Oak Story: Improving Learner Outcomes with LLM-Mediated Interactive Narratives

Alan Y. Cheng
Stanford University
Stanford, California, USA
ayc@stanford.edu

Carolyn Q. Zou
Stanford University
Stanford, California, USA
cqz@stanford.edu

Anthony Xie
Stanford University
Stanford, California, USA
anthonyx@stanford.edu

Matthew Hsu
Stanford University
Stanford, California, USA
matthsu@stanford.edu

Felicia Yan
Stanford University
Stanford, California, USA
feliciay@stanford.edu

Felicity Huang
Stanford University
Stanford, California, USA
huangfe@stanford.edu

David K. Zhang
Stanford University
Stanford, California, USA
dkzhang@stanford.edu

Arjun Sharma
Stanford University
Stanford, California, USA
arsharma@stanford.edu

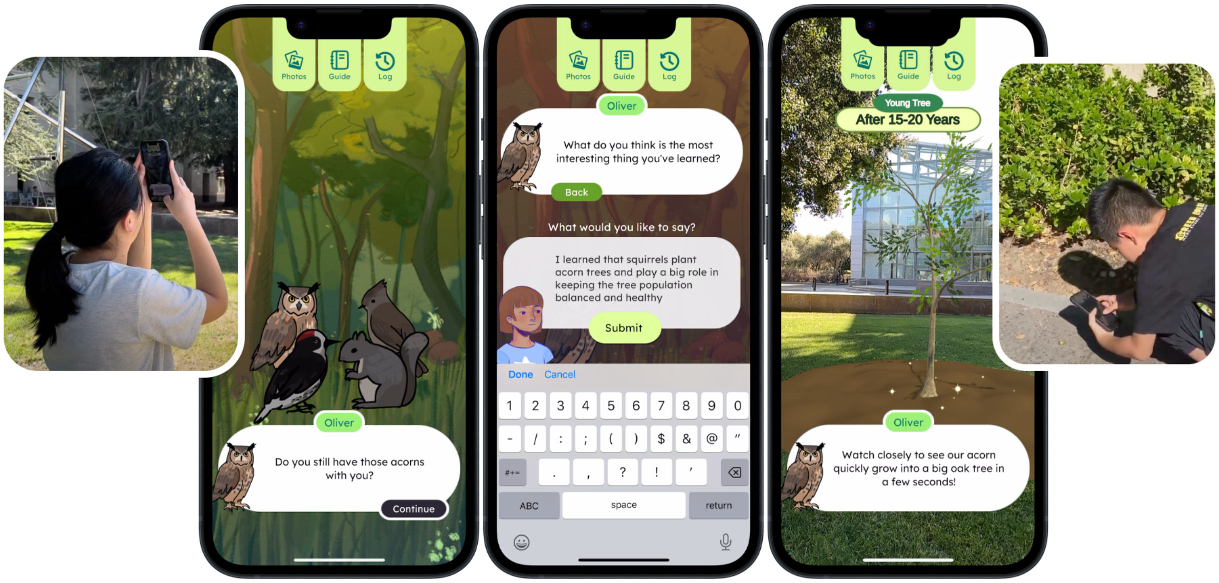
Rashon Poole
Stanford University
Stanford, California, USA
rashonp@stanford.edu

Daniel Wan Rosli
Stanford University
Stanford, California, USA
danwr@stanford.edu

Andrea Cuadra
Olin College
Needham, Massachusetts
USA
acuadra@olin.edu

Roy D. Pea
Stanford University
Stanford, California, USA
roypea@stanford.edu

James A. Landay
Stanford University
Stanford, California, USA
landay@stanford.edu



Oliver: Do you still have those acorns with you?

Oliver: What do you think is the most interesting thing you've learned?

What would you like to say?

I learned that squirrels plant acorn trees and play a big role in keeping the tree population balanced and healthy

Oliver: Watch closely to see our acorn quickly grow into a big oak tree in a few seconds!

Figure 1: *Oak Story* is a mobile app that engages children in real-world learning about ecology through an interactive narrative featuring animal characters. It uses a custom large language model architecture to generate character dialogue (center left) and gives the learner agency to influence the story through freeform responses (center), all while ensuring that the story and lesson remain on track. *Oak Story* also uses augmented reality activities to tie the narrative to the learner's physical surroundings (left, center right, and right).

ABSTRACT

Narrative-based education engages children in learning, but traditional approaches offer limited adaptability to individual preferences. Although large language models (LLMs) offer promising opportunities for interactive narratives, balancing their unpredictability with structured learning objectives remains challenging. To answer this challenge, we designed and built *Oak Story*, an educational mobile application for 4th–6th graders centered on local oak woodland ecosystems. *Oak Story* employs a learning-goal-directed LLM architecture that adapts the narrative, as well as multimodal real-world activities, to each individual student while ensuring that learning goals are met. In a between-participants study ($N = 47$), we find that *Oak Story* produces statistically significant increases in learning gains, engagement, and perceived agency compared to a control with static sequencing within and between scenes. These findings demonstrate an effective architectural approach for LLM-based educational systems that successfully balances learner agency with pedagogical structure.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *Interactive systems and tools*; Mixed / augmented reality; Natural language interfaces; • **Applied computing** → **Education**; **Interactive learning environments**; • **Computing methodologies** → *Artificial intelligence*;

ACM Reference Format:
Alan Y. Cheng, Carolyn Q. Zou, Anthony Xie, Matthew Hsu, Felicia Yan, Felicity Huang, David K. Zhang, Arjun Sharma, Rashon Poole, Daniel Wan Rosli, Andrea Cuadra, Roy D. Pea, and James A. Landay. 2025. *Oak Story*: Improving Learner Outcomes with LLM-Mediated Interactive Narratives. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28-October 1, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3746059.3747698

1 INTRODUCTION

Educational experiences that engage children as active participants have long been recognized as powerful catalysts for learning [22, 63], yet must also ensure students meet specific educational objectives. The challenge lies in designing systems that respond to learners' individual preferences while ultimately guiding them toward established learning goals. Narrative-based approaches naturally balance these competing priorities by providing both structured educational progression and opportunities for personalized exploration.

Interactive narratives extend these benefits by allowing learners to influence story progression according to their interests and curiosities [46] . Recent advances in artificial intelligence, particularly large language models (LLMs), create unprecedented opportunities to make educational narratives more responsive to individual learners while maintaining pedagogical direction. Similarly, augmented reality (AR) technologies enable these narratives to extend beyond screens into the physical world, creating possibilities for situated learning experiences that connect abstract concepts with tangible environments [31].

To bring interactive narratives into a child's immediate physical environment, we created *Oak Story*, an educational mobile application that integrates AI and AR technologies to facilitate learning about local oak woodlands for 4th–6th graders (Figure 1). *Oak Story* introduces learners to a colorful cast of animal characters (Figure 2) to illustrate the history and ecology of oak woodlands. To enable self-directed learning, *Oak Story* employs a learning-goal-oriented architecture, leveraging a large language model (LLM) to dynamically interpret learner interests, generate responsive dialogue, and sequence narrative events adaptively. This architecture maintains educational coherence by ensuring all specified learning objectives are met, while simultaneously providing the flexibility to accommodate individual exploration through prioritizing areas and dialogue that align with a learner's expressed interests.

We evaluate *Oak Story* through a between-participants experimental study addressing two research questions:

RQ1: How do learning outcomes (learning gains, motivation, and engagement) and sense of agency compare between a dynamic narrative adapted to a learner (*Oak Story*) and a static linear narrative (control)?

RQ2: What are the opportunities of using LLM-mediated interactions to improve user autonomy and engagement with interactive narratives and AR?

We make two main contributions to the literature. First, we present the design, architecture, and implementation of *Oak Story*, an educational app that combines an interactive story with real-world activities and dynamically adapts narrative progression based on user preferences. In particular, we contribute a novel LLM architecture that steers the narrative toward *specific learning goals*.

Second, we provide empirical evidence that the LLM-supported adaptivity of *Oak Story* promoted learning gains, improved engagement, and increased learners' sense of agency among 4th–6th grade students, even when compared to a control that also featured LLM-generated dialogue and the identical AR activities.

2 RELATED WORK

Our work builds upon prior research in AI-supported storytelling, interactive narratives, and student agency.

2.1 AI-Supported Storytelling in Education

Narrative-based learning environments integrate elements of storytelling into educational technologies. In particular, narratives can provide scaffolding that engages learners and supports their learning [5, 35]. For instance, prior work has demonstrated that narrative-based learning can lead to a variety of positive outcomes, such as increasing self-efficacy [37], promoting interest in subjects [37, 45], and fostering a sense of identity [45]. Narrative-based learning environments can take many forms, but most relevant to the present work are those that use generative AI to create or co-create educational stories, especially for children.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

UIST '25, September 28-October 1, 2025, Busan, Republic of Korea

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-2037-6/2025/09
https://doi.org/10.1145/3746059.3747698

Many existing approaches focus on vocabulary lessons due to their compatibility with the strong natural language capabilities of large language models. A common design pattern observed in these applications is the use of an LLM to insert target vocabulary words into children's stories. For example, Storyfier [44] helps learners acquire new vocabulary by generating a story based on a list of target words and then co-writing a new story with AI assistance. *Open Sesame? Open Salami!* [32] embeds vocabulary from a child's daily life into a customized storybook, adapting the story to the child's everyday context. *Mathemyths* [69] targets mathematical vocabulary and co-creates stories with children while embedding mathematical language. Another recent project, *StoryMate* [9], engages children in active reading through conversation with an AI companion and extends these conversations to real-world topics.

The aforementioned works primarily use storytelling as a vehicle to support young learners' literacy skills. However, narrative can be used effectively to engage children in a variety of topics, including science and mathematics (e.g., [10, 39, 49]). Furthermore, these tools rely heavily on the LLM to generate new story events, and LLMs struggle with long-term narrative coherence [27], making it difficult to create narrative-based learning experiences where the educational content is deeply intertwined with the story. Our approach draws inspiration from a long line of work in interactive narratives, which we describe in the next subsection.

2.2 Interactive Narratives

Interactive narratives are experiences in which the user actively takes on the role of a character in the story, and their actions influence plot events [12, 21]. These experiences can take many forms, from text-based adventures to complex game environments, but share the common characteristic of allowing users to shape narrative outcomes through their choices and actions.

The field of interactive narrative has evolved substantially from early hypertext fiction to structured authoring tools like Twine [18]. These traditional approaches relied on branching narratives with pre-authored content, limiting the possible story variations to what authors had explicitly defined. Early work established foundational concepts like experience management [46] and demonstrated the practical viability of generative storytelling systems like StoryAssembler [19], as well as innovating techniques in artificial intelligence to tailor story events to the user's decisions [20, 47]. However, these approaches required extensive knowledge engineering and technical expertise to author narratives.

The rise of LLMs in recent years has sparked a new wave of interest in integrating AI with interactive narratives, from driving interactive games solely with generative AI [55] to using AI to co-write screenplays [38] or co-facilitate tabletop role-playing games [71]. This line of work has focused primarily on interactive narratives for entertainment, emphasizing the generation of interesting, emergent behaviors and narrative changes that arise as a result of the user's choices, such as StoryVerse [64] and WhatELSE [34].

One recent interactive narrative authoring tool, Drama Llama [56], combines human-authored storylets [28] with LLM-driven generation. Some of its stated aims include "playthrough diversity" and "responsiveness to player-introduced narrative ideas." These aims, however, can be detrimental in educational settings: educators want learners to engage with an established lesson plan and instead require a system with sufficient structure to proactively guide learners through narrative content, educational content, and self-evaluation. While our system, *Oak Story*, employs a similar hybrid architecture that combines pre-written storylet-like scenes with LLM generation, it additionally aligns the generated content with learning objectives and employs retrieval-augmented generation to improve the factual accuracy of the dialogue. Furthermore, we demonstrate the efficacy of our approach on learning outcomes.

2.3 Student Agency in Education

A primary advantage of interactive narratives is that they increase the user's control and sense of responsibility [21, 58]. In an educational context, this can lead to better learning outcomes because of increased *learner agency*. Learner agency, defined as students taking ownership of their learning [62], is associated with increased engagement, motivation, and academic achievement. The pivotal role of learner agency in education is amply documented [15]. For example, Hattie and Anderman [23] identify learner agency as a critical factor for student achievement, emphasizing the importance of feedback and teacher-student relationships in its development. Bandura's social cognitive theory [3, 4] emphasizes the role of self-efficacy, positing that individuals with a higher sense of agency, tied to self-efficacy, may be more likely to seek help when facing challenges.

Despite the amount of attention that the concept of learner agency has received from scholars in the learning sciences, there is a relatively small body of work exploring how to design *technology* to support learner agency, especially for out-of-classroom contexts where an educator is not available yet guidance is still needed. Such studies have shown that increased student agency in technology-based learning environments is generally associated with improved learning outcomes [48, 52]. However, there is also the danger of providing too much freedom, which can lead to more time spent off task [50]. We contribute to this body of work by presenting an approach that balances freedom and structure: our system, *Oak Story*, employs an interactive narrative and an LLM-based architecture that allows learners to shape their own path through the narrative while proactively guiding them through a defined lesson plan.

3 *OAK STORY* APPLICATION

Oak Story is an educational mobile app for 4th–6th graders that employs an interactive narrative to engage children in learning within their local surroundings. In this section, we describe the educational goals of *Oak Story*, explain the design of its narrative, and provide an overview of the system's features.

3.1 Educational Objectives

Oak Story teaches the ecology and environmental science of oak trees and their ecosystems. Oaks are keystone species that are environmentally, historically, and culturally significant in our local area and are the subject of active revitalization efforts [53]. We use oak trees and their surrounding ecosystems as an anchoring phenomenon to explore broader concepts in ecology, such as seed dispersal and symbiosis. We devised the following learning objectives for *Oak*

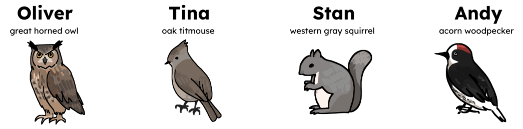


Figure 2: The characters of *Oak Story*: Oliver the owl, Tina the oak titmouse, Stan the squirrel, and Andy the acorn woodpecker. These are all local species found in oak woodlands.

Story, which were chosen to align with Next Generation Science Standards (NGSS) [54] and target different cognitive levels within Bloom's Taxonomy [1]:

1. Identify an oak tree by its structures (leaves and acorns) and distinguish oaks from other trees.
2. Explain why oak trees are important to both natural ecosystems and human society.
3. Analyze the relationships within an oak ecosystem and explain how different organisms in that ecosystem interact with each other.
4. Understand how an animal's structures and behaviors help it survive.
5. Give an example of how human activity affects natural ecosystems.

3.2 Narrative Design

The *Oak Story* system draws inspiration from the 5E model of instruction for inquiry learning [7, 17], an approach rooted in constructivist theories of learning that provides a set of recommended instructional phases (Engage, Explore, Explain, Elaborate, and Evaluate) to promote effective active learning. Because the steps in 5E model are designed to span a whole curriculum rather than a single lesson, we chose to align the narrative with just the Engage, Explore, Explain, and Evaluate phases of the 5E model.

3.2.1 Engage. After providing their name and creating an avatar to represent themselves, the learner meets Oliver the owl (see Figure 2). Oliver describes himself as a longtime resident of the local area and **engages** the learner in an introductory discussion of oak woodlands, giving background information on animals in the area and the trees' historical significance. Oliver then asks the learner whether any of those topics particularly interests them. The learner's response influences which options Oliver suggests next.

3.2.2 Explore. After the introduction is complete, the learner spends most of their time **exploring** the various scenes available in the narrative. Because the explore phase in the 5E model is meant to be learner-directed, we wanted to give the learner multiple options within the story to investigate. The learner can visit each scene in any order, as well as ask follow-up questions about topics mentioned by characters, and Oliver's suggestions depend on what preferences the learner expressed earlier.

3.2.3 Explain. In each scene, the learner is introduced to a new topic, where they are asked to **explain** their current understanding before developing a more complete understanding through interactions with the characters. The following scenes are available:

- Visit a magic portal to see what local oak woodlands might have looked like 200 years ago and understand their rich history. The learner describes the differences they notice and speculates about the reasons for the change.
- Meet Tina the oak titmouse, a dispirited bird who recently lost her home to deforestation, illustrating the consequences of habitat loss due to human activity. The learner is asked to speculate why Tina's tree may have been cut down.
- Meet Andy the acorn woodpecker, whose specialized beak enables a unique survival strategy. The learner is asked to guess how Andy might be affected by competition.
- Meet Stan the squirrel, who shares a symbiotic relationship with oak trees by caching acorns. The learner is encouraged to describe what how the relationship between squirrels and oak trees is symbiotic.
- Find and identify an oak tree in real life (which might be the perfect new home for Tina!). The learner is guided through hints for accurate tree identification.

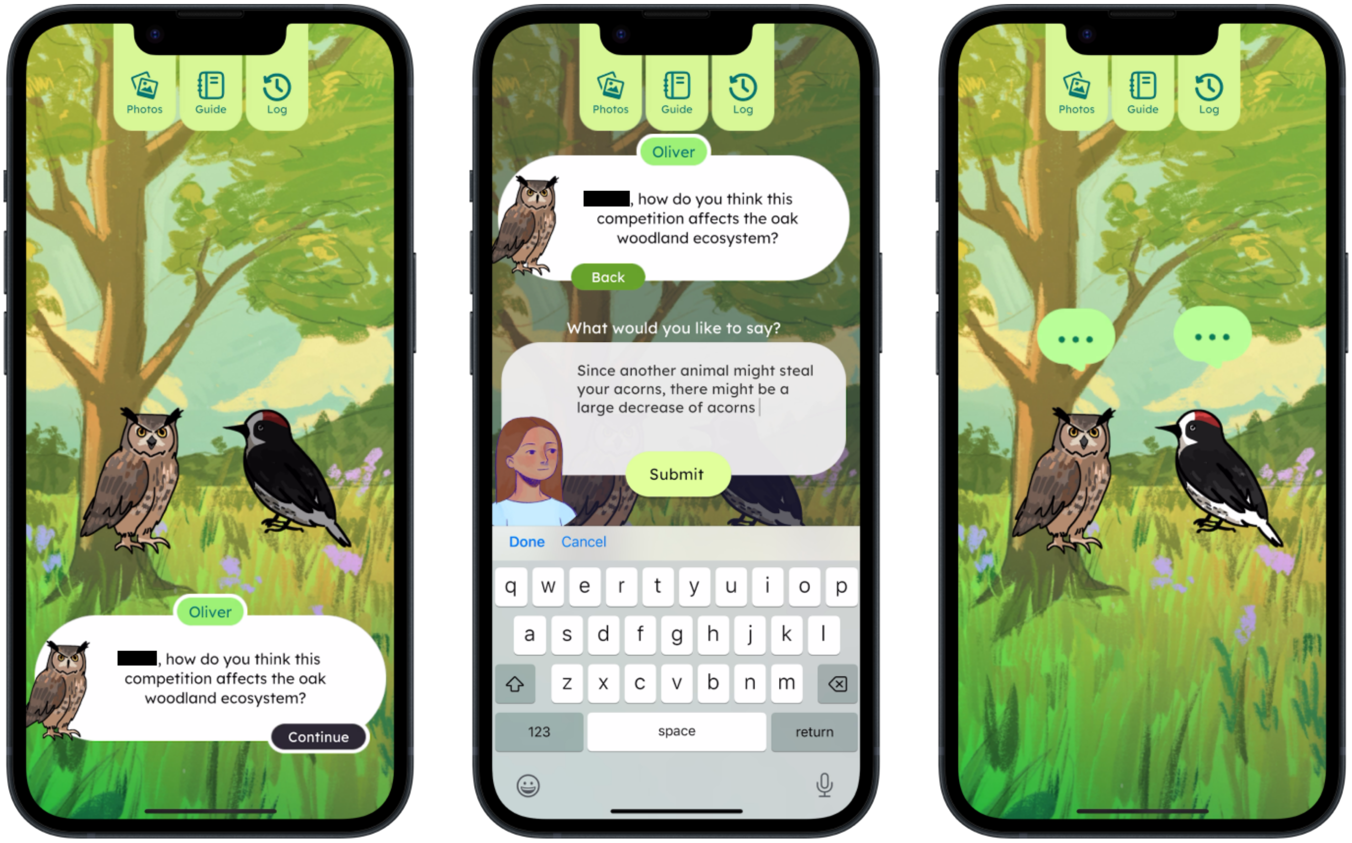
3.2.4 Evaluate. Upon completing a scene, Oliver the owl asks the learner to **evaluate** what they have learned by asking them to reflect on what they have seen and read so far. As suggested by the 5E model, these reflections are embedded throughout the learning experience as lightweight, formative assessments. Afterward, Oliver asks the learner what they want to do next, providing recommendations and repeating this process until the learner has visited all scenes.

3.2.5 End. Once all other scenes are complete, the learner participates in the conclusion of *Oak Story*, where all the characters they've met join forces to brainstorm how to find Tina a new home. For this scene to be completed successfully, the learner must have already identified an oak tree in real life, met Andy the woodpecker who can peck a new hole for Tina to live in, and collected acorns with Stan the squirrel that can be traded for Andy's help. The learner takes the acorns they received from Stan the squirrel and gives them to Andy the acorn woodpecker. In exchange, Andy pecks a large AR hole in the real-life oak tree the learner previously identified, providing Tina the oak titmouse a new home. Finally, after Tina has successfully been rehomed, the learner participates in a final tree planting activity, which concludes *Oak Story* (Figure 5, right).

3.3 Core Features

To inform the design of the core features in *Oak Story*, we wanted to leverage the opportunities afforded by LLMs for narrative-based learning. Two of *Oak Story*'s core features are (1) dynamic dialogue, enabled by the LLM's conversational capabilities, and (2) interactive real-world activities, some of which utilize images captured directly from learners' environments.

Throughout *Oak Story*, the learner converses with animal characters (Figure 2) to learn about oak ecosystems and progress the story. At specific points throughout the story, characters encourage the learner to participate in real-world learning activities, such as finding a real-life oak tree in their vicinity or walking through an augmented reality portal to the past to visualize the oak woodlands that used to cover the area.



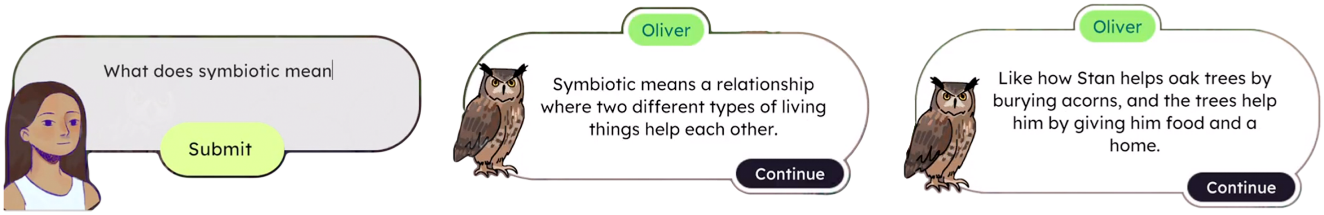
[redacted], how do you think this competition affects the oak woodland ecosystem?

[redacted], how do you think this competition affects the oak woodland ecosystem?

What would you like to say?

Since another animal might steal your acorns, there might be a large decrease of acorns

Figure 3: An example of the dialogue flow in *Oak Story* from one of our study participants. After a character prompts the learner (left), a text field appears where the learner can compose a freeform text response (center). Once the learner submits the response, the characters will "think" as the system generates the next dialogue (right).



What does symbiotic mean

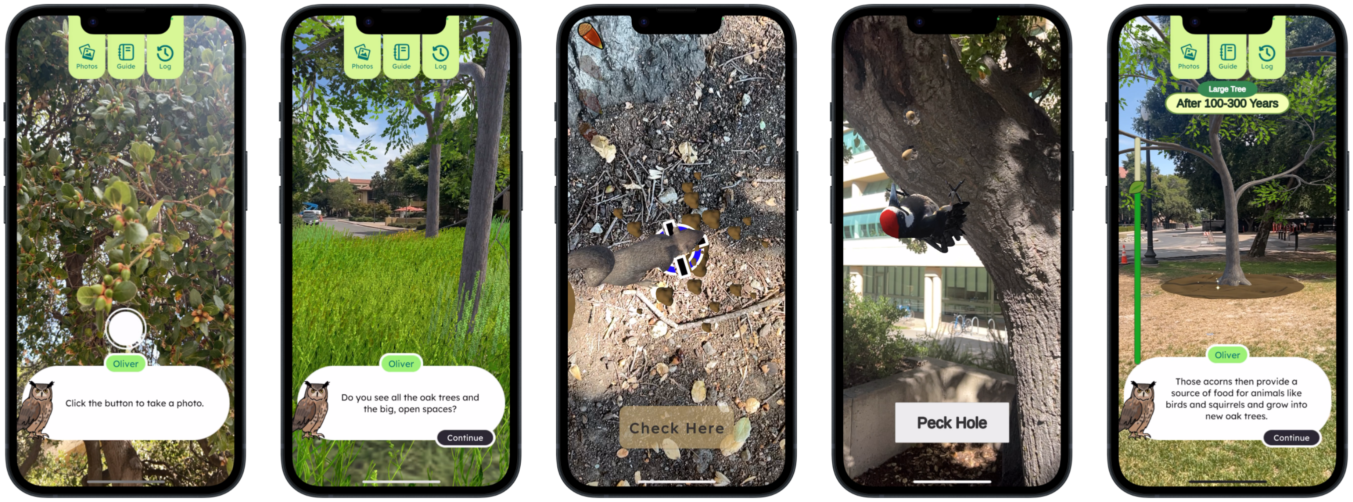
Symbiotic means a relationship where two different types of living things help each other.

Like how Stan helps oak trees by burying acorns, and the trees help him by giving him food and a home.

Figure 4: Another example of a dialogue exchange in *Oak Story*. Here, the learner chose to ask a question (left). In response, a character will answer the question (center) and connect that answer back to the narrative (right).

3.3.1 Dynamic dialogue. The primary interaction in *Oak Story* consists of dialogue between the learner and the characters in the narrative. Specifically, a character speaks to the learner and asks for their response (Figure 3, left), positioning the learner as a participant in the story while encouraging active learning. Next, the learner writes a freeform text response to the character(s) (Figure 3, center) by typing or using dictation. They are free to ask follow-up questions (Figure 4) and express their creativity. Once the learner submits their reply, a thinking animation is displayed (Figure 3, right) for a few seconds before the character's response is shown, upon which this cycle repeats. All of the characters' dialogue is created by our LLM architecture, which is described in greater detail in Section 4.

The responsive nature of these character messages is reflective of the dialogic framework of teaching and learning [67]: characters ask and answer open-ended questions, provide relevant knowledge and feedback, and guide the learner to learn about ecological concepts in their own environment through exploration that matches their desired pace.



Click the button to take a photo.

Do you see all the oak trees and the big, open spaces?

Those acorns then provide a source of food for animals like birds and squirrels and grow into new oak trees.

Figure 5: Real-world activities in *Oak Story*. From left to right: (a) Tree Identification: take pictures of trees to find an oak tree; (b) Portal: walk through an AR portal to see oak woodlands in the past; (c) Squirrel: help Stan the squirrel dig up acorns; (d) Woodpecker: guide Andy the acorn woodpecker to peck a new home for Tina the oak titmouse; (e) Oak Lifecycle: plant and grow an AR oak tree.

3.3.2 Real-world activities.

Some scenes in the *Oak Story* narrative have associated real-world activities that prompt the learner to interact with their physical surroundings. These activities, illustrated in Figure 5, are described below.

Tree identification (Figure 5a). In this activity, the learner is asked to find and identify a real-world oak tree. Oliver the owl instructs the learner to go outside to a nearby tree. Once the learner has located a tree, Oliver provides a description of oak leaves and asks whether the leaves on the tree match this description. Next, the learner takes a photograph of the leaves and Oliver asks them to reflect on whether their tree's leaves resemble an oak's. Oliver then repeats this process with acorns. Finally, the learner synthesizes their observations and decides whether the tree is an oak. If they decide the tree is not an oak, the learner is prompted to find another tree and try again. This activity deliberately positions the learner as the decision-maker, rather than having the system identify the tree's species on the learner's behalf.

Portal (Figure 5b). In this activity, the learner places an AR portal on the ground and walks through it to be transported into a 3D environment of a historical oak woodland. This activity is accompanied by an explanation from Oliver on the long history of native oaks in the local area, how indigenous peoples used and lived with the oaks, and the negative impact of colonial settlement on the oak population.

Squirrel (Figure 5c). In this activity, the learner helps Stan the squirrel dig up acorns from his stash. The learner uses an AR crosshair to point at spots where an acorn may have been buried. An AR version of Stan the squirrel will run to the selected spot and attempt to dig up a virtual acorn. The learning goal of this activity is to show how squirrels engage in caching behavior for survival.

Woodpecker (Figure 5d). In this activity, the learner guides Andy the acorn woodpecker to make a new home for Tina the oak titmouse. The learner uses a crosshair to place an AR version of Andy on the real-life oak tree they previously identified. Then, the learner taps the screen to make Andy peck at the tree, creating an AR hole in the tree that grows in size with each tap. Once the hole is large enough, Tina the oak titmouse flies into her new home. The goal of this activity is to visualize how woodpeckers use their specialized beaks to help them survive, and how different animals in the ecosystem interact.

Oak lifecycle (Figure 5e). In this activity, the learner measures out a large space in their surroundings, plants a virtual acorn, and observes an AR visualization of an oak tree's growth through its lifecycle, from seed to sprout to sapling to young tree to great oak.

4 IMPLEMENTATION AND LLM ARCHITECTURE

We built a custom large language model (LLM) architecture to operate as an experience manager [46] that mediates the learner's interactions with the narrative. Our architecture generates appropriate character dialogue, understands open-ended input from the learner, and advances the state of the narrative based on the student's progress towards learning goals. In this section, we elaborate on the implementation of *Oak Story*, with a particular focus on our LLM architecture. The prompt template we use is provided in the Appendix.

4.1 Scenes

To ensure the inclusion of all the content required for learners to meet the learning objective in *Oak Story*, we structured *Oak Story*'s narrative as a set of distinct, pre-defined story scenes, each oriented toward a learning goal (outlined in Section 3.1). Each scene contains a high-level description of its events, as well as a list of **goals, actions, and characters**, which are described in detail in the following subsections. An example is the scene where "the learner is introduced to Stan the squirrel."

4.2 Goals

To ensure that the LLM facilitates all learning objectives, we define a list of goals at each scene for the LLM to pursue. This allows characters in the story to proactively introduce new concepts and activities. A goal is a natural language description of what a character or learner should do within a scene. In the scene where Stan the squirrel is introduced, the goals include "Stan the squirrel mentions his stash of acorns" (an event related to the narrative), "Stan or Oliver describes the symbiotic relationship between oaks and squirrels" (a learning goal), and "The learner helps Stan collect acorns" (a goal related to a real world activity).

To produce the desired learning outcomes, the LLM is given a list of incomplete goals and is instructed to bring about the completion of all goals. To invite the learner's continued engagement, the characters' responses typically contain a question for the learner to answer. Thus, goals often guide the generation of questions that give the learner an opportunity to explain their current understanding of the topic or reflect on their learning progress. Notably, the order in which goals are met within a state is not deterministic; the conversation history is used to select the goal that is most appropriate for each exchange. By orienting the LLM toward specific goals, our architecture requires that all goals are met before ending the scene, ensuring that the learner completes all intended objectives.

4.3 Actions

The action mechanism enables the LLM to trigger changes within a scene and to transition between scenes in a dynamic order that adapts to learner preferences (see Figure 7). Each action is defined by a name and natural language description, which informs the LLM of the narrative events that should cause an action to be triggered, a set of prerequisite goals, which must be met before the action can be triggered, and a set of effects that the action brings about. These effects can include the adding or removing of a character, introducing new goals, initiating a real-world activity, ending the current scene, and starting a new scene. For example, once the learner has finished meeting Stan the squirrel, the `LEAVE_STAN` action has three effects: Stan is **added** to the list of characters that appear in a later scene, the action to meet Stan is **removed** so that it cannot be triggered again, and the current scene is **ended** to represent the learner leaving.

The *Oak Story* architecture uses a process analogous to function calling [42] to trigger actions. It prompts the LLM with the learner's messages and a list of available actions. The LLM compares both the learner's most recent message and the preferences expressed in previous messages with the available action list to determine which action, if any, should be triggered. This allows for explicitly expressed (e.g., direct responses to "where do you want to go next?") and implicitly expressed (e.g., time spent discussing different parts of an oak ecosystem) preferences to be taken into account when determining which eligible action is triggered.

In some cases, there may be many available actions, such as when the learner has finished a scene and is deciding where to go to next. Because the input is open-ended, the available affordances may not be immediately obvious. To support user sensemaking, *Oak Story*'s characters actively suggest relevant actions to the learner based on the conversation history and the description of each action.

4.4 Characters

Finally, each scene contains a list of characters present. *Oak Story* invites the learner into the narrative by incorporating contextual information about their responses and real-world interactions into character messages. Information like learning goals and available actions is delivered in the voices of characters. Similarly, when the learner expresses that they would like to take an action, character messages provide feedback that something has happened.

Each character has a unique personality, background, and area of expertise. In *Oak Story*, Oliver the owl is an inquisitive guide who is well-informed about history, but not much about oak trees in the present day; Andy is an irritable woodpecker frustrated with the competition he experiences against squirrels for acorns; Tina the oak titmouse is anguished after losing her home, showcasing the ecological impacts of habitat destruction. By having these characters interact with each other and the learner, complex concepts can be presented in a engaging way that emphasizes the connections between topics. This character-based approach not only makes the learning experience more enjoyable, but also aids in comprehension by associating lesson concepts with distinct personalities [68].

4.4.1 Retrieval-augmented generation. To ensure that the system provides accurate information while maintaining the individuality of different characters, *Oak Story* uses a form of retrieval-augmented generation [33] that supplements a natural language description of each character with differing knowledge bases that reflect their limited knowledge based on their role in the ecosystem. The documents in each character's knowledge base is curated by the experience designer. For instance, *Oak Story*'s Andy the acorn woodpecker is presented as an expert on the oak trees' acorns. As such, whenever the learner is speaking to Andy, the LLM will have access to certain information about acorns, and will produce messages for Andy that reference this information. The information available at each time step is pertinent to the unmet goals and user messages, e.g., Andy's knowledge base contains answers to questions like "How do oak trees disperse their acorns?" and "How did Native Americans use and manage acorns as a food source?" but only the most relevant items are retrieved. This method also supports multiple characters in a scene: at each step, a text embedding of the learner's input and unmet goals is used to retrieve relevant knowledge for each character, and specialized knowledge is attributed to the appropriate speaker.

4.5 Miscellaneous Details

Lastly, we provide details on the specific tools and libraries used to build *Oak Story*. *Oak Story* is a mobile application targeting iOS devices that uses a client-server architecture (see Figure 6). The client is built with the Unity [60] game engine and deployed on iOS devices with original artwork. The UI for dialogue was implemented with the Yarn Spinner dialogue engine [30]. The augmented reality

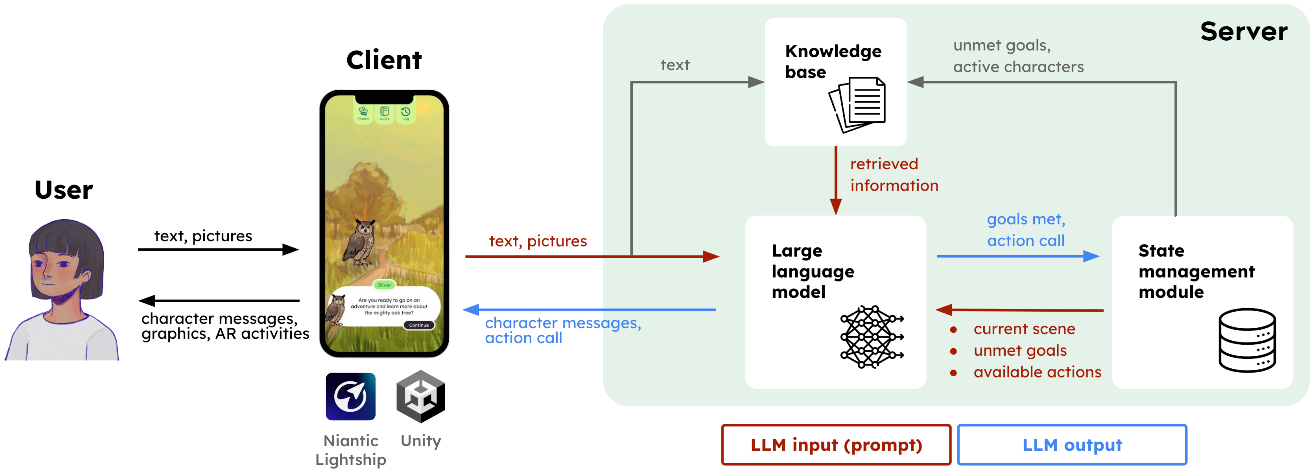


Figure 6: System diagram for *Oak Story*. The learner interacts directly with the mobile client to read character dialogue, provide user input, and engage in AR activities. The server-side architecture combines (1) user input from the client, (2) information about the current narrative state, and (3) any relevant factual information from the knowledge base into a single prompt that is sent to the LLM. The LLM then outputs the appropriate character dialogue and updates the narrative state accordingly. See Section 4 for more details on the LLM architecture.

(AR) features for real-world activities were developed using Niantic Lightship ARDK [40].

The Unity client connects to our LLM architecture over a WebSocket connection. *Oak Story* is implemented in Python and hosted on Heroku [24], and it runs on Uvicorn [61], an Asynchronous Server Gateway Interface (ASGI) webserver. Under the hood, *Oak Story* uses OpenAI's GPT-4o model [43], a state-of-the-art multimodal LLM with a low-latency inference provider.

5 EXPERIMENT

We conducted an IRB-approved study with fourth–sixth graders ($N = 47$). In this section, we will describe how participants were assigned to a specific condition, our participants and recruitment approach, and our experimental measures. In this section and beyond, *Oak Story* refers to the app used by participants in the treatment condition (described fully in Sections 3 and 4), while those in the control condition used a modified version of *Oak Story* (described in Section 5.1).

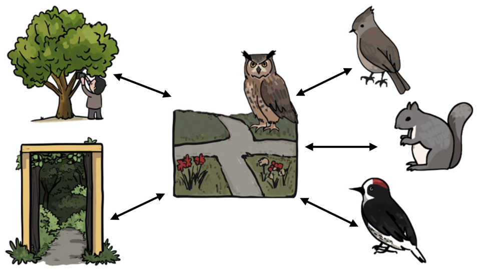


Figure 7: Characters in *Oak Story* present the learner with choices that offer different paths through the scenes in the narrative. Each image in this diagram represents a different scene in the narrative that can be traversed in a different order.

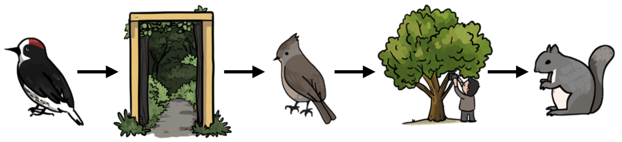


Figure 8: In the control variant, learners progress through scenes in a fixed order that does not adapt to their individual preferences.

5.1 Control Condition

Members of the control group used an ablated version of *Oak Story*. Our experimental design aims to isolate the effects of **learner control over narrative progression**: while *Oak Story* and the control variant share the same user interface, story elements, learning goals, and real-world activities (as described in Section 3), the control presents scenes and activities in a fixed order that does not take the learner's preferences into account (see Figure 8). To preserve the conversational interface between conditions, we still allow learners using the control variant to write open-ended messages. *Oak Story* characters reply to these messages in a manner that is *responsive*, but the story is not *adaptive*. Instead of being generated entirely by an LLM, control variant replies are each composed of two parts: an LLM-generated prefix that responds to an individual learner's message, followed by a static suffix[1], shared by all learners in the control group, that moves along to the next learning goal in a predefined order.

By setting up the control in this way, we ensure that both the control and treatment conditions appear essentially identical to participants, as they both feature AR activities and responsive dialogue. The only distinction between the two experiences is whether the system tailors the learner's progression in the story based on their input and preferences.

5.2 Participants and Condition Assignment

We recruited a total of 51 participants. Children were recruited through mailing lists, physical flyers, and word of mouth. Mailing lists included pre-existing contacts gathered from previous studies, institutional listservs for parents working at our institution, and mailing lists forwarded to via outreach to local schools and afterschool organizations.

Using systematic stratified randomization, participants were assigned to either a control group (n = 21) or a treatment group (n = 26), maintaining a roughly even distribution across grades and gender. Condition was first randomized by strata, with each stratum being defined by a grade-gender pairing, and then alternated within each stratum between the control and treatment conditions. Two participants (P10, P27) did not finish the study due to irrecoverable technical issues, one participant (P50) declined to finish using *Oak Story*, and one participant (P34) declined to complete the final surveys and interview.[2]

Excluding these cases, the remaining 47 participants (30 male and 17 female) were in fourth (n = 13), fifth (n = 15), and sixth grade (n = 19). Participants' parents or guardians identified them as East Asian (n = 16), White (n = 15), Hispanic or Latino (n = 6), South Asian (n = 2), Southeast Asian (n = 1), and mixed (n = 7).

Two participants' families did not have the means to drive to our institution. Consequently, those two studies were conducted at a nearby park. We also used these sessions as an opportunity to test whether *Oak Story* could operate in other locations, and because they worked appropriately, we include these sessions in our analysis.

5.3 Procedure

Interested parents brought their children on site for the study for a 90-minute individual session. We first obtained consent from the participants' parents or guardians (adults) and assent from each participant. Participants' parents/guardians were asked to fill out a demographics survey on behalf of the participant. Parents/guardians also could optionally agree to their child's session being filmed. When video-recording, we aimed to record interactions while avoiding capturing the child's face.

Participants then filled out a pre-study questionnaire and a pre-test (knowledge quiz) on the iPad. Participants completed the knowledge quiz primarily using dictation for consistency, as not all were familiar with typing on an iPad; however, they could manually correct their responses as needed.

After completing the pre-study questionaire and pre-test, we launched *Oak Story* on an iPhone, started screen recording, and handed it to the participant. When the first request for user input appeared, we let the participant know that they could use the microphone (dictation) feature to input their answer. During AR activities, a researcher took the participant outside. There were two to three researchers at each session to moderate, observe, and record the study.

Once participants were done using our app, we asked the participant if they wanted to take a break. They were then asked to complete the same knowledge quiz from the beginning in the same manner. Next, they completed a post-study questionnaire. Finally, participants completed a semi-structured interview that was audio-recorded.

A follow-up knowledge quiz was sent a week after the in-person session to be completed at home, and participants were asked to complete it as soon as possible. Upon completing the follow-up survey, we sent them a $25 Amazon gift card to thank them for their participation.

5.4 Measures

In this section we describe our quantitative measures: learning gains, pre-post measures, engagement, and perceptions of the characters and story. In addition, we describe our semi-structured interview and qualitative data analysis approach. Copies of our questionnaires and interview guide are included in the Supporting Files.

5.4.1 *Pre-post measures.*

We used a pre- and post-app knowledge quiz to measure learning gains. The quiz contained a multiple-choice question that asked participants to select all the oak trees out of a variety of tree photos, then several open-ended questions where participants were asked to verbally describe: 1) the importance of oak trees, 2) their role in today's cities, 3) the benefits they can provide to people, and 4) how different parts of an oak tree's ecosystem work together (e.g., what roles do oak trees, acorns, squirrels, and acorn woodpeckers play in the ecosystem). The rubric for knowledge quiz grading is included in the Supporting Files.

In the pre- and post- tests we also included a scale to measure learners' motivation toward science learning (**SMTSL**) [59], the Environmental (**2-MEV**) Scale questionnaire for children [25], and a scale to measure interest levels. For the SMTSL, we adapted six items on a five-point Likert scale encompassing the following categories: encompassing self-efficacy, active learning strategies, and perceived value of science learning. We also used ten Likert items from the 2-MEV Scale to measure children's attitudes toward the environment. These items encompass preservation categories such as intent of support and enjoyment of nature. Finally, to measure interest levels, we used a five-point Likert scale to measure interest on seven items: animals, plants, trees, oak trees, biology, ecology, and regional wildlife.

5.4.2 *Post-only measures.*

In the post-completion questionnaire we asked for feedback on *Oak Story*. To measure engagement, we adapted the seven items from Dietz, et al.'s Giggle Gauge [16] from a four-point scale to a five-point Likert scale for consistency with our other scales. We also developed a set of 12 items on a five-point Likert scale to evaluate interactions with *Oak Story*. A set of eight

---

[1] To ensure a consistent writing style between the *Oak Story* and the control, these static suffixes were also written by the same LLM that generates dialogue in *Oak Story*.

[2] We use an excerpt from P34's in-app conversations to demonstrate our system's response to off-topic inputs but otherwise exclude their data from analysis.

items dealt with the conversations with the animals in the story, and a set of four questions dealt with the story overall (shown in Table 2).

5.4.3 *Interview.*

Once participants had finished the in-person questionnaires, they completed a short, semi-structured interview about their experience. In this interview, we asked questions such as *"How was your overall experience with the app?"* or *"How did this experience compare to how you usually learn science?"*

6 FINDINGS

We first present a technical evaluation of our system. Then, we describe the results from our controlled study and the findings from our participant interviews.

6.1 Technical Evaluation

Achieving our goal of building adaptive learning experiences requires that our LLM architecture accurately bridge the learner and the narrative state. To that end, we coded the logs of user and character messages from all user studies to validate system behavior. In total, 1405 turns from the treatment condition and 1060 turns from the control were analyzed, with each turn consisting of a user message and the response from the characters, as well as any goals met, actions taken, and knowledge retrieved. Below, we provide standalone analyses of our LLM architecture (which was only used in the *Oak Story* treatment condition), as well as comparative analysis between the treatment and control versions of the app. The codes used can be found in the appendix.

We first focus on our LLM architecture's ability to make the correct action calls when learners express interest in an action. Our analysis finds that **98.03% of user-requested actions out of 203 total instances were successfully executed**, demonstrating the robustness of the system in correctly interpreting user intent from natural language and translating it into appropriate state changes. Since actions become available when the prerequisite goals are completed, we also evaluate detection of goal completion. **Our architecture accurately marked 98.51% of goals as completed when users successfully achieved them**, out of 875 total goals completed. This indicates that it effectively tracks learners' progress through the narrative and learning objectives.

Our architecture also contains a knowledge base with curated educational materials, sections of which were delivered to the LLM for incorporation into character responses when any sections met a threshold of cosine similarity to the user input or unmet goals. This threshold was met, and consequently retrieval was used, in 40.00% of LLM-generated responses. Out of these 562 instances, **character messages incorporated externally validated knowledge to provide accurate information to the learner in 64.59% of responses**, demonstrating effective integration of the knowledge base into the dialogue.

Next, we conducted comparative analyses between *Oak Story* (treatment) and the control variant. Both versions are able to answer student questions and provide specific responses due to the use of an LLM in the control. Our first analysis showed that **100.00% of 39 total questions posed by learners interacting with *Oak Story* were adequately answered**, compared to 76.47% (13 of 17) in the control, indicating a strong ability to provide relevant information in response to user queries. Furthermore, 58.97% of answers from *Oak Story* were supplemented with relevant retrieved knowledge, reducing the likelihood of LLM hallucination by incorporating curated information.

Additionally, coders were instructed to label messages as "specific" if they directly quoted or referenced the student's input and "generic" if the response failed to acknowledge the student's input at all, for instance:

> (specific)
> *Andy*: [...] What do you think squirrels do to get their acorns?
> *Participant*: I think they climb the oak tree.
> *Oliver*: That's right, [name]! Squirrels are great climbers and use their strong legs and claws to scurry up oak trees quickly [...]
>
> (generic)
> *Oliver*: [...] What do you think is the most interesting thing you've learned?
> *Participant*: That animals have always lived in oak trees
> *Oliver*: It's fascinating to see what this area used to look like with so many oak trees. They played a big role in the ecosystem [...]

Messages that were not flagged as 'specific' nor 'generic' comprised 79.15% of responses in *Oak Story*, and 83.21% of responses in the control. Messages that respond to short inputs from the student, like "yes" and "okay," comprise the majority of these neutral messages.

We find that *Oak Story* enables a higher degree of narrative personalization, with **20.36% of characters' responses tailored specifically to the learner's input compared to 10.85% in the control**. Characters respond to the learner's expressed interests, prior narrative events, and real world actions nearly twice as often.

These results demonstrate the effectiveness of *Oak Story* in creating personalized, responsive learning experiences. The high accuracy in action execution and goal tracking ensures that the narrative progresses coherently based on user choices. The remaining metrics provide evidence that *Oak Story* successfully manages the complex task of concurrently interpreting user input, updating the narrative state, and generating responsive character dialogue.

6.2 Learner Outcomes

We gathered data on learner outcomes from our pre-test and post-test, as well as the pre- and post-study questionnaires. For pre-post measures, we compare whether there was a shift from pre- to post-, and if there was, we tested for a difference between the treatment and the control conditions. For post-only measures, we test for differences between conditions only.

6.2.1 *Knowledge quiz.*

The answers to the knowledge quiz were graded by two researchers who were blind to the participants' condition and to whether the responses were from the pre-test or post-test. They achieved high inter-rater reliability, with an intraclass correlation coefficient (ICC) of 0.918, indicating excellent reliability [26].

	Quiz Score Mean [std dev]		
	Pre-test	Post-test	Gain (Post - Pre)
Control (n = 20)	4.23 [2.49]	5.18 [2.49]	0.95 [2.23]
Treatment (n = 26)	3.08 [1.98]	6.15 [1.94]	3.08 [2.11]
Overall (N = 46)	3.58 [2.26]	5.73 [2.22]	2.15 [2.39]

Table 1: Learning gains based on the knowledge quiz, separated by condition. The sample size is 46 because we remove one outlier who scored greater than 2.5 standard deviations above the mean in the pre-test (the outlier was assigned to the control condition).

To measure immediate learning gains after using *Oak Story*, we scored each pre-test and post-test (range: [-2, 14])[3]. For our analysis, we removed one outlier who scored over 2.5 standard deviations higher than the mean on the pre-test. Summary statistics for pre-test scores, post-test scores, and gain (difference between pre and post) are provided in Table 1. Based on a paired t-test, we observed a significant post-test increase in scores across both conditions, $t(45) = 6.11$, $p < 0.001$, indicating learning gains. The effect size, using Cohen's d calculated with change score variability [13], was $d = 0.609$, suggesting a medium-to-large effect size.

We further analyzed the difference in learning gains between conditions using a one-way ANCOVA [41] and found a **significant effect of the condition on the post-test score after controlling for pre-test performance**, $F(1, 43) = 7.25$, $p = 0.010$. The effect size, measured with the partial eta-squared statistic, was $\eta^2 = 0.144$, indicating a large effect [11]. This finding suggests that **the treatment led to increased learning gains** compared to the control.

6.2.2 *Follow-up knowledge quiz.*

The answers to the follow-up knowledge quiz, which were emailed to families a week after the study session, were graded by the same two researchers who scored the original quizzes. Not all participants completed the follow-up quiz, so we restrict our analysis to the 42 participants who completed it. Using a paired t-test, we observed a significant increase in scores (range: [-2, 14]) from pre ($M = 3.89, SD = 2.17$) to follow-up ($M = 6.49, SD = 2.17$), $t(40) = 5.24$, $p < 0.001$, suggesting that **participants retained what they learned**. The effect size, using Cohen's d calculated with change score variability [13], was $d = 0.82$, suggesting a large effect size. Interestingly, the scores on the follow-up quiz were higher on average than the respective post-test scores ($M = 6.13$) for the same participants.

We also compared the difference in learning retention between the treatment and control conditions. We conducted a one-way ANCOVA [41] and found no significant effect of the condition on the follow-up score, $F(1, 38) = 1.21$, $p = 0.28$.

6.2.3 *Engagement.*

To assess whether there was a difference in engagement between the treatment and control groups, we sum the items in our modified Giggle Gauge scale for each participant (range: [0, 30]). Only six out of seven items in the scale are summed because one item was mistakenly excluded from the survey for early participants. We used an unpaired t-test to compare the results and found that engagement in the treatment group ($M = 26.1$, $SD = 2.47$) was significantly higher than the control group ($M = 24.2$, $SD = 2.99$), $t(45) = 2.41$, $p = 0.020$, suggesting that **the treatment made *Oak Story* more engaging** compared to the control.

6.2.4 *Conversation- and story-related survey items.*

To analyze the 12 Likert-type items we developed to evaluate participants' perceptions of the characters and story, we clustered them into groups of 2-3 similar items and summed them, resulting in five separate Likert scales with the following themes: system responsiveness, characters' distinctiveness, feelings toward characters, connection to real-life activities, and story enjoyment 11 of these items are shown in Table 2. The remaining item ("The conversations encouraged me to think deeper.") did not fit well into any of the themes, so we omit it from hypothesis testing, as testing a single Likert item can be unreliable [8].

We performed an unpaired t-test on each scale and correct for multiple comparisons using the Holm–Bonferroni method. Table 2 shows the p-value for each scale after correction. We found that participants' attitudes toward system responsiveness were significantly higher in the treatment group compared to the group who received the control, suggesting that **interactions in the treatment feel more responsive** to learners' inputs.

6.2.5 *Motivation to learn science.*

We summed the items in the Student Motivation Towards Science Learning (SMTSL) scale for each participant (range: [6, 30]) and used a paired t-test to compare participants' scores between the pre-study survey and post-study survey. We did not find evidence to suggest that this change in motivation to learn science differed between the two conditions.

6.2.6 *Attitudes toward the environment.*

For each participant, we summed the items in the Environmental (2-MEV) Scale questionnaire for children (range: [10, 50]) and used a paired t-test to compare for differences between pre and post. We did not find a significant difference from pre ($M = 38.4$, $SD = 5.61$) to post ($M = 39.0$, $SD = 6.43$). The effect size, measured with Cohen's d calculated with the change score variability [13], was $d = 0.161$, suggesting a small effect size.

6.2.7 *Interest in scientific topics.*

Similar to the conversation- and story-related survey items, we condense seven Likert-type items into three scales: interest in animals (animals and local wildlife), interest in plants (plants, trees, and oak trees), and interest in science (biology and ecology). Then, we performed a paired t-test on each scale to compare how scores changed from pre to post and correct for multiple comparisons using the Holm–Bonferroni method. We found that the scores for all three scales significantly increased

---

[3] The minimum possible score is -2 because it is possible to lose up to two points by guessing incorrectly on the multiple choice question on the quiz. For more details on scoring, see the rubric in the Supporting Files.

Scale	Items	Range	Condition	Mean [std dev]	p
System responsiveness	• "The characters responded to what I specifically said or asked." • "My actions, interests, and messages directly influenced the story."	[2, 10]	treatment	8.92 [0.89]	0.009*
			control	7.76 [1.48]	
Characters' distinctiveness	• "The characters were unique compared to each other." • "I remember a lot about each character and what I learned from them."	[2, 10]	treatment	8.96 [1.00]	0.164
			control	8.43 [0.81]	
Feelings toward characters	• "I felt happy when I read what the characters said to me." • "The characters spoke in a kind way." • "I wanted to have more conversations with the characters."	[3, 15]	treatment	12.62 [1.96]	0.908
			control	12.19 [1.86]	
Connection to real-life activities	• "Seeing the characters in activities (where they interact with the real world) made me more interested in talking to them." • "The activities (where characters interact with the real world) felt connected to the conversations I had with the characters."	[2, 10]	treatment	9.00 [1.10]	0.908
			control	9.05 [0.97]	
Story enjoyment	• "I felt excited to find out what was going to happen next." • "The story made learning more fun."	[2, 10]	treatment	9.23 [0.95]	0.088
			control	8.43 [1.36]	

Table 2: Groupings of our Likert-type items into scales. Asterisks on p-values indicate that learners in the treatment group gave statistically significantly higher responses than learners using our control app.

from pre to post, indicating that *Oak Story* and the control increased learners' interest in the topics covered.

The increase in interest in each of these topics appears larger in the treatment over the control, but only one was significant: the increase in interest in oak trees is significantly larger in the treatment group ($M = 1.12$, $SD = 0.91$) compared to the control group ($M = 0.52$, $SD = 1.08$) using a Mann–Whitney U test: $Z = 163.5$, $p = 0.014$.

These findings suggest that **the treatment leads to a greater increase in interest in the topic than the control**.

6.3 Qualitative Observations

Here, we share some common themes in interview responses, including differences we found between participants in the *Oak Story* and control groups. We provide the condition (Treatment or Control) and grade level (4th–6th) for each quoted participant. Quotes are lightly edited for clarity.

6.3.1 Oak Story *supported participants' sense of agency.* *Oak Story* allowed participants to write freeform responses to the questions posed by characters in the narrative, allowing them more freedom to express themselves. P20 (Treatment, 5th) explained, "I actually loved [writing responses] because **I get to type my own opinions** and stuff".

While participants across both groups appreciated the open-ended response format, those in the treatment group noticed that the system was giving them opportunities to make meaningful choices to shape both their own learning and the direction of the story. P17 (Treatment, 5th) described their feelings as follows:

> Well, it asked me what I would want to do next. So it's really generous. That helped me think that **it gives me the choice and gives me the power to think about it**, and when I say what I want, it responds with a really helpful answer, like it just responds exactly to what I want. I felt like I had a lot of choices.

6.3.2 Oak Story *generated dialogue that felt responsive to participants.* Participants in the treatment group regularly emphasized that the characters responded meaningfully to their inputs. For example, P14 (Treatment, 6th) remarked, "I liked how the animals took a minute to process the information, and then they actually gave me extremely useful answers, and **they knew exactly what I was asking**, and they gave me very perfect answers. **So then I learned better."**

The precision of *Oak Story* in responding to inputs not only entertained participants, but it also motivated them to learn more. Another participant expressed increased interest, saying, "I like that the characters are really responsive, and the responses are actually pretty nice, like they can easily take what I said and make it like, normally talking to a real person... **I hate science, but I can learn a lot better from this**" (P6, Treatment, 6th).

	Range	Pre	Post	p
Interest in animals and local wildlife	[2, 10]	7.57 [1.70]	7.98 [1.70]	$p = 0.026$
Interest in plants, trees, and oak trees	[3, 15]	10.1 [2.64]	12.0 [2.45]	$p < 0.001$
Interest in biology and ecology	[2, 10]	7.09 [1.85]	7.68 [1.87]	$p = 0.026$

Table 3: Ranges, pre-app scores, and post-app scores for participant interest in scientific topics related to *Oak Story*.

Meanwhile, participants in the control group sometimes expressed reservation about system comprehension: "Well, **sometimes it doesn't understand my response very well**" (P54, Control, 4th). This suggests that, even though the control version of *Oak Story* also used an LLM to respond to participants' messages, its inability to meaningfully act on those messages hurt the overall experience.

6.3.3 Oak Story *was able to handle off-topic inputs.*

While the vast majority of participants engaged earnestly with *Oak Story*, a few participants tested the limits of the system with unexpected inputs. The guardrails in *Oak Story*'s LLM architecture could handle these situations, either by ignoring the input altogether (if they were non sequiturs) or steering the conversation back on track. We provide an example conversation where this occurred:

> Oliver: Now that we're back at the central area, we have two things we could do. We could look for a tree and see if it's an oak, or we could start thinking of how to help Tina find a new home. What would you like to do next?
> P34 (Treatment, 5th): Tina Home give her boba and in n out [a fast food restaurant]
> Oliver: That's a fun idea, [name]! **While boba and In-N-Out might make us happy, Tina needs a new home more than a snack.** We can help by gathering ideas from our animal friends, or we can explore a nearby tree and see if it's an oak. What do you think we should do?

In this example, our system was able to recognize P34's off-topic input, acknowledge their answer, and connect it back to the task at hand.

6.3.4 *Participants valued* Oak Story*'s context-awareness and wanted more.*

In certain scenarios, the open-ended prompting allowed participants to make requests or observations that the system was not fully equipped to handle due to limited contextual awareness. One such example came with the final rehoming activity, where we intended for participants to ask the acorn woodpecker to peck a new hole (in augmented reality) for the titmouse in the oak tree they'd previously identified—some participants, however, recognized holes that were already present on the tree as a potential home for Tina. The app was not advanced enough to be aware of this and was unable to effectively respond to these observations.

While the freedom afforded to participants allowed limitations like these to be exposed, these unexpected outcomes indicate the variety of responses among participants and the natural manner with which they interacted with the app. In other words, *Oak Story* encouraged them to be creative and think for themselves, to the extent that they made unexpected discoveries. Occasionally, participants would even incorrectly assume context-awareness. For example, P42 (Treatment, 6th) prescribed an ecological meaning to the randomness of the acorn-digging activity: "Finding the acorns was kind of challenging... and then **I remembered that squirrels usually bury their acorns by oak trees**. So I went to the oak tree, and they're all really close."

6.3.5 *Participants in both conditions enjoyed the outdoor, real-world activities.*

Finally, almost all participants across both conditions expressed that they enjoyed the real-world activities. Many participants highlighted *Oak Story*'s active approach as refreshingly different from their usual repetitive school experiences. P38 (Control, 6th) remarked, "I liked that you had to go outside... **Because that way you actually get to do something, it would help you focus more**. Like fidgeting with something, like doing something, not just staring at a textbook." Specifically, they appreciated that active interaction replaced idleness; one observed, "It included a lot of physical stuff instead of just sitting around and not doing anything and just watching it instead of interacting with it" (P39, Control, 4th).

Participants also positively compared *Oak Story* to other mobile apps, noting that the outdoor portions prevented boredom and made the app more challenging. P56 (Treatment, 6th) emphasized, "Well, [going outside] is more interesting. It's like, the other games, you're usually just like sitting on your computer in the classroom... I liked how you had to walk around outside, scavenging around, trying to find things like acorns and oak trees." Another participant described how *Oak Story* prevented sedentary behavior, a common problem in children when using screens: "**I feel like not enough apps really take into consideration that kids shouldn't be just glued onto the screen**; even if they're learning something they should also be able to go outside and interact with the real world" (P43, Treatment, 6th).

7 DISCUSSION

Designing and evaluating *Oak Story* provided insights into how agency and interactivity support children's learning and engagement. In this section, we synthesize our findings and discuss the implications of our work.

7.1 Improved Learning Outcomes

To reiterate our quantitative findings, *Oak Story* is superior to the control on three core measures: learning gains, engagement, and interest in the subject matter. Additionally, participants in the treatment group feel that the characters are more responsive and that their own actions and messages influence the story more. Because we designed the control to be almost identical to the treatment (they share the same interface, narrative scenes, and real-world activities), we conclude that subtle difference of having the system adapt the progression of the story to the learner significantly enhances participants' learning gains and feelings of control. We isolate and attribute these increases to the greater *agency* that *Oak Story* affords over the control. While both conditions had dialogue that was responsive to participants, only the treatment gave participants any meaningful control within the narrative. This increase in agency is further supported by our qualitative findings, where we observe that participants in the treatment condition express excitement about the tangible impact of their messages and the specificity of character responses to their questions and interests.

While *Oak Story* does not afford total freedom, it provides enough agency over the control to demonstrate a measurable effect on learners. Indeed, too much agency, such as an open-ended chatbot that can steer the conversation in any direction, can hinder learning as learners may struggle to self-regulate without appropriate structure and support. Our findings align with prior work that supporting a *moderate* level of agency in narrative-driven learning environments can lead to improved learning outcomes and increased cognitive stimulation [57].

7.2 Design Implications

Here, we discuss implications for design based on the findings from our study with *Oak Story*.

7.2.1 Design LLM-based experiences to support learner agency. A fundamental principle of student-centered learning is learner agency: empowering students to make meaningful decisions about their learning trajectory and methodology. The emergence of LLMs opens new possibilities for personalized learning but also raises concerns about potential overreliance and the erosion of learner autonomy. [14]. Our study contributes to the growing discourse on AI's impact on learner agency by demonstrating how LLMs can be thoughtfully integrated to enhance rather than diminish student autonomy. Recent work has explored similar approaches, such as using LLMs as dynamic writing coaches [2] and adaptive question generators [51], while carefully preserving learner initiative and decision-making. *Oak Story* offers a way to take advantage of LLMs' affordances to accept and make sense of open-ended input from the learner to shape their learning experience. In the real world, the space of actions that the learner can take is broad—perhaps too broad–and *Oak Story* supports agency by giving learners the same sense of open-endedness, while pairing the learner's input with the narrative state to translate their intent into narrative progression and characters' reactions. This result suggests important implications for interaction design: we show that it is possible for learner conversations with LLMs to move beyond the chatbot interface, and that LLMs can instead be integrated with interactive narratives and augmented reality to enable responsive interaction and meaningful incorporation of learner input. In this way, learners' interactions are more effortful, but their effort is rewarded.

7.2.2 Align LLM-driven narratives with learning goals. When designing interactive narratives, where the user makes choices to impact the story, there is a central tension between coherent story progression (or authorial intent) and the user's agency [46]. Recent work has explored building story management architectures on top of LLMs to address this challenge (e.g., [56]). Our work extends this line of research by focusing on the domain of education. When these narratives also need to be educational, another constraint emerges: the interactive story must now also support a given curriculum and allow the learner to achieve certain learning objectives.

Oak Story addresses these challenges through a novel learning-goal-directed LLM architecture that strikes a balance between narrative structure, learning objectives, and learner freedom. Each scene within the narrative provides specific goals expressed in natural language. These goals are flexible, supporting both *system-driven* goals that can orient the story towards specific narrative needs (e.g., "`Tina has mentioned that she lost her home`") or topics to teach (e.g., "`Oliver explains how oak trees benefit humans and cities`") and *learner-initiated* goals (e.g., "`The learner proposes a method to protect oak trees`").

Our architecture tracks both completed and unmet goals and prompts the LLM to select and pursue an unmet goal. This allows the system to prioritize scenes that align with a learner's preferences while ensuring all educational objectives are eventually addressed. By maintaining a comprehensive view of which goals have been met and which remain outstanding, the system can make informed decisions about producing character dialogue that balance responsiveness to learner interests with educational requirements.

We found our architecture to be successful at balancing narrative and educational constraints while still promoting self-paced exploration, and provide our prompt template in the Appendix to support the future development of this method.

While *Oak Story* was created to be a tailored learning experience for a single domain (ecology) and focused on a particular phenomenon (oak ecosystems), we believe that our approach can generalize to other AI-supported learning environments. By orienting an LLM toward specific goals and creating an architecture to manage those goals as they change throughout an experience, a designer can create an interactive, narrative-driven learning experience where the learner can make meaningful choices while achieving desired learning objectives.

7.2.3 Create characters with specialized knowledge bases. To ensure that characters provide accurate information and mitigate hallucinations [70], our architecture incorporates a retrieval-augmented generation approach, similar to recent work [9]. Additionally, we distinguish between characters by providing each one with a different knowledge base. This allows each character to be the "expert" on the subtopics relating to their niche in the ecosystem.

Our approach offers several advantages over relying solely on the LLM's parametric knowledge. First, it substantially improves factual accuracy by grounding results in validated content. Second, it creates a more responsive conversational experience, as characters can directly build on the topics mentioned by the learner. Finally, it helps foster a richer world of characters with not only diverse personalities, but also diverse areas of expertise.

7.2.4 Explore bringing narratives into the real world. While narrative on its own is an effective way to engage learners, we found in the post-study interviews that participants tended to highlight the real-world augmented reality activities as their favorite. We also observed from the interviews that some participants enjoyed the fact that events in the narrative, such as finding Tina the oak titmouse a new home, connected directly to the real-world activities. This suggests that engagement through narrative and engagement through physical exploration could have an amplifying effect when employed together.

Finally, we also note that the responsiveness of the conversational aspects of the narrative increases the immersion and expectations of learners with regards to other parts of the app. For example, one AR activity involves finding Tina a new home and requires the learner to create a (virtual) hole in a (real) oak tree. Some participants observed a real hole in the tree, and attempted to use it for Tina's new home. We thus suspect that, because one aspect of our app (text input) allows for a high degree of freedom and understands their intent, participants may incorrectly assume that this level of understanding extends to all aspects of the experience.

8 LIMITATIONS AND FUTURE WORK

While *Oak Story* demonstrates significant potential in enhancing learner agency in narrative-based education, we acknowledge the limitations of our research and offer potential areas for future research.

8.0.1 Accuracy, safety, and transparency. The application of language models in children's education amplifies concerns about accuracy, bias, and safety [29, 66]. The potential for model hallucinations or inconsistencies [36, 66] becomes particularly concerning when learners may not fully understand that they are interacting with a tool prone to these errors. Interviews with participants revealed that several learners believed they were communicating with a human or a system capable of real-time web searches, highlighting the need for additional safeguards or transparent communication about the nature of the interaction without breaking immersion. Exploring techniques to improve the factual accuracy of language model outputs, such as improved retrieval-augmented generation or fact-checking mechanisms, is thus crucial for reliability in educational settings.

8.0.2 Understanding learner needs. Effectively adapting to diverse learner needs remains challenging. While *Oak Story* aims to cater to varying grade levels and engagement styles, the nuanced decision-making required to balance challenging learners and allowing uncertainty is an area requiring further refinement. The system's ability to make effective judgments in this regard impacts both the learning outcomes and the learner's overall experience.

Future research might focus on advancing the adaptive capabilities of *Oak Story* through improved learner understanding, such as capturing and storing "memories" about the learner's learning preferences and progress over time. Additionally, investigating methods for real-time assessment of learner engagement and comprehension could accompany the existing self-assessment, further supporting learner autonomy and growth.

8.0.3 Language model constraints. At the core of *Oak Story* is a large language model, which, while powerful, introduces a set of complex challenges. The use of prompting techniques like chain-of-thought reasoning [65] and few-shot learning [6] improves the consistency of outputs but comes with tradeoffs. Chain-of-thought reasoning, for instance, increases the coherence of the model's responses but produces more output tokens, potentially introducing latency that can disrupt the flow of interaction. Similarly, few-shot examples, while helpful in guiding the model's behavior, risk overfitting to specific scenarios, potentially limiting the system's flexibility.

8.0.4 Multimodal and extended reality integration. In its present form, *Oak Story* utilizes AR activities as complementary components to the text-based narrative interactions. These AR elements offer learners the opportunity to engage with concepts introduced in the story through immersive, real-world interactions. However, the connection between the language model driving the narrative and the AR experiences remains somewhat disconnected. The language model receives a text description of what the learner sees in AR during the activity or after its completion, but it lacks real-time insight into the learner's actions and experiences during the interaction.

Exploring the continued use of multimodal language models could enable real-time processing of visual and interactive data from AR activities. This would allow the narrative to adapt dynamically based on the learner's actions and discoveries in the AR environment. For instance, if a learner struggles with a particular AR task, characters could immediately offer guidance or adjust the narrative to provide additional context.

Developing mechanisms for bidirectional communication between the AR components and the language model could also create a more cohesive learning experience—currently, *Oak Story* can launch activities on the client but does not receive any synchronous data about it. Improvements in this area might involve designing protocols for the AR system to send real-time updates to the language model, allowing it to incorporate immediate feedback or observations into the narrative flow. Advancing the LLM's ability to interpret the semantic meaning of AR interactions could enhance the educational value of these experiences. This might involve integration of existing semantic segmentation capabilities in Niantic Lightship or methods for the language model to influence or generate AR content in real-time.

8.0.5 Educator involvement and authoring tools. The current iteration of *Oak Story* requires substantial involvement from educational practitioners in shaping the learning experience. Moreover, the sensitivity of language models to minor variations in prompts poses a challenge for non-expert users in eliciting desired behaviors consistently.

To address these limitations, future work might involve the development of user-friendly authoring tools that empower educators to create and customize narrative-based learning experiences with minimal technical expertise. Drawing inspiration from platforms like Twine [18], these tools could provide intuitive interfaces for defining character personalities and knowledge, setting educational goals, and structuring narrative flows.

8.0.6 Limitations of our study. Our experimental study was conducted at our institution's campus, which is located in an relatively affluent area with a skewed population distribution relative to national (and international) demographics. Although we made great efforts to recruit widely across nearby cities and counties with different demographic makeups, the children in our sample may not be representative of the broader population. For example, none of our participants identified as Black.

Also, lower-income families within driving distance did not always have the means to commute to our institution to participate in the study. While we were able to run two sessions in families' neighborhoods for this current study, we hope to conduct future studies closer to individual families given evidence that *Oak Story* is flexible enough to operate in alternative locations.

9 CONCLUSION

In this work, we presented *Oak Story*, a designed to support learner agency through interactive, narrative-driven experiences. *Oak Story* facilitates and interprets learners' conversations with story characters, uses this input to direct the progression of the story, and extends the narrative into the learner's real-world environment through augmented reality activities.

Our contributions include (1) the design and implementation of a narrative framework that bridges LLM-mediated dialogue with AR-based activities to support situated learning, and (2) empirical evidence showing that adaptive narrative sequencing, made possible through our novel learning-goal-oriented architecture, leads to improved learning gains, engagement, and a heightened sense of agency among children. Importantly, our architecture balances learner agency with system-driven structure, addressing a key tension in the design of educational interactive narratives.

As educators and interaction designers continue to explore the use of LLMs in educational and exploratory settings, we believe our work offers a potential framework for building interactive systems that are both expressive and pedagogically grounded.

ACKNOWLEDGMENTS

We are grateful to TAL Education Group and the Stanford Institute for Human-Centered Artificial Intelligence (HAI) for providing funding for this research. We also thank Jane E for providing feedback on this manuscript. Finally, we express our gratitude to the wonderful families who participated in our study.

REFERENCES

[1] Lorin W. Anderson and David R. Krathwohl (Eds.). 2001. *A Taxonomy for Learning, Teaching, and Assessing. A Revision of Bloom's Taxonomy of Educational Objectives* (2 ed.). Allyn & Bacon, New York.

[2] Riku Arakawa and Hiromu Yakura. 2024. Coaching copilot: blended form of an LLM-powered chatbot and a human coach to effectively support self-reflection for leadership growth. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 1–14.

[3] Albert Bandura. 2001. Social cognitive theory: An agentic perspective. *Annual review of psychology* 52, 1 (2001), 1–26.

[4] Albert Bandura. 2006. Toward a psychology of human agency. *Perspectives on psychological science* 1, 2 (2006), 164–180.

[5] Fredrik S Breien and Barbara Wasson. 2021. Narrative categorization in digital game-based learning: Engagement, motivation & learning. *British Journal of Educational Technology* 52, 1 (2021), 91–111.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *ArXiv* abs/2005.14165 (2020). https://arxiv.org/abs/2005.14165

[7] Rodger W Bybee. 1990. Science for life & living: An elementary school science program from biological sciences curriculum study. *The American Biology Teacher* 52, 2 (1990), 92–98.

[8] James Carifio and Rocco J Perla. 2007. Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of social sciences* 3, 3 (2007), 106–116.

[9] Jiaju Chen, Minglong Tang, Yuxuan Lu, Bingsheng Yao, Elissa Fan, Xiaojuan Ma, Ying Xu, Dakuo Wang, Yuling Sun, and Liang He. 2025. Characterizing LLM-Empowered Personalized Story-Reading and Interaction for Children: Insights from Multi-Stakeholder Perspectives. *arXiv preprint arXiv:2503.00590* (2025).

[10] Alan Y Cheng, Meng Guo, Melissa Ran, Arpit Ranasaria, Arjun Sharma, Anthony Xie, Khuyen N Le, Bala Vinaithirthan, Shihe Luan, David Thomas Henry Wright, et al. 2024. Scientific and Fantastical: Creating Immersive, Culturally Relevant Learning Experiences with Augmented Reality and Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–23.

[11] J. Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

[12] Chris Crawford. 2013. Interactive storytelling. In *The video game theory reader*. Routledge, 259–273.

[13] Scott J Dankel and Jeremy P Loenneke. 2021. Effect sizes for paired data should use the change score variability rather than the pre-test variability. *The Journal of Strength & Conditioning Research* 35, 6 (2021), 1773–1778.

[14] Ali Darvishi, Hassan Khosravi, Shazia Sadiq, Dragan Gašević, and George Siemens. 2024. Impact of AI assistance on student agency. *Computers & Education* 210 (2024), 104967.

[15] Edward L Deci and Richard M Ryan. 2008. Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian psychology/Psychologie canadienne* 49, 3 (2008), 182.

[16] Griffin Dietz, Zachary Pease, Brenna McNally, and Elizabeth Foss. 2020. Giggle gauge: a self-report instrument for evaluating children's engagement with technology. In *Proceedings of the Interaction Design and Children Conference*. 614–623.

[17] Lena Ballone Duran and Emilio Duran. 2004. The 5E instructional model: A learning cycle approach for inquiry-based science teaching. *Science Education Review* 3, 2 (2004), 49–58.

[18] Interactive Fiction Technology Foundation. 2024. Twine: An open-source tool for telling interactive, nonlinear stories. https://twinery.org/. Accessed 09/13/2024.

[19] Jacob Garbe, Max Kreminski, Ben Samuel, Noah Wardrip-Fruin, and Michael Mateas. 2019. StoryAssembler: an engine for generating dynamic choice-driven narratives. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*. 1–10.

[20] Stefan Göbel, Viktor Wendel, Christopher Ritter, and Ralf Steinmetz. 2010. Personalized, adaptive digital educational games using narrative game-based learning objects. In *Entertainment for Education. Digital Techniques and Systems: 5th International Conference on E-learning and Games, Edutainment 2010, Changchun, China, August 16-18, 2010. Proceedings 5*. Springer, 438–445.

[21] Melanie C Green and Keenan M Jenkins. 2014. Interactive narratives: Processes and outcomes in user-directed stories. *Journal of Communication* 64, 3 (2014), 479–500.

[22] Michael J Hannafin and Susan M Land. 1997. The foundations and assumptions of technology-enhanced student-centered learning environments. *Instructional science* 25 (1997), 167–202.

[23] John Hattie and Eric M Anderman. 2013. *International guide to student achievement*. Vol. 711. Routledge New York, NY.

[24] Inc. Heroku. 2024. Heroku. https://www.heroku.com/. Accessed: 2024-09-12.

[25] Bruce Johnson and Constantinos C Manoli. 2010. The 2-MEV scale in the United States: a measure of children's environmental attitudes based on the theory of ecological attitude. *The Journal of Environmental Education* 42, 2 (2010), 84–97.

[26] Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine* 15, 2 (2016), 155–163.

[27] Max Kreminski and Chris Martens. 2022. Unmet creativity support needs in computationally supported creative writing. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*. 74–82.

[28] Max Kreminski and Noah Wardrip-Fruin. 2018. Sketching a map of the storylets design space. In *International Conference on Interactive Digital Storytelling*. Springer, 160–164.

[29] Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2022. Language generation models can cause harm: So what can we do about it? An actionable survey. *arXiv preprint arXiv:2210.07700* (2022).

[30] Secret Lab. 2022. Yarn Spinner. https://yarnspinner.dev/ Accessed: 2022-09-13.

[31] Jean Lave and Etienne Wenger. 1991. *Situated learning: Legitimate peripheral participation*. Cambridge university press.

[32] Jungeun Lee, Suwon Yoon, Kyoosik Lee, Eunae Jeong, Jae-Eun Cho, Wonjeong Park, Dongsun Yim, and Inseok Hwang. 2024. Open Sesame? Open Salami! Personalizing Vocabulary Assessment-Intervention for Children via Pervasive Profiling and Bespoke Storybook Generation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–32.

[33] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel,

et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[34] Zhuoran Lu, Qian Zhou, and Yi Wang. 2025. WhatELSE: Shaping narrative spaces at configurable level of abstraction for AI-bridged interactive storytelling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[35] Areej Mawasi, Peter Nagy, Ed Finn, and Ruth Wylie. 2022. Narrative-based learning activities for science ethics education: an Affordance perspective. *Journal of Science Education and Technology* 31 (2022), 16–26.

[36] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661* (2020).

[37] Scott W McQuiggan, Jonathan P Rowe, Sunyoung Lee, and James C Lester. 2008. Story-based learning: The impact of narrative on learning experiences and outcomes. In *International conference on intelligent tutoring systems*. Springer, 530–539.

[38] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–34.

[39] Aquiles Negrete. 2005. *Fact via Fiction: Stories that Communicate Science*. The Pantaneto Press, UK, 95–102. https://doi.org/10.13140/RG.2.1.5110.1207

[40] Niantic Lightship. 2023. Lightship ARDK. https://lightship.dev/products/ardk/. [Online; accessed 10-September-2023].

[41] Nathaniel S O'Connell, Lin Dai, Yunyun Jiang, Jaime L Speiser, Ralph Ward, Wei Wei, Rachel Carroll, and Mulugeta Gebregziabher. 2017. Methods for analysis of pre-post data in clinical research: a comparison of five common methods. *Journal of biometrics & biostatistics* 8, 1 (2017), 1.

[42] OpenAI. 2024. Function Calling - OpenAI API. https://platform.openai.com/docs/guides/function-calling. Accessed 09/13/2024.

[43] OpenAI. 2024. Introducing GPT-4o. https://openai.com/index/hello-gpt-4o/ Accessed: 2024-09-12.

[44] Zhenhui Peng, Xingbo Wang, Qiushi Han, Junkai Zhu, Xiaojuan Ma, and Huamin Qu. 2023. Storyfier: Exploring vocabulary learning support with text generation models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–16.

[45] Nichole Pinkard, Sheena Erete, Caitlin K Martin, and Maxine McKinney de Royston. 2017. Digital youth divas: Exploring narrative-driven curriculum to spark middle school girls' interest in computational activities. *Journal of the Learning Sciences* 26, 3 (2017), 477–516.

[46] Mark Owen Riedl and Vadim Bulitko. 2013. Interactive narrative: An intelligent systems approach. *AI Magazine* 34, 1 (2013), 67–67.

[47] Jonathan P Rowe and James C Lester. 2015. Improving student problem solving in narrative-centered learning environments: A modular reinforcement learning framework. In *Artificial Intelligence in Education: 17th International Conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings 17*. Springer, 419–428.

[48] Jonathan P Rowe, Lucy R Shores, Bradford W Mott, and James C Lester. 2011. Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education* 21, 1-2 (2011), 115–133.

[49] Sherry Ruan, Jiayu He, Rui Ying, Jonathan Burkle, Dunia Hakim, Anna Wang, Yufeng Yin, Lily Zhou, Qianyao Xu, Abdallah AbuHashem, et al. 2020. Supporting children's math learning with feedback-augmented narrative technology. In *Proceedings of the interaction design and children conference*. 567–580.

[50] Robert Sawyer, Andy Smith, Jonathan Rowe, Roger Azevedo, and James Lester. 2017. Is more agency better? The impact of student agency on game-based learning. In *Artificial Intelligence in Education: 18th International Conference, AIED 2017, Wuhan, China, June 28–July 1, 2017, Proceedings 18*. Springer, 335–346.

[51] Griffin Dietz Smith, Siddhartha Prasad, Matt J Davidson, Leah Findlater, and R Benjamin Shapiro. 2024. ContextQ: Generated Questions to Support Meaningful Parent-Child Dialogue While Co-Reading. *arXiv preprint arXiv:2405.03889* (2024).

[52] Erica L Snow, Laura K Allen, Matthew E Jacovina, and Danielle S McNamara. 2015. Does agency matter?: Exploring the impact of controlled behaviors within a game-based environment. *Computers & Education* 82 (2015), 378–392.

[53] E. Spotswood, R. Grossinger, and S. Hagerty. 2017. *Re-Oaking Silicon Valley: Building Vibrant Cities with Nature*. San Francisco Estuary Institute. https://books.google.com/books?id=zJUotAEACAAJ

[54] Next Generation Science Standards. 2022. Next generation science standards. https://www.nextgenscience.org/

[55] Yuqian Sun, Zhouyi Li, Ke Fang, Chang Hee Lee, and Ali Asadipour. 2023. Language as reality: a co-creative storytelling game experience in 1001 nights using generative AI. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 19. 425–434.

[56] Yuqian Sun, Phoebe J Wang, John Joon Young Chung, Melissa Roemmele, Taewook Kim, and Max Kreminski. 2025. Drama Llama: An LLM-Powered Storylets Framework for Authorable Responsiveness in Interactive Narrative. *arXiv preprint arXiv:2501.09099* (2025).

[57] Michelle Taub, Robert Sawyer, Andy Smith, Jonathan Rowe, Roger Azevedo, and James Lester. 2020. The agency effect: The impact of student agency on learning, emotions, and problem-solving behaviors in a game-based learning environment. *Computers & Education* 147 (2020), 103781.

[58] David Thue, Vadim Bulitko, Marcia Spetch, and Eric Wasylishen. 2007. Interactive storytelling: A player modelling approach. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 3. 43–48.

[59] Hsiao-Lin Tuan*, Chi-Chin Chin, and Shyang-Horng Shieh. 2005. The development of a questionnaire to measure students' motivation towards science learning. *International journal of science education* 27, 6 (2005), 639–654.

[60] Unity. 2023. Unity Real-Time Development Platform. https://unity.com/. [Online; accessed 10-September-2023].

[61] Uvicorn. 2024. Uvicorn. https://www.uvicorn.org/. Accessed: 2024-09-12.

[62] Margaret Vaughn. 2021. *Student agency in the classroom: Honoring student voice in the curriculum*. Teachers College Press.

[63] Lev S Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Vol. 86. Harvard university press.

[64] Yi Wang, Qian Zhou, and David Ledo. 2024. StoryVerse: Towards co-authoring dynamic plot with LLM-based character simulation via narrative planning. In *Proceedings of the 19th International Conference on the Foundations of Digital Games*. 1–4.

[65] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv* abs/2201.11903 (2022). https://arxiv.org/abs/2201.11903

[66] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[67] C Gordon Wells. 1999. *Dialogic inquiry*. Vol. 10. Cambridge University Press Cambridge.

[68] Ying Xu, Joseph Aubele, Valery Vigil, Andres Sebastian Bustamante, Young-Suk Grace Kim, and Mark Warschauer. 2021. Dialogue with a conversational agent promotes children's story comprehension via enhancing engagement. *Child Development* 93 (2021), e149 – e167. https://srcd.onlinelibrary.wiley.com/doi/10.1111/cdev.13708

[69] Chao Zhang, Xuechen Liu, Katherine Ziska, Soobin Jeon, Chi-Lin Yu, and Ying Xu. 2024. Mathemyths: leveraging large language models to teach mathematical language through Child-AI co-creative storytelling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–23.

[70] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219* (2023).

[71] Andrew Zhu, Lara Martin, Andrew Head, and Chris Callison-Burch. 2023. CALYPSO: LLMs as Dungeon Master's Assistants. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 19. 380–390.