Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 25-1119

GenAI-101: What Undergraduate Students Need to Know and Actually Know About Generative AI

Sina Rismanchian
University of California, Irvine

Eesha Tur Razia Babar
University of California, Irvine

Shayan Doroudi
University of California, Irvine

In November 2022, OpenAI released ChatGPT, a groundbreaking generative AI chatbot backed by large language models (LLMs). Since then, these models have seen various applications in education, from Socratic tutoring and writing assistance to teacher training and essay scoring. Despite their widespread use among high school and college students in the United States, there is limited research on students' understanding and perception of these technologies. This study aims to fill that gap by developing a novel framework for generative artificial intelligence (GenAI) literacy, focusing on what undergraduate students know about generative AI and how they perceive the capabilities of AI chatbots. We designed a GenAI literacy survey to measure students' knowledge and perceptions, collecting data from 568 undergraduate students. The results show that about 60% of students use AI chatbots regularly for academic tasks, but they often overestimate the capabilities of these tools. However, increased knowledge about how generative AI works correlates with more accurate estimation of its capabilities in real-world tasks. Our findings highlight the need for enhanced GenAI and AI literacy to ensure students use these tools effectively and responsibly. This research underscores the importance of developing educational strategies and policies that prepare students for critical and informed engagement with AI technologies.

VERSION: January 2025

Suggested citation: Rismanchian, Sina, Eesha Tur Razia Babar, and Shayan Doroudi. (2025). GenAI-101: What Undergraduate Students Need to Know and Actually Know About Generative AI. (EdWorkingPaper: 25 -1119). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/1087-pv22

GenAI-101: What Undergraduate Students Need to Know and Actually Know About Generative AI

Sina Rismanchian, Eesha Tur Razia Babar, Shayan Doroudi

School of Education, University of California, Irvine

Irvine, CA 92697, USA

Corresponding Author: Sina Rismanchian

October 29, 2024

Abstract

In November 2022, OpenAI released ChatGPT, a groundbreaking generative AI chatbot backed by large language models (LLMs). Since then, these models have seen various applications in education, from Socratic tutoring and writing assistance to teacher training and essay scoring. Despite their widespread use among high school and college students in the United States, there is limited research on students' understanding and perception of these technologies. This study aims to fill that gap by developing a novel framework for generative artificial intelligence (GenAI) literacy, focusing on what undergraduate students know about generative AI and how they perceive the capabilities of AI chatbots. We designed a GenAI literacy survey to measure students' knowledge and perceptions, collecting data from 568 undergraduate students. The results show that about 60% of students use AI chatbots regularly for academic tasks, but they often overestimate the capabilities of these tools. However, increased knowledge about how generative AI works correlates with more accurate estimation of its capabilities in real-world tasks. Our findings highlight the need for enhanced GenAI and AI literacy to ensure students use these tools effectively and responsibly. This research underscores the importance of developing educational strategies and policies that prepare students for critical and informed engagement with AI technologies.

Keywords: AI literacy, generative artificial intelligence, folk theory

GenAI-101: What Undergraduate Students Need to Know and Actually Know About Generative AI

Introduction

In November 2022, OpenAI released ChatGPT, a general-purpose generative artificial intelligence chatbot powered by large language models (LLMs) trained on extensive text data from the web, books, and academic papers (Hötte et al., 2022). Since the introduction of LLMs, educational researchers, technologists, practitioners, and policymakers have sought to address both the potential benefits and critical challenges posed by generative AI in education. These challenges include concerns such as cheating and academic dishonesty, while the opportunities focus on improving educational practices. For instance, these models have been applied in Socratic tutoring (e.g., Khan Academy's Khanmigo), writing assistance (Shi et al., 2022), teacher assistant training (Markel et al., 2023), and essay scoring (Tate et al., 2023). Additionally, new educational technologies have emerged that enable teachers to create customized chatbots for their students (e.g., Playlab). Meanwhile, several reports (Baek, Tate, & Warschauer, 2023; Digital Education Council, 2024; Hopelab, Common Sense Media, & Center For Digital Thriving, 2024) indicate widespread use of these models by high school and college students in the United States. In spite of known problems with these models such as cultural bias (Atari et al., 2023) and hallucinations (Rawte et al., 2023), the adoption of LLMs by students appears inevitable.

While much work has focused on creating tools and environments to enhance students' learning experiences using these models, there is limited research on what students know about these models and how they perceive LLM-based AI chatbots. Assessing students' knowledge about AI more generally is a key focus within AI literacy. Researchers have developed tests

(Hornberger, Bewersdorff, & Nerdel, 2023; Yau et al., 2022) and pedagogical practices (Ng et al., 2021; Williams et al., 2023) to evaluate and enhance students' understanding of AI. Given that AI encompasses a wide range of algorithms developed over the past six decades, education researchers approach AI literacy from various perspectives and contents, including heuristic algorithms, machine learning, and AI ethics (Ng et al., 2021). However, since the introduction of generative AI tools and their adoption by students, there have been limited efforts to explore students' knowledge of these tools and their perceptions of them. For instance, researchers have looked into students' use of these tools for cheating (Lee et al., 2024) and their attitudes toward and adoption of ChatGPT (Chang et al., 2024; Abdalla et al., 2024); however, there has been little focus on measuring students' knowledge of generative AI.

AI chatbots, particularly ChatGPT, have gained worldwide attention, with ChatGPT reaching one million users faster than any other technological product in history (Gordon, 2023). An essential aspect of the interaction between students and generative AI is students' ways of perceiving genAI's capabilities, limitations, and fundamentals. For instance, students might overestimate its abilities and consequently over-rely on it (Klingbeil, Grützner, & Schreck, 2024), or they might underestimate its capabilities and incorrectly assume they cannot use it for some tasks. More importantly, the beliefs students develop shape the ways they use AI (Williams et al., 2018).

In this work, we propose a novel framework for GenAI literacy, GenAI-101, built upon research in AI literacy and folk theory, and we develop a new GenAI literacy survey to measure students' knowledge about large language models and their perceptions of AI chatbots' capabilities in different tasks. We gather data from 568 undergraduate students to validate the test and measure students' knowledge and perceptions of GenAI. We find that while about 60% of the students report using AI chatbots for their academic tasks daily or weekly, they tend to overestimate these chatbots' capabilities. However, we find that knowledge about GenAI impacts the way students perceive its capabilities; when students know more about AI, they tend to more accurately estimate its capabilities. These results underscore the importance of enhancing students' GenAI literacy to shape their utilization of these tools in a more informed way. Understanding these dynamics is crucial for developing effective educational strategies and policies that prepare students to critically and responsibly engage with AI technologies as digital citizens (Touretzky & Gardner-McCune, 2022).

Our findings have implications for educational practices, policy development, and the future of AI integration in education. By understanding and addressing students' misconceptions and knowledge gaps, educators and policymakers can create more effective AI literacy programs, ensuring students are better prepared to engage with AI technologies critically and responsibly. Moreover, this research contributes to the broader discourse on AI ethics and societal impact, highlighting the need for informed public engagement with AI advancements.

Theoretical Framework

Researchers have posited AI literacy as encompassing various types of literacies, such as computational, digital, classical, and media literacy (Logan, 2024). In this study, we are not focused on computational literacy, which involves measuring students' knowledge of technical concepts in deep learning. Instead, our aim is to assess how students perceive these tools and the extent to which they understand concepts that are important for informed usage. This approach aligns more closely with digital and media literacy.

Long and Magerko's (2020) prominent conceptual framework identified five themes that researchers perceive as AI literacy, with seventeen different competencies associated with the themes that are needed by students to not only avoid misconceptions about AI but also prepare for a future where AI is ubiquitous. This framework has become one of the most widely adopted approaches for framing research on AI literacy and addressing students' knowledge needs. While this work and that of others (e.g., Ng et al., 2021) have served as great ways to think about AI literacy, there is a need to update our conceptualizations about AI literacy because the forms of artificial intelligence that students encounter have gone beyond search engines or voice assistants. Nowadays, students can use these models to seek solutions for their homework, request advice about their personal decisions, etc. Hence, our framework adapts previous work to address the emerging needs with the advent of generative AI and students' knowledge about these tools.

The introduction of LLMs represents a transformative change in how artificial intelligence is accessed and used by everyday individuals. Unlike earlier AI applications such as Alexa, which were built for specific tasks or to respond to voice commands, LLM-based chatbots enable users to have more natural, open-ended conversations with AI. This shift has made AI more accessible and user-friendly, allowing a much wider audience to leverage the capabilities of generative AI. These models are based on an advanced architecture that can demonstrate a better context awareness and a more sophisticated ability to generate human-like responses. This architectural advancement distinguishes LLMs from previous AI technologies by offering a more versatile and dynamic interaction experience.

Given this unique interaction paradigm, we argue that AI literacy in the context of generative AI differs from the traditional understanding of AI or machine learning techniques taught in educational settings, such as K-nearest neighbors (Ng et al., 2021). For example, while it is not necessary to grasp complex deep learning concepts[1] to use these chatbots effectively, it is crucial to understand how these chatbots differ fundamentally from search engines like Google. This understanding is essential because it affects how users interpret and trust the information generated by AI, highlighting the need for a different kind of literacy—one that emphasizes understanding the unique characteristics and limitations of generative AI tools.

Although not everyone fully understands the technical workings of these chatbots, users often develop "folk theories"—intuitive, informal explanations about the outcomes, effects, or consequences of these technologies. These folk theories guide users' reactions to—and interactions with—chatbots (DeVito, Gergle, & Birnholtz, 2017). The concept of folk theories is well-established in human-computer interaction (HCI) research and has been applied to understanding various technological phenomena, such as social media, voice assistants (e.g., Alexa), and other digital tools (DeVito, Gergle, & Birnholtz, 2017; Druga et al., 2017; Eslami et al., 2015). Given that these chatbots and large language models are relatively new, education and HCI researchers need to understand students' current beliefs and perceptions about these chatbots. This understanding can inform future instruction and help educators address potential misconceptions about these models.

A comprehensive framework for GenAI literacy not only tries to examine students' knowledge about generative AI but also aims to understand the way they perceive these models so it can facilitate future AI literacy efforts and curricula to be responsive to common presumptions, misconceptions, and knowledge that students possess. This is akin to Long and Magerko's (2020) inclusion of the question "How do people perceive AI?" as one of their five themes. We note that while their other four themes are questions that we would like students to answer, the question of

[1] For example, technical terms such as multi-head attention (Vaswani et al., 2017).

how people perceive AI is one that would inform educators, designers, and researchers in how we study and design for AI literacy.

As such, GenAI-101 consists of two intertwined components: Knowledge and Perception of Generative AI. The Knowledge component is composed of three constructs: "What is generative AI?", "Emergent capabilities and open problems in generative AI," and "GenAI and societal impact." The Perception component is composed of a single construct: "Perceptions of AI chatbots." While we have separated these components out, we note that they are actually intertwined insofar as students' knowledge of AI informs their perceptions of AI and vice versa. Figure 1 provides a depiction of GenAI-101 and Table 1 shows how the constructs in GenAI-101 build upon prior work in AI literacy and folk theory.

In the knowledge component, for the first construct, we focus on measuring participants' ability to distinguish between different types of intelligent systems (e.g., human-level understanding, search engines, and basic algorithmic functions) and their ability to qualitatively explain how LLMs work. The second construct aims to assess participants' awareness of the limitations and strengths of this technology. Finally, the third construct targets the ethical and societal challenges associated with these models.

Table 1

The interrelationships between GenAI-101, Long & Magerko (2020), and folk theory (DeVito, Gergle, & Birnholtz, 2017)

Theoretical Frameworks	Themes	GenAI-101	keywords
How do researchers define AI literacy? What are the essential competencies that students need regarding AI? Long & Magerko (2020)	What is AI?	What is generative AI?	Emergence vs. by training, dataset size, RLHF, probabilistic next token prediction, prompting
	What can AI do?	Emergent capabilities and open problems in generative AI	Hallucinations, unlearning, logical reasoning, prompt injection,
	How does AI work?		
	How should AI be used?	GenAI and societal impact	Guardrails, jailbreaking, RLHF, Bias, AI alignment
	How do people perceive AI?	Perceptions of AI chatbots	Search engines, human-like reasoning, databases
Folk theory (DeVito, Gergle, & Birnholtz, 2017)	What are people's informal theories about technology?		

In this work, we design a novel survey on GenAI literacy to not only measure students' conceptual understandings in AI literacy but also further characterize their overt understandings related to their folk theories and possible misconceptions about GenAI in the domain of chatbots. Overall, we use the GenAI-101 framework and survey to address the following research questions:

RQ1) How much GenAI literacy do university students have? What are the factors that influence their GenAI literacy?

RQ2) How do students perceive the abilities of AI chatbots? How often do they tend to overestimate these models' abilities and what are the factors that impact it?

Methods

3.1. Survey Design

We designed a survey on generative AI literacy with a focus on large language models, in which there are two major types of questions: 1) *knowledge questions*, including 14 questions with four options (except for one question with three options) where only one is correct and is designed to capture what general knowledge about generative AI students possess and 2) *perception questions*: including 17 5-point Likert items which are designed to investigate how participants estimate strengths and limitations in prevalent AI chatbots and to what extent their perceptions are close to chatbots' actual capabilities. The knowledge questions and perception questions naturally relate to the Knowledge and Perception components of GenAI-101 respectively. However, given the intertwined nature of knowledge and perceptions, we note that the perception questions also largely probe users' knowledge of generative AI, especially with relation to the "Emergent capabilities and open problems in generative AI" construct. On the other hand, as we describe below, we also use the knowledge questions alongside the perception questions to better understand participants' beliefs about AI.

To ensure the validity of our questions, we held a session with an expert in the field of natural language processing and employed a think-aloud technique to gain feedback on the correctness and accuracy of our questions and their options. Based on the session we updated two of our questions and removed one. We further pilot-tested our knowledge questions with a group of undergraduate students and collected their feedback on the readability and wording of the questions and updated the text of a few of our questions based on it.

3.1.1. Knowledge Questions

We designed our knowledge questions based on our theoretical framework. For each of the three constructs in the Knowledge component of our framework (e.g., what is generative AI?), we designed a few questions to assess participants' knowledge in that specific construct. We designed the questions and the distractor options based on recent literature in computer science that include research that we believe should inform students' AI literacy across the three constructs. Table 2 shows each knowledge question, its corresponding construct in GenAI-101, and related papers in the computer science literature that introduce the associated ideas or issues. In addition to the types of questions, the distractor options were crafted to reflect common misunderstandings identified in prior literature on AI literacy (e.g., Druga et al., 2017; Eslami et al., 2015), capturing misconceptions people may have about GenAI, such as believing that ChatGPT reasons like humans or that these models simply search the web to generate responses.

Table 2

Knowledge questions in our survey and corresponding constructs. Note that the wording of questions in this table differs from what is represented in the survey for the sake of brevity.

Construct	Question	Related Literature
What is generative AI?	How do large language models generate text?	Vaswani et al., 2017; Wei et al., 2022
	How do large language models generate programming code?	Vaswani et al., 2017; Roziere et al., 2023
	What is a prompt?	Wei et al., 2022
	What is the reasoning procedure in LLMs when they play chess?	Duan et al., 2024
Emergent capabilities and applications and open problems in generative AI	What does hallucination or confabulation mean in the domain of LLMs?	Rawte et al., 2023;

	How can we have an LLM forget data?	Eldan & Russinovich, 2023
	What is the reason for an LLM to answer a math problem wrong?	Razeghi et al., 2022
	What is the reason for an LLM to identify a non-sensitive question as a sensitive question?	Weidinger et al., 2022
Generative AI and societal impact	How do LLMs avoid answering sensitive questions?	Chao et al., 2023; Yong et al., 2023
	Is it possible to have an AI chatbot generate problematic or sensitive text?	Chao et al., 2023; Yong et al., 2023
	Do LLMs work well in all different languages?	Robinson et al., 2023
	Is it possible for public or independent researchers to investigate the data used for training proprietary chatbots such as ChatGPT?	Balloccu et al., 2024
	Can LLMs be biased toward some cultures?	Atari et al., 2023
	Does an AI chatbot answer a specific sensitive question about the best professions for different genders?	Ghosh, & Caliskan, 2023

3.1.2. Perception Questions

We designed 17 different perception questions where we asked participants how they assign a likeliness to their AI chatbot to come up with a correct response to a prompt. For instance, one of the questions in the survey is as follows: “How likely do you think an AI chatbot is to answer the following question correctly? ‘Where was the current President of the United States born?’” The domain of these prompts spans a wide range of areas including information retrieval

(4 tasks), safety measures (1 task), commonsense questions (3 tasks), math problems (3 tasks), counting tasks (2 tasks), text transformation (1 task), and automatic text generation (3 tasks). The tasks were chosen based on related literature that reported the strengths of these models in text generation tasks (Bubbeck et al., 2023) and those that reported specific challenges and limitations such as mathematical or commonsense problems (Cherian et al., 2023; Mitchell, 2023).

For each perception question, we evaluated the performance of the most commonly used chatbots—ChatGPT-3.5, GPT-4, and Gemini/Bard—against specific prompts to determine their success rates. Each chatbot was tested five times with the same prompts, and we recorded how often they successfully answered the questions[2]. The performance of each chatbot on each question was then calculated by averaging the number of successful attempts. To obtain an overall performance score for the chatbots, we averaged the individual performances of the three chatbots, assigning scores on a scale of 1 to 5. These actual performance scores served as a reference point for comparing participants' estimations of a chatbot's ability, helping to identify overestimations or underestimations. In addition to measuring participants' perception of AI chatbots, in terms of how likely they are to overestimate and underestimate their abilities, these questions probe further into students' understanding of the emerging capabilities and open problems (or limitations) in generative AI, the third construct in the Knowledge component of our framework.

A potential question regarding this approach is why we did not compare each participant's answers to the performance of their most-used chatbot. While it seems reasonable to assume that users of different chatbots might have significantly different perceptions, our data did not support this. We conducted a two-sided t-test to examine whether users of different chatbots had significantly different perceptions, and out of 17 questions, significant differences were found in

[2] The tests took place in March 2024.

only 4. Notably, in 3 of those 4 questions, the chatbots had the same performance. Therefore, we chose not to personalize this measure and instead used an aggregated measure of chatbot performance in our analyses.

3.2. Data Collection

To have a relatively diverse sample of undergraduate students, we collected data from three different groups of participants, two from intact classes and one from the online participant recruitment platform, Prolific (Peer et al., 2022). In the first phase, we gathered data from students in two classes in a large public R1 university in the United States[3]: 1) computer science students who were enrolled in a course on introduction to artificial intelligence (which did not cover state-of-the-art techniques in natural language processing such as LLMs) and 2) psychological sciences and education sciences students enrolled in a course on cognition and learning in K-12 educational settings. Participants were recruited by email and were offered 2% extra credit for their participation. As per IRB requirements, an alternative task was provided for those who did not wish to participate in the research; they could watch a movie on generative AI and write a short summary paragraph to receive the extra credit. This phase of the study was conducted during March 2024. Overall, 257 students were recruited in this phase.

In the second phase, during May 2024, we used Prolific to gather data from a wider range of students throughout the US. In this phase, we gathered data from undergraduate students in the United States with an exclusion criteria of age under 18 or higher than 40. The median time it took each participant to complete the survey was around 15 minutes and we compensated each participant with an average rate of $17.52 per hour. In this phase, 334 participants were recruited.

[3] The study was conducted in accordance with ethical guidelines and Institutional Review Board (IRB) approval was obtained prior to the commencement of the research.

We excluded 23 participants who failed to answer two attention questions correctly, resulting in a final sample of 568.

The final sample demographic of our participants comprises 44% computer science students, 13% STEM students except for computer sciences, and 43% non-STEM students. Regarding participants' gender, 54% are male, 44% are female, and 1% identify themselves as having other gender identities.

Participants' demographics as well as their reported chatbot usage frequency and the type of chatbots that they use are reported in Table 3. Among usage statistics, we did not find a significant difference between reported data from the groups of our participants.

Table 3

Descriptive demographic data of participants in our survey

	CS majors	Other STEM majors	Non-STEM majors	Male	Female	Never used chatbot	Daily or Weekly usage	Using ChatGPT	Using GPT-4
R1 University	77.2%	0%	22.8%	62.4%	37.1%	19.8%	66%	64.1%	19.5%
Prolific	21.1%	22.1%	56.8%	47.7%	49.2%	21%	55.4%	72.8%	19.5%
Total	44.5%	12.9%	42.6%	53.9%	44.2%	20.5%	60.2%	67.9%	19.5%

We asked participants what chatbots they used the most and 68% percent of participants reported using ChatGPT3.5, 19% using GPT4 (the paid version of ChatGPT with higher capabilities), 8% using Gemini/Bard, and the rest using other chatbots. This finding is notable as the differences in capabilities of ChatGPT3.5 and GPT4 are considerable (Bubbeck et al., 2023), and most of our participants are using free versions that lag behind the state-of-the-art models.[4]

Furthermore, 17% of our participants reported using AI chatbots daily, 43% weekly, 20% monthly, and the rest 20% rarely or never on average which shows the widespread usage of these tools among undergraduate students in the US. However, we want to note that as the most of participants are recruited from either a CS class or from Prolific–itself a technology platform–it seems possible that the adoption rate of AI chatbots in our sample may be higher than for undergraduate students in large

3.3 Analysis

3.3.1. Students' Overestimation

Following the excitement that the introduction of ChatGPT brought to students, educators, and technology leaders, many researchers warned about the ethical challenges of these tools as well as the hype around them that might mislead people to overestimate the abilities these models have (Rudolph, Tan, & Tan, 2023). Given the circumstances, we hypothesized that university students might tend to overestimate the abilities of current chatbots. We aimed to first answer the question about the proportion of participants overestimating abilities in the chatbots. To measure if a participant is overestimating a chatbot's specific ability, we first calculated three prominent chatbots' actual abilities in successfully responding to a prompt in the range of 1 to 5 as described in the methods section. Subsequently, we subtract the average actual ability of the chatbots from students' assigned ability using Likert scores. If the result is larger than 1, then the student is

[4] Recently, OpenAI has made GPT-4 model free to users, but at the time that we conducted the survey, only ChatGPT3.5 was available with no payment.

counted as overestimating that specific ability, and if the result is smaller than -1, counted as underestimating.

We further used linear regression models with overestimation percentage as the dependent variable and demographic variables as well as students' derived abilities from knowledge questions as predictors.

3.3.2. Students' Beliefs

To explore different patterns of beliefs and theories that participants develop about GenAI, we need to delve into their choices in knowledge and perception questions. We employed the K-modes clustering method (Chaturvedi, Green, & Caroll, 2001). K-modes clustering is an unsupervised machine learning algorithm tailored for categorical data. Unlike k-means clustering, which relies on calculating the mean, K-modes employs the mode to determine cluster centroids, making it well-suited for non-numeric data. This algorithm iteratively assigns data points to clusters based on the similarity of their categorical attributes, updating the centroids to minimize the total distance within clusters.

After running the elbow method to determine the optimal number of clusters, we found that the elbow point appeared at k=8 for within-cluster simple-matching distance for each cluster. However, considering that eight clusters seemed excessive for our dataset, we decided to manually test for other indices of clustering using NbClust package (Charrad et al., 2014) in R. Through this process, we calculated the dissimilarity matrix using Gower method and using the matrix we found that by majority voting of five different indices, k = 3 was chosen as the optimal number for our clustering. Especially, k = 3 had highest scores in Silhouettes (Rousseeuw, 1987) and Dunn (1974) indices, showing that it is an appropriate number of clusters, balancing interpretability with within-cluster homogeneity. This allowed for a more coherent analysis of participants' patterns.

Results

4.1. Measurement

4.1.1. Analysis of The Test

Similar to related work on AI literacy tests (Hornberger, Bewersdorff, & Nerdel; 2023), we use Item Response Theory (IRT) models to analyze our test data and its reliability using *mirt* package in R (Chalmers, 2012). In these models, participants' abilities and item characteristics are estimated on a joint scale (Embretson & Reise, 2013). IRT offers several advantages over classical test theory, such as the ability to analyze test data at the item level. The three most prominent IRT models are the Rasch (1-PL) model, the 2-PL model, and the 3-PL model (Embretson & Reise, 2013). These models differ in the number of parameters they estimate for item characteristics. The Rasch model estimates only one parameter for item difficulty. The 2-PL model additionally estimates parameters for item discrimination and the 3-PL model extends the 2-PL model by adding the guessing parameter (γ) which reflects the likelihood of participants with very low abilities guessing the correct answer, enhancing the model's ability to differentiate between varying levels of participant ability.

Before fitting the IRT models for the analysis of our survey, we tested for unidimensionality and local independence assumptions in our data and we found that the unidimensionality assumption held true; unidimensionality was measured using a confirmatory factor analysis (CFA) model, resulting in root mean square error of approximation (RMSEA) < 0.03, standardized root mean squared residuals (SRMR) < 0.05, $\chi 2/df < 2$.

4.1.2. Model Selection

In selecting the appropriate IRT model for our analysis, we considered the underlying assumptions and characteristics of each model. Given the confirmation of unidimensionality in our dataset, we focused on models that best accommodate this structure. While the Rasch (1-PL) model offers simplicity and strict requirements for measurement equivalence, its assumption that all items have the same discrimination parameter may not fully capture the nuances in our data. Notably, one item showed poor fit in the Rasch model, suggesting that a more flexible model might be necessary. The 2-PL and 3-PL models, which allow for varying item discriminations and, in the case of the 3-PL model, incorporate a guessing parameter, appeared more suitable.

In our analysis, we compared the Rasch, 2-PL, and 3-PL models using fit indices such as AIC, BIC, log-likelihood, Tucker-Lewis Index (TLI), and Comparative Fit Index (CFI) to determine the best fit for our data. Initially, the 2-PL model showed superior fit compared to the Rasch model, with lower AIC and significantly improved log-likelihood (χ2(13)=45.882, p<.001), indicating the added benefit of item discrimination parameters. Comparing the 2-PL and 3-PL models, the 3-PL model exhibited further improvement with significantly lower AIC and continued enhancement in log-likelihood (χ2(14)=44.028, p<.001), accommodating both item discrimination and guessing effects. Additionally, the 3-PL model met the threshold criteria for TLI and CFI, which assess model fit relative to a baseline model, further supporting its suitability despite higher BIC.

Table 4

Model fit indices for all Rasch, 2-PL, and 3-PL models. The 3-PL model suggests the best fit.

Model	**M2**	**RMSEA**	**SRMSR**	**TLI**	**CFI**	**AIC**	**BIC**

Rasch	143.896	0.043	0.071	0.89	0.891	5478.521	5535.552
2-PL	98.876	0.029	0.05	0.948	0.956	5458.639	5565.098
3-PL	71.075	0.02	0.048	0.977	0.984	5442.611	5602.300

Note. M2 = RMSEA =root mean square error of approximation; SRMR =standardized root mean squared residuals, TLI =Tucker–Lewis index; CFI =comparative fit index; AIC =Akaike information criterion; BIC =Bayesian information criterion.

Item fit statistics of all models for the knowledge questions are included in Table 5. With 2-Pl and 3-PL models all items show good fit while in the Rasch model, two items show poor fit.

Using the 3-PL model, we examined the assumption of local independence, which posits that the responses to test items are independent of each other given the participant's ability level. We assessed this assumption using correlations of residuals for the Q3 index across all items. According to Chen and Thissen (1997), local independence is upheld when these correlations are smaller than 0.2. Our analysis confirmed that all correlations fell below this threshold, affirming the assumption of local independence within our test data.

On the other hand, for the combination of knowledge and perception questions we used Cronbach Alpha's statistic to measure the internal consistency of our survey, and with alpha = 0.82 it suggests that the survey is highly reliable in measuring the underlying construct.

Table 5

Item fit statistics. 2-PL and 3-PL models show a good fit for all items.

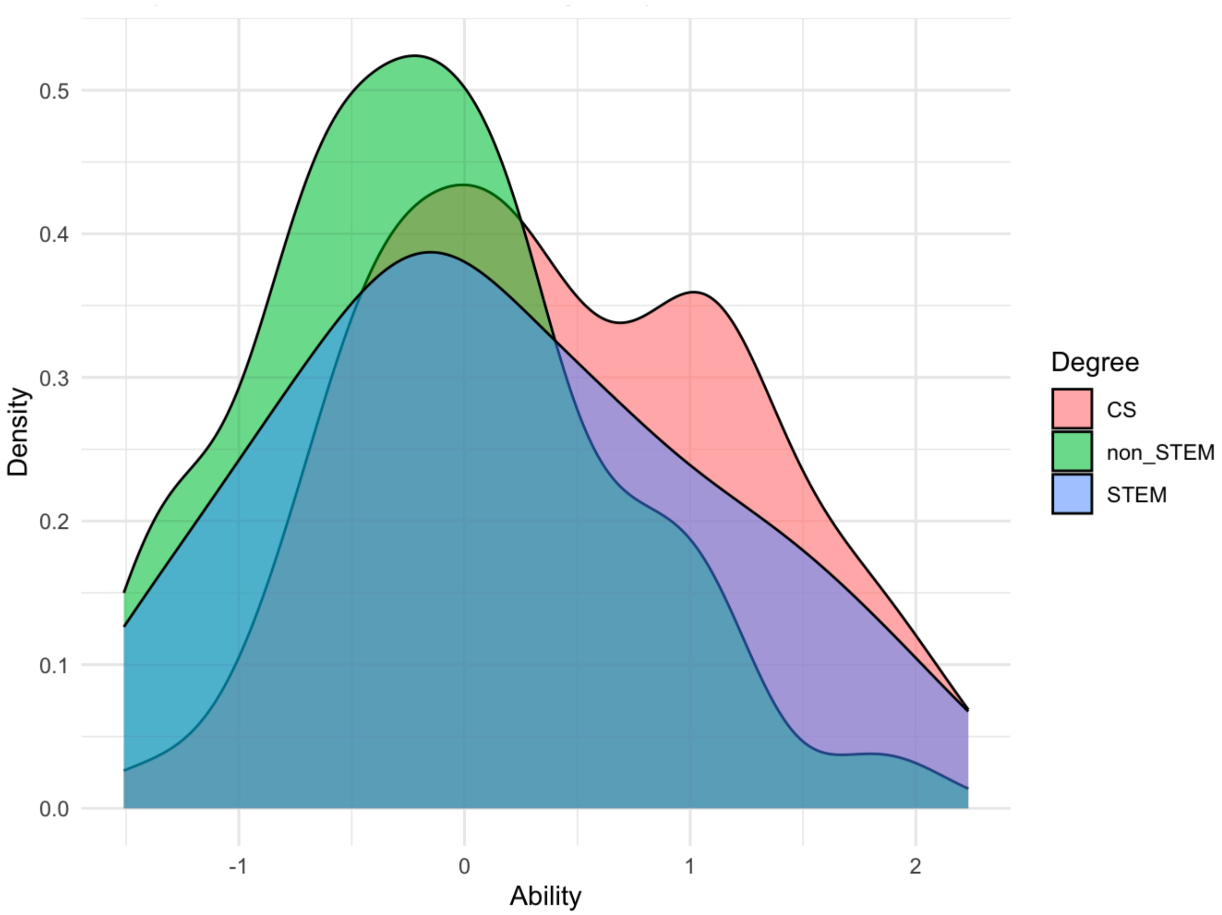
	Rasch			2-PL			3-PL		
Item	**S-X2**	**df**	**p**	**S-X2**	**df**	**p**	**S-X2**	**df**	**p**
TextGen	11.034	9	0.273	10.529	9	0.309	8.715	7	0.274
CodeGen	12.654	8	0.124	6.295	7	0.506	3.712	6	0.716
IncorrectMult	11.376	9	0.251	6.355	9	0.704	7.901	7	0.341
RefuseReq	20.748	8	**0.008**	6.597	9	0.679	3.454	8	0.903
PromptDef	6.214	8	0.623	5.451	8	0.708	6.123	7	0.525
Hallucination	4.745	8	0.784	3.314	8	0.913	1.421	5	0.922
ForgetData	16.989	8	**0.03**	10.098	9	0.343	6.923	8	0.545
Professions	12.492	8	0.131	12.683	8	0.123	12.742	7	0.079
ChessReasoning	1.346	9	0.998	2.069	8	0.979	2.426	8	0.965
SensitiveAns	8.5	8	0.386	3.383	7	0.847	3.44	6	0.752
ReligionQ	8.404	8	0.395	10.272	9	0.329	10.528	8	0.23
LangPerf	10.398	9	0.319	9.8	8	0.279	10.943	7	0.141
LLMBias	12.307	7	0.091	5.671	6	0.461	8.444	6	0.207
LLMData	4.993	8	0.758	5.366	8	0.718	4.288	7	0.746

4.2. How do university students perform in our GenAI literacy survey?

Using the ability variable from the 3-PL model, we can derive participants' knowledge in Generative AI. As Figure 1 shows, students in computer science majors tend to perform better than students in other majors as they have a smaller number of low-achieving students. On the other hand, non-STEM students have a lower number of high-achieving students compared to other groups. The difference between groups here might seem natural, but as the concept of generative AI is relatively new, the difference here might not be explained by the different courses that those students take. Using linear models in regression analysis, we tried to explain the differences among these groups better. Not surprisingly, we found that computer science students who were recruited from the AI course performed significantly better than the rest of the students (and even than other CS students from different universities). Additionally, the frequency of students using chatbots is associated with higher abilities in the test. Table 6 shows how different

Figure 1

Distribution of abilities in different degree groups



demographic variables are associated with students' derived abilities from knowledge questions. As the table shows, female students performed significantly better than male students, and students who are not English native speakers performed lower than native English speakers. This difference is particularly noteworthy because researchers had hypothesized that generative AI tools could help close the achievement gaps between native speakers and English language learners (Warschauer et al., 2023). However, the persistence of these gaps in students' knowledge about these models suggests that the benefits of generative AI tools may not fully extend to mitigating existing disparities.

Table 6

Regression results using ability as the dependent variable

Predictor	*b*	*b*	*sr^2*	*sr^2*	Fit
		95% CI		95% CI	
		[LL, UL]		[LL, UL]	
(Intercept)	0.33	[-0.03, 0.68]			
Major (Baseline: CS)					
non-STEM	-0.15	[-0.38, 0.07]	.00	[-.01, .01]	
STEM	0.04	[-0.22, 0.31]	.00	[-.00, .00]	
TimeTaken	0.00	[-0.00, 0.00]	.00	[-.00, .00]	
Chatbot Usage Frequency	0.10**	[0.03, 0.17]	.01	[-.00, .03]	
Chatbot used (Baseline: Gemini)					
GPT4	-0.18	[-0.45, 0.10]	.00	[-.00, .01]	
ChatGPT3.5	-0.14	[-0.39, 0.11]	.00	[-.00, .01]	
Bing/Other	-0.17	[-0.57, 0.23]	.00	[-.00, .01]	
Gender (Baseline: Male)					
Female	0.29**	[0.14, 0.44]	.02	[.00, .04]	
Other Gender	0.25	[-0.24, 0.74]	.00	[-.00, .01]	
Non-native English Speaker	-0.28*	[-0.52, -0.05]	.01	[-.01, .02]	

R1-Computer Science	0.46**	[0.13, 0.79]	.01	[-.00, .03]	
R1 University	-0.06	[-0.30, 0.19]	.00	[-.00, .00]	
					R^2 = .176** 95% CI[.11,.22]

Note. A significant *b*-weight indicates the semi-partial correlation is also significant. *b* represents unstandardized regression weights. sr^2 represents the semi-partial correlation squared. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.
* indicates p < .05. ** indicates p < .01.

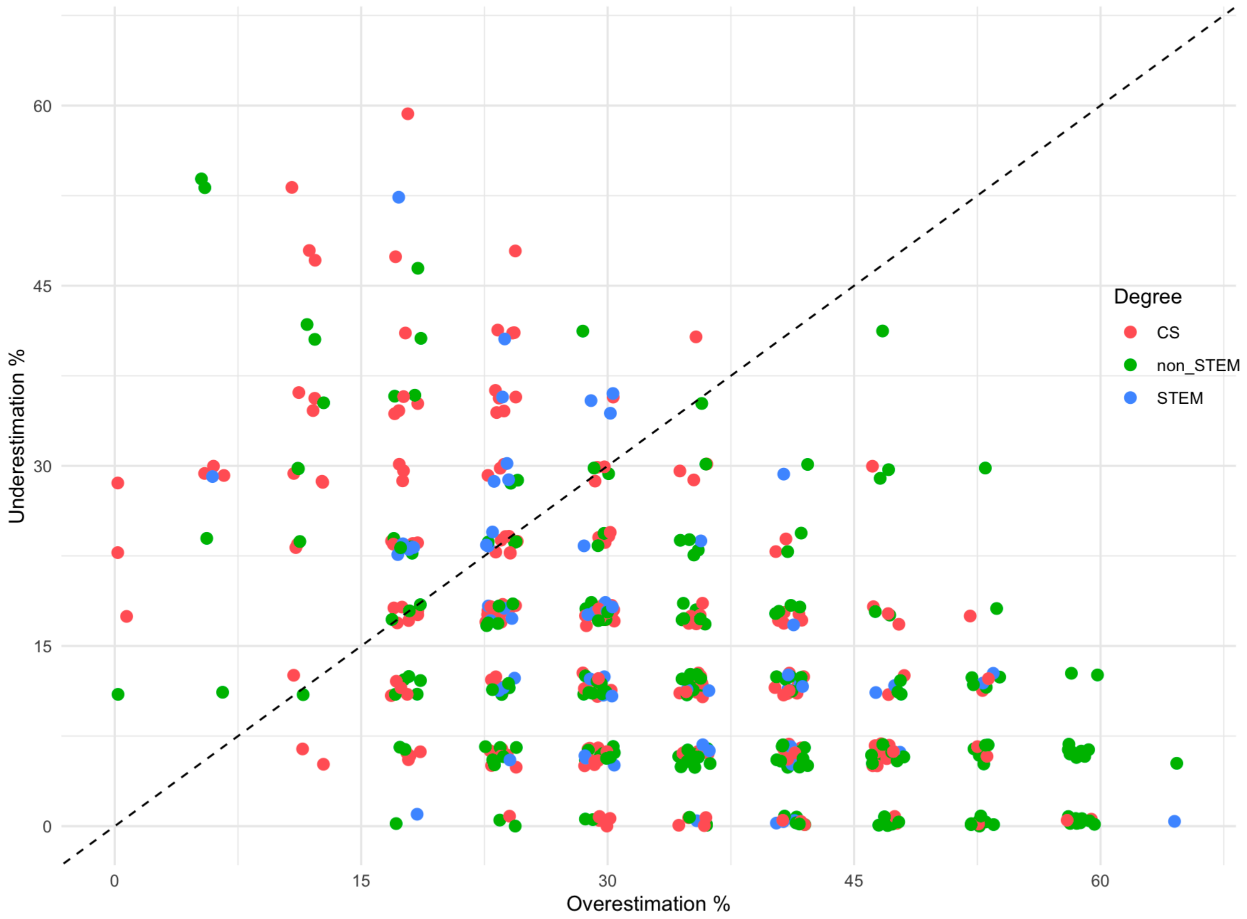
4.3. What are students' perceptions of generative AI?

4.3.1. Students Overestimate GenAI's Capabilities Too Often

We first hypothesized that most students tend to overestimate more often than underestimate, as Figure 2. Shows, where each student is depicted as a dot on the scatter plot where the x is their overestimation percentage and y is their underestimation percentage, most of the participants lie under the x=y line which shows that our hypothesis holds true. Perhaps in resonance with the results in previous sections, we find that non-STEM students tend to overestimate more compared to other groups (36.9% vs. 31.5% on average, p < 0.001), and interestingly, computer science students tend to underestimate more than other groups (15.01% vs. 11.74% on average, p = 0.001).

Figure 2

Scatter plot of students' overestimation vs. underestimation in perception questions. Students tend to overestimate more frequently than underestimate



Note: we have added noise of uniform(-1, 1) to the percentages to prevent identical dots from overlaying.

One of the common types of questions that students overestimate is the ones that ask about AI chatbots' abilities to count. For instance, one question was asking about the ability of an AI chatbot to come up with five words that each have two “a”s as in “again”. Although the question should be simple for humans, it is not the case for LLMs as they lack the ability to count. Usually, when this question is asked by an AI chatbot they come up with words such as “banana”, but most of the users (84%) assigned “very likely” for their chatbot to come up with the correct response to the question. On the other hand, we asked if their AI chatbot can write a paragraph that only includes the word "the" seven times; in this case, similarly, students assigned high likeliness for the chatbots to respond successfully to this prompt (45% assigned “very likely”) but as our tests show, the models are not accurate in these text generation regimes as they are probabilistic prediction models and they lack the accurate ability of counting.

On the other hand, when the questions were about tasks such as information retrieval about past events (e.g., winner of Superbowl 2022), participants were successful in determining the abilities of chatbots. However, users’ of ChatGPT3.5 also predicted a high likeliness of their chatbot to answer a question about recent events such as the winner of Superbowl 2024 (44% of ChatGPT3.5 users), while we know that this model does not have internet access and its data is updated until 2023, so it will not be able to respond to the prompt successfully.

4.3.2. More Knowledge Reduces Overestimation

As previously mentioned, one of the assumptions underlying our work is that increased knowledge about GenAI can enable students to use these tools in a more constructive and informed manner. This awareness includes understanding the strengths and limitations of these models, thereby allowing students to make informed decisions about when to utilize them and when to refrain from relying on them. To examine our assumption about the impact of students’ knowledge, we ran a linear regression model where the dependent variable is a student’s overestimation percentage and we incorporate their derived abilities from the knowledge question as an independent variable as well as other demographic variables such as their degree.

Table 7

Regression results using overestimation_percentage as the criterion

Predictor	*b*	*b* 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	Fit
(Intercept)	37.34**	[31.93, 42.75]			
Major (Baseline: CS)					
non-STEM	2.94	[-0.45, 6.34]	.00	[-.01, .01]	
STEM	0.48	[-3.54, 4.50]	.00	[-.00, .00]	
TimeTaken	0.00	[-0.00, 0.00]	.00	[-.00, .00]	
Chatbot Usage Frequency	1.27*	[0.18, 2.36]	.01	[-.01, .02]	
Chatbot used (Baseline: Gemini)					
GPT4	0.03	[-4.19, 4.24]	.00	[-.00, .00]	
ChatGPT3.5	-1.69	[-5.42, 2.03]	.00	[-.00, .01]	
BingAI/Other	2.27	[-3.81, 8.36]	.00	[-.00, .00]	
Gender (Baseline: Male)					
Female	1.92	[-0.34, 4.17]	.00	[-.01, .01]	
Other Gender	-8.57*	[-15.97, -1.17]	.01	[-.01, .02]	
Non-native English Speaker	-1.29	[-4.88, 2.30]	.00	[-.00, .00]	
ability	-6.35**	[-7.62, -5.09]	.14	[.09, .19]	
R1-Computer Science	0.85	[-4.25, 5.95]	.00	[-.00, .00]	
R1 University	-1.92	[-5.69, 1.85]	.00	[-.00, .01]	
					R^2 = .208** 95% CI[.13,.25]

Note. A significant *b*-weight indicates the semi-partial correlation is also significant. *b* represents unstandardized regression weights. sr^2 represents the semi-partial correlation squared. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively. * indicates p < .05. ** indicates p < .01.

As the results in Table 7 show, students' derived ability is a statistically significant predictor and a unit increase in students' derived abilities is associated with a 6.34% decrease in students' overestimation percentage on average.

4.3.3. What are participants' beliefs about GenAI?

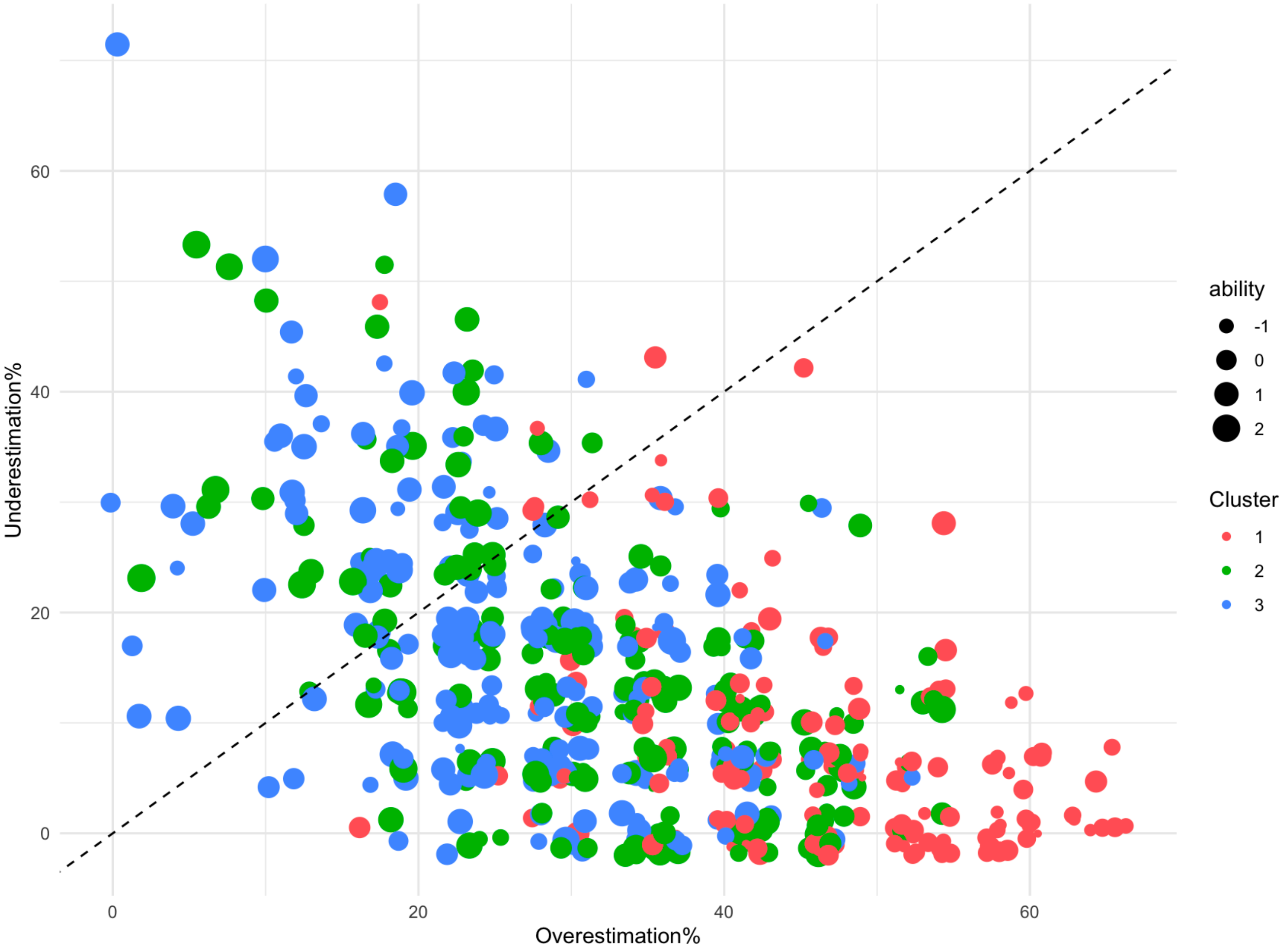
The k-modes clustering method provides insightful results about participants' understandings of these models that are related to their folk theories. For instance, cluster 1 in our data, which consists of 28% of participants—60% of whom are non-STEM students—is an outstanding case. This cluster has chosen options in response to knowledge questions that demonstrate a combination of anthropomorphism of AI and confusing GenAI tools with search engines like Google. For example, in response to why LLMs do not solve multiplication problems correctly, this cluster chose the option: "Similar to humans, they calculate the multiplication step-by-step and may make mistakes along the way." Additionally, when explaining these models' abilities to generate answers to questions or programming code, they selected: "LLMs search the data available to them and generate answers based on all the question-answer data they have," and "They search their database to find the best code that matches a problem," respectively. They have chosen the "very likely" option in most of the perception questions which shows their high expectations of these models. More specifically, Figure 3 shows where this cluster usually lies on the high-estimation side of the spectrum. In contrast, the other cluster can be seen on both sides of the over- and under-estimation.

Clusters 2 and 3 show some similarity in their responses to knowledge questions, differing in only three questions (21%). However, Cluster 3 exhibits more skepticism in their responses to perception questions. In 8 out of 16 questions (50%), participants in Cluster 2 chose the "very likely" option, whereas those in Cluster 3 opted for "likely." This suggests that participants in Cluster 2 are less skeptical in their perception of these models. We will discuss the implications of these beliefs in the discussion section in more detail.

Our results here resonate with previous literature on folk theories of conversational agents (Xu & Warschauer, 2020), voice assistants (Sciuto et al., 2018; Druga et al., 2017), etc. where children assigned anthoropomorphic attributes to conversatinoal agents or thought of voice assistants as Google search engines. However, the fact that participants in this work, as opposed to previous work, are adult learners shows that in the absence of accessible content on generative AI, and its capabilities and limitations people may think of these tools in such ways that in the meantime hinder them from making the most out of these tools (as LLMs have capabilities that exceed but are also different from those of search engines) and fall for the hype (Markelius et al., 2024) that exists around new technologies (assuming GenAI tools have human-level intelligence).

Figure 3

Overestimation vs. Underestimation for our K-modes clusters.



Note: Cluster 1 lies on the high-estimation side of the spectrum.

Discussion and Limitations

In this work, we introduced a novel measure of GenAI literacy based on prominent AI literacy frameworks and folk theory and validated it using IRT techniques. We provided insights into the status quo of undergraduate students' knowledge and perceptions of generative AI technologies.

Our results indicate significant overestimations, particularly among less technical demographics. While AI chatbot users are typically general users and laypeople, only a small proportion of our participants—specifically computer science students from a large R1 university in California—performed well on the knowledge test. Most other participants exhibited modest performance and frequently overestimated the capabilities of AI chatbots. Our findings indicate that higher levels of general knowledge regarding these tools and more frequent usage of these tools are associated with higher levels of knowledge. Although we cannot establish a causal link, this probably suggests that policymakers and educational leaders should consider encouraging students to use and learn about these tools and have more opportunities to use them responsibly to promote better, more informed usage.

Our work has limitations in terms of the number of knowledge test questions. Future work can build upon our work to extend the number and variety of the questions to measure students' GenAI more deeply. Furthermore, the relationship between students' knowledge of and theories about these models with their use cases can be explored, as different perceptions can impact the ways people utilize technology. Lastly, we only scratched the surface of folk theories in this study. We did not explore participants' beliefs about generative AI in depth but instead identified common patterns of belief. A more thorough investigation of students' folk theories would require qualitative studies, which could be pursued in future work.

As we discussed some exploratory folk theories in Section 4.3, we want to note that some folk theories can be harmful. Anthropomorphism might increase overreliance on and trust in AI while thinking of GenAI as search engines might lead students to believe that the results are always accurate and referenced, which is far from reality given the issue of hallucinations in AI-generated responses. However, educators can identify and harness these beliefs among their students, using them as starting points to encourage discussions and raise awareness about generative AI. This approach can help students develop a more accurate understanding of these models. For example, an AI literacy educator could prompt students to consider situations where humans and generative AI tools might behave differently, and what those differences reveal about the fundamental nature of each.

We believe that our work offers educators a solid foundation for teaching AI literacy, particularly GenAI literacy. By using our theoretical framework, instructors can target various knowledge areas and leverage students' folk theories to create meaningful learning experiences that resonate with their existing beliefs about AI. Also, using our survey, educators can find how optimistic or pessimistic their students' are about GenAI. While there are many approaches to developing GenAI literacy, we suggest that informed interactions with these tools might be an effective way to situate one's understanding of this new technology. Moreover, engaging practices that encourage learners to differentiate between different types of intelligent systems (e.g., humans, search engines, etc.) can be fruitful for enhancing knowledge and awareness about GenAI. The survey questions we developed, included as supplementary material, can serve not only as an evaluative assessment but also as a useful resource for educators and students to facilitate discussions around the capabilities and limitations of GenAI tools.

Given the reported usage of K-12 teachers' usage of GenAI tools in classrooms (Diliberti et al., 2024), future research can also explore teachers' GenAI literacy, especially those integrating AI tools into education. Better understanding teachers' perceptions of GenAI is essential for enhancing educational practices and ensuring students receive effective guidance in AI literacy. This assessment can guide professional development efforts to equip educators with the knowledge needed to navigate and teach with and about these technologies in classrooms.

The area of GenAI literacy is a burgeoning field of research. As this area evolves—indeed, as our understanding of generative AI and the underlying technologies evolve—we can expect researchers to conduct further studies targeting different aspects of GenAI literacy among diverse groups, particularly students. Given that students, as future digital citizens, have the potential to shape our future, it is crucial to provide them with learning opportunities to understand this new technology, its nature, and its potential impact on our future.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to check for grammatical error and improve wording. After using this tool/service, the authors reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Abdalla, A. A., Bhat, M. A., Tiwari, C. K., Khan, S. T., & Wedajo, A. D. (2024). Exploring ChatGPT Adoption among Business and Management Students through the Lens of Diffusion of Innovation Theory. Computers and Education: Artificial Intelligence, 100257.

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023, September 22). Which Humans?. https://doi.org/10.31234/osf.io/5b26t

Baek, C., Tate, T., & Warschauer, M. (2023, December 12). “ChatGPT Seems Too Good to be True”: College Students’ Use and Perceptions of Generative AI.

Balloccu, S., Schmidtová, P., Lango, M., & Dušek, O. (2024). Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. arXiv preprint arXiv:2402.03927.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. Research Methods in Applied Linguistics, 2(3), 100068.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., ... & Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv preprint arXiv:2307.15217.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. Journal of statistical Software, 48, 1-29.

Chang, H., Liu, B., Zhao, Y., Li, Y., & He, F. (2024). Research on the acceptance of ChatGPT among different college student groups based on latent class analysis. Interactive Learning Environments, 1-17.

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., & Wong, E. (2023). Jailbreaking black box large language models in twenty queries. arXiv preprint arXiv:2310.08419.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a data set. Journal of statistical software, 61, 1-36.

Chaturvedi, A., Green, P. E., & Caroll, J. D. (2001). K-modes clustering. Journal of classification, 18, 35-55.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics, 22(3), 265-289.

Cherian, A., Peng, K. C., Lohit, S., Smith, K. A., & Tenenbaum, J. B. (2023). Are deep neural networks SMARTer than second graders?. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10834-10844).

del Rio-Chanona, M., Laurentsyeva, N., & Wachs, J. (2023). Are large language models a threat to digital public goods? evidence from activity on stack overflow. arXiv preprint arXiv:2307.07367.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., ... & Lakhani, K. R. (2023). Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper, (24-013).

DeVito, M. A., Gergle, D., & Birnholtz, J. (2017, May). " Algorithms ruin everything" # RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In Proceedings of the 2017 CHI conference on human factors in computing systems (pp. 3163-3174).

Digital Education Council (2024). AI or Not AI: What Students Want. Digital Education Council Global AI Student Survey 2024.

Diliberti, M., Schwartz, H. L., Doan, S., Shapiro, A. K., Rainey, L., & Lake, R. J. (2024). Using Artificial Intelligence Tools in K-12 Classrooms. RAND.

Dipaola, D., Payne, B. H., & Breazeal, C. (2022). Preparing children to be conscientious consumers and designers of AI technologies. Computational thinking education in K-12: Artificial intelligence literacy and physical computing, 181-205.

Druga, S., Williams, R., Breazeal, C., & Resnick, M. (2017, June). " Hey Google is it ok if I eat you?" Initial explorations in child-agent interaction. In Proceedings of the 2017 conference on interaction design and children (pp. 595-600).

Duan, J., Zhang, R., Diffenderfer, J., Kailkhura, B., Sun, L., Stengel-Eskin, E., ... & Xu, K. (2024). Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. arXiv preprint arXiv:2402.12348.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. Journal of cybernetics, 4(1), 95-104.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jian, L., Lin, B. Y., ... & Choi, Y. (2023). Faith and Fate: Limits of Transformers on Compositionality. arXiv preprint arXiv:2305.18654.

Eldan, R., & Russinovich, M. (2023). Who's Harry Potter? Approximate Unlearning in LLMs. arXiv preprint arXiv:2310.02238.

Embretson, S. E., & Reise, S. P. (2013). Item response theory. Psychology Press.

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., ... & Sandvig, C. (2015, April). " I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds. In Proceedings of the 33rd annual ACM conference on human factors in computing systems (pp. 153-162).

Ghosh, S., & Caliskan, A. (2023, August). Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (pp. 901-912).

Gordon, C. (2023). "ChatGPT Is The Fastest Growing App In The History Of Web Applications." *Forbes*. Retrieved from https://www.forbes.com/sites/cindygordon/2023/02/02/chatgpt-is-the-fastest-growing-ap-in-the-history-of-web-applications/

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models. arXiv preprint arXiv:2302.12173.

Henning, J. (2007). The art of discussion-based teaching: Opening up conversation in the classroom. Routledge.

Hopelab, Common Sense Media, & Center For Digital Thriving. (2024). Teen and Young Adult Perspectives on Generative AI: Patterns of use, excitements, and concerns.

Hornberger, M., Bewersdorff, A., & Nerdel, C. (2023). What do university students know about Artificial Intelligence? Development and validation of an AI literacy test. Computers and Education: Artificial Intelligence, 5, 100165.

Hötte, K., Tarannum, T., Verendel, V., & Bennett, L. (2022). Exploring Artificial Intelligence as a General Purpose Technology with Patent Data--A Systematic Comparison of Four Classification Approaches. arXiv preprint arXiv:2204.10304.

Kang, D., Li, X., Stoica, I., Guestrin, C., Zaharia, M., & Hashimoto, T. (2023). Exploiting programmatic behavior of llms: Dual-use through standard security attacks. arXiv preprint arXiv:2302.05733.

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. Computers in Human Behavior, 108352.

Lee, V. R., Pope, D., Miles, S., & Zárate, R. C. (2024). Cheating in the age of generative AI: A high school survey study of cheating behaviors before and after the release of ChatGPT. Computers and Education: Artificial Intelligence, 7, 100253.

Leivada, E., Murphy, E., & Marcus, G. (2023). DALL· E 2 fails to reliably capture common syntactic processes. Social Sciences & Humanities Open, 8(1), 100648.

Li, M., & Suh, A. (2021, January). Machinelike or humanlike? A literature review of anthropomorphism in AI-enabled technology. In 54th Hawaii International Conference on System Sciences (HICSS 2021) (pp. 4053-4062).

Logan, C. (2024). Learning About and Against Generative AI Through Mapping Generative AI's Ecologies and Developing a Luddite Praxis. In Proceedings of the 18th International Conference of the Learning Sciences-ICLS 2024, pp. 362-369. International Society of the Learning Sciences.

Mahmood, R., Wang, G., Kalra, M., & Yan, P. (2023, October). Fact-Checking of AI-Generated Reports. In International Workshop on Machine Learning in Medical Imaging (pp. 214-223). Cham: Springer Nature Switzerland.

Markel, J. M., Opferman, S. G., Landay, J. A., & Piech, C. (2023). GPTeach: Interactive TA Training with GPT Based Students.

Markelius, A., Wright, C., Kuiper, J., Delille, N., & Kuo, Y. T. (2024). The mechanisms of AI hype and its planetary and social costs. AI and Ethics, 1-16.

Mitchell, M. (2023). AI's challenge of understanding the world. Science, 382(6671), eadm8175.

Moskvichev, A., Odouard, V. V., & Mitchell, M. (2023). The ConceptARC Benchmark: Evaluating Understanding and Generalization in the ARC Domain. arXiv preprint arXiv:2305.07141.

Nargundkar, S., Samaddar, S., & Mukhopadhyay, S. (2014). A guided problem-based learning (PBL) approach: Impact on critical thinking. Decision Sciences Journal of Innovative Education, 12(2), 91-108.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, 100041.

Peer, E., Rothschild, D., Gordon, A., & Damer, E. (2022). Erratum to Peer et al.(2021) Data quality of platforms and panels for online behavioral research. Behavior Research Methods, 54(5), 2618-2620.

Rawte, V., Sheth, A., & Das, A. (2023). A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922.

Razeghi, Y., Logan IV, R. L., Gardner, M., & Singh, S. (2022). Impact of pretraining term frequencies on few-shot reasoning. arXiv preprint arXiv:2202.07206.

Robinson, N. R., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023). ChatGPT MT: Competitive for high-(but not low-) resource languages. arXiv preprint arXiv:2309.07423.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., ... & Synnaeve, G. (2023). Code llama: Open foundation models for code. arXiv preprint arXiv:2308.12950.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. Journal of applied learning and teaching, 6(1), 342-363.

Sallam, M. (2023, March). ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In Healthcare (Vol. 11, No. 6, p. 887). MDPI.

Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018, June). " Hey Alexa, What's Up?" A Mixed-Methods Studies of In-Home Conversational Agent Usage. In Proceedings of the 2018 designing interactive systems conference (pp. 857-868).

Shan, G., & Qiu, L. (2023). Examining the Impact of Generative AI on Users' Voluntary Knowledge Contribution: Evidence from A Natural Experiment on Stack Overflow. Available at SSRN 4462976.

Shi, S., Zhao, E., Tang, D., Wang, Y., Li, P., Bi, W., ... & Ma, D. (2022). Effidit: Your ai writing assistant. arXiv preprint arXiv:2208.01815.

Smits, J., & Borghuis, T. (2022). Generative AI and Intellectual Property Rights. In Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice (pp. 323-344). The Hague: TMC Asser Press.

Tate, T. P., Steiss, J., Bailey, D. H., Graham, S., Ritchie, D., Tseng, Warschauer, M. (2023, December 5). Can AI Provide Useful Holistic Essay Scoring?. https://doi.org/10.31219/osf.io/7xpre

Touretzky, D. S., & Gardner-McCune, C. (2022). Artificial intelligence thinking in K–12.

Valmeekam, K., Olmo, A., Sreedharan, S., & Kambhampati, S. (2022). Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change). arXiv preprint arXiv:2206.10498.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Wang, Z. M., Peng, Z., Que, H., Liu, J., Zhou, W., Wu, Y., ... & Peng, J. (2023). Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. arXiv preprint arXiv:2310.00746.

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for writers of english as a second or foreign language. Journal of Second Language Writing, 62.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., ... & Gabriel, I. (2022, June). Taxonomy of risks posed by language models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 214-229).

Williams, R., Ali, S., Devasia, N., DiPaola, D., Hong, J., Kaputsos, S. P., ... & Breazeal, C. (2023). AI+ ethics curricula for middle school youth: Lessons learned from three project-based curricula. International Journal of Artificial Intelligence in Education, 33(2), 325-383.

Williams, R., Machado, C. V., Druga, S., Breazeal, C., & Maes, P. (2018, June). " My doll says it's ok" a study of children's conformity to a talking doll. In Proceedings of the 17th ACM Conference on Interaction Design and Children (pp. 625-631).

Xu, Y., & Warschauer, M. (2020, April). What are you talking to?: Understanding children's perceptions of conversational agents. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-13).

Yau, K. W., Chai, C. S., Chiu, T. K., Meng, H., King, I., Wong, S. W. H., ... & Yam, Y. (2022, December). Developing an AI literacy test for junior secondary students: The first stage. In 2022 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE) (pp. 59-64). IEEE.

Yong, Z. X., Menghini, C., & Bach, S. H. (2023). Low-resource languages jailbreak gpt-4. arXiv preprint arXiv:2310.02446.