Do You Think GPT Will Be Correct?: Measuring and Improving Generative AI Literacy Through Metacognition Training

Emily Su
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada

Jessica Bo
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada

Siyi Wu
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada

Ashton Anderson
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada

Beth Coleman
Institute of Communication, Culture,
Information Technology
Faculty of Information
University of Toronto
Toronto, Ontario, Canada

ABSTRACT

With the widespread use of AI chatbots like ChatGPT amongst students, it raises the concern about how they are using these tools in their studies. In our study, we focused on increasing AI literacy among students, which refers to the ability to not only understand how AI works and its limitations, but also how to evaluate its output and when to use it appropriately. We created a quiz with metacognitive activities and compared it with a traditionally formatted AI literacy module as the control. We found that with the metacognitive quiz, students felt more confident in writing effective prompts and knew how to correctly self-rely on themselves in situations where an AI chatbot gave them an incorrect answer. Students given the metacognitive quiz were able to write prompts with more context overall and, on average, obtain a response from GPT-4o that was closer to the ground truth in a prompting task compared to the control group. However, participants given the intervention felt less confident in verifying the outputs of AI chatbots and when to use them. Future studies can look into developing metacognition-based AI literacy material for other groups like seniors.

1 INTRODUCTION

As large language models (LLMs) become increasingly used by students in their studies since the emergence of LLM-based chatbots like ChatGPT, concerns are raised about how effectively and appropriately they utilize these tools. With this, AI literacy has become a critical skill to develop. AI literacy is the ability to not only understand how AI works and AI's capabilities and limitations, but also know how to evaluate its output and when to use it appropriately [3]. With a lack of understanding of how GenAI tools works, the uncritical use of these tools can lead to not only inappropriate use with consequences such as academic offences for students but also fears about skills becoming irrelevant, such as coding.

1.1 Motivation

Currently there are existing materials such as the University of Toronto Centre for Teaching Support & Innovation's GenAI Literacy Course Modules (OER) to support AI literacy training [6]. However, these materials are often static and lack activities that train people's metacognition, which refers to the ability to monitor and evaluate their own actions and thinking [6, 13]. Tankelevitch et al (2024) raise the concern about how the metacognitive demands of GenAI systems needed from users are increasing, and current work in human-AI interaction often do not concern itself with solving this issue [14]. They suggested that one way to counteract the metacognitive demands of GenAI systems was to improve user's metacognition of GenAI systems, which includes having supports for "planning, self-evaluation, and self-management" [14]. Having a developed metacognition in the context of AI literacy does not only allow students to be more effective in terms how they are using GenAI tools for their own learning, but also it empowers students to make informed decisions on when to appropriately use these tools [12].

Our research aims to fill in this gap with the lack of AI literacy materials grounded in developing one's metacognition about GenAI systems. We established an intervention, a quiz containing interactive metacognitive activities, to improve AI literacy in students. The activities in our intervention had the following goals in mind: 1) bring awareness of one's goals, context, and available resources, 2) calibrate one's expectation of GenAI's abilities and limitations, 3) learn when to use GenAI appropriately based on context and situations.

In our study, we sought to answer the following questions:

- **RQ1**: How does training metacognition for Generative AI impact user's prompting and verification skills?
- **RQ2**: How do the interventions affect people's perceptions and confidence with using AI chatbots?

2 RELATED WORK

Previous works in AI literacy have looked into developing materials to increase AI literacy across different groups. Cao et al (2025) developed a series of short videos AI literacy videos for adults to increase AI understanding, use, and evaluation [4]. During their intervention, they gave participants a short multiple-choice quiz after each video that measures how much their knowledge of AI changes after each video [4]. They found that their intervention increased participants' self-efficacy rating of knowing when to use AI but not their understanding of how AI works or how to evaluate AI technologies [4]. Taking a more gamification approach, Ma et al (2025) developed an educational game to teach adults how GenAI

tools work [10]. The goal of their game was to help adults understand and be more aware of bias in GenAI tools, the capabilities and limitations of GenAI tools, and how to write effective prompts [10]. Ko et al (2025) created an AI literacy education chatbot name Litti for seniors (age 65 and up) focused on improving AI knowledge, improving AI detection, and understanding AI ethics, which are based on the Meta AI Literacy Scale (MAILS) framework [9]. They specifically looked at these three metrics due to their relevance to seniors and their lives [9]. They measured changes to the three metrics with participants before and after interacting with the chatbot [9]. Ko et al (2025) did find an increase in AI literacy scores across these three metrics [9]. However, Ko et al (2025)'s study lacked a control group similar to Ma et al (2025) [9, 10], which makes it difficult to evaluate if there were confounding variables in the study or if their results were by chance.

Our study takes inspiration from Ma et al (2025)'s study in terms of its gamificiation aspect as well as Carlini (2023)'s "A GPT-4 Capability Forecasting Challenge" [5, 10]. In this challenge, participants forecast their predictions about the correctness of GPT-4 when asked different questions [5]. Our work adapted this idea of forecasting from Carlini [5] for the context of AI literacy in our quiz design. With this, our intervention focuses on improving functional AI literacy based on Becker et al (2024)'s AI literacy framework. More specifically, we are focused on increasing students' awareness of the affordances and limitations of GenAI tools and how to effective prompt GenAI tools [1].

3 METHODOLOGY

We had a total of $n = 44$ students ages 18 and above participating in our randomized online study through Qualtrics. These students were recruited through various means such as social media, classroom announcements, and word of mouth. However, we only analyzed our results using 21 participants ($n = 11$ females, $n = 10$ males) after filtering out participants based on an attention check question. This study is still ongoing, and through power analysis we conducted, we are planning to obtain approximately 400 participants (200 people per condition) in order to achieve a statistical power of 0.8 with a small effect size of 0.25. We currently have $n = 11$ participants in the intervention condition and $n = 10$ participants in the OER module condition.

3.1 Study Design

The study was divided into three sections: pre-intervention, intervention, and evaluation. During the pre-intervention stage, participants were given 4 statements on their confidence and perceptions of using AI chatbots ("AI chatbots are reliable", "I feel confident about choosing when to use AI chatbots", "I feel confident about writing effective prompts", "I can confidently verify the outputs of AI chatbots") on Likert scale where 1 is strong disagree and 7 is strongly agree with 4 being neutral. These statements were created based on Jian et al (2000)'s "Trust in Automation" scale [8].

During the intervention stage, participants were randomized to be either in the megacognitive intervention condition or the OER module condition. The megacognitive intervention group worked on two activities. In the first activity, they are asked to evaluate 2 prompts for a given student context, with one being a good prompt for the given scenario and the other being a not optimal prompt. In the next activity, they are asked to verify the corresponding GPT output for each of the prompt. In both activities, they were asked to make evaluations by rating statements on a scale from 1 (highly disagree) to 7 (highly agree). The statements are as follows, given the student's name in the scenario is Sarah: "The AI chatbot's response to her prompt will be accurate", "The AI chatbot's response will help Sarah achieve her goal", "Sarah's prompt demonstrates that she is appropriately engaged in achieving her goal", and "Sarah is using the AI chatbot appropriately". After each of the evaluations, participants are given an explanation of what makes the prompts or GPT response different. The OER group went through reading material from the OER modules that goes over how one can write effective prompts and evaluate the results from GPT with small quizzes of what they read.

During the evaluation stage, we had a task to evaluate their prompting skills, which ask participants to write a prompt for ChatGPT to guess the number of chickpeas in a jar in two images showing the bottom and side of the jar. The ground truth for the number of chickpeas is 403. We also had a task to evaluate their verification skills where participants are given a question from the LSAT and asked about their confidence in their own answer. Afterwards, we gave them ChatGPT's response to the question (either the wrong or right answer) and asked them again what they think the right answer is and asking them their confidence in their answer. Lastly, we asked them the same statements we asked them before the intervention to observe any changes in their confidence and perception in using AI chatbots.

3.2 Methods for Analysis

When conducting semantic analysis on the prompts given by participants for the open-ended task of the study, we used a combination of n-grams, more specifically trigrams and the Term Frequency - Inverse Document Frequency (TF-IDF) to measure for any important sequences of words across all prompts and look for patterns.

TF-IDF is a statistic for computing the significance of a word or a sequence of words relative to the word corpus it comes from, which in our case is our prompts. It is given by the function $f(t, d)$ [7]:

$$f(t, d) = tf(t, d) \times idf(t)$$

$tf(t, d)$ represents the number of times term t shows up in a prompt d while $idf(t)$ is the inverse document frequency that represents the importance of a word where the less it shows up in other prompts, the more rare or important the term is [7]. The $idf(t)$ is defined as follows:

$$idf(t) = log(\frac{N}{1 + df(t)})$$

where N is the number of prompts and $df(t)$ is the number of prompts that has the term t [7].

We first tokenized the prompts and filtered out words that only show up in only 1 prompt and English stop words using a TF-IDF vectorizer. We also reduced the number of dimensions in our data by only taking at most the top 10 features. This vectorizer would then compute scores for the significance of trigrams of words from the prompts.

For the open-ended prompt, we used GPT-4o through OpenAI's API [11] to evaluate the quality of the prompts across the two different groups. The dimensions we looked at were "relevance" (is the prompt relevant to the task and does the prompt reference the goal?), "quality" (is the prompt free of spelling and grammar mistakes? Does the prompt include in-context examples or use prompting techniques like chain-of-thought), and "coherence" (are the ideas in the prompt connected?). We ran GPT-4o 4 times on the prompts. For the prompting challenge, we also ran GPT-4o across all the prompts and extracted the final number of chickpeas given by GPT-4o using GPT-4o-mini.

To analyze participants reliance on ChatGPT's response for the LSAT question, we used Bo et al (2024)'s appropriate reliance metrics [2]. We used the methodology from the paper in order to calculate 4 reliance outcome: "Appropriate Reliance on LLM", "Appropriate Self-Reliance", "Under-Reliance", and "Over-Reliance". "Appropriate Reliance on LLM" is defined as the scenario when the LLM's response is correct and the participant switches their answer from an incorrect answer or they become more confident in their own correct answer [2]. "Appropriate Self-Reliance" is the outcome where a participant sticks with their original answer (it can be incorrect or correct) when LLM provides an incorrect response [2]. With "Under-Reliance", it's when a participant sticks with their incorrect answer despite the LLM's response being correct and with "Over-Reliance, it is when a participant switches their answer (whether it is correct or not) to the LLM's response when it's incorrect [2].

4 RESULTS

4.1 Prompting Skills

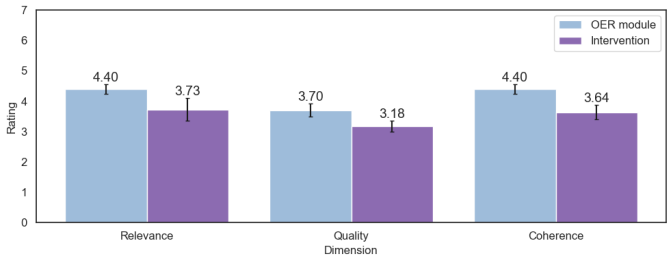

Figure 1: Mean ratings (1 to 7) given by GPT-4o for prompts across different conditions (OER module, Intervention) and dimensions (Relevance, Quality, and Coherence)

Figure 1 shows that GPT-4o rated the prompts written by the OER module group higher on average than the prompts written by the intervention group. For relevance, we found that the OER module group had an average rating of 4.40 while for the intervention group they had an average rating of 3.73. In terms of quality, the intervention group had an average rating of 3.18 while with the OER module group it was 3.70. In terms of the coherence of their prompts, the OER module group and the intervention group had average ratings of 4.40 and 3.64, respectively.

It appears that both groups overall received an average rating or below average rating across the dimensions during the evaluation stage. Since the data is not normally distributed, we conducted a Mann-Whitney U test to look at the significance of our results and found that while the difference between the group for the dimensions, relevance and quality, have p-values of $p = 0.228$ and $p = 0.099$, respectively, the difference for the dimension, coherence, was statistically significant where $p = 0.015$, however the statistical power of the result is still pending. The results we see might have been due to outliers in our current dataset and our limited sample size.

Figure 2: Trigram of important words used by the intervention group

Figure 2 and 3 shows the highest TF-IDF scoring sequence of words used by the intervention and OER module group while filtering out sequences that only show up in 1 prompt. The sub-groups represent a prompt written by a person from that group and the prompt that are shown are the ones that met our filters indicated in section 3.2.

With figure 2, it shows that the 6 trigrams with the highest TF-IDF scores for the intervention group were ("estimate chickpeas jar"), ("estimate number chickpeas"), ("estimate total number"), ("image view jar"), ("number chickpeas jar"), and ("total number chickpeas"). Figure 3 shows the 5 trigrams with the highest TF-IDF scores for the OER module group, which were ("chickpea picture shows"), ("estimate number chickpeas"), ("glass jar chickpeas"), ("number chickpeas jar"), and ("total number chickpeas"). Both groups overall used similar vocabulary when prompting GPT to estimate the number of chickpeas in the jars. However, with the intervention group, some prompts described the view of the jar that the two images were showing. This indicates that the intervention group were more specific overall in terms of helping GPT learn about the images.

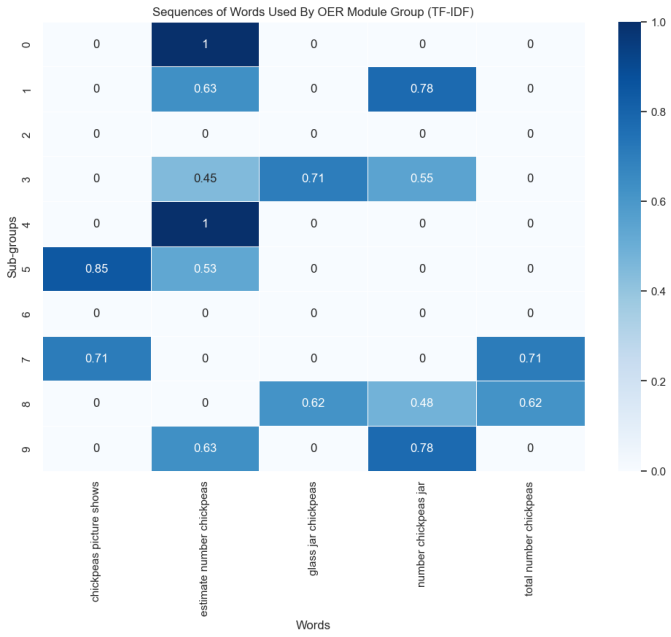

Figure 3: Trigram of important words used by the OER module group

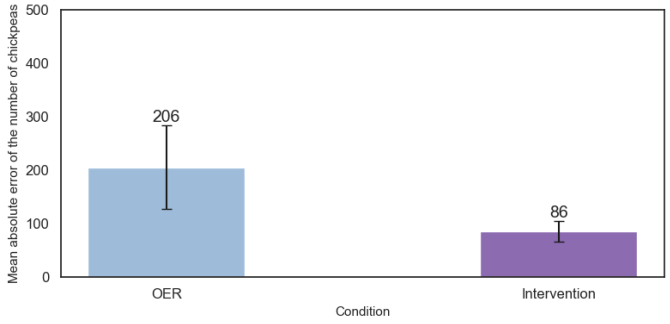

Figure 4: Mean absolute error of the number of chickpeas obtained from GPT-4o from open-ended prompts across conditions

When looking at the GPT-4o output of all the prompts given by the participants in figure 4, we found that the intervention condition had a lower mean absolute error of 86 from the ground truth of 403 chickpeas than the OER condition, which was 206. When we conducted a one-sided t-test, we found that the p-value was $p = 0.053$. This means that on average, the intervention condition received answers from GPT-4o that was closer to the ground truth than the OER condition.

4.2 Verification Skills

Table 1 shows that there was an increase in Appropriate Self-Reliance for the intervention condition after being shown ChatGPT's response to the LSAT question. This means that participants

Table 1: Reliance on LLMs across conditions

Outcome	OER Modules	Intervention
Appropriate Reliance on LLM	2	2
Appropriate Self-Reliance	3	5
Under-Reliance	1	1
Over-Reliance	1	1

correctly relied on themselves when ChatGPT provided wrong advice. The overall appropriate reliance ratio [2] for the intervention is therefore higher than the OER Modules, but significance is pending due to the limited sample size.

4.3 Confidence and Perception on using AI chatbots

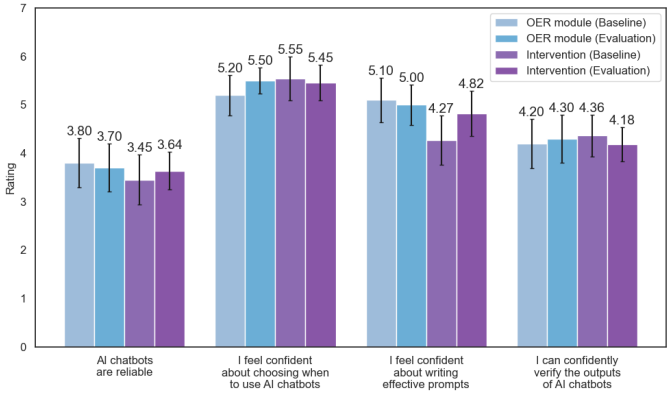

Figure 5: Mean ratings (1 to 7) given by participants on confidence and perception of using AI chatbots before (baseline) and after intervention (evaluation)

We observed in figure 5 an increasing trend in the intervention group in their rating regarding their view of AI chatbots being reliable and confidence in writing effective prompts. However, we saw a decrease in confidence for the intervention group in regards to verifying the outputs of AI chatbots and when they can choose to use AI chatbots. When we conducted a Mann-Whitney U test across all statements, comparing the change in rating between the two conditions, we found that the results were not statistically significant with the statement "AI chatbots are reliable", having a p-value of $p = 0.299$, the statement "I feel confident about choosing when to use AI chatbots" having a p-value of $p = 0.187$, the statement "I feel confident about writing effective prompts" having a p-value of $p = 0.443$, and the statement with a p-value of $p = 0.202$.

4.4 Evaluation of Learning Materials

Figure 6 showed that the intervention condition found the material to be not too long compared to the OER module condition. However, people in the intervention condition were slightly, on average, less likely to remember what they learned about AI chatbots in the future compared to the OER module condition. The intervention

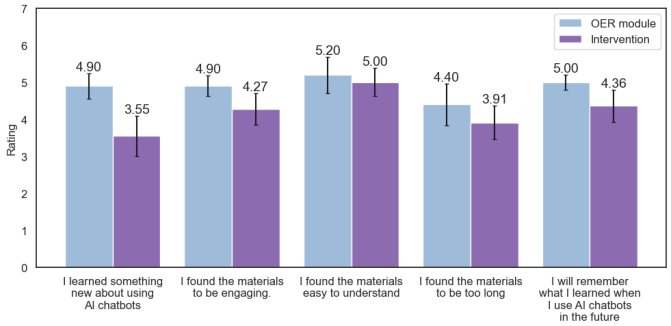

Figure 6: Mean ratings (1 to 7) given by participants evaluating the learning material

condition also on average are less likely to feel they learned something new about using AI chatbots than the OER module, which we found to be the only statement that is close to significant at a p-value of $p = 0.054$.

5 DISCUSSION

Our study is currently ongoing, and so results from the study are pending in significance and power.

We found that both the metacognitive intervention and OER modules had strengths that the other fell short with. We observed that students who performed metacognitive activities tend to correctly rely on themselves when an AI chatbot provided wrong advice and they felt more confident in writing effective prompts. This indicates that incorporating metacognitive activities could be beneficial for increasing AI literacy. However, the metacognitive activities did not help students feel more confident with verifying outputs from AI chatbots and their confidence actually decrease after the intervention. On the other hand, we saw that with the OER module condition, participants felt more confident with choosing when to use AI chatbots and verify their outputs. We also saw that the intervention condition gave a lower rating in terms of having learned something new about using AI chatbots from the intervention material. This could be because the intervention material did not include advice on applying the explanation to other fields, which the OER module touched upon. These findings indicate AI literacy can be improved with a combination of interactive metacognitive activities and static content teaching effective prompt writing and verification skills initially.

However, we found with the prompting task during the evaluation stage, everyone did average or below average in terms of prompt quality when evaluated by GPT-4o. We believe it was due to the fact that for both conditions, it lacked a practical component where participants can practice writing prompts and receive feedback for them. For next steps, we hope to include this to our intervention. We also found that the intervention group got closer to the ground truth than the OER group, and we observed that the prompts in the intervention condition tend to specify the view of the image, which could have impacted the accuracy of how GPT-4o responded to their prompts. Future works can look into how metacognition-based AI literacy activities can be applied to not only students but also other groups like seniors.

6 APPENDIX

The GitHub repository containing the scripts to run the GPT models for the various prompts in the study, running the prompting challenge, and using an LLM as a judge can be found here: https://github.com/moonsdust/llm-automation-prompt-eval

REFERENCES

[1] Kimberly P Becker, Jessica L Parker, and Desi Richter. 2024. Framework for the Future. (2024). https://moxielearn.ai/wp-content/uploads/2024/06/Ai-literacies-white-paper.docx.pdf

[2] Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To rely or not to rely? evaluating interventions for appropriate reliance on large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.

[3] Huajie Jay Cao, Hee Rin Lee, and Wei Peng. 2025. Empowering Adults with AI Literacy: Using Short Videos to Transform Understanding and Harness Fear for Critical Thinking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 203, 8 pages. doi:10.1145/3706598.3713254

[4] Huajie Jay Cao, Hee Rin Lee, and Wei Peng. 2025. Empowering Adults with AI Literacy: Using Short Videos to Transform Understanding and Harness Fear for Critical Thinking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–8.

[5] Nicholas Carlini. 2023. A GPT-4 Capability Forecasting Challenge. (2023). https://nicholas.carlini.com/writing/llm-forecast/question/Capital-of-Paris

[6] University of Toronto Centre for Teaching Support Innovation. 2025. Teaching with Generative AI at U of T. *University of Toronto* (2025). https://teaching.utoronto.ca/teaching-uoft-genai/

[7] GeeksforGeeks. [n. d.]. How to store a TfidfVectorizer for future use in scikit-learn? /. *GeeksforGeeks* ([n. d.]). https://www.geeksforgeeks.org/nlp/how-to-store-a-tfidfvectorizer-for-future-use-in-scikit-learn/

[8] Jiun-Yin Jian, Ann Bisantz, and Colin Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4 (03 2000), 53–71. doi:10.1207/S15327566IJCE0401_04

[9] Eunhye Grace Ko, Shaini Nanayakkara, and Earl W Huff Jr. 2025. " We need to avail ourselves of [GenAI] to enhance knowledge distribution": Empowering Older Adults through GenAI Literacy. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[10] Qianou Ma, Anika Jain, Jini Kim, Megan Chai, and Geoff Kaufman. 2025. ImaginAItion: Promoting Generative AI Literacy Through Game-Based Learning. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–9.

[11] OpenAI. [n. d.]. API Platform. *OpenAI* ([n. d.]). https://openai.com/api/

[12] Julie Dangremond Stanton, Amanda J Sebesta, and John Dunlosky. 2021. Fostering metacognition to support student learning and performance. *CBE—Life Sciences Education* 20, 2 (2021), fe3.

[13] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 680, 24 pages. doi:10.1145/3613904.3642902

[14] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The metacognitive demands and opportunities of generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–24.