NeuroSymRead: Symbolic Governance of Neural Generation for Adaptive Dialogic Reading

Md Biplob Hosen[1,*,†], Houbing Herbert Song[1,†], Shuling Yang[1,†] and Lujie Karen Chen[1,†]

[1]University of Maryland Baltimore County, Baltimore, MD, USA

Abstract

The deployment of Large Language Models (LLMs) in early childhood education promises scalable personalization but is hindered by stochastic behaviors, specifically the tendency to lose track of learner proficiency and disrupt long-term story consistency. While LLMs excel at fluent text generation, they often lack the executive logic required for consistent, adaptive reading instruction. To address these reliability and interpretability challenges, we introduce NeuroSymRead, a neuro-symbolic architecture that strictly decouples pedagogical governance from content generation. The system employs a deterministic rule-based controller to provide architectural transparency, enforcing hard constraints on difficulty adaptation and narrative progression, while a separate neural engine handles question generation, scaffolding, and semantic evaluation within these symbolic bounds. We validated the system through an LLM-based synthetic user simulation ($N = 60$ sessions, 600 turns total) using a LLM-as-a-Judge protocol. Results demonstrate that NeuroSymRead effectively eliminates proficiency tracking errors by ensuring all instructional shifts align with predefined pedagogical logic, maintaining high generative quality with an average age appropriateness of 4.66/5 and scaffolding quality of 4.75/5. By generating narrative session audits grounded in the controller's logic, NeuroSymRead enhances interpretability for educators, providing a verifiable glass-box alternative to end-to-end neural tutors for early reading assistance.

Keywords

Neuro-symbolic AI, Large Language Models, Explainable AI, Adaptive Dialogic Reading, Early Literacy.

1. Introduction

The integration of generative artificial intelligence into Intelligent Tutoring Systems (ITS) offers promising capabilities for addressing the 2-sigma problem—the quest to scale the effectiveness of one-on-one human tutoring to all learners [1, 2]. Unlike traditional heuristic-based tutors, LLMs offer fluid, natural language interactions that can adapt to a child's proficiency [3]. However, during the formative early years, the very randomness that enables this fluency presents significant pedagogical risks [4, 5], as young learners are developmentally prone to trusting hallucinated outputs as factual authorities.

A fundamental tension exists in current AI architectures: while LLMs excel at content generation, they often lack the executive control required for consistent planning [6], a deficiency that in educational settings manifests as a failure of pedagogical logic. When deployed end-to-end, this deficiency manifests in two primary failure modes. The first is proficiency tracking instability, where the model loses track of the learner's mastery state over time, causing difficulty levels to fluctuate randomly rather than adapting to the learner's Zone of Proximal Development (ZPD) [7]. The second challenge is narrative inconsistency, a failure of causality where the model inadvertently reveals future plot points or violates sequential dependencies, shattering the immersion required for effective dialogic reading. Furthermore, purely neural systems operate as black boxes, failing to provide educators with the requisite transparency (the ability to inspect the system's underlying logic) or interpretability (the ability for humans to understand the reasoning behind specific interventions).

XAI-Ed 2026: Demystifying AI in Education and Learning Analytics through Explainability, Agency, and Transparency Workshop (XAI-Ed@LAK26), 27 April, 2026, Bergen, Norway

*Corresponding author.

†These authors contributed equally.

mhosen1@umbc.edu (M. B. Hosen); songh@umbc.edu (H. H. Song); syang@umbc.edu (S. Yang); lujiec@umbc.edu (L. K. Chen)

0000-0002-6755-867X (M. B. Hosen); 0000-0003-2631-9223 (H. H. Song); 0000-0002-7185-8405 (S. Yang); 0000-0002-4813-3406 (L. K. Chen)

To resolve this conflict, we introduce NeuroSymRead, a neuro-symbolic architecture that unifies the transparency of symbolic rules with the flexibility of neural networks. We posit that while learner interaction should be fluid, pedagogical governance must be precise [8]. Building on the framework of explainable AI for education, our specific contributions are twofold. First, regarding architecture, we utilize a glass-box approach [9] to propose a symbolic controller that strictly decouples pedagogical state management from the neural generation engine. This enforces hard constraints on difficulty scaling and narrative pacing, ensuring architectural transparency through verifiable logic. Second, regarding evaluation, we implement a validation protocol combining strictly typed user simulations with an LLM-as-a-Judge audit [10]. This evaluates not just response quality, but also pedagogical alignment—the degree to which the system's actions remain faithful to established educational goals and are interpretable to educators via grounded session summaries.

2. Related Work

The development of NeuroSymRead situates itself at the intersection of five distinct research streams: the foundational principles of dialogic reading in early literacy, the historical evolution from model tracing to generative tutors, the critical reliability challenges posed by proficiency tracking instability and narrative inconsistency, the emerging application of neurosymbolic architectures for educational control, and the use of simulated learners for automated evaluation.

2.1. Dialogic Reading in Early Literacy

Research establishes that while reading aloud to children is the single most important activity for building early reading success [11], the quality of interaction is what ultimately drives cognitive and linguistic growth [12]. Unlike traditional shared reading where children listen passively, dialogic reading transforms the interaction into an active dialogue by shifting the child's role from listener to storyteller. However, the efficacy of this method hinges on adaptivity; the facilitator must continuously and dynamically scaffold questions to match the evolving proficiency zone of the child. This requirement for real time pedagogical adjustment poses a significant cognitive load on parents and teachers, often leading to inconsistent implementation in non expert settings.

To mitigate these challenges and structure adaptive scaffolding, dialogic reading relies on two complementary heuristics: the PEER sequence and the CROWD prompting framework [13]. The PEER sequence governs the interaction loop consisting of Prompt, Evaluate, Expand, and Repeat, while the CROWD framework offers specific prompt types (Completion, Recall, Open-ended, Wh-prompts, and Distancing) designed to systematically engage children at varying cognitive levels. Empirical reviews confirm that strict adherence to these frameworks consistently enhances vocabulary acquisition and narrative skills [14], with studies showing that children trained via these methods exhibit significantly higher expressive vocabulary gains compared to control groups [15]. Despite these proven benefits, structural barriers such as limited caregiver training and the unique challenges of multilingual households often prevent widespread access to such high quality literacy practices [16][17].

2.2. From Model Tracing to Generative Tutors

For decades, the gold standard in educational technology has been to replicate the effectiveness of individual human tutoring [1]. Classic ITS, anchored by Anderson's ACT-R cognitive architecture [18] and operationalized through Cognitive Tutors [19], relied on explicit, hand-crafted cognitive models to track learner states. These systems utilize a technique known as model tracing, which compares student actions against a predefined library of production rules, representing both expert and buggy problem-solving paths. While these systems provide high instructional granularity and pedagogical consistency [20], they are severely constrained by an authoring bottleneck stemming from the need for extensive knowledge engineering and cognitive task analysis [21]. The integration of LLMs has the potential to alleviate this bottleneck by enabling the efficient generation of diverse educational content

[3]. However, this shift from structured cognitive architectures to probabilistic generation introduces a new control crisis: the stochastic nature of the model threatens the instructional integrity previously guaranteed by symbolic logic.

2.3. Proficiency Tracking Instability and Narrative Inconsistency

The deployment of end-to-end LLMs in K-12 education introduces specific failure modes that differ fundamentally from traditional software errors. The most significant is proficiency tracking instability, where the model struggles to maintain a persistent, accurate estimate of learner mastery over long interaction windows due to a lack of explicit state management. Without external constraints, this results in the loss of the specific scaffolding strategy required to keep tasks within the learner's Zone of Proximal Development, as purely neural models may fail to track the cold start or evolving mastery of specific skills [22]. Additionally, generative models are prone to narrative inconsistency. In dialogic reading tasks, this manifests as causal hallucinations where the model references plot points that have not yet occurred or contradicts established story logic, disrupting the mental model formation essential for early literacy [4]. Finally, purely neural systems suffer from the black-box problem, failing to provide educators with verifiable audit trails for interventions and violating the transparency requirements proposed for AI in education [23].

2.4. Neurosymbolic Architectures for Educational Control

Neurosymbolic AI seeks to bridge the gap by combining the robust learning and abstraction capabilities of neural networks (system 1) with the reasoning and constraints of symbolic logic (system 2), a paradigm often described as the third wave of AI [24, 25]. Recent studies categorize these hybrid approaches based on the structural integration of their components [26], providing a framework for selecting architectures that balance flexibility with control. In the context of AI in education, this offers a rigorous solution to the reliability problems of generative AI. NeuroSymRead adopts a symbolic governance pattern where a deterministic module acts as a hard constraint on a probabilistic generator [8]. Crucially, while retrieval augmented generation typically restricts what a model knows regarding content boundaries, our architecture restricts how the model behaves pedagogically regarding interaction logic. This aligns with recent efforts to create hybrid systems that leverage the fluency of LLMs while retaining the verifiable logic of traditional ITS [27].

2.5. Simulated Learners and Automated Evaluation

Evaluating adaptive learning systems is difficult due to the ethical and logistical constraints of testing experimental algorithms with vulnerable populations. To address this, the field has increasingly adopted simulated learners to stress test pedagogical strategies. Early work focused on statistical simulations to model learning trajectories and assess policy effectiveness under uncertainty, prioritizing low risk interventions [28]. Building on this, recent research utilizes LLMs to simulate diverse cognitive profiles and human-like misconceptions [29]. Concurrently, the use of LLM-as-a-Judge has emerged as a scalable alternative to human annotation. Rather than relying solely on rigid programmatic metrics, researchers increasingly deploy advanced LLMs as independent evaluators to assess complex qualitative dimensions without the prohibitive costs of human annotation loops [10, 30]. Aligning with these methodological advancements, we utilize Gemini-2.5-Pro as the automated evaluator in our simulated experiments.

3. Methodology

The NeuroSymRead architecture operates as a closed-loop control system that strictly decouples pedagogical decision-making (system 2) from content generation (system 1). The system consists of two primary components: a deterministic symbolic controller that evaluates logical production rules,

and a probabilistic neural engine responsible for natural language generation and semantic evaluation. The workflow is visualized in Figure 1.

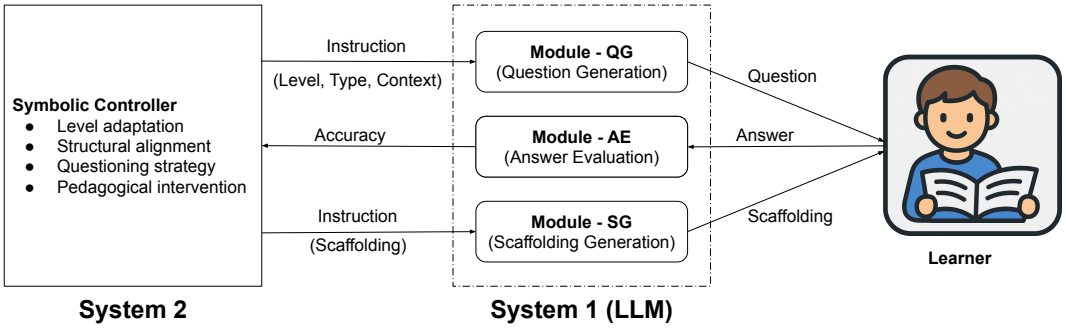

Figure 1: The NeuroSymRead closed-loop architecture. The symbolic controller maintains the context vector and issues deterministic instructions to the LLM. The LLM generates interactive content and semantically evaluates the child's natural language response, returning a discrete accuracy metric (acc_t) that updates the controller's internal state for the next turn.

3.1. The Symbolic Controller (System 2)

To maintain deterministic control over the stochastic neural engine, the controller tracks a state vector C_t at each turn t of a session, defined as:

$$C_t = (\theta_t, P_t, E_t, k_t) \tag{1}$$

where θ_t is the proficiency probability, P_t is the narrative phase, $E_t \in \{0, 1\}$ is a boolean state flag indicating whether the PEER expansion phase has been completed for the current question, and k_t is the rotation counter ensuring orthogonal content coverage aligned with the CROWD framework.

3.1.1. Probabilistic Level Adaptation

Instead of a rigid switch between levels, the system uses a soft handover mechanism governed by θ_t. The difficulty level L_t is determined stochastically based on the learner's dynamic proficiency estimate. We define the levels as follows: Level 1 (Recall) focuses on explicit text retrieval, while Level 2 (Inference) requires synthesizing information to draw conclusions.

Rule R1 (Initialization): The session initializes with an equal probability of Level 1 or Level 2 to calibrate the learner's baseline.

$$\theta_0 = 0.5 \tag{2}$$

Rule R2 (Stochastic selection): At each turn, the system selects the difficulty level by sampling a discrete decision $L_t \sim Bernoulli(\theta_t)$. This converts the soft probabilistic estimate into a hard constraint for the LLM.

$$P(L_t = L2) = \theta_t, \quad P(L_t = L1) = 1 - \theta_t \tag{3}$$

Rule R3 (Proficiency update): The probability of receiving inference questions (θ) adapts based on both accuracy (acc_t) and the number of attempts (att_t). This distinguishes between immediate mastery and scaffolded success, bounded within $[0.1, 0.9]$. These constraints ensure the learner is never entirely deprived of foundational questions nor permanently excluded from challenging ones, maintaining engagement even at performance extremes.

$$\theta_{t+1} = \begin{cases} \min(0.9, \theta_t + 0.10) & \text{if } acc_t = 1 \wedge att_t = 1 \\ \min(0.9, \theta_t + 0.05) & \text{if } acc_t = 1 \wedge att_t = 2 \\ \max(0.1, \theta_t - 0.10) & \text{if } acc_t = 0 \wedge att_t = 2 \end{cases} \tag{4}$$

3.1.2. Questioning Strategy

The controller manages content alignment and question types to ensure narrative coherence and pedagogical variety.

Rule R4 (Structural alignment): Narrative progression is strictly gated by the session progress ratio r_t, calculated as the fraction of the current turn t relative to the maximum allowed turns T_{max} ($r_t = t/T_{max}$). This ratio prevents narrative inconsistency by enforcing phase boundaries:

$$P_t = \begin{cases} \text{Beginning} & \text{if } r_t \leq 0.3 \\ \text{Middle} & \text{if } 0.3 < r_t \leq 0.7 \\ \text{End} & \text{otherwise} \end{cases} \tag{5}$$

Rule R5 (Strategy rotation): To prevent contextual saturation, the question type τ_t is selected via a deterministic round-robin rotation using index k_t. We define the strategy sets for Level 1 (Recall) and Level 2 (Inference) as:

$$\mathcal{S}_{L1} = \{\text{Recall}, \text{Completion}, \text{Wh-prompt}\}$$

$$\mathcal{S}_{L2} = \{\text{Open-ended}, \text{Distancing}\}$$

The specific strategy is then selected from these sets:

$$\tau_t = \begin{cases} \mathcal{S}_{L1}[k_t \pmod 3] & \text{if } L_t = L1 \\ \mathcal{S}_{L2}[k_t \pmod 2] & \text{if } L_t = L2 \end{cases} \tag{6}$$

The rotation counter increments ($k_{t+1} \leftarrow k_t + 1$) after each selection to ensure diverse prompting.

3.1.3. Pedagogical Interaction

The interaction rigorously follows the PEER cycle: the system generates a prompt, evaluates the learner's response, and dynamically provides an explanation or repeat request based on performance.

Rule R6 (Interaction flow): The system dictates the subsequent pedagogical action A_t based on the learner's immediate accuracy ($acc_t \in \{0, 1\}$), their current attempt count (att), and the boolean expansion flag ($E_t \in \{0, 1\}$). To strictly enforce the dialogic loop, the action space is constrained as follows:

$$A_t = \begin{cases} \text{Scaffold} & \text{if } acc_t = 0 \wedge att = 1 \\ \text{Model\_Answer} & \text{if } acc_t = 0 \wedge att \geq 2 \\ \text{Expand\_Repeat} & \text{if } acc_t = 1 \wedge E_t = 0 \\ \text{Affirm\_Next} & \text{if } acc_t = 1 \wedge E_t = 1 \end{cases} \tag{7}$$

To operationalize this framework, these actions execute distinct pedagogical interventions. Scaffold prompts the neural engine to generate a semantic hint without prematurely revealing the answer. If the learner struggles a second time, Model_Answer explicitly states the correct solution and asks the child to repeat it to ensure cognitive encoding. Conversely, an initially correct answer initiates Expand_Repeat, wherein the system elaborates on the learner's response, requests a repetition, and crucially, updates the state flag to $E_t = 1$. Finally, once the expanded concept is successfully repeated, Affirm_Next validates the input, resets the local state variables, and advances the session to the subsequent narrative node.

3.1.4. Transparency Audit

To ensure accountability, the system concludes by converting the session logs into an interpretable format for guardians.

Rule R7 (Parent reporting): Upon reaching the session time limit T_{max}, the controller aggregates the complete state history sequence to synthesize a session report. We define the final report generation $\mathcal{R}$ as a deterministic mapping of the state sequence:

$$\mathcal{R} = \text{Audit}(C_{0 \ldots T_{max}}) \tag{8}$$

This process translates the quantitative metrics, specifically the learning trajectory (θ_t) and intervention points ($acc_t = 0$), into a natural language summary. This provides a verifiable glass-box audit trail of the learner's progress and struggles.

3.2. The Neural Engine (System 1)

While the symbolic controller orchestrates the pedagogical state and interaction flow, the neural engine translates these deterministic instructions into coherent, age-appropriate natural language discourse. Utilizing a foundational large language model, System 1 operates within a modular framework comprising four discrete operational modules that execute dynamically during a reading session. Initially, the context retrieval module ensures narrative coherence by parsing the foundational text into semantically distinct, child-friendly segments. By enforcing strict JSON schemas, this module maintains structural and temporal consistency throughout the session. Once the context is established, the question generation module (Module-QG) formulates queries based on the target difficulty level and specific CROWD strategy designated by System 2. To accommodate early cognitive profiles, this generative phase is strictly bounded by length constraints (e.g., a maximum of eight words) and vocabulary limitations.

Upon receiving the learner's response, the answer evaluation module (Module-AE) functions as the primary sensory mechanism for the closed-loop architecture. It conducts a semantic analysis of the child's unstructured natural language against the narrative context, distilling the evaluation into a discrete binary accuracy classification (correct or wrong) that updates the controller's internal state. Finally, depending on the controller's PEER cycle logic, the scaffolding and feedback module (Module-SG) operationalizes pedagogical interventions. Whether supplying a subtle cognitive cue, explicitly modeling the correct response, elaborating on a concept, or offering positive reinforcement, Module-SG synthesizes supportive dialogue that strictly adheres to the prescribed instructional intent without prematurely disclosing solutions.

4. Experimental Setup

To evaluate the adaptive capabilities and pedagogical efficacy of NeuroSymRead prior to human-subject deployment, we designed an LLM-based synthetic learner simulation.

4.1. Simulation Design

We implemented a closed-loop simulation using two concurrent Gemini-2.5-Pro instances interacting in real-time. The first functioned as the NeuroSymRead tutor, operating without access to the user's hidden state. The second acted as a simulated learner, constrained by a persona prompt to replicate the cognitive and linguistic behaviors of a five-year-old child. To rigorously test the system's adaptive logic across diverse cognitive profiles, we initialized the learner profile with three proficiency tiers: Tier 1 (struggler) with a baseline success probability of $P_{\text{base}} = 0.20$, Tier 2 (improver) with $P_{\text{base}} = 0.50$, and Tier 3 (performer) with $P_{\text{base}} = 0.75$.

To model the pedagogical impact of interventions, we applied a dynamic conditional probability: receiving a scaffold yielded a net performance boost of 0.15 on the subsequent attempt. This parameter represents a conservative adjustment from Zuo et al. [31], whose meta-analysis of digital scaffolding established an effect size of $d = 0.53$ (approximately a 20% percentile gain). We deliberately attenuated this theoretical gain to 15% to account for the hallucination penalty observed by Herklotz et al. [32], acknowledging that imperfect LLM-generated hints may occasionally fail to facilitate cognitive uptake. We executed this setup across $N = 60$ complete interaction sessions, totaling 600 dialogic turns.

4.2. Evaluation Metrics & Auditing

To systematically assess the pedagogical efficacy and reliability of the architecture, we employed an automated semantic auditing protocol using Gemini-2.5-Pro operating under an LLM-as-a-Judge framework. Because the nuances of dialogic reading and remedial scaffolding require deep contextual comprehension rather than simple binary validation, this approach allowed us to rigorously review the complete interaction traces of all evaluation sessions. As detailed in Table 1, the semantic auditing evaluated structural alignment with the intended narrative phases, the age-appropriateness of the generated vocabulary and syntax, the pedagogical quality of the provided scaffolding, and the factual accuracy of the system's internal evaluation module. By deploying this LLM evaluator, we were able to scalably validate the overall quality of the generated educational content, ensuring the system's outputs consistently adhered to the foundational principles of early childhood education.

Table 1
Evaluation metrics used for system validation with LLM-as-a-Judge.

Metric	Type	Definition
Structural alignment	Binary	Strict adherence to narrative phase avoiding inconsistency.
Age appropriateness	Likert (1–5)	Suitability of vocabulary and sentence structure for ages 4–7.
Scaffolding quality	Likert (1–5)	Provision of valid guidance without leaking the answer.
Evaluation quality	Binary	Factual accuracy of the system's judgment of learner answers.

4.2.1. Learning Trajectory Analysis

Beyond static turn-level metrics, we analyzed the system's dynamic responsiveness by tracking its internal proficiency estimates over time. We define the learning trajectory ($\bar{\theta}_t$) as the mean proficiency estimate at turn t, averaged across all N simulation sessions for a given learner profile:

$$\bar{\theta}_t = \frac{1}{N} \sum_{i=1}^{N} \theta_{t,i} \tag{9}$$

where $\theta_{t,i}$ represents the estimated proficiency at turn t in session i. For a proficient learner, $\bar{\theta}_t$ should rapidly ascend toward the mastery threshold (0.9). Conversely, for a struggling learner, the trajectory should safely stabilize near the baseline (0.1), demonstrating that the symbolic controller correctly prioritizes supportive scaffolding over premature difficulty escalation.

5. Findings

5.1. System Performance Metrics

Table 2 presents the system's performance metrics across the 60 evaluation sessions, stratified by learner profile. Performance regarding binary technical constraints varied by category. Structural alignment remained relatively consistent across cohorts, averaging 64.0%, with the highest adherence observed in Tier 1 (66.0%). Conversely, evaluation quality exhibited a strong positive correlation with learner proficiency; as the simulated learner's responses became more coherent, the accuracy of the system's internal judgment improved markedly from 40.8% in Tier 1 to 74.5% in Tier 3. Both age appropriateness (averaging 4.66) and scaffolding quality (averaging 4.75) consistently maintained a good score, displaying minimal variance across the three proficiency tiers. This scaffolding quality directly validates the deterministic enforcement of the PEER cycle. Rather than allowing the neural engine to prematurely disclose correct answers, the symbolic controller successfully forced the model into systematic evaluate and scaffold sequences, ensuring consistent pedagogical interventions regardless of the learner's baseline proficiency.

Table 2
NeuroSymRead performance across learner profiles ($n = 60$ sessions).

Metric	Tier 1 (20 sessions)	Tier 2 (20 sessions)	Tier 3 (20 sessions)	Average (60 sessions)
Age Appropriateness (1–5)	4.72 (SD = 0.74)	4.66 (SD = 0.74)	4.61 (SD = 0.92)	4.66 (SD = 0.80)
Scaffolding Quality (1–5)	4.76 (SD = 0.79)	4.75 (SD = 0.88)	4.73 (SD = 0.78)	4.75 (SD = 0.82)
Structural Alignment (%)	66.0	62.5	63.5	64.0
Evaluation Quality (%)	40.8	59.0	74.5	56.3

**Standard Deviation - SD*

5.2. Post-Session Analytics

Extending beyond real-time interaction metrics, the system architecture generates comprehensive and interpretable summative analytics, as illustrated in Figure 2. This dashboard serves to synthesize raw interaction data into actionable pedagogical insights. The prominent natural language summary, 'Emerging Independence,' effectively operationalizes the current position of the learner within their ZPD. This qualitative assessment is triangulated by granular performance data, specifically an unassisted independent success rate of 60% (6/10 tasks). Crucially, rather than framing the remaining tasks from a deficit perspective, the system highlights a 100% assisted success rate (4/4 tasks), providing empirical evidence of the high receptivity of the learner to dynamic instructional scaffolding. Analytically, this assisted success metric serves as the quantitative validation of the PEER interventions. It demonstrates that when an initial prompt resulted in failure, the subsequent deterministic transition into the expand and repeat phases of the PEER loop successfully bridged the cognitive gap, converting initial struggle into assisted mastery. Furthermore, the learning trajectory visualization models this developmental progression, mapping the probability of mastering inference levels from an initial baseline to a stable proficiency plateau by the conclusion of the session. Ultimately, this visual summary allows to clearly identify what the learner has already mastered and where they still need support.

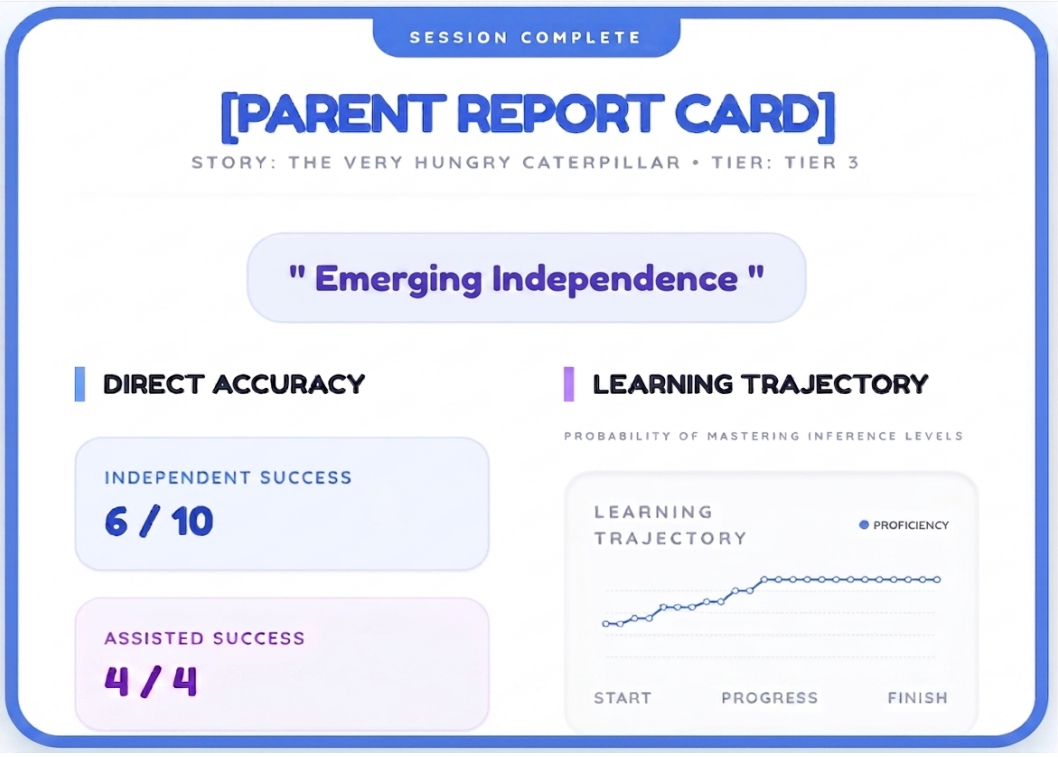

Figure 2: Example of a parent-facing summary generated by the system upon session completion.

5.3. Learning Trajectories

Figure 3 illustrates the evolution of the proficiency parameter θ_t. The improver (Tier 2) profile began at $\theta_0 = 0.5$ and demonstrated a consistent upward gradient, crossing the mastery threshold ($\theta_t > 0.75$) by Turn 8. In contrast, the struggler (Tier 1) profile's trajectory decayed toward the lower bound, stabilizing the difficulty distribution. This stabilization directly dictated the distribution of CROWD strategies; because Tier 1 learners remained at a lower proficiency threshold, the controller adaptively concentrated their interactions on foundational CROWD prompts (Completion, Recall, and Wh-prompts). Furthermore, the advanced learner (Tier 3) profile demonstrated rapid and superior proficiency compared to Tier 2. While the improver profile crossed the 0.75 mastery threshold by Turn 8, the Tier 3 trajectory surpassed this same threshold by Turn 4. Critically, this profile established a very high, stable proficiency plateau near $\theta_t = 0.90$ from Turn 6 onwards, exhibiting accelerated and sustained mastery over the course of the interaction. Consequently, the underlying CROWD distribution for Tier 3 dynamically shifted away from basic recall toward higher-order cognitive tasks (Open-ended and Distancing prompts), proving that the system successfully scales not just raw difficulty, but the specific dialogic mechanisms required to stimulate advanced learners.

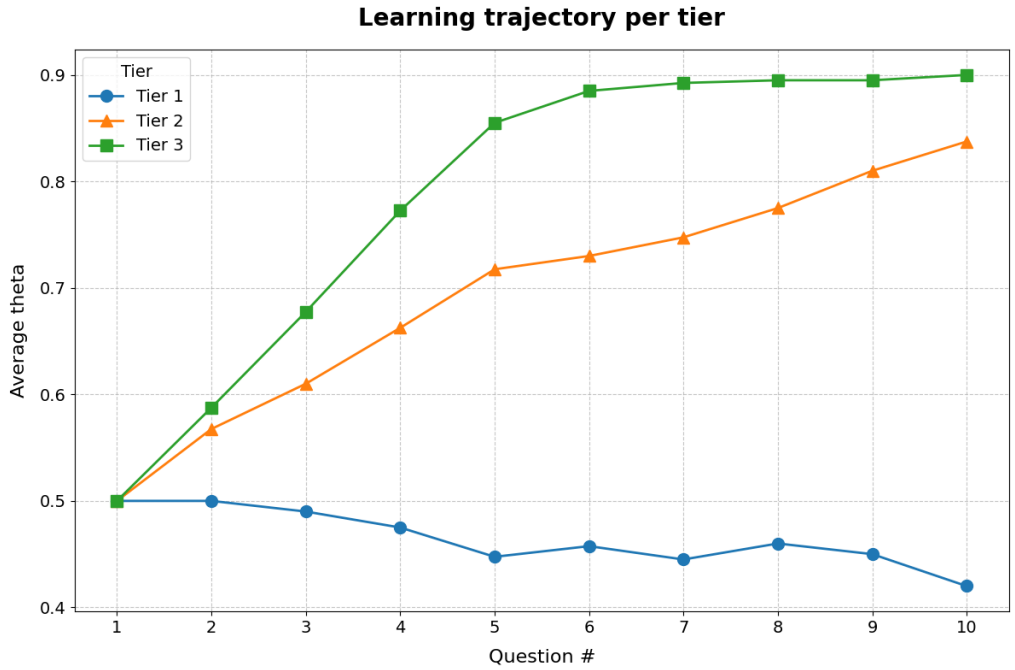

Figure 3: Evolution of the learning trajectory (θ_t) across the three synthetic profiles.

6. Discussion

6.1. Architecture Analysis

The findings validate that our glass-box architecture successfully decouples pedagogical governance from neural generation, preventing stochastic noise from corrupting session logic. By mathematically enforcing the CROWD strategy rotation (Rule R5), the symbolic controller systematically mitigates contextual saturation—a common failure mode wherein purely neural tutors collapse into repetitive questioning loops. The learning trajectories further highlight the system's adaptability: advanced (Tier 3) profiles were consistently challenged with inference-based questions (level 2) to prevent boredom, while improver (Tier 2) profiles demonstrated a steady climb toward mastery through effective scaffolding. Crucially, for struggling (Tier 1) students, the system prioritized motivation over difficulty by locking interactions to foundational questions (level 1). However, despite this restriction in cognitive difficulty,

the underlying CROWD framework ensured these learners still received a diverse pedagogical rotation of completion, recall, and wh-prompts, thereby sustaining engagement and plot comprehension rather than inducing frustration.

Discrepancies in the evaluation metrics reveal the nuances of automated auditing within complex pedagogical environments. The seemingly lower structural alignment scores must be contextualized through the strict enforcement of the PEER framework. While the LLM-as-a-Judge frequently penalized the system for narrative delays, these pauses were not stochastic timing errors; rather, they were deliberate instructional interventions triggered by the expand and repeat phases of the PEER cycle. The architecture appropriately prioritizes remedial dialogic scaffolding over rigid story pacing. Similarly, the divergence in evaluation quality for Tier 1 arises because our child-friendly system intentionally accepts partially correct answers to initiate these exact PEER scaffolding loops, whereas the rigid LLM judge penalizes such responses as binary failures. In contrast, Tier 3 learners predominantly provided fully accurate independent responses, bypassing the extended PEER loops and resulting in high evaluator agreement.

6.2. Limitations and Future Work

While the NeuroSymRead architecture demonstrates strong internal consistency, its current evaluation relies entirely on synthetic simulations ($N = 60$) using Gemini-2.5-Pro, which introduces potential self-preference bias. The simulation depends on heuristic, hard-coded thresholds for proficiency state updates (e.g., +10% for first-attempt success, -10% for repeated failure) and probabilistic constraints (e.g., a 15% scaffolded success rate) that lack grounding in empirical human data. Additionally, we have not yet benchmarked individual modules—such as the neural engine's semantic scoring—against expert human annotations or pure LLM baselines. Future work will prioritize ecological validation through human pilot studies to calibrate these pedagogical parameters, alongside comparative ablation studies to quantify the specific reliability gains of our neuro-symbolic approach.

Furthermore, the current prototype is restricted to text-only interaction, presenting a significant barrier for our pre-literate target demographic (ages 4–7). Authentic digital engagement for this age group fundamentally requires voice-driven interfaces. Future iterations will address this by integrating multimodal inputs while concurrently expanding the proficiency model into a multi-dimensional vector. This enhancement will enable the system to track emotional engagement alongside cognitive mastery, facilitating more dynamic, responsive content expansion during remedial loops.

7. Conclusion

In this work, we introduced NeuroSymRead, a neuro-symbolic framework that resolves the tension between the generative fluency of large language models and the rigid safety constraints required for early literacy. By architecting a glass-box symbolic controller to govern a black-box neural engine, we achieved strict adherence to curriculum logic without sacrificing narrative engagement. Our results demonstrate that this deterministic governance enables a dynamic, soft handover mechanism that adapts to learner proficiency far more reliably than pure prompt engineering. Ultimately, NeuroSymRead establishes a verifiable blueprint for Intelligent Tutoring Systems where pedagogical intent is mathematically guaranteed, rather than merely hallucinated.

References

[1] B. S. Bloom, The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring, Educational researcher 13 (1984) 4–16.

[2] N. Slijepcevic, A. Yaylali, Leveraging "Khanmigo" Generative AI-Powered Tool for Personalized Tutoring to Learn Scientific Concepts, Journal of Teaching and Learning 19 (2025) 155–178.

[3] 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.
[4] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM computing surveys 55 (2023) 1–38.
[5] J. Jiao, S. Afroogh, K. Chen, A. Murali, D. Atkinson, A. Dhurandhar, LLMs and childhood safety: Identifying risks and proposing a protection framework for safe child-LLM interaction, arXiv preprint arXiv:2502.11242 (2025).
[6] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, Advances in neural information processing systems 36 (2023) 11809–11822.
[7] L. S. Vygotsky, M. Cole, Mind in society: Development of higher psychological processes, Harvard university press, 1978.
[8] C. D. Jaldi, E. Ilkou, N. Schroeder, C. Shimizu, Education in the era of Neurosymbolic AI, Journal of Web Semantics 85 (2025) 100857.
[9] X. Ochoa, A. F. Wise, Supporting the shift to digital with student-centered learning analytics, Educational Technology Research and Development 69 (2021) 357–361.
[10] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, Advances in neural information processing systems 36 (2023) 46595–46623.
[11] R. C. Anderson, et al., Becoming a nation of readers: The report of the commission on reading. (1985).
[12] D. H. Arnold, C. J. Lonigan, G. J. Whitehurst, J. N. Epstein, Accelerating language development through picture book reading: replication and extension to a videotape training format., Journal of educational psychology 86 (1994) 235.
[13] C. Pillinger, E. J. Vardy, The story so far: A systematic review of the dialogic reading literature, Journal of Research in Reading 45 (2022) 533–548.
[14] G. J. Whitehurst, F. L. Falco, C. J. Lonigan, J. E. Fischel, B. D. DeBaryshe, M. C. Valdez-Menchaca, M. Caulfield, Accelerating language development through picture book reading., Developmental psychology 24 (1988) 552.
[15] A. C. Hargrave, M. Sénéchal, A book reading intervention with preschool children who have limited vocabularies: The benefits of regular reading and dialogic reading, Early Childhood Research Quarterly 15 (2000) 75–90.
[16] L. Reese, C. Goldenberg, Community literacy resources and home literacy practices among immigrant Latino families, volume 43, Taylor & Francis, 2008, pp. 109–139.
[17] K. He, A. Gastón-Panthaki, D. Zhuo, J. Munsey, M. Zhang, M. Warschauer, StoryPal: Supporting Young Children's Dialogic Reading with Large Language Models, in: Proceedings of the 24th Interaction Design and Children, 2025, pp. 494–511.
[18] J. R. Anderson, ACT: A simple theory of complex cognition., American psychologist 51 (1996) 355.
[19] J. R. Anderson, A. T. Corbett, K. R. Koedinger, R. Pelletier, Cognitive tutors: Lessons learned, The journal of the learning sciences 4 (1995) 167–207.
[20] K. VanLehn, The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems, Educational psychologist 46 (2011) 197–221.
[21] K. R. Koedinger, E. Brunskill, R. S. Baker, E. A. McLaughlin, J. Stamper, New potentials for data-driven intelligent tutoring system development and optimization, Ai Magazine 34 (2013) 27–41.
[22] A. Scarlatos, R. S. Baker, A. Lan, Exploring knowledge tracing in tutor-student dialogues using llms, in: Proceedings of the 15th international learning analytics and knowledge conference, 2025, pp. 249–259.
[23] D. Touretzky, C. Gardner-McCune, F. Martin, D. Seehorn, Envisioning AI for K-12: What should every child know about AI?, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 9795–9799.

[24] A. d. Garcez, L. C. Lamb, Neurosymbolic AI: The 3rd wave, Artificial Intelligence Review 56 (2023) 12387–12406.
[25] G. Marcus, The next decade in AI: four steps towards robust artificial intelligence, arXiv preprint arXiv:2002.06177 (2020).
[26] M. K. Sarker, L. Zhou, A. Eberhart, P. Hitzler, Neuro-symbolic artificial intelligence: Current trends, Ai Communications 34 (2022) 197–209.
[27] R. J. Tong, X. Hu, Future of education with neuro-symbolic AI agents in self-improving adaptive instructional systems, Frontiers of Digital Education 1 (2024) 198–212.
[28] S. Doroudi, E. Brunskill, Fairer but not fair enough on the equitability of knowledge tracing, in: Proceedings of the 9th international conference on learning analytics & knowledge, 2019, pp. 335–339.
[29] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in: Proceedings of the 36th annual acm symposium on user interface software and technology, 2023, pp. 1–22.
[30] J. Wang, Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, J. Xu, J. Qu, J. Zhou, Is chatgpt a good nlg evaluator? a preliminary study, in: Proceedings of the 4th New Frontiers in Summarization Workshop, 2023, pp. 1–11.
[31] M. Zuo, S. Kong, Y. Ma, Y. Hu, M. Xiao, The effects of using scaffolding in online learning: A meta-analysis, Education Sciences 13 (2023) 705.
[32] M. Herklotz, N. Ippisch, A.-C. Haensch, Can we trust LLMs as a tutor for our students? Evaluating the Quality of LLM-generated Feedback in Statistics Exams, arXiv preprint arXiv:2511.04213 (2025).