FROM EDUVISBENCH TO EDUVISAGENT: A BENCHMARK AND MULTI-AGENT FRAMEWORK FOR REASONING-DRIVEN PEDAGOGICAL VISUALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While foundation models (FMs), such as diffusion models and large vision-language models (LVLMs), have been widely applied in educational contexts, their ability to generate pedagogically effective visual explanations remains limited. Most existing approaches focus primarily on textual reasoning, overlooking the critical role of structured and interpretable visualizations in supporting conceptual understanding. To better assess the visual reasoning capabilities of FMs in educational settings, we introduce EduVisBench, a multi-domain, multi-level benchmark. EduVisBench features diverse STEM problem sets requiring visually grounded solutions, along with a fine-grained evaluation rubric informed by pedagogical theory. Our empirical analysis reveals that existing models frequently struggle with the inherent challenge of decomposing complex reasoning and translating it into visual representations aligned with human cognitive processes. To address these limitations, we propose EduVisAgent, a multi-agent collaborative framework that coordinates specialized agents for instructional planning, reasoning decomposition, metacognitive prompting, and visualization design. Experimental results show that EduVisAgent substantially outperforms all baselines, achieving a 40.2% improvement and delivering more educationally aligned visualizations.

1 INTRODUCTION

"To truly teach is not to tell the answer, but to illuminate the path."

While foundation models (FMs), such as diffusion models and large vision-language models (LVLMs), have been extensively adopted in educational domains (Chu et al., 2025; Wang et al., 2024), including pedagogical agents providing automated classroom assistance and science learning agents offering textual explanations of problem-solving processes (Wu et al., 2023), their applications have predominantly focused on text-based interactions (Wu et al., 2023; Xu et al., 2024). However, in education, especially K-12 settings, creating compelling visualizations is crucial for cognitive comprehension and overall learning effectiveness (Presmeg, 2006). Despite its importance, there is currently limited understanding of how FMs can effectively generate visually grounded elements (e.g., *diagrams*, *interactive education tools*, *illustrative graphics*) to support the pedagogical illustration of problem-solving processes.

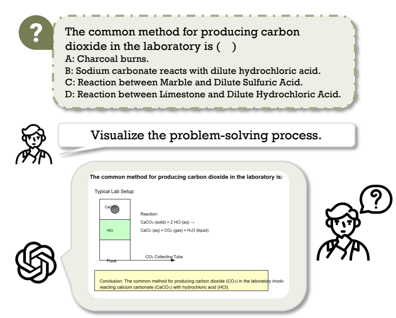


Figure 1: GPT-4o fails to illustrate its problem-solving with high-quality, logical, and explanatory visualization.

Currently, generating visually grounded elements for pedagogical reasoning poses several challenges: (1) decomposing complex reasoning into representable steps that align closely with human cognitive processes is non-trivial (Yang et al., 2024; Chen et al., 2024d); (2) precisely producing visual aids for each sub-step to optimally support learners is challenging (Hong et al., 2025); and (3) different educational domains require distinct visualization styles and formats, which makes consistent and adequate visual aid delivery difficult (Pandey & Ottley, 2025).This difficulty stems not just from technical rendering challenges, but from the complex task of translating abstract pedagogical concepts into intuitive visual narratives. Addressing these obstacles first requires a picture of how current FMs perform, so that future models can be purpose-built to close the gaps. Consequently, a comprehensive evaluation platform is critical for systematically assessing FMs on visual pedagogical reasoning.

To bridge this gap, we introduce **EduVisBench**, a multi-domain, multi-level benchmark designed to evaluate the capacity of FMs to generate pedagogically effective, step-by-step visual reasoning. EduVisBench comprises structured problem sets across diverse domains, each requiring multimodal-centric reasoning and solutions that prioritize visualization principles such as *interpretability*, *cognitive alignment*, and *instructional clarity* to achieve high evaluation score. To facilitate a detailed evaluation, we further develop a fine-grained rubric enabling multidimensional assessments of AI-generated visual outputs, focusing explicitly on pedagogical criteria such as *contextual relevance*, *visual clarity*, *multimodal coherence*, *reasoning support*, and *interactive engagement*.

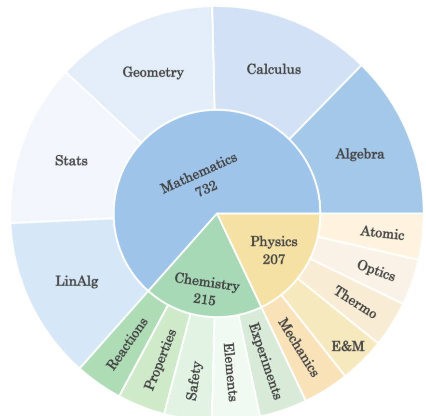


Figure 2: Dataset distribution of EduVisBench. Each domain encompasses various sub-domains, collectively covering 15 comprehensive pedagogical scenarios.

Utilizing this benchmark, we conduct extensive evaluations on a variety of FMs and agents. Our findings reveal that although current models achieve predominantly correct step-by-step textual analyses, they frequently fail to generate useful or faithful visualizations, as depicted in Figure 1. Specifically, our systematic analysis highlights recurring challenges including (1) semantic misalignments between textual explanations and visual components, (2) omissions of critical steps within rendered diagrams, and (3) structural inconsistencies in code-based visual outputs, collectively undermining accuracy, clarity, and interactivity.These shortcomings collectively compromise the pedagogical utility of the generated content, often leading to more confusion than clarity for the learner.

To address these limitations, we introduce a multi-agent collaborative framework, **EduVisAgent**, designed to simulate the complete learning journeyfrom initial problem exposure to deep conceptual understanding. Specifically, a central planning agent orchestrates five specialized expert agents dedicated to *visualization design*, *cognitive scaffolding*, and *metacognitive regulation*. A synthesis module then integrates these expert outputs into interactive, personalized learning webpages tailored specifically to human learners. Experimental results demonstrate that our proposed method EduVisAgent achieves an average improvement of 40.2% than current SOTA method. This underscores the effectiveness of our approachleveraging modular specialization and collaborative integration to produce robust and visually grounded learning solutions.

2 EduVisBench Benchmark

2.1 Overview

In this section, we introduce EduVisBench, a novel and challenging benchmark designed to evaluate the capability of models to generate logical and explanatory visualizations for educational purposes. As shown in Figure 2, EduVisBench comprises 1,154 carefully curated STEM questions across three academic subjects and 15 distinct domains, organized into three levels of difficulty. In addition to assessing accuracy in step-by-step problem solving, EduVisBench places particular emphasis on a models ability to communicate the reasoning process clearly and visuallyhelping students understand problems through structured, interpretable visual outputs, as illustrated in Figure 3.

Specifically, EduVisBench adopts a multimodal setting in which models are provided with both textual and visual inputs and are tasked with producing diverse output formats, including interactive web pages and visual diagrams. Beyond evaluating the correctness of final answers, we introduce a fine-grained evaluation framework that assesses the quality of visualizations across five key dimensions: (1) the logical sequencing of visual elements, (2) the structural richness of the visuals, (3) semantic

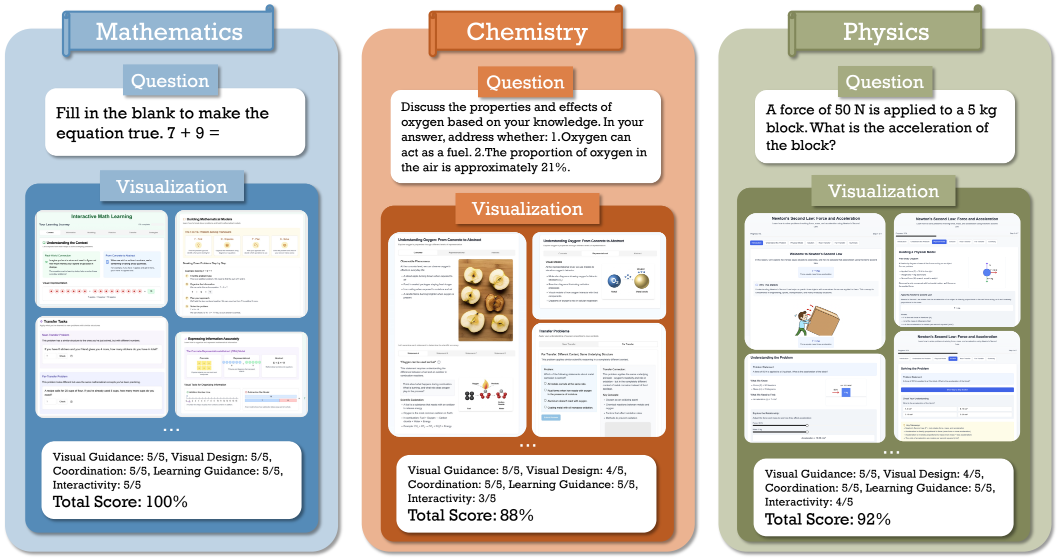


Figure 3: Representative examples from EduVisBench, featuring questions from Maths, Chemistry, and Physics alongside their corresponding high-scoring visual explanations. These interactive visualizations, generated by our multi-agent system EduVisAgent, exemplify well-designed, pedagogically effective outputs for STEM problems.

alignment with the underlying subject matter, (4) the clarity and guidance provided for problem-solving, and (5) the level of interactivity and engagement. In the following subsections, we describe our dataset curation process and the design of the evaluation rubric in detail.

2.2 Dataset Curation

EduVisBench is built from several high-quality public educational resources that we carefully curated, translated, and adapted to support multimodal visualization learning tasks. Specifically, the chemistry questions are sourced from the *C-MHChem-Benchmark* (Zhang et al., 2024), originally presented in Chinese and meticulously translated into English with careful attention to scientific accuracy and terminology. The physics questions are drawn from the *high-school-physics* (Rohith, 2023) dataset, which includes a range of conceptual and quantitative exercises suitable for secondary-level learners. The mathematics component combines easy-level problems from the Illustrative Mathematics curriculum with medium- to hard-level questions selected from the *MATH-500* (Lightman et al., 2023) dataset. Furthermore, each domain encompasses diverse sub-domains, collectively covering 15 comprehensive scenarios, as illustrated in Figure 2. All data sources were standardized into a unified format and consolidated to enable consistent and comprehensive evaluation across subjects.

2.3 Evaluation Metric

In this subsection, we detail the performance evaluation rubrics in EduVisBench.

Evaluation Dimensions. To comprehensively evaluate the quality of generated visualizations in supporting student understanding and learning, we introduce a fine-grained scoring metric grounded in five pedagogically motivated dimensions: **(1) Context Visualization**: evaluates how clearly the visualization situates the problem within a relevant context; **(2) Diagram Design**: assesses the clarity, accuracy, and effectiveness of the diagrams used to represent information; **(3) Text–Graphic Integration**: measures the coherence between textual explanations and visual elements, ensuring mutual interaction; **(4) Thought Guidance**: examines the extent to which the visualization supports reasoning processes and highlights critical thinking steps; **(5) Interactivity**: evaluates whether and how the visualization invites students engagement, reflection, or active manipulation. Each dimension captures a distinct aspect of effective multimedia learning, with detailed rubrics provided in Appendix A.1 to guide the scoring process.

Evaluation Protocol. As shown in Figure 4, models are provided with a visualization prompt together with a question and are asked to generate visual outputs. To enable fair comparison across heterogeneous outputs, we first canonicalize every model result to a raster image prior to scoring. This standardization is a crucial step that ensures all systems are evaluated on a level playing field, independent of their native modality or file format, and prevents format-specific rendering artifacts from biasing the assessment. Visuals produced directly as SVG or PNG are used as-is. Web pages (HTML or Next.js) are rendered in a headless browser and captured as screenshots of the primary view; when lightweight interactivity is present (e.g., buttons, tabs, or toggles), we systematically traverse the reachable states and retain one representative screenshot per state. All resulting images are then evaluated by GPT-4o along five dimensions defined in Appendix A.2 to compute an overall performance score. Each dimension is rated on a 0-5 scale; the ratings are summed (0-25) and, when appropriate, normalized to a percentage to yield the final overall score.

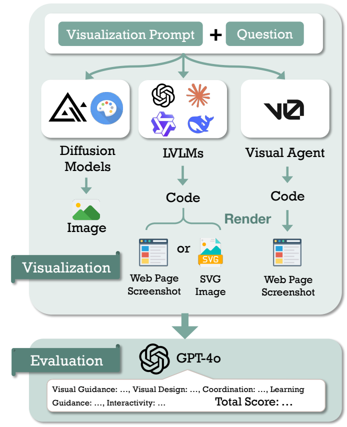


Figure 4: Workflow for evaluation.

3 EDUVISAGENT

Using EduVisBench we systematically evaluate the performance of existing text-to-image models and LVLMs (see detailed results in Table 1 in Section 4). We find that most models perform poorly, with average scores below 50 on a 0-100 scale. This significant performance gap serves as the primary motivation for developing a novel approach that moves beyond monolithic model architectures. This underperformance underscores the inherent challenge of decomposing complex reasoning and translating it into visual representations that align with human cognitive processes to effectively support educationa task that remains highly non-trivial.

To address these challenges, we propose a multi-agent system, EduVisAgent, inspired by pedagogical theories and designed to emulate the division of labor and collaborative reasoning found in expert instructional design. EduVisAgent consists of five specialized yet interdependent agents: a **Task Planning Agent**, which structures the instructional objective; a **Conceptual Mapping Agent**, which extracts and organizes key information; a **Reasoning Decomposition Agent**, which constructs step-by-step problem-solving logic; a **Metacognitive Reviewer**, which encourages summarization and learner reflection; and a **Visualization Agent**, which generates appropriate visual representations. This design introduces modularity and pedagogical interpretability by embedding distinct instructional roles directly into the agent workflow. The overall operation of EduVisAgent proceeds in two stages: (1) instructional flow construction and (2) collaborative solution generation, as detailed below.

3.1 INSTRUCTIONAL FLOW CONSTRUCTION

The first stage of EduVisAgent focuses on formulating a well-structured instructional task based on the original problem. A key challenge lies in analyzing the underlying reasoning structure, identifying implicit logical dependencies, and associating each reasoning step with relevant conceptual knowledge. To address this, we employ the **Task Planning Agent**, which systematically organizes the problem into an instructional format suitable for multimodal visualization. Its main functions include: (1) breaking down the problem into coherent subgoals, (2) clarifying the reasoning expected at each step, (3) aligning each step with domain-specific principles or formulas, and (4) anticipating potential student misconceptions or cognitive needs. This structured formulation provides a pedagogically grounded foundation that guides the downstream agents in generating coherent, targeted, and educationally effective visual explanations.

3.2 COLLABORATIVE SOLUTION GENERATION

In this stage, EduVisAgent executes the instructional task constructed by sequentially activating a set of specialized agents, each responsible for completing a specific aspect of the task. As shown in Figure 5, these agents operate in a coordinated manner to enhance the coherence of instructional logic, improving the clarity of visual representation, and ensure alignment with educational objectives. Specifically,we detail each agent as follows.

Conceptual Mapping Agent. This agent is responsible for extracting and organizing the core components of the input problem. Drawing on the ConcreteRepresentationalAbstract (CRA) instructional model (Nugroho & Jailani, 2019), it classifies information into three categories: concrete entities, representational elements, and abstract constructs. This structured classification helps bridge the gap between the concrete elements of a problem and the abstract principles required to solve it. This progression from concrete to abstract is particularly valuable for an AI system, as it provides a structured pathway to ground complex concepts in relatable terms before generating symbolic representations.The agent conducts fine-grained categorization and semantic summarization to support downstream visualization modules.

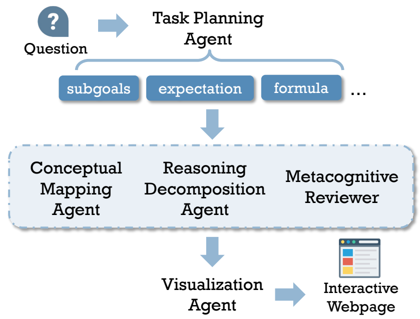


Figure 5: The structure of EduVisAgent.

Reasoning Decomposition Agent. This agent decomposes complex problems into manageable subcomponents and provides step-specific instructional guidance. It applies the memory-oriented FOPS strategy (Miller & Cohen, 2020)find the problem type (e.g., equation solving, conceptual reasoning, commonsense application, or graphical interpretation), organize the structure via equations or diagrams, plan the solution path, and solve the task. Based on the decomposed steps, the agent also identifies critical instructional points that require additional support, especially those that benefit from visual scaffolding or interactive guidance.

Metacognitive Reviewer. Grounded in metacognitive theory (Schraw & Moshman, 1995), this agent supports learners in monitoring their comprehension and reasoning processes. It generates reflective prompts that foster self-questioning and self-correction, encouraging learners to evaluate the soundness of their problem-solving approaches.

Visualization Agent. This agent is responsible for constructing the visual guidance component of the instructional output. Instead of relying on decorative visuals, it emphasizes the use of abstract yet pedagogically effective representationssuch as number lines, bar charts, schematic object illustrations, graphic organizers, sketch diagrams, and structured data tables. The agent ensures that each visualization is tightly aligned with the underlying abstract concept being taught. All visuals are rendered using the `v0` (Vercel, 2025) system for web-based deployment.

4 EXPERIMENTS

This section outlines the experimental setup for benchmarking various foundation models on EduVisBench. We evaluate Diffusion Models, LVLMs, a specialized visualization agent (`v0`), and our proposed EduVisAgent. Our investigation seeks to address the following key questions: (1) How proficient are existing models at generating high-quality, explanatory visualizations within EduVisBench? (2) Can the proposed EduVisAgent system outperform current models? (3) What distinct performance patterns emerge across different model architectures, academic disciplines, and evaluation dimensions in EduVisBench?

4.1 EXPERIMENT SETUP

Baseline Models. Our experimental evaluation encompasses a range of FMs, categorized as follows: (1) Image Generation Models: This category includes Flux.1-dev (Labs, 2024), Stable Diffusion 3.5 Large (SD3.5) (IT Admin, 2024), and Stable Diffusion XL Base 1.0 (SDXL) (Podell et al., 2023). These models are tasked with generating static images directly from textual or visual inputs. (2) Large Vision-Language Models (LVLMs): We evaluate Deepseek-VL2 (Wu et al., 2024), GLM-4V-9B (GLM et al., 2024), MiniCPM-V2.6 (Yao et al., 2024), Mistral-Small-3.1-24B-Instruct-2503 (Mistral AI, 2025), Phi-3.5-Vision-Instruct (Abdin et al., 2024), Phi-4-Multimodal-Instruct (Abouelenin et al., 2025), Qwen2.5-VL-72B (Team, 2025), GPT-4o (Hurst et al., 2024), Claude 3.7 Sonnet (Anthropic, 2025), and Gemini 2.0 Flash (Mallick & Kilpatrick, 2025). These models are prompted

Table 1: Performance of Diffusion Models, Large Vision Language Models and v0 on EduVisBench.

Method	Vis. Type	Maths			Physics			Chemistry			Avg
		Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	
Diffusion Model											
Flux.1-dev	Image	13.8	13.4	13.2	11.7	8.5	10.0	20.0	16.6	16.0	13.8
SD3.5	Image	17.3	20.3	18.8	16.8	13.0	12.0	22.8	21.7	34.0	18.4
SDXL	Image	17.3	23.3	25.5	18.9	15.4	24.0	33.6	30.2	24.0	21.8
Large Vision Language Model											
Deepseek VL2	Webpage	20.3	17.1	15.7	17.9	17.0	20.0	16.4	13.8	14.0	17.5
GLM-4V-9B	Webpage	22.3	21.1	19.4	24.5	21.5	24.0	22.3	21.5	16.0	21.9
MiniCPM-V-2.6	Webpage	24.1	17.3	15.5	19.1	17.4	20.0	14.5	15.2	12.0	19.3
Mistral-Small-3.1	Webpage	29.1	31.6	32.2	32.3	33.5	20.0	30.6	27.5	24.0	30.2
Phi-3.5	Webpage	25.3	20.7	19.1	21.2	19.5	12.0	20.0	18.6	20.0	21.8
Phi-4	Webpage	26.1	25.1	22.9	27.8	25.5	24.0	31.2	27.5	12.0	26.4
Qwen2.5-VL-72B	Webpage	24.3	18.1	15.8	19.7	17.1	24.0	18.2	16.4	12.0	20.0
Claude 3.7 Sonnet	SVG	61.2	26.7	23.6	18.5	16.9	14.0	47.5	47.2	18.0	42.0
Claude 3.7 Sonnet	Webpage	56.2	57.5	55.6	44.8	42.6	24.0	61.1	60.6	64.0	54.6
GPT-4o	Webpage	47.6	39.3	37.9	25.7	24.2	24.0	34.3	32.6	36.0	38.1
GPT-4o	SVG	36.1	19.7	19.5	13.0	12.8	4.0	30.0	27.5	22.0	26.3
Gemini 2.0 Flash	Webpage	46.9	9.5	15.7	31.7	26.5	24.0	32.0	25.8	30.0	43.6
Visualization Agent											
v0	Webpage	63.0	37.6	47.2	53.3	58.5	52.0	74.7	52.8	68.0	58.2

to generate SVG or HTML code, which is then rendered into visual outputs for evaluation. (3) Specialized Visualization Agent: We also assess v0 (Vercel, 2025), an AI agent specifically designed to create interactive web pages based on instructional content.This diverse selection of models was chosen to represent the current state-of-the-art across different architectural paradigms.

Evaluation Setups. During evaluation, all generated visualizations are standardized into image format. For interactive web pages containing buttons, an automated script navigates through all accessible sub-pages, capturing individual screenshots of each. This automated approach ensures that the evaluation is both scalable and free from subjective human bias during the rendering process. Performance is assessed using the evaluation metric described in Section 2.3, where GPT-4o scores the visual outputs based on predefined rubrics, assigning a score from 0 to 5 for each of the five dimensions. The cumulative score (maximum 25 points) is then normalized to a 0-100 scale for standardized reporting and comparison.

Reliability of GPT-based Scoring. To validate the reliability of our automatic judge, we compared GPT-based evaluations with human evaluations. Specifically, we selected 50 samples from each subject category (Chemistry, Math, and Physics), and had both GPT and human evaluators independently rate them. Human evaluators were undergraduate students from top universities. We measured agreement using Cosine Similarity and Mean Squared Error (MSE). As shown in Table 2, high agreementaverage cosine similarity 0.9655 and MSE 0.5702 across subjects indicates negligible practical discrepancy .

Table 2: Cosine similarity and mean squared error across subjects. Math is the average of Math500 and IllustrativeMath, each with 50 samples.

Metric	Chemistry	Math	Physics	Average
Cosine Similarity	0.9742	0.9557	0.9666	**0.9655**
MSE	0.3895	0.7093	0.6118	**0.5702**

4.2 Baseline Benchmarking

The performance of all evaluated baseline models is detailed in Table 1. Across all evaluated models, the average scores indicate significant room for improvement. Diffusion Models generally exhibited the lowest performance, with average scores ranging from 13.8% (Flux.1-dev) to 21.8% (SDXL). This suggests that direct static image generation, while capable of producing visual elements, struggles substantially with the nuanced requirements of explanatory and guiding visualizations for complex logical problems in our benchmark.

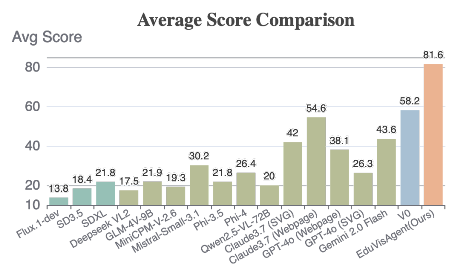
LVLMs typically scored between 17.5% (Deepseek VL2) and 30.2% (Mistral-Small-3.1). Notable exceptions include Gemini 2.0 Flash (43.6%) and Claude 3.7 Sonnet; the latter's significantly better performance with Webpages (54.6%) over SVG (42.0%). GPT-4o also showed a preference for Webpage generation (38.1%) over SVG (26.3%), suggesting that prompting advanced LVLMs for structured interactive webpages can yield more effective visual explanations. Nevertheless, even these top-tier LVLMs face considerable challenges in consistently meeting all of evaluation criteria. The visualization agent `v0`, specifically designed for webpage generation, achieved the highest average score among all baseline models at 58.2%. This result highlights the advantage of a specialized agent in this task over more general-purpose FMs.

Table 3: Overall comparison of models: left is our EduVisAgent performance, right is the bar chart. EduVisAgent achieves the highest average score among all models.

(a) Performance of our EduVisAgent on EduVisBench.

Method	**Vis. Type**	**Easy**	**Medium**	**Hard**	**Avg**
EduVis Agent	Webpage	**Maths**			**81.6**
		90.2	64.5	65.0	
		Physics			
		85.3	81.7	84.0	
		Chemistry			
		69.0	76.3	76.0	

(b) Comparison of average score across all models.



4.3 Performance Analysis of EduVisBench

Building upon the insights gained from the baseline evaluations, we assessed our proposed multi-agent system, EduVisAgent. The results in Table 3 demonstrate a substantial leap in performance for generating explanatory and logically valuable visualizations for STEM problems. EduVisAgent achieved an impressive overall average score of 81.6%. Specifically, EduVisAgent surpasses the best-performing baseline `v0` (58.2%), by a remarkable 23.4 percentage points. This constitutes an approximately 40.2% relative improvement, underscoring the efficacy of our multi-agent architecture and the integration of educational methodologies. Compared to the best performing LVLM (Claude 3.7 Sonnet Webpage at 54.6%) and the top diffusion model (SDXL at 21.8%), the advancement offered by EduVisAgent is even more pronounced. These results clearly indicate that the design principles underlying EduVisAgent, which incorporate a multi-agent structure and pedagogical strategies, effectively address many of the limitations observed in existing generative models.

4.4 Case Analysis

To further illustrate the limitations of existing baselines and how our approach addresses these challenges, we present two case studies in Figure 6. On the left, for a chemistry question, the GPT-4o-generated solution lacks intuitive visualization of the chemical processes, resulting in fragmented information without visual guidancereflected in a low score of just 28%. In contrast, EduVisAgent begins by displaying background images of the relevant chemical elements, activating students prior knowledge. This strategy effectively connects abstract chemical concepts to tangible, everyday experiences, a well-established method for enhancing comprehension and retention. It then contextualizes each of the four answer options with real-world scenarios, thereby enhancing students' understanding of the underlying chemical transformations.

Conversely, for the Carnot cycle efficiency physics problem (right side of Figure 6), the Gemini solution presents a single, flawed chart. Its depiction of 300K and 400K temperatures with identical heights introduces visual misinformation, failing to accurately represent data differences and thereby diminishing its pedagogical value. In stark contrast, EduVisAgent employs a multi-agent collaborative approach: it first generates a concrete factory scene to activate students' working memory of the "heat engine" concept. Subsequently, it constructs an accurate Carnot cycle diagram and offers a step-by-step problem breakdown, fostering clear conceptual understanding. Crucially, EduVisAgent

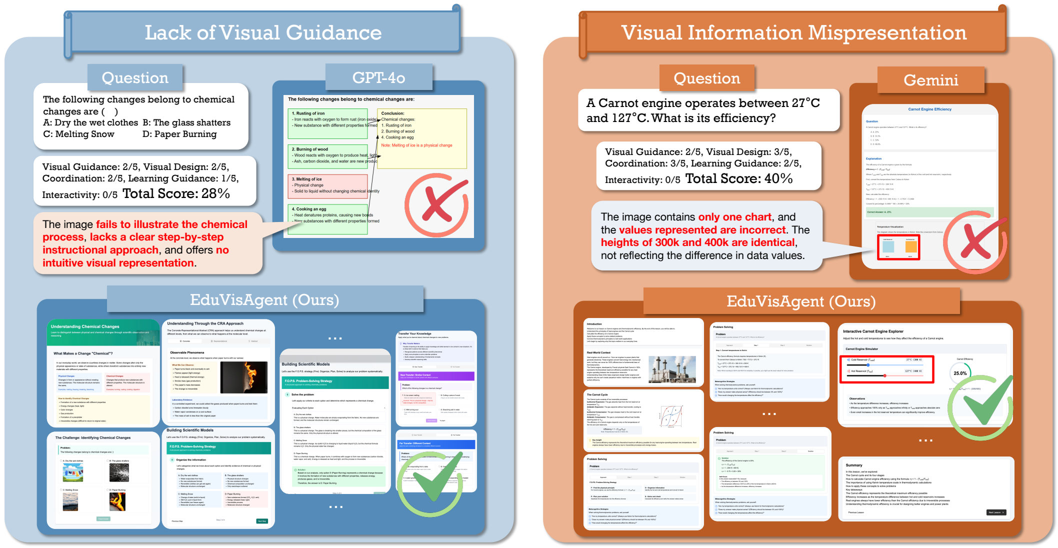


Figure 6: Baseline models versus our EduVisAgent. These examples clearly demonstrate the often poor output quality of baseline models, contrasting sharply with the high-quality, effective visualizations produced by EduVisAgent.

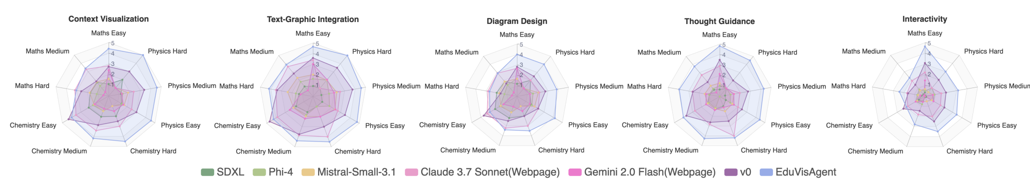


Figure 7: Fine-grained performance comparison across our five key evaluation dimensions.

provides interactive visualization components, enabling users to dynamically adjust temperatures via sliders and observe real-time changes in heat engine efficiency. This interactive element transforms the learner from a passive observer into an active participant, which is known to deepen engagement and learning. This interactive engagement significantly facilitates higher-order thinking skills.

Overall, through coordinated multi-agent optimization of image design, instructional structure, and learning pathways, EduVisAgent significantly outperforms traditional single-model approaches in accuracy, guidance, and interactivity.

4.5 Fine-Grained Analysis on Five Evaluation Dimensions

Figure 7 reveals distinct performance profiles for eight high-performing evaluated models. In Context Visualization and Diagram Design, most baselines, including SDXL, Claude 3.7, and `v0`, exhibit moderate to low scores, often struggling with providing rich situational cues or pedagogically sound visual structures, especially for complex problems. `v0` and Claude show relatively better capabilities in Text-Graphic Integration and Thought Guidance compared to other FMs, which generally offer minimal support in these areas. However, all baseline models, including `v0`, are significantly limited in the Interactivity dimension, primarily due to their output format (static images/SVG or less dynamic webpages). In contrast, our EduVisAgent demonstrates consistently strong performance across all five dimensions. It particularly excels in creating rich context visualizations, well-structured diagram designs, and ensuring seamless text-graphic integration. Furthermore, EduVisAgent provides superior thought guidance and achieves notably high scores in Interactivity, areas where baseline models significantly lag. This comprehensive strength highlights EduVisAgent's advanced ability to generate not just visualizations, but truly effective and interactive pedagogical tools.

5 Related Work

LLM for Pedagogical Assistance. Foundation models (FMs), including diffusion models and large vision-language models (LVLMs), have been increasingly adopted in educational contexts (Chu et al., 2025; Wang et al., 2024) to support teaching and classroom interactions. EduAgent (Xu et al., 2024) and Teachtune (Jin et al., 2025) enhance the problem-solving process through automated simulations of student-teacher dialogues, collaborative learning, and task-oriented reasoning. Agents such as SEFL (Zhang et al., 2025) and PROF (Nair et al., 2024) synthesize immediate, on-demand feedback to support large-scale instructional scenarios. Furthermore, domain-specific agents such as MathChat (Wu et al., 2023), NEWTON (Wang et al., 2023b), and MEDCO (Wei et al., 2024) further provide textual explanations tailored to scientific and medical education. While these systems address diverse pedagogical needs, their focus remains largely on text-based interactions (Wu et al., 2023; Xu et al., 2024; Cui et al., 2024), overlooking the critical role of visualization in fostering conceptual understanding and improving learning outcomes (Presmeg, 2006). While valuable, these text-centric systems do not address the large body of educational research highlighting the unique cognitive benefits of visual learning. Despite its pedagogical importance, the capacity of FMs and agents to generate logical, explanatory visual illustrations remains underexplored. EduVisBench is the first comprehensive benchmark designed to systematically evaluate FMs ability to produce pedagogically effective, step-by-step visual reasoning, covering 15 diverse visually grounded educational scenarios with multi-level problem sets and multimodal-centric solutions.

LLM for Scientific Visualization. While some existing works have preliminarily explored the potential of FMs in supporting visual scaffolding (Podo et al., 2024; Chen et al., 2024c; Pandey & Ottley, 2025; Hong et al., 2025), they are typically fragmented, lack pedagogical grounding, and fail to generalize across diverse educational tasks (Wang et al., 2023a; Ku et al., 2025). For instance, Visual Sketchpad (Hu et al., 2024) attempts to illustrate problem-solving processes with sketches generated from code. However, these visuals are often low in quality, lack logical coherence, and fall short in explanatory depth (Wang et al., 2025). Other approaches like MatplotAgent (Yang et al., 2024), PlotGen (Goswami et al., 2025), and OmniSVG (Yang et al., 2025) leverage plotting and SVG tools to produce more accurate, data-grounded visualizations. Still, these methods are limited in scope, often addressing only isolated steps rather than providing systematic, end-to-end visual explanations of multi-step problem-solving tasks (Vázquez, 2024; Chen et al., 2024a; 2025b). Our framework, in contrast, is designed to manage the entire pedagogical workflow, from problem deconstruction to the final interactive explanation. To overcome these limitations, we propose a multi-agent collaborative framework, EduVisAgent, that simulates the full learning journeyfrom initial problem exposure to deep conceptual understandingby coordinating specialized agents to generate coherent, pedagogically aligned visualizations throughout the reasoning process.

LLM-based Education Agents. Recent advancements in LLM-based agents have led to the development of specialized architectures capable of long-horizon planning, tool use, and memory management across a range of real-world domains (Yao et al., 2023; Chan et al., 2024; Chen et al., 2024b; 2025a; Nie et al., 2025; Han et al., 2025; Zhou et al., 2025). In the educational domain, AI agents such as EduAgent (Xu et al., 2024) and Teachtune (Jin et al., 2025) simulate student-teacher dialogues, collaborative learning activities, and task-oriented reasoning to enhance problem-solving instruction. Agents like SEFL (Zhang et al., 2025) and PROF (Nair et al., 2024) generate on-demand feedback for large-scale educational settings, while domain-specific tools such as MathChat (Wu et al., 2023), NEWTON (Wang et al., 2023b), and MEDCO (Wei et al., 2024) provide textual explanations for scientific and medical learning. Despite these advances, limited research has investigated collaborative, multi-agent approaches tailored to educational reasoning and visualization. EduVisAgent is the first systematic multi-agent framework that coordinates specialized agents and provides a comprehensive approach to supporting step-by-step pedagogical problem-solving.

6 Conclusion

This paper addressed the challenge of generating pedagogically meaningful visual explanations with AI systems. We introduced EduVisBench, a benchmark revealing that existing models often produce inadequate visual outputs. This work provides a quantitative baseline for the field, clearly identifying the key areas where current technologies fall short. To overcome this, we proposed EduVisAgent, a collaborative multi-agent framework. Experiments show EduVisAgent significantly outperforms all baselines, demonstrating the potential of agent-based systems for advancing educational visualization.

Ethics Statement

The primary goal of this research is to advance educational technology by improving the pedagogical quality of AI-generated visualizations, aiming for a positive societal impact. The benchmark developed, EduVisBench, is curated from publicly available and high-quality educational resources, including C-MHChem-Benchmark, high-school-physics, Illustrative Mathematics, and MATH-500. To validate our automated evaluation metric, we conducted a comparative study involving human evaluators, who were undergraduate students from top universities. All data used in the study was handled with care to ensure anonymity and was used solely for the purpose of validating the scoring system. The models and methods proposed are intended for beneficial educational applications. The authors are not aware of any other ethical issues and declare no competing interests or conflicts of interest associated with this research.

Reproducibility Statement

To ensure the reproducibility of our work, we have provided detailed descriptions of our methodology and resources. The curation process for our benchmark, EduVisBench, is detailed in Section 2.2, with data sources explicitly cited. Our comprehensive evaluation framework, including the five key dimensions and scoring protocol, is described in Section 2.3 and Section 4.1.The detailed scoring rubrics and the exact prompt used for our GPT-4o-based evaluation are available in Appendix A.2 and Appendix A.3, respectively. The architecture of our proposed EduVisAgent and the roles of each specialized agent are thoroughly explained in Section 3. A complete list of all baseline models and their versions used in our experiments is provided in Section 4.1.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Singh Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio C'esar Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allison Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Young Jin Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xianmin Song, Olatunji Ruwase, Praneetha Vaddamanu, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Andre Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Cheng-Yuan Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, Xiren Zhou, and Yifan Yang. Phi-3 technical report: A highly capable language model locally on your phone. ArXiv, abs/2404.14219, 2024. URL https://api.semanticscholar.org/CorpusID:269293048.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. arXiv preprint arXiv:2503.01743, 2025.

Anthropic. Claude 3.7 sonnet and claude code. https://www.anthropic.com/news/claude-3-7-sonnet, February 2025. Accessed: 2025-05-16.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. Mle-bench: Evaluating machine learning agents on machine learning engineering. arXiv preprint arXiv:2410.07095, 2024.

Jiawen Chen, Jianghao Zhang, Huaxiu Yao, and Yun Li. Celltypeagent: Trustworthy cell type annotation with large language models. arXiv preprint arXiv:2505.08844, 2025a.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? arXiv preprint arXiv:2407.04842, 2024a.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. Advances in Neural Information Processing Systems, 37:130185–130213, 2024b.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. arXiv preprint arXiv:2403.00425, 2024c.

Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. Autoprm: Automating procedural supervision for multi-step reasoning via controllable question decomposition. arXiv preprint arXiv:2402.11452, 2024d.

Zhaorun Chen, Mintong Kang, and Bo Li. Shieldagent: Shielding agents via verifiable safety policy reasoning. arXiv preprint arXiv:2503.22738, 2025b.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, et al. Llm agents for education: Advances and applications. arXiv preprint arXiv:2503.11733, 2025.

Chenhang Cui, An Zhang, Yiyang Zhou, Zhaorun Chen, Gelei Deng, Huaxiu Yao, and Tat-Seng Chua. Fine-grained verifiers: Preference modeling as next-token prediction in vision-language alignment. arXiv preprint arXiv:2410.14148, 2024.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

Kanika Goswami, Puneet Mathur, Ryan Rossi, and Franck Dernoncourt. Plotgen: Multi-agent llm-based scientific data visualization via multimodal feedback. arXiv preprint arXiv:2502.00988, 2025.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. Mdocagent: A multi-modal multi-agent framework for document understanding. arXiv preprint arXiv:2503.13964, 2025.

Jiayi Hong, Christian Seto, Arlen Fan, and Ross Maciejewski. Do llms have visualization literacy? an evaluation on modified visualizations to test generalization in data interpretation. IEEE Transactions on Visualization and Computer Graphics, 2025.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. arXiv preprint arXiv:2406.09403, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024.

IT Admin. Introducing Stable Diffusion 3.5. https://stability.ai/news/introducing-stable-diffusion-3-5, October 2024. Updated October 29, 2024; Accessed: 2025-05-20.

Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, pp. 1–28, 2025.

Max Ku, Thomas Chong, Jonathan Leung, Krish Shah, Alvin Yu, and Wenhu Chen. Theoremexplainagent: Towards multimodal explanations for llm theorem understanding. arXiv preprint arXiv:2502.19400, 2025.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. arXiv preprint arXiv:2305.20050, 2023.

Shrestha Basu Mallick and Logan Kilpatrick. Gemini 2.0: Flash, flash-lite and pro. https://developers.googleblog.com/en/gemini-2-family-expands/, February 2025. Accessed: 2025-05-20.

Chad M. Miller and Jonathan D. Cohen. Metacognitive prompts in multimedia learning: A meta-analysis. Educational Psychology Review, 32(3):979–1003, 2020. doi: 10.1007/s10648-020-09525-3. URL https://files.eric.ed.gov/fulltext/EJ1262620.pdf.

Mistral AI. Mistral small 3.1: Sota. multimodal. multilingual. apache 2.0. https://mistral.ai/news/mistral-small-3-1, March 2025. Accessed: 2025-05-20.

Inderjeet Nair, Jiaye Tan, Xiaotian Su, Anne Gere, Xu Wang, and Lu Wang. Closing the loop: Learning to generate writing feedback via language model simulated student revisions. arXiv preprint arXiv:2410.08058, 2024.

Fan Nie, Lan Feng, Haotian Ye, Weixin Liang, Pan Lu, Huaxiu Yao, Alexandre Alahi, and James Zou. Weak-for-strong: Training weak meta-agent to harness strong executors. arXiv preprint arXiv:2504.04785, 2025.

Satria Nugroho and Jailani Jailani. The effectiveness of concrete representational abstract approach (cra) approach and problem solving approach on mathematical representation ability at elementary school. KnE Social Sciences, 06 2019. doi: 10.18502/kss.v3i17.4620.

Saugat Pandey and Alvitta Ottley. Benchmarking visual language models on standardized visualization literacy tests. arXiv preprint arXiv:2503.16632, 2025.

Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. ArXiv, abs/2307.01952, 2023. URL https://api.semanticscholar.org/CorpusID:259341735.

Luca Podo, Muhammad Ishmal, and Marco Angelini. Vi (e) va llm! a conceptual stack for evaluating and interpreting generative ai-based visualizations. arXiv preprint arXiv:2402.02167, 2024.

Norma Presmeg. Research on visualization in learning and teaching mathematics. Handbook of research on the psychology of mathematics education, pp. 205–235, 2006.

M. Rohith. High school physics. https://huggingface.co/datasets/mrohith29/high-school-physics, 2023. Accessed: 2025-05-15.

Gregory Schraw and David Moshman. Metacognitive theories. Educational Psychology Review, 7(4): 351–371, 1995. ISSN 1040726X, 1573336X. URL http://www.jstor.org/stable/23359367.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Pere-Pau Vázquez. Are llms ready for visualization? In 2024 IEEE 17th Pacific Visualization Conference (PacificVis), pp. 343–352. IEEE, 2024.

Vercel. v0: Ai chat interface for web automation. https://v0.dev, 2025. Accessed: 2025-05-20.

Lei Wang, Songheng Zhang, Yun Wang, Ee-Peng Lim, and Yong Wang. Llm4vis: Explainable visualization recommendation using chatgpt. arXiv preprint arXiv:2310.07652, 2023a.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. Large language models for education: A survey and outlook. arXiv preprint arXiv:2403.18105, 2024.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. arXiv preprint arXiv:2503.12605, 2025.

Yi Ru Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. Newton: Are large language models capable of physical reasoning? arXiv preprint arXiv:2310.07018, 2023b.

Hao Wei, Jianing Qiu, Haibao Yu, and Wu Yuan. Medco: Medical education copilots based on a multi-agent framework. arXiv preprint arXiv:2408.12496, 2024.

Yiran Wu, Feiran Jia, Shaokun Zhang, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Qingyun Wu, and Chi Wang. Mathchat: Converse to tackle challenging math problems with llm agents. arXiv preprint arXiv:2306.01337, 2023.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

Songlin Xu, Xinyu Zhang, and Lianhui Qin. Eduagent: Generative student agents in learning. arXiv preprint arXiv:2404.07963, 2024.

Yiying Yang, Wei Cheng, Sijin Chen, Xianfang Zeng, Jiaxu Zhang, Liao Wang, Gang Yu, Xingjun Ma, and Yu-Gang Jiang. Omnisvg: A unified scalable vector graphics generation model. arXiv preprint arXiv:2504.06263, 2025.

Zhiyu Yang, Zihan Zhou, Shuo Wang, Xin Cong, Xu Han, Yukun Yan, Zhenghao Liu, Zhixing Tan, Pengyuan Liu, Dong Yu, et al. Matplotagent: Method and evaluation for llm-based agentic scientific data visualization. arXiv preprint arXiv:2402.11453, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In International Conference on Learning Representations (ICLR), 2023.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800, 2024.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, Shufei Zhang, Mao Su, Hansen Zhong, Yuqiang Li, and Wanli Ouyang. Chemllm: A chemical large language model, 2024.

Mike Zhang, Amalie Pernille Dilling, Léon Gondelman, Niels Erik Ruan Lyngdorf, Euan D Lindsay, and Johannes Bjerva. Sefl: Harnessing large language model agents to improve educational feedback systems. arXiv preprint arXiv:2502.12927, 2025.

Yiyang Zhou, Zhaoyang Wang, Tianle Wang, Shangyu Xing, Peng Xia, Bo Li, Kaiyuan Zheng, Zijian Zhang, Zhaorun Chen, Wenhao Zheng, et al. Anyprefer: An agentic framework for preference data synthesis. arXiv preprint arXiv:2504.19276, 2025.

A APPENDIX

A.1 VISUALIZATION DISCIPLINES

Table 4 illustrates the disciplines and types in our EduVisBench.

Discipline	Common Visualization Types
Mathematics	Number lines, function graphs, and other formalized visual tools.
Physics	Diagrams involving levers, rigid body motion, forces, and fields.
Chemistry	Molecular structures and schematic representations of standard laboratory apparatus.

Table 4: Representative Visualization Types Across Academic Disciplines

A.2 EVALUATION METRIC

Visual Scenario Design Guidance The category of "Visual Scenario Design Guidance" outlines different levels of visualizing mathematical concepts, progressing from basic text-only representations to highly integrated visual-text formats. Through five defined levels, the framework demonstrates how visual elements can enhance students' understanding and engagement with abstract ideas, guiding instructional designers to gradually enrich scenarios, add annotations, and strengthen contextual connectionsultimately achieving the goal of visually presenting the full flow and conceptual structure of the content.The five levels of Visual Scenario Design Guidance are as follows:

Level	Description
Level 1	The image contains no scenes or illustrations, presenting only text and formulas. It lacks contextual visual cues, failing to spark interest or connect the concepts to real-life situations.
Level 2	The image includes a single static illustration or low-fidelity mockup with minimal labeling that does not highlight variables or key objects, offering limited context and poor immersion.
Level 3	Multiple static schematic diagrams or sketch-style illustrations appear in the image, labeling core objects, variables, and simple steps, providing basic visual guidance but lacking layered coherence.
Level 4	The image integrates scenario illustrations, storyboard panels, and info-graphics to present the process in multiple views and steps, with annotations and captions guiding students through mapping abstract concepts to context.
Level 5	Storyboard-style illustrations and infographics are fused into a single image, including overview, detailed close-ups, and key pathway diagrams with comprehensive annotations, allowing students to grasp the entire flow and conceptual network at a glance.

Table 5: Five Levels of Visual Scenario Design Guidance

Visual Illustration Design The category of "Visual Illustration Design" describes progressive levels of visual elements used to support students systematic understanding of quantities and relationships. It ranges from no visual aids to complex integrated dashboards that deeply connect data and model structures. Through five levels, the framework guides designers to improve clarity, coherence, and contextual richness of visual illustrations, enhancing students analytic and comparative abilities.

Level	Description
Level 1	The image contains no charts, axes, or flow diagramsonly text. Without embedded visual tools, students cannot systematically organize or analyze quantities and relationships.
Level 2	The image presents a static number line and colored bar chart with complete scales and a legend, helping students gain a basic understanding of numerical changes. However, it lacks comparison and contextual layering.
Level 3	The image presents a static number line and colored bar chart with complete scales and legends, helping students grasp basic numerical changes visually, though comparison and context layering are absent.
Level 4	The image combines number lines, flowcharts, infographics, and arrow annotations; multiple visuals are juxtaposed or overlaid to show processes and variable changes for a coherent modeling view.
Level 5	The image presents a dashboard-style visualization integrating axes, bar charts, flow diagrams, heatmaps, etc., with linked elements that deeply visualize data relationships and model structure.

Table 6: Five Levels of Visual Illustration Design

TextIllustration Coordination The category of "TextIllustration Coordination" describes levels of alignment and integration between textual content and visual elements within images. This progression ranges from complete disconnection to seamless fusion, enabling students to effectively map and synthesize text, formulas, and graphics. The framework guides designers in strengthening links between verbal and visual information to enhance comprehension and structural understanding.

Level	Description
Level 1	Text and illustrations in the image are completely disconnected, with no labels, legends, or connectorsstudents cannot use visuals to understand text or formulas.
Level 2	Text occasionally prompts see diagram or refer to the illustration, but the image lacks legends or clear labels, so mapping between text and graphics remains ambiguous.
Level 3	Text descriptions and image elements share consistent numbering, color blocks, or arrows linked to a simple legend, explaining core symbols and variables to support initial mapping.
Level 4	Text paragraphs are laid out alongside corresponding visuals within the same image, with detailed legends and color-coded annotations enabling simultaneous reading and mapping.
Level 5	Text, formulas, and legends are fully integrated in one image, using consistent colors, numbering, and layered layout to achieve seamless textgraphic fusion for complete structural understanding.

Table 7: Five Levels of TextIllustration Coordination

Learning Thought Guidance The category of "Learning Thought Guidance" describes the progressive inclusion of visualized problem-solving strategies and reflective cues in images. From presenting only problem statements to complex integrated dashboards, this framework guides designers to scaffold students strategic thinking and metacognitive reflection through visual tools, enabling deeper reasoning and transfer of learning.

Level	Description
Level 1	The image offers no visualized problem-solving guidance, showing only the problem statement and formulas, leaving students without strategic cues or reflection prompts.
Level 2	The image embeds a simple flowchart or two title-style hints (e.g., Identify problem type, Check result), but the flowchart is overly simplistic and hints lack hierarchical detail.
Level 3	The image displays a step-by-step flowchart template with key thinking nodes and self-check checkpoints, leaving annotation space for students to visually record their reasoning.
Level 4	The image combines a near-transfer exercise with a comparative thought diagram, visually highlighting strategy differences so students can apply existing reasoning to a new context.
Level 5	The image fuses near- and far-transfer exercises, concept mind maps, and a reflection panel into a dashboard-style layout, allowing students to review and extend their problem-solving network visually.

Table 8: Five Levels of Learning Thought Guidance

Interactivity and Personalized Support The category of "Interactivity and Personalized Support" outlines levels of incorporating feedback, hints, and tailored assistance into images, evolving from static presentations to dynamic, student-responsive visual supports. This framework encourages designers to embed interactive elements that adapt to learner needs, promoting engagement and personalized problem-solving.

Level	Description
Level 1	The image includes no feedback or support componentsonly a static problem statement and answer fieldoffering no hints, examples, or error cues and resulting in a nonresponsive visual.
Level 2	The image shows fixed hint boxes (e.g., Hint: draw a number line, Hint: check rounding), but hints are not tailored to student responses, limiting personalized guidance.
Level 3	The image integrates multiple static correction tips and example solution modules (common mistakes and standard approaches), which students can reference visually but without intelligent recommendations.
Level 4	The image presents example solution workflows, text hints, and a common-errors analysis section highlighted with color blocks and arrows, providing diverse visual support in a single layout.
Level 5	The image displays a comprehensive visual support panel with difficulty suggestions, personalized hints, worked examples, and extension resource links, enabling students to select tailored guidance directly from the visual layout.

Table 9: Five Levels of Interactivity and Personalized Support

A.3 Evaluation Prompt

The instructional web page evaluation prompt is structured as follows:

Evaluation Prompt

As a **professional evaluator of instructional web pages**, your task is to determine whether the generated web page meets expectations across five specific categories.

Instructions:

- Assign an **integer score from 0 to 5** for each of the five categories (15).
- **0** = completely missing or extremely poor
  5 = fully meets the highest standard
- Evaluation should be based solely on the specified aspect: *{category}*.
  The definition of *{category}* is: *{description}*.
- **Do not include any explanation, justification, or additional commentary. Refusing to provide a score is not allowed.**

Evaluation Output Format

```
{{RATING: {"1":score, "2":score, "3":score, "4":score, "5":score}}
```

A.4 Generation Prompts for Baseline Models

The following prompts are used to instruct baseline models to generate instructional visualizations for multiple-choice questions.

HTML Webpage Generation Prompt

Please generate a fully structured and styled HTML webpage for the following {subject} multiple-choice question, with a focus on clearly presenting the problem, visualizing key scientific concepts, and explaining the reasoning behind the correct choice.
The page should include:

- A prominently displayed question section.
- A clearly formatted list of answer choices (A, B, C, D).
- A step-by-step explanation section that helps users understand why the correct answer is right, and why the others are not. This section may include:
  - Diagrams or illustrations (e.g., molecules, environmental impact visuals),
  - Charts or data comparisons (e.g., particulate levels),
  - Flowcharts or labeled process diagrams,
  - Any other visual representation that supports comprehension.
- A clear highlight of the final correct answer (e.g., a visual cue or box).

Requirements:

- Use modern CSS styling (inline `<style>` block or external).
- Ensure layout is responsive and readable across devices.
- Use semantic HTML structure with headings and subheadings (<h1>, <h2>, etc.).
- Render scientific symbols or formulas correctly (e.g., MathJax or KaTeX).

- Visualization libraries such as Chart.js, D3.js, or SVG may be used to enhance explanations.

Question: {question}
Please output the full HTML + CSS + JavaScript code only, without any extra explanation or comments.

SVG Diagram Generation Prompt

Generate a **stand-alone SVG diagram** that visually explains the following question as a reasoning task.
The SVG must include:

- The full question text;
- Key reasoning steps, visual annotations, or illustrations that aid in understanding or solving the problem;
- If applicable, highlight the final answer or conclusion clearly.

Requirements:

- Output only a single `<svg>...</svg>` block, with no extra text outside it;
- SVG width between 800–1200px, layout should be clean and adaptive;
- Use `<text>` elements with readable font sizes for all text;
- Use arrows, symbols, and diagrams if they help communicate the solution process;
- You may omit multiple-choice options if not relevant.

Question: {question}

Visualization Agent Generation Prompt

Your task is to generate a complete webpage solution for the following problem. The page should include:

- Introduce a scenario to engage students.
- Explain the problem background and provide a step-by-step walkthrough of the solution.
- Give another similar problem to assess students' understanding.

Question: {question}

A.5 Multi-Agent System Prompts

A.5.1 Task Planning Agent Prompt

System Prompt

Transform learner's question into structured instructional task.

1. **Scenario & Understanding**: Real-world context + givens/goals/constraints
2. **Solution Strategy**: Analysis + step-by-step solution + reasoning
3. **Transfer Tasks**:

  - Near-transfer: same structure, minor changes
  - Far-transfer: different surface, same logic

4. **FOPS Structure** (per step):
   - **F**: Problem type
   - **O**: Diagram/equation structure
   - **P**: Solution path
   - **S**: Execute + verify
   - Specify: goal, action, concepts, pitfalls

5. **UI Layout**: Shadcn/UI structure + sections + meta-prompt placement + math rendering

JSON Output Format

```
{
  "scenario": {
    "context": "Real-world scenario description",
    "givens": ["given1", "given2"],
    "goals": ["goal1"],
    "constraints": ["constraint1"]
  },
  "solution_strategy": {
    "analysis": "Problem analysis",
    "steps": [
      {
        "step_number": 1,
        "fops_label": "F/O/P/S",
        "description": "What to do",
        "reasoning": "Why this step"
      }
    ]
  },
  "transfer_tasks": {
    "near_transfer": {
      "problem": "Near-transfer problem statement",
      "explanation": "Why it's near-transfer"
    },
    "far_transfer": {
      "problem": "Far-transfer problem statement",
      "explanation": "Why it's far-transfer"
    }
  },
  "instructional_steps": [
    {
      "step_number": 1,
      "goal": "What learner should achieve",
      "action": "Expected reasoning/action",
      "concepts": ["concept1", "concept2"],
      "pitfalls": ["potential misconception1"]
    }
  ],
  "ui_layout_suggestion": {
    "structure": "Overall page flow",
    "key_components": ["Card", "Accordion", "Alert", "Table"],
    "content_organization": "Organize scenario/solution/practice",
    "math_rendering": "KaTeX/MathJax"
  }
}
```

User Prompt

Give your analysis to the following question in JSON format.
{Question}

A.5.2 CONCEPTUAL MAPPING AGENT PROMPT

System Prompt

Map concepts using **CRA framework** (ConcreteRepresentationalAbstract).

1. **Concrete**: Objects, quantities, situations (directly experienced)
2. **Representational + Visual Design**:
   - Tools: number lines, bar graphs, diagrams, organizers, sketches, tables
   - Visual design: form, purpose, concrete→abstract bridge
   - Avoid decorative; focus pedagogical
3. **Abstract**: Formulas, principles, theorems + connection to concrete
4. **Think Aloud**: Verbalization prompts

JSON Output Format

```
{
  "cra_mapping": [
    {
      "step_number": 1,
      "step_name": "Step name",
      "concrete": {
        "entities": ["object1", "quantity1"],
        "description": "What can be directly experienced"
      },
      "representational": [
        {
          "tool": "Number line/Bar graph/etc",
          "purpose": "Why it helps",
          "represents": "What it shows",
          "visual_design": {
            "form": "Specific visual form to use",
            "elements": ["key visual elements"],
            "bridge": "How it bridges concrete to abstract"
          }
        }
      ],
      "abstract": {
        "constructs": ["formula1", "principle1"],
        "connection": "How abstract connects to concrete"
      },
      "think_aloud": ["prompt1", "prompt2"]
    }
  ]
}
```

User Prompt

According to the following planning, give your analysis in JSON format.
{Task Planning JSON}

A.5.3 Reasoning Decomposition Agent Prompt

System Prompt

Decompose reasoning using design **scaffolded practice**.

1. **Step Guidance**:
   - Action (what learner does)
   - Link to CRA concepts
   - **Visual support**: Which CRA visual + interaction (draggable, reveals, etc.)
2. **Practice Activities**:
   - Types: fill-in, choice, judgment, explanation
   - Gradual release strategy
3. **Math**: LaTeX in $...$ or $$...$$

JSON Output Format


```
{
  "fops_reasoning": [
    {
      "action": "What learner does",
      "concepts": ["linked concept1", "linked concept2"],
      "scaffolding_needed": true,
      "scaffolding_notes": "Why and how to scaffold",
      "visual_support": {
        "needed": true,
        "which_visual": "Reference to CRA mapping visual",
        "interaction": "Suggested interactive element if any"
      }
    }
  ],
  "practice_activities": [
    {
      "activity_number": 1,
      "name": "Activity name",
      "type": "fill-in/choice/judgment/explanation",
      "task": "Problem or task content",
      "gradual_release": "How independence builds"
    }
  ]
}
```


User Prompt

According to the following planning and conceptual mapping, give your analysis in JSON format.
{Task Planning JSON}
{Conceptual Mapping JSON}

A.5.4 Metacognitive Reviewer Prompt

System Prompt

Generate metacognitive prompts for **monitor, evaluate, regulate**.

1. **Reflective Prompts**: Check understanding, comprehension, strategy
2. **Self-Questioning**:
   - "What did I do?"
   - "Why does this work?"
   - "What did I miss?"
   - "How does this connect?"
   - "What if...?"
3. **Self-Correction**: Checkpoints + error identification
4. **Organize by Phase**:
   - **Before**: Understanding + planning
   - **During**: Monitoring + checking
   - **After**: Evaluating + reflecting
5. **Math**: LaTeX in $...$ or $$...$$

JSON Output Format

```
{
  "metacognitive_prompts": {
    "before_solving": [
      "Understanding check prompt 1",
      "Planning prompt 1"
    ],
    "during_solving": [
      "Step monitoring prompt 1",
      "Strategy adjustment prompt 1"
    ],
    "after_solving": [
      "Verification prompt 1",
      "Reflection prompt 1",
      "Improvement prompt 1"
    ]
  },
  "general_strategies": [
    {
      "strategy": "Self-monitoring technique",
      "when_to_use": "During which phase",
      "how_to_apply": "Specific actions"
    }
  ]
}
```

User Prompt

According to the following planning, conceptual mapping and reasoning decomposition, give your analysis in JSON format.
{Task Planning JSON}
{Conceptual Mapping JSON}
{Reasoning Decomposition JSON}

A.5.5 Visualization Agent Prompt

System Prompt

As the **Visual Representation Specialist**, generate the visualization teaching webpage of the given question.

1. **Abstract Visualization Tools**: Encourage number lines, bar graphs, object diagrams, organizers, sketches, and data tables.
2. **Avoid Decorative Imagery**: No photorealistic or cartoon-style images; visuals must remain schematic and pedagogical.
3. **Conceptual Mapping**: Explicitly explain how each visual corresponds to the underlying abstract concept (e.g., number line for quantity change).
4. **UI Integration**: Specify how visuals should appear in the web UI (placement, sequencing, interaction).
5. **Rendering**: math rendered with KaTeX/MathJax.

User Prompt

Using all of the given data, generate a visualized teaching webpage for the following question.
{Question}
{Task Planning JSON}
{Conceptual Mapping JSON}
{Reasoning Decomposition JSON}
{Metacognitive Reviewer JSON}

B Additional Results

B.1 Ablation Studies on EduVisAgent

We conduct ablation studies on the multi-agent teaching system in Table 10.

We compare the overall Total score of the complete system (**Full**) against six variants: **-TP** (without Task Planning), **-CM** (without Conceptual Mapping), **-RD** (without Reasoning Decomposition), **-MR** (without Metacognitive Review), **-VIS** (Using `Claude 3.7 Sonnet` as the underlying model while keeping the same multi-agent pipeline), and **Single** (a single-agent baseline that uses the same set of prompts without explicit modularization or inter-agent coordination).

Impact of Removing Individual Modules. The ablation results reveal clear and systematic evidence that each module in the multi-agent architecture contributes a distinct cognitive function whose removal leads to measurable degradation. Across all subjects and difficulty levels, dropping any single module results in a consistent decline of approximately 4–6 points relative to the full model, indicating that no component is redundant.

Task Planning (-TP). Removing the Task Planning module reduces the system's ability to structure the solution trajectory at the outset. This manifests in weaker performance particularly on Medium and Hard problems, where multi-step planning is essential. Without this stage, the problem-solving process becomes more linear and less globally coherent, leading to incremental reasoning errors that accumulate throughout the solution.

Conceptual Mapping (-CM). The Conceptual Mapping module provides the system with an intermediate representational scaffold—a way to organize domain concepts, formulas, and symbolic relations before detailed reasoning begins. Its removal produces one of the largest degradations across domains, showing that the system heavily relies on this symbolic schema construction phase. Without CM, the

Table 10: Ablation study on EduVisAgent. **Full** denotes our complete multi-agent system. **Single** denotes a single-agent baseline using all prompts without modularization. The highest scores are in **bold**.

Subject	Difficulty	Full	-TP	-CM	-RD	-MR	-VIS	Single
Chemistry	Easy	**69.00**	64.60	63.80	63.00	64.20	64.80	59.30
	Medium	**76.27**	71.87	71.07	70.27	71.47	72.07	67.42
	Hard	**76.00**	71.60	70.80	70.00	71.20	71.80	66.85
Physics	Easy	**85.33**	80.94	80.14	79.34	80.54	81.14	76.10
	Medium	**81.71**	77.31	76.51	75.71	76.91	77.51	72.05
	Hard	**84.00**	79.60	78.80	78.00	79.20	79.80	74.32
Maths	Easy	**90.20**	85.80	85.00	84.20	85.40	86.00	81.05
	Medium	**64.50**	60.10	59.30	58.50	59.70	60.30	55.27
	Hard	**65.00**	60.60	59.80	59.00	60.20	60.80	53.88

reasoning tends to jump directly into procedural steps without establishing the underlying conceptual structure, which leads to misapplied rules, inconsistent variable usage, and missing constraints.

Reasoning Decomposition (-RD). This module creates explicit fine-grained segmentation of logical steps, and its absence consistently yields the largest drop among all ablations. Without RD, the model tends to condense multi-step reasoning into single large leaps, increasing the likelihood of hidden errors that are never surfaced or corrected. RD therefore serves as the backbone of reliable multi-hop reasoning, ensuring that intermediate steps are interpretable, verifiable, and less prone to compounding mistakes.

Metacognitive Review (-MR). The MR module acts as an internal critic, performing self-evaluation and error checking. Removing MR reduces the models ability to detect calculation inconsistencies, missing assumptions, or logical conflicts within its own output. While the primary reasoning path remains intact, the absence of this reflective layer leads to small but systematic accuracy losses, especially on tasks requiring unit checking, boundary conditions, or verification of derived results.

Model Replacement (-VIS). Replacing the underlying model with `Claude 3.7 Sonnet` while preserving the multi-agent structure produces a moderate drop across categories. The decline is smaller than removing core cognitive modules, demonstrating that the multi-agent pipeline itself contributes substantial performance stability. However, it also indicates that the pipeline and the underlying model are jointly responsible for peak performance.

Single-Agent Baseline. The **Single** baseline, which collapses all prompts into a single monolithic agent, performs the worst across all conditions. Its scores are consistently below every multi-agent ablation, averaging 8–15 points lower than the full system. This sharp decline highlights that the benefits of the multi-agent design arise not merely from the prompts themselves, but from the explicit architectural separation of planning, conceptual structuring, stepwise reasoning, and self-review.

B.2 Evaluation on non-STEM datasets

Table 11: Performance of Visualization Agents on EduVisBench.

Method	Vis. Type	Prehistory	Sociology	Avg
`v0`	Webpage	61.8	68.0	64.9
EduVisAgent	Webpage	79.4	84.8	82.1

To further explore EduVisAgent's performance in non-STEM disciplines, we selected 100 problems from MMLU's prehistory and sociology categories to evaluate EduVisAgent's adaptability to narrative and conceptual reasoning tasks. Results in Table 11 show that EduVisAgent significantly outperforms the current SOTA baseline (v0) by 17.2 points, demonstrating that our multi-agent framework effectively generalizes to humanities education.

B.3 Evaluation robustness

We tested two other evaluators, Gemini 2.5 pro and Claude 3.7 Sonnet to score visualization results of SDXL, GPT-4o(SVG) and EduVisAgent. The results in Table 12 show consistent scoring patterns across different evaluator models. While there are minor numerical variations in the absolute scores, the overall trends and relative rankings remain highly consistent with the GPT-4o results.

Table 12: Evaluated by GPT-4o, Gemini 2.5 Pro and Claude 3.7 Sonnet on EduVisBench.

Method	Vis. Type	GPT-4o	Gemini 2.5 Pro	Claude 3.7 Sonnet
SDXL	Image	21.8	18.9	24.7
GPT-4o	SVG	25.4	26.3	27.4
EduVisAgent	Webpage	78.9	81.6	84.4

C Additional Mathematics Examples in EduVisBench

Figure 8, 9, 10 and 11 presents four representative problems from the Hard subset (MATH-500) in EduVisBench, illustrating the range and difficulty of the mathematical reasoning challenges in the benchmark.

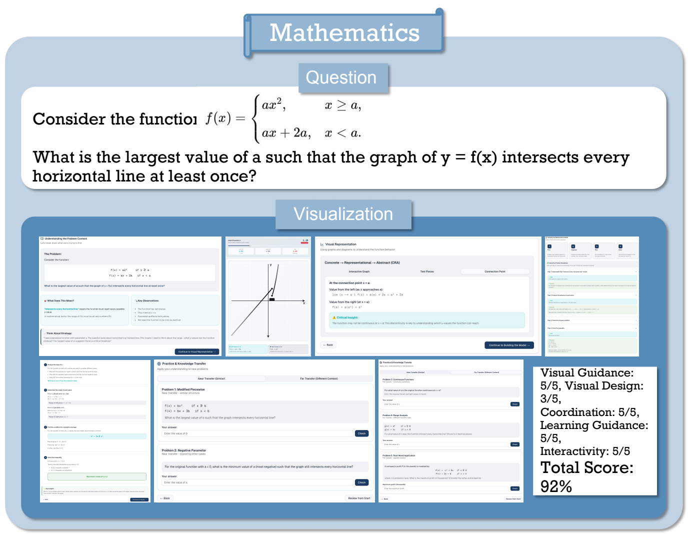


Figure 8: Representative example 1 from the Hard subset (MATH-500) of EduVisBench.

D Real-world deployment optimization

For real-world deployment, we identify several optimization directions: (1) reducing token consumption through more efficient prompt engineering, (2) optimizing the metacognitive reasoning depth to balance quality and efficiency, and (3) implementing caching mechanisms for frequently used educational concepts. Additionally, the system could operate asynchronously. Teachers prepare materials in advance rather than real-time generation, making latency less critical. Integration with existing LMS platforms would involve API-based deployment where visualization generation occurs server-side.

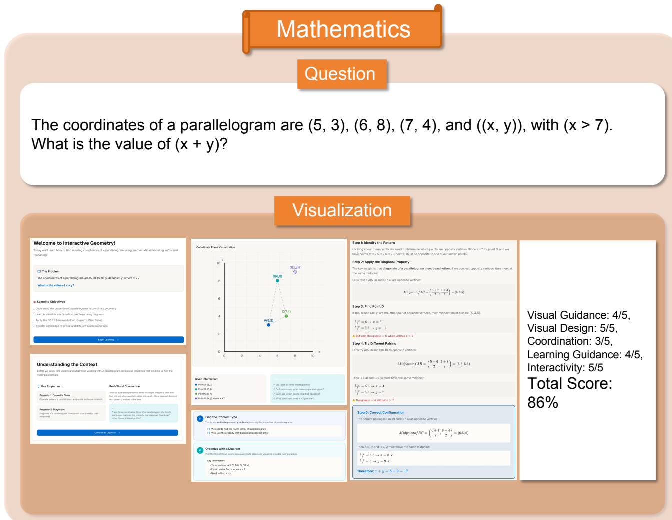


Figure 9: Representative example 2 from the Hard subset (MATH-500) of EduVisBench.

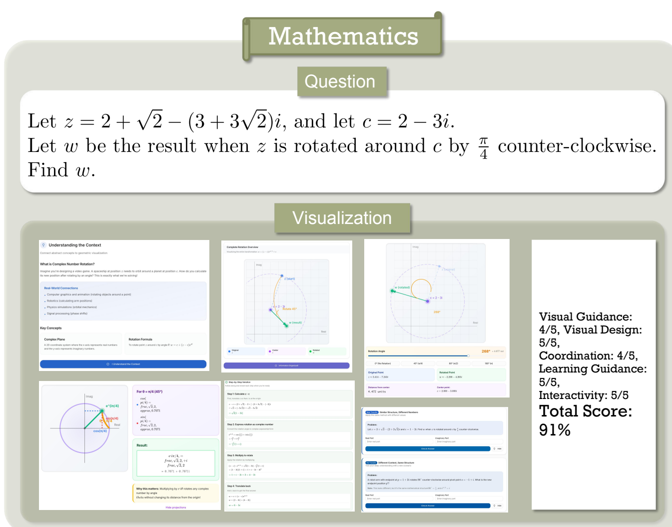


Figure 10: Representative example 3 from the Hard subset (MATH-500) of EduVisBench.

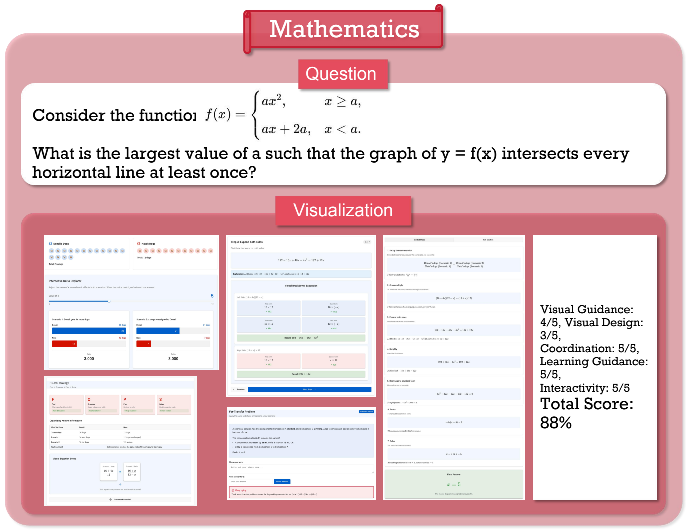


Figure 11: Representative example 4 from the Hard subset (MATH-500) of EduVisBench.

D.1 Task planning representation for an angular motion problem

Scenario, givens, goals, constraints

Scenario. A wheel has given angular acceleration and initial angular velocity; we want its angular velocity after a specific time.

Givens.
- Angular acceleration: $\alpha = 4\ \mathrm{rad/s^2}$
- Initial angular velocity: $\omega_0 = 2\ \mathrm{rad/s}$
- Time: $t = 5\ \mathrm{s}$

Goal.
- Find the angular velocity ω after 5 seconds.

Constraint.
- Use the kinematic equation for angular motion: $\omega = \omega_0 + \alpha t$.

Solution strategy

Analysis. Angular motion with constant angular acceleration; use $\omega = \omega_0 + \alpha t$ directly, since ω_0, α, t are given.

Step 1.
FOPS label:
- F: Angular motion with constant acceleration
- O: $\omega = \omega_0 + \alpha t$
- P: Substitute values
- S: Solve for ω

Description: Choose the kinematic equation $\omega = \omega_0 + \alpha t$.
Reasoning: This equation links ω, ω_0, α, and t, matching the given data.

Step 2.
FOPS label:
- F: Substitution and calculation
- O: $\omega = \omega_0 + \alpha t$
- P: Perform arithmetic
- S: Verify result

Description: Substitute $\omega_0 = 2$ rad/s, $\alpha = 4$ rad/s^2, $t = 5$ s: $\omega = 2 + (4 \times 5)$.
Reasoning: Direct substitution turns the symbolic relation into a numerical value.

Step 3.
FOPS label:
- F: Final calculation
- O: $\omega = 2 + 20$
- P: Simplify
- S: Verify correctness

Description: Compute $\omega = 2 + 20 = 22$ rad/s.
Reasoning: Ensures the arithmetic is correct and consistent with possible answer options.

Result: $\omega = 22\ \mathrm{rad/s}$.

Transfer tasks

Near transfer.
Problem: A wheel rotates with an angular acceleration of $3\ \mathrm{rad/s^2}$. If its initial angular velocity is $5\ \mathrm{rad/s}$, what is its angular velocity after 4 seconds?
Explanation: This is a near-transfer task because it has the same structure as the original problem but with different numerical values (still using $\omega = \omega_0 + \alpha t$).

Far transfer.
Problem: A car accelerates uniformly from rest at $3\ \mathrm{m/s^2}$. What is its velocity after 6 seconds?
Explanation: This is a far-transfer task because it involves linear motion instead of angular motion, but uses the analogous kinematic equation $v = v_0 + at$; the reasoning pattern is the same.

Instructional steps

Step 1. Understand the problem and identify the given values.
Concrete.
Entities:
- wheel
- angular acceleration
- initial angular velocity
- time

Description: A rotating wheel with a given angular acceleration and initial angular velocity over a specific time.

Representational.
Tool: Table.
Purpose: Organize the given values clearly for reference.
Represents: Angular acceleration (α), initial angular velocity (ω_0), and time (t).
Visual design:
- Form: a simple table with columns for quantity, symbol, and value.
- Elements:
  - headers for quantity, symbol, and value
  - rows for α, ω_0, and t
- Bridge: helps transition from the problem statement to identifying the variables needed for the equation.

Abstract.
Constructs:
- kinematic equation for angular motion: $\omega = \omega_0 + \alpha t$

Connection: The equation relates the given quantities (α, ω_0, t) to the unknown angular velocity (ω).

Think-aloud.
- What quantities are given in the problem?
- What is the goal of the problem?

Step 2.
Goal: Apply the kinematic equation for angular velocity.
Action: Write $\omega = \omega_0 + \alpha t$ and substitute the given values.
Concepts:
- Kinematic equation for angular motion

Pitfalls:
- Incorrect substitution of values
- Forgetting to include the initial angular velocity

Step 3.
Goal: Calculate the final angular velocity.
Action: Perform the arithmetic calculation to find ω.
Concepts:
- Basic arithmetic
- Angular velocity

Pitfalls:
- Arithmetic errors
- Misinterpreting the result

UI layout suggestion

Structure.
Step-by-step instructional layout with separate sections for scenario, solution strategy, and practice problems.

Content organization.
Put the scenario at the top, followed by the solution strategy with stepwise explanations, and practice problems at the end.

Key components.
- Card
- Accordion
- Alert
- Table

Math rendering.
Use KaTeX or MathJax for equations such as $\omega = \omega_0 + \alpha t$.

Figure 12: Two-column task-planning view for the angular motion example.

E FULL PROMPT AND INTERMEDIATE STRUCTURES FOR THE ANGULAR VELOCITY EXAMPLE

To make the behavior of each agent more transparent, we include below the full prompt structure for the angular velocity example discussed in reviewer v6KF Q3, covering `task_planning`, `conceptual_mapping`, `reasoning_decomposition`, `metacognitive_reviewer`, and the `visualization_page` design.

AGENT 1: TASK PLANNING (`task_planning`)

```
task_planning
  scenario
    context
      A wheel is rotating with a given angular acceleration and initial angular
          velocity.
      The task is to determine its angular velocity after a specific time.
    givens
      - Angular acceleration (\alpha) = 4 rad/s^2
      - Initial angular velocity (\omega_0) = 2 rad/s
      - Time (t) = 5 seconds
    goals
      - Find the angular velocity (\omega) after 5 seconds
    constraints
      - Use the kinematic equation for angular motion
  solution_strategy
    analysis
      This problem involves angular motion with constant angular acceleration.
      The angular velocity can be calculated using the kinematic equation for
          angular motion \omega = \omega_0 + \alpha t.
    steps
      - step_number 1
        fops_label F Angular motion with constant acceleration / O \omega = \
            omega_0 + \alpha t / P Substitute values / S Solve for \omega
        description
          Use the kinematic equation for angular velocity \omega = \omega_0 + \
              alpha t.
          Substitute the given values into the equation.
        reasoning
          This equation relates angular velocity, initial angular velocity,
              angular acceleration, and time.
          It is directly applicable since all required values are provided.
      - step_number 2
        fops_label F Substitution and calculation / O \omega = \omega_0 + \alpha
             t / P Perform arithmetic / S Verify result
        description
          Substitute \omega_0 = 2 rad/s, \alpha = 4 rad/s^2, and t = 5 s into
              the equation \omega = 2 + (4 x 5).
        reasoning
          Substituting the values allows us to calculate the final angular
              velocity directly.
      - step_number 3
        fops_label F Final calculation / O \omega = 2 + 20 / P Simplify / S
            Verify correctness
        description
          Simplify the equation \omega = 2 + 20 = 22 rad/s.
```

```
      reasoning
        This step ensures the final value is calculated correctly and matches
            the given options.
transfer_tasks
  near_transfer
    problem
      A wheel rotates with an angular acceleration of 3 rad/s^2.
      If its initial angular velocity is 5 rad/s, what is its angular velocity
           after 4 seconds?
    explanation
      This is a near-transfer task because it has the same structure as the
          original problem
      but with different numerical values.
  far_transfer
    problem
      A car accelerates uniformly from rest at a rate of 3 m/s^2.
      What is its velocity after 6 seconds?
    explanation
      This is a far-transfer task because it involves linear motion instead of
           angular motion,
      but the logic of using the kinematic equation remains the same.
instructional_steps
  - step_number 1
    goal Understand the problem and identify the given values.
    action Extract angular acceleration, initial angular velocity, and time
        from the problem statement.
    concepts
      - Angular motion
      - Kinematic equations
    pitfalls
      - Confusing angular velocity with linear velocity
      - Misidentifying the given values
  - step_number 2
    goal Apply the kinematic equation for angular velocity.
    action Write the equation \omega = \omega_0 + \alpha t and substitute the
        given values.
    concepts
      - Kinematic equation for angular motion
    pitfalls
      - Incorrect substitution of values
      - Forgetting to include the initial angular velocity
  - step_number 3
    goal Calculate the final angular velocity.
    action Perform the arithmetic calculation to find \omega.
    concepts
      - Basic arithmetic
      - Angular velocity
    pitfalls
      - Arithmetic errors
      - Misinterpreting the result

ui_layout_suggestion
  structure
    A step-by-step instructional layout with clear sections for scenario,
        solution strategy, and practice problems.
  key_components
    - Card
    - Accordion
    - Alert
```

```
    - Table
  content_organization
    Organize the scenario at the top, followed by the solution strategy with
        step-by-step explanations,
    and practice problems at the end.
  math_rendering
    Use KaTeX or MathJax to render equations like \omega = \omega_0 + \alpha t
        .
```

AGENT 2: CONCEPTUAL MAPPING (conceptual_mapping)

```
conceptual_mapping
  cra_mapping
    - step_number 1
      step_name Understand the problem and identify the given values
      concrete
        entities
          - wheel
          - angular acceleration
          - initial angular velocity
          - time
        description
          A rotating wheel with a given angular acceleration and initial angular
              velocity over a specific time.
      representational
        - tool Table
          purpose Organize the given values clearly for reference.
          represents Angular acceleration (\alpha), initial angular velocity (\
              omega_0), and time (t).
          visual_design
            form A simple table with columns for quantity, symbol, and value.
            elements
              - Headers for quantity, symbol, and value
              - Rows for \alpha, \omega_0, and t
            bridge
              Helps transition from the problem statement to identifying the
                  variables needed for the equation.
      abstract
        constructs
          - Kinematic equation for angular motion \omega = \omega_0 + \alpha t
        connection
          The equation relates the given quantities (\alpha, \omega_0, t) to the
               unknown angular velocity (\omega).
      think_aloud
        - What quantities are given in the problem?
        - What is the goal of the problem?

    - step_number 2
      step_name Apply the kinematic equation for angular velocity
      concrete
        entities
          - angular velocity
          - kinematic equation
        description
          Using the equation \omega = \omega_0 + \alpha t to calculate the
              angular velocity.
      representational
```

```
      - tool Equation
        purpose Show the relationship between angular velocity, initial
            angular velocity, angular acceleration, and time.
        represents The kinematic equation for angular motion.
        visual_design
          form Equation written in standard mathematical notation.
          elements
            - \omega = \omega_0 + \alpha t
          bridge
            Links the abstract formula to the concrete problem by substituting
                 the given values.
      - tool Dynamic animation (optional)
        purpose Illustrate how angular velocity changes over time with
            constant acceleration.
        represents The growth of angular velocity as time progresses.
        visual_design
          form A graph or animation showing \omega increasing linearly with
              time.
          elements
            - Time on x-axis
            - Angular velocity on y-axis
          bridge
            Visualizes the relationship described by the equation.
    abstract
      constructs
        - Substitution of values into \omega = \omega_0 + \alpha t
      connection
        Substituting concrete values into the abstract formula to find the
            solution.
    think_aloud
      - What is the equation we need to use?
      - How do we substitute the given values into the equation?

  - step_number 3
    step_name Calculate the final angular velocity
    concrete
      entities
        - arithmetic calculation
        - final angular velocity
      description
        Performing the calculation to find the final angular velocity.
    representational
      - tool Step-by-step arithmetic
        purpose Break down the calculation into smaller steps for clarity.
        represents The process of solving \omega = 2 + (4 x 5).
        visual_design
          form Sequential steps showing each part of the calculation.
          elements
            - Substitute values
            - Multiply \alpha and t
            - Add to \omega_0
          bridge
            Connects the abstract substitution to the concrete result.
      - tool Alert or feedback box
        purpose Highlight common errors in arithmetic or substitution.
        represents Potential pitfalls in the calculation process.
        visual_design
          form A colored box with tips or warnings.
          elements
```

```
          - Error examples
          - Corrective suggestions
        bridge
          Prevents errors and reinforces correct calculation methods.
    abstract
      constructs
        - Final result \omega = 22 rad/s
      connection
        The calculated result is the concrete realization of the abstract
            formula.
    think_aloud
      - What is the next step in the calculation?
      - Does the result make sense given the problem context?
```

AGENT 3: REASONING DECOMPOSITION (reasoning_decomposition)

```
reasoning_decomposition
  fops_reasoning
    - action
      Extract angular acceleration, initial angular velocity, and time from
          the problem statement.
    concepts
      - Angular motion
      - Kinematic equations
    scaffolding_needed true
    scaffolding_notes
      Learners may confuse angular velocity with linear velocity or
          misidentify the given values.
      Provide a table to organize the values.
    visual_support
      needed true
      which_visual Table for organizing given values
      interaction Draggable labels to match quantities with their symbols and
          values

    - action
      Write the equation \omega = \omega_0 + \alpha t and substitute the given
           values.
    concepts
      - Kinematic equation for angular motion
    scaffolding_needed true
    scaffolding_notes
      Learners may struggle with identifying the correct equation or
          substituting values correctly.
      Provide a guided substitution exercise.
    visual_support
      needed true
      which_visual Equation with placeholders for substitution
      interaction Reveals for each substitution step

    - action
      Perform the arithmetic calculation to find \omega.
    concepts
      - Basic arithmetic
      - Angular velocity
    scaffolding_needed true
    scaffolding_notes
```

```
      Arithmetic errors are common. Break down the calculation into smaller
          steps and provide feedback on errors.
    visual_support
      needed true
      which_visual Step-by-step arithmetic breakdown
      interaction Dynamic reveals for each calculation step

practice_activities
  - activity_number 1
    name Identify Given Values
    type fill-in
    task
      Fill in the table with the given values for angular acceleration, initial
          angular velocity, and time.
    gradual_release
      Start with guided examples where learners match quantities to symbols,
          then move to independent identification.

  - activity_number 2
    name Substitute Values into Equation
    type choice
    task
      Choose the correct substitution for \omega = \omega_0 + \alpha t given \
          alpha = 4 rad/s^2, \omega_0 = 2 rad/s, and t = 5 s.
    gradual_release
      Begin with hints and feedback for incorrect choices, then transition to
          independent substitution.

  - activity_number 3
    name Calculate Angular Velocity
    type judgment
    task
      Perform the calculation \omega = 2 + (4 x 5) and verify if the result is
          correct.
    gradual_release
      Provide step-by-step guidance initially, then allow learners to perform
          the calculation independently with feedback.

  - activity_number 4
    name Near Transfer Problem
    type explanation
    task
      Solve a similar problem
      A wheel rotates with an angular acceleration of 3 rad/s^2.
      If its initial angular velocity is 5 rad/s, what is its angular velocity
          after 4 seconds?
    gradual_release
      Start with guided problem-solving steps, then transition to independent
          problem-solving.

  - activity_number 5
    name Far Transfer Problem
    type explanation
    task
      Solve a far-transfer problem
      A car accelerates uniformly from rest at a rate of 3 m/s^2.
      What is its velocity after 6 seconds?
    gradual_release
```

```
    Provide a comparison to angular motion initially, then allow learners to
        solve independently.
```

AGENT 4: METACOGNITIVE REVIEWER (metacognitive_reviewer)

```
metacognitive_reviewer
  metacognitive_prompts
    before_solving
      - What quantities are given in the problem, and what do they represent?
      - What is the goal of the problem, and how can I achieve it?
    during_solving
      - Am I using the correct equation for this problem?
      - Have I substituted the values correctly into the equation?
    after_solving
      - Does my result make sense given the context of the problem?
      - What could I improve in my approach to solving similar problems?

  general_strategies
    - strategy Self-monitoring technique
      when_to_use During the substitution and calculation phases
      how_to_apply
        Pause after each step to verify the accuracy of substitutions and
            calculations.
    - strategy Error-checking routine
      when_to_use After completing the calculation
      how_to_apply
        Revisit the problem statement and verify that the result aligns with the
             given values and context.
    - strategy Visualization
      when_to_use During the understanding phase
      how_to_apply
        Organize the given values in a table or diagram to clarify relationships
             between quantities.
```

AGENT 5: VISUALIZATION PAGE DESIGN (visualization_page)

```
visualization_page
  layout_overview
    structure
      description >
        The webpage uses a step-by-step instructional layout with clearly
            separated sections.
      sections
        - id scenario_overview
          title Scenario Overview
          content Problem statement and given values.
        - id step_by_step_solution
          title Step-by-Step Solution
          content Interactive visuals for each step of the solution.
        - id practice_problems
          title Practice Problems
          content Near-transfer and far-transfer tasks.
        - id reflection_prompts
          title Reflection Prompts
```

```
        content Metacognitive questions to guide learners.
  ui_components
    - Cards for each step in the solution strategy.
    - Accordion for expandable explanations and hints.
    - Alerts for common pitfalls and corrections.
    - Interactive visuals including tables, equations, and animations.

static_visuals
  step_1_organizing_values
    goal Organize given values for reference.
    visual_design
      type table
      columns
        - Quantity
        - Symbol
        - Value
      rows
        - Angular acceleration (\alpha)
        - Initial angular velocity (\omega_0)
        - Time (t)
    interaction
      description Learners drag and drop labels to match quantities with
          symbols and values.
  step_2_applying_equation
    goal Make the kinematic equation explicit and concrete.
    visual_design
      equation_form \omega = \omega_0 + \alpha t
      substitution_example \omega = 2 + (4 x 5)
    dynamic_support
      description Equation initially shows placeholders; each placeholder
          reveals the substituted value when clicked.
  step_3_calculation
    goal Make the arithmetic process transparent.
    visual_design
      breakdown_steps
        - 4 x 5 = 20
        - 2 + 20 = 22
      representation flowchart or sequential steps.

dynamic_behavior
  animation_angular_velocity_graph
    purpose >
      Show how angular velocity increases linearly over time with constant
          angular acceleration.
    design
      axes
        x_axis Time t (0 to 5 s)
        y_axis Angular velocity \omega (0 to 22 rad/s)
      behavior
        start_value 2 rad/s
        end_value 22 rad/s
    interaction
      on_load Line animates from t = 0 to t = 5.
  animation_stepwise_substitution
    purpose Visualize the substitution process in the equation.
    design
      base_equation \omega = \omega_0 + \alpha t
      placeholders
        - \omega_0 = 2
```

```
        - \alpha = 4
        - t = 5
      timing
        effect short fade-in for each reveal.

  user_interactions
    interactive_table
      description >
        Drag-and-drop interface to match quantities, symbols, and numerical
            values.
    equation_substitution
      description >
        Clickable placeholders in the equation that reveal each substituted
            value in sequence.
    arithmetic_breakdown
      description >
        "Next step" interaction revealing 4 x 5 = 20, followed by 2 + 20 = 22,
            with optional feedback messages.
```

F Qualitative Success and Failure Cases

Success Cases

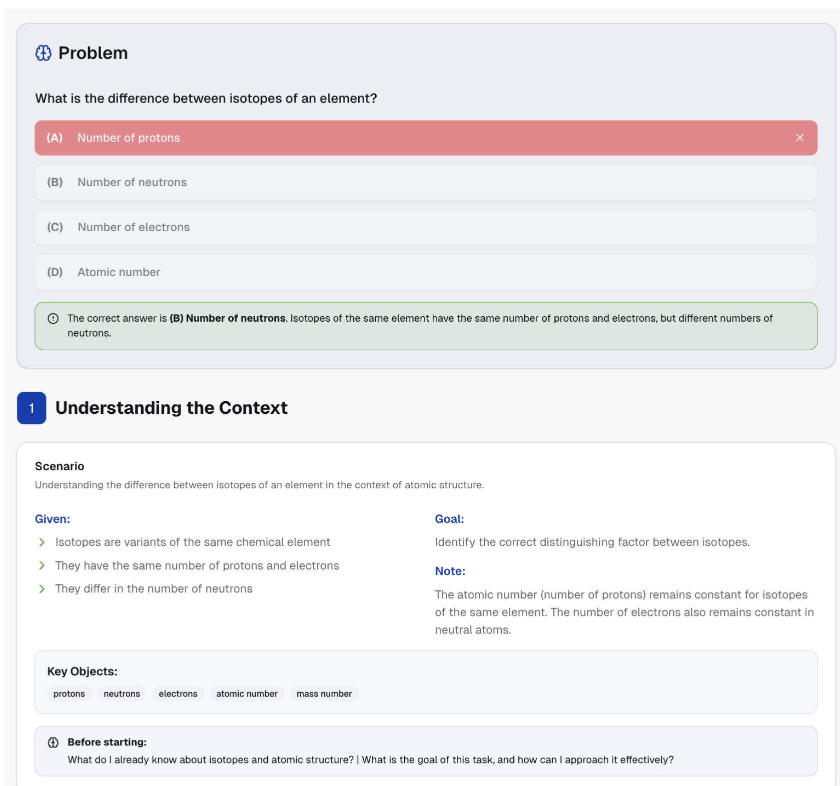
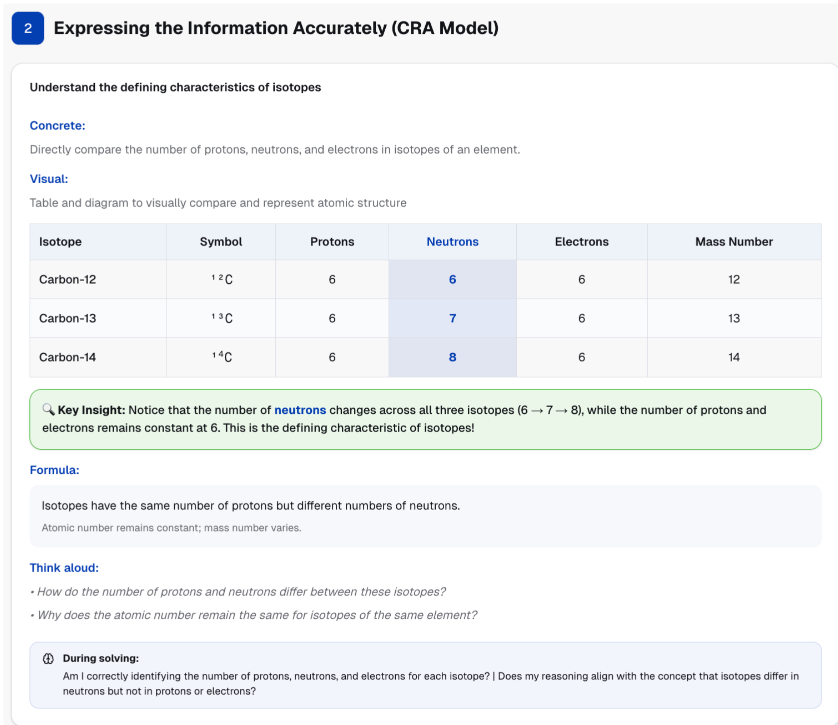


Figure 13: Success case 1 illustrating coherent multi-agent collaboration.



2 **Expressing the Information Accurately (CRA Model)**

Understand the defining characteristics of isotopes

Concrete:

Directly compare the number of protons, neutrons, and electrons in isotopes of an element.

Visual:

Table and diagram to visually compare and represent atomic structure

Isotope	Symbol	Protons	Neutrons	Electrons	Mass Number
Carbon-12	^{12}C	6	**6**	6	12
Carbon-13	^{13}C	6	**7**	6	13
Carbon-14	^{14}C	6	**8**	6	14

Key Insight: Notice that the number of **neutrons** changes across all three isotopes (6 → 7 → 8), while the number of protons and electrons remains constant at 6. This is the defining characteristic of isotopes!

Formula:

Isotopes have the same number of protons but different numbers of neutrons.

Atomic number remains constant; mass number varies.

Think aloud:

- *How do the number of protons and neutrons differ between these isotopes?*
- *Why does the atomic number remain the same for isotopes of the same element?*

During solving:

Am I correctly identifying the number of protons, neutrons, and electrons for each isotope? | Does my reasoning align with the concept that isotopes differ in neutrons but not in protons or electrons?

Figure 14: Success case 2 illustrating coherent multi-agent collaboration.

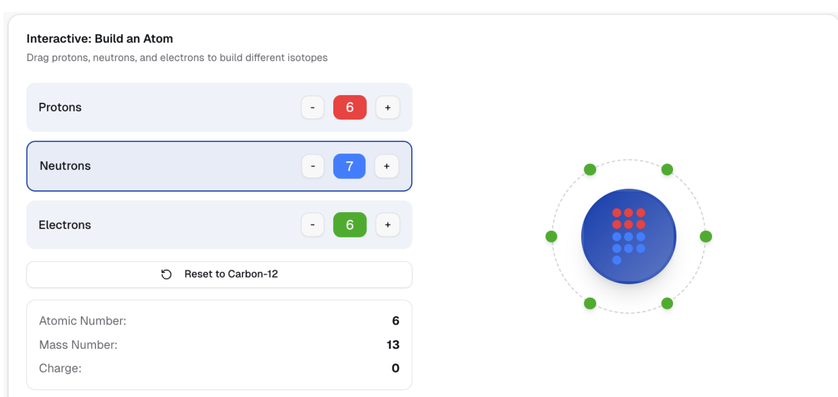


Figure 15: Success case 3 illustrating coherent multi-agent collaboration.

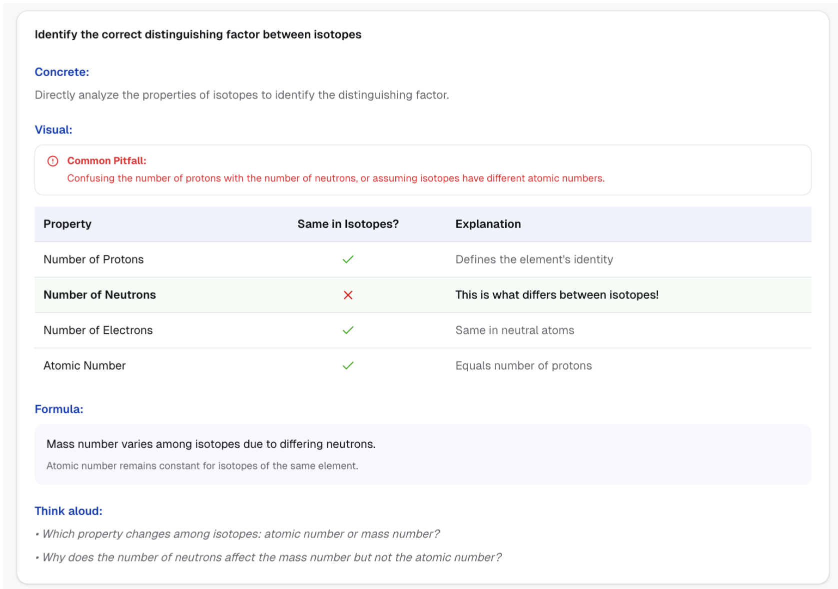


Figure 16: Success case 4 illustrating coherent multi-agent collaboration.

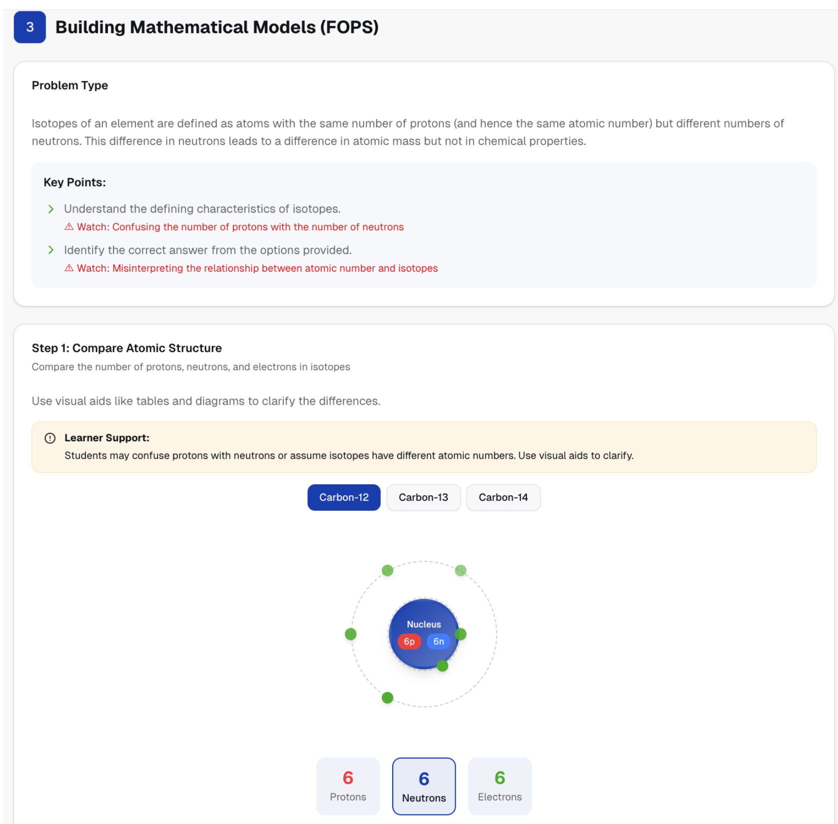


Figure 17: Success case 5 illustrating coherent multi-agent collaboration.

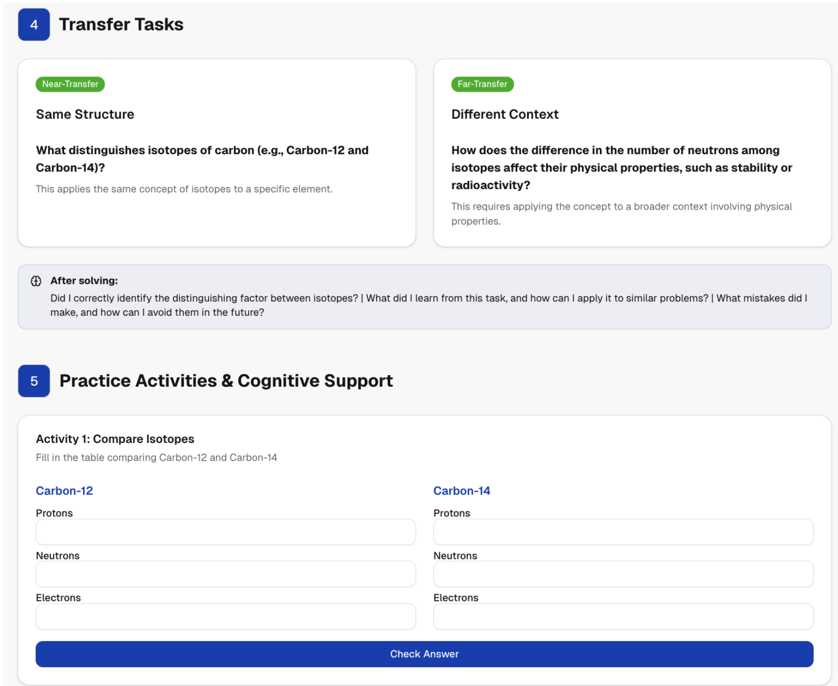


Figure 18: Success case 6 illustrating coherent multi-agent collaboration.

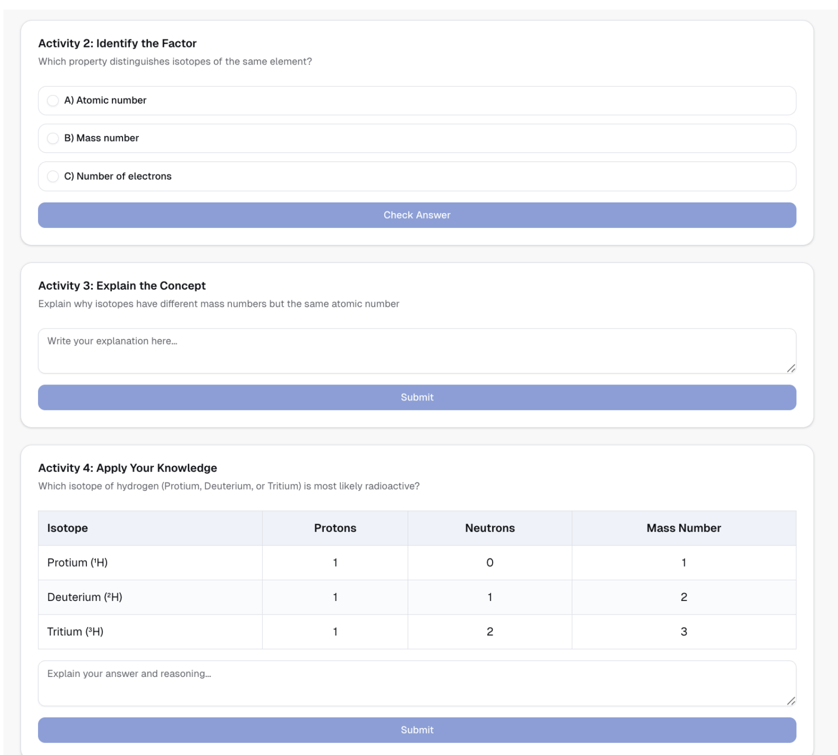


Figure 19: Success case 7 illustrating coherent multi-agent collaboration.

FAILURE CASE: ERROR ACCUMULATION IN THE MULTI-AGENT PIPELINE

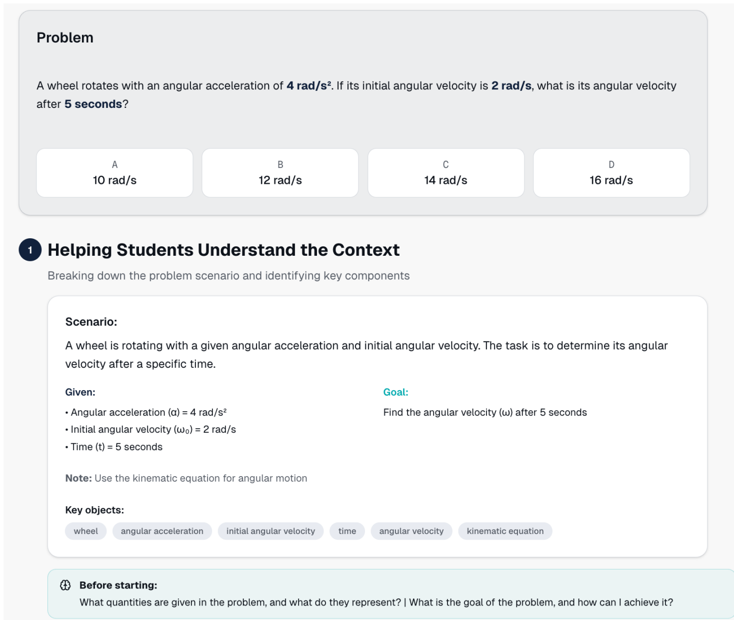


Figure 20: Failure case step 1. The initial `task_planning` agent proposes an inappropriate `ui_layout_suggestion` where elements such as `Card`, `Accordion`, `Alert`, and `Table` have no meaningful relation to the underlying concept of angular velocity.

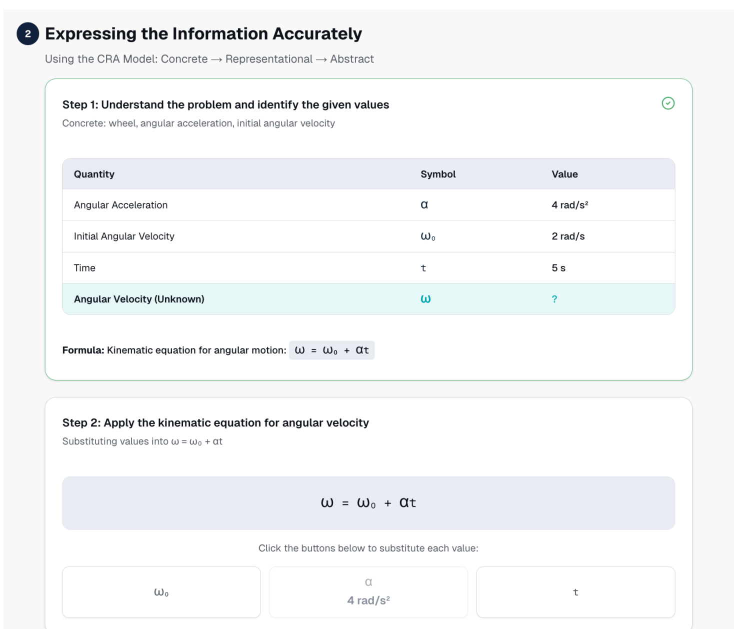


Figure 21: Failure case step 2. Downstream reasoning agents infer the need for a rotating wheel visualization, but the earlier erroneous layout suggestion continues to impose irrelevant UI constraints.

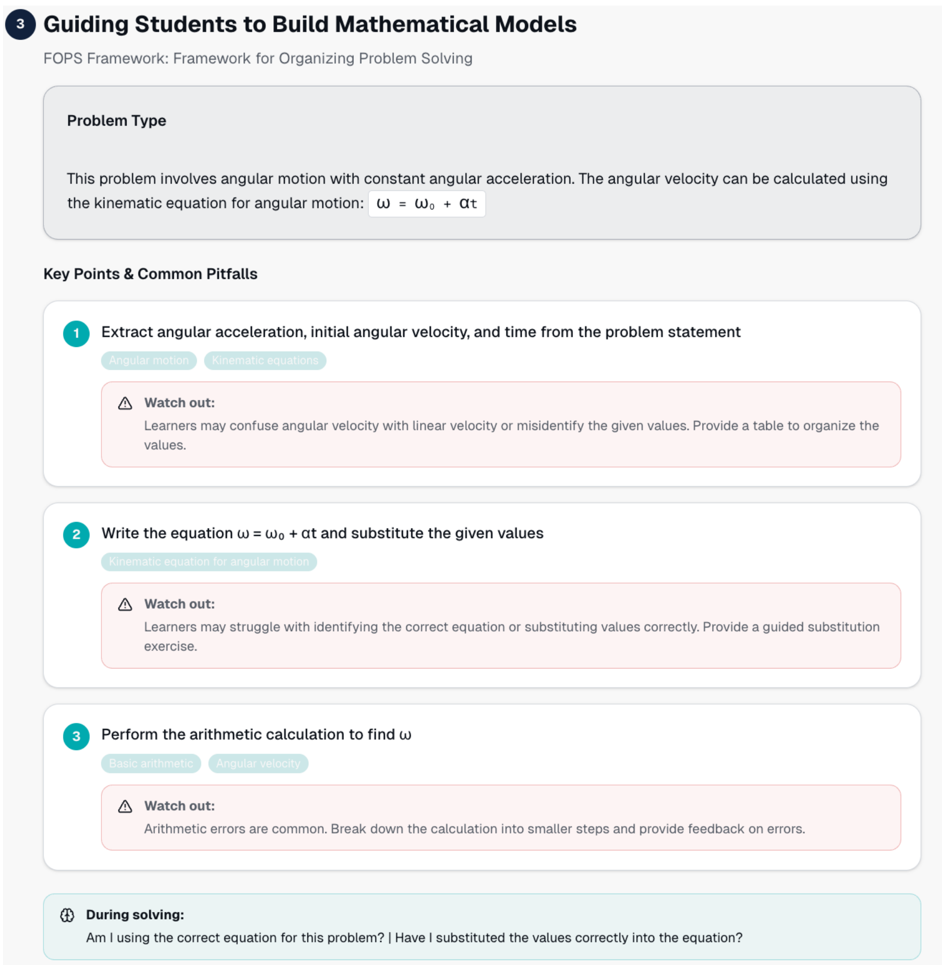


Figure 22: Failure case step 3. Visualization Agent attempts to comply with upstream layout directives, resulting in non-physics-related components instead of correct angular-motion visuals.

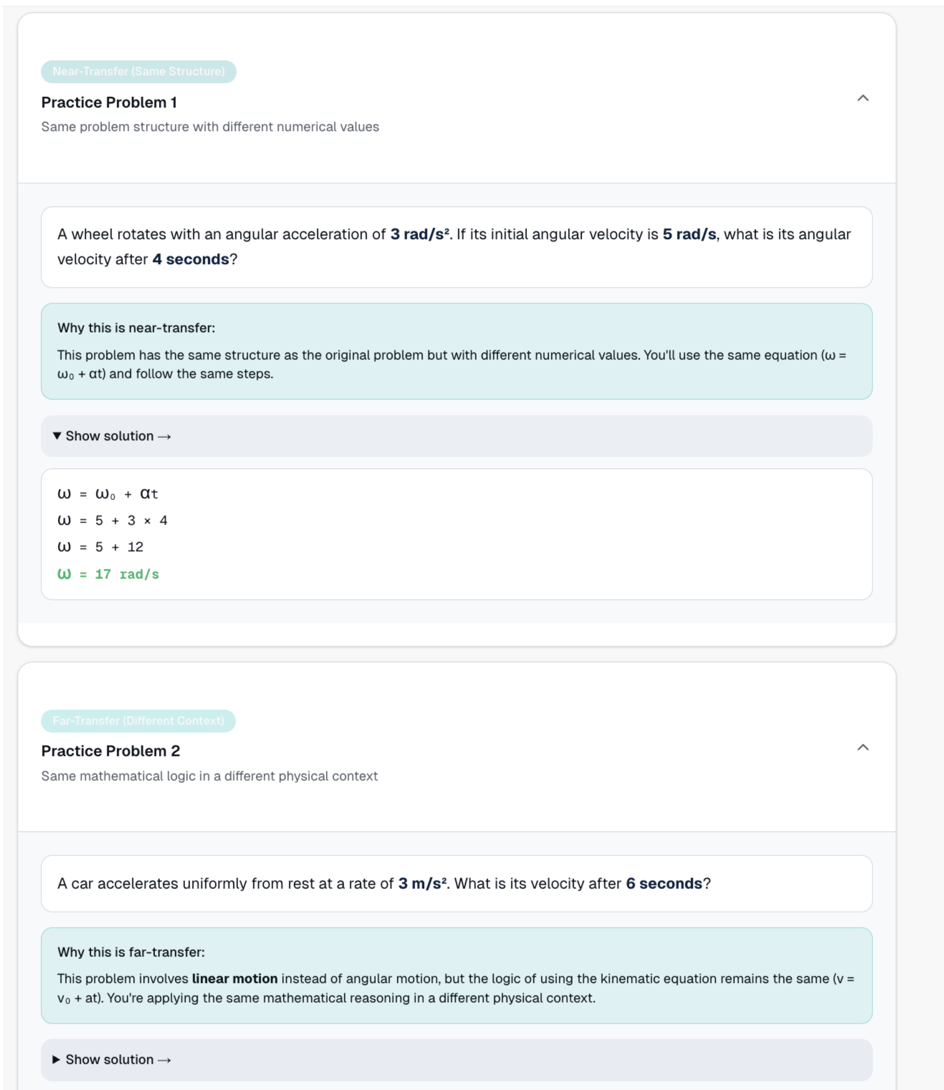


Figure 23: Failure case step 4. The multi-agent chain amplifies the initial planning mistake: despite correct domain inference, the final UI is structurally misaligned.

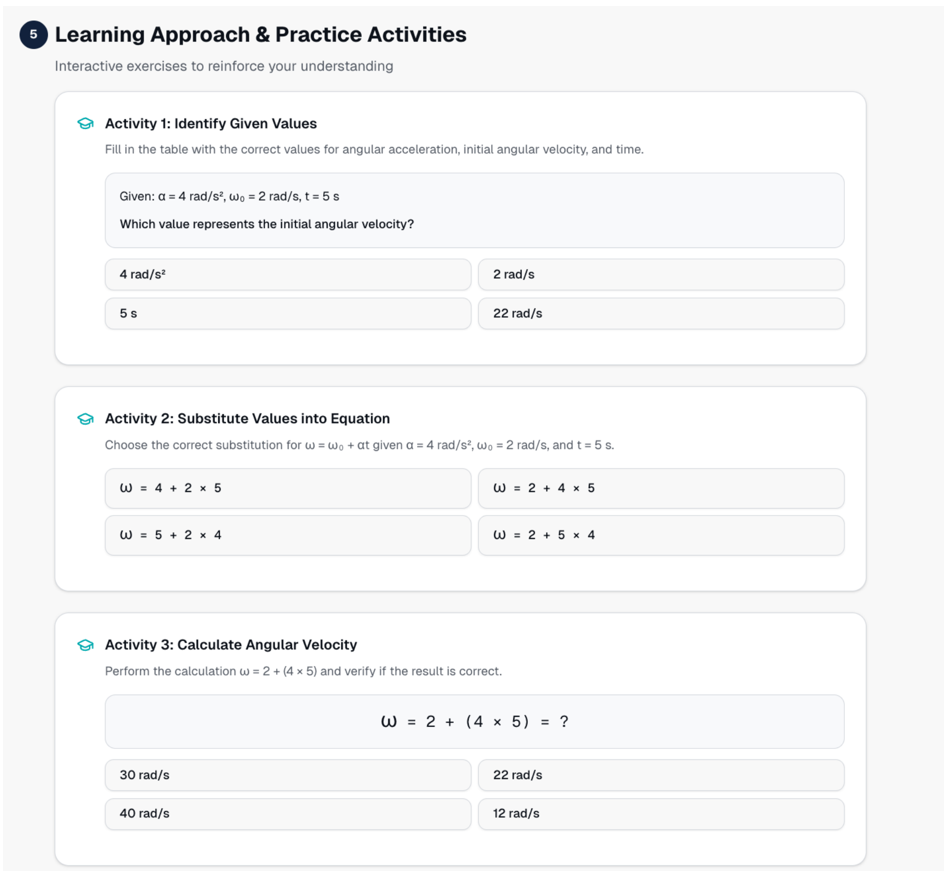


Figure 24: Failure case final outcome. Due to accumulated constraint errors, the system fails to produce any meaningful visualization related to angular velocity, demonstrating error propagation in multi-agent settings.