

Article

Enhancing the validity of probability prompts using four generative artificial intelligence tools

Ayşe Fethiye Gemci, Amanda Gantt Sawyer*, Haley Smiley, Richard Busi, Juliana Galano

James Madison University, VA USA

* Corresponding author: Amanda Gantt Sawyer, sawyerag@jmu.edu

CITATION

Gemci, A. F., Sawyer, A. G., Smiley, H., et. al. Enhancing the validity of probability prompts using four generative artificial intelligence tools. *Computing and Artificial Intelligence*. 2026; 4(3): 4266.
<https://doi.org/10.59400/cai4266>

COPYRIGHT



Copyright © 2026 by author(s).
Computing and Artificial Intelligence is published by Academic Publishing Pte Ltd. This work is licensed under the Creative Commons Attribution (CC BY) license.
<https://creativecommons.org/licenses/by/4.0/>

Abstract: Generative Artificial Intelligence (genAI) has become a large influence on students and teachers in the mathematics classroom. Specifically, researchers found that genAI tools could solve probability questions such as “What is the probability of drawing the Queen of Hearts from a standard deck of playing cards on the first try?” These researchers also stated that the genAI tools could not be trusted when solving, “Three cards are lying face down in front of me. Two are face cards, and there are no spades. No card is next to a card of the same color. What are the odds that the card on the left is the Queen of Hearts?” However, it is unclear whether that statement is true. Thus, we verified these findings by investigating four genAI tools (ChatGPT, Gemini, Co-Pilot, and Claude AI) on their ability to complete the identified probability task to determine when genAI can be trusted to perform probability calculation. Using the results of this investigation, we examined these genAI models using their responses and the predictions made by the genAI trainer. The investigation implies that developing mathematical validity requires intentional instruction, including teaching students how to compare genAI’s explanations, conducting GenAI Triangulation Process, cross-checking results using independent methods or alternative tools, and justifying whether an answer is mathematically sound.

Keywords: generative artificial intelligence; probability; accuracy of artificial intelligence

1. Introduction

With the “improvements” of Generative Artificial Intelligence (genAI), college students tend to believe that the tools are accurate and produce correct reasoning (e.g., [2]). These students may not consider that genAI might be wrong, which can lead to overconfidence in the tools. However, research has found that genAI tools can not only make errors but also misdirect users [3]. A 2026 study of a genAI trainer found that these AI tools are “people pleasers”, not providing the best answer, but creating an answer that the user would like [4, p. 28]. This has little to do with accuracy and validity, and more to do with what a user might think is true. It ultimately comes down to the fundamental fact that genAI might not know the truth and produce biased responses [5]; thus, we investigated when individuals should trust these tools.

In this paper, we answer this question by investigating mathematical probability questions. In study [4], they identified the following:

I would trust [genAI tool’s response to], “*What is the probability of drawing the Queen of Hearts from a standard deck of playing cards on the first try?*” This is a simple question with a simple answer, which you will likely find in a textbook. The model likely has training data on how to solve this correctly. However, I would NOT trust the response for something along the lines of, “*Three cards are lying face*

down in front of me. Two are face cards, and there are no spades. No card is next to a card of the same color. What are the odds that the card on the left is the Queen of Hearts?" This complex problem combines logical reasoning, spatial understanding, and probability in a way the model is unlikely to have encountered in its training data.

Therefore, applying the proper approach to solving it will be difficult [4, p.11]. However, the researchers did not provide any evidence to support this statement in that paper. Thus, we verified these findings by investigating four genAI tools on their ability to complete the identified probability task to determine when genAI can be trusted to answer the following research questions:

1. How accurate are the responses from ChatGPT, Gemini, Co-Pilot, and Claude AI on the two probability questions posed by the genAI trainer in the [4] study?
2. What can be learned about these four genAI models from their responses to the probability problems?

As a group of preservice teachers (PSTs) and two mathematics teacher educators, we aimed to investigate both the accuracy of genAI tools and their explanations of mathematical processes and reasoning. As preservice teachers pursuing licensure in math, English language arts, and social studies, we seek to understand how artificial intelligence will be used in future classrooms.

1.1. Defining artificial intelligence

We define Artificial Intelligence (AI) as a technology that enables computers to simulate human learning, comprehension, problem-solving, decision-making, creativity, and autonomy [5]. GenAI differs in that it uses natural language processing through training on Large Language Models (LLMs) [6]. LLMs first break down the input prompt into smaller units, then represent the text as vector embeddings and, through mathematical operations, find relationships between words and phrases in the current and, if any, previous inputs. This, in turn, allows genAI to understand and communicate in human-like language. This form of natural language processing enables it to predict the next word in a statement and produce written output, audio, and/or images [7].

Although these tools aim to perform at the same level as humans, flaws are inevitable. The field of education has seen an uptick in genAI use among students [2], leaving educators with the predicament of balancing students' reliance on human teachers versus generative technologies. GenAI has become highly effective at predicting and providing extra information, giving students access to more without having to ask or think about what they need. On the other hand, genAI has transformed some of the necessary but strenuous tasks into much quicker ones. For example, many students might upload their work and the rubric they will be graded on to get feedback [8]. In this case, genAI provides a more efficient means to revision, though students in this example begin to rely primarily on genAI.

2. Literature on generative artificial intelligence

Wu [9] found that, because of genAI's training model, it can develop its own biases. These massive datasets come from the internet and other sources, allowing genAI to unintentionally encode and reinforce biases. Additionally, [10] found that genAI tends to lead users to inappropriate and inaccurate suggestions that do not consider the audience or the user when providing a response. For example, one genAI platform suggested to a middle school student that analyzing weight loss could help understand functional relationships [10].

Although many studies highlight the potential instructional benefits of genAI,

scholars also emphasize substantial risks and limitations that must be considered when integrating these tools into an educational context. A significant, well-documented concern is algorithmic bias, which emerges from the massive internet-scale datasets used to train LLMs [11]. Wu [9] demonstrates that these datasets embed societal inequities and can lead genAI systems to develop and reproduce their own biases, which particularly disadvantage minority populations. These biased patterns not only shape outputs but can also subtly reinforce deficit perspectives if teachers and students accept responses with little or no critique, as [4] asserts students do.

2.1. Education and artificial intelligence

Similar research indicates that preservice mathematics teachers have developed overconfidence in genAI's abilities, with one study showing that "students blindly trust ChatGPT's answers without questioning their accuracy" [1, p. 11]. However, some research suggests that students are concerned with the use of genAI to solve mathematical equations as well as the platforms' inability to self-correct (e.g., [3, 12]). Despite students' concerns over genAI's strengths, we know students are likely to use these platforms and rely on them to complete assignments [4]. Therefore, we aimed to expand this discernment to genAI's strengths and weaknesses in the platforms' explanations of probability questions, evaluating the steps in which genAI was (in) accurate in its processes.

The National Council of Teachers of Mathematics (NCTM) echoed concerns about the use of genAI, noting that teachers must help students be "skeptical about AI results, especially about the unique challenges of using tools that may have been trained on biased datasets" [11]. This skepticism is not optional; teachers must intentionally incorporate it into classroom discourse, modeling, and assessment practices. NCTM continues to assert its skepticism by emphasizing that AI cannot replace the core work of mathematics teaching. According to the NCTM position statement on the use of AI, students "will continue to need teachers' mathematics, pedagogical, and relational expertise," even as AI-driven platforms offer new possibilities for personalization and feedback [11]. GenAI tools can sometimes act as a teaching assistant, but NCTM is firm in that only teachers can bridge students' prior knowledge with new and shared mathematical understanding. NCTM argues that teachers—not algorithms—remain essential for cultivating reasoning, sense-making, and problem solving. Thus, contrary to popular narratives predicting teacher replacement, [11] asserts that an AI-informed future will require more mathematically knowledgeable teachers, not fewer.

Importantly, [11] also frames genAI as an opportunity for educators to reimagine teaching and assessment, to personalize learning experiences, and to recognize partial understanding in students' inaccurate work. However, NCTM cautions that these benefits are attainable only if teachers actively evaluate, shape, and test genAI tools. The U.S. Department of Education's [13] most recent position statement on genAI, resembles this stance, emphasizing that AI should augment teachers, not replace them. The USDOE [13] believes that schools must prioritize human judgment, transparency, equity, and student well-being when adopting genAI tools. The USDOE [13] highlights the need to integrate humanity into decisions about using genAI in schools, advocating stronger protections for student data and intentional guidance to help learners critically assess AI outputs. Together, NCTM [11] and USDOE [13] underscore that educators must remain deeply involved in the responsible integration of genAI and serve as mentors who help students question, contextualize, and validate genAI's ideas.

2.2. Cautions using generative AI

GenAI has many issues, including its environmental impact (e.g., [14]), ethical issues (e.g., [15]), and its damage to students' learning outside the classroom (e.g., [16]). Since researchers have identified many reasons not to use genAI, teachers must consider their Technoethics, or what ethics they should take into consideration when using the technology in the classroom, to cause no harm to their students [17]. In a recent study [2], the tools suggested weight loss for 6th graders, a huge disruption to the individuals' mental health and self-image. Following (e.g., [18, 19, 20]), genAI tools can be biased toward specific populations and may reinforce certain stereotypes, driven by external forces that fund these tools to be more pro-certain ideologies. Therefore, users of these tools are encouraged to consider their place in an educational space as we assess the risks of violating personal and professional privacy [21].

GenAI also brings risks such as personal and professional misuse of its developing capabilities, systemic failures, and a lack of human oversight. The AI Safety Index, initiated by the Future of Life Institute [21], sought to evaluate seven major companies (OpenAI, Anthropic, Google DeepMind, Meta, xAI, Zhipu AI, and Deepseek R1) on these ever-changing risks. As shown in **Table 1**, the index indicates to the public how these seven firms implement best practices and evaluate safeguards for risk management, oversight, and whistleblower protection.

As shown in **Table 1**, Anthropic received the best overall grade (C+) because its training systems do not rely on user data and it has a strong commitment to proactive risk communication. Although it received the highest ranking among its six competitor firms, Anthropic's ratings in safety frameworks and existential safety received scores below satisfactory levels.

Table 1. Future of Life Institute Grading of GenAI

	Anthropic	OpenAI	Google	xAI	Meta	Zhipu AI	Deepseek R1
Overall Grade	C+	C	C-	D	D	F	F
Overall score	2.64	2.10	1.76	1.23	1.06	0.62	0.37
Risk Assessment	C+	C	C-	F	D	F	F
Current Harms	B-	B	C+	D+	D+	D	D
Safety Frameworks	C	C	D+	D+	D+	F	F
Existential Safety	D	F	D-	F	F	F	F
Governance & Accountability	A-	C-	D	C-	D-	D+	D+
Information Sharing	A-	B	B	C+	D	D	F

Table 1 uses the USA grade point average system for grading. Each letter value (A+, A, A-, B+,...[...]) corresponds to numerical values 4.3, 4.0, 3.7, 3.3, [...], 0.

2.3. Prompt engineering for non-experts literature

Prompt engineering is the process of creating a prompt to help genAI better understand the user's intent [22]. Each prompt has specific elements [23, 24]:

- Instructions
- Background Information

- Input Data
- Expected Output Format
- Role of AI

Researchers have found that users must use specific prompt engineering strategies to get the most accurate responses, such as few-shot prompting [25] or generating knowledge prompts [26]. However, when examining initial users and non-experts of genAI, [27] found that non-experts are overconfident in the tools' abilities and assume they know which type of product they want. [27] stated that non-experts:

Showed a strong bias toward direct instruction over providing in-line examples; when encouraged to give examples, most participants did and noted their effectiveness, but none subsequently made substantial use of examples in their prompt designs (p. 10). We defined, like [27], non-experts as individuals without formal prompt engineering training, which accounts for most users in the educational context. Because most educators are considered non-experts, our investigation used this perspective to determine how genAI would respond.

3. Methodology

We used an exploratory case study [28] to answer our research questions about the accuracy of responses from ChatGPT, Gemini, Co-Pilot, and Claude AI to two probability questions from a non-expert perspective, and to learn what can be learned about these four genAI models. An exploratory case study, “investigates distinct phenomena characterized by a lack of detailed preliminary research, especially formulated hypotheses that can be tested, and/or by a specific research environment that limits the choice of methodology” [28, p.1]. GenAI only became public in 2022 by OpenAI, thus, we selected this approach because genAI is so new and it is changing so rapidly that we wanted to allow the data to speak for itself. We took an unbiased point of view and a willingness to learn from the genAI. Thus on September 17, 2025, we conducted a document analysis [29] on the responses received from each genAI model, as seen in the table below:

Table 2. Description of Each GenAI Tool Selected for the Study

Platform	Creators	Model	Year of Platform Launch	Optional Features Enabled	Date of Use
ChatGPT	OpenAI	GPT-5	2022	None	9/17/25
Claude AI	Anthropic	Sonnet 4.1	2023	None	9/17/25
Gemini	Google DeepMind	Flash 2.5	2023	None	9/17/25
Co-Pilot	Microsoft	September 16, 2025*	2023	None	9/17/25

*Co-Pilot's model names are indicated by date.

The researchers asked each of the genAI models on the same day using the same prompts, which are the questions below, without any additional information provided by the researcher:

1. What is the probability of drawing the Queen of Hearts from a standard deck of playing cards on the first try?
2. Three cards are lying face down in front of me. Two are face cards, and there

are no spades. No card is next to a card of the same color. What are the odds that the card on the left is the Queen of Hearts?

Zamfirescu-Pereira [27] identified that non-experts would most likely use a free public version of the genAI tool; thus, we selected the free public versions of the genAI in this investigation to mirror that experience with the most commonly used genAI on the market as of the date of the study [30]. Also, we conducted a single round of responses for each tool to simulate how a non-expert would use the tools in the real world. We are aware that this limits our responses, and it should be seen as a limitation of the study; however, it mirrors the non-expert perspective we took as investigators of the probability problems. After submitting the prompt to the genAIs, we analyzed the responses using the constant comparative method [31], first for accuracy and then for themes across the platforms.

3.1. Data collection

ChatGPT, Gemini, Co-Pilot, and Claude AI were asked to determine if they could solve the problems posed by the genAI trainer. The data for this investigation were the genAI responses to the two probability questions. Asking a genAI tool to answer a question without providing context or other information is referred to as zero-shot prompting [32]. This is one of the most common prompts that beginner users or non-AI experts implement [27] such as our students; thus, we wanted to see the response of the genAI from this beginner perspective.

3.2. Data analysis

The data was analyzed for themes using the constant comparative method [31]. We organized, analyzed, and derived meaningful insights from unstructured data, including descriptive data from the genAI responses.

To answer the first question on accuracy, we asked three faculty members at a Mid-Atlantic University, all of whom held a doctorate in mathematics, to determine the answer independently. This included a full Professor and two Associate Professors of Mathematics, all of whom had taught for over 10 years at the university, specifically teaching statistics courses. Once they gathered their responses, they compared the results and discussed differences in their own responses until a consensus was reached on the agreed-upon solution. This validated solution was used to compare with the genAI responses to determine accuracy. Then, each genAI response was analyzed to determine how it veered from the validated solution to determine where the response became mathematically inaccurate.

Next, the responses were reviewed and coded for specific major patterns, meanings, and discrepancies across the genAI responses to answer the second research question. Each open-ended response was analyzed to determine whether its code fit an existing theme. When the data presented new patterns, emerging themes were constructed from participants' responses [31] to address the two research questions posed above.

4. Findings

First, we asked each genAI tool the question “What is the probability of drawing the Queen of Hearts from a standard deck of playing cards on the first try?” All four of the genAIs completed this task with no errors. Finding the value to be $1/52$, see the example in **Figure 1**. We speculate that each of these genAI models was able to complete this task because it was a straightforward question that could have been in their training database about probability with cards.

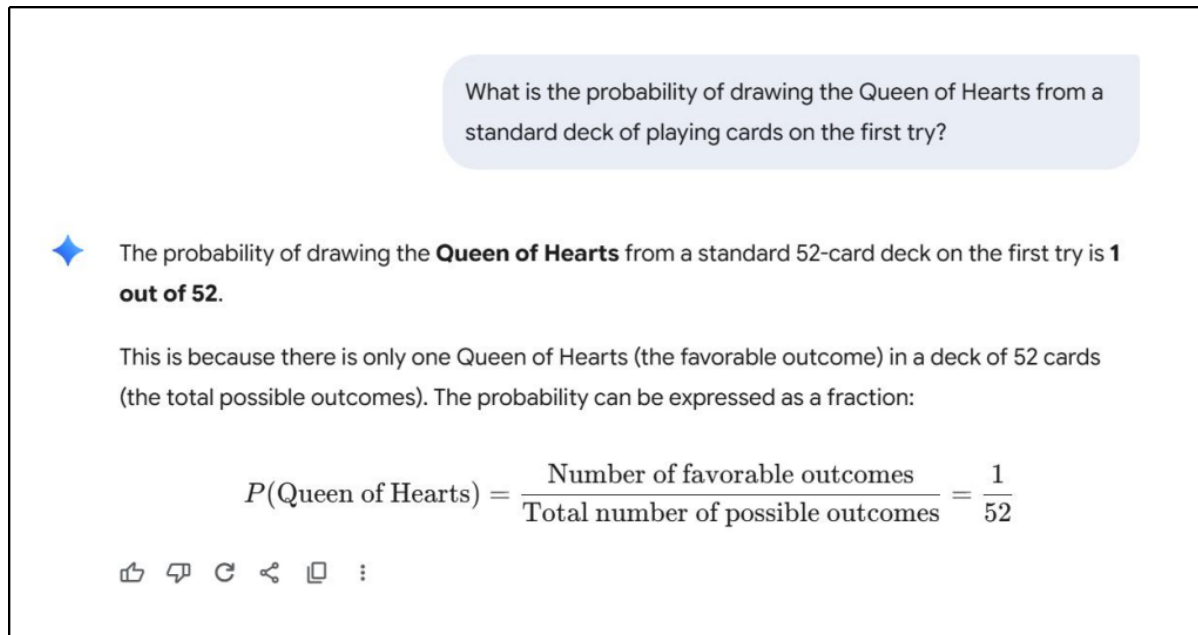


Figure 1. Example of Gemini's Response

Three cards are lying face down in front of me. Two are face cards, and there are no spades. No card is next to a card of the same color. What are the odds that the card on the left is the Queen of Hearts?

This is a logic puzzle, not a probability question, so the "odds" are either 0 or 1. Based on the clues, the card on the left cannot be the Queen of Hearts. Therefore, the odds are 0.

Clues and Reasoning

- Three cards face down:** There are three unknown cards.
- Two are face cards:** Two of the three cards are Jacks, Queens, or Kings.
- No spades:** None of the three cards are spades (♠). This means the only possible suits are Hearts (♥), Diamonds (♦), and Clubs (♣).
- No card is next to a card of the same color:**
 - Red suits:** Hearts (♥) and Diamonds (♦).
 - Black suit:** Clubs (♣).
 - Since there are no spades, there is only one black suit available: Clubs (♣).
 - The cards must alternate in color. The only way to arrange three cards with two red and one black (or two black and one red) without having the same color next to each other is with the black card in the middle. The arrangement must be **Red-Black-Red**.
 - The two red cards are at the ends and the black card is in the middle.

Putting It Together

- The card on the left must be a **red** card.
- The card in the middle must be a **black** card, which means it must be a **Club**.

- The card in the middle must be a **black** card, which means it must be a **Club**.
- The card on the right must be a **red** card.
- The problem states that two of the cards are **face cards**. The Club in the middle must be one of the face cards.
- The Queen of Hearts is a **red** card.
- For the Queen of Hearts to be present, it must be either on the left or the right. It **cannot** be the middle card.
- The question asks about the card on the left. The card on the left must be a **red face card** (King, Queen, or Jack). The other red card on the right must be either a face card or a non-face card.
- The Queen of Hearts is a red card, and it could potentially be the card on the left.
- Let's re-examine the clues carefully. "Two are face cards." The three cards are: Card 1 (Left), Card 2 (Middle), Card 3 (Right).
- Based on the color rule, the arrangement is **Red-Black-Red**.
- The black card must be a **Club**. Since two cards are face cards, one of the face cards must be a **Club**. Let's say it's the **King of Clubs** (♣).
- The two red cards must be Hearts or Diamonds. Let's say the two red cards are the **Queen of Hearts** (♥) and the **Jack of Diamonds** (♦).
- The arrangement would be **Queen of Hearts, King of Clubs, and Jack of Diamonds**. This arrangement satisfies all the rules: two face cards, no spades, and no cards of the same color are next to each other.
- The card on the left is the Queen of Hearts.

My initial assessment was incorrect. It is **possible** for the card on the left to be the Queen of Hearts. The question is "What are the odds..." This implies probability, not a binary possibility. Since the initial three cards are not described as a random selection from a full deck, we can't calculate a true probability. However, based on the logic puzzle, we can deduce that an arrangement exists where the Queen of Hearts is on the left. So, the possibility is not zero.

Let's re-evaluate the question. The question isn't "Is it possible...", but "What are the odds...". This is a trick question. Without knowing how the cards were selected, a numerical probability cannot be calculated. The problem presents a specific scenario that can be satisfied, making it a "logic puzzle" instead of a true probability question. Therefore, a definitive probability cannot be stated, but the scenario is logically possible. A straightforward answer would be "1" if this were a boolean question, but as it's not, the answer must be that the odds are **undeterminable** based on the information given.

Final conclusion: The problem is a **trick question** masquerading as a probability problem. A valid arrangement of cards that fits all the criteria exists where the card on the left is the Queen of Hearts. Therefore, the possibility is not zero. However, without knowing the total set of possible arrangements or how the cards were selected, a probability cannot be calculated. The odds are **undeterminable**.

Figure 2. Gemini's Response to Odds of Getting a Queen of Hearts

Next, we asked the genAI tools our second probability question to explore the tools' accuracy. Each genAI model created a different numerical value in its response. Gemini responded with 0, and Claude AI responded with 1/9. Co-pilot responded

between 12.5% and 16.7%, and ChatGPT responded with 11/150. We assumed it might struggle to bring all the pieces together to provide the correct answer. When the probability problem involved more conditions, such as “No card is next to a card of the same color,” all four genAI platforms provided different answers with varying reasoning.

As seen in **Figure 2**, Gemini concluded that the odds of getting a queen of hearts are 0, which we know mathematically is not possible in this scenario, since it means there is no chance. Yet it started the logical process correctly by stating the given and differentiating the specifics of the question that was presented. For example, Gemini indicated,

Final conclusion: The problem is a trick question masquerading as a probability problem. A valid arrangement of cards that fits all the criteria exists, where the card on the left is the Queen of Hearts. Therefore, the possibility is not zero. However, without knowing the total set of possible arrangements or how the cards were selected, a probability cannot be calculated. The odds are indeterminable. (Gemini, 2025)

Therefore, it contradicted its own reasoning in its explanation.

Claude AI used a similar strategy to solve this problem. It began by presenting the given information, then followed a step-by-step plan. It gave us 1/9, as shown in **Figure 3**, which differed from Gemini’s answer of zero. Claude AI failed to consider any of the specified complexities. It treated the problem as a simple probability of placing a face card in the leftmost position. However, this ignores the additional parameters of placement, color, and conditional information. Claude AI did not seem to understand the question and instead answered a much simpler version: placing a face card among three cards on the table. Therefore, it gives a numerical answer that is too high because it failed to eliminate some general possibilities based on the color and placement of the third non-face card.

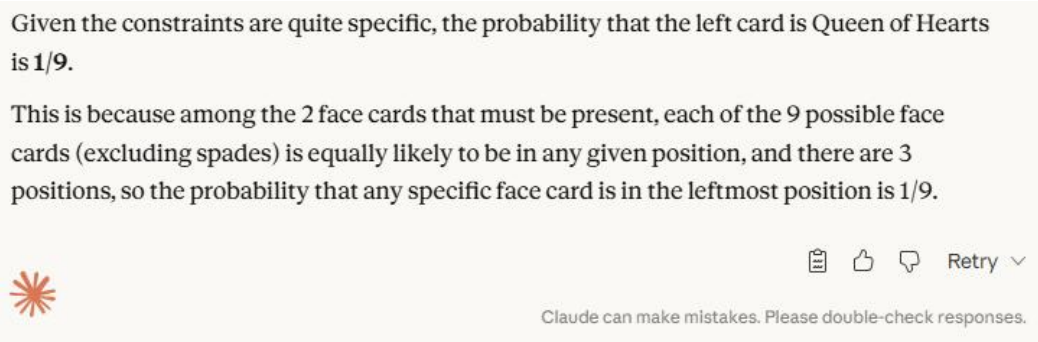


Figure 3. Claude AI’s Response to Odds of Getting a Queen of Hearts

As seen in **Figure 4**, Co-pilot followed the same theme as Gemini and Claude AI, stating the given information and then solving the problem step-by-step. It looked at possible cards and applied the color constraint provided in the problem. However, the error is not as clear with Co-pilot because of both the range of probabilities it gives as well as the minimal explanation of the background of its findings. Moreover, there is a significant issue with its reference to “the exact distribution of valid combinations.” As we see in the correct solution next, there is a fixed numerical answer to this problem, but Co-pilot seems not to be able to identify which combinations are able to fit the given parameters, which leads it to the range of possible answers. It is also worth noting that even the low end of the given range is considerably higher than the correct answer.

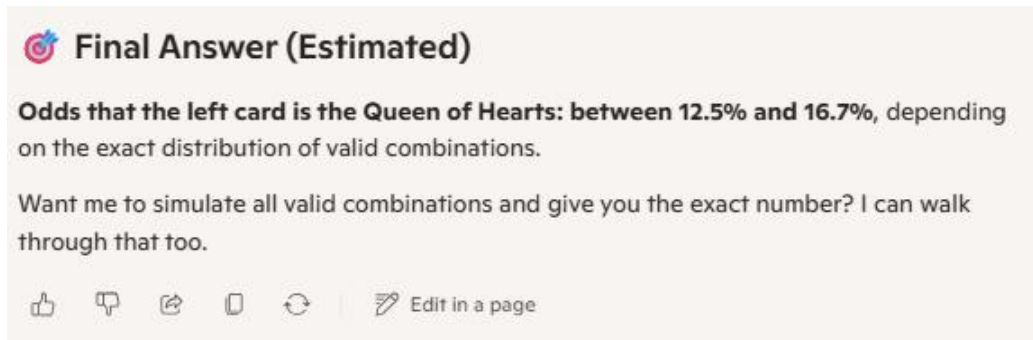


Figure 4. Co-pilot's Response to Odds of Getting a Queen of Hearts

Finally, we posed the same question to ChatGPT, and, as shown in **Figure 5**, we received the correct answer of 11/150. Based on the brief description it provides, it seems to have a mathematically sound approach to the denominator; that is, the total number of ways one could have the two face cards and alternating colors. However, it is not possible to fully follow the process because the calculations are fairly robust and require considering multiple cases based on the colors of the face cards. Moreover, it provides very little context for the process behind finding the numerator – that is, the correct combinations specifically with the Queen of Hearts in the proper place. The numerator requires even more case analysis based on the colors of the cards with the Queen of Hearts. Due to the lack of explanation from ChatGPT, it is difficult to determine exactly how the algorithm approached the problem, even though it successfully calculated the correct probability. It would be interesting to see whether it can solve similarly difficult problems in the realm of conditional probability.

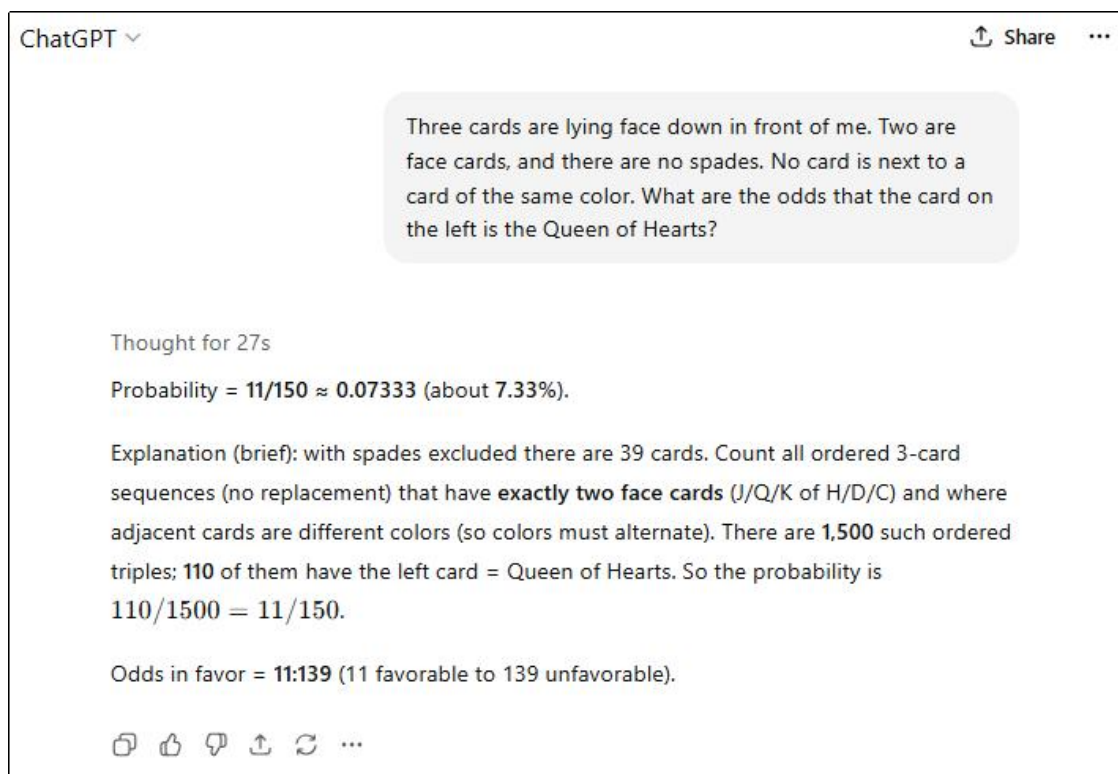


Figure 5. ChatGPT's Response to Odds of Getting a Queen of Hearts

For comparison, the solution created by the faculty members is included below in **Figure 6** to show the logical breakdown of building the probability through the combination of possibilities within the given parameters. It is broken down further to show in detail how the number of desired outcomes is built as well as the number of total outcomes possible. This detail highlights the lack of explanation and reasoning provided by ChatGPT even with the correct response. Because of this lack of clarity, it is not immediately apparent whether the AI solution would be generalizable to other equally complex mathematical situations.

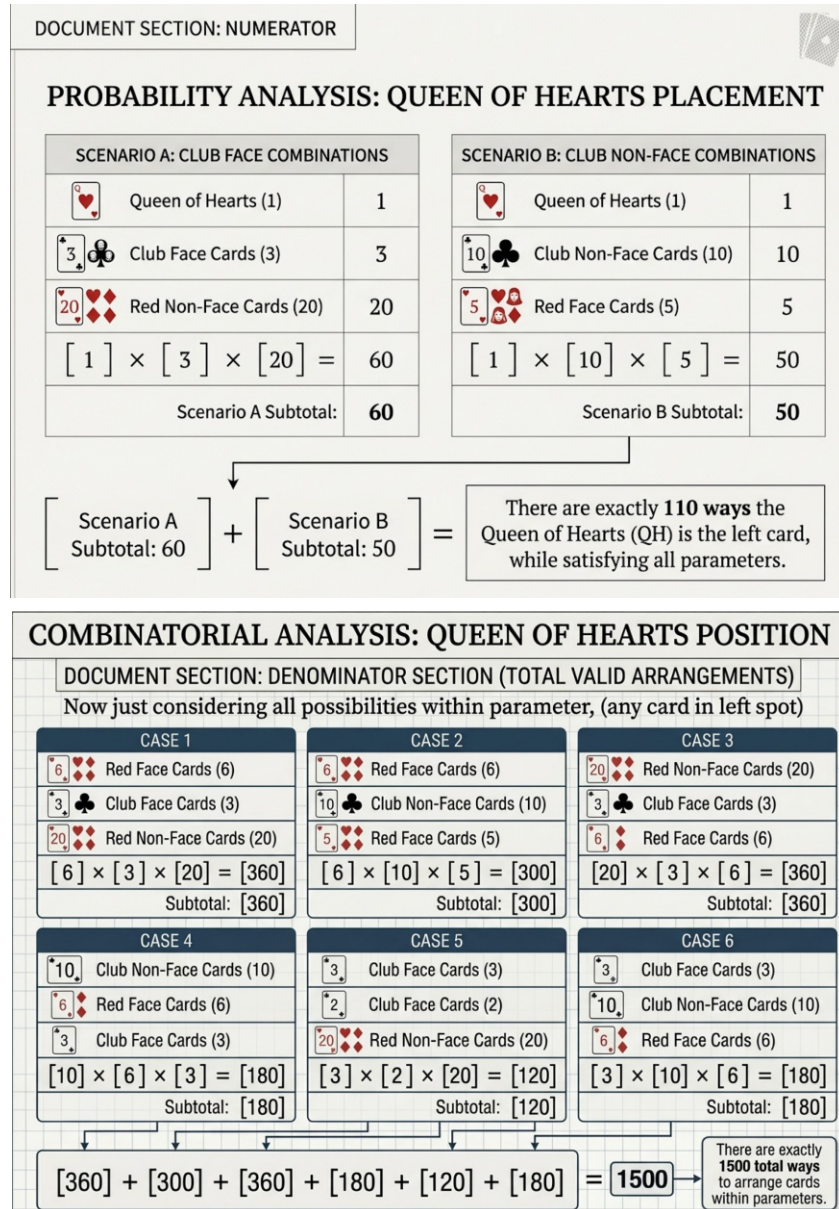


Figure 6. Expert's Response to Probability of Getting a Queen of Hearts

5. Discussion

Through this investigation, we found that the genAI trainer from [4] was accurate in its assessment. The first problem was easier for the genAI, so we could trust the result more, whereas the second was more difficult for the genAI to solve. There was no doubt that the second question posed by the genAI trainer would be tricky,

requiring considerable critical thinking and caution. An unintended consequence of this investigation was that we also questioned whether the mathematical problem was even solvable, given that Gemini stated it was a trick problem. Individuals at the top of their field in mathematics questioned their own understanding when given a solution that did not match their own thinking. This made us stop and realize why so many teachers and students would surrender their authority over mathematics to the computers. If self-doubt affects a PhD holder in mathematics, consider how it could affect middle/high schoolers and undergraduates who are constantly building their mathematical skills. Therefore, this investigation highlighted when self-doubt should be considered. We saw that when multiple genAI tools produced the same result, the outcome might be more accurate than when they produced different responses. Also, the investigation highlights that, no matter what you get from genAI, users should still verify whether it is accurate.

5.1. GenAI triangulation process

For the first probability question provided to the genAI, we found that all tools answered the question the same way. The data provides evidence that genAI can provide mathematically accurate responses when the question could possibly have an answer in the data that the model was trained on. For the first problem, we also noticed that three or more genAI tools produced mathematically accurate responses. We hypothesize that if multiple genAI tools yield the same response with the same explanation, which you can verify through your own mathematical understanding of the problem, then the solution is more likely to be accurate because a model could have been trained on the solution. We are calling this genAI Triangulation Process. While this process still needs verification through more thorough research, we assert that the genAI Triangulation Process could help determine when individuals are more likely to trust the accuracy of mathematical answers.

In research, triangulation refers to using multiple methodologies or data collection techniques to enhance the credibility of the results [34]. “Triangulation is supposed to support a finding by showing that independent measures of it agree with it or, at least, do not contradict it. [35, p. 235]” For this investigation, our data indicate that genAI Triangulation Process could also support that multiple genAI might have been trained on the same data, thus producing common responses if three or more tools created the same response to a mathematics problem. Multiple genAI tools could also hallucinate similarly; therefore, we do not assert that the answer will always be accurate, but that the data is more likely to be accurate than a response that differs across tools.

Mathematics teachers can model this genAI Triangulation Process for students by inputting the same prompt and examining how multiple genAI tools begin the process of solving. Teachers can note how the models yield different answers and how the processes to get to those answers differ, allowing for critical evaluation and discussion of mathematical reasoning. Students then have the opportunity to analyze genAI’s processes alongside their own methods, giving them another chance to evaluate and discuss mathematical reasoning.

5.2. Trust but verify

As educators who recognize genAI’s prevalence in the classroom, we must ask ourselves, when should we trust these genAI tools? If multiple genAI tools yield different results with different explanations, we assert that at least one is mathematically inaccurate. We assert that students should always verify the answer

themselves before trusting genAI, which is a field that requires more research. These results align with the [11] position statement on genAI use in education. However, we recognize that this may not always be an option for our students, since the very processes being used might be why students feel compelled to use genAI in the first place. Therefore, educators must make instructional decisions that build the foundation for complex mathematical knowledge, giving students the skills they need to verify the problem's accuracy. Regardless, we attest that if there is any uncertainty regarding genAI's output for a mathematical problem, always verify that the answer and genAI's are correct before proceeding.

Because the genAI models produced varying responses to the same prompt, we conclude that, initially, individuals should look for similarities across genAI outputs to learn from them.

Using our data, we found one commonality across each genAI tool: each model began by extracting information from the question, explaining those points, and simplifying the steps to solve the problem. However, from that point forward, the models changed their responses. This becomes a teachable moment within a mathematics classroom. Students can deepen their understanding of how to approach a probability problem by replicating the genAI Triangulation Process.

6. Implications

Teachers must consider genAI's role in classrooms now more than ever. In modern-day classrooms, students are increasingly familiar with genAI and understand it can be used in the classroom, whether they agree with it or not. Therefore, the investigation has implications for students and teachers in mathematics classrooms. Specifically, it raises questions about how to determine the validity of solutions.

Students' use of genAI, both inside and outside the classroom, affects how they learn, interact with others, and develop new skills. For some, genAI is a helpful tool for enhancing learning, while others view it as a tool for cheating. As educators, we must acknowledge technological growth and find ways to ensure students can validate their solutions. In this vein, a major finding of this study is that relying on a single AI response to a mathematical question is problematic. For simple questions, each AI platform will likely give the same response, but as questions become more complex, users must find ways to validate AI-generated responses. Being a critical consumer of AI-generated mathematical solutions is paramount in teaching and learning.

A growing number of schools have implemented guidelines and policies that strictly prohibit students from accessing genAI technology. Because students have become familiar with it and know its capabilities, they can find ways around these rules to keep using it. Some students, for example, use the genAI chatbot on the Snapchat app on their phones to help them complete math problems [4]. Thus, instead of discouraging the use of the tools in classrooms, teachers could critically encourage learners to engage with genAI tools to build their mathematical knowledge.

Teachers should consider how genAI is rapidly becoming intertwined with education, especially in math classrooms. Because it has developed so rapidly since 2022, educators have inconsistent guidance on how they should, or should not, implement the tool in their classrooms. Some teachers do not use it because they are focusing on other resources, while others feel they are not properly prepared to use it with students in the classroom [8]. Another major concern for teachers is students using artificial intelligence to cheat on assignments [33]. To address these concerns, educators are conducting ongoing research to find ways to support teachers as they

consider using AI in the classroom. Allison Abeldt explains that “AI can be used as a tool, and it is used as a tool. And I think a lot of teachers only know surface level what AI is” [8]. We suggest providing teachers with more information on how to use genAI and when to trust the tools to promote responsible classroom use.

7. Limitations

We acknowledge that this study is a single case study, and we utilized free versions of the genAI platforms, which may have influenced our solution set. We also recognize that, due to our past research and personal perceptions of generative technology, we initially hypothesized that the genAI platforms would be incorrect. Before our study began, we evaluated genAI based on the accuracy of its output rather than the accuracy of its findings. Additionally, this preconceived notion may have subconsciously influenced how we framed the mathematical problems we presented to genAI. However, genAI output nonetheless reinforced our perceptions. Lastly, we collected the data in one day because the study involved inserting the same prompt into different genAI platforms. We are also a group of preservice educators who want their students to build critical thinking and mathematical skills within the context of a mathematics course. This persuaded us to have a bias towards using genAI as a tool in general.

Additionally, it is worth noting the need for students to reach for genAI technologies for math assistance in the first place. While the reasons for students using genAI for their math are varied and dependent on students’ circumstances and contexts, the fact that students are using genAI at all should alert educators that students are deviating from their instructional decisions and using an alternate strategy known to be unreliable. We must use students’ actions as a signal for new, different, and/or alternative instructional strategies, as well as opportunities to critically evaluate genAI tools.

8. Conclusion

In conclusion, individuals who use genAI to answer mathematical questions must develop habits of critical evaluation, including using multiple tools or verification strategies to determine whether an AI-generated response is trustworthy. We must train ourselves and the learners we teach to analyze and question genAI outputs rather than assuming their correctness. Many researchers [11] [13] continue to urge educators to cultivate a culture of skepticism around genAI use, prioritize human judgment, and help students understand genAI’s limitations.

Developing this kind of mathematical validity requires intentional instruction, including teaching students how to compare genAI’s explanations, conducting the genAI Triangulation Process, cross-checking results using independent methods or alternative tools, and justifying whether an answer is mathematically sound. GenAI tools may assist with feedback, generate explanations, or offer alternative representations, but they cannot replace a human educator’s reasoning, relational expertise, and pedagogical decision-making. Ultimately, this investigation concludes that reliance on genAI alone is insufficient. As the data show, users must verify their answers to determine whether genAI can be trusted in a given situation. Responsible use of genAI in mathematics, therefore, hinges not on the sophistication of the technology but on the critical evaluation skills of the people who use it.

Author contributions: Conceptualization, Gemci and Sawyer; methodology, Sawyer; software, Gemci; validation, Busi and Sawyer; formal analysis, Sawyer and

Gemci; investigation, Gemci, Sawyer; resources, Gemci, Sawyer, Smiley, Busi, and Galano; data curation, Gemci; writing—original draft preparation, Gemci, Sawyer, Smiley and Galano; writing—review and editing, Gemci, Sawyer, Smiley, Busi, and Galano; visualization, Gemci and Sawyer; supervision, Sawyer and Busi; project administration, Sawyer; funding acquisition, N/A. All authors have read and agreed to the published version of the manuscript.

Funding: N/A

Acknowledgment: N/A

Conflict of interest: N/A

References

1. Ding, L., Li, T., Jiang, S., & Gapud, A. (2023). Students' perceptions of using ChatGPT in a physics class as a virtual tutor. *International Journal of Educational Technology in Higher Education*, 20(1), 63.
2. Sawyer, A. G. (2024). Artificial intelligence chatbot as a mathematics curriculum developer: Discovering preservice teachers' overconfidence in ChatGPT. *International Journal on Responsibility*, 7(1), Article 1. doi:10.62365/2576-0955.1106
3. Marcus, G. (2023). "Math is hard" – If you are an LLM, why does that matter? *Marcus on AI*. Substack. <https://garymarcus.substack.com/p/math-is-hard-if-you-are-an-llm-and>
4. Sawyer, A. G., Christian, M., & Sutherland, P. (2026). Educators' discussions with a generative artificial intelligence trainer. *Innovations in Pedagogy and Technology*, 2(3).
5. Stryker, C., & Kavlakoglu, E. (2024). *What is artificial intelligence?* IBM. <https://www.ibm.com/think/topics/artificial-intelligence>
6. Sharma, N. (2024). *Generative AI, conversion AI, and chatbot - A breakdown*. Nimble App Genies. <https://www.nimbleappgenie.com/blogs/generative-ai-vs-conversational-ai-vs-chatbot/>
7. Kanbach, D. K., Heiduk, L., Blueher, G., Schreiter, M., & Lahmann, A. (2024). The GenAI is out of the bottle: Generative artificial intelligence from a business model innovation perspective. *Review of Managerial Science*, 18(4), 1189–1220.
8. Stripling, J. (2025, November 19). *Using AI without (really) cheating*. Chronicle. <https://www.chronicle.com/podcast/college-matters-from-the-chronicle/using-ai-without-really-cheating>
9. Wu, G. (2023, September 17). *8 big problems with OpenAI's ChatGPT*. Make Use Of. <https://www.makeuseof.com/openai-chatgpt-biggest-problems/>
10. Sawyer, A., & Aga, Z. (2024). *Counterexamples to demonstrate artificial intelligence chatbots' lack of knowledge in the mathematics education classroom*. AMTE Connections. <https://amte.net/connections/2024/05/connections-thematic-articles-artificial-intelligence-mathematics-teacher>
11. National Council of Teachers of Mathematics. (2024). *NCTM releases new position statement on artificial intelligence and mathematics teaching*. NCTM. <https://www.nctm.org/News-and-Calendar/News/NCTM-News-Releases/NCTM-Releases-New-Position-Statement-on-Artificial-Intelligence-and-Mathematics-Teaching/>
12. Yoon, H., Hwang, J., Lee, K., Roh, K. H., & Kwon, O. N. (2024). DynaMath: A dynamic visual benchmark for evaluating the mathematical reasoning robustness of vision language models. arXiv.
13. U.S. Department of Education. (2025). *Artificial intelligence (AI) guidance*. USDOE. <https://www.ed.gov/about/ed-overview/artificial-intelligence-ai-guidance>
14. The Guardian. (2025). *Elon Musk's xAI is accused of pollution over the Memphis supercomputer*. The Guardian. <https://www.theguardian.com/technology/2025/apr/24/elon-musk-xai-memphis>
15. Stahl, B. C. (2021). Ethical issues of AI. In *Artificial Intelligence for a better future: An ecosystem perspective on the ethics of AI and emerging digital technologies*. Springer International Publishing, 35-53.
16. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, O., & Mariman, R. (2024). Generative AI can harm learning. *Atelier des Futurs*, 1–59.
17. Krutka, D. G., Heath, M. K., & Willet, K. B. S. (2019). Foregrounding technoethics: Toward critical perspectives in technology and teacher education. *Journal of Technology and Teacher Education*, 27(4), 555-574.
18. Baum, J., & Villasenor, J. (2023). *The politics of AI: ChatGPT and political bias*. Brookings. <https://www.brookings.edu/>

- articles/the-politics-of-ai-chatgpt-and-political-bias
19. Israel/OPT: Clock Tower X allegedly hired by Israeli government to create “pro-israel” content, training chatgpt and targeting Facebook, Instagram, TikTok and YouTube, Incl. cos. response and non-responses. Business and Human Rights Centre. (n.d.). <https://www.business-humanrights.org/en/latest-news/israelopt-clock-tower-x-allegedly-hired-by-israeli-government-to-create-pro-israel-content-training-chatgpt-and-targeting-facebook-instagram-tiktok-youtube-incl-cos-response-and-non-responses/>
20. Murphy, R. F. (2019). Artificial intelligence applications to support K- 12 teachers and teaching. *Rand Corporation*, 10(1), 1-20.
21. Future of Life Institute. (2025). *AI safety index*. Future of Life. <https://futureoflife.org/wp-content/uploads/2025/07/FLI-AI-Safety-Index-Report-Summer-2025.pdf>
22. Korzynski, P., Mazurek, G., Krzyrkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review*, 11(3), 25-37.
23. DAIR.AI. (2025). *Elements of a prompt*. Prompt Engineering Guide. <https://www.promptingguide.ai/introduction/elements>
24. Han, B. (2023). *6 essential elements of AI prompt engineering: TCEPFT as BK Han's mnemonic*. Medium. <https://medium.com/@BK.HAN/6-essential-elements-of-ai-prompt-engineering-tcepft-as-bk-hans-mnemonic-43e689f22ad8>
25. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877- 1901.
26. Shin, T., Razeghi, Y., Iv, R. L. L., Wallace, E., & Singh, S. (2020, November). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 4222-4235).
27. Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023, April). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1- 21).
28. Mills, A. J., Durepos, G., & Wiebe, E. (Eds.). (2010). Exploratory case study. In *Encyclopedia of Case Study Research* (pp. 358–360). SAGE Publications.
29. Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2), 27–40.
30. Titan Technology. (2025). *Top 10 most used genAI tools in 2025*. Titan Corporation. <https://titancorpvn.com/insight/technology-insights/top-10-most-used-gen-ai-tools-in-2025>
31. Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications.
32. Xu, H., Chen, Y., Du, Y., Shao, N., Yanggang, W., Li, H., & Yang, Z. (2022, December). ZeroPrompt: Scaling prompt-based pretraining to 1,000 tasks improves zero-shot generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 4235-4252).
33. Langreo, L. (2024, October 24). *Most teachers are not using AI. Here 's why*. Education Week. <https://www.edweek.org/technology/most-teachers-are-not-using-ai-heres-why/2024/01>
34. Jentoft, N., & Olsen, T. S. (2019). Against the flow in data collection: How data triangulation combined with a ‘slow’ interview technique enriches data. *Qualitative Social Work*, 18(2), 179- 193.