"I would have completely trusted the LLM's response": Changes in Students' Understanding and Behaviors After an AI Literacy Hallucination Lesson

GRACE LI, University of Chicago, USA
JONATHAN LIU, University of Chicago, USA
JAEMARIE SOLYST, University of Washington, USA
DIANA FRANKLIN, University of Chicago, USA
MINA LEE, University of Chicago, USA

Many students use LLMs for their academic work, leading to significant concerns about the reliability of LLM outputs since LLMs are prone to generating hallucinations or inaccurate information. While existing AI literacy research aims to inform users about these risks, little is known about how increased knowledge about hallucinations translates to behavioral changes. We conduct a mixed-methods study to explore the effects of a hallucination lesson on student behaviors when performing two common information-search activities—current events fact-finding and citation-finding—that are particularly susceptible to hallucinations. After the lesson, there is a strong improvement of students' usage of hallucination mitigation behaviors for the fact-finding activity but little observable change for the citation-finding activity. This reveals that students are not consistently vigilant about using hallucination mitigation behaviors. We contextualize our findings within the domain of behavioral change to understand factors that might influence students' adoption of these techniques and provide recommendations for teaching about hallucinations.

CCS Concepts: • **Applied computing → Education**; • **Social and professional topics → K-12 education**; **Computing literacy**.

Additional Key Words and Phrases: AI literacy, AI in K-12, Interaction logs

ACM Reference Format:
Grace Li, Jonathan Liu, Jaemarie Solyst, Diana Franklin, and Mina Lee. 2018. "I would have completely trusted the LLM's response": Changes in Students' Understanding and Behaviors After an AI Literacy Hallucination Lesson. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 32 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Students' use of generative AI, and particularly Large Language Models (LLMs) like ChatGPT, has become more widespread. In a 2024 global survey, 86% of students are regularly using generative AI in their studies, and over 2 in 3 students are using AI for information search [12]. Prior work has shown that students are particularly vulnerable to accepting LLM outputs without scrutiny, which can result in students accepting false or inaccurate information [30, 47]. While there have been growing efforts in AI literacy to teach students to critically evaluate LLMs, these studies largely

Authors' addresses: Grace Li, grace_li@uchicago.edu, University of Chicago, Chicago, Illinois, USA; Jonathan Liu, jonliu@uchicago.edu, University of Chicago, Chicago, Illinois, USA; Jaemarie Solyst, jaemarie@cs.washington.edu, University of Washington, Seattle, Washington, USA; Diana Franklin, dfranklin@uchicago.edu, University of Chicago, Chicago, Illinois, USA; Mina Lee, mnlee@uchicago.edu, University of Chicago, Chicago, Illinois, USA.

Manuscript submitted to ACM

focus on measuring student *knowledge* about AI concepts (e.g., bias) and not student *behaviors* after the lesson [40, 41]. Measuring student knowledge is not enough to assess whether students will actually critically and responsibly engage with these LLMs in practice.

Given that many students use LLMs for information search and often accept LLM outputs without verification, we investigate the effect that a hallucination AI literacy lesson has on their information search behaviors. Hallucination mitigation behaviors refer to the concrete actions users can take to reduce the risk of using incorrect, fabricated, or misleading outputs from LLMs. In our work, we break down hallucination mitigation behaviors into two steps: (1) the choice of tool used and (2) the verification behavior.

There are different types of hallucinations, most notably factuality hallucinations, where a model generates information that is factually incorrect, and faithfulness hallucinations, where the LLM output is inconsistent with or unsupported by the source material [20]. In this work, we focus on factuality hallucinations because of the prevalence of students using LLMs to help them with information search tasks [12]. We examine students on two information search tasks: (1) **current events fact-finding** which entails answering questions about current events and (2) **citation-finding** which focuses on finding specific sources to support claims. These two activities are designed to elicit hallucinations from LLMs by focusing on tasks that LLMs cannot accomplish well due to inherent limitations in their training process [20].

We investigate the relationship between students' self-reported knowledge about hallucinations (what students think they know), their demonstrated knowledge about hallucinations (what students actually know), and their LLM usage behaviors (how students use an LLM). Furthermore, we use a lesson on hallucinations to determine whether a change in knowledge is accompanied by a change in behavior.

Concretely, we seek to understand:

- RQ1: How do students' self-reported knowledge of hallucinations align with their demonstrated knowledge and use of hallucination mitigation behaviors in practice?
- RQ2: How do students' knowledge and behaviors change as a result of learning about LLM hallucination mitigation techniques for (1) current events fact-finding and (2) citation-finding activities?

To answer these questions, we performed a classroom intervention during one 2.5-hour hallucination lesson in a 3-week college-level, credit-bearing summer intensive course focused on teaching students AI literacy. We use a mixed-methods approach to integrate qualitative and quantitative data to understand why and how student behaviors change and the relationship between students' knowledge and their behaviors.

Before a hallucination lesson (RQ1), we find that all students self-reported a high understanding that LLMs can generate inaccurate information, but when asked how and why this occurs, students lacked an in-depth understanding of hallucinations. Furthermore, we find that students' demonstrated knowledge does not correspond to their use of hallucination mitigation behaviors. In Section 7.1, we discuss the importance of explicitly teaching students about hallucinations and hallucination mitigation behaviors and provide recommendations for doing so.

After the hallucination lesson (RQ2), students' knowledge of hallucinations improved, but that did not always translate to the adoption of hallucination mitigation behaviors. To contextualize these findings, we use theories of behavior change to understand the inconsistent effect the hallucination lesson has on student usage of hallucination mitigation behaviors.

2 RELATED WORK

2.1 Hallucinations in LLMs

In Natural Language Processing (NLP), *hallucinations* refer to the phenomenon where LLM output is nonsensical or unfaithful to the provided source content [35]. Hallucinations, especially in the context of LLMs, are important to understand because these models can present factually inaccurate outputs in a fluent and convincing way [37]. As different high stakes domains like medicine, finance [22], education [13], and law [14] begin to incorporate LLMs, hallucinations can have significant impacts and lead to potentially harmful real-world consequences. For example, the attorney general of New York City warned resident voters not to rely on AI chatbots to get voting information for an upcoming election due to inaccuracies and missing information in LLM outputs [2, 36]. In another, a lawyer used ChatGPT to write court filings and referenced non-existent cases [42]. These two examples illustrate different reasons LLMs generate hallucinations in common potential real-world applications. LLMs can generate hallucinations on tasks that rely on up-to-date factual knowledge (e.g., recent voting information) or tasks that require citing specific sources (e.g., citing legal cases).

For a systematic way to classify different categories of hallucinations, Huang et al.'s taxonomy of LLM hallucinations uses the stages of the training process of LLMs (training data, training, and inference) to classify the different types of hallucinations [20]. For this paper, we focus on hallucinations due to training data (the data that is used to train the LLM) and inference (the process of generating output text), as these two stages do not require a lot of prerequisite knowledge into different training processes [20]. In the following paragraphs, we provide details about hallucinations due to training data and hallucinations due to inference.

Hallucinations as a result of training data refer to the specific types of data that are used (or are missing) in the training data of LLMs. There are 3 main areas of this type of hallucination: flawed pre-training data sources that contain misinformation and bias [7], limitations due to the knowledge cutoff, where the model lacks access to more recent information and therefore fabricates details to fill the gap [34], and issues with alignment with human preferences where the LLM learns to sound convincing to humans even when its answers aren't actually correct.[18]. By understanding the limitations of the training data that is used to training LLMs, people can be more informed about what questions LLMs are (and are not) able to answer.

Hallucinations can also occur during inference, the process of a model generating the next word.[1] LLMs are statistical models that use probabilities to select the next word in a sequence. There are different methods to select the next word, such as always choosing the most likely word to come next or randomly selecting the from a set of likely words. These design decisions around how a model should generate the next word can have impacts on the types of hallucinations that can occur. For example, many methods to select the next-best-word rely on randomness to generate more creative and diverse content, but introducing randomness has also been found to be positively correlated with an increased risk of hallucinations [11]. Other issues that arise during inference are caused by how probabilities are assigned to words. For example, when predicting what word will come next in a sentence, language models use the attention mechanism to prioritize which words are the most important to consider when generating the next word. These mechanisms generally prioritize nearby words when generating words, which can result in text that is fluent, but not accurate [31]. Given the breadth of types of hallucinations and how they occur, in Section 3.3.2 we detail our rationale for the specific focus for our class.

[1]For this explanation, we use the abstraction of "words" instead of tokens to explain how LLMs generate outputs.

2.2 Generative AI Literacy

AI literacy is focused on supporting people in understanding, using, and critiquing AI [29, 32]. Before the advent of LLMs, AI literacy was focused on teaching people about topics like supervised machine learning and robotics through online simulations, unplugged activities, and formalized curricula [15, 17, 48, 50]. Traditional methods of lecture and assessments have also been used to teach K-12 students about technical and socio-ethical dimensions of AI literacy [23, 25]. Recently, more effort has focused on developing new AI literacy materials that support students in critically evaluating the outputs of AI tools [40] and using different activities like prompt auditing to elicit bias from these models [41].

One of the main metrics that these studies employ is student learning and how different interventions might support student ability to critically evaluate AI outputs and limitations, but not the effect on student usage of AI tools [27, 33, 49]. Incorporating behavioral measures into the assessment of AI literacy lessons is important because students are already frequent users of AI technologies in their everyday academic and personal lives. Behavioral data capture how they actually apply (or fail to apply) responsible practices in these authentic contexts.

2.3 Knowledge and Behavior Gap

The ability for students to critically evaluate LLM outputs for hallucination or bias is important for mitigating potentially harmful LLM usage, but as described above the prior studies have not examined whether students apply these skills to their everyday LLM uses. In other domains like health and security HCI, researchers have discovered that user awareness of an issue and knowledge of how to fix it are not sufficient to prompt behavioral change.

In security research, many users report that they value their online privacy and recognize when situations are potentially security-compromising, but then not take even basic privacy-preserving measures. This disconnect was first coined as the "Privacy Paradox" during an investigation into social media habits [6], but has since been identified across various everyday practices like password creation [4] and online shopping [8]. Here, knowledge seems to only be a partial solution: though misconceptions about security were a common reason for not adopting privacy-preserving practices [3, 21], many users also prioritize convenience over privacy [5, 8]. Interestingly, Sawaya et al. find that user confidence in security knowledge is a much stronger predictor of secure behaviors than actual security knowledge [39], further complicating the discourse around the adoption of good habits.

In terms of health, many people start exercise programs with the intentions of improving their health, but 50% drop out of the program within the first 6 months [38, 43]. Many individuals might be motivated to change their exercise habits but struggle to realize those plans when confronted with barriers like lack of time and motivation [10, 24, 44].

Given that an understanding of desired behaviors is not sufficient for their adoption, as observed in both the security and health domains, it is important to understand the extent that teaching students about AI literacy translates to the actual adoption of good critical usage of LLMs. Gaps between knowledge and behavior in AI literacy would demonstrate a need for AI literacy efforts to go beyond learning-focused outcomes and explore how to support learners in applying their AI literacy understanding to their interactions with LLMs in practice.

3 COURSE CONTEXT AND LESSON DESCRIPTION

In this section, we provide context for the course that the hallucination lesson was taught in, as well as details on the hallucination lesson itself.

3.1 Course Context

The lesson was part of the 3-week summer intensive high school program we ran through an R1 institution. The course was taught in June 2025 and had a total enrollment of 30 students. During the beginning of the course, we collected assent and consent forms from students and parents, respectively. Out of 30 students, 19 consented to participate in the study. This study was approved by our IRB.

No prior knowledge about AI or computing was required to participate in the course. Even so, many (n=16) students had taken at least one computing course offered by their school. 3 students did not have any computing experience. The majority of students were incoming high school juniors and seniors, with one participant being an incoming sophomore. The majority of students (n=13) identify as boys, 4 identify as girls, and 2 preferred not to say. In terms of how often students use LLMs in their everyday lives, most students used them **always** (n=4), **often** (n=10), or **sometimes** (n=3), with only 1 student **rarely** using it. For more context about the structure of the course and other topics covered, see Appendix A.

3.2 Hallucination Lesson Topics

For the hallucination lesson, we discuss one type of hallucination that occurs at each step of the training process (training data, training, and inference) based on Huang et. al.'s taxonomy of hallucinations [20]. This provides students a better understanding of how the design of LLMs informs the types of hallucinations that occur. For hallucinations that occur due to the training data, we focus on the knowledge cutoff, the latest date that is included in the training data set for the LLM. We selected the knowledge cutoff because of the prevalence of people using LLMs to get information about current events, like 2025 polling information [36]. For hallucinations due to training, we discuss sycophancy, the tendency of the model to excessively agree with or flatter the user, often at the expense of accuracy. We selected this topic because it uncovers the tension between making AI agreeable and making it truthful. For hallucinations due to inference, we talk about issues with decoding strategies, how the next word is generated by an LLM, because they can result in LLMs generating citations or article titles that seem accurate but do not exist [42].

3.3 Lesson Structure

Figure 1 provides an overview of the lesson structure. In the following sections, we provide more detail about the different pre-lesson, lesson, and post-lesson components.

3.3.1 Pre-Lesson Activity and Reflections. Before the start of the lesson, students independently complete two, 10-minute warm-up *pre-lesson activities* that are designed to highlight two different types of hallucinations: knowledge cutoff (training data) and imperfect decoding strategies (inference). These two activities were chosen because they are common use cases for LLMs. After each activity, students respond to a prompt to reflect on the tools and process they used to complete the activity. In this section, we detail the design of these activities and reflections.

We design a current events fact-finding activity to illustrate the knowledge cutoff hallucination. This activity contains four short current events questions that students are asked to answer (See Appendix 4 and 5). For example, one question was *"True or False: Beyoncé came to Chicago during her Cowboy Carter Tour."* Since Beyoncé announced her tour in February 2025 [26], this information would not be in the training dataset of the provided LLMs. This activity was based off recent news articles detailing instances where hallucinations due to the knowledge cutoff resulted in inaccurate election information [36], chatbots generating misleading outputs when asked about current events [19], and Meta's AI

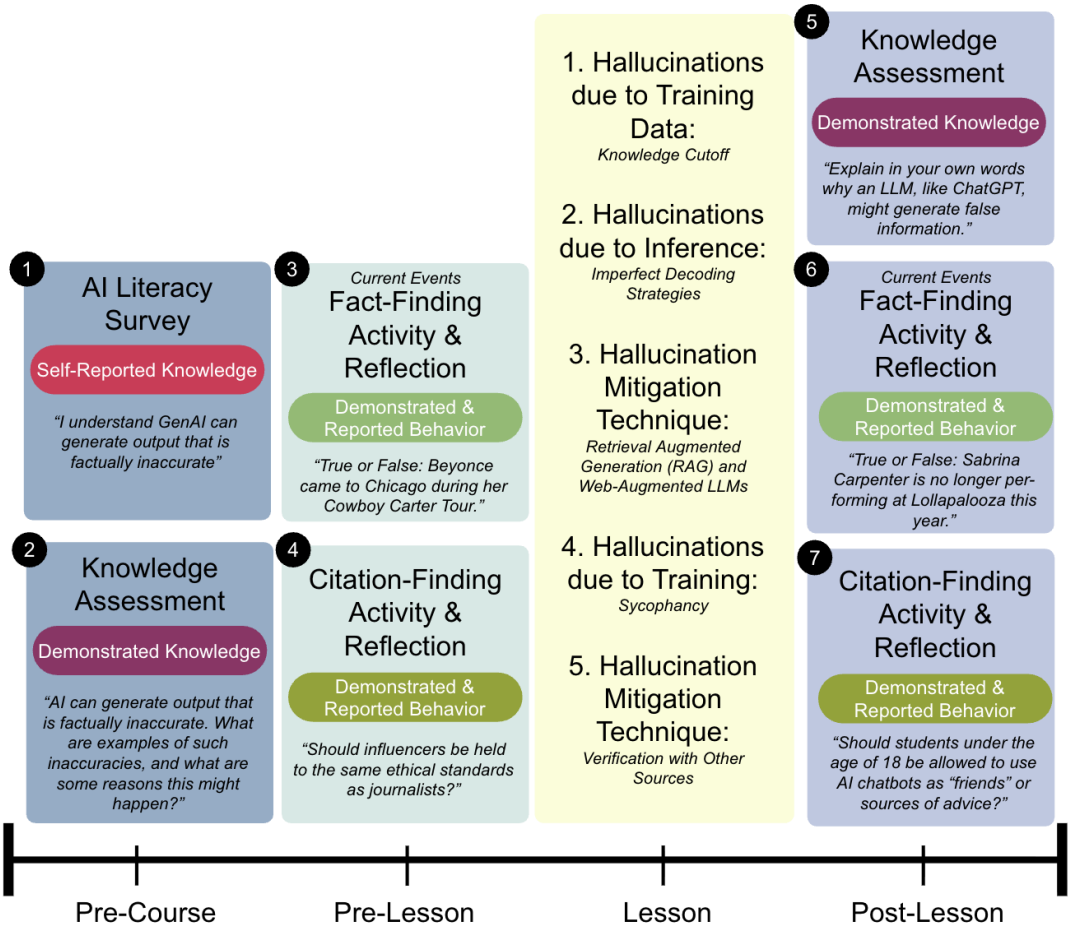


Fig. 1. Overview of hallucination AI literacy lesson structure and data collection, as well as the specific types of data collected at each step. ❶ captures students' self-reported knowledge, ❷ and ❺ captures students' demonstrated knowledge, ❸ and ❹ capture demonstrated and reported behavior. Demonstrated behavior refers to the interaction logs of students' process when completing the activity. Reported behavior refers to students' self-reported reflections on what they did to complete the task. Changes in (❷ and ❺) demonstrate knowledge change. Changes between (❸ and ❻) and (❹ and ❼) capture behavior change.

responding with the incorrect president after the 2024 election [45]. Additionally, grading current events fact-finding activities is straightforward because there is only one correct answer to the designed questions.

We design a citation-finding activity to illustrate how LLMs often hallucinate citations by producing plausible but non-existent article titles, authors, or website links because the model generates text by extending statistical patterns in language rather than retrieving from a verified database. In our activity, students are presented with a debate topic, then asked to choose a side and find 10 links that they would use to support their argument (See Appendix 6). For the pre-lesson citation finding activity, students were provided the prompt: *"Influencers in tech and politics use their social media platforms to share information, which raises the question: "Should influencers be held to the same ethical standards*

Topic	Description
Hallucination due to Training Data: Knowledge Cutoff	The latest date at which training data was gathered. For example, Meta's Llama3.1-405B has a knowledge cutoff date of December 2023 which means that the model was not trained on any articles that were published after December 2023 [1]. This topic illustrates the limitations of the data LLMs are trained on. Students complete a short exercise using different prompts to try and identify the knowledge cutoffs of different models.
Hallucination due to Inference: Imperfect Decoding Strategies	Students learn about why the method of generating an output from an LLM may yield inaccurate information. This topic highlights how LLMs are statistical, next-word prediction machines without any conceptualization of "truth."
Hallucination Mitigation Techniques 1: Retrieval Augmented Generation (RAG) and Web-augmented LLMs	Students learn about using retrieval augmented generation (RAG) as a method to use a knowledge base as a source of factual information. This concept is then extended to web-augmented LLMs, where the model first scrapes the top results from an internet search and appends it to the user's prompt as context before generating a response. Students also are informed that these techniques are not a fool-proof way of preventing hallucinations.
Hallucination due to Training: Sycophancy	Students learn about the difference between misinformation and disinformation. Students learn about sycophancy, how it occurs in LLMs, and complete a short exercise to use an LLM to generate disinformation.
Hallucination Mitigation Techniques 2: Verification with Other Sources	Finally, students learn about general verification strategies of LLM-generated outputs and using alternative sources like internet search to assess the accuracy of LLM generated information. It is re-stated that different tools like RAG and web-augmented LLMs can mitigate hallucinations, but not guarantee that no hallucinations occur. Thus, using alternative sources to cross-check the outputs is important.

Table 1. Overview of topics covered during the 2.5 hour hallucination AI literacy lesson.

as journalists?" This activity design was based on current events where a lawyer cited nonexistent court cases [42] and when a judge found nine hallucinations in a filing about a high-profile case [46].

After each activity, students are asked to write a short reflection based on this prompt: "Please describe the process you used to complete the activity. What tools did you use? How did you evaluate the outputs of the LLM?"

3.3.2 Hallucination Lesson Content. During the lesson, we use a combination of instructor-led lectures and in-class exercises. Table 1 provides an outline of the sequence of topics covered, topic descriptions, and any corresponding activities that were included in the lesson.

3.3.3 Post-Lesson Knowledge Check. After the lesson, students then complete 4 short-answer knowledge check questions. The four questions were (1) "Explain in your own words why an LLM, like ChatGPT, might generate false information." (2) "If LLMs are trained solely on factual data, would that eliminate misinformation? Why or why not?" (3) "Your grandfather is using LLMs to get the latest news. You are concerned, what concrete actions would you give him about misinformation when using LLMs?" and (4) "Explain how training a language model on user preferences might lead to sycophantic responses. Why might this be bad?"

3.3.4 Post-Lesson Activity and Reflection. Students then complete the current events fact-finding and citation-finding activities, but with different questions and prompts. For example, in the post-lesson current events fact-finding activity, students are asked "True or False: Sabrina Carpenter is no longer performing at Lollapalooza this year." Due to time

constraints, there are only 2 current events questions in the post-lesson activity (See Appendix D Table 5). For the citation activity, the prompt was: "Should students under the age of 18 be allowed to use AI companions as "friends" or sources of advice?" (See Appendix D Table 6).

After each activity, students completed a short reflection based on the tools and process they used. The same prompt from the pre-lesson activity reflection was used.

4 DATA COLLECTION & ANALYSIS

In this section, we describe the specific tools (e.g., activity interface and interactive polling platform) that we use during the lesson to support data collection (Section 4.1). Then, we outline the types of data that were collected during the pre-course (Section 4.2), the pre- and post-lesson (Section 4.3) sections, as well as the data analysis conducted for each data source. Finally, we conclude this section detailing how each of the data sources are used to answer our research questions.

For all qualitative analysis, two coders independently applied the codebook, compared results, and then resolved any disagreements to come to a consensus.

4.1 Tool and Platform

To understand students' behaviors during the pre- and post-lesson activity, we build a custom web interface that tracks student interaction behaviors when completing each activity (Figure 2, Section 4.1.1). In addition, we use PollEverywhere, an interactive polling platform, to facilitate gathering student responses throughout the lesson for student reflections and knowledge checks (Figure 3, Section 4.1.2).

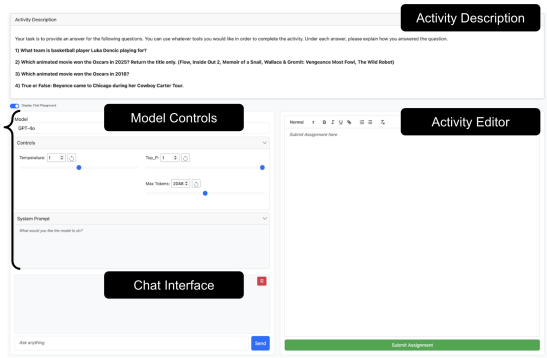


Fig. 2. The activity interface contains the (1) activity description, (2) model controls, (3) chat interface, and (4) activity editor.

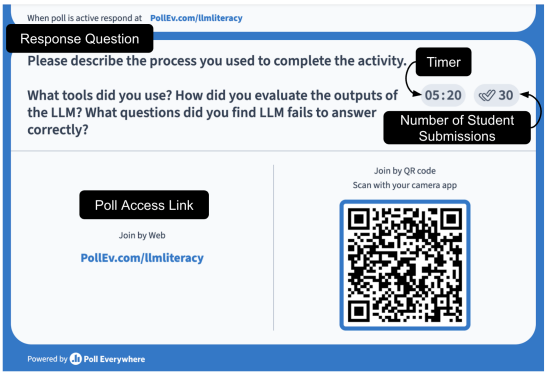


Fig. 3. The Polleverywhere (PollEv) interface contains (1) response question, (2) a timer, (3) the number of student responses, and (4) the link to access the poll.

4.1.1 Activity Interface. To track how a student completes an activity, we built a custom activity interface (Figure 2). The interface has 4 main components: 1) the activity description provides a short description regarding the task and what students are expected to do as part of the activity, 2) model controls which provide students the ability to change the LLM model, decoding parameters (e.g., temperature, top_p, max_tokens etc.) and system prompt, 3) a chat interface to query an LLM and a container to see prior chat messages, and 4) an activity editor where students draft and submit their response to the activity. For more details about the features of the interface, see Appendix B.

To balance privacy and scalability, we capture interaction logs via the activity interface instead of screen recordings. The interface tracks all the actions that a student performs on the activity interface such as copying text from the activity description (1), writing a system prompt in the model controls (2), prompts to the LLM in in the chat interface (3), and the drafting process of typing, deleting, pasting in the activity editor (4). We also track when students leave the activity interface by logging when the page becomes inactive. One of the limitations of only tracking student interactions through our interface is that we lose visibility of exactly what happens outside of the activity interface tab. We can infer details when students navigate to a different webpage through the interaction logs and then use students' self-reported reflections to provide additional information such as specific websites or tools students used outside of the activity interface (see Section 8). In Section 4.3.2, we detail how we store and process these interaction logs for our analysis.

The activity interface was built using HTML/CSS/Javascript and used Quill[2] to track changes in the text-input areas. We used Flask as the backend and hosted the site on a university-affiliated server. We include more details about the interface and models provided in Appendix B.

4.1.2 Interactive Polling. We use Polleverywhere (PollEv),[3] an interactive polling platform, as a low-lift way to survey the students during class (Figure 3). Specifically, we use PollEv to collect student activity reflections and knowledge checks during the lesson.

4.2 Pre-Course Data Collection

4.2.1 Pre-Course Survey. To assess student confidence in understanding hallucinations in LLMs, we used an AI literacy questionnaire to measure self-reported AI literacy, which included the 5-point Likert scale question "I understand LLMs can generate output that is factually inaccurate" (1 - strongly disagree to 5 - strongly agree). This question is part of the Expectancy-Value-Theory survey instrument for LLM tools [9].

Using the pre-course survey, we can assess students' self-reported knowledge about LLMs ability to generate inaccurate information.

4.2.2 Pre-Course Knowledge Check. To understand the prior knowledge that students have before starting the course, we use a pre-course knowledge check to assess what students know. For the scope of this paper, we specifically focus on the following question: "AI can generate output that is factually inaccurate. What are examples of such inaccuracies, and what are some reasons this might happen?"

To code student responses, we inductively create a codebook of all the types of hallucinations that students mentioned or implied in their response: reasoning, sycophancy, inaccurate training data, knowledge cutoff, and hallucinations from inference (see Appendix D Table 7) for the code book). Then, for each code, we deductively code for whether students demonstrate an (1) **explicit** understanding, where the student names the specific type of hallucination (knowledge cutoff or imperfect decoding strategies) and describes it, an (2) **implicit** understanding, where the student describes but does not name the phenomenon, or (3) **none**, where the student did not mention or describe the hallucination type in their response.

[2]https://quilljs.com/
[3]https://www.polleverywhere.com/

4.3 Pre- and Post-Lesson Data Collection

4.3.1 Post-Lesson Knowledge Check. Each of the four questions in the post-lesson knowledge check (as described in Section 3.3.3) was worth 2 points. Students are given 2 points for a correct answer, 1 point for a partially-correct answer, and 0 points for an incorrect answer.

After the lesson, we code student responses to the post-lesson knowledge check based on whether students demonstrate an explicit, implicit, or no stated understanding of the topic, as described in Section 4.2.2.

By comparing the pre-course knowledge check with the post-lesson knowledge check, we can evaluate student learning as a result of the lesson.

4.3.2 Pre- and Post-Lesson Activity Interaction Logs. Pre- and post-lesson activities were targeted towards activities that LLMs were more likely to hallucinate on, so they were analyzed to determine the extent to which students utilize hallucination mitigation techniques. In this section, we detail how we process and analyze the interaction logs collected via the custom activity interface.

Similar to prior work in user-LLM interaction analysis, we first process the activity sessions for each participant [28]. Each activity session consists of sequential, key-stroke level events that were logged via the interface. These events are turned into event-blocks, which are deterministic, non-overlapping abstractions of a sequence of events that occur within the same location (e.g. text-insert(a, locationA), text-insert(b, locationA) → text-insert(ab, locationA).

Since interaction logs can only capture the behaviors of students on the activity interface website, we use students' reflections to supplement the missing information from the interaction logs. For example, from the interaction log, we might observe that a student leaves the page for 30 seconds before typing in an answer to Q2. From the student reflection, the student might say "I used Google AI's Overview to answer Q2." Using these two data sources, we can create a more specific **interaction process**. Two authors independently coded the interaction logs and student reflection, then compared, and resolved any disagreements.

The students' interaction processes were then deductively coded for the tools that the student used: Provided LLM (student used the provided chatbot), outside-LLM (student reported using an LLM but didn't prompt a provided LLM), web-augmented LLM (student reported using a specific model that supports web-browsing like GPT-4 with web search enabled), search (student reported using an internet search), or brain (student used their own knowledge).

We also coded the students' interaction processes for the sequence of actions that students performed to accomplish the task. We track what tool(s) were used and the order the tools were used in. Processes were also tagged for the presence of hallucination mitigation behaviors. For the fact-finding activity, the hallucination mitigation behavior of interest is the use of any tool other than a non-web-augmented-LLM at any point. For the citation-finding activity, the hallucination mitigation behaviors of interest are either finding citations through search engines or the individual verification of LLM-generated citations. Student reflections, interaction logs, and submissions were all utilized to classify the adoption of these behaviors – see Appendix D Table 9 for the codebook.

By comparing the interaction processes of students before and after the lesson, we can observe changes in (1) the tools that students use and (2) the overall process that students follow, which can provide insight into student adoption of hallucination mitigation behavior.

4.3.3 Pre- and Post-Lesson Activity Submitted Work. Another metric by which student hallucination mitigation behaviors were measured is through the work they submitted for the pre- and post-lesson activities. Because activities were selected to elicit hallucinations, student submissions were evaluated for the extent to which these hallucinations manifested.

For the current events fact-finding activity, we compute the percentage of correct answers. For the citation-finding activity, submissions are evaluated by citation quality. The authors inductively identified four types of quality issues among submissions, all of which can be caused by LLM limitations during citation generation, and coded each citation provided by students for those issues.

The four issues were sorted into levels in terms of increasing effort required for students to check: (1) the source is a hallucination and does not exist, (2) the citation points to an organization/journal/collection rather than a single usable source, (3) the source is not relevant to the topic of interest, and (4) the source is not credible. The codebook for how these were objectively identified can be found in Appendix D Table 10.

We are interested in capturing student *citation-checking* behavior, so each student's overall submission (of 10 citations) was coded for the lowest-level issue identified anywhere in the submission. For example, if a student submits 10 hyperlinks generated by an LLM, and only two of the hyperlinks do not point to a real website, we still conclude that the student did not engage in the hallucination-mitigation behavior of checking their links for existence, and their submission is tagged as such.

By comparing the submission quality before and after the lesson, we can assess whether students utilize better hallucination mitigation behaviors on these tasks.

4.3.4 Post-Activity Student Reflections. In addition to using student reflections for filling in gaps in interaction logs, as described above, we also use the reflections to provide additional insights into why students choose to certain the tools/process to complete the activity or to understand students' rationale for changing their behaviors.

5 LIMITATIONS

There are a few limitations to the data collection methods that should be considered. First, our small sample size of 19 students limits the generalizability of the findings. Although the students were representative of various countries and cultural backgrounds, they all were in a college prep summer program, reflecting a degree of self-selection toward academically motivated and resource-supported students.

Second, due to logistical and student privacy constraints, our analysis requires us to make some inferences about student behavior. Primarily, we use interface interaction logs to track student behaviors when completing pre- and post-lesson activities. Though this log data provides an objective record of student behavior in the interface, students were allowed to access external tools (e.g., search engines or external LLMs) to complete the activities. We did not use screen recordings to monitor the sites students access outside of the activity interface due to the potential risk of capturing personal information unrelated to the study. We use students' written reflections to fill in these gaps, but these reflections vary in detail and may be less reliable than the interaction logs. As such, though we measure and report student behavior to the best of our ability, inaccuracies are possible.

6 RESULTS

In the following section, we detail our findings for each research questions: RQ1 (Section 6.1), RQ2a about the current events fact-finding activity (Section 6.2), and RQ2b about the citation-finding activity (Section 6.3.

6.1 RQ1: How do students' self-reported knowledge of hallucinations align with their demonstrated knowledge and hallucination mitigation behaviors in practice?

To answer RQ1, we use ❶ AI literacy survey to understand students' self-reported knowledge based on their response to the question "I understand GenAI can generate output that is factually inaccurate." We use ❷ pre-course knowledge assessment to evaluate the demonstrated knowledge that students have relating to hallucinations. Then, we use the current events fact-finding and citation-finding pre-lesson activities (❸ and ❹) to understand the types of hallucination mitigation behaviors that students demonstrate on these two activities.

Overall, prior to any instruction, we find that students know that LLMs can generate factually inaccurate information (self-reported knowledge), but students are largely unaware about how and when inaccurate information occurs (demonstrated knowledge). Furthermore, across both the current events fact-finding and citation-finding pre-lesson activities, we see inconsistent use of hallucination mitigation behaviors.

Students know that LLMs can generate factually inaccurate information. For the statement, "I understand LLMs can generate output that is factually inaccurate," the majority of students responded "strongly agree" (n=14) and "slightly agree" (n=5) demonstrating that all students know that LLMs can generate inaccurate information.

But students do not know how or when different types of hallucinations occur. During the pre-course knowledge assessment, students answered an open-response question of "AI can generate output that is factually inaccurate. What are examples of such inaccuracies, and what are some reasons this might happen?" We code student responses for whether they mentioned or described the following hallucination categories: reasoning, sycophancy, inaccurate training data, knowledge cutoff, and imperfect decoding strategies (See Appendix Table 7 for the codebook). We use student responses to measure their demonstrated knowledge of different types of hallucinations and students' understanding of why these hallucinations happen.

In Figure 4, we find that the most common types of hallucinations students were familiar with were hallucinations due inaccurate training data (47%, n=9), reasoning (42%, n=8), imperfect decoding strategies (16%, n=3), sycophancy (16%, n=3), and knowledge cutoff(5%, n=1). Students were prompted to list reasons, but most students only mention one type of hallucination (68%, n=13), and a few mention two types (16%, n=3). 3 students do not mention any specific types of hallucinations in their response (16%, n=3). Students know that LLMs can generate inaccurate information, but lack deeper knowledge to understand when and how these hallucinations occur.

Despite awareness of LLM hallucinations, students' use of hallucination mitigation behaviors is inconsistent. Even though students have a high self-reported knowledge that LLMs can generate inaccurate information, students often do not engage in hallucination mitigation behaviors during the pre-lesson activities.

For the current events fact-finding activity (Figure 5), 42% of students do not engage in the hallucination mitigation behavior of checking an LLM responses (n=8). 57% of students do engage in hallucination mitigation behaviors like (1) choosing to not use an LLM to complete this activity or (2) checking the LLM response, if they do choose to use a LLM (n=11). Additionally, we find that student use of hallucination mitigation behaviors does not depend on their demonstrated knowledge. Of the 18 students who do not know about the knowledge cutoff, 11 of them (61%) still employ hallucination mitigation behaviors during the current events fact-finding activity which was targeted at knowledge cutoff hallucinations. Conversely, the one student who described the knowledge cutoff limitation in the pre-course knowledge assessment did not engage in hallucination mitigation behaviors (Figure 5).

For the citation-finding activity (Figure 6), 53% of students engage in hallucination mitigation behaviors by checking all the links for existence–either through searching for links themselves or verifying the links that an LLM generates

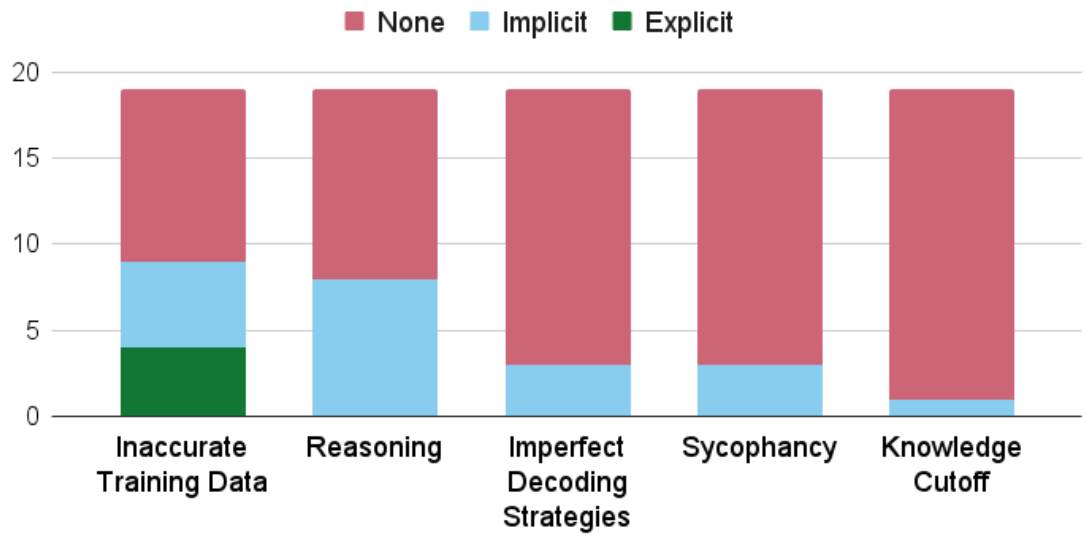


Fig. 4. Students demonstrated little knowledge about different types of hallucinations before the hallucination lesson. **None** means that the student did not mention or describe that given type of hallucination in their response. **Implicit** denotes that the students' ability to describe the phenomenon. **Explicit** denotes the students' ability to name the phenomena of the knowledge cutoff and describe it. Students could mention multiple types of hallucinations in their response.

and only submitting working links (n=10). 42% of students did not engage fully (some links checked, n=3) or at all (no links checked, n=5), and one student was unclear. Like the current events activity, there is no relationship between students' demonstrated knowledge of imperfect decoding strategies and their use of hallucination mitigation behaviors. 16 students did not describe hallucinations from imperfect decoding strategies, but 9 of them still exhibited the hallucination mitigation behavior of checking all their links. The remaining 7 students checked some (n=2), none (n=4) of the links, and for one student it was unclear. For the 3 students that had an implicit understanding of imperfect decoding strategies, one checked no links, one checked some links, and one checked all the links, demonstrating variance.

Despite the fact that a majority of students in each activity perform hallucination mitigation behaviors, only 21% (n=4) of students do both, so students who adopt these techniques in one activity do not appear to be more likely to adopt them in the other activity.

Overall, we find a misalignment between students' demonstrated knowledge of different types of hallucinations (knowledge cutoff and imperfect decoding strategies) and their use of hallucination mitigation techniques. Specifically, we find that students do not need to know about a certain type of hallucination in order to perform hallucination mitigation behaviors. Conversely, students' demonstrated knowledge of a certain type of hallucination does not necessarily mean that they will use hallucination mitigation techniques. Furthermore, we find that students are inconsistent in their use of these techniques, verifying LLM outputs in some activities but not others. Broadly, despite high awareness that hallucinations occur, we find little evidence that students are consistently vigilant about hallucinations in their LLM usage.

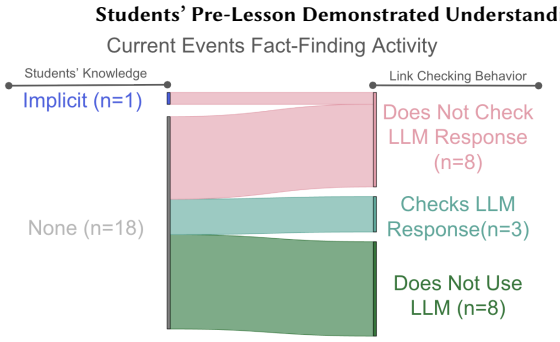


Fig. 5. Students' relationship for the current event fact-finding activity between 1) students' demonstrated understanding of the knowledge cutoff a pre-lesson knowledge check, and 2) students' demonstrated hallucination mitigation behavior based how they complete the current events fact-finding activity.

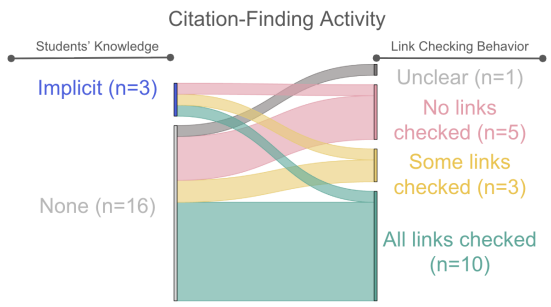


Fig. 6. Students' workflow for the citations activity between 1) self-reported awareness of hallucinations on a Likert scale, 2) demonstrated understanding of hallucinations on a pre-lesson knowledge check, and 3) demonstration of hallucination mitigation behavior based on their process to complete the task.

6.2 RQ2a: How do students' understanding and behaviors change as a result of learning about LLM hallucination mitigation techniques when completing a fact-finding task?

To answer RQ2a, we evaluate changes in student understanding by comparing the ❷ pre-course knowledge assessment with the ❺ post-lesson knowledge assessment. We then evaluate change in student behavior by comparing the students' behaviors from the pre- and post-lesson current events fact-finding activity (❸, ❻). We use reflections to contextualize why students selected their approach to completing the activity. Finally, we compare the overall quality of student submission between the pre- and post-lesson activities.

Students' demonstrated knowledge of the knowledge cutoff increased. In Figure 7, we show the change in students' demonstrated knowledge of the knowledge cutoff between the pre- and post-lesson knowledge checks. From the pre-course knowledge check, only one student demonstrated an implicit understanding of the knowledge cutoff, and all other students (n=18) did not indicate any awareness. After the lesson, most students either explicitly (n=11) or implicitly (n=5) mentioned the knowledge cutoff, though three students did not mention the knowledge cutoff in their post-lesson knowledge check.

After the lesson, students used more internet search and web-augmented LLM tools to complete the current events fact-finding activity compared to the pre-lesson activity. During the post-lesson current events fact-finding activity, the majority of students did not use the provided LLM as their first tool to answer the question. In Figure 8, compared to the pre-lesson fact-finding activity, there was an 8% increase in percentage of questions answered with internet search and a 14% increase in percentage of questions answered with a web-augmented LLM. In the post-lesson activity reflection, students referenced hallucination mitigation practices like using a web-augmented LLM or awareness of the knowledge cutoff to inform their use of tools. One student used the knowledge cutoff of different models as a way to answer specific questions: *"This time I used different models of AI depending on the knowledge cutoff dates"* (P17).

Because the answers to the questions on the pre-lesson activity were reviewed afterwards, students had the chance to get immediate feedback on whether they had answered correctly or not. Four students switched from using a LLM in the

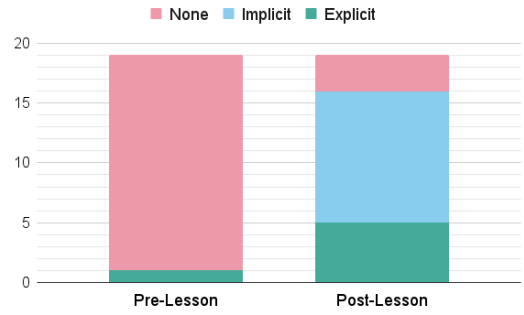


Fig. 7. Changes in students' knowledge of the knowledge cutoff before and after the hallucination lesson. **Explicit** denotes the students' ability to name the phenomena of the knowledge cutoff and describe it. **None** means that the student did not mention or describe that given type of hallucination in their response.

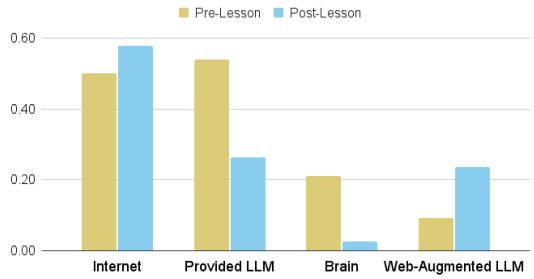


Fig. 8. Comparison of percentage of questions for which each tool was used in completing the current events fact-finding activity.

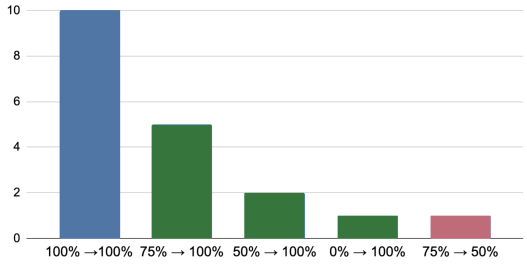


Fig. 9. Changes in student submission quality in the pre- and post-lesson activities on the current events fact-finding activity before and after the lesson. **Blue** bars denote that there was no change, **Green** bars denote an increase, and **Red** bars denote a decrease in student submission quality.

pre-lesson to using internet search in the post-lesson because they realized they submitted an incorrect LLM-provided answer during the pre-lesson: *"I used google search [in the post-lesson activity]. Last time, LLM lied to me"* (P7).

In the post-lesson activity, far fewer students used the provided LLM as their first information source. In Figure 10, we show the process that students use to answer each question for the pre- and post-lesson current events fact-finding activity. In the pre-lesson current events fact-finding activity, students immediately consulted the provided LLM for 42% of the questions. When the LLM provided an answer, students only corroborated with another source 26% of the time, even though the LLM was incorrect 71% of the time. On the other hand, in the post-lesson version of the same activity, only 16% of questions were immediately queried to the provided LLM, and two of the three responses were checked with another source.

Different LLM responses affect how students perform with hallucination mitigation behaviors. For students that directly prompt an LLM, we code LLM responses based on whether they (1) directly state the answer, (2) imply an incomplete knowledge base, (3) explicitly recommend additional verification, or (4) refuse to answer (see Appendix 8 for the codebook). We only evaluate this subset of students to see whether they perform hallucination mitigation

behaviors such as verifying the LLM generated output against a different source. For students who consulted another source first, they have already done a desired behavior.

For LLM outputs that directly state the answer, only 19% (3/16) of students checked the LLM's answer with another source. For LLM outputs that implied a knowledge cutoff, 50% (3/6) of students checked the answer, and for LLM outputs that explicitly suggested verification, 75% (3/4) of students checked the answer with another source. When the LLM outputs did not provide an answer, all students used another source to get the answer. These findings imply that differing presentations of LLM outputs may trigger different verification behaviors (which we discuss further in Section ??).

Student performance on the task improved after the lesson. Student fact-finding activity submissions were graded as the percentage of correct answers on each activity. As seen in Figure 10, we see a dramatic improvement in student performance on the fact-finding activity as a result of all but one student adopting hallucination mitigation behaviors. 10 students maintained their perfect score between the pre- and post-lesson activity. Of the 9 students who did not get a perfect score on the pre-lesson activity, 8 improved their score from a 75% (n=5), 50% (n=2), and 0% (n=1) to a 100% score on the post-lesson activity. For one student, we do not see an increase in their success on the student submission. This result is closely tied to changes in tool use and process: 21/23 of the incorrect answers in the pre-lesson activity came from students who consulted the LLM and did not verify with another source, but only one student followed this process in the post-lesson activity (and submitted the only incorrect answer).

6.3 RQ2b: How do students' understanding and behaviors change as a result of learning about LLM hallucination mitigation techniques when completing a citation-finding task?

To answer RQ2b, we follow the same process as RQ2a, but for the citation-finding task (❹, ❼). We evaluate changes in student understanding, behavior (tool use and process), and submission quality between the pre-lesson and post-lesson citation activity.

We found that while there was a clear change in student understanding of imperfect decoding methods for LLMs, there was only a marginal change in the tools and process students used to complete the tasks. Additionally, there was only a slight increase in student submission quality during the post-lesson activity.

Students' understanding increased as a result of the lesson. The citation-finding activity was targeted at hallucinations caused by imperfect decoding strategies. Figure 11 compares students' understanding of imperfect decoding strategies before and after the hallucination lesson. Before the lesson, 3 students were implicitly aware of imperfect decoding strategies in their response to the question of "AI can generate output that is factually inaccurate. What are examples of such inaccuracies, and what are some reasons this might happen?" The remaining 16 students did not mention imperfect decoding. After the lesson, 94% of students were able to either explicitly (n=13) or implicitly (n=5) describe hallucinations caused by imperfect decoding strategies. Only one student did not.

Students' tool use varied slightly between the pre- and post-lesson citation-finding activity. Though we see a large change in student understanding, there is little change in student tool usage. From Figure 12, there was a decrease in students using the provided LLM (5%, n=1) and search (5%, n=1). There was an increase in the use of a web-augmented LLM (5%, n=1). There was no change in the number of students that used an outside LLM.

Students who used an LLM often said that they were more efficient in helping them complete the task even if the LLM generated incorrect links (n=3): *"The LLM-generated outputs still had some faulty links however I used the ones that are valid and then recycled my prompt to get some new sources. I then verified those as well and repeated the process until I reached 10 credible and verifiable sources"* (P8). On the flip side, two student who used internet search for both the pre-

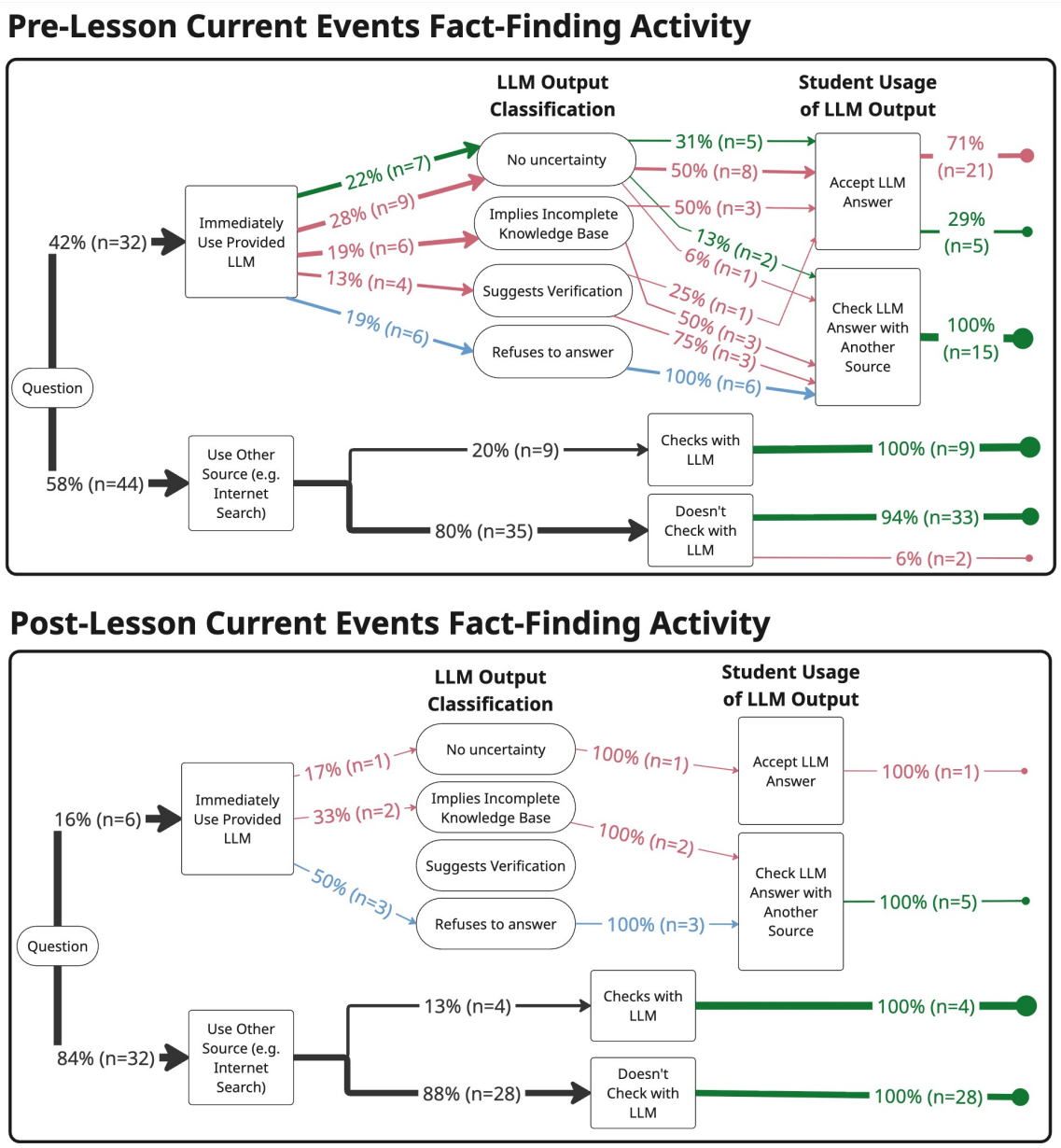


Fig. 10. Pre- and post-lesson student workflow that each student used to answer each question for the current event fact-finding task. **Green** arrows denote that the answer is correct. **Red** arrows denote that the answer is incorrect. **Blue** arrows denote that the LLM did not provide an answer in its response.

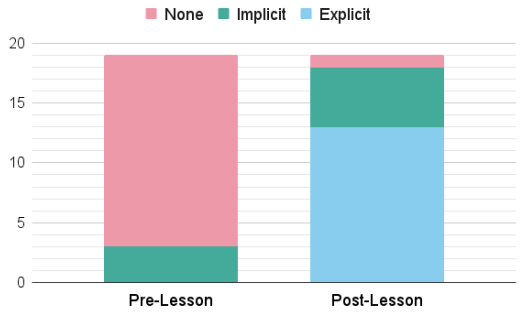


Fig. 11. Comparison of students' knowledge about hallucinations due to imperfect decoding strategies before and after the lesson. **Explicit** denotes the students' ability to name the phenomenon and describe it. **None** means that the student did not mention or describe that given type of hallucination in their response.

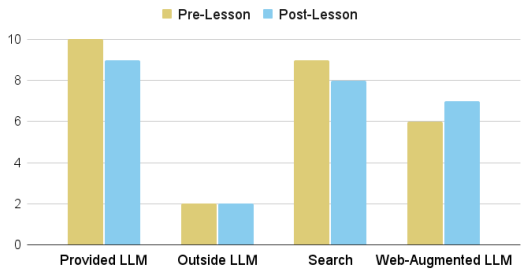


Fig. 12. Comparison of students' use of tools in the citation-finding activity before and after the lesson.

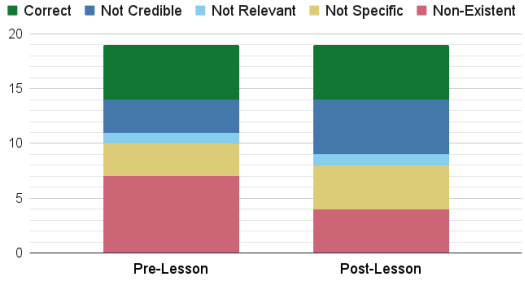


Fig. 13. Comparison of citation quality in the citation-finding activity before and after the lesson. (See Section 4.3.3).

and post-lesson activity said that the task itself was easy enough to complete without the help of a LLM: *"I didn't use an LLM and used the same method. As someone who has to research and find articles quickly as a debater, this task was a no-brainer for me to just look things up online"* (P4). One student who switched from using internet search for the pre-lesson activity to the web-augmented LLM for the post-lesson activity also cited efficiency as a reason for switching: *"Instead of going into every single credible sites and searching relevant key words, I asked the LLM to list credible sources about why AI Chatbots can be harmful, then copy/paste the titles it gave in google search to ensure that source exists. This way it was way faster, and allowed me to find essays I didn't know was conducted before"* (P6).

Students' processes were similar the same when completing the post-lesson citation-finding activity, with some increased citation verification behavior. We code the process that a student uses to complete the activity by sequence of tools and verification behaviors that we observe through the interaction logs and student reflections (See Appendix D Table 9 for codebook). Figure 14 shows the processes used by students to complete this activity in the pre- and post-lessons. Comparing the flowcharts of the pre-lesson and post-lesson activity in Figure 14, students largely followed similar processes to complete the activity before and after the lesson. Three students explicitly mentioned in their reflection that they did the same process: *"I did the same thing as before."* (P15), and *"I used Perplexity again*

to find me these sources" (P16). Though tool selection and most verification behaviors remain relatively similar, we see a moderate increase in the number of students who check all LLM-generated citations, which was one desired hallucination mitigation behavior.

Student submissions contain fewer broken links during the post-lesson citation-finding activity. As discussed in Section 4.3.3, citation quality was analyzed as a partial proxy for hallucination-mitigation behaviors. Submission quality in the pre- and post-lesson citation-finding activities are presented in Figure 13.

We do not explicitly teach students general information literacy, so we do not expect to see much changes in the amount of specific, relevant or credible sources. The overall presence of not specific and not relevant links stayed the same with a slight increase in the number of not credible links between the pre- and post-lesson activities.

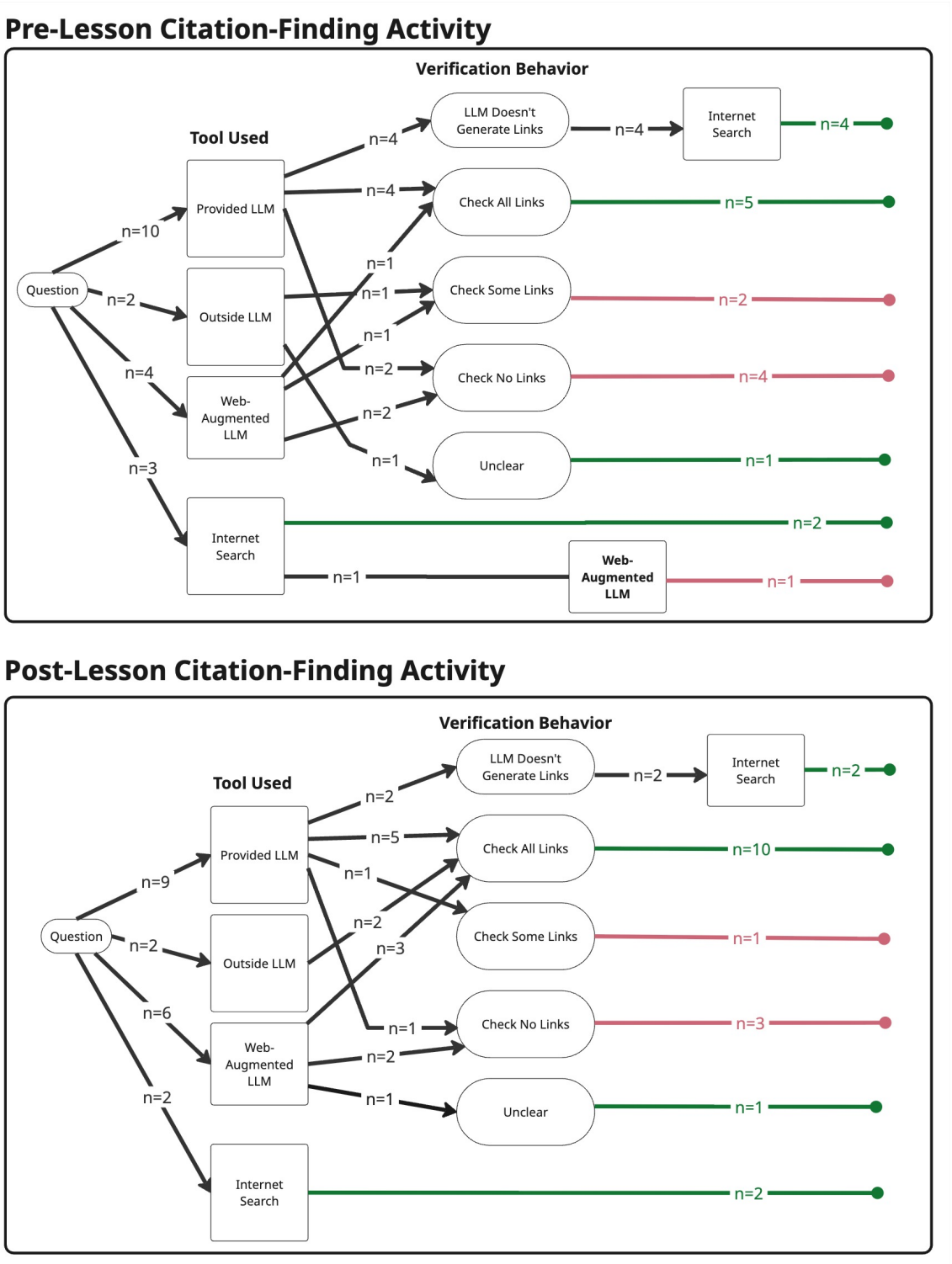
However, as we did explicitly teach students about the ability of LLMs to hallucinate links, we did expect to see a difference in the number of student submissions that contain non-existent citations. To that end, from Figure 13, we do see a moderate decrease in submissions with non-existent sources (7 in the pre-lesson, 4 in the post-lesson), in parallel with the increase in students who checked all LLM-provided citations. Of the four students who submitted non-existent sources in the post-lesson activity, three also submitted nonexistent sources in the pre-lesson activity. On the other hand, three of these four students verified their answers in the fact-checking activity, illustrating an incomplete adoption of hallucination mitigation techniques.

7 DISCUSSION

7.1 Students are aware that LLMs can generate inaccurate information, but can not recognize specific instances in theory or practice

While students are aware that LLMs can generate inaccurate information, this awareness does not translate to an understanding of how or when hallucinations can occur, nor an ability to identify these inaccuracies in practice. Students' behaviors on the current events fact-finding and citation-finding activities provide insight into real-world student usage. Students tend to trust the LLM-generated outputs and accept/utilize inaccurate information without question, which can lead to students internalizing incorrect information and passively consuming information. We conclude that existing warnings about potential LLM inaccuracies are insufficient for motivating critical usage of LLMs. The gap between awareness of and action towards hallucinations likely reflects an underestimation by students of the situations or frequency that these hallucinations occur. As such, explicit teaching of hallucination causes and hallucination mitigation behaviors is important to improve students' responsible use of LLMs.

7.1.1 Recommendations for future design of generative AI literacy lessons. Explicit instruction about AI literacy topics is important to ensure that students understand the limitations of LLM-based technologies. As such, students need to understand how different types of hallucinations occur. In this paper, we use three different types of hallucinations (training data, training, and inference) to teach students about when and why hallucinations occur. We provide students both *conceptual* and *actionable* knowledge to understanding why hallucinations occur and how they can apply that knowledge. We designed realistic activities (fact-finding and citation-finding) to illustrate the commonness of these occurrences in real-world tasks. In the following subsection, we explore which aspects of the lesson and activity design may have been particularly helpful for supporting students in adopting hallucination mitigation behaviors.



Manuscript submitted to ACM

Fig. 14. Pre- and post-lesson student workflow that students' completed the citation-finding activity with. **Red** lines denote submissions with broken links. **Green** lines denote submissions that do not contain broken links. For simplicity, some secondary tools used by students are omitted. These tools do not impact verification behavior.

7.2 Students' demonstrated knowledge doesn't always align with their behaviors.

After the lesson, the majority of students demonstrate an increase in their knowledge about when and why hallucinations happen. However, students' adoption of hallucination mitigation does not always increase as a result of the lesson. For the current events fact-finding activity, we observe that 9 students accepted an LLM-generated answer without verification in the pre-lesson activity, whereas only one student did so in the post-lesson activity. On the other hand, for the citation-finding activity, the changes were more moderate: 7 students submit unchecked sources in the pre-lesson and 4 do in the post-lesson. Considering further that of those four students, three students use hallucination mitigation behaviors in the fact-finding activity, **we find a task-dependent disconnect between student knowledge and behavior. While increased awareness caused some adoption of hallucination mitigation behaviors, a more thorough adoption of these techniques may require more than just knowledge about how and when to use them.**

To further explore this disconnect between knowledge and technique usage, we use the Fogg Behavior Model (FBM) to understand how factors of motivation, ability, and triggers may cause the desired behavior of hallucination mitigation to occur [16]. The FBM argues that a user's motivation, ability, and the existence of a trigger are the 3 necessary components for a target behavior to take place. Additionally, ability and motivation are inversely related in FBM. For example, if the targeted behavior is difficult and the existing ability is low, a high amount of motivation is necessary to change behavior. On the other hand, if the targeted behavior is easy, then a trigger can incentivize a user with low motivation to perform the task.

In the following sections, we use the lens of the FBM to explore different reasons why students may or may not have adopted more hallucination mitigation behaviors.

7.2.1 Students' use of hallucination mitigation behaviors depends on the cost of adopting the behavior. In FBM, ability refers to how easy or difficult a behavior is for a user to complete. In the current-event fact-finding activity, one reason we observed a clear shift in the tools students used may be because using an LLM and conducting an internet search require a similar level of effort, so switching tools or verifying an answer are behaviors that students have high ability to do. For example, copying the question "True or False: Beyoncé came to Chicago for her Cowboy Carter tour" into an LLM or into a Google search require a similar amount of effort. The effort required to adopt the hallucination behavior of using internet search, instead of the provided LLM, was low, so by the FBM the required motivation to change was low as well. Thus, when presented with a trigger to change (the lesson and following post-lesson activity), we see a high rate of behavior change, with all but one student adopting hallucination mitigation behaviors in the activity.

On the other hand, finding citations with only the internet can be a challenging task for students because it requires more refined use of keywords to find relevant articles. Using an LLM to find the sources by just pasting into the prompt required less effort than a traditional internet search. This may explain why tool use across the pre- and post-lesson activity for citations remains largely consistent: students have some motivation to change tools, as shown by the change in the fact-finding activity, but because the increased difficulty of changing in this case, the motivation is insufficient to incentivize change. This is corroborated by the fact that students often said cited efficiency as the reason why they chose to use certain tools: '*'Using an LLM to find sources was a more efficient way than actually finding on my own."* (P17). This was not just for LLMs either: for one student familiar with finding sources for debate, it was a *"no-brainer"* to just use internet search (P4). As predicted by FBM, we see that the difficulty of performing a task is a strong indicator for behavior, and that the presence of a trigger like a lesson is insufficient to create change.

For students that used an LLM to find citations, we can also use FBM's ability construct to understand why students were not checking all the generated links. Even though students have some motivation to succeed, the more motivation is needed when the task is more difficult to perform. The tedium in checking all 10 sources generated may increase the task difficulty to the point where, for some students, the motivation to succeed on the task is insufficient to adopt the behavior. Comparatively, for the current events fact-finding activity, students only have to check 4 current events questions at most, suggesting that the scale of the task may affect the extent to which students adopt hallucination mitigation behaviors as well.

7.2.2 Immediate feedback on student answers may incentivize behavior change. Answers to the current events fact-finding activity were reviewed after students completed the activity. As such, students who used the LLM and did not verify the output with another source received immediate feedback that their submitted answers were incorrect. Four students explicitly stated that they changed their behaviors when completing the post-lesson activity because *"last time, LLM lied to me"* (P17). Due to the immediate feedback that they received, students may have been motivated to change their behaviors in order to achieve a better grade. Comparatively, for the citation-finding activity, students did not receive immediate feedback on whether or not they submitted working links. This may have resulted in lower motivation, which could contribute to the lower changes in behavior.

7.2.3 LLM-generated outputs may serve as a trigger for adoption of hallucination mitigation behaviors. From student interactions with the LLM, the LLM outputs can have an effect on how students utilize the LLM-generated output. For example, in the current events fact-finding pre-lesson activity, of the 10 LLM outputs that contained information about the LLM's limitations or encouraged the student to consult other sources, 6 were checked with another source. Comparatively, of the 16 LLM-generated answers that expressed no uncertainty, only 3 were verified by students. Furthermore, among students that used the provided LLM for all the questions, students only engaged in hallucination mitigation behaviors for the questions where the LLM expressed uncertainty. For example, P6 directly accepts the LLM-generated incorrect answer (with no uncertainty) for the pre-lesson question about Luka Dončić (Appendix 4), but chooses to verify the LLM-generated answer (that contains an answer and the stated limitation of the knowledge cutoff). This demonstrates that the design of LLM outputs can have an impact on how students' use and perform hallucination mitigation behaviors. Using the Fogg Model of Behavior (FBM), we can see that uncertainty expressed in the LLM-generated output could act as a trigger to engage in the hallucination mitigation behavior of checking the LLM-generated answer with another source.

8 FUTURE WORK

8.1 Future Data Analysis Plans

Our findings present the effects of the hallucination lesson on student behaviors directly after the lesson. Future work can explore the longitudinal effects of the hallucination AI literacy lesson on student behavior and usage of hallucination mitigation behaviors. Longitudinal analysis can reveal patterns in when and where students use hallucination mitigation strategies. These longitudinal analyses can be conducted in research-focused classes such as independent study seminars, capstone design courses, or project-based research classes, where students repeatedly engage with information-seeking, writing, and verification tasks over extended periods. Embedding AI literacy interventions in these contexts would allow researchers to identify whether hallucination mitigation behaviors are sustained or diminished over time.

8.2 Development of Future Lessons on Hallucination

Our findings also indicate that explicit instruction about hallucinations is important, but does not need to occur in formal learning environments. Future work can explore how alternative educational methods such as asynchronous modules or short tutorials can help students gain more knowledge about hallucinations and hallucination mitigation behaviors. Furthermore, integrated learning opportunities embedded within chat interfaces could provide just-in-time reminders to students to verify LLM-generated outputs on tasks that are particularly sensitive to hallucinations.

In our study, we designed a custom interface that allowed post-hoc analysis to take place to understand students' use of hallucination. Our activity interface only tracks student interactions within the interface page. Future work can more comprehensively capture detailed information regarding students behaviors beyond just the activity interface. This would provide more detailed insights into the tools and process that students use when completing activities.

In our study, we use a post-hoc method to analyze students' use of hallucination mitigation behaviors. This information may be helpful to provide to educators in real-time to gain an understanding of how students are completing the tasks and tailoring lesson content to specific hallucination mitigation strategies that students may be unaware of. A dashboard of students' tool usage could provide visualizations such as students' use of hallucination mitigation behaviors and other metrics such as time on task and accuracy of submission. This system could also be used to support student reflection after completing the activity. Seeing visualizations such as accuracy of the submission, number of prompts to the LLM, and use of hallucination mitigation behaviors could help students gain more awareness of how they using these tools, and perhaps motivate them to engage in more responsible use in the future.

8.3 Extension of Fogg Behavior Model to Future Work

In the discussion, we use the Fogg Behavior Model (FBM) as a lens to understand students' use of hallucination mitigation behaviors. The Fogg Behavior Model also provides insights into different types of triggers that can cause the user to perform the behavior: facilitators (ones that increase the user's ability), sparks (triggers that increase a user's motivation), and signals (triggers that indicate or remind users of the behavior) [16]. In this subsection, we explore future work targeted towards each of these dimensions.

Increasing students' ability to perform hallucination mitigation behaviors We suggest that one reasons students' may not have widely adopted hallucination mitigation behaviors for the citation-finding activity was due to how switching tools or checking each link might require more effort. Future work can explore the design of systems that make using hallucination mitigation behaviors easier for students to adopt. One method is to design tools that integrate hallucination mitigation behaviors directly into the students'. Companies like Google and OpenAI have begun to integrate hyperlinked icons into LLM generated outputs, so users can click on the icon and open the referenced article in a new page. Future work can explore whether these design choices promote hallucination mitigation. Additionally, collaboration between designers, researchers, and educators is important in designing systems that help develop important skills such as information, digital, and AI literacy that persist as these technologies continue to develop.

Increasing students' motivation to use hallucination mitigation behaviors In our study, students are graded based on the accuracy of their activity submissions. When students received immediate feedback on the current event fact-finding activity, students may have been more motivated to change their behaviors when they realized that not verifying LLM outputs resulted in them submitting incorrect answers. Automated grading and feedback can be helpful in provided instantaneous feedback on whether students' usage of LLMs aligns with the desired goal. Additionally, there may be social factors that influence whether students engage with hallucination mitigation behaviors such as

norms established in the classroom, peer expectations, and perceived instructor emphasis. Peer influence can act as both a motivator and a deterrent — when students see classmates actively cross-checking sources, they may be more inclined to adopt similar practices. Future work can explore how these different dimensions affect students' motivation to adopt hallucination mitigation behaviors.

Reminding students to use hallucination mitigation behaviors In our study, we provide a naturalistic exploration of how different LLM generated outputs may affect the hallucination mitigation behaviors of students. We found that different ways that the LLM would express certainty/uncertainty or reminders about verification did impact whether students used hallucination mitigation behaviors. Future work can more systematically explore the impacts that different types of LLM generated responses have on student usage of hallucination mitigation behavior. This type of just-in-time intervention could provide reminders about limitations of LLMs or suggest recommended hallucination mitigation behaviors. While platforms like ChatGPT have static disclaimers at the bottom of their page like "ChatGPT can make mistakes. Check important info," these reminders may only contribute to student awareness of hallucinations without creating understanding of potential causes or incentivizing actual mitigation behaviors. More integrated and responsive approaches can be explored to remind users about hallucinations and recommend behaviors to verify LLM-generated outputs.

9 CONCLUSION

In this paper, we sought to understand the impacts that a hallucination lesson has on students' knowledge and hallucination mitigation behaviors. Though we find that the lesson properly increased students' understanding of when and how hallucinations occur, the extent to which this understanding spurred hallucination mitigation behaviors was more mixed. Specifically, we find that student performance on a fact-finding activity drastically improved, whereas their performance on a citation-finding activity saw only moderate change. We contextualize these findings with existing literature on behavioral change to explore the factors that might have influenced student use of hallucination mitigation strategies. Finally, we describe future work in the domain of AI literacy focused on the development of hallucination lessons and methods to increase students' usage of hallucination mitigation behaviors. Overall, our findings emphasize that training critical and responsible users of AI tools will require more than simply increasing user understanding of AI.

ACKNOWLEDGMENTS

We thank the students who participated in the class, TAs who supported the course and data collection, and the institution's computational resources to support LLM access.

REFERENCES

[1] [n. d.]. meta-llama/Llama-3.1-405B · Hugging Face — huggingface.co. https://huggingface.co/meta-llama/Llama-3.1-405B. [Accessed 25-08-2025].
[2] 2024. Attorney General James Warns Voters Against Relying on AI Chatbots for Election Questions — ag.ny.gov. https://ag.ny.gov/press-release/2024/attorney-general-james-warns-voters-against-relying-ai-chatbots-election. [Accessed 26-08-2025].
[3] Ruba Abu-Salma, M. Angela Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. 2017. Obstacles to the Adoption of Secure Communication Tools. In *2017 IEEE Symposium on Security and Privacy (SP)*. 137–153. https://doi.org/10.1109/SP.2017.65
[4] Anne Adams and Martina Angela Sasse. 1999. Users are not the enemy. *Commun. ACM* 42, 12 (1999), 40–46.
[5] Erinn Atwater, Cecylia Bocovich, Urs Hengartner, Ed Lank, and Ian Goldberg. 2015. Leading Johnny to Water: Designing for Usability and Trust. In *Proceedings of the Eleventh USENIX Conference on Usable Privacy and Security* (Ottawa, Canada) *(SOUPS '15)*. USENIX Association, USA, 69–88.
[6] Susan B Barnes. [n. d.]. A privacy paradox: Social Networking in the United States. https://firstmonday.org/ojs/index.php/fm/article/view/1394/1312
[7] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[8] Alastair R. Beresford, Dorothea Kübler, and Sören Preibusch. 2012. Unwillingness to pay for privacy: A field experiment. *Economics Letters* 117, 1 (2012), 25–27. https://doi.org/10.1016/j.econlet.2012.04.077

[9] Cecilia Ka Yuk Chan and Wenxin Zhou. 2023. An expectancy value theory (EVT) based instrument for measuring student perceptions of generative AI. *Smart Learning Environments* 10, 1 (2023), 64.

[10] David J Chinn, Martin White, Jane Harland, Christopher Drinkwater, and Simon Raybould. 1999. Barriers to physical activity and socioeconomic position: implications for health promotion. *Journal of epidemiology and community health* 53, 3 (1999), 191.

[11] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883* (2023).

[12] Digital Education Council. 2024. What students want: Key results from DEC Global AI Student Survey 2024. https://www.digitaleducationcouncil.com/post/what-students-want-key-results-from-dec-global-ai-student-survey-2024

[13] Wildemarkes de Almeida da Silva, Luis Carlos Costa Fonseca, Sofiane Labidi, and José Chrystian Lima Pacheco. 2024. Mitigation of hallucinations in language models in education: A new approach of comparative and cross-verification. In *2024 IEEE International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 207–209.

[14] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis* 16, 1 (2024), 64–93.

[15] Daniella DiPaola, Blakeley H Payne, and Cynthia Breazeal. 2020. Decoding design agendas: an ethical design activity for middle school students. In *Proceedings of the interaction design and children conference*. 1–10.

[16] Brian J Fogg. 2009. A behavior model for persuasive design. In *Proceedings of the 4th international Conference on Persuasive Technology*. 1–7.

[17] D Ganesh, M Sunil Kumar, P Venkateswarlu Reddy, S Kavitha, and D Sudarsana Murthy. 2022. Implementation of AI Pop Bots and its allied Applications for Designing Efficient Curriculum in Early Childhood Education. *International Journal of Early Childhood Special Education* 14, 3 (2022).

[18] Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904* (2024).

[19] https://www.theguardian.com/profile/matthewweaver. [n. d.]. AI chatbots distort and mislead when asked about current affairs, BBC finds — theguardian.com. https://www.theguardian.com/technology/2025/feb/11/ai-chatbots-distort-and-mislead-when-asked-about-current-affairs-bbc-finds. [Accessed 26-08-2025].

[20] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55.

[21] Iulia Ion, Rob Reeder, and Sunny Consolvo. 2015. "...No one Can Hack My Mind": Comparing Expert and Non-Expert Security Practices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*. USENIX Association, Ottawa, 327–346. https://www.usenix.org/conference/soups2015/proceedings/presentation/ion

[22] Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of large language models in finance: An empirical examination of hallucination. *arXiv preprint arXiv:2311.15548* (2023).

[23] Kashifa Khalid, Netta Iivari, Marianne Kinnula, and Sumita Sharma. 2022. Familiarizing children with artificial intelligence. In *Proceedings of the 25th International Academic Mindtrek Conference*. 372–376.

[24] Abby C King, Cynthia Castro, Sara Wilcox, Amy A Eyler, James F Sallis, and Ross C Brownson. 2000. Personal and environmental factors associated with physical inactivity among different racial–ethnic groups of US middle-aged and older-aged women. *Health psychology* 19, 4 (2000), 354.

[25] Siu-Cheung Kong, William Man-Yin Cheung, and Olson Tsang. 2023. Evaluating an artificial intelligence literacy programme for empowering and developing concepts, literacy and ethical awareness in senior secondary students. *Education and Information Technologies* 28, 4 (2023), 4703–4724.

[26] Joe Kottke. [n. d.]. Beyoncé announces 'Cowboy Carter' tour dates after Grammys win — nbcnews.com. https://www.nbcnews.com/pop-culture/pop-culture-news/beyonce-announces-cowboy-carter-tour-ahead-grammys-night-rcna190318. [Accessed 25-08-2025].

[27] 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.

[28] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[29] 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.

[30] Tammy McCausland. 2020. The bad data problem. , 68–71 pages.

[31] Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. 2021. Prevent the language model from being overconfident in neural machine translation. *arXiv preprint arXiv:2105.11098* (2021).

[32] 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.

[33] Chinasa T Okolo. 2024. Beyond AI hype: A hands-on workshop series for enhancing AI literacy in middle and high school students. In *Proceedings of the 2024 on RESPECT Annual Conference*. 86–93.

[34] Yasumasa Onoe, Michael JQ Zhang, Eunsol Choi, and Greg Durrett. 2022. Entity cloze by date: What LMs know about unseen entities. *arXiv preprint arXiv:2205.02832* (2022).

[35] Gabrijela Perković, Antun Drobnjak, and Ivica Botički. 2024. Hallucinations in llms: Understanding and addressing challenges. In *2024 47th MIPRO ICT and electronics convention (MIPRO)*. IEEE, 2084–2088.

[36] Aimee Picchi. 2024. AI chatbots are serving up wildly inaccurate election information, new study says — cbsnews.com. https://www.cbsnews.com/news/ai-chatbots-inaccurate-election-information-proof-news/. [Accessed 26-08-2025].

[37] G Pradeep Reddy, YV Pavan Kumar, and K Purna Prakash. 2024. Hallucinations in large language models (LLMs). In *2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*. IEEE, 1–6.

[38] James F Sallis and Neville Owen. 1998. *Physical activity and behavioral medicine*. SAGE publications.

[39] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. 2017. Self-Confidence Trumps Knowledge: A Cross-Cultural Study of Security Behavior. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2202–2214. https://doi.org/10.1145/3025453.3025926

[40] Jaemarie Solyst, Emily Amspoker, Ellia Yang, Motahhare Eslami, Jessica Hammer, and Amy Ogan. 2025. RAD: A Framework to Support Youth in Critiquing AI. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*. 1071–1077.

[41] Jaemarie Solyst, Alexis Axon, Angela EB Stewart, Motahhare Eslami, and Amy Ogan. 2023. Investigating girls' perspectives and knowledge gaps on ethics and fairness in Artificial Intelligence in a Lightweight workshop. *arXiv preprint arXiv:2302.13947* (2023).

[42] Josh Taylor. [n. d.]. Australian lawyer caught using ChatGPT filed court documents referencing 'non-existent' cases — theguardian.com. https://www.theguardian.com/australia-news/2025/feb/01/australian-lawyer-caught-using-chatgpt-filed-court-documents-referencing-non-existent-cases. [Accessed 26-08-2025].

[43] Gershon Tenenbaum and Robert C Eklund. 2020. *Handbook of sport psychology*. John Wiley & Sons.

[44] Stewart G Trost, Neville Owen, Adrian E Bauman, James F Sallis, and Wendy Brown. 2002. Correlates of adults' participation in physical activity: review and update. *Medicine & science in sports & exercise* 34, 12 (2002), 1996–2001.

[45] U.S. News & World Report. 2025. Meta seeks urgent fix to AI Chatbot's confusion on name of US president. https://www.usnews.com/news/top-news/articles/2025-01-23/meta-seeks-urgent-fix-to-ai-chatbots-confusion-on-name-of-us-president. Accessed: 2025-08-25.

[46] Gaby Del Valle. [n. d.]. Why do lawyers keep using ChatGPT? — theverge.com. https://www.theverge.com/policy/677373/lawyers-chatgpt-hallucinations-ai. [Accessed 02-09-2025].

[47] Krzysztof Walczak and Wojciech Cellary. 2023. Challenges for higher education in the era of widespread access to Generative AI. *Economics and Business Review* 9, 2 (2023).

[48] Randi Williams. 2021. How to train your robot: project-based ai and ethics education for middle school classrooms. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 1382–1382.

[49] Helen Zhang, Irene Lee, Safinah Ali, Daniella DiPaola, Yihong Cheng, and Cynthia Breazeal. 2023. Integrating ethics and career futures with technical learning to promote AI literacy for middle school students: An exploratory study. *International Journal of Artificial Intelligence in Education* 33, 2 (2023), 290–324.

[50] Xiaofei Zhou, Jessica Van Brummelen, and Phoebe Lin. 2020. Designing AI learning experiences for K-12: Emerging works, future opportunities and a design framework. *arXiv preprint arXiv:2009.10228* (2020).

A COURSE OVERVIEW

Students were in class everyday from 9am-3pm, received a letter grade, and college credit. The course is an introductory-level course that provides students a foundational understanding of the technical, socio-ethical, and career development dimensions of AI literacy, specifically on Large Language Models (LLMs). Each week of the course covered a different dimension of AI literacy and each day focused on different topics within that dimension (Table 2).

Week	Topics Covered by Day
Week 1: Technical Dimension	Day 1: Introduction to GenAI and LLMs
	Day 2: Data Collection and Processing
	Day 3: Pre-Training
	Day 4: Post-Training (Decoding Methods, RLHF, Supervised Finetuning)
	Day 5: Prompt Engineering and Student Project Presentations
Week 2: Socio-Ethical Dimension	Day 6: Bias in LLMs
	Day 7: Misinformation
	Day 8: Environmental Impact
	Day 9: Attribution, Copyright, and Governance
	Day 10: Student Project Presentations
Week 3: Career Development Dimension	Day 11: Impact of LLMs in SWE and Finance Careers
	Day 12: LLMs on Education
	Day 13: LLMs in Art
	Day 14: Student Project Presentations

Table 2. Course Overview by Week and Daily Lesson Topics

B ACTIVITY INTERFACE SETTINGS

In table 3 below, we provide a description of all the features that are available in the interface, the default settings, as well as the design decisions behind the interface design.

Icon/Label	Name	Description	Features	Default
Activity Description	Activity Description	Provides instructions and context for the activity.	Collapsible textbox	Open
Model	LLM Model Selection	Students choose from available models (e.g., GPT-4o, GPT-4o-mini, GPT-4, GPT-3.5-turbo, DeepSeek-R1/V3, Llama3.1-405B/80B, Llama2-70B).	Dropdown menu	GPT-4o
Controls	Parameter Controls	Students adjust model parameters (top_p, top_k, temperature, max_tokens).	Expandable section; sliders	Provider defaults
System Prompt	System Prompt	Allows students to specify a custom system prompt.	Expandable section	Closed
Red Trashcan Icon	Clear Chat History	Clears the current chat conversation and resets context.	Button	N/A
Display Chat Playground	Chat Playground Toggle	Show or hide the LLM chat-playground interface.	Toggle	Display on
Chat-Input	Chat Input	Textbox for students to draft prompts to the LLM.	Free-text input	Empty
Assignment Editor	Assignment Editor	Text editor where students draft responses to the activity.	Text editor	Empty
Submit Assignment	Submit Assignment	Button for students to submit their assignment response.	Submission button	N/A

Table 3. Overview of Student Interface Features

C PRE- AND POST-LESSON ACTIVITY PROMPTS

C.1 Pre-Lesson Current Event Fact-Seeking Prompts

Pre-Lesson Current Event Question	Rationale
1. What team is basketball player Luka Dončić playing for?	In February 2025, Luka Dončić was traded from the Dallas Mavericks to the Los Angeles Lakers. This trade would not be in the training data set for the provided models.
2. Which animated movie won the Oscars in 2025? Return the title only. (Flow, Inside Out 2, Memoir of a Snail, Wallace & Gromit: Vengeance Most Fowl, The Wild Robot)	This question would not be in the training data set for the provided models.
3. Which animated movie won the Oscars in 2018?	This question would be in the training set of the provided models since it occurred before 2023. We used this question to insights into the types of time-based questions an LLM might be able to answer correctly.
4. True or False: Beyoncé came to Chicago during her Cowboy Carter Tour.	In February 2025, Beyoncé announced her tour dates. These dates would not be in the training data set of LLMs.

Table 4. Set of questions students were asked to answer for the **pre-lesson** current events fact-seeking activity, with rationale.

C.2 Post-Lesson Current Event Fact-Seeking Prompts

Post-Lesson Current Event Question	Rationale
1. True or False: Sabrina Carpenter is no longer performing at Lollapalooza this year.	In March 2025, organizers revealed the lineup for Lollapalooza 2025. This information, or Sabrina's potential withdrawal, would not be included in the training dataset of the provided models.
2. Which fiction book won the Pulitzer Prize for Fiction in 2024? Return the title only. (Night Watch, Wednesday's Child, Same Bed Different Dreams)	This question would not be in the training data set for the provided models.

Table 5. Set of questions students were asked to answer for the **post-lesson** current events fact-seeking activity, with rationale.

C.3 Pre- and Post-Lesson Citation Seeking Prompts

Pre-Lesson Prompt	Post-Lesson Prompt
This week, we are exploring how you can enact change in your local community through the community advocacy project to share knowledge with people in your communities. In our daily lives, there are many different people who take on the role of information-sharers, like journalists and media companies. Recently, many people have been using social media as a way to stay up-to-date on current events. Influencers in tech and politics use their social media platforms to share information, which raises the question: "Should influencers be held to the same ethical standards as journalists?" For this activity, state the position you want to take (e.g., influencers should or should not be held to the same ethical standards as journalists). Then, cite 10 different sources that you might use in your argument to justify your point.	As artificial intelligence becomes more integrated into daily life, many teenagers interact with AI chatbots for social support, companionship, and advice. These chatbots are powered by large language models trained through techniques like Reinforcement Learning from Human Feedback (RLHF). RLHF is designed to help AI respond more helpfully and appropriately by learning from human preferences. However, this raises concerns about the influence these AI "friends" may have on young, impressionable users, especially since they may not challenge harmful ideas or provide balanced, critical feedback. Should students under the age of 18 be allowed to use AI chatbots as "friends" or sources of advice? For this activity, state the position you want to take (e.g., people under 18 should not have access to AI chatbots designed to be friends). Then, cite 10 different sources you might use to justify your point.

Table 6. Pre and post-lesson activity prompts for the citation-finding activity.

D QUALITATIVE CODING CODEBOOKS

Category	Description	Example Student Quote
Hallucination due to Inference: Reasoning	Student describes LLM's inability to understand, reason about, or evaluate the logical consistency of statements.	Since there is no way for AI to... think beyond, they can generate inaccurate responses.
Hallucination due to Training: Sycophancy	Student describes LLM's tendency towards responses that prioritize the user over logical soundness or truthfulness.	The AI is coded so that it does whatever it can to make the user happy, and so it may generate some fake stuff if it can't find what the user is looking for.
Hallucination due to Training Data: Inaccurate Training Data	Student describes the potential for inaccurate training data to affect LLM response correctness.	One reason this might happen is the AI is utilizing inaccurate information from sources that are not credible.
Hallucination due to Training Data: Knowledge Cutoff	Student describes how an LLM may have been trained on now outdated information.	It could also be when the AI database is not updated, like how the very first publicized ChatGPT still assumes Queen Elizabeth is alive.
Hallucination due to Inference: Imperfect Decoding Strategies	Student describes LLM generating inaccurate or incorrect information due to decoding methods not explicitly checking for accuracy.	The primary goal of an LLM is to output coherent text, which may involve hallucinating false information.

Table 7. Tagging Scheme for Student Pre-/Post-lesson Understanding of Hallucinations

Code	Description	Example
No uncertainty	LLM outputs an answer to the question with no further elaboration related to training data or verification.	Luka Dončić plays for the Dallas Mavericks in the NBA.
Implies incomplete knowledge	LLM outputs an answer to the question and mentions the scope of its training data, but does not express explicit doubt or suggest verification.	False. Beyoncé has not had a tour called "Cowboy Carter Tour" **as of my knowledge cutoff in 2023**.
Suggests verification	LLM outputs an answer to the question, along with an explicit suggestion to check with other sources.	Luka Dončić currently plays for the Dallas Mavericks in the NBA. He has been with the Mavericks since being drafted in 2018 and has established himself as one of the league's top players. **For the most accurate and up-to-date information, feel free to double-check with recent news sources!**
Refuses to answer	LLM does not output an answer to the question, stating that the question requires information beyond its knowledge cutoff.	**I'm unable to provide information about the 2025 Oscars as my knowledge cutoff is October 2023**.

Table 8. Tagging Scheme for LLM Responses to Current Events Fact-Seeking Questions

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Code	Description
All checked	Links are pasted in one at a time. Student reflection describes verification. OR Links are pasted in batches, but student leaves page multiple times after each batch. Faulty links (if they existed) are then deleted and replaced with new ones.
Some checked	Links are pasted in batches. Behavior changes by batch. Some batches have evidence of being checked, but others do not and contain faulty links.
None checked	Student pastes all links in at once. After links are pasted in, student submits assignment. OR Student pastes in links one at a time. Some links are faulty. Reflection does not describe verification.
Unclear	Links are pasted in one at a time. All links work. Student reflection does not describe verification.

Table 9. Tagging Scheme for Citation Verification Behavior

Code	Description	Example from Student Submission
Missing	Student cites a resource that does not exist.	"Blurred Lines: Exploring the Critical Intersection Between Journalism and Marketing" (Journal of Media Ethics, Volume 34, Issue 3.)
Not Specific	Student cites a resource in a way that is too vague to be useful, like a journal or organization name instead of a specific article.	https://www.pbs.org/
Not Relevant	Student cites a resource that is clearly unrelated to the assigned topic.	Why Yale Law School Left the U.S. News & World Report Rankings (Article by *The Atlantic*)
Not Credible	Student cites a resource that does not need to follow high standards of correctness, like a personal website or blog, and does not cite sources.	[Permalink to Linkedin Post with no sources cited – link redacted by authors]

Table 10. Tagging Scheme for Citation Quality. Citations were checked for the following four issues from top to bottom, with higher levels being easier for students to check. Citations that had none of the four issues were marked as correct.