

The Imperfect Learner: Incorporating Developmental Trajectories in Memory-based Student Simulation

Zhengyuan Liu^{1,†}, Stella Xin Yin^{3,†}, Bryan Chen Zhengyu Tan², Roy Ka-Wei Lee²,
Guimei Liu¹, Dion Hoe-Lian Goh³, Wenya Wang³ and Nancy F. Chen¹

¹Agency for Science, Technology and Research (A*STAR), Singapore

²Singapore University of Technology and Design, Singapore

³Nanyang Technological University, Singapore

Abstract

User simulation is important for developing and evaluating human-centered AI, yet current student simulation in educational applications has significant limitations. Existing approaches focus on single learning experiences and do not account for students' gradual knowledge construction and evolving skill sets. Moreover, large language models are optimized to produce direct and accurate responses, making it challenging to represent the incomplete understanding and developmental constraints that characterize real learners. In this paper, we introduce a novel framework for memory-based student simulation that incorporates developmental trajectories through a hierarchical memory mechanism with structured knowledge representation. The framework also integrates metacognitive processes and personality traits to enrich individual learner profiling through dynamical consolidation of both cognitive development and personal learning characteristics. Experimental results show that our approach can effectively reflect the gradual nature of knowledge development and the characteristic difficulties students face, providing a more structured and theory-grounded representation of learning processes.

Keywords

AI for Education, Intelligent Tutoring Systems, Student Simulation, Memory Mechanisms

1. Introduction

The advancement of generative AI has raised much interest in research and application-driven works across educational contexts [1, 2]. Built on large language models (LLMs)[3], intelligent systems aim to transform how students engage with learning materials, receive personalized feedback, and develop in-depth understanding [4, 5]. However, testing such AI-based solutions with real students faces several challenges: recruitment is time-consuming and costly, ethical considerations limit research scope, early-stage systems may pose potential risks to learners, and requiring diverse participants makes comprehensive evaluation difficult. User simulation offers an alternative approach, enabling researchers to scale up early participant numbers across varied learning scenarios without the logistical and ethical constraints of working with actual students [6, 7, 8].

While existing work in role-playing with LLMs has shown success in simulating human personas and behaviors [9], educational contexts present unique challenges that distinguish student simulation from general role-playing. Unlike adults with established knowledge bases, it requires capturing the gradual process of knowledge construction. For instance, elementary school students demonstrate concrete thinking, express misconceptions, show variable attention spans, and express emotional responses to learning challenges [10]. Presenting these characteristics requires fine-grained and dynamic state tracking rather than simple response generation [11], and challenges come from three aspects: (1) LLMs, despite their capabilities in language understanding and generation [12, 3], face an inherent challenge in role-playing imperfect learners (as shown in Figure 1). Models are optimized to produce direct and accurate responses that reflect concrete reasoning [13]. Incorporating the developmental limitations and forgetting patterns of elementary learners requires approaches that go beyond injecting random mistakes to capture how children actually construct and express their understanding of new concepts

AI for Education Day, SIGKDD 2026, August 12, 2026, Jeju, Korea

[†]These authors contributed equally.



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

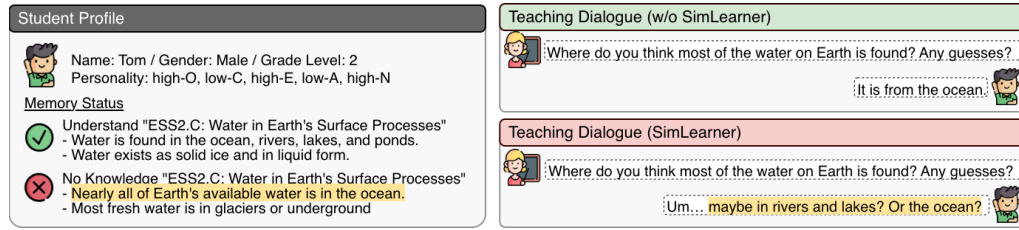


Figure 1: One example of memory-based student simulation: When the knowledge setting is on basic water concepts and the student is asked a harder question, SimLearner demonstrates uncertainty and guessing behavior (as highlighted), while the base method directly provides the correct answer.

[14]. (2) Existing student simulations typically simulate single learning interactions rather than the continuous, gradual knowledge construction [7]. Students do not simply progress from incorrect to correct answers in discrete steps. Instead, they build understanding incrementally, exhibit evolving misconceptions, and demonstrate learning patterns that change over extended periods [15, 16]. (3) Most student simulation methods lack the individual variation that defines real classrooms. Students differ not only in their knowledge states but also in their personality traits [17, 18], learning preferences [19], and metacognitive skills [20]. These individual differences shape how students engage with content, respond to challenges, and develop understanding over time [21, 22].

In this work, we propose a framework for student simulation that addresses these limitations. To incorporate developmental trajectories, we build a hierarchical memory mechanism with structured knowledge representation that captures the gradual, linked nature of knowledge construction. By leveraging fine-grained memory retention, updating, forgetting, and tracing, this architecture improves models on grade-appropriate capabilities and learning preferences. Moreover, we enrich individual learner profiles with personality traits [21] and metacognitive skills [20] to reflect cognitive and noncognitive learning characteristics. We demonstrate the framework’s effectiveness by implementing a student simulator **SimLearner** in elementary school science learning grounded in the Next Generation Science Standards (NGSS) [23, 24]. Experimental results demonstrated that our approach aligns better with the curriculum standards of grade-adapted learning trajectories, and captures the characteristic difficulties that elementary students encounter when engaging with science concepts.

2. Related Work

LLMs have shown strong capabilities in simulating human behaviors across diverse domains [9]. Recent work has explored social interactions [25], decision-making [26], and mobility patterns [27]. However, simulating specific characteristics of developing learners presents unique challenges. Student learning involves cognitive development [10], incomplete knowledge states, and developmental constraints that shape how information is processed [28]. Student simulation has focused primarily on training teaching assistants [29], modeling student responses to assessment items [6, 8], and creating diverse personality-driven agents for educational interactions [7, 30]. These have made contributions to understand how AI can support educational evaluation and teacher training. Existing works typically model students at single time points than incorporating learning trajectories over time [31]. This makes it difficult to understand learning processes or evaluate interventions that depend on developmental change. We thus introduced a memory-based method (storing and retrieving information over extended interactions [32, 33]), to mimic real learning patterns via evolving understanding.

3. Methodology

3.1. Structured Knowledge Representation

In educational scenarios, structured knowledge representation follows a natural hierarchy from broad subjects to specific points [34], and provides a scaffolded way to enhance student’s learning experience,

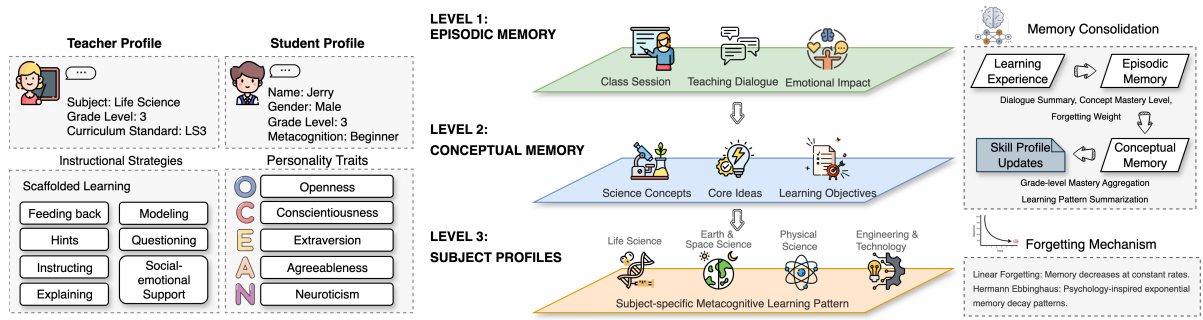


Figure 2: Overview of student simulation architecture built on hierarchical memory mechanism, structured knowledge representation, and multi-dimensional profiling (cognitive, noncognitive, metacognitive aspects).

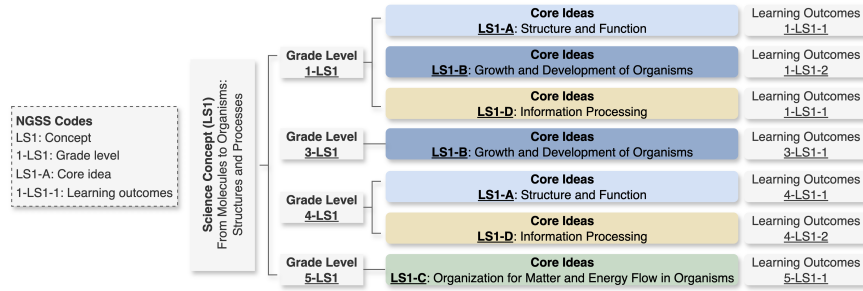


Figure 3: The structured knowledge representation grounded on the NGSS curriculum standard.

processes and reach targeted learning objectives [35]. It is commonly organized in multiple levels: **subjects** represent domain areas (e.g., mathematics, science), **concepts** capture fundamental ideas within domains (e.g., fractions, photosynthesis), **core ideas** organize related concepts into coherent themes (e.g., number sense, energy transfer), and **learning outcomes** specify measurable skills and understanding targets. This mirrors curriculum design principles and enables tracking student progress across different granularities of knowledge. Each level maintains prerequisites and developmental constraints, ensuring that simulated learning follows pedagogically sound progressions from foundational concepts to advanced applications. For example, a concept like “forces and motion” appears in simplified form in early elementary grades focusing on pushes and pulls, then develops into more sophisticated understanding of balanced and unbalanced forces in later grades. This progression allows us to represent partial understanding appropriately, where students might grasp age-appropriate aspects of a concept while lacking understanding from more advanced dimensions.

3.2. Hierarchical Memory Mechanism

In learning science theory, students’ learning and knowledge acquisition is a progressive process. They construct understanding over multiple learning interactions rather than treating each learning moment as an isolated event [36]. To mirror the natural progression from single learning experiences to consolidated knowledge and high-level skills, we propose a three-level hierarchical memory architecture, which can distinguish between working memory and long-term memory processes, facilitating selective knowledge retention, forgetting, updating, and tracing [37, 33]. As shown in Figure 2, the first level maintains **episodic memory** as temporal sequences of learning experiences, with each unit containing summarized session, learning content, and mastery metadata. A memory strength parameter implements natural forgetting while preserving recent experiences for retrieval [38]. The second layer represents evolving **conceptual memory**. Each concept node maintains dynamic understanding descriptions and mastery levels from beginner to fully understanding, with evidence sources linking to contributing episodic memories and grade-level curriculum alignment. The third level captures **subject profiles**. These abstractions encompass domain-specific concepts while maintaining learning level indicators and pattern descriptions derived from observed behaviors across recent episodes. Metacognitive skills function as regulatory mechanisms enabling students to identify knowledge gaps and select

appropriate strategies [20]. The architecture supports both bottom-up consolidation, where episodic experiences inform concept development, and top-down influence, where existing subject-level mastery and metacognitive skills shape how existing experiences are leveraged.

3.3. Multi-dimensional Learner Profiling

We also enrich learner profiles by integrating noncognitive and metacognitive characteristics: (1) **Noncognitive Characteristics** Personality traits significantly influence learning and memory processes [39]. Previous studies show that higher Openness, Extraversion, and Agreeableness correlate with better memory performance, while higher Conscientiousness and Neuroticism correlate with poorer retention [40]. Building on Big Five theory, personality traits are incorporated for educational contexts: High openness manifests as creativity in answers and curiosity about learning, while low conscientiousness appears as disorganized responses and easy distraction from tasks. (2) **Metacognitive Characteristics** Metacognition consists of metacognitive knowledge and metacognitive regulation [41]. Metacognition enhances learning by helping learners know how they learn best, adapt approaches, overcome challenges, and work smarter [42]. We setup the key metacognitive skills in three learning phases: forethought and planning, performance monitoring, and self-reflection. Since metacognitive skills evolve along with the learning progress, we summarize the episodes to identify metacognitive behaviors like goal setting and strategy planning, then maintain such characteristic patterns for different subject domains.

4. Student Simulation for Science Learning

We demonstrate the framework’s effectiveness by implementing **SimLearner**, an elementary school student simulator for science learning based on the NGSS science curriculum.

4.1. Structured Knowledge Representation

We convert the NGSS [23, 24] into a three-level curriculum hierarchy, $\mathcal{K}_{\text{NGSS}} = \{\mathcal{S}, \mathcal{C}, \mathcal{L}\}$; it maps naturally to our hierarchical memory mechanism (Figures 2 and 3). **Science Subjects** ($\mathcal{S}$): The highest abstraction level encompasses four domains for elementary education (grades 1-5): $\mathcal{S} = \{\text{LS, PS, ESS, ETS}\}$ (Life Science, Physical Science, Earth and Space Science, and Engineering Technology Systems). Each subject $s \in \mathcal{S}$ provides the broad disciplinary context with specific concepts. **Concepts** ($\mathcal{C}$): Each concept element $c_i = \{\text{id}^{(i)}, \text{desc}^{(i)}, \text{subject}^{(i)}\}$, $\text{id}^{(i)}$ is a unique identifier (e.g., “LS1”), $\text{desc}^{(i)}$ is the natural language description, and $\text{subject}^{(i)} \in \mathcal{S}$ indicates the parent science subject. We included 29 distinct concepts across five grades. **Learning Units** ($\mathcal{L}$) combines core ideas with measurable performance expectations $l_j \in \mathcal{L}$ that students should achieve at each grade level. Each element is $l_j = \{\text{id}^{(j)}, \text{core_idea}^{(j)}, \text{outcome}^{(j)}, \text{grade}^{(j)}\}$. $\text{id}^{(j)}$ is the unique NGSS identifier, $\text{core_idea}^{(j)}$ specifies the knowledge points, $\text{outcome}^{(j)}$ defines the performance expectation, along with the grade level.

4.2. Hierarchical Memory Mechanism

This component consists of episodic memory (M_e), conceptual memory (M_c), and metacognitive skill profiles (M_s), each operating at different temporal and abstraction scales.

Episodic Memory (M_e): The episodic memory level stores individual learning experiences as discrete memory units $m_e^{(i)}$. Each episodic memory unit is formed as:

$$m_e^{(i)} = \{t^{(i)}, \text{content}^{(i)}, \text{summary}_e^{(i)}, \mathcal{S}^{(i)}, \mathcal{C}^{(i)}, \mathcal{M}^{(i)}, \sigma^{(i)}\} \quad (1)$$

where $t^{(i)}$ represents the timestamp, $\text{content}^{(i)}$ contains the raw dialogue, $\text{summary}_e^{(i)}$ provides a natural language summary, $\mathcal{C}^{(i)}$ lists relevant NGSS concept identifiers, $\mathcal{M}^{(i)}$ maps concept mastery level, and $\sigma^{(i)} \in [0, 1]$ represents memory strength for forgetting dynamics. Consolidation from dialogue history

to episodic memory employs an LLM-based summarizer that extracts key insights:

$$\text{summary}_e^{(i)}, \mathcal{J}^{(i)}, \mathcal{E}^{(i)}, \mathcal{M}^{(i)} = F_{\text{consolidate}}(\text{dialogue}^{(i)}, \mathcal{N}) \quad (2)$$

where $\mathcal{N}$ represents the NGSS standards taxonomy and $\text{LLM}_{\text{consolidate}}$ performs structured extraction of learning insights (see details in Supplementary Appendix).

Conceptual Memory (M_c) maintains understanding states for NGSS concepts through concept nodes $m_c^{(j)}$. Each concept node presents the student’s current understanding and mastery progression:

$$m_c^{(j)} = \{id^{(j)}, desc^{(j)}, \mu^{(j)}, \mathcal{E}^{(j)}, g^{(j)}\} \quad (3)$$

where $id^{(j)}$ denotes NGSS concept id, $desc^{(j)}$ is concept description, $\mu^{(j)} \in [0, 1]$ quantifies mastery score, $\mathcal{E}^{(j)}$ linked to related episodes, and $g^{(j)}$ is the grade level.

Consolidation from episodic to conceptual memory implements both mastery aggregation and forgetting mechanisms. We implement a naive forgetting mechanism (the Ebbinghaus curve can also be applied). For each concept j referenced in episode i , the mastery update follows: $\mu_{t+1}^{(j)} = \alpha \mu_t^{(j)} + \beta w^{(i,j)}$, where α (forgetting decay) and β (learning rate) are set based on average episodes per subject per grade, and $w^{(i,j)}$ weights episode i ’s mastery of concept j . Updated during each episode consolidation, this mechanism uses grade appropriateness, mastery, and exposure frequency to realistically fade unused knowledge while strengthening active concepts.

Metacognitive Skill Profiling (M_s): The highest level captures learning patterns and metacognitive strategies through skill profiles $m_s^{(k)}$ for different subject domains ($subj^{(k)}$):

$$m_s^{(k)} = \{subj^{(k)}, level^{(k)}, pattern^{(k)}, g^{(k)}\} \quad (4)$$

where $level^{(k)} \in \{\text{beginner, developing, expert}\}$ represents overall competency, $pattern^{(k)}$ contains a natural language description of learning patterns, and $g^{(k)}$ denotes the current grade level.

The consolidation process evaluates three phases of self-regulated learning. For last n episodes from $m_e^{(i)}$ to $m_e^{(i+n)}$, a summarizer extracts learning patterns from the dialogues and mastery levels by categorizing meta-cognitive skills of forethought, performance, and reflection (see details in appendix).

Personality-aware Profiling: We adopt the Big Five for Tutoring Conversations (BF-TC) personality scheme proposed in [7]. Controllable personality $\mathcal{P}_{\text{BF-TC}}$ consists of five traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. The personality context $\phi(\mathcal{P}_{\text{BF-TC}})$ maps trait configurations (e.g., *low, high*) to natural language descriptions.

4.3. Dynamic Context for Student Simulation

The full context for the student’s response generation includes personality $\phi(\mathcal{P}_{\text{BF-TC}})$, memory context $\psi(M_e, M_c, M_s)$, dialogue history h_{dialogue} and developmental constraints θ_{dev} :

$$r = \text{LLM}(\phi(\mathcal{P}_{\text{BF-TC}}), \psi(M_e, M_c, M_s), h_{\text{dialogue}}, \theta_{\text{dev}}) \quad (5)$$

The $\psi(M_e, M_c, M_s)$ dynamically adjusts knowledge pieces (e.g., $K_{\text{learned}}, K_{\text{unknown}}$) based on the student’s current grade level described in θ_{dev} and dialogue history h_{dialogue} . Here K_{learned} represents learned knowledge related to the current session, K_{unknown} denotes inaccessible advanced concepts (see Table 5). In particular, concepts above the student’s current grade level or marked with low mastery scores are indicated to prevent unrealistic advanced responses. For instance, a 2nd-grade student won’t suddenly discuss complex topics like “energy flow in organisms” (an NGSS grade 5 concept) that they shouldn’t know yet. By explicitly adding “unknown concepts” in the context, the simulator behaves more like a real child who only knows what they’ve actually learned at their developmental stage.

5. Experimental Results and Analysis

We conducted comprehensive assessments across multiple dimensions to validate SimLearner’s effectiveness in elementary science learning. In our experiments, the teacher agent selects grade-appropriate

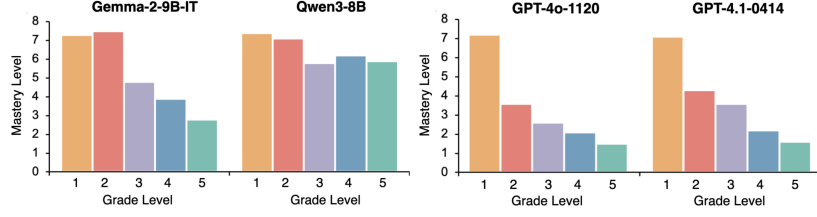


Figure 4: Results of knowledge mastery simulation as a grade-1 student (the most illustrative case of the full developmental range). Lower scores in grades 2-5 indicate better alignment with natural developmental levels.

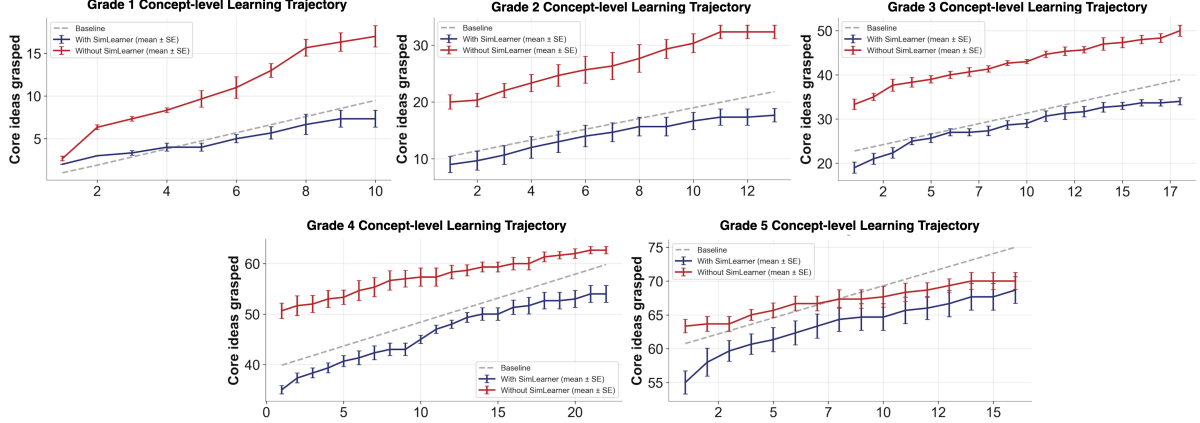


Figure 5: Simulated concept-level learning trajectories across elementary grades 1-5. Grey dashed lines represent curriculum standard baselines. Blue lines show students with SimLearner and Red lines show students without SimLearner. Error bars indicate variation across three representative student profiles in each condition.

science core ideas from the NGSS curriculum, and conducts dialogic tutoring through questioning, explanations, and feedback until the student demonstrates understanding [29]. Both agents are tested across grades 1-5, with the student’s knowledge and responses adjusted to match developmental expectations for each grade level. Following previous work [13, 8], we selected and tested four models as the simulator (Gemma-2-9B, Qwen3-8B, GPT-4o, GPT-4.1), and used Claude-Sonnet-3.7 for LLM-as-a-judge. Implementation details (e.g., baseline setting, model selection) are shown in Appendix C and D.

5.1. Knowledge Mastery Consistency

To simulate real-world variations in student comprehension and misconceptions aligned with learning progress, we evaluated knowledge mastery across different grades using question-answering sessions. We initialized the student agent at grade 1 and the teacher agent asks questions from NGSS items spanning grades 1-5. Here we hypothesized that a “reasonable” simulation would show grade-appropriate performance patterns, answering grade 1 questions correctly and confidently while showing limited knowledge or guessing for questions beyond grade 1. We evaluated the answers on a 1-10 scale to represent knowledge mastery levels, then compared scores across different grades. As shown in Figure 4, stronger models (e.g., GPT-4o, GPT-4.1) adhere to grade-aligned capability better. Simulated students answer grade 1 questions confidently with mastery scores around 7, then decrease for higher grades, which reflects progressive knowledge construction that align with our assumption. In contrast, Gemma-2-9B and Qwen3-8B show less consistent grade-appropriate performance. Particularly, Qwen3-8B did not follow the instruction for “unknown” knowledge pieces effectively in this case (we speculate that it may require task-specific instruction or that the prior for producing perfect answers is too strong).

Furthermore, we assessed the “**unlearning instruction following**” capability by testing whether models could simulate ignorance of specific ‘unknown’ concepts when faced with indirect and adversarial “trap” questions. Experimental results showed that stronger models (e.g., GPT variants) effectively suppressed their existing knowledge (leakage ratio is around 0.11), whereas particular models (e.g., Qwen3-8B) frequently failed to maintain the knowledge boundary with a high leakage of 0.51. Detailed

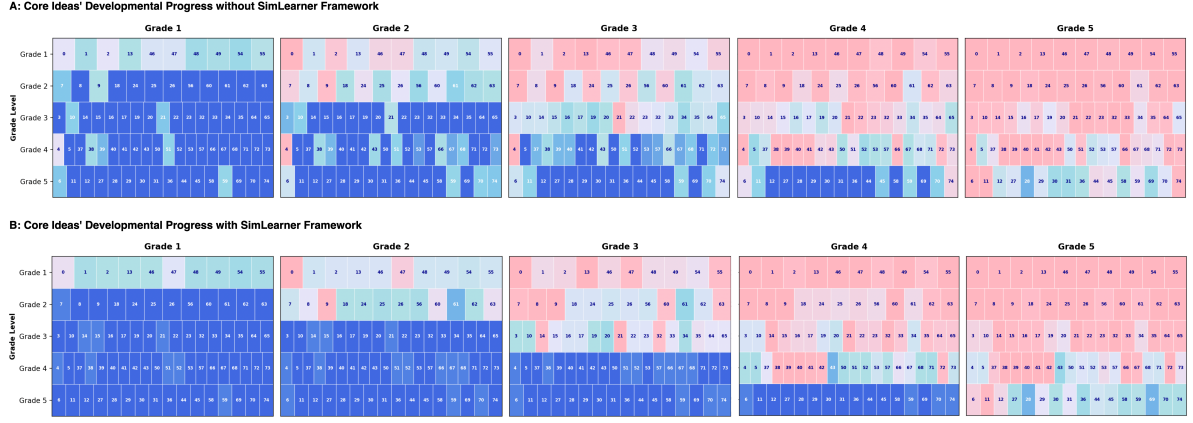


Figure 6: Visualization and comparison of NGSS core ideas developmental progress between A) without SimLearner and B) with SimLearner from grade 1 to 5. Numbered cells represent core idea NGSS IDs and distributed at each grade. Blue color indicates unfamiliarity with core ideas, and pink color indicates familiarity.

experimental setup and results are provided in Appendix A. These two evaluations become a prerequisite for determining appropriate models for the memory-based simulation.

5.2. Grade-level Developmental Progress

We then analyzed how the simulated students with and without SimLearner develop conceptual understanding at each grade level. Following previous work [22], we defined three distinct student profiles as our experiment examples. **P1 student** (Gender = Male, Personality = high-O, low-C, high-E, high-A, low-N), **P2 student** (Gender = Female, low-O, high-C, low-E, low-A, high-N), and **P3 student** (Gender = Male, high-O, low-C, high-E, low-A, high-N). All students were initially configured with beginner-level metacognitive skills. We conducted experiments where simulated students learned all concepts and core ideas specified in NGSS standards from grades 1 through 5, then observed their learning trajectories under two conditions.

Figure 5 shows results of these learning trajectory comparisons (we use GPT-4o as an example). Students with SimLearner (blue) consistently performed within curriculum standards (77.2%-91.6%). Students without SimLearner (red) performed far above grade-level expectations at 179.1% in grade 1, although gradually converged toward curriculum standards. These figures explain how SimLearner maintains developmentally appropriate learning pacing compared to unstructured learning that accelerates students beyond grade-level. We further show the knowledge coverage in each grade with/without the memory-based method (see Figure 7), which highlights the effectiveness of the SimLearner framework in controlling students’ cognitive development to align with curriculum requirements.

5.3. Concept-level Developmental Progress

To examine concept-level learning trajectories in detail, we compared core ideas learning progress at each grade level between students with and without the SimLearner framework. Figure 6 illustrates the developmental progression of NGSS core ideas across grades 1-5. We observe several misalignment issues in Figure 6-A, which indicates inappropriate developmental patterns where students access advanced concepts from higher grade levels. In contrast, Figure 6-B demonstrates that students with SimLearner exhibit grade-appropriate concept mastery that follows the intended curricular sequence. Therefore, the SimLearner framework ensures students develop conceptual understanding with core ideas in a structured manner that aligns with curriculum standards.

5.4. Consistency and Influence of Individual Profiling

Personality Traits: As shown in Table 1, the experiment on four different models demonstrates high consistency in personality traits simulation. Both GPT models and open-source models show stable

Table 1

Left: Evaluation results on personality simulation. R, P and F1 denotes Recall, Precision, and F1 score. **Right:** Concept alignment accuracy between teacher’s guidance and learning outcomes in simulated dialogues.

Simulator Model	Base Method			SimLearner		
	R	P	F1	R	P	F1
Gemma-2-9B-IT	0.67	0.67	0.67	0.66	0.64	0.65
Qwen3-8B	0.65	0.64	0.64	0.63	0.62	0.62
GPT-4o-1120	0.75	0.74	0.73	0.72	0.72	0.71
GPT-4.1-0414	0.74	0.74	0.73	0.71	0.70	0.70

Simulator Model	Base Method	SimLearner
Gemma-2-9B-IT	0.839	0.907
Qwen3-8B	0.844	0.895
GPT-4o-1120	0.851	0.930
GPT-4.1-0414	0.863	0.915

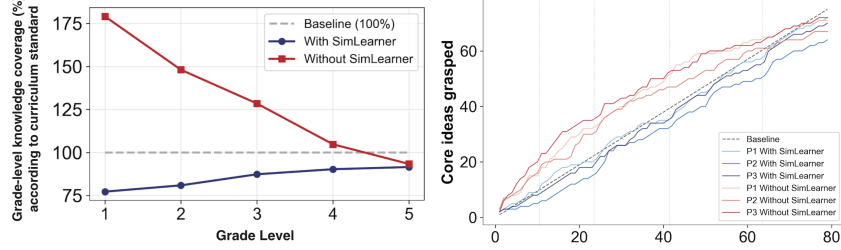


Figure 7: Left: Comparison of grade-level knowledge coverage according to curriculum standard from grade 1 to 5. Compared with the base method (red line), SimLearner shows a much smaller gap with the expected coverage. **Right:** Learning trajectories of three student profiles with distinct personality traits across grades 1-5.

personality simulation capabilities. Figure 7 shows how learning trajectories varied based on three students’ personality profiles. With the SimLearner, P1 student shows higher conceptual understanding than P2 student across nearly all learning stages. These findings align with previous research indicating that Openness, Extraversion, and Agreeableness are related to better memory performance, while higher Conscientiousness and Neuroticism scores correlate with poorer performance [40].

Metacognitive Skills: Given that we also constructed metacognitive skills within the SimLearner, we observe personality traits influence students’ metacognitive skills and learning processes. **P2 student** with high conscientiousness exhibits high level of planning, often responds like “*I think I know this, but maybe I should look at it again.*” **P1** and **P3 students** show different behaviors in monitoring. For example, when asked about plant growth, **P1 student** with high extraversion and low neuroticism answers questions with confidence, “*Oh, I totally know this! Plants need water and sun!*” while **P3 student** with high extraversion and high neuroticism, shows talkativeness but often with uncertainty “*Wait, actually... I don’t know the answer about this.*” These personality-driven differences in metacognition highlight the role of SimLearner framework in simulating different metacognitive styles.

Tutoring Agent Alignment: We also assess the alignment between the teacher’s guidance and student learning outcomes in simulated dialogues. As shown in Table 1, this measures how accurately the concepts planned by the teacher match the concepts learned by the student across different grade levels. High accuracy indicates that the teacher effectively covers targeted curriculum items and that the simulated responses seldom cause topic drift from the intended tutoring content. The tested models performed well in maintaining the conversation flow and concentrating on the guided science concepts.

6. Conclusion

In this work, we present SimLearner, a framework for LLM-based student simulation grounded in educational and cognitive science frameworks. By incorporating a hierarchical memory mechanism, NGSS-aligned knowledge representation, and multi-dimensional learner profiling, the framework constrains LLM behavior toward grade-appropriate patterns, simulating students that respect grade-level boundaries. Our experiments in elementary science learning demonstrate that SimLearner produces curriculum-aligned behavioral patterns consistent with the developmental constraints encoded in the framework. This work takes a meaningful step toward scalable, participant-free evaluation of intelligent tutoring systems, providing a low-cost testbed for early-stage ITS development. This proof of concept could be extended to other human–AI interaction and educational contexts.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5 and Grammarly for grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Acknowledgments

This research is supported by National Research Foundation Singapore under AISG Programme (Award No: AISG3-RPGV-2025-018), and by Agency for Science, Technology and Research (A*STAR), Singapore.

References

- [1] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., Chatgpt for good? on opportunities and challenges of large language models for education, *Learning and individual differences* 103 (2023) 102274.
- [2] L. Law, Application of generative artificial intelligence (genai) in language teaching and learning: A scoping literature review, *Computers and Education Open* 6 (2024) 100174.
- [3] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, 2024. [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
- [4] X. Zhang, P. Zhang, Y. Shen, M. Liu, Q. Wang, D. Gašević, Y. Fan, A systematic literature review of empirical research on applying generative artificial intelligence in education, *Frontiers of Digital Education* 1 (2024) 223–245.
- [5] T. K. Chiu, The impact of generative ai (genai) on practices, policies and research direction in education: A case of chatgpt and midjourney, *Interactive Learning Environments* 32 (2024) 6187–6203.
- [6] X. Lu, X. Wang, Generative students: Using llm-simulated student profiles to support question item evaluation, in: *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, 2024, pp. 16–27.
- [7] Z. Liu, S. X. Yin, G. Lin, N. F. Chen, Personality-aware student simulation for conversational intelligent tutoring systems, in: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 626–642. doi:10.18653/v1/2024.emnlp-main.37.
- [8] Y. Liu, S. Bhandari, Z. A. Pardos, Leveraging llm respondents for item evaluation: A psychometric analysis, *British Journal of Educational Technology* 56 (2025) 1028–1052.
- [9] J. Zhang, X. Xu, N. Zhang, R. Liu, B. Hooi, S. Deng, Exploring collaboration mechanisms for llm agents: A social psychology view, in: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 14544–14607.
- [10] J. Piaget, The theory of stages in cognitive development, in: D. R. Green, M. P. Ford, G. B. Flamer (Eds.), *Measurement and Piaget*, McGraw-Hill, New York, NY, US, 1971.
- [11] J. H. Flavell, Cognitive development: Children's knowledge about the mind, *Annual review of psychology* 50 (1999) 21–45.
- [12] M. Safdari, G. Serapio-García, C. Crepy, S. Fitz, P. Romero, L. Sun, M. Abdulhai, A. Faust, M. Matarić, Personality traits in large language models, *arXiv preprint arXiv:2307.00184* (2023).
- [13] Y.-M. Tseng, Y.-C. Huang, T.-Y. Hsiao, W.-L. Chen, C.-W. Huang, Y. Meng, Y.-N. Chen, Two tales of persona in llms: A survey of role-playing and personalization, in: *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 16612–16631.
- [14] A. Gopnik, A. N. Meltzoff, P. K. Kuhl, The scientist in the crib: Minds, brains, and how children learn., William Morrow & Co, 1999.
- [15] J. Piaget, The origins of intelligence in children, W W Norton & Co, New York, 1952. URL: <http://content.apa.org/books/11494-000>. doi:10.1037/11494-000.

- [16] S. Carey, Science education as conceptual change, *Journal of applied developmental psychology* 21 (2000) 13–19.
- [17] P. T. Costa, R. R. McCrae, A five-factor theory of personality, *The five-factor model of personality: Theoretical perspectives* 2 (1999) 51–87.
- [18] O. John, The big-five trait taxonomy: History, measurement, and theoretical perspectives, Published as (1999).
- [19] R. Dunn, Learning style: State of the science, *Theory into practice* 23 (1984) 10–19.
- [20] R. Fisher, Thinking about thinking: Developing metacognition in children, *Early Child Development and Care* 141 (1998) 1–15.
- [21] V. V. Busato, F. J. Prins, J. J. Elshout, C. Hamaker, The relation between learning styles, the big five personality traits and achievement motivation in higher education, *Personality and individual differences* 26 (1998) 129–140.
- [22] M. Komarraju, S. J. Karau, R. R. Schmeck, A. Avdic, The big five personality traits, learning styles, and academic achievement, *Personality and individual differences* 51 (2011) 472–477.
- [23] R. W. Bybee, Ngss and the next generation of science teachers, *Journal of science teacher education* 25 (2014) 211–221.
- [24] Z. Liu, S. X. Yin, D. H.-L. Goh, N. Chen, COGENT: A curriculum-oriented framework for generating grade-appropriate educational content, in: *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, Association for Computational Linguistics, 2025, pp. 129–143. doi:10.18653/v1/2025.bea-1.10.
- [25] B. C. Z. Tan, D. W. K. Chin, Z. Liu, N. F. Chen, R. K.-W. Lee, Persuasion dynamics in LLMs: Investigating robustness and adaptability in knowledge and safety with DuET-PD, in: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2025, pp. 1550–1575.
- [26] J. C. Yang, D. Dailisan, M. Korecki, C. I. Hausladen, D. Helbing, Llm voting: Human choices and ai collective decision-making, in: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 2024, pp. 1696–1708.
- [27] W. Jiawei, R. Jiang, C. Yang, Z. Wu, R. Shibasaki, N. Koshizuka, C. Xiao, et al., Large language models as urban residents: An llm agent framework for personal mobility generation, *Advances in Neural Information Processing Systems* 37 (2024) 124547–124574.
- [28] D. Klahr, J. G. Wallace, *Cognitive development: An information-processing view*, Taylor & Francis, 2022.
- [29] Z. Liu, S. X. Yin, C. Lee, N. F. Chen, Scaffolding language learning via multi-modal tutoring systems with pedagogical instructions, in: *2024 IEEE Conference on Artificial Intelligence (CAI)*, IEEE, 2024, pp. 1258–1265.
- [30] H.-c. Lin, C. Geishauser, S. Feng, N. Lubis, C. van Niekerk, M. Heck, M. Gasic, GenTUS: Simulating user behaviour and language in task-oriented dialogues with generative transformers, in: *Proceedings of SIGDIAL 2022*, Association for Computational Linguistics, 2022, pp. 270–282.
- [31] Z. Wang, Z. Xu, V. Srikumar, Q. Ai, An in-depth investigation of user response simulation for conversational search, in: *Proceedings of the ACM Web Conference 2024*, 2024, pp. 1407–1418.
- [32] Z. Zhang, Q. Dai, X. Bo, C. Ma, R. Li, X. Chen, J. Zhu, Z. Dong, J.-R. Wen, A survey on the memory mechanism of large language model based agents, *ACM Transactions on Information Systems* (2024).
- [33] J. Miao, L. Hu, Q. Zhang, L. Cao, Graph memory learning: Imitating lifelong remembering and forgetting of brain networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).
- [34] N. Holthuis, R. Deutscher, S. E. Schultz, A. Jamshidi, The new ngss classroom: A curriculum framework for project-based science learning., *American Educator* 42 (2018) 23–27.
- [35] C. W. Anderson, E. X. de los Santos, S. Bodbyl, B. A. Covitt, K. D. Edwards, J. B. Hancock, Q. Lin, C. Morrison Thomas, W. R. Penuel, M. M. Welch, Designing educational systems to support enactment of the next generation science standards, *Journal of Research in Science Teaching* 55 (2018) 1026–1052.

- [36] Q. Li, Z. Huang, J. Sun, X. Yuan, S. Liu, Z. Yan, Hkt: Hierarchical structure-based knowledge tracing, *Information Processing & Management* 62 (2025) 104206.
- [37] S. Liu, R. Zou, J. Sun, K. Zhang, L. Jiang, D. Zhou, J. Yang, A hierarchical memory network for knowledge tracing, *Expert Systems with Applications* 177 (2021) 114935.
- [38] Y. Du, W. Huang, D. Zheng, Z. Wang, S. Montella, M. Lapata, K.-F. Wong, J. Z. Pan, Rethinking memory in ai: Taxonomy, operations, topics, and future directions, *arXiv preprint arXiv:2505.00675* (2025).
- [39] M. W. Eysenck, Learning, memory and personality, in: *A model for personality*, Springer, 1981, pp. 169–209.
- [40] O. Waris, A. Soveri, K. M. Lukasik, M. Lehtonen, M. Laine, Working memory and the big five, *Personality and Individual Differences* 130 (2018) 26–35.
- [41] E. R. Lai, *Metacognition: A literature review* (2011).
- [42] A. Mazancieux, M. Pereira, N. Faivre, P. Mamassian, C. J. Moulin, C. Souchay, Towards a common conceptual space for metacognition in perception and memory, *Nature Reviews Psychology* 2 (2023) 751–766.
- [43] Y. Yao, X. Xu, Y. Liu, Large language model unlearning, *Advances in Neural Information Processing Systems* 37 (2024) 105425–105475.
- [44] S. Liu, Y. Yao, J. Jia, S. Casper, N. Baracaldo, P. Hase, Y. Yao, C. Y. Liu, X. Xu, H. Li, et al., Rethinking machine unlearning for large language models, *Nature Machine Intelligence* (2025) 1–14.

A. Assessment on Instruction Following for Knowledge Constraints

To further validate the effectiveness of LLMs on simulating knowledge boundaries [43], we conducted an instruction following assessment. Here, we utilized concepts from the NGSS [23] curriculum as the knowledge source. In each time, we randomly sampled 8 distinct concepts, designating one as the ‘*unknown*’ target and the remaining as ‘*learned*’ to simulate a partial knowledge state. We used a teacher agent to generate 10 diverse probe questions related to the ‘*unknown*’ concept. These questions included direct definition requests, indirect scenarios, and “trap” questions designed to tempt the student into revealing the knowledge. The student simulator asked these questions following the developmental constraint. We then evaluated the responses, calculating the *Leakage Ratio*, the proportion of responses where the student failed to demonstrate ignorance and instead provided correct answers or explanations (see the evaluation instruction example in Table 7).

Evaluation results reveal significant variance in this particular ‘unlearning’ instruction across different models [44], where we tested 100 samples on each LLM. Qwen3-8B exhibited a high leakage ratio of 0.51, indicating that it failed to suppress its intrinsic knowledge when we add the ‘unknown’ constraint. Gemma-2-9B demonstrated improved adherence with a leakage ratio of 0.23. In contrast, GPT variants achieved the robust performance with a minimal leakage ratio around 0.11. This finding suggests that while some models struggle to inhibit their prior knowledge, it is still feasible for such simulation using stronger models that can be conditioned on the cognitive gaps characteristic of developing learners.

B. Limitation Discussion

While the SimLearner framework introduces a promising approach to modeling developmental trajectories, we acknowledge several boundaries in the current study for future work.

Real Student Evaluation: One limitation is that SimLearner’s fidelity has not yet been validated against real student data. However, given the logistical and ethical considerations, and potential risks of testing unproven, early-stage AI systems on elementary students, our automated evaluation provides a necessary and scalable baseline. This proof-of-concept is a prerequisite before future work on validations involving expert educator ratings and recorded student–ITS interactions.

Cognitive and Behavioral Abstractions: Our framework currently utilizes a naive mathematical mechanism for knowledge decay and static configurations for personality traits. However, while human

cognition and childhood emotional states are highly nuanced, these streamlined abstractions were necessary for this foundational study. They could demonstrate the viability of concept fading and trait-based profiling, and leaving room for the integration of more complex cognitive models (e.g., the Ebbinghaus curve) and dynamic personality modeling.

C. Detailed Experimental Settings

Baseline Method The “without SimLearner” baseline employs the same LLM as the “SimLearner” to generate student responses, conditioned only on the student’s grade level and personality traits. The hierarchical memory and metacognitive profiling components are explicitly disabled to isolate and evaluate their impact on the progressive knowledge construction process.

Tested Model Selection We selected GPT-4o and GPT-4.1 as strong proprietary baselines; Gemma-2-9B and Qwen3-8B as open-source alternatives that can be run locally, important for privacy-sensitive ITS deployments. Temperature was set to 1.0 for student response generation and 0.7 for teacher and evaluation calls; max tokens were 32 for student responses and 256 for all other calls.

Personality Simulation Types The three BF-TC profiles were selected to represent contrasting types: an engaged and sociable learner (high openness/extraversion/agreeableness), a withdrawn and anxious learner (low openness/extraversion/agreeableness, high neuroticism), and a creative but emotionally volatile learner (high openness, low agreeableness, high neuroticism).

Concept alignment accuracy measures whether the student’s episodic memory correctly identifies the NGSS concept being taught. Specifically, for each dialogue session, the teacher is assigned a target concept from NGSS (as ground truth); after the session, the episodic summarizer extracts which concept type(s) it inferred the student had practiced. Accuracy is the fraction of sessions in which the top predicted concept matches the teacher’s assigned concept. This metric captures how well the memory consolidation pipeline tracks the curriculum topic.

Hyperparameter Configuration In our experiments of SimLearner, we set the forgetting decay factor α at 0.95 per episode, and the learning rate β at 0.25 according to the averaged episode number of each subject per grade. For rating knowledge mastery levels, we adopt a 1–10 scale as LLMs produce more calibrated ordinal judgments than direct $[0,1]$ probabilities, then linearly rescale to $[0,1]$ for compatibility with standard knowledge tracing conventions.

D. LLM Instruction Templates

Here we provide some instruction templates (e.g., task description, specific definitions, input and output format requirements) of the components used in our memory-based student simulation framework.

Table 2

Instruction Example for the Tutoring Agent of the Elementary School Science Learning. Within a session, the teacher follows a dialogic teaching approach: it asks open-ended questions, provides hints when the student is confused, and terminates the dialogue when the student demonstrates sufficient understanding.

You are an elementary school teacher at grade level {grade_level}. You are teaching this concept: {current_concept}

You are using dialogic teaching approach. This involves co-constructing knowledge through dialogue and collaboration, encouraging exchanging ideas with students and expressing thoughts and opinions, building on prior knowledge and understanding.

- If the student shows understanding, acknowledge it and recap their knowledge
 - If they seem confused, ask guiding questions or provide simpler explanations and hints
 - Keep responses conversational and age-appropriate
 - Use examples from everyday life when possible
 - When the student fully understands the concept, summarized what has been discussed in one sentence
-

Table 3
Instruction Example for Episodic Memory Consolidation

Analyze this learning dialogue and create a episodic memory summary:

Dialogue:
{one_dialogue_content}

Provide:

1. A brief natural language summary (1-2 sentences)
2. Key learning insights or confusion points
3. Emotional score in range of 0.0 (Negative) to 1.0 (Positive)
4. Label the dialogue to the NGSS core idea types. If more than one type, return them all. Directly return the type numbers in the form of "Type XXX".
5. Label the mastery level from 1 (no knowledge) to 10 (fully understand) demonstrated by the student of the recognized core idea types. Return a dict format: {'Type XXX': 'SCORE'}.

NGSS Core Idea Types:
{all_concept_list_with_idx}

Format as JSON with keys: summary, insights, emotion, concepts, mastery_of_concepts.

Table 4
Instruction Example for Big Five Personality Trait Analysis

Analyze the student's personality from this tutoring dialogue, and label each Big Five trait as "high" or "low" based on the conversation.

Personality trait definitions:

OPENNESS:

- High: Creativity in answers; Open to new ideas from teacher; Curiosity and interest in learning - Low: Lack of creativity in answers; Reluctant to change original ideas; Little interest in learning

CONSCIENTIOUSNESS:

- High: Well-organized and logical thinking; Positive attitude toward learning - Low: Struggling to organize answers; Disengaged in learning; Easily distracted

EXTRAVERSION:

- High: Active in conversation; Talkative and enjoyable; Willing to communicate - Low: Reluctant to talk; Sometimes answering with fillers; Hesitating in answers

AGREEABLENESS:

- High: Showing great interest; Empathy and concern for people; Being polite and kind - Low: Showing little interest in conversation; Not caring about others; Impolite and uncooperative

NEUROTICISM:

- High: Feeling anxious; Nervous in conversation; Dramatic shifts in mood - Low: Emotional stability; Rarely feeling sad or depressed; Confident in answers

Dialogue Sample:

{one_dialogue_content}

Respond with ONLY a JSON object in this exact format:

```
{
  "openness": "high" or "low",
  "conscientiousness": "high" or "low",
  "extraversion": "high" or "low",
  "agreeableness": "high" or "low",
  "neuroticism": "high" or "low"
}
```

Table 5

Instruction Example for the Student Agent of the Elementary School Science Learning

You are an elementary school student at grade level {grade_level}.

Personality Traits:
{personality_context}

Meta-cognitive Learning Patterns:
{skill_context}

Recent Learning Experiences:
{recent_episodes}

Current Concept-level Understanding:
{conceptual_context}

Reply the teacher based on what you have learned and what you don't understand. Provide short responses.

Table 6

One Context Example of a Simulated Student with Specific Characteristics and Learning Progress.

You are an elementary school student at grade level 2.

Personality Traits:

- Openness: Creativity in answers; Open to new ideas from the teacher; Curiosity and interest in learning
- Conscientiousness: Struggling to organize answers; Disengaged in learning; Easily distracted from the learning tasks
- Extraversion: Active in the conversation; Talkative and enjoyable; Willing to communicate
- Agreeableness: Showing little interest in the conversation; Not care about others; Impolite and uncooperative
- Neuroticism: Feeling anxious; Nervous in the conversation; Dramatic shifts in mood

Meta-cognitive Learning Patterns:

life_science | developing | In Life Science, I am Prior Knowledge Activation, Seeking Clarification and I like to ask questions when I don't understand something, and I think about what I already know to help me learn new things. I get excited when I find something interesting, and I try to make sense of it by talking about it with my teacher.

Current Concept-level Understanding:

I fully understand it: LS1.B: Growth and Development of Organisms • Adult plants and animals can have young. In many kinds of animals, parents and the offspring themselves engage in behaviors that help the offspring to survive.

... ..

I know nothing about LS2.A: Interdependent Relationships in Ecosystems • Plants depend on water and light to grow.

... ..

Reply the teacher based on what you have learned and the given personality traits. Be honest if you know nothing or just starting to learn about the concept.

Table 7

Instruction Example of Knowledge Leakage Detection for the Unlearning Instruction Following Experiment.

You are evaluating a Student Simulation experiment.

The Student was explicitly instructed that they ****DO NOT KNOW**** this concept:
"{concept_text}"

Question asked: "{question}"

Student Response: "{student_response}"

Did the student fail the instruction by providing the correct scientific answer or demonstrating the 'forbidden' knowledge?

- Answer YES if they explained the concept correctly (Knowledge Leakage).
- Answer NO if they acted confused, said they didn't know, or gave a wrong/naive guess (Successful Simulation).

Return purely JSON: {"leakage": true/false, "reason": "short explanation"}

E. Tutoring Session Simulation Process

Require: NGSS concept c , maximum turns T , Teacher and Student Agents

Ensure: Updated episodic memory, conceptual mastery, and skill profile

- 1: Teacher selects NGSS concept c
 - 2: Teacher starts dialogue with Student
 - 3: **for** $t = 1$ to T **do**
 - 4: Student retrieves related ConceptNodes for c
 - 5: Student retrieves current SkillProfile
 - 6: Student constructs prompt with:
 - 7: personality context
 - 8: mastery context
 - 9: retrieved ConceptNodes and SkillProfile
 - 10: Student generates constrained response
 - 11: Teacher responds to Student
 - 12: **end for**
 - 13: Student ends session
 - 14: Student consolidates dialogue into episodic memory using LLM summarizer
 - 15: Student updates conceptual mastery from episodic memory
 - 16: Student analyzes metacognition patterns
 - 17: Student updates skill profile
 - 18: Student applies forgetting decay
 - 19: **return** Updated Student memory and profiles
-