Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Context-Aware Personalized Education Using Artificial Intelligence

Mariam Mohamed Hussien Sayed

mariam.sayedmohamed@student.guc.edu.eg

German University in Cairo

Aya Salama Abdelhady

German University in Cairo

Research Article

Keywords: intelligent tutoring systems, multi-agent systems, large language models, narrative generation, adaptive scaffolding, elementary science education, educational technology, agentic AI

Posted Date: October 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11039716/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Context-Aware Personalized Education Using Artificial Intelligence

Mariam Mohamed Hussien Sayed[1*] and Aya Salama Abdelhady[1]

[1*]Computer Science and Engineering, German University in Cairo, Cairo, Cairo, Egypt.

*Corresponding author(s). E-mail(s): mariam.sayedmohamed@student.guc.edu.eg;
Contributing authors: aya.abd-elhady@guc.edu.eg;

Abstract

Intelligent educational systems increasingly support personalised learning, yet the integration of narrative-based instruction with mastery-driven adaptation remains underexplored. This paper presents TaleTutor AI, an adaptive AI-driven framework that uses pedagogically structured narrative progression as its primary instructional mechanism. A controller-driven multi-agent pipeline generates educational narratives from ScienceQA lessons following a Noticing–Testing–Using sequence grounded in established learning theory. Chapter-level formative assessment updates a lightweight mastery estimate, which determines subsequent instructional support from extensive scaffolding to increasingly independent learning. Evaluation across seven model configurations on twenty ScienceQA lessons assessed pedagogical compliance, factual consistency, semantic alignment, and readability. The selected GPT-4o outline-first configuration demonstrated strong phase progression, plan adherence, and scaffolding differentiation. Following style-guided simplification, phase progression remained at 4.96, while scaffolding differentiation remained high at 4.90 and 4.95 for the Testing and Using chapters, respectively. Qualitative analysis further showed structurally distinct mastery-conditioned narratives and phase-aligned assessment questions. These findings demonstrate the integration of narrative generation, formative assessment, and mastery-conditioned scaffolding within a unified adaptive tutoring pipeline and provide a foundation for future learner-centred evaluation.

Keywords: intelligent tutoring systems, multi-agent systems, large language models, narrative generation, adaptive scaffolding, elementary science education, educational technology, agentic AI

1 Introduction

Artificial intelligence has substantially expanded the possibilities for personalised education, enabling systems that adapt instruction to individual learner needs, provide formative feedback, and generate educational content at scale. Intelligent Tutoring Systems (ITS) have evolved from rule-based systems toward adaptive environments that model learner knowledge and adjust instructional support accordingly [1]. In parallel, advances in Large Language Models (LLMs) have enabled flexible content generation, open-ended dialogue, and context-sensitive explanation [2], creating new opportunities for narrative-based instruction that embeds educational concepts within meaningful situations.

Despite these advances, an important challenge remains: connecting narrative-based instruction with systematic mastery-driven adaptation. Existing narrative learning systems incorporate different forms of adaptation, assessment, and learner interaction; however, comparatively limited attention has been devoted to using an evolving estimate of learner mastery to systematically shape scaffolding across a theoretically grounded narrative progression.

This paper addresses that gap by presenting and evaluating TaleTutor AI, an adaptive narrative tutoring framework for elementary science education. The framework organises each lesson as a fixed three-phase pedagogical arc grounded in established educational theory; maintains a continuously updated mastery estimate derived from chapter-level formative assessment; and coordinates narrative generation, assessment, and adaptation through a controller-driven multi-agent architecture in which pedagogical decision-making is handled deterministically rather than left to emergent language model behaviour.

The study addresses three research questions:

- **RQ1:** Can an agentic AI tutoring framework reliably generate instructional narratives that maintain a structured three-phase pedagogical arc and produce structurally distinct adaptive scaffolding across mastery tiers?
- **RQ2:** To what extent do AI-generated educational narratives maintain scientific grounding, semantic alignment with lesson content, and age-appropriate readability for elementary learners?
- **RQ3:** What strengths and limitations emerge when evaluating narrative intelligent tutoring across multiple complementary educational quality dimensions?

The remainder of this paper is organised as follows. Section 2 reviews related work in ITS, multi-agent tutoring, narrative ITS, neural story generation, and the pedagogical foundations that motivate the framework's design. Section 3 describes the system architecture, evaluation dataset, and evaluation framework. Section 4 presents results and discussion across all evaluation dimensions. Section 5 concludes with limitations and directions for future work.

2 Background and Related Work

2.1 Intelligent Tutoring Systems

Intelligent Tutoring Systems (ITS) are AI-driven educational systems designed to provide personalised instruction and feedback approximating that of a human tutor [1]. Their primary objective is to adapt instructional content and teaching strategies to the individual learner without requiring continuous human intervention. Early ITS relied on rule-based reasoning and expert-authored knowledge representations; later systems introduced statistical learner models such as Bayesian Knowledge Tracing (BKT) [3] and Deep Knowledge Tracing (DKT) [4] to estimate learner mastery and support adaptive instructional decisions. For instance, OATutor [5] implements a mastery-based ITS loop driven by BKT for problem selection and skill-targeted scaffolding.

The integration of LLMs has substantially expanded ITS capabilities by enabling natural-language interaction, context-aware explanation generation, and flexible educational dialogue [1, 2]. Unlike earlier systems that adapted primarily through problem sequencing and predefined feedback, LLM-based tutors can generate explanations, engage in open-ended questioning, and produce instructional content responsive to specific learner needs. For instance, Open TutorAI [6] demonstrates a modular four-layer architecture, incorporating Retrieval-Augmented Generation (RAG), adaptive quiz flow, and engagement analytics, though validation remains at proof-of-concept level. CourseAssist [7] uses intent classification and course-content retrieval with pedagogical post-processing and reported improved usefulness and accuracy over GPT-4 on a representative question set, though it maintains no explicit learner model. However, concentrating all tutoring responsibilities—planning, generation, assessment, and adaptation—within a single model introduces role conflation, where competence at generation does not reliably imply competence at tutoring [8].

Recent work therefore increasingly adopts agentic and multi-component approaches that distribute instructional responsibilities across specialised modules. PETRA [9], for example, presents a personalised educational tutoring tool that leverages multiple LLM-based components to support learning in recursion and dynamic programming. AgentTutor [10] decomposes tutoring into curriculum planning, learner assessment, strategy revision, reflection, and memory management, achieving 96.9% Pass@1 on HumanEval with GPT-4. GenAL [11] separates long-term learner modelling from local instructional recommendation through a two-agent architecture, outperforming GEHRL, CSEAL, DQN, and KNN baselines. SageJavon [12] combines a Learn–Practice–Evaluate–Support (LPES) framework with Retrieval-Augmented Generation, adaptive knowledge tracing, exercise recommendation, and a multidimensional automatic code-assessment mechanism termed FUPS-Score. The system reported reduced cognitive load and improved knowledge transfer relative to conventional learning tools.

Table 1 summarises these systems across their core architecture, pedagogical basis, assessment and adaptation mechanisms, and key reported results and limitations.

Table 1 Comparison of representative intelligent tutoring systems across architecture, pedagogical basis, assessment and adaptation mechanisms, and key results and limitations

System	Architecture & Mechanism	Pedagogical Basis	Assessment & Adaptation	Results / Limitations
AgentTutor [10] *Multi-turn interactive tutoring*	Curriculum decomposition; learner assessment; strategy search; reflection; memory	Bloom's Taxonomy; cognitive progression; personalised teaching	Cognitive-level assessment; dynamic teaching-strategy adaptation	96.9% Pass@1 on HumanEval with GPT-4; 92.7% with GPT-3.5-turbo; no controlled learner study
GenAL [11] *LLM learning-path recommendation*	Two-agent architecture; learner profiling; reflection; sequential recommendation	Adaptive learning	Deep Knowledge Tracing; profile updates; next-item adaptation	Outperformed evaluated baselines across three real-world datasets; learning-path recommendation focus
Open TutorAI [6] *Open-source modular tutoring*	Learner configuration; conversational tutoring; RAG; learning analytics	Personalised and adaptive learning	Learner-profile-based personalisation; feedback and engagement analytics	Platform-oriented evaluation; limited controlled learning-outcome validation
OATutor [5] *Open-source adaptive tutoring*	Mastery-based ITS loop; skill targets; BKT-driven problem selection	Mastery learning; BKT; scaffolding; immediate feedback	BKT mastery updates; adaptive problem selection	Open research platform; evaluation focuses primarily on system design
CourseAssist [7] *LLM tutor for CS education*	Intent classification; question decomposition; content retrieval; pedagogical post-processing	Hint-based guidance	Intent-dependent response generation; pedagogical constraints	0.94 average evaluation score vs. 0.79 for GPT-4 on 50 Q&A pairs; deployed across six CS courses to 500+ students; no explicit mastery model
SageJavon [12] *Java programming ITS*	LPES framework; enhanced RAG; follow-up tutoring; adaptive knowledge tracing; FUPS-Score	Constructivism; Cognitive Load Theory; Bloom's mastery; Zone of Proximal Development	Knowledge tracing; adaptive exercise recommendation; feedback loops	12-week deployment with 85 students; exercise-recommendation satisfaction above 80%; improved knowledge transfer (5.09 vs. 4.82); evaluated within a Java programming context

2.2 Theoretical and Pedagogical Foundations

The design of TaleTutor AI is grounded in three complementary bodies of educational theory, each motivating a specific design decision in the framework. These theoretical clusters are presented below.

2.2.1 How Learners Build Understanding

The framework's first chapter phase is motivated by constructivist learning theory and Bruner's discovery-oriented perspective. Constructivism, associated with Piaget and Inhelder [13], holds that learning is an active process of knowledge construction rather than passive reception: learners integrate new experiences into existing cognitive structures and reorganise those structures when new experiences do not fit. From this perspective, effective instruction should present learners with opportunities to observe, interpret, and reorganise concepts rather than only receive completed explanations.

Bruner's account of discovery-oriented learning [14] argues that learners benefit when guided to uncover relationships and principles through their own exploration rather than direct exposition. Bruner's Enactive–Iconic–Symbolic (EIS) framework [15] further proposes that understanding develops through progressively more abstract modes of representation, from action-based and image-based experience toward language and other symbolic systems. This framework implies that effective instruction should begin with concrete observable experience before moving toward abstract conceptual explanation.

2.2.2 How Learners Change Incorrect Ideas

The second chapter phase is motivated by misconception research, conceptual change theory, and the productive failure paradigm. Research on prior conceptions in science learning [16] documents that learners frequently hold intuitive explanations of natural phenomena that differ from accepted scientific accounts, and that these prior conceptions are resistant to direct correction because they are grounded in everyday experience. Driver et al. established that effective science teaching must engage learners' existing understanding rather than simply replacing it with correct information.

Conceptual change theory [17] proposes that learners are more likely to adopt a new explanation when they have first become dissatisfied with their current one: the new explanation must be intelligible, plausible, and fruitful relative to the old. This dissatisfaction is a necessary precondition for conceptual change, not a side effect of good instruction. Kapur's productive failure paradigm [18] provides a complementary perspective, demonstrating that initial problem-solving attempts before formal instruction can, under appropriate conditions, deepen conceptual understanding by making knowledge gaps cognitively visible. The core insight is that struggle, when followed by consolidating instruction, prepares the learner to receive that instruction more effectively.

Vygotsky's Zone of Proximal Development (ZPD) [19] establishes the broader scaffolding principle: effective instruction should target tasks that are not yet manageable alone but are achievable with guidance. Wood, Bruner, and Ross [20] operationalised this as contingent scaffolding, in which support is calibrated precisely to what the learner cannot yet do independently and withdrawn as competence increases.

2.2.3 How Support Should Vary with Expertise

The mastery-conditioned scaffolding applied across the Testing and Using phases is informed by Cognitive Load Theory, the worked-example effect, scaffolding fading, and the expertise-reversal effect.

Cognitive Load Theory (CLT), associated with Sweller [21], holds that learning is constrained by the limited capacity of working memory. Effective instructional design should minimise extraneous cognitive load while preserving the generative processing required for actual learning. The worked-example effect [22] is a direct application of this principle: for novice learners, studying fully worked solutions is more effective than unaided problem solving, because worked examples make all reasoning steps explicit and reduce unproductive search that would otherwise consume working memory.

As expertise increases, the value of explicit instructional support diminishes. Scaffolding fading approaches [23] involve gradually reducing the completeness of worked examples as learners develop competence, transitioning from full worked examples to completion problems and eventually to independent problem solving. The expertise-reversal effect [24] documents the boundary condition: instructional guidance that benefits novices can become redundant or even counterproductive for advanced learners, since it imposes cognitive load processing information the learner already possesses. Bjork's desirable difficulties framework [25] provides further support, suggesting that manageable additional effort during learning can strengthen long-term retention and transfer, particularly for more capable learners.

2.2.4 Learner Modelling and Formative Assessment

The mastery-tracking mechanism is grounded in formative assessment theory and lightweight computational approaches to learner modelling. Formative assessment, as characterised by Black and Wiliam [26], refers to assessment whose evidence is used *during* instruction to shape subsequent teaching decisions, rather than only to evaluate achievement after the fact. This principle implies that assessment should be embedded within the instructional loop, with learner responses directly informing adaptation.

Bayesian Knowledge Tracing [3] and Deep Knowledge Tracing [4] provide influential computational approaches to learner modelling in ITS, each estimating skill mastery from response histories using probabilistic or neural methods. However, these approaches require per-skill parameter calibration from historical data that is often unavailable in novel deployment contexts, and their complexity may be disproportionate for short, session-bounded interactions with few assessment opportunities. Prior work has investigated exponential moving average (EMA) approaches as a computationally lightweight alternative for adaptive instructional decisions, finding that simple update rules can serve as practical mastery criteria without the parameter requirements of BKT or DKT [27].

2.3 Narrative Learning and Narrative Intelligent Tutoring

2.3.1 Narrative Intelligent Tutoring Systems

Narrative ITS occupy a middle ground between conventional tutoring and storytelling systems, embedding pedagogical support within a narrative experience. MatheMYTHS [28] uses a conversational agent for co-creative mathematical storytelling and reported learning gains comparable to a human-partner condition in a study with 35 children; however, adaptation remains local to the current turn and does not maintain a persistent mastery estimate across the session, so scaffolding is generic rather than targeted to the specific source of difficulty. AIStoryBot [29] uses story authoring as a medium for AI-literacy education and reported gains in learner confidence, but assessment functions only as a reflective check at the end of the activity. A financial-literacy platform [30] embeds agentic advisory support within gameplay and reported reduced risky behaviour across 134 children, but adaptation responds to behavioural and gameplay state rather than an explicit model of conceptual mastery. Kahaani [31] combines specialised generation and rendering components for multimodal co-creative storytelling but does not maintain a learner model or adapt story progression to educational performance.

Table 2 summarises representative narrative intelligent tutoring systems across their core architecture, pedagogical basis, assessment and adaptation mechanisms, and reported results and limitations. These systems illustrate different approaches to integrating narrative with educational support, ranging from turn-level conversational scaffolding to behavioural adaptation and reflective assessment.

2.3.2 Neural Story Generation and Planning

Long-form neural story generation requires maintaining thematic consistency, preserving entity relationships, and producing a coherent plot that unfolds in a sensible order [32, 33]. Fan et al. [32] introduced hierarchical story generation, demonstrating that first generating a high-level premise before producing the full narrative substantially reduces topic drift relative to non-hierarchical approaches. Yao et al. [33] extended this with the plan-and-write framework, which first generates an explicit storyline representing the intended plot structure and then writes the story conditioned on that plan; their static schema, in which the full plan is generated before any text is written, outperformed dynamic interleaving in producing coherent and topically relevant narratives. These findings suggest that educational storytelling benefits from explicit structural organisation that balances narrative engagement with instructional objectives.

2.4 Research Gap

The literature demonstrates substantial progress in both adaptive tutoring and narrative-based learning; however, their integration remains underexplored. Existing narrative learning systems provide different forms of adaptation and instructional support, but these are typically local, reflective, or behaviour-driven rather than

Table 2 Comparison of representative narrative intelligent tutoring systems across architecture, pedagogical basis, assessment and adaptation mechanisms, and reported results and limitations

System	Core Architecture	Pedagogical Basis	Assessment & Adaptation	Results / Limitations
MatheMYTHS [28] *Voice-based co-creative math-language story tutor*	GPT-4 tutoring loop; story opening and question generation; in-context explanation; separate scaffolding path for weak responses	Question–feedback–scaffolding framework	In-session adaptation; story continuation responds to the child's most recent narrative input; support varies in response to silence, confusion, or brief answers	35 children aged 4–8; AI group pre-test 17.105 → post-test 19.500 vs. control 16.187 → 18.344; definition scores AI +0.631 vs. control +0.031; turn-local adaptation; generic scaffolding; logical inconsistencies and hallucinations reported
AIStoryBot [29] *AI-literacy story-authoring platform*	Introductory AI instruction; story-theme selection; LLM-supported story writing; reflection and knowledge-check stage	Bloom's Taxonomy	Reflection-based checking at the end of the activity	Pre/post gains in perception, confidence, and motivation; students reported narrative steering; limited support for deeper AI understanding; small exploratory sample
John and the Rainbow Coins [30] *Adaptive financial-literacy game / agentic advisory system*	2D game with Quality-Diversity (QD) policy archive; controller agent and 10 advisor agents; real-time state-based advice; post-game reflective analysis	Constructivist game-based learning; adaptive scaffolding; behavioural decision learning; reflection after action	8-dimensional state-space adaptation; advice responds to learner behaviour and gameplay state	134 children aged 7–12; 37% fewer gambling instances; higher learning scores; mixed cognitive-load effects; behavioural rather than conceptual-mastery adaptation; static archive; increased extraneous load

systematically conditioned on an evolving estimate of conceptual mastery. Consequently, comparatively limited attention has been devoted to how learner mastery can progressively shape the pedagogical role and scaffolding of successive narrative chapters.

This gap is particularly relevant given established pedagogical principles of conceptual change, productive failure, Cognitive Load Theory, expertise reversal, and formative assessment, which motivate varying instructional support according to learner understanding. TaleTutor AI addresses this underexplored intersection by combining mastery-driven scaffolding with a structured Noticing–Testing–Using narrative progression and chapter-level formative assessment, and evaluates the resulting narratives across complementary educational quality dimensions.

3 Methodology

3.1 Overview

The proposed framework adopts a controller-driven multi-agent architecture consisting of three cooperating components: a Story Agent, an Assessment Agent, and a Controller Agent. As illustrated in Fig. 1, the Controller Agent orchestrates the overall tutoring workflow by coordinating narrative generation, formative assessment, and adaptive pedagogical decision-making. The workflow proceeds in three stages. First, the Controller Agent initialises the tutoring session by retrieving the selected ScienceQA lesson and establishing an initial learner mastery estimate. Second, inspired by hierarchical plan-and-write story generation [32, 33], the Story Agent generates a global three-chapter narrative outline while the Controller Agent selects the initial pedagogical plan. Third, the framework enters an iterative generation loop in which the Story Agent generates each chapter under the selected pedagogical plan, the Assessment Agent produces a chapter-aligned formative assessment, the learner's response updates the mastery estimate, and the Controller Agent selects the pedagogical plan for the subsequent chapter.

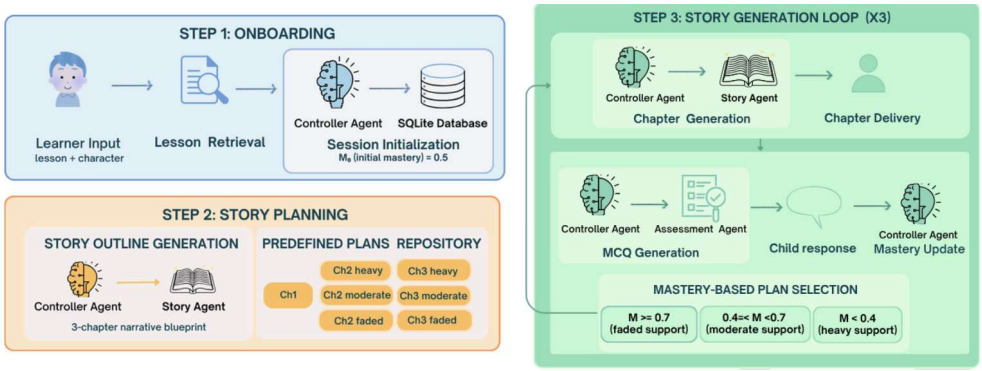

Fig. 1 Overview of the TaleTutor AI tutoring pipeline

3.2 Dataset

The system is grounded in ScienceQA [34], a multimodal multiple-choice question-answering dataset spanning K–12 science topics, including a lecture, reference solution, hint, and optional image for many items. For the purposes of this study, each selected ScienceQA instance together with its associated instructional context is referred to as a lesson. Lessons were filtered to the natural-science domain and grades 2 through 6, retaining only instances that include a lecture, hint, and visual context, and organised hierarchically by topic, category, and skill to support controlled lesson retrieval.

3.3 Three-Chapter Pedagogical Structure

The educational theories presented in Section 2.2 are operationalised through a fixed three-phase instructional structure consisting of *Noticing*, *Testing*, and *Using*. While the generated narrative varies according to the lesson content and learner mastery, the pedagogical objective of each phase remains constant.

The Noticing phase operationalises constructivist learning and discovery-oriented instruction by presenting a concrete phenomenon without immediate explanation, encouraging learners to observe and formulate initial hypotheses.

The Testing phase is informed by research on prior conceptions, conceptual change, and productive failure. It introduces an incomplete or incorrect explanation that is challenged through evidence while withholding the correct scientific concept, with the degree of scaffolding adapted to the learner's estimated mastery.

The Using phase guides successful application of the target concept, with instructional support progressively varied according to mastery. Low-mastery learners receive structured reasoning support resembling worked examples, moderate-mastery learners receive partially guided reasoning, and high-mastery learners encounter largely independent application.

3.4 Predefined Pedagogical Plans

Rather than relying on the language model to infer appropriate pedagogical behaviour at generation time, the system encodes instructional strategy explicitly through a library of predefined chapter plans. Each plan specifies, at a conceptual level, the chapter's pedagogical role, required narrative elements, and constraints that must not be violated, such as withholding premature explanation of the target concept. The Controller Agent selects an appropriate plan according to the current mastery estimate and supplies it to the Story Agent as generation context. Three support tiers are defined for the Testing and Using chapters: *heavy support*, providing the most explicit scaffolding for low-mastery learners; *moderate support*, leaving a final inferential step to the learner; and *faded support*, removing explicit scaffolding for high-mastery learners. Across all tiers, the correct answer is withheld until the Using chapter, operationalising productive failure and conceptual change at the level of narrative structure rather than only at the level of explanation. This separates pedagogical decision-making from narrative generation, allowing instructional strategy to be explicitly specified and reproduced independently of the language model.

3.5 Story Agent

The Story Agent is responsible for transforming lesson content into pedagogically grounded instructional narratives. Inspired by hierarchical story-generation approaches, narrative generation is decomposed into two stages: outline generation and chapter realization. First, the Story Agent constructs a global three-chapter outline (*Noticing*, *Testing*, and *Using*) grounded in the ScienceQA lesson and the selected pedagogical objective. Subsequently, individual chapters are generated using the corresponding outline segment together with the pedagogical plan selected by the Controller Agent. This hierarchical generation strategy separates global narrative planning from

local text generation, improving coherence while ensuring that each chapter satisfies its intended instructional role.

3.6 Controller Agent

The Controller Agent is the orchestration component of TaleTutor AI. Unlike the Story Agent and Assessment Agent, it does not generate instructional content directly. Instead, it manages session state, retrieves the selected lesson, routes lesson and chapter context to the appropriate agent, updates the session-level mastery estimate after each assessment, and selects the pedagogical plan used for the next chapter. This design makes adaptation explicit and centralised: the Story Agent generates under a plan selected before generation, rather than deciding the scaffolding level implicitly during open-ended text generation.

3.7 Mastery Tracking

The Controller Agent maintains a scalar mastery estimate $M \in [0, 1]$, initialised at a neutral prior $M_0 = 0.5$. This places the learner in the moderate-support tier at session start, avoiding the unsupported assumption that the learner has either no prior understanding or complete mastery. After each chapter assessment, the estimate is updated using an exponential moving average:

$$M_t = 0.7M_{t-1} + 0.3\text{score}_t \tag{1}$$

where $\text{score}_t \in \{0, 1\}$ indicates whether the learner answered correctly. This update rule was selected for its simplicity, interpretability, and suitability to short, session-bounded interactions, consistent with prior work on lightweight mastery estimation as an alternative to more elaborate knowledge-tracing models [27]. The weighting gives greater influence to the previous mastery estimate while still allowing the latest assessment response to affect the next scaffolding decision. From the neutral prior, a correct answer raises mastery to 0.65, keeping the learner in moderate support, while an incorrect answer lowers mastery to 0.35, triggering heavy support for the next chapter. This prevents a single correct response from immediately moving the learner to faded support, while still allowing recent performance to influence adaptation. The resulting estimate is mapped to one of three support tiers: $M < 0.4$ selects heavy support, $0.4 \leq M < 0.7$ selects moderate support, and $M \geq 0.7$ selects faded support. This mapping determines the plan selected for the subsequent chapter. Plan selection occurs before chapter generation, so adaptation is proactive rather than a post-hoc transformation of already-generated content.

3.8 Assessment Agent

The Assessment Agent receives the generated instructional chapter together with the corresponding lesson context and produces a grounded formative assessment consisting of a multiple-choice question, answer options, the correct answer, and an explanation. Assessment generation is constrained to information contained within the lesson material and generated narrative to reduce unsupported claims. Generated outputs

undergo structural validation, with fallback to the original ScienceQA assessment when validation repeatedly fails.

3.9 Evaluation Design

The evaluation addresses the three research questions established in Section 1: whether the framework reliably generates narratives that follow the intended three-phase pedagogical arc (RQ1); whether AI-generated narratives maintain scientific grounding, semantic alignment with lesson content, and age-appropriate readability (RQ2); and what strengths and limitations emerge when educational quality is assessed across multiple complementary dimensions (RQ3). Evaluation was conducted across seven model configurations on 20 lessons sampled from ScienceQA across Grades 2–6, with chapters generated at three fixed representative mastery levels (0.3, 0.5, and 0.8) to probe each support tier. Following the comparative evaluation perspective also adopted in recent natural-language processing studies contrasting alternative AI approaches [35], the configurations were evaluated across multiple dimensions rather than selected on the basis of a single metric.

Four complementary metric families were used. *Factual consistency* was assessed using MiniCheck [36], which evaluates whether generated claims are supported by the lesson source through entailment-style checking. *Semantic alignment* was assessed using Sentence-BERT (SBERT) cosine similarity [37], which measures conceptual closeness between generated text and lesson content using sentence embeddings. Learned embedding representations have also been applied as task-specific representations in contemporary natural-language processing classification settings [38]. In the present work, however, embeddings are used exclusively as an evaluation mechanism for measuring semantic alignment rather than as features for a predictive classifier. *Readability* was assessed using three formula-based indices—the Coleman-Liau index, the Gunning Fog index, and the Flesch-Kincaid grade level—interpreted as indicators of surface-level linguistic complexity rather than complete measures of age-appropriateness [39]. The grade associated with each ScienceQA lesson was used as the target reading level, and readability alignment was measured as the absolute difference between the estimated readability grade of the generated text and the corresponding lesson grade.

Pedagogical quality was assessed using Atla Selene Mini, a small language-model-based rubric judge [40], across three criteria: *plan adherence*, whether a chapter satisfies the structural requirements of its assigned tier; *phase progression*, whether the three-chapter arc preserves the Noticing–Testing–Using sequence; and *scaffolding differentiation*, whether the three mastery-tier versions of a chapter are structurally distinguishable. Plan adherence was evaluated for each generated chapter and mastery condition individually, with scores averaged across the evaluated outputs. Phase progression was evaluated over complete three-chapter trajectories to assess preservation of the Noticing–Testing–Using sequence. Scaffolding differentiation was evaluated separately for Chapters 2 and 3 by comparing the heavy-support ($M = 0.3$), moderate-support ($M = 0.5$), and faded-support ($M = 0.8$) versions, with scores averaged across the evaluated lessons. All pedagogical criteria were scored on a five-point rubric, with higher scores indicating stronger satisfaction of the corresponding criterion.

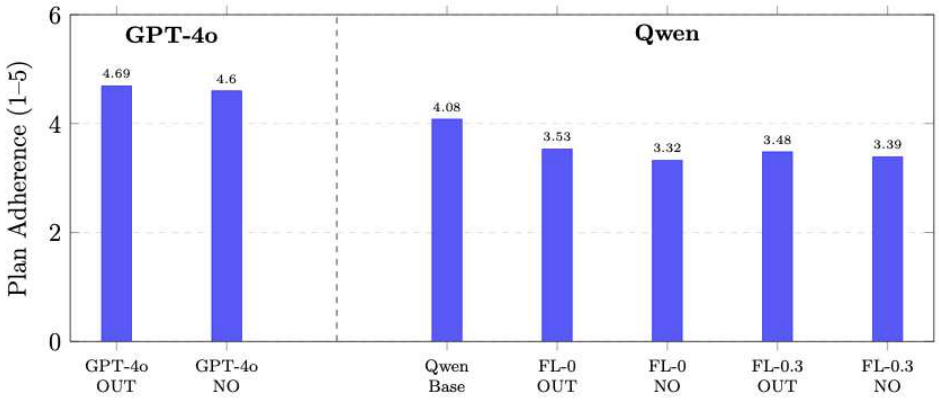

Fig. 2 Plan adherence scores across the seven evaluated Story Agent configurations. OUT denotes outline-first generation, NO denotes direct generation, and FL denotes Fact-Lock prompting. Higher scores indicate stronger compliance with the predefined pedagogical plan

Seven Story Agent configurations were evaluated to isolate the effects of model choice, hierarchical outline generation, and grounding strategy. Two configurations used GPT-4o with and without outline-first generation (GPT-4o OUT and GPT-4o NO). Five configurations used Qwen2.5-VL-7B: a baseline outline-first configuration (Qwen Base) and four Fact-Lock variants combining outline-first (OUT) or direct (NO) generation with two decoding temperatures ($T = 0.0$ and $T = 0.3$), denoted FL-0 OUT, FL-0 NO, FL-0.3 OUT, and FL-0.3 NO. Fact-Lock refers to a structured grounding mechanism that injects lesson constraints into the generation prompt before story generation.

Assessment quality was subsequently evaluated using SBERT similarity at two levels: concept alignment between the generated assessment target and the corresponding ScienceQA skill, and question–answer pair alignment between the generated assessment and the source lesson assessment.

4 Results and Discussion

4.1 RQ1: Pedagogical Coherence and Scaffolding Differentiation

The primary question is whether the framework reliably generates narratives that maintain the intended three-phase pedagogical arc and produce structurally distinct scaffolding across mastery tiers. Figures 2–4 compare the seven Story Agent configurations across the three pedagogical evaluation metrics.

GPT-4o with outline-first generation achieved the strongest overall pedagogical performance, with a plan adherence score of 4.69, perfect phase progression (5.00), and the highest Chapter 3 scaffolding differentiation (4.90). Removing the outline resulted in only a slight reduction across all three metrics, indicating that GPT-4o maintained

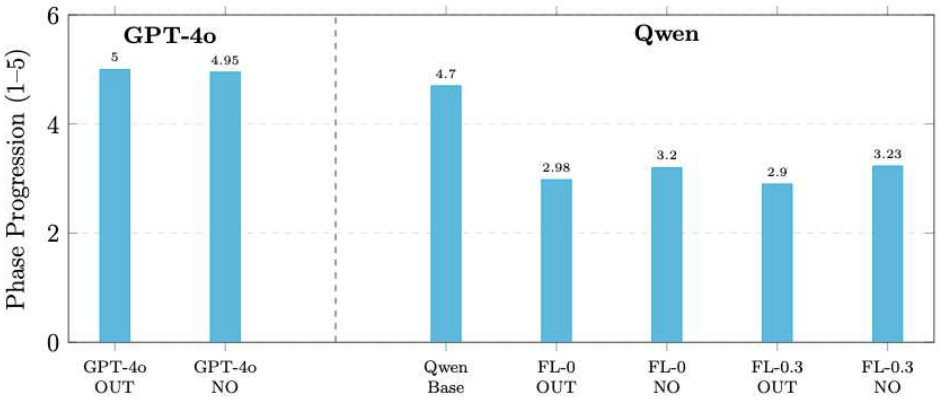

Fig. 3 Phase progression scores across the evaluated Story Agent configurations. GPT-4o maintained nearly perfect execution of the Noticing–Testing–Using pedagogical sequence, whereas Qwen configurations showed lower consistency

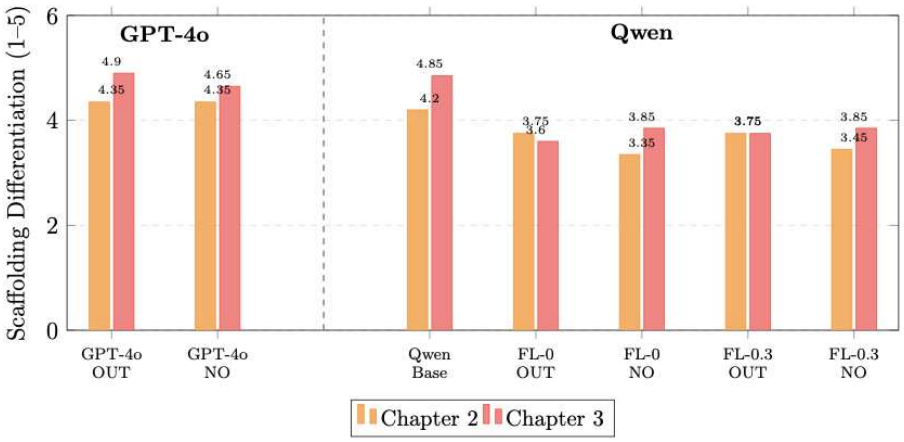

Fig. 4 Scaffolding differentiation scores for Chapters 2 and 3 across all evaluated Story Agent configurations. Higher scores indicate clearer separation between heavy, moderate, and faded-support pedagogical plans

the intended Noticing–Testing–Using structure even without explicit hierarchical planning. Nevertheless, the consistent improvement provided by outline-first generation supports the use of a planning stage for maximising pedagogical consistency.

Within the Qwen2.5-VL-7B family, the baseline configuration, which used outline generation without Fact-Lock prompting, consistently outperformed all Fact-Lock variants. It achieved a plan adherence score of 4.08, phase progression of 4.70, and Chapter 3 scaffolding differentiation of 4.85, approaching GPT-4o performance on

some pedagogical dimensions. Introducing Fact-Lock, however, substantially reduced pedagogical quality across all evaluated settings.

These results suggest that the primary limitation is not decoding strategy but the interaction between the additional grounding constraints imposed by Fact-Lock and the model's capacity to simultaneously maintain narrative coherence and pedagogical structure. When the model is required to satisfy increasingly restrictive prompt constraints while generating a structured instructional narrative, the resulting chapters become less consistent with the intended pedagogical plans. This supports the architectural decision to externalise pedagogical decision-making to the Controller Agent rather than relying solely on prompt engineering within the generative model.

Comparing the two GPT-4o configurations further shows that outline-first generation consistently improves pedagogical performance, particularly for phase progression and Chapter 3 scaffolding differentiation. This observation aligns with previous work on hierarchical and plan-and-write story generation [32, 33], where global planning improves structural consistency during long-form generation. In the present framework, the intermediate outline primarily strengthens higher-level pedagogical organisation rather than individual chapter quality, resulting in clearer separation between instructional support tiers while preserving the intended three-phase narrative progression.

4.2 RQ2: Scientific Grounding, Semantic Alignment, and Readability

4.2.1 Semantic Alignment and Factual Consistency

The GPT-4o variants exhibited a diagnostically meaningful SBERT trajectory across the three chapters: a dip in the Testing chapter followed by recovery in the Using chapter. Figure 5 reports mean SBERT scores per chapter for selected configurations.

The dip-and-recovery pattern in GPT-4o configurations is consistent with the intended pedagogical design: the Testing phase deliberately distances itself from lecture vocabulary to support productive struggle, before the Using phase returns to applying the correct concept with its associated terminology. Among the Qwen2.5-VL-7B configurations, the Fact-Lock variants consistently achieved higher Chapter 2 SBERT similarity than the GPT-4o configurations, with the strongest result obtained by FL-0 OUT (0.586). The Qwen baseline exhibited a similar trend, although to a lesser extent, indicating that increased lexical overlap was primarily associated with the additional grounding constraints introduced by Fact-Lock rather than with the underlying model alone.

These results indicate that higher semantic similarity is not necessarily associated with higher pedagogical quality. Although several Qwen Fact-Lock configurations achieved greater lexical and semantic overlap with the source lesson during the Testing phase, they simultaneously obtained substantially lower scores for plan adherence and phase progression. Conversely, the GPT-4o configurations exhibited lower semantic similarity during the Testing phase while achieving the strongest pedagogical performance. This suggests that semantic similarity metrics should be interpreted together with pedagogical evaluation rather than in isolation.

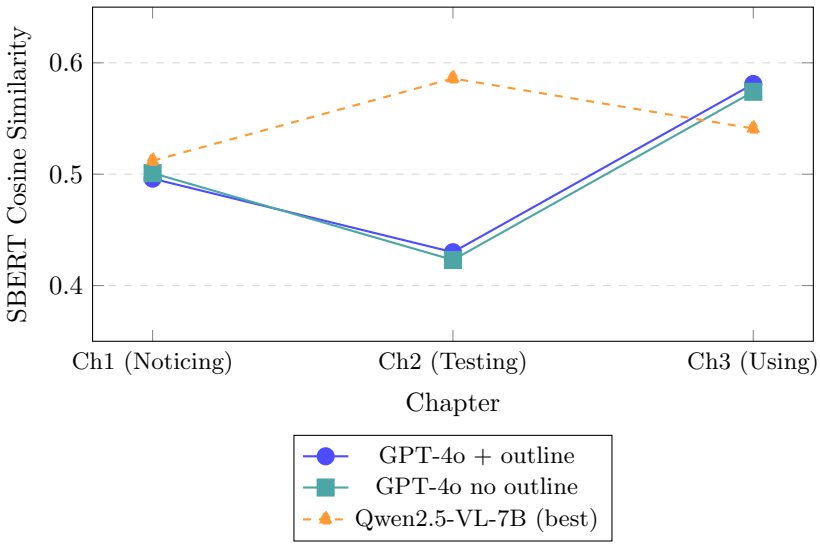

Fig. 5 SBERT cosine similarity between generated chapter text and lesson source across the three chapters, for selected configurations

MiniCheck factual consistency scores for the GPT-4o configurations in the Testing chapter were low in absolute terms (0.010–0.055). Closer inspection revealed that mean entailment probability remained near 0.47–0.48, suggesting uncertainty in the entailment judgments. This pattern may reflect, in part, a mismatch between narrative paraphrase and strict claim-level entailment checking. These results reinforce the methodological point above: entailment-based fact-checking tools may require careful interpretation in narrative educational settings where paraphrase and creative retelling are intentional design choices.

4.2.2 Readability

Readability presented a different trade-off from the pedagogical evaluation. Several Qwen2.5-VL-7B configurations produced text closer to the target reading level than GPT-4o. However, because GPT-4o consistently achieved substantially higher pedagogical quality across plan adherence, phase progression, and scaffolding differentiation, readability optimisation was performed only for the selected GPT-4o configuration using a lightweight style-guidance instruction in the prompt.

Figure 6 shows the absolute readability difference from the target grade level before and after style-guided simplification.

Style-guided simplification reduced the readability gap across all mastery tiers, with the largest improvements occurring in the Chapter 3 application phase. The readability difference decreased by approximately 0.9–1.5 grade levels while preserving the pedagogical properties of the selected configuration. After simplification, phase progression remained at 4.96, and scaffolding differentiation remained high at 4.90 and 4.95 for Chapters 2 and 3, respectively. These results suggest that linguistic register can be adjusted largely independently of pedagogical structure.

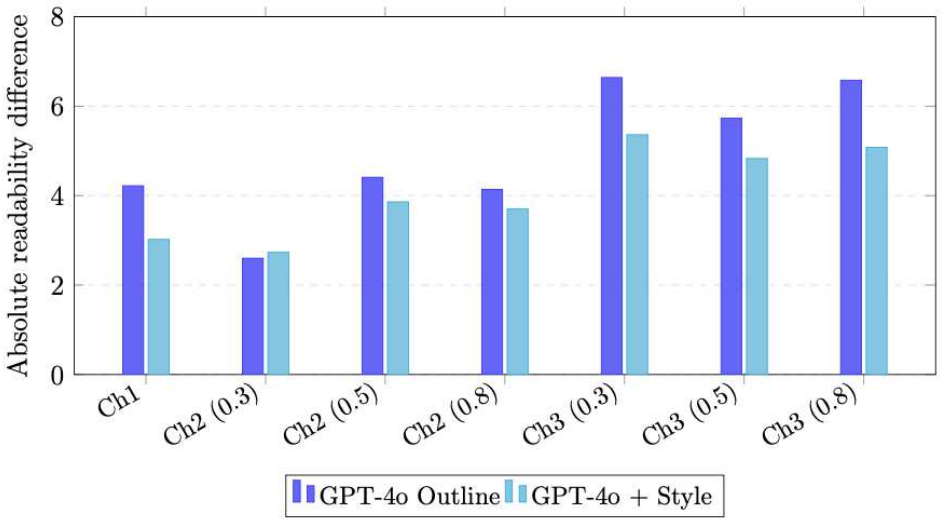

Fig. 6 Absolute Coleman-Liau readability difference from the target grade level before and after style-guided simplification. Lower values indicate closer alignment with the target reading level. Style guidance consistently reduced the readability gap across all mastery tiers while preserving pedagogical quality

Despite these improvements, the *Using* chapter remained above the target reading level. This may reflect a trade-off in the present narrative tutoring design: the worked-example-like structure provided to low-mastery learners requires explicit reasoning steps and concept terminology, which may increase linguistic complexity. Completely resolving this tension may require a dedicated post-generation simplification stage or vocabulary-controlled planning, and ultimately requires validation through learner studies to determine whether the remaining readability gap meaningfully affects comprehension.

4.3 Assessment Agent Alignment

Figure 7 illustrates the SBERT alignment scores for Assessment Agent outputs across the three pedagogical phases.

Concept alignment—between the generated target concept and the skill in the dataset—was highest in Chapters 1 and 3, with mean SBERT similarities of 0.848 and 0.835, respectively, while Chapter 2 obtained a lower but still moderate similarity of 0.635. This pattern is consistent with the pedagogical design of the framework, in which the first and final chapters focus on introducing and applying the target concept, whereas the Testing chapter emphasizes reasoning through an intermediate misconception.

Question–answer pair alignment followed a similar pattern, with lower similarity in Chapter 2 than in Chapters 1 and 3. This is expected because the generated assessment is designed to evaluate the reasoning developed within the current chapter rather than

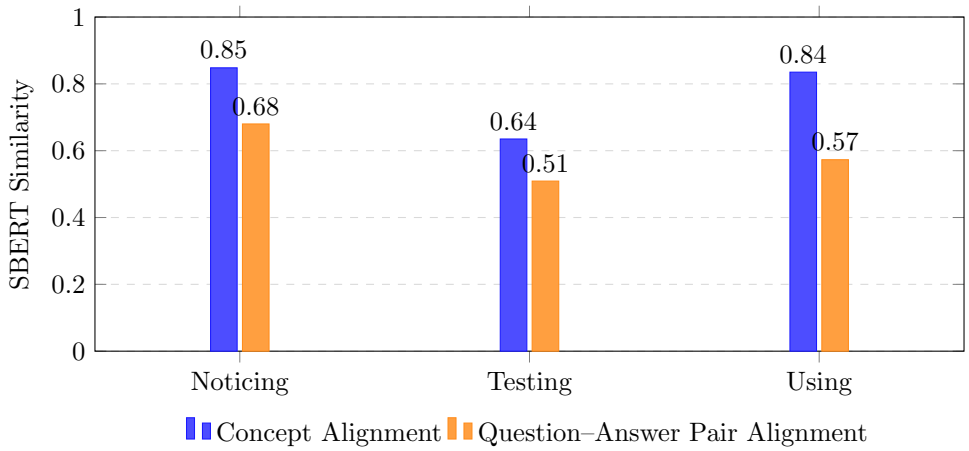

Fig. 7 SBERT alignment scores for Assessment Agent outputs across the tutoring session

reproduce the original ScienceQA question. Consequently, lower similarity does not necessarily indicate lower assessment quality, but may instead reflect the intended pedagogical adaptation of the generated question.

Overall, these findings suggest that reference-based similarity metrics should be interpreted together with the pedagogical role of each chapter. Assessments generated for different instructional purposes are not expected to exhibit the same degree of similarity to the original dataset while still remaining aligned with the underlying lesson concept.

4.4 RQ3: Qualitative Evidence and Cross-Dimensional Synthesis

To examine whether the quantitative metrics corresponded to observable pedagogical differences, a qualitative case study was conducted on a Grade 2 biology lesson requiring learners to distinguish a bird from an amphibian based on observable traits. This case was selected because the concept relies on identifying and comparing distinguishing animal traits, providing a clear setting in which differences in instructional support and reasoning across mastery tiers can be examined.

The generated narratives exhibited clear structural differences across the three mastery tiers that were consistent with their assigned pedagogical plans. In the Testing chapter, the heavy-support version explicitly guided the learner's attention toward the relevant distinguishing feature, the moderate-support version left one inferential step unresolved, and the faded-support version provided no explicit guidance, requiring the learner to infer the scientific distinction independently from the narrative. Similar differences were observed in the Using chapter, where the three versions progressively transitioned from a fully worked reasoning process to independent concept application. Likewise, the Assessment Agent generated questions that reflected the instructional objective of each chapter, progressing from initial concept recognition to misconception correction and finally concept application.

Taken together, the quantitative and qualitative analyses reveal both strengths and limitations of evaluating narrative intelligent tutoring systems across complementary educational quality dimensions. A key strength is that the pedagogical evaluation captured differences that were also observable qualitatively in the generated narratives. Differences in rubric-based scaffolding scores corresponded to meaningful changes in the amount of guidance provided across mastery tiers, while the Assessment Agent produced questions consistent with the instructional purpose of each phase. The readability analysis further showed that linguistic complexity could be reduced through style-guided simplification while largely preserving the underlying pedagogical structure.

At the same time, the analysis exposes important limitations of individual automatic metrics. Grounding and semantic similarity scores cannot independently characterise pedagogical quality because deliberate instructional strategies, such as productive struggle, may reduce similarity to the source material without indicating poorer instruction. Readability and pedagogical quality likewise represent distinct dimensions: text that is easier to read is not necessarily better aligned with the intended instructional strategy. Consequently, no individual metric used in this study provides a complete measure of educational quality.

Overall, the findings indicate that the principal strength of a multi-dimensional evaluation is its ability to capture distinct and complementary properties of educational narratives, while its principal limitation is that individual automatic metrics remain sensitive to only particular aspects of quality. Evaluating narrative intelligent tutoring systems therefore benefits from combining grounding and semantic measures with pedagogical evaluation and qualitative inspection rather than interpreting any single metric in isolation.

5 Conclusion

This paper presented an empirical evaluation of TaleTutor AI, an adaptive agentic narrative tutoring framework that combines a controller-driven multi-agent architecture with pedagogically grounded narrative generation and mastery-conditioned scaffolding. The framework was designed to investigate whether explicit pedagogical planning and adaptive orchestration enable the generation of educational stories that are pedagogically coherent, scientifically grounded, and responsive to learner mastery.

The evaluation showed that GPT-4o with outline-first generation achieved the strongest overall pedagogical performance, consistently maintaining the intended Noticing–Testing–Using instructional structure while producing clearly differentiated scaffolding across mastery levels. The results further showed that grounding and semantic similarity metrics should be interpreted alongside pedagogical evaluation, as lower similarity during the Testing phase may arise from deliberate instructional design rather than generation errors. Although several alternative configurations achieved better readability, style-guided simplification improved the readability of the selected configuration while largely preserving its pedagogical properties.

The primary limitation of this study is the absence of a human learner evaluation. While the experiments demonstrate that the framework generates pedagogically structured, adaptive narratives, they do not establish whether these design choices translate into improved learning outcomes or learner engagement. Evaluating the framework with elementary-aged learners therefore remains the most important direction for future work.

Overall, this work demonstrates that combining explicit pedagogical planning with controller-driven multi-agent orchestration provides a promising direction for generating adaptive educational narratives whose quality can be evaluated across multiple pedagogical and linguistic dimensions.

6 Future Work

Human learner evaluation. A controlled study with elementary-aged learners, incorporating pre- and post-assessments, engagement measures, and qualitative feedback, is needed to determine whether the pedagogically grounded adaptations translate into measurable improvements in learning outcomes and learner experience.

Enhanced learner modelling. Future versions of the framework could replace the neutral mastery prior with a diagnostic pre-assessment and extend the current session-based exponential moving average to persistent cross-session learner models, for example using Bayesian Knowledge Tracing or other lightweight mastery-estimation approaches.

Evaluation and readability. The findings suggest opportunities for developing phase-aware evaluation metrics that better account for pedagogically staged narratives, as well as investigating post-generation simplification and vocabulary-controlled planning to further reduce linguistic complexity while preserving instructional structure.

Generalisation. Future work will evaluate the framework across additional subjects beyond elementary science and investigate branching narrative structures in which learner responses influence subsequent story events while maintaining pedagogical coherence.

Acknowledgements. The authors thank the German University in Cairo for supporting this research. The authors also thank the developers of ScienceQA for making the dataset publicly available, and the developers of Atla Selene Mini and MiniCheck for making their evaluation tools accessible to the research community.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Funding
The authors received no funding for this research.

Authors' contributions
M.M.H.S. designed and implemented the framework, conducted the experiments, analysed the results, and drafted the manuscript. A.S.A. supervised the research, contributed to the study methodology and interpretation of the results, and reviewed and revised the manuscript. Both authors read and approved the final manuscript.

Data availability
The ScienceQA dataset used in this study is publicly available at https://scienceqa.github.io.

Code availability
The implementation code for TaleTutor AI is not publicly available. The code may be shared upon reasonable request to the corresponding author.

References

[1] Zerkouk, M., Mihoubi, M., Chikhaoui, B.: A Comprehensive Review of AI-Based Intelligent Tutoring Systems: Applications and Challenges. Preprint at https://arxiv.org/abs/2507.18882 (2025)

[2] Dahal, C., Chen, J., Ye, M., Xiong, Z.: Multi-Agent Systems in Education: A Survey from the Trustworthiness Perspective. TechRxiv preprint. https://doi.org/10.36227/techrxiv.176704820.01980102/v1 (2025)

[3] Corbett, A.T., Anderson, J.R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction **4**(4), 253–278 (1995) https://doi.org/10.1007/BF01099821

[4] Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., Sohl-Dickstein, J.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems, vol. 28, pp. 505–513. Curran Associates, Inc., Red Hook, NY (2015)

[5] Pardos, Z.A., Tang, M., Anastasopoulos, I., Sheel, S.K., Zhang, E.: Oatutor: An open-source adaptive tutoring system and curated content library for learning sciences research. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), pp. 1–17. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3581574

[6] El Hajji, M., Ait Baha, T., Dakir, A., Fadili, H., Es-Saady, Y.: Open tutorai: An open-source platform for personalized and immersive learning with generative ai. arXiv preprint arXiv:2602.07176 (2026)

[7] Feng, T., Liu, S., Ghosal, D.: Courseassist: Pedagogically appropriate ai tutor for computer science education. In: Proceedings of the 2024 ACM Virtual Global

Computing Education Conference (SIGCSE Virtual), pp. 310–311. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3649409.3691094

[8] Macina, J., Daheim, N., Hakimi, I., Kapur, M., Gurevych, I., Sachan, M.: Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 204–221. Association for Computational Linguistics, Suzhou, China (2025). https://doi.org/10.18653/v1/2025.emnlp-main.11

[9] Hewedy, D.W., Abdelhady, A.S.: Petra: A personalized educational tutoring tool for recursive assistance leveraging multi-model llms for dynamic programming learning. In: 2025 15th International Conference on Electrical Engineering (ICEENG), pp. 1–6. IEEE, Cairo, Egypt (2025). https://doi.org/10.1109/ICEENG64546.2025.11031311

[10] Liu, Y., Song, Z., Lou, J., Wu, C., Li, J.: AgentTutor: Empowering Personalized Learning with Multi-Turn Interactive Teaching in Intelligent Education Systems. Preprint at https://arxiv.org/abs/2601.04219 (2026)

[11] Lv, R., Liu, Q., Gao, W., Zhang, H., Lu, J., Zhu, L.: GenAL: Generative agent for adaptive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 577–585 (2025). https://doi.org/10.1609/aaai.v39i1.32038

[12] Zhao, H., Wu, Y., Lu, Z., Yu, X., Miao, W., Chen, L.: Sagejavon: A scalable ai tutor for personalized programming learning. Information Processing & Management **63**(5), 104605 (2026) https://doi.org/10.1016/j.ipm.2025.104605

[13] Piaget, J., Inhelder, B.: The Psychology of the Child. Basic Books, New York (1969)

[14] Bruner, J.S.: The act of discovery. Harvard Educational Review **31**(1), 21–32 (1961)

[15] Bruner, J.S.: Toward a Theory of Instruction. Harvard University Press, Cambridge, MA (1966)

[16] Driver, R., Squires, A., Rushworth, P., Wood-Robinson, V.: Making Sense of Secondary Science: Research Into Children's Ideas. Routledge, London (1994)

[17] Posner, G.J., Strike, K.A., Hewson, P.W., Gertzog, W.A.: Accommodation of a scientific conception: Toward a theory of conceptual change. Science Education **66**(2), 211–227 (1982) https://doi.org/10.1002/sce.3730660207

[18] Kapur, M.: Examining productive failure, productive success, unproductive failure, and unproductive success in learning. Educational Psychologist **51**(2), 289–299 (2016) https://doi.org/10.1080/00461520.2016.1155457

[19] Vygotsky, L.S.: Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA (1978)

[20] Wood, D., Bruner, J.S., Ross, G.: The role of tutoring in problem solving. Journal of Child Psychology and Psychiatry **17**(2), 89–100 (1976) https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

[21] Sweller, J.: Cognitive load theory. In: Mestre, J.P., Ross, B.H. (eds.) Psychology of Learning and Motivation vol. 55, pp. 37–76. Academic Press, Oxford (2011). https://doi.org/10.1016/B978-0-12-387691-1.00002-8

[22] Sweller, J., Cooper, G.A.: The use of worked examples as a substitute for problem solving in learning algebra. Cognition and Instruction **2**(1), 59–89 (1985) https://doi.org/10.1207/s1532690xci0201_3

[23] Renkl, A., Atkinson, R.K., Maier, U.H., Staley, R.: From example study to problem solving: Smooth transitions help learning. Journal of Experimental Education **70**(4), 293–315 (2002) https://doi.org/10.1080/00220970209599510

[24] Kalyuga, S., Ayres, P., Chandler, P., Sweller, J.: The expertise reversal effect. Educational Psychologist **38**(1), 23–31 (2003) https://doi.org/10.1207/S15326985EP3801_4

[25] Bjork, R.A.: Memory and metamemory considerations in the training of human beings. In: Metcalfe, J., Shimamura, A.P. (eds.) Metacognition: Knowing About Knowing, pp. 185–205. MIT Press, Cambridge, MA (1994)

[26] Black, P., Wiliam, D.: Inside the black box: Raising standards through classroom assessment. Phi Delta Kappan **80**(2), 139–148 (1998)

[27] Pelánek, R., Řihák, J.: Experimental analysis of mastery learning criteria. In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17), pp. 156–163. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3079628.3079667

[28] Zhang, C., Liu, X., Ziska, K., Jeon, S., Yu, C.-L., Xu, Y.: MatheMYTHS: Leveraging large language models to teach mathematical language through child-AI co-creative storytelling. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, pp. 1–14. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3613904.3642647

[29] Han, A., Han, S.: Empowering children's AI literacy through co-creating stories with LLM. In: Proceedings of the 24th Annual ACM Interaction Design and

Children Conference, pp. 994–999. Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3713043.3731520

[30] Santos, G.M., Silva Julia, R.M., Nascimento, M.: Personalizing financial literacy learning for children through GenAI and agentic design: A game-based experiment. Expert Systems with Applications **309**, 131149 (2026) https://doi.org/10.1016/j.eswa.2026.131149

[31] Arif, S., Haroon, M.S., Khan, A.J., Arif, T., Raza, A.A., Athar, A.: Kahaani: A multimodal co-creative storytelling system. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pp. 347–365. Association for Computational Linguistics, Rabat, Morocco (2026). https://doi.org/10.18653/v1/2026.eacl-srw.25

[32] Fan, A., Lewis, M., Dauphin, Y.: Hierarchical neural story generation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 889–898. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1082

[33] Yao, L., Peng, N., Weischedel, R., Knight, K., Zhao, D., Yan, R.: Plan-and-write: Towards better automatic storytelling. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 7378–7385 (2019). https://doi.org/10.1609/aaai.v33i01.33017378

[34] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, Ø., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: Advances in Neural Information Processing Systems, vol. 35, pp. 2507–2521 (2022)

[35] Abdelhady, A.S., Shaker, M.A.: Comprehensive email spam detection using large language models: Exploring the limits of llm-based and traditional methods. In: 2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA), pp. 1–6. IEEE, Piscataway, NJ (2025). https://doi.org/10.1109/ACDSA65407.2025.11166175

[36] Tang, L., Laban, P., Durrett, G.: Minicheck: A lightweight and efficient fact-checking model for llm-generated content. arXiv preprint arXiv:2404.10774 (2024) https://doi.org/10.48550/arXiv.2404.10774

[37] Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1410

[38] Mounir, M., Abd El Hady, A.: Fake profile detection using fine-tuned embedding models. In: 2025 Twelfth International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 172–177. IEEE, Piscataway, NJ (2025). https://doi.org/10.1109/ICICIS66182.2025.11313189

[39] Rooein, D., Röttger, P., Shaitarova, A., Hovy, D.: Beyond flesch-kincaid: Prompt-based metrics improve difficulty classification of educational texts. In: Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), pp. 54–67. Association for Computational Linguistics, Mexico City, Mexico (2024)

[40] Alexandru, A., Calvi, A., Broomfield, H., Golden, J., Dai, K., Leys, M., Burger, M., Bartolo, M., Engeler, R., Pisupati, S., Drane, T., Park, Y.S.: Atla selene mini: A general purpose evaluation model. arXiv preprint arXiv:2501.17195 (2025) https://doi.org/10.48550/arXiv.2501.17195