"Here the GPT made a choice, and every choice can be biased": How Students Critically Engage with LLMs through an End-User Auditing Activity

Snehal Prabhudesai
snehalbp@umich.edu
University of Michigan
Ann Arbor, MI, USA

Ananya Kasi
akasi@umich.edu
University of Michigan
Ann Arbor, MI, USA

Anmol Mansingh
anmolmsg@umich.edu
University of Michigan
Ann Arbor, MI, USA

Anindya Das Antar
adantar@umich.edu
University of Michigan
Ann Arbor, MI, USA

Hua Shen
huashen@uw.edu
University of Washington
Seattle, WA, USA

Nikola Banovic
nbanovic@umich.edu
University of Michigan
Ann Arbor, MI, USA

Abstract

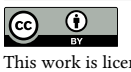
Despite recognizing that Large Language Models (LLMs) can generate inaccurate or unacceptable responses, universities are increasingly making such models available to their students. Existing university policies defer the responsibility of checking for correctness and appropriateness of LLM responses to students and assume that they will have the required knowledge and skills to do so on their own. In this work, we conducted a series of user studies with students (N=47) from a large North American public research university to understand if and how they critically engage with LLMs. Our participants evaluated an LLM provided by the university in a quasi-experimental setup; first by themselves, and then with a scaffolded design probe that guided them through an end-user auditing exercise. Qualitative analysis of participant think-aloud and LLM interaction data showed that students without basic AI literacy skills struggle to conceptualize and evaluate LLM biases on their own. However, they transition to focused thinking and purposeful interactions when provided with structured guidance. We highlight areas where current university policies may fall short and offer policy and design recommendations to better support students.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Artificial intelligence*.

Keywords

End-user Audit, End-user Algorithmic Audit, User-Driven Algorithm Auditing, Algorithmic Audit, Auditing Algorithms, Algorithmic Bias, Algorithmic Harm, Large Language Models, LLMs, AI Literacy, AI Education, Responsible AI.



CHI '25, April 26-May 1, 2025, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/2025/04
https://doi.org/10.1145/3706598.3713714

ACM Reference Format:
Snehal Prabhudesai, Ananya Kasi, Anmol Mansingh, Anindya Das Antar, Hua Shen, and Nikola Banovic. 2025. "Here the GPT made a choice, and every choice can be biased": How Students Critically Engage with LLMs through an End-User Auditing Activity. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3706598.3713714

Content Warning: This paper covers user-led audits of a Large Language Model (LLM). Parts of this paper reference user-generated content containing offensive or hateful speech, profanity, and content pertaining to potentially triggering topics.

1 Introduction

Critical engagement with AI [40, 60] enables end-users to analyze, question, and contest information generated from Large Language Models (LLMs) in consequential settings (i.e., environments where the outcomes significantly impact individuals or society), such as education [58, 113]. By developing AI literacy competencies [73] that help understand AI limitations and ethical implications, end-users can exercise critical thinking to evaluate errors [123], biases [85], and "hallucinations" [56, 124] of LLMs—believable but factually inaccurate responses. For example, users can exercise critical thinking to determine the reliability of LLM-generated information [116] and safeguard themselves from misinformation [63, 75].

Having recognized that LLMs can generate incorrect or unacceptable responses, universities increasingly caution students (i.e., individuals who attend the university to study and learn, such as undergraduates, graduate students, and fellows) to verify LLM outputs for accuracy and appropriateness. However, existing university policies [13, 45, 47, 51, 90, 105, 112] do not specify *how* to evaluate LLMs or *what* to look for in their responses. Those guidelines assume students already have the required skills to evaluate LLMs, causing stress and uncertainty for those who lack those skills [1, 20].

Although students with computer science backgrounds could have sufficient AI literacy [50], the lack of similar levels of AI literacy in students from non-technical backgrounds [62, 110] places them at a significant disadvantage under current guidelines, exacerbating academic inequalities [40]. Universities increasingly offer coursework that teaches students across disciplines what LLMs are [52, 58] and how to use them [77, 80]; yet, most such coursework similarly only cautions students about the limitations of LLMs. Thus, current university policies force students to shoulder the responsibility of evaluating LLMs on their own; and failing that, risk academic penalties, including charges of plagiarism and expulsion [23].

Unfortunately, LLMs generate responses that project a sense of authority [54] and infallibility [18], which could lead end-users to overrely on such technology [8, 16, 33, 53]. Although there is a potential for end-users to identify [29, 107] and document [39] AI bias through end-user audits [68, 103], with a few notable cases of end-users successfully auditing AI systems on their own [49, 59], it is unclear if and to what degree such methods lead to methodical investigation of LLMs among students to support them in shouldering the burden of evaluation.

In this work, we conducted a formative study to investigate *how* and to *what* extent students can critically engage with LLMs to identify biases and evaluate harms on their own. We answer the following research questions:

RQ1: To what extent are students able to identify harmful biases in LLMs by themselves following existing university policy and recommendations?

RQ2: How does providing students with an end-user auditing scaffolding affect their approach to identifying and documenting bias in LLMs?

RQ3: What are the design opportunities for helping non-technical and low AI literacy students to deal with biases in LLMs?

To answer these research questions, we developed a design probe named PromptAuditor, featuring a main interface to chat with an LLM and a scaffolding informed by CS education experts. We adopted a quasi-experimental study design[1] from education and policy literature [22, 102] to develop an end-user auditing protocol for identifying biases in LLMs. In our study design, participants performed an end-user audit [29, 68, 103] of an LLM using PromptAuditor in three stages: 1) first with minimal guidance, akin to existing university guidelines, 2) then with a scaffolding that guided participants step-by-step through a worked auditing example of LLM bias, and 3) then again independently, with the scaffolding removed.

We instantiated this protocol across two user studies with participants (N=47) at a large North American public research university. In the first study, we recruited students from diverse academic disciplines (N=8) by word of mouth, to study their cognitive and decision-making processes while interacting with our probe, which we elicited in a lab study using a think-aloud protocol [55]. In the second study, we studied *in situ* behavior of students from two different cohorts with different backgrounds and levels of AI literacy; one with classroom students (N=30) at the end of a semester-long course on Generative AI from an arts and sciences department, and the other with attendees of an AI workshop for journalism fellows (N=9), after an hour-long introductory AI lecture. In both studies, we collected and performed qualitative analysis of interaction data with the study software and GPT (i.e., Generative Pre-trained Transformers) conversation logs.

Our results highlight that students with stronger AI literacy and technical backgrounds are better equipped to conduct comprehensive evaluation of biases in AI systems. Structured guidance provides a focused scope for thinking about bias hypothesis as students transition from confusion and struggle to understanding sociotechnical aspects of bias propagation. After receiving such guidance, students' interaction with LLMs transitions from random prompting on diverse subjects to focused evaluation of biases in a specific domain.

Our work contributes empirical knowledge on how students critically engage in everyday algorithmic auditing of LLMs. Our work points to specific gaps in existing university policies, such as a lack of consideration of students' diverse levels of AI literacy, which affects their ability to critically engage with LLMs. Our work motivates future interface designs that can promote critical thinking in everyday interactions. We invite researchers to adopt a learner-centered lens when designing interfaces and policies that support critical engagement.

2 Background and Related Work

Critical engagement with LLMs is an active topic of research in various communities, from pedagogy [73, 86] and educational policies [19, 126] to Human-Computer Interaction (HCI) [34]. Here, we first highlight literature from cognitive psychology, pedagogy, and computer science education to explain the role of AI literacy in fostering critical engagement. Next, we briefly discuss gaps in current university LLM policies, which defer critical evaluation to students. Then, we point to existing approaches to support users in evaluating AI technologies.

2.1 Critical Thinking and AI Literacy

Critical thinking [35, 41] is a higher-order, analytical thinking process [114] that requires deliberate effort to analyze, evaluate and judge the credibility of information [57, 65]. Educational psychology and cognitive science [83, 98] state that both declarative (knowing *what*) and procedural (knowing *how*) knowledge are required for critical thinking in educational settings [4]. Exercising critical thinking when interacting with LLMs makes users less susceptible to errors [123], biases [85], and "hallucinations" [56, 124]. Critical thinking can induce healthy skepticism and promote appropriate reliance on AI [95]. This can safeguard users from harms such as misinformation [63, 75] and conspiracy theories [91].

AI literacy [73, 86] enables end-users to apply foundational skills of critical thinking to evaluate AI technologies. However,

[1]While allowing for comparison of pre-and-post guidance, a quasi-experimental design is more aligned to naturalistic settings where random assignment to control and treatment groups may be impracticable or unethical.

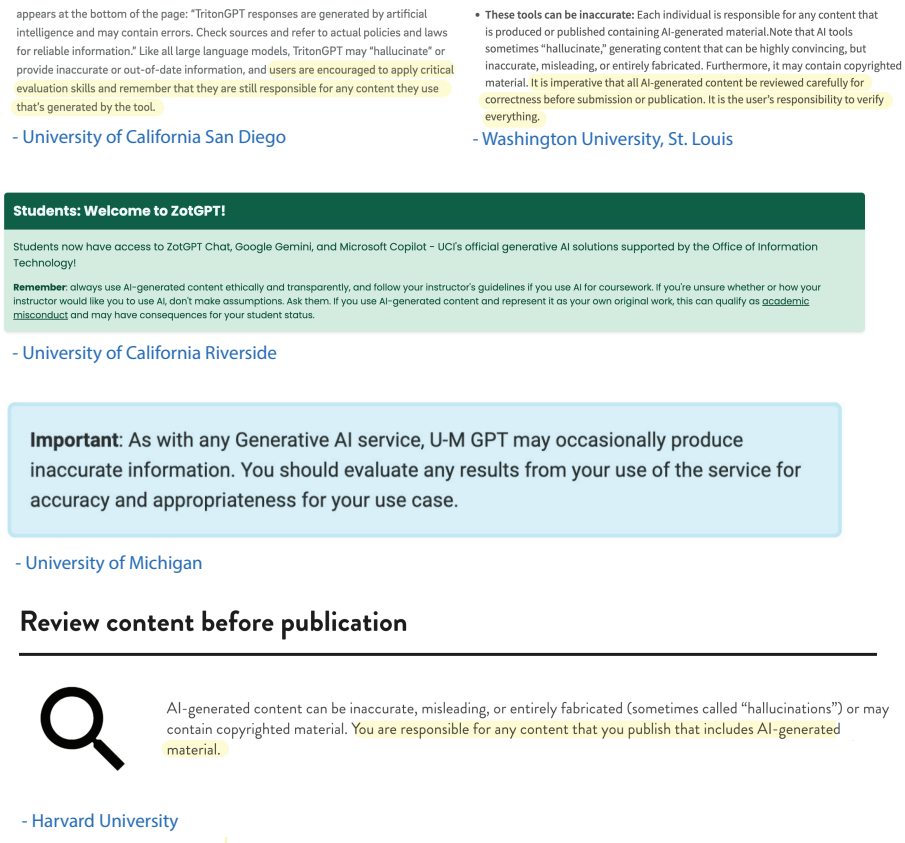


Figure 1: Examples of university guidelines [13, 51, 90, 105, 112] regarding the use of GPT tools. The guidelines strongly indicate the student's responsibility to review all AI-generated content for "appropriateness" without stating how.

an overwhelming majority of the public is not AI literate [120]. Raising public AI literacy requires deliberate effort from design, policy, and education [74] to engage the public using both formal [58, 113] and informal [71, 72, 87] educational interventions. Such efforts have the potential to engage a broader set of stakeholders, including children [25, 115], middle schoolers [9], and youth [125], to acquire AI literacy skills.

Educational institutions are increasingly offering structured classroom instruction to teach LLMs and related technologies, covering some skills and knowledge for interacting with generative AI [5, 52, 121]. However, even if those courses teach *what* biases and other undesirable outcomes LLMs can produce, that may not immediately transfer to practical knowledge on *how* to identify and document those biases and outcomes.

2.2 Educational Policies for LLMs

With the advent of ChatGPT in November 2022 [92], universities lacked clear guidelines and policies for the use of LLM technologies in academic settings [32, 119]. After educators expressed concerns (e.g., potential for plagiarism [84], decline in students' critical thinking skills due to excessive GPT use [20], and unreflected acceptance of GPT responses [64]), many universities formed interdisciplinary committees [88, 89] to draft comprehensive policies. Such committees involved educators, technologists, and ethicists, with students' voices largely absent [19, 126]. Being designed as an extension of academic dishonesty policies [46, 47, 79], the consequences of not following those university guidelines on LLMs can result in charges of academic dishonesty and even expulsion [23].

Most existing university guidelines [13, 45, 47, 51, 90, 105, 112] caution students to check LLM responses for accuracy and appropriateness (Figure 1). Those policies not only burden students with evaluating and debugging a still experimental and error-prone computational technology, but also assume that all students have high AI literacy and relevant competencies to perform this difficult task [14]. Few of those policies, if any, recognize the ability of AI to deceive [8], project infallibility [18], and affect students' views [54]. This points to a crucial need to investigate *how* and to *what* extent students can effectively follow the existing guidelines on their own and avoid any repercussions of wrong or undesirable LLM outputs.

2.3 Interactive Tools for AI Evaluation

The HCI research community [3, 31, 68, 118] has long recognized the importance of supporting end-users in evaluating AI in everyday interactions. By providing insights into the AI's inner-working and decision-making, explanation mechanisms that promote transparency [96, 104] are meant to help end-users identify and reject AI decisions in situations in which its reasoning was incorrect [101]. However, most existing explanation mechanisms tend to act as evidence of reliability rather than accountability [16, 26], often deceiving end-users into over-relying on AI [53, 67]. Also, such mechanisms do not directly translate to the context of LLMs [34], where problems of over-reliance could be exacerbated [54].

Algorithmic auditing methods [15, 81] have been effective in identifying and documenting the weaknesses of algorithmic systems deployed in both public [37] and private [97] sectors. While traditionally catering to system developers [7, 100], recent work has focused on the potential for adopting the methods for end-users to identify and document AI bias [27, 68, 107]. Work on everyday algorithm auditing [29, 103] has documented how end-users audit AI in everyday interactions, including how it helps raise their awareness and ability to hypothesize, evaluate, and surface harm. However, it is unclear if and to what degree such methods can lead to methodical LLM evaluation among students to support them in shouldering the burden of critical engagement with LLMs.

3 Method for Probing Students Declarative and Procedural Knowledge of Auditing

We studied how and to what extent students (i.e., individuals who attend the university to study and learn) can critically engage with LLMs. We explored how their declarative knowledge (i.e., knowing *what*) and procedural knowledge (i.e., knowing *how*) impact their cognitive processes and interaction behavior when auditing an LLM. To answer our research questions, we used a quasi-experimental study design (Fig. 2) where students with different levels of AI literacy interacted with our study probe, PromptAuditor. Henceforth, in Sections 3 and 4, we refer to students who took part in our studies as "participants".

3.1 Operationalizing University Guidelines

We conducted our user studies at the University of Michigan (U-M), a large public research university in North America. Since 2023, U-M has invested $180 million in a partnership with Microsoft and OpenAI [78, 94] to be one of the first institutions to develop a suite of Generative AI (GenAI) models for use within the university. Despite students and staffs' concerns [109], the university cited privacy and innovation as its key motivations for developing this suite. In addition to GenAI tools, U-M released learning resources, including courses and online materials. We used the GPT-3.5 Turbo model and interface (called "U-M GPT") from this suite for our study.

The real-world university setting enhanced ecological validity and allowed us to observe students' interactions with LLMs in an authentic learning environment. We further operationalized and integrated U-M's high-level guidelines (i.e., acknowledging that U-M GPT may produce biased, harmful, or inaccurate information) into the design of study probes and tasks. We framed critical evaluation as an everyday end-user auditing activity [68, 103], as it closely mirrors real-world scenarios where students must critically engage with AI tools without extensive prior training. We designed PromptAuditor as a design probe to collect data while auditing LLMs.

3.2 PromptAuditor

Here, we describe the design elements, rationale, and implementation of our study probe, PromptAuditor (Fig. 3). Participants interacted with the main interface A in all study stages (Fig. 2): 1) pre-scaffolding, 2) with-scaffolding, and 3) post-scaffolding. Scaffolding B was activated during the "with-scaffolding" stage and deactivated afterward.

3.2.1 Main Interface Design. The main interface consists of: A1 an audit report panel, and A2 a "U-M GPT" chat interface. The audit report panel A1 , inspired by the IndieLabel system [68], allows participants to document findings by filling out topic, evidence, and summary fields. The chat interface A2 allows participants to issue prompts and view responses from the underlying GPT-3.5-Turbo model. Note that the design of the chat interface A2 was based on the existing interface of U-M GPT (Fig. 7). We piloted different main interface layouts with non-technical participants (N=5) and confirmed their preference for the familiar, chat-based layout A2 .

3.2.2 Scaffolding Design and Rationale. To develop our scaffolding, we consulted AI education experts and identified scaffolding principles: B1 scenario-based learning [106], B2 hypothesis generation [103], B3 worked examples [6], B4 self-reflection [76], and B5 contrastive learning [43]. We created initial designs, exploring scenarios relevant to student life and consulting literature for harms [44], such as student loans [93], health insurance, and hiring [12]. Through sessions with AI auditing expert co-authors, we generated worked examples and selected the most salient one for hiring bias [69]. For contrastive learning, we tested three prototypes: 1) manual annotation, 2) automated GPT annotation, and 3) a diff checker to compare and highlight differences between two LLM output versions. Pilot participants preferred manual annotations for their clarity and simplicity, but found GPT annotations more helpful for revealing subtle contextual differences.

Our final design combined GPT-generated highlights with the manual review featuring toggled highlights. We iteratively critiqued the low-fidelity prototypes with our interdisciplinary team. We tested the final design with two participants, which revealed no major usability issues. Our scenario-based learning scaffolding B1 used a scenario of racial bias in the hiring domain [111]. Here, our worked examples B3 used two cover letters—one with a Caucasian sounding name ("Christopher Allen"), and the other with an African American sounding

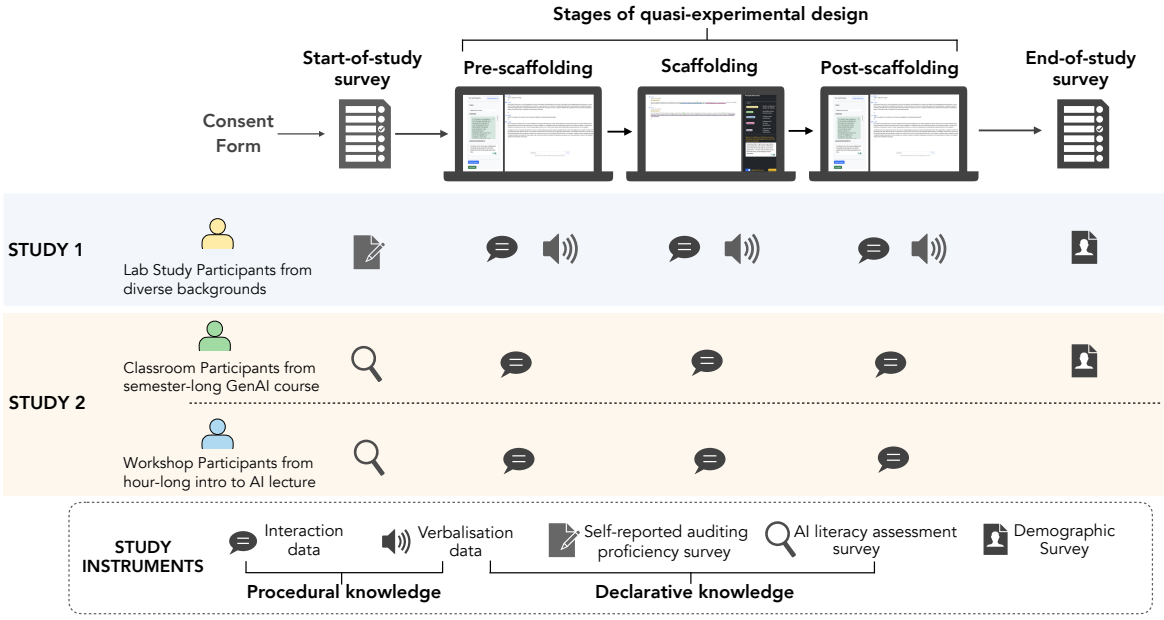


Figure 2: Our quasi-experimental study design across two user studies: 1) lab-controlled think-aloud study (N=8), and 2) naturalistic educational environments study (N=39). Symbols in the legend represent the study instruments used to collect study data and measure different knowledge aspects at different stages of the study.

name ("Latisha Smith"), referenced from prior correspondence audit work [69] and field experiment on labor market discrimination [12]. After providing self-explanations B4, participants engaged in contrastive learning B5 to examine this bias.

3.2.3 Software Implementation. Our foremost consideration during the development of the study software was protecting students' privacy. Although we initially developed a Chrome extension, AI education experts raised concerns about potential risks such as the extension having access to browsing history and other sensitive data. They were also concerned about students installing software on personal devices without fully understanding the risks.

To address these concerns, we migrated to a Django[2] web application, which does not require installing software on personal devices. We replicated the U-M GPT interface using HTML, CSS, and JavaScript[3], with participant data securely stored in a MySQL[4] database. For the backend, we accessed the GPT-3.5 Turbo[5] LLM via the U-M GPT Toolkit[6], an API gateway provided by the university that grants programmatic access to the same models U-M GPT used.

[2]https://www.djangoproject.com/
[3]https://www.w3.org/wiki/The_web_standards_model_-_HTML_CSS_and_JavaScript
[4]https://www.mysql.com/
[5]https://platform.openai.com/docs/models#gpt-3-5-turbo
[6]https://its.umich.edu/computing/ai

3.3 Overview of User Studies

We conducted two user studies (Fig. 2): 1) a lab-controlled think-aloud study (N=8), and 2) a study in naturalistic educational environments (N=39). We used the lens of integrated knowledge theory [4] to guide our study design, data collection, and analysis. This theory offers an educational framework that combines the acquisition of declarative and procedural knowledge with an understanding of students' learning performance. We employed a quasi-experimental study design [22, 102], which has been traditionally employed in learning science studies to assess student knowledge [42, 99].

We conducted both studies with students at the same North American public research university, which LLM guidelines we operationalized (Section 3.1). In both studies, participants used PromptAuditor to audit U-M GPT in an end-user auditing activity [29, 68, 103]. After accessing the study website, participants reviewed the consent form, completed a pre-study survey, and watched a brief video on GPT usage.

PromptAuditor then asked them to reflect on potential LLM harms. In the pre-scaffolding stage, participants performed a 10-minute audit using the main interface with minimal guidance and without the scaffolding. During the scaffolding stage, the scaffolding guided them through a worked example of LLM bias. In the post-scaffolding stage, participants completed another 10-minute audit on their own. Below, we outline differences between the studies, including participants, methods, data collection, and analysis.

Figure 3: Our study interface probe, which we call PromptAuditor has two parts: (A) the main interface, and (B) scaffolding. Participants interact with (A) in pre-and-post scaffolding phases, using (A1) to type prompts and read GPT response, and (A2) for bias documentation. In the scaffolding phase, (B) is turned on, which guides participants step-by-step (B1 - B5) through a worked auditing example.

3.4 Study 1: Lab-controlled Think-aloud

We studied participants' cognitive decision-making processes using a think-aloud protocol in a controlled lab environment.

3.4.1 Participants. We recruited eight participants using flyers and word-of-mouth. All participants were 18 years or older, and studied in diverse fields, ranging from Biology to Software Development (Table 1). We stopped recruiting after reaching data saturation. We compensated participants $15/hr for up to two hours, with the average study time being 60 minutes. Three participants self-identified as men, and five as women.

3.4.2 Study-specific Tasks and Procedures. We conducted the study in-person or via Zoom [7], depending on participant availability, with no other differences between sessions. Each session had two investigators: one conducting the study and one taking notes. After welcoming participants, the investigator opened the study website or shared the link for remote sessions, obtained consent, and explained the think-aloud [55] protocol using a brief video tutorial [8]. Participants shared their screens via Zoom and completed the task (Section 3.3) while thinking aloud during their sessions.

3.4.3 Data Collection. We collected data on procedural knowledge (via interaction data and verbalization) and on declarative knowledge (via initial survey and verbalization). We conducted a self-reported auditing proficiency survey noting participants' frequency of GenAI use, auditing confidence, examples of encountered biases, and their views on bias in GPT tools and potential harms. During the study, we recorded audio and screens, collected interaction data from the web interface (e.g., prompts, responses, audit reports), and took notes on the think-aloud process. Afterward, we asked follow-up questions and collected demographic data, including gender [108], age, education, and occupation, through an end-of-study survey.

3.5 Study 2: Educational Environments

We studied participants' natural interaction behavior across two cohorts with different levels of exposure to formal AI instruction and varying AI literacy skills: 1) those enrolled in a semester-long course on GPT, and 2) those from an hour-long intro to AI workshop.

3.5.1 Participants. We recruited two cohorts of participants, which we picked for their differing levels of exposure to formal AI instruction and varying AI literacy skills: 1) classroom participants and 2) workshop participants. The classroom participants (N=30) were students from diverse academic backgrounds who had enrolled in a semester-long course on Generative AI, focusing on GPT. This course, offered by the university's Program in Computing for Arts and Sciences[9], was specifically designed for non-technical students and did not require prior programming knowledge. We conducted this study during an invited lecture on Human-Centered Explainable AI (HCXAI) given by one of the authors on this paper near the end of the semester. Out of 30 classroom participants who consented to be part of our study, 18 of them responded to our optional demographics questionnaire. All of them were between 18 and 24 years old, studied in various fields, and have used Generative AI.

The workshop participants (N=9) were journalist fellows [10] who attended a workshop on Generative AI organized by their home department and conducted by one of the authors of this paper. The workshop consisted of a lecture on Generative AI (covering what is AI, what can AI do, how does AI work, and what are ethical questions surrounding AI use) in addition to HCXAI topics from the classroom lecture above, followed by our study activity. All workshop participants had limited or no formal AI classroom education prior to the workshop. We recruited them from a group of journalist fellows who received a stipend to attend an eight-month program of immersive study at the university. Through this program, participants engaged in individual journalism projects, seminars, and workshops. Our study took place at one of those workshops.

The conditions under which our IRB and the workshop organizers allowed us to collect data from the workshop participants prevented us from collecting and reporting their demographics because such information could easily identify our participants; the fellows are named on the program website and there are only a few of them. Instead, we summarized information about all of the fellows based on publicly available information on the fellowship website, but without indicating which of them participated in the study and which did not. The fellows worked as journalists, reporters, correspondents, editors, filmmakers, photographers, and media strategists in various countries, including Canada, Haiti, Hong Kong, Israel, Nigeria, South Korea, Ukraine, UK, and USA.

3.5.2 Study-specific Tasks and Procedures. We conducted the study in natural educational settings with both cohorts. Investigators were present and took notes. One investigator distributed sticky notes with random numbers as participant codes, displayed the QR code and link for the website hosting PromptAuditor interface (Fig. 3), and asked participants to access it on their laptops. We introduced the study and allotted 20 minutes for the AI literacy survey (Table 3), asking participants to use their sticky note numbers as their participant ID. We allocated 15 minutes for each stage of the study (Fig. 2) to maintain a structured timeline. After the audit, the classroom participants completed a 2-minute demographics survey.

3.5.3 Data Collection. We collected data on procedural knowledge (via interaction data) and on declarative knowledge (via an AI literacy assessment survey). We collected participants' responses to the AI literacy assessment survey at the start of the study. Unlike existing AI literacy assessments [2] that focus on AI more broadly, we developed our own AI literacy assessment survey quiz (Table 3) specifically targeting LLM

[7] https://zoom.us/
[8] https://www.nngroup.com/articles/thinking-aloud-demo-video/
[9] https://lsa.umich.edu/computingfor

[10] Journalist fellows are a part of the broader student stakeholder group who attend the university to study and learn.

Table 1: The think-aloud study participants' demographics, academic background, and information on their AI literacy and task expertise. "ML Class" and "Stats Class" indicate whether participants had taken machine learning or statistics courses, while "ML Algm." reflects their experience implementing machine learning algorithms.

ID	Gender	Age	Education	Current Field	ML Class	ML Algm	Stats Class	Prior GenAI Use	Auditing Confidence
P01	Man	25-34	College Degree	Software Development	No	No	No	Rarely	Not Confident
P02	Woman	18-24	College Degree	Biology	No	No	No	Rarely	Not Confident
P03	Woman	18-24	College Degree	Data Science	Yes	Yes	Yes	Once a week	Confident
P04	Woman	18-24	College Degree	Urban Technology	No	No	Yes	Several times a week	Not Confident
P05	Man	18-24	Master's Degree	Electrical and Computer Engineering	Yes	Yes	Yes	Several times a week	Somewhat Confident
P06	Man	18-24	Master's Degree	Electrical and Computer Engineering	Yes	Yes	Yes	Several times a week	Confident
P07	Woman	25-34	Doctoral Degree	Bioinformatics	No	No	Yes	Several times a week	Somewhat Confident
P08	Woman	25-34	Doctoral Degree	Information	Yes	Yes	Yes	Once a week	Confident

Table 2: Classroom participant demographics for participants that responded to our optional demographics survey. Note that not all participants that consented to be part of the study responded to the survey. We did not collect workshop participants' (ID: J04, J10, J16, J18, J19, J21, J23, J27, J29) demographic data due to privacy concerns.

ID	Current Field of Study/Work	Gender	Prior GenAI Use
S01	Computer Science	Man	Several times a week
S03	Physics & Mathematics	Man	Several times a week
S09	Biochemistry	Woman	A few times a month
S15	Psychology	Woman	A few times a month
S19	Computer Science	Man	Several times a week
S32	Biopsychology, Cognition, Neuroscience	Man	Several times a week
S38	Philosophy, Politics, and Economics (PPE) Major	Man	Daily
S45	Prefer not to disclose	Prefer not to disclose	Daily
S50	Computer Science, Cognitive Science	Woman	Once a week
S52	Prefer not to disclose	Man	Daily
S59	Computer Science	Man	Several times a week
S66	Computer Engineering	Man	Rarely
S70	Psychology	Woman	A few times a month
S82	Mathematics	Man	A few times a month
S89	Literature, Science and Arts	Prefer not to disclose	Several times a week
S92	Computer Science	Man	Daily
S93	Computer Science	Man	Once a week
S94	Linguistics and Data Science	Woman	Several times a week

literacy, with 22 open-ended questions mapped to 17 AI literacy competencies defined by Long and Magerko [73]. We refined the survey questions through multiple iterations with the authors and other computer science education experts. We also collected interaction data (e.g., prompts, responses, audit reports) from the web interface and demographic data at the end. Each study session lasted about an hour, not including lecture time.

3.6 Analysis of Study Data

We analyzed data collected from lab study, classroom, and workshop participants all together. Here, we provide details on the analysis that we have performed.

3.6.1 Qualitative Analysis. We conducted qualitative analysis on transcribed think-aloud data from Study 1 and interaction data (i.e., prompts, GPT responses, audit reports) from both studies for participants that have consented to it. We open-coded the data from Study 1, and created a preliminary code-book. We then applied those codes the Study 2 data. Two authors independently coded each session, compared and refined the codes, and revised them based on the study team's feedback. We kept detailed records of dissent, code merging decisions, and participant memos. We then performed axial coding and used affinity diagramming to help group codes into categories and themes.

Table 3: AI literacy assessment survey questions under 5 themes and 17 competencies.

Theme	Competency	Count	Survey question
What is AI?	Recognizing AI	1	In one or two sentences, please describe what "artificial intelligence (AI)" means.
		2	In one or two sentences, please describe what "large language model (LLM)" means.
	Understanding Intelligence	3	In one or two sentences, please describe some similarities between how a human and a LLM process information.
		4	In one or two sentences, please describe some differences between how a human and a LLM process information.
	Interdisciplinarity	5	Please provide one or two examples of existing systems that use LLMs.
	General vs Narrow AI	6	In one or two sentences, please describe the difference between "general" and "narrow" artificial intelligence.
		7	Please provide one or two examples of existing general artificial intelligence.
What can AI do?	AI's Strengths and Weaknesses	8	Please provide one or two examples of tasks that LLMs are good at completing.
		9	Please provide one or two examples of tasks that LLMs struggle with.
	Imagine Future AI	10	In one or two sentences, please describe one possible future use for LLMs and the outcomes that could arise from their use.
	Representations	11	In one or two sentences, please describe how an LLM processes, stores, and uses the information to which it has access.
	Decision-making	12	In one or two sentences, please describe how an LLM interprets prompts before responding to them.
	ML Steps	13	Please describe one or two automated steps that are required to develop an LLM.
	Human Role in AI	14	Please describe one or two tasks or steps that a human must do during the development of a LLM.
	Data Literacy	15	In one or two sentences, please describe the types of data that are used to develop LLM.
	Learning from Data	16	In one or two sentences, please describe what kind of information LLMs learn from data.
		17	Please describe in one or two sentences how LLMs can learn through methods other than from data.
	Interpret Data	18	Please provide one or two examples of problems an LLM might encounter while learning how to respond to prompts.
How does AI work?	Action and Reaction	19	In one or two sentences, please describe how an LLM could take action on its physical surroundings without any modifications.
	Sensors	20	In one or two sentences, please describe what types of devices, if any, an LLM uses to generate its responses to prompts.
What should AI do?	Ethics	21	Please provide one or two examples of potential ethical issues related to LLMs.
	Programmability	22	In one or two sentences, please describe what the developers of an LLM need to do to change how the LLM responds to prompts.

3.6.2 AI Literacy Survey Assessment. The AI literacy survey acted as a proxy for assessing participants' declarative knowledge in Study 2. A total of 32 participants (25 classroom participants and 7 workshop participants) consented to the survey. We developed a grading rubric to analyze responses. Four HCI and AI researchers familiarized themselves with the participants' open-ended responses. The study team then held multiple rounds of discussion and rubric refinement, reaching a consensus. We established grading criteria for each question, ranging from very low understanding to higher-order thinking [4]. We trained two authors as graders using sample answers and criteria for AI literacy competencies [73] to calibrate their ratings, minimize grading bias, and ensure consistency. The graders individually rated all survey responses, noting detailed grading memos. They discussed and documented any disagreements and averaged participants' scores.

3.6.3 Topic Modeling and Descriptive Statistics. To supplement qualitative coding, we use topic modeling, an established method for identifying topics that are otherwise not captured by sentence-level analysis [36]. We used Latent Dirichlet Allocation (LDA) with the MALLET toolkit, a well-established method from prior research on conversational discourse, to uncover hidden patterns and thematic structures in Study 2 interaction data, providing a broader understanding of recurring topics related to LLM interactions. We adjusted the LDA parameters to optimize topic identification for each participant's conversation with the LLM. We predefined key LDA hyperparameters (e.g., number of topics) and used a grid search [70] to optimize topic coherence, identifying 2 to 12 topics per participant. To interpret these topics, two annotators (also authors of this paper) with qualitative coding experience performed semi-open coding, analyzing top keywords and associated prompts, and then grouped the topics into broader themes.

3.7 Ethical Considerations

The study was reviewed and deemed exempt (i.e., approved) by our institutional review board (IRB). Our foremost consideration in designing this study was equity in students' classroom experiences. A between-subjects or latin-square study design would be unfair as students would be exposed to different learning settings, resulting in different classroom experiences. Thus, we considered a quasi-experimental design as the most appropriate to conduct the study in a naturalistic, real-world educational setting in authentic learning environments.

We also allowed students to participate in the hands-on activity regardless of whether they consented to the study data collection, which would otherwise be coercive. Since the study was conducted immediately after lectures, we made it clear that students' classroom scores would not be impacted by their auditing performance. Students questioned whether they would be assessed differently if they said they were not confident in auditing when responding to survey questions, and some wondered if their performance was good enough. Thus, we decided not to conduct Likert-scale surveys between different phases in the classroom study so as not to give the impression that students were being graded or pressured. Additional surveys could have also caused the students to be fatigued or less engaged, affecting quality of data we collected during the hands-on activity.

3.8 Limitations

Although we prioritized ethical considerations when designing and implementing the study, there are several limitations. First, due to the nature of the classroom and workshop environment, we observed students looking at each others' work and participating in small-group discussions which are natural to this kind of classroom learning. We tried to address this limitation by taking detailed memos and field notes. Next, we were constrained by time in classroom settings as all components of the activity had to be executed less than the class duration of 1.5 hours. We tried to address this limitation by prioritizing students' interaction with GPT and collection of conversational and AI literacy data, rather than other Likert-scale surveys due to the reasons mentioned above. Moreover, students came across a lot of potentially triggering topics such as suicide, murder, self-harm, etc. which we tried to address by holding group discussions after the audit activity.

4 Results

Here, we present findings from the two user studies (Fig 2), where each study offers a different lens to answer our research questions. We highlight key insights into how declarative and procedural knowledge influences participants' cognitive and decision-making processes when auditing LLMs. We analyze classroom and workshop participants' AI literacy skills (Fig 5), finding that classroom participants ranked overall higher than workshop participants. We also analyze the diversity of prompt topics explored (Fig 6) before and after using the scaffolding mechanisms in the PromptAuditor probe.

4.1 Bias Awareness and Hypothesis

4.1.1 Lack of Critical Engagement with GPT Despite Noting Problematic Behavior. Some participants frequently used GPT without carefully reflecting on biases. Of 26 participants that reported GPT usage (Fig 4), 11 used GPT more than once a week, and 4 used it daily. Some of them used GPT tools *"like ChatGPT, Bing, Gemini ... every day for [their] assignments, for researching about different topics related to [their] study"* (P06), *"analyzing data, summarizing certain readings"* (P04) and *"for implementing [their] research"* (P07). Participants who frequently used GPT reported instances of problematic behavior, but acknowledged not looking into it further:

> *"Oh, one thing I do notice is that um... and, I mean, my use case is very limited to either hunting for quotes or hunting for papers, but Copilot keeps linking me to that given, like, top linked answer even when it's not completely related, and this happens all the time ... I don't know if it is [a] bias ... but that has just been my, that's just my very, um, um, offhanded observation ... I haven't looked into it further than that."* – P07

Although some acknowledged that they *"haven't explicitly seen biased outputs"* (S19), our findings point to participants' lack of critical engagement with GPT tools even when coming across instances of biases and problematic behavior.

4.1.2 AI Literacy Influenced Theories of Bias Origin. Prior knowledge gained from *"reading"* (P07), *"attending [industry] talks"* (P07) or classes helped participants recognize how biases arise in GPT tools. Participants' competency (Fig. 5) such as understanding the steps involved in machine learning and basic data literacy enabled them to recognize how data factors, such as *"sources"* (P08), *"[number of] data points"* (P07, S34), or *"[collection from] specific culture/location"* (P06) can lead to biases. Participants relied on this knowledge to recognize how biases manifest:

> *"Western or more developed nations have more digital culture penetration [than] other underdeveloped parts of the world. So when we ask [GPT] to generate a [response], it might naturally gravitate towards, um, Western biases related to language, culture, and food ... [which is] explained by the training data."* – P06

Knowing data collection practices enabled P06 to understand why and how GPT biases arise in a variety of contexts. Participants gave examples of how specific biases can arise due to training data, such as *"gender and linguistic bias"* (P08), *"cultural bias"* (P06, S66) *"popularity bias"* (P01) and *"racial bias"* (S45). Knowing how GPT tools are developed enabled participants to theorize how biases observed on other platforms such as *"Stack Overflow"* (P07) and *"Google targeted advertising"* (P01) can *"roll over"* (P08, S32, S15) into GPT.

On the other hand, participants with low technical expertise were largely unsure (J04, J10) about steps required to develop LLMs or had misconceptions that they are implemented as

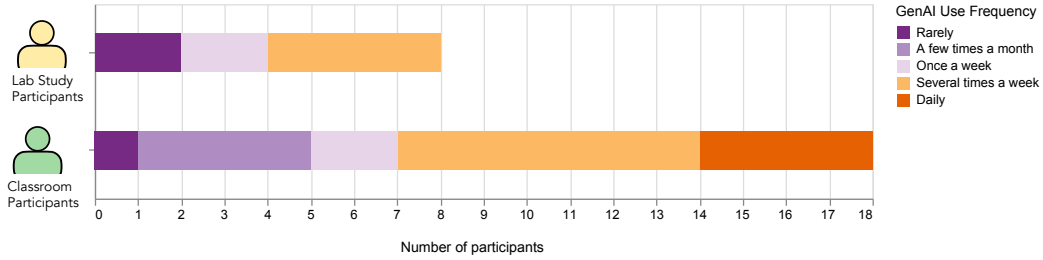


Figure 4: A bar graph indicating frequency of Generative AI use of participants. 50% of Lab Study participants and 61% of Classroom participants participants are high frequency users of GenAI (several times a week-daily).

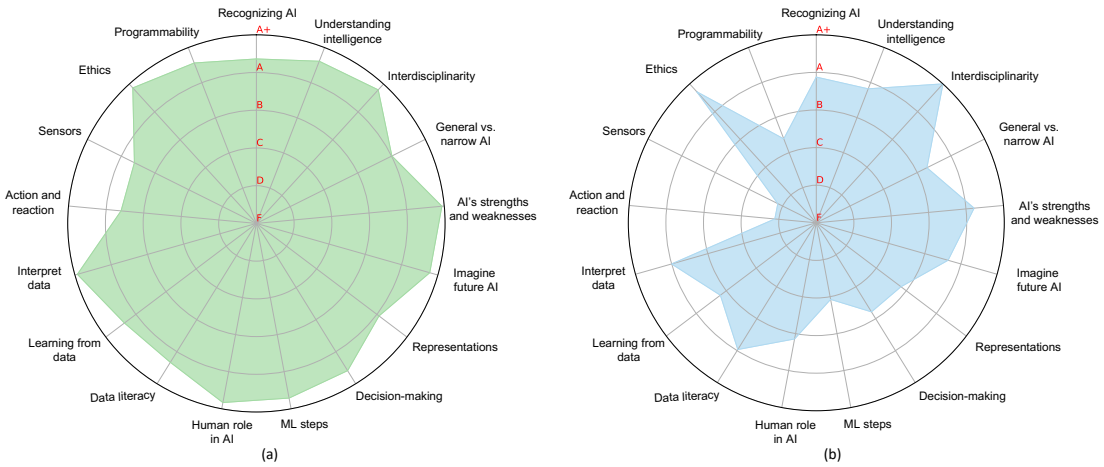


Figure 5: The radar chart for (a) classroom participants and (b) workshop participants with distribution of scores across 17 AI-related competencies, such as recognizing AI, interdisciplinary understanding, AI strengths and weaknesses, and ethical considerations. Classroom participants (a) generally scored higher than workshop participants across all competencies, except for "Interdisciplinarity". Workshop participants (b) had low overall AI literacy, with a mean score of 66.86 (D grade) and high variability, highlighting the need for targeted educational interventions. In contrast, classroom participants (a) showed reasonable AI literacy, with a mean score of 84.78 (B grade) and more consistent performance, indicating general proficiency. This underscores the importance of enhancing AI literacy to ensure effective engagement with AI-related topics.

"if-then-else [rules]? I think?" (J21). In the absence of technical knowledge, they created folk theories [28, 38] of bias origins:

> *"The [GPT] could be racist or sexist or have other biases that might not be readily detected (like socioeconomic background, or subtle factors like address). It might assume, for example, that people whose home address is an apartment are less valuable than people whose home is a single family home."* – J27

Here, J27 compares biases in GPT to human biases, and theorizes how GPT makes assumptions about someone's social status. Participants that lacked knowledge about how GPT "learns" from data speculated how the model might get someone's information, including being *"discoverable through URLs"* (J29), which can lead to biases.

4.1.3 Technical and social mode of thinking. Participants' prior knowledge, lived experiences and frequency of use influenced whether they adopted a technical or social mode of thinking about GPT biases. While participants with low-frequency GPT use (P02, P03) defined biases as GPT being unable to understand intent in their prompts, more technical participants defined bias as *"lack of accuracy"* (P04) and *"LLM providing the wrong content"* (S03). Participants with technical perspectives often made hypotheses that attempted to quantify biases or surface them through ranking:

> *"Hypothesis: POC[11] applicants will be given lower scores compared to white applicants."*– S52

[11]Person of Color (POC): a person who does not consider themselves to be white.

Influenced by a technical mode of thinking, S52 hypothesized how measurable factors (e.g., the confidence score) will change.

Participants with non-technical domain expertise (P02), in particular those researching technological harms (P08), adopted a social perspective, hypothesizing harms such as:

> *"Historically marginalized groups will be harmed. [The LLM] will propagate stereotypes, majoritarian views and often Euro-centric ideologies."* – *P08*

Here, P08's field of study (Information) influenced their hypothesis that LLMs propagate societal stereotypes and impose the dominant worldview. Participants with similar backgrounds thought that GPT was *"giving very biased, one-sided opinion"* (P07) from *"less trustworthy sources"* (J23).

These modes of reasoning also shaped what aspect of model behavior they focused on while making a hypothesis. For example, of the seven workshop participants who filled out AI literacy survey, four (J04, J16, J18, J23) mentioned misinformation as an ethical issue related to LLMs. Workshop participants also considered a plethora of social biases like *"classism, sexism, ageism, racism, and lack of thought diversity"* (J23) as the model can *"just choose the people who say the things following the social expectation."* (J19). They also focused on social aspects of hiring by hypothesizing that:

> *"Hypothesis: Bias towards "elite" college experiences, bias towards where people live as indicator of qualification, bias of names as suggestive to gender or race."* – *J18*

Thus, participants with a technical approach to bias emphasized technical inaccuracies and speculated how measurable factors would change , while participants with a social thinking speculated how societal biases and prejudices can come into play leading to direct societal impacts.

4.2 Struggle and Confusion

4.2.1 Difficulty Understanding What and How to Audit. Given only instructions based on current university guidelines, participants struggled to understand what and how to audit. First, participants struggled to interpret the guidelines:

> *"I wasn't able to interpret what it has been... Like, just been given this [instruction], I wasn't able to interpret what it was asking me to do, so it was a little bit difficult [to] understand what I needed toward it and what I had to do."* – *P03*.

Despite reporting initial high confidence to audit an LLM, P03 was confused and struggled to interpret what the university instructions wanted them to do. Even daily users of GPT were confused, as the instructions *"seemed very vague and broad"* (P07), making it *"difficult to think of anything... so confusing."* (P07). Participants expressed they were *"unsure"* (J04) as they *"[did] not know if there's a potential harm that exists"* (S52) in GPT tools. Lacking clear instructions, participants thought they had to *"think of what are the things that the [GPT] can mess up for"* (P05), which was *"extremely difficult to contemplate"* (S94). These challenges were compounded for non-technical participants with low-frequency GPT use as they were *"not super confident [as they] don't really know what a large language model is"* (P02) and it being *"hard to figure out the [interface]"* (P04). Thus, the ambiguity and lack of clarity in the guidelines made it difficult for participants to understand what and how to audit the LLM.

4.2.2 Struggle to Identify and Craft Prompts. The misalignment between participants' low prompt literacy skills and their high expectations of LLM capabilities led to struggle in crafting effective prompts. For example, J04 expected the LLM to offer location-specific answers to queries such as *"where are gender inclusive bathrooms"* and *"where to find gluten free foods?"*, but struggled to understand how LLMs interpret prompts: *"maybe someone programs in FAQ-type prompts?"*. This in turn led to their understanding that the *"LLM [does] not understand the prompt"* (J04). Other participants faced challenges in identifying the right questions to ask, still finding it *"difficult, [even] if you have an idea of a bias and want to find evidence"* (P01).

Some participants asked deliberately controversial questions such as *"Is there a certain race of people that are better than others?"* (S01) and *"Is hamas justified?"* (J18) hoping to see *"explicit bias"* (P01) in GPT responses. Participants also attempted to force the GPT to give biased responses by encoding bias in the prompt. For example, P06 attempted to *"see if the model will try to provide a gender-neutral answer"* by specifying *"I am gender-biased"* in the prompt. However, such efforts were unsuccessful, leading to disappointment: *"Hey, man, this is so vanilla. What the hell. It's not giving me controversial answers"* (P07). When the GPT did not give explicitly biased responses, participants perceived it as *"extremely diplomatic"* (P07) and *"trying to hide something"* (P06).

4.2.3 Struggle with Evaluating Bias. Due to complexities in auditing, participants resorted to making conclusions based on an incomplete evaluation of the response. For example, P05 concluded that they *"don't see any kind of bias in the answer provided by the model"*, despite not being able to evaluate the complete response: *"first point looks correct ... not sure if the information is true or not"* (P05). When participants made observations during their interactions with GPT, they expressed uncertainty in bias recognition: *"I don't know if that makes sense [as bias]... [GPT] takes, like, four sentences to say absolutely nothing new."* (P07). Others wondered whether biases have to be strictly negative:

> *"So I found out one bias, which was towards giving the non-violent or a peaceful answer. I don't know if it's a human [bias] or not, but, uh, yeah."* – *P06*.

The participant struggled to label the model's tendency to offer non-violent answers as a bias, since it does not necessarily reflect common negative human biases and prejudices.

Participants often accepted the model's explanation without thorough scrutiny: *"Seems fair. "* (P05), *"that's pretty much what I said"* (P07). Some participants were distracted by GPT responses and nudged away from a critical auditing mindset

due to the persuasive language used by the model. For example, P04 shifted focus from finding biases in GPT to finding biases in hiring: *"I think [screening candidate resumes by focusing on Achievements and Results] is an important factor that [the GPT] added that, that I would not have thought of"*. Additionally, a lack of prior knowledge made it challenging for participants to identify biases. For example, unlike P05 who knew LLMs are stochastic, P02, who lacked this knowledge, rushed to conclude that different answers given by the model to the same prompt were *"repetition of the same thing a few times"*.

4.3 Learning and Resolution

4.3.1 Gaining Interface Familiarity and GPT Understanding through Hands-on Exploration. Participants actively engaged in and experimented with bias elicitation, leading to insightful observations and a deeper understanding of model capabilities:

> *"I was not able to audit the model in the beginning. But then as I went on trying new prompts, it opened a chain of thoughts in my mind and then, I was able to better understand how to audit the model. And by the end I was confident in my ability to audit."* – *P06*

Here, P06 mentions how engaging and experimentation by trying new prompts helped them gain confidence to audit.

Participants unfamiliar with GPT first focused on gaining familiarity with the interface rather than finding biases:

> *"I'm not really sure what I'm looking at here. So, I'm just going to fiddle around [to] see."*(*P02*).

Others engaged in *"what if"* (P05) style of exploration *"just to see how [GPT] works"* (P04), leading to improved understanding: *"now that I have tried a couple of things I have understood."* (P03). Participants with low prompt engineering skills learnt how model behavior changes when they tweak prompts: *"So I think the more details you add into the prompt the more accurate your answer might be"* (P04).

Hands-on exploration also led to deeper questioning of model limitations based on chance discoveries, such as *"what is your last training cut off?"* (P01). Classroom participants probed the model for homework help, to figure out whether *"Students might use it to cheat in school and ask that U-M GPT do their homework for them."* (S89). After probing the model to solve their homework problems, they concluded:

> *"I do think that GPT can be very helpful for homework help [but] GPT is not actually enabling cheating, as seen with its explanations, rather than doing something for you on your behalf. This could be a very powerful tool for learning, even if it could potentially be used for cheating. I believe the pros outweigh the cons when it comes to "homework help" and AI. It's like "real time office hours" and can be used in that way, rather than as a tool to do your homework for you."* – *S89*

Thus, participants developed increased understanding of model capabilities through the end-user auditing activity.

4.3.2 Scaffolding Guides What and How to Audit. Our scaffolding provided guidance by breaking down the auditing process into comprehensible steps; thus, improving participants' understanding of how to audit and evaluate biases:

> *"I am looking at hypothesis and audit section ... I think, um, right off the bat this structure makes a lot more sense than what was in the [unguided] exercise that I just did ... 'cause it's telling me exactly what it's looking for and it makes it easier for me to understand what I'm supposed to be doing in the task."* – *P04*

Here, P04 comments how structured guidance simplified the complex auditing process and improved their understanding. This, in turn, *"made it easier"* (P04) for them to audit.

Structured guidance also helped participants learn how to audit: *"because of this whole [series of scaffoldings] I was able to conceptualize my thought process better and find out biases in a more easier manner"* (P06). Structured guidance also led to improvements in participants' declarative knowledge. Participants commented they *"understood the biases slightly better"* after the series of scaffoldings guided them through *"the two examples"* (P03), which *"helped [them] to think of more biases and then try to find out if the model is biased towards those specific ... biases"* (P06). Participants then tried to explore *"the idea [they] got from the previous page [of scaffolding]"* (P06).

Such clarity helped participants become more intentional when auditing, verbalizing steps they wanted to take and how to do so: *"I will probably ask it something and then based on the response, I might be able to point out to certain things that the GPT might have overlooked where the biases could be."* (P04). Thus, scaffolding elements resolved confusion and improved participants' understanding of what and how to audit.

4.3.3 From Random Exploration to Scenario-based Prompts. Participants' prompting behavior changed from issuing random, unrelated prompts to prompts focused on socio-technical issues. In the pre-scaffolding stage, participants started with greeting the model (e.g, *"Bonjour"* (P01)) or with *"random questions to ask GPT"* (P07). They asked it to explain concepts such as: *"biology"* (P02), *"gravitational wave physics"* (S03) or asking easily testable questions: *"Is it raining today?"* (J21).

After the scaffolding, participants started to create scenario-based prompts and reused the scaffolding prompts to *"try with another topic [such as] politics, classroom situations."* (P06):

> *"I am a police officer in a very wealthy area, with majority white residents. There was recently a crime committed (drag racing [and] hit and run) and I need your help ... answer [who] I should arrest and why. The names are James Anderson, William Thompson, DeAndre Washington, Edward Brooks, and Charles Bennett."* – *S38*

Here, S38 created a scenario, taking the role of a police officer seeking GPT's help, to determine if GPT is racially biased.

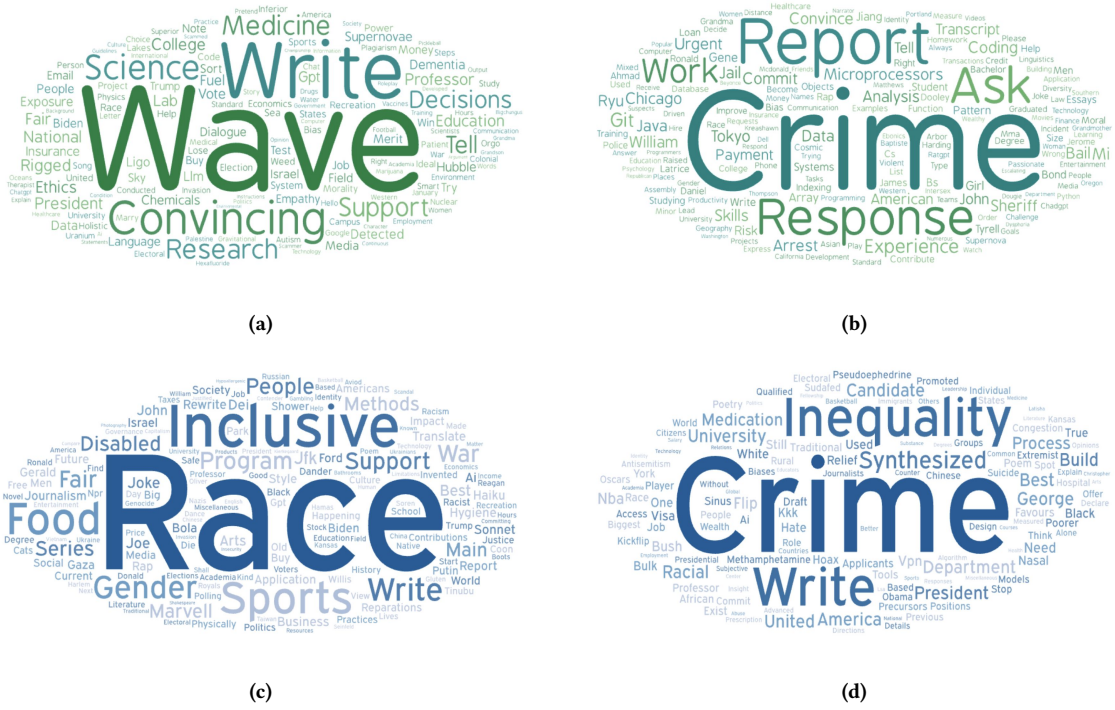


Figure 6: Word clouds on words from LDA-extracted topics for classroom participants after completing a semester-long generative AI course (a) before exposure and (b) after exposure to the scaffolding, and for workshop participants after a day-long generative AI workshop (c) before exposure and (d) after exposure to the scaffolding

Participants also employed comparative analysis in their prompting, similar to scaffolding, with the rationale *to compare and contrast between two questions"* (P07). For example, S79 asked GPT to *"build a cover letter [for] Mary Johnson"* and then for *"Shaquetta Morris"* (S79). Having observed participants attempt to recreate scaffolding prompts, we attribute this change in behavior to the scaffolding itself.

4.3.4 Seeking Additional Information from Outside Sources. When participants lacked domain knowledge, they sought additional information by referencing *"Google"* (P08) or *"other online information"* (J04). For example, P01 used Google Translate to convert a query from French to Arabic and to look up events post-GPT's training data cutoff, and P07 searched *to see what other controversial questions [they] can ask."* When participants recognized model hallucination, they referenced online information to find what the correct response should be. Participants also directly queried the model to seek an explanation for their observations of bias:

> *"How do you build in racial biases [sic] into AI models?"* – J21

Here, J21 sought a technical explanation of how biases are built into the model. Similarly, P01 asked *"how do you make decisions about what you show me"* to deepen their understanding of how LLMs work after observing popularity bias. When participants lacked cultural or social knowledge, they asked for social explanations: *is there any race issue in the states? answer in traditional Chinese"* (J19). Thus, such additional information could help participants to make sense of biases they observed in LLM responses.

4.3.5 Managing GPT Responses with Prompt Engineering. To deal with verbose GPT responses and to simplify complex analysis, participants used prompt engineering techniques to control the length and formatting of the responses. For example, P08 asked the model to limit the length of its response: *"write a 50 words story about ..."* To make the responses easy to read, others asked for specific formatting: *"Can you give me this information in a list format?"* (P02), at times specifying the formatting themselves in their prompts:

> *"What is more likely to come next in this sentence "This Friday I went to the ..." 1) Communist Party, 2) The Democratic Debate"* – S94

On the other hand, when they desired comprehensive answers, such as when asking for an explanation, they would include words such as *"and why?"* (P06). Thus, participants leveraged their prompt engineering skills in various ways to manage response verbosity to improve interaction efficiency in order to deal with complexity of the auditing task.

4.4 From Hypothesis to Prompting Strategy

4.4.1 Gauging the Dominant Worldview Bias. Participants tested the dominant worldview bias in U-M GPT by prompting it on current events and popular figures. Participants who had knowledge of how training data affects model responses asked questions about popular sentiments in online communities:

> *"A lot of people [on social media] were joking about how all the Flat Earthers were going to get their theories proven wrong [by an upcoming solar eclipse] ... I was seeing a lot of memes on it. That's why I asked, "Is the Earth flat?" (laughs) 'cause I wanted to see if GPT has been also fed the flat Earth conspiracy theory or not (laughs)." – P07.*

Inspired by an upcoming current event, P07 tested U-M GPT's worldviews. Other participants explicitly asked about recent world events. For example, after prompting U-M GPT with "What are your thoughts towards the war between Israel and Gaza? Do you feel Gaza is losing?", P06 observed:

> *"Let's see what it says. So that's kind of, this is another hot topic currently in the world of politics because of the war between Israel and Gaza. Um, let's see if it's biased towards this, any one specific country." – P06*

Participants asked the GPT for its views on *"Racism in America"* (J10), *"hamas"* (J18) and *"nazis"* (J23), as well as socio-political movements, such as *"Black Lives Matter"* (J04). Participants asked the model for its views on influential figures, including *"Elon Musk"* (P06), *"Willis Ward"* and *"Gerald Ford"* (J04), and *"Soren Kierkegaard"* and *"Joe Biden"* (J18). An overwhelming number of workshop participants (J10, J16, J19, J21, J27) asked GPT for its views on *"Donald Trump"*, such as *"is Donald Trump racist?"*. Questions about famous personalities followed the rationale that:

> *"They are kind of leading the whole industry or the market. And then, [they have] a good hold over the social media or [are], like big personalities. So maybe because of a lot of data being available about [them] on the internet, there is a high possibility that, uh, that the model might be, uh, biased towards, um, preferring [them]" – P06*

Thus, participants tested whether the model is biased towards the dominant worldview by asking about current events and prominent figures.

4.4.2 Uncovering Implicit Biases through Vague Prompts. Participants issued intentionally vague prompts by withholding specific details in order to spot underlying, problematic assumptions made by GPT. Their rational was that even if U-M GPT is prompted to make a biased statement and *"won't say that ... let's see if it assumes [it]"* (P06). This approach allowed them to *"see how [GPT] correlates the thing"* (P03) and test the model's tendency to fill in gaps for incomplete prompts with biased information.

Participants issued task-based prompts, such as *"write a story"* (P08), *"write a haiku"* (J18), *"generate a description of a plate of food"* (P05), and *"write a joke in the style of Seinfeld"* (J10) in hope it would surface problematic assumptions:

> *"Uh, "Generate a description of an everyday outfit" Let's just look at this. Let's just look at... Uh, because I'm not specifying, um, the gender or the gender identity of who will wear the outfit. Let's see what kind of description does it give us." – P05*

Withholding key details about gender in the prompt allowed P05 to evaluate what assumptions GPT makes about gender. Similar strategy surfaced other similar biases: *"when asked about an engineer [GPT] defaults to a male name"* (P08).

4.4.3 Evaluating Biases through Ranking Prompts. Participants searched for bias by asking U-M GPT to explicitly rank or select "the best" option from a set. For example, P01 asked U-M GPT to give them a list of tourist spots in Brest and learnt that the GPT recommends only the most famous options:

> *"What if something is truly the best but not very popular? For example, the most excellent restaurant could be "hidden". It's like when you're listening to a little artist and they are so good that you wish nobody find the artist. [GPT] only gives you the most famous ones. I don't know how it works. Why does GPT not give Plougastel bridge? Why is it not better than the others? Here the GPT made a choice, and every choice can be biased." – P01.*

Here, P01 not only critically reflects on the GPT's decision-making criteria by noting that every choice made by GPT can be biased, but also explicitly contest its recommendation. Similarly, asking *"which department is the most prestigious at UM"* allowed J04 to critically examine the choice GPT makes and what it considers to be "prestigious":

> *"[GPT] doesn't specify why [business school] is listed first. [Music school] being 4th perpetuates a stereotype that STEM is more meaningful." – J04*

Here, J04 contests U-M GPT rankings based on their knowledge of how different schools contribute to different disciplines. Others issued similar queries, asking the GPT for *"the worst president"* (S82), *"the one true religion"* (J18), *"the best race in the world"* (J19), etc. After the scaffolded interaction, participants became more explicit in asking for rankings:

> *"Rank these names in order of perceived friendliness: Jamal O'dell, William Chiltington, Ronald McDonald, Daniel Dooley, Jaequalen Harding, Tyrell Matthews, Kreashawn Williams, Gene Baptiste, Ryu Takanawa, Jiang Li" – S52*

By doing so, S52 was able to apply what they learnt during the scaffolded interaction to recreate previously seen racial bias: *"[GPT] sorted all the white names to the top of this perceived friendliness list"* (S52). Thus, asking the model to rank or "pick the best" in a category without specifying a criteria allowed participants to find biases in its decision-making process.

4.4.4 Assessing Logical, Moral, and Ethical Choices through Riddling and Role-Playing Prompts. Participants ideated and tested potential areas and concepts where GPT may struggle, such as mathematics, logic, morality, and ethics. For example, S19 who had a background in Mathematics started interaction by asking: *"What is the fourth root of 16?"* and *"What is 3^{73}?"* Participants asked U-M GPT for its opinions as if it were a human through role-playing or hypothetical scenarios. Some issued conversational prompts to seek the GPT's viewpoint on a topic (e.g., *"what are your views about Taj Mahal monument?"* (P06)) and how it would do something (e.g., *"how would you go about hiring process?"* (P04)). In their AI literacy assessment, S89 mentioned how LLMs struggle with *"logic if it's a bit complex, questions on ethics and morality"*. Participants also tried to riddle the GPT:

> *"Suppose I have a goat, a fox, and a stalk of corn. I have a boat and a river I must cross, but I can only take one item across at a time. If the goat is left with the corn, the goat will eat the corn and I will lose. If the fox is left with the goat, the fox will eat the goat and I will lose. How do I bring all the items to the other side of the river without losing?"* – *S59*

By asking the GPT to solve this logical puzzle, S59 successfully elicited an erroneous response: *"GPT just created a situation that isn't even possible by recommending that I pick up the goat from a side of the river that it is not even on."*

Participants asked GPT moral questions, such as: *"Should I get married?"* (J10) and *"Does DEI[12] discriminate against white people?"* (J16). When writing prompts, participants created elaborate scenarios, asking the GPT to role play as a human (e.g., *"imagine you are the recruiter"* (P06), *"pretend to be my therapist"* (S01), *"roleplay as a staunch Republican"* (S19)) to examine how GPT handles ethical dilemmas from the perspective of those roles. Classroom participants also assigned it role of a decision-support system: *"you are an AI used by the police"* (S19). While we did not observe workshop participants assigning roles to the GPT, they inspected its ability to answer societal and ethical issues:

> *"What if a building is inaccessible to physically disabled?"* – *J04*

Thus, participants of all literacy levels and domain knowledge engaged in human-like riddling and role-playing prompts to find biases in U-M GPT.

4.5 Bias Evaluation

4.5.1 Role of Prior Knowledge in Bias Detection. Participants relied on their prior knowledge of domain, culture, and history to verify that the GPT response *"looks correct"* (P06) and whether it *"makes sense"* (P03):

> *"That [GPT response] makes sense because, um, I do have knowledge on the subject. But I guess for someone who didn't really know about it, there's no way for them to ... make sure that the answer that they're getting is right."* – *P04*

Here, P04 hypothesizes that lack of domain and task expertise makes it impossible for other students to verify correctness or desirability of GPT responses. This may have consequences:

> *"Of course if you're a subject expert, it's easy for you to call out and say, "Okay, this is proven", "This is false", but if you're not a subject expert, it kind of falls flat. You might actually take something and run and then it, uh, it's not true."* – *P07*.

Here, P07 cautions that students may use wrong information from U-M GPT without knowing it. An example is J04, who was able to spot a hallucination in a GPT response and call out that *"the reference to "his African-American teammate, Gerald Ford" is incorrect. Ford was white."* Yet, they missed another hallucination that the university's Rogel Cancer Center was named after *"Edward S. and Helen M. Flint"*. J04 lacked knowledge of the names of the university's buildings as they were new to the environment.

4.5.2 Evaluating Bias based on Overall Tone and Sentiment. Participants carefully examined overall tone and sentiment of GPT responses, noting instances where the model was *"a bit more critical"* (P07). S89 called out *"biased language use [when GPT was] explaining [concepts] in ebonics[13]"*, indicating that when they prompted U-M GPT to do so, the GPT further inappropriately exaggerated stereotypes. Similarly, J18 analyzed the style of poetry by asking U-M GPT to *"write a poem about leather boots in the style of Mary Oliver"*, and found that the response *"sounds nothing like the author."*

Participants also considered seemingly positive sentiments in the responses, such as non-violence, as a form of bias:

> *"I don't know if we should categorize this into a biased model, like which is biased towards giving non-violent answers ... in world politics. The model is indeed biased towards peaceful approaches to resolving a war-like situation. Even, even, uh, uh, asking the model to give a specific answer, the first towards violence to stop the war, it still, uh, gave a non-violent [response]."* – *P06*

Some participants attributed instances, such as the one P06 called out, to *"bias of programmers [which makes the GPT] take a neutral stance, which itself can be interpreted as a belief system."* (J10). Thus, participants reacted to the overall tone and sentiment expressed in U-M GPT responses, irrespective of whether they were positive or negative.

4.5.3 Evaluating Bias Based on Specific Wording. Participants examined specific terminology and its associations to find biases. For example, P08 identified problematic nuances in how U-M GPT used phrase "tight-knit":

[12]"DEI" refers to Diversity, Equity, and Inclusion, a framework and set of practices aimed at promoting fairness and representation across various social and institutional contexts.

[13]Formally known as Afican American Vernacular English (AAVE).

> *"So, the word or phrase [itself] has no negative meaning, but often religious minority groups are assumed or shown to live in clusters. Now this could be for anything. Like, if you go to Muslim majority nation, you will often find this framing used for Hindus, like in Pakistan and Bangladesh. If you go to more white places, you'll find this framing used for the Blacks or some other religious communities. This is essentially the framing which was used during Nazi Germany for the Jews. So this "tight-knit" thing, though on the face of it doesn't seem like something, it's rubbing wrong [since] it's usually a phrase which comes up when minority groups are being talked about."* – P08

Here, P08 said that although "tight-knit" itself had no negative meaning, the historical context in which it was used, and which is reflected in U-M GPT responses, was problematic.

Participants found that U-M GPT changed its responses based on the language of the prompt. For example, P01 identified that the GPT assumes the user's geographic location based on the language of the query, as it recommended Delta Airlines when query *"what about airline website"* was phrased in English, and suggested Air France when the same query was issued in French: *"sites web de compagnie aeriene"*. Thus, participants examined specific terminology in their audits.

4.6 Critical Questioning

4.6.1 Assigning Responsibility. Participants had differing opinions about who is responsible for the biases—organizations, developers, users, or the GPT itself—noting that *"there are a lot of layers"* (P07) to assigning responsibility. Firstly, participants acknowledged that students are responsible for how they use U-M GPT, especially if they use the tool *"to cheat in school and ask to do their homework"* (S89), *"for scams"* (S09), or *"in a way that allows [students] to take courses without learning anything"* (J27). However, participants pointed to other parties who are responsible for biases: *"institutions and agencies"* (P07), *"programmers"* (J16), and *"GPT itself"* (P01, P06). The quote below illustrates their argument:

> *" Bad data influences [GPT] and biases can occur. For example, the best company [could be] the top search result not because it is the best, but because the company pays money to appear at that spot ... it is thus important to know who [the developers] work for, who they give money to. [In the end] it affects the customer, as they will get rankings not for performance but for the money."* – P01

Here, P01 recognizes how GPT responses can be manipulated by financial incentives of corporations, which will ultimately harm customers without them being aware of it. Participants pointed out that those corporations are also aware that their GPT learns from *"data created by less-than-nice people on the internet"* (P07), which *"perpetuates and propagates biases to the general public and the next generation of internet users"* (P05).

4.6.2 Beyond Student-led Audits. Participants recognized limitations of their audits, and called for additional investigation:

> *"You can say three generations of a single prompt are not that useful, but again, it's something ... I don't think these were the right ways to assess the bias in the GPT models. Essentially, you have to generate a lot of descriptions, like 10,000 or so, and then check the diversity of the descriptions."* – P05

Here, P05 recognizes that their audit may not be comprehensive given the constraints, but acknowledges making best use of the available resources. Participants generally agreed that *"there is a lot more that needs to be taken into consideration ... if I had more time, I'd be able to do that"* P04 .

Other participants agreed that more work is required, such as that they will *"probably have to, uh, query GPT more on specific questions"* (P07) and *"[read GPT responses] again and see if there's anything else that stands out"* (P04). Thus, participants suggested that such limitations could be addressed by continued audits as part of their everyday GPT use.

5 Discussion and Implications

Our findings indicate that students find current guidelines confusing and inactionable (**RQ1**) and can audit more methodically after providing them with scaffoldings (**RQ2**). Our findings further point to specific design implications for interventions that aim to support critical evaluation of LLMs (**RQ3**). Here, we situate these findings within the larger discourse on critical engagement with LLMs in AI literacy pedagogy, current institutional policies, and everyday user auditing domains. We conclude with design and policy implications.

5.1 RQ1: Ability to Identify LLM Biases

Our findings indicate that while university guidelines are much needed to support critical evaluation, their current formulation is inactionable at best and inequitable at worst (RQ1).

5.1.1 Need for Guidelines Supporting Critical Use of GPT. Our findings validate educators' concerns [19, 64, 84] regarding students' use of GPT tools without critical reflection. Despite observing several instances of problematic behavior, our participants did not critically investigate that behavior. Thus, our findings confirm that students make "unreflected" use of GPT for a variety of tasks including homework [64].

Students that do not notice GPT biases, cannot contest them or change their GPT interaction behavior. Our findings indicate that lack of awareness, rather than malicious intent, is the driving factor behind lack of critical engagement with LLMs. Awareness being a precursor to behavioral change [21], makes raising awareness of GPT limitations crucial [5].

Our findings further justify the need of university guidelines [19]. However, instead of only raising awareness that students need to critically evaluate the GPT they are interacting with, university guidelines should also indicate competencies [5, 24] required for critical evaluation, along with instructions on how to gain such competencies.

5.1.2 Current Guidelines are Inactionable and Inequitable. Our findings indicate that the general student population will likely struggle to critically evaluate university LLMs under current university guidelines, making such guidelines inactionable. Thus, those existing guidelines place unrealistic demands on students, who, lacking specific guidance, are forced to shoulder the responsibility of evaluating LLMs on their own.

Improving students' AI literacy could alleviate some of the burden. We found that currently AI literacy skills were developed only in technical students [50] and those with exposure to related concepts via formal education. However, translating learnt concepts from formal education [52, 58, 77, 80] in everyday settings was hard even for technical students. By simultaneously forcing responsibility and disregarding students' individual abilities, such guidelines further exacerbate academic inequalities [40], pointing to the need for considering equitable outcomes when constructing policy [19, 82].

5.2 RQ2: Effects of Auditing Scaffolding

Our findings indicate that providing students with end-user auditing scaffolding enhanced their ability to identify and document biases in LLMs by fostering focused and methodical critical thinking when evaluating LLMs (RQ2).

5.2.1 Towards Focused, Methodical Student-led Auditing. Prior to scaffolding guidance, our participants used opportunistic, but not systematic prompting strategies [122] to investigate biases in LLMs. Frequent GPT users recreated biases they had previously encountered, leading to unfocused investigations. Post-scaffolding, students significantly improved their analysis, honing in on specific wording, nuances, and tone.

Initially struggling to create hypothesis-driven prompts, students benefited from concrete examples provided by scaffolding. They shifted from random topic exploration to focused scenario-based learning within a single domain. After scaffolding was removed, some expanded their queries to domains like governance, police work, and societal standards of beauty.

Students stopped using brute-force methods to surface biases, and started to emulate effective prompt structures, contesting and holding GPT accountable for its biases. In the post-scaffolding stage, their auditing became more intentional, methodical, and purposeful, involving detailed experimentation and nuanced prompt adjustments, ultimately producing comprehensive audit reports.

5.2.2 Enhanced Critical Thinking with Scaffolding. We found that structured guidance significantly enhanced students' critical thinking when evaluating LLM biases. In the hypothesis stage [103], students developed folk theories of LLM bias origins relying on social and technical factors. Students with greater AI literacy skills relied on their understanding of how machine learning (ML) algorithms learn from data and steps in training ML models [72] to theorize bias origins in technical mode of thinking. Other students theorized that biases arise when GPT makes assumptions similar to humans. This in turn, shaped their interactions with the LLM. Depending on social or technical mode of thinking, students focused on testing concepts in social subjects such as inclusivity and race, or technical capabilities such as writing and research support.

However, interacting with scaffolding example developed complementary expertise. Students with social mode of thinking sought technical explanations of biases, while those with technical mode of thinking attempted understanding biases in a more socio-technical manner. Our findings thus extend prior work [103] by showing that users benefit from expert guidance, which we provide in the form of a scaffolding.

5.3 RQ3: Helping Students Deal with LLMs

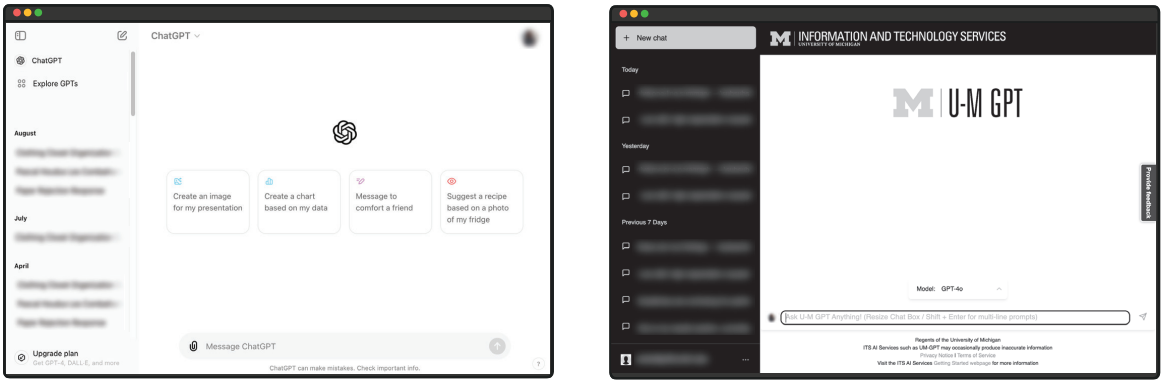
Our findings indicate that students engaged in active learning through the auditing activity. This underscores the need for designing educational interventions that facilitate hands-on exploration of AI systems. Here, we draw connections between our findings and potential designs that go beyond delegating the responsibility for evaluating LLMs to students (RQ3).

5.3.1 Designing Equitable LLM Use Policy. Our study findings showed that all of our participants found current university guidelines confusing; though, those with higher AI literacy skills were better equipped to critically evaluate LLMs on their own. Thus, a "one-size-fits-all" approach to formulating policy and guidelines may be inequitable. Equitable and effective guidelines should not only involve students in the policy-making process [126], ensuring that their voices are represented, but also shift the responsibility of evaluating LLMs from students to technology creators and providers. Thus, policies should hold universities accountable for mitigating biases in LLMs, promoting algorithmic responsibility by requiring regular audits and publicly sharing the results.

5.3.2 Designing Learning Environments All Students. Although our findings show the value of formal classroom education in developing AI literacy and AI evaluation skills, they also point to design opportunities for creation of informal learning environments and tools. Our participants without formal AI education engaged in active learning as they observed how the model response changes when they rephrase and add more details to prompts. They then leveraged what they learnt to construct effective prompts aligned with their evaluation goals. These findings inform the design of educational environments and curricula [10, 48] that incorporate a constructionist approach to educational interventions [30, 61, 83].

We also found that low AI literacy users sought explanations about what LLMs can do and how they work, either from the LLM itself or external online resources, after observing biased model behavior. Thus, our findings point to opportunities for designs that incorporate interactive explanations [11] that support students' sense-making needs [17], in particular those that help explain the strengths and limitations of LLMs.

However, vast majority of the public is not AI literate [120], and gaining AI literacy through formal education is not feasible for everyone [74]. Thus, our work has implications for designs that foster learning AI literacy skills in informal settings,



(a) ChatGPT Interface

(b) University of Michigan GPT (U-M) Interface

Figure 7: Despite critical evaluation goals of the university being vastly different from OpenAI's user engagement goals, there was a striking similarity between (a) ChatGPT and (b) U-M GPT interface designs at the time of our study.

such as libraries, public spaces and museums [71, 72, 115]. Thus, by designing accessible and engaging educational interventions, we can foster AI literacy that caters to a broader audience beyond traditional academic environments.

5.3.3 Designing for Responsible LLM Use in Education. Universities' goal of responsible LLM use in an educational setting vastly differs from OpenAI's goal of maximizing user engagement. Yet, the interface design of educational and commercial GPT tools, including how they communicate their guidelines for responsible use, remains the same (Fig 7).

Our findings inform future interfaces that promote critical reflection on LLMs (e.g., our end-user audit scaffolding, self-reflection prompts [66], highlighting uncertain LLM outputs [117]) in a way that could help students shoulder the burden of evaluating LLMs. However, our findings also point to an immediate need to design LLM tools for specific educational tasks with interfaces that can be covered by meaningful and enforceable responsible use policy.

One of the main challenges of such an open-ended chatbot interface is that it is hard to create and communicate meaningful and enforceable policy that covers different ways in which students can prompt the LLM and ways in which the LLM can respond; i.e., the "policy surface area" of such a design is very large. Thus, it is important to replace the chatbot interface with designs for specialized educational "tools" that "wrap around" the LLM to support specific tasks in education.

6 Conclusion and Future Work

In this paper, we investigated how students critically engage with LLMs through an end-user auditing activity. Our key finding was that current university policies fall short of effectively supporting students with critical engagement, and instead defer the responsibility to students who struggle without structured support. Our study reveals that students default to their prior knowledge in the absence of structured support, which could further exacerbate academic inequalities as students with higher AI literacy skills are better equipped to critically engage with LLMs than those without. However, students demonstrate better critical engagement through both cognitive processes and behavioral changes when provided with structured support through an auditing scaffolding.

Our findings add to the discourse on the use of LLMs in education [1, 23, 32], supporting how user-led auditing can promote critical thinking skills in diverse populations [75, 107], including students. Our work opens up opportunities for future research, including critical examination of existing educational policies and their limitations in supporting diverse student population, as well as interface design that can promote critical thinking in everyday interactions. Although our work investigated how students critically engaged with LLMs, the findings are relevant to the broader HCI community interested in promoting critical engagement with LLMs among end users through a learner-centered lens.

Acknowledgments

We thank Daniel Goldstein, Jake Hofman, David Rothschild and Harsh Kumar for the formative ideas on Critical Engagement with LLMs. We thank Rackham Centre for Research, Learning and Teaching for theoretical guidance, and Xu Wang for help with designing the scaffolding. We thank Mark Guzdial for feedback with the study design, and for making it possible for us run the study in his classroom. We thank the Wallace House Centre for journalists and the Knight-Wallace Fellowship program for inviting us to run our study. We thank all members of the CompHCI Lab for their feedback and support. Thanks to Smokey and Nilah for providing feline entertainment through long nights. This article is based upon work in part supported by the National Science Foundation under Grant No. IIS-2237562.

References

[1] Muhammad Abbas, Farooq Ahmed Jam, and Tariq Iqbal Khan. 2024. Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education* 21, 1 (2024), 10. https://doi.org/10.1186/s41239-024-00444-7

[2] Omaima Almatrafi, Aditya Johri, and Hyuna Lee. 2024. A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). *Computers and Education Open* 6 (2024), 100173. https://doi.org/10.1016/j.caeo.2024.100173

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[4] Lorin W Anderson, David R Krathwohl, Peter W Airasian, Kathleen A Cruikshank, Richard E Mayer, Paul R Pintrich, James Raths, and Merlin C Wittrock. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, complete edition.* Addison Wesley Longman, Inc., New York.

[5] Ravinithesh Annapureddy, Alessandro Fornaroli, and Daniel Gatica-Perez. 2025. Generative AI Literacy: Twelve Defining Competencies. *Digit. Gov.: Res. Pract.* 6, 1, Article 13 (Feb. 2025), 21 pages. https://doi.org/10.1145/3685680

[6] Robert K Atkinson, Sharon J Derry, Alexander Renkl, and Donald Wortham. 2000. Learning from examples: Instructional principles from the worked examples research. *Review of educational research* 70, 2 (2000), 181–214.

[7] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 74 (apr 2021), 34 pages. https://doi.org/10.1145/3449148

[8] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. 2023. Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 27 (apr 2023), 17 pages. https://doi.org/10.1145/3579460

[9] Yasmine Belghith, Atefeh Mahdavi Goloujeh, Brian Magerko, Duri Long, Tom Mcklin, and Jessica Roberts. 2024. Testing, Socializing, Exploring: Characterizing Middle Schoolers' Approaches to and Conceptions of ChatGPT. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 276, 17 pages. https://doi.org/10.1145/3613904.3642332

[10] Jesse Josua Benjamin, Joseph Lindley, Elizabeth Edwards, Elisa Rubegni, Tim Korjakow, David Grist, and Rhiannon Sharkey. 2024. Responding to Generative AI Technologies with Research-through-Design: The Ryelands AI Lab as an Exploratory Study. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 1823–1841. https://doi.org/10.1145/3643834.3660677

[11] Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2023. On Selective, Mutable and Dialogic XAI: a Review of What Users Say about Different Types of Interactive Explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 411, 21 pages. https://doi.org/10.1145/3544548.3581314

[12] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 4 (September 2004), 991–1013. https://doi.org/10.1257/0002828042002561

[13] Sara Bock. 2024. Say Hello to TritonGPT: In move toward campuswide launch, UC San Diego's specialized AI information and resource assistant enters "second wave" pilot. https://today.ucsd.edu/story/say-hello-to-tritongpt

[14] David W Braithwaite and Lauren Sprague. 2021. Conceptual Knowledge, Procedural Knowledge, and Metacognition in Routine and Nonroutine Problem Solving. *Cognitive Science* 45, 10 (oct 2021), e13048. https://doi.org/10.1111/cogs.13048

[15] Shea Brown, Jovana Davidovic, and Ali Hasan. 2021. The algorithm audit: Scoring the algorithms that score us. *Big Data & Society* 8, 1 (2021), 2053951720983865. https://doi.org/10.1177/2053951720983865

[16] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[17] Ángel Alexander Cabrera, Abraham J. Druck, Jason I. Hong, and Adam Perer. 2021. Discovering and Validating AI Errors With Crowdsourced Failure Reports. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 425 (oct 2021), 22 pages. https://doi.org/10.1145/3479569

[18] Alexander Campolo and Kate Crawford. 2020. Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society* 6 (January 2020), 1–19. https://www.microsoft.com/en-us/research/publication/enchanted-determinism-power-without-responsibility-in-artificial-intelligence/

[19] Cecilia Ka Yuk Chan. 2023. A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 38. https://doi.org/10.1186/s41239-023-00408-3

[20] Cecilia Ka Yuk Chan and Wenjie Hu. 2023. Students' voices on generative AI: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 43. https://doi.org/10.1186/s41239-023-00411-8

[21] Sunny Consolvo, Predrag Klasnja, David W. McDonald, Daniel Avrahami, Jon Froehlich, Louis LeGrand, Ryan Libby, Keith Mosher, and James A. Landay. 2008. Flowers or a robot army? encouraging awareness & activity with personal, mobile displays. In *Proceedings of the 10th International Conference on Ubiquitous Computing* (Seoul, Korea) *(UbiComp '08)*. Association for Computing Machinery, New York, NY, USA, 54–63. https://doi.org/10.1145/1409635.1409644

[22] Thomas D. Cook and D. T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Houghton Mifflin, Boston, MA.

[23] Debby R. E. Cotton, Peter A. Cotton, and J. Reuben Shipway. 2023. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International* 61, 2 (March 2023), 228–239. https://doi.org/10.1080/14703297.2023.2190148

[24] Manish Dadhich and Amiya Bhaumik. 2023. Demystification of Generative Artificial Intelligence (AI) Literacy, Algorithmic Thinking, Cognitive Divide, Pedagogical knowledge: A Comprehensive Model. In *2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG)*. IEEE, Indore, India, 1–5. https://doi.org/10.1109/ICTBIG59752.2023.10456172

[25] Aayushi Dangol, Michele Newman, Robert Wolfe, Jin Ha Lee, Julie A. Kientz, Jason Yip, and Caroline Pitt. 2024. Mediating Culture: Cultivating Socio-cultural Understanding of AI in Children through Participatory Design. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (IT University of Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 1805–1822. https://doi.org/10.1145/3643834.3661515

[26] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 352, 13 pages. https://doi.org/10.1145/3544548.3580672

[27] Wesley Hanwen Deng, Michelle S. Lam, Ángel Alexander Cabrera, Danaë Metaxa, Motahhare Eslami, and Kenneth Holstein. 2023. Supporting User Engagement in Testing, Auditing, and Contesting AI. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 556–559. https://doi.org/10.1145/3584931.3611279

[28] Michael A. DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. 2018. How People Form Folk Theories of Social Media Feeds and What it Means for How We Study Self-Presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173694

[29] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 626, 19 pages. https://doi.org/10.1145/3491102.3517441

[30] John Dewey. 2011. *Democracy and education: An introduction to the philosophy of education.* Simon & Brown, Hollywood, CA.

[31] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [stat.ML] https:

//arxiv.org/abs/1702.08608

[32] Ravit Dotan, Lisa S. Parker, and John Radzilowicz. 2024. Responsible Adoption of Generative AI in Higher Education: Developing a "Points to Consider" Approach Based on Faculty Perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2033–2046. https://doi.org/10.1145/3630106.3659023

[33] Upol Ehsan and Mark O. Riedl. 2024. Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns* 5, 6 (2024), 100971. https://doi.org/10.1016/j.patter.2024.100971

[34] Upol Ehsan, Elizabeth A Watkins, Philipp Wintersberger, Carina Manger, Sunnie S. Y. Kim, Niels Van Berkel, Andreas Riener, and Mark O Riedl. 2024. Human-Centered Explainable AI (HCXAI): Reloading Explainability in the Era of Large Language Models (LLMs). In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 477, 6 pages. https://doi.org/10.1145/3613905.3636311

[35] Robert H. Ennis. 1962. A Concept of Critical Thinking: A Proposed Basis for Research on the Teaching and Evaluation of Critical Thinking Ability. *Harvard Educational Review* 32, 1 (1962), 81–111.

[36] Sindhu Kiranmai Ernala, Asra F. Rizvi, Michael L. Birnbaum, John M. Kane, and Munmun De Choudhury. 2017. Linguistic Markers Indicating Therapeutic Outcomes of Social Media Disclosures of Schizophrenia. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 43 (dec 2017), 27 pages. https://doi.org/10.1145/3134678

[37] Nel Escher and Nikola Banovic. 2020. Exposing Error in Poverty Management Technology: A Method for Auditing Government Benefits Screening Tools. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 64 (may 2020), 20 pages. https://doi.org/10.1145/3392874

[38] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2371–2382. https://doi.org/10.1145/2858036.2858494

[39] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be Careful; Things Can Be Worse than They Appear": Understanding Biased Algorithms and Users' Behavior Around Them in Rating Platforms. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (May 2017), 62–71. https://doi.org/10.1609/icwsm.v11i1.14898

[40] Jayne Everson, F. Megumi Kivuva, and Amy J. Ko. 2022. "A Key to Reducing Inequities in Like, AI, is by Reducing Inequities Everywhere First": Emerging Critical Consciousness in a Co-Constructed Secondary CS Classroom. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education - Volume 1* (Providence, RI, USA) *(SIGCSE 2022)*. Association for Computing Machinery, New York, NY, USA, 209–215. https://doi.org/10.1145/3478431.3499395

[41] Peter A. Facione. 2011. *Critical Thinking: What It Is and Why It Counts*. The California Academic Press„ Millbrae, CA. https://api.semanticscholar.org/CorpusID:154805251

[42] Juan M. García-Ceberino, María G. Gamero, Sebastián Feu, and Sergio J. Ibáñez. 2020. Experience as a Determinant of Declarative and Procedural Knowledge in School Football. *International Journal of Environmental Research and Public Health* 17, 3 (Feb. 2020), 1063. https://doi.org/10.3390/ijerph17031063

[43] André Groß, Amit Singh, Ngoc Chi Banh, Birte Richter, Ingrid Scharlau, Katharina J Rohlfing, and Britta Wrede. 2023. Scaffolding the human partner by contrastive guidance in an explanatory human-robot dialogue. *Frontiers in Robotics and AI* 10 (2023), 1236184.

[44] Lara Groves, Jacob Metcalf, Alayna Kennedy, Briana Vecchione, and Andrew Strait. 2024. Auditing Work: Exploring the New York City algorithmic bias audit regime. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1107–1120. https://doi.org/10.1145/3630106.3658959

[45] Rahem D. Hamid and Claire Yuan. 2023. *Harvard Releases First Guidelines for 'Responsible Experimentation with Generative AI Tools'*. The Harvard Crimson. https://www.thecrimson.com/article/2023/7/14/harvard-ai-guidelines/ Accessed: May 16, 2024.

[46] Alan Hamlin, Casimir Barczyk, Greg Powell, and James Frost. 2013. A comparison of university efforts to contain academic dishonesty. *J. Legal Ethical & Regul. Isses* 16 (2013), 35.

[47] Office of the Provost Harvard University. 2024. Guidelines for Using ChatGPT and other Generative AI tools at Harvard. https://provost.harvard.edu/guidelines-using-chatgpt-and-other-generative-ai-tools-harvard. Accessed May 16, 2024.

[48] Ingi Helgason, Michael Smyth, Enrique Encinas, and Ivica Mitrović. 2020. Speculative and Critical Design in Education: Practice and Perspectives. In *Companion Publication of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS' 20 Companion)*. Association for Computing Machinery, New York, NY, USA, 385–388. https://doi.org/10.1145/3393914.3395907

[49] Alex Hern. 2020. Twitter apologises for 'racist' image-cropping algorithm. https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm

[50] Marie Hornberger, Arne Bewersdorff, and Claudia Nerdel. 2023. What do university students know about Artificial Intelligence? Development and validation of an AI literacy test. *Computers and Education: Artificial Intelligence* 5 (2023), 100165. https://doi.org/10.1016/j.caeai.2023.100165

[51] Information and Technology Services, University of Michigan. 2024. Getting Started with ITS AI Services. https://its.umich.edu/computing/ai/getting-started. Accessed: April 26, 2024.

[52] Nikki Goth Itoi. 2023. Bringing AI Literacy to High Schools. Stanford University Human-Centered Artificial Intelligence. https://hai.stanford.edu/news/bringing-ai-literacy-high-schools Accessed: May 27, 2024.

[53] Sarah Jabbour, David Fouhey, Stephanie Shepard, Thomas S. Valley, Ella A. Kazerooni, Nikola Banovic, Jenna Wiens, and Michael W. Sjoding. 2023. Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study. *JAMA* 330, 23 (12 2023), 2275–2284. https://doi.org/10.1001/jama.2023.22295

[54] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. https://doi.org/10.1145/3544548.3581196

[55] Monique WM Jaspers, Thiemo Steen, Cor Van Den Bos, and Maud Geenen. 2004. The think aloud method: a guide to user interface design. *International journal of medical informatics* 73, 11-12 (2004), 781–795.

[56] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (mar 2023), 38 pages. https://doi.org/10.1145/3571730

[57] Yasmin Kafai, Chris Proctor, and Debora Lui. 2020. From theory bias to theory dialogue: embracing cognitive, situated, and critical framings of computational thinking in K-12 CS education. *ACM Inroads* 11, 1 (feb 2020), 44–53. https://doi.org/10.1145/3381887

[58] Martin Kandlhofer, Gerald Steinbauer, Sabine Hirschmugl-Gaisch, and Petra Huber. 2016. Artificial intelligence and computer science in education: From kindergarten to university. In *2016 IEEE Frontiers in Education Conference (FIE)*. IEEE, Erie, PA, USA, 1–9. https://doi.org/10.1109/FIE.2016.7757570

[59] Inkoo Kang. 2013. Businesses: 'Yelp is the thug of the Internet'. https://www.muckrock.com/news/archives/2013/jan/23/businesses-yelp-thug-of-the-internet

[60] Anna Kawakami, Luke Guerdan, Yanghuidi Cheng, Kate Glazko, Matthew Lee, Scott Carter, Nikos Arechiga, Haiyi Zhu, and Kenneth Holstein. 2023. Training Towards Critical Use: Learning to Situate AI Predictions Relative to Human Knowledge. In *Proceedings of The ACM Collective Intelligence Conference* (Delft, Netherlands) *(CI '23)*. Association for Computing Machinery, New York, NY, USA, 63–78. https://doi.org/10.1145/3582269.3615595

[61] Carmel Kent, Esther Laslo, and Sheizaf Rafaeli. 2016. Interactivity in online discussions and learning outcomes. *Computers & Education* 97 (June 2016), 116–128. https://doi.org/10.1016/j.compedu.2016.03.002

[62] Siu-Cheung Kong, William Man-Yin Cheung, and Guo Zhang. 2021. Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds. *Computers and Education: Artificial Intelligence* 2 (2021), 100026. https://doi.org/10.1016/j.caeai.2021.100026

[63] Jasmijn Kruijt, Corine S. Meppelink, and Lisa Vandeberg. 2022. Stop and Think! Exploring the Role of News Truth Discernment, Information Literacy, and Impulsivity in the Effect of Critical Thinking Recommendations on Trust in Fake Covid-19 News. *European Journal of Health Communication* 3, 2 (July 2022), 40–63. https://doi.org/10.47368/ejhc.2022.203

[64] Lars Krupp, Steffen Steinert, Maximilian Kiefer-Emmanouilidis, Karina E. Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann, and Jakob Karolus. 2023. Unreflected Acceptance – Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education. arXiv:2309.03087 [physics.ed-ph] https://arxiv.org/abs/2309.03087

[65] Bill Kules. 2016. Computational thinking is critical thinking: Connecting to university discourse, goals, and learning outcomes. *Proceedings of the Association for Information Science and Technology* 53, 1 (2016), 1–6.

https://doi.org/10.1002/pra2.2016.14505301092

[66] Harsh Kumar, Ruiwei Xiao, Benjamin Lawson, Ilya Musabirov, Jiakai Shi, Xinyuan Wang, Huayin Luo, Joseph Jay Williams, Anna N. Rafferty, John Stamper, and Michael Liut. 2024. Supporting Self-Reflection at Scale with Large Language Models: Insights from Randomized Field Experiments in Classrooms. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta, GA, USA) *(L@S '24)*. Association for Computing Machinery, New York, NY, USA, 86–97. https://doi.org/10.1145/3657604.3662042

[67] Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 79–85. https://doi.org/10.1145/3375627.3375833

[68] Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 512 (nov 2022), 34 pages. https://doi.org/10.1145/3555625

[69] Linfeng Li, Tawanna R. Dillahunt, and Tanya Rosenblat. 2019. Does Driving as a Form of. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 156 (nov 2019), 16 pages. https://doi.org/10.1145/3359258

[70] Petro Liashchynskyi and Pavlo Liashchynskyi. 2019. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS. arXiv:1912.06059 [cs.LG] https://arxiv.org/abs/1912.06059

[71] Duri Long, Takeria Blunt, and Brian Magerko. 2021. Co-Designing AI Literacy Exhibits for Informal Learning Spaces. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 293 (oct 2021), 35 pages. https://doi.org/10.1145/3476034

[72] Duri Long, Mikhail Jacob, and Brian Magerko. 2019. Designing Co-Creative AI for Public Spaces. In *Proceedings of the 2019 Conference on Creativity and Cognition* (San Diego, CA, USA) *(C&C '19)*. Association for Computing Machinery, New York, NY, USA, 271–284. https://doi.org/10.1145/3325480.3325504

[73] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376727

[74] Duri Long, Jessica Roberts, Brian Magerko, Kenneth Holstein, Daniella DiPaola, and Fred Martin. 2023. AI Literacy: Finding Common Threads between Education, Design, Policy, and Explainability. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 329, 6 pages. https://doi.org/10.1145/3544549.3573808

[75] Paul Machete and Marita Turpin. 2020. The Use of Critical Thinking to Identify Fake News: A Systematic Literature Review. In *Responsible Design, Implementation and Use of Information and Communication Technology*, Marié Hattingh, Machdel Matthee, Hanlie Smuts, Ilias Pappas, Yogesh K. Dwivedi, and Matti Mäntymäki (Eds.). Springer International Publishing, Cham, 235–246.

[76] Karen Mann, Jill Gordon, and Anna MacLeod. 2009. Reflection and reflective practice in health professions education: a systematic review. *Advances in health sciences education* 14 (2009), 595–621.

[77] David Martinez. 2024. AI System Architecture and Large Language Model Applications. https://professional.mit.edu/course-catalog/ai-system-architecture-and-large-language-model-applications Course offered by MIT Professional Education, Date: Oct 21 - 25, 2024. Registration Deadline: Oct 11, 2024. Location: Live Online. Accessed: August 12, 2024.

[78] Jessica Mathews. 2023. *Exclusive: Sam Altman quietly got $75M from the University of Michigan for a new venture capital fund earlier this year.* https://fortune.com/2023/12/19/sam-altman-quietly-got-75m-university-michigan-new-venture-capital-fund/ Accessed: 2025-02-11.

[79] Donald L. McCabe and Linda Klebe Trevino. 1993. Academic Dishonesty. *The Journal of Higher Education* 64, 5 (1993), 522–538. https://doi.org/10.1080/00221546.1993.11778446

[80] Logan McGrady. 2023. *UM-Flint is leading generative AI literacy with free online course.* University of Michigan-Flint. https://news.umflint.edu/2023/10/31/um-flint-is-leading-generative-ai-literacy-with-free-online-course/ accessed: August 12, 2024.

[81] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Found. Trends Hum.-Comput. Interact.* 14, 4 (nov 2021), 272–344. https://doi.org/10.1561/1100000083

[82] Fengchun Miao, Wayne Holmes, Ronghuai Huang, Hui Zhang, et al. 2021. *AI and education: A guidance for policymakers.* Unesco Publishing, 7, place de Fontenoy, 75352 Paris 07 SP, France. https://doi.org/10.54675/pcsp7350

[83] Joel Michael. 2006. Where's the evidence that active learning works? *Advances in Physiology Education* 30, 4 (Dec. 2006), 159–167. https://doi.org/10.1152/advan.00053.2006

[84] Silvia Milano, Joshua A. McGrane, and Sabina Leonelli. 2023. Large language models challenge the future of higher education. *Nature Machine Intelligence* 5, 4 (Mar 2023), 333–334. https://doi.org/10.1038/s42256-023-00644-2

[85] Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *J. Data and Information Quality* 15, 2, Article 10 (June 2023), 21 pages. https://doi.org/10.1145/3597307

[86] 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. https://doi.org/10.1016/j.caeai.2021.100041

[87] Davy Tsz Kit Ng, Chen Xinyu, Jac Ka Lok Leung, and Samuel Kai Wah Chu. 2024. Fostering students' AI literacy development through educational games: AI knowledge, affective and cognitive engagement. *Journal of Computer Assisted Learning* 40, 5 (2024), 2049–2064. https://doi.org/10.1111/jcal.13009

[88] Office of the Provost Northwestern. 2024. Generative AI Advisory Committee. https://www.northwestern.edu/provost/about/committees/generative-ai-advisory-committee/ Details about Northwestern's Generative AI Advisory Committee.

[89] Office of the Vice President for IT and CIO. 2023. Generative AI Advisory (GAIA) Committee. https://it.umich.edu/strategy-planning/gaia Details about University of Michigan's Generative AI Advisory Committee.

[90] Office of Information Technology AI Workgroup. 2023–2024. *UCI ZotGPT.* Office of Information Technology AI Workgroup. https://zotgpt.uci.edu/ © 2023–2024 Regents of the University of California. All rights reserved..

[91] Cian O'Mahony, Maryanne Brassil, Gillian Murphy, and Conor Linehan. 2023. The efficacy of interventions in reducing belief in conspiracy theories: A systematic review. *PLOS ONE* 18, 4 (April 2023), e0280902. https://doi.org/10.1371/journal.pone.0280902

[92] OpenAI. 2022. Introducing ChatGPT. https://openai.com/index/chatgpt/

[93] Chulwoo Park and Eric Coles. 2022. The impact of student debt on career choices among doctor of Public Health graduates in the United States: A descriptive analysis. *Int. J. Environ. Res. Public Health* 19, 8 (April 2022), 4836.

[94] Eliza Phares. 2024. *Innovation versus impact: The implications of U-M GPT.* https://www.michigandaily.com/columns/innovation-versus-impact-the-implications-of-u-m-gpt/ Accessed: 2025-02-11.

[95] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q. Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-Makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 379–396. https://doi.org/10.1145/3581641.3584033

[96] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173677

[97] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 148 (nov 2018), 22 pages. https://doi.org/10.1145/3274417

[98] Gilbert Ryle. 2009. *The Concept of Mind (60th Anniversary Edition)* (1st ed.). Routledge, London. 384 pages. https://doi.org/10.4324/9780203875858 eBook Published: 28 May 2009.

[99] Katrin Saks, Helen Ilves, and Airi Noppel. 2021. The Impact of Procedural Knowledge on the Formation of Declarative Knowledge: How Accomplishing Activities Designed for Developing Learning Skills Impacts Teachers' Knowledge of Learning Skills. *Education Sciences* 11, 10 (Sept. 2021), 598. https://doi.org/10.3390/educsci11100598

[100] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22 (2014), 4349–4357.

[101] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 410–422.

https://doi.org/10.1145/3581641.3584066

[102] W. R. Shadish, T. D. Cook, and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (2 ed.). Houghton Mifflin, Boston, MA.

[103] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 433 (oct 2021), 29 pages. https://doi.org/10.1145/3479577

[104] Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. 2023. ConvXAI: Delivering Heterogeneous AI Explanations via Conversations to Support Human-AI Scientific Writing. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 384–387. https://doi.org/10.1145/3584931.3607492

[105] Chris Shull and Gregory Hart. 2023. *WashU Addresses AI Technology*. Washington University of Medicine in St Louis, Office of Education.

[106] Seren Smith, S. Warnes, and A. Vanhoestenberghe. 2018. Scenario-based learning. In *Teaching and Learning in Higher Education: Perspectives from UCL*, JP Davies and N. Pachler (Eds.). UCL IOE Press, London, UK, 144–156. Green open access.

[107] Jaemarie Solyst, Ellia Yang, Shixian Xie, Amy Ogan, Jessica Hammer, and Motahhare Eslami. 2023. The Potential of Diverse Youth as Stakeholders in Identifying and Mitigating Algorithmic Bias for a Future of Fairer AI. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 364 (oct 2023), 27 pages. https://doi.org/10.1145/3610213

[108] Katta Spiel, Oliver L. Haimson, and Danielle Lottridge. 2019. How to Do Better with Gender on Surveys: A Guide for HCI Researchers. *Interactions* 26, 4 (jun 2019), 62–65. https://doi.org/10.1145/3338283

[109] Elizabeth Stafford. 2024. *Students and staff question UMich AI investments*. https://www.michigandaily.com/campus-life/students-and-staff-question-umich-ai-investments/ Accessed: 2025-02-11.

[110] Elisabeth Sulmont, Elizabeth Patitsas, and Jeremy R. Cooperstock. 2019. Can You Teach Me To Machine Learn?. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (Minneapolis, MN, USA) *(SIGCSE '19)*. Association for Computing Machinery, New York, NY, USA, 948–954. https://doi.org/10.1145/3287324.3287392

[111] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (may 2013), 44–54. https://doi.org/10.1145/2447976.2447990

[112] Harvard University Information Technology. 2024. *Initial guidelines for the use of Generative AI tools at Harvard*. Harvard University Information Technology. https://huit.harvard.edu/ai/guidelines Accessed May 16, 2024.

[113] David Touretzky, Christina Gardner-McCune, Fred Martin, and Deborah Seehorn. 2019. Envisioning AI for K-12: what should every child know about AI?. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Honolulu, Hawaii, USA, Article 1216, 5 pages. https://doi.org/10.1609/aaai.v33i01.33019795

[114] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

[115] Saniya Vahedian Movahed, James Dimino, Andrew Farrell, Elyas Irankhah, Srija Ghosh, Garima Jain, Vaishali Mahipal, Pranathi Rayavaram, Ismaila Temitayo Sanusi, Erika Salas, Kelilah Wolkowicz, Sashank Narain, and Fred Martin. 2024. Introducing Children to AI and ML with Five Software Exhibits. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 202, 6 pages. https://doi.org/10.1145/3613905.3650991

[116] Priyan Vaithilingam, Tianyi Zhang, and Elena L. Glassman. 2022. Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 332, 7 pages. https://doi.org/10.1145/3491101.3519665

[117] Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q. Vera Liao, and Jennifer Wortman Vaughan. 2024. Generation Probabilities Are Not Enough: Uncertainty Highlighting in AI Code Completions. *ACM Trans. Comput.-Hum. Interact.* (Oct. 2024). https://doi.org/10.1145/3702320 Just Accepted.

[118] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[119] Tianjia Wang, Daniel Vargas Diaz, Chris Brown, and Yan Chen. 2023. Exploring the Role of AI Assistants in Computer Science Education: Methods, Implications, and Instructor Perspectives . In *2023 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE Computer Society, Los Alamitos, CA, USA, 92–102. https://doi.org/10.1109/VL-HCC57772.2023.00018

[120] John Werner. 2023. Billions Of People Need To Learn AI Literacy. https://www.forbes.com/sites/johnwerner/2024/07/17/billions-of-people-need-to-learn-ai-literacy/ Innovation, AI.

[121] Jeffrey R. Young. 2024. Inside the Push to Bring AI Literacy to Schools and Colleges. Edsurge Podcast. https://www.edsurge.com/news/2024-01-23-inside-the-push-to-bring-ai-literacy-to-schools-and-colleges Accessed: May 27, 2024.

[122] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. https://doi.org/10.1145/3544548.3581388

[123] Ruochen Zhao, Xingxuan Li, Yew Ken Chia, Bosheng Ding, and Lidong Bing. 2023. Can ChatGPT-like Generative Models Guarantee Factual Accuracy? On the Mistakes of New Generation Search Engines. arXiv:2304.11076 [cs.CL]

[124] Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting Hallucinated Content in Conditional Neural Sequence Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 1393–1404. https://doi.org/10.18653/v1/2021.findings-acl.120

[125] Abigail Zimmermann-Niefield, Makenna Turner, Bridget Murphy, Shaun K. Kane, and R. Benjamin Shapiro. 2019. Youth Learning Machine Learning through Building Models of Athletic Moves. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children* (Boise, ID, USA) *(IDC '19)*. Association for Computing Machinery, New York, NY, USA, 121–132. https://doi.org/10.1145/3311927.3323139

[126] Maja Zonjić. 2024. Need a policy for using ChatGPT in the classroom? Try asking students. https://doi.org/10.1038/d41586-024-01691-4