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

ANONYMOUS AUTHOR(S)

Recognizing that Large Language Models (LLMs) can hallucinate or generate unacceptable responses, universities increasingly caution students to check responses for accuracy and appropriateness. However, existing university policies defer the responsibility of critical evaluation to students and assume that they will have the required knowledge and skills to do so on their own Here, we conducted a series of user studies (N=47) to understand how students critically engage with LLMs at a large North American public research university. In our studies, participants evaluated an LLM 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 indicate that students with higher AI literacy skills are better equipped, while those without these 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**.

Additional Key Words and Phrases: Critical Thinking, Auditing LLM, AI literacy, Scaffolding and Interactive Methods

ACM Reference Format:

Anonymous Author(s). 2018. "Here the GPT made a choice, and every choice can be biased": How Students Critically Engage with LLMs through End-User Auditing Activity. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 27 pages. https://doi.org/XXXXXXX.XXXXXXX

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 [38, 48, 73] enables end-users to analyze, question, and contest information generated from Large Language Models (LLMs) in consequential settings, such as education [71, 114, 130]. By developing AI literacy competencies [3, 85, 86] that help understand AI limitations and ethical implications, end-users can exercise critical thinking to evaluate errors [56], biases [26, 69, 112, 144–146], and "hallucinations" of LLMs—believable but factually inaccurate responses [60, 94, 134, 137, 147]. For example, users can exercise critical thinking to determine the reliability of LLM-generated information [133] and safeguard themselves from misinformation [77, 89, 107].

Recognizing that Large Language Models (LLMs) can hallucinate or generate incorrect or unacceptable responses, universities increasingly caution students to verify LLM outputs for accuracy and appropriateness. However, existing university policies [1, 18, 54, 57, 63, 105, 129] do not specify *how* to evaluate LLMs or *what* to look for in their responses. These guidelines assume students already have the required skills to interact with LLMs, causing stress and uncertainty for those who lack those skills [2, 25].

Although students with computer science backgrounds could have sufficient AI literacy [62], the lack of similar levels of AI literacy in students from non-technical backgrounds [76, 127] places them at a significant disadvantage under current guidelines, exacerbating academic inequalities [48]. Universities increasingly offer coursework that teaches students across disciplines what LLMs are and how to use them [64, 71, 92, 96]; 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, potentially risking academic penalties, including charges of plagiarism and expulsion [30].

Manuscript submitted to ACM

Growing calls for user-centered evaluation of AI [5, 39, 81, 135] has led to the development of evaluation tools [11, 67, 121, 140], explainability mechanisms [16, 42, 141], and end-user auditing methods [37, 81, 122, 125]. However, existing publicly available explainability mechanisms [24, 51] often mislead end-users to overly rely on AI [13, 21, 33, 41, 65, 80]. Despite the potential for end-users to identify and document AI bias [37, 122, 125] and a few notable cases of end-users successfully auditing AI systems on their own [47, 59, 72, 122], it is unclear if and to what degree such investigation can lead to methodical LLM evaluation 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 critically reflect on and identify biases in LLMs?

To answer these research questions, we developed a study 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 [29, 120] to develop an end-user auditing protocol for identifying biases and harms in LLMs. In our study design, participants interacted with PromptAuditor to perform an end-user auditing [37, 81, 122] of an LLM 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 (N=47) at a large North American public research university. In the first study, we focused on students (N=8) recruited by word of mouth from diverse academic disciplines and their cognitive and decision-making processes, which we elicited using a think-aloud method [68]. In the second study, we focused on analyzing the in-the-wild behavior of students from two different cohorts with different backgrounds and levels of AI literacy; one with students (N=30) at the end of a semester-long course on Generative AI from an arts and sciences department, and the other in a journalist workshop (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 conversation logs.

Our results highlight that students with stronger AI literacy and technical backgrounds are better equipped to identify biases, allowing for more comprehensive verification and 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 socio-technical aspects of bias propagation. Finally, students' behavioral interactions also transition from random exploration of diverse subjects in their prompts to focused evaluation of biases within a specific domain after scaffolded guidance.

Our work contributes empirical knowledge on how students critically engage in different processes of everyday algorithmic auditing. 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 support users in critically engaging with LLM technologies and provide concrete examples and concepts. We highlight future research opportunities, including interactive LLM tutorials for user onboarding, targeted scaffolding for interface design, and bias awareness through reporting biased LLM behavior and dissemination of knowledge in the academic community.

2 BACKGROUND AND RELATED WORK

Critical engagement with LLMs is an active topic of interest in a variety of communities, from pedagogy and educational policies to Human-Computer Interaction (HCI). 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 sketch how universities stress the need for LLM evaluation and gaps in their current LLM policies and guidelines. We then point to existing approaches to support users in evaluating AI technologies and documenting their limitations.

[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.

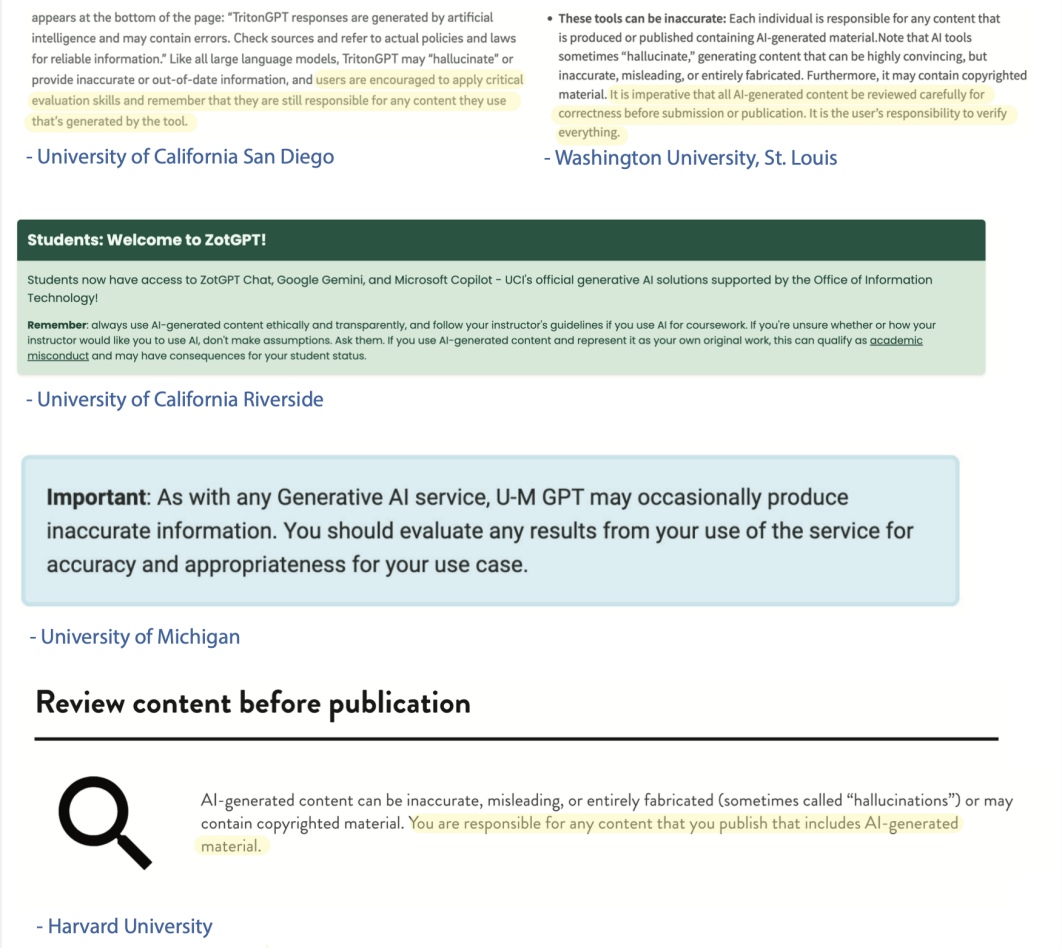


Fig. 1. Screenshots of guidelines stated by various universities [1, 18, 63, 105, 129] regarding the use of GPT tools. Guidelines strongly state that it is the student's responsibility to review all AI-generated content for "appropriateness" without stating how.

2.1 Importance of Critical Thinking and AI Literacy

Critical thinking [6] is a higher-order, analytical thinking process [131] that requires deliberate effort to analyze, evaluate and judge the credibility of information [43, 70, 79]. Educational psychology and cognitive science [99, 115] state both declarative (knowing *what*) and procedural (knowing *how*) knowledge are required for critical thinking in educational settings [6]. Exercising critical thinking when interacting with LLMs makes users less susceptible to errors [9, 61, 88, 145], biases [26, 69, 112, 144–146] and "hallucinations" [60, 94, 134, 147]. This further promotes appropriate reliance [8, 49, 91, 110, 124], which can safeguard users from harms such as misinformation and fake news [77, 89, 107].

AI literacy [86, 102] enables users to apply foundational skills of critical thinking to evaluate AI technologies. However, an overwhelming majority of the public is not AI literate [138], and developing these skills is a slow and ongoing process [71, 87, 114, 130] requiring deliberate effort. Such efforts include engaging diverse user groups, including children, by designing informal learning spaces [84, 85, 132] and activities [14, 32, 101, 148]. Educational institutions are increasingly offering structured classroom instruction to teach LLMs and related technologies, covering some skills and knowledge for interacting with generative AI [7, 64, 142]. 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 [106], universities lacked clear guidelines and policies for the use of LLM technologies in academic settings [40, 136]. To address educators' concerns regarding the use of these technologies for academic

dishonesty and plagiarism [25, 27, 78, 100], many universities formed interdisciplinary committees [103, 104] involving educators, technologists, and ethicists (with students' voices largely absent [25, 149]) to draft comprehensive policies. Being designed as an extension of academic dishonesty policies [55, 57, 95], the consequences of not following those university guidelines on LLMs can result in charges of academic dishonesty and even expulsion [30].

Most existing university guidelines [1, 18, 54, 57, 63, 105, 129] 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 assume that all students have high AI literacy and relevant competencies to perform this difficult task [19]. Few of those policies, if any, recognize the ability of AI to deceive [13], project infallibility [23], and affect students' views [66]. 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

Human-Computer Interaction (HCI) research community [5, 39, 81, 135] has long recognized the importance of supporting end-users in evaluating AI in everyday interactions. By providing users insights into the AI's working and decision-making criteria and processes, explanation mechanisms that promote transparency[111, 123] are meant to help end-users identify and reject AI decisions in situations in which its reasoning was incorrect [21]. However, most existing explanation mechanisms tend to act as evidence of reliability rather than accountability [21, 33], often deceiving end-users into over-relying on AI [65, 80]. Also, such mechanisms do not directly translate to the context of LLMs [42], where problems of over-reliance could be exacerbated [66].

Algorithmic auditing methods [20, 45, 97, 113] have been effective in identifying and documenting the weaknesses of different algorithmic systems. While traditionally catering to system developers [12, 117], recent work has focused on the potential for adopting the methods for end-users to identify and document AI bias [35, 37, 81, 122, 125]. Work on everyday algorithm auditing [37, 122] has documented how end-users audit AI in everyday interactions by raising awareness, hypothesizing, and testing to 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 DECLARATIVE AND PROCEDURAL KNOWLEDGE IN AUDITING

Here, we study how students' declarative knowledge (i.e., knowing *what*) and procedural knowledge (i.e., knowing *how*)—impacts their cognitive processes and interaction behavior as they audit an LLM. To answer our research questions, we adopted a quasi-experimental study design where students with different levels of AI literacy interacted with our study probe, PromptAuditor.

3.1 Operationalizing University Guidelines

We conducted a series of user studies at a large public research university in North America. This institution was among the first to offer students a suite of Generative AI (GenAI) tools and learning resources, including courses and online materials. We use the GPT-3.5 Turbo model and interface from this suite for our study, which we refer to as "U-GPT" for anonymity. 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 the university's high-level guidelines (i.e., acknowledging that U-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 [81, 122], as it closely mirrors real-world scenarios where students must critically engage with AI tools without extensive prior training. We designed PromptAuditor as a 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). Students interacted with the A main interface across all stages: pre-scaffolding, with-scaffolding, and post-scaffolding (Fig. 2). B Scaffolding 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-GPT" chat interface. The A1 audit report panel, inspired by the IndieLabel system [81], allows users to document findings by filling out topic, evidence, and summary fields. The A2 chat interface allows users to issue prompts and view responses from the underlying GPT-3.5-Turbo

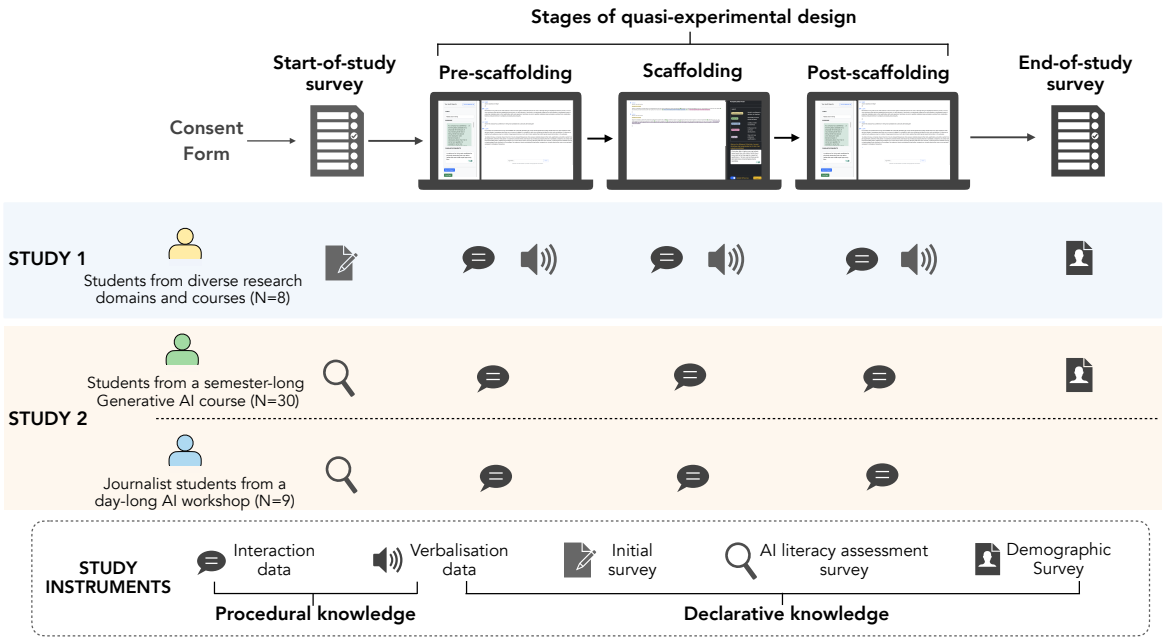


Fig. 2. We conduct a quasi-experimental study design across three contexts: a think-aloud study with 8 students from diverse domains, a one-day workshop with 9 journalists after a brief introductory AI lecture, and a classroom deployment with 30 students at the end of a semester-long AI literacy course. Symbols in the legend represent the study instruments used to measure declarative and procedural knowledge at different stages of the study.

model. Note that the design of A2 is based on the existing interface of the university's GPT (Fig. 7). We piloted sequential and chat-based layouts for the main interface with non-expert users (N=5) and found a preference for the chat-based layout, though users struggled to create audit reports using the original U-GPT interface A2 on its own.

3.2.2 Scaffolding Design and Rationale. To develop our scaffolding, we consulted AI education experts and identified scaffolding principles – B1 scenario-based learning [119], B2 hypothesis generation [122], B3 worked examples [10], B4 self-reflection [90] and B5 contrastive learning [52]. We created initial designs, exploring scenarios relevant to student life and consulting literature for harms [53], such as student loans [108], health insurance, and hiring [17]. Through sessions with AI auditing expert co-authors, we generated worked examples and selected the most salient one for hiring bias [82]. 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 limited highlights. Our final design combined GPT-generated highlights with the manual review featuring toggled highlights. Low-fidelity prototypes were critiqued iteratively with interdisciplinary input, leading to a final design tested with another two pilot participants, revealing no usability issues. Finally, we created the B scaffolding in B1, a scenario of racial bias in the hiring domain [128]. Here, B3 generates two cover letters—one with a Caucasian sounding name ("Christopher Allen"), and the other with an African American sounding name ("Latisha Smith"), referenced from prior correspondence audit work [82] and field experiment on labor market discrimination [17]. After self-explanation based on B4, participants engage in B5 contrastive learning to carefully examine this bias.

3.2.3 Software Implementation. We initially built a Chrome extension but later migrated to a Django[2] web application for privacy concerns raised by AI education experts. We implemented the interface using HTML, CSS, and JavaScript[3], with user data securely stored in a MySQL[4] database.

[2] https://www.djangoproject.com/
[3] https://www.w3.org/wiki/The_web_standards_model_-_HTML_CSS_and_JavaScript
[4] https://www.mysql.com/

Fig. 3. Our study interface probe, which we call PromptAuditor has 2 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.

3.3 Overview of User Studies

In this work, we use the lens of integrated knowledge theory [6] to study students' declarative knowledge (i.e., knowing *what*) and procedural knowledge (i.e., knowing *how*) to guide our study design, data collection, and analysis. We employed a quasi-experimental study design [29, 120], which has been traditionally employed in learning science studies to assess declarative and procedural knowledge [50, 116]. We conducted two user studies with students at the same North American public research university, which LLM guidelines we operationalized (Section 3.1).

In both studies, participants used the PromptAuditor design probe to audit U-GPT in an end-user auditing activity [37, 81, 122] across three stages (Fig. 2). 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 with A with minimal guidance. During the scaffolding stage, B guided them through a worked example of LLM bias. In the post-scaffolding stage, the scaffolding B was deactivated, and participants completed another 10-minute audit on their own. Below, we outline differences between the studies, including participants, methods, data collection, and analysis.

3.4 Study 1: Lab-controlled Think-aloud

We investigated students' cognitive and decision-making processes using a think-aloud protocol in a controlled lab environment.

3.4.1 Participants. We recruited (N=8) participants above the age of 18 from diverse fields, including Biology and Software Development (Table 1). To ensure a diverse participant pool, we employed random sampling and word of mouth, and stopped recruiting after reaching data saturation. We compensated participants at $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.

Table 1. The think-aloud study collected 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	Prior GenAI Use	Current Field	ML Class	ML Algm	Stats Class	Auditing Confidence
P01	Man	25-34	College Degree	Rarely	Software Development	No	No	No	Not Confident
P02	Woman	18-24	College Degree	Rarely	Biology	No	No	No	Not Confident
P03	Woman	18-24	College Degree	Once a week	Data Science	Yes	Yes	Yes	Confident
P04	Woman	18-24	College Degree	Several times a week	Urban Technology	No	No	Yes	Not Confident
P05	Man	18-24	Master's Degree	Several times a week	ECE	Yes	Yes	Yes	Somewhat Confident
P06	Man	18-24	Master's Degree	Several times a week	ECE	Yes	Yes	Yes	Confident
P07	Woman	25-34	Doctoral Degree	Several times a week	Bioinformatics	No	No	Yes	Somewhat Confident
P08	Woman	25-34	Doctoral Degree	Once a week	Information	Yes	Yes	Yes	Confident

3.4.2 Study-specific Tasks and Procedures. We conducted the study in-person or via Zoom [5], depending on participant availability, with no 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 [68] protocol using a brief video tutorial [6]. Participants shared their screens via Zoom and completed the task (Section 3.3) while thinking aloud during 60-minute sessions.

3.4.3 Data Collection. We conducted a brief start-of-study survey (Fig. 2) 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 [126], including gender, age, education, and occupation, through an end-of-study survey.

[5] https://zoom.us/
[6] https://www.nngroup.com/articles/thinking-aloud-demo-video/

3.5 Study 2: Naturalistic Educational Environments

We investigated students' in-the-wild interaction behavior in two classroom settings: at the end of a semester-long GenAI course with its students and during a day-long workshop with journalists following a brief lecture on generative AI.

3.5.1 Participants. We recruited two cohorts (N=39) with different levels of exposure to classroom instruction on AI literacy concepts. The first cohort was students (N=30) from a semester-long Generative AI course offered by the university's arts and sciences department, designed for students from diverse backgrounds with no programming requirements. We conducted this study during an invited lecture near the end of the semester. Out of this, 18 students responded to the demographics questionnaire, were all aged between 18 to 24, studied in various fields, and used Generative AI less frequently than a few times a month to daily (Table ??). The second cohort consisted of journalists (N=9) from various countries attending a fellowship program at the same university. We did this study immediately after an hour-long AI lecture as part of a workshop. Due to the small number of participants, we did not collect or report their demographic information to protect their anonymity, including any quotes in this paper.

Table 2. In-the-wild study participant demographics. Data was collected from university undergraduate students (18-24 years). We did not collect journalists' (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	Student	Prefer not to disclose	Daily
S50	Computer Science, Cognitive Science	Woman	Once a week
S52	College	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	LSA	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

3.5.2 Study-specific Tasks and Procedures. We conducted the study in classroom settings with both cohorts. Investigators were present and took field notes. One investigator distributed sticky notes with random numbers as participant codes, displayed the QR code and link for the PromptAuditor website (Fig. 3), and asked participants to access it on their laptops. We introduced the study and allotted 20 minutes for the AI literacy survey, asking participants to use their sticky note numbers as their participant ID. We followed the study protocol (Fig. 2), allocating 10 minutes per stage to maintain a structured timeline. After the audit, GenAI course students completed a 2-minute demographics survey.

3.5.3 Data Collection. We collected participants' responses to an AI literacy assessment survey completed before the auditing activity with PromptAuditor. Unlike existing AI literacy assessments [4] that focus on AI more broadly, we developed our own AI literacy assessment survey quiz specifically targeting LLM literacy, with 22 open-ended questions mapped to 17 AI literacy competencies defined by Long et al [86]. 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 1.5 hours.

3.6 Analysis of User Studies' Data

We analyzed data collected from Study 1 and Study 2 (Table **??**) together. Here, we provide details on the analysis we perform.

3.6.1 Interpretive Qualitative Analysis. We conducted interpretative qualitative analysis on think-aloud data from Study 1 and interaction data (i.e., prompts, GPT responses, audit reports) from both studies. We open-coded the think-aloud data from study 1, merged codes through axial coding, and then applied those themes to close-code the study 2 data. We transcribed [7] Study 1 audio recordings, annotated them with interaction details, and collected GPT conversation logs of participants from Study 2 who consented. 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. Finally, we used a MIRO [8] board for affinity diagramming to group codes into categories and themes.

3.6.2 AI Literacy Survey Assessment. The AI literacy survey acted as a proxy for assessing participants' declarative knowledge in Study 2 (Fig. 2). A total of 32 participants (25 journalists and 7 students) 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 [6]. We trained two authors as graders using sample answers and criteria for AI literacy competencies [86] 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 then averaged participants' scores.

3.6.3 Topic Modeling and Descriptive Statistics. To supplement qualitative coding, we use topic modeling, an established method for identifying themes that are otherwise not captured by sentence-level analysis [44]. 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 [83] 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.

Table 3. A total number of 39 students participated in the study. We conducted all data analyses separately, without matching participants or discarding data if the participant IDs didn't align. This approach allowed us to utilize all available data, maximizing the insight gained from each aspect of the study.

Cohort	Total Participants	Consented for AI Literacy Survey	Consented for Interaction Data Collection
Students	30	25	23
Journalists	9	7	9

3.7 Ethical Considerations

The study was reviewed and deemed exempt (i.e., approved) by our institutional ethics 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

[7] https://www.rev.com/
[8] https://miro.com/

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 our two studies, where each study offers a different "lense" to answer our research questions. We highlight key insights into how procedural knowledge influences the cognitive and decision-making processes that students and journalists undergo while interacting with and auditing LLMs. We connect these insights to their declarative knowledge, as assessed through our AI literacy assessment. Additionally, we mention key summary statistics and the diversity of relevant topics explored by each group, noting the similarities and differences observed in their interactions before and after using the scaffolding mechanisms in the PromptAuditor probe.

4.1 Bias Awareness and Hypothesis

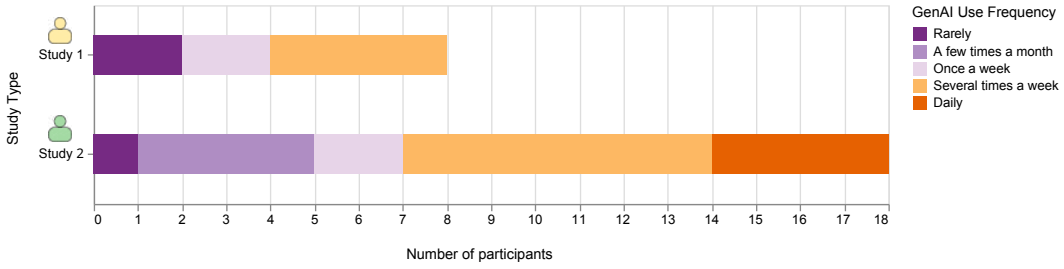


Fig. 4. A bar graph indicating frequency of Generative AI use of participants. 50% of Study 1 and 39% of Study 2 participants are low frequency users (rare-once a week). 50% of Study 1 and 61% of Study 2 participants are high frequency users of GenAI (several times a week-daily).

4.1.1 Students lack critical engagement with GPT despite noting problematic behavior. Students frequently use GPT without carefully reflecting on biases. Of the 18 students that answered the demographic survey (Fig 4), 11 used GPT more than once a week, and 4 used it daily, Some of those students used GenAI 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). Such frequent users reported problematic instances of GPT 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

Others acknowledged that they *"haven't explicitly seen biased outputs"* (S19). This points to a lack of critical engagement with GPT, even when frequent users come across instances of problematic behavior.

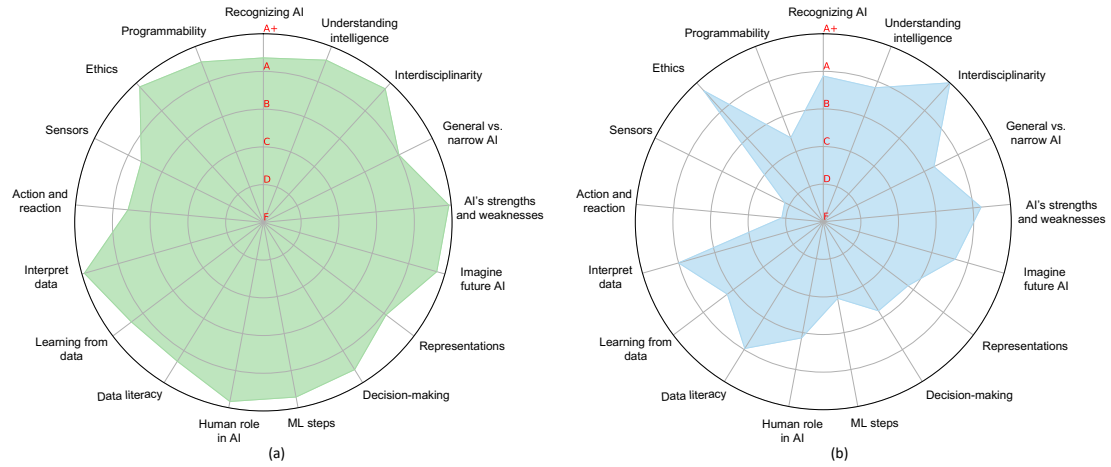


Fig. 5. The radar chart for (a) students and (b) journalists with distribution of scores across 17 AI-related competencies, such as recognizing AI, interdisciplinary understanding, AI strengths and weaknesses, and ethical considerations. Students (a) generally scored higher in competencies like "Recognizing AI" and "AI's Strengths and Weaknesses." Journalists (b) had a 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, students (a) showed reasonable AI literacy, with a mean score of 84.78 (B grade) and more consistent performance, indicating general proficiency. These findings underscore the importance of enhancing AI literacy, particularly for journalists, to ensure effective engagement with AI-related topics. .

4.1.2 AI literacy influenced how participants theorized origins of bias. Prior knowledge and AI literacy gained from *"reading"*(P07), *"attending [industry] talks"* (P07) or classes helped participants recognize how biases arise in GPT tools. Prior knowledge of AI Literacy skills such as ML Steps (students: 4.71; journalists: 2.07) and Data Literacy (students: 4.33; journalists: 3.96) enabled participants 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

Recognizing data collection practices enabled the participant to understand why and how GPT biases arise in a variety of contexts. Others gave examples of how specific biases can arise due to it's training data, such as *"gender and linguistic bias"*(P08), *"cultural bias"*(P06, S66) *"popularity bias"*(P01) and *"racial bias"*(S45), etc. Knowledge of 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) in GPT tools. 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 *"if-then-else [rules]? I think?"*(J21). In the absence of technical knowledge, participants created folk theories [36, 46] 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

The participant compares biases in GPT to human biases, and theorizes how GPT can make assumptions about someone's social status. When participants lacked data literacy, they 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 low-frequency users (P02, P03) defined biases as GPT being unable to understand intent in their prompts, more technical participants bias as *"lack of accuracy"*(P04) and *"LLM providing the wrong content"*(S03). Of the seven journalists who filled out AI literacy survey, four (J04, J16, J18, J23) mentioned misinformation as an ethical issue related to LLMs. Participants with lived experiences such as developing software

in industry settings (P01), researching technological harms (P08), or non-technical domain expertise (P02) adopted a social perspective, defining GPT *"giving very biased, one-sided opinion"* (P07) from *"less trustworthy sources"* (J23) as a bias, leading to harms such as:

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

Due to their field of study, P08 was deeply aware how LLMs propagate societal stereotypes and impose the dominant worldview. Thus, participants' modes of thinking about GPT biases were shaped by their backgrounds and experiences, emphasizing either technical inaccuracies or societal impacts. These modes also shaped what aspect of model behaviour they focused on while making a hypothesis. For example, students demonstrated technical thinking by making hypotheses based on numerical factors:

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

Here, the participant hypothesized how measurable factors, such as the confidence score will change, influenced by a technical mode of thinking. Conversely, journalists 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 speculated how numerical information would change, while participants with a social thinking speculated how societal biases and prejudices can come into play.

4.2 Struggle and Confusion

4.2.1 Students had difficulty understanding what and how to audit. Under instructions based on current university guidelines, participants struggled to understand what and how to audit. First, participants struggled to interpret what the task instructions (i.e., guidelines) mean:

> *"Um, I haven't... 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 *" [very] 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 low-frequency and non-technical users 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 Students struggled to identify and craft prompts. Journalists had a lower knowledge of how LLMs interpret data (average score: 4.0) than students (average score: 4.93) and were unable to construct effective prompts. This 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) as *"it doesn't answer ... exactly what I asked."*(P02). 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).

To resolve these issues, 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"*(P06) 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). Thus, participants struggled to craft effective prompts to elicit biases.

4.2.3 Students struggled 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 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"*. When participants made observations during their interactions with GPT, they expressed uncertainty in recognition of biases: *"I don't know if that makes sense [as bias]... it 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 human biases and prejudices. Participants exhibited a lack of critical evaluation during the interaction when they accepted the model's explanation without thorough scrutiny: *"Seems fair. "* P05 , *"that's pretty much what I said"* P07 . Some participants were distracted and nudged by GPT responses away from a critical auditing mindset due to the persuasive language used by the model. For instance, P04 shifted focus from finding biases in GPT to finding biases in hiring: *"I think 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

Here, we present how participants resolved their struggles. Note that the data supporting these findings is primarily from the think-aloud and understanding of participants' cognitive processes.

4.3.1 Gaining interface familiarity and model understanding through hands-on exploration. Actively engaging and experimenting with bias elicitation, leading to insightful observations and a deeper understanding of model capabilities. For example:

> *"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 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. Users 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 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). Users with low prompt engineering skills learnt how model behaviour 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 explorations also led to deeper questioning of model limitations based on chance discoveries, such as *"what is your last training cut off?"*(P01). Students such as S89 specifically 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."*. After hands-on interaction and probing 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 outweight 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 elements provide guidance on what and how to audit. Scaffolding elements provided guidance by breaking down the auditing process into comprehensible steps, thus improving participants' understanding of how to audit and evaluate biases in LLMs. The quote captures this:

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

Structured guidance led to improvements in knowing how to audit: *"because of this whole [scaffolding] I was able to conceptualize my thought process better and find out biases in a more easier manner"*(P06). Due to structured guidance also led to improvements in participants' declarative knowledge. Participants commented they *"understood the biases slightly better"* after scaffolding 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 set intentionality for auditing, as they verbalized 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.

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

4.3.3 Transitioning from random exploration to scenario-based prompts. Participants' prompting behaviour transitioned from issuing random, unrelated prompts to focusing on socio-technical aspects when issuing prompts. To resolve uncertainty in the pre-scaffolding stage, participants started by asking *"random questions to ask GPT"*(P07), such as greeting (P01, J19) the model: *"你好"*, asking it to explain known concepts such as: *'biology"*(P02), *"gravitational wave physics"*(S03) or easily testable questions: *"Is it raining today?"*(J21). However, after the scaffolding, participants focused on creating more social scenario type questions and re-used the scaffolding prompts to *"try with another topic [such as] politics, classroom situations."*(P06). For example:

> *"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, the participant creates a scenario as a police officer seeking GPT's to determine if GPT is racially biased. Participants also employed comparative analysis in their prompting, similar to scaffolding. For example, S79 asked GPT to *"build a cover letter [for] Mary Johnson"* and then for *"Shaquetta Morris"*(S79) with the rationale *to compare and contrast between two questions"*(P07). Thus, participants with higher prompt literacy, such as students, transitioned from random exploration to eliciting GPT biases by recreating scaffolding prompts. However, we did not observe this phenomenon in journalists.

4.3.4 Seeking additional information for bias evaluation and sense-making. When participants lacked domain knowledge or AI literacy, they sought additional information to support them in bias evaluation. Participants obtained such information by referencing *"Google"*(P08) or *"other online information"*(J04). For instance, 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 (J04) 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, participants attempted to seek additional information to make sense of observed biases in LLM responses.

4.3.5 Managing complexity with prompt engineering. To deal with fatigue caused by verbosity in responses and complexity of analysis, participants used prompt engineering techniques to control the length and formatting of the response. For example, P08 asked the model to limit the response:*"write a 50 words story about ..."*. Others asked for a specific formatting:*"Can you give me this information in a list format?"*(P02) to make the response easy to read. When they desired verbosity, such as when asking for an explanation, they would include words such as *"and why?"*(P06) to get longer responses. Knowledge of model's internal processes, such as *"limit of inputs"*(P03) shaped participants' interaction with the model. For example:

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

Here, S94 leveraged their knowledge of model's capabilities to issue a cloze task [109] and control response verbosity. Thus, participants leveraged their prompt engineering skills to manage response verbosity and interaction efficiency in order to deal with complexity and fatigue.

4.4 Translating Hypothesis to Prompting Strategies

4.4.1 Gauging Dominant Worldview Bias by Prompting Based on Current Events and Popular Figures. Participants who had knowledge of how training data affects model responses asked questions about recent events to see whether model echoes popular sentiment such as the *"eclipse"*(P07) and *"a lot of people [on social media] were joking about how all the Flat Earthers were going to get their theories proven wrong ... 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). Other participants asked about recent world events, such as war:

> *"I can ask about Israel and Gaza. [Types: What are your thoughts towards the war between Israel and Gaza? Do you feel Gaza is losing?] ... 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

Journalists asked the GPT for it's views on socio-political movements such as*"Black Lives Matter"* (J04), *"Racism in America"* (J10), *"hamas"* (J18) and *"nazis"* (J23). Participants asked the model for its views about influential figures, including *"Elon Musk"* (P06) and if *"Prompt: Do you think Tesla company would be a big hit in the near future?"*, *"Willis Ward* and *"Gerald Ford*(J04), and

"Soren Kierkegaard" and *"Joe Biden"* (J18). An overwhelming number of journalists(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:

> *"Currently .. 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

4.4.2 Uncovering Implicit Biases Through Vague and Unspecified Prompts. Participants issued vague prompts by withholding specific details in order to spot underlying, problematic assumptions made by GPT in the response :*"won't say that ... but let's see if it assumes"*(P06). This approach allowed them to *"see how it 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 50 words story"*(P08), *"write a haiku"* (J18), *"write a poem about leather boots in the style of Mary Oliver"* (J18), *"Generate a description of a plate of food"*(P05), and *"write a joke in the style of Seinfeld"* (J10) to see how model executes the task and find 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 such as gender in the prompt allowed participants to evaluate assumptions made by GPT and find hypothesized biases like gender bias:*"when asked about an engineer ChatGPT defaults to a male name."*(P08),

4.4.3 Evaluating GPT's Decision-Making Biases Through Ranking and Selection Prompts. Asking the model to give them a list of objects or attributes allowed participants to uncover biases in the way GPT ranks objects and uncover the model's problematic decision-making processes. For example, P01 asked the model to give them a list of touristic places and *"recommended activities in Brest"*, and learnt that the model recommends only the most famous options. They questioned:

> *"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. ChatGPT 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 critically questioned the GPT's decision-making criteria by asking "why not this" questions to contest its decisions about the things it recommends. They also note that every choice made by GPT can be biased. 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":*"this answer doesn't specify why [business school] is listed first. The [music school] being 4th perpetuates a stereotype that STEM is more meaningful."*. Others issued similar queries, asking the GPT for *"the worst president"*(S82), *"the one true religion"* (J18), *"the best race in the world?"* (J19). After the scaffolded interaction, they modified their approach to use such prompting to examine and elicit specific biases, such as racial bias:

> *"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, the participant was able to apply what they learnt during the scaffolded interaction to recreate previously seen racial bias:*"U-GPT sorted all the white names to the top of this perceived friendliness list"*. Thus, asking the model to rank based on some criteria or or "pick the best" in a category allowed students to find biases in its decision-making process.

4.4.4 Assessing Logical, Moral, and Ethical Biases through Riddling and Role-Playing Prompts. Students, ideating a potential area of GPT struggle, tested GPT on human concepts 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 interacted with GPT in a human-like manner, asking it for opinions in role-playing or hypothetical scenarios. Some issued conversational prompts to seek GPT's viewpoint on some topic:*"what are you views about Taj Mahal monument?"* (P06) and how it would do something:*"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"*. We also observed students 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 a logical puzzle, S59 successfully elicited an erroneous response: *"GPT just created a situation that isn't even possible by recommending that I take the pick up the goat from a side of the river that it is not even on"*.

Journalists asked GPT questions on themes of morality: *"Should I get married?"*(J10) and *"Does DEI discriminate against white people?"*(J16). When writing prompts, participants created elaborate scenarios, asking the GPT to role play as a human: *"imagine you are the recruiter"* (P06), *"pretend to be my therapist"*(S01), and *"roleplay as a staunch Republican"*(S19) to examine how GPT handles ethical dilemmas and figure out its values. Students also also assigned it role of a decision-support system: *"you are an AI used by the police"*(S19). While we did not observe journalists assigning the GPT roles, they inspected it's 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 GPT.

4.5 Bias Evaluation

4.5.1 Role of Prior Knowledge in Bias Detection. Prior knowledge helped students verify accuracy of information. Participants' relied on their prior knowledge of domain, culture and history to evaluate biases in the response and identify the response *"looks correct"* (P06 and whether it *"makes sense"*(P03). Others calling out the advantage of subject experts in verifying information with prior knowledge said,

> *"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. You either find that it's not true or it's not completely verifiable"* – P07.

Expertise supported prior knowledge did indeed come handy for P04 when evaluating biases as *"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 double check and make sure that the answer that they're getting is right."*. Knowing subject matter helped J04 spot hallucinations in the GPT's response: *"the reference to "his African-American teammate, Gerald Ford" is incorrect. Ford was white."*. However, they missed certain hallucinations when they did not possess subject knowledge like in the case of: *"Edward S. and Helen M. [last name] were long-time supporters of [university], particularly in the field of cancer research. Their significant contributions included gifts totaling $225 million, which led to the renaming of [the university's] cancer center to the [redacted] Cancer Center."*. The journalist J04 lacked knowledge of the names of university's buildings as they were new to the environment, leading them to miss this hallucination. Thus, while they were easily able to spot instance where model hallucinated due to having prior knowledge, a lack of such knowledge leads to missed biases.

4.5.2 Evaluating bias based on overall tone and sentiment. Participants carefully examined overall tone and sentiment of response to find biases, noting instances where the model was *"a bit more critical"* (P07), and *"how it wrapped around the [prompt] and focus more on [a particular aspect]"* (P03). Similarly, S89 noticed *"biased language use [when GPT was] explaining [concepts] in "ebonics" "*, indicating that while ebonics itself is not wrong, the model's application of it inappropriately exaggerated stereotypes. P05 similarly analyzed tone, commenting how *"when asked to generate a plate of food, GPT gravitates towards describing more western style food, majority of the time"*, finding Western bias in GPT responses. Similarly, J18 analyzed the style of poetry on asking model to *"Write a poem about leather boots in the style of Mary Oliver"*, and found that the response *"sounds nothing like the author prompted"*. Surprisingly, 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*

Instances like *"bias of programmers [manifests as] the model seems to take a neutral stance, which itself can be interpreted as a belief system."*(J10), support the claim that participants examine the overall tone and sentiment expressed in the response to evaluate biases, irrespective of whether they are positive or negative.

4.5.3 Evaluating bias based on specific wording. Students and participants examined specific terminology and its associations to find biases. For example, P08 identified problematic nuances in the terminology used by GPT:

> *"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*

P08 mentioned that while the word "tight-knit" itself had no negative meaning, the historical context in which it was used, and which is reflected in model response, was indeed problematic. P01 identified that GPT assumes the user's geographic location based on the language of the query, as GPT 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"*. Other participants focused on specific wording to examine inaccuracies: *"inaccurate information [given by model]: "[redacted] S. Knight" is the wrong middle initial. Can't easily tell from other online information whether it should be "[redacted] S. Knight" or "[redacted] L. Knight.""* (J04). Thus, participants examined specific words and terminology to evaluate biases.

4.6 Critical Questioning

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

> *" ... bad data influences ChatGPT 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. P07 commented how GPT lears from *"data created by less-than-nice people on the internet"* and *"picks up on those trends"*, leading to biases which can be observed when GPT makes ranking choices:*"here the GPT made a choice, and every choice can be biased"*(P01). P05 commented how GPT harms not only individuals *"using it to learn something or as a source of information"* but also *"perpetuates and propagates biases to the general public and the next generation of internet users"*. Others perceived bias as GPT providing *"harmful advice"* (S38) or *"dangerous information"*(S52).

4.6.2 Need for additional resources and support. Participants recognized limitations in their auditing approach, and expressed need for additional support:

> *"You can say three generations of a single prompt are not that useful, but again, it's something ... [but] 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, so... " – P05*

Here, the participant recognizes they weren't able to do a thorough audit given the constraints, but acknowledges making best use of the available resources. P04 agrees:*"there is a lot more that needs to be taken into consideration ... if I had more time, I'd be*

able to do that". Other participants attempted corrective measures, such as *"I'll probably have to, uh, query GPT more on specific questions"* (P07) and *"[reading the response] it again and see if there's anything else that stands out to me."* (P04). This highlights how participants recognized their auditing limitations and attempted to address them by planning for more extensive querying and iterative review processes, even when immediate resources were limited. This points to a larger need for additional support and resources to support their auditing efforts.

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

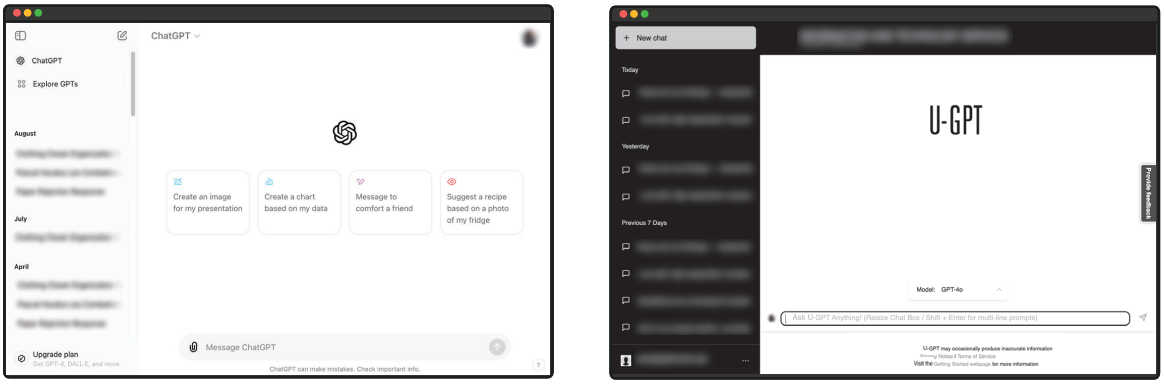
Our findings indicate that while guidelines are much needed to support critical evaluation, their current formulation is inactionable at best and inequitable at worst.

5.1.1 Need for guidelines supporting critical use of GPT. Firstly, our findings validate educators' concerns [25, 27, 78, 100] regarding students' use of GPT tools without critical reflection. Despite observing several instances of problematic behaviour, we found that students do not critically investigate these problematic behaviour. Our findings confirm prior work [78] which states that students make "unreflected" use of GPT for a variety of tasks including homework. As students do not notice these biases, they do not reflect on whether they should change their interaction in response to these biases. Our findings also indicate that lack of awareness, rather than malicious intent, is the factor driving lack of critical engagement with LLMs. The ubiquitious computing community has long established [28] that awareness leads to behavioural change and thus, greater awareness of problems and biases in GPT is crucial [7]. Our findings thus provides justification for the need of these guidelines [25]. However, instead of simply dictating that students need to critically evaluate, guidelines should instead raise awareness of competencies [7, 31] that students need for interacting with GenAI tools and how to implement them.

5.1.2 Current guidelines are inactionable at best and inequitable at worst. Students struggled to understand *what* and *how* to audit under current university guidelines (Fig 1) regardless of their AI literacy skills or technical background, making such guidelines inactionable. Despite lack of evidence [118] for users' ability to spot hallucinations, such guidelines place unrealistic demands on students. We found that students were able to spot hallucinations only when they possess prior factual knowledge, a factor that such guidelines do not consider. Lacking specific guidance, students were forced to shoulder the responsibility of evaluating LLMs on their own as they had to rely on their prior AI literacy skills. We found that AI literacy skills were more developed in technical students [62] and those with more exposure to these concepts via formal education. However, translating learnt concepts from formal education [64, 71, 92, 96] in everyday settings was hard even for technical students. By simultaneously forcing responsibility and disregarding students' individual abilities, such guidelines further exacerbate academic inequalities [48], pointing to the need for considering equitable outcomes [25, 98] in their construction.

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

5.2.1 Enhanced critical thinking with scaffolding. We found that structured guidance significantly enhanced participants' critical thinking when evaluating LLM biases. In the hypothesis stage [122], participants developed folk theories of LLM bias origins relying on social and technical factors. Particpants with greater AI literacy skills relied on their Data Literacy and ML Steps [85] competencies to theorize bias origins in technical mode of thinking, while participants with field and lived experience 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 (Fig 6). However, interacting with scaffolding example developed



(a) ChatGPT Interface

(b) Anonymized University GPT (U-GPT) Interface

Fig. 7. Interface Similarity between (a) ChatGPT and (b) University GPT (U-GPT). While the critical evaluation goals of the university vastly differ from OpenAI's user engagement goals, their interface design remains the same.

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 [122] by showing that users benefit from expert guidance, which we provide in the form of a scaffolding.

5.2.2 Behavioral Changes Towards Focused, Methodical Auditing. Prior to scaffolding guidance, students employed random and unrelated prompts to investigate biases in Large Language Models (LLMs). Frequent GPT users recreated biases they had previously encountered, leading to unfocused investigations. Post-scaffolding, participants 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. Initially using brute-force methods, participants evolved to emulate effective prompt structures, holding GPT accountable for biases and understanding its decision-making process. Post-scaffolding, their auditing became more intentional, methodical, and purposeful, involving detailed experimentation and nuanced prompt adjustments, ultimately producing comprehensive audit reports.

5.3 RQ3: What are the design opportunities for helping non-technical and low AI literacy students to critically reflect on and identify biases in LLMs?

Our findings indicate that low AI literacy students engage in active learning through the auditing activity, which underscores the need for designing educational interventions that facilitate hands-on exploration and critical evaluation of AI systems.

5.3.1 Need for critical design. Despite the diverging goals of universities and OpenAI, their interface design remains the same. Companies such as OpenAI seek to enhance user engagement to collect data for LLM training, and the ChatGPT interface supports these goals. However, the goal of universities is completely different—that of reducing over-reliance and encouraging critical evaluation of LLMs. Despite these diverging goals, we raise the provocative question of why interface design is the same. The scaffolding probe we designed for the purposes of this study was based on the goal of end-user auditing [81], and mapped to the stages of everyday algorithmic auditing [37, 122] led to auditing awareness, hypothesis generation, testing and evaluation, and reflection on bias responsibility. Thus, careful thought needs to be given to what goals the interface is supporting, and whose guidelines are being operationalized. Moreover, we found that low AI literacy and low-frequency users are easily influenced by how useful GPT is, and diverted from a critical mindset. Our findings support calls in the CHI community [139] to investigate the user effects of dark patterns [93](i.e.,manipulative design choices influence users on a cognitive level) in GenAI tools. By leveraging critical design [75] to develop interface elements, universities can move away from the current paradigm that prioritizes engagement and convenience, towards one that fosters critical evaluation [34] and deeper understanding.

5.3.2 Designing hands-on activities to foster AI literacy. We found that interactive exploration of LLM biases in auditing activity led to active learning [38], especially in low AI literacy students. For example, students with low AI literacy initially struggled to construct effective prompts to find biases, confirming prior work [143]. However, students engaged in active learning as they observed how the model response changes when they rephrase and add more details to prompts. Students then leveraged what they learnt to then construct effective prompts aligned to their goals. These findings support the need for designing educational environments and curricula [15, 58] that incorporate a constructionist approach [38, 74, 99]. We also found that low AI literacy users sought technical explanations, either from the model itself or external online resources, after observing biased model behaviour, supporting prior work [22] in AI sense-making. Future work can conduct detailed investigation to design interactive explanations [16] that support users' sense-making needs. Furthermore, our low AI literacy cohort were journalists with brief exposure to AI literacy lecture in a North American research university. However, a vast majority of the public is not AI literate [138], and gaining such skills through formal education is not always feasible [87]. Thus, our work points to efforts to foster AI literacy skills in informal settings, such as libraries, public spaces and museums [84, 85, 132]. Thus, by designing accessible and engaging educational interventions, we can foster AI literacy that caters to a broader audience beyond traditional academic environments.

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 current university policies fall short in 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 behavioural changes when provided with scaffolding and structured support.

Our findings add to the overarching discussion on the use of LLMs in education [2, 30, 40], supporting how end user auditing activity can promote critical thinking skills in diverse populations [89, 125], including students. Our work opens up opportunities for the future research, including critical examination of existing educational policies and their limitations in supporting diverse AI literacy in the student population, as well as interface design that can promote critical thinking in everyday interactions. While this work investigated how students critically engaged with LLMs, the findings are relevant to the larger CHI and HCI community interested in promoting critical engagement with LLMs in end users. We conclude by inviting researchers to adopt a learner-centered lens when designing interfaces and policies that support critical engagement.

REFERENCES

[1] 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..

[2] 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 (Feb. 2024). https://doi.org/10.1186/s41239-024-00444-7

[3] Safinah Ali, Blakeley H. Payne, Randi Williams, Hae Won Park, and Cynthia Breazeal. 2019. Constructionism, Ethics, and Creativity: Developing Primary and Middle School Artificial Intelligence Education. In *Proceedings of IJCAI 2019.*

[4] 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* (2024), 100173.

[5] 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

[6] 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.

[7] Ravinithesh Annapureddy, Alessandro Fornaroli, and Daniel Gatica-Perez. 2024. Generative AI Literacy: Twelve Defining Competencies. *Digit. Gov.: Res. Pract.* (aug 2024). https://doi.org/10.1145/3685680 Just Accepted.

[8] Theo Araujo. 2018. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior* 85 (2018), 183–189. https://doi.org/10.1016/j.chb.2018.03.051

[9] Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask Me Anything: A simple strategy for prompting language models. arXiv:2210.02441 [cs.CL]

[10] 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.

[11] Raghav Awasthi, Shreya Mishra, Dwarikanath Mahapatra, Ashish Khanna, Kamal Maheshwari, Jacek Cywinski, Frank Papay, and Piyush Mathur. 2023. HumanELY: Human evaluation of LLM yield, using a novel web-based evaluation tool. (Dec. 2023). https://doi.org/10.1101/2023.12.22.23300458

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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. *UC San Diego Today* (19 March 2024). https://today.ucsd.edu/story/say-hello-to-tritongpt

[19] 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). https://doi.org/10.1111/cogs.13048

[20] 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 arXiv:https://doi.org/10.1177/2053951720983865

[21] 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

[22] 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

[23] 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/

[24] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8 (2019). https://doi.org/10.3390/electronics8080832

[25] 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 (July 2023). https://doi.org/10.1186/s41239-023-00408-3

[26] Qingyu Chen, Jingcheng Du, Yan Hu, Vipina Kuttichi Keloth, Xueqing Peng, Kalpana Raja, Rui Zhang, Zhiyong Lu, and Hua Xu. 2023. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. arXiv:2305.16326 [cs.CL]

[27] B. Civil. 2023. ChatGPT can hinder students' critical thinking skills: Artificial intelligence is changing how students learn to write. *The Queen's Journal* (March 16 2023). https://www.queensjournal.ca/story/2023-03-16/opinions/chatgpt-can-hinder-students-critical-thinking-skills/

[28] 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

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

[30] 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

[31] 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)*. 1–5. https://doi.org/10.1109/ICTBIG59752.2023.10456172

[32] 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

[33] 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

[34] Teresa Datta and John P. Dickerson. 2023. Who's Thinking? A Push for Human-Centered Evaluation of LLMs using the XAI Playbook. arXiv:2303.06223 [cs.HC] https://arxiv.org/abs/2303.06223

[35] 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

[36] 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

[37] 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* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 626, 19 pages. https://doi.org/10.1145/3491102.3517441

[38] John Dewey. 2001. *Democracy and Education.* Pennsylvania State University Press.

[39] 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

[40] 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

[41] Upol Ehsan and Mark O. Riedl. 2021. Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. In *Proceedings of the NeurIPS Workshop on Human Centered AI*. https://arxiv.org/abs/2109.12480

[42] 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

[43] 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.

[44] 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

[45] 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

[46] 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

[47] 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

[48] 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

[49] Peter A. Facione. 2011. Critical Thinking: What It Is and Why It Counts. https://api.semanticscholar.org/CorpusID:154805251

[50] 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

[51] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[52] 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.

[53] 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

[54] 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.

[55] 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.

[56] Elizabeth Harkavy. 2022. *Accessible AI That's Out of This World: Globalizing AI Literacy through Problem-Based Learning and Deep Learning Models in a Low Code Environment*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[57] Office of the Provost Harvard University. [n. d.]. Guidelines for Using ChatGPT and other Generative AI tools at Harvard. https://provost.harvard.edu/guidelines-using-chatgpt-and-other-generative-ai-tools-harvard. Sincerely, Alan M. Garber, Meredith Weenick, Klara Jelinkova.

[58] 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

[59] Alex Hern. 2020. Twitter apologises for 'racist' image-cropping algorithm. *The Guardian* (Sept. 2020). https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm

[60] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. arXiv:1904.09751 [cs.CL]

[61] Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2022. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right. arXiv:2104.08315 [cs.CL]

[62] 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

[63] 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.

[64] 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.

[65] 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

[66] 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

[67] Anniek Jansen and Sara Colombo. 2022. Wizard of Errors: Introducing and Evaluating Machine Learning Errors in Wizard of Oz Studies. 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 426, 7 pages. https://doi.org/10.1145/3491101.3519684

[68] 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.

[69] 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

[70] 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

[71] 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)*. 1–9. https://doi.org/10.1109/FIE.2016.7757570

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

[73] 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

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

[75] Goda Klumbyte, Phillip Lücking, and Claude Draude. 2020. Reframing AX with Critical Design: The Potentials and Limits of Algorithmic Experience as a Critical Design Concept. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (Tallinn, Estonia) *(NordiCHI '20)*. Association for Computing Machinery, New York, NY, USA, Article 67, 12 pages. https://doi.org/10.1145/3419249.3420120

[76] 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

[77] 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

[78] 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

[79] 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 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.2016.14505301092

[80] 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

[81] 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

[82] 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

[83] Petro Liashchynskyi and Pavlo Liashchynskyi. 2019. Grid search, random search, genetic algorithm: a big comparison for NAS. *arXiv preprint arXiv:1912.06059* (2019).

[84] 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

[85] 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) *(CC '19)*. Association for Computing Machinery, New York, NY, USA, 271–284. https://doi.org/10.1145/3325480.3325504

[86] 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* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376727

[87] 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

[88] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 8086–8098. https://doi.org/10.18653/v1/2022.acl-long.556

[89] 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.

[90] 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.

[91] Charles F. Manski. 2020. The lure of incredible certitude. *Economics and Philosophy* 36, 2 (2020), 216–245. https://doi.org/10.1017/S0266267119000105

[92] 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.

[93] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 360, 18 pages. https://doi.org/10.1145/3411764.3445610

[94] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

[95] 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

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

[97] 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

[98] Fengchun Miao, Wayne Holmes, Ronghuai Huang, Hui Zhang, et al. 2021. *AI and education: A guidance for policymakers*. Unesco Publishing.

[99] 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

[100] 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

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

[102] Kenney Ng, Uri Kartoun, Harry Stavropoulos, John A. Zambrano, and Paul C. Tang. 2021. Personalized treatment options for chronic diseases using precision cohort analytics. *Scientific Reports* 11, 1 (Jan. 2021). https://doi.org/10.1038/s41598-021-80967-5

[103] 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.

[104] 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.

[105] Chris Shull (Chief Information Security Officer) and Gregory Hart (Chief Technology Officer). 2023. *WashU Addresses AI Technology*. Washington University of Medicine in St Louis, Office of Education. URL_of_the_article Rebecca.

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

[107] 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

[108] 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.

[109] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language Models as Knowledge Bases? arXiv:1909.01066 [cs.CL]

[110] Snehal Prabhudesai, Nicholas Chandler Wang, Vinayak Ahluwalia, Xun Huan, Jayapalli Rajiv Bapuraj, Nikola Banovic, and Arvind Rao. 2021. Stratification by Tumor Grade Groups in a Holistic Evaluation of Machine Learning for Brain Tumor Segmentation. *Frontiers in Neuroscience* 15 (2021). https://doi.org/10.3389/fnins.2021.740353

[111] 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*. 1–13.

[112] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling Language Models: Methods, Analysis Insights from Training Gopher. arXiv:2112.11446 [cs.CL]

[113] 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

[114] Juan David Rodríguez-García, Jesús Moreno-León, Marcos Román-González, and Gregorio Robles. 2021. Evaluation of an Online Intervention to Teach Artificial Intelligence with LearningML to 10-16-Year-Old Students. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (Virtual Event, USA) *(SIGCSE '21)*. Association for Computing Machinery, New York, NY, USA, 177–183. https://doi.org/10.1145/3408877.3432393

[115] Gilbert Ryle. 2009. *The Concept of Mind: 60th Anniversary Edition*. Routledge. https://doi.org/10.4324/9780203875858

[116] 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

[117] 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.

[118] Christian A. Schiller. 2024. The Human Factor in Detecting Errors of Large Language Models: A Systematic Literature Review and Future Research Directions. arXiv:2403.09743 [cs.CL] https://arxiv.org/abs/2403.09743

[119] M Seren Smith, Sarah Warnes, and Anne Vanhoestenberghe. 2018. Scenario-based learning. UCL IOE Press.

[120] W R Shadish, Thomas D Cook, and D T Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (2 ed.). Cengage Learning.

[121] Shreya Shankar, J. D. Zamfirescu-Pereira, Björn Hartmann, Aditya G. Parameswaran, and Ian Arawjo. 2024. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. arXiv:2404.12272 [cs.HC] https://arxiv.org/abs/2404.12272

[122] 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

[123] 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*. 384–387.

[124] Meredith Skeels, Bongshin Lee, Greg Smith, and George Robertson. 2008. Revealing Uncertainty for Information Visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Napoli, Italy) *(AVI '08)*. Association for Computing Machinery, New York, NY, USA, 376–379. https://doi.org/10.1145/1385569.1385637

[125] 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

[126] 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

[127] 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

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

[129] 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#:~:text=Data%20Privacy%20office.-,Review%20content%20before%20publication,that%20includes%20AI-generated%20material Accessed May 16, 2024.

[130] 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* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Article 1216, 5 pages. https://doi.org/10.1609/aaai.v33i01.33019795

[131] 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

[132] 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

[133] Priyan Vaithilingam, Tianyi Zhang, and Elena L. Glassman. 2022. Expectation vs.nbsp;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

[134] Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 Grasp Metaphors? Identifying Metaphor Mappings with Generative Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 1018–1032. https://aclanthology.org/2023.acl-long.58

[135] 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

[136] Tianjia Wang, Daniel Vargas Díaz, 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. https://doi.org/10.1109/vl-hcc57772.2023.00018

[137] Benjamin Weiser and Nate Schweber. 2023. The ChatGPT Lawyer Explains Himself. The New York Times. https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html [Accessed on 11 June 2023].

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

[139] Robert Wolfe and Alexis Hiniker. 2024. Expertise Fog on the GPT Store: Deceptive Design Patterns in User-Facing Generative AI. In *Proceedings of the Workshop Mobilizing Research and Regulatory Action on Dark Patterns and Deceptive Design Practices (DDPCHI 2024) co-located with the CHI Conference on Human Factors in Computing Systems (CHI 2024), Hybrid Event, Honolulu, HI, USA, May 11-16, 2024 (CEUR Workshop Proceedings, Vol. 3720)*, Colin M. Gray, Johanna Gunawan, René Schäfer, Nataliia Bielova, Lorena Sánchez Chamorro, Katie Seaborn, Thomas Mildner, and Hauke Sandhaus (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-3720/paper15.pdf

[140] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2019. Errudite: Scalable, Reproducible, and Testable Error Analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 747–763. https://doi.org/10.18653/v1/P19-1073

[141] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 6707–6723. https://doi.org/10.18653/v1/2021.acl-long.523

[142] 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.

[143] 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

[144] 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]

[145] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. arXiv:2102.09690 [cs.CL]

[146] Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why Does ChatGPT Fall Short in Providing Truthful Answers? arXiv:2304.10513 [cs.CL]

[147] 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

[148] 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

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