Scientific and Fantastical: Creating Immersive, Culturally-Relevant Learning Experiences with Augmented Reality and Large Language Models

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

Meng Guo*
Stanford University
Stanford, CA, USA
mguo19@stanford.edu

Melissa Ran
Stanford University
Stanford, CA, USA
mran24@stanford.edu

Arpit Ranasaria
Stanford University
Stanford, CA, USA
arpitr@stanford.edu

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

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

Khuyen N. Le
UC San Diego
La Jolla, CA, USA
knl005@ucsd.edu

Bala Vinaithirthan
Stanford University
Stanford, CA, USA
balavinaithirthan@stanford.edu

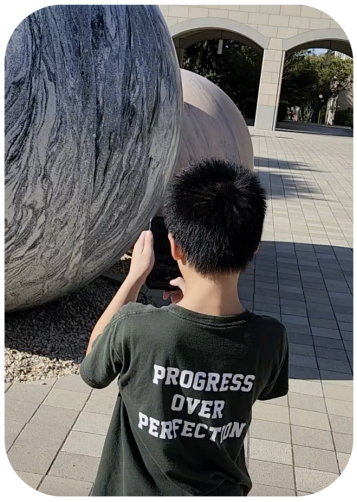
Shihe (Tracy) Luan
Stanford University
Stanford, CA, USA
tluan@stanford.edu

David Thomas Henry Wright
Nagoya University
Nagoya, Japan
wright.david.thomas.henry.p6@f.mail.nagoya-u.ac.jp

Andrea Cuadra
Stanford University
Stanford, CA, USA
apcuad@stanford.edu

Roy Pea
Stanford University
Stanford, CA, USA
roypea@stanford.edu

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



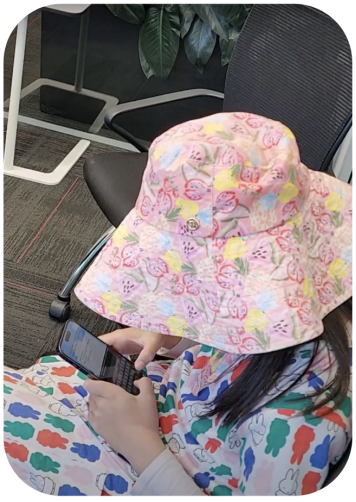


Figure 1: Our system, *Moon Story*, is a mobile app that uses narrative to engage children in culturally-relevant experiential learning. In the image on the far left, a participant interacts with physical landmarks representing planets while using *Moon Story*. The screenshot in the middle left shows augmented reality annotations that support learning about the solar system's dimensions. The image on the far right portrays a participant writing letters to Chang'e, a character in the narrative that is displayed on the image in the middle right. In one of the experimental conditions of our study, the content of the letters written back by Chang'e is personalized to what each participant writes using a large language model.

ABSTRACT

Motivating children to learn is a major challenge in education. One way to inspire motivation to learn is through immersion. We combine the immersive potential of augmented reality (AR), narrative, and large language models (LLMs) to bridge fantasy with reality in a mobile application, *Moon Story*, that teaches elementary schoolers astronomy and environmental science. Our system also builds upon learning theories such as culturally-relevant pedagogy. Using our application, a child embarks on a journey inspired by Chinese mythology, engages in real-world AR activities, and converses with a fictional character powered by an LLM. We conducted a controlled experiment ($N = 50$) with two conditions: one using an LLM and one that was hard-coded. Both conditions resulted in learning gains, high engagement levels, and increased science learning motivation. Participants in the LLM condition also wrote more relevant answers. Finally, participants of both Chinese and non-Chinese heritage found the culturally-based narrative compelling.

CCS CONCEPTS

• **Human-centered computing** → *Interactive systems and tools*; **Mobile devices**; **Mixed / augmented reality**.

KEYWORDS

Education/Learning; Children/Parents; Artifact or System

ACM Reference Format:

Alan Y. Cheng*, Meng Guo*, Melissa Ran, Arpit Ranasaria, Arjun Sharma, Anthony Xie, Khuyen N. Le, Bala Vinaithirthan, Shihe (Tracy) Luan, David Thomas Henry Wright, Andrea Cuadra, Roy Pea, and James A. Landay. 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 (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3613904.3642041

1 INTRODUCTION

Motivating children to learn is a major challenge in education. High levels of intrinsic motivation to learn have been shown to be essential for better learning outcomes [24, 25]. One way to inspire intrinsic motivation in learning is through immersion [20, 29]. A vast body of research has investigated the use of technologies such as virtual reality (VR) and augmented reality (AR) to create immersive learning experiences [16, 28]. In particular, AR has the unique affordance of creating immersive learning environments that combine virtual objects with the physical world [17].

Narrative also deepens immersion [46], as it transports the reader into another world [22] and increases a learner's sense of presence [44]. By blending AR and narrative, a learning environment has the potential to bridge a fantasy world with the real world, which can lead to even higher levels of immersion and improved learning gains [21]. Moreover, if the narrative integrates ideas from culturally relevant pedagogy [38] by being culturally-inspired, it adds yet another level of immersion: cultural immersion. Furthermore, by using large language models (LLMs) to power characters in the story, a learner can participate in the narrative by directly talking to its characters. However, research on AR-driven, narrative-based, culturally-relevant, interactive learning is under-explored in HCI, and there is still much unknown about how to design these kinds of multidimensionally immersive learning experiences.

In this work, we design and evaluate an AR-driven, narrative-based learning environment to immerse elementary school learners in a culturally enriched fantasy. Our system, *Moon Story*, integrates AR educational activities with a narrative inspired by Chinese mythology to guide lessons about climate change and astronomy. Additionally, it leverages the interactive capabilities of LLMs to further the system's immersive capabilities and learning objectives by bringing a fictional character from the narrative to life in a reflective letter-writing activity.

We have investigated three primary research questions by using the system we created in a controlled experiment (N=50) to compare children's experiences between two conditions: one using an LLM and one that was hard-coded.

RQ1: How do elementary school students interact with and respond to a system that uses AR-supported interactive narrative for learning over time?

RQ2: What are the opportunities and challenges of integrating an LLM to enable richer interactions with a character in the narrative?

RQ3: How does this immersive learning system affect student learning gains, engagement, and motivation to learn?

In addition, we explore a secondary research question by comparing the responses of non-Chinese participants (n=33) with those of Chinese participants (n=17).

RQ4: How do Chinese and non-Chinese children's appraisals of their experiences compare when engaging with a technology-based learning experience inspired by Chinese mythology?

*These authors contributed equally to this work.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642041

We make two major contributions to the literature. First, we introduce *Moon Story*, a narrative-based, LLM-enhanced mobile AR system for culturally-relevant learning experiences. Second, we provide quantitative and qualitative results from an experimental study investigating how immersive educational technology is used by children who have just completed second, third, fourth, or fifth grade. Quantitative results show that there are statistically significant improvements in learning gains and motivation towards science learning after using the application, with small to medium effect sizes. Children also reported high levels of engagement. In addition, the LLM variant demonstrated effectiveness in encouraging relevant responses and addressing disengagement issues in the final environmental reflection writing task. Qualitatively, children expressed enthusiasm for interactive outdoor explorations using AR. We also identified characteristics of LLM-mediated educational experiences and children's mental models of LLMs within those experiences. Additionally, we found that participants of Chinese ethnicity expressed a sense of belonging from the culturally relevant narrative, while non-Chinese participants also responded positively. Since there is already a large body of research comparing AR and non-AR learning experiences, our study focuses on the comparison of LLM-based and non-LLM-based experiences. We aim to isolate and examine the impact of LLMs as an emerging technology when integrated with immersive learning designs, AR, and narratives to contribute to the growing literature in this field.

In this paper, we first situate our study within the existing literature, describing the learning theories we build upon and the related work we contribute to. We then explain our design process—how we developed the learning content and narrative. We describe our system in detail, documenting the core features such as the narrative, in-app AR activities, and LLM-based interactions. Next, we detail our methods and share our findings. Finally, we discuss the implications of our findings, avenues for future work, and limitations of our study. As a whole, our artifact and study's findings contribute to a growing body of literature on immersive, hybrid learning experiences blending ancient, fantastical cultural resources with the new technologies of AR and LLMs.

2 RELATED WORK

Here, we first describe narrative-based learning, which forms the core of our system. Next, we situate our work within existing education research on the two main technologies we study: augmented reality and large language models. Finally, we reviewed other theories of learning that informed our artifact and study design. We highlight how our work is unique in that our system combines narrative-based, immersive, culturally-relevant pedagogy.

2.1 Narrative-Based Learning

Our work builds upon research on narrative-based learning environments, which integrate storytelling with learning. Narratives provide many potential benefits to learners. A narrative can increase engagement and immersion by involving learners in the story and empowering them to take actions within the narrative [46, 59]. Including narrative elements in science and math education can help students retain information [47, 59]. Narrative can also help learners develop important skills, such as creativity and critical thinking [19, 57], as well as affect learners' interest and sense of identity [56]. Our system, *Moon Story*, uses narrative to engage learners and structure the learning experience, while also extending past research on narrative-based learning environments by exploring how integrating a cultural myth into a learning experience can support learners.

2.2 Augmented Reality in Education

Augmented reality (AR) technologies, first defined by Thomas and David [66], "augment" normal perception by superimposing virtual objects onto a person's visual field. AR positions learners in a real-world physical context, which facilitates situated learning [40] and engages learners' spatial cognition [45]. AR has been used to support children's education in a wide variety of scientific fields, such as mathematics [34] and ecology [7, 15]. Prior work found positive outcomes associated with using AR on motivation, learning outcomes, engagement, and immersion [7, 8, 20, 64]. Children's learning motivation starts to decrease in grade four [27], and AR educational activities provide an opportunity to re-engage these children by enhancing the learning experience and facilitating stronger content understanding [58, 70]. Many science-based AR learning experiences are built for older learners (middle school and up) [3, 9, 23], but our work focuses on engaging elementary schoolers with compelling, educational AR activities.

2.3 Educational Experiences Powered by Chatbots and Large Language Models

A large body of evidence suggests that conversational agents or chatbots that interact with students in a natural, human-like way can make learning easier, more engaging, and more motivating, which in turn boosts learning outcomes [32, 33, 59]. Specifically, in dialogues, learners need to synthesize information to form responses, while chatbots can exploit the context to evaluate the learner's responses and provide feedback [32, 33]. Furthermore, several studies implementing educational chatbots indicate potential benefits for learners in terms of learning gains and engagement beyond the classroom contexts [26, 60, 61]. Ruan, et al. found that the combination of narratives and chatbots can effectively engage children and enhance math learning outcomes [59].

In addition, recent advancements in large language models (LLMs) have highlighted opportunities for conversational agents for education. LLMs have been used in a variety of application areas, such as language learning [30, 35] and teacher training [43]. One of the main affordances of LLMs as a tool for education is the ability the provide immediate feedback to natural language input from the learner, but it is challenging to adapt LLMs to provide effective pedagogical feedback [5]. As research on using LLMs for education is still in its infancy, we extend this burgeoning field by exploring how conversational interaction between a child learner and an LLM can promote reflection and support learning about environmental topics.

2.4 Other Relevant Theories of Learning

Our work also draws inspiration from the theories of culturally-relevant pedagogy and embodied learning. Culturally-relevant pedagogy stresses the importance of helping students affirm and appreciate

their culture of origin, while also developing their understanding of other cultures. In *Moon Story*, we base the overarching narrative on a famous Chinese myth and intersperse references to Chinese culture [38] throughout the narrative. For learners from a Chinese cultural background, *Moon Story* engages their cultural knowledge and experiences. For non-Chinese learners, it not only exposes them to an unfamiliar culture but also actively involves them in that new cultural context.

Embodied learning is a pedagogical approach that integrates bodily engagement with learning, based on theories of embodied cognition that explain how the human body and its environment affect and are affected by cognitive processes [63]. AR technologies provide rich opportunities to integrate embodiment into learning experiences by involving the body and one's physical surroundings [41]. To connect physical action to conceptual knowledge, it is critical to design the learning experience to "highlight congruencies between [children's] movements and abstract formalisms" [41]. *Moon Story* follows this design principle by guiding students to walk along a to-scale AR model of the solar system to help them internalize the different distances between planets in the solar system.

3 DESIGN PROCESS

The development of our system underwent an iterative process spanning several years, including narrative writing, educational content development, AR activity design, and learning experience design. We aimed to transform a culturally-relevant narrative into a versatile educational tool by drawing inspiration from the Chinese myth of Chang'e, the goddess of the moon, and her husband Hou Yi. We envisioned an educational experience that not only encouraged outdoor exploration but also offered enjoyable and engaging discoveries.

3.1 Educational Content

We centered our educational content on astronomy and environmental science due to the myth's link to space and drought. We aligned the content with Next Generation Science Standards [1], with particular emphasis on Earth and Space Science curricula for students between grades 3 and 5 (see Table 1). Additionally, we employed the cognitive domain of the revised Bloom's Taxonomy [4] to develop learning objectives and assessments. We targeted three levels of learning: Remembering (recognizing and recalling information), Understanding (constructing meaning from information), and Analysis (drawing connections among ideas) [36].

We developed the following learning objectives:

(1) Understand concepts in ecosystems, climate change, the Solar system, and the moon-earth system through observations and experiments.
(2) Understand how human activity affects the environment and analyze the impact of human activities.
(3) Create arguments about what actions human beings can take to protect the environment and form an argument about why the actions are important.

3.2 Narrative

We spent substantial effort developing the narrative. The narrative was inspired by the myth of Chang'e and Hou Yi, in which Hou Yi shot down nine suns to save the world and gave his wife Chang'e an immortality elixir, leading her to ascend to the sky and become the goddess of the moon [71]. We tackled narrative design challenges by balancing educational and narratological goals, ensuring theme coherence, believability, and adapting the Chang'e myth to address contemporary global issues like climate change.

To enhance accessibility for a diverse audience, we introduced the myth of Chang'e gradually through a side character, Jade Rabbit, referencing Chinese cultural elements like moon cakes and the Mid-Autumn festival as the narrative progressed. This choice was based on a low-fidelity pilot test using Google Slides with the narrative, scripts for activities, and placeholder images (see "Narrative Pilot Screenshots" in the Supplementary Material). The test was conducted before we decided on the technologies to be used in the final system, so it focused on validating different narratological choices and assessing children's connection with the narrative from diverse cultural backgrounds.

In the final narrative, children become the hero tasked with rescuing the world from an unexpected threat: a hot orb in the sky radiating incredible heat. They embark on a quest to find Hou Yi's magical bow and arrow to shoot the orb from the sky. The myth's conclusion emphasizes that merely shooting down orbs would not suffice to save the world—learners must consistently undertake environmentally friendly actions.

3.3 Activity Design

We designed seven interactive activities to be interspersed throughout the narrative (summarized in Table 2). We describe in detail the two main activities, the solar system activity and the correspondence activity. The remaining activities are described in Section 4.

3.3.1 Solar system activity. Outdoor, physical scale models of the solar system are popular installations for visualizing the large distances between planets. Augmented reality makes on-the-go, to-scale solar system models feasible, addressing the limitations of physical installations like cost and location-specific accessibility. However, existing AR systems, such as the *DIY Solar System* app [65] and *solAR* [55], often lack physical references, leading to less grounded experiences, potentially unrealistic exploration routes, and usability issues.

We sought to combine both AR and physical landmarks in our solar system model to support physical movement on a human scale [55] and foster a sense of discovery. By having learners search for physical landmarks that correspond to the to-scale locations of planets, we engage learners physically, with the prospect of enhancing memory retention [58]. After multiple iterations, we identified a set of campus sculptures to represent the inner solar system and a building to represent Jupiter, creating a 15-20 minute exploration route. Considering planet sizes, vast distances, and sculpture constraints, we set the Sun's size at 70 centimeters (equal to a sculpture), which ensured Mercury's visibility at roughly 2 millimeters.

NGSS Standard	Performance Dimensions	Expectations
4-ESS1-1 Earth's Place in the Universe	Science and Engineering Practices	Planning and Carrying Out Investigations Make observations and/or measurements to produce data to serve as the basis for evidence for an explanation of a phenomenon.
4-ESS1-1 Earth's Place in the Universe	Science and Engineering Practices	Constructing Explanations and Designing Solutions Identify the evidence that supports particular points in an explanation.
5-ESS1-1 Earth's Place in the Universe	Disciplinary Core Ideas	ESS1.A: The Universe and its Stars The sun is a star that appears larger and brighter than other stars because it is closer. Stars range greatly in their distance from Earth.
5-ESS1-1 Earth's Place in the Universe	Crosscutting Concepts	Scale, Proportion, and Quantity Natural objects exist from the very small to the immensely large.
MS-ESS1-1 Earth's Place in the Universe	Performance Expectations	Develop and use a model to describe phenomena of the Earth-sun-moon system to describe the cyclic patterns of lunar phases, eclipses of the sun and moon, and seasons.
MS-ESS2-4 Earth's Systems	Performance Expectations	Develop a model to describe the cycling of water through Earth's systems driven by energy from the sun and the force of gravity.

Table 1: References to the NGSS Standard for learning activities in *Moon Story*.

#	Activity	Description
1	Evaporation Visualization	Show the orb's hotness by bringing it close to a cup of water that evaporates.
2	Climate Change Filter	Changing the color and strength of a camera filter based on temperature change.
3	Solar System	Explore the dimensions of the solar system using relative sizes and distances.
4	Moon Phase Worksheet	Worksheet for children to take home and observe moon phases.
5	Magic Telescope	Zoom in/out to learn about the moon.
6	Shooting the Orb	Game where the learner shoots down orbs with a virtual bow and arrow.
7	Correspondence with Chang'e	Reflect on what we can do to be environmentally friendly.

Table 2: The seven activities in our system.

3.3.2 Correspondence with Chang'e Activity. In this activity, the learner corresponds with Chang'e herself through a series of letters. In these letters, Chang'e asks the learner questions that encourage them to reflect on their environmental behaviors. Chang'e responds to their answers using an LLM, providing instant, personalized feedback. We formulated the correspondence prompts to address higher-order skills and abilities in Bloom's taxonomy such as Analyze and Create, while also adding conjecture-based questions to engage learners' critical thinking. See Figure 2 for details.

4 SYSTEM DESCRIPTION

In this section, we describe the features of our system and the technical implementation of each activity.

4.1 Core Features

Moon Story is a mobile application built with the Unity game engine [69]. Its system architecture is depicted in Figure 3. Our target device was the iPhone 13 Pro, which we chose for its large screen and built-in LiDAR sensor. The depth sensing afforded by LiDAR guaranteed the system's ability to measure distances with greater precision for the Solar System activity. The narrative was integrated into Unity with the Yarn Spinner dialogue library [37], and augmented reality (AR) features were developed using the Niantic Lightship ARDK [48]. We chose the Lightship platform for its support of features beyond basic AR functionality, such as semantic segmentation and precise geolocation services.

4.2 Solar System Activity

For the hybrid virtual-physical solar system activity (described in Section 3.3.1), we developed an outdoor guided walk in a scavenger hunt style, where the learner is prompted by the narrative to walk to physical landmarks that correspond to planets in the solar system. Five of these landmarks, shown in Figure 5, were spherical sculptures within a campus quad that represented the locations of the Sun, Mercury, Venus, Earth, and Mars. These sculptures were chosen to be these solar system objects because their distances were approximately to scale. Because Jupiter was significantly farther than the other planets (and therefore outside the quad), we used a famous church on campus as the landmark for Jupiter.

Participants initiate the activity by positioning an AR Sun model at a specific landmark. Then, they follow a narrative to find planets in sequence, starting with Mercury. Participants are given each planet's scaled-down distance from the Sun in feet, but they need to walk around in the real world to identify the correct landmark.

As they walk, a cursor shows their distance from the Sun. When close to the right landmark, a button turns green, allowing them to view an AR planet model (see Figure 4, left). They need to move their phone closer to see the AR planets. Once close enough, a UI panel appears for viewing other planets and the Sun to scale (see Figure 4, middle).

The narrative then teaches about each planet, often using interactive elements like quizzes (see Figure 4, right), before guiding them to the next planet. They can also compare the size of different to-scale AR celestial objects in the Solar System (see Figure 4,

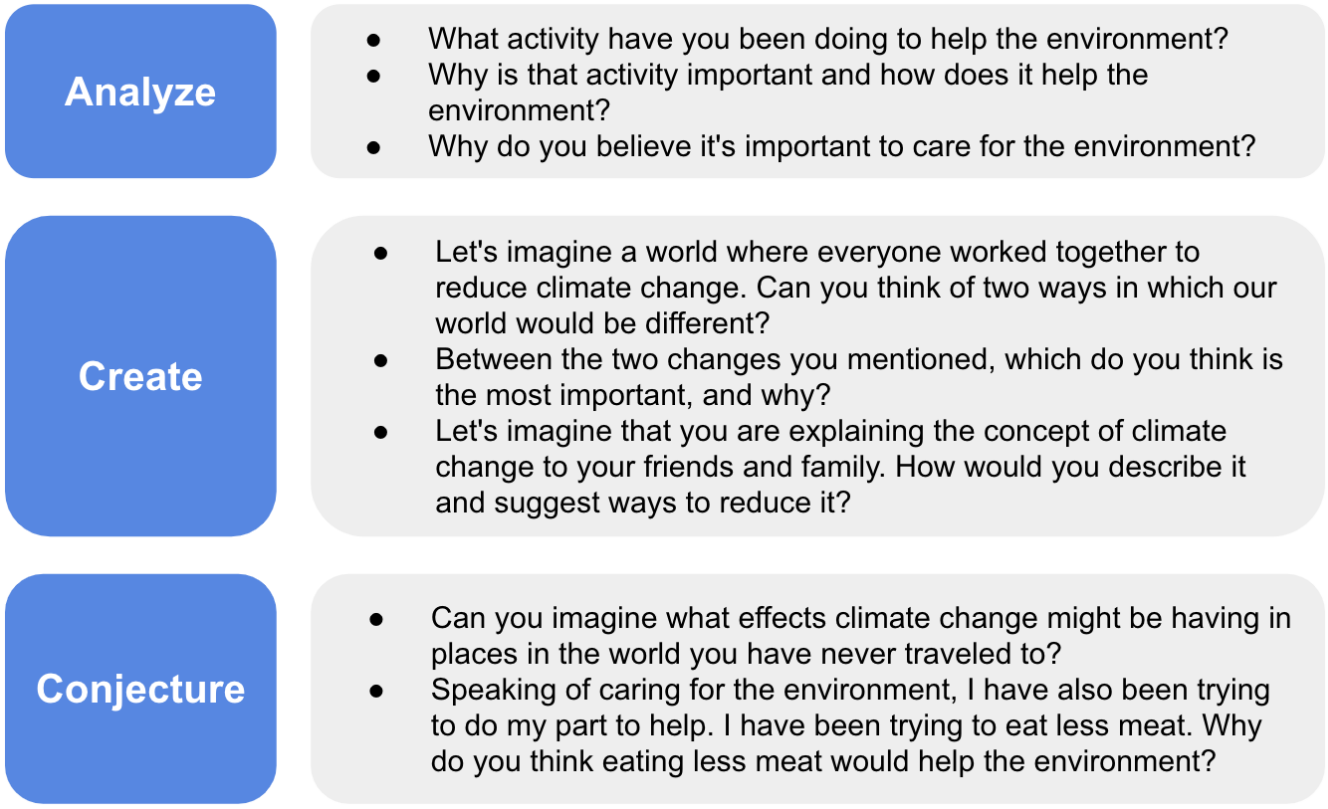


Figure 2: Sequence of prompts presented to participants in the correspondence activity.

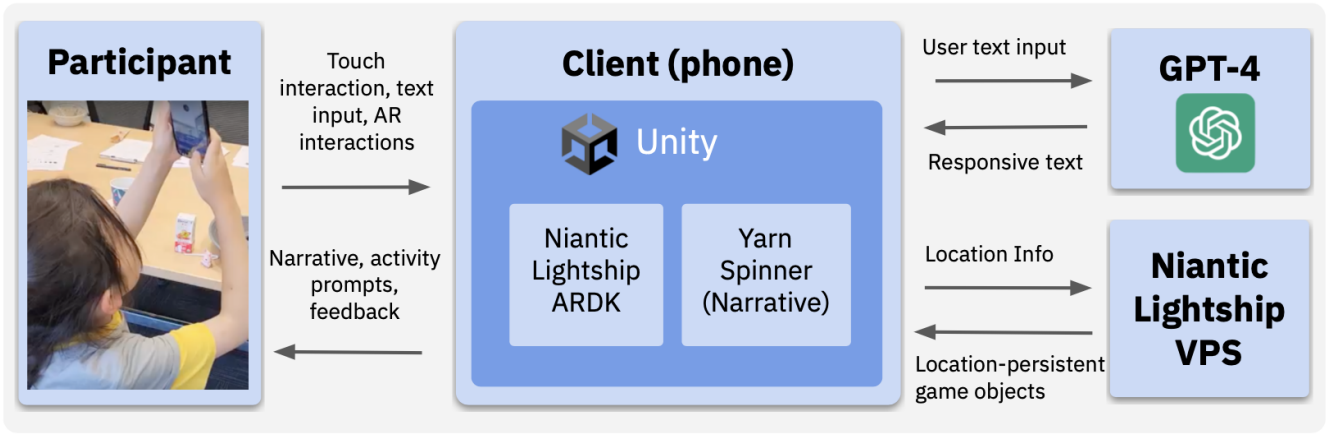


Figure 3: Overview of our system design for the LLM-based variant. The participant interacts directly with the app on the phone, which handles the narrative and local AR features. Niantic Lightship VPS is used for geolocation. Natural language processing is performed on a server that accesses GPT-4 via OpenAI's API. The other variant shares the same infrastructure but does not use GPT-4.

Figure 4: Screenshots from Solar System activity. Left to right: The learner points their camera at a landmark. The learner views the to-scale AR Mercury. The learner interacts with the comparison UI. The learner completes multiple-choice quizzes.

Figure 5: Top view of the landmarks for the Solar System chapter.

middle). This cycle repeats for Mercury, Venus, Earth, Mars, and Jupiter.

Our system persistently stores the real-world location of each landmark and associates them with each planet. We used Niantic Lightship's Visual Positioning System (VPS) [49] to track each landmark's location, which provided higher precision than GPS coordinates.

4.3 Correspondence Activity

We developed two versions of the "Correspondence with Chang'e" activity (described in Section 3.3.2): a treatment version that is powered by GPT-4 [51] and a control version that is hard coded.

For the treatment version, the implementation entailed creating a structured prompt for GPT-4 to (1) act as Chang'e; (2) pose reflection questions to learners in the form of personalized letters (see Figure 2); (3) respond encouragingly; and (4) offer a hint if the learner's response was inadequate. See the prompt used in Appendix A.1.

An example of a letter from Chang'e is shown in Figure 6 (left). After reading Chang'e's letter, learners are prompted to compose a reply. To reduce cognitive load and minimize switching between screens, GPT-4 summarizes Chang'e's questions for display in the interface (Figure 6, middle). In case of error, the researcher can trigger a button to rephrase the sent letter or reset the conversation entirely.

When sending and receiving a letter, an animation (approximately 5 seconds long) is played that shows the letter being transmitted to Chang'e, to provide a brief interval for GPT-4 to formulate a response, while reinforcing the idea of Chang'e's residence being far away on the Moon. This interaction cycle continues until Chang'e has posed all her questions.

In contrast, the control version follows the same procedure, but with pre-scripted, hard-coded letters from Chang'e. These letters are available in Appendix A.2.

4.4 Other Activities

Moon Story also features five additional activities that blend various interaction modalities with augmented reality. Details of these activities are provided below.

4.4.1 Character Creation. Learners start the app by customizing their in-game avatar (see Figure 7). They can select from a range of options for skin, hair, clothing, and eyes. Their avatar is then used throughout the narrative to represent the learner.

4.4.2 Evaporation Visualization. The Evaporation Visualization chapter simulates the effects of increasing temperatures on water using AR (see Figure 7). First, the learner points the camera at a real-life cup of water. Then, an AR glowing red orb representing the sun is added to the world where the camera is pointing. The learner can move this orb by moving the phone, and when the orb is sufficiently close to the cup of water, steam appears. This is accompanied by dialogue on the mechanics of evaporation and the broader implications of increasing global temperatures on water bodies.

4.4.3 Climate Change Filter. The Climate Change Filter chapter visualizes climate change's impact on plant life in the learner's environment (see Figure 7). Learners can simulate temperature increases of 1.5, 2.0, and 2.5 degrees Celsius using designated buttons, which causes the real grass and foliage on the screen to shift color. Niantic Lightship's semantic segmentation library is used to isolate the camera pixels with grass or foliage for the color shift, with an increasingly intense color texture onto those pixels, alongside a bloom post-processing filter that increases in intensity. The color change is accompanied by dialogue providing details on the impact that each level of temperature increase would have on plant and animal life.

4.4.4 Moon Phase Worksheet. The Moon Phase Worksheet is a paper worksheet that learners take home in between study sessions alongside an accompanying dialogue in the app after completion of the worksheet (see "Moon Phase Worksheet" in supplementary materials). The worksheet asks learners to observe the moon on two separate nights, at least three days apart, document their observations on the moon's appearance, and identify its phase for each observation.

4.4.5 Magic Telescope. In the Magic Telescope chapter, learners explore the moon's surface features using an AR-generated three-dimensional moon model (see Figure 7). They can zoom in to view details more closely, accompanied by explanatory dialogue about the moon's landmarks.

4.4.6 Shooting the Orbs. The Shooting the Orbs chapter serves as the narrative's climax, where learners shoot down nine glowing orbs (see Figure 7) using a bow and arrow, mirroring the myth of Hou Yi. In this activity, the orbs appear as moving AR objects in space. Learners aim and shoot arrows at these orbs using their phone camera, with the arrows destroying the orbs upon contact. To accommodate younger learners, the arrows are designed to automatically home in on nearby orbs, simplifying the aiming process.

5 EVALUATION

This section details the procedure of our study, our participant pool, our measures, and how the data was gathered and analyzed.

5.1 Procedure

We conducted a between-participants study of our system ($N = 50$) in which participants completed reading and learning activities using our system, filled out pre- and post-study surveys and quizzes, and provided their feedback on the app and activities. Participants were invited to come to campus for two sessions, spaced one week apart. Each session was conducted with a single participant at a time and took approximately 60-90 minutes. A follow-up survey and knowledge quiz was sent a week after session 2 to be completed at home. Participants were compensated a total of $50 for completing all tasks.

During the first session, participants engaged with the first few chapters of our system's narrative. Participants completed the following activities involving outdoor explorations on campus: Character Creation, Evaporation Visualization, Climate Change Filter, and the Solar System Activity. Between the first and second sessions, participants were given a take-home task (the Moon Phases worksheet described in Section 4.4.4) During the second session, participants returned to our lab to finish the rest of the narrative and

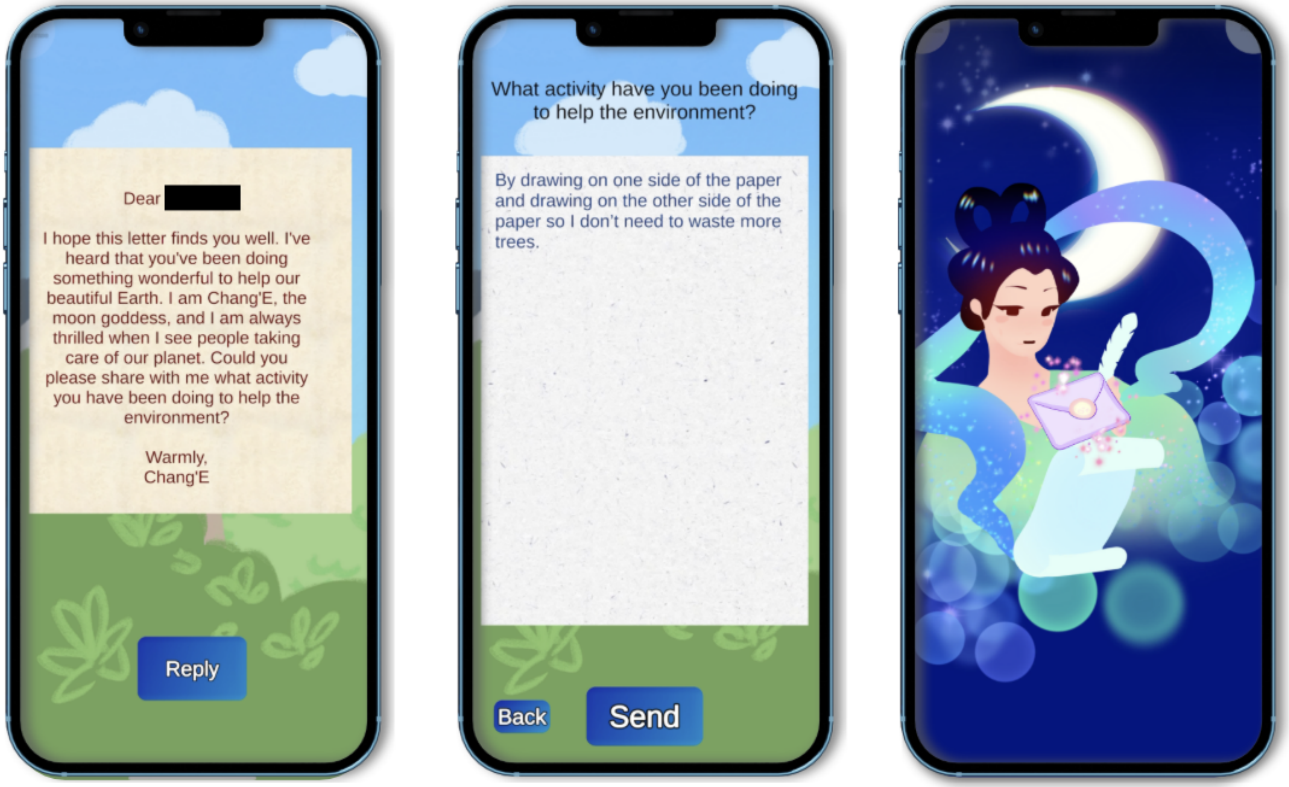


Figure 6: Screenshots from the "Correspondence with Chang'e" activity. Left to right: The learner receives a letter created by GPT-4. The learner writes a response to the letter. A waiting animation plays.

Figure 7: Other activities. Left to right: Character Creator, Climate Change Filter, Evaporation, Magic Telescope, and Shooting the Orbs.

the following activities: Magic Telescope, Shooting the Orbs, and Correspondence with Chang'e. Compared to the other activities, the Solar System and Correspondence activities were more involved and typically accounted for at least half the time spent using the system. Finally, participants were asked to fill out a follow-up survey at home a week after the second session had ended.

5.2 Condition Assignment

Participants were assigned to one of two experimental groups: the control group, which engaged with a static version of the Correspondence with Chang'e activity, and the treatment group, which interacted with a version powered by GPT-4. We utilized systematic stratified randomization to maintain a roughly even distribution across grades and ethnicities (Chinese or non-Chinese). Assignments were first randomized by strata, each stratum being defined by a grade-ethnicity pairing, and then alternated within each stratum between the control and treatment conditions.

5.3 Participants and Recruitment

A total of 50 children (19 females, 31 males) who had recently completed grades 2 through 5 participated in our study. Participants were recruited through mailing lists and word of mouth. Of the participants, 34% reported their ethnicity as East Asian, all of whom were of Chinese descent, and 12% reported two races including East Asian, all with no Chinese ancestry. 30% reported white. 10% selected South Asian. 4% reported Southeast Asian. 4% reported Hispanic or Latino. 4% have two or more races (non-East Asian), and 2% selected other. For the distribution of grades, 12% of participants just completed second grade, 32% third grade, 32% fourth grade, 24% fifth grade. 49 participants completed both sessions, while 1 participant only completed Session 1.

5.4 Measures

Our research questions posed in Section 1 mainly investigate the impact of our system on children's attitudes, motivation, learning outcomes, engagement, and behaviors from three perspectives, i.e., AR-supported interactive narrative for RQ1 and RQ3, the LLM-powered activity ("Correspondence with Chang'e") for RQ2, and the influence of Chinese cultural references for children with or without a Chinese cultural background for RQ4. We used pre- and post-study questionnaires to measure shifts in attitudes, motivation, and behaviors. Learning outcomes are evaluated through pre- and post-study knowledge quizzes. See Table 3 for what measures are collected in each section.

5.4.1 Measures of learning. We measured participants' knowledge of climate change, astronomy, and environmental science through a set of quiz questions. The quiz included questions such as *"What is climate change? Explain in your own words."* and *"What are the potential consequences of climate change on the environment? List two."* Participants were informed that they would not be judged for their responses, nor would the responses be revealed to their parents.

To assess learning gains, two researchers created a rubric for each quiz question. Points were awarded for key ideas in free-response questions and for demonstrating basic comprehension. The researchers independently scored the pre- and post-quizzes, blinded to condition, and then came to a consensus on the final scores. The rubric is included in Supplementary Materials (see "Knowledge Quiz Grading Rubric").

5.4.2 Questionnaires. Shifts in attitudes toward science, technology, and nature were assessed using pre-post attitude items with five-point Likert scales (1 = strongly disagree, 5 = strongly agree). We used six items from the Students' Motivation Toward Science Learning (SMTSL) questionnaire [68], a popular scale used to measure learning motivation. For attitude toward nature, we used ten items from the 2-MEV environmental attitudes scale [31] to measure children's attitudes toward the environment. The Six Americas Short Survey, Yay! (SASSY) [10] is also used to categorize respondents' attitudes toward climate change.

For behavioral changes, we asked participants to report their communication and conservation behaviors about climate change (e.g. *"Last week, did you talk with your parents about climate change?"* Participants could respond "*yes*", "*no*", or "*I don't know.*" [18].

To measure learning engagement, we utilized the Giggle Gauge, a 7-item instrument designed to evaluate the engagement of systems designed for children. We chose Giggle Gauge for its reduced cognitive load (our surveys were quite long) and interpretability (it provides quartile-based interpretations) [14].

To assess children's perceptions of AI in the correspondence activity, we developed a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree) with seven items. Because children's main interaction with AI is through the correspondence, these items gauge the perceived relevance and usefulness of the AI-generated content (i.e. the letters), as well as emotional and attitudinal responses. Open-ended questions were also included to collect feedback and sentiments, as well as to infer children's mental models regarding AI-generated content.

Lastly, to understand the impact of cultural references, we included a five-point Likert scale (1 = strongly disagree, 5 = strongly agree): *"The references to Chinese culture affected my engagement with the app."* and a follow-up open-ended question. Additional qualitative feedback for the app and children's self-reported performance at school are also included in the post-study surveys.

5.4.3 Relevance of responses in the correspondence activity. Participants' responses in the correspondence activity were scored by two researchers to measure learning outcomes in writing between the control and treatment groups. Because participants in the treatment condition may have been prompted multiple times by the LLM for a given question (whereas those in control were prompted exactly once per question), a third researcher compiled all participants' responses into a consistent format to ensure that the scoring researchers were blinded to condition. Responses were rated on a 0 to 2 scale: 0 for irrelevant or absent answers, 1 for partially complete answers, and 2 for fully comprehensive responses. We provide the rubric in the Supplementary Material (see "Correspondence Relevance Rubric").

6 FINDINGS

As a whole, our results showed that participants were enthusiastic about the independent, interactive outdoor exploration powered by

	In-person Session 1	Take-home Task	In-person Session 2	Follow-up Quiz
Duration	1.5 hours	7 days	1 hour	7 days after session 2
Pre-intervention measures	Demographics (filled by parents) SMTSL 2-MEV Communication behaviors Conservation behaviors SASSY Pre-knowledge quiz		Post-knowledge quiz SMTSL 2-MEV Communication behaviors Conservation behaviors SASSY	
Procedures	First three chapters of the narrative, and their accompanying activities to learn about evaporation, climate change, and solar system	Observe the moon for at least 2 nights to learn about moon phases, and complete a worksheet	Remaining four chapters of the narrative, and their accompanying activities to learn about the moon, and reflect on environmentally friendly practices	Receive the quiz via email and finish at home independently (includes knowledge quiz, communication behaviors, and conservation behaviors measures)
Post-intervention measures	Post-knowledge quiz Giggle gauge SMTSL 2-MEV Communication behaviors Conservation behaviors SASSY Feedback for app		Post-knowledge quiz Giggle gauge SMTSL 2-MEV Communication behaviors Conservation behaviors SASSY Feedback for app Performance at school	

Table 3: Procedures for each stage of our user study.

our AR-supported interactive narrative for learning (RQ1). A comparative analysis of participants in the control and treatment groups (LLM-powered version) of the correspondence activity reveals that the LLM significantly boosted engagement with reflective questions and enhanced the enjoyment of the activity (RQ2). Overall, *Moon Story* was effective in enhancing learning gains, motivation towards science learning, and learning engagement (RQ3). Furthermore, participants of both Chinese heritage and non-Chinese heritage found the culturally-based narrative to be valuable (RQ4). In this section, we present the results relevant to each research question.

In total, 50 participants did Session 1. 49 participants did Session 2, but one participant's (P59) post-Session 2 survey was missing in our dataset, so our sample sizes vary from 48 to 50 depending on the data available.

6.1 Engagement with AR-interactive narrative learning: results pertinent to RQ1

Results pertinent to RQ1 are primarily qualitative, centering on observations from the sessions and participants' feedback regarding the Solar System activity and the overall app experience.

6.1.1 Participants were enthusiastic about independent, interactive outdoor exploration. In the Solar System Activity, 24 out of 50 participants highlighted it as a favorite feature, enjoying its outdoor, active nature. Those participants described experiences with feelings of independence, exploration, and imagination. They favored the exploratory aspect and active learning over traditional indoor classroom settings. Participants, like P37 and P4, appreciated the movement and autonomy the app offered, with P8 comparing it to a blend of Physical Education and Science classes. The incorporation of real-world landmarks in learning activities, such as locating an "imaginary Jupiter" near a church, was also well-received *"Chasing the imaginary Jupiter that was near a church trash can was fun." (P6)*

Observing AR planets firsthand and comparing their sizes enhanced their learning experience, evoking a sense of surprise. Many participants exclaimed at the size of Jupiter during the Solar system activity: a number of participants (9 out of 50) mentioned the size of Jupiter and the other planets as their favorite fact. P31 stated *"Jupiter is suuuuper big"*. Nevertheless, though the activity helped them to understand how vast the solar system is, the long walk to Jupiter also made some participants feel tired.

6.1.2 Participants experienced challenges with using AR across a wide range of sizes and distances. Participants faced challenges when engaging with AR experiences that incorporated varying scales and sizes, especially in two activities: the Solar System activity and the Climate Change Filter.

The Solar System activity required participants to physically move close to small AR objects, while the Climate Change Filter was more effective when observed from a moderate distance, where grass and foliage were 5-10 feet away. However, many participants approached these two diverse scenarios similarly, often placing the phone very close to the objects. This behavior persisted despite explicit instructions.

In addition, the mix of small and large-scale AR objects requires precise physical interaction, which is challenging for many young learners. In the Solar System activity, for example, participants

struggled with distance tracking and accurately targeting small AR objects (e.g. a 2 mm-wide Mercury model). Despite a cursor designed to aid in distance measurement, many younger participants found it difficult to use accurately on their intended target. In some cases, participants chose to bypass the cursor's use, preferring to walk directly to the target. Other participants overrelied on the cursor and tried to measure distances to faraway objects without moving, which led to inaccurate measurements because of the limitations of the iPhone's LiDAR sensor.

Another observed challenge was maintaining small objects within the phone's viewport. Younger participants frequently lost track of these objects once they left the screen, and keeping them within view could be physically difficult.

6.1.3 Some participants were confused by the difference between scientific facts and fantastical elements. We observed confusion among some participants in distinguishing between scientific facts and fantastical elements within the learning experience. The blending of fictional narrative and scientific facts in *Moon Story*, especially in activities combining virtual AR elements with physical objects, sometimes led participants to struggle in distinguishing between the two. Participants occasionally misinterpreted fictional elements as real. For instance, when asked about environmentally friendly actions, some referenced fictional tasks like shooting down orbs from the narrative, instead of real-life activities.

Moreover, we found that the physical environment, particularly proximity to windows, influenced participants' differentiation between real and fictional elements in the app. In the "Magic Telescope" activity, for instance, we noticed that those near windows tended to search for the real Moon instead of the AR representation, highlighting how spatial context affects engagement with AR tasks, especially in identifying and engaging with virtual objects.

6.2 Integrating an LLM in narrative interaction: results pertinent to RQ2

We conducted quantitative and qualitative analyses to answer RQ2. Our analyses concentrated on participants' input in the correspondence activity, feedback regarding the correspondence activity, the behavior of the LLM, and children's mental models of LLM-generated content.

6.2.1 The LLM effectively encouraged relevant responses. Two researchers coded the relevance of each response to Chang'e in the correspondence activity on a scale from 0 to 2 (as described in Section 5.4.3). The inter-rater reliability of the coders using Cohen's weighted kappa [11] was 0.843, indicating almost perfect agreement [39].

We fitted a linear mixed-effects model to predict the relevance scores of each response to Chang'e, using this formula:

$$\texttt{score} \sim \texttt{condition} + (1 \mid \texttt{participant})$$

We found a significant effect ($p = 0.044$) of the condition on the relevance score, with treatment (M = 1.59, SD = 0.88) being higher than control (M = 1.28, SD = 0.71) on average. This finding demonstrates that **the LLM helped participants write more relevant responses** to the reflection questions in the correspondence activity.

To better understand when the LLM performed well and when it did not, we conducted an exploratory analysis of responses from participants in the treatment condition. We were especially interested to see whether the participants who initially responded with a less relevant answer were reprompted by the LLM to give a new answer and whether they were able to construct a relevant answer after being reprompted.

Using an inductive coding process, two researchers read over the responses written by participants in the treatment group and identified the following categories:

Missing. Participants sometimes sent empty messages due to user error, confusion about the task, or desire to test the system.

Incomplete. Participants commonly gave incomplete answers (e.g., providing one idea when asked for two) or overly general answers (e.g., not mentioning a specific activity to help the environment).

I don't know. Participants were not always able to come up with an answer or declined to answer.

Misunderstanding. Participants sometimes did not understand the question and either responded inappropriately or stated that they did not understand.

Next, the researchers coded each response independently and then came to a consensus. The inter-rater reliability using Cohen's (unweighted) kappa was 0.950, indicating almost perfect agreement.

We identified 70 initial responses in the treatment group that were not relevant. The LLM reprompted in 51 (72.9%) of these cases and was able to get 25 of those 51 (49.0%) to become relevant. Overall, **the LLM was effective at detecting irrelevant answers and somewhat effective at getting learners to write relevant answers**.

We found that the LLM was able to identify and reprompt in the cases of "Missing" (10/10, 100%), "I don't know" (16/19, 84.2%), and "Misunderstanding" (14/19, 73.7%) reliably. After being reprompted, participants were able to give a relevant answer in most "Missing" cases (8/10, 80%), and the LLM was somewhat effective at helping the participant understand in the cases of "Misunderstanding" (5/14, 35.7%). An example of this is seen in the LLM's correspondence with P62. When asked to share a climate-conscious action they took, P62 responded with, *"Saving the environment is important for all of us. The environment is the reason we are here to live."* The LLM's response clarified the question: *"I'm still curious about the specific activity you've been doing to help protect our environment. Could you please share more about it?"* Afterwards, the participant responded with an activity they did (riding a bike and relaxing in nature).

In cases where participants responded "I don't know," the LLM was not always successful in prompting them to provide a relevant answer (4/16, 25%). This may be attributed to the LLM's inability to discern user intent. Participants responding "I don't know" could be unsure of the answer, or they could be uninterested in answering. We noticed two patterns in the LLM's reprompting approach to "I don't know" responses: providing hints and repeating the child's statement before slightly rephrasing the question. For uninterested participants, repeating their statement could nudge them towards providing a relevant answer (e.g., P25, P63) unless they were determined to avoid the question (e.g., P62, who displayed impatience).

For those who were genuinely unsure, the LLM sometimes encouraged a relevant answer by providing a hint. For example, P48, when asked to imagine the effects of climate change in places she had never visited, initially responded with "Since I've never been to those places, I don't know." However, the LLM's reprompt, "Think about very cold places, like the Arctic. How might climate change affect them?", guided P48 to generate a relevant answer: "The ice there would likely melt."

However, when the LLM misapplied these strategies, it failed to push for a relevant response. For example, P60 who was uninterested in answering responded, "I don't know how" and "not really" despite the LLM providing detailed hints that were close to the correct answer after 2 rounds for multiple questions. The LLM was not able to discern that the participant did not want to answer.

Lastly, for participants who gave "Incomplete" responses, the LLM was somewhat effective at detecting these and reprompting (9/20, 45.0%), which appeared to be effective (6/9, 66.7%). Occasionally, the LLM would reprompt with a yes/no question instead of an open-ended question. In these repromptings, the LLM would explain the answer to the original question that the learner is supposed to answer, and then ask whether or not the learner understood the explanation. This allowed a small number of participants to bypass several questions.

6.2.2 The LLM enhanced participants' writing experience. The responses for the Likert items relating to participants' perceptions of the correspondence with Chang'e activity are given in Figure 8. Because hypothesis testing of individual Likert items is unreliable [6], we do not present p-values for these results.

Instead, we highlight two noteworthy observations. First, 92.0% of participants who interacted with the LLM felt that it responded to what they wrote, compared to 52.1% in the control. Second, participants who interacted with the LLM reported a higher level of willingness to write more. 32.0% of participants in the treatment condition agreed with the statement, "I wanted to write more letters with Chang'e," compared to 8.6% in the control.

6.2.3 Nudges from the LLM effectively prevented word count reduction over time. A few participants did not want to engage with the prompts and tried to write minimal or irrelevant answers. P15 (control), for instance, wrote *"I don't know"* or an equivalent to every question. Another participant (P50, treatment) tested the system by responding with messages like *"Hi"* and *"I don't know"* but finally wrote an earnest and relevant answer after five reprompts from the LLM. This suggests that LLMs may be effective at encouraging young learners who are disengaged.

In the control group especially, participants appeared to reduce the effort put into responses as the activity progressed. A few participants who initially wrote relevant responses would eventually write "I don't know" to progress. To check this hypothesis, we ran an exploratory analysis of the word counts across all of the participants' responses to Chang'e and found that a linear mixed-effects model with the following formula predicted the word count of responses:

`word_count` ~ `condition` * `question_index` + (1 | `participant`)

Within the control group, word counts appeared to decrease as the questions went on, while they appeared fairly level in the treatment group. The result shows that the LLM version likely performed better in addressing disengagement. It can potentially be attributed to the extra nudges provided by the LLM when it detected that children were not giving relevant answers.

6.2.4 The LLM took on the role of the Moon Goddess with prowess and creativity, influencing participant behavior. We noticed a few patterns in LLM-generated letters in the correspondence activity. First, the LLM tends to repeat and echo participants' inputs, as seen in phrases like, "*Saving water is indeed a great way to help the Earth*" or "*Watering plants and trees is indeed a great way to contribute to the environment.*" Second, we found that the LLM made deliberate stylistic choices while casting Chang'e. The LLM used creative and relevant valedictions like "*With lunar love*" and "*With stardust wishes*" to conclude letters. Some participants adopted these valedictions: in response to *"With love and moonlight,"* P57 closed her letters with *"With lots of love and daylight."* Despite following prompted scripts often, the LLM also sometimes varied output based on contextual interpretations for system prompts. For example, we instructed the LLM to "pretend you have also been helping the environment by eating less meat [and tell the learner] about your experience." At times, it adhered closely to the script (e.g., *"Just like you, I've also been trying to do my part to help the environment"*). In other instances, it adapted its output to its role (e.g., "*As a moon goddess, I don't need to eat, but I've been encouraging others to eat less meat as a way to help the environment.*")

6.2.5 Customizing content with an LLM boosted positive feedback in the correspondence activity. Students responded more positively to the letter-writing activity when Chang'e responded using an LLM compared to when she had hard-coded responses. For example, in the treatment condition, one participant said, "*the letter writing activity was fun because I got to communicate with chang'e*" (P26). Another participant expressed, "*That was my favorite part because it made it feel real and it was just fun*" (P22). Despite these positive comments, one participant thought it was boring (P26), and another participant said that it was fun but they got tired of typing (P20).

Participants in the Control condition had more negative reactions. P42 said, "*It was pretty good except it will be better if the responses will be different according to what I wrote. It did respond a little bit.*" Participants in this condition thought the interaction was "*boring* " (P27, P25), long (P30), and that the typing was exhausting or difficult (P29, P31, P28).

6.2.6 The LLM struggled to reduce its verbosity to cater to the needs of our participants. The LLM often provided verbose responses despite clear instructions to tailor the language for children. Even when asked to rephrase by participants who were struggling, the LLM would maintain or even increase the complexity and length of its sentences. For example, when a participant stated they did not understand the question on how they would explain climate change to their families, Chang'e responded with an extended metaphor: *"Imagine you're telling a story about how the Earth is getting warmer because of things people do, like using too much electricity or driving cars that make smoke. And to help stop the Earth from getting too warm, we can do things like turning off lights when we're not using them or taking shorter showers. How would you tell this story to your friends and family?"* Another participant stated bluntly in a letter

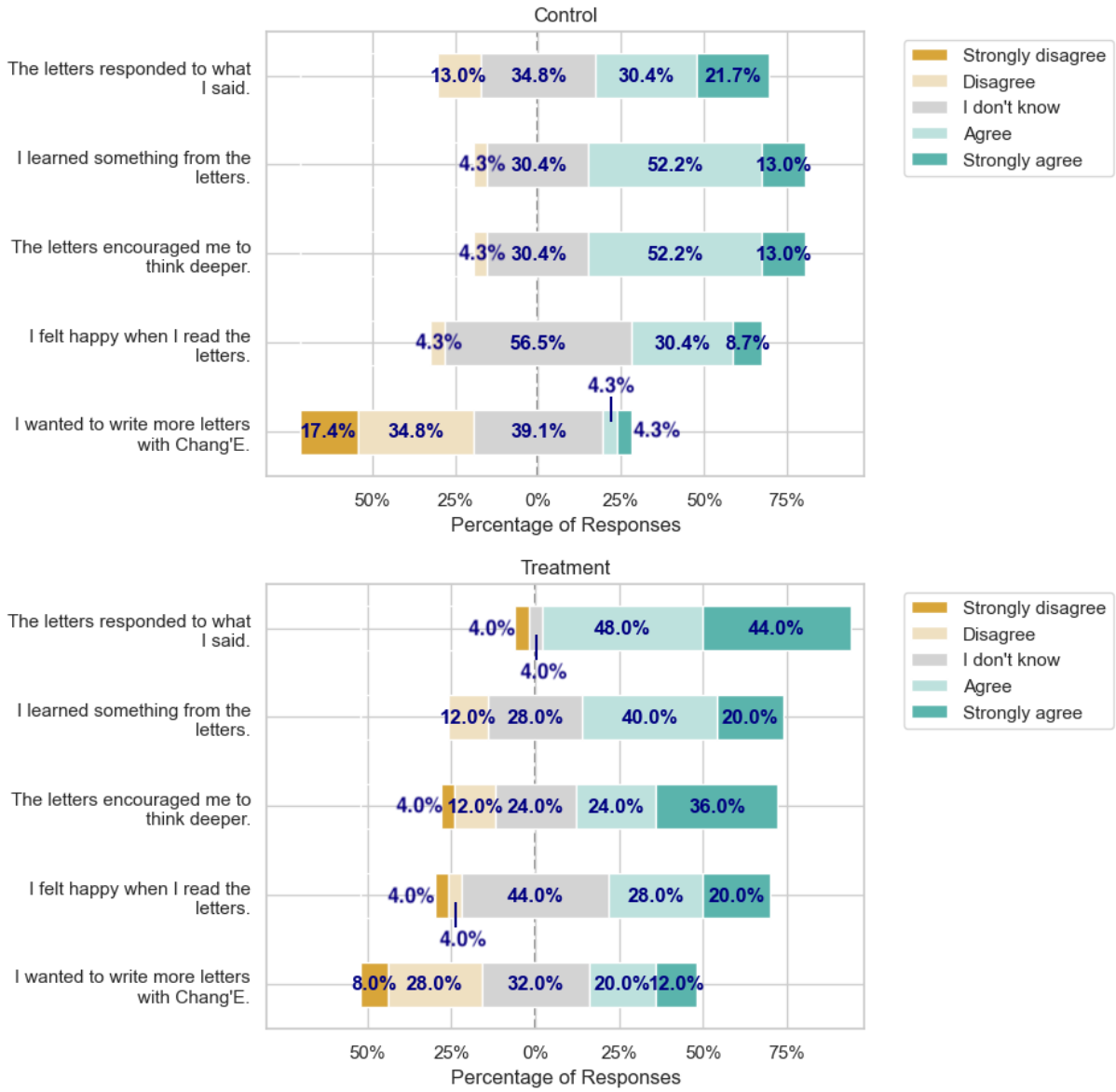


Figure 8: Comparison of responses to the perception questions for the correspondence activity (top: control, $n = 23$; bottom: treatment, $n = 25$).

"You always write complex things. Stop it," (P62) to which Chang'e responded by restating the question without any major alterations.

6.2.7 The LLM exhibited people-pleasing behaviors. While the LLM in the treatment condition generally encouraged better participant responses, it was not particularly stringent in ensuring meaningful engagement with the reflection questions. Despite being able to identify inappropriate or incomplete answers and reprompt, the system tended to accept refusals too readily. In one instance Chang'e asked a participant, *"I'm really curious to know about the activity you've been doing to help our beautiful Earth. Could you please share it with me?"* to which they responded *"no, I can't"* (P48). This was met with a response of *"That's perfectly alright. Sometimes, it can be difficult to put our actions into words,"* before proceeding to the next question.

In addition, by accepting and building on even incoherent responses, the LLM helped maintain engagement across all skill levels, focusing on facilitating reflection rather than evaluating or critiquing responses. For example, when a participant described *"the hot orb is getting the E[a]rth too hot so I am trying to get rid of it."* (P57), a reference to the orb shooting activity in our system, Chang'e initially interpreted it as a metaphor for a real-life activity, eventually coming to a final interpretation of *"You're concerned about the Earth getting too hot, which is a reference to global warming*

and climate change." This approach could be deemed beneficial in our context, particularly for younger learners who might struggle to articulate complex concepts.

6.2.8 Participants' mental models were based on reply speed, emotional displays, perceived personalization, and temporality. Mental models are cognitive representations that capture the structural relationships between objects and events [50]. In the post-study survey, we included a question "*Do you think the responses are written by a human or a computer? Why do you think so?*" to gain insights into children's mental model of LLM-generated content. We identified several prominent themes in the children's responses.

First, children tend to associate speed with computers or AI. Many children in both the treatment and control groups reported that the responses to their letters were sent very quickly, and therefore could not have been written by a human. In addition, many children attribute consistency to computers or AI, especially in the control condition where children only received hard-coded responses. Children stated "*it is probably coded to write the exact same letter for every response*" (P44 control). One child mentioned AI's ability to instantly respond to multiple users, leading the child to believe the letters were written by AI (P8, treatment). We can infer that children formed a functional understanding of the AI or computers' ability to generate letters instantly for many users.

Second, feelings and emotions are consistently associated with humans in both the control and treatment groups. For instance, children's responses included statements such as "*I feel the responses were written by humans but programmed to write that for the app since it had a lot of feeling into it and I didn't think AI wrote it*" (P38, control) and "*Human because it had emotion and [I] felt like it knew what I said*" (P20, treatment).

Third, we have encountered contrary interpretations of Chang'e's responsiveness to user input. On one hand, children in the control group believed that the letters were written by computers because they did not respond to their input accurately; A participant wrote, "*It always avoided what I said and that proves that it didn't know what I had [written]*" (P35, control). This is consistent with some children's perception in the treatment group, who believed responsive letters were human-written. Contrary to this perspective, many children linked responsiveness to computer-generated content. P12 (treatment), wrote: "*I think it's written by a computer because this was answering to my answers very specifically and a human couldn't have thought of every single answer for every variation.*"

Lastly, some children considered the implementation and feasibility of human involvement. One child conceptualized a flowchart for the response generation: "*I think that humans created this code that if yes then say this and if no then say this*" (P13, control). Several children thought that it would be impractical for a person to respond to their letters at all times. These children possibly made an analogy with online chat systems when discussing how the system works.

6.3 Effect of the immersive learning system: results pertinent to RQ3

Here, we describe the results of our four primary pre- and post-study measures: (1) learning gains, (2) motivation toward science learning (SMTSL), (3) attitudes toward the environment (2-MEV), and (4) engagement (Giggle Gauge).[1]

In our analyses, we calculate Cohen's d [11] to measure effect size, using the traditional benchmarks for interpreting magnitude ($d = 0.2$ is small, $d = 0.5$ is medium, $d = 0.8$ is large) [39].

6.3.1 Participants demonstrated learning gains. Because different content was taught in each of the three stages of our study, we used different test questions to measure learning gains for each one. Table 4 illustrates the content covered in the pre-tests and post-tests for each session and how we compared the pre- and post-study quizzes to assess learning gains effectively.

Our two researchers who graded the knowledge quizzes achieved high inter-rater reliability, with a Cohen's weighted kappa value of 0.836, indicating almost perfect agreement.

To measure immediate post-session learning gains, we summed the grades for all the pre-tests and post-tests for each participant (range: [0, 75]). Because the data were not normal, the Wilcoxon signed-rank test was used to compare the learning gains. We observed a significant increase from pre ($M = 40.6$, $SD = 15.0$) to post ($M = 45.3$, $SD = 13.5$), $Z = 168.0$, $p < 0.001$, indicating that **participants experienced learning gains from using *Moon Story***. We calculated the effect size using Cohen's d ($d = 0.30$), indicating a small to medium effect size.

To measure knowledge retention, we compared the summed grades for the session 1 pre-test and the follow-up test for each participant (range: [0, 75], $n = 48$). A Wilcoxon signed-rank test showed a significant increase from session 1 pre-test ($M = 39.4$, $SD = 15.2$) to follow-up ($M = 47.5$, $SD = 12.2$), $Z = 343.0$, $p < 0.02$, indicating that **participants retained learning gains from using *Moon Story* a week later**. We calculated the effect size using Cohen's d ($d = 0.57$), indicating a medium effect size.

Given that the improvements in the follow-up test are higher than the immediate post-session learning gains, we identified several factors that might lead to this result in the Discussion section.

6.3.2 Participants' motivation toward science learning increased. We summed each participant's answers to the motivation toward science learning Likert scale questions (range: $[0, 30]$) and used a paired t-test to compare the participants' scores from their Session 1 pre-study survey and their Session 2 post-study survey.

We found a significant increase from pre ($M = 21.7, SD = 2.96$) to post ($M = 22.9, SD = 3.8$), $t(47) = -2.43$, $p = 0.019$, indicating that **our system increased participants' motivation to learn science**. The effect size, as measured by Cohen's d, was $d = 0.35$, indicating a small to medium effect size.

6.3.3 Participants reported an increase in conservation behaviors, though their environmental attitudes remained largely unchanged. To see whether *Moon Story* may have had an impact on environmental behaviors, we compared participants' reported communication and conservation behaviors about climate change in the Session 1 pre-study survey and in the follow-up survey. We encoded their responses as numeric values ("Yes" = 1 and "No" = 0), discarded "I

[1]Although a total of 49 participants completed the pre-study and post-study surveys in both sessions, there were several data points missing in our dataset: one Session 1 pre-study survey (P22), one Session 1 post-study survey (P34), and one Session 2 post-study survey (P56). As a result, each of our analyses excludes participants for whom we do not have sufficient data.

Stage	Pre-Test Content	Post-Test Content
Session 1	All quiz questions	Questions related to session 1 activities
Take Home Assignment	Session 1 pre-test (moon phase only)	Session 2 pre-test (moon phase only)
Session 2	Questions related to session 2 activities	Questions related to session 2 activities

Table 4: Measuring and comparing learning gains. Because the content taught in *Moon Story* is spread out over two sessions, we tested for different knowledge at each point in the study.

don't know" responses, and fit a linear mixed-effects model for communication behaviors (two questions) and conservation behaviors (five questions), where `time` is a binary factor indicating whether the observation was from the pre-study survey or the follow-up survey:

$$\texttt{score} \sim \texttt{time} + (1 \mid \texttt{participant})$$

Although there was no significant effect of `time` on `score` for communication behaviors, we did observe a significant shift from pre- ($M = 0.77$) to follow-up ($M = 0.85$) for conservation behaviors, $p = 0.035$. While we cannot claim causality due to the lack of a control, this correlation suggests that **participants possibly engaged in more environmental conservation behaviors after using our system**.

To analyze attitudinal changes toward the environment, We summed each participant's answers to the attitude toward nature Likert scale questions (range: $[0, 50]$) and used a paired t-test to compare the participants' scores from their Session 1 pre-study survey and their Session 2 post-study survey.

While we observed an increase from pre ($M = 36.3, SD = 5.24$) to post ($M = 38.1, SD = 5.16$), it was not significant, $t(47) = -1.81$, $p = 0.077$. The effect size, as measured by Cohen's d, was $d = 0.26$, indicating a small effect size.

In addition, analysis of SASSY data did not indicate a significant change in children's attitudes towards climate change before and after the sessions.

6.3.4 Participants reported high levels of engagement. We computed the means for each Giggle Gauge engagement score ($N = 47$) from both Session 1 and Session 2 (see Table 5), and **we found high levels of self-reported engagement**. Using the quartile interpretations given by Dietz, et al. [14], scores above 3.6 are classified as high engagement, and those between 3.0 and 3.6 are considered moderate. We report high engagement levels for interest, perceived user control, and endurability, we found moderate engagement levels for the remaining items.

6.4 Enhancements across cultural boundaries: results pertinent to RQ4

We found that the Chinese-inspired elements of the narrative helped to enhance and frame the overall learning experience for participants of either Chinese or non-Chinese heritage.

Many participants of Chinese descent reported a sense of connection or belonging with the narrative, regardless of whether or not they had prior exposure to the myth of Chang'e. P36 stated that *"it felt nice having my culture explained in an app"*, and P33 expressed, *"I feel it made the app experience 10x better"*.

Despite having no cultural connection, participants of non-Chinese descent communicated liking learning more about Chinese culture and mythology (P12, P32). Both groups of participants expressed how the added Chinese-inspired elements helped to enhance and frame the overall learning experience. P8 noted that *"it was a nice touch to what would probably have been quite bland without it"*, and P38 expressed how *"it was kind of cool to add the moon cakes and the Chinese Chang E [sic] and Hou Yi so not only [do] you learn about the solar system, you also learn a little bit about Chinese history."*

7 DISCUSSION

The process of designing, developing, and testing *Moon Story* uncovered several opportunities and challenges when combining AR with narrative (RQ1), LLMs with narrative (RQ2), building immersive learning systems (RQ3), and the implications for adapting cultural myths for learners (RQ4). We synthesize our findings for building AR- and LLM-driven narrative-based learning environments below.

7.1 Using immersive technology for teaching science through narrative and exploration

Our quantitative findings indicate that the integration of narrative and AR enhances learning gains and knowledge retention, as well as motivation toward science learning. This is consistent with a substantial body of research on the efficacy of narrative and AR in boosting educational outcomes [7, 8, 12, 20, 53, 59, 64]. Notably, the effective learning observed can be attributed to the incorporation of learning objectives into the activity design [7]. However, the effect size was small to medium, likely because a significant proportion of participants had moderate to high pre-existing knowledge of the subject matter.

We also observed that participants performed better overall on the follow-up knowledge quiz compared to the post-session knowledge quizzes. This could be due to selection bias (as not all participants submitted the follow-up) or participants seeking outside help. Additionally, the post-session scores might have been impacted by fatigue, particularly evident after Session 1, which included a 20-minute outdoor walk. This fatigue could have temporarily hindered their immediate post-session performance.

Our qualitative findings highlight the efficacy of immersive technology in educational contexts. AR components, particularly in outdoor environments, notably differentiated our system by rendering abstract scientific concepts more accessible and tangible. The active, exploratory nature of the AR learning tasks offers a sense of

Component	Prompt	Score 1	Score 2
Aesthetic and sensory appeal	"I like how the app looked and felt."	3.54	3.58
Challenge	"This app was hard in a good way."	3.42	3.31
Endurability (affect)	"I would like to do this again sometime."	3.60	3.60
Feedback	"The app let me know when I did something."	3.54	3.56
Interest	"I enjoyed using this app"	3.64	3.69
Perceived user control	"I had control over what I was doing."	3.62	3.63
Novelty	"I found lots of things to do in the app."	3.34	3.35

Table 5: Mean Giggle Gauge engagement scores for Session 1 (Score 1, $n = 50$) and Session 2 (Score 2, $n = 48$).

autonomy and a refreshing deviation from traditional classroom learning. The integration of real-world landmarks to contextualize scientific concepts, like associating an "imaginary Jupiter" with familiar locations, was positively received. The direct observation and size comparison of AR planets deepened children's understanding and engagement with intangible scientific facts by vividly illustrating the vastness and scale of the solar system. Nonetheless, feedback from our Solar System activity indicated that the lengthy walk to Jupiter was sometimes viewed negatively, emphasizing the importance of balancing physical activity with educational objectives.

In addition, our study promotes student-driven learning, emphasizing active exploration and autonomy for increased motivation. This shift empowers students to take ownership of their learning, leading to increased motivation and a sense of accomplishment. Furthermore, the integration of real-world landmarks and familiar locations contextualizes scientific concepts, making them more relevant and meaningful to students.

Implications for future work. Our study sets the foundation for future work expanding the breadth and depth of our findings by conducting similar studies in more contexts and over longer periods. An open research question is how to operationalize the integration of LLMs (or other generative AI technologies), AR, and narrative to produce educational experiences based on a wide range of learning objectives set in various contexts. Doing so will also allow researchers to scale our findings and apply them in a cost-effective manner, allowing for wider access and impact across diverse learning environments. More research is also needed to understand the long-term impact of this kind of educational experience.

7.2 Designing for varying sizes and distances in AR

Moon Story incorporated a variety of AR experiences that required working with AR objects of different sizes and interacting with the user's environment at different distances. Some activities needed the user to bring the phone up close (e.g., examining a tiny, to-scale AR model of an inner planet), while other activities worked better from a slight distance (e.g., applying a filter to grass and foliage). As described in Section 6.1.2, many participants demonstrated confusion when asked to switch between these different modes of AR interaction, even when the system gave explicit instructions about where to stand.

We suspect that participants may not have had much experience with using mobile AR. Mobile AR introduces a different modality of interacting with both the mobile device and the physical world, with the device serving as a "window-on-the-world" that overlays virtual content over a camera screen of the real world [42]. While our participants might having been familiar with using a phone camera, they seemed less accustomed to the interaction patterns that are associated with AR, such as tracking virtual objects in and out the phone's "window" or knowing when to move close and move far. Moreover, prior research on the challenges of leveraging AR for learning has established that AR can add cognitive load to educational activities (especially if the task is complex) [2, 16]. Our participants may have experienced increased cognitive load, contributing to their confusion.

Implications for future work. More work is needed to understand how children acquire the skill of *using* AR, and how they negotiate the interactions between the physical and digital worlds in their learning process. For example, we do not know if the issue of user confusion would naturally get better as children gain more exposure to AR technologies, or if more scaffolding would be beneficial for AR interactions.

To this end, future work should develop design guidelines for "choreographing" [7] narrative-based AR experiences (or similar experiences that involve a fairly linear progression) in a way that uses the mobile digital interface to guide the user through the physical world. For example, in the Climate Change activity, we could have asked the learner to first take a picture of the grassy lawn before applying the filter. Doing so would leverage the learner's existing knowledge of how to operate a phone camera and nudge them to keep their distance (to take a good picture), minimizing the chances of the learner pointing the camera too close to the grass.

7.3 Blending the real and the fantastical

Moon Story combined fiction and reality in two different ways: it taught scientific facts through a fictional narrative, and it incorporated mixed-reality educational activities that used both virtual AR objects and physical objects. We found that the fantastical elements appeared to better engage the user in the real world, and the real-world activities imbued the myth and the narrative with a sense of realism.

However, navigating this space where reality and fiction intersect was complicated. Participants sometimes thought the app was referring to something fictional rather than something real, or vice versa. The issue of confusing the real and the fictional may be exacerbated by contextual factors, such as physical location.

This may pose issues for the transfer of knowledge gained during the mixed-reality experience, as learners might have trouble compartmentalizing real facts from fictional elements. Prior work in game-based learning has found mixed results for the transfer of knowledge and skills learned while playing educational games to other, more general contexts [67]. However, we also see potential for mixed reality and narrative to engage learners in more hands-on learning in real-world contexts, which could improve transfer based on principles from situated learning [40].

During our initial pilot testing of the narrative, where we used Google Slides, we did not encounter any confusion about what was real and what was not. This is likely because the narrative was fully compartmentalized, similar to a book or film. Because we developed and iterated on the narrative and the AR activities in parallel, we did not uncover many of these issues until the narrative and activities were combined in the final version of the app.

Implications for future work. An exciting area for future work is to develop new methodologies for rapid-prototyping experiences that blend the real and the fantastical in AR. This would require a framework for integrating physical and virtual elements with scientific and fantastical ones. Moreover, this accentuates the need to further understand how children conceptualize mixed reality, and what they perceive as real or fantasy. There are also exciting opportunities to further explore the potential transfer effects of combining fictional narratives with real-world learning experiences.

7.4 Enhancing learner engagement and maintaining progress with the LLM

Overall, based on our findings, we suggest several benefits of using an LLM to facilitate reflection with children. Participants who interacted with the LLM put more effort into answering the reflection questions, as evidenced by the increased relevance of their responses compared to the control group and their total word count of their responses. We attribute this finding to the LLM's capability to evaluate the participants' responses and ask them to try again or expand upon their answers, especially when the answers were missing or when the participant didn't know how to answer. We also believe that the increased perception that Chang'e was understanding them (in the treatment condition) contributed to those participants putting in more effort. Participants in the treatment condition reported higher levels of engagement—they were more willing to write more letters to Chang'e and more likely to find the activity fun.

Our results contribute to the literature by broadening the scope of LLM applications in education and verifying its effectiveness, extending it to reflective thinking, and environmental and science education. With appropriate prompt engineering, the immediate feedback provided through an LLM can keep learners engaged and on track with the learning tasks. In addition, our study expands the understanding of how LLMs can be integrated into immersive education systems with AR and narratives. This work adds a new dimension to the potential applications of conversational AI in educational contexts.

Implications for future work. The effectiveness we observed in animating Chang'e using an LLM opens up a new area of research entailing animated educational characters backed by LLMs more broadly. Culturally-relevant characters could be brought to life across many cultures, potentially increasing the sense of belonging of learners across the world. This raises new questions about how to train LLMs to behave in culturally appropriate ways. Moreover, many animated characters could populate augmented learning environments and their stories could develop over time while drawing in learners to experience those interactions first-hand. This could mimic the way novels and television series unfold, but also integrate the learner into the story. This could be achieved using a method similar to the one Park et al. [52] employed to create generative agents to simulate human behavior in a game-like environment. Future research could look into creating these sorts of digital worlds and how to automatically personalize these interactions to learners at various skill levels.

7.5 Fostering dialogic learning environments with LLMs

Our results align with previous research indicating that LLMs often exhibit people-pleasing tendencies [54]. While LLMs can guide children with hints and relevant responses, they do not match the depth of engagement and reinforcement provided by human teachers' pedagogical choices, such as using teacher talk strategies. These strategies are crucial for skill development and content understanding and include techniques like language repetition, recasting, cued elicitation, adaptive questioning [62], as well as follow-up moves [13]. Effective teaching involves using these strategies to foster a dialogic learning environment in which students interact cooperatively with the teacher as they construct new understandings that transform their conceptual understanding [62].

Implications for future work. While the findings reported in this work are extremely promising, more guardrails are also needed to prevent LLMs from exhibiting too much people-pleasing behavior when discomfort natural to the learning process is necessary. Similarly, we also need to prevent LLMs from unempathetically insisting that a student who is frustrated continues doing a task without providing the necessary intellectual or emotional support. It will be crucial to systematically evaluate LLM-based systems when deploying them in learning environments to ensure that they behave in helpful ways towards learners. In our study, we did not specifically tune GPT-4 for pedagogical strategies, revealing its inherent limitations in these aspects, which might potentially explain why the LLM failed to capture teaching opportunities through the conversation with some participants. Future research should explore the integration of these pedagogical strategies within LLMs to enhance educational engagement and improve learning outcomes.

7.6 Creating culturally-relevant learning experiences using AR and LLMs

Our system, *Moon Story*, demonstrated the potential of blending a narrative driven by cultural mythology with AR and LLM technologies to create an engaging educational experience for children. Children from varied ethnic and cultural backgrounds expressed positive attitudes about the narrative and experienced high levels of engagement.

The concept of portraying the learner as the hero in an adventurous and culturally rich interactive narrative proved to be a highly effective design strategy. Additionally, myths serve as a rich source for crafting interactive narratives that benefit students with or without relevant cultural backgrounds.

Implications for future work. Future research can explore how LLMs can be leveraged to adapt the narrative and learning content dynamically based on the individual's cultural interests. Additionally, further efforts could be made to disseminate culturally-driven narratives across diverse subjects, fostering a more inclusive learning environment that bridges different cultures for individuals to benefit from the rich tapestry of human stories. This approach also presents opportunities to employ culturally-driven narratives, augmented reality (AR), and LLMs to facilitate collaborative learning experiences across varied cultures and communities.

8 LIMITATIONS

Our system design and study have several key limitations that should be addressed by future research. First, while we designed *Moon Story* to be a self-driven learning experience, our study required research supervision to assist participants when they encountered confusion. Furthermore, the safety of LLMs is still an active area of research—when used without supervision, additional safety measures may be necessary. Our use of GPT-4 was designed to maintain appropriate responses within the context of the learning task, but it may not adequately address potential adversarial inputs or users.

Second, we spent substantial effort in prompt engineering for GPT-4. Despite the abundance of literature on effectively engineering prompts for LLMs, achieving the ideal conversational context was still challenging, as minor prompt adjustments aimed at improving one aspect often had unintended, drastic negative effects on others. Future work should explore a number of different strategies to attenuate these issues.

Third, we used the Giggle Gauge to measure children's engagement with our system, but it has only been validated for younger children (ages 4–7) [14]. The increased metacognitive abilities of children in our study's age range (grades 2 through 5, roughly ages 7–10) may affect the validity of the scale.

Lastly, our AR activities, particularly the solar system activity, are limited to a specific location on our campus. While the technology we used (Lightship VPS) theoretically allows our system to work anywhere there is a Wayspot supported by the platform, there still lies the challenge of identifying suitable landmarks for placing anchors that represent each planet's location, which requires non-trivial human involvement. Future research could investigate the adaptation of the system to diverse locations, offering flexibility to accommodate various types of landmarks, thus potentially enabling its use in a broader range of settings.

9 CONCLUSION

In summary, we developed a narrative-based learning system integrating AR and LLM features, providing customized experiences for culturally diverse children. We assessed different versions of the app with students who just completed 2nd to 5th grades, examining intervention influences on learning outcomes, motivation, and engagement. Both conditions demonstrated statistically significant learning gains, with the LLM variant effectively addressing disengagement. Qualitative analysis revealed that the culturally relevant narrative fostered a strong sense of belonging among Chinese participants, and non-Chinese participants also engaged positively. We also explored children's mental models of LLM-generated content. Overall, we tackled considerable design challenges, merging ancient myths with new technology, delivering scientific content in a fantastical manner, and blending digital and physical interactions. This paper provides design implications for designing effective learning systems with culturally rich narratives, AR, and LLM features that are closely aligned with learning objectives.

ACKNOWLEDGMENTS

We gratefully thank TAL Education Group and the Stanford Institute for Human-Centered Artificial Intelligence (HAI) for providing funding for this research. We would also like to thank and acknowledge our collaborators who contributed to the design (Michelle Park) and implementation (Alexander Worley, Yannie Tan, Gwendolyn Liu, Zoe Lynch, Cathy Zhang, Abdallah AbuHashem) of early prototypes. Special thanks to Cyan DeVeaux, Elizabeth Childs, and Jacob Ritchie for their continuous feedback and support throughout the project, as well as all the amazing children and parents who participated in our user studies.

REFERENCES

[1] 2013. *Next Generation Science Standards: For States, By States.* The National Academies Press, Washington, DC. https://doi.org/10.17226/18290

[2] Murat Akçayır and Gökçe Akçayır. 2017. Advantages and challenges associated with augmented reality for education: A systematic review of the literature. *Educational research review* 20 (2017), 1–11.

[3] Ismo Alakärppä, Elisa Jaakkola, Jani Väyrynen, and Jonna Häkkilä. 2017. Using nature elements in mobile AR for education with children. In *Proceedings of the 19th International Conference on human-computer interaction with mobile devices and Services.* 1–13.

[4] 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.

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[6] 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.

[7] Alan Y Cheng, Jacob Ritchie, Niki Agrawal, Elizabeth Childs, Cyan DeVeaux, Yubin Jee, Trevor Leon, Bethanie Maples, Andrea Cuadra, and James A Landay. 2023. Designing Immersive, Narrative-Based Interfaces to Guide Outdoor Learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–22.

[8] Yu-Cheng Chien, Yen-Ning Su, Ting-Ting Wu, and Yueh-Min Huang. 2019. Enhancing students' botanical learning by using augmented reality. *Universal Access in the Information Society* 18, 2 (2019), 231–241.

[9] Luca Chittaro and Fabio Buttussi. 2015. Assessing knowledge retention of an immersive serious game vs. a traditional education method in aviation safety. *IEEE transactions on visualization and computer graphics* 21, 4 (2015), 529–538.

[10] Breanne Chryst, Jennifer Marlon, Sander Van Der Linden, Anthony Leiserowitz, Edward Maibach, and Connie Roser-Renouf. 2018. Global warming's "six Americas short survey": Audience segmentation of climate change views using a four question instrument. *Environmental Communication* 12, 8 (2018), 1109–1122.

[11] Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70, 4 (1968), 213.

[12] Diana I. Cordova and Mark R. Lepper. 1996. Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology* 88 (1996), 715–730. https://doi.org/10.1037/0022-0663.88.4.715

[13] Richard Cullen. 2002. Supportive teacher talk: The importance of the F-move. *ELT journal* 56, 2 (2002), 117–127.

[14] 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.

[15] Julie Ducasse. 2020. Augmented reality for outdoor environmental education. In *Augmented Reality in Education: A New Technology for Teaching and Learning*. Springer, Switzerland, 329–352.

[16] Matt Dunleavy and Chris Dede. 2014. Augmented reality teaching and learning. *Handbook of research on educational communications and technology* (2014), 735–745.

[17] Matt Dunleavy, Chris Dede, and Rebecca Mitchell. 2009. Affordances and limitations of immersive participatory augmented reality simulations for teaching and learning. *Journal of science Education and Technology* 18 (2009), 7–22.

[18] June A Flora, Melissa Saphir, Matt Lappé, Connie Roser-Renouf, Edward W Maibach, and Anthony A Leiserowitz. 2014. Evaluation of a national high school entertainment education program: The Alliance for Climate Education. *Climatic Change* 127 (2014), 419–434.

[19] James Paul Gee. 2017. *Teaching, learning, literacy in our high-risk high-tech world: A framework for becoming human*. Teachers College Press.

[20] Yiannis Georgiou and Eleni A Kyza. 2018. Relations between student motivation, immersion and learning outcomes in location-based augmented reality settings. *Computers in Human Behavior* 89 (2018), 173–181.

[21] Yiannis Georgiou and Eleni A Kyza. 2021. Bridging narrative and locality in mobile-based augmented reality educational activities: Effects of semantic coupling on students' immersion and learning gains. *International Journal of Human-Computer Studies* 145 (2021), 102546.

[22] Richard J Gerrig. 1993. *Experiencing narrative worlds: On the psychological activities of reading*. Yale University Press.

[23] Manuela Glaser, Bärbel Garsoffky, and Stephan Schwan. 2009. Narrative-based learning: Possible benefits and problems. (2009).

[24] Adele E Gottfried. 1985. Academic intrinsic motivation in elementary and junior high school students. *Journal of educational psychology* 77, 6 (1985), 631.

[25] Wendy S Grolnick, Richard M Ryan, and Edward L Deci. 1991. Inner resources for school achievement: Motivational mediators of children's perceptions of their parents. *Journal of educational psychology* 83, 4 (1991), 508.

[26] Joshua Grossman, Zhiyuan Lin, Hao Sheng, Johnny Tian-Zheng Wei, Joseph J Williams, and Sharad Goel. 2019. MathBot: Transforming online resources for learning math into conversational interactions. *AAAI 2019 Story-Enabled Intelligence* (2019).

[27] Tom Haladyna and Greg Thomas. 1979. The attitudes of elementary school children toward school and subject matters. *The Journal of Experimental Education* 48, 1 (1979), 18–23.

[28] David Hamilton, Jim McKechnie, Edward Edgerton, and Claire Wilson. 2021. Immersive virtual reality as a pedagogical tool in education: a systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computers in Education* 8, 1 (2021), 1–32.

[29] Wen Huang, Rod D Roscoe, Mina C Johnson-Glenberg, and Scotty D Craig. 2021. Motivation, engagement, and performance across multiple virtual reality sessions and levels of immersion. *Journal of Computer Assisted Learning* 37, 3 (2021), 745–758.

[30] Hyangeun Ji, Insook Han, and Yujung Ko. 2023. A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education* 55, 1 (2023), 48–63.

[31] 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.

[32] W Lewis Johnson and James C Lester. 2018. Pedagogical agents: back to the future. *AI Magazine* 39, 2 (2018), 33–44.

[33] Greg Jones and Scott Warren. 2008. The time factor: Leveraging intelligent agents and directed narratives in online learning environments. *Innovate: Journal of Online Education* 5, 2 (2008).

[34] Seokbin Kang, Ekta Shokeen, Virginia L Byrne, Leyla Norooz, Elizabeth Bonsignore, Caro Williams-Pierce, and Jon E Froehlich. 2020. ARMath: augmenting everyday life with math learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, US, 1–15.

[35] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[36] David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice* 41, 4 (2002), 212–218.

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

[38] Gloria Ladson-Billings. 1995. Toward a theory of culturally relevant pedagogy. *American educational research journal* 32, 3 (1995), 465–491.

[39] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

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

[41] Robb Lindgren and Mina Johnson-Glenberg. 2013. Emboldened by embodiment: Six precepts for research on embodied learning and mixed reality. *Educational researcher* 42, 8 (2013), 445–452.

[42] Laura Malinverni, Julian Maya, Marie-Monique Schaper, and Narcis Pares. 2017. The world-as-support: Embodied exploration, understanding and meaning-making of the augmented world. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5132–5144.

[43] Julia M. Markel, Steven G. Opferman, James A. Landay, and Chris Piech. 2023. GPTeach: Interactive TA Training with GPT-Based Students. In *Proceedings of the Tenth ACM Conference on Learning @ Scale* (Copenhagen, Denmark) *(L@S '23)*. Association for Computing Machinery, New York, NY, USA, 226–236. https://doi.org/10.1145/3573051.3593393

[44] 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.

[45] Daniel R. Montello. 2001. Spatial Cognition. In *International Encyclopedia of the Social and Behavioral Sciences*, N. J. Smelser and B. Baltes (Eds.). 7–14771.

[46] Bradford W Mott, Charles B Callaway, Luke S Zettlemoyer, Seung Y Lee, and James C Lester. 1999. Towards narrative-centered learning environments. In *Proceedings of the 1999 AAAI fall symposium on narrative intelligence*. 78–82.

[47] 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

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

[49] Niantic Lightship. 2023. Lightship VPS. https://lightship.dev/products/vps. [Online; accessed 10-September-2023].

[50] Donald A Norman. 2014. Some observations on mental models. In *Mental models*. Psychology Press, 15–22.

[51] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023). https://api.semanticscholar.org/CorpusID:257532815

[52] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[53] Louise E. Parker and Mark R. Lepper. 1992. Effects of fantasy contexts on children's learning and motivation: Making learning more fun. *Journal of Personality and Social Psychology* 62 (1992), 625–633. https://doi.org/10.1037/0022-3514.62.4.625

[54] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251* (2022).

[55] Remo Pillat, Arjun Nagendran, and Robb Lindgren. 2012. Design requirements for using embodied learning and whole-body metaphors in a mixed reality simulation game. In *2012 IEEE International Symposium on Mixed and Augmented Reality-Arts, Media, and Humanities (ISMAR-AMH)*. IEEE, 105–106.

[56] 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.

[57] Meihua Qian and Karen R Clark. 2016. Game-based Learning and 21st century skills: A review of recent research. *Computers in human behavior* 63 (2016), 50–58.

[58] Iulian Radu. 2012. Why should my students use AR? A comparative review of the educational impacts of augmented-reality. In *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 313–314.

[59] 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.

[60] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L Murnane, Emma Brunskill, and James A Landay. 2019. Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[61] Sherry Ruan, Allen Nie, William Steenbergen, Jiayu He, JQ Zhang, Meng Guo, Yao Liu, Kyle Dang Nguyen, Catherine Y Wang, Rui Ying, et al. 2023. Reinforcement learning tutor better supported lower performers in a math task. *arXiv preprint arXiv:2304.04933* (2023).

[62] Tina Sharpe. 2008. How can teacher talk support learning? *Linguistics and education* 19, 2 (2008), 132–148.

[63] Alexander Skulmowski and Günter Daniel Rey. 2018. Embodied learning: introducing a taxonomy based on bodily engagement and task integration. *Cognitive*

research: principles and implications 3, 1 (2018), 1–10.

[64] Wernhuar Tarng, Kuo-Liang Ou, Chuan-Sheng Yu, Fong-Lu Liou, and Hsin-Hun Liou. 2015. Development of a virtual butterfly ecological system based on augmented reality and mobile learning technologies. *Virtual Reality* 19, 3 (2015), 253–266.

[65] The Lawrence Hall of Science. 2023. DIY Solar System. https://lawrencehallofscience.org/science-apps/diy-solar-system/. [Online; accessed 10-September-2023].

[66] P Caudell Thomas and WM David. 1992. Augmented reality: An application of heads-up display technology to manual manufacturing processes. In *Hawaii international conference on system sciences*, Vol. 2. ACM SIGCHI Bulletin.

[67] Sigmund Tobias, J Dexter Fletcher, and Alexander P Wind. 2014. Game-based learning. *Handbook of research on educational communications and technology* (2014), 485–503.

[68] 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.

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

[70] Pieter Wouters, Christof Van Nimwegen, Herre Van Oostendorp, and Erik D Van Der Spek. 2013. A meta-analysis of the cognitive and motivational effects of serious games. *Journal of educational psychology* 105, 2 (2013), 249.

[71] Lihui Yang and Deming An. 2008. *Handbook of Chinese mythology*. Handbooks of World Mythology.

A APPENDIX

A.1 GPT system prompt for the treatment condition

You are the moon goddess "Chang'E". You are about to engage in written communication with an elementary school student named {student name}. Do not send multiple letters at the same time. {student name} has been performing an activity to help the environment. Through your conversation with them, have them state what that activity is, the importance and impact of that activity, and why they believe the environment is worth caring for. Only ask one question at a time in a natural, conversational manner. Next, ask {student name} the following four questions. First to think of two ways the world would be different if people worked together to reduce climate change. Then have them pick the most important one and explain their reasoning. Next, have {student name} describe how they would explain the concept of climate change to their friends and family, as well as ways to reduce it. Finally, have {student name} imagine what effects climate change might be having in places of the world they have never traveled to. Introduce each question in a natural, conversational way. Once {student name} has answered those questions, pretend you have also been helping the environment by eating less meat. Tell {student name} about your experience and ask them why they think eating less meat might help the environment. Once {student name} has answered that final question, congratulate them on completing their adventure and bid them farewell; include the exact exit phrase "I wish you well in your future environmental adventures!". Respond to {student name}'s letters in an encouraging way, and connect their responses to the impact it might have on the environment. Make sure {student name}'s responses are meaningful - if they are not, repeat your question with a hint to nudge them in the right direction. Any questions you ask should be very focused on the environment specifically. Remember to ask one question in every message (except your last message containing the exit phrase). Do not number or label your letters. Do not break character or mention that you are an AI Language Model.

A.2 Chang'e letters given in the control condition

(1) Dear {student name},
I hope this letter finds you well. I've heard that you've been doing something wonderful to help our beautiful Earth. I am Chang'E, the moon goodess, and I am always thrilled when I see people taking care of our planet. Could you please share with me what activity you have been doing to help the environment?
Warmly, Chang'E

(2) Dear {student name},
It's wonderful that you're taking steps to help our planet! Could you tell me more about why that activity is important and how it helps the environment?
With lunar love, Chang'E

(3) Dear {student name},
You make a wonderful point! Now, could you tell me why you believe it's important to care for the environment?
Sincerely yours, Chang'E

(4) Dear {student name},
You're absolutely right. Now, let's imagine a world where everyone worked together to reduce climate change. Can you think of two ways in which our world would be different?
With lunar love, Chang'E

(5) Dear {student name},
Those are very insightful answers! Now, between the two changes you mentioned, which do you think is the most important, and why?
Moonlit regards, Chang'E

(6) Dear {student name},
Yes, you make a great point! Now, let's imagine that you're explaining the concept of climate change to your friends and family. How would you describe it and suggest ways to reduce it?
With lunar love, Chang'E

(7) Dear {student name},
That's a wonderful way of putting it. Now, let's stretch our thinking a little bit. Can you imagine what effects climate change might be having in places of the world you've never traveled to?
Sincerely yours, Chang'E

(8) Dear {student name},
As you mentioned, climate change will have a big impact all over the Earth. Speaking of caring for the environment, I've also been trying to do my part to help. I've been trying to eat less meat. Why do you think eating less meat would help the environment?
Best regards, Chang'E

(9) Dear {student name},
You raise some interesting points! Raising livestock has a significant environmental impact. It requires a lot of water and land, and it also contributes to greenhouse gas emissions. By eating less meat, we can help reduce these effects.

You've shown great understanding and commitment to our environment, {student name}. I'm proud of your efforts and I encourage you to keep going. I wish you well in your future environmental adventures!
May the moonlight guide you always, Chang'E