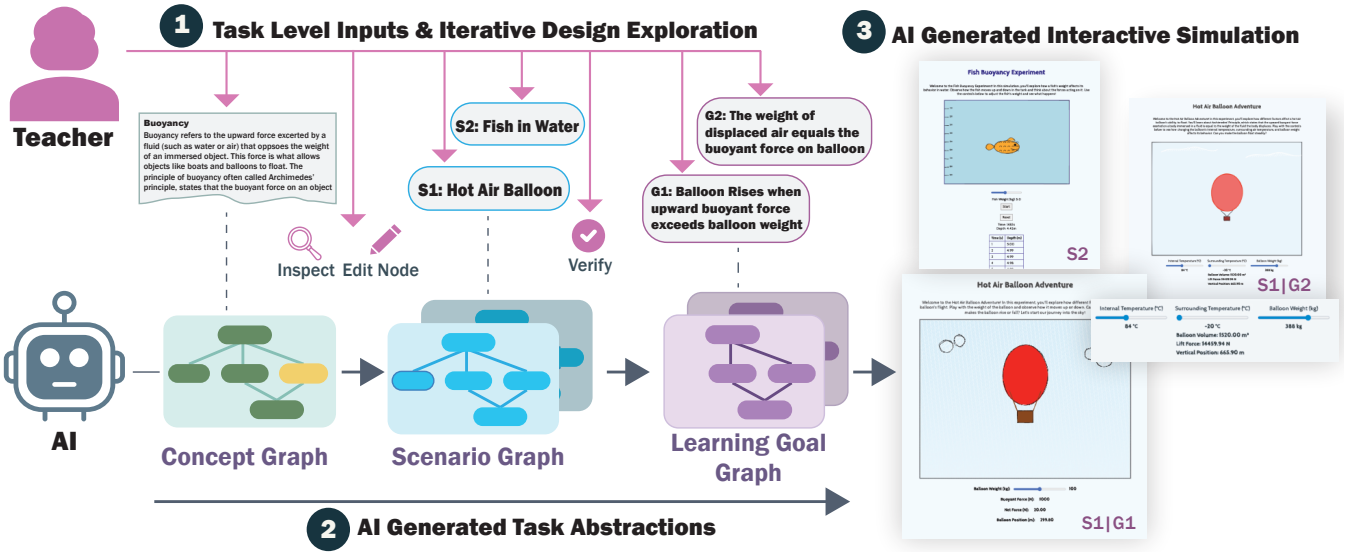


Figure 1: SimStep authoring workflow: (1) A teacher specifies a concept, scenario, and learning goal (top). After each input, (2) the AI generates and surfaces respective intermediate task-level abstractions: Concept Graph, Scenario Graph, and Learning Goal Graph that the teacher can inspect, edit, and verify (bottom). These abstractions guide the (3) generation of an interactive simulation (right), enabling human–AI collaboration through a Chain-of-Abstractions that preserves programming affordances while aligning with teaching tasks.

Abstract

Generative AI enables educators to create interactive learning content by describing goals in natural language. However, without programming affordances such as traceability, refinement, and debugging, teachers struggle to align simulations with learners' needs, refine them step by step, or verify that they reflect intended learning concepts. We propose a task-level abstraction approach that structures authoring as a sequence of representations, mirroring how teachers plan lessons and providing checkpoints for specification, inspection, and refinement. We instantiate this approach in SimStep, an authoring environment that scaffolds simulation design with four abstractions, including Concept Graph, Scenario Graph, Learning Goal Graph, and UI Graph, and introduces an inverse correction process to revise hidden model assumptions without requiring code manipulation. A technical evaluation shows that these abstractions preserve fidelity across transformations, while a user study with educators demonstrates their effectiveness in authoring simulations. Our work reframes AI-assisted programming as human–AI co-authoring through structured, domain-aligned abstractions.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

Keywords

Program Synthesis, Task-level abstractions, Educational Simulation Authoring

ACM Reference Format:

Zoe Kaputa, Anika Rajaram, Vryan Almanon Feliciano, Zhuoyue Lyu, Maneesh Agrawala, and Hari Subramonyam. 2018. SimStep: Human-in-the-Loop Authoring of Interactive Educational Simulations Through Task-Level Abstractions. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 32 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Programming-by-prompting with generative AI promises to democratize code creation by shifting the focus from writing code to expressing semantic intent [37, 64]. For non-technical experts such as educators, this approach is potentially transformative. Rather than learning formal programming constructs, teachers can describe learning content, goals, and personalized scenarios using natural language to create interactive learning experiences [10]. For instance, a teacher might prompt *"give me a simulation to teach about buoyancy by showing hot-air balloons rise and fall,"* and receive a functional interactive simulation aligned with their lesson (see Figure 1).

However, when natural language becomes the primary programming interface (e.g., a tool like ChatGPT [57]), important affordances for testing, debugging, and refining the code are lost. When using generative AI, users struggle to envision concrete intentions [65]. Consequently, the generative model fills in missing information to resolve underspecifications, and these inferences may not always align with user goals. For example, a high-level prompt requesting a buoyancy simulation might yield a visually appealing animation, but *omits* essential causal relationships, *misrepresents* scientific principles, or *fails* to support student interactions that deepen understanding. While traditional programming interfaces afford ways to check and correct implementation through explicit program structure, natural language collapses these structures into opaque textual abstractions. This makes it challenging to isolate errors or understand how prompt changes affect generated code.

Our goal is to provide a better solution for educators authoring interactive STEM simulations with AI while keeping the process *grounded* in tasks teachers already perform, such as articulating goals, identifying concepts, designing scenarios, and specifying interactions. To realize this goal, we draw on two theoretical foundations. First, *Distributed Cognition* [36] postulates that when cognitive work is distributed between people (or teachers and AI in our case), external representations serve as the *shared substrate* where otherwise invisible inferences become inspectable and negotiable. Second, models of pedagogical reasoning converge towards a *staged transformation* approach [63]. Teaching is not a single-step translation from content to activity, but a sequence of transformations from content to examples to goal- and learner-adapted activities.

In SimStep, we realize these frameworks by adapting the Chain-of-Abstractions (CoA) approach into a human-AI authoring framework [25]. Concretely, rather than generating a simulation from content as a single-step, we support the staged transformation of content to simulation through a series of task abstractions. As shown in Figure 1, the abstractions are explicit, domain alinged artifacts including — Concept Graph, Scenario Graph, Learning Goal Graph — that teachers and AI co-create. The AI generates each abstraction, serving as a checkpoint for teachers to validate the AI's interpretation and manually make corrections to align with teachers' pedagogical intentions. When the AI fills in unspecified details, an inverse correction process surfaces these hidden assumptions and maps them to the appropriate abstraction for repair, thus enabling teachers to fix errors at the level of concepts or goals rather than code. SimStep thus recovers core programming affordances, including traceability, testability, and control, through representations teachers already understand.

Our key contributions are: (1) a conceptual framework that repositions programming-by-prompting as authoring through task-level abstractions, emphasizing the role of human-in-the-loop reasoning and correction; (2) the design and implementation of SimStep, an instantiation of this framework that enables educators to incrementally specify, test, and revise AI-generated simulations without writing code; and (3) technical and empirical evaluation demonstrating how SimStep supports pedagogical alignment, reduces authoring complexity, and provides control in content creation.

2 Related Work

SimStep supports the authoring of interactive simulations for scientific discovery learning, where learners develop understanding through hypothesis testing and experimentation [17]. Here, we review relevant literature in learning experience authoring tools, interfaces for prompt-based programming, and end-user approaches to error handling.

2.1 Authoring tools for Learning Experiences

Prior research in Intelligent Tutoring Systems (ITS) have enabled non-programmers to author learning experiences [77]. Tools such as CTATs [5, 40] support both *example-tracing*, which encodes step-by-step demonstrations of program solving, and cognitive modeling, which captures teacher-defined production rules for tutoring. Specifically, CTAT developed specialized debugging interfaces, including Conflict Tree, which visualizes which rules fired and why, while the Why-Not Window lets authors query why expected behavior did not occur. These tools recognize that non-programmers need domain-aligned representations of system logic.

Beyond these systems, ITS research has explored constraint-based authoring, example- and demonstration-based authoring, and component- or template-based approaches to reduce the cognitive burden on educators[13, 31, 50, 53]. For instance, ASSISTments [31] uses templates to let teachers rapidly create their own problems and feedback for students without specifying any formal underlying cognitive model. This lowers barriers but limits expressiveness; teachers work within predefined structures rather than defining novel interactions. More broadly, although these systems have significantly broadened access, challenges with high cognitive load in defining conditions and models, debugging, and exploring authoring pathways remain [71]. SimStep extends these approaches; similar to CTAT, it makes pedagogical logic explicit and editable, but through task abstractions rather than formal production rules.

Further, like the Why-Not Window, SimStep surfaces hidden assumptions, but maps them to task abstractions for easy repair.

2.2 Abstractions in Prompt-Based Programming

Programming-by-prompting enables users to generate code using natural language [18, 37, 61], but users frequently struggle with prompt ambiguity, underspecified behavior, and limited control over generated outputs [18]. End-user programming research has long addressed similar challenges through *representational abstractions* that help non-programmers express and refine complex behavior [55], including concept graphs for high-level relationships [72], block-based structures for syntax-free logic [85], and scene graphs for semantic information [8, 59]. These serve not just as simplifications but as cognitive scaffolds that support reasoning about program behavior, consistent with distributed cognition theory [36].

Recent systems combine natural language interaction with structured abstractions, clustering into two approaches. First is *pre-code abstractions* structure intent specification before generation, e.g., [32, 47, 70, 82]. For instance tools such as CoLadder [82] decompose prompts hierarchically; ProgramAlly [32] uses filter blocks and by-example specification, Sprout [47] scaffolds tutorial authoring through tree-of-thought structures. Second, *Post-code abstractions* support inspection and repair after generation. For instance, DynaVis [69] synthesizes widgets for editing specifications; Misty [48] and CodeShaping [81] enable repair through semantic diffs. Other systems blend approaches: Spellburst [6] combines node-based prompting with output visualization; Biscuit [16] offers ephemeral scaffolds in notebooks.

Table 1 surveys this landscape and compares SimStep to these other approaches. SimStep differs in three ways. First, it offers *multi-layer* pre-code abstractions (Concept → Scenario → Learning Goal → UI) that mirror stages of pedagogical reasoning rather than programming decomposition. Second, it provides *cross-layer* inverse correction in which errors map to the abstraction level where repair makes sense, rather than forcing a single repair mechanism. Third, it enables *end-to-end traceability*. Teachers can follow how a concept manifests through scenario instantiation, goal filtering, and final UI, with validation at each checkpoint grounded in pedagogical intent.

2.3 Error Correction via Human-AI Interaction

Even with well-formed prompts, LLM-generated code is susceptible to errors from hallucinations, incomplete specifications, or semantic mismatches [9, 64, 79]. Correction strategies include test-based evaluation [35], execution-trace validation [42], and iterative self-critique [19]. Systems like Rectifier [83] automate validation using test suites, while Fan et al. [23] demonstrate that program repair techniques can fix common LLM errors. However, automated testing does not reveal all hidden bugs, and runtime feedback is not always helpful for LLM debugging [68].

These methods often treat users as passive recipients of model output. In contrast, hybrid systems incorporate human-in-the-loop workflows by externalizing model assumptions for inspection and correction [46, 73]. Tools like Whyline [39] and Hypothesizer [4] reimagine debugging as explanation and hypothesis-testing rather than syntactic fixing, allowing users to query program behavior through natural language or runtime traces. SimStep combines automated testing (JavaScript error detection, behavioral validation) with human-in-the-loop repair through *inverse correction*. When teachers identify unexpected behavior, the system surfaces relevant assumptions. It maps them to the appropriate abstraction level, i.e., a concept relationship, a scenario instantiation, or a UI binding, rather than requiring code-level debugging.

3 Conceptual Framework for Task-Level CoA

To ground prompt-to-code authoring as a human-in-the-loop process, we formalize our approach using a distributed cognition lens [36]. Distributed cognition theory views cognitive processes as extending beyond individual minds, operating across people, artifacts, and representational media. In Hutchins' study of ship navigation, for example, navigational charts, instruments, and logs serve not merely as information displays, but as cognitive artifacts that structure and distribute reasoning over time and across individuals [36].

Further, we connect this framing to the Chain-of-Abstractions (CoA) idea from Gao et al. [25], which encourages models to plan at an abstract level and only later ground those plans with domain knowledge. We transpose CoA from *internal model planning* to *task-level authoring*: in SimStep, the abstractions are explicit domain-aligned artifacts. We treat intermediate representations not simply as steps in a pipeline but as structured task checkpoints where reasoning is transformed and shared between humans and AI. Each abstraction in the CoA supports distinct types of cognitive work—such as articulating domain knowledge, defining causal structure, or specifying interface logic—and affords both interpretability for the human and tractability for the model. What qualifies this decomposition as a form of distributed cognition is not merely that the process is staged, but that each stage enables coordination between agents (human and AI), externalizes internal thought processes, and provides a representational substrate for validation, revision, and semantic control. In this sense, CoA reflects a system of representations that scaffolds joint reasoning across agents, aligning with core principles of distributed cognitive systems. Furthermore, our framework enables both forward synthesis (intent → abstractions → code) and inverse correction (code → targeted abstraction → revised code), with end-to-end links that make assumptions legible, testable, and correctable.

3.1 Human Guided Forward Transformations

Let P denote a user's initial natural language prompt and C the final executable code. Between them lies a sequence of representational abstractions $\mathcal{A} = \{A_1, A_2, ..., A_n\}$, each representing a task checkpoint where intent is transformed, clarified, and refined. In prior work, several key types of abstractions have been proposed including concept graphs, scene graphs, task decomposition structures, etc. [8, 59, 72, 85]. These abstractions are derived through transformations $T = \{T_1, ..., T_{n+1}\}$, and we propose they involve a collaboration between human agents (H) and machine agents (M) powered by large language models. We formalize the collaborative

System	Pre-code abstractions	Post-code abstractions	Traceability	Validation	Repair scope
SimStep (ours)	Multi-layer (Goals → Concepts → Scenarios → UI)	Cross-layer (UI & Assumption graphs)	End-to-end	Cognitive	Learning goals, concepts, scenarios, UI assumptions
Devy [12]	None	FSM workflows	Local	Visual	Workflow intent steps
ProgramAlly [32]	Filter blocks / by-example	Filter/scene graph over detectors	Structured	Provenance	Filter conditions, detector blocks
Spellburst [6]	None	Node triples (prompt+code +output +sliders)	Local	Visual	Prompt text, sliders, branch variants
DynaVis [69]	None	Widgets on Vega-Lite spec	Structured	Visual	Chart properties via widgets
WaitGPT [78]	None	Operation graph (steps)	Structured	Stepwise	Pipeline operations, step nodes
Data Formulator 2 [70]	Shelves + builder	Spec + data threads	Structured	Provenance	Chart design parameters, data bindings
Biscuit [16]	Ephemeral scaffolds	Scaffold ↔ notebook code	Local	Visual	Scaffolded UI components
Misty [48]	Example UIs for blending	Semantic diffs over UI code	Structured	Stepwise	UI components via semantic diffs
DynEx [49]	Design matrix	Modular stepwise modules	Structured	Stepwise	Module steps in design matrix
Dango [15]	Demo tables	DSL steps + explanations	Structured	Provenance	Data transformation steps
Sprout [47]	Tutorial steps (tree-of-thought)	Step nodes linked to code/text	Structured	Stepwise	Tutorial steps, branching nodes
MoGraphGPT [80]	Drawn proxies for elements	Element modules + central interaction	Structured	Visual	Scene elements, interactive proxies
CoLadder [82]	Hierarchical prompt tree	Prompt blocks ↔ code segments	Structured	Stepwise	Prompt blocks, hierarchical segments
CodeShaping [81]	Sketch annotations on code/output	Semantic diffs	Structured	Stepwise	Code edits via sketch diffs
InstructPipe [84]	None	Pseudocode → DAG pipeline	Structured	Stepwise	Pipeline nodes, DAG modules
InterLink [44]	None	Links between text, code, and outputs	Structured	Visual	None (read-only)
What It Wants Me To Say [45]	None	Code Summary	Structured	Visual/Interpretive	Utterances rephrased into code

Table 1: Comparison of existing program synthesis systems by pre-code and post-code abstractions (types), traceability, validation, and repair scope.

transformation process as:

$$P \xrightarrow{T_1^{(M)}} A_1 \xrightarrow{\text{Refine}^{(H)}} A_1' \xrightarrow{T_2^{(M)}} A_2 \xrightarrow{\text{Direct}^{(H)}} A_2' \xrightarrow{\cdots} C$$

Each abstraction A_i serves as a *task checkpoint* that exposes the model's current understanding of the task. Functionally, each abstractions is characterized by its *visibility*—how observable and legible its structure is to the user, its *manipulability*—how easily the user can adjust or modify its components, and its *fidelity*—how accurately it encodes the user's goals or task-relevant system logic. These properties collectively determine the abstraction's effectiveness in supporting meaningful human oversight and intervention within the generative workflow.

Further, we define four core *forward operations* that the human agent H performs at each task checkpoint A_i:

- **Inspect**: Examine the abstraction A_i to understand what the model has inferred, identify mismatches with intent, and diagnose possible errors or oversights.
- **Refine**: Modify A_i to produce A_i', correcting assumptions, adding missing components, adjusting relationships, or removing irrelevant elements.
- **Validate**: Assess whether A_i' meets instructional or behavioral goals, ensuring semantic coherence, completeness, and alignment with domain expectations.
- **Direct**: Provide guidance for the next transformation $T_{i+1}^{(M)}$ by specifying priorities, constraints, or features that must be preserved in downstream abstractions.

3.2 Inverse Correction

The forward transformation pipeline produces a sequence of intermediate abstractions from a natural language prompt. Each abstraction A_i represents a progressively refined interpretation of user intent, narrowing ambiguity and structuring information for downstream synthesis. Ideally, underspecified details would be surfaced and resolved at appropriate points along this chain. However, in practice, generative models often defer filling in those details until the final stages, particularly at the code level where the model must commit to specific values, logic, or visual behavior. This can lead to implicit assumptions being introduced without the user's awareness, as those decisions were not made explicit in earlier abstractions.

While introducing intermediate abstractions helps reduce ambiguity and support human-in-the-loop correction, it also introduces

a design tradeoff. Increasing the number of abstractions, or lowering their level of specificity, can overwhelm users with too many representational layers or expose them to implementation-level detail. For instance, block-based environments like Scratch [51] provide visual structure but often require users to reason about program flow, variable state, and low-level operations. This reintroduces syntactic and procedural complexity, defeating the purpose of semantic-level prompting. In contrast, our approach favors task-level abstractions that align with how users naturally organize their thinking, aiming for a balance between expressive control and cognitive accessibility.

However, to recover and revise these hidden assumptions, we introduce a complementary **inverse process**, which includes a new set of *targeted* abstractions $\mathcal{B} = \{B_1, B_2, ..., B_m\}$. The same abstraction X may exist in both $\mathcal{A}$ and $\mathcal{B}$, but these sets can also include unique abstractions. The inverse correction process begins by realizing the B_i that is most suitable for correcting the system's assumption. Then, the user can **Inspect**, **Refine**, and **Validate** B_i as described for abstractions in $\mathcal{A}$. After producing B_i', a direct translation $T^{(M)}$ exists to transform B_i' into C'.

$$C \xrightarrow{\text{Realize}} B_i \xrightarrow{\text{Refine}^{(H)}} B_i' \xrightarrow{T^{(M)}} C'$$

This revision occurs at the abstraction level rather than the code level, allowing users to correct high-level misunderstandings without needing to inspect low-level syntax. Let $\Omega(X)$ denote the set of all valid implementations (e.g., simulation code) that are compatible with a representation X, and let $U(X)$ represent the degree of underspecification in X. If a representation is vague or abstract, $\Omega(X)$ will be large, indicating that many different implementations are possible, and $U(X)$ will be high, reflecting the ambiguity of the representation.

As the synthesis pipeline progresses from the natural language prompt P to the final code C, each transformation incrementally reduces both the ambiguity and the number of compatible implementations:

$$\Omega(P) \supset \Omega(B_1) \supset \Omega(B_1') \supset \cdots \supset \Omega(C)$$

$$U(P) > U(B_1) > U(B_1') > \cdots > U(C)$$

This formalism reflects how intent becomes increasingly concrete. Note that this process is not intended to surface all hidden details but to expose and resolve only those underspecifications that result in incorrect behavior. It turns abstractions into bi-directional interfaces that can be used for forward synthesis as well as targeted recovery.

3.3 Deriving the Abstraction Sequence for SimStep

To operationalize this framework in SimStep, we draw on established instructional design theories as well as insights from four years of teaching and refining a graduate-level course on designing interactive simulations for learning [1]. At a high level, the abstraction sequence must satisfy the following two constraints: (1) it must support *progressive formalization* of conceptual knowledge into a simulation specification, and (2) it must capture the *design elements* that simulation research shows are critical for effective discovery learning including conceptual models, contextualized scenarios, focused learning goals, and hypothesis-testing interactions [2, 17, 74]. Classic instructional design frameworks, such as Backward Design [75] and Shulman's model of pedagogical reasoning [63], articulate how teachers move from content understanding to goal-setting and activity design. However, they offer little guidance on which *intermediate* representations should structure an authoring pipeline or how the *function* of those representations should evolve across transformation stages. We therefore adopt Gravemeijer's notion of progressive formalization [28] as the sequencing principle for SimStep, because it uniquely conceptualizes learning as a representational trajectory in which artifacts shift between *models-of* situated activity to *models-for* increasingly general reasoning.

Extending this idea, SimStep's abstraction becomes a deliberate shift in model function that simultaneously aligns with simulation-design requirements. The **Concept model** serves as the initial abstract representation of the phenomenon, capturing and organizing the causal relationships that structure teachers' understanding. The **Scenario** then grounds this conceptual structure in experientially meaningful situations, rendering them as comparable, revisable contexts in which the domain relationships can be explored. Culturally responsive pedagogy [26] emphasizes that this contextualization should connect to students' lived experiences. The **Learning Goal** identifies which aspects of the concept–scenario structure should become stable invariants of student inquiry, refining the abstraction into a pedagogical focus. Finally, the **Interactivity** specifies how these invariants are enacted through manipulable affordances, defining the activity structures through which learners test hypotheses, receive feedback, and engage with the underlying model. To visualize these abstractions in SimStep, we adopt node–link diagrams because they are a well-established representation for causal structures in STEM and align with teachers' existing practice of framing scientific phenomena as interconnected systems [66].

4 User Experience

To realize the CoA framework in Section 3, we developed SimStep, a tool that allows educators to author interactive simulations through a human-guided, step-by-step approach. SimStep's interface follows a wizard-style design pattern using the four main stages discussed in Section 3.3, providing appropriate representational abstractions in each stage. To explore the specific authoring and testing features, let us follow Mr. Carlos, a high school science teacher, as he creates a simulation using SimStep.

4.1 CoA Code Generation

Mr. Carlos is working on his lesson plan to teach about *buoyancy* and wishes to use an interactive simulation to promote active student engagement. Mr. Carlos opens SimStep on his web browser, which shows him the authoring interface with the *text input step* open (Figure 2a). This step has a text input box on the left and a collapsible panel to display the concept graph on the right. Using the prescribed science textbook for the course, Mr. Carlos copies the text about core principles of buoyancy and pastes it in the text box. Based on this text, SimStep automatically generates a **Concept**

[1] Anonymized for review

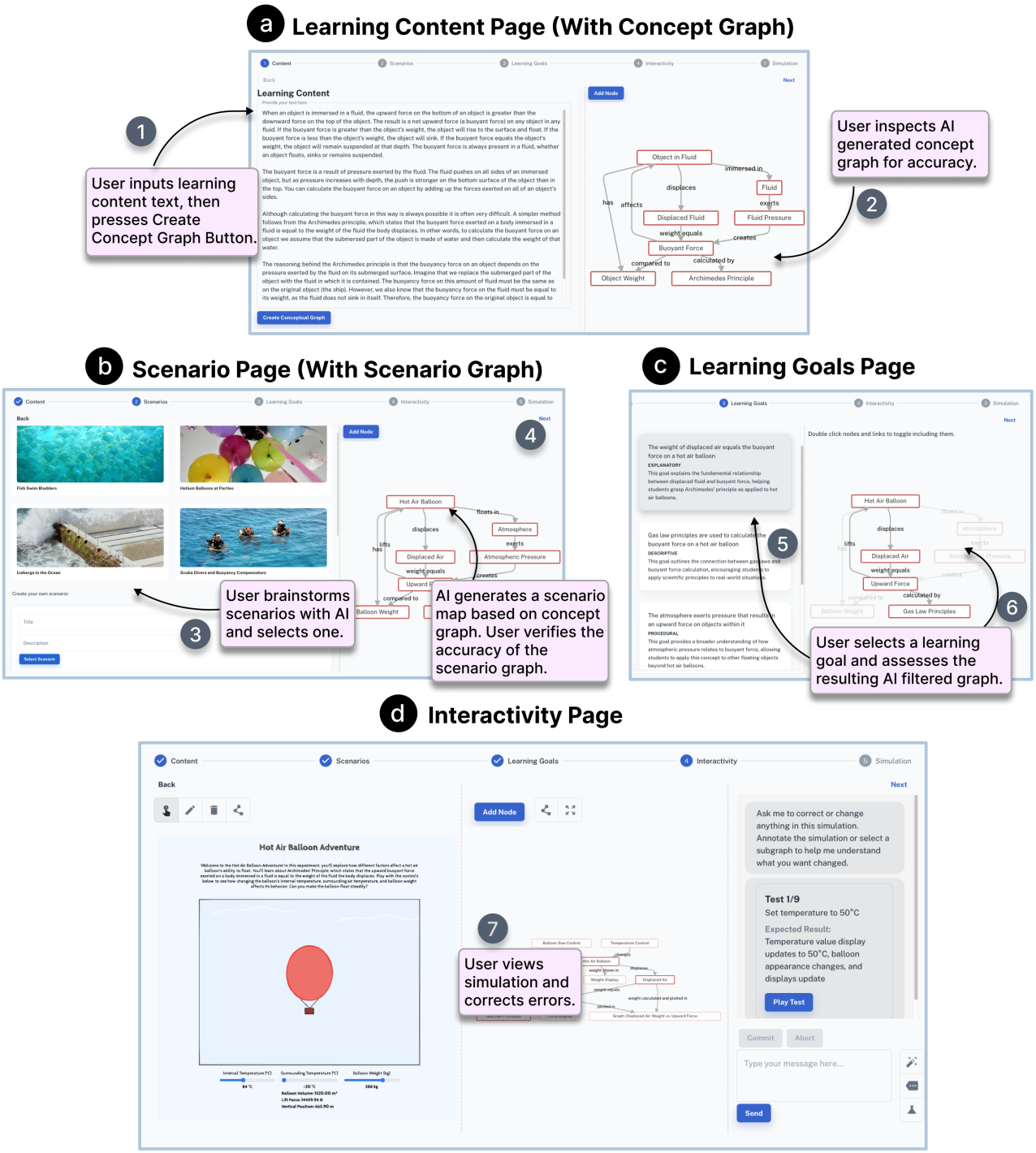


Figure 2: SimStep user interface walkthrough: (a) Teachers input content and generate a Concept Graph, then inspect and refine it. (b) They brainstorm scenarios with AI support, select one, and verify the Scenario Graph. (c) Teachers select a learning goal, and AI generates a filtered Learning Goal Graph using relevant concepts from the scenario graph. (d) The workflow culminates in an Interactivity Page where teachers view and correct the simulation. Each page provides task-level abstractions as checkpoints for human–AI collaboration.

Graph—a visual model made up of nodes representing key concepts like Object's Weight, Buoyant Force B, Fluid Density ρ, and Displaced Volume V, connected by edges that encode relationships and equations (e.g., $m = \rho_{\text{object}} \times V$, $W = m \times g$).

As Carlos **Inspects** the graph, he notices an issue: Buoyant Force is linked directly to Object's Mass, skipping the required relationship with Fluid Density and Displaced Volume. Using the graph editor, he deletes the incorrect link and adds two new nodes and connecting links to represent the correct equation (i.e., **Refine**): $B = \rho \times V \times g$. Once the concept graph accurately reflects the scientific model, Carlos proceeds to the next step.

In the second step, Mr. Carlos is prompted to specify a desired experimental scenario (just like a human expert would do), which serves as the contextual foundation for generating the simulation code. To facilitate scenario grounding, SimStep uses the conceptual model to generate and display a set of potential scenarios for teaching about buoyancy (Figure 2a). Mr. Carlos notices that one of the scenarios is using *hot air balloon*, and realizes that since his students recently saw hot air balloons rise at the annual city festival, it would be a great way to connect the abstract concept of buoyancy to something they had all experienced. Alternatively, Mr. Carlos can define his own scenario using the text input box, tailored to his understanding of his students. Once Mr. Carlos provides a scenario, system generates a **Scenario Graph**, a situation model representation (Figure 2b) in which all the nodes and links in the conceptual model are instantiated with the context of hot air balloon. As before, Carlos **Inspects** the instantiated graph and proceeds to the learning goal selection by clicking the 'Next' button.

The learning goal selection panel offers Mr. Carlos a range of objectives tailored to the chosen scenario. For instance, SimStep might presents several objectives around (1) conceptual understanding focusing on key concepts and describing phenomenon (e.g., *understand the relationship between air temperature inside the balloon and its buoyancy, and be able to describe how equilibrium is achieved*), causal or explanatory understanding (e.g., *explain how the decrease in air density inside the balloon leads to an increase in buoyancy, allowing the balloon to rise*), and procedural knowledge (e.g., *effects of different heating methods on the rate of the balloon's ascent and procedural relationship between gradual heating and smooth altitude control*). On this panel (Figure 2c), Mr. Carlos can select a learning goal (or create his own) to *direct* next steps in the code generation, and as before explore the generated **Learning Goal Graph** which is derived from the situation model. From here, Mr. Carlos simply clicks the 'Next' Button to see the final interactive simulation (Figure 2d). To generate the simulation, SimStep uses the learning goals sub-graph and provides an **Interactivity Graph** including all of the controls and how controls link to the behavior of elements in the situation model to meet the intended learning goals.

4.2 Interactive Debugging and Refinement

While the forward authoring path provides several affordances for aligning steering code generation, potential errors can still emerge in realizing the final interaction graph as simulation code, i.e., syntactic inference errors. The errors might include misaligned layout elements or mislabeled controls, errors where the model incorrectly fills in unspecified details like slider ranges or animation timings, or errors where UI elements fail to trigger the expected behaviors due to missing or flawed logic. To support testing, debugging, and correcting these issues, SimStep offers several features, including guided testing and widget-based error correction.

4.2.1 Guided Testing and Automated Repair: SimStep supports a form of guided UI testing in which it automatically generates test cases from the abstraction pipeline and simulates learner (end-user) interactions within the final simulation UI. This feature was informed by our user study (Section 6). For instance, as seen in Figure 3b, in the buoyancy simulation, the system executes scripted interactions—such as adjusting the temperature slider or releasing the balloon—and prompts for Mr. Carlos's feedback. Specifically, Mr. Carlos is asked to evaluate whether the observed behavior aligns with his instructional intent, such as whether increasing the temperature causes the balloon to rise faster. His response indicates whether the test has passed or failed, enabling it to refactor the code to fix any issues through human feedback.

4.2.2 Manual Debugging and Repair. Beyond guided testing and automated repair, Mr. Carlos can directly inspect the simulation on his own and identify points of misalignment between intents and simulation behavior. Even with state-of-the-art models, underspecification in earlier stages can result in missing or incorrect logic. SimStep provides affordances for interactive debugging and refinement through a rich chat-based interface along with direct annotations on the generated simulation and interaction graph (Figure 4). For instance, Mr. Carlos can circle around a region of interest on the simulation and inspect the relevant nodes in the interaction graph, which is filtered automatically based on the annotation, and then use natural language to describe the problem. For example, he might observe that adjusting the weight slider does not affect the balloon's altitude, and describe the correction in terms of missing connections between relevant concepts.

To support such targeted correction, SimStep implements an underspecification resolution engine that interprets the context of the chat message and generates prefilled widgets that represent candidate fixes at the appropriate level of abstraction. Mr. Carlos can confirm, edit, or reject these suggestions, allowing him to refine the simulation without needing to modify code directly. Or, Mr. Carlos can manually select an abstraction and modify it, as seen in Figure 3a. All changes are shown as drafts until he chooses to commit or discard them, enabling iterative refinement grounded in his instructional intent. Figure 5 shows example simulations generated with SimStep.

5 System Architecture

Here we describe the technical details for (1) the CoA pipeline, (2) the undespecification resolution approach, (3) automated code testing, and (4) affordances for referential conversational interactions with the LLM.

5.1 Chain-of-Abstractions Pipeline

SimStep employs a chain-of-abstractions (CoA) technique that allows teachers' design decisions to be easily integrated into the previous steps of the simulation design process. Figure 1 shows the abstractions used in this chain. SimStep's CoA forward abstractions

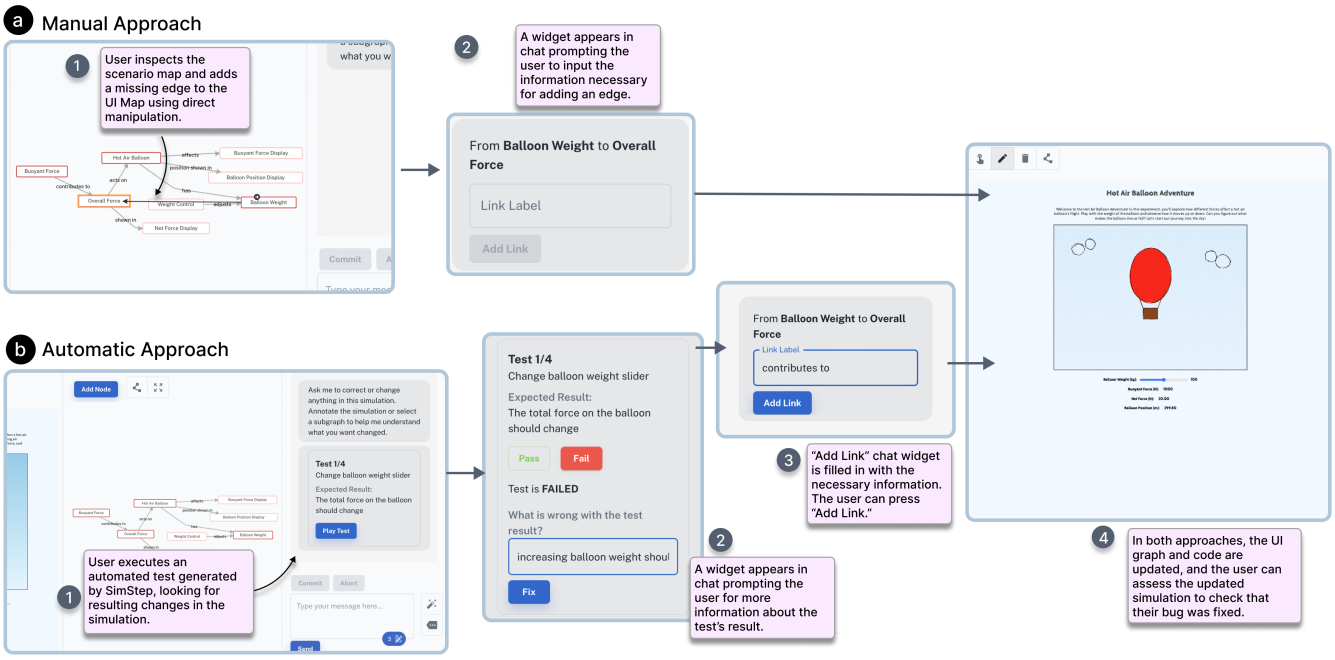


Figure 3: The debugging process can be split into two main approaches: the manual approach and the automatic approach, which provide the user varying levels of control and automation.

are node-link diagrams representing process and UI information. SimStep's set of forward transformation abstractions is

$$\mathcal{A} = \{\text{Concept Graph, Scenario Graph,}$$

$$\text{Learning Goal Graph, User Interaction Graph}\}$$

.

SimStep's set of transformations $\mathcal{T}$ therefore includes five transformations, one for the generation of each abstraction in $\mathcal{A}$ and one for generation of C. All abstractions in $\mathcal{A}$ can be refined through a set of six direct manipulation widgets. These forward abstraction refining widgets and their implementations are outlined in Table 2.

5.1.1 Concept Graph. The teacher's initial learning content prompt is translated into a knowledge graph, or node-link diagram, via LLM prompting. We call this knowledge graph abstraction the "Concept Graph,". In the Concept Graph, objects in the input text are become nodes, and the relationships between objects become directed links between nodes. This graph is visually displayed to the user upon generation. Knowledge graphs like this are commonly used for the representation of concepts in an educational setting, so teachers will be familiar with this form of abstraction [3, 14].

When prompting for this graphical abstraction, SimStep uses natural language request (including user inputted learning content) along with a list of requirements for the form of the generated graph.

5.1.2 Scenario Graph. Once an initial Concept Graph is generated, the teacher then selects a scenario. SimStep generates the Scenario Graph by prompting an LLM to update the nodes in the Concept Graph to be specific to the chosen scenario. Links remain the same. Narrowing a simulation down to a specific scenario or example does not change the conceptual relationships presented in the simulation, but may change the objects that the simulation is engaging with.

For example, a Concept Graph representing the states of matter may state that the nodes Solid and Liquid are connected via the link $\longrightarrow$ *melting* $\longrightarrow$. If the user selects the scenario "The Water Phase Transition", SimStep would generate a scenario graph where Ice is connected to Water via $\longrightarrow$ *melting* $\longrightarrow$.

5.1.3 Learning Goal Graph. The Scenario Graph can be complex, with many nodes and links that are not relevant to the student's hypotheses-verification process. In the end simulation, students should have the ability to prove or disprove hypotheses to achieve scenario-specific learning goals without grappling with extraneous information. To address this, SimStep lets the user select a scenario-specific learning objective, then translates the Scenario Graph to a Learning Goal Graph by prompting an LLM to remove any nodes and links that are unnecessary to the chosen learning goal.

5.1.4 User Interaction Graph. The User Interaction Graph (UI Graph) represents the full simulation, including conceptual information about the content, UI elements, and visuals. It is generated once the teacher has made all three necessary simulation design decisions (Learning Content, Scenario, and Goals).

The UI Graph generation process involves identifying key objects and relationships in the learning goals and generating an experimental procedure based on this information. Appendix C explains the process of procedure generation.

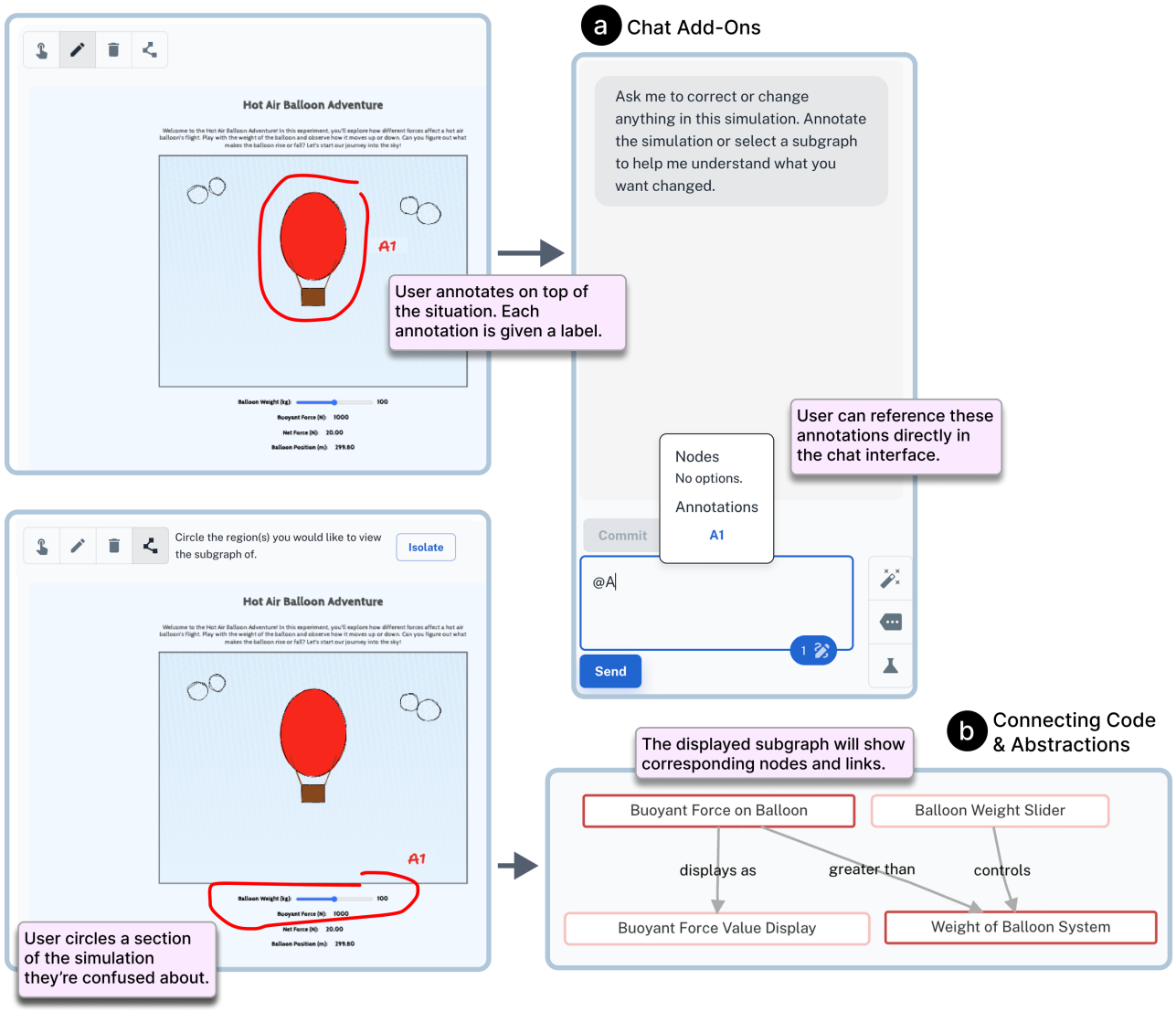


Figure 4: SimStep allows for collaborative interactions allowing users to explore and understand their simulations. (a) shows the process of a user annotating their simulation and referencing their annotation in conversation with the chat. (b) shows a user circling a section of the simulation with the "Subgraph Selector" tool to reveal the subgraph of the UI Map associated with the visuals they circled.

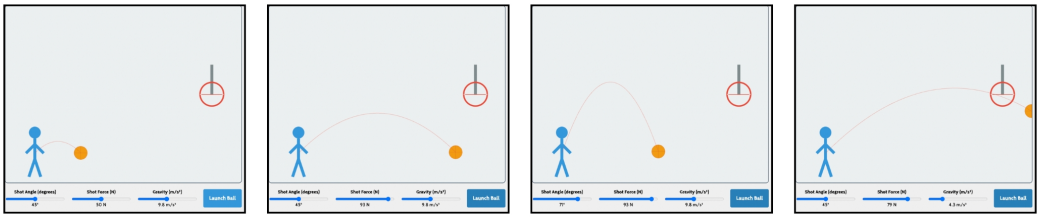
5.1.5 Simulation Code. The UI Graph is then directly translated into simulation code. This simulation code includes HTML, CSS, and JavaScript for all functionality and visual elements included in the UI Graph.

5.2 Underspecification Resolution Approach

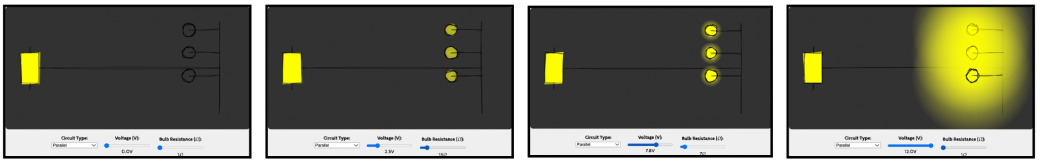
In SimStep, inverse correction is implemented using the Underspecification Resolution Engine. This process allows the user to correct any assumptions made by the LLM revealed at the final code level by refining abstractions in $\mathcal{B}$:

$$\mathcal{B} = \{\text{UI Graph}, \text{Code Assumptions Abstraction}, \text{Redraw Abstraction}\}$$

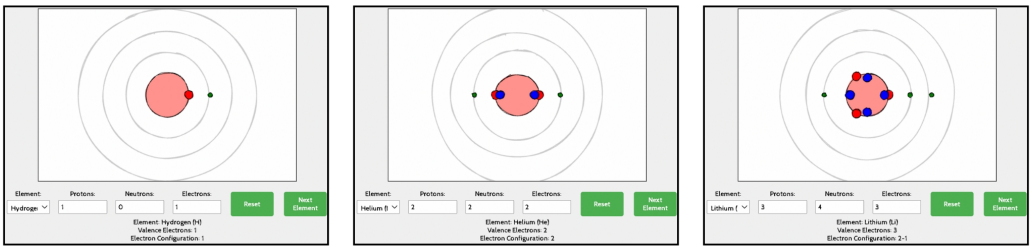
The **UI Graph** maintains the same forward operation implementations as in the forward pass. In the **Code Assumptions Abstraction**, the user uses a chat widget to select a node in the UI Graph and inspect a list of assumptions that the code is making about that node. The user can then refine these assumptions through text editing. Once they've made edits, SimStep prompts an LLM to correct the code's implementation to reflect these updated assumptions. And the **Redraw Abstraction** requires the user to circle an object in the end simulation and create a rough sketch of what they want that object to look like visually using a chat widget.



How angle, shot force, and gravity affect a basketball's trajectory.

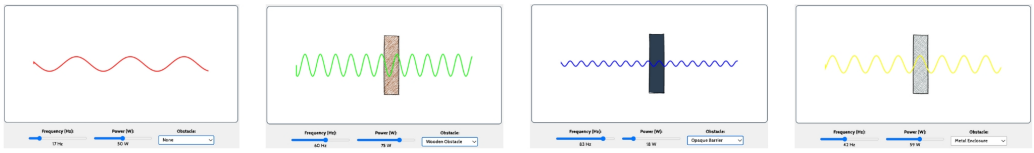


How voltage and resistance change light bulb intensity.

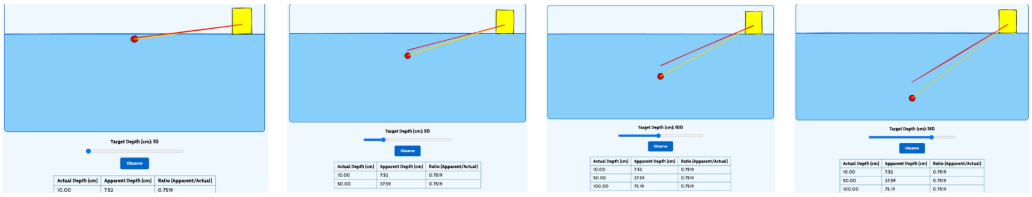


How protons, electrons, and neutrons determine an element's identity.

How different elements form different types of bonds.



How wave transmission depends on frequency, power, and obstacles.



How apparent depth depends on a target's true depth underwater.

Figure 5: Example Simulations Generated with SimStep based on topics in Next Generation Science Standards [56]

SimStep Forward Abstraction Refinement Widgets	
Widget	Description
Add Node	Requires the user to input the name of their new node and adds that node to the current abstraction.
Add Link	Requires the user to draw a new link between two nodes and input the label of their new link. Adds a new link between the provided nodes with the provided label to the current abstraction.
Remove Node	Removes the selected node from the current abstraction.
Remove Link	Removes the selected link from the current abstraction.
Edit Node Label	Changes the label of the selected node to a new, user inputted label.
Edit Link Label	This widget changes the label of the selected link to a new, user-inputted label.

Table 2: The widgets used to refine forward abstractions in SimStep.

The system then prompts an LLM to update the circled visual to more closely match the user's rough sketch.

Using these abstractions, SimStep implements two approaches to underspecification resolution.

5.2.1 Guided Testing and Automated Repair. This approach combines automated code testing and abstraction modification suggestions. As described in Section 5.3.3, SimStep automatically tests and resolves code issues. However, tests involving user interface components are displayed to the user as chat widgets instead of automatically being resolved. Using these widgets, the user can automatically "play" UI actions and assess whether they produce expected results. When the user indicates that an unexpected result has occurred, SimStep invokes inverse correction. SimStep uses an LLM to select the previous abstraction in $\mathcal{B}$ that is most closely aligned with the assumption of interest, then automatically inspects and refines that abstraction. All the user must do is validate the updated abstraction B_i'. This updated abstraction is shown to the user using a chat widget..

In this approach, a pre-filled assumption modification widget is returned from the LLM as a JSON object with all necessary information. See Figure 6 for an example of this response for the "Edit Assumptions" abstraction.

5.2.2 Manual Debugging and Repair. The user can also manually select an abstraction from $\mathcal{B}$ to inspect, refine, and validate. For example, if the user notices that the concept graph is missing a node, they can directly select and fill out the "Add Node" widget to add it in. Or they can prompt the chat explaining the error, in which case an LLM will refine the affected abstraction for the user. The user can verify and accept this refinement.

5.3 Automated Code Testing

We noticed that some of the generated simulations do not work due to JavaScript, logical or User Interface issues. In order to fix this, we use a headless browser approach, combined with detailed logging. The detailed logging is achieved by capturing the UI and logic state of the simulation in the form of log messages, as well as taking screenshots after UI actions. We automate this using Puppeteer, a Node library used to control headless Chrome. Figure 7 shows the sequence diagram for the automated testing workflow, which happens seamless to the user.

5.3.1 JavaScript error resolution. We first ensure that the simulation has no JavaScript errors. This is crucial since JavaScript errors can prevent the UI from working properly. Since button clicks could also result in JavaScript errors, we go through all buttons and perform a click action for every button. All JavaScript errors are then captured using Puppeteer and sent to the LLM to fix.

5.3.2 Test case generation. Once the JavaScript errors are resolved, we ask the LLM to generate test cases. Each test case is defined as a JSON object with the following structure:

```
// Identifier of the UI element
"ID": "slider-weight",
// Action to perform
"action": "set_value",
// Value to set
"value": 80,
// Description of what's being tested
"description": "Adjust weight to observe
    balloon response",
// Expected outcome
"expectedOutcome": "Balloon altitude
    decreases",
// Whether this is UI-specific
"isUIVerification": true
```

5.3.3 Automated test case execution and verification. For tests that do not relate to UI components, SimStep automatically executes, verifies, and updates the code based on the test. In order to provide a comprehensive context to the LLM, we enable debug logging (e.g., capturing SVG coordinates and action timestamps) before executing the test cases with Puppeteer. We run each test case by executing the specified action and capturing a screenshot at the end of each test step. We also capture an initial and final screenshots of the simulation.

The LLM is then invoked with the test case execution results from Puppeteer, the simulation code and the learning goal. We ask the LLM to verify each test case and update the simulation code if the tests fail or if the learning goal is not met. If the LLM finds that a test case is not successful, or a learning goal is not satisfied, it will attempt to update the HTML with an updated version that addresses the underlying problem. We found that it does a very reliable job of ensuring logical errors related to the simulation are fixed. The detected UI errors are fixed based on the limited vision capabilities of the LLM. Tests involving UI components are displayed in the chat for the user to verify.

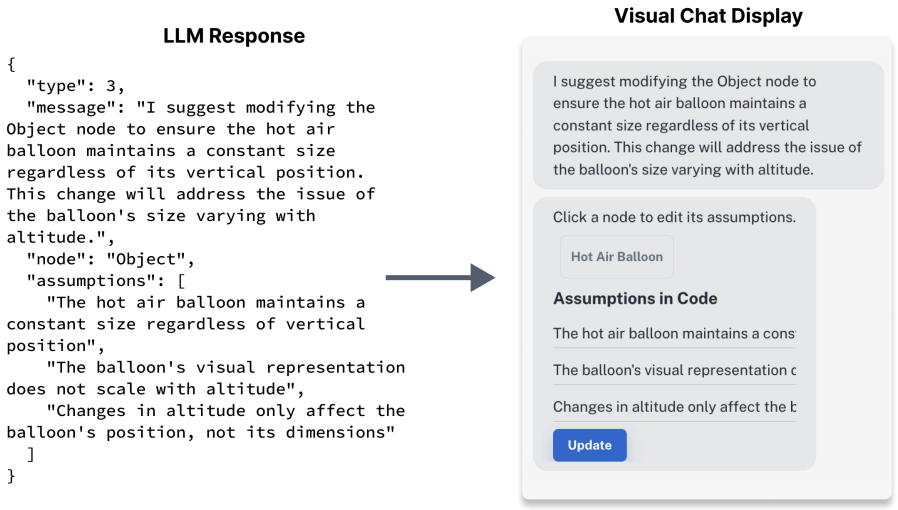


Figure 6: The JSON LLM response representing the "Edit Assumptions" abstraction, along with its visual display in the chat. "type" in the JSON object indicates the type of chat widget returned. Note that the "node" attribute corresponds to the ID of the relevant node, while the value displayed in the widget displays the label on that node.

5.4 Improving Collaborative Interactions Between LLM and User

A side effect of the CoA is that the user must keep maintaining an understanding of multiple, often complex, abstractions of their authored simulation. This can make communicating with the LLM during the Underspecification Resolution Process cognitively expensive. As a means of improving the user experience of SimStep during this process, we have also implemented features that allow users to more clearly communicate their ideas to the LLM, along with allowing users to understand the assumptions of the LLM.

5.4.1 Chat Add-Ons. When prompting via the chat, the user often would like to reference simulation abstractions to explain their desired changes. In order to aid the user in doing this, the user can draw annotations on top on the current simulation. Every time the user annotates, the annotation is labeled. In chat, the user can type "@" to get a list of all nodes in the UI Graph, along with all annotations on the simulation. The user can select from this list to reference a specific annotation or node. When prompting for the desired change to the code, the LLM is provided with these annotations and nodes as context. Figure 4a shows this feature when used with an annotation.

5.4.2 Connecting Code and Abstraction. The translation from UI Graph to HTML Code can be cognitively intense for users. In order to help users understand the connection between the two abstractions, we've implemented a "Subgraph Selector" tool, in which the user can circle one or multiple sections of the end simulation and the tool will display the sub-graph of the UI Graph that this region corresponds to. This is implemented by prompting the LLM with an image of the annotated simulation along with the context of the simulation code and UI Graph. Figure 4b shows this interaction.

5.5 Implementation Details

5.5.1 Frontend. The frontend of the SimStep application is built using `React` [52]. For many of the UI elements and visual components, we use Material UI [54], a UI component library. In order to visually and directly manipulate the abstractions that have a graphical structure, we use `joint.js` [22]. While for parsing these graphs, we use a `mermaid.js` [67] format. We also employ `tldraw` [60] for annotating in the Interactivity Page.

5.5.2 Backend. The backend is built using Node.js [24] and Express.js [34]. We use Anthropic's Claude [7] claude-3.5-sonnet model for all large language model prompting, which we do throughout the process of generating and modifying abstractions in our system. In the design of prompts in this system, we employed several prompting techniques in a trial-and-error approach, resulting in outputs that have consistent form and quality. We also store user-generated simulation code in a Firebase [27] database. This allows users to access the simulations they've created by link.

5.5.3 Prompt Engineering. By aligning the interface with the strengths and limitations of the LLM, we created a system that simplifies the authoring process while maintaining the necessary balance between AI capabilities and user control. As Adar puts it *"Your UI shouldn't write checks your AI can't cash, and your AI shouldn't write checks your UI can't cash*[2]*."*.

6 User Evaluation of Simulation Authoring with SimStep

To understand the user experience of SimStep's Chain-of-Abstractions (CoA) approach, we conducted a user study with $N = 11$ science teachers. Table 3 shows the teaching experience of all participants. Participants had an average of 10.18 years of teaching experience

[2]https://www.youtube.com/watch?v=11UKXaELg8M

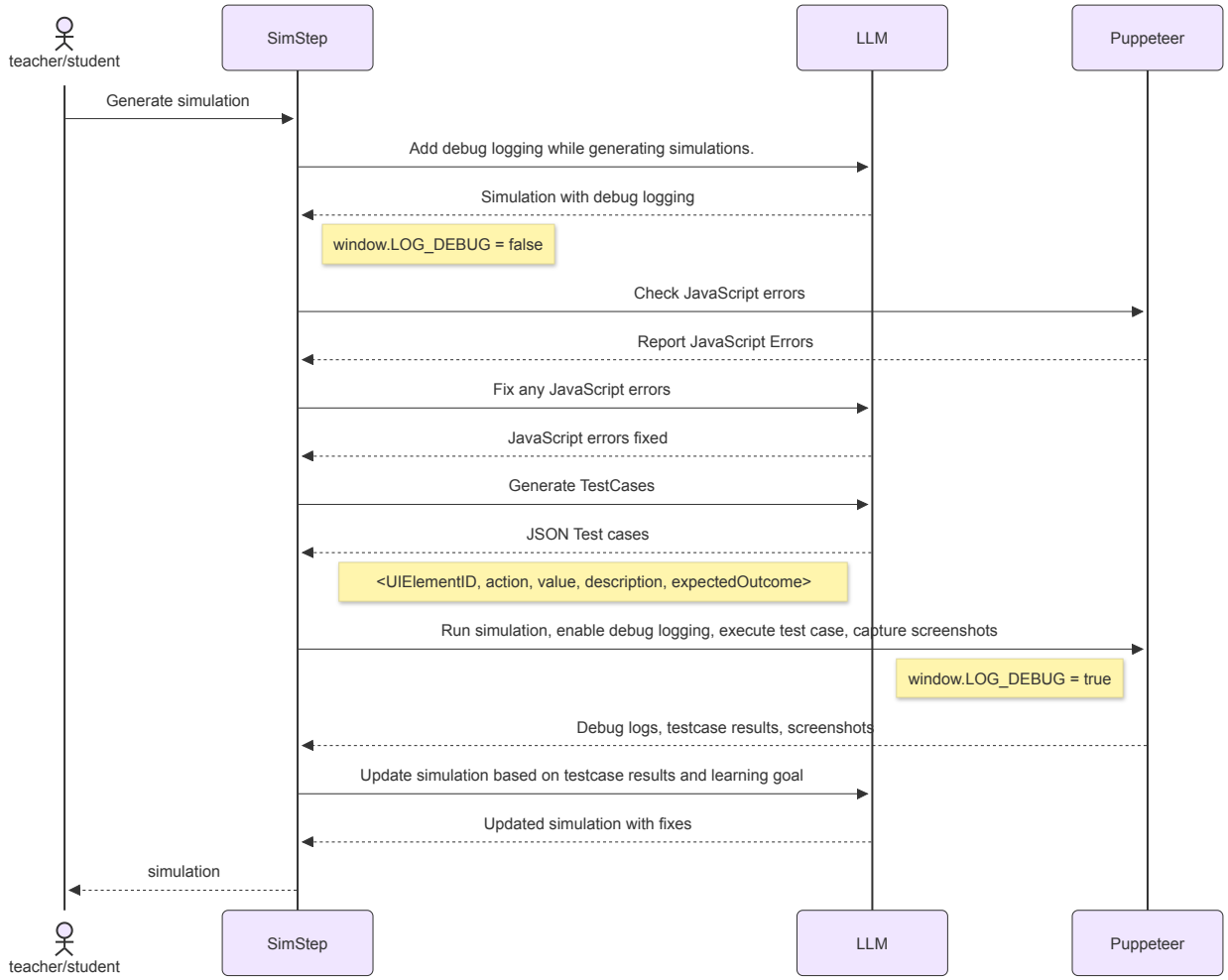


Figure 7: Automated Testing and Error Resolution Workflow for Simulation Verification

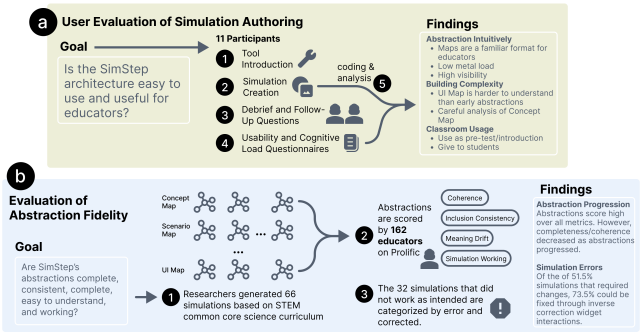


Figure 8: Summary of the goals, methods, and findings of the user evaluation and abstraction evaluation.

($\mu = 10.18$, $\sigma = 6.21$) and were recruited through social media and the researchers' professional networks. The researchers screened participants based on their experience teaching STEM subjects and the grade band they teach. All participants had prior experience teaching STEM subjects, spanning grade levels from middle school to undergraduate college. Each study session lasted approximately 90 minutes and was conducted one-on-one over Zoom, during which participants interacted with a deployed version of SimStep on their web browser. Participants received a $40 honorarium for their time. The study protocol was approved by the institution's Institutional Review Board (IRB). After each interview, the authors reviewed and collected qualitative feedback through inductive coding on the interview transcripts and survey responses.

6.1 Method

In each session, one of the researchers gave a participant a 15-minute demo of SimStep, going through the process of generating

Participant	Subjects Taught	Grade Level	Years of Experience
P1	Chemistry, Physics	9th through 12th	12
P2	Science	9th and 10th	22
P3	Science	9th through 12th	18
P4	Physics	6th and 7th	2
P5	STEM, Liberal Arts	6th through 8th	12
P6	English	7th and 8th	7
P7	Math	9th through 12th	6
P8	Chemistry, Environmental Science, Physics, Biology	9th through 12th	14
P9	Chemistry	9th through 12th	3
P10	Psychology, Neuroscience	Undergraduate	6
P11	Physics, Living Earth	1st through 6th	10

Table 3: The teaching background of all educators in the user evaluation.

a simulation about *States of Matter and Phase Change*. The participant could ask any questions they had about the tool. Then, the researcher allowed the participant to use SimStep to generate their own simulation about either a subject matter of the participant's choice or one of two pre-prepared learning content (Gravitational Force and Buoyancy). When time permitted, a second simulation was generated using a second learning content. After the participant generated and corrected at least one simulation, they were asked to reflect on their experience using SimStep through an unstructured interview. Finally, participants filled out a questionnaire on the usability of the system, the task load, and the Cognitive Dimensions of Notations [29]. The usability of the system was evaluated using the Post-Study System Usability Questionnaire (PSSUQ) [43], and the task load was assessed using the NASA Task Load Index (NASA-TLX) [30].

6.2 Results

6.2.1 System Usability. Overall, teachers found SimStep intuitive to use and reported that they would feel comfortable using this system in a classroom setting. In the PSSUQ questionnaire, we anchored 1 as "Strongly Disagree" and 6 as "Strongly Agree." Our overall usability score was 4.66 (σ = 0.36), and in qualitative feedback, multiple participants commented that the system felt "natural to use." The System Usefulness (SYSUSE) sub-scale was 4.8 (σ = 0.15), indicating that the teachers viewed this tool as useful to their teaching process. Teachers also remarked that they could see themselves using this tool in a variety of ways, including to create introductory material for students, exploratory expansions of their class subjects, and replacement materials for in-class activities such as experiments. Further, the Interface Quality (INTERQUAL) sub-scale had a score of 4.78 (σ = 0.15), indicating the teachers felt that they had the ability to navigate and alter the CoA and the underspecification resolution process with ease. However, the Information Quality (INFOQUAL) sub-scale had a score of 4.44 (σ = 0.48), indicating that we can improve the quality of information displayed to the user. We note that several users found that error messages were not present or were not useful during their process. There is room for integration of more intuitive error messages when the user encounters an unexpected error in the system. Figure 9 shows each participants

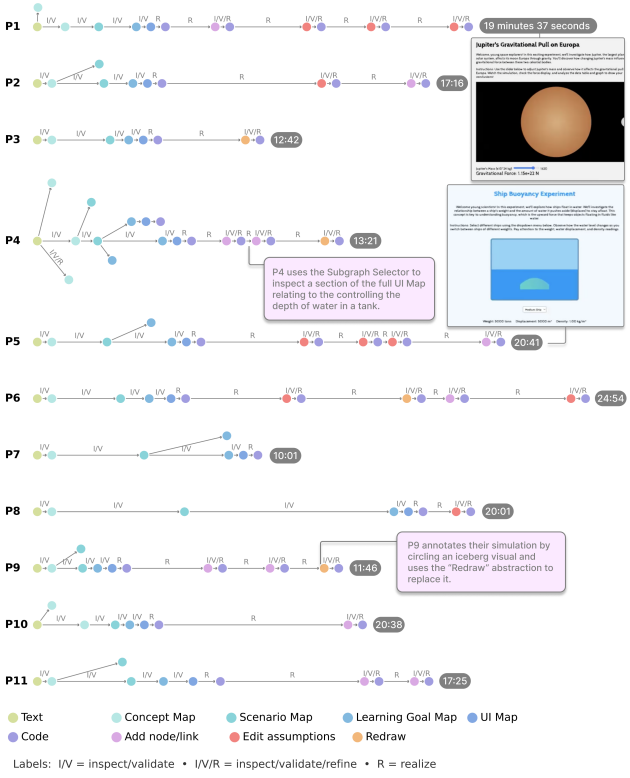


Figure 9: Teacher participant actions while authoring a simulation using the CoA framework. The length of arrows indicates the time between actions.

actions during their authoring process. Participants took an average of 17.13 minutes (σ = 4.40) to completely author their simulation. The majority of users' time was spent evaluating abstractions and making design decisions. Table 4 shows the average time spent on generation for abstractions in SimStep during this evaluation. The UI Map and Simulation Code take substantially longer to generate that other abstractions with an average of 36.27 seconds.

6.2.2 Task Load. Although SimStep does not eliminate the task load of simulation generation, our system provides an interface that imparts a load onto users that is not high. We evaluated task load using NASA-TLX with a 1-6 scale. We anchored 1 with "Very Low" and 6 with "Very High." We did not directly assess the physical demand of our tool, considering the digital nature of our tool. The unweighted TLX score for SimStep was 2.64, indicating that although teachers felt that the task of generating a simulation required work, this load was not high. Participants felt that they were successful in the process of authoring a simulation. Section 5.4 discusses several features added to SimStep to improve collaboration and reduce task load. During this study, we found participants using these features with varying degrees. All 11 participants used the chat feature, while 5 of 11 used the Chat Add-Ons and 2 used the Subgraph Selector. Figure 9 includes examples of how teachers used these collaborative features during their process, along with some of the end simulations that teachers authored. In qualitative feedback, participants remarked that there was low mental demand in the initial steps of the simulation generation process. Mental demand increased at the step of correcting behavior and errors at the simulation code abstraction level (i.e., the Interactivity Page).

Based on our analysis of SimStep's log data, for all but two educators, the majority of actions are performed on the Interactivity Page. Participants most often engaged in the underspecification resolution process by prompting the chat. In association with the task load of this activity, P2, who has been teaching 9th and 10th grade science for 22 years, remarked *"I wanted to convey my thoughts specifically enough that they would be understood, so I struggled at first with how to describe things."* In response to participants' feelings of being lost in the under-specification resolution process, the researchers augmented the process with automated code testing and guided testing functionality, as discussed in section 5.3.3.

6.2.3 Cognitive Dimensions of Notations. Teachers also found that the notations used in the SimStep's CoA were intuitive and mirrored their own internal process when planning to teach about a subject. More specifically, we assessed these notations (abstractions) using the Cognitive Dimensions of Notations [29]. These dimensions allow us to evaluate whether or not these abstractions are useful to teachers for the task of simulation generation, and are often used to evaluate representational tools [11, 41]. We found that SimStep meets all the most relevant dimensions. In this study, we evaluated users' experiences of the system specifically for the dimensions of Visibility, Viscosity, Premature Commitment, Role-Expressiveness, Abstraction, Closeness of Mapping, Consistency, and Diffuseness using Likert scale questions with a range of 1-6. We anchored 1 with "Strongly Disagree" and 6 with "Strongly Agree." The average score across all included questions was 4.61 (μ = 4.61, σ = 0.34). Questions related with visibility had a notable average score of 5.14 (μ = 5.14, σ = 0.27), which indicated success in our goal of generating a tool that easily displays different abstractions of the same end simulation. In qualitative feedback, teachers did, however, indicate that they focused less on the UI Graph on the Interactivity Page, which is expected, considering this page's primary focus on the simulation's behavior itself.

6.2.4 Qualitative Feedback. In a discussion period after using the tool, participants remarked that they enjoyed the graphical abstractions of the input learning content. They said they found this form of abstraction to be intuitive and useful for even their own knowledge formation process. P3, who has been teaching high school science for 18 years, remarked that *"The graphical representations help to show the underlying concepts that are running the simulation. It is also nice to be able to adjust those that would adjust the simulation's behavior."* Some teachers even commented that they would consider giving these maps, or the tool in general, to students to help them learning the concepts at hand. Likewise, teachers found value in the Scenario Page specifically. They appreciated that they could choose a topic specific to their students. One participant also commented that they often struggle to express learning content through different examples when students do not understand the original example. This participant said they appreciated that the tool provides example scenarios to choose from, which may present the content in ways they had not thought of. Similarly, P1 mentioned that using simulations that they've customized using SimStep would be a useful starting point for students to "springboard" off of during group discovery sessions, especially in classrooms with a many different types of learners. In general, the educators in this study were excited at the prospect of using SimStep to generate interactive material specific to their classroom.

7 Evaluation of Abstraction Fidelity

The effectiveness of SimStep in supporting distributed cognition is dependent on the fidelity of the implemented abstractions and the transitions between them along the chain. We evaluated the fidelity of forward transformation abstractions in SimStep with 66 curated simulations. To determine the topic and scenarios, we collected a variety of STEM learning content, culturally relevant scenarios, and learning goals from online STEM curriculum resources. Based on this, we first generated initial simulations and all intermediate abstractions. Since we aim to collect feedback on the generated abstraction in its original state, we omitted making any corrections to the abstractions at this stage. We then recruited 162 participants on Prolific [58] who collectively evaluated all of the abstractions for the 66 simulations. Each abstraction was evaluated by 10 educators. Participants were screened based on their employment as teachers and compensated $12 hourly for their contribution. Specifically we asked educators to rate each abstraction on (1) **Conceptual Completeness/Coherence** — whether the concepts introduced at each abstraction are understandable in explaining the input for each forward abstraction, (2) **Inclusion Consistency** — whether the concepts introduced early are carried through all the way to the final simulation, (3) **Meaning Drift** — whether the concepts don't change meaning in ways that confuse or contradict earlier stages), and (4) **Simulation Working** — whether the final simulation works and is usable. We used a three-point scale to rate each dimension. Complete details of the evaluation is provided in Appendix B. In addition, two of the authors analyzed the simulations for specific breakdowns and attempted to fix the simulations using SimStep's abstractions. Here we report on these results to characterize the fidelity of our task level abstractions.

Time to Generate Abstractions during User Evaluation		
Page	**Time** (seconds)	
	μ	σ
Concept Map	3.54	1.21
Scenario Map	3.45	1.04
Learning Goal Map	3.27	1.68
UI Map & Simulation Code	36.27	9.53

Table 4: The average time to generate each forward abstraction for all 11 participants in the user evaluation. UI Map and Simulation Code are combined because they are generated consecutively in practice.

7.1 Results

7.1.1 Abstraction and CoA Quality. Table 5 reports the average scores across all dimensions of evaluation on a 3 point scale. Overall, the results show that the abstractions produced through the CoA approach maintain high fidelity across multiple dimensions. Educators rated the abstractions highly on **completeness/coherence**, **inclusion consistency**, and **meaning drift**, indicating that the abstractions reliably captured the conceptual content of the input texts and preserved these concepts across stages without significant shifts in meaning. Specifically, the **Concept Map** stage achieved the highest completeness/coherence score ($\mu = 2.61, \sigma = 0.61$), with only a modest decrease observed as abstractions progressed, culminating in the **UI Map** stage ($\mu = 2.53, \sigma = 0.65$). The **Scenario Map** ($\mu = 2.60, \sigma = 0.59$) and **Learning Goal Map** ($\mu = 2.57, \sigma = 0.60$) maintained similarly strong coherence, underscoring the robustness of conceptual fidelity across intermediate steps. In terms of broader abstraction fidelity, **Inclusion Consistency** scored well ($\mu = 2.53, \sigma = 0.63$), reflecting that concepts introduced early were generally carried through to the final simulation.

Meaning Drift was also highly rated ($\mu = 2.57, \sigma = 0.61$), showing that concepts retained stable meanings across transformations. The lowest-performing metric was **Simulation Working** ($\mu = 2.31, \sigma = 0.80$), suggesting that while abstractions were conceptually sound, technical execution of the final interactive simulation occasionally suffered from glitches or incomplete behavior. Together, these findings suggest that the CoA framework in SimStep yields abstractions that are conceptually coherent, consistent, and stable across stages, with the main challenges arising in the reliability of simulation execution rather than in the fidelity of abstraction itself.

7.1.2 Simulation Quality. Based on our manual inspection of the 66 simulations, 48.5% of the simulations worked as intended without any human intervention. To better understand the failures within the remaining simulations, we analyzed the simulations and categorized them into three error types:

(1) **UI/Data Disconnect Errors:** Simulations where the underlying data calculations were correct, but the visual user interface (UI) failed to reflect these changes (e.g., a static or blank UI). This also includes the inverse, where the UI functioned correctly but the data tables were not updated properly.
(2) **Logic & Correctness Errors:** Simulations that were functional but contained logical flaws where their behavior did not correctly model the scientific principles.
(3) **Non-Functional Simulations:** Simulations that loaded but were entirely non-functional, where neither the UI nor the underlying data responded to user input.

Table 6 shows the distribution of types of errors found across all the generated simulations. The researchers then corrected all broken simulations using the CoA steps in SimStep. The correction process often involved a combination of automated and manual interventions. For instance, to fix a UI/Data Disconnect Error in a circular motion simulation, the system identified a UI-related test failure and suggested an "Edit Node Assumptions" action, which resolved the issue by adding the necessary UI for the bike and track. Similarly, for the Refraction Simulation, the automated testing process prompted the LLM to automatically correct the code, thereby fixing the visualization to correctly render the light source, pool, and refraction angle. In a third example, a Non-Functional simulation of Earth's tectonic plates (Figure 10) was partially fixed by automated tests, but a redundant "Measure Tool" button remained. To resolve this, a researcher manually edited the UI graph, selected the unnecessary node, and removed it. 73.5% of the simulations were fixed using such inverse correction widget interactions. Table 7 shows the number of widget interactions necessary to fix the broken simulations. These results suggest that, with the incorporation of distributed cognition, the CoA approach facilitates the effective authoring of cognitively aligned simulations. The Appendix has a full list of simulation topics used in the study.

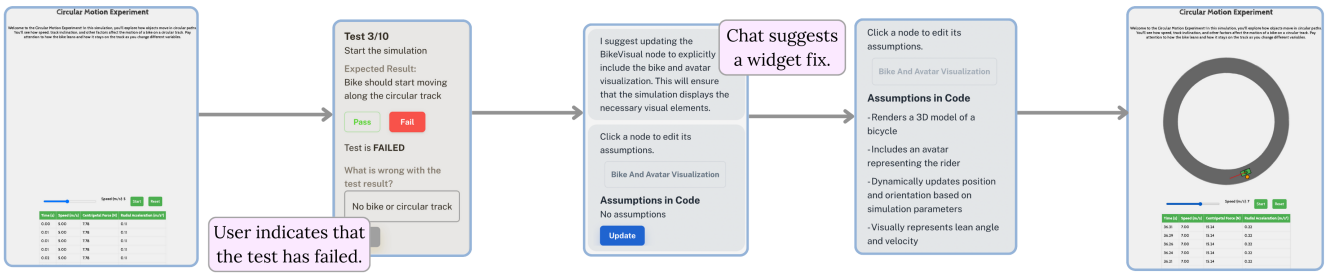
8 Discussion

8.1 Desiderata for Abstractions in CoA

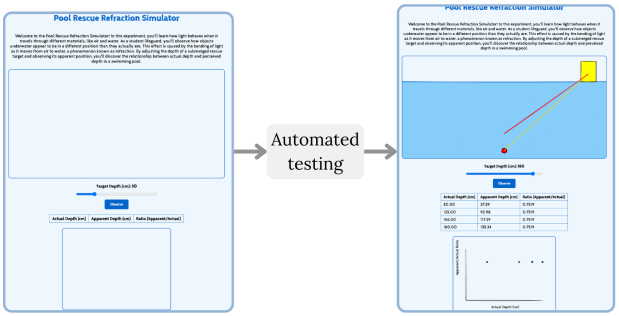
In software engineering, abstractions are essential for managing complexity, modularizing functionality, and enabling reasoning at multiple levels of the system [62]. Our CoA framework applies this perspective to programming-by-prompting by introducing intermediate representations that formalize partial intent and guide the step-by-step transformation from natural language to code. Ideally, the abstractions selected in CoA should **reflect the representational structures of the domain and the decomposition of the task**. In SimStep, for example, the Concept Graph captures the core domain-specific knowledge and relationships, the Scenario Graph expresses contextualized learning situations, the Learning Goal Graph defines desired educational outcomes, and the UI Interaction Graph specifies how learners will interact with the system. Each of these representations corresponds to a meaningful task boundary in the teacher's workflow and provides a different lens for inspecting and refining intent.

SimStep CoA Approach Evaluation		
Metric	**Score**(1-3)	
	μ	σ
Concept Map: Completeness/Coherence	2.61	0.61
Scenario Map: Completeness/Coherence	2.60	0.59
Learning Goal Map: Completeness/Coherence	2.57	0.60
UI Map: Completeness/Coherence	2.53	0.65
Simulation Working	2.31	0.80
Inclusion Consistency	2.53	0.63
Meaning Drift	2.57	0.61

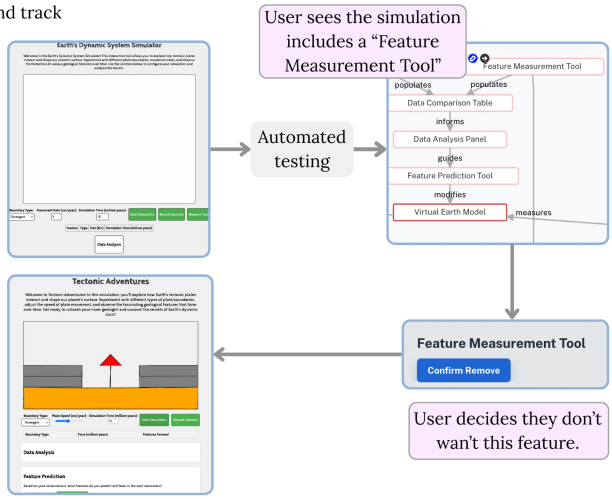
Table 5: The CoA approach is scored using various metrics over 66 curated simulation specifications, and mean (μ) and standard deviation (σ) scores are reported per metric over all simulation specifications and all expert scorers.



Example 1: SimStep's guided testing identified a UI-related test failure and suggested an "Edit Node Assumptions" action, which resolved the issue by adding the necessary UI for the bike and track



Example 2: SimStep's automated testing process prompted the LLM to update the visualization which adds the light and pool along with refraction angle.



Example 3: SimStep's automated testing process automatically fixed the visualization, but left a button, so the researcher used the remove node widget to remove it.

Figure 10: Examples of Fixing a Simulation

Across these contexts, designing effective abstractions requires attention to several key properties. In educational simulation authoring, abstractions should exhibit **visibility** by making key relationships and behaviors perceptible. For instance, a Concept Graph should clearly show that *temperature affects air density*, helping a teacher reason about the cause-and-effect dynamics of buoyancy. Abstractions should also maintain **interpretive continuity** across levels; for example, a concept like "heat" introduced in the Concept Graph should persist through the Scenario Graph and be reflected in UI elements such as a "heating switch" or "temperature slider." They must be **actionable**, allowing users to modify content directly such as enabling a teacher to revise the range of a slider or change the linked behavior of a button without touching the generated code. Abstractions should also support **propagation**, so that edits to a node in the Learning Goal Graph (e.g., changing "students should understand buoyant force" to "students should compare hot vs. cold

Table 6: Distribution of Initial Simulation Evaluation Results (N=66).

Error Type	Count	Percentage (%)
Working	32	48.5
UI/Data Disconnect Errors	19	28.8
Non-Functional	9	13.6
Logic & Correctness Errors	6	9.1
Total	**66**	**100.0**

Table 7: Aggregates of widget interactions needed to fix the simulations

Action Type	Frequency
Edit node assumptions	68
Add node	8
Automated test fix	7
Add edge	5
Remove node	3
Change label	1

air") trigger meaningful updates in downstream abstractions and ultimately in the simulation logic.

Further, the sequence of abstractions should follow a logic of **progressive formalization**. Early abstractions like Concept and Scenario Graphs align with the teacher's domain expertise and planning practices, allowing them to express ideas in familiar pedagogical terms. Later abstractions like the Interaction Graph introduce more structure such as conditional logic or state transitions—providing a bridge between educational intent and executable behavior. At each level, users should be able to validate whether the representation reflects their goals (e.g., "does this scenario match my intended classroom example?"), detect where the system has introduced incorrect or missing assumptions, and revise the abstraction accordingly before continuing to code generation.

8.2 Broader Utility

While SimStep exemplifies the CoA framework by helping educators author interactive simulations, the framework (Section 3) itself has the potential to support a broad range of scenarios. The central construct of CoA is its structuring of the code generation process into a series of task-aligned abstractions. This approach connects to longstanding goals in end-user programming [38], which aims to empower non-programmers to create, adapt, and control computational artifacts. Compared to current approaches that fall along the spectrum of tradeoffs between expressive power and ease of use, the CoA framework offers a middle path: by representing programs as structured, editable abstractions that reflect end-users' task logic, it preserves expressive power while maintaining accessibility and semantic clarity.

A distinctive affordance of CoA is that once a simulation is expressed through staged abstraction graphs, individual components can be isolated, modified, or recombined to serve different purposes. This enables easy generation of multiple simulation variants that would be far more difficult to achieve if working directly in natural language or raw code. For example, educators can readily produce versions of the same conceptual graph tailored for different grade levels (K–6 versus 7–12), adjust the learning objective, or introduce conceptual variance by swapping scenarios (e.g., buoyancy illustrated with submarines versus balloons). Abstraction graphs also make it straightforward to vary the level of complexity by adding or removing detail, creating simplified versions for introductory learners or enriched versions for advanced learners. In this way, CoA supports flexible curriculum design by enabling systematic exploration of how concepts, scenarios, and representational complexity can be varied to match pedagogical needs.

Furthermore, the task abstractions approach can potentially be helpful in other domains in which task-level control of authoring is desirable. Our earlier table 1 shows systems that implement isolated versions of program abstractions. For instance, a data scientist might move from analysis goals to transformation logic to visual outputs [76]; a game designer might articulate core mechanics, progression rules, and interaction feedback [1]. In each case, abstractions can be chosen to reflect natural task decompositions and representational practices in the domain, ensuring that the pipeline aligns with how users already think.

Lastly, while SimStep targets teachers as authors, a promising direction is whether students could use it to create their own simulations. Research on troubleshooting in physics laboratories demonstrates that diagnosing malfunctioning equipment, such as tracking unexpected behavior back to faulty components or incorrect assumptions, engages students in model-based reasoning and deepens experimental and critical thinking competencies [20, 21, 33]. SimStep's abstraction graph could allow students to trace simulation misbehavior back to specific concept relationships, scenario instantiations, or UI bindings, developing causal reasoning skills through representational debugging. This would extend CoA from an authoring framework into a learning environment where the act of repair becomes a site for deepening domain understanding.

8.3 Limitations and Future Work

While the CoA framework offers a principled structure for programming-by-prompting, several limitations point towards future extensions to the framework. The current implementation, though effective for educational simulation authoring, may limit flexibility in domains where task workflows are less standardized or abstractions are harder to formalize. Designing methods for customizing or synthesizing domain-appropriate abstractions remains an open challenge. Second, while CoA reduces the need for syntactic programming, it introduces new forms of cognitive load: users must interpret and manipulate layered representations. As evidenced in the user study, the quality of outputs depends not only on the model's capabilities but also on the user's ability to structure intent within the abstraction pipeline. Future work should explore adaptive interfaces that scaffold abstraction complexity based on user expertise. Current LLMs are not inherently abstraction-aware and often fail to maintain consistency across transformation stages. We also do not yet understand how sensitive SimStep's transformations are to small perturbations in input graphs; minor edits such

as adding or deleting a single node may disproportionately alter downstream outputs, and it remains unclear whether such changes reflect genuine structural differences or shifts in the model's pretrained priors. Future work should systematically evaluate this robustness, particularly for well-known topics where the model may override user-provided structure. Errors from hallucination, misalignment, or representational gaps also remain difficult to detect and correct, suggesting opportunities for model fine-tuning on abstraction-preserving transformations.

9 Conclusion

This work introduces the Chain-of-Abstractions (CoA) framework as a principled approach to programming-by-prompting, one that treats code generation not as a single-shot translation, but as a structured process of task-level semantic articulation. By decomposing the synthesis process into cognitively meaningful, domain-aligned representations, CoA enables users to externalize, inspect, and iteratively refine their intent. We instantiate this approach in SimStep, a tool that supports educators in authoring interactive simulations through a scaffolded, human-in-the-loop workflow. SimStep's inverse correction process addresses underspecification by surfacing assumptions and guiding revision at abstraction checkpoints, recovering key affordances of traditional programming such as traceability, testability, and control. CoA provides a foundation for more controllable, expressive, and domain-sensitive code generation.

References

[1] Ernest Adams. 2014. *Fundamentals of game design.* Pearson Education.

[2] Wendy K Adams, Sam Reid, Ron LeMaster, Sarah B McKagan, Katherine K Perkins, Michael Dubson, and Carl E Wieman. 2008. A study of educational simulations part I-Engagement and learning. *Journal of Interactive Learning Research* 19, 3 (2008), 397–419.

[3] Garima Agrawal, Yuli Deng, Jongchan Park, Huan Liu, and Ying-Chih Chen. 2022. Building knowledge graphs from unstructured texts: Applications and impact analyses in cybersecurity education. *Information* 13, 11 (2022), 526.

[4] Abdulaziz Alaboudi and Thomas D Latoza. 2023. Hypothesizer: A Hypothesis-Based Debugger to Find and Test Debugging Hypotheses. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[5] Vincent Aleven, Bruce M Mclaren, Jonathan Sewall, and Kenneth R Koedinger. 2009. A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education* 19, 2 (2009), 105–154.

[6] Tyler Angert, Miroslav Suzara, Jenny Han, Christopher Pondoc, and Hariharan Subramonyam. 2023. Spellburst: A node-based interface for exploratory creative coding with natural language prompts. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[7] Anthropic. 2023. Claude: Language Model by Anthropic. https://www.anthropic.com/.

[8] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 2019. 3D Scene Graph: A Structure for Unified Semantics, 3D Space, and Camera. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 5663–5672. https://api.semanticscholar.org/CorpusID:203837042

[9] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732* (2021).

[10] Michael P Barnett and WM Ruhsam. 1968. A natural language programming system for text processing. *IEEE transactions on engineering writing and speech* 11, 2 (1968), 45–52.

[11] Michael Bostock and Jeffrey Heer. 2009. Protovis: A graphical toolkit for visualization. *IEEE transactions on visualization and computer graphics* 15, 6 (2009), 1121–1128.

[12] Nick C Bradley, Thomas Fritz, and Reid Holmes. 2018. Context-aware conversational developer assistants. In *Proceedings of the 40th International Conference on Software Engineering*. 993–1003.

[13] Tommaso Calo and Christopher Maclellan. 2024. Towards educator-driven tutor authoring: generative AI approaches for creating intelligent tutor interfaces. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. 305–309.

[14] Penghe Chen, Yu Lu, Vincent W Zheng, Xiyang Chen, and Boda Yang. 2018. Knowedu: A system to construct knowledge graph for education. *Ieee Access* 6 (2018), 31553–31563.

[15] Wei-Hao Chen, Weixi Tong, Amanda Case, and Tianyi Zhang. 2025. Dango: A Mixed-Initiative Data Wrangling System using Large Language Model. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–28.

[16] Ruijia Cheng, Titus Barik, Alan Leung, Fred Hohman, and Jeffrey Nichols. 2024. BISCUIT: Scaffolding LLM-generated code with ephemeral UIs in computational notebooks. In *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 13–23.

[17] Ton De Jong and Wouter R Van Joolingen. 1998. Scientific discovery learning with computer simulations of conceptual domains. *Review of educational research* 68, 2 (1998), 179–201.

[18] Paul Denny, Viraj Kumar, and Nasser Giacaman. 2023. Conversing with Copilot: Exploring Prompt Engineering for Solving CS1 Problems Using Natural Language. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1* (Toronto ON, Canada) *(SIGCSE 2023)*. Association for Computing Machinery, New York, NY, USA, 1136–1142. https://doi.org/10.1145/3545945.3569823

[19] Shihan Dou, Haoxiang Jia, Shenxi Wu, Huiyuan Zheng, Weikang Zhou, Muling Wu, Mingxu Chai, Jessica Fan, Caishuang Huang, Yunbo Tao, et al. 2024. What's Wrong with Your Code Generated by Large Language Models? An Extensive Study. *arXiv preprint arXiv:2407.06153* (2024).

[20] Dimitri R. Dounas-Frazer and H. J. Lewandowski. 2017. Electronics lab instructors' approaches to troubleshooting instruction. *Physical Review Physics Education Research* 13 (2017), 010102.

[21] Dimitri R. Dounas-Frazer, Kevin L. Van De Bogart, MacKenzie R. Stetzer, and H. J. Lewandowski. 2016. Investigating the role of model-based reasoning while troubleshooting an electric circuit. *Physical Review Physics Education Research* 12 (2016), 010137.

[22] David Durman. 2024. joint.js. https://www.jointjs.com/.

[23] Zhiyu Fan, Xiang Gao, Martin Mirchev, Abhik Roychoudhury, and Shin Hwei Tan. 2023. Automated repair of programs from large language models. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 1469–1481.

[24] OpenJS Foundation. 2024. Node.js. https://nodejs.org/en/.

[25] Silin Gao, Jane Dwivedi-Yu, Ping Yu, Xiaoqing Ellen Tan, Ramakanth Pasunuru, Olga Golovneva, Koustuv Sinha, Asli Celikyilmaz, Antoine Bosselut, and Tianlu Wang. 2024. Efficient tool use with chain-of-abstraction reasoning. *arXiv preprint arXiv:2401.17464* (2024).

[26] Geneva Gay. 2013. Culturally responsive teaching principles, practices, and effects. In *Handbook of urban education*. Routledge, 391–410.

[27] Google. 2024. Firebase. https://firebase.google.com/.

[28] Koeno Gravemeijer. 1999. How emergent models may foster the constitution of formal mathematics. *Mathematical thinking and learning* 1, 2 (1999), 155–177.

[29] Thomas RG Green. 1989. Cognitive dimensions of notations. *People and computers V* (1989), 443–460.

[30] SG Hart. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload/Elsevier* (1988).

[31] Neil T Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education* 24, 4 (2014), 470–497.

[32] Jaylin Herskovitz, Andi Xu, Rahaf Alharbi, and Anhong Guo. 2024. ProgramAlly: Creating Custom Visual Access Programs via Multi-Modal End-User Programming. *ArXiv* abs/2408.10499 (2024). https://api.semanticscholar.org/CorpusID:271909496

[33] Natasha G. Holmes and Carl E. Wieman. 2018. Introductory physics labs: We can do better. *Physics Today* 71, 1 (2018), 38–45.

[34] TJ Holowaychuk. 2024. Express.js. https://expressjs.com/.

[35] Zichao Hu, Francesca Lucchetti, Claire Schlesinger, Yash Saxena, Anders Freeman, Sadanand Modak, Arjun Guha, and Joydeep Biswas. 2024. Deploying and evaluating llms to program service mobile robots. *IEEE Robotics and Automation Letters* 9, 3 (2024), 2853–2860.

[36] Edwin Hutchins. 1995. *Cognition in the Wild.* MIT press.

[37] Ellen Jiang, Edwin Toh, A. Molina, Kristen Olson, Claire Kayacik, Aaron Donsbach, Carrie J. Cai, and Michael Terry. 2022. Discovering the Syntax and Strategies of Natural Language Programming with Generative Language Models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022). http://dl.acm.org/citation.cfm?id=3501870

[38] Amy J. Ko, Robin Abraham, Laura Beckwith, Alan F. Blackwell, Margaret M. Burnett, Martin Erwig, Chris Scaffidi, Joseph Lawrance, Henry Lieberman, Brad A. Myers, M. Rosson, Gregg Rothermel, Mary Shaw, and Susan Wiedenbeck. 2011. The state of the art in end-user software engineering. *ACM Computing Surveys (CSUR)* 43 (2011), 1 – 44. https://api.semanticscholar.org/CorpusId:128364433

[39] Amy J Ko and Brad A Myers. 2004. Designing the whyline: a debugging interface for asking questions about program behavior. In *Proceedings of the SIGCHI*

conference on Human factors in computing systems. 151–158.
[40] Kenneth R Koedinger, John R Anderson, William H Hadley, and Mary A Mark. 1997. Intelligent tutoring goes to school in the big city. *International journal of artificial intelligence in education* 8 (1997), 30–43.
[41] Benjamin Lee, Arvind Satyanarayan, Maxime Cordeil, Arnaud Prouzeau, Bernhard Jenny, and Tim Dwyer. 2023. Deimos: A grammar of dynamic embodied immersive visualisation morphs and transitions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.
[42] Teemu Lehtinen, Charles Koutcheme, and Arto Hellas. 2024. Let's Ask AI About Their Programs: Exploring ChatGPT's Answers To Program Comprehension Questions. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training*. 221–232.
[43] James R Lewis. 1992. Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proceedings of the human factors society annual meeting*, Vol. 36. Sage Publications Sage CA: Los Angeles, CA, 1259–1260.
[44] Yanna Lin, Leni Yang, Haotian Li, Huamin Qu, and Dominik Moritz. 2025. InterLink: Linking Text with Code and Output in Computational Notebooks. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–15.
[45] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D Gordon. 2023. "What it wants me to say": Bridging the abstraction gap between end-user programmers and code-generating large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–31.
[46] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin G. Zorn, J. Williams, Neil Toronto, and Andrew D. Gordon. 2023. "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://api.semanticscholar.org/CorpusID:258107840
[47] Yihan Liu, Zhen Wen, Luoxuan Weng, Ollie Woodman, Yi Yang, and Wei Chen. 2024. SPROUT: an interactive authoring tool for generating programming tutorials with the visualization of large language models. *IEEE Transactions on Visualization and Computer Graphics* (2024).
[48] Yuwen Lu, Alan Leung, Amanda Swearngin, Jeffrey Nichols, and Titus Barik. 2025. Misty: Ui prototyping through interactive conceptual blending. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
[49] Jenny GuangZhen Ma, Karthik Sreedhar, Vivian Liu, Pedro A Perez, Sitong Wang, Riya Sahni, and Lydia B Chilton. 2025. Dynex: Dynamic code synthesis with structured design exploration for accelerated exploratory programming. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–27.
[50] Christopher J MacLellan and Kenneth R Koedinger. 2022. Domain-general tutor authoring with apprentice learner models. *International Journal of Artificial Intelligence in Education* 32, 1 (2022), 76–117.
[51] John Maloney, Mitchel Resnick, Natalie Rusk, Brian Silverman, and Evelyn Eastmond. 2010. The scratch programming language and environment. *ACM Transactions on Computing Education (TOCE)* 10, 4 (2010), 1–15.
[52] Meta. 2024. React. https://react.dev/.
[53] Antonija Mitrovic. 2012. Fifteen years of constraint-based tutors: what we have achieved and where we are going. *User modeling and user-adapted interaction* 22, 1 (2012), 39–72.
[54] MUI. 2024. Material UI. https://mui.com/material-ui/.
[55] Brad A Myers, Amy J Ko, and Margaret M Burnett. 2006. Invited research overview: end-user programming. In *CHI'06 extended abstracts on Human factors in computing systems*. 75–80.
[56] NGSS Lead States. 2013. Next Generation Science Standards: For States, By States. https://www.nextgenscience.org. Accessed: 2025-03-04.
[57] OpenAI. 2025. ChatGPT. https://chat.openai.com/. Large language model, version GPT-5.
[58] Prolific. 2024. Prolific. https://www.prolific.com. Version: month/year of use; accessed: YYYY-MM-DD.
[59] Nico Ritschel, Felipe Fronchetti, Reid Holmes, Ronald Garcia, and David C. Shepherd. 2022. Can guided decomposition help end-users write larger block-based programs? a mobile robot experiment. *Proceedings of the ACM on Programming Languages* 6 (2022), 233 – 258. https://api.semanticscholar.org/CorpusID:253239351
[60] Steve Ruiz. 2021. tldraw: A tiny little drawing app. https://tldraw.com/.
[61] Advait Sarkar, Andrew D Gordon, Carina Negreanu, Christian Poelitz, Sruti Srinivasa Ragavan, and Ben Zorn. 2022. What is it like to program with artificial intelligence? *arXiv preprint arXiv:2208.06213* (2022).
[62] Ahmed Seffah, Jan Gulliksen, and Michel C Desmarais. 2005. *Human-centered software engineering-integrating usability in the software development lifecycle*. Vol. 8. Springer Science & Business Media.
[63] Lee Shulman. 1987. Knowledge and teaching: Foundations of the new reform. *Harvard educational review* 57, 1 (1987), 1–23.
[64] Da Song, Zijie Zhou, Zhijie Wang, Yuheng Huang, Shengmai Chen, Bonan Kou, Lei Ma, and Tianyi Zhang. 2023. An empirical study of code generation errors made by large language models. In *7th Annual Symposium on Machine Programming*.
[65] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.
[66] Hariharan Subramonyam, Colleen Seifert, Priti Shah, and Eytan Adar. 2020. Texsketch: Active diagramming through pen-and-ink annotations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[67] Knut Sveidqvist and Mermaid.js contributors. 2014. Mermaid: Generation of diagrams and flowcharts from text in a similar manner as markdown. https://mermaid.js.org/.
[68] Runchu Tian, Yining Ye, Yujia Qin, Xin Cong, Yankai Lin, Yinxu Pan, Yesai Wu, Haotian Hui, Weichuan Liu, Zhiyuan Liu, et al. 2024. Debugbench: Evaluating debugging capability of large language models. *arXiv preprint arXiv:2401.04621* (2024).
[69] Priyan Vaithilingam, Elena L Glassman, Jeevana Priya Inala, and Chenglong Wang. 2024. Dynavis: Dynamically synthesized ui widgets for visualization editing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.
[70] Chenglong Wang, Bongshin Lee, Steven M Drucker, Dan Marshall, and Jianfeng Gao. 2025. Data Formulator 2: Iterative Creation of Data Visualizations, with AI Transforming Data Along the Way. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
[71] Daniel Weitekamp, Erik Harpstead, and Ken R Koedinger. 2020. An interaction design for machine teaching to develop AI tutors. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–11.
[72] Martin Weyssow, Houari A. Sahraoui, and Bang Liu. 2022. Better Modeling the Programming World with Code Concept Graphs-augmented Multi-modal Learning. *2022 IEEE/ACM 44th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)* (2022), 21–25. https://api.semanticscholar.org/CorpusID:245837887
[73] Jules White, Sam Hays, Quchen Fu, Jesse Spencer-Smith, and Douglas C Schmidt. 2024. Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. In *Generative ai for effective software development*. Springer, 71–108.
[74] Carl E Wieman, Wendy K Adams, and Katherine K Perkins. 2008. PhET: Simulations that enhance learning. *Science* 322, 5902 (2008), 682–683.
[75] Grant Wiggins and Jay McTighe. 2005. *Understanding by Design* (expanded 2nd edition ed.). ASCD.
[76] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2015. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 649–658.
[77] Beverly Park Woolf. 2010. *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann.
[78] Liwenhan Xie, Chengbo Zheng, Haijun Xia, Huamin Qu, and Chen Zhu-Tian. 2024. Waitgpt: Monitoring and steering conversational llm agent in data analysis with on-the-fly code visualization. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–14.
[79] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817* (2024).
[80] Hui Ye, Chufeng Xiao, Jiaye Leng, Pengfei Xu, and Hongbo Fu. 2025. MoGraphGPT: Creating Interactive Scenes Using Modular LLM and Graphical Control. *arXiv preprint arXiv:2502.04983* (2025).
[81] Ryan Yen, Jian Zhao, and Daniel Vogel. 2025. Code Shaping: Iterative Code Editing with Free-form AI-Interpreted Sketching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
[82] Ryan Yen, Jiawen Zhu, Sangho Suh, Haijun Xia, and Jian Zhao. 2023. Coladder: Supporting programmers with hierarchical code generation in multi-level abstraction. *arXiv preprint arXiv:2310.08699* (2023).
[83] Xin Yin, Chao Ni, Tien N Nguyen, Shaohua Wang, and Xiaohu Yang. 2024. Rectifier: Code translation with corrector via llms. *arXiv preprint arXiv:2407.07472* (2024).
[84] Zhongyi Zhou, Jing Jin, Vrushank Phadnis, Xiuxiu Yuan, Jun Jiang, Xun Qian, Jingtao Zhou, Yiyi Huang, Zheng Xu, Yinda Zhang, et al. 2023. InstructPipe: Building Visual Programming Pipelines with Human Instructions. *arXiv preprint arXiv:2312.09672* (2023).
[85] Yifeng Zhu, Jonathan Tremblay, Stan Birchfield, and Yuke Zhu. 2020. Hierarchical Planning for Long-Horizon Manipulation with Geometric and Symbolic Scene Graphs. *2021 IEEE International Conference on Robotics and Automation (ICRA)* (2020), 6541–6548. https://api.semanticscholar.org/CorpusID:229156165

A Architecture Prompting

Our chain-of-abstractions architecture utilizes LLM prompting for the generation and modification of abstractions. In this section, we document the prompts used in SimStep.

A.1 Forward Abstraction Generation

All graphical abstractions are represented as mermaid.js graphs. In prompts that include the graphical structure outputs, we specify this format:

```
Use mermaidjs. The format of the mermaidjs
    graph should have each node with a label
    surrounded by square brackets, and each
    link with a label surrounded by vertical
    bars. Each line defining a node or link
    should begin with 4 spaces. Don't add any
     addition styling nor any blank lines.
    More specific instructions: The diagram
    must start with graph LR (or graph TD).
    Each node should be defined as NodeID[
    Node Label] with a single space before it
    . Each edge should be defined as SourceID
     -->|Edge Label| TargetID without any
    extra spaces between the arrow and the |.
     Do not include any Markdown formatting (
    no triple backticks, no language tags).
    Only output the raw Mermaid code.
```

A.1.1 Concept Graph. The following prompt is used to generate the concept graph using the user's learning content.

```
Given the following learning content,
    generate a conceptual diagram of this
    text.

The graph should meet the following
    requirements:
- The nodes of the graph represent physical
    objects presented in the learning content
    .
- The links represent the relationships
    connecting them.
- The graph contains relationship that are
    not necessarily stated in the learning
    content but can be inferred.
- The graph is as abstract and simple as
    possible.
- If the graph has math or numbers involved,
    they are included.
- Each node has a description in square
    brackets.
- Each link a description that explains its
    meaning.

NO Explanation.
${mermaidDirections}
 ${learningContent}
```

A.1.2 Scenario Graph. The following prompt is used to generate the scenario graph using the concept graph and the user's chosen scenario.

```
Given the concept graph and scenario, change
    the names of all the nodes in the graph
    to represent the specific objects that
    the scenario involves. I'm trying to
    understand how the scenario can represent
     the relationship that the graph explains
    . Each node should have an updated is
    that is specific to this scenario.

A good response:
- Keeps the ids of the nodes in the graph the
     same (ids being the values in square
    brackets)
NO Explanation.
${mermaidDirections}
Concept graph:${graph}
Scenario:${scenario}
```

A.1.3 Learning Goal Graph. The following prompt is used to generate the learning goal graph using the scenario graph and the user's chosen learning goal (labeled hypothesis below).

```
Given the concept graph and learning goal,
    give me an updated graph (also using
    mermaidjs) that only contains nodes and
    links that pertain to the given learning
    goal (in other words, remove unnecessary
    nodes and links).
NO Explanation
${mermaidDirections}
Concept Graph:${graph}
Learning Goal:${hypothesis}
```

A.1.4 User Interaction Graph. The following prompt is used to generate the user interaction graph using the learning goal graph and the experimental procedure. Subsequent subsections describe how this procedure is generated.

```
Given an experimental procedure and concept
    graph, create a UI interaction graph of a
     web-based interactive simulation for
    this procedure.
Consider what the main experimental object is
     in this experiment, along with the
    dependent and independent variables.
What is the best way to visually show the
    experimental object within the procedure?
```

```
Include nodes for:
- all nodes in the given concept graph that
    you think are necessary for following the
     procedure and learning the conclusion of
     the learning goal. Keep the same ids and
     labels for these nodes as were in the
    original concept graph.
- UI controls necessary for following the
    procedure
Include edges for:
- all edges in the given concept graph that
    you think are necessary for following the
     procedure and learning the conclusion of
     the learning goal.
- any more relationships between visuals
- relationships between the UI controls and
    the visuals

Give the graph using mermaid js. NO
    explanation.
${mermaidDirections}

Graph:${graph}
Procedure:${proc}
Learning Goal:${hypothesis}
```

Descriptive Learning Goals: When the selected learning goal is descriptive, procedure is generated by first identifying the independent and dependent variables in question:

```
In an experiment testing the provided
    hypothesis using the laws explained in
    the provided concept graph, what is the
    main experimental object that we are
    interacting with in this experiment? NO
    explanation.

Concept Graph:${graph}
Hypothesis: ${hypothesis}
```

```
In an experiment testing the provided
    hypothesis using the laws explained in
    the provided concept graph, what is the
    dependent variable of this experiment? NO
     explanation and don't put a period at
    the end.

Concept Graph:${graph}
Hypothesis: ${hypothesis}
```

Using this information, we prompt for a procedure:

```
You are an expert in designing experimental
    procedures for interactive simulations.
    Given the concept graph and learning goal
     that you are trying to test, generate a
    simple procedure testing what effect ${
    indep} has on ${dep} as indicated by the
    learning goal using the concept graph.

A good procedure:
- Comes to the conclusion of the learning
    goal
- Is simple to follow
- Outlines any data collection that you want
    to perform

NO explanation

Concept Graph:${graph}
Learning Goal:${hypothesis}
```

Explanatory: When the selected learning goal is explanatory, procedure is generated by again identifying the independent and dependent variables in question, then prompting for a procedure:

```
Based on the concept graph, what underlying
    process explains why ${indep} has an
    effect on ${dep}? Give me one single
    phrase. NO explanation and don't put a
    period at the end.

Concept Graph:${graph}
```

```
You are an expert in designing experimental
    procedures for interactive simulations.
    Given the concept graph and learning goal
     that I am trying to test, give me a
    simple procedure testing how ${exp} as
    indicated by the learning goal using the
    concept graph.

A good procedure:
- Comes to the conclusion of the learning
    goal
- Is simple to follow
- Outlines any data collection that you want
    to perform

NO explanation.

Concept Graph:${graph}
Learning Goal:${hypothesis}
```

Procedural: When the selected learning goal is procedural, we first prompt for the experimental object:

```
In an experiment testing the provided
    hypothesis using the laws explained in
    the provided concept graph, what is the
    main experimental object that we are
    interacting with in this experiment? NO
    explanation.

Concept Graph:${graph}
Hypothesis: ${hypothesis}
```

Then we prompt for the process in question:

```
In an experiment testing the provided
    hypothesis using the laws explained in
    the provided concept graph, what is the
    porocess that the ${obj} goes through? NO
     explanation and don't put a period at
    the end.

Concept Graph:${graph}
Hypothesis: ${hypothesis}
```

And finally, we use this information to prompt for a procedure:

```
You are an expert in designing experimental
    procedures for interactive simulations.
    Given the concept graph and learning goal
     that I am trying to test, give me a
    simple procedure for running an
    interactive simulation showing how ${obj}
     goes through ${proc}.

A good procedure:
- Comes to the conclusion of the learning
    goal
- Is simple to follow
- Outlines any data collection that you want
    to perform

NO explanation.

Concept Graph:${graph}
Learning Goal:${hypothesis}
```

A.1.5 Code. We purposefully want SimStep simulations to look low-fidelity. In order to achieve this look, we use rough.js. Therefore, any prompts generating code include the following information about rough.js:

```
Imports and uses roughjs (https://unpkg.com/
    roughjs@latest/bundled/rough.js) for the
    entire app, including UI controls. Make
    sure to include a canvas element in the
    html to reference as the rough.js canvas
    in your script.
```

The following prompt is used to generate the simulation code:

```
Create a web-based interactive simulation
    based on the UI Interface Graph. Do your
    best to interpret which nodes represent
    visuals, which represent UI controls &
    data collection mechanisms, and give each
     UI control its desired function.
Include:
- A title and descripton of the experiment.
    The description should act as an
    introduction for the students to the
    experiment and what it teaches. In this
    description include any definitions that
    directly relate to the learning goal. Don
    't just outright state the learning goal,
     we want students to figure this out on
    their own.
- Integrate any instruction necessary. It
    should be clear how students should
    interact with the simulation.
- All nodes representing visuals and
    phenomena in a visual display of the
    experiment.
- All nodes representing UI controls & data
    collection below the visual display. You
    must label what each UI control
    represents. If a control represent an
    amount, make sure to provide concrete
    experimental units.
- ${roughDirections}

Make sure to include ALL nodes in the UI
    interface graph and give them their
    desired purpose
When inplementing the relationship between
    any two nodes, think about what
    attributes of each node define the
    functional reationship between them.
Keep track of these attributes in your script
    .

The simulation is for middle school kids, so
    make the visuals fun and engaging.
The UI controls should invoke expressive
    animations in the visual display.
```

```
Generate SVGs that are as realistic as
    possible and use gradients and additional
     shapes if needed for any necessary
    experimental objects. Rather than placing
     SVGs directly in the code, clearly
    define each as a variable.
For all text, either use the font cabin-
    sketch-regular or cabin-sketch-bold. In
    order to use these fonts, add <link rel="
    preconnect" href="https://fonts.
    googleapis.com"> <link rel="preconnect"
    href="https://fonts.gstatic.com"
    crossorigin> <link href="https://fonts.
    googleapis.com/css2?family=Cabin+Sketch:
    wght@400;700&family=Londrina+Sketch&
    family=Roboto:ital,wght@0
    ,100;0,300;0,400;0,500;0,700;0,900;1,100;1,300;1,4
    display=swap" rel="stylesheet"> to the
    HTML code.


Give the HTML, CSS and any necessary JS
    interactivity.
Make sure the code displays all visual
    elements correctly and no uncaught errors
     occur when it is run.

Add comprehensive logging for debugging
    purposes ONLY. Include the following
    debug logging system in your code:
    // Debug logging system - only active
        when window.LOG_DEBUG is true
    function logDebug(message) {
        if (window.LOG_DEBUG) {
            console.log(\`DEBUG [\${new Date
                ().toISOString()}]: \${
                message}\`);
        }
    }

    // Add this in your initialization
    document.addEventListener('
        DOMContentLoaded', () => {
        if (window.LOG_DEBUG) {
            logDebug('Simulation initialized
                ');
            // Log all SVG elements
            document.querySelectorAll('svg').
                forEach((svg, index) => {
                const rect = svg.
                    getBoundingClientRect();
                logDebug(\`SVG #\${index}:
                    Position {x: \${rect.x},
                    y: \${rect.y}}, Size {w:
                    \${rect.width}, h: \${
                    rect.height}}\`);
            });
            // Log UI controls
            document.querySelectorAll('input,
                 button, select').forEach((
                control) => {
                logDebug(\`Control \${control
                    .id || 'unnamed'} (\${
                    control.tagName}):
                    Initial value: \${control
                    .value || 'N/A'}\`);
            });

            // Add event listeners for
                logging interactions
            document.querySelectorAll('button
                ').forEach((button) => {
                button.addEventListener('
                    click', () => logDebug(\`
                    Button \${button.id || '
                    unnamed'} clicked\`));
            });
            document.querySelectorAll('input
                ').forEach((input) => {
                input.addEventListener('
                    change', (e) => logDebug
                    (\`Input \${input.id || '
                    unnamed'} changed to \${e
                    .target.value}\`));
            });
        }
    });

    // For animations and calculations, add
        logging in those functions
    // Example: logDebug(\`Calculated \${
        variable} = \${formula} = \${result
        }\`);
    // Example: logDebug(\`Element \${id}
        moved to {x: \${newX}, y: \${newY
        }}\`);

IMPORTANT: This logging should only be active
     when window.LOG_DEBUG is true, so it won
    't affect normal usage.
The default state MUST be window.LOG_DEBUG =
    false; in your code.
Make sure to call logDebug() in key places of
     your code to track:
- All SVG elements' positions and sizes
- UI control state changes
```

```
- User interactions with timestamps
- Animation frame-by-frame updates
- Mathematical calculations with formulas
- Any errors or warnings

NO explanation.

UI Interface Graph::${graph}
Learning Goal:${hypothesis}
```

A.2 User Direction

A.2.1 Scenario Options. Below is the prompt used to generate potential scenarios given the concept graph.

```
Give me 8 contexts that I can use to teach
    the concepts involved in the concept
    graph.
By context, I mean an instantiation of the
    concepts involved in the concept graph.
For each context, write a title enclosed in
    double curly braces {{title}} without
    numbering and a few words on how it
    presents the concept and how the teacher
    should use the contexts to teach high
    school kids about this concept.
Don't talk about classroom activities or
    student activities.
MAKE SURE to separate each context with "|".

Concept graph:${graph}
```

A.2.2 Learning Goal Options. Below is the prompt used to generate potential learning goals given the scenario graph.

```
Give me 6 concise and testable learning goals
     that I can use to teach high schoolers
    about the concept involved in the concept
     map. When forming goals, please focus on
     a few specific nodes in the concept
    graph and how the links between them
    represent relationships. Phrase these
    goals as a single statement that you want
     the student to learn by looking at the
    concept map, sort of like a hypothesis.
    For each learning goal, write the goal in
     double curly braces {{title}} without
    numbering and a brief description of what
     learning gains that goal will lead to.
    If the goal just describes a process,
    begin the desciption with "1.", if the
    goal explains why a process works in the
    way it does, begin the description with
    "2.", and if the goal explains an overall
     process, begin the description with
    "3.". MAKE SURE to separate each learning
     goal from the next with "|".

Concept Graph: ${graph}
```

A.3 Abstraction Modification

A.3.1 Suggesting a Code Change. The following prompt is used to suggest a class of code change based on a chat or error description message.

```
In a previous prompt, you generated the
    provided HTML code for an interactive
    simulation based on the provided UI Map.

However, there's an issue with this code/UI
    map: ${prompt}.

This issue may also be outlined via the red
    annotations in the provided image, which
    are labeled with labels such as "A1" or "
    A2" near the annotation.
These labels may be referenced in the above
    issue description.

Given the issue mentioned above, what type of
     change to the code would you would need
    to make in order to fix this issue?
    Choose from the following types of
    changes:
1. Add new variable/visual component to the
    code.
2. Add new function that acts on/using
    specific variables/visual components to
    the code.
```

```
3. Remove a variable/visual component from
    the code.
4. Remove a function that acts on/using
    specific variables/visual components from
     the code.
5. Completely re-implement a variable/visual
    component in the code.
6. Completely re-implement a function in the
    code.
7. Change an SVG in the code.
8. Change the implementation of some
    preexisting functionality.

Given me just the number associated with the
    type of change you would make.
NO EXPLAINATION.
```

A.3.2 Populate Remove Edge. If the chat suggests removing an edge, we populate a chat widget describing the change to the user. The following prompt is used to get the necessary information.

```
In a previous prompt, you generated HTML code
     for an interactive simulation based on
    the provided UI Map.

However, there's an issue with this code/UI
    map: ${prompt}.

This issue may also be outlined via the red
    annotations in the provided image, which
    are labeled with labels such as "A1" or "
    A2" near the annotation.
These labels may be referenced in the above
    issue description.

We think removing an edge from the UI Map
    might solve the above issue conceptually.
We want some more information on the edge
    that should be removed.

Please respond in the following format:
{ type: 8, message: <SHORT message explaining
     your suggestion. Use full, eloquent
    sentences and first person.>, source: <id
     of the source node>, target: <id of the
    target node>, label: <the label of the
    edge to delete>}

NO EXPLAINATION.
```

A.3.3 Populate Remove Node.

```
In a previous prompt, you generated HTML code
     for an interactive simulation based on
    the provided UI Map.

However, there's an issue with this code/UI
    map: ${prompt}.

This issue may also be outlined via the red
    annotations in the provided image, which
    are labeled with labels such as "A1" or "
    A2" near the annotation.
These labels may be referenced in the above
    issue description.

We think removing a node from the UI Map
    might solve the above issue conceptually.
We want some more information on the node
    that should be removed.

Please respond in the following format: {
    type: 7, message: <SHORT message
    explaining your suggestion. Use full,
    eloquent sentences and first person.>,
    node: <id of the node to delete>}

NO EXPLAINATION.
```

A.3.4 Populate Edit Edge.

```
In a previous prompt, you generated HTML code
     for an interactive simulation based on
    the provided UI Map.

However, there's an issue with this code/UI
    map: ${prompt}.

This issue may also be outlined via the red
    annotations in the provided image, which
    are labeled with labels such as "A1" or "
    A2" near the annotation.
These labels may be referenced in the above
    issue description.

We think changing the label of an edge in the
     UI Map might solve the above issue
    conceptually.
We want some more information on the edge
    that should be changed.

Please respond in the following format:
```

```
{ type: 6, message: <SHORT message explaining
    your suggestion. Use full, eloquent
    sentences and first person.>, source: <id
    of the source node>, target: <id of the
    target node>, oldLabel: <the old label of
    the edge to change>, newLabel: <the new
    label>}

NO EXPLAINATION.
```

A.3.5 Populate Edit Node.

```
In a previous prompt, you generated HTML code
    for an interactive simulation based on
    the provided UI Map.

However, there's an issue with this code/UI
    map: ${prompt}.

This issue may also be outlined via the red
    annotations in the provided image, which
    are labeled with labels such as "A1" or "
    A2" near the annotation.
These labels may be referenced in the above
    issue description.

We think changing the label of a node in the
    UI Map might solve the above issue
    conceptually.
We want some more information on the node
    that should be changed.

Please respond in the following format:
{ type: 5, message: <SHORT message explaining
    your suggestion. Use full, eloquent
    sentences and first person.>, node: <id
    of the node to change>, oldLabel: <the
    old label of that node>, newLabel: <the
    new label>}

NO EXPLAINATION.
```

A.3.6 Populate Redraw.

```
In a previous prompt, you generated the
    provided HTML code for an interactive
    simulation.

However, there's an issue with this code/UI
    map: ${prompt}.

This issue may also be outlined via the red
    annotations in the provided image, which
    are labeled with labels such as "A1" or "
    A2" near the annotation.
These labels may be referenced in the above
    issue description.

We think editing one of the SVGs in the code
    might solve the above issue.
We want some more information on how we
    should update the SVG.

Please respond in the following format:
{ type: 4, message: <SHORT message explaining
    your suggestion. Use full, eloquent
    sentences and first person.>, box: <a box
    surrounding the object that is being
    redrawn in the provided image. First
    consider which object needs to be redrawn
    . Then, find the EXACT location and size
    of this object by looking at both the
    provided image and the html code. Based
    on this location and size, generate a box
    that surrounds this object using the
    format [start_x, start_y, width, height]
    that uses pixels as its units (e.g.
    [0,0,50,50] would be a 50x50px box in the
    top left corner of the image). When
    applied on top of the image, the box
    should completely encompass the object.>,
    svg: <simple svg representing the new
    visual, approx 100px by 100px> }

NO EXPLAINATION.
```

A.3.7 Populate Edit Assumptions.

```
In a previous prompt, you generated HTML code
    for an interactive simulation based on
    the provided UI Map.

However, there's an issue with this code/UI
    map: ${prompt}.

This issue may also be outlined via the red
    annotations in the provided image, which
    are labeled with labels such as "A1" or "
    A2" near the annotation.
These labels may be referenced in the above
    issue description.

We think editing the details of one of the
    nodes in the UI Map might solve the above
    issue.
```

```
We want some more information on how we
    should edit these details to fix the
    issue.

Please respond in the following format:
{ type: 3, message: <SHORT message explaining
    your suggestion. Use full, eloquent
    sentences and first person.>, node: <id
    of the node whose assumptions are being
    updated>, assumptions: <list of
    assumptions updated to explicitly fix the
    issue> }

NO EXPLAINATION.
```

A.3.8 Populate Add Edge.

```
In a previous prompt, you generated HTML code
    for an interactive simulation based on
    the provided UI Map.

However, there's an issue with this code/UI
    map: ${prompt}.

This issue may also be outlined via the red
    annotations in the provided image, which
    are labeled with labels such as "A1" or "
    A2" near the annotation.
These labels may be referenced in the above
    issue description.

We think adding an edge to the UI Map might
    solve the above issue conceptually.
We want some more information on the edge
    that should be added.

Please respond in the following format:
{ type: 2, message: <SHORT message explaining
    your suggestion. Use full, eloquent
    sentences and first person.>, source: <id
    of the source node>, target: <id of the
    target node>, label: <new edge name>}

NO EXPLAINATION.
```

A.3.9 Populate Add Node.

```
In a previous prompt, you generated HTML code
    for an interactive simulation based on
    the provided UI Map.

However, there's an issue with this code/UI
    map: ${prompt}.

This issue may also be outlined via the red
    annotations in the provided image, which
    are labeled with labels such as "A1" or "
    A2" near the annotation.
These labels may be referenced in the above
    issue description.

We think adding a node to the UI Map might
    solve the above issue conceptually.
We want some more information on the node
    that should be added.

Please respond in the following format:
{ type: 1, message: <SHORT message explaining
    your suggestion. Use full, eloquent
    sentences and first person.>, label: <new
    node name> }

NO EXPLAINATION.
```

```
Your designers want to know what nodes and
    links in the given UI interactivity graph
    correspond to the region(s) circled in
    red in your prototype.
Given an image, the code used to generate the
    prototype in the image, and a graph, you
    respond with the subgraph corresponding
    to the elements and relationships these
    region(s) represents.

A good subgraph has the:
- nodes corresponding to the visuals/elements
    circled in red
- nodes corresponding to the attributes of
    the elements circled in red
- links corresponding to the relationships
    between all included nodes
- original label for EVERY node and edge

NO explanation. Only respond with the
    mermaidjs graph.
```

A.3.10 Code Assumptions. The following prompt is used to indentify the current code assumptions abstraction.

```
Given the following html code and the UI
    Interactivity Map that it represents,
    please give me a json object with an
    attribute for each node in the UI Map.
    The name of the each attribute should be
    the label, or value in square brackets,
    of each node (DO NOT NAME IT THE NAME IN
    SQUARE BRACKETS). The value for each of
    these attributes should be a list of all
    of the details and assumptions that the
    code is making about this node. For
    example, the numeric range of a node
    representing a slider, attributes of a
    phypical display or graph, all formulas
    used to caluculate displayed or output
    values, and the way physical objects are
    visually represented.

A good list of assumptions:
- Doesn't reference specific variables in the
    code but instead speaks generically
- Pays close attention to the implementation
    in order to identify assumptions that are
    not intended (i.e. bugs)

NO explaination.

Html Code: ${htmlCode}
UI Map: ${UIMap}
```

The following prompt is used to update the code based on the user's changes to the code assumptions.

```
The given html code is the implementation of
    an experiment based on the following
    concept graph. Please update the html
    code so that it implements the details
    outlined in the given details list for
    the ${node} object.

A good response:
- contains all the same functionality of the
    original html code
- updates the code so that ${node} now abides
    by the provided list of details
- ${roughDirections}

NO explanation.

Graph:${graph}
Html Code:${htmlCode}
Node of interest: ${node}
Details:${JSON.stringify(newAssumptions)}
```

Redraw

The following prompt is used to redraw a visual when the LLM identifies that the user wants to update it (from a chat or error description message).

```
Given the above low-fidelity hand-drawn image
    , create a high fidelity SVG representing
    the object that the hand-drawn image is
    of.

A good SVG response:
- interprets the object that the hand-drawn
    image is of (typically this object is one
    of the objects included as nodes in the
    provided UI Map)
- has similar shape and features to the hand
    drawn image
- uses gradients and additional shapes if
    needed
- ignores the fact that the hand-drawn sketch
    is drawn in red

NO EXPLAINATION
```

The following prompt updated the simulation code to use a new svg for a specific visual.

```
Given the above image and html code, replace
    the SVG circled in red in the image with
    the provided SVG in the code. Respond
    with the ENTIRE updated HTML code and
    nothing selse.

A good response MUST:
- have a the old circled SVG replaced with
    the new provided SVG
- resize the new SVG so it is the same size
    as the original (not just cropping the
    new SVG, but actaully reducing/increasing
    its attributes)
- have all of the functionality of the
    original code
- ${roughDirections}

NO EXPLAINATION
```

A.3.11 Auto-Add Edges. Below is the prompt used to automatically add links when a new node is added by the user.

```
Please add a new node to the following
    mermaid graph with the provided label.
    Add any edge that you think reasonably
    connect this new object to the rest of
    the graph.

NO explaination.
```

```
${mermaidDirections}

Graph: ${UIMap}
Label: ${newNodeName}
```

A.4 Automated Testing

The following prompt is used to generate test cases for automated testing.

```
You had previously created the following HTML
     code based on the UI Map and learning
    goal.
More info:
- HTML Code: ${htmlCode}
- Learning Goal: ${selectedHypothesis}
- UI Map: ${UIMap}.
Please generate a JSON array of structured
    test cases for the interactive simulation
    .
Each test case should be a JSON object with
    the following keys:
- uiElementId: The ID of the UI element to
    interact with.
- actionType: The type of action (one of "
    click", "set_value", "toggle", "
    verify_content").
- actionValue (optional): The value to set (
    if applicable).
- description: A brief description of what is
     being tested for.
- expectedOutcome: A brief description of the
     expected result.
- isUIVerification: A boolean indicating
    whether this test case is related to
    changes in any of the UI components or
    visual elements in the simulation.
Return only the JSON array, with no
    additional text, and ensure that the JSON
     starts between <START> and <STOP> tags.
```

Below is the prompt used to verify test results using debug log information and Puppeteer screenshots.

```
You had previously created the following HTML
     code based on the UI Map and learning
    goal.
More info:
- HTML Code: ${htmlCode}
- UI Map: ${UIMap}
- Learning Goal: ${selectedHypothesis}
- JavaScript Errors captured: ${.join('; ')}.
${debugLogsText}
${initialScreenshotNote}

Note: An initial screenshot of the UI (before
     any interactions) is available.

Based on the above context, including the
    test case results, screenshots and
    runtime logs, do the following:
1) Verify whether the actual outcomes in each
     test case (as described below) match the
     expected outcomes.
2) If discrepancies or errors remain, update
    the HTML code so that all test cases pass
     and errors are resolved.
3) Verify that the simulation satisfies the
    learning goal.
IMPORTANT: Return only the updated HTML code,
     starting between <START> and <STOP> tags
    , or return "PASS" if no changes are
    needed.

Here is the:
1) Structured list of test case results and
    verification details,
2) Indication of whether the HTML code needs
    to be changed.
If no changes are needed, simply return "PASS
    ".
If changes are required, return the updated
    HTML code starting between <START> and <
    STOP> tags.
```

The following prompt is used to fix js errors in the simulation code when they are present.

```
You had previously created the following HTML
     code based on the UI Map and learning
    goal.
More info:
- HTML Code: ${htmlCode}
- UI Map: ${UIMap}.
However, the HTML code is generating these
    JavaScript errors: ${errorMessages.join
    ('; ')}.
Please update the HTML code so that these
    errors are fixed while preserving all the
     original functionality.
Return only the updated HTML code, starting
    between <START> and <STOP> tags.
```

B Visualizations of Survey Responses from User Evaluation

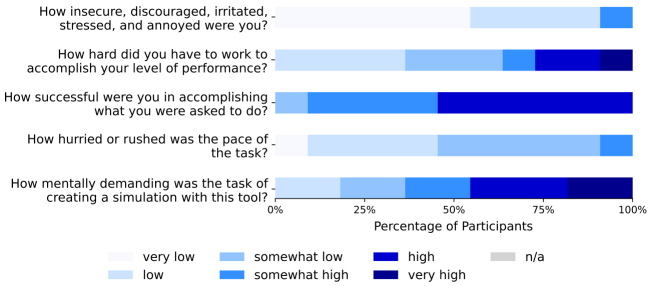


Figure 11: Teacher Participant Responses to NASA's Task Load Index (NASA-TLX) Questions

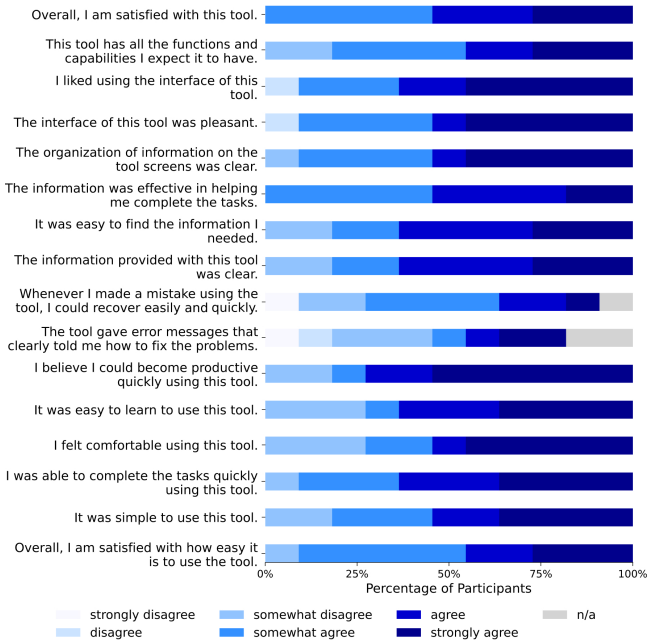


Figure 12: Teacher Participant Responses to Post-Study System Usability Questionnaire (PSSUQ)

Appendix: Evaluation Instructions

Answer the four questions below by reviewing the stages in order.

1. Conceptual Completeness/Coherence (1–3)

Goal: Check that the concepts introduced at each stage are understandable in explaining the content that the input texts describe.

What to do:

(1) Start with the Concept Graph. Consider the text describing the content of this graph. Make a mental list of key terms and relationships (e.g., "buoyant force," "mass," "fluid"). How well does this list align with the graph itself? Are there extra objects or relationships, or missing ones?

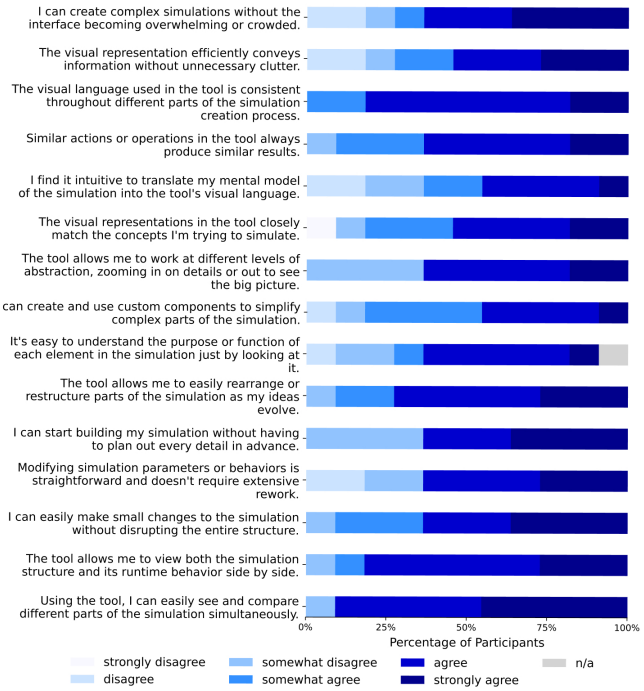


Figure 13: Teacher Participant Responses to a Questionnaire on the Cognitive Dimensions of Notations

(2) Look at the Learning Goal Graph and compare it to the included learning goal: Do the ideas in this stage represent this distinct learning goal?
(3) Then check the UI Graph: Are the added UI controls useful for helping a student understand the content?

Rate:

3: Either all included concepts and relationships are important and none are missing, *or* I can tell what concepts and relationships need to be added.
2: A few concepts or relationships are missing or extraneous *and* I cannot tell which should be added or removed.
1: Many missing or extraneous concepts or relationships *and* I cannot tell which should be added or removed.

2. Inclusion Consistency (1–3)

Goal: Check that the concepts introduced early are carried through all the way to the final simulation.

What to do:

(1) Start with the Concept Graph. Make a mental list of key terms (e.g., "buoyant force," "mass," "fluid").
(2) Look at the Scenario Graph: Do those same ideas appear, just with different names or examples? (e.g., "buoyant force" → "lift force on balloon" is okay).
(3) Then check the Learning Goal, UI Graph, and Simulation: Are these concepts still present and used meaningfully?

Rate:

3: All key ideas appear in each stage.
2: A few missing or unclear connections.
1: Many dropped or added concepts not explained earlier.

3. Meaning Drift (1–3)

Goal: Check that the concepts don't change meaning in ways that confuse or contradict earlier stages.

What to do:

(1) Start again with the Concept Graph: What does each idea mean?
(2) Check the Scenario Graph and Learning Goal: Did anything get misused or redefined?
(3) In the UI Graph or Simulation, does the behavior reflect the correct meaning? (e.g., "increase weight" should lower height if gravity is being modeled).

Rate:

3: Concepts stay consistent.
2: Some minor changes in meaning.
1: Concepts shift or are used incorrectly.

4. Simulation Working (1–3)

Goal: Check if the final simulation works and is usable.

What to do:

(1) Try the simulation (if interactive), or inspect the behaviors described in the code or preview.
(2) Do sliders/buttons respond? Do variables update? Are values and charts shown correctly?

Rate:

3: Everything works as expected.
2: Some small bugs or glitches.
1: Simulation is broken or hard to understand.

C Experimental Procedure Generation

Simulations addressing **descriptive knowledge** generate their UI Graph by first identifying (1) the independent variable of the learning goal, (2) the dependent variable, and (3) the relationship between these two variables. The third characteristic is identified as the learning goal itself, but SimStep's process also prompts for the independent and dependent variables. Then all these are characteristics are used to generate an experimental procedure through LLM prompting. Finally, SimStep prompts to translate this procedure into an UI Graph with all necessary interactions to complete this procedure. For instance, if the learning goal is to understand how sunlight affects the height of plants, the procedure will involve the independent variable — amount of sunlight, the dependent variable — the height of the plant, and the relationship: more sunlight leads to taller plants. The specific procedure might be as follows:

```
(1) Select the plant for observation.
(2) Expose the plant to full sunlight
    for a week and measure its
    height daily.
(3) Change the condition to partial
    sunlight for the next week and
    continue measuring its height
    daily.
(4) Reduce sunlight to no sunlight
    for the final week and again
    measure its height daily.
(5) Record and compare the data to
    determine how the different
    sunlight conditions affected the
    plant's growth.
```

For **explanatory knowledge**, we find four main characteristics that must be identified. First, SimStep prompts for the main experimental object that the learning goal is exploring. Then SimStep again identifies the independent and dependent variables of the learning goal, along the explanatory object, or the object that explains the relationship between the independent and dependent variables. All four of these characteristics are then used to prompt for an experimental procedure, which is then translated to the UI Graph.

Learning goals that address **procedural knowledge** use two main characteristics. Namely, SimStep prompts for both the main experimental object that the learning goal is investigating, along with the underlying process that is explained through the learning goal. Using both of these features, SimStep then generates an experimental procedure to uncover the process presented in the learning goal. And again, this is translated into a UI Graph including all necessary experimental objects and processes.