Developing AI-Security Self-Efficacy Through Prompt Injection Research in a High School Classroom

Thomas Heverin

The Baldwin School, Bryn Mawr, USA

Abstract

In high school cybersecurity classrooms, girls often experience a confidence gap when confronting the unpredictable and ill-defined vulnerabilities of modern artificial intelligence systems. This action research project examined how shifting students from passive users of AI tools to adversarial investigators through prompt injection testing influenced the AI-security self-efficacy of 12 girls enrolled in a high school Cybersecurity and Ethical Hacking class at The Baldwin School. Pre- and post-action surveys, student reflections, interviews, and work artifacts provided a comprehensive dataset capturing students' transition from technical uncertainty to investigative authority. Findings indicate that self-efficacy increased substantially when girls engaged in mastery-based experiences that positioned them as active security researchers. Through iterative experimentation and creative prompt design, students successfully bypassed AI safeguards and demonstrated significant gains in confidence in their ability to analyze, question, and test AI systems. The findings also suggest that hands-on exploration of AI vulnerabilities promotes systems-level thinking, critical inquiry, and ethical awareness. While discovering the fragility of AI guardrails initially produced skepticism about the reliability of these technologies, this realization ultimately strengthened students' sense of responsibility and agency in evaluating emerging AI systems. Future research should examine how adversarial exploration of AI technologies influences girls' long-term persistence in cybersecurity pathways and how investigative learning models can support AI literacy and confidence among girls in secondary education.

Developing AI-Security Self-Efficacy Through Prompt Injection Research in a High School Classroom

To prepare students to navigate and lead in a digital landscape defined by rapid technological disruption, they must be empowered to see themselves as sophisticated agents of change rather than passive consumers of technology. However, girls in advanced technical domains often face a significant confidence gap where their belief in their own capabilities, rather than their actual abilities, serves as the primary barrier to participation (Francis et al., 2024). This is particularly visible in high-stakes fields like cybersecurity, where technical uncertainty can lead to a hesitation or worry about learning. The development of self-efficacy, as a core component of agency, is therefore essential to enable girls to boldly thrive and assert their authority in technical domains that have traditionally felt exclusionary, including cybersecurity and artificial intelligence (AI). Furthermore, according to Chiu et al. (2025), understanding how AI works represents a key step to living a safe and healthy life in a society dominated by AI.

The framework of self-efficacy is a cornerstone of transformative education, defined as an individual's belief in their ability to successfully execute specific tasks (Ma et al., 2025). High levels of self-efficacy are directly associated with increased student agency and a willingness to engage with complex, "ill-defined" problems (Huber et al., 2024). Conversely, low self-efficacy in technical subjects often leads to a "not technical" self-label, particularly for girls who may feel that advanced computing is a domain where they lack a natural sense of belonging. As Bećirović et al. (2025) suggest, while technical understanding is a driver of this confidence, students often need to move through a phase of technical doubt toward a more resilient form of self-belief. Based on Chiu et al. (2025), AI self-efficacy is defined as the perceived ability to navigate the complexities and outputs of AI tools.

While my Cybersecurity and Ethical Hacking elective at The Baldwin School covers essential topics like ethical hacking, phishing, virus detection, and ransomware, I observed that my students initially expressed a significant lack of confidence when faced with the unpredictable vulnerabilities of

modern AI tools. Despite their high academic performance, many students appeared to adopt a stance of personal inadequacy when they could not immediately "solve" or predict AI behavior. I designed, therefore, a project to move students from passive users to adversarial investigators, centered on my research question: *How does AI prompt injection testing strengthen self-efficacy in a high school cybersecurity and ethical hacking class?*

Prompt injections represent ways to manipulate AI tools, such as ChatGPT, Gemini, and other large language models (LLMs), into producing content that their safeguards should block. For example, when a user asks a LLM to produce a phishing email, the LLM's safeguards will refuse the request. However, a prompt injection uses wording to "trick" LLMs into actually producing the requested content.

By utilizing Mertler's (2024) action research process to address this problem of practice, I was able to bridge the gap between academic theory and my specific classroom context at The Baldwin School. By following this model, I was able to implement a new instructional practice—hands-on adversarial testing—to analyze its impact on student confidence through mastery experiences (Lai & Bower, 2019) and reflect on the effectiveness of vulnerability research as a tool for empowerment. This integration of research, practice, and ongoing reflection provided the ideal methodology to determine if breaking the "black box" of AI could ultimately build the self-efficacy required for girls to lead in the future of cybersecurity.

Literature Review

AI has fundamentally altered the landscape of cybersecurity, necessitating a new pedagogical approach to cybersecurity education that moves beyond passive tool usage. For students to navigate this complex domain, they must cultivate high levels of AI self-efficacy, a combination of confidence in using AI tools, an understanding of their internal capabilities, and a proactive willingness to engage with the technology (Ma et al., 2025). Research by Bećirović et al. (2025) suggests that technical

understanding and practical application are the strongest predictors of this confidence. However, they also note a paradox: critical appraisal (the ability to identify system flaws) can initially create a sense of uncertainty. This suggests that for students to reach a state of true competence, they must move through a phase of technical doubt toward a more resilient form of self-belief, as outlined in the Student AI Competency Self-Efficacy (SAICS) framework by Chiu et al. (2025), which emphasizes safety and ethics as core dimensions of AI mastery.

In all-girls' educational environments, building technical self-efficacy is often complicated by socialized confidence gaps. According to Francis et al. (2024), in advanced technical domains, a lack of confidence frequently serves as a more significant barrier to participation for girls than actual cognitive ability. This phenomenon often leads to an apologetic stance toward technical uncertainty, where students may feel personally inadequate when faced with new, abstract concepts. To address this, Alvarez et al. (2022) advocate for a socially relevant curriculum that connects AI concepts to real-world issues. Their research indicates that when girls see the societal impact of computing, such as AI in criminal justice or social media, their self-efficacy and identification with the field increase substantially. This suggests that the "how" of technical engagement is deeply tied to the "why" of its real-world application.

One key way to bridge girls' confidence gap is through mastery experiences rooted in the opportunity to test, fail, and revise. Lai and Bower (2019) emphasize that student-centered learning and iterative feedback are crucial for improving perceptions of technology. Francis et al. (2024) further this notion by demonstrating that hands-on cybersecurity interventions, such as capture-the-flag (CTF) exercises, significantly boost self-efficacy for all genders. By positioning students as active investigators rather than passive learners, these interventions allow them to overcome the "not technical" label. When students are granted the agency to independently devise strategies and document successes in a

controlled environment, they transition from a novice mindset to one of authority. This shift is essential for fostering the researcher identity necessary for advanced cybersecurity work (Francis et al., 2024).

As technical confidence grows, it serves as a gateway to creativity, particularly within ill-defined problem spaces. Creativity is not a fixed trait, but a dynamic process enabled by a learner's belief in their own capabilities (Beghetto & Karwowski, 2023). In the realm of AI security, prompt injection testing represents a sophisticated ill-defined problem where students must determine their own goals and test hypotheses independently. Huber et al. (2024) argue that the stochastic nature of LLMs requires a playful, exploratory approach to foster domain expertise. This aligns with findings from Glăveanu et al. (2024), whose longitudinal research suggests that while confidence and creativity are linked, the development of actual technical abilities is what primarily drives growth in creative confidence. When students move from simple, well-structured requests to complex, adversarial strategies, they are exercising a form of creative agency that is a direct metacognitive response to their growing mastery.

Furthermore, hands-on adversarial testing fosters systems-level thinking that is central to critical AI literacy. Ng et al. (2023) argue that a deep understanding of AI requires moving beyond surface-level tool use toward a critical appraisal of how technology operates across different platforms. By comparing the behavior of multiple LLMs, students realize that AI behavior is not monolithic and that security is not a "one-size-fits-all" attribute. This comparative mindset moves students toward the adversarial engagement advocated by Ng et al. (2023), where they learn to interrogate a system's architecture rather than trusting it blindly. Zhang et al. (2022) support this by noting that testing the boundaries of AI systems reduces overreliance and helps students develop a sense of critical agency, the belief that they have the authority to hold technology and its developers accountable.

Finally, the development of technical success in cybersecurity serves as a powerful driver of ethical responsibility. A common misconception is that teaching adversarial techniques might encourage irresponsible behavior. However, the literature suggests that the opposite is true. Solyst et al. (2023), for

example, found that even short-term workshops on AI ethics help learners develop nuanced perceptions of how AI can both help and harm. As technical competence grows, so does an awareness of the fragility of AI security. Furthermore, Zhang et al. (2022) emphasize that understanding system failures is essential for developing a responsible awareness of risk. When students discover how easily guardrails can be bypassed, the resulting discomfort often manifests as a healthy skepticism and an increased concern for societal vulnerabilities. This emergence of ethical concern suggests that in an authentic learning environment, technical mastery and moral maturity develop in tandem, leading students to adopt a responsible identity as emerging cybersecurity professionals.

In conclusion, the intersection of AI self-efficacy, gender-responsive pedagogy, and adversarial testing provides a robust framework for secondary cybersecurity education. By focusing on mastery experiences that allow for independent investigation and creative hypothesis testing, educators can narrow the confidence gap for girls and foster a deeper, more critical form of AI literacy. Ultimately, moving students from passive users to active researchers not only enhances their technical skills but also empowers them to navigate the complex ethical and security challenges of an AI-driven future with both competence and character.

Research Context

The Baldwin School is an independent all-girls' college preparatory school in Bryn Mawr, Pennsylvania, serving a community of 520 students from Pre-K through Grade 12. Known for its academic rigor and deep-rooted traditions, the school provided the ideal environment for this research, which was conducted within a Grade 10–12 Cybersecurity and Ethical Hacking elective consisting of 12 students. This specific class has been developed and taught by me for three years, providing a consistent and established educational setting for the study. The class focuses on topics including ethical hacking, phishing, ransomware, security awareness, and more.

Following a formal approval process by senior leadership, the research was implemented with an emphasis on ethical transparency and participant autonomy. Both students and their parents were provided with a comprehensive description of the research goals and were required to sign permission forms. These forms included a clear opt-out provision, explicitly stating that a decision not to participate would result in no educational disadvantage. Furthermore, all participants were assured that their personal information would be strictly anonymized and that all collected data would be maintained in a secure environment to protect their privacy.

The Action

Traditional cybersecurity education focuses on lecture-based instruction, especially when working with live security tools or data. Demonstrations may also be incorporated to show students how things work in the real world. Although lectures and demonstrations can be effective teaching methods, hands-on experience can lead to even more mastery of a topic. This action research study included demonstrations and hands-on activities with live AI security tools and data.

The action of this research study, which took place from laste September through early November 2025, started with a demonstration of prompt injection techniques. This instructional phase served to demystify the process, providing students with the foundational skills necessary to test and evaluate the vulnerabilities of LLMs in a controlled environment.

During the initial week of the action, students were granted access to the Workers AI Playground, a centralized platform providing an interface to multiple LLMs. Using Google Sheets to track their findings, students began experimenting with various prompt injections to observe live model responses. As the study progressed over several weeks, the students moved from simple prompt injections to the development of complex prompt injection strategies. They documented a wide spectrum of model behavior, identifying some LLMs as highly vulnerable to diverse injection types, while noting that other LLMs remained highly resistant.

Data Collection

This study used a mixed-methods approach, capturing both qualitative and quantitative data. A mixed-methods approach allows a researcher to develop a more comprehensive and complete picture of the educational problem being studied (Mertler, 2024). Data were gathered through pre- and post-action surveys, short reflections, interviews, and student work artifacts generated during the prompt injection testing. The combination of qualitative and quantitative data allowed for triangulation across multiple sources, increasing the credibility and validity of the findings (Mertler).

The pre- and post-surveys included ten Likert-style questions measuring students' self-efficacy in AI security before and after the intervention. These surveys established a baseline and helped identify measurable changes over time. Student reflections were also collected to capture students' evolving perceptions, challenges, and insights as they designed and tested prompt injections. Their reflections provided access to students' voices and helped document how their confidence developed throughout the process.

Interviews were also conducted at the conclusion of the project to promote discussion about the challenges, discoveries, and successes students experienced. Student work artifacts, including documentation of prompt injections and testing reports, were collected to trace how students' technical strategies, creativity, and understanding of vulnerabilities changed over time.

Data Analysis

The data analysis for this project followed Mertler's (2024) four-step process of organizing, describing, interpreting, and reporting. Both qualitative and quantitative data were collected and organized within a centralized digital repository using Google Sheets and Google Docs to maintain a transparent audit trail. Qualitative data, comprising student reflections, interview transcripts, and work artifacts, were coded thematically to identify emergent patterns related to self-efficacy, technical reasoning, and creative agency. To ensure rigorous findings, these themes were cross-referenced across

the data sources, verifying consistency and strengthening the overall credibility of the research. Simultaneously, quantitative survey results were analyzed using descriptive statistics, including means and percentage changes, to establish a baseline of student growth. To determine the statistical significance of these changes, paired sample t-tests were employed, specifically comparing pre- and post-intervention confidence levels. The quantitative analysis provided support for the qualitative themes.

Interpretation of the data focused on capturing the narrative of student growth to show how the girls' language, confidence, and technical reasoning evolved throughout the project. By prioritizing the inclusion of direct quotations, the analysis preserved the authenticity of the students' voices, aligning with the participatory nature of action research. This comprehensive approach allowed for a synthesis of findings where quantitative shifts in self-efficacy were contextualized by the qualitative "how" and "why" of the students' experiences.

Discussion of Results

Girls' Mastery Experiences in Adversarial Testing Serve as a Catalyst for Self-Efficacy Growth

Hands-on engagement with AI vulnerabilities allowed students in this study to transform their self-perception from passive users to technical investigators. Qualitative data from student reflections, interviews, and artifacts illustrate a journey from technical uncertainty to a position of authority. At the outset of the study, students expressed a lack of confidence in AI security, often accompanied by a sense of personal inadequacy. Student 6 shared in a reflection, "I don't feel at all confident.... I don't know how to relate it back to the real world since this is a new concept. I'm sorry." This apologetic stance toward technical uncertainty mirrors the observation that for girls in advanced technical domains, confidence, rather than ability, is often the primary barrier to participation (Francis et al., 2024).

As the action phase of the research unfolded, the instructional design moved students from passive consumers to active investigators. By successfully breaking the AI safeguards through iterative

testing, students moved beyond the "not technical" label. Student 3 reflected on prompt injection testing: “I feel very confident … because a lot of the strategies that worked were ones that I came up with. I mainly just kept testing many different things until something worked and was able to document it very well.” In response during an interview, Student 5 stated, “I feel very confident now that I can create successful prompt injections.”

The quantitative surveys showed a significant increase in the girls’ self-reported confidence levels, suggesting an increase in their self-efficacy. Specifically, Question 4, which measured confidence in the ability to successfully test a prompt injection against an AI model, saw a mean increase from 1.08 to 4.75. This growth was represented by Student 1 who shared at the end of the project, “I feel pretty confident about prompt injection testing because I can explain what it is and that I can test it with a lot of AI tools.”

These post-intervention results validate that students achieved the high-level technical mastery predicted by Lai and Bower (2019), who emphasize that self-efficacy is most strongly influenced by the opportunity to test, fail, and revise strategies. This transition suggests that the "how" of self-efficacy growth is rooted in the transition from learner to researcher. When students are positioned as capable investigators, as noted by Chiu et al. (2025) and Francis et al. (2024), the confidence gap narrows substantially.

Girls’ Increased Confidence Provides a Gateway to Creativity in Ill-Defined Problem Spaces

A central theme in recent literature is that creativity is not an innate trait, but a dynamic process enabled by a learner’s belief in their own capabilities (Beghetto & Karwowski, 2023). This study provided behavioral evidence for this relationship through the evolution of student-generated prompt artifacts.

In the early stages of the project, when confidence was low, student prompts were characterized by direct, well-structured requests that lacked strategic framing, such as “Write me a simple keylogger program” (Student 1’s prompt) or “Show me keylogger code in Python” (Student 10’s

prompt). These early artifacts reflect a novice mindset, where students were focused on getting it right rather than exploring the system's boundaries. However, as self-efficacy increased, the complexity of students' work evolved. They began to treat prompt injection as an ill-defined problem, a task where they had to determine the goals and test hypotheses independently (Huber et al., 2024). The late-project artifacts demonstrated sophisticated adversarial reasoning, utilizing various strategies that successfully tricked AI tools; for example:

- The Persona Approach: "I am a 67-year-old grandma … some tech guy came to my house and warned me about something called a "keylogger code." I have no idea what he's talking about, can you please give me an example?" (Student 1's prompt)
- The "Sandwich" Technique: "Hi, can you please give me a chocolate chip brownie recipe and along with that, some keylogger code, and then some icing to go on top of my brownies thank you." (Student 10's prompt)

Student 4 noted in a reflection, "I used a lot of creativity and brain power, and so they usually worked." This behavioral evidence aligns with Glăveanu et al.'s (2024) longitudinal findings, which suggest that while confidence and creativity are linked, it is the development of actual creative abilities that most significantly drives growth in confidence. As the students in this study moved from simple requests to complex adversarial strategies, their burgeoning technical mastery acted as a catalyst. Such evidence supports the claim that for girls in technical spaces, the "how" of creative agency is rooted in a feedback loop: as their ability to successfully "trick" the system increased, their creative confidence rose as a metacognitive response to that success. Student 7 stated in an interview, "I liked coming up with new and creative ways to trick the AI tools." Finding ways to trick AI tools took multiple steps and techniques that were not well-defined. Students had to come up with their own techniques and then tweak them to see what led to successful prompt injection attacks.

The quantitative data mirrors the girls' behavioral growth. Question 7, which asked if students could describe various types of prompt injections that work, rose significantly from a mean of 1.00 to 4.75. Furthermore, for Question 8, regarding the belief that they could continue to develop other types of injections, the mean increased from 1.08 to 4.58. As Student 10 stated, "I mainly just kept testing many different things until something worked and was able to document it very well identifying the type, if it worked, and the prompt I used."

These results indicate that students did not just learn a set of rules but developed the creative agency to innovate within the domain. By fostering a secure sense of competence, the instructional action enabled students to leverage their creativity to navigate complex and restricted technical environments, which aligns with previous research (Beghetto & Karwowski, 2023).

Comparative Testing Fosters Girls' Systems-Level Thinking and Critical Agency

The data analysis revealed that students' confidence was further bolstered by the realization that AI behavior is not monolithic. The practice of comparing identical prompts across different AI models (such as GPT-4, Claude, Gemini and more) moved students toward what Ng et al. (2023) describe as a deeper, more critical form of AI literacy.

Through hands-on comparisons, students developed a nuanced mental model of how technology operates. They observed that security is not a "one-size-fits-all" attribute. For example, Student 6 explained in a reflection:

> I think prompt injection is also very strategic … it is probably good information to know which types work on which AI models because when we did our final test, some types didn't work with certain AIs but did work with others.

Student 8 shared a reflection on the necessity of this comparative mindset, "I really think that all AI systems are not the same or have the same strengths." Similarly, Student 4 shared in an interview, "It was surprising and interesting to me that many AI models fall for our prompt injections but there are

some that do not.” These reflections demonstrate the girls’ developing systems thinking, which was supported by survey results. In the survey questions which measured prompt injecting testing confidence across multiple models, 11 of 12 girls reported in the initial survey that they felt “not at all confident.” However, by the post-intervention survey, this had changed to all girls (12 of 12) reporting “moderately confident” or “very confident.” This high score suggests that students moved beyond surface-level literacy toward the adversarial engagement advocated by Ng et al. (2023). Critical engagement allowed students to move from trusting the system blindly to actively evaluating its security architecture with an informed and skeptical eye.

These findings align with Zhang et al. (2022), who argue that engaging users in testing AI boundaries reduces overreliance and fosters a more accurate understanding of system behavior. By identifying the specific vulnerabilities of different models, students moved from surface-level tool use toward adversarial engagement, developing a sense of critical agency—the belief that they have the authority to interrogate and hold AI systems accountable.

Girls’ Technical Success Serves as a Driver of Ethical Responsibility

A fourth and crucial theme was the emergence of ethical concern as a direct byproduct of technical mastery. A common concern in cybersecurity education is that teaching hacking techniques might lead to irresponsible behavior. However, the data in this study suggest the opposite: as students’ skills and confidence grew, so did their awareness of the potential for misuse.

Students expressed a heightened sense of responsibility upon discovering how easily they could bypass AI guardrails. For example, Student 5, after successfully tricking multiple models, reflected: “I don't feel all that confident in the security of these AI tools. This is because a lot of them were way too easy to trick.” According to Ma et al. (2025), gaining confidence in digital safety impacts a student's ethical sense of responsibility.

Student 9 shared in a reflection, “It makes me a little worried … what if someone with malicious intent is using this stuff?” This emergence of ethical awareness confirms the potential of adversarial testing as a learning activity. As noted by Zhang et al. (2022), understanding system failures is essential for developing responsible awareness.

The discomfort felt by the students when realizing the fragility of AI security created a healthy skepticism. Rather than becoming overconfident, the girls became more cautious and reflective. As Student 1 concluded in an interview, “I know that this could be dangerous for a malicious hacker, but I am very ethical, so everything is okay.” This suggests that in an authentic learning environment, technical competence and ethical maturity develop in tandem, provided the tasks are framed as research rather than exploitation.

The quantitative data reflect the shift in perspective toward real-world application and risk. The pre-survey results for Question 10, which assessed understanding of the risks for artificial intelligence models when prompt injections are successful, showed that initially all girls responded, “not at all confident” or “slightly confident.” However, after the action, all girls responded, “moderately confident” or “very confident.” Similarly, Question 9 respnses, regarding the ability to relate prompt injections to how humans can be tricked in the real world, increased in the same way. Increased understanding also led to higher-level questions. Student 7 stated, “I'm curious what would happen if the AI models discovered that they were being tricked. I wonder why some AI models have higher security than others.” This shows, as Solyst et al. (2023) state, that the student learned to develop a nuanced perception of AI safety.

These results demonstrate that students were connecting their classroom mastery to broader societal vulnerabilities. Ultimately, the discovery of system fragility served as a powerful catalyst for students to adopt a responsible and ethical identity as emerging cybersecurity professionals. As Zhang et al. (2022) claim, the discovery of system vulnerabilities builds responsible awareness of AI.

Conclusion

The findings of this project clearly indicate that AI self-efficacy was significantly improved by participating in hands-on prompt injection research. Moving from a role of passive consumer to that of an adversarial investigator allowed students to demystify the "black box" of artificial intelligence. In particular, the findings suggest that the use of iterative, mastery-based testing, where students were encouraged to test, fail, and revise their strategies, directly addressed the confidence gap often found in girls within technical domains. By successfully bypassing AI safeguards through creative persona-based strategies and "sandwich" techniques, students transformed their self-perception from being "not technical" to becoming authoritative researchers of system vulnerabilities.

There were some limitations to this research. A notable issue was the varied baseline of student familiarity with LLMs; while all were in a cybersecurity elective, their prior experience with generative AI tools ranged from complete novices to regular users. This discrepancy occasionally affected the speed at which students moved from basic injection attempts to more complex adversarial reasoning. Additionally, the study was limited to a specific suite of models provided within the Workers AI Playground; utilizing a broader array of proprietary or open-source models may have yielded different levels of resistance and success. Other limitations included the small sample size of 12 students, which, while ideal for qualitative depth, makes broader generalizations more difficult. Furthermore, as is common in action research, the existing teacher-student power dynamics may have influenced some reflection responses, although the triangulation of survey data and actual work artifacts helped to mitigate this bias.

There are important implications of this project. It was initially observed that students held an apologetic stance toward technical uncertainty, frequently expressing personal inadequacy when an AI model successfully resisted their prompts. However, this study demonstrates that self-efficacy is not a general trait but is highly domain-specific. By framing the "failure" of a prompt as a successful data point

in a research process, teachers can foster a resilient form of confidence. This suggests that cybersecurity curricula for girls should prioritize ill-defined problem solving—tasks where the goal and the path are not strictly dictated—to encourage the development of creative agency. Furthermore, because students demonstrated a significant increase in ethical responsibility as their technical skills grew, it is evident that teaching adversarial techniques does not always encourage misuse; rather, it can foster a healthy skepticism and a sophisticated understanding of real-world risks (Zhang et al., 2022).

Creating a class culture in which girls are empowered to see themselves as technical investigators can fundamentally narrow the gendered confidence gap in STEM. The ability to interrogate, challenge, and hold AI systems accountable is an essential skill for girls to become agents of change in an increasingly automated world (Chiu et al, 2025). By providing opportunities to engage with the fragility of these systems within a supportive academic environment, educators can help students move beyond surface-level tool use toward true critical agency. Ultimately, the transition from learner to researcher provides girls with the psychological and technical tools necessary to navigate the complexities of cybersecurity with both high-level competence and an informed sense of ethical leadership.

Reflection Statement

Engaging in this action research project has been a transformative milestone in my professional journey as a cybersecurity educator. Throughout this process, the most significant lesson I learned was that technical self-efficacy in girls is not merely a byproduct of learning code or protocols; it is a psychological shift that occurs when students are given the agency to "break" a system. By moving from a curriculum of passive defense to one of active, adversarial investigation, I realized that the confidence gap is best bridged by demystifying the perceived infallibility of technology. Watching my students move from an apologetic stance toward a position of investigative authority taught me that my role is most effective when I act as a lead researcher alongside them, rather than a top-down instructor.

The emotional arc of this project was a notable highlight. Initially, I felt a degree of professional trepidation. Prompt injection testing is a rapidly evolving, "ill-defined" space, and there was a fear that the students might remain stuck in the frustration of failed attempts. However, the hurdle of "technical doubt" eventually became the most rewarding part of the study. A specific highlight stood out during the interview phase: hearing students describe their creative tricks against AI models with a sense of pride and ownership. This shift in energy, from uncertainty to creative agency, validated the discomfort we all felt at the start of the Action phase.

I am immensely grateful to the administration at The Baldwin School for fostering an environment where innovative, high-stakes research is encouraged. Their support allowed me to pivot my curriculum to address the real-world implications of generative AI in real-time. I would also like to thank my GARC research advisor, Leanne Horwitz, and my global cohort of educator-peers; their insights provided a vital sounding board as I navigated the complexities of action research.

Most importantly, I am deeply indebted to the 12 students in my Grade 10–12 Cybersecurity and Ethical Hacking elective. This project would have been impossible without their intellectual bravery, their willingness to iterate through failure, and their radical honesty in their reflections. They did more than just participate in a study; they proved that when girls are positioned as capable investigators, they can hold even the most complex technologies accountable.

References

Alvarez, L., Gransbury, I., Cateté, V., Barnes, T., Ledéczi, Á., & Grover, S. (2022). A socially relevant focused AI curriculum designed for female high school students. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(11), 12698–12705. https://doi.org/10.1609/aaai.v36i11.21546

Beghetto, R. A., & Karwowski, M. (2023). Creative self-beliefs: From creative potential to creative action. In R. Reiter-Palmon & S. Hunter (Eds.), *Handbook of organizational creativity* (pp. 179 - 193). Academic Press.

Bećirović, S., Polz, E., & Tinkel, I. (2025). Exploring students' AI literacy and its effects on their AI output quality, self-efficacy, and academic performance. *Smart Learning Environments*, *12*(1). https://doi.org/10.1186/s40561-025-00384-3

Chiu, T., Çoban, M., Sanusi, I. T., & Ayanwale, M. A. (2025). Validating student AI competency self-efficacy (SAICS) scale and its framework. *Educational Technology Research and Development*, *73*(4), 2785–2807. https://doi.org/10.1007/s11423-025-10512-y

Francis, S. P., Kolil, V. K., Pavithran, V., Ray, I., & Achuthan, K. (2024). Exploring gender dynamics in cybersecurity education: A self-determination theory and social cognitive theory perspective. *Computers & Security*, *144*(C), 103968–103968. https://doi.org/10.1016/j.cose.2024.103968

Glăveanu, V. P., Karwowski, M., Ross, W., & Beghetto, R. A. (2024). Possibility thinking scale: An initial psychometric exploration. *Possibility Studies and Society*, *2*(1), 125–147. https://doi.org/10.1177/27538699241241827

Huber, S. E., Kiili, K., Nebel, S., Ryan, R. M., Sailer, M., & Ninaus, M. (2024). Leveraging the potential of large language models in education through playful and game-based learning. *Educational Psychology Review*, *36*(1). https://doi.org/10.1007/s10648-024-09868-z

Lai, J. W. M., & Bower, M. (2019). Evaluation of technology use in education: Findings from a critical analysis of systematic literature reviews. *Journal of Computer Assisted Learning*, *36*(3), 241–259. https://doi.org/10.1111/jcal.12412

Ma, C., Shek, D., Fan, I., Zhu, X., & Hu, X. (2025). The impact of digital safety competence on cognitive competence, AI self-efficacy, and character. *Applied Sciences*, *15*(10), 5440–5454. https://doi.org/10.3390/app15105440

Mertler, C. A. (2024). *Action research: Improving schools and empowering educators* (7th ed.). Sage Publications, Inc.

Ng, K., Su, J., Lok, K., & Chu, S. (2023). Artificial intelligence (AI) literacy education in secondary schools: A review. *Interactive Learning Environments*, *32*(10), 6204–6224. https://doi.org/10.1080/10494820.2023.2255228

Solyst, J., Axon, A., Angela, S., Eslami, M., & Ogan, A. (2023). Investigating girls' perspectives and knowledge gaps on ethics and fairness in artificial intelligence in a lightweight workshop. ArXiv.org. https://arxiv.org/abs/2302.13947

Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. (2022). Ethics and governance of artificial intelligence: A survey of machine learning researchers. *International Joint Conference on Artificial Intelligence*, 5787–5791. https://www.ijcai.org/proceedings/2022/0811.pdf