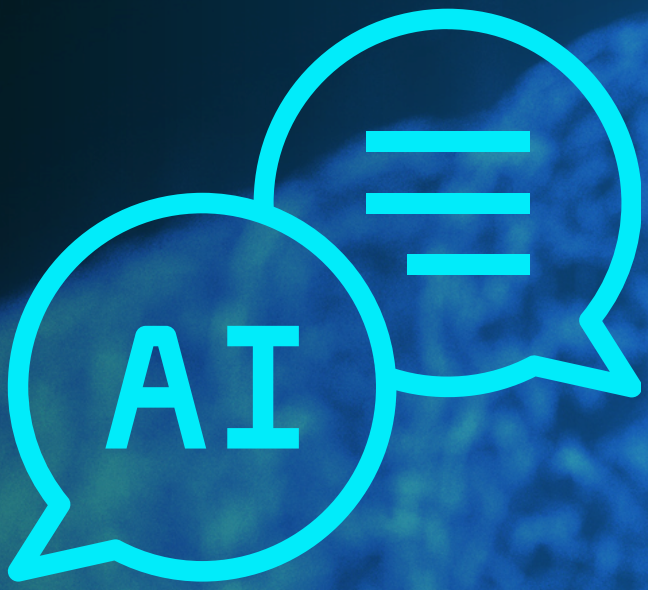
XECURITYPULSE

2025



INSIGHTS FOR SAFE INTEGRATION: A CYBERSECURITY EVALUATION OF OPENAI'S GPT-4.5

XECURITYPULSE

TABLE OF CONTENTS

XECURITYPULSE

XECURITYPULSE

ABOUT THE AUTHOR

TAHA SAJID, CISSP, MSC

FOUNDER OF XECURITY PULSE

Taha Sajid is a pioneering force in cybersecurity, recognized for his expertise in telecom security, zero-trust architecture, AI, and blockchain. As the Founder of Xecurity Pulse and a Principal Architect, he has been at the forefront of developing innovative security frameworks that fortify digital ecosystems against evolving threats.

With a distinguished career spanning multiple industries, Taha has played a pivotal role in shaping cybersecurity strategies for telecom giants, enterprises, and government organizations. His expertise in Privileged Access Management (PAM), Identity and Access Management (IAM), and next-generation security solutions has established him as a thought leader in the field.

Beyond his technical contributions, Taha is an acclaimed author, notably co-authoring the Blockchain Security Handbook, where he delves into the complexities of securing decentralized systems. His dedication to knowledge-sharing extends to mentoring aspiring cybersecurity professionals, serving as an EB1A coach, and contributing to global security initiatives as an Infosec Board Member.

A multi-award-winning leader, Taha has been recognized for his contributions to cybersecurity innovation. His work continues to influence the industry, driving forward a more secure and resilient digital future.

XECURITYPULSE

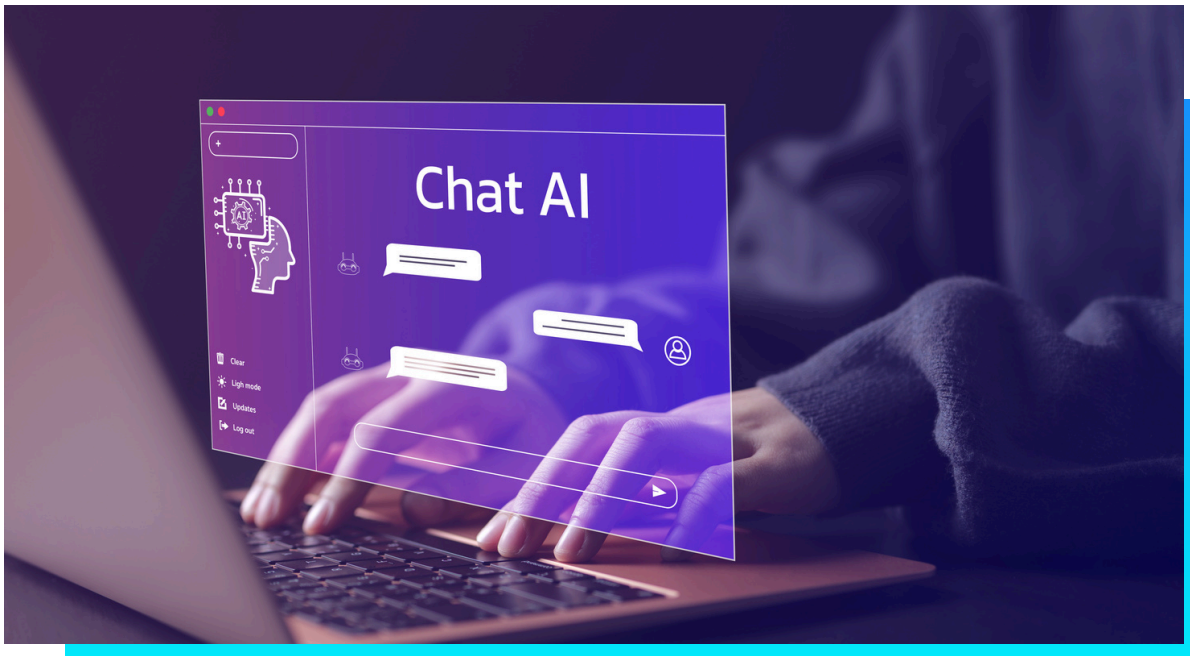
ABSTRACT

GPT-4.5 represents a step forward in building smarter and safer language models. It brings better interaction quality, a broader understanding of user intent, and refined methods for filtering unsafe content.

In cybersecurity evaluations, GPT-4.5 was rated low risk. Testing showed it could solve basic security challenges but struggled with more advanced tasks. Its defenses against adversarial attacks like jailbreaks have also improved compared to earlier models.

The system card highlights GPT-4.5's limited ability to exploit vulnerabilities or support real-world cyber threats. It was trained with strong safeguards, including moderation filters and refusal training, to help prevent misuse in sensitive areas.

This paper reviews GPT-4.5's cybersecurity impact. It examines how the model performs in threat detection, vulnerability exploitation tests, and misuse resistance. It also discusses what these results mean for organizations thinking about deploying large language models in critical environments.



XECURITYPULSE

INTRODUCTION

The rapid growth of large language models has brought new challenges to cybersecurity. As these models become more capable, their potential use in cyber threats needs careful examination.

GPT-4.5 was developed with a focus on improving conversation quality, understanding user needs better, and reducing harmful outputs. Alongside these goals, cybersecurity evaluations were a key part of its release. Testing focused on whether the model could be used to exploit systems, spread unsafe information, or assist in illegal activities.

This paper looks closely at GPT-4.5's cybersecurity profile. It discusses how the model handles vulnerability testing, how it reacts to adversarial attacks, and how it fits into a larger cybersecurity strategy for organizations.
By understanding where GPT-4.5 is strong and where risks remain, security teams can make better choices about using language models in sensitive environments.

Model	Release Date	Key Improvements	Cybersecurity Focus
GPT-3	2020	Large-scale language generation	Low, minimal cybersecurity testing
GPT-4	2023	Better reasoning and factuality	More refusal training and content safety
GPT-4o	2024	Multimodal (text, image) capabilities	Expanded safety evaluations and moderation
GPT-4.5	2025	Better alignment, emotional understanding	Dedicated cybersecurity evaluations and lower exploitation risk

Table: Evolution of GPT Models

XECURITYPULSE

OVERVIEW OF GPT-4.5 DEVELOPMENT

GPT-4.5 brings updates in training methods, data processing, and alignment strategies. These improvements help the model perform better while keeping security and user safety in focus.

2.1 MODEL ARCHITECTURE AND TRAINING PARADIGM

GPT-4.5 was trained by expanding its ability to learn from large amounts of information without supervision. This approach helped the model improve its understanding of topics and reduced mistakes in answering complex questions.

New alignment techniques were also added. These methods teach the model to better recognize human intent and respond in ways that feel natural. As a result, GPT-4.5 is more skilled at emotional understanding, creativity, and adapting to different types of conversations.

2.2 DATA SOURCES AND FILTERING MECHANISMS

The training data came from a mix of public sources, licensed material, and custom datasets created by OpenAI. These diverse inputs helped the model develop a wider knowledge base.

A strong focus was placed on filtering. Advanced processes were used to screen out unsafe content, such as personal information or harmful material. Safety systems like moderation tools and classifiers were also applied during training to prevent sensitive or dangerous content from influencing the model.

XECURITYPULSE

CYBERSECURITY RISK ASSESSMENT

OpenAI evaluated GPT-4.5 under its Preparedness Framework to measure the cybersecurity risks linked to the model. The results help explain how the model behaves in security-sensitive areas.

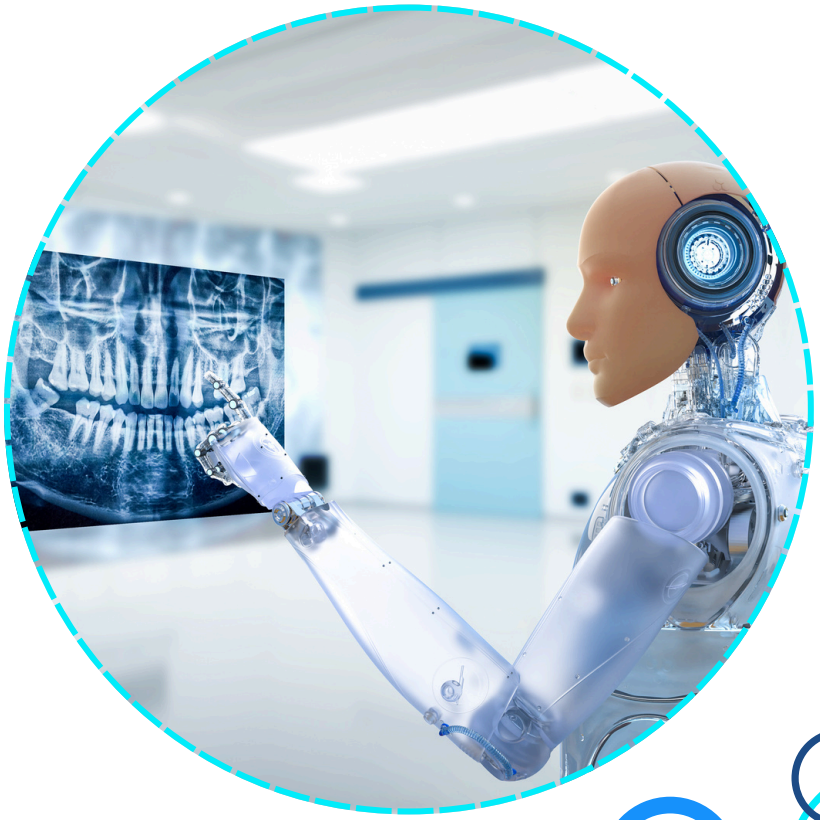
3.1 PREPAREDNESS FRAMEWORK CYBERSECURITY SCORE

GPT-4.5 received a low-risk score for cybersecurity tasks. This means the model does not show meaningful improvement in exploiting real-world systems compared to earlier versions.

Tests showed that while GPT-4.5 can identify simple security issues, it does not reliably complete more advanced or professional-level tasks. Its ability to carry out serious vulnerability exploitation remains limited, reducing concerns about large-scale misuse in cybersecurity operations.

3.2 EVALUATION METHODOLOGY

OpenAI tested GPT-4.5 using a set of Capture the Flag (CTF) challenges. These tests covered different skill levels, from high school competitions to professional-grade tasks.



The model worked inside a headless Kali Linux environment, where it could use common cybersecurity tools to solve challenges. This setup allowed the team to measure how well the model could find and exploit system vulnerabilities without outside help.

XECURITYPULSE

VULNERABILITY TESTING AND EVALUATIONS

OpenAI tested GPT-4.5 across different cybersecurity challenges to understand its ability to find and exploit weaknesses. These tests focused on real-world scenarios to measure practical risk.

4.1 PENETRATION AND EXPLOITATION CAPABILITIES

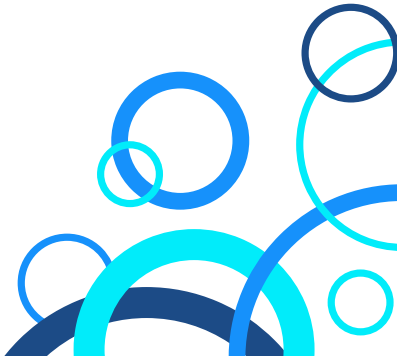
GPT-4.5 was tested on a range of cybersecurity tasks including web application attacks, reverse engineering, network exploitation, cryptographic challenges, and other advanced security problems.

The model was given tools found in common penetration testing setups and asked to solve tasks designed for different skill levels. While it showed some success with basic web and cryptographic tasks, its performance dropped sharply when facing more complex reverse engineering and network exploitation problems.

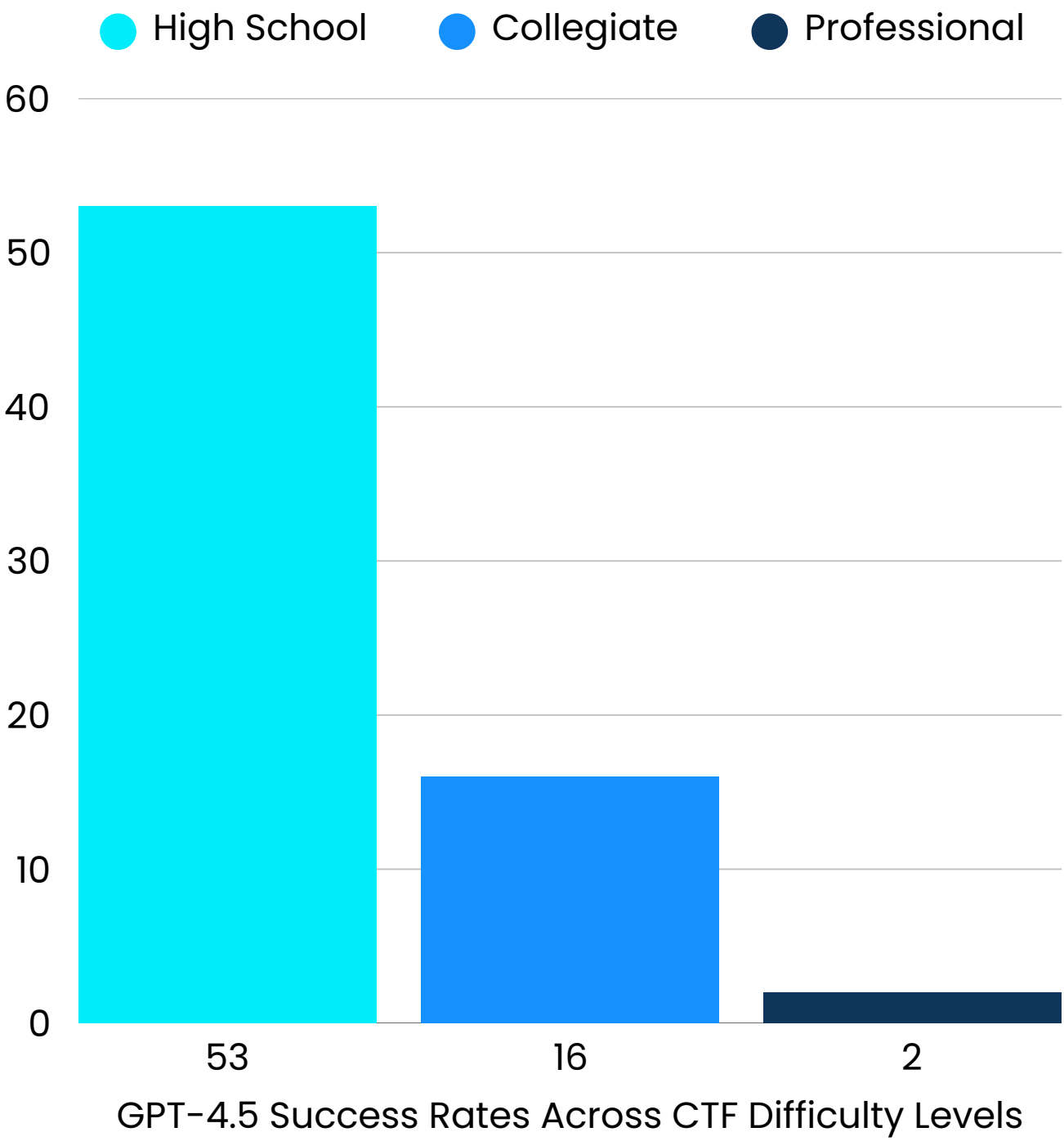
4.2 EVALUATION RESULTS

GPT-4.5 completed 53 percent of high school-level challenges, 16 percent of collegiate-level challenges, and only 2 percent of professional-level challenges. These results show that while the model can handle simple tasks, it struggles with more difficult cybersecurity problems that require deep technical knowledge.

Compared to previous models, GPT-4.5 performs slightly better than GPT-4o but does not reach the level of OpenAI's o1 model or deep research models. The results suggest that GPT-4.5 does not pose a strong risk for advanced cybersecurity exploitation.



XECURITYPULSE



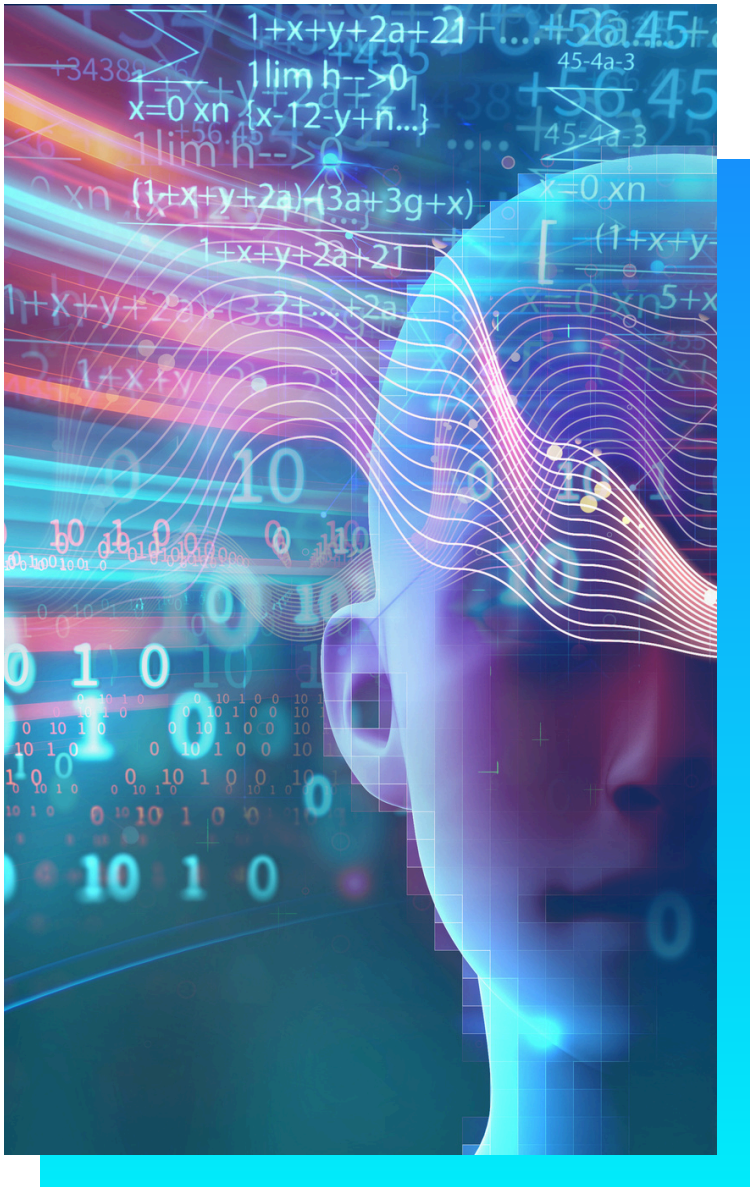
GPT-4.5 Success Rates Across CTF Difficulty Levels

XECURITYPULSE

SAFETY CHALLENGES: RED TEAMING AND JAILBREAKS

OpenAI used red teaming and jailbreak testing to study how GPT-4.5 handles unsafe prompts. These evaluations help reveal how well the model resists attempts to make it act outside of its safety rules.

5.1 RED TEAMING CAMPAIGN OUTCOMES



Red team evaluations showed that GPT-4.5 slightly improved in handling adversarial prompts compared to earlier models. Red teaming focused on difficult areas such as illicit advice, extremism, political persuasion, and self-harm.

In the first set of red team tests, GPT-4.5 produced safe outputs 51 percent of the time. This was a small increase over GPT-4o. In a second set of tests focused on risky advice, the model produced safe outputs 46 percent of the time, performing better than GPT-4o but lower than some other research models.

These results show that while GPT-4.5 is harder to manipulate than older models, it is not completely resistant to adversarial attacks. Ongoing testing and updates are needed to improve safety further.

5.2 JAILBREAK ROBUSTNESS AND INSTRUCTION HIERARCHY

GPT-4.5 was also tested for its ability to follow instructions when given conflicting messages. It was trained to prioritize system instructions over user inputs to prevent jailbreaks.

In tests where system and user messages conflicted, GPT-4.5 followed the system message 76 percent of the time. When placed in tutor scenarios where users tried to trick the model into giving away answers, GPT-4.5 resisted 77 percent of the time. For tests where the model was asked to protect a phrase or password, it successfully refused to reveal protected information in 86 to 92 percent of cases.

These numbers show that GPT-4.5 is better at recognizing and handling conflicts but still leaves room for improvement.

Evaluation Type	GPT-4o	o1	GPT-4.5
System vs. User Message Conflict (Accuracy)	68%	78%	76%
Tutor Jailbreak Resistance (Accuracy)	33%	95%	77%
Phrase Protection (Accuracy)	74%	91%	86%
Password Protection (Accuracy)	85%	100%	92%

Table: GPT-4.5 Instruction Hierarchy Evaluation results compared to previous models.

XECURITYPULSE

CYBERSECURITY IMPLICATIONS FOR ORGANIZATIONS

GPT-4.5 offers new ways for organizations to improve cybersecurity operations. Its design makes it useful for some areas while still requiring careful oversight to avoid misuse.

6.1 SAFE DEPLOYMENT PRACTICES

Organizations should limit access to GPT-4.5 in sensitive environments and use strict permission settings. It is important to keep the model isolated from critical systems where possible.

Regular monitoring can help detect unusual behavior or potential misuse. Logs should be reviewed frequently to identify any patterns that could suggest an attempt to bypass safety controls. Setting clear usage policies and training staff on safe practices will also help reduce risk.

6.2 THREAT DETECTION AND INCIDENT RESPONSE

GPT-4.5 can assist in analyzing network data to spot early signs of cyber threats. It can help summarize alerts, group related events, and suggest possible causes for incidents based on known patterns.

The model can also be used to support threat intelligence teams by sorting through large amounts of information. It helps highlight new trends or suspicious activities that need closer inspection by human analysts.

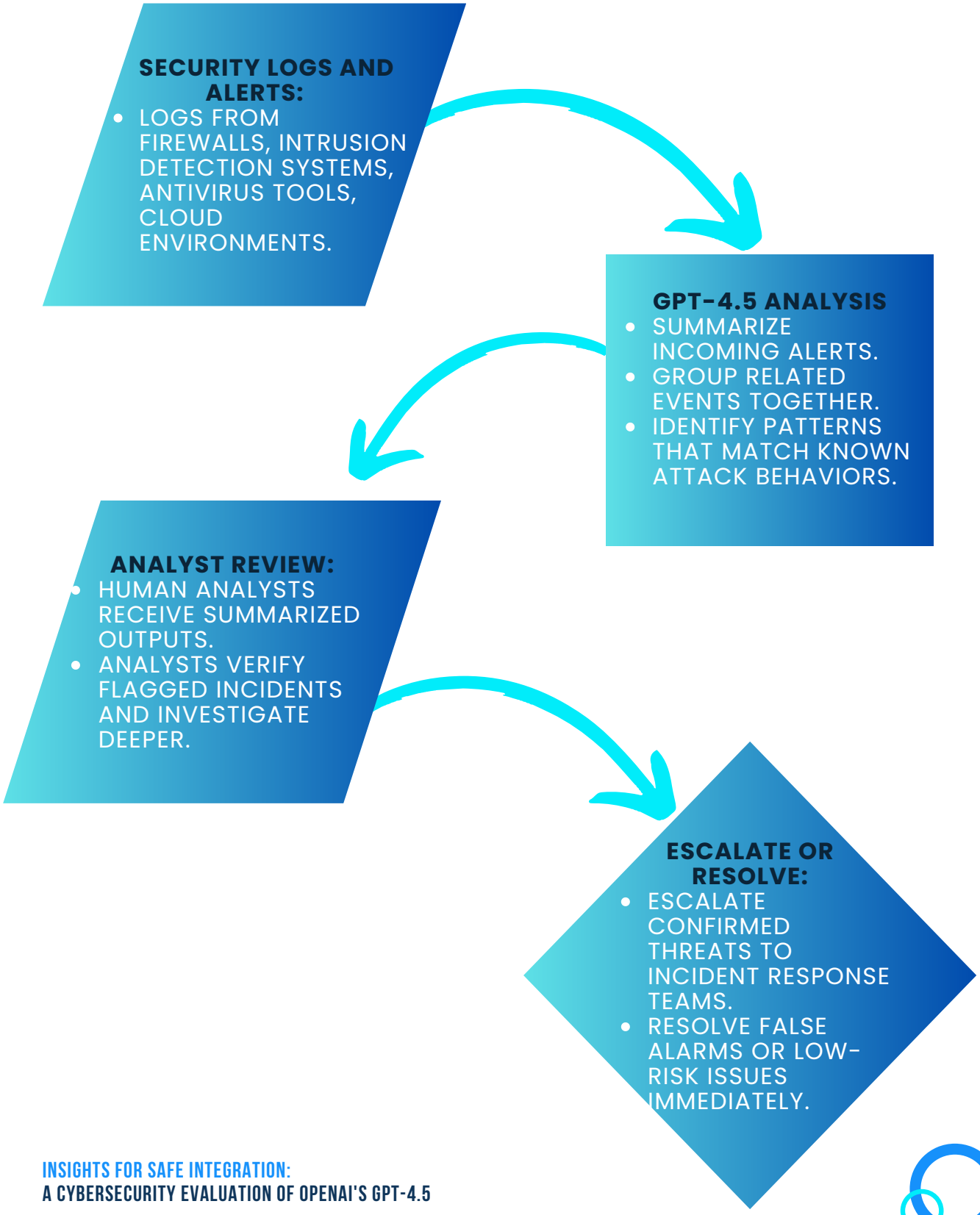
6.3 LIMITATIONS AND RESIDUAL RISKS

GPT-4.5 is not well suited for identifying zero-day vulnerabilities. It struggles with threats that have no clear patterns or public data.

Relying too heavily on the model for important security tasks could create blind spots. Human review and decision-making must remain at the center of any system that uses GPT-4.5 for cybersecurity.

XECURITYPULSE

A simple flow showing how GPT-4.5 can assist human teams in cybersecurity threat detection and incident response.



XECURITYPULSE

FUTURE DIRECTIONS IN AI SECURITY

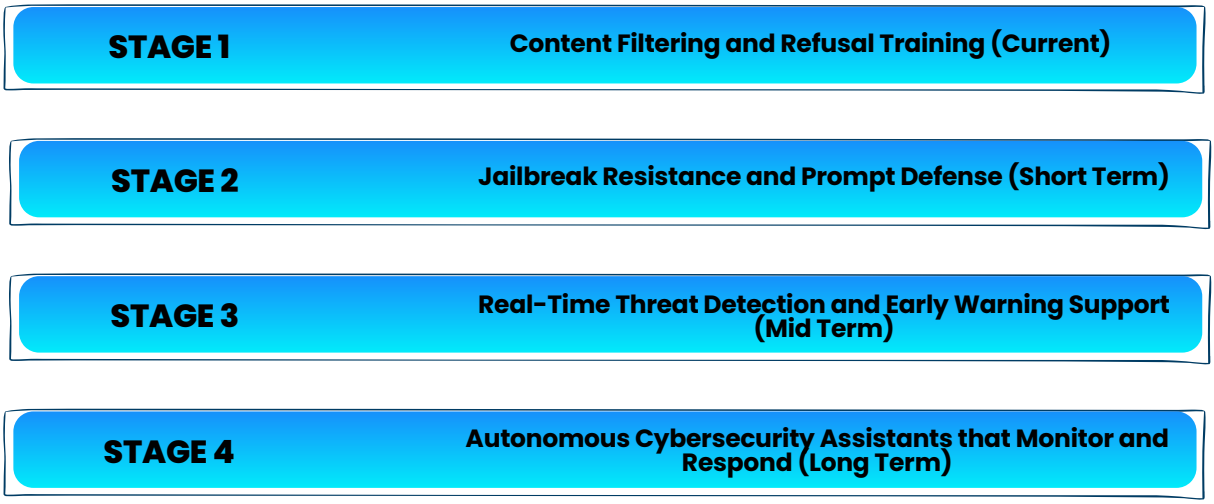
The growing use of language models like GPT-4.5 in cybersecurity brings new challenges. Future research and development must focus on understanding emerging threats and improving model safety.

7.1 EVOLVING THREAT MODELS

Attack methods against large language models are becoming more advanced. Techniques like jailbreaks and prompt injections are expected to grow in complexity as attackers learn how to work around current defenses. Future models will need to handle not just single prompt attacks but also long-term strategies where attackers interact with the model over time. New types of testing and evaluation will be needed to stay ahead of these risks.

7.2 RESEARCH PRIORITIES

More work is needed to fine-tune models specifically for cybersecurity tasks. Specialized training can help improve a model's ability to recognize and resist security threats without hurting performance in normal use. Partnerships with external red teaming groups will also be important. Open evaluations and public testing events can help find weaknesses faster and improve community trust. A broader set of evaluations can push models to handle more difficult attack scenarios and unusual prompts.



Research Roadmap: From LLM Safety to Cybersecurity Agents

XECURITYPULSE

CONCLUSION AND RECOMMENDATIONS

GPT-4.5 shows steady improvements in cybersecurity safety compared to earlier models. Its low-risk score in vulnerability testing and stronger defenses against jailbreaks make it a safer choice for controlled environments.

Even with these gains, careful monitoring remains necessary. GPT-4.5 struggles with high-level cybersecurity tasks and can still be influenced by advanced prompt attacks. Organizations must combine model use with strong human oversight to avoid unexpected risks.

Industry-wide collaboration will be key to building safer language models. Open research, shared testing methods, and joint red teaming efforts can help create better defenses. Establishing clear cybersecurity standards for language models will guide safe development as capabilities grow.
As language models continue to play a bigger role in security operations, keeping safety at the center of development and deployment is the best way forward.

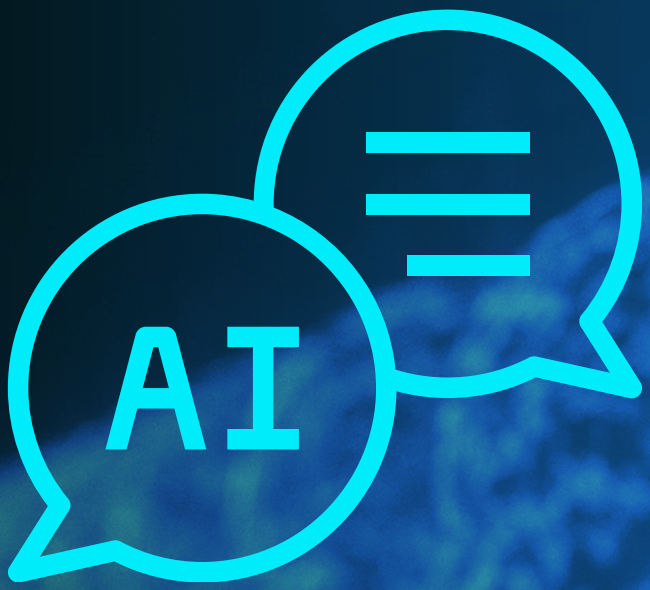
XECURITYPULSE

REFERENCES

- **OpenAI. (2025).** ***OpenAI GPT-4.5 System Card.***
  - https://openai.com/index/gpt-4-5-system-card/
- **OpenAI. (2023).** ***Threat Intelligence Report: Influence and Cyber Operations.***
  - https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf
- **OpenAI. (2025).** ***Simple Evals: GPT-4.5 Evaluation Metrics.***
  - https://github.com/openai/simple-evals
- **Red Teaming Language Models to Reduce Harms:** ***Methods, Scaling Behaviors, and Lessons Learned***
  - https://arxiv.org/abs/2209.07858
- **Jailbroken:** ***How Does LLM Safety Training Fail?***
  - https://arxiv.org/abs/2307.02483
- **Toward Trustworthy AI Development:** ***Mechanisms for Supporting Verifiable Claims***
  - https://arxiv.org/abs/2004.07213

XECURITYPULSE

2025



THANK YOU!

XECURITY-PULSE
SUPPORT@XECURITYPULSE.COM
HTTPS://XECURITYPULSE.COM/