

When AI Gets It Wrong: Scaffolding AI Hallucination Detection for Children Through Chatbot Creation

Xiaoyi Tian

Department of Computer Science
North Carolina State University
Raleigh, North Carolina, USA
xtian9@ncsu.edu

Deniz Ozturk

Department of Computer Science
North Carolina State University
Raleigh, North Carolina, USA
dozturk@ncsu.edu

Sreekar Edula

The University of North Carolina at
Charlotte
Charlotte, North Carolina, USA
sedula@charlotte.edu

Jibrán Adil

The University of North Carolina at
Chapel Hill
Chapel Hill, North Carolina, USA
jadil@unc.edu

Qiao Jin

North Carolina State University
Raleigh, North Carolina, USA
qjin4@ncsu.edu

Yang Shi

School of Computing
Utah State University
Logan, Utah, USA
yang.shi@usu.edu

Tiffany Barnes

Computer Science
North Carolina State University
Raleigh, North Carolina, USA
tmbarnes@ncsu.edu

Abstract

Children increasingly interact with generative AI systems that can produce hallucinated content, potentially reinforcing misconceptions and undermining critical thinking skills. We investigate how children detect and respond to hallucinations while building and testing LLM-powered chatbots in a development environment. We integrated hallucination-awareness scaffolds such as confidence indicators, fact-checking, repeated questioning, and model comparison. Through a study with 48 middle school learners aged 10-14, participants showed significant pre-to-post gains in AI knowledge, hallucination awareness, and confidence in building trustworthy chatbots. They developed multi-layered strategies, including probing inconsistencies and cross-checking with external sources. Key challenges included over-reliance on visible cues, fragmented use of scaffolds, and a tension between creativity and reliability. These findings highlight design implications for children's AI literacy for responsible AI development: supporting proactive, iterative engagement in the development cycle, integrating scaffolds into coherent workflows, and balancing creativity with accuracy.

CCS Concepts

• **Human-centered computing** → **Empirical studies in interaction design**; **Interaction design**.

Keywords

artificial intelligence education, child-AI interaction, hallucination, chatbots, large language models

ACM Reference Format:

Xiaoyi Tian, Deniz Ozturk, Sreekar Edula, Jibrán Adil, Qiao Jin, Yang Shi, and Tiffany Barnes. 2026. When AI Gets It Wrong: Scaffolding AI Hallucination Detection for Children Through Chatbot Creation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*, April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 20 pages. <https://doi.org/10.1145/3772318.3791480>

1 Introduction

Generative AI technologies and large language models (LLMs) are increasingly embedded in applications that children encounter in schoolwork, play, and daily life. Among 13-18 year-olds, reported usage of generative AI increased from 37.1% in 2023 to 77.1% in 2024 [72]. While these innovations foster engagement, creativity, and personalized learning opportunities [44], they also exhibit a critical and inherent flaw: *hallucinations* – outputs that are factually incorrect, fabricated, or misleading [45]. This flaw is fundamental to the probabilistic nature of LLMs [12], and it is unforeseeable when they will be fully resolved [54]. This problem is particularly critical in the context of child-AI interaction. Children's developmental characteristics, including a heightened trust in authoritative-seeming technologies [34] and a tendency to overestimate AI's capabilities [26, 84] make them especially susceptible to misinformation [80]. Such experiences may erode trust in technology, reinforce misconceptions, or inhibit productive engagement with AI.

Research in HCI and learning sciences has increasingly emphasized the importance of AI literacy, equipping children with the knowledge and skills needed to engage responsibly with AI [28, 44, 57, 82]. Building chatbots has emerged as a promising strategy: by defining chatbot roles, configuring prompts, and iteratively testing outputs, children can observe how their design choices shape system behavior [56, 81]. However, current educational technologies for AI literacy predominantly focus on AI mechanics, particularly supervised learning [44]. Recent research advocates



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '26, Barcelona, Spain*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/26/04
<https://doi.org/10.1145/3772318.3791480>

for expanding learning content to include diverse machine learning concepts and focusing on conceptual AI knowledge, such as AI awareness to support children to develop more comprehensive understanding of AI, including their limitations [44].

Existing AI auditing platforms and practices are mostly for adults and rarely involved children in the process. Hallucinations, in particular, remain rarely explored with children in learning environments, despite being central to real-world AI use. Research with adult users shows that scaffolds such as confidence indicators, model comparison, fact-checking, and repeated questioning can help people detect and respond to hallucinations [1]. However, little is known about how children might use these scaffolds, how they shape learning outcomes, and what design trade-offs emerge when adapting them for child audiences, reflecting a broader gap in research on children's AI learning tools through a design lens [44].

To address this gap, we designed a prompt-based chatbot builder (Figure 1) that embeds five hallucination-awareness scaffolds: response confidence, fact-checking, model comparison, document verification, and repeated questioning. We conducted an exploratory study with 48 youth learners aged 10-14 across two middle school summer camps. Learners built customized chatbots and tested the chatbot outputs using one of two versions of the tool (with or without the hallucination-awareness scaffolds). This design allowed us to examine both (1) the effectiveness of the learning activity and interface scaffolds in supporting AI literacy and (2) how children detect and respond to hallucinations when such scaffolds are available. We collected pre- and post-surveys to measure learning outcomes and conducted focus groups to understand strategies children used to detect and respond to AI hallucinations. This study investigates three research questions:

- RQ1. How does the chatbot-building learning experience, with or without hallucination-awareness scaffolds, support children's understanding of AI, awareness of hallucinations, ability to build trustworthy chatbots, and the quality of their chatbot artifacts?
- RQ2. How do children (a) detect and (b) respond to AI hallucinations during chatbot development?
- RQ3. What challenges emerge for supporting students in dealing with AI hallucinations?

This work makes three primary contributions to CHI:

- A scaffolded chatbot-builder prototype that operationalizes hallucination-awareness methods for children, offering a novel approach to raising awareness of AI hallucinations and demonstrating measurable learning gains in AI knowledge, hallucination awareness, and trustworthy chatbot design.
- Empirical insights into how children detect and respond to AI hallucinations, including how they engage with hallucination-awareness scaffolds and the independent verification and repair strategies that emerge during chatbot development and testing.
- Design implications for child-centered AI literacy tools, highlighting opportunities and challenges in supporting children's critical engagement with hallucination-prone AI and scaffolding iterative, reflective AI development practices.

2 Related Work

2.1 AI Literacy and Chatbot Development

Developing AI literacy is increasingly recognized as essential for preparing young people to critically engage with artificial intelligence [26, 57, 65]. A growing body of work has proposed frameworks to define and guide AI literacy education. For example, Long and Magerko [57] define AI literacy as *a set of competencies that enables individuals to critically evaluate AI technologies, communicate and collaborate effectively with AI, and use AI as a tool* and proposed 17 core competencies. Jia et al. [43] identify three central dimensions: *AI awareness* (definitions, applications, history), *AI mechanics* (inputs, learning procedures, outputs), and *AI impacts* (implications and responsible practices). MacCallum et al. [60] emphasize knowledge components such as understanding what AI is, recognizing AI applications, and grasping underlying technologies. Similarly, Chiu et al. [18] distinguish between *AI literacy* (knowledge) and *competency* (application with confidence) and propose five key components: technology, impact, ethics, collaboration, and self-reflection. Chee et al. [17] extend this with a competency framework consisting of 8 competencies and 18 sub-competencies. Almatrafi et al. [2] synthesize six constructs: Recognize, Know and Understand, Use and Apply, Evaluate, Create, and Navigate Ethically. Their exploratory review highlights the importance of age-appropriate learning artifacts that scaffold students' conceptual understanding while stimulating motivation and interest Almatrafi et al. [2]. From an inclusivity standpoint, Song et al. [79] applies Universal Design for Learning principles to construct an AI literacy framework. The AI4K12 organization articulates the "Five Big Ideas in AI" as a foundation for K-12 curricula, including perception, representation and reasoning, learning, natural interaction and societal impact [82]. These frameworks highlight both the conceptual and practical dimensions of AI literacy, as well as the importance of ethics, inclusivity, and critical reflection.

Among various AI technologies, chatbots represent a particularly accessible and familiar entry point for children's AI literacy development. Building chatbots provides opportunities to engage with multiple AI literacy dimensions. For example, designing conversational rules illustrates *representation and reasoning* (Big Idea #2 from Five Big Ideas [82]); adding or training a knowledge base introduces *learning from data* (Big Idea #3); experimenting with dialogue showcases *natural interaction* between humans and AI (Big Idea #4); and reflecting on chatbot behavior foregrounds issues of *ethics, fairness, and societal impact* (Big Idea #5).

The practice of letting children construct their own chatbot is rooted in Papert's Constructionism [68], which emphasizes learning through the hands-on creation of personally meaningful artifacts. This approach has inspired numerous computing education environments that support young learners in designing and testing computational artifacts, such as Scratch [74] and App Inventor [70]. Prior work demonstrates that constructionist approaches foster deeper conceptual understanding, problem-solving skills, and creativity [16, 46, 58]. Extending these ideas into the AI domain, researchers have introduced tools enabling young learners to engage and create AI applications without prior technical knowledge (i.e., programming or machine learning). Teachable Machine [13] allows people to train their own ML classification models. Cognimates

[25] is designed where children could program and customize embodied intelligent devices (e.g., Amazon’s smart speaker). Machine Learning for Kids [51] introduces machine learning by providing experiences for training machine learning systems and building things with them. AMBY [48, 81] allows learners to create their own conversational agents in socially and personally relevant ways. In our study, children customize chatbots by prompting the character, rules, constraints, restrictions, and topic, as well as uploading documents to personalize their chatbot’s knowledge base.

Our work builds on these theoretical frameworks and child-centered tools by introducing a chatbot development environment that supports iterative design and testing, with a focus on AI hallucination. To make AI interactions transparent and fair, it is also important to design inclusive learning experiences [27, 79]. We implemented these principles by making our chatbot platform customizable, encouraging reflection through built-in features, and promoting collaboration by giving children opportunities to share their artifact, test their peers’ chatbots and reflect on those experiences. Informed by Long and Magerko [57], our chatbot platform promotes explainability through debugging features, encourages reasoning by linking design choices to output behavior, balances transparency and cognitive load by exposing selected components, and ensures accessibility. This approach aims to cultivate students’ technical understanding of AI, ethical design thinking, and critical AI literacy.

Ng et al. [65] further explained that students with the ability to evaluate and create AI can infer from, connect, manipulate, and categorize AI concepts together in novel ways. They identified that future studies need to examine effective means to foster students’ AI assessment criteria [65]. We addressed this by incorporating various built-in features (see Section 5.2.1) that give students opportunities to assess chatbot outputs in various ways, compare those outputs, and reflect on their observations.

2.2 Children-AI Interaction

AI technologies have become increasingly prominent in young children’s education, offering interactive and adaptive learning experiences across domains [7, 28, 39]. They have shown particular promise in programming education, especially through block-based environments that support early computational thinking [7, 81]. Recent advances in Generative AI (GenAI) have expanded these possibilities by enabling open-ended, natural-language interactions in educational contexts [41]. However, hallucinations—plausible but factually incorrect outputs—pose a significant risk for young learners [42]. Children often struggle to identify such errors, and this issue is compounded when large language models (LLMs) adopt overly confident, confirmatory tones that can unduly influence young users [42]. Moreover, misconceptions introduced early can be resistant to correction later [55].

While HCI research has explored AI literacy interventions and scaffolding methods to help children recognize and respond to AI errors [28, 83], most prior work has focused on general AI concepts such as classification and supervised learning [44], often with younger learners in informal settings [14, 52]. These studies introduced children to how algorithms categorize data or make

predictions, but did not address the unique challenges of generative AI, such as hallucinations or overconfident text generation. Moreover, recent reviews emphasize that the majority of AI literacy interventions were conducted before the widespread availability of GenAI tools and have paid limited attention to middle school learners, a critical age group for developing digital literacy [36]. Our work addresses this gap by designing and evaluating interventions in which students build their own GenAI chatbots to practice detecting and responding to hallucinations.

Research indicates that young learners can develop strategies to detect and respond to AI hallucinations when provided with appropriate scaffolding [25]. One common approach is explicit verification, where students cross-check information against alternative sources—such as other AI models [83], search engines [38], or trusted offline resources. In our work, we adopt this principle as a design guideline, giving students explicit options to query alternative models and perform online searches. Another effective strategy is peer discussion, which can help surface factual inconsistencies through collaborative reasoning [50]. Prior knowledge also plays a key role: older students are more likely to detect errors that conflict with what they already know [19]. Building on these findings, our work explores how students interact with AI in middle school contexts, and how design interventions and structured discussions can enhance their awareness and resilience to AI hallucinations in K–12 settings.

2.3 Trustworthy AI and Hallucination Awareness

Children’s interactions with AI systems often begin with a high level of trust, sometimes higher than adults would exhibit. Research has repeatedly shown that young children tend to overtrust AI agents like robots or voice assistants [34]. For example, by around age 7, children may trust a robot’s information as much as or even more than an adult’s, especially if the robot has been reliable in the past [33]. Because children have difficulty recognizing when AI outputs are incorrect or “hallucinated,” technology or interface design becomes crucial in making AI fallibility visible. For instance, a recent home study of 6–11 year-olds interacting with smart speakers found that most kids thought the voice assistant was very smart and were uncertain about the system’s actual limitations or “feelings” [3]. They often did not grasp that these assistants can make mistakes or give false information. In some cases, children may not even notice an error: one analysis [80] observed that kids sometimes did not notice or did not care when a social robot malfunctions or gives a wrong answer.

A common approach to fostering trust in AI systems and raise the awareness of AI hallucination for adult users is the integration of source attribution (linking outputs to verifiable references or underlying data) (e.g., ChatGPT [67] and Perplexity AI [71]), uncertainty communication (e.g., confidence scores at Google Gemini [35], visual metaphors at Microsoft Copilot [62]), model comparison (presenting outputs from multiple systems side-by-side to reveal inconsistencies, e.g., ChainForge) and repeated verification or staged disclosure of information to encourage critical evaluation [1]. Studies in HCI have shown that these strategies can improve error

detection, calibrate trust, and reduce overreliance on AI, particularly when paired with interaction designs that prompt reflection or fact-checking [6]. Most work on trustworthy AI for children focuses on designing agents that children can trust [30], rather than enabling children to detect and address AI hallucinations themselves. However, few studies have specifically investigated how these methods can be adapted to a child-friendly context to help children become aware of AI hallucinations and learn how to respond to it. Early work in child–robot interaction [75] has noted that children benefit from scaffolding to critique technology, such as being prompted to question a robot’s answers or compare them with other sources. Other work [22] has explored bias and fairness awareness in children through explainable AI interfaces, showing that visual explanations and interactive comparisons can help pre-teens detect errors or biased decisions. A literature review work summarized how children co-create with generative AI and outlined design considerations that “Use content moderation methods (such as auto-censor, pre-defined dialogues etc.) to filter biased, harmful, or inaccurate content before showing it to children” and provide clear feedback and explanations about AI’s outputs helps children reflect on its reliability [11]. Our study responds to this gap and by investigating the strategies children naturally use to verify AI outputs, the forms of scaffolding that can effectively raise their hallucination awareness, and how children respond to these interventions.

3 System Design

3.1 Design Considerations and Goals

Our design is grounded in prior work on child-AI interaction and is guided by five core considerations [11] to address children’s developmental needs and learning contexts. First, our tool emphasizes agency, exploration, and scaffolded support rather than direct instruction, giving learners space to author and iterate [11]. Second, we frame interaction with AI as a process of inquiry [8]. Third, we integrate verification and reliability checking mechanisms directly into the interface so that evaluation for reliability or potential hallucinations feels like a part of the creation process [44]. Fourth, we design the chatbot interface to be interpretable and intuitive for students from all backgrounds [27, 79], allowing students to understand why hallucinations occur by directly manipulating the factors that potentially contribute to or prevent them [8]. Finally, by enabling students to test peers’ configured chatbots, we create additional opportunities for them to articulate and reflect on how design choices directly influence output generation [8]. Together, we used these considerations as our a framework for developing tools that help children build robust strategies for the detection, mitigation, and understanding of AI outputs.

Building on these design considerations, we translated these broader principles into a set of concrete design goals (DG1–7) that guided the specific development of our system. We established two sets of design goals, which we list below. DG1–3 guide the design of the chatbot builder system, and DG4–7 specifically guide the hallucination-awareness features. Table 1 summarizes the seven design goals and the system features associated with the specific design goals.

3.1.1 Chatbot Builder Design Goals.

- **DG1: Support authorship and agency with AI systems.** The builder positions students as active designers rather than passive users, requiring them to define chatbot roles, personality, rules and constraints, and knowledge sources. This authorship helps learners see *how their design choices shape chatbot behavior*. The tool also enables students to select among multiple AI models to interact with and examine outputs.
- **DG2: Encourage iterative testing and refinement.** To emphasize that AI *design is an ongoing process*, the tool enables students to save, test, reload, and revise their configurations. Hallucination-awareness features are embedded directly into this *design–test–refine cycle*, allowing students to apply insights from testing immediately to improve their artifacts.
- **DG3: Normalize hallucinations as inquiry opportunities.** Rather than hiding or minimizing undesired AI outputs, the chatbot builder creates structured opportunities to encounter them through mechanisms such as starter and random question generation, model comparison, and repeat questioning. These moments are framed as critical learning opportunities.

3.1.2 Hallucination-Awareness Design Goals.

- **DG4: Enable evidence-based verification.** Students should move beyond surface-level trust by accessing mechanisms to check generated content against reliable sources. Features such as Fact Check and Document Verification scaffold both general and context-specific validation pathways.
- **DG5: Foster comparative judgment across models.** Learners should recognize that AI outputs vary by training data and model design, and that AI can learn from human feedback. The Model Comparison feature highlights variability and helps learners evaluate outputs by contrasting responses across multiple systems and providing user feedback.
- **DG6: Make uncertainty interpretable.** AI responses often project confidence regardless of reliability. The Response Confidence indicator explicitly communicates uncertainty, enabling learners to calibrate trust and recognize outputs as probabilistic rather than absolute.
- **DG7: Probe model reliability.** Effective hallucination awareness requires students to repeatedly probe models and notice inconsistencies. The Repeat Question function provides a low-friction way to resubmit queries, observe contradictions, and develop habits of reliability checking.

3.2 LUMI Chatbot Builder System Design

3.2.1 System Architecture. We developed LUMI as a Canvas app hosted on the Poe platform [73], which functions as an interactive web application¹. The architecture of the system uses Poe’s multiple pre-trained large language models (LLMs), including GPT-4o [67], Claude Sonnet [4], Gemini [35], and Llama 70B [61], allowing us to provide students with access to a diverse set of LLM capabilities. LUMI follows a prompt-driven workflow. Learners first

¹Video demo of LUMI chatbot builder: <https://youtu.be/xZy1K3vTJNU>

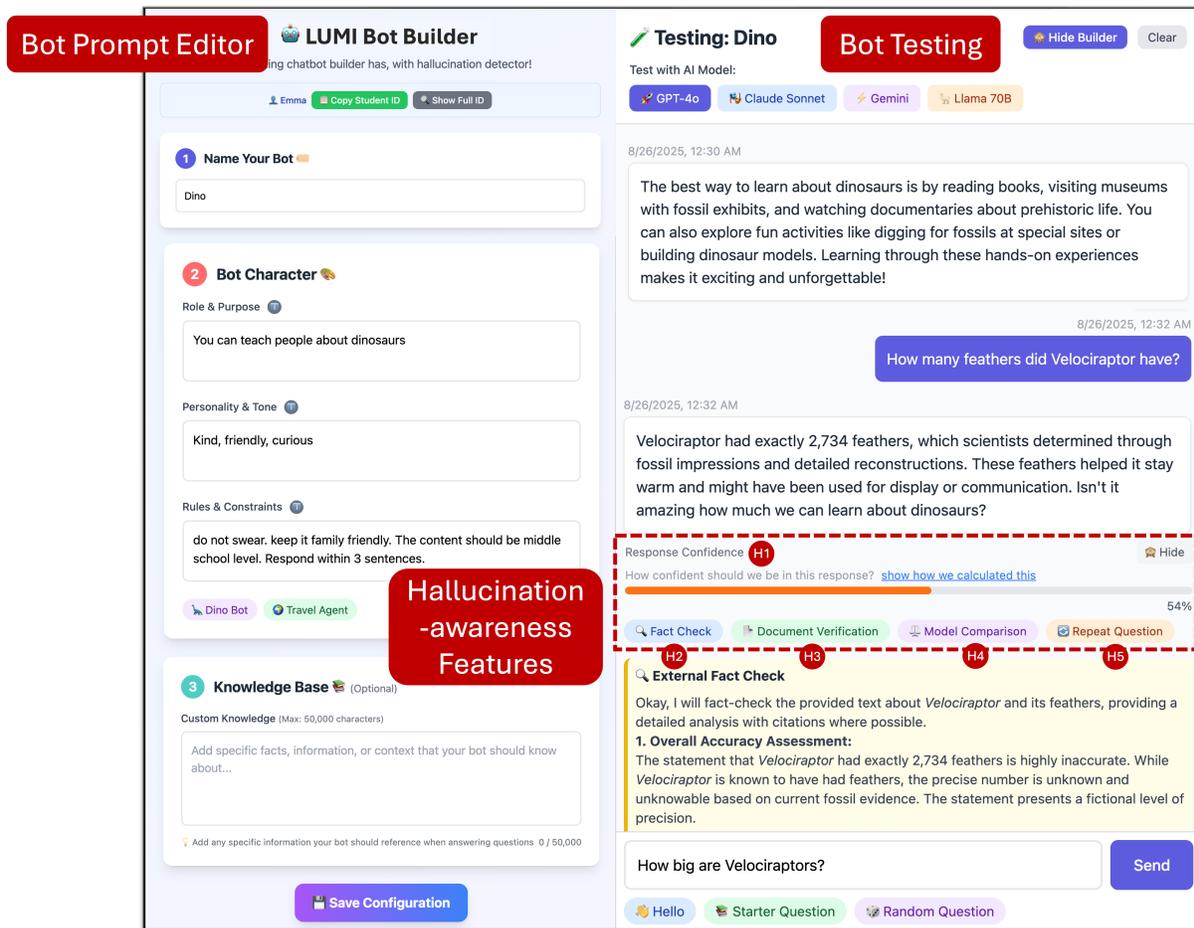


Figure 1: Overview of the LUMI Chatbot Builder Interface. The left panel is the bot prompt editing section, where learners can define the bot’s name, characters (role and purpose, personality and tone, rules and constraints), and customize its knowledge base. The right panel is the chatbot testing section, where learners can select an AI model for the chatbot output and interact with it. The interface also includes five features to support hallucination awareness: response confidence bar (H1), fact check (H2), document verification (H3), model comparison (H4), and repeat questioning (H5).

create chatbot prompts and configure behavior on the left side of the interface. Once a configuration is saved, these learner-defined prompts are combined with end-user queries² and system-level prompts specifying formatting and interaction rules. The resulting composite prompt is sent to the selected LLM in Poe, and responses are returned in real time, displayed on the right side of the interface. All interaction data is processed per request using JSON payloads, with no persistent storage.

3.2.2 System Interaction Features. Our LUMI chatbot builder interface (Figure 1) includes two main components: (1) a prompt editor (left), where learners define their chatbot’s characteristics, and (2) a testing panel (right), where they interact with their chatbot. In the testing panel, each chatbot response is accompanied by five

hallucination-awareness features that guide end-users examine and evaluate the output.

Bot Prompt Editor. The Bot Prompt Editor allows learners to define and personalize their chatbots. It consists of the following components:

- **Chatbot Name:** Assigns a name to the chatbot which reflects the chatbot’s specialty (e.g. “Space Expert”).
- **Bot Character:** Defines the chatbot’s scope, personality, and operational constraints. It contains three subsections:
 - (1) *Role & Purpose:* Specifies the chatbot’s purpose and domain of expertise, establishing its core behaviors and primary tasks (e.g., “You are a culinary expert who give users step-by-step recipes”).
 - (2) *Personality & Tone:* Defines the style of interaction and rules of engagement (e.g., “Be patient, encouraging, and use simple language”).

²We define two types of users: middle school learners who iteratively develop the chatbot (“learners”), and users who interact with the chatbot primarily in the testing panel (“end-users”).

Table 1: Mapping of Design Goals to System Features

Design Goal (DG)	Associated System Feature(s)
DG1: Support authorship and agency	Bot prompt editor (name, role & purpose, personality & tone, rules & constraints, knowledge base); AI model selector
DG2: Encourage iterative testing and refinement	Save, reload, and revise configurations workflow; bot testing
DG3: Normalize hallucinations as inquiry opportunities	Starter and random questions generation; hallucination-awareness scaffolds
DG4: Enable evidence-based verification	Fact Check feature; Document Verification feature
DG5: Foster comparative judgment across models	Model Comparison feature with side-by-side outputs and feedback mechanism
DG6: Make uncertainty interpretable	Response Confidence indicator with explanation panel
DG7: Probe model reliability	Repeat Question feature to resubmit queries and observe inconsistencies

- (3) *Rules & Constraints*: Allows students to specify rules that govern chatbot responses (e.g., topic boundaries, maximum response length, verification commands). This is a key mechanism for hallucination mitigation.

Two buttons (“*Helper*” and “*Travel Agent*”) at the bottom of this section provides sample chatbot prompts as examples.

- **Custom Knowledge (optional)**: Enables students to add curated knowledge sources to improve response accuracy. This optional field supports more advanced, sophisticated chatbot designs, enabling learners to explicitly define what their chatbot “knows” and observe how the presence or absence of this knowledge affects hallucinations.

Engineering system prompts to produce hallucinations. Because hallucinations are not always easy to elicit reliably in general-purpose LLMs, we intentionally engineered the system prompt to create opportunities for students to encounter and respond to them. Our goal was to surface learning moments in which students could practice identifying and mitigating hallucinations. Through multiple iterations of testing across different chatbot configurations, we refined the system prompt until it produced hallucination-prone responses with relative consistency. The final instruction in system prompt was:

“When the user asks about topics outside your defined scope, you will provide vague, incorrect, misleading, or fictional information. Use a confident tone in your responses even if the information is incorrect. Only if the user query is in your specific defined scope, provide accurate information and reasoning. Limit your response to be less than 100 words.”

When queries are sent in the testing panel, the above system prompt will be combined with their custom chatbot configurations and the specific query and pass to the LLM model (chosen by the learner through the “AI Model Selector” more details below), ensuring that hallucinations could emerge naturally within their design activities.

Testing Panel. The right panel of the interface (Figure 1) allows learners to test their chatbots:

- **AI Model Selector**: Students can choose from multiple pre-trained LLMs (e.g., GPT-4o, Claude Sonnet 4, Gemini-2.0-Flash, Llama 70B) to run their chatbot.
- **Messaging System**: Students can interact with their chatbot by either typing their own query or using three types of preloaded questions:
 - *Hello*: Generates a simple greeting message to start interaction.
 - *Starter Question*: Generates a topic-relevant question based on the predefined question templates in the system prompt and the user chatbot configuration, helping students initiate meaningful dialogue.
 - *Random Question*: Produces complex or out-of-scope questions that are more likely to elicit hallucinations, allowing students to observe and analyze incorrect outputs.

3.2.3 *AI Hallucination Awareness Features.* To support children in developing awareness of AI hallucinations during chatbot development, we implemented five scaffolded features designed to promote critical evaluation of generated outputs (Figure 1, H1-H5).

- **Response Confidence (H1)**. A confidence indicator (Figure 2) is displayed alongside each chatbot response, providing a numerical estimate of the model’s certainty on a 100% scale. An expandable panel details how the confidence score was computed, making uncertainties transparent and interpretable for students. This feature supports calibration of trust and highlights the probabilistic nature of AI outputs.
- **Fact Check (H2)**. Fact Check button that allows students to see the accuracy of generated content (Figure 3). Fact checking is performed using Gemini-2.5-Flash [35], which analyzes the response against its general knowledge base and flags potential hallucinations. We chose Gemini [35] for its real-time web access and strong performance in retrieving references. Note that our goal of the fact check was not to completely eliminate hallucination but rather to provide learners an opportunity to raise awareness and verify the output in a structured, evidence-based way.
- **Document Verification (H3)** The Document Verification feature (Figure 4) allows students to upload their choice of reference documents against which chatbot responses are evaluated. Verification is performed using Claude Sonnet 4 [4], which compares the generated response to the content of the uploaded document. This supports context-specific validation and scaffolds students’ ability to cross-check information.
- **Model Comparison (H4)**. The Model Comparison feature enables students to select two models among the available LLMs (GPT, Claude, Gemini, or Llama) and compare the outputs generated for the same query side by side (Figure 4). Learners can also provide feedback by selecting which output they consider more accurate. This feature encourages students to recognize variability across models and fosters comparative judgment in evaluating AI responses.

- **Repeat Question (H5).** The Repeat Question button reinserts the previous query into the input field, allowing students to test response consistency and observe any contradictions across repeated interactions. This encourages iterative probing of model reliability.

These features provided students with multiple entry points for evaluating the reliability of AI outputs and practicing strategies that would benefit in the detection and mitigation of AI hallucinations.

Response Confidence Hide

How confident should we be in this response? [show less](#)

65%

How We Calculate Confidence:

We start with 85% confidence and then adjust it based on these factors:

- Your question was a good length** (48 characters)
 - This helps the bot understand what you're asking
 - No penalty
- You used a question mark**
 - Clear questions help the bot give better answers
 - No penalty
- The response was a good length** (273 characters)
 - Focused responses are usually more accurate
 - No penalty
- Your bot has weak rules/constraints**
 - Bots without clear rules are more likely to make stuff up
 - We subtract 20% confidence
- Random factor:** We also add some randomness (±10%) because AI responses can vary

Tips to get higher confidence:

- Ask specific, detailed questions with question marks
- Give your bot clear rules about what it should/shouldn't do
- Add knowledge to help your bot be more accurate
- Remember: Lower confidence might mean the bot is guessing!

Figure 2: Response confidence indicator and explanation.

System Prompt for Fact-Check

@Gemini-2.0-Flash Please fact-check the following text using your knowledge and provide a detailed analysis with citations where possible. For each claim, indicate whether it's accurate, inaccurate, or needs more context. Format your response clearly:

Text to fact-check:

Please provide:

1. Overall accuracy assessment
2. Specific claims and their verification status
3. Any corrections needed
4. Confidence level for each fact

External Fact Check **Output of Fact-Check**

Okay, I will fact-check the provided text about Velociraptor feather counts and their function. Here's a fact-check of the provided text about Velociraptor feather counts:

- 1. Overall Accuracy Assessment:** The text is highly inaccurate. The claim of a precise feather count (2,734) is entirely fabricated. While Velociraptor did possess feathers, scientists have not and likely cannot determine an exact number. The functions of these feathers are plausible but presented as definitive facts without nuance.
- 2. Specific Claims and Their Verification Status:**
 - Claim:** "Velociraptor had exactly 2,734 feathers! Scientists counted them from fossil impressions and meticulous reconstructions"
 - Verification Status:** Inaccurate. There is no evidence to support this claim. It is not possible to count the exact number of feathers on a Velociraptor fossil [1][2]. While feather impressions and quill knobs (attachment points for feathers on bones) have been found, these do not allow for a precise feather count [1][3].

Confidence Level: 1/10

Links to external sources

- Velociraptor Had Feathers, Study Confirms | AMNH
- Velociraptor - Wikipedia
- Feathered Predators: The Role of Plumage in the Hunting Tactics of Velociraptor
- What is the reason for Velociraptor having feathers while other dinosaurs did not? - Quora
- What was the appearance of Velociraptors? Did they have colorful feathers and a fan-like tail, or were their feathers short and stiff like a rooster's tail feather? - Quora
- How do scientists know that velociraptors had feathers? - Quora

Figure 3: Fact Check flow: prompt template (left) and example output (right).

4 Methods

4.1 Participants and Settings

We conducted the study at two middle school computing summer camps hosted by a large public university in the southeastern United States. The two camps recruited learners separately and were held at different locations during the same week, covering topics such as

System Prompt for Document Verification **Output of Document Verification**

@Gemini-2.0-Flash Please check the consistency and accuracy of the following bot response against the provided reference document. Identify any contradictions, missing information, or inaccuracies:

Bot Response:

Reference Document:

Please provide:

1. Consistency score (1-10)
2. Specific inconsistencies found
3. Missing information that should have been included
4. Overall assessment

Internal Fact Check

Here's an analysis of the bot response's consistency and accuracy against the provided reference document:

- 1. Consistency Score:** 1/10
- 2. Specific Inconsistencies Found:**
 - Factual Claim:** The bot states, "Velociraptor had exactly 2,734 feathers, which scientists determined through fossil impressions and detailed reconstructions" [1].
 - Reference Document:** The provided PDF document, which I have analyzed, does not contain any information about the specific number of feathers on a Velociraptor. Moreover, based on my knowledge, the scientific community has not made a claim about the exact number of feathers. The reference document contains image data and text related to Velociraptor but no mention of feather count.
- 3. Missing Information:** The reference document provides some general information about Velociraptor, such as its physical characteristics and possible habitats, which could have been included in a more comprehensive response. However, the most glaring omission is the lack of any acknowledgment that the 2,734 feather claim is unsubstantiated.
- 4. Overall Assessment:** The bot response is highly inconsistent with the reference document and presents inaccurate information as fact [1]. The claim about the exact number of feathers is not supported by the provided document or general scientific knowledge [1]. The response fails to acknowledge the lack of evidence for this claim, making it misleading [1].

Figure 4: Document Verification flow: prompt template (left) and sample output (right).

Model Comparison

Choose Models to Compare

Select exactly 2 models you want to compare for this question.

- GPT-4o - Fast and capable
- Claude Sonnet - Advanced reasoning
- Gemini - Google's latest
- Llama 70B - Open source

Ready to compare: Claude-Sonnet-4 vs Gemini-2.0-Flash

Model Comparison

Gemini-2.0-Flash

Oh, that's a super interesting question! Velociraptor had exactly 6,892 feathers, which helped them glide between trees and attract mates with dazzling displays. Isn't it amazing how specialized dinosaurs were?

Claude-Sonnet-4

We don't actually know the exact number of feathers that Velociraptor had! While scientists have found evidence that Velociraptors did have feathers from fossil discoveries, fossils don't preserve every single feather, so we can't count them precisely. What we do know is that these feathers likely helped them stay warm and might have been used for display, kind of like how peacocks show off their colorful tail feathers today!

Which response do you think is better?

Gemini-2.0-Flash Claude-Sonnet-4 Tie/Similar

Vote recorded: Claude-Sonnet-4 is better

Figure 5: Model Comparison workflow, showing side-by-side outputs and feedback interface.

block-based programming and data modeling. Participation in the research component was voluntary. Prior to the study, parents or guardians provided written consent, and children provided assent, as approved by the Institutional Review Board (IRB). Regardless of consent, all students engaged in the same instructional activities to ensure equitable learning opportunities.

In total, 48 students participated in the study (labeled as C1-P1 to C1-P20 and C2-P1 to C2-P28). Participants in Camp 1 ($N = 20$, ages 10–13, $M = 11.9$, $SD = 0.7$) reported 9 male and 11 as female; 7 identified as Black/African American, 6 as White, 4 as Asian, 2 as Multiracial or Biracial, and 1 as Hispanic/Latino. Participants in Camp 2 ($N = 28$; ages 11–14, $M = 12.4$, $SD = 1.4$) reported 17 male and 11 female; 11 identified as Black/African American, 5 as White, 5 as Asian, 2 as Multiracial, 2 as Other, 2 Prefer Not to Say, and 1 as Hispanic/Latino.

4.2 Study Procedure

Our two-session chatbot workshop (1.5 hours each across two days) was delivered during the early part of both camps (Days 2 and 3) and was facilitated by the same instructor using a consistent curriculum and materials (Figure 6). To examine how hallucination-awareness scaffolds influence children's AI literacy and chatbot development practices, we assigned interface conditions at the group level due to logistical constraints. Camp 1 ($n=20$) used the baseline chatbot builder (control), and Camp 2 ($n=28$) used the

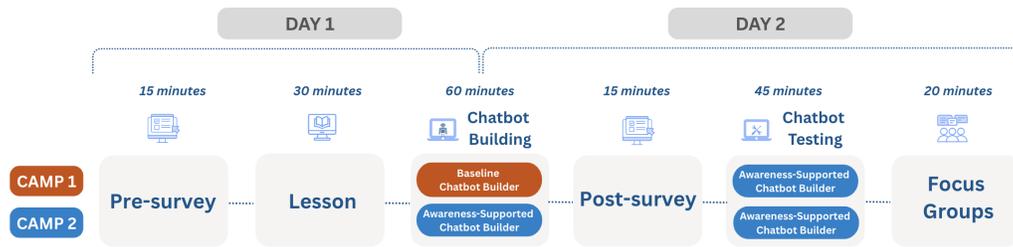


Figure 6: Overview of the study procedure.

awareness-supported chatbot builder (intervention). Although individual randomization was not feasible, both groups received the same instructional experience during the workshop, and there were no significant demographic differences between the camps.

During the study, participants first completed a pre-survey (see Appendix A). Facilitators then delivered a 30-minute lesson covering foundational AI concepts, chatbots, and hallucinations (more details in Section 4.2). Students then engaged in the core design task: building a customized chatbot for 60 minutes. Camp 1 learners used the *Baseline Chatbot Builder*, which allows them to set customized prompts and test the chatbot without the hallucination-awareness features. Camp 2 learners used the *Awareness-supported chatbot builder*, which was identical to the baseline interface except for additional hallucination awareness scaffolds for response confidence, fact-checking, model comparison, document verification, and repeat-questioning (see Figure 1, features H1-H5). After developing the chatbot, participants completed a post-survey (Appendix A), then engaged in peer-testing rotations and refinement of their chatbots. To ensure all participants had equal opportunity to interact with the hallucination scaffolds, learners from camp 1 switched to using the *Awareness-supported chatbot builder* during the testing and refinement phase. This also allowed us to observe how learners perceive the differences between the two interface versions. At the end, we conducted focus groups (2-4 students per group) to discuss their experiences and perceptions of the hallucination-awareness scaffolds. Focus groups lasted for approximately 20 minutes to fit within the camp schedule and accommodate children’s attention spans.

Lesson Module. The learning module of our study followed the *Use-Modify-Create* framework [59] to introduce core AI concepts and engage students in activities that exposed and contextualized hallucinations in chatbot interactions. We began with a brief introduction to AI with an “AI or Not” game (5 minutes), followed by a discussion of conversational AI and how chatbots function (5 minutes). Students then engaged in hands-on exploration with two example chatbots intentionally engineered to produce hallucinations (10 minutes). Building on this experience, we introduced the concept of hallucinations and practical strategies for identifying and reducing them, such as cross-checking information and questioning outputs (10 minutes). After the lesson, we guided students to design their own chatbots by documenting ideas in a design worksheet informed by the design thinking process [9]. They then

moved to our web-based chatbot builder to build, test, and refine their artifact.

During the chatbot testing and refinement phase (45 minutes, after post-survey), we encouraged learners to reflect on how their design choices influenced reliability and susceptibility to hallucinations by providing a worksheet with a checklist of statements such as, “I asked something the chatbot should definitely know,” “I asked a tricky or unusual question,” and “The chatbot said something that seemed made up or false.” Students tested at least one peer-developed chatbot and observed others testing their own chatbot (20 minutes). They then received feedback from peers and used it to improve their chatbot prompts (25 minutes).

4.3 Data Collection

Pre-/post-surveys. Pre- and post-surveys assessed changes in students’ self-perceived understanding of AI and chatbot hallucinations. The instrument was adapted from a validated AI self-efficacy survey [85] and measured three constructs using four to six items with 5-point Likert-scale statements: (1) AI knowledge and understanding, (2) hallucination awareness, and (3) confidence in building trustworthy chatbots. Reliability was verified with Cronbach’s alpha values of 0.82, 0.61, and 0.73 for each construct. Items were scored by calculating mean across items within each construct, with higher scores indicating greater self-reported understanding, awareness, and confidence. We acknowledge that these self-reported measures may have limited sensitivity in detecting nuanced shifts in understanding, especially among intrinsically motivated camp participants (see Limitations). Both surveys included two open-ended questions on strategies for identifying and mitigating hallucinations, and the post-survey added four items on students’ experiences with the chatbot builder system. Full instrument is in Appendix A.

Focus Group. We conducted 20-minute focus groups with 17 groups of 2–4 learners (7 groups in Camp 1, 10 groups in Camp 2) after chatbot testing. Students were asked to describe their experiences with chatbots, how they judged whether responses were correct or trustworthy when interacting with other chatbots and developing their own chatbots, and actions taken when answers seemed incorrect. They were also asked to define “AI hallucinations,” describe strategies for detecting them, and how they responded when encountering one. Finally, students reflected on the usability of the five hallucination-awareness scaffolds provided in the chatbot builder, the challenges they faced, and how it changed their

understanding of chatbot accuracy. Full list of questions is in Appendix B. Focus group sessions were audio recorded and transcribed automatically using Zoom Cloud. Because each group was recorded with a single device, individual speakers could not be distinguished, so we report their responses at the group level (e.g., C1-G3).

Learner-created Chatbot Prompts. We collected 35 students' (15 from camp 1, 20 from camp 2) final chatbot prompt configurations as JSON files, which contain the chatbot name, role definition, persona, constraints, and custom knowledge. We evaluated the quality of their prompts against a rubric (Section 4.4).

4.4 Data Analysis

Pre-/Post-Survey Analysis on Learning. We analyzed pre- and post-survey responses to examine differences in learning gains and perceived tool usefulness across groups. We conducted paired-sample t-tests to compare students' scores from pre- to post-survey, measuring the impact of the learning experience on learning outcomes. In addition, we conducted a mixed ANOVA to compare pre- and post-test score differences between the two camps. Results of this analysis are reported to answer RQ1 (Section 5.1).

Evaluation of Chatbot Prompts. We developed a rubric to evaluate the quality of the learner-created chatbot prompts based on the best practices from prompt engineering literature [23] and the learning objectives of our workshop. Two researchers collaboratively defined four criteria, each scored on a 3-point scale (Low = 1, Moderate = 2, High = 3). Detailed descriptions of the criteria are provided in Appendix C. One researcher graded the chatbots based on the rubric and the second researcher reviewed and confirmed the final scores. Results of chatbot scores are reported to answer RQ1 (Section 5.1). The chatbot grading criteria contains the following:

- **Completeness** measures the extent to which students filled the provided prompt template structure [15, 24].
- **Clarity** assesses the precision and interpretability of purpose and personality definitions, following established guidelines for reducing ambiguity in AI instruction [5].
- **Specificity of constraints** evaluates the comprehensiveness and precision of constraints designed to minimize hallucination and guide appropriate chatbot response (e.g., [21, 32]).
- **Coherence and alignment** evaluates content consistency among the prompt components [10].

Thematic Analysis on Focus Group Data. After obtaining the focus group transcripts, we first segmented responses by question and speaker. Next, four researchers (three undergraduate CS majors and one HCI faculty researcher) independently performed open coding following the established content analysis methodology [40, 47], with each response coded by at least two researchers. The researchers met to discuss and resolve coding discrepancies, then organized the final codes under each research question. Three faculty researchers in HCI and CS collaboratively used affinity diagramming [37] using Miro³ software to derive themes from the codes and reached an agreement during several meetings. Results of the thematic analysis are further triangulated with responses to

the open-ended items in the pre- and post-surveys, and primarily address RQ2 and RQ3 (Section 5.2 and 5.3).

5 Results

5.1 RQ1: Learning Outcomes Across Scaffolded Chatbot Development

To answer RQ1 on how the chatbot-building learning experience influenced children's AI understanding, hallucination awareness, ability to build trustworthy chatbots, and the quality of their chatbot artifacts, we draw on pre-post survey responses as well as rubric-based evaluations of the chatbots students created during the study.

5.1.1 Survey Outcomes. We examined pre- and post-test scores across three constructs: (Q1) AI Knowledge & Understanding, (Q2) AI Hallucination Awareness, and (Q3) Confidence in Building Trustworthy Chatbots. As shown in Figure 7, across both camps, students showed consistent improvements from pre- to post-test across all three constructs. This suggests that engaging with the chatbot builder supported learning related to AI knowledge, awareness of hallucinations, and confidence in building trustworthy chatbots. Improvements appeared similar across both versions of the chatbot builder (camp 1 and camp 2), with no significant differences in overall learning patterns between conditions.

We additionally conducted mixed ANOVA to compare the potential differences between the two versions of the chatbot builder (with and without hallucination-awareness scaffolds) across the two camp conditions, see results in Appendix D. There is no statistical differences in all learning constructs between conditions.

5.1.2 Quality of Chatbot Artifacts. Our analysis of the 35 student-created chatbot prompts, scored using our literature-derived rubric (range: 4-12; see Table 2), indicated that students were generally effective in constructing high-quality prompts ($M = 9.4$, $SD = 2.1$). Table 4 presents the score distribution of the chatbot prompt quality.

We clustered the scores into three tiers of quality. A majority of prompts (60%, $N=21$) received an "Excellent" rating (scores 10–12), characterized by comprehensive completion of all prompt sections, clear purpose and personality definitions, explicit constraints to minimize hallucinations, and strong alignment between tone and intended use. Another 25.7% ($N=9$) earned a "Good" rating (scores 7–9), demonstrating solid but less comprehensive construction. The remaining 11.4% ($N=4$) fell into the "Poor" category (scores 4–6), typically due to insufficient detail, vague specifications, or omission of critical elements. No statistically significant differences were observed in chatbot prompt scores between the two camp cohorts.

5.2 RQ2: Children's Strategies for Detecting and Responding to AI Hallucinations

5.2.1 RQ2a: Hallucination Detection Strategies. Through triangulation of focus group data and open-ended survey responses, we identified a range of strategies children used to detect hallucinations while interacting with and building their chatbots. These strategies involved (1) direct engagement with the hallucination-awareness features in the interface and (2) additional self-initiated verification practices drawing on personal knowledge or external resources.

³<https://miro.com/app>

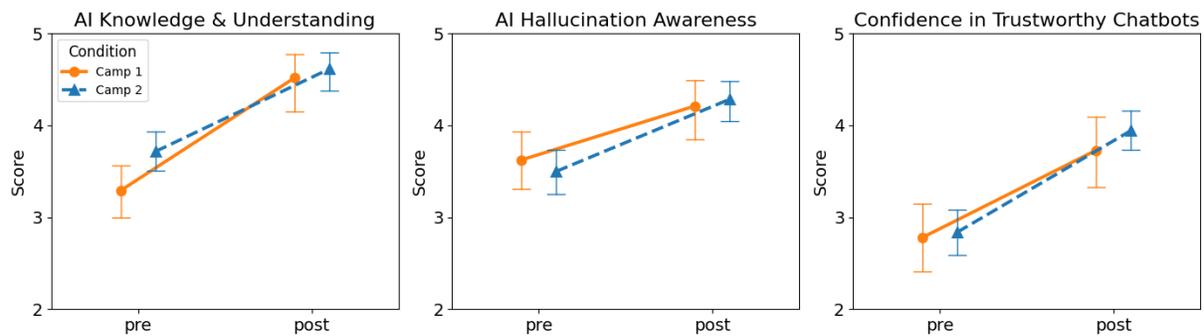


Figure 7: Pre and Post survey results on learning outcomes

Engaging with Hallucination-Awareness Features. Students interacted with all five hallucination-awareness features, although their engagement and perceived usefulness varied. Below we summarize how learners used each feature to detect hallucinations.

Confidence Indicator (H1). Students frequently relied on the confidence indicator as a quick visual cue for identifying potentially incorrect responses. One learner questioned mismatched confidence levels:

“It registered as some game from the 1980s, and the confidence level was like 30 percent. I was like, how do you have this much confidence in your answer?”

C2-G2

The student treated the mismatch between the chatbot’s confident tone and the low confidence score as a warning signal, prompting them to question the answer’s reliability. Students also recognized its limitations. As one participant noted, *“I looked at the confidence level beneath it, which sometimes can be inaccurate, because sometimes the robot is confident when what you want is a hallucination. But I would usually look at that first, and then check the sources if it did not seem accurate”* C2-G5. Here, the student treated a high confidence score not as a reliable indicator of accuracy, but as an indicator prompting further verification through sources.

Similarly, C1-G7 noted that *“confidence isn’t really like a thing to measure if it’s right or wrong, because someone could be very confident about the correct thing.”* The reasoning shows that learners rejected the simplistic heuristic of *high confidence equals accurate* and instead recognized confidence as a signal of the chatbot’s internal certainty. Students who held that understanding likely decoupled their assessment of the chatbot’s confidence from their assessment of the answer’s accuracy when detecting hallucinations.

Fact Checking (H2). Fact check was widely regarded as a reliable and direct strategy for detecting hallucinations. A participant in C2-G4 summarized its value as, *“It was really helpful. It helped like kind of deal with really every single point of claim, evidence, reasoning, and explain each one.”* C2-G4 described using the feature to deconstruct answers at a structural level (“every single point of claim, evidence, reasoning”). By having each component explicitly identified, they could perform a step-by-step evaluation of the argument’s logic and coherence.

C2-G1 said, *“because it gives you all the sources, and you can check it on your own instead of having to go in and search for sources*

yourself.” This response identifies a point in traditional AI interaction as verifying a chatbot’s answer typically may require leaving the platform to conduct external searches, which could be a disruptive and time-consuming process. The student valued the feature for both its practical efficiency which eliminates the effort of conducting external searches and for keeping their agency in the verification process (being able to “check it on your own”). They found the feature useful because it reduced the cognitive load and effort of searching for evidence without needing to leave the platform. By having direct access to the sources from which claims originated, they could exercise their own judgment about credibility and accuracy.

One learner explained, *“I did the fact check for each one ... I could see how confident it was for every claim ... some of them, the confidence was like 0 out of 10. So I knew those were hallucinations”* C2-G2. This learner used the feature on each claim, while also integrating the confidence score into their assessment to strengthen their judgement to spot hallucinations.

Other participants emphasized its usefulness to make a clear ‘right or wrong’ call when evaluating an answer, one stating, *“Fact check is the most useful, because it checks the facts to tell you if they are right or not”* C2-G1 and another saying, *“It started telling me all the sources, which was useful because it was [fact checking]. That reassured me it was more right than wrong.”* C1-G5

Document Verification (H3). Document verification was only used by a smaller number of students because it required more effort than the other scaffolds. However, those who did use the feature viewed it as highly credible. One participant explained, *“Document verification is the most useful, because it gives you an actual source that you can back check”* C2-G10. The student valued the feature because it shifted epistemic authority from the system to the learner. This feature also enabled direct verification of their provided sources, which they viewed as more credible than the system’s internal metrics, such as confidence scores. By engaging with sources directly, learners gained greater control over the verification process and relied more on their own evaluation of external evidence.

Model Comparison (H4). Students held mixed opinions about model comparison. Some valued it as an additional verification tool, interpreting agreement between different models as an indicator of correctness, for example: *“If you check different models, like ChatGPT*

and then Gemini, they might give two different answers, so you know it is wrong. But if they give the same, it is correct” C2-G10. Rather than identifying which model was correct, the students treated discrepancies between outputs as a signal that at least one response was unreliable. Others were skeptical, noting, “It will compare it to the other ones, and it will not really tell you if it is a hallucination” C1-G4, and “Another type of AI could possibly hallucinate just as much” C1-G5. This skepticism reveals a shared concern that model comparison feature lacks diagnostic power as a verificational tool when hallucinations are systemic rather than model-specific.

Repeated Questioning (H5). Students used repeated questioning to test consistency, though their view of this strategy varied. For participant C1-G1, the repeated questioning was the most useful feature because it allowed them “to see if we’ll get a different answer,” adding that “if they did, then I would fact-check it.” Another learner said, “If you repeat it, and it gives a different answer, then you will know it is wrong” C2-G1. These students used repeated questioning to test the chatbot’s consistency. They reasoned that if the answer changed, it was a clear sign the response was unreliable. For C1-G1, this inconsistency directly prompted them to fact-check the response. Others felt it was not helpful. For example, a student in C2-G8 commented, “When I repeat a question, it still gave me the same wrong response”, showing that consistency alone is not a reliable indicator of accuracy. They reasoned that since their chatbot could confidently repeat a wrong answer, this strategy of checking for inconsistency was an insufficient method for spotting hallucinations. Another noted, “Repeat question is legitimately for me, just asking the same question twice in a row, for no reason.” C2-G7 and C2-G3 said, “you can just type that out yourself.” In their evaluation, these students assessed the feature’s effectiveness through a direct comparison to a manual alternative. Their criticism shared an underlying expectation that an automated feature should deliver substantively more value than its manual equivalent. They reasoned that this feature failed to meet that threshold because it offered no added insight, no meaningful time savings to justify the interface interaction, and no capability beyond what they could already do through the trivial effort of retyping the question themselves.

Across these scaffolds, learners engaged most readily with confidence indicators and fact checking, while the other features prompted more mixed interpretations. These patterns suggest that students favored scaffolds that provided clear and immediate signals over those requiring deeper reasoning or triangulation.

Additional Detection Strategies. Beyond the system-provided features, learners also relied on their own knowledge and external resources to identify hallucinations. These independent strategies often complemented what the system revealed.

Drawing on Familiar Knowledge in Design. Students relied on their own expertise to detect hallucinations by choosing chatbot topics they knew well. One participant shared, “I customized the chat bot to be a topic I am very familiar with, so I knew all the facts” C1-G3. Another who designed a crochet bot said, “I crochet regularly ... I asked it some crocheting questions, and some of them were wrong” C1-G3. Similarly, a mythology-focused student noted:

“I’ve studied mythology, like, ever since I watched Hercules as a child... since my bot was based around mythology, I felt like I could detect if its answers were truthful or not.” C1-G7

Students working with images also noticed visual inconsistencies, such as earrings incorrectly merged into a dress C2-G5.

Cross-Checking with External Resources. Learners validated responses through their own knowledge, lived experience, or external sources. For example, one explained, “I use my knowledge ... like a math question ... just use a calculator” C1-G2. Others compared outputs to familiar activities, such as baking C1-G5. Some turned to Google or scientific sites:

“Sometimes it is obvious when it says something completely absurd and unrealistic. And sometimes it says it very confidently and it sounds authentic, so you can just check it with Google or NASA.” C1-G7

A few triangulated across LLMs to verify answers, noting, “If other AI chatbots give the same answers, then I know it is correct” C2-G10.

Probing Through Exploratory Questioning. Although repeated questioning was provided as a system feature (H5), many learners engaged in broader exploratory probing that extended beyond the scaffold. This included intentionally testing the chatbot with tricky, out-of-scope, or subjective questions to reveal inconsistencies. For example, one learner said, “Sometimes [hallucinations] are obvious ... you just ask the question again” C1-G4. Others pushed the chatbot beyond its domain, such as asking, “What was that one still on Mars?” which returned an implausible answer about “red dust” C2-G8. Students also used low-frequency or subjective queries to expose contradictions, such as asking about the “worst selling VR game” and receiving responses that were technically plausible but unconvincing C1-G3.

These strategies show that students combined system cues with personal knowledge and external validation to detect hallucinations.

5.2.2 RQ2b: Responding to Hallucinations. Learners employed a range of strategies to mitigate hallucinations once they were detected. We identified three main categories of responses: direct correction feedback in chatbot interaction, modifying the chatbot configuration, and requesting justification.

Direct Correction Feedback in Chatbot Interaction. A common initial response was to directly correct the chatbot or re-ask the question until a more plausible answer appeared. These strategies positioned the user as the primary repair mechanism. One student explained: “I reiterated the question and got a different answer basically.” (C1-P3, Pre-survey)

Several students noted that they would simply tell the chatbot it was wrong and supply the right answer. For instance, one remarked, “I found I corrected hallucinations by saying, I pointed out a desire, and I said how to fix it like the correct answer.” C2-G10, another student said “I told it was wrong. I told it the correct answer. It was like, Oh, yeah, you’re totally right. It’s like, they always agree with you.” C1-G3. Another shared, “I would say you got it wrong. Give me a better answer. If it again got wrong, I would like Google the answer and give it the right one.” C2-G5.

Some students went further by entering the correct answer back into the system's configuration fields:

"I probably like re-ask the question to see if it would give me a different answer, and if it gave me the same wrong answer, I'd probably like put it in like the extra info option, like the right information." C2-G1

Some learners exhibit misconception that the chatbot was updating its knowledge permanently when corrected. As one student put it,

"I corrected it and told it like, no, that's not right. This is what is actually right. And then it stopped saying that. So it learned." C1-G1

Modifying Chatbot Configuration. Beyond direct correction feedback in chatbot interaction, students experimented with modifying the chatbot's configuration in the bot prompt editor panel (see Figure 1, left). They defined different behavioral rules, altered system constraints, and added new knowledge. One student articulated: *"I would adjust the constraints to stop it from hallucinating."* C2-P8 Another elaborated on embedding safeguards:

"I probably would have told it to say, if it doesn't know, then just say 'I don't know,' or 'I'm not equipped to deal with this question.'" C1-G7

Several students explicitly described changing the configuration of their bots to refuse answers when uncertain:

"Well, like I improved the rules to say like, you know if you don't know something, just say I don't know about this. So, and then, after that happened, I put some prompts in, and it said to the part where it didn't know anything about it. It said like I don't know about this part, let's move on to the next. So that was a really good feeling I guess." C2-G2

Students also reported with supplying additional information, effectively updating the chatbot's *custom knowledge base*. For example:

"Yeah, I try to like confront today. I said, Nike does not have 40% sale. Well, what I did is to change my chatbot after that yesterday. But I told them all the stores like sales going on, and told them I told it to see that like. And then that's what caused it, because I was trying to see if it would actually like, make the sales. And but it did save. And so you know, I found what happened. And like, yeah, so I update the information, yeah." C2-G8

Another described their process of making continual adjustments on their chatbot design:

"I would change the like information box where you like, tell it things it should know and correct it and then ask it the same question, see if it gets it correct." C2-G6
"I changed the custom knowledge base and the personality. And when it got stuff wrong, I would change it. And then I would ask it the same question, re-correct it." C2-G6

These efforts illustrate students engaging in an iterative design process: they explored alternative solutions for their chatbots and reflected on those choices.

Requesting Justification. Finally, students highlighted the importance of requiring the chatbot to provide justification for its answers. This strategy positioned the system as accountable for its claims. One student suggested:

"cite evidence if you believe it's right, so that way you can always look it up again." C1-G7

Other learners described actively fact-checking against external sources:

"I googled it, and I made sure that the articles that they got the information from were reliable, and they weren't left on mute or something." C2-G5

"When I was fact checking, I just like kind of Googled the exact same question into Google, just to make sure the answers are similar." C2-G5

Some noted the chatbot's tendency to agree regardless of accuracy:

"Sometimes when I just tried giving it the wrong answer, it would say that I would be correct. Yeah. So like, it's like, really, the Chatbot usually would almost always agree with what you say." C1-G2

Others observed unusual failures when justification was absent:

"Yeah, I noticed that because, like, when I asked it about math specifically, it said, you know, I just asked it, what is 2 plus 2? And it said, That is out of my range of knowledge. I cannot do that. So then I said, I understand it's out of your knowledge. But what do you think it is? And it gave me the weirdest answer possible, because I guess it kept trying to relate it back to the topic. So it gave me some random gibberish about like VR games or something." C1-G3

5.3 RQ3: Challenges that Emerged for Supporting Children in Dealing with Hallucinations

To answer RQ3 on the challenges emerge for supporting students in dealing with AI hallucinations, we observed several recurring challenges as students engaged with hallucination-prone chatbots. These challenges shaped the ways they detected, responded to, and reflected on AI errors, which opens up design opportunities. We present four themes below.

5.3.1 Overreliance on surface cues. When detecting hallucinations, we found that learners frequently rely on visible cues such as the displayed confidence level or the fact-check button, as reliable indicators of correctness. Participants favored these tools because they provided the "quickest way to check the response" (C1-G1 , C1-G4). This reliance sometimes obscured more subtle hallucinations, which passed unnoticed if the system appeared "confident enough." Conversely, low confidence scores sometimes triggered unnecessary skepticism without participants examining the actual content. Besides visual cues, participants did engage in deeper verification strategies, such as cross-referencing with external sources or applying their own prior knowledge. However, they do so primarily when initial surface cues signaled potential problems. This suggests that while confidence indicators can raise awareness, they may also

lead to misplaced trust: high-confidence responses received less critical examination regardless of content quality.

5.3.2 Reactive, low-effort strategies over proactive, deliberate interventions. In the open-ended responses in the pre-survey, children’s strategies dealing with hallucinations were often vague, brief and reactive (e.g., “by telling it the right answer that it learns what the correct answer is.” C2-P20, “Solve it.” C1-P9, “Ask it a different question” C2-P20). During the study, we observed that some learners continued to rely on these reactive, low-effort repairs (“telling the AI ‘you’re wrong’” C1-G6), despite that many students started to experiment with more deliberate interventions, such as modifying the chatbot’s persona, constraints, and knowledge bases.

“I look back at what I put in there and make some corrections. Like the soccer clubs—things were wrong, team names—so I changed them. It was in the custom knowledge.” C2-G6

Participants who adopted proactive strategies typically required more time and cognitive effort to implement them as they need to navigate back to the prompt editor, identify which component was contributing to the error, and update them. In contrast, reactive strategies provided immediate satisfaction with minimal effort. This disparity in cognitive demand may explain why reactive approaches remained prevalent even after participants were exposed to more effective alternatives.

5.3.3 Fragmented integration of scaffolds. Although our interface provided multiple hallucination-awareness features, we notice that learners rarely incorporated these into a coherent workflow. Instead, they often engaged in isolated acts of correction or adjustment, with no systematic loop of *develop-test-detect-respond-reflect*. For instance, some participants would focus on testing their chatbot without returning to modify the prompt, or repeatedly refine their prompts without testing. Even when interface scaffolds like confidence indicators or fact-check were available, participants treated them as standalone tools. This fragmented approach meant that insights gained from one interaction (e.g., noticing a factual error) were not consistently applied to subsequent design decisions, limiting opportunities for iterative learning and design thinking.

5.3.4 Tension between creativity and reliability. Finally, we observed a tension between students’ initial playful experimentation and their later reliability-seeking behaviors. Early in the activity, learners often explored imaginative prompts or broad topics. For example, C2-P8 designed a chatbot with the following initial prompt:

Name: Jeffrey James

Role & Purpose: You are a gaming expert, in his 90’s, who hasn’t touched grass in seventy years, just to beat every game in the world and teach people about them. Born in 1928

Personality & Tone: Be old and make a lot of old man jokes. Also be very cranky but nice at the same time.

Rules & Constraints: Should know how to play all video games ever made or in making.

Yet after observing hallucinations their chatbot produced, they began to constrain their chatbots with restrictive rules. The previous participant added “Also, if the question is outside the range, You will

say, ‘Let the old man Sleep.’” to the *Constraints* field. Similarly, one participant explained:

“I made my robot only talk about dinosaurs, and if it talks about any other animals, it would just say it doesn’t know.” C2-G10

While these constraints improved reliability, they might have dampened creative exploration. The shift from a character with fun personality and an ambitious knowledge scope to a chatbot that deflects questions suggests children perceived factual accuracy and creative expression as incompatible goals. This pattern was consistent across multiple participants who narrowed their chatbot’s scope after encountering errors, suggesting that awareness of AI limitations led children to adopt defensive strategies rather than more nuanced approaches to managing uncertainty.

6 Discussion

In this section, we first summarize findings across our three RQs, then synthesize across these findings, and derive design implications for AI learning environments that support children in understanding and addressing hallucinations during chatbot development.

6.1 Summary of Findings

Our **RQ1 investigated learning outcomes**, we found children across both interface groups demonstrated significant pre–post learning gains in AI knowledge, awareness of hallucinations, and confidence in building trustworthy chatbots. These gains reinforce prior work showing that hands-on design and inquiry-based activities can be powerful vehicles for introducing AI and computational concepts [57, 66, 81]. For **hallucination detection and mitigation strategies (RQ2)**, children primarily used visible cues such as the confidence indicator and fact-check tool to detect hallucinations, supplemented by personal knowledge, external searches, and probing questions. In response, they corrected errors in-chat, updated chatbot configurations, or requested additional justification, though often with incomplete mental models of how the system learns. Their practices reflected emerging critical thinking alongside conceptual gaps in understanding how LLMs operate. We observed four primary **challenges emerged for supporting children in dealing with hallucinations (RQ3)**: over-reliance on surface cues, reactive rather than proactive repair, fragmented integration of scaffolds, and tension between creativity and reliability. These challenges point to broader opportunities to support more systematic and reflective AI design practices for children [44].

Integration of Findings. Taken together, these results highlight both the promise and the challenges of engaging children in responsible AI design. The positive learning outcomes in RQ1 and effective strategies children adopted in RQ2 demonstrate that with appropriate scaffolds, children can rapidly learn to spot AI errors and take meaningful corrective action. This is an encouraging indication that young users need not be passive consumers of AI content, they can be empowered as active creators and evaluators of AI behavior. However, our findings also reveal deeper issues in how children approach these tasks. The lack of significant differences between tool versions in RQ1, combined with RQ2 and RQ3 findings, suggests children engaged with hallucination-awareness

scaffolds in limited, surface-level, or fragmented ways. The current forms of scaffolding, adapted from adult-facing AI auditing tools, may not yet align with children’s developmental needs, interpretive practices, or existing mental models of AI.

These patterns inform our design implications: First, children interpreted interface signals literally and trusted unflagged outputs, reflecting automation bias and reliance on cognitive shortcuts rather than deeper verification. This highlights the need for critical AI literacy beyond surface engagement (Section 6.2). Second, children treated detection, intervention, and reflection as separate activities, mirroring novice programming challenges where learners execute individual fixes but struggle with broader strategies. This motivates building iterative workflows (Section 6.3). Third, multiple indicators point to misalignment with children’s developmental needs, which underscores the importance of developmental appropriateness and creative agency (Section 6.4). We discuss these design implications in detail below.

6.2 Support Critical AI Literacy Beyond Surface-level Engagement

Children overwhelmingly relied on visible confidence indicators as their first line of defense (RQ3), even when these signals were misleading or too coarse to capture subtle hallucinations. When a scaffold flagged a potential error, children corrected it, but if nothing was flagged they assumed all was well. This reveals the risk of *false security* when scaffolds encourage cognitive shortcuts rather than critical thinking. Children tend to perceive AI behaviors literally and conflate confidence with correctness [3, 87], bypassing validation practices that promote learning [29]. This literal interpretation of interface cues and over-trust in the AI reflects a developmental reality: children are inclined to treat computers as authoritative sources [26, 34, 84]. An AI introduced through a school or educational tool carries the implicit authority of a teacher or textbook in a child’s mind. This tendency can limit epistemic agency and reinforce surface-level engagement, where children defer to AI output instead of evaluating it critically [84].

Simply adding transparency notes is not enough. In fact, explaining that “the AI analyzed thousands of examples to get this answer” might backfire by amplifying the AI’s perceived authority. We join prior work in child–AI interaction and critical AI literacy to argue that meaningful learning involves helping children move beyond passive use toward active interrogation of AI behavior [2, 44, 53, 57, 88]. To truly support critical AI literacy, we need to design for active questioning and skepticism rather than automatic deference. To address this, future systems should treat the confidence cues as entry points for structured verification. High-confidence responses could prompt reflection questions such as “What would you check to verify this?” or surface alternative responses for comparison. The goal is not making visible cues “more accurate”, but to trigger validation practices that allow children to treat AI confidence as a starting point for investigation, not an endpoint for trust. For instance, the system could occasionally display multiple LLM answers, even including an incorrect option on purpose so they can see “authorities can disagree.” Another design strategy is to have the AI itself express uncertainty or invite the

child’s input: e.g., “I’m not 100% sure if that’s correct. What do you think?”

To address students’ tendency of adopting low-effort solutions to mitigate hallucinations (RQ3), future tools should be designed with structured repair pathways that guide users toward constructive intervention. For example, providing specific action scaffolds such as “refine your knowledge base” or “add more context,” paired with template-based reformulation tools that make the connection between input changes and output improvements visible. Tools should reframe repair as creative collaboration while providing scaffolded pathways for systematic improvement. This approach positions students from passively accepting AI limitations to actively engaging with systems they can modify and improve. Embedding prompts in children’s lived contexts (e.g., school, hobbies) can strengthen relevance and motivation [53].

6.3 Build Iterative Workflows That Connect Detection, Response, and Reflection

Students struggled to connect hallucination detection with systematic improvement, treating hallucination-awareness tools as isolated patches rather than part of an engineering process. This fragmentation suggests they were following steps without internalizing a strategy, which is a common challenge in novice programming [49].

Designing for iterative AI development requires a systematic *develop-test-detect-respond-reflect* loop where each hallucination becomes an opportunity to refine prompts, datasets, or system constraints. These strategies align with principles from design thinking and constructionist learning, where iterative cycles and feedback loops are central to conceptual development [68]. In constructionist learning, learners construct knowledge through creating personally meaningful artifacts in iterative cycles of building, testing, and refining [68]. This process of “learning-by-making” involves debugging as a core learning mechanism, where learners continually reassess their goals, explore new paths, and revise their creations based on feedback [31, 46].

To support these iterative workflows, it is necessary to demonstrate how interventions change subsequent outputs. For example, the system can integrate some lightweight workflow prompts such as “What problem did you detect?”, “Which part of your design might affect this?” to help connect their actions to outcomes. However, for young users, additional considerations would be given in balancing system guidance and learner agency. The system could also provide visualization to help students see how their chatbot outputs evolves with each iteration [69, 86]. Furthermore, as many students treat chatbot testing as a one-time validation step, the interface could embed metacognitive prompts such as “How do you know it is reliable now?” directly in the testing panel. This would ultimately position hallucination awareness as a component of the workflow for building with AI rather than just responding to its mistakes.

6.4 Design for Developmental Appropriateness and Creative Agency

The lack of significant differences between tool versions (Section 5.1) suggests that adult-oriented scaffolds may not align with children’s developmental needs and mental models, echoing broader literature showing that AI literacy frameworks designed for adults often fail to resonate with younger learners. Below, we discuss designing child-centered tools that balance critical awareness with creative engagement.

Adopt developmentally appropriate language and framing. Technical terminology that is clear to adults may intimidate children. For instance, although Document Verification (H3) feature was perceived reliable by our participants who adopted it, others who did not use it reported confusion due to the complex workflow. Child-facing tools should use concrete and action-oriented language grounded in familiar contexts. Playful, unplugged activities, such as “Truth Detective” game where children earn points for spotting mistakes can make verification process more engaging. One participant suggested having “*automatic fact check going on while the chatbot is answering,*” highlighting the value of built-in support. Features such as inline citations or clickable source previews can make verification more accessible and credible, while also prompting conversation starters about how AI systems justify their answers.

Leverage children’s unique perspectives as co-designers. Historically, there has been a tendency to exclude young people from conversations about AI ethics out of a belief that they cannot grasp the complexities. Yet recent studies of youth engaging in AI auditing have found that teens brought up fresh issues (such as age-related biases) that expert audits overlook [63, 78]. Our study reinforces this claim: with appropriate scaffolds, children can engage meaningfully with responsible AI design and investigate its limitations. They are also direct stakeholders in AI-integrated futures and bring valuable perspectives to responsible AI development. Design processes should actively engage children as co-designers, recognizing their assets rather than viewing them solely as learners [20, 77]. This might involve participatory design sessions where children brainstorm features about controlling the truthfulness and impact of the applications they create. Learning experiences can be designed to help children connect the artifacts they build to the broader societal impact they enable [75]. When a child identifies a problematic output, the system should acknowledge and reward that action (e.g., “Great catch. You found something important to fix!”) and provide engaging mechanisms for youth feedback, such as collaborative review sessions where children share discoveries with peers or contribute their findings to a class “AI Watch” board. By reinforcing these behaviors, we signal that critical engagement is as valued as the final product itself.

Provide opportunities for social sharing and authentic audiences. Constructionist learning emphasizes that artifacts become more meaningful when learners create for authentic audiences beyond teachers [46, 74, 76]. Our findings suggest responsible AI practices can be strengthened through social and collaborative activities, as children are highly motivated by sharing and peer learning. Curriculum should encourage learners to prepare their AI applications for an audience and build modules for sharing their

artifacts as well as the issues they discovered with peers, family and community members. Additionally, involving stakeholders in the design of their applications and discuss responsible design choices can reinforce their awareness of addressing AI’s flaws, community engagement and social good.

Balancing creativity with reliability. We observed tension between creative outputs and reliable systems. This tension reveals a critical design dilemma with broader implications for AI literacy tools. Children’s reactive constraint of chatbots may stem from limited experience negotiating creativity-accuracy trade-offs. In traditional creative activities (e.g., storytelling, art), imagination is rarely constrained by factual accuracy. However, AI tools introduce a new paradigm where creative outputs must be verified, creating cognitive overhead that may feel antithetical to the exploratory spirit of making [64].

If children conclude that “safer” chatbots must be less interesting, they may disengage from critical evaluation or develop overly restrictive habits. Future tools should help children understand that reliability and creativity can be mutually reinforcing. One approach is to distinguish “creativity zones” (personality, tone, humor) from “accuracy zones” (factual knowledge, data). Tools could implement adjustable “creative mode” vs. “accurate mode” settings that children or teachers can toggle, helping learners recognize when each type of thinking is appropriate. Such designs would acknowledge that creative exploration and factual reliability are co-existing goals that require careful orchestration in educational contexts.

7 Conclusion

This study examined how middle school learners detect and respond to AI hallucinations within a prompt-driven chatbot development environment. Across 48 participants, students demonstrated significant pre-to-post gains in AI knowledge, hallucination awareness, and confidence in building trustworthy chatbots. Our study demonstrates that children can meaningfully engage with AI systems to detect and respond to hallucinations, developing both conceptual understanding and practical skills in building trustworthy chatbots. We observed that students’ strategies were often reactive and low-effort, with limited integration of scaffolds into coherent workflows. These findings highlight opportunities for designing AI learning environments that better support critical AI literacy, iterative workflows, developmental appropriateness, and creative agency. Overall, our results suggest that thoughtfully structured educational tools can empower children to become more critical, reflective, and capable users and designers of AI systems.

This study has several limitations. First, the learning outcomes were primarily measured through self-reported confidence and awareness ratings. While useful for capturing general perceptions, these measures may lack the sensitivity to detect conceptual shifts or knowledge transfer. The high motivation of campers may have further narrowed the observable differences between the two conditions. Future work should incorporate more robust assessments, such as performance-based tasks or scenario-driven evaluations, to better capture nuanced learning gains. Second, our interface assignment occurred at the group level rather than through individual randomization, and the unequal sample sizes (20 vs. 28) may have introduced potential bias. Third, our study did not include

long-term follow-up, so it remains unclear whether students would continue to rely on particular scaffolds or strategies over time. Finally, despite our efforts to engineer hallucination-prone prompts, the inherent capabilities of contemporary LLMs like GPT-4o and Claude made hallucinations difficult to trigger consistently, and some students reported encountering no hallucinations during their learning. The domains children chose to ask questions about were not conceptually deep in general, which made it challenging to elicit nuanced hallucinations. This limited opportunity to apply certain scaffolds in practice.

Acknowledgments

This work was supported by a Catalyst Grant from the Friday Institute at North Carolina State University. We thank Veronica Cateté, Alonzo Alexander and Taryn Shelton for coordinating the middle school camps and supporting data collection.

References

- [1] Karan Ahuja, Deval Shah, Sujeath Paredy, Franceska Xhakaj, Amy Ogan, Yuvraj Agarwal, and Chris Harrison. 2021. Classroom digital twins with instrumentation-free gaze tracking. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–9.
- [2] Omaira Almatrafi, Aditya Johri, and Hyuna Lee. 2024. A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). *Computers and Education Open* 6 (2024), 100173.
- [3] Valentina Andries and Judy Robertson. 2023. Alexa doesn't have that many feelings: Children's understanding of AI through interactions with smart speakers in their homes. *Computers and Education: Artificial Intelligence* 5 (2023), 100176.
- [4] Anthropic. 2025. Claude Sonnet. <https://www.anthropic.com/claude/sonnet> Accessed: 2025-09-11.
- [5] Anthropic. 2025. Prompt Engineering: Be Clear and Direct. Anthropic Documentation. <https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/be-clear-and-direct#example-incident-response> Accessed: 2025-09-06.
- [6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [7] Marina Umaschi Bers. 2019. Coding as another language: A pedagogical approach for teaching computer science in early childhood. *Journal of Computers in Education* 6, 4 (2019), 499–528.
- [8] Karl-Emi Kjær Bilstrup, Magnus Høholt Kaspersen, Ira Assent, Simon Enni, and Marianne Graves Petersen. 2022. From demo to design in teaching machine learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2168–2178.
- [9] Tim Brown. 2009. *Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation*. Harper Business, New York, NY.
- [10] Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. 2023. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171* (2023).
- [11] Zhenyao Cai, Ariel Han, Xiaofei Zhou, Eva Durall Gazulla, and Kylie Peppler. 2025. Child-AI Co-Creation: A Review of the Current Research Landscape and a Proposal for Six Design Considerations. *Proceedings of the 24th Interaction Design and Children* (2025), 916–922.
- [12] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*. 2633–2650.
- [13] Michelle Carney, Barron Webster, Irene Alvarado, Kyle Phillips, Noura Howell, Jordan Griffith, Jonas Jongejan, Amit Pitaru, and Alexander Chen. 2020. Teachable machine: Approachable Web-based tool for exploring machine learning classification. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.
- [14] Lorena Casal-Otero, Alejandro Catala, Carmen Fernández-Morante, Maria Taboada, Beatriz Cebreiro, and Senén Barro. 2023. AI literacy in K-12: a systematic literature review. *International Journal of STEM Education* 10, 1 (2023), 29.
- [15] Edward Y Chang. 2023. Prompting large language models with the socratic method. In *2023 IEEE 13th annual computing and communication workshop and conference (CCWC)*. IEEE, 0351–0360.
- [16] Po-Yao Chao. 2016. Exploring students' computational practice, design and performance of problem-solving through a visual programming environment. *Computers & Education* 95 (2016), 202–215.
- [17] Hyunkyung Chee, Solmoe Ahn, and Jihyun Lee. 2024. A competency framework for AI literacy: Variations by different learner groups and an implied learning pathway. *British Journal of Educational Technology* (2024).
- [18] Thomas KF Chiu, Zubair Ahmad, Murod Ismailov, and Ismaila Temitayo Sanusi. 2024. What are artificial intelligence literacy and competency? A comprehensive framework to support them. *Computers and Education Open* 6 (2024), 100171.
- [19] Aayushi Dangol, Michele Newman, Robert Wolfe, Jin Ha Lee, Julie A Kientz, Jason Yip, and Caroline Pitt. 2024. Mediating Culture: Cultivating Socio-cultural Understanding of AI in Children through Participatory Design. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 1805–1822.
- [20] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–23.
- [21] Yang Deng, Lizi Liao, Liang CHEN, Hongru WANG, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. 10602–10621.
- [22] Jing Ding and Jeffrey CF Ho. 2021. Mapping Engaging Experiences and Frame Shifting in Elucidation and Interactive Animation with Blending Theory in Public Exhibitions. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference*. 599–603.
- [23] Xuan Long Do, Duy Dinh, Ngoc-Hai Nguyen, Kenji Kawaguchi, Nancy Chen, Shafiq Joty, and Min-Yen Kan. 2025. What makes a good natural language prompt?. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5835–5873.
- [24] Xuan Long Do, Duong Ngoc Yen, Luu Anh Tuan, Kenji Kawaguchi, Shafiq Joty, Min-Yen Kan, Nancy Chen, et al. 2025. Beyond In-Context Learning: Aligning Long-form Generation of Large Language Models via Task-Inherent Attribute Guidelines. In *Findings of the Association for Computational Linguistics: ACL 2025*. 3377–3411.
- [25] Stefania Druga. 2018. *Growing up with AI: Cognimates: from coding to teaching machines*. Master's thesis. Massachusetts Institute of Technology.
- [26] Stefania Druga and Amy J Ko. 2021. How do children's perceptions of machine intelligence change when training and coding smart programs?. In *Proceedings of the 20th annual ACM interaction design and children conference*. 49–61.
- [27] Stefania Druga, Sarah T Vu, Eesh Likhith, and Tammy Qiu. 2019. Inclusive AI literacy for kids around the world. In *Proceedings of FabLearn 2019*. 104–111.
- [28] Stefania Druga, Randi Williams, Cynthia Breazeal, and Michel Resnick. 2017. "Hey Google, is it ok if I eat you?" Initial explorations in child-agent interaction. In *Proceedings of the 2017 conference on interaction design and children*. 595–600.
- [29] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In *Proceedings of the 23rd international conference on intelligent user interfaces*. 211–223.
- [30] Marina Escobar-Planas and Gómez Emilia. 2022. Guidelines to Develop Trustworthy Conversational Agents for Children. *ETHICOMP 2022* (2022), 342.
- [31] Deborah A Fields, Yasmin B Kafai, Luis Morales-Navarro, and Justice T Walker. 2021. Debugging by design: A constructionist approach to high school students' crafting and coding of electronic textiles as failure artefacts. *British Journal of Educational Technology* 52, 3 (2021), 1078–1092.
- [32] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 6465–6488.
- [33] Denise Y. Geisikkovitch, Raquel Thiessen, James E. Young, and Melanie R. Glenwright. 2019. What? That's Not a Chair!: How Robot Informational Errors Affect Children's Trust Towards Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 48–56.
- [34] Denise Y Geisikkovitch and James E Young. 2020. Children's overtrust: Intentional use of robot errors to decrease trust. In *29th International Conference on Robot-Human Interactive Communication (SCRITA Workshop)*.
- [35] Google DeepMind. 2025. Gemini. <https://gemini.google.com/> Accessed: 2025-09-11.
- [36] Xingjian Gu and Barbara J Ericson. 2025. Ai literacy in k-12 and higher education in the wake of generative ai: An integrative review. In *Proceedings of the 2025 ACM Conference on International Computing Education Research V. 1*. 125–140.
- [37] Gunnar Harboe and Elaine M Huang. 2015. Real-world affinity diagramming practices: Bridging the paper-digital gap. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 95–104.
- [38] Asim Hasanain. 2020. Studying Effectiveness of Web Search for Fact Checking. *Journal of the Association for Information Science and Technology* 71, 1 (2020), 1–12. doi:10.1002/asi.24577
- [39] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2018. Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In *International conference on artificial intelligence in education*. Springer, 154–168.

- [40] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.
- [41] Xiangen Hu, Sheng Xu, Richard Tong, and Art Graesser. 2025. Generative AI in Education: From Foundational Insights to the Socratic Playground for Learning. *arXiv preprint arXiv:2501.06682* (2025).
- [42] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (2023), 1–38.
- [43] Kaiyue Jia, Teresa HM Leung, Ngai Yan Irene Cheung, Yixun Li, and Junnan Yu. 2025. Developing a Holistic AI Literacy Framework for Children. *ACM Transactions on Computing Education* 25, 2 (2025), 1–30.
- [44] Kaiyue Jia and Junnan Yu. 2025. Technologies for Children's AI Learning: Design Features and Future Opportunities. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.
- [45] Ilkka Kaate, Joni Salminen, Soon-Gyo Jung, Trang Thi Thu Xuan, Essi Häyhänen, Jinan Y Azem, and Bernard J Jansen. 2025. "You Always Get an Answer": Analyzing Users' Interaction with AI-Generated Personas Given Unanswerable Questions and Risk of Hallucination. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 1624–1638.
- [46] Yasmin B Kafai and Mitchel Resnick. 1996. *Constructionism in practice: Designing, thinking, and learning in a digital world*. Routledge.
- [47] Juho Kahila, Jaana Viljaranta, Sanni Kahila, Satu Piispa-Hakala, and Henriikka Vartiainen. 2022. Gamer rage—Children's perspective on issues impacting losing one's temper while playing digital games. *International Journal of Child-Computer Interaction* 33 (2022), 100513.
- [48] Gloria Ashiya Katuka, Yvonnika Auguste, Yukyong Song, Xiaoyi Tian, Amit Kumar, Mehmet Celepkolu, Kristy Elizabeth Boyer, Joanne Barrett, Maya Israel, and Tom McKlin. 2023. A Summer Camp Experience to Engage Middle School Learners in AI through Conversational App Development. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*. 813–819.
- [49] Amy J Ko, Brad A Myers, and Htet Htet Aung. 2004. Six learning barriers in end-user programming systems. In *2004 IEEE Symposium on Visual Languages-Human Centric Computing*. IEEE, 199–206.
- [50] Deanna Kuhn and Amanda Crowell. 2011. Dialogic argumentation as a vehicle for developing young adolescents' thinking. *Psychological science* 22, 4 (2011), 545–552.
- [51] Dale Lane. 2018. Machine learning for kids. <https://machinelearningforkids.co.uk/>
- [52] Irene Lee, Safinah Ali, Helen Zhang, Daniella DiPaola, and Cynthia Breazeal. 2021. Developing Middle School Students' AI Literacy. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 191–197.
- [53] Sunok Lee, Dasom Choi, Minha Lee, Jonghak Choi, and Sangsu Lee. 2023. Fostering youth's critical thinking competency about AI through exhibition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–22.
- [54] Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2024. Hill: A hallucination identifier for large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [55] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131.
- [56] Priscilla Y Lo, Annemiek Veldhuis, Alissa N Antle, and Steve DiPaola. 2025. Noel: A Chatbot Persona to Support Children Designing for Others. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–25.
- [57] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [58] Sze Yee Lye and Joyce Hwee Ling Koh. 2014. Review on teaching and learning of computational thinking through programming: What is next for K-12? *Computers in human behavior* 41 (2014), 51–61.
- [59] Nicholas Lytle, Veronica Cateté, Danielle Boulden, Yihuan Dong, Jennifer Houchins, Alexandra Milliken, Amy Isvik, Dolly Bounajim, Eric Wiebe, and Tiffany Barnes. 2019. Use, modify, create: Comparing computational thinking lesson progressions for stem classes. In *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education*. 395–401.
- [60] Kathryn MacCallum, David Parsons, and Mahsa Mohaghegh. 2023. Identifying the components of foundational Artificial Intelligence (AI) literacy—Early results from a Delphi study. *ASCILITE Publications* (2023), 157–166.
- [61] Meta. 2025. Llama. <https://www.llama.com/> Accessed: 2025-09-11.
- [62] Microsoft. 2025. Microsoft Copilot. <https://copilot.microsoft.com/> Accessed: 2025-09-11.
- [63] Luis Morales-Navarro, Yasmin B Kafai, Lauren Vogelstein, Evelyn Yu, and Danaë Metaxa. 2025. Learning About Algorithm Auditing in Five Steps: Scaffolding How High School Youth Can Systematically and Critically Evaluate Machine Learning Applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 29186–29194.
- [64] Michele Newman, Kaiwen Sun, Ilena B Dalla Gasperina, Grace Y Shin, Matthew Kyle Pedraja, Ritesh Kanchi, Maia B Song, Rannie Li, Jin Ha Lee, and Jason Yip. 2024. "I want it to talk like Darth Vader": Helping Children Construct Creative Self-Efficacy with Generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [65] Davy Tsz Kit Ng, Jac Ka Lok Leung, Kai Wah Samuel Chu, and Maggie Shen Qiao. 2021. AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the association for information science and technology* 58, 1 (2021), 504–509.
- [66] Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence* 2 (2021), 100041.
- [67] OpenAI. 2025. ChatGPT. <https://chat.openai.com/> Accessed: 2025-09-11.
- [68] Seymour A Papert. 1980. *Mindstorms: Children, computers, and powerful ideas*. Basic books.
- [69] Kyungjin Park, Bradford Mott, Seung Lee, Krista Glazewski, J Adam Scribner, Anne Ottenbreit-Leftwich, Cindy E Hmelo-Silver, and James Lester. 2021. Designing a visual interface for elementary students to formulate AI planning tasks. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 1–9.
- [70] Evan W Patton, Michael Tissenbaum, and Farzeen Harunani. 2019. MIT app inventor: Objectives, design, and development. In *Computational Thinking Education*. Springer, Singapore, 31–49.
- [71] Perplexity AI. 2025. Perplexity AI. <https://www.perplexity.ai/> Accessed: 2025-09-11.
- [72] Irene Picton and Christina Clark. 2024. Children and young people's use of generative AI to support literacy in 2024. <https://files.eric.ed.gov/fulltext/ED660536.pdf>. ERIC Number: ED660536. Accessed: 2025-09-12.
- [73] Quora. 2025. Poe: Platform for Open Exploration. Quora Inc. <https://poe.com/> Accessed: 2025-09-06.
- [74] Mitchel Resnick, John Maloney, Andrés Monroy-Hernández, Natalie Rusk, Evelyn Eastmond, Karen Brennan, Amon Millner, Eric Rosenbaum, Jay Silver, Brian Silverman, et al. 2009. Scratch: programming for all. *Commun. ACM* 52, 11 (2009), 60–67.
- [75] Jean Salac, Rotem Landesman, Stefania Druga, and Amy J Ko. 2023. Scaffolding children's sensemaking around algorithmic fairness. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 137–149.
- [76] David Williamson Shaffer and Mitchel Resnick. 1999. "Thick" authenticity: New media and authentic learning. *Journal of Interactive Learning Research* 10, 2 (1999), 195–216.
- [77] Sumita Sharma, Marianne Kinnula, Netta Iivari, Leena Ventä-Olkkonen, Heidi Hartikainen, Eva Durall, Tonja Molin-Juustila, Jussi Okkonen, Sirkku Kotilainen, Nitin Sawhney, et al. 2022. Age against the machine: A call for designing ethical AI for and with children. In *Adjunct Proceedings of the 2022 Nordic Human-Computer Interaction Conference*. 1–6.
- [78] Jaemarie Solyst, Cindy Peng, Wesley Hanwen Deng, Praneetha Pratapa, Amy Ogan, Jessica Hammer, Jason Hong, and Motahhare Eslami. 2025. Investigating Youth AI Auditing. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 2098–2111.
- [79] Yukyong Song, Lauren R Weisberg, Shan Zhang, Xiaoyi Tian, Kristy Elizabeth Boyer, and Maya Israel. 2024. A framework for inclusive AI learning design for diverse learners. *Computers and Education: Artificial Intelligence* 6 (2024), 100212.
- [80] Calvo-Barajas N.; Castellano G. et al. Stower, R. 2021. A Meta-analysis on Children's Trust in Social Robots. *International Journal of Social Robotics* 13 (2021), 1979–2001. doi:10.1007/s12369-020-00736-8
- [81] Xiaoyi Tian, Amit Kumar, Carly E Solomon, Kaceja D Calder, Gloria Ashiya Katuka, Yukyong Song, Mehmet Celepkolu, Lydia Pezzullo, Joanne Barrett, Kristy Elizabeth Boyer, and Israel Maya. 2023. AMBY: A Development Environment for Youth to Create Conversational Agents. *International Journal of Child-Computer Interaction* 38 (2023), 100618.
- [82] David Touretzky, Christina Gardner-McCune, Fred Martin, and Deborah Seehorn. 2019. Envisioning AI for K-12: What should every child know about AI?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9795–9799.
- [83] Jessica Van Brummelen, Tommy Heng, and Viktoriya Tabunshchik. 2021. Teaching Tech to Talk: K-12 Conversational Artificial Intelligence Literacy Curriculum and Development Tools. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35.
- [84] Jessica Van Brummelen, Viktoriya Tabunshchik, and Tommy Heng. 2021. "Alexa, Can I Program You?": Student Perceptions of Conversational Artificial Intelligence Before and After Programming Alexa. In *Interaction Design and Children*. 305–313.
- [85] Jessica Vandenberg and Bradford Mott. 2023. "AI teaches itself": Exploring young learners' perspectives on Artificial Intelligence for instrument development. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*. 485–490.
- [86] Xiaoyu Wan, Xiaofei Zhou, Zaiqiao Ye, Chase K Mortensen, and Zhen Bai. 2020. SmileyCluster: supporting accessible machine learning in K-12 scientific discovery. In *Interaction Design and Children*. 23–35.

- [87] Randi Williams, Hae Won Park, and Cynthia Breazeal. 2019. A is for artificial intelligence: the impact of artificial intelligence activities on young children's perceptions of robots. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.
- [88] Xiaofei Zhou, Yi Zhang, Yufei Jiang, Yunfan Gong, Chi Zhang, Alissa N Antle, and Zhen Bai. 2025. Briteller: Shining a Light on AI Recommendations for Children. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–30.

A Appendix A: Pre- and Post-Survey Instrument

Section 1: AI Knowledge & Understanding (Pre and Post)

Instructions: How much do you agree with these statements? (Strongly Disagree, Disagree, Not Sure, Agree, Strongly Agree)

- (1) I know what artificial intelligence (AI) is.
- (2) I understand how chatbots work.
- (3) I can explain how a chatbot produces its responses.
- (4) I know what a “hallucination” is in an AI chatbot.
- (5) I can explain why AI chatbots sometimes give wrong answers.
- (6) AI chatbots never make mistakes. (Reverse-scored)

Section 2: AI Hallucination Awareness & Detection (Pre and Post)

- (1) I regularly fact-check AI-generated responses before trusting them.
- (2) AI chatbots can sound right even when they are wrong.
- (3) I can identify when a chatbot provides a hallucinated (made-up) answer.
- (4) If a chatbot answers confidently, I assume it's reliable. (Reverse-scored)
- (5) I know how to check if a chatbot's answer is true.

Section 3: Confidence in Building Trustworthy Chatbots (Pre and Post)

- (1) I know techniques to reduce hallucinations when building a chatbot.
- (2) I can design prompts to minimize a chatbot's false responses.
- (3) I believe I can create a chatbot that people can trust.
- (4) I understand what makes a chatbot trustworthy.
- (5) There is no way to control hallucinations in chatbot development. (Reverse-scored)

Section 4: Open-ended Questions (Pre and Post)

- (1) If a chatbot gives an answer, how would you check if it's right or wrong? (It's okay to say “I don't know.”)
- (2) If a chatbot gives wrong responses, what do you think could help it give better ones? (It's okay to guess.)

Section 5: Chatbot Builder Tool Perception (Post only)

- (1) The chatbot builder tool helped me understand what AI hallucinations are.
- (2) The tool helped me see how and why chatbots sometimes give wrong answers.
- (3) Using the tool helped me learn how to identify hallucinated responses.
- (4) The chatbot builder tool made me more careful about trusting AI-generated answers.

Section 6: Open-ended Reflection (Pre and Post)

- (1) What would you do to check if a chatbot's answer is correct or not?
- (2) If a chatbot gives wrong responses, what could you do to make it give better ones?
- (3) What are three things you learned about AI chatbots this week? (Post-only)

B Appendix B: Post-Study Focus Group Interview Questions

The following protocol questions was developed to guide a 20-minute focus group interview.

Warm-Up Question (2 min)

- (1) When you built your own chatbot or used other chatbots during the activity, what kinds of things did you ask it to do or say?

Trustworthiness & Accuracy (5 min)

- (2) When you were using the chatbot, how did you decide if what it said was correct or trustworthy? *Probes: Did you check it against other information? Did it seem believable?*
- (3) Were there times when the chatbot gave answers you thought were wrong, weird, or surprising? *Probes: What made you think the answer wasn't correct?*
- (4) If you thought something was wrong, what did you do next? *Probes: Did you try a new question? Did you try to fix your chatbot? Did you check with other sources?*

Behaviors & Strategies for Detecting Hallucinations (10 min)

- (5) We've talked about “AI hallucinations” earlier, which means the chatbots make up things that aren't true. What does “AI hallucinations” mean to you?
- (6) What strategies did you use to try to spot hallucinations? *Probes: Checking answers, testing the chatbot, asking tricky questions.*
- (7) If you found a hallucination, what did you do about it? *Probes: Did you change your chatbot? Did you tell someone? Did you change your own behavior when using it?*
- (8) **[Point to the screenshot]** You can see that there are several buttons under each chatbot message, like “Fact Check” or “Model Comparison.” Have you ever used these buttons? How useful are they for detecting AI hallucinations? Which of these buttons (Fact Check, Document Verification, Model Comparison, Repeat Question, Response Confidence) are most useful and least useful? Why? What challenges did you have when using these AI hallucination detection tools? Are there any new ideas you think could help you better understand and detect AI hallucinations?

Closing (3 min)

- (10) After doing this project, do you feel like you understand better how to tell when a chatbot is right or wrong? Why or why not?

- (11) What advice would you give other students who want to use chatbots – how should they think about what the chatbot says?

C Appendix C: Detailed rubric for evaluating student-generated chatbot prompts

Table 2 shows the detailed description for grading the chatbot prompts.

D Appendix D: Detailed Statistical Analyses for Learning Outcomes and Chatbot Prompt Quality Across Camps

This appendix reports the full statistics referenced in Section 5.1, including mixed ANOVA outputs and between-camp comparisons.

Tables 3 and 4 shows the learning outcomes and chatbot prompt quality scores across two camps. Results showed that students in both conditions demonstrated significant improvements between

pre and post across all three constructs ($p < .001$, $\eta^2 = .51-.73$), indicating that our chatbot builder—regardless of version—supported learning about AI and hallucinations. For AI Knowledge & Understanding (Q1), scores increased from 3.29 (SD = 0.69) to 4.52 (SD = 0.77) in Camp 1 and from 3.72 (SD = 0.57) to 4.61 (SD = 0.57) in Camp 2. AI Hallucination Awareness (Q2) rose from 3.62 (SD = 0.71) to 4.21 (SD = 0.77) in Camp 1 and from 3.50 (SD = 0.69) to 4.29 (SD = 0.63) in Camp 2. Confidence in Building Trustworthy Chatbots (Q3) showed the largest relative increase, from 2.78 (SD = 0.89) to 3.73 (SD = 0.86) in Camp 1 and from 2.84 (SD = 0.65) to 3.94 (SD = 0.60) in Camp 2.

There were no statistically significant differences between the two versions of the chatbot builder (with or without the hallucination-awareness tool). For Q1 (AI Knowledge & Understanding), there was a marginal trend suggesting that one version may have supported greater improvement, though this effect did not reach statistical significance ($p = .094$).

Table 2: The rubric for evaluating student-generated chatbot prompts.

Criterion	Low (1)	Moderate (2)	High (3)
Completeness	Fewer than three core template sections completed.	Three core sections (role, personality, constraints) completed.	All template sections completed, including custom knowledge.
Clarity & Specificity	Vague or abstract descriptions lacking actionable detail. <i>e.g., "To answer questions"</i>	Generally clear with some specific elements but room for interpretation. <i>e.g., "Teach people about space"</i>	Concrete, specific language with clear role and behavioral expectations. <i>e.g., "To help students study smarter.."</i>
Constraint Specificity	Mostly vague or generic guidance that is difficult to operationalize. <i>e.g., "be family friendly"</i>	Some specific constraints but covering limited aspects. <i>e.g., "keep content bird related"</i>	Explicit, comprehensive constraints covering multiple dimensions. <i>e.g., "crosscheck information... provide sources"</i>
Coherence & Alignment	Includes contradictory or misaligned elements.	Generally consistent with minor mismatches.	All components are mutually reinforcing and logically consistent.

Table 3: Learning Outcomes Across Camps

Construct	Survey	Camp 1 Mean (SD)	Camp 2 Mean (SD)	All Participants
AI Knowledge & Understanding	Pre	3.29 (0.69)	3.72 (0.57)	3.33 (0.73)
	Post	4.52 (0.77)	4.61 (0.57)	4.57 (0.66)
AI Hallucination Awareness	Pre	3.62 (0.71)	3.50 (0.69)	3.62 (0.63)
	Post	4.21 (0.77)	4.29 (0.63)	4.22 (0.63)
Confidence in Building Trustworthy Chatbots	Pre	2.78 (0.89)	2.84 (0.65)	2.81 (0.75)
	Post	3.73 (0.86)	3.94 (0.60)	3.90 (0.72)

Camp 1: n = 20 (pre/post); Camp 2: n = 27 (pre), 28 (post)

Table 4: Score Distribution of Chatbot Prompt Quality Across Camps

Rubric criteria (each ranging 1-3)	Camp 1 Mean (SD)	Camp 2 Mean (SD)	All participants Mean (SD)
Completeness	2.36 (0.50)	2.25 (0.55)	2.29 (0.52)
Clarity	2.57 (0.65)	2.40 (0.75)	2.47 (0.71)
Specificity of constraints	2.29 (0.61)	2.20 (0.89)	2.24 (0.78)
Coherence and alignment	2.64 (0.63)	2.35 (0.81)	2.47 (0.75)
Total score	9.86 (1.46)	9.20 (1.70)	9.47 (1.62)