Do You Think GPT Will Be Correct?

Measuring and Improving Generative AI Literacy Through Metacognition Training

Emily Su, Jessica Bo, Siyi Wu, Ashton Anderson, and Beth Coleman

Computational Social Science Lab, Schwartz Reisman Institute

Introduction

- Increasing concerns about how students are using AI chatbots in their studies.

> **AI literacy**
>
> 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 [1].

- With a lack of understanding of how Generative AI (GenAI) tools works, the uncritical use of these tools can lead to not only inappropriate use of them but also fears about what they are learning in school being irrelevant.

Motivation

- Current AI literacy materials for students are static and often lack interactive activities to train student's metacognition.

> **Metacognition**
>
> The ability to monitor and evaluate one's own actions and thinking [2].

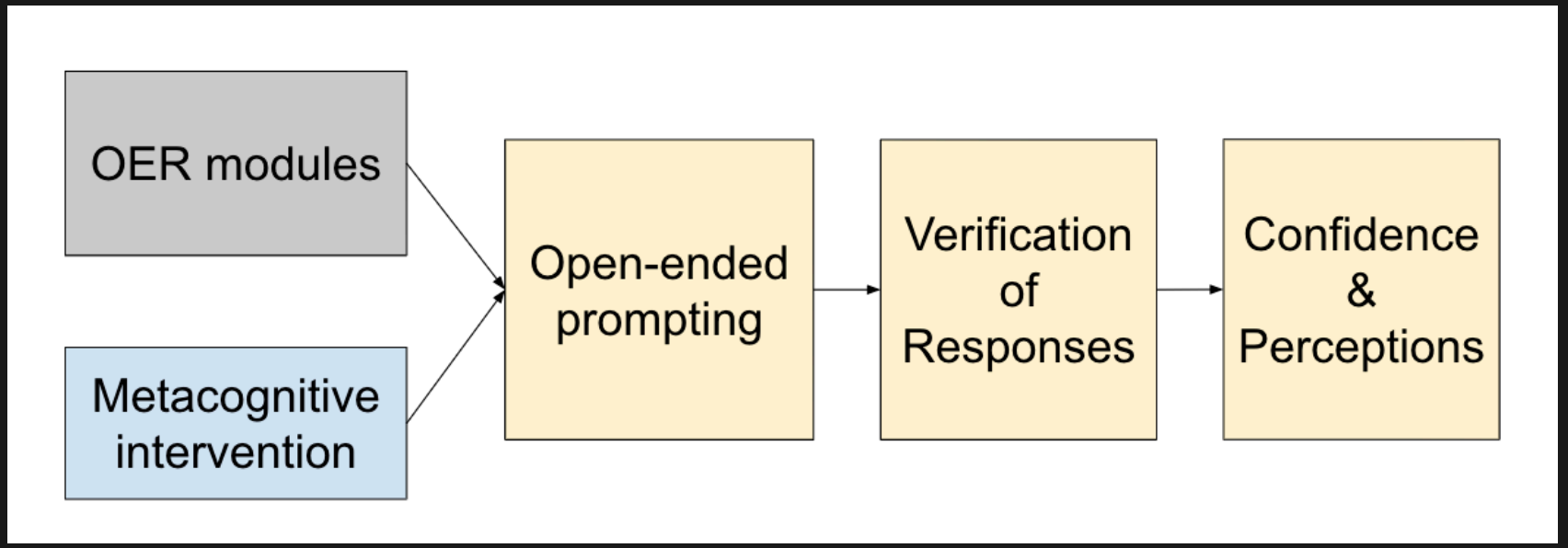
- Having a developed metacognition in the context of AI literacy -> students become more effective in terms of how they use GenAI tools for their own learning & empowers students to make informed decisions on when to appropriately use these tools [3]

Our Intervention

- Quiz containing interactive metacognitive activities, to improve GenAI literacy in students. The activities 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

Methods



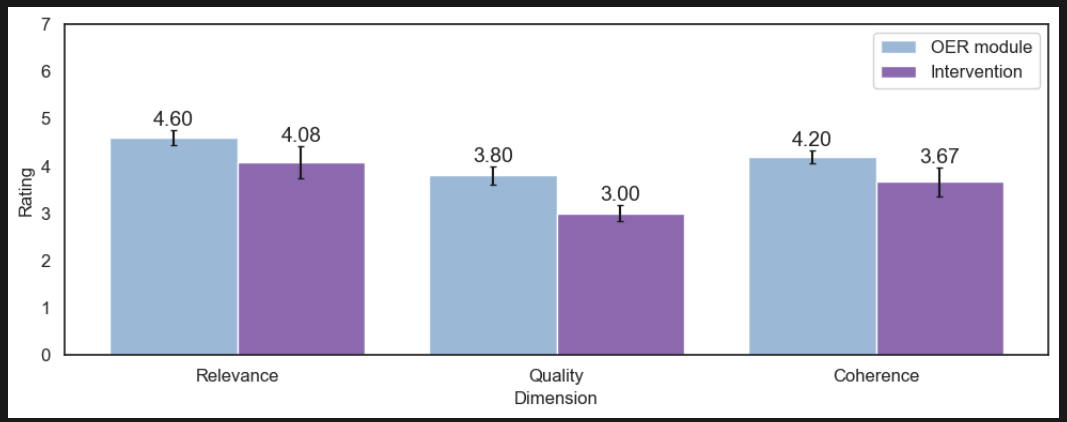
- Study ran on Qualtrics
- Between-subjects: The University of Toronto's GenAI Literacy Course Modules (OER) & Metacognitive intervention
- Participants: 22 participants (12 in intervention & 10 in OER)
- We hope to obtain ~400 participants (200 per condition) in order to achieve a statistical power of 0.8 with a small effect size of 0.25 with a t-test.

Effects on Prompting and Verification Skills

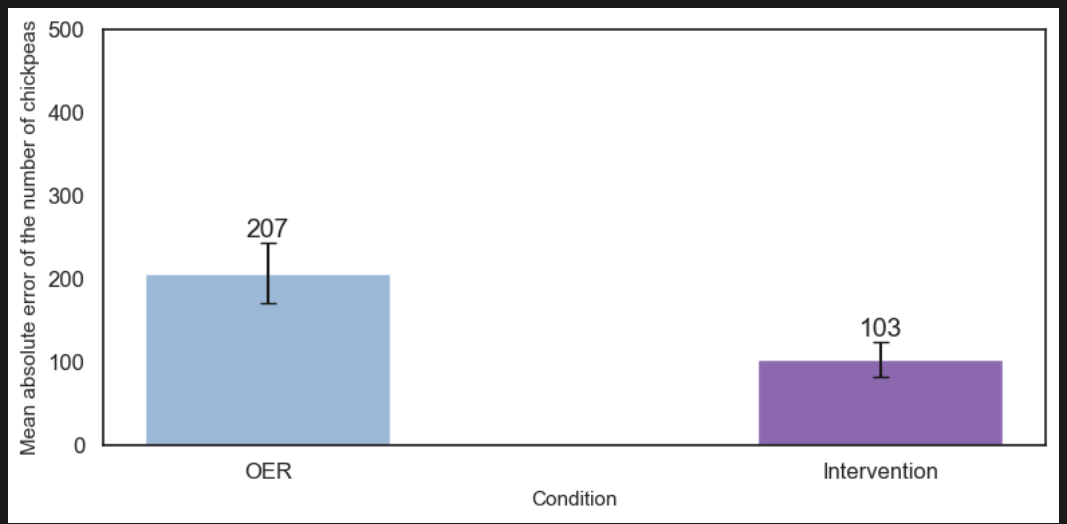
How does training metacognition for Generative AI impact user's prompting and verification skills?

Open-Ended Prompting Results

Prompt Quality Rating using LLM as a judge (GPT-4o)



Mean absolute error of the number of chickpeas obtained from GPT-4o



Open-Ended Prompting Results (Part 2)

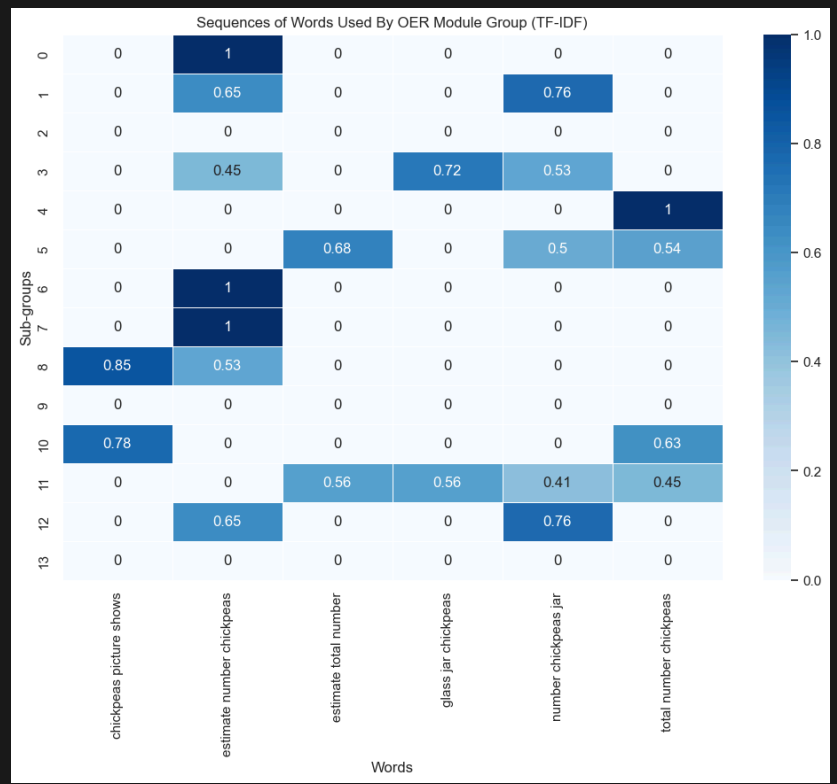
- We saw the intervention condition **received poorer ratings** across all prompt dimensions by GPT-4o ($p = 0.287$ (relevance), $p = 0.010$ (quality), $p = 0.1920$ (coherence)). However, the prompts written by the intervention condition was **closer to the ground truth/true value** in the open-ended prompting task (ex: Get GPT to estimate the number of chickpeas in a jar from 2 photos) ($p = 0.008$). Why could that be the case?

Preliminary qualitative analysis

- Saw some outlier prompts for intervention -> Lower rating across dimensions by GPT-4o
- Intervention condition **didn't use prompting techniques as often (ex: chain-of-thought)** compared to the OER condition -> Lower rating across dimensions by GPT-4o.
- Intervention condition **tend to assist with calculations, express their thought process, and specify the view** of the chickpea jar the images are showing -> Hypothesis of why they were closer to ground truth

Open-Ended Prompting Results (Part 3)

- Linguistic analysis shows that both conditions used **similar vocabulary**, however the intervention conditions **specified the views of the images**.



Prompts by Intervention Condition

Prompt 1

“i am providing 2 images of a cylindrical jar of beans, your goal is to estimate the total number of beans in the jar. you may assume the beans are uniformly spherical. image 1 is the circular bottom of the jar, while image 2 is a side view of the jar i am considering two approaches, take the average of the results… approach 1: using image 1, estimate the diameter of the jar using beans as the unit of measurement… simply multiply the count per layer by the number of layers since both approaches rely heavily on careful accounting of beans in both photos, ensure this step has absolutely no mistakes”

Prompt 2

“Predict how many chickpeas are in this circular container. It seems around 8 chickpeas can fit in its diameter, and around 8 chickpeas in its height.”

Prompt 3

“estimate the number of chickpeas in the pictures. here you see the view from the bottom and from the side you may count how many are along each side (radius and height) and use the formula for the volume of a cylinder to help get a closer estimate. if you know any better way to get the estimate of items in a jar than this, feel free to do that.”

Verification Skills

Reliance on LLMs across conditions

Using Bo et al (2024)'s appropriate reliance metrics [4].

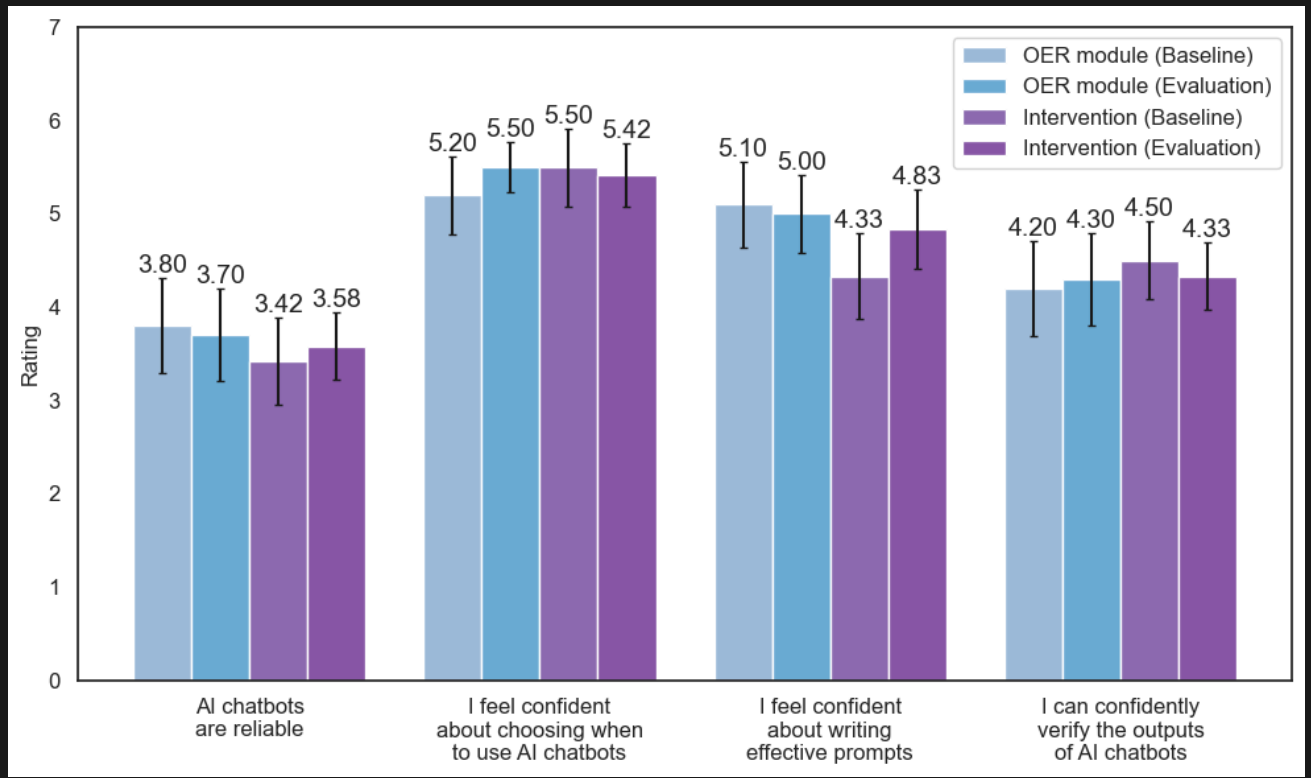
Outcome	OER Modules	Interventions
Appropriate Reliance on LLM	2	2
Appropriate Self-Reliance	3	5
Under-Reliance	1	2
Over-Reliance	1	1

- Intervention condition tend to **correctly rely on themselves** when chatbot gave the wrong answer compared to OER module condition.

Effects on Perception and Confidence with AI chatbots

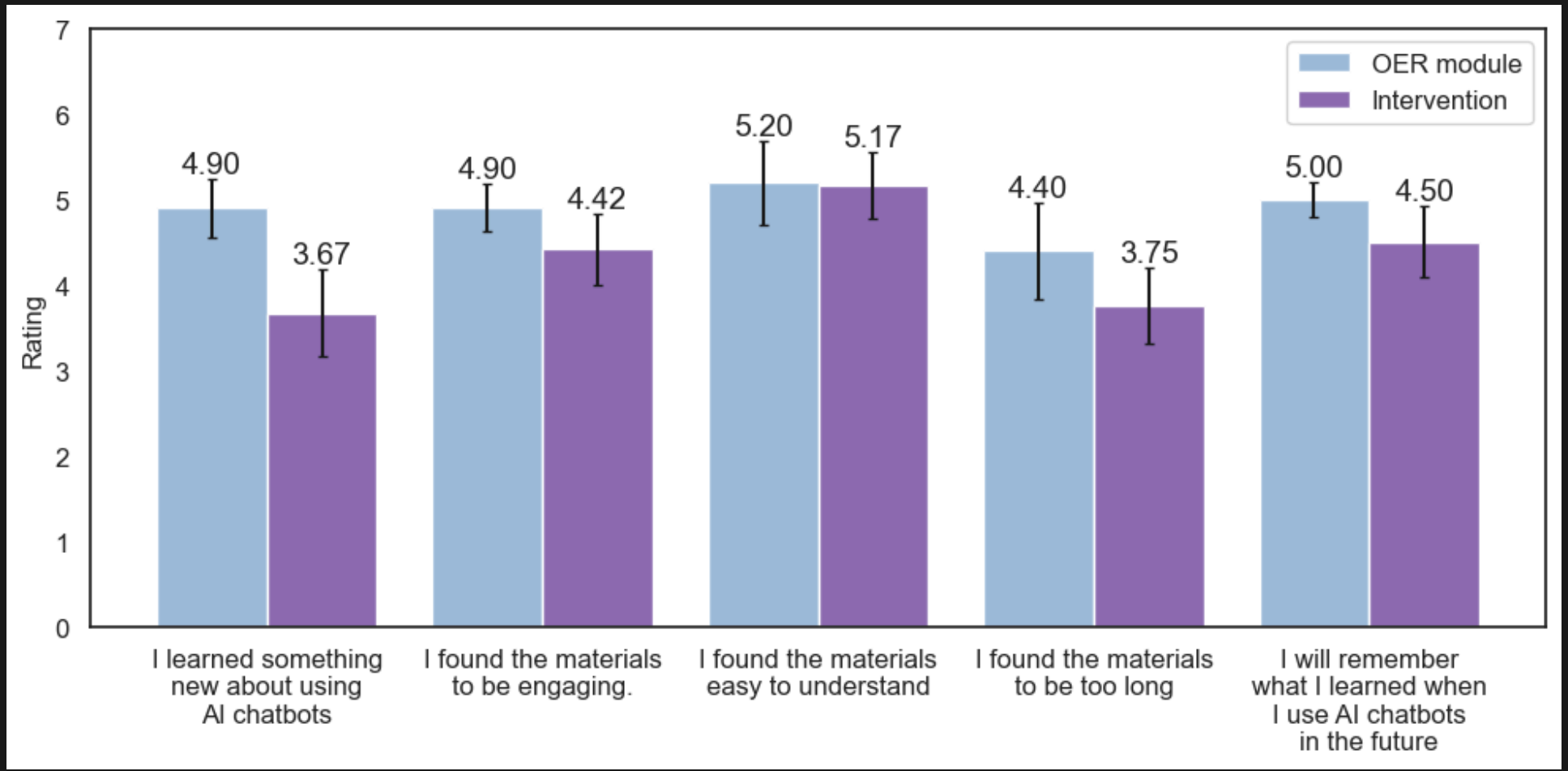
How do the interventions affect people’s perceptions and confidence with using AI chatbots?

Results



- We saw an **increasing trend** in the intervention condition in their rating regarding their view of **AI chatbots being reliable** ($p = 0.2915$) and **confidence in writing effective prompts** ($p = 0.4908$).
- However, we saw a **decrease in confidence** for the intervention in regards to **verifying the outputs of AI chatbots** ($p = 0.2075$) and **when to use chatbots** ($p = 0.1776$).

Feedback



- The intervention condition felt that they didn't learn something new about chatbots compared to the OER condition ($p = 0.070$) but they didn't find the material too long ($p = 0.3269$).

Thanks for listening!

References

[1] 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

[2] 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

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

[4] Bo, J. Y., Wan, S., & Anderson, A. (2025, April). 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 (pp. 1-23).