International Society of the Learning Sciences

Critical AI Literacy through Exploring Generative AI Limitations

Jaemarie Solyst, University of Washington, jaemarie@cs.washington.edu
Mandy Yiliu Pan, Carnegie Mellon University, yiliup@andrew.cmu.edu
Abigail Andam, Carnegie Mellon University, aandam@andrew.cmu.edu
Irene Portillo Poblete, Carnegie Mellon University, iportill@andrew.cmu.edu
Motahhare Eslami, Carnegie Mellon University, meslami@andrew.cmu.edu
Jessica Hammer, Carnegie Mellon University, hammerj@andrew.cmu.edu
Amy Ogan, Carnegie Mellon University, aeo@andrew.cmu.edu
Angela E.B. Stewart, University of Pittsburgh, angelas@pitt.edu

Abstract: Critical AI literacy enables understanding of the limitations of AI. In this work, we investigated how Black girls (N=11, ages 9-12) critically engaged with generative AI (genAI) through exploring ChatGPT's limitations. Learners used various approaches and leveraged their funds of knowledge (e.g., knowledge of pop culture) to investigate where genAI did not perform satisfactorily. We discuss how taking an asset-based approach can support critical AI literacy.

Introduction

In this study, we investigate critical AI literacy. Such AI literacy provides learners with knowledge beyond functional understandings, e.g., addressing ethical and social aspects of AI (Pinski & Benlian, 2024; Long & Magerko, 2020). Critical AI literacy could support youth in thoughtfully navigating everyday AI systems and avoid pitfalls like overtrust of AI's capabilities (Solyst et al., 2024). This is especially important with the rise of new generative AI (genAI) technologies. GenAI has had recent and drastic innovation, enabling generation of text, images, and media. Innovation in genAI has permeated youths' lives as well, as they gain access to technologies like ChatGPT, DALL-E, Gemini, and more. Youth are using genAI in various ways, from educational support to creative endeavors. However, youths' use of AI has raised concerns, as adult stakeholders such as teachers and parental guardians are wary about children's potential overreliance (e.g., on homework) (Blose, 2023) or exposure to harmful algorithmic bias (e.g., gender and race injustice embedded in AI-driven technologies) (Epps-Darling, 2020). Supportive educational efforts can help embolden youth in the age of AI.

Computing education has had growing emphasis on critical aspects of computing technologies (Ko et al., 2020, Morales-Navarro & Kafai, 2023). Critical computing encourages learners to examine computing technologies not only as technical tools but as socio-technical systems that reflect and reinforce societal values and biases. This approach aligns with a need for critical AI literacy, which includes an explicit focus on critique as an essential skill. One framework supports youth in understanding AI as a socio-technical system and provides a structured approach to critique by helping learners recognize the potential harms of AI, analyze the social dimensions of these harms, and deliberate on ways to create more responsible AI (Solyst et al., 2025). Furthermore, culturally responsive computing emphasizes engaging students in ways that respect and affirm their cultural identities, providing experiences that are relevant and empowering (Ashcraft et al., 2017; Scott et al., 2014). In this work, we focus on Black girls, whose cultural funds of knowledge have been systemically ignored or penalized in computing education. However, research has shown that culturally responsive approaches can disrupt this inequity by creating educational spaces where Black girls feel seen and valued, in turn allowing them to explore and shape technology in ways that resonate with their experiences (Erete et al., 2021).

With inspiration from these bodies of work, we ran an out-of-school computing workshop with Black girls (N = 11, aged 9-12). In this study, we asked: How do Black girls explore the limitations of generative AI? We analyzed data from the girls' investigations of ChatGPT's limitations in an activity designed to support fostering critical AI literacy. Through our thematic analyses, we found that they held varied perceptions of genAI and employed their funds of knowledge in prompting the AI.

Methods

Participants and recruitment

We worked with 11 Black girls ages 9-12 (avg = 10.5 years) during a two-week computing education summer camp. The girls were recruited through the community organization with whom we have had an ongoing multi-year partnership. The school reported that 99% Black youth enrolled, and most students qualified for free lunches. This study took place in a mid-sized city on the East Coast of the United States. Recruitment happened through working with the school staff, who distributed recruitment text and reached out to families. All youth who

expressed interest in joining and were in fourth to eighth grade could enroll in the camp. Of the youth who participated, all but two reported some degree of prior experience learning about programming or robotics. Learners were compensated $100 for attending any number of sessions in the camp.

Workshop content

Each day of the summer camp had three hours in total with a fifteen-minute break with snacks. Since the camp was multifaceted and included many activities about different aspects of computing, for the scope of this paper, we detail one activity, which took place on the very first day. This session focused on critical exploration via recognition of the limitations of genAI. We next detail the session in more detail.

Introduction to Generative AI: We first introduced the basic definition of AI with examples, including selfie-filters, voice recognition, and language translation apps. We then introduced the idea of genAI, including chat-based and text-to-image AI like OpenAI's ChatGPT (chatgpt.com) and DALL-E 3 (openai.com/index/dall-e-3/). We showed images of DALL-E image outputs and asked learners to guess what they thought the input was. We noted that all learners mentioned that they had experience using ChatGPT before.

Break ChatGPT: We then tasked learners to try and "break" ChatGPT by making it output something that the learners knew to be incorrect or something they disagreed with. With adult supervision, learners then individually tinkered with ChatGPT by inputting prompts that they thought could result in such outputs. At the beginning of the activity, a facilitating researcher shared some examples of types of questions to ask ChatGPT, including a difficult long multiple question (which ChatGPT got wrong) and a factual question, including how many bridges there were in the city where the study took place (which ChatGPT got right). Overall, this activity encouraged participants to engage critically with genAI by designing prompts that tested the limits of its capabilities. Learners then debriefed in a group conversation how they tried to prompt ChatGPT to break it.

Throughout the activities, the girls had control over the keyboard and mouse, although facilitators assisted them as needed. A researcher was present solely to facilitate conversations and answer questions.

Data capture and analysis

Throughout the camp, there were at least four researchers taking notes, in addition to one main facilitating researcher. To better protect participants' privacy, we did not have a running audio or video recording of the entire sessions, but researchers aimed to transcribe what was said in group conversations. Short conversations one-on-one between researchers and learners were also transcribed and sometimes recorded with consent. When learners used laptops, their screens were recorded to capture most of their activities. Additionally, log data from using genAI, as well as learners' digital and physical artifacts were saved for analysis. We used consensus-based (Hammer & Berland, 2014) thematic analysis (Clarke & Braune, 2017), continuously discussing as a team.

For the Break ChatGPT data specifically, two researchers independently reviewed the 145 prompts submitted by the girls and camp and coded them into six categories, e.g., Anthropomorphizing the AI, Cultural fluency assessment, and Fact checking. Two additional researchers did a further pass to finalize all the distinct categories in conversation with two senior researchers on the project. Two researchers then categorized the AI response types into three main broad categories based on whether ChatGPT was 'broken': (1) Satisfactory (addresses prompt fully), (2) Unsatisfactory (does not address prompt fully or very well), and (3) Biased (contains harmful bias). Satisfactory and Unsatisfactory were exclusive categories, while Biased was not. For example, Satisfactory and Biased responses could occur in the image generation where ChatGPT generated the images but were biased in some way (e.g., erasing Black representation). Satisfactory meant ChatGPT correctly responded to the learner, and the answer did not 'break' ChatGPT. Unsatisfactory referred to responses that were irrelevant, cut short, or contained misinformation, suggesting that it did break ChatGPT. Responses coded as Biased indicated that the output may have reinforced harmful biases, such as stereotypes. The researchers reached consensus on the categorization of both the prompts and ChatGPT's responses by regularly conversing with one another and working out any disagreements, supporting reliability in the analysis.

Limitations: Our approach had a few limitations. First, we could not capture all learner IDs to match with the data collection (e.g., quotes) at all times. Second, we could not always capture learners' interpretations of, reasonings about, or reactions to the AI, especially in the ChatGPT task. This means that our analyses focused on the conversation logs from ChatGPT for the Break ChatGPT task, and we did not have much data on learners' hypotheses and motivations for different prompts. Lastly, learners who signed up for the camp were self-selected.

Researcher Positionality: Our team of researchers all identified as women, with varying backgrounds. Our diverse backgrounds acted as lenses as we created the study protocol and analyzed the data, as we were in conversation with one another. Our academic backgrounds, too, brought various perspectives to the study design and analysis. These included human-computer interaction, culturally responsive teaching and computing, computer science, responsible AI, and games. While our team did not have anyone specifically from the

community we worked with, we had a community coordinator, who is from the neighborhood in which our community partner is housed, and who focused on maintaining our research relation with the school and advocating for their needs. Over the years, we have had the opportunity to learn with the community partner.

Findings

Shifting perceptions of generative AI

The learners' prior knowledge and perceptions of ChatGPT and AI varied. Many students were familiar with AI in their lives—their understanding included knowing that AI was embedded in general technology, although they were not familiar with underlying mechanisms. For example, during the classroom observation, almost all students raised their hands when asked if they had used Instagram, showing familiarity with popular apps but not necessarily understanding how AI powers these platforms. Additionally, many learners tended to associate AI with robots, e.g., P6 suggested that *"in some places, people use AI as waiters in restaurants."* When it came to genAI, although the majority of learners were not familiar with the term *"generative AI,"* they had prior experience using ChatGPT outside the camp, often for tasks such as homework, using it potentially in ways that may hinder learning. P6 mentioned, *"I type the math problem word for word, and it gives me the answer."*

Learners' initial beliefs about ChatGPT's capabilities evolved as they interacted with the system, highlighting shifts in understanding and confidence in the technology. At the start, most girls believed that ChatGPT could not be broken or make mistakes, one mentioning, *"I don't think you can break it."* Another learner suggested the reason behind her thought was *"maybe because actual people made it, and they're intelligent,"* implying trust in the developers and the AI they created. However, during the session where students were encouraged to 'break' ChatGPT, they began to realize its limitations. We observed that this activity impacted their perception by allowing them to see ChatGPT as fallible.

Exploring generative AI's limitations and capabilities

Learners engaged critically with ChatGPT, as they employed a variety of strategies to challenge the capabilities of the genAI. These approaches revealed their curiosities and critical reasoning with the tool. Their strategies fell into six main types: Fact checking, Philosophical reasoning evaluation, Cultural fluency assessment, Generating images with nuance, Noise injection, and Anthropomorphizing (Table 1). Of these six major categories, the most frequently employed prompt types were Fact checking (63 messages), Noise injection (27 messages), Cultural fluency assessment (21 messages), Anthropomorphizing (21 messages).

Table 1

Coding of Prompts when Learners Tried to "BreakGPT"

Prompt Type	Satisfactory	Unsatisfactory	Bias or Erasure	Prompt Ratio
Fact checking	52	11	0	63 (43.4%)
Philosophical reasoning	5	0	0	5 (3.4%)
Cultural fluency	9	12	4	21 (14.5%)
Anthropomorphizing	15	6	0	21 (14.5%)
Noise injection	23	4	0	27 (18.6%)
Generating images	7	1	5	8 (5.5%)
Response Type Ratio	111 (76.6%)	34 (23.4%)	9 (6.2%)	145

Fact checking: In some prompts, learners sought objective, verifiable information based on established facts. For example, seven learners tested ChatGPT's ability to handle factual information, e.g.: *"What's the biggest animal in the world?"* Potentially inspired by the example of long multiplication we gave, five learners challenged ChatGPT with basic math problems. For instance, P1 asked *"What is 12*6?"* Using their own creativity, some learners suggested false information. For example, P7 tried to trick ChatGPT by asking *"Is a dog heavier than an elephant?"* to test if ChatGPT would affirm false information.

Philosophical reasoning evaluation: Learners also asked philosophical questions without definitive answers. For example, P4 asked *"What is the meaning of life?"* and P5 asked *"Why do we fight?"* ChatGPT responded to P5*, "Fighting can happen for a lot of reasons—misunderstandings, conflicting interests, or even just a clash of personalities. Sometimes, it's just about communication and finding common ground ..."* going on to provide a generally comprehensive and satisfactory response.

Cultural fluency assessment: Learners also experimented with trendy cultural references and terms, asking ChatGPT questions like *"What does clock that tea mean?"* (P4), *"What is the fattest anime?"* (P10), or *"What does rizz mean?"* (P6) to assess its ability to understand trending language and pop culture.

Generating images with nuance: Although we did not introduce image generation at this stage, some learners were aware that ChatGPT could also generate images and asked the AI to create visual content, e.g., *"Make me a photo of a mix[ed] 90's couple,"* (P6). This suggests they were testing the AI's ability to capture details and questioning its capacity for nuanced racial and temporal information, in this case a couple from the 1990s that consisted of two people from different races. These prompts tested how ChatGPT handled artistic tasks.

Noise injection: Five learners also attempted to 'break' ChatGPT by spamming it with random text or emojis or nonsensical comments, e.g., *"he went he the milk."* We interpret this as potentially testing if this would confuse AI or cause it to respond unusually, revealing its limitations.

Anthropomorphizing: These prompts involved treating ChatGPT as if it were a human. Learners asked questions to investigate ChatGPT's limitations around human-specific prompts, e.g., P3 asked*, "Will you be my boyfriend?"* and P6 asked, *"Do you have a family?"* While ChatGPT responded by, *"I don't have a family or personal relationships, since I'm an AI. I'm here to assist and chat with you, though! If you have any questions or need help with something, just let me know."* Another approach was putting down ChatGPT (i.e., aiming to 'break' it the same way a human may be put to the limits through bullying), such as by calling the AI *"ugly."*

Based on learners' prompts, we categorized ChatGPT's responses into three categories: Satisfactory, Unsatisfactory, and Biased. Satisfactory responses occurred when ChatGPT provided accurate and relevant answers to learners' questions. For example, when a learner (P3) asked for the product of *"298,048,555*8,"* ChatGPT correctly returned *"2,384,388,440."* Similarly, when the learner (P2) asked *"What color are my eyes?"* ChatGPT responded, *"I can't see or determine physical attributes like eye color. If you'd like to share more about yourself or if there's anything else you'd like to discuss, feel free to let me know!"* Unsatisfactory responses occurred when ChatGPT's output was incomplete, irrelevant, or incorrect. For example, when ChatGPT would define trendy terms incorrectly. Or, in another instance, a learner submitted a string of emojis, which led ChatGPT to cut off its response after a few words. Biased responses arose when ChatGPT displayed bias like erasure (e.g., of Black representation) or reinforced stereotypes (Shelby et al., 2023). For instance, P5 prompted, *"Can I have a pic of a boy with dreads,"* and ChatGPT generated an image of a White boy with dreads, despite dreads being a commonly Black hairstyle (Kuumba & Ajanaku, 1998). P6 prompted ChatGPT to generate *"a guy with a dress on who has a pet werewolf,"* but ChatGPT returned an image of a man in a suit (not a dress) with a werewolf, demonstrating its limitations and bias toward gender norms.

Prompts related to trendy pop culture and image generation sometimes generated bias through stereotypes and erasure in the responses. Trendy pop culture and anthropomorphizing prompts most often produced unsatisfactory responses that were either incorrect or irrelevant to the learners' prompt. While we noted these limitations rather than the learners themselves, these instances showcase the kinds of issues we can encourage learners to notice, question, and reflect on when they engage with genAI.

Discussion and conclusion

We found that the learners were familiar with genAI before coming into the camp and sometimes were regular users. When it came to their understanding of the limitations of ChatGPT, their approaches to "breaking" it suggest various believed limitations, but one of the most successful approaches was prompting it with assessing its cultural fluency through trending terms and pop culture references. Through critiquing AI in investigating the limits of ChatGPT, we saw that learners were able to tap into their knowledge and cultural assets. For example, the learners' use of terms rooted in Black culture revealed how ChatGPT failed to accurately represent the origins and meaning of terms. Instead ChatGPT produced outputs that erased the language of its cultural significance or misattributed its origins (Shelby et al., 2023). Ultimately, we saw that Black girls were active agents in their interactions with AI, as their actions went beyond passive interaction as they explored the limitations.

Prior work has highlighted the importance of computing education in providing access to career pathways (Scott et al., 2014), and equipping learners with technical skills (Lee et al., 2021, Zhang et al., 2023). However, many AI literacy approaches have often overlooked the role of identity, self-expression, and critical engagement in empowering underrepresented learners in technology-focused spaces. This gap leaves room for educational experiences that foster learners' agency. Emphasis on criticality supported the girls in realizing the limitations of AI, further highlighting problematic bias (e.g., by showing AI bias through stereotypes or erasure). Further iterations of this work could also investigate how Black girls create new AI systems are more culturally responsive rather than focusing on engaging with existing systems as was the case in our work. These shifts are essential to empower Black girls to see themselves as creators and visionaries, with critique as a starting point.

This work contributes new understandings of critical AI literacy by emphasizing actively exploring the limitations of genAI systems as a way of asset-based learning. We observed how Black girls used their funds of knowledge to successfully investigate the shortcomings of genAI.

References

Ashcraft, C., Eger, E. K., & Scott, K. A. (2017). Becoming technosocial change agents: Intersectionality and culturally responsive pedagogies as vital resources for increasing girls' participation in computing. Anthropology & Education Quarterly, 48(3), 233-251.

Blose, A. (2023). As CHATGPT enters the classroom, teachers weigh pros and cons. NEA.

Clarke, V., & Braun, V. (2016). Thematic analysis. The Journal of Positive Psychology, 12(3), 297–298.

Epps-Darling, A. (2020). How the racism baked into technology hurts teens. The Atlantic, 24.

Erete, S., Thomas, K., Nacu, D., Dickinson, J., Thompson, N., & Pinkard, N. (2021). Applying a transformative justice approach to encourage the participation of Black and Latina girls in computing. ACM Transactions on Computing Education, 21(4), Article 27, 1–24.

Hammer, D., & Berland, L. K. (2014). Confusing claims for data: A critique of common practices for presenting qualitative research on learning. Journal of the Learning Sciences, 23(1), 37–46.

Kuumba, M., & Ajanaku, F. (1998). Dreadlocks: The hair aesthetics of cultural resistance and collective identity formation. Mobilization: An International Quarterly, 3(2), 227–243.

Ko, A. J., Oleson, A., Ryan, N., Register, Y., Xie, B., Tari, M., Davidson, M., Druga, S., & Loksa, D. (2020). It is time for more critical CS education. Communications of the ACM, 63(11), 31–33.

Lee, I., Ali, S., Zhang, H., DiPaola, D., & Breazeal, C. (2021). Developing middle school students' AI literacy. Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE '21), 191–197.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. Proceedings of the CHI Conference on Human Factors in Computing Systems, 1–16.

Morales-Navarro, L., & Kafai, Y. B. (2023). Conceptualizing approaches to critical computing education: Inquiry, design, and reimagination. In Apiola, M., López-Pernas, S., & Saqr, M. (Eds.), Past, present and future of computing education research. Springer, Cham.

Pinski, M., & Benlian, A. (2024). AI literacy for users: A comprehensive review and future research directions of learning methods, components, and effects. Computers in Human Behavior: Artificial Humans.

Scott, K. A., Sheridan, K. M., & Clark, K. (2014). Culturally responsive computing: A theory revisited. Learning, Media and Technology, 40(4), 412–436.

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23) (pp. 723–741).

Solyst, J., Amspoker, E., Yang, E., Eslami, M., Hammer, J., & Ogan, A. (2025). RAD: A framework to support youth in critiquing AI. Proceedings of the SIGCSE Technical Symposium.

Solyst, J., Yang, E., Xie, S., Hammer, J., Ogan, A., & Eslami, M. (2024). Children's overtrust and shifting perspectives of generative AI. In Lindgren, R., Asino, T. I., Kyza, E. A., Looi, C. K., Keifert, D. T., & Suárez, E. (Eds.), Proceedings of the 18th International Conference of the Learning Sciences - ICLS 2024, 905–912.

Zhang, H., Lee, I., Ali, S., DiPaola, D., Cheng, Y., & Breazeal, C. (2023). Integrating ethics and career futures with technical learning to promote AI literacy for middle school students: An exploratory study. International Journal of Artificial Intelligence in Education, 33(2), 290–324.

Acknowledgments

We thank the school we worked with, community leaders and staff, and youth participants that made this study possible. We also thank additional members of our research team, including Chun Li, Paras Sharma, and Erin Walker for supporting data collection and feedback. This work is supported by the Jacobs Foundation's CERES Network and the NSF DRL-1811086.