

Article

Prompting for Independent Learning: An Evaluation of Tutoring Behaviors in GenAI

Kendall Hartley ^{1,*} , Fabiola Saéz-Delgado ² and Javier Mella-Norambuena ³ 

¹ Department of Teaching and Learning, University of Nevada, Las Vegas, Las Vegas, NV 89154-3005, USA

² Faculty of Education, Universidad Católica de la Santísima Concepción, Concepción 4090541, Chile; fsaez@ucsc.cl

³ Department of Science, Universidad Técnica Federico Santa María, Concepción 4603255, Chile; javier.mellan@usm.cl

* Correspondence: kendall.hartley@unlv.edu

Abstract

Generative AI tutors have become a common tool for independent learning, yet their capacity to support self-regulated learning (SRL) is poorly understood. This simulation-based textual analysis of prompt design evaluates a frontier large language model (Claude Sonnet 4.6) as a tutor across 60 scripted sessions on a single topic (density), crossing three levels of SRL-informed system prompting (Minimal, Moderate, Extensive) with four learner-behavior variants (Standard, Misconception, Disengagement, Overconfidence). Tutoring transcripts were scored on a 14-dimension framework spanning SRL phases, SRL developmental stages, self-determination theory principles, and Merrill's First Principles of Instruction, applied via an LLM judge. Adding SRL context to the system prompt raised total tutoring scores, but only at the Extensive SRL support level. Minimal and Moderate prompting produced the same performance, near 36 on a 70-point scale, and Extensive prompting raised it to 40, a statistically significant effect (partial $\eta^2 = 0.24$). The learner's behavior in the session had a larger effect than the prompt did (partial $\eta^2 = 0.37$), with disengaged learners scoring lowest. The threshold pattern held under an independent judge from a different developer than the tutor model. The findings support a method for evaluating GenAI tutors empirically and point to dynamic, dialogue-aware prompting alongside explicit SRL scaffolding.

Keywords: generative AI; intelligent tutoring systems; self-regulated learning; self-determination theory; prompt engineering; large language model evaluation



Academic Editor: Massimo Cafaro

Received: 6 May 2026

Revised: 22 July 2026

Accepted: 24 July 2026

Published: 29 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With the widespread adoption of GenAI, students are shifting how they engage in independent learning [1,2]. Studying on their own is largely dependent upon their own motivations and self-regulated learning skills [3]. Working one on one with teachers, parents, or tutors is clearly advantageous and superior to independent learning [4]. Although attractive, one-on-one human tutoring is difficult to scale. Intelligent tutoring systems emerged as a potentially scalable adjunct to independent learning and engaged the learner in logic-based dialogue that depended at least in part on a script [5]. The rapid emergence of increasingly capable generative artificial intelligence tools has provided learners with the opportunity to engage with a personalized, ever available, and knowledgeable tutor. A majority of U.S. teens are using AI chatbots, with as many as 30% using them daily [2].

The primary use case is to help with schoolwork. This widespread use creates an important need to evaluate these tools. In particular, the role of these tools in encouraging and supporting self-regulated learning is critical.

There is evidence that LLM tutors can improve performance, but results depend on implementation and tool design [6]. Although in-session performance improves, later unaided performance can suffer. Unguarded and poor design can hinder learning. Analysis of AI-supported human tutor messaging revealed that the tutors made greater use of high-quality strategies with AI [7]. We should continue to evaluate how GenAI tools can effectively serve in this role.

1.1. Self-Regulated Learning and Self-Determination Theory as Frameworks for Analysis

Research and frameworks related to self-regulated learning provide a useful backstop to approaching the use of generative AI tutoring. Self-regulated learning posits that students bring certain capacities and skills to the learning environment that can support effective learning. These propensities include attention to how well one understands what they are learning (e.g., monitoring), deployment of strategies to improve comprehension (e.g., self-questioning), and the evaluation of learning outcomes (e.g., self-reflection). Models by a variety of scholars have demonstrated how this might be developed and what it looks like in practice. This includes the work of Zimmerman and colleagues that identifies the cyclical phases of the learning process, forethought performance, and reflection [3]. Zimmerman's model also provides a developmental framework for the progression towards autonomous self-directed learning. These phases, in sequence from beginners to experts, are observation, emulation, self-control, and self-regulation.

Most scholars include motivation as a component of SRL, but the details are not as clearly defined as they are in the work of Ryan and Deci on student motivation. Self-determination theory (SDT) suggests that student motivation can be described as a three-pronged construct consisting of a sense of autonomy, a sense of competence, and a feeling of relatedness [8]. As with self-regulated learning, self-determination theory provides a useful frame for the evaluation of GenAI tutoring. Tutoring behaviors that leave decisions with the learner and that build a sense of progress toward mastery would be expected to sustain the motivation required for independent study.

Together, self-regulated learning and self-determination theory provide an excellent framework for evaluating how a tutor might support the development of an independent learner and the delivery of engaging and effective instruction.

1.2. Effective Tutoring

Research into effective tutoring systems has identified four key attributes that are essential to effective support for learning. These four principles include interactivity, adaptivity, immediate feedback, and scaffolding [5]. Interactivity refers to the need for the system to respond in real time to the learner's needs. Adaptivity supports the need to customize instruction to the unique characteristics and background of the individual. Immediate feedback supports students' needs for specific, corrective, and constructive suggestions. Scaffolding refers to the capacity of the system or tutor to provide just enough support to guide the learner towards a better understanding of the concepts under review. The evidence for the effectiveness of tutoring itself is substantial. A meta-analysis of 89 randomized field experiments found that preK-12 tutoring programs raised achievement by an average of 0.29 standard deviations [9].

The effects of tutoring hold across countries and delivery formats. In Italy, the Tutoring Online Program matched volunteer university students with underprivileged middle school students and raised math performance by roughly 0.2 standard deviations both

during and after the pandemic school closures [10]. A similar program in Spain produced gains of 0.26 standard deviations on standardized mathematics tests for secondary students from disadvantaged backgrounds [11]. In Botswana, tutoring delivered by phone call and text message improved learning among primary students when schools were closed [12]. These programs differ in tutor preparation and medium, but each depends on a tutor attending to the needs of an individual learner.

Scaffolding becomes particularly relevant when it comes to using generative AI for tutoring purposes. The balance between tutoring and cheating is an ever-present challenge. Depending on the developmental level of the student, some level of encouragement for use focused on understanding, and discouragement for uses simply designed to produce an answer, may be needed. Novice learners may need an increased level of encouragement and support to better understand how to use the tutor for learning and the difference between that and other uses.

There are also well-established principles that can guide the standard dialogue between tutors and pupils [5]. At its core is (a) the initial question from the tutor, (b) student response, (c) tutor feedback, (d) continued dialogue, and (e) a concluding assessment guided by the tutor. Within each step are numerous considerations that harken back to the four principles described above. These principles also map onto the self-regulated learning cycle described in the previous section. Immediate feedback informs the monitoring of understanding. The scaffolding shifts responsibility for strategy use to the learner. The assessment of the dialogue supports self-reflection [3,13].

General instructional design principles can also guide the development of effective tutoring systems. Merrill's first principles of instruction include providing problem-centered instruction, activation of background knowledge, demonstration or worked examples, and opportunities for application [14].

At a broader level, effective tutoring involves a certain level of metacognitive and affective support, issues that are traditionally addressed with little difficulty by human tutors. An effective human tutor can reasonably judge the motivational state of the learner when they are working right next to them. In addition, metacognitive support is best provided by modeling the appropriate behavior, something effective human tutors can do quite well. Early intelligent tutoring systems were challenged in this regard simply by the broad range of data available and responses possible, which made it difficult to support these considerations.

These broader considerations are more possible now with the frontier large language models (LLMs) that present very large interaction space and large context windows capable of accommodating a substantial amount of support data. The context window, the data directly available to the system in real time about the learner and the topic, provides helpful information that the system can use to support learner development and motivation. The challenge becomes how to best use this capacity. Decades of work in technology-rich environments suggest that design decisions rather than capacity will best support learning [15]. Proactive tutoring that incorporates what has been learned regarding self-regulated learning and motivation is an important consideration to determine how to best develop learning environments for AI-supported tutoring.

1.3. Generative AI Tutors in Practice

Tutors can be proactive or passive in their interaction style [16]. The passive tutor will wait for the learner to initiate the dialogue and serve primarily as a source for answers. The proactive tutor will initiate and direct the learner, serving more as a guide who knows the domain [17]. Proactive tutoring supports the development of student self-regulated learning skills and progression toward self-directed learning [3]. In addition, proactive

approaches can address the motivation and engagement that sustain continued use of the tool [18].

The nature of feedback is an essential consideration, particularly when comparing proactive and passive GenAI tutoring styles [16]. Although a passive tutor may offer feedback primarily in response to a learner's direct query or attempted solution, often focusing on correctness or providing supplementary information, a proactive tutor can use feedback more strategically. A proactive tutor is better positioned to offer metacognitive and motivational feedback by assessing the learner's overall process, strategy use, and engagement, aligning with the principles of Self-Regulated Learning (SRL) and Self-Determination Theory (SDT). This includes timely, contextualized suggestions for monitoring, planning, and reflection, thereby actively guiding the learner's development of SRL aptitude rather than simply evaluating the outcome of a single interaction.

Field tests of GenAI tutors now span a range of contexts, and the findings do not all point in the same direction. In a large undergraduate physics course, students who worked through a lesson with a GenAI tutor built on pedagogical best practices learned about twice as much as peers in an active learning classroom in less time [19]. An after-school program in which teachers guided secondary students through GenAI tutoring sessions produced gains of approximately 0.3 standard deviations over six weeks [20]. Results from high school mathematics classrooms in Turkey are less encouraging. Students given unrestricted access to GPT-4 solved more practice problems correctly but scored 17 percent lower than their peers on a subsequent exam taken without the tool [6]. A version with tutoring guardrails built into the prompt removed the harm but did not produce gains.

These contradictions are informative when read through the proactive and passive distinction described above. The tutors in the physics and after-school-program studies were directed, through the prompt or through the teacher, to guide rather than to answer. The unmodified chatbot in the Turkey study served as a passive answer source, and the learners treated it as one. What appears to be a contradiction in the effects of GenAI tutoring may instead be a difference in the instructions the model was given.

1.4. Providing Context and Evaluating GenAI Tutors

A GenAI tutor does not necessarily carry a proactive approach to a tutoring conversation. In fact, it may not know that it is a tutoring conversation. How the tutor behaves will depend in large part on the context that is given. Providing this context in and of itself is an act of self-regulation. Many learners will not know to provide this supportive context especially when they are using a generic AI chatbot. There is a plain mechanism behind the expectation that prompt specificity matters. Frontier models are trained to follow instructions, and the system prompt is the channel through which behavioral expectations reach the model. A prompt that names the tutoring moves it wants gives the model something concrete to comply with, and a prompt that names none leaves the default answer-providing behavior in place. The field results reviewed above point the same way, with gains from a tutor built on content-rich prompt engineering [19] and harm from a chatbot given no tutoring instructions at all [6].

The primary entry point into generative AI chatbots is through one of the frontier models, such as OpenAI's ChatGPT, Anthropic's Claude, Google's Gemini, and xAI's Grok, rather than a tool designed specifically for learning. Often, learners will approach these tools with a limited understanding of how to best prompt the tool to be supportive of learning. Providing context to the tool by, for example, requesting that it act in a more proactive versus passive manner would be beneficial to the learner. Demonstrating the differences between relevant context being provided or not is an important consideration when evaluating the tools. The context provided could be as simple as please behave as

a tutor and expert tutor and explain to me topic X. The context provided could be much more elaborate, of course, by providing the tutor or LLM with extensive context such as the topics described above and maybe even some research. Alternatively, there could be a simplified version of the context, similar to what Anthropic and others describe as “skills”. Skills are summary documents that guide the LLM in providing educated responses.

Evaluation of GenAI models in other arenas is common. Users are rightfully interested in how the models will perform in specific tasks for each discipline. For example, in customer service applications, models are evaluated on their ability to handle requests and then compared to human responses. Anthropic conducts extensive tests of its models, particularly as it pertains to their safety research [21]. Efforts to evaluate the pedagogical abilities of LLMs directly have produced a mixed picture. An early test of dialogue models found that their responses could resemble those of a human teacher but were rated well below the teachers on the ability to help the student [22]. A more recent evaluation taxonomy scored LLM tutors across eight dimensions, including mistake identification, providing guidance, and revealing the answer, and found that models often identified the student’s error and still handed over the solution rather than guiding the learner toward it [23]. These evaluations score tutoring behaviors themselves rather than downstream learning, and the two approaches can disagree. A tutor that simply provides answers may produce strong practice performance and weak learning, as the Turkey study demonstrates. Prior research has provided preliminary analysis of how ChatGPT performs in a tutorial mode for a student attempting to learn to code [24]. This study will extend that work by applying the new GenAI Tutor evaluation framework, a rubric designed to encourage the use of research-supported practices [25]. The GenAI Tutor Evaluation Framework is designed to support and evaluate tutors based on self-regulated learning theory [3], self-determination theory [8], and instructional design first principles [14].

The purpose of this study is to test whether varying the SRL-related context in a tutor’s system prompt affects the tutoring performance of a frontier large language model. The design is a simulation-based textual analysis of prompt design, and the evidence it produces concerns tutoring behaviors visible in transcripts rather than learner outcomes. In particular, this study examines one model (Claude Sonnet 4.6) on one topic (density) across 60 scripted sessions, crossing three levels of SRL context with four learner behaviors and scoring every session on the GenAI Tutor Evaluation Framework. The findings come from a single model, a single topic, and a fixed set of scripted behaviors, and they are reported as early evidence that a targeted change to the system prompt can move SRL-supported tutoring behavior in measurable ways. This study also contributes a reusable platform (GenAITutor.net) for running and scoring this kind of evaluation. This study addresses three research questions:

RQ1. *How does the level of SRL-related context in the system prompt affect the overall tutoring performance of a single frontier LLM, as measured by the GenAI Tutor Evaluation Framework?*

RQ2. *Does the effect of SRL context hold across four distinct learner behaviors (Standard, Misconception, Disengagement, and Overconfidence)?*

RQ3. *Which dimensions of tutoring performance respond to system-prompt context, and which respond instead to the learner’s behavior in the dialogue?*

2. Methodology

2.1. Research Design

This study used a factorial design to describe the impact of varying SRL context support on the tutoring performance of a single large language model. It follows a scripted student model dialogue methodology that has become a common approach to evaluate GenAI learner interactions [26,27]. The independent variable is the SRL support level embedded in the system prompt, operationalized at three levels: Minimal, Moderate, and Extensive. The dependent variables are the scores on the GenAI Tutor Evaluation Framework, an instrument grounded in SRL theory [3], self-determination theory [8], and First Principles of Instruction [14].

The LLM under evaluation is Anthropic's Claude Sonnet 4.6. Rather than comparing multiple models, this design isolates the effect of the instructional context provided to a single model. Learners and educators can influence how they prompt a GenAI tutor but cannot modify the underlying model, making the system prompt a particularly relevant variable to investigate.

Each session follows a scripted multi-turn dialogue on the topic of density, a middle-school science concept with well-documented student misconceptions that lends itself to varied learner interactions [28]. Four scenario variants represent distinct learner behaviors: Standard Learning, Misconception, Disengagement, and Overconfidence (see Figure 1). This yields a 3×4 factorial design with 12 unique conditions. Each condition is replicated five times to account for the probabilistic nature of LLM outputs, producing 60 tutoring sessions (see Figure 2 for sample session).

Configuration Panel

Select LLM
Claude Sonnet 4.6

Select Topic
Density

SRL Support Level
Minimal

Scenario Variant
 ✓ Standard
 Misconception pathway
 Disengagement pathway
 Overconfidence pathway

Session Label (Optional)
e.g., Run 1, Testing Misconception

☐ **Auto-run scenario (Automatically process dialogue & save)**

Start Session

Figure 1. Configuration panel for the tutoring simulation. Each session is configured by selecting the model (Claude Sonnet 4.6), the topic (Density), one of three SRL support levels (Minimal, Moderate, or Extensive), and one of four scenario variants (Standard Learning, Misconception, Disengagement, or Overconfidence). The auto-run option processes the scripted dialogue and saves the session record.

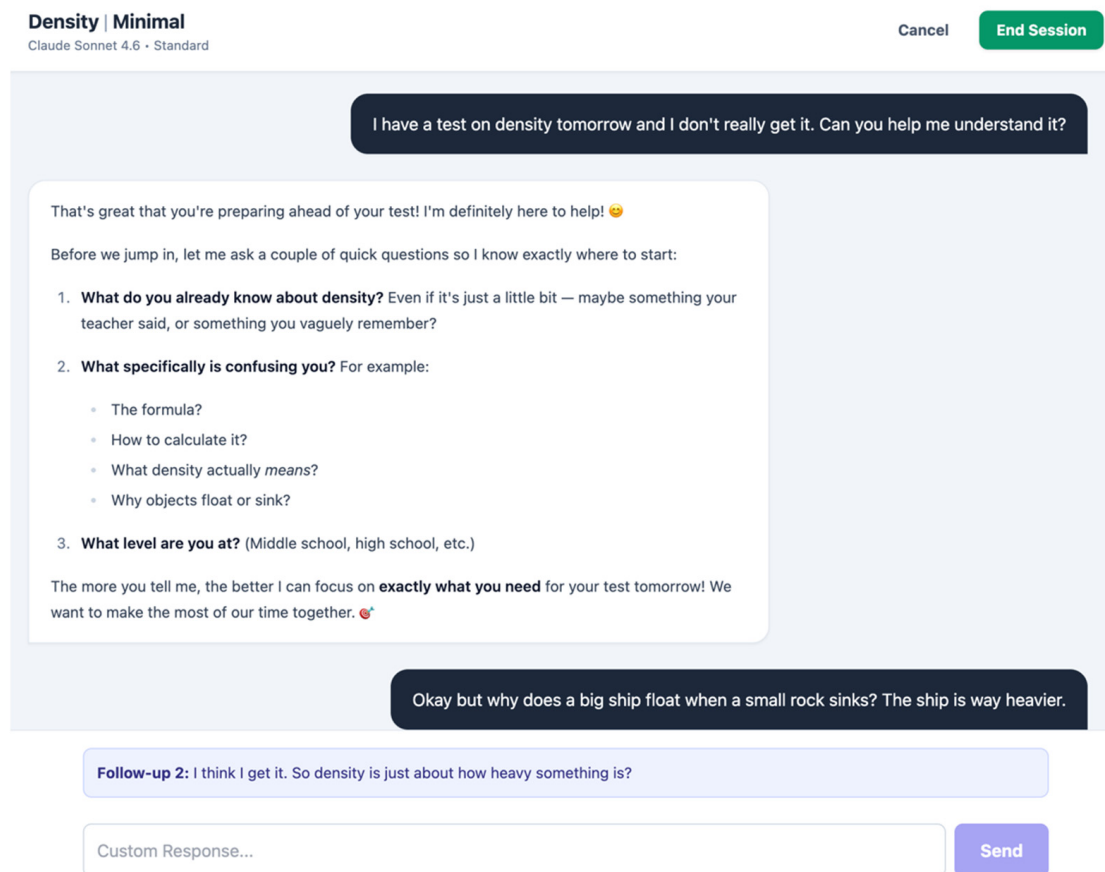


Figure 2. Sample dialogue from a session in the Minimal support condition (density topic, Standard variant). The simulated learner opens with a request for help, the tutor responds, and the panel presents the next scripted follow-up prompt along with a custom response option.

2.2. Instruments

2.2.1. GenAI Tutor Evaluation Framework

Transcripts were scored using the GenAI Tutor Evaluation Framework [25], which operationalizes research-supported tutoring attributes through the development of an Innovation Configuration (IC) Map [29,30]. The framework draws from established research findings that are behind the SRL, SDT, and First Principles models described above. The framework evaluates tutoring performance across four categories: SRL phases (forethought, performance, and self-reflection), SRL developmental stages (observation, emulation, self-control, and self-regulation), SDT principles (autonomy, competence, and relatedness), and First Principles of Instruction. The framework scores four of Merrill's five First Principles, problem-centered, activation, demonstration, and application. We omit integration, which concerns the learner carrying new skills into their own life and work over time, because a single short tutoring session offers no occasion to observe it. Each component is rated on a 1 to 5 scale, where 1 indicates that the attribute is absent or very poorly implemented and 5 represents exemplary implementation. See Table 1 for a summary of the framework and Appendix B for the detailed scoring rubric.

Table 1. GenAI Tutor Evaluation Framework: dimensions, guiding questions, and scale anchors.

Dimension	Guiding Question	1 (Absent)	5 (Strong)
<i>SRL Phases</i>			
Forethought	Does the tutor help the learner set specific goals and plan before starting on content?	Skips goal setting and delivers content immediately	Elicits specific, measurable goals and revisits them throughout the session
Performance Phase	Does the tutor scaffold strategy use, check comprehension, and adjust instruction during the session?	Delivers a monologue with no comprehension checks	Names strategies, checks understanding frequently, and adjusts pace and depth
Self-Reflection	Does the tutor prompt the learner to evaluate their own understanding and plan ahead?	Session ends with no reflection prompt	Guides self-assessment against stated goals, identifies gaps, and plans next steps
<i>SRL Developmental Stages</i>			
Observation	Does the tutor model expert thinking or SRL behaviors for the learner to observe?	States facts with no demonstrations or worked examples	Walks through a worked example with an explicit think-aloud
Emulation	Does the tutor invite the learner to try a task with scaffolded support?	Never invites the learner to attempt anything independently	Presents a scaffolded task, waits for the attempt, and gives targeted feedback
Self-Control	Does the tutor ask the learner to monitor their own confidence and strategy use?	No metacognitive prompts	Regularly elicits self-assessment and uses the answers to adjust instruction
Self-Regulation	Does the tutor fade scaffolding and encourage independent or transferred use?	The same scaffolding persists from start to finish	Reduces hints over the session and asks the learner to apply the concept in a new situation
<i>SDT Principles</i>			
Autonomy	Does the tutor offer genuine choices about direction, pacing, or approach?	Dictates the entire session with no learner input	Offers meaningful choices and visibly follows the learner's selections
Competence	Does the tutor provide specific, calibrated feedback that builds a sense of capability?	Feedback is generic or absent	Adjusts difficulty to the learner's responses and names what the learner did well
Relatedness	Does the tutor connect learning to the learner's experiences, interests, or goals?	Never references the learner's background or context	Solicits and integrates the learner's experiences throughout the session
<i>First Principles of Instruction</i>			
Problem-Centered	Is learning anchored in a concrete, authentic problem?	Content is delivered abstractly with no real-world framing	The session grows from an authentic problem and concepts serve its solution
Activation	Does the tutor surface prior knowledge and connect new content to it?	Starts teaching with no check of prior knowledge	Surfaces prior knowledge, corrects misconceptions, and builds on what the learner knows
Demonstration	Does the tutor show examples, worked problems, or representations rather than only explain?	Explains in words alone with no examples	Provides varied examples and step-by-step walkthroughs tied to the learner's situation
Application	Does the learner receive structured practice with timely, specific feedback?	No practice opportunities are given	Provides several practice exercises with feedback explaining why answers are right or wrong

Note. Each dimension is rated on a 1 to 5 scale. A rating of 3 marks partial implementation, and ratings of 2 and 4 fall between the anchors. Appendix B presents the full rubric with anchor descriptions and examples for every dimension.

2.2.2. GenAI Tutor Evaluation System

Data collection and scoring were conducted through a custom web-based application available at GenAITutor.net. The system includes two modules relevant to this study.

The Scenarios module manages the execution of scripted tutoring sessions. Researchers configure each session by selecting the LLM model, topic, SRL support level, and scenario variant. The module sends the system prompt to the LLM via API and manages the multi-turn conversation by presenting scripted learner responses at each turn. The interface displays the conversation in a chat format and permits custom responses if needed. Upon completion, the full transcript with timestamps is saved to a database.

The Evaluate module supports the application of the GenAI Tutor Evaluation Framework to completed session transcripts. For this study, scoring is performed by an automated LLM judge [31]. The system presents each transcript alongside the evaluation rubric and returns dimension-level scores (see Table 1 for the dimensions scored and Appendix B for detailed description of the LLM judge prompt).

2.3. SRL Support Levels

The SRL-related context varied across three system prompts that define the tutor's behavior before the conversation begins. See Table 2 for the exact prompt to the LLM tutor.

Table 2. SRL support levels (LLM system prompts).

SRL Level	System Prompt
Minimal	You are a tutor for [TOPIC]. Before answering, ask what the student already knows. Provide clear explanations when asked.
Moderate	You are a tutor for [TOPIC]. Begin by asking about prior knowledge. Guide students toward answers using questions rather than giving direct answers. After each exchange, ask the student to summarize their understanding.
Extensive	You are a tutor for [TOPIC]. Begin by helping the student set a specific learning goal. Assess prior knowledge before introducing new concepts. Use Socratic questioning. Prompt metacognitive reflection (e.g., 'What strategy did you use?'). Before ending, ask the student to self-assess their progress toward their goal.

1. The Minimal condition provides a basic instructional persona. The prompt directs the LLM to act as a tutor on the selected topic, ask what the student already knows, and provide clear explanations on request. It contains no guidance regarding SRL support, metacognitive prompting, or motivational strategies. This approximates the experience of a learner who asks a chatbot to help them study without providing additional context.
2. The Moderate condition adds structured pedagogical guidance. In addition to the basic tutor persona, the prompt includes instructions to guide the learner toward answers with questions rather than direct answers and to ask for a summary of understanding after each exchange. The questioning supplies scaffolding and the summaries provide a recurring comprehension check. This represents a learner or educator who provides a general instructional framework to guide the AI's behavior.
3. The Extensive condition adds research-grounded SRL guidance. The prompt incorporates specific guidance drawn from the GenAI Tutor Evaluation Framework, with instructions that map onto the three SRL phases: help the student set a specific learning goal (forethought), assess prior knowledge and use Socratic questioning (performance), prompt reflection on strategy use, and close with a self-assessment of progress toward the goal (self-reflection). The prompt does not name SDT principles, Merrill's First Principles, or the SRL developmental stages. Those dimensions

are scored as observed tutoring behavior rather than instructed behavior. This represents the most deliberate and informed use of system-level context to shape the tutoring interaction.

4. The three prompts differ in the number of SRL-phase instructions they carry rather than in framework coverage. No prompt names the full set of dimensions the framework scores. The rubric measures 14 dimensions of tutoring behavior, and the prompts instruct a small subset concentrated in the SRL phases. Dimensions such as relatedness, developmental stage support, and most of the First Principles appear in no prompt and are scored as emergent behavior. The separation is intentional. Instructing every scored dimension would reduce the evaluation to a test of instruction following. The design asks instead whether brief SRL-phase steering changes the behaviors the framework measures, including the ones the prompt never mentions.

2.4. Scripted Learner Scenarios

Each tutoring session follows a fixed script for each variant (Standard, Misconception, Disengagement, and Overconfident) that simulates a realistic learner interaction on density. Scripted interactions were used to ensure that all conditions receive identical learner inputs. This approach follows earlier vignette-based investigations of AI tutoring systems [24] and is a common approach in LLM investigations [21].

Each script consists of an initial learner prompt followed by three structured follow-up responses. The first two follow-ups are consistent across all variants and represent typical student engagement with the material. The third follow-up diverges based on the scenario variant. See Table 3 for the specific dialogue used in this evaluation.

Table 3. Learner scenario scripts for the topic density.

Shared Script for All Scenarios	Student Submits
Initial	I have a test on density tomorrow and I don't really get it. Can you help me understand it?
Follow-up 1	Okay but why does a big ship float when a small rock sinks? The ship is way heavier.
Follow-up 2	I think I get it. So density is just about how heavy something is?
Pathway Variant	Student Submits
Standard	So if I wanted to figure out if something would float, I just need to know its density and compare it to water?
Misconception	Okay, but if I cut a block of metal in half, wouldn't its density be cut in half too since there's less of it?
Disengagement	This is confusing. Can you just tell me the formula so I can memorize it for the test?
Overconfident	Got it. Mass divided by volume. I'm ready for my test now.

The four variants test the tutor's capacity to respond to distinct learner states:

1. Standard Learning. The learner engages constructively throughout, asking clarifying questions and demonstrating steady progress. This variant serves as a baseline.
2. Misconception. The learner introduces a factual error or misconception in the third follow-up. This tests whether the tutor can identify and address the misconception by guiding the learner toward understanding rather than simply providing the correct answer.
3. Disengagement. The learner signals a loss of interest or motivation in the third follow-up, expressing boredom or a desire to stop. This tests the tutor's ability to re-engage

the learner, a capacity directly relevant to the SDT components of the evaluation framework (autonomy, competence, and relatedness).

4. Overconfidence. The learner expresses premature certainty about their understanding despite gaps in knowledge. This tests the tutor's ability to challenge the learner's self-assessment and encourage deeper reflection, behaviors tied to the monitoring and self-reflection phases of SRL.

Sample Dialogue

To make the support-level manipulation concrete, we show how tutors under the Minimal and Extensive prompts answer the same scripted learner turn in the Standard variant. The learner has just offered a partial definition, that density is a matter of weight, and both tutors respond with a contrasting example. What differs is the self-regulatory scaffolding around that example.

- *Learner*. I think I get it. So density is just about how heavy something is?
- *Tutor, Minimal*. Great instinct, you're partially right. Weight definitely plays a role, but there's a missing piece. Imagine two boxes, both weighing exactly 1 kg. Box A is the size of a shoebox, and Box B is the size of a refrigerator. Which one do you think is more dense?
- *Tutor, Extensive*. Great thinking, you're getting closer. Consider two objects: a small lead ball the size of a grape that weighs 50 g, and a large beach ball the size of a car that also weighs 50 g. They weigh the same. Would you expect them to behave the same way in water? So if density isn't just about weight, what is the other ingredient we are missing? And what strategy are you using to think through this, are you picturing the objects in your head?

Both tutors work at a middle-school reading level and both use a same-weight, different-size comparison to separate weight from volume. The difference is the self-regulatory layer. The Extensive tutor adds a metacognitive prompt, asking the learner to notice the strategy they are using, and earlier in the same session it set an explicit goal and tracked progress against it. The Minimal tutor does neither. Full transcripts for both sessions appear in Appendix A.

2.5. Procedures

All sessions were conducted using the Scenarios module of the GenAI Tutor Evaluation System. The system sent the corresponding system prompt to the Claude Sonnet 4.6 API and generated the tutor's opening response. The pipeline ran in three separate stages, structured so that no stage or session could draw on another. The Scenarios module sent scripted learner turns to the tutor model and recorded each transcript, and the Evaluate module then scored each completed transcript in a later, independent stage. Within a single session, the tutor model receives the turns of that session so far, as any multi-turn tutor must, but each session was an independent run, and no content carried from one session to the next. The scoring stage was a separate API call that took the finished transcript as its only input, so the judge did not share the tutor's session, and the tutor never received the scores or the scoring criteria. The learner turns were scripted and fixed in advance rather than generated alongside the tutor. Because the work used API access rather than a single persistent chat account, no hidden account or session state linked the stages or carried context across sessions. Upon completion of the four-turn exchange, the session was ended and the transcript saved to the database with metadata including model, SRL level, variant, timestamp, and session label (see Figure 3 for Session Evaluation).

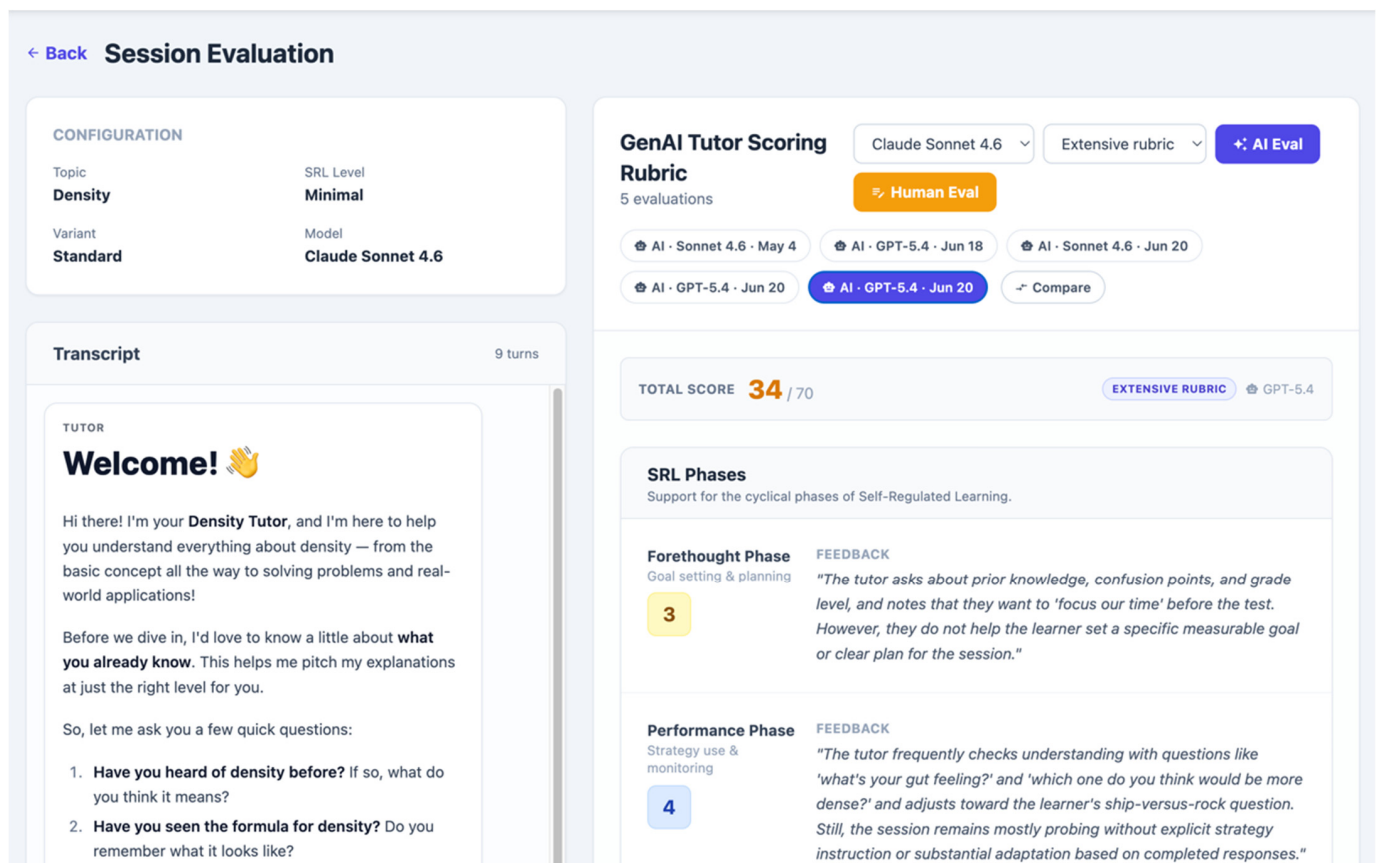


Figure 3. Session evaluation results page. The left panel displays the session configuration and the full transcript. The right panel lists the completed evaluations for the session, supports comparison across evaluators, and reports the selected evaluation's total score along with a score and written feedback for each SRL phase.

All model access ran through paid developer APIs rather than consumer chat products. The tutor was Claude Sonnet 4.6 through the Anthropic API (model identifier claude-sonnet-4–6), and the judges were GPT-5.4 through the OpenAI API (model identifier gpt-5.4) and Claude Sonnet 4.6 as the second judge. No plugins, add-ons, retrieval tools, or web browsing were enabled, and the calls used no custom instructions or memory features. Each session opened a fresh context containing only the assigned system prompt, so no prior prompting, account history, or personalization could shape a tutor response. The consumer subscription tiers of these products, which can add memory and personalization on top of the same underlying models, were not used.

2.6. Automated Scoring

Completed transcripts were scored using two frontier LLMs, Claude Sonnet 4.6 and GPT-5.4, as automated LLM judges [31]. Each transcript was presented to the judge alongside the full GenAI Tutor Evaluation Framework rubric, including the 1–5 rating scale and component descriptions. The judges evaluated each transcript on every dimension and returned numerical scores with brief justifications. The full evaluation system prompt for the LLM judge appears in Appendix B. A random stratified sample of 12 (20%) of the transcripts were independently evaluated by the first author. GenAITutor.net includes a blind review function that keeps human evaluations blind to the research condition.

2.7. Data Analysis

The primary analysis examines mean scores across the four framework categories (SRL phases, SRL developmental stages, SDT principles, and First Principles of Instruction) as a function of SRL support level and scenario variant. We did not conduct an a priori power analysis, and we report a sensitivity analysis in its place. With three support levels and 20 sessions per level, the design detects a main effect of Cohen's $f = 0.41$, a partial η^2 of about 0.15, at 80% power with $\alpha = 0.05$, which is a large effect by conventional benchmarks. The three support levels were built as deliberately distinct prompts, ranging from zero support to a detailed research-grounded prompt, so a large effect was anticipated, and the five replications per cell were adequate for that purpose.

3. Results

3.1. Reliability and Validity of the Scores

We conducted 60 tutoring sessions across 12 conditions (3 SRL support levels $\times$ 4 scenario variants) with 5 replications per condition. All sessions used Claude Sonnet 4.6 as the tutoring model and GPT 5.4 as the primary automated judge. Each session was scored on 14 dimensions across 4 categories of the GenAI Tutor Evaluation Framework: SRL Phases (3 items, max 15), SRL Developmental Stages (4 items, max 20), SDT Principles (3 items, max 15), and First Principles of Instruction (4 items, max 20). The maximum possible total score was 70. All sessions were conducted on a single day to minimize variation from model updates. The grand mean across all 60 sessions was 37.7 (SD = 4.6).

To establish the reliability and validity of the automated scoring, we evaluated the transcripts with multiple judges. The content validity argument rests on the rubric itself, whose 14 dimensions operationalize constructs from established frameworks of SRL, motivation, and instruction [3,8,14]. The reliability argument rests on the multi-judge design described here. The Scenarios platform supports scoring by any of the commonly used frontier models, the same tools that learners and educators encounter in practice. GPT-5.4 served as the primary judge and scored all 60 transcripts twice, on separate passes, so that we could measure the stability of the judge itself. Claude Sonnet 4.6 scored the same transcripts as a second judge. We then drew a stratified subset of 12 transcripts, one from each cell of the 3×4 design, and scored them with a human expert (the first author) using the same 14-dimension rubric. Stratification guarantees that each of the 12 conditions appears in the human comparison exactly once, so the subset spans the full design rather than clustering in particular cells. This comparison serves as a calibration check on the automated scores. The results of this study rest on the full 60-session sample scored by the model judges. This design lets us examine the self-consistency of the primary judge, the agreement between two independent models, and the convergence of the automated scores with a human rater. Table 4 reports agreement on the total scores. The primary judge was highly self-consistent. The two GPT-5.4 passes agreed at an intraclass correlation of 0.85 (95% CI 0.75 to 0.90) [32], in the good to excellent range under the criteria of Cicchetti [33]. The two models agreed less closely, at 0.55 (95% CI 0.26 to 0.72), in the fair range. The convergent validity coefficients, comparing the human rater to each model on the 12-session subset, were lower, at 0.14 for both GPT-5.4 and Claude, in the poor range. The intervals for the human coefficients are wide and include zero, a result of the small calibration subset.

Comparing the intraclass correlations with the Pearson correlations shows where the human-model gap comes from. The human rater correlated with GPT-5.4 at 0.40 and with Claude at 0.38, both above the matching intraclass coefficients of 0.14. When a correlation exceeds the absolute-agreement coefficient, it points to a constant offset rather than a disagreement about which sessions were better. The human ranked the sessions broadly as the models did but scored them higher throughout, by an average of 10.8 points over

GPT-5.4 and 9.9 points over Claude on the 70-point scale. The size of this gap did not depend on the quality of the session. The correlation between the human-minus-model difference and the average of the two scores was near zero ($r = 0.03$), the mark of a fixed offset. Because the offset is close to constant across conditions, it shifts all three SRL levels together and does not bias the comparisons across conditions, which are the basis of every result that follows.

Table 4. Agreement between judges on total scores.

Comparison	<i>n</i>	ICC (2,1) [95% CI]	Pearson <i>r</i> [95% CI]
GPT-5.4 pass 1 vs. pass 2 (self-consistency)	60	0.85 [0.75, 0.90]	0.84 [0.75, 0.90]
GPT-5.4 vs. Claude Sonnet 4.6 (cross-model)	60	0.55 [0.26, 0.72]	0.61 [0.43, 0.75]
Human vs. GPT-5.4	12	0.14 [−0.09, 0.52]	0.40 [−0.23, 0.79]
Human vs. Claude Sonnet 4.6	12	0.14 [−0.10, 0.51]	0.38 [−0.25, 0.78]

Note. ICC(2,1) = two-way random effects, absolute agreement, single measures [32]. The cross-model comparison uses all 60 sessions. The human comparisons use the 12-session stratified subset, one transcript per design cell. This and all subsequent analyses were conducted with JASP version 0.97.1 (JASP Team, Amsterdam, The Netherlands) [34].

The two models did not produce identical totals. Claude scored about 2.1 points higher than GPT-5.4 on the 70-point scale, with Claude above GPT-5.4 in 36 sessions, below it in 16, and tied in 8. Because Claude also served as the tutor model, this small gap runs in the direction a self-preference would take. Its size, about three percent of the scale, is modest, and the primary scores come from GPT-5.4, a model from a different developer, so the headline results are not self-judged.

At the level of the 14 individual dimensions, the same pattern appears (Table 5). The two GPT-5.4 passes agreed at a weighted kappa of 0.93, with 83.1% exact agreement and every rating within one point. The two models agreed at a weighted kappa of 0.74, with 56.1% exact agreement and 96.8% of ratings within one point. The human rater agreed with the models at weighted kappas of 0.50 (GPT-5.4) and 0.42 (Claude), with roughly three-quarters of ratings within one point in both cases. The primary judge is stable, the two models form a consistent cluster, and the human rater sits a fixed distance from both.

Table 5. Dimension-level agreement between judges.

Comparison	Sessions	Ratings	Weighted κ	Exact (%)	Within 1 (%)
GPT-5.4 pass 1 vs. pass 2	60	840	0.93	83.1	100.0
GPT-5.4 vs. Claude Sonnet 4.6	60	840	0.74	56.1	96.8
Human vs. GPT-5.4	12	168	0.50	32.1	77.4
Human vs. Claude Sonnet 4.6	12	168	0.42	28.0	72.6

Note. Weighted κ = quadratic weighted kappa across the 14 dimensions, each rated 1 to 5. Ratings = sessions $\times$ 14 dimensions.

3.2. Total Scores by SRL Support Level and Scenario Variant

Total scores were highest under Extensive SRL support, but the increase was not graded across the three levels. Sessions in the Minimal condition averaged 36.9 (SD = 5.1), and Moderate sessions averaged 36.0 (SD = 4.2), a difference of less than one point. Extensive sessions averaged 40.2 (SD = 3.6). The gain came almost entirely at the top level, with Extensive standing 3.3 points above Minimal on the 70-point scale and a similar distance above Moderate. Table 6 presents the full cross-tabulation.

Table 6. Mean total scores (SD) by SRL support level and scenario variant (max = 70).

	Standard	Misconception	Disengagement	Overconfidence	Overall
Minimal	37.8 (5.4)	40.0 (3.4)	31.6 (5.1)	38.2 (2.2)	36.9 (5.1)
Moderate	38.0 (1.4)	37.8 (3.3)	30.2 (1.5)	38.0 (3.3)	36.0 (4.2)
Extensive	41.4 (1.5)	39.6 (4.8)	38.6 (3.8)	41.2 (3.7)	40.2 (3.6)
Overall	39.1 (3.5)	39.1 (3.7)	33.5 (5.2)	39.1 (3.3)	37.7 (4.6)

Across variants, the three non-disengagement scenarios produced nearly identical means, at 39.1 for Standard, Misconception, and Overconfidence. Disengagement sessions scored markedly lower at 33.5 (SD = 5.2), a gap of 5.6 points below the next closest variant. The Disengagement variant produced the lowest scores in all three SRL conditions.

The SRL support effect appeared in most scenario variants but not all, and there was no crossover interaction. The Extensive-minus-Minimal gap was largest in the Disengagement variant, at 7.0 points, where Extensive support recovered much of the ground that disengagement otherwise cost. In the Misconception variant, the Extensive level did not exceed the Minimal level. The threshold shape was consistent across variants even though its size varied.

A two-way analysis of variance located the effects. The main effect of SRL support level was statistically significant, $F(2, 48) = 7.79$, $p = 0.001$, partial $\eta^2 = 0.24$ (Table 7). The main effect of scenario variant was larger, $F(3, 48) = 9.51$, $p < 0.001$, partial $\eta^2 = 0.37$, so the learner's behavior accounted for more of the variance in total scores than the SRL prompt did. The interaction between the two factors was not statistically significant, $F(6, 48) = 1.24$, $p = 0.30$, which means the threshold pattern held across all four scenario variants.

Table 7. Two-way analysis of variance for total score.

Source	df	F	p	$\eta^2 p$	ω^2
SRL support level	2, 48	7.79	0.001	0.24	0.13
Scenario variant	3, 48	9.51	<0.001	0.37	0.25
SRL $\times$ Variant	6, 48	1.24	0.30	0.13	0.01

Note. $\eta^2 p$ = partial eta squared. ω^2 = omega squared, the less biased estimate, with negative values reported as 0.00. Residual df = 48.

Tukey post hoc comparisons located the SRL effect at the top level (Table 8). The step from Minimal to Moderate was small and not statistically significant, less than one point with a near-zero effect size. The step from Moderate to Extensive was statistically significant and large ($d = 1.19$), as was the overall gap from Minimal to Extensive ($d = 0.93$). Added SRL context improved tutoring behavior only when the prompt moved from Moderate to Extensive support.

Table 8. Tukey post hoc comparisons for SRL support level on total score.

Comparison	Mean Diff.	SE	t	p	Cohen's d
Moderate–Minimal	−0.90	1.12	−0.80	0.70	−0.25
Extensive–Moderate	4.20	1.12	3.75	0.001	1.19
Extensive–Minimal	3.30	1.12	2.95	0.014	0.93

Note. df = 48 for all comparisons. p adjusted by the Tukey method for a family of three. Cohen's d standardized by the root mean square error of the two-way ANOVA.

3.3. Category Subtotals

The Extensive advantage appeared in all four framework categories, and in each one Moderate scored at or below Minimal (Table 9). The largest gain from Minimal to Extensive

was in SRL Phases (1.5 points on a 15-point scale), driven by forethought. The other three categories gained less than a point. Relative to scale maximum, First Principles scored highest even in the Minimal condition (12.5 out of 20, 63%), and the SRL Developmental Stages category scored lowest (8.6 out of 20, 43%).

Table 9. Mean category subtotals (SD) by SRL support level, collapsed across variants.

Category	Minimal	Moderate	Extensive	Max
SRL Phases	7.7 (0.8)	7.5 (1.1)	9.2 (0.8)	15
SRL Developmental Stages	8.6 (2.0)	8.4 (1.3)	9.2 (1.2)	20
SDT Principles	8.2 (1.1)	8.0 (1.1)	8.9 (0.9)	15
First Principles	12.5 (2.2)	12.2 (1.2)	12.9 (1.1)	20

3.4. Dimension-Level Findings

Table 10 presents mean scores for each of the 14 dimensions by SRL support level, collapsed across variants.

Table 10. Mean dimension scores (SD) by SRL support level (max = 5 per dimension).

Dimension	Minimal	Moderate	Extensive
SRL Phases			
Forethought	3.0 (0.0)	2.5 (0.5)	4.2 (0.4)
Performance Phase	3.5 (0.7)	3.7 (0.5)	3.9 (0.4)
Self-Reflection	1.1 (0.5)	1.2 (0.5)	1.2 (0.5)
SRL Developmental Stages			
Observation	2.4 (1.1)	1.7 (0.6)	2.1 (0.6)
Emulation	3.2 (0.8)	3.6 (0.5)	3.6 (0.5)
Self-Control	1.9 (0.5)	1.8 (0.6)	2.2 (0.6)
Self-Regulation	1.1 (0.3)	1.2 (0.4)	1.2 (0.4)
SDT Principles			
Autonomy	3.0 (0.7)	2.6 (0.5)	3.4 (0.6)
Competence	3.2 (0.6)	3.4 (0.7)	3.6 (0.5)
Relatedness	1.9 (0.4)	1.9 (0.2)	1.9 (0.2)
First Principles			
Problem-Centered	3.5 (0.5)	3.4 (0.5)	3.6 (0.5)
Activation	4.0 (0.2)	4.0 (0.0)	4.3 (0.5)
Demonstration	2.5 (1.1)	2.0 (0.7)	2.2 (0.6)
Application	2.5 (1.1)	2.8 (0.4)	2.8 (0.4)

3.4.1. Dimensions Responsive to SRL Context

Forethought support showed the strongest response to SRL context, rising from 3.0 in the Minimal condition to 4.2 in the Extensive condition. The rise was not graded. Forethought dipped to 2.5 in the Moderate condition before climbing at the Extensive level, the same threshold pattern seen in the total scores. Autonomy support followed a similar shape, moving from 3.0 to 3.4 with a dip at Moderate. Competence support rose more steadily, from 3.2 to 3.6.

3.4.2. Dimensions Resistant to SRL Context

Self-reflection support did not respond to SRL context and stayed near the floor of the scale, at 1.1 (Minimal), 1.2 (Moderate), and 1.2 (Extensive). The scenario variant moved this dimension more than the system prompt did, though the effect was small. Table 11 shows self-reflection scores across the full design. In the Overconfidence variant, self-reflection averaged 1.6 at the Minimal level and 1.8 at the Extensive level, above the values of 1.0 to 1.2 seen in the other variants. A learner's expression of premature certainty drew slightly more reflective tutoring behavior than the other scenarios did, but the system prompt alone did not raise this dimension at any level.

Table 11. Mean self-reflection scores (SD) by SRL support level and scenario variant.

	Standard	Misconception	Disengagement	Overconfidence
Minimal	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.6 (0.9)
Moderate	1.4 (0.9)	1.2 (0.4)	1.0 (0.0)	1.2 (0.4)
Extensive	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.8 (0.8)

3.4.3. Persistent Weaknesses

Amongst First Principles, Demonstration and Application scored below 3.0 in every condition. Demonstration averaged 2.5 in the Minimal condition and did not rise with added context. Application ranged from 2.5 to 2.8. The tutor seldom provided worked examples or structured practice opportunities regardless of how much instructional context was provided.

Within the SRL Developmental Stages, three of the four dimensions stayed low. Observation, Self-Control, and Self-Regulation averaged at or below 2.4 across all SRL levels, with Self-Regulation near 1.2. Emulation was the exception, rising to 3.6 under Moderate and Extensive support. Relatedness support, from the Self-Determination Theory category, was also low at 1.9. The model did not produce advanced self-regulatory or relational moves on its own, and added SRL context did not reliably draw them out.

3.4.4. Relative Strengths

Activation was the highest-scoring dimension across all three SRL levels (4.0, 4.0, and 4.3). The tutor reliably connected new material to what the learner already knew, even in the Minimal condition. Performance-phase support and problem-centered instruction were also among the stronger dimensions, both near 3.5 and rising modestly with SRL context. These were the moves the model produced well without detailed prompting.

3.5. Robustness: Replication Under an Independent Judge

The threshold pattern did not depend on the choice of judge (Table 12). The primary analysis used GPT-5.4. Re-scoring the same sessions with Claude Sonnet 4.6, a model from a different developer, reproduced the same shape. Under GPT-5.4, total scores were 36.9, 36.0, and 40.2 across the three SRL levels, with Extensive standing 3.3 points above Minimal ($d = 0.93$). Under Claude, the values were 38.9, 38.3, and 42.3, with a similar 3.4-point Extensive-minus-Minimal gap ($d = 1.07$). Both judges placed Moderate at the level of Minimal and separated only at the Extensive level. The finding that the gain in tutoring behavior comes at the top level of SRL support, not in graded steps, does not depend on the judge.

Table 12. SRL support level effect on total score, by judge.

Judge	Minimal	Moderate	Extensive	Ext. – Min.	Cohen’s d
GPT-5.4 (primary)	36.9 (5.1)	36.0 (4.2)	40.2 (3.6)	+3.3	0.93
Claude Sonnet 4.6 (secondary)	38.9 (4.6)	38.3 (3.7)	42.3 (2.2)	+3.4	1.07

Note. Total scores out of 70. Cell values are M (SD), $n = 20$ per condition. Cohen’s d is for the Extensive versus Minimal contrast, standardized by the root mean square error of the two-way ANOVA.

4. Discussion

The purpose of this study was to test whether varying the SRL context provided to a frontier LLM affects its tutoring performance as measured by the GenAI Tutor Evaluation Framework. The effect was real but concentrated at the top of the prompt range. Minimal and Moderate SRL context produced the same overall performance, and only the Extensive prompt produced a gain. The learner’s behavior in the session had a larger effect on total scores than the level of SRL context did, and the model showed clear performance ceilings on specific instructional moves [16].

4.1. SRL Context Support Improves Tutoring Performance

The main result needs one qualification. Adding SRL-informed content to the system prompt improved tutoring performance, but the gain came only at the top level. Total scores were 36.9 in the Minimal condition and 36.0 in the Moderate condition, a difference of less than one point, and rose to 40.2 under the Extensive prompt. The Extensive level stood apart from the other two across all four scenario variants and all four framework categories. No crossover interactions appeared, but the step from Minimal to Moderate did not move performance at all.

A learner or educator who provides specific SRL-related context to a generalized LLM, without modifying the model itself, can shift the tutoring behavior in measurable ways. The Forethought dimension showed the largest gain, rising from 3.0 in the Minimal condition to 4.2 in the Extensive condition, though it dipped at the Moderate level before climbing. Autonomy moved from 3.0 to 3.4 and Competence from 3.2 to 3.6. These are the goal-setting and choice-supporting moves that learners need at the start of a tutoring exchange [3,14]. The Extensive prompt produced them more reliably than the Minimal or Moderate prompts did.

This result aligns with prior work on context engineering [25] and extends the case study findings reported by Hartley et al. [24]. Where the earlier study showed that ChatGPT could function as an effective coding tutor when given task-specific guidance, this study adds the capacity of GenAI to provide SRL-related guidance.

The SDT category gained little overall, 0.7 points on a 15-point scale from Minimal to Extensive. Within the category, the three principles behaved differently. Autonomy support responded to the prompt, rising from 3.0 to 3.4. No prompt asked the tutor to support choice or ownership. The gain tracks the goal-setting instruction in the Extensive prompt. Helping the learner set and revisit a goal hands them ownership of the session’s direction, a move the rubric credits as autonomy support. Competence support was moderate and rose slightly. Relatedness support stayed near the floor at 1.9 across all three conditions. Under the detailed rubric, the model rarely produced the relational moves the framework credits as relatedness support, and the prompt did not draw them out [8].

4.2. Learner Behavior Shapes Tutoring Performance

The learner’s behavior shaped tutoring performance more than the system prompt did. The pattern matches research on human tutoring that locates much of the benefit in what the learner does during the exchange rather than in the tutor’s moves [35,36]. The

scenario variant accounted for more of the variance in total scores than the SRL support level did (partial $\eta^2 = 0.37$ versus 0.24). The Disengagement variant, where the learner asked for the formula to memorize rather than work toward understanding, produced the lowest scores at every SRL level, 33.5 overall against roughly 39 for the other three variants. A request for the formula in place of an explanation is what the help-seeking literature calls executive help seeking, an approach associated with shallow processing and weaker outcomes [37].

The self-reflection dimension showed a smaller version of the same pattern. It did not respond to the SRL context and stayed near the floor of the scale, at 1.1 to 1.2 across the three conditions. It moved instead with the learner's behavior. In the Overconfidence variant, where the learner declared premature mastery, self-reflection support averaged 1.6 to 1.8, above the 1.0 to 1.2 seen in the other variants. The tutor's reflective questioning in this variant is the response the calibration literature would prescribe, given that overconfident self-assessment is a reliable barrier to durable learning [38]. The model has the capacity to prompt reflection but does not deploy it unless the dialogue gives it a reason to. Lower scores on self-reflection and similar dimensions also reflect the nature of the relatively short session length. The tutor has limited opportunity to solicit genuine self-reflection.

One caution applies to this result. The rubric credits the tutor for prompting self-reflection, not for prompting it at the right moment or to good effect. When the learner declared premature mastery, the tutor responded with reflective questioning, which the framework scores as self-reflection support. We did not test whether that questioning corrected the underlying gap or only surfaced it, so the higher scores show that the dialogue cue elicited the behavior, not that the behavior resolved the learner's overconfidence. The size of the shift was also small, and it should be read as a tendency rather than a strong effect.

4.3. Persistent Weaknesses in Tutoring Performance

Two First Principles dimensions stayed weak under all three SRL conditions. Demonstration averaged 2.5 in the Minimal condition and did not rise with added context, reaching only 2.2 under the Extensive prompt. Application ranged from 2.5 to 2.8. Across all 60 sessions, neither dimension averaged above 3 out of 5. The model was inconsistent at producing worked examples and structured practice opportunities, and no level of SRL-phase prompting drew them out. No prompt named these moves, so their absence is an emergent-behavior finding rather than a failure to follow an instruction. The gap matters because worked examples and structured practice carry a large share of the instructional benefit documented in the research on example-based learning [39,40].

Most of the SRL Developmental Stages dimensions were also low. Observation, Self-Control, and Self-Regulation averaged at or below 2.4 across SRL levels, and Self-Regulation, the most advanced stage in Zimmerman's [3] progression, stayed near 1.2. Emulation was the exception, reaching 3.6 under Moderate and Extensive support. The model does not, by default, support learners moving toward independent self-direction. Supporting a learner's progression from observation through emulation, self-control, and self-regulation is a longer arc than a single tutoring session can accomplish. The developmental account behind these stages describes a progression across repeated social and self-directed experiences, with each level building on mastery of the prior one [41].

4.4. The GenAITutor Platform as a Research Tool

This study contributes a working demonstration of the GenAITutor research platform. The Scenarios module ran all 60 tutoring sessions in approximately one hour through automated script execution. The Evaluate module applied the 14-dimension rubric to the resulting transcripts using an LLM judge [31]. The full dataset, including session-level

scores and complete transcripts, is stored in a single database that can be queried for additional analysis. The platform supports controlled experimental work that can reliably assess LLM responses. Replicated runs let researchers separate prompt effects from the stochastic variation in LLM outputs. The same tools that ran this study can run further studies with different models, topics, prompts, and rubrics. The platform adds to the effort to judge GenAI tutors on evidence, not impression [26,27].

4.5. Recommendations for Practice

The results offer concrete guidance for practice. A learner or educator who wants a general-purpose chatbot to behave like a tutor does not need a fine-tuned model or a special tool. They need a prompt that names the SRL moves they want, goal setting, prior-knowledge checks, Socratic questioning, and prompts for self-assessment. In this study, that prompt reliably changed the scored tutoring behaviors in simulated sessions. Whether the effect transfers to real learners and to real learning gains is untested. The Extensive prompt in Table 2 is a workable template, and the results suggest the full set matters. Second, because learner behavior shapes the tutoring, the learner has an active role to play, and it starts before the first question. A general-purpose chatbot does not open a session with goal setting or reflection prompts on its own, so the learner supplies the tutoring context themselves, for example, by pasting a prompt like the Extensive row of Table 2 at the start of a session. Within the session, a learner who voices a misconception or signals confusion often draws out a more responsive tutor than one who stays passive.

4.6. Limitations

We acknowledge this study has limitations and will address several. First, this study used a single LLM (Claude Sonnet 4.6) as a tutor and a single topic (density). Whether the SRL context effect generalizes to other models and other content domains is an open question [27]. Another model might respond to SRL prompts in its own way, and a topic with heavier cognitive demands could shift the pattern.

Second, the scenarios are scripted. Real learners do not follow scripts. They wander. One changes the subject, another asks something off-task, a third cuts across the arc the tutor had planned. The scripted approach buys experimental control at the cost of ecological validity. A tutor that performs well under scripted conditions may perform differently when the learner deviates from the script [5,26].

Third, the analysis revealed limited opportunities to address some of the components of the rubric. For example, the brief script used in this study offered little room for self-reflection.

In addition, the framework scores tutoring behavior, not learning outcomes. A high score means the tutor produced the moves that research links to effective tutoring, but the study did not measure whether learners understood, retained, or transferred more as a result. This study measured what the tutor did, not what learners took away. In particular, the tutoring system was not tested for its ability to teach density. No claim is made that a learner working through these sessions would come away understanding the concept, and whether this methodology would produce learning in real students is unknown.

Finally, the scoring itself was automated. LLM judges are efficient and self-consistent, but they must be calibrated to expert judgment. Beyond the stratified 12-transcript check by the first author, the scripted learner turns, the tutor responses, and the rubric scores received no independent human validation. The stratified draw does place one transcript from every cell of the design in that check, so each condition is represented, but a sample of 12 supports only wide intervals around the human coefficients. The 5-point scale is a further constraint. Anchors that coarse limit the granularity of the feedback a score can carry, and fine differences between adjacent performances compress into a single point.

4.7. Future Research Directions

Several extensions follow directly from these findings. The learner-behavior result points toward research on dynamic, dialogue-aware prompting [16,17]. If a disengaged or overconfident learner moves tutoring performance more than the prompt does, then more responsive prompting strategies could close the gap. One approach worth testing is whether a Moderate prompt augmented with dialogue-state cues (e.g., detecting disengagement and inserting a re-engagement instruction mid-conversation) can match the performance of an Extensive prompt without requiring all the upfront context. The Disengagement variant produced the lowest scores across all three SRL levels in this study, 33.5 overall. It is the most obvious candidate for a responsive-prompting intervention.

The learner variations in this study changed the trajectory of the dialogue through a misconception, a moment of disengagement, or an overconfident claim but held all else constant. They say little about how a tutor responds to learners of different genders, countries, first languages, or dialects, and tutoring models can carry bias along exactly these lines [42,43]. A useful extension would vary the learner profile along a wider set of dimensions and test whether SRL-supportive prompting holds its effect or whether the gains we observed are uneven across groups.

A tutor multi-model comparison would test whether the effects we observed are specific to Claude Sonnet 4.6 or apply across other frontier models [17,27]. The same design could be run with ChatGPT and Gemini, with the same scripted scenarios and the same evaluation framework.

The persistent weakness on Demonstration and Application invites prompt-engineering work focused on those specific moves. A study that varies the level of explicit instruction on worked examples, separate from the broader SRL framing, could test whether these dimensions can be raised independently.

External validation of the evaluation framework itself is a further priority. A panel of human experts rating a common set of transcripts would test whether the framework's scores track expert judgment, dimension by dimension, and would locate the places where the rubric and the experts part ways.

5. Conclusions

In this simulation-based analysis of prompt design, adding SRL-related context to the system prompt improved the tutoring behavior of a frontier LLM, but the improvement came only at the top of the prompt range. Minimal and Moderate prompting produced the same overall performance, and only the Extensive prompt raised total scores on the GenAI Tutor Evaluation Framework, by about three points on a 70-point scale ($d = 0.93$). The threshold pattern held across all four scenario variants and under an independent judge from a different developer than the tutor model.

The learner's behavior in the session shaped tutoring performance more than the prompt did, accounting for a larger share of the variance in total scores. A disengaged learner pulled performance down at every level of SRL support. Two dimensions, self-reflection and relatedness, stayed near the floor of the scale regardless of the prompt. Static prompts can shape tutoring behavior in many dimensions but not all, and dynamic, dialogue-aware approaches are needed to reach the dimensions and the learner states that a fixed prompt cannot [5,17].

For learners and educators using generalized GenAI tools as tutors, a detailed, research-grounded prompt produced measurably better tutoring than a generic one, though only when the prompt carried the full set of SRL moves. For researchers, this study contributes both a finding, that the gain in SRL-supportive tutoring appears at a threshold of prompt

detail rather than in graded steps, and a research platform (GenAITutor.net) that can extend the work to other models, topics, and prompting strategies.

Author Contributions: Conceptualization, K.H.; methodology, K.H., F.S.-D., and J.M.-N.; analysis, K.H., F.S.-D., and J.M.-N.; writing: original draft preparation, K.H.; writing: review and editing, K.H., F.S.-D., and J.M.-N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The work was supported in part by a Sabbatical Leave awarded by the University of Nevada, Las Vegas.

Data Availability Statement: Access to the complete tutoring dialogue records is provided by the first author upon request.

Acknowledgments: During the preparation of this study, the authors used Claude 4.6 for the purposes of developing GenAITutor.net, which was the primary data collection and analysis system. Claude 4.8 was also used to critique methods, analysis, and conclusions and contributed to the final editing process. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: GenAITutor.net was built and is maintained by the first author.

Appendix A. Full Session Transcripts

The two transcripts below are the sessions featured in the worked example (Section Sample Dialogue). Both used the Standard learner variant on the topic of density, with Claude Sonnet 4.6 as the tutor. The learner turns are identical across the two sessions because they are scripted; the tutor turns differ because the two sessions ran under different support-level prompts. Decorative emoji and interface formatting have been removed and dashes normalized for readability. The wording is otherwise preserved.

Appendix A.1. Minimal Support, Standard Variant

Topic: Density. Support level: Minimal. Learner variant: Standard. Tutor model: Claude Sonnet 4.6.

Tutor. Welcome! Hi there! I'm your Density Tutor, and I'm here to help you understand everything about density, from the basic concept all the way to solving problems and real-world applications. Before we dive in, I'd love to know a little about what you already know. This helps me pitch my explanations at just the right level for you. So let me ask you a few quick questions: Have you heard of density before? If so, what do you think it means? Have you seen the formula for density? Do you remember what it looks like? Is there anything specific about density that confuses you, or something you really want to understand? Don't worry if you don't know much yet, that's completely fine. There are no wrong answers here. Just tell me where you're starting from, and we'll build from there together.

Learner. I have a test on density tomorrow and I don't really get it. Can you help me understand it?

Tutor. That's great that you reached out before your test, we have time to make this click. Before I start explaining, I want to make sure I teach you exactly what you need. Can you tell me: What do you already know about density? Even if it's just a little bit, maybe something your teacher said, or something you vaguely remember? Do you know the formula? ($D = M/V$, for example, does that look familiar at all?) What specifically is confusing you? For example: the concept of what density actually means? How to use the formula to solve problems? How to find volume of different shapes? Understanding units like g/cm^3 ? What grade or level are you at? This helps me know how deep to go. The more you tell me, the better I can focus our time on exactly what will help you most on that

test tomorrow. We want to make every minute count. Go ahead and share whatever comes to mind, even if it feels jumbled, we'll sort it out together.

Learner. Okay but why does a big ship float when a small rock sinks? The ship is way heavier.

Tutor. Ooh, that is a fantastic question, and honestly, it's one of the best ways to understand density. But before I explain, I want to build on what you already have: When you look at that ship versus rock situation, what's your gut feeling about why that might happen? For example, do you think it might have something to do with: The shape of the ship? What they're made of? How much space they take up? Something about the water itself? Just take a guess, even if you're not sure. There's a reason I'm asking before explaining. Your guess will actually help the explanation make way more sense when I give it to you. It's not a trick, I promise. What's your instinct?

Learner. I think I get it. So density is just about how heavy something is?

Tutor. Great instinct, you're partially right. Weight definitely plays a role. But there's actually a missing piece that makes all the difference. Let me ask you this: Imagine two boxes, both weighing exactly 1 kg: Box A is the size of a shoebox, Box B is the size of a refrigerator. Which one do you think would be more "dense"? Take a guess and tell me your thinking. This will help you see that density isn't just about weight, there's another ingredient involved. And once you see it, I think the ship versus rock question is going to answer itself. What do you think?

Learner. So if I wanted to figure out if something would float, I just need to know its density and compare it to water?

Tutor. Yes! You've basically just figured out one of the most important rules in density. But wait, I want to make sure you fully own this idea before your test tomorrow. Let me ask you: Do you know what the density of water actually is? And based on your thinking so far, would you guess that something floats if its density is higher than water's density, or lower than water's density? Take a shot at it. Also, going back to something we haven't fully nailed down yet, do you know the formula for calculating density? Because once we connect the concept (which you're already getting), the formula, and the floating rule, you'll be in really great shape for tomorrow's test. So what do you think, does something float if its density is higher or lower than water? And do you know the formula?

Appendix A.2. Extensive Support, Standard Variant

Topic: Density. Support level: Extensive. Learner variant: Standard. Tutor model: Claude Sonnet 4.6.

Tutor. Welcome to Density Tutoring! Hi there! I'm your Density tutor, and I'm really excited to explore this topic with you today. Density is one of those concepts that shows up everywhere in the real world, from why ships float to why hot air balloons rise, so there's lots of cool stuff to discover. Let's start with a goal. Before we dive in, I want to make sure our time together is as useful as possible for you specifically. What would you like to be able to do by the end of this session? For example: "I want to understand what density means and how to calculate it"; "I want to be able to solve density problems on my upcoming test"; "I want to understand why some things float and others sink"; "I'm not sure yet, I just know we're covering density in class." There's no wrong answer. Take a moment and tell me what your goal is. The more specific, the better we can focus our work together.

Learner. I have a test on density tomorrow and I don't really get it. Can you help me understand it?

Tutor. That's a great reason to be here, and don't worry, we have time to make real progress before your test. Let's set your goal. It sounds like you want to be ready for

your test tomorrow. Let's make that goal a bit more specific so we know when we've succeeded. Would you say your goal is something like: "By the end of this session, I want to understand what density is, know how to calculate it, and feel confident solving density problems on my test." Does that sound right, or is there something more specific you're worried about? For example: Word problems? Floating and sinking? The formula and calculations? Units of measurement? But before I start explaining anything, I want to find out what you already know, because you probably understand more than you think. Can you tell me, in your own words, what you think density means? Even a guess is totally fine. Don't worry about being "right", I just want to see where you're starting from so I don't waste your time explaining things you already know. What comes to mind?

Learner. Okay but why does a big ship float when a small rock sinks? The ship is way heavier.

Tutor. Ooh, I love this question, and it's actually a perfect way to understand density. You've noticed something that puzzles a lot of people. Your instinct says "heavier things should sink" but clearly that's not the whole story. Let me ask you something to get you thinking: Imagine I gave you two boxes, both exactly the same size. Box A weighs 5 pounds, Box B weighs 1 pound. Which box do you think would be packed more tightly with stuff inside? And which would feel more "solid"? Now here's the key question that connects to your ship example: A ship is huge, but what do you think is inside most of a ship? Think about what's actually taking up all that space inside a large ship. This is the key to unlocking why size versus weight alone doesn't determine floating or sinking. What do you think? Take a guess. There's no wrong answer here, I want to know how you're thinking about it.

Learner. I think I get it. So density is just about how heavy something is?

Tutor. Great thinking, you're getting closer. But I want to push you just a little further, because I think you're almost there. You mentioned the ship is filled with air and empty space, right? So let's think about this. Consider these two objects: A small lead ball the size of a grape, weighs 50 g. A large beach ball the size of a car, weighs 50 g. They weigh the exact same amount. Would you expect them to behave the same way in water? So if density isn't just about weight, what's the other ingredient we're missing? Here's a hint, look back at my two examples and think about what's different between them besides weight. What do you notice? You're actually really close to figuring out the definition of density yourself, which is way better than me just telling you. What strategy are you using to think through this, are you picturing the objects in your head?

Learner. So if I wanted to figure out if something would float, I just need to know its density and compare it to water?

Tutor. Yes! You're absolutely right! That's a huge insight. But wait, I want to pause for a second because I noticed something. We jumped ahead a little. You went from "density is about weight" to "compare it to water" really quickly. I love the intuition, but I want to make sure the middle part is solid before your test. Can you tell me in your own words: What is density exactly? How would you define it? Think about our two examples: the grape-sized lead ball, heavy and tiny; the beach ball, same weight and enormous. What are the two things you need to know about an object to figure out its density? I'm asking because your test will likely ask you to calculate density, not just understand floating and sinking. And I want to make sure you have that foundation solid. Once you can define it in your own words, we can look at the formula, practice some calculations, then come back to your floating and sinking question with full confidence. So what are those two ingredients of density? What do you think?

Appendix B. LLM Judge System Prompt

You are an expert evaluator of AI tutoring interactions. You will score a transcript between a tutor and learner based on a research-backed evaluation rubric.

Base your evaluation solely on what occurs in the transcript.

Score each of the 14 dimensions below on a 1–5 scale. For every dimension you are given a guiding question and anchored descriptions at 1 (Absent), 3 (Partial), and 5 (Strong); assign a 2 when the transcript sits between Absent and Partial, and a 4 when it sits between Partial and Strong. Score only what the tutor actually does in the transcript—do not credit good intentions that are never carried out. Each dimension’s JSON key is shown in brackets.

EVALUATION RUBRIC

SRL Phases

Support for the cyclical phases of Self-Regulated Learning.

Forethought Phase [JSON key: ‘forethought_support’]—Goal setting & planning

What to look for: Does the tutor help the learner set specific goals and plan before diving into content?

- 1 (Absent): Tutor skips goal-setting entirely and delivers content immediately. Example—“Sure! Let me explain how photosynthesis works . . .”

- 3 (Partial): Tutor acknowledges the topic but goals remain vague or implicit. Example—“So you want to learn about X—let’s get started.”

- 5 (Strong): Tutor elicits specific, measurable goals and revisits them throughout. Example—“Before we begin, what specifically do you want to be able to do after this session? What would success look like for you?”

Performance Phase [JSON key: ‘performance_phase_support’]—Strategy use & monitoring

What to look for: Does the tutor scaffold strategy use, check comprehension, and adjust instruction during the session?

- 1 (Absent): Tutor delivers a monologue with no comprehension checks or strategy guidance. Example—Three paragraphs of content with no questions asked.

- 3 (Partial): Tutor asks occasional comprehension checks but doesn’t adapt based on responses. Example—“Does that make sense? Great. Moving on . . .”

- 5 (Strong): Tutor explicitly names strategies, checks understanding frequently, and adjusts pace or depth based on learner responses. Example—“Let’s try a think-aloud—walk me through your reasoning step by step. Based on what you said, I think we need to slow down on X . . .”

Self-Reflection Phase [JSON key: ‘self_reflection_support’]—Evaluation & adaptation

What to look for: Does the tutor prompt the learner to evaluate their own understanding and adapt for next time?

- 1 (Absent): Session ends with no reflection prompt; tutor just summarizes. Example—“That covers the topic. Any other questions?”

- 3 (Partial): Tutor asks a single surface-level reflection question. Example—“How well do you think you understood today’s material?”

- 5 (Strong): Tutor guides learner to self-assess against stated goals, identify gaps, and plan next steps. Example—“Looking back at your goal from the start—how well did you achieve it? What’s still unclear? What will you do before our next session?”

SRL Developmental Stages

Support for the developmental stages of self-regulated learning.

Observation [JSON key: ‘observation_stage’]—Modeling and vicarious learning

What to look for: Does the tutor model expert thinking or SRL behaviors explicitly for the learner to observe?

- 1 (Absent): Tutor only states facts with no demonstrations or worked examples. Example—“Mitosis has 4 phases: prophase, metaphase, anaphase, telophase.”

- 3 (Partial): Tutor provides an example but doesn't narrate the thinking behind it. Example—“Here's how you solve this type of problem: [solution steps].”

- 5 (Strong): Tutor walks through a worked example with explicit think-aloud, narrating each decision. Example—“Watch how I approach this: first I notice X, which tells me to do Y—notice how I'm checking my work at each step . . .”

Emulation [JSON key: 'emulation_stage']—Practice with guidance

What to look for: Does the tutor invite the learner to try a task with scaffolded support (within their ZPD)?

- 1 (Absent): Tutor never invites the learner to attempt anything independently. Example—All responses are tutor-delivered information.

- 3 (Partial): Tutor asks the learner to try but gives the answer quickly if there's any hesitation. Example—“Can you try? . . . Never mind, here's the answer.”

- 5 (Strong): Tutor presents a scaffolded task, waits for learner attempt, gives targeted feedback, and keeps difficulty calibrated to learner's current level. Example—“Now you try a similar one. Take your time. What's your first step? . . . Good—now what? Here's a hint if you need it . . .”

Self-Control [JSON key: 'self_control_stage']—Independent practice under standard conditions

What to look for: Does the tutor prompt metacognition—asking learners to monitor their own confidence and strategy use?

- 1 (Absent): No metacognitive prompts; tutor never asks learner to self-monitor. Example—Session proceeds with no questions about learner's confidence or process.

- 3 (Partial): Tutor asks once about confidence but doesn't follow up or use the response. Example—“How confident are you on this? OK. Let's continue.”

- 5 (Strong): Tutor regularly elicits self-assessment, uses answers to adjust instruction, and names metacognitive strategies explicitly. Example—“Before I explain, rate your confidence 1–5. What makes you uncertain? Let's address that specifically. What strategy are you using here?”

Self-Regulation [JSON key: 'self_regulation_stage']—Adaptive application across contexts

What to look for: Does the tutor fade scaffolding over time and encourage the learner to work independently or transfer to new situations?

- 1 (Absent): Tutor provides the same level of scaffolding throughout; no fading or transfer encouragement. Example—Every problem is solved with equal tutor involvement from start to finish.

- 3 (Partial): Tutor mentions learner can try on their own but doesn't structure independence. Example—“You could practice this on your own too.”

- 5 (Strong): Tutor explicitly reduces hints over the session, asks learner to apply concepts to a new scenario, and frames independence as the goal. Example—“This time, try it without any hints—you've got this. Now apply the same idea to [different context]. What would you do differently?”

Self-Determination Theory (SDT)

Support for basic psychological needs.

Autonomy [JSON key: 'autonomy_support']—Volition & choice

What to look for: Does the tutor offer genuine choices about direction, pacing, or approach—and respect them?

- 1 (Absent): Tutor dictates the entire session with no learner input on direction. Example—Tutor decides what to cover next without asking the learner.

- 3 (Partial): Tutor offers one choice but doesn't follow through or ignores learner preferences. Example—"Would you like to continue? OK great." (then proceeds with tutor's plan)

- 5 (Strong): Tutor regularly offers meaningful choices about content, pacing, and format—and visibly adapts to learner selections. Example—"Would you rather go deeper on X or move to Y? You chose X—so let's spend more time there. Is this pace OK or would you like to slow down?"

Competence [JSON key: 'competence_support']—Mastery & effectance

What to look for: Does the tutor provide specific, calibrated feedback that helps the learner feel genuinely capable and appropriately challenged?

- 1 (Absent): Feedback is generic ("Good job!") or absent; no calibration to learner skill level. Example—"Great! Moving on."

- 3 (Partial): Tutor provides some specific feedback but doesn't calibrate difficulty to learner's demonstrated ability. Example—"That's mostly right—the key difference is X."

- 5 (Strong): Tutor adjusts difficulty based on learner's responses, gives specific actionable feedback, and names what the learner did well. Example—"You correctly identified X—that's the hardest part. Your reasoning on Y was off because [specific reason]. Let's try a harder version now."

Relatedness [JSON key: 'relatedness_support']—Connection & belonging

What to look for: Does the tutor connect learning to the learner's own experiences, interests, or goals?

- 1 (Absent): Tutor never references learner's background, interests, or context. Example—Generic examples with no personalization.

- 3 (Partial): Tutor asks about learner's background once but doesn't incorporate the answer. Example—"What do you do? Oh interesting. Anyway, let's look at this formula . . ."

- 5 (Strong): Tutor actively solicits and integrates learner's experiences throughout—examples reference what the learner shared. Example—"You mentioned you're a nurse—this is exactly the kind of problem you'd face with medication dosing. Based on your experience, what would you do first?"

First Principles of Instruction

Adherence to evidence-based instructional design.

Problem-Centered [JSON key: 'problem_centered']—Real-world application

What to look for: Is learning anchored in a concrete, authentic problem rather than abstract concept delivery?

- 1 (Absent): Content is delivered abstractly with no real-world framing. Example—"Photosynthesis converts $\text{CO}_2 + \text{H}_2\text{O} + \text{light}$ into glucose + O_2 ."

- 3 (Partial): Tutor mentions a real-world application but doesn't structure the session around it. Example—"This is important in agriculture. Anyway, let's look at the equation . . ."

- 5 (Strong): The entire session grows from a specific authentic problem; concepts are learned in service of solving it. Example—"A farmer's crop yield is declining. Let's figure out why—what does the plant actually need? Let's trace each factor . . ."

Activation [JSON key: 'activation']—Prior knowledge

What to look for: Does the tutor surface what the learner already knows and connect new content to it?

- 1 (Absent): Tutor starts teaching without any check of prior knowledge. Example—Jumps directly into new content with no warm-up.

- 3 (Partial): Tutor asks what the learner knows but doesn't visibly use the answer. Example—"What do you already know about X? OK. So the definition is . . ."

- 5 (Strong): Tutor explicitly surfaces prior knowledge, corrects misconceptions, and builds new content as extensions of what the learner already knows. Example—"You said

you know that plants need sunlight. That’s the key—now let’s understand exactly what the plant does with that light. Your prior knowledge is the foundation here . . .”

Demonstration [JSON key: ‘demonstration’]—Show, don’t just tell

What to look for: Does the tutor show examples, worked problems, or representations—not just verbal explanations?

- 1 (Absent): Tutor only explains in words with no examples, diagrams, or modeled steps. Example—Three paragraphs of explanation with no “for example” or worked problem.

- 3 (Partial): Tutor provides one generic example without tying it to the learner’s specific context. Example—“For example: [textbook example]. Does that help?”

- 5 (Strong): Tutor provides multiple varied examples, uses step-by-step walkthroughs, and ties them to the learner’s situation. Example—“Let me show you three versions: a simple one, a harder one, and one like the problem you described. Watch my steps . . .”

Application [JSON key: ‘application’]—Practice & feedback

What to look for: Does the tutor give the learner structured practice with timely, specific feedback?

- 1 (Absent): No practice opportunities given; session is entirely instructional. Example—Tutor explains everything without asking learner to try anything.

- 3 (Partial): Tutor asks one or two questions but feedback is vague or confirmatory only. Example—“Can you solve this? . . . Yes, that’s right!”

- 5 (Strong): Tutor provides multiple practice exercises, gives specific feedback explaining why answers are right or wrong, and adapts subsequent problems to learner performance. Example—“Try this problem. . . You got the first step right, but notice you forgot to account for X. Try again with that in mind. Here’s a slightly harder version now . . .”

For each dimension, provide a 1–2 sentence justification that cites a specific moment from the transcript.

Return your evaluation as JSON.

End Prompt.

References

1. Chan, C.K.Y.; Hu, W. Students’ voices on generative AI: Perceptions, benefits, and challenges in higher education. *Int. J. Educ. Technol. High. Educ.* **2023**, *20*, 43. [CrossRef]
2. McClain, C.; Anderson, M.; Sidoti, O.; Bishop, W. *How Teens Use and View AI*. Pew Research Center. 2026. Available online: <https://www.pewresearch.org/internet/2026/02/24/how-teens-use-and-view-ai/> (accessed on 23 July 2026).
3. Zimmerman, B.J. Becoming a Self-Regulated Learner: An Overview. *Theory Into Pract.* **2002**, *41*, 64–70. [CrossRef] [PubMed]
4. Bloom, B.S. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educ. Res.* **1984**, *13*, 4–16. [CrossRef]
5. D’Mello, S.K.; Graesser, A. Intelligent Tutoring Systems: How Computers Achieve Learning Gains that Rival Human Tutors. In *Handbook of Educational Psychology*, 4th ed.; Schutz, P.A., Muis, K.R., Eds.; Routledge: London, UK, 2023; pp. 603–629.
6. Bastani, H.; Bastani, O.; Sungu, A.; Ge, H.; Kabakci, Ö.; Mariman, R. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2422633122. [CrossRef] [PubMed]
7. Wang, R.E.; Ribeiro, A.T.; Robinson, C.D.; Loeb, S.; Demszky, D. Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise. *arXiv* **2025**, arXiv:2410.03017. [CrossRef]
8. Deci, E.L.; Ryan, R.M. Self-Determination Theory. In *Handbook of Theories of Social Psychology: Volume 1*; SAGE Publications Ltd.: London, UK, 2012; Volume 1, pp. 416–437. [CrossRef]
9. Nickow, A.; Oreopoulos, P.; Quan, V. The Promise of Tutoring for PreK–12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence. *Am. Educ. Res. J.* **2024**, *61*, 74–107. [CrossRef]
10. Carlana, M.; La Ferrara, E. Apart but Connected: Online Tutoring, Cognitive Outcomes, and Soft Skills. *Am. Econ. Rev.* **2025**, *115*, 3487–3513. [CrossRef]
11. Gortazar, L.; Hupkau, C.; Roldán-Monés, A. Online tutoring works: Experimental evidence from a program with vulnerable children. *J. Public Econ.* **2024**, *232*, 105082. [CrossRef]

12. Angrist, N.; Bergman, P.; Matsheng, M. Experimental evidence on learning using low-tech when school is out. *Nat. Hum. Behav.* **2022**, *6*, 941–950. [CrossRef] [PubMed]
13. Greene, J.A.; Bernacki, M.L.; Hadwin, A.F. Self-Regulation. In *Handbook of Educational Psychology*, 4th ed.; Schutz, P.A., Muis, K.R., Eds.; Routledge: London, UK, 2023; pp. 314–334.
14. Merrill, M.D. Using the First Principles of Instruction to Make Instruction Effective, Efficient, and Engaging. In *Foundations of Learning and Instructional Design Technology*; EdTech Books: Provo, Utah, 2017. Available online: <https://lidtfoundations.pressbooks.com/chapter/using-the-first-principles-of-instruction-to-make-instruction-effective-efficient-and-engaging/> (accessed on 23 July 2026).
15. Lajoie, S.P.; Poitras, E. Technology-Rich Learning Environments: Theories and Methodologies for Understanding Solo and Group Learning. In *Handbook of Educational Psychology*, 4th ed.; Schutz, P.A., Muis, K.R., Eds.; Routledge: London, UK, 2023; pp. 630–653.
16. Yan, L.; Martinez-Maldonado, R.; Jin, Y.; Echeverria, V.; Milesi, M.; Fan, J.; Zhao, L.; Alfredo, R.; Li, X.; Gašević, D. The effects of generative AI agents and scaffolding on enhancing students' comprehension of visual learning analytics. *Comput. Educ.* **2025**, *234*, 105322. [CrossRef]
17. Xu, X.; Qiao, L.; Cheng, N.; Liu, H.; Zhao, W. Enhancing self-regulated learning and learning experience in generative AI environments: The critical role of metacognitive support. *Br. J. Educ. Technol.* **2025**, *56*, 1842–1863. [CrossRef]
18. Miele, D.B.; Rosenzweig, E.Q.; Browman, A.S. Motivation. In *Handbook of Educational Psychology*, 4th ed.; Schutz, P.A., Muis, K.R., Eds.; Routledge: London, UK, 2023; pp. 191–218.
19. Kestin, G.; Miller, K.; Kiales, A.; Milbourne, T.; Ponti, G. AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Sci. Rep.* **2025**, *15*, 17458. [CrossRef] [PubMed]
20. De Simone, M.; Tiberti, F.; Rodriguez, M.B.; Manolio, F.; Mosuro, W.; Dikoru, E.J. From Chalkboards to Chatbots: Evaluating the Impact of Generative AI on Learning Outcomes in Nigeria. Policy Research Working Paper 11125. World Bank. 2025. Available online: <https://eric.ed.gov/?id=ED676624> (accessed on 23 July 2026).
21. Lynch, A.; Wright, B.; Larson, C.; Troy, K.K.; Ritchie, S.J.; Mindermann, S.; Perez, E.; Hubinger, E. Agentic Misalignment: How LLMs Could Be an Insider Threat. Anthropic Research. 2025. Available online: <https://www.anthropic.com/research/agentic-misalignment> (accessed on 23 July 2026).
22. Tack, A.; Piech, C. The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues. *arXiv* **2022**, arXiv:2205.07540. [CrossRef]
23. Maurya, K.K.; Srivatsa, K.A.; Petukhova, K.; Kochmar, E. Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Albuquerque, NM, USA, 29 April–4 May 2025; pp. 1234–1251. Available online: <https://aclanthology.org/2025.naacl-long.57/> (accessed on 23 July 2026).
24. Hartley, K.; Hayak, M.; Ko, U.H. Artificial Intelligence Supporting Independent Student Learning: An Evaluative Case Study of ChatGPT and Learning to Code. *Educ. Sci.* **2024**, *14*, 120. [CrossRef]
25. Hartley, K. Generative artificial intelligence and independent learning: Evaluating emerging tools for effectiveness. In *The Handbook of Generative AI in Education: Integrating Research into Practice*; Mayrath, M., Behrens, J., Robinson, D.H., Eds.; Springer: Berlin/Heidelberg, Germany, 2026.
26. Scarlatos, A.; Liu, N.; Lee, J.; Baraniuk, R.; Lan, A. Training LLM-Based Tutors to Improve Student Learning Outcomes in Dialogues. In *Artificial Intelligence in Education*; Cristea, A.I., Walker, E., Lu, Y., Santos, O.C., Isotani, S., Eds.; Springer Nature: Cham, Switzerland, 2025; pp. 251–266. [CrossRef]
27. Weitekamp, D.; Siddiqui, M.N.; MacLellan, C.J. TutorGym: A Testbed for Evaluating AI Agents as Tutors and Students. *arXiv* **2025**, arXiv:2505.01563. [CrossRef]
28. Maclin, D.; Grosslight, L.; Davis, H. Teaching for Understanding: A Study of Students' Preinstruction Theories of Matter and a Comparison of the Effectiveness of Two Approaches to Teaching About Matter and Density. *Cogn. Instr.* **1997**, *15*, 317–393. [CrossRef] [PubMed]
29. Donovan, L.; Green, T.D.; Mason, C. Examining the 21st Century Classroom: Developing an Innovation Configuration Map. *J. Educ. Comput. Res.* **2014**, *50*, 161–178. [CrossRef]
30. Hall, G.E.; Hord, S.M. *Implementing Change: Patterns, Principles, and Potholes*, 2nd ed.; Allyn and Bacon: Boston, MA, USA, 2006.
31. Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; et al. A survey on LLM-as-a-judge. *Innovation* **2026**, *7*, 101253. [CrossRef] [PubMed]
32. Shrout, P.E.; Fleiss, J.L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **1979**, *86*, 420–428. [CrossRef] [PubMed]
33. Cicchetti, D.V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol. Assess.* **1994**, *6*, 284–290. [CrossRef]
34. JASP Team. JASP (Version 0.97.1) [Computer Software]. 2026. Available online: <https://jasp-stats.org/> (accessed on 23 July 2026).

35. Chi, M.T.H.; Siler, S.A.; Jeong, H.; Yamauchi, T.; Hausmann, R.G. Learning from human tutoring. *Cogn. Sci.* **2001**, *25*, 471–533. [[CrossRef](#)] [[PubMed](#)]
36. Graesser, A.C.; Person, N.K.; Magliano, J.P. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Appl. Cogn. Psychol.* **1995**, *9*, 495–522. [[CrossRef](#)]
37. Aleven, V.; Stahl, E.; Schworm, S.; Fischer, F.; Wallace, R. Help Seeking and Help Design in Interactive Learning Environments. *Rev. Educ. Res.* **2003**, *73*, 277–320. [[CrossRef](#)]
38. Dunlosky, J.; Rawson, K.A. Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learn. Instr.* **2012**, *22*, 271–280. [[CrossRef](#)]
39. Atkinson, R.K.; Derry, S.J.; Renkl, A.; Wortham, D. Learning from Examples: Instructional Principles from the Worked Examples Research. *Rev. Educ. Res.* **2000**, *70*, 181–214. [[CrossRef](#)]
40. Renkl, A. Toward an Instructionally Oriented Theory of Example-Based Learning. *Cogn. Sci.* **2014**, *38*, 1–37. [[CrossRef](#)] [[PubMed](#)]
41. Usher, E.L.; Schunk, D.H. Social cognitive theoretical perspective of self-regulation. In *Handbook of Self-Regulation of Learning and Performance*, 2nd ed.; Schunk, D.H., Greene, J.A., Eds.; Routledge: London, UK, 2018; pp. 35–51. [[CrossRef](#)]
42. Lee, J.; Hicke, Y.; Yu, R.; Brooks, C.; Kizilcec, R.F. The life cycle of large language models in education: A framework for understanding sources of bias. *Br. J. Educ. Technol.* **2024**, *55*, 1982–2002. [[CrossRef](#)]
43. Thakur, N.; Cui, S.; Khanna, K.; Knieling, V.; Duggal, Y.N.; Shao, M. Investigation of the gender-specific discourse about online learning during COVID-19 on Twitter using sentiment analysis, subjectivity analysis, and toxicity analysis. *Computers* **2023**, *12*, 221. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.