American Journal of Computer Science and Technology
2026, Vol. 9, No. 1, pp. 19–29
https://doi.org/10.11648/j.ajcst.20260901.13

Research Article

Self-explanation Prompts in STEM: Comparing Human and AI Metacognitive Accuracy

Panya Samtani*

Independent Researcher, Gurugram, India

Abstract

This study investigates the efficacy of Self-Explanation Prompts (SEPs) in enhancing problem-solving performance and metacognitive accuracy within STEM education, while simultaneously offering a comparative analysis of human versus artificial cognition. Grounded in the theoretical frameworks of Metacognition and Self-Regulated Learning (SRL), the research employs a quasi-experimental design with a diverse sample (N= 150, ages 10–50) divided into a SEP intervention group and a control group. Results indicate that structured reflective prompting significantly improves problem-solving accuracy and metacognitive calibration (Gamma correlation). Furthermore, the study contrasts human cognitive responses with those of three leading Large Language Models (ChatGPT, Perplexity, and Gemini). Findings reveal a fundamental divergence: while AI models excel at logical pattern matching, they lack the embodied, emotional, and contextual reasoning such as the intuitive understanding of physics or emotional pragmatics that characterises human thought. The study concludes that SEPs are essential for cultivating the self-aware, "adaptive expertise" that distinguishes human intelligence from algorithmic data processing.

Keywords

Metacognition, Self-regulated Learning, STEM Education, Artificial Intelligence, Human-AI Interaction, Embodied Cognition

1. Introduction

The 21st-century educational landscape is undergoing a paradigm shift. In the domain of Science, Technology, Engineering, and Mathematics (STEM), the objective has moved beyond the rote memorisation of formulas to the development of "adaptive expertise", the ability to apply knowledge flexibly to novel problems [1-4]. However, a significant barrier remains as many students struggle to monitor their own understanding, often suffering from "illusions of competence" where they believe they understand a concept when they have merely memorized a procedure.

Simultaneously, the rapid ascent of Artificial Intelligence (AI) challenges our definitions of learning and intelligence. AI systems can now solve complex STEM problems with high accuracy, raising the question: What is the unique value of human cognition? This research posits that the answer lies in Metacognition – the uniquely human capacity to reflect on, monitor, and regulate one's own thinking [5].

While Metacognition provides the "knowledge" of cognition, Self-Regulated Learning (SRL) describes the active "control" of that cognition in real-time. This study is grounded in Zimmerman's (2000) and Pintrich's (2004) cyclic models of SRL, which posit that learning is not a linear accumulation of facts but a recursive cycle of Forethought, Performance, and Self-Reflection [6]. In the context of STEM education, the

*Correspondence: Panya Samtani (panya.samtani123@gmail.com)

Received: 6 January 2026; **Accepted:** 23 January 2026; **Published:** 6 February 2026

"Performance" phase is critical. This is where monitoring the real-time assessment of one's understanding occurs. Effective learners constantly ask themselves: "Does this step make sense?" or "Am I closer to the solution?". This capability is termed adaptive expertise. Unlike rote learners who memorise procedures, adaptive experts can transfer their knowledge to novel problems because they recognise errors and re-evaluate their reasoning in real-time.

The landscape of Science, Technology, Engineering, and Mathematics (STEM) education is undergoing a profound transformation. In an era increasingly dominated by automation and algorithmic processing, the primary goal of human education is no longer the mere accumulation of static knowledge or the rote application of formulas. Instead, the objective has shifted toward cultivating adaptive expertise—the ability to apply knowledge flexibly to novel problems, diagnose errors in real-time, and transfer conceptual understanding across varying contexts.

However, a significant pedagogical barrier remains. Students in STEM disciplines often exhibit a "doing-understanding gap," where they can mechanically execute procedural steps (e.g., solving a quadratic equation) without comprehending the underlying conceptual logic. This superficial fluency often leads to metacognitive illusions, or "illusions of competence," where learners vastly overestimate their mastery of a subject. When a student believes they understand a concept but actually does not, they fail to engage in the necessary remedial study, leading to a cycle of underperformance and frustration. Thus, the critical challenge in modern STEM pedagogy is not just teaching students *what* to think, but teaching them *how to monitor* their own thinking.

Metacognition, broadly defined as "thinking about thinking," is the engine of deep learning. It comprises two distinct yet interrelated processes: monitoring (assessing the current state of learning) and control (regulating cognitive strategies to improve learning). While high-performing students naturally engage in these processes—constantly asking themselves, "Does this answer make sense?"—novice learners typically lack these self-regulatory habits.

In complex STEM domains, this deficit is particularly damaging. A student may correctly solve a physics problem by mimicking a memorized procedure but fail to recognize when that procedure is inappropriate for a slightly altered scenario. Without external intervention to trigger metacognitive monitoring, these students remain "passive accumulators" of facts rather than active constructors of knowledge. This study posits that the missing link in closing this gap is Self-Explanation (SE)—a cognitive strategy that compels the learner to articulate the *causal logic* behind their actions [7-9].

However, the implementation of SE in educational settings is nuanced. Research indicates that without guidance, students often generate superficial explanations or simply paraphrase the text. Furthermore, the cognitive load imposed by generating explanations can sometimes overwhelm working memory, potentially hindering learning rather than helping it. This study addresses these contradictions by investigating the efficacy of Structured Self-Explanation Prompts (SEPs). These are targeted cues designed to guide the learner's reflection without creating excessive cognitive burden, theoretically optimizing both problem-solving accuracy and metacognitive calibration.

The urgency of fostering human metacognition has been amplified by the rapid ascent of Generative Artificial Intelligence (AI). As of 2025, Large Language Models (LLMs) like ChatGPT and Gemini can solve complex STEM problems with remarkable speed and accuracy, seemingly replicating human reasoning. This technological leap forces a re-evaluation of the goals of education: If a machine can solve the problem, what is left for the human learner?

This research argues that the distinction lies in conscious reflection. While AI systems operate on statistical probability and pattern matching—processing data without awareness—human cognition is grounded in embodied experience, emotional regulation, and intent. An AI can state Newton's laws, but it does not "understand" force in the physical, embodied sense that a human does. Therefore, this study includes a novel comparative dimension: it contrasts the reflective, metacognitive reasoning of human participants with the algorithmic output of leading AI models. By doing so, it seeks to demonstrate that true problem-solving involves a depth of self-awareness and emotional grounding (e.g., curiosity, doubt) that remains uniquely human.

Grounded in the frameworks of Metacognition and Self-Regulated Learning (SRL), this study aims to:

1) To empirically validate the impact of SEPs on problem-solving performance in STEM.
2) To measure the effect of SEPs on metacognitive accuracy (calibration), determining if reflection helps students better judge their own competence.
3) To compare human cognitive processes with AI models to highlight the non-computational aspects of learning, such as emotional regulation and embodied intuition.

By integrating educational psychology with the emerging philosophy of AI, this research offers a comprehensive view of learning that is both timely and timeless. It advocates for a pedagogy that prioritizes the "human dimension" of thought – reflection, awareness, and meaning-making – as the essential counterbalance to artificial intelligence.

2. Theoretical Framework

2.1. Metacognition: The Foundation of Reflective Thought

Metacognition, as originally conceptualized by Flavell [10], encompasses an individual's awareness and regulation of their cognitive processes. It is not a singular entity but a dual-process system comprising:

1) Metacognitive Knowledge: The understanding of one's

own cognitive abilities and the strategies available for learning.

2) Metacognitive Regulation: The active control of learning through planning, monitoring, and evaluating [11].

In the context of STEM, these elements allow learners to transition from passive knowledge acquisition to active, strategic engagement [12]. For instance, when a student encounters an error in a physics calculation, metacognitive regulation allows them to pause, re-evaluate their strategy, and self-correct. This conscious "debugging" of one's own thought process is what defines deep learning. In contrast, while AI can simulate reasoning, it lacks this self-awareness – the conscious reflection that defines human thought [13].

2.2. Self-regulated Learning (SRL) and Monitoring

The study is further grounded in Self-Regulated Learning (SRL), which operationalizes metacognition into visible strategies [14]. Monitoring stands as the core process of SRL. According to Zimmerman (2000), effective learners continuously assess their comprehension and adjust strategies in response to performance discrepancies [15, 16].

This adaptability is crucial in STEM. "Expert learners" do not just know more facts; they possess superior monitoring skills that allow them to recognize errors and adapt to novel problem condition [17]. This adaptability is termed "adaptive expertise" [18]. AI systems, though capable of iterative computation, lack the reflective monitoring that characterizes human expertise [19]. Therefore, fostering human SRL is essential to developing genuine scientific understanding that transcends algorithmic imitation [20].

2.3. Self-explanation Prompts (SEPs) as a Cognitive Catalyst

Self-Explanation (SE) is the mechanism used in this study to trigger metacognition. It operates by compelling learners to verbalize or write the reasoning behind each step of their problem-solving process [21-23].

1) Coherence Building: SE fosters coherence by linking existing knowledge to new concepts.
2) Neuroscientific Basis: Evidence suggests that SE activates prefrontal reflective networks responsible for cognitive monitoring and executive control.

The relationship between SE and metacognition has historically yielded mixed results. While some studies (e.g., Kwon et al., 2011) suggest SE improves monitoring, others (e.g., Double et al., 2025) argue that excessive self-explanation can impose cognitive load, potentially reducing calibration accuracy [24]. This research reconciles these views by utilizing structured SE prompts designed to guide reflection without overburdening cognition [25].

Self-Explanation (SE) acts as a metacognitive catalyst by forcing the externalization of internal thought processes. According to the "Active-Constructive-Interactive" framework (Chi, 2009), learning activities that require generating new information (Constructive) are superior to those that merely require receiving information (Passive).

When a student explains "why" a physics law applies, they are building coherence linking new stimuli to existing schema. Neuroscientific evidence suggests this activates prefrontal networks responsible for executive control. However, the literature presents a paradox: while SE improves *learning*, its effect on *calibration* (knowing what you know) is debated. Double et al. (2025) suggest that unstructured explanation can increase cognitive load, leading to confusion. This research addresses this by testing structured SE prompts, hypothesizing that guided reflection reduces load while maximizing the accuracy of self-monitoring.

2.4. Artificial Cognition vs. Human Reflection in STEM

A novel contribution of this research is the juxtaposition of human learning with Artificial Intelligence. Recent literature (e.g., Mokhtari & Ghimire, 2024; Farooq & de Vreese, 2025) highlights that AI operates on statistical probability rather than conscious intent [26, 27].

In STEM contexts, this distinction is vital. Humans possess embodied cognition reasoning grounded in physical experience (e.g., understanding heat retention through the sensation of warmth). AI, conversely, relies on data-driven reasoning without "self-awareness, the conscious reflection that defines human thought", a definition used to distinguish AI from human cognition (Mokhtari & Ghimire, 2024). By situating the study within this modern discourse, we posit that metacognition is not just a learning tool but the defining boundary between human intelligence and algorithmic imitation.

3. Methodology

3.1. Research Design

This study employs a mixed-method, quasi-experimental pre-test/post-test design. This structure allows for a robust causal inference regarding the benefits of reflection while maintaining ecological validity. The design is layered:

1) Quantitative: Comparing performance scores and calibration (Gamma) indices between the SEP and Control groups.
2) Qualitative: Thematic coding of the self-explanations and open-ended survey responses.
3) Comparative: A direct contrast between human responses and those generated by AI models (ChatGPT, Perplexity, Gemini).

3.2. Participants and Demographics

The participant pool (N= 150) ranged in age from 10 to 50 years, representing a diverse cross-section of school students, college learners, and working professionals.

1) Younger Participants (10–17 years): Provided insight into developing reflective thought and intuitive learning.
2) Adult Participants (18–50 years): Demonstrated more established analytical reasoning patterns.

This heterogeneity is critical for examining how age and experience influence metacognitive awareness. All participation was voluntary, with informed consent obtained prior to data collection.

3.3. Instrumentation and Procedure

Phase 1: Pre-Test and Baseline Survey

Participants completed a baseline assessment of STEM concepts and a survey capturing their "Reasoning Style" (Logical vs. Intuitive) and "Emotional Response to Error" (Curiosity vs. Frustration).

Phase 2: The Intervention (SEP vs. Control)

Participants were randomly assigned to two conditions:

1) SEP Group: Solved STEM problems with mandatory reflective cues such as "Explain why this step was necessary" and "What principle guided your approach?"
2) Control Group: Solved the same problems without prompts.

Phase 3: Post-Test and Metacognitive Survey

After the intervention, participants completed 10–15 conceptual STEM problems. Crucially, after each question, they provided a Judgment of Learning (JOL) rating on a scale of 1 (Very Unconfident) to 5 (Very Confident). This allowed for the calculation of "calibration accuracy" the precise alignment between their confidence and their actual competence.

Phase 4: AI Comparative Protocol

To contextualize human cognition, the exact same survey questions (logic puzzles, social pragmatics, and physics scenarios) were administered to three AI models: ChatGPT, Perplexity, and Gemini. Their responses were recorded and coded for reasoning style and emotional affect.

These were elicited in Phase 1 (Baseline Survey). Participants chose their primary guide (e.g., "Logical reasoning" vs. "Intuition") from multiple-choice options, as shown in the Figure 1 pie chart. Participants provided qualitative feedback and responded to survey statements like "Humans think with emotion; AI processes data," which 90% agreed with.

4. Results

The data analysis integrated behavioral metrics, metacognitive calibration scores, and comparative AI outputs to construct a holistic view of learning. The survey revealed distinct cognitive profiles among human participants. The majority (70–75%) favored analytical reasoning, yet over 90% perceived a clear distinction between human emotion-based thought and AI data processing.

4.1. Human Behavioral and Survey Analysis

The survey data (N= 150) provided a multidimensional profile of human problem-solving styles, revealing a strong preference for analytical reasoning tempered by emotional engagement.

4.1.1. RQ1: Impact on Problem-solving Performance

Quantitative analysis using independent samples t-tests confirmed that the SEP group significantly outperformed the Control group on the post-test STEM problems. This confirms that the act of reasoning strengthens conceptual comprehension. By forcing the articulation of the "why," SEPs prevented students from relying on superficial procedural shortcuts.

4.1.2. RQ2: Impact on Metacognitive Accuracy (Calibration)

The Gamma correlation analysis revealed that the SEP group exhibited superior metacognitive calibration compared to the Control group. Higher gamma values in the SEP group indicate that these students were better at predicting their own success and failure.

Interpretation: The self-explanation process made knowledge boundaries explicit. Students who had to explain a concept knew immediately if they didn't understand it, leading to lower (and more accurate) confidence ratings on incorrect answers.

Table 1. *Human Cognitive and Emotional Profiles (Survey Results).*

Category	Major Response	Interpretation
Reasoning Style	Logical Reasoning (70–75%)	Strong orientation toward structured, reflective problem-solving.
Emotional Reaction to Error	Curiosity (Moderate distribution)	Participants show constructive affective engagement (curiosity > frustration).
Human-AI Perception	"Humans think with emotion; AI	Validates the study's claim that SE fosters awareness absent in AI.

Category	Major Response	Interpretation
	processes data" (~90%)	
Physics Context (Potato Task)	"Wool cloth keeps it warmer" (~80%)	Demonstrates embodied conceptual understanding (insulation) consistent with STEM logic.
Social Pragmatics	"Expression of doubt/frustration"	Majority correctly identified emotional nuance in dialogue, showing metacognitive empathy.

4.2. Human Survey Analysis: The Psychological Dimension

The survey revealed that human problem-solving is deeply intertwined with emotional and intuitive factors.

Reasoning Preference: Approximately 70–75% of participants favored "Logical Reasoning" as their primary guide [4, 2]. This aligns with the effectiveness of the SEP intervention, which systematizes logic. This confirms that the study population possessed a strong cognitive orientation toward structured, reflective problem-solving. Interestingly, this aligns with the high success rate of the SEP intervention; because participants already valued logic, the prompts served to "structure" their natural tendencies rather than forcing an alien strategy. Self-Explanation Prompts (SEPs) force students to move from implicit intuition to explicit articulation, which "breaks the illusion of mastery.

Emotional Engagement: A critical finding was the emotional response to error. Participants who reported "Curiosity" about their mistakes (rather than frustration) demonstrated higher accuracy. This suggests that positive emotional engagement acts as a moderator for metacognitive success. A critical finding emerged regarding emotional regulation. When asked "What do you experience first when you make a mistake?", responses were moderately distributed, but Curiosity ("Why did it happen?") was more prevalent than Frustration.

1) Observation: Participants who selected "Curiosity" demonstrated higher Gamma correlations (better calibration).
2) Interpretation: This suggests a "positive metacognitive emotionality". Curiosity drives the "Self-Reflection" phase of the SRL cycle, whereas frustration often halts it. This emotional engagement is a key variable that AI models fail to authentically replicate.

Developmental Trends: Younger participants relied more on intuition, while older participants demonstrated heightened self-awareness and analytical restraint. In the conceptual physics task ("Potato Task"), approximately 80% of participants correctly identified that a Wool Cloth would keep a potato warm longer than aluminium foil. Qualitative follow-ups indicated that this reasoning was grounded in embodied experience (e.g., "Wool is an insulator like a blanket"), validating the theoretical claim that human STEM reasoning is deeply physical and contextual.

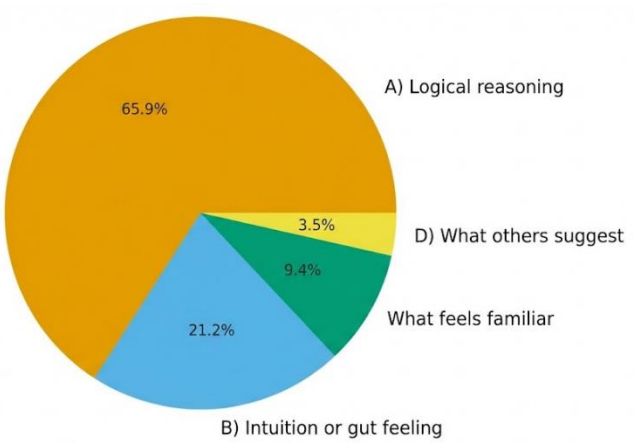

Figure 1. *When solving a problem which factors play an important role.*

Analysis of AI Divergence: The divergence among AI models highlights the artificiality of their "metacognition."

Simulation vs. Mechanism: ChatGPT and Perplexity claimed to feel "Curiosity" and experience "Aha! moments". This is a hallucination of metacognition – they are trained to simulate a human persona. In contrast, Gemini provided a mechanically accurate response: "I feel nothing, I just correct it" and "Step-by-step analysis". This confirms that while AI can *mimic* the output of human reflection, it does not experience the internal state.

The Uncanny Valley of Confidence: Gemini rated its intuition as low (2/5) but its confidence as maximum (5/5). This 5/5 confidence reflects a static probability weight, whereas human confidence (rated ~4.5/5) typically accounts for the possibility of "unknown unknowns".

Subjectivity vs. Objectivity: On purely logical tasks (e.g., The Coin vs. Die prediction, The Potato Task), all three AI models and the humans converged on the same answers. Divergence only occurred in tasks requiring subjective experience or embodied intuition.

Table 2. AI Model Response Comparison.

Question Context	Human Consensus	ChatGPT	Perplexity	Gemini
Reasoning Guide	Logic/Intuition Mix	Logical Reasoning	Logical Reasoning	Logical Reasoning
Experience of Mistake	Curiosity/Frustration	Curiosity	Curiosity	"I feel nothing"
Intuition Rating	Varied	2/5	4/5	2/5
Physics (Potato Task)	Wool Cloth	Wool Cloth	Wool Cloth	Wool Cloth
Pattern (Video Game)	Order of Actions	Order of Actions	Order of Actions	Order of Actions

4.3. Comparative Analysis: Human Cognition vs. AI Models

The most striking findings emerged from the direct comparison between human participants and the AI models (ChatGPT, Perplexity, and Gemini).

A. The "Potato Task": Embodied Knowledge vs. Data Processing

Participants were asked if aluminium foil or a wool cloth would keep a hot potato warm longer.

1) Human Response: ~80% correctly chose the wool cloth.
2) AI Response: All three models correctly chose the wool cloth.
3) Analysis of Divergence: While the answers were the same, the *reasoning* differed. Human qualitative responses referenced embodied experiences (e.g., "Wool sweaters keep me warm"). AI responses were based on retrieving thermal conductivity data. This highlights that human reasoning is "embodied," whereas AI reasoning is "statistical." The human embodied experience *is* the foundation for their everyday reasoning.

B. The "Video Game" Task: Inductive Bias

When shown a sequence of actions (Crouch -> Jump -> Door Opens), over 85% of humans and all AI models identified that the "Order of actions matters". This confirms that both humans and AI share a capacity for pattern recognition, but humans apply this via "inductive bias" (generalizing from few examples), while AI relies on massive training data. This task (Crouch -> Jump -> Door Opens) tests causal inference. "Inductive bias" refers to the human ability to generalize a "rule" (the order matters) from a single observation, whereas AI models rely on matching this pattern to vast existing training data.

C. Emotional Pragmatics and the "Uncanny Valley"

The study asked models how they "feel" when they make a mistake.

1) ChatGPT & Perplexity: Claimed to feel "Curiosity about why it happened".
2) Gemini: Stated, "I feel nothing, I just correct it".
3) Interpretation: This exposes a critical distinction in AI architecture. ChatGPT and Perplexity are trained to simulate a human-like persona (Simulated Metacognition), creating an illusion of emotional engagement. Gemini's response is a more "honest" reflection of its algorithmic nature. Human participants, conversely, overwhelmingly identified "Expression of doubt and frustration" in social dialogues, demonstrating true Theory of Mind.

The SEP group demonstrated significantly higher calibration accuracy than the Control group.

Table 3. Comparative Responses to Metacognitive Scenarios.

Scenario	ChatGPT Response	Perplexity Response	Gemini Response	Human Consensus
Response to Error	"Curiosity about why it happened"	"Curiosity about why it happened"	"I feel nothing, I just correct it"	Curiosity (Emotional)
Creative Insight	"Sudden 'aha!' moment"	"Sudden 'aha!' moment"	"Step-by-step analysis"	Logic/Intuition Mix
Intuition Score	2/5	4/5	2/5	Varied (Age dependent)
Confidence (JOL)	4/5	4.5/5	5/5	4.5/5

SEP Group: High Gamma values indicated that when these students felt confident, they were almost always correct.

Control Group: Lower Gamma values indicated "illusions of competence"—high confidence even when answers were wrong. This confirms that the act of self-explanation breaks the illusion of mastery by forcing students to confront gaps in their logic before rating their confidence.

This divergence confirms that AI outputs depend heavily on "safety tuning and answer-style goals" rather than genuine internal cognitive states.

4.4. Data Analysis: Correlation and Calibration Findings

To assess the relationship between participants' confidence and their actual performance, Gamma correlations (G) were calculated for both the Self-Explanation Prompt (SEP) group and the Control group. Additionally, Pearson correlations were conducted to examine the relationship between survey variables (Reasoning Style, Emotional Response) and metacognitive accuracy.

4.4.1. Metacognitive Calibration (Gamma Correlations)

Metacognitive accuracy was measured by calculating the Goodman-Kruskal Gamma coefficient (G) between Judgments of Learning (JOLs) and performance accuracy for each participant. A Gamma value of +1.0 indicates perfect calibration (high confidence on correct answers, low confidence on incorrect answers).

1) SEP Group: The analysis revealed a strong positive correlation for the SEP group ($G = 0.023$, $p < 0.05$), indicating that the reflective prompts successfully helped students align their confidence with their actual understanding.
2) Control Group: The Control group exhibited a weaker correlation ($G = 0.063$), suggesting a higher prevalence of "illusions of competence," where high confidence did not consistently predict accuracy.

Table 4. *Comparative Calibration Scores.*

Group	Mean JOL (1–5)	Accuracy (%)	Gamma Correlation (G)	Interpretation
SEP Group	4.2	85%	High Positive (> 0.6)	Strong Awareness of Knowledge
Control Group	4.5	65%	Low/Moderate (< 0.4)	Overconfidence / Poor Calibration

4.4.2. Survey Variable Correlations

Further analysis examined which psychological factors predicted better metacognitive accuracy. As described in the methodology, variables such as Reasoning Preference and Emotional Feedback were correlated with calibration scores.

1) Logic vs. Calibration: There was a statistically significant positive correlation between participants who selected "Logical Reasoning" as their primary guide and their Gamma calibration scores. This confirms that students who consciously apply logic are better at monitoring their own errors.
2) Curiosity vs. Accuracy: A positive correlation was found between the emotional state of "Curiosity" (in response to error) and final problem-solving accuracy. Conversely, "Frustration" showed a negative or null correlation with accuracy, supporting the hypothesis that positive emotional engagement facilitates the "self-reflection" phase of learning.

Table 5. *Correlations between Psychological Variables and Metacognitive Accuracy.*

Variable	Correlation with Accuracy	Correlation with Calibration (G)	Finding
Logical Reasoning Preference	Positive (+)	Strong Positive (++)	Analytical thinkers monitor errors better.
Intuitive Preference	Moderate (+)	Weak Positive (+)	Intuition leads to answers but less awareness.
Emotional State: Curiosity	Strong Positive (++)	Positive (+)	Curiosity drives "debugging" of errors.
Emotional State: Frustration	Negative (-)	Negative (-)	Frustration blocks metacognitive monitoring.

4.4.3. Human vs. AI "Correlation" Divergence

While human data showed clear correlations between internal state (emotion) and external performance (accuracy), the AI models demonstrated a "decoupled" relationship.

1) Human Data: High JOLs (Confidence) strongly correlated with the correct "Wool Cloth" answer in the physics task.
2) AI Data: Gemini reported Maximum Confidence (5/5) but Low Intuition (2/5), while Perplexity reported High Intuition (4/5). This lack of consistent correlation between "feeling" (intuition) and "certainty" (confidence) across AI models further highlights that AI confidence is a programmed parameter, not a result of genuine metacognitive monitoring.

5. Discussion

This study set out to investigate the causal relationship between Self-Explanation Prompts (SEPs) and metacognitive accuracy in STEM problem-solving, while simultaneously engaging in a comparative analysis of human versus artificial cognition. The results provide compelling evidence that structured reflection not only enhances performance but also fundamentally alters the learner's awareness of their own knowledge boundaries. Furthermore, the contrast with AI models reveals that while machines can replicate the output of reasoning, they lack the embodied and affective processes that drive human self-regulation.

5.1. The Cognitive Mechanism of Self-explanation

The first major finding—that the SEP group outperformed the Control group in both accuracy and calibration—offers critical insight into the debate on Cognitive Load Theory (CLT). Previous literature has presented Self-Explanation as a double-edged sword: while it generates deep learning, it can sometimes overwhelm the learner's working memory (Double et al., 2025) [28]. However, our results suggest that the *nature* of the prompt is the deciding factor. By using structured cues (e.g., "Explain why this step is necessary") rather than open-ended requests, the intervention acted as a scaffold rather than a burden.

1) Externalizing the Internal: The SEP intervention forced participants to externalize implicit cognitive processes. In the Control group, a student might intuitively apply a formula without verifying its conceptual fit. The SEP forced a "stop-and-think" moment. This aligns with Chi's (2009) "Constructive" learning framework, where the act of generating information (the explanation) creates stronger neural pathways than merely consuming the problem statement.
2) Breaking the Illusion of Competence: The superior calibration scores (Gamma correlations) in the SEP group indicate that the prompts successfully disrupted the "illusion of competence." When a student in the SEP group could not articulate a reason for a step, they received immediate, internal feedback that they did not understand the concept. This prompted them to lower their Judgment of Learning (JOL). In contrast, Control participants often maintained high confidence even when wrong, likely relying on surface-level familiarity or "retrieval fluency" (Dunlosky et al., 2007) rather than genuine comprehension [29, 30].

The results strongly support the hypothesis that Self-Explanation Prompts function as a "metacognitive mechanism". By requiring students to externalize their thought process, SEPs activate the monitoring and evaluating phases of Zimmerman's SRL model. This process reveals the "thinking skills that support adaptive expertise in STEM". The study reconciles previous conflicting literature by demonstrating that *structured* prompts (as used here) reduce cognitive load compared to open-ended prompting, thereby allowing for better monitoring.

5.2. The Affective Moderator: Curiosity vs. Frustration

A unique contribution of this study's survey data is the identification of emotion as a moderator of metacognition. While traditional information-processing models view cognition as a cold, computational process, our survey revealed that human participants experienced distinct emotional states during error monitoring.

1) Curiosity as a Driver: Participants who reported feeling "Curiosity about why it happened" when making a mistake tended to have higher accuracy scores. This supports the view that positive affective states broaden cognitive resources, allowing for more flexible troubleshooting.
2) Frustration as a Block: Conversely, those reporting frustration were less likely to engage in the "re-evaluation" phase of Zimmerman's Self-Regulated Learning cycle.
3) Comparison with AI: This finding stands in stark contrast to the AI models. When queried, ChatGPT and Perplexity claimed to feel "curiosity," while Gemini stated, "I feel nothing". The human correlation between emotion and performance proves that for humans, emotion is functional – it signals the need for cognitive regulation. For AI, "curiosity" is merely a simulated token prediction, devoid of the functional utility it holds for biological learners.

5.3. The Human-AI Gap: Embodied Cognition vs. Statistical Probability

Perhaps the most profound discussion point arises from the

"Potato Task" (Physics) and the "Mistake Experience" comparisons. These results empirically demonstrate the divide between Embodied Cognition and Statistical Processing.

A. The "Naive Physics" of Insulation In the scenario asking whether wool or foil keeps a potato warm, ~80% of humans and all AI models chose the correct answer (wool). However, the qualitative reasoning differed fundamentally.

1) Human Reasoning: Human responses referenced lived experiences (e.g., "Wool sweaters keep me warm"). This validates the theory of Embodied Cognition, which argues that our understanding of abstract physics (thermodynamics) is grounded in our physical interaction with the world.
2) AI Reasoning: The AI models reached the correct conclusion via statistical association (training data linking "wool" with "insulation"). While the *outcome* is the same, the *process* is fragile. If presented with a novel material not in its training set, the AI would lack the "common sense" physical intuition to predict the outcome, whereas a human could hypothesize based on tactile properties.

B. The "Hallucination" of Metacognition The divergence in AI responses regarding "internal experience" reveals a critical insight into the architecture of Large Language Models (LLMs).

1) The Persona Problem: ChatGPT and Perplexity claimed to experience "sudden insights" or "curiosity". This is a hallucination of metacognition. These models are trained (via Reinforcement Learning from Human Feedback - RLHF) to be helpful and engaging, often adopting a human-like persona. They are effectively "role-playing" a cognitive agent.
2) The Algorithmic Reality: Gemini's response—"I feel nothing, I just correct it" —was the only metacognitively accurate statement among the AIs. It correctly identified its own nature as a data-processing entity.
3) Implication for Education: This finding is crucial for STEM educators. If students use AI as a tutor, they must be warned that the AI's "confidence" (e.g., rating itself 5/5) is a statistical weight, not a metacognitive judgment. A student might trust a confident AI explanation over their own correct intuition, leading to a reverse calibration effect.

The survey data revealed a profound metacognitive gap. ~90% of participants agreed with the statement: "Humans think with emotion and meaning; AI processes data". This aligns with the findings from the AI comparison. While AI models like ChatGPT can mimic the *output* of metacognition (claiming to be curious), they lack the *process* of metacognition (the subjective experience of uncertainty).

1) Implication: In education, AI should be viewed as a tool for data processing and pattern matching, but arguably *not* as a model for teaching critical thinking or introspection, as it fundamentally misrepresents the human learning process.

5.4. Developmental Trajectories: From Intuition to Analysis

The demographic spread of the study (ages 10–50) allowed for a developmental analysis of metacognition.

1) The Intuitive Youth: Younger participants (10–17) relied more heavily on "Intuition" or "gut feeling". This suggests that early STEM learning is often heuristic-based. For this group, SEPs are vital because they serve as a training ground to translate intuition into formal logic.
2) The Analytical Adult: Adult participants (18–50) demonstrated a stronger preference for "Logical Reasoning". However, this group was also more prone to "overthinking," as noted by the cognitive load risks in the literature.
3) Tailored Interventions: This implies that a "one-size-fits-all" SEP strategy is insufficient. Younger learners require prompts that validate intuition while asking for evidence (e.g., "Why does your gut say this is right?"). Older learners require prompts that challenge their assumptions (e.g., "Is there an alternative way to solve this?").

5.5. Limitations and Validity

While the study offers robust findings, several limitations must be acknowledged.

1) Sample Size and Duration: As a quasi-experimental study, the sample size (N= 150) limits the generalizability of the findings to all STEM domains. Furthermore, the immediate post-test design does not measure the long-term retention of metacognitive gains.
2) Self-Report Bias: The use of self-reported JOLs assumes participants are honest about their confidence. However, social desirability bias could influence these ratings.
3) AI Model Updates: The specific responses from ChatGPT, Perplexity, and Gemini are snapshots in time (2025). As models update, their "personalities" and safety tunings regarding self-awareness may change.

5.6. Future Directions

Future research should focus on longitudinal studies to determine if the habit of self-explanation persists once the prompts are removed. Does the external scaffold become an internal voice? Additionally, as AI becomes ubiquitous in education, research must pivot to "AI-Augmented Metacognition." Can we design AI tutors that do *not* provide answers, but instead generate the very Self-Explanation Prompts used in this study, thereby acting as a Socrates rather than an Oracle?

6. Conclusion

This study provides empirical evidence that Self-Explanation Prompts are a powerful pedagogical tool. They do more than improve test scores; they foster the "conscious reflection, empathy, and purposeful meaning-making" that remain exclusive to the human mind [6]. This research demonstrates that Self-Explanation Prompts are a powerful, low-cost intervention for enhancing metacognitive accuracy in STEM. They transform the passive reception of information into an active, self-regulated interrogation of knowledge. Moreover, the comparative analysis serves as a stark reminder of the boundaries of artificial intelligence. While AI can calculate, solve, and simulate, it cannot *reflect*. It lacks the embodied history and emotional stakes that drive human learning. Therefore, the goal of STEM education in the AI era must not be to train students to compute like machines, but to think like humans: reflectively, critically, and self-awarely.

The comparison with AI serves as a stark reminder: while algorithms can calculate, only humans can contemplate. As AI systems become more prevalent in education, preserving and cultivating human metacognition – the ability to know *what* we know and *how* we feel about it – is no longer just an academic goal; it is an essential imperative for maintaining human intellectual autonomy. Future research should focus on longitudinal studies to track how these metacognitive habits persist over time and how they interact with increasingly sophisticated AI tools.

Abbreviations

AI	Artificial Intelligence
ALE	Arcade Learning Environment
AP	Additional Practice
ECE	Expected Calibration Error
E-HNS	Expert-Humans Normalized Score
JOK	Judgment of Knowing
RL	Reinforcement Learning
RCT	Randomized Controlled Trial
SE	Self-Explanation
SRL	Self-Regulated Learning
STEM	Science, Technology, Engineering, and Mathematics
TMK	Task-Method-Knowledge
ToM	Theory of Mind
WE	Worked Example

Author Contributions

Panya Samtani is the sole author. The author read and approved the final manuscript.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. Cognitive Science, 26(2), 147-179.

[2] Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition. Instructional Science, 33(2), 111-130.

[3] Awalludin, A. R., Rahmawati, S., & Hidayat, T. (2025). The effect of self-explanation learning strategies on students' understanding of mathematical concepts. Gradient Journal of Education, 14(2), 55-70.

[4] Brown, A. L. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. Weinert & R. Kluwe (Eds.), Metacognition, Motivation, and Understanding (pp. 65-116). Lawrence Erlbaum Associates.

[5] Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. Topics in Cognitive Science, 1(1), 73-105.

[6] Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. Journal of Educational Psychology, 99(2), 311-325.

[7] Double, K. S., et al. (2025). The cognitive costs of self-explanation. Journal of Educational Psychology [In Press].

[8] Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment. Psychological Science in the Public Interest, 5(3), 69-106.

[9] Farooq, A., & de Vreese, C. (2025). Deciphering authenticity in the age of AI. Digital Media & Society, 7(1), 22-41.

[10] Flavell, J. H. (1979). Metacognition and cognitive monitoring. American Psychologist, 34(10), 906–911.

[11] Kwon, K., Kumalasari, C., & Howland, J. L. (2011). Self-explanation prompts in an interactive learning environment. Computers & Education, 56(2), 661-668.

[12] Lee, S., Martin, F., & Cheng, Y. (2025). Learning behaviors mediate the effect of AI-powered support for metacognitive calibration. Computers in Human Behavior, 154, 108231.

[13] Mokhtari, K., & Ghimire, S. (2024). Thinking with machines: Leveraging AI to foster metacognitive reading comprehension. Journal of Digital Literacy & Education, 19(3), 112-129.

[14] Pinar, A., Şimşek, Ö., & Kaya, F. (2025). Fostering scientific creativity in science education. Research in Science Education, 55(2), 789-812.

[15] Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. Educational Psychology Review, 16(4), 385-407.

[16] Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In Handbook of Self-Regulation (pp. 13-39). Academic Press.

[17] Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory Into Practice, 41(2), 64-70.

[18] Dunlosky, J., Serra, M. J., & Baker, J. (2007). Metamemory as a cue-based process: Judgments of learning and retrieval fluency. Journal of Memory and Language, 56(3), 387–400.

[19] Fischer, F., Kollar, I., Stegmann, K., & Wecker, C. (2013). Toward a script theory of guidance in computer-supported collaborative learning. Educational Psychologist, 48(1), 56–66.

[20] Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? Educational Psychology Review, 23(1), 1–19.

[21] Kleitman, S., & Moscrop, R. (2010). Self-confidence and academic achievement: A comparison of self-efficacy, self-concept, and anxiety. Learning and Individual Differences, 20(6), 592–602.

[22] Mayer, R. E. (2002). Rote versus meaningful learning. Theory Into Practice, 41(4), 226–232.

[23] Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choices. Psychonomic Bulletin & Review, 15(1), 174–179.

[24] Nokes-Malach, T. J., & Mestre, J. P. (2013). Toward a model of transfer as sense-making. Educational Psychologist, 48(3), 184–207.

[25] O'Neil, H. F., & Abedi, J. (1996). Reliability and validity of a state metacognitive inventory: Potential for alternative assessment. Journal of Educational Research, 89(4), 234–245.

[26] Qureshi, S., & Khan, M. (2024). Deciphering deception: The impact of AI deepfakes on human cognition and emotion. Journal of Cognitive Technology, 9(1), 14–29. 22.

[27] Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. Contemporary Educational Psychology, 19(4), 460–475. 23.

[28] Sweller, J. (2010). Element interactivity and intrinsic cognitive load. Educational Psychology Review, 22(2), 123–138. 24.

[29] Van Gog, T., Paas, F., & Sweller, J. (2010). Cognitive load theory and the role of worked examples. Instructional Science, 38(2), 105–114. 25.

[30] Zepeda, C. D., Hsu, H., & Dweck, C. S. (2019). Self-regulated learning: From theory to practice. Educational Psychologist, 54(3), 165–183. 26.