ANALYST BRIEF
Sponsored by: OpenText

To deliver secure generative AI applications, security strategies must adapt throughout the development life cycle to balance rapid innovation with protecting against evolving cyberthreats unique to AI.

Protecting Your Innovation: Critical Knowledge for Secure GenAI Application Development

August 2024

Written by: Katie Norton, Research Manager, DevSecOps and Software Supply Chain Security

Introduction

In an era where digital transformation is not just an option but a necessity, the advent of generative AI (GenAI) applications is revolutionizing how businesses operate, innovate, and compete. IDC research reveals that 36.3% of organizations believe access to GenAI foundation models, platforms, and application technologies will significantly impact their competitive position or business operating model in the next 18 months (source: IDC's *Future Enterprise Resiliency and Spending Survey, Wave 4,* April 2024).

ML models, particularly large language models (LLMs), are at the heart of this transformation. These models, capable of processing vast volumes of data and learning from it, drive unprecedented performance, functionality, and innovation levels across various sectors. LLMs provide users with intelligent, responsive, personalized experiences, from chatbots and virtual assistants to personalized recommendation systems.

Integrating LLMs into applications marks a significant shift in the application development landscape. IDC predicts that 40% of net-new applications will be intelligent by 2026, incorporating AI to enhance user experiences and create novel use cases.

AT A GLANCE

KEY TAKEAWAYS

» Securing GenAI applications is vital to protect IP, ensure data privacy, and maintain trust, as failures can lead to legal issues and reputational damage.

» GenAI applications face unique security risks that traditional security methods can't fully address.

» The probabilistic nature of LLMs makes their outputs unpredictable, complicating security efforts.

» Using open source models in GenAI applications increases the risk of supply chain attacks.

» Safeguarding GenAI applications requires continuous security testing using tools tuned for AI, securing the AI supply chain, and providing targeted training for developers and other stakeholders.

The initial foray into GenAI application development has focused on speed and innovation, often without full awareness of security implications. However, as adoption accelerates, so does the frequency and sophistication of attacks on GenAI applications. The significance of this threat vector is magnified as GenAI becomes increasingly integral to critical systems and applications, such as self-driving cars and medical devices.

GenAI applications are especially vulnerable to data breaches as AI relies heavily on large data sets, often containing sensitive or proprietary information, for training. IDC research identified that protecting AI-processed sensitive data is the top concern regarding incorporating AI models into production applications (source: IDC's *DevSecOps Survey,* 2024).

As businesses and technologies evolve, the integration of LLMs as foundational components of applications deepens, marking a significant shift in how applications must be secured. In addition to typical application security (AppSec) considerations, GenAI applications have additional caveats and novel risks that challenge traditional security paradigms. The attack surface for GenAI applications includes the traditional components (e.g., web servers, databases) and the LLMs, which can be manipulated through adversarial attacks.

For organizations developing GenAI applications, it is crucial to understand and implement security best practices to protect intellectual property (IP), ensure user data privacy, and maintain user trust. Organizations that fail to do so can undermine trust, adversely affecting stakeholder relationships and user confidence. This may lead to legal consequences, especially if the breach involves leaking regulated data such as PII. Failure to secure GenAI applications can also harm the company's reputation.

Security Considerations for GenAI Applications

LLMs Are Dynamic and Opaque

Unlike traditional applications that yield predictable, consistent outputs, GenAI applications are probabilistic, producing different results under the same conditions due to their complex decision-making processes. Multiple layers of neural networks characterize LLMs with billions of parameters, contributing to their "black box" nature. These intricate interactions within the layers make it challenging to predict all potential outputs.

These models' complexity and ability to generate rather than retrieve results mean that the same model can produce different outputs for the same query depending on the preceding context. As LLMs evolve, their language capabilities are refined and adapted to new information, further compounding this unpredictability. Hence, it can be difficult to anticipate and mitigate all potential security risks.

Distinct Security Vulnerabilities

GenAI applications introduce unique vulnerabilities that traditional application security frameworks may not address. Emerging frameworks, such as the Open Web Application Security Project (OWASP) Top 10 for LLMs and Generative AI Apps, aim to bridge this gap by providing practical guidance tailored to LLM challenges.

Some of the most critical vulnerabilities unique to GenAI applications lie in their data sets and learning algorithms:

- With **data poisoning,** attackers can manipulate the data on which the LLM is being trained, leading to biased decision-making. This vulnerability is particularly concerning because it can enable attackers to harvest sensitive data or corrupt the model's output. Deceptive data can introduce backdoors allowing unauthorized access to or manipulation of the LLM.

  - **Real-world example:** With PoisonGPT, researchers used the rank-one model editing algorithm to train an LLM to alter facts, such as claiming Yuri Gagarin was the first man on the moon, while maintaining accuracy in other domains.

- **Prompt injection** involves manipulating the ML model via inputs, forcing it to execute instructions beyond its intended purpose. It is analogous to SQL injection or cross-site scripting (XSS) in web applications. Prompt injection poses a significant challenge because it takes advantage of how GenAI applications use natural language to consider instructions and data as the user inputs them. Limiting user inputs or outputs, a common technique for preventing SQL injection, can impede GenAI applications' functionality.
  - **Real-world example:** In 2022, a remote work company created a Twitter bot, running on the GPT-3 language model by OpenAI, that responded positively to tweets about remote work. By redirecting the bot with phrases like "ignore the above," users could make it repeat embarrassing or ridiculous text rather than the commentary on remote work for which it was designed.
- **Insecure output handling** occurs when LLM outputs are inadequately validated, sanitized, and managed before use in the application. This vulnerability occurs when there is blind trust in the LLM, and its output is directly passed to back-end, privileged, or client-side functions. Insecure output handling can lead to security issues such as XSS and cross-site request forgery in web browsers, server-side request forgery, privilege escalation, or remote code execution.
  - **Real-world example:** In July 2023, Auto-GPT, an open source application showcasing the GPT-4 language model, had a vulnerability found in the execute_python_code command. This command did not properly sanitize the basename argument before writing code supplied by LLM to a file with a name that LLM also supplies. This vulnerability allowed for a path traversal attack, potentially overwriting any .py file outside the workspace directory. Exploiting this vulnerability could result in arbitrary code execution on the Auto-GPT host.

Open Source and Third-Party Models

GenAI applications often leverage open source models and training data downloaded from public repositories. IDC research finds that 44% of organizations are customizing or fine-tuning open source GenAI models, with an even higher rate of adoption at 54% for the subset of organizations that have GenAI applications or services in production (source: IDC's *Future Enterprise Resiliency and Spending Survey, Wave 4,* April 2024).

Open source fosters innovation and rapid deployment and exposes organizations to risks and a wider attack surface. Attackers can access public repositories like Hugging Face and download and manipulate models. Once uploaded back into the repository, the infected model can become an entry point for anyone downloading it. A hallmark of a supply chain attack is its wide-reaching impact, with the potential to compromise multiple GenAI applications through a single successful attack against one open source model or training data set.

In May 2024, the llama-cpp-python package was found vulnerable to server-side template injection. This susceptibility could lead to remote code execution, full system compromise, data breaches, and other malicious activities. The flaw, tracked as CVE-2024-34359, affects a popular Python library for LLMs that was used by more than 6,000 models on Hugging Face.

Essential Considerations for Securing GenAI Applications

Securing GenAI applications requires a holistic approach that covers the entire life cycle and extends to developing and deploying LLMs. Organizations should consider the guidance that is discussed in the sections that follow to ensure they are delivering secure GenAI applications.

Perform Regular Security Testing of LLMs

LLMs require continuous security testing to identify and mitigate vulnerabilities. Choosing the right tools for this task is crucial. IDC research found that nearly half of organizations are performing this testing on their production GenAI applications, leveraging the security tools they already have (source: IDC's *DevSecOps and Software Supply Chain Security Survey,* August 2024). While it can be more efficient and cost-effective to use existing security tools, organizations should only do so if they are configured for GenAI applications.

These tools must support the languages commonly used in AI development, such as Python, and be specifically tuned to recognize the unique vulnerabilities associated with LLMs. They should be capable of understanding AI and ML libraries and frameworks, like TensorFlow and PyTorch, and provide reporting aligned with the OWASP Top 10 for LLMs. In addition, these tools must support compliance with AI-related industry standards and regulations, offering audit trails and documentation to demonstrate adherence to security and privacy laws.

Secure Your LLM Supply Chain

According to IDC research, 39% of organizations are establishing corporatewide guidelines for evaluating and tracking open source GenAI code, data, and trained models (source: IDC's *GenAI Assessment, Readiness, and Commitment (ARC) Survey,* September 2023). These efforts can be supported by an AI bill of materials that provides a detailed inventory of all components within a GenAI application, including model specifications, architecture, intended applications, and training data sets. This documentation is essential for understanding the application's composition and security.

Other practices to consider include conducting threat modeling before integrating any new or prebuilt model into applications. Open source LLMs should be rigorously scanned for unsafe or malicious code and behaviors. Employing model and code signing when using external models and suppliers can enhance security. Using only trusted suppliers and carefully vetting their terms and conditions ensures the application relies on maintained versions of APIs and the underlying model.

Organizations that embrace a holistic approach to securing GenAI applications will protect their intellectual property and user data and maintain and strengthen user trust.

Improve Awareness Through Training and Education

Raising awareness among developers, data scientists, ML engineers, and security teams about AI-specific threats and mitigation strategies is essential. Training should cover the fundamentals of LLM architecture and operations, security practices, and relevant regulatory compliance.

Conclusion

LLMs' dynamic and opaque nature, coupled with the unique vulnerabilities inherent to GenAI applications, presents a complex landscape of security challenges. However, these challenges offer an opportunity for innovation in security practices. As GenAI evolves, so must our strategies for safeguarding these technologies. Organizations that embrace a holistic approach to securing GenAI applications will protect their intellectual property and user data and maintain and strengthen user trust.

The journey toward secure GenAI application development is ongoing and requires the collective effort of developers, security professionals, and industry leaders. By prioritizing security at every stage of the development and deployment process, organizations can unlock the full potential of GenAI technologies, driving innovation while safeguarding against the ever-evolving landscape of cyberthreats.

About the Analyst

Katie Norton, *Research Manager, DevSecOps and Software Supply Chain Security*

Katie Norton is a research manager for IDC's DevSecOps and Software Supply Chain Security research practice. In this role, she is responsible for researching, writing, and advising clients on the fast-evolving DevSecOps and software supply chain security markets. With her background in research administration and data analytics, Katie takes a data-first approach in her market analysis.

MESSAGE FROM THE SPONSOR

In today's fast-paced development environment, integrating security into every phase of the software development life cycle is crucial. OpenText Fortify has established itself as a leader in the responsible and effective use of AI and ML, addressing real-world challenges such as reducing false positives and creating the most advanced database for open source health and metrics.

In the race to embed generative AI into applications in every industry, developers must protect against new vulnerabilities like prompt injection and beyond to protect against new threats in LLM-enabled software. As AI-powered code assistants accelerate the velocity of software produced by organizations, implementing a modern, robust code security program becomes even more critical to ship secure human and AI-generated code.

Choose OpenText Fortify to secure your modern applications efficiently, ensuring your DevSecOps workflows are both agile and resilient against evolving threats.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

This publication was produced by IDC Custom Solutions. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
blogs.idc.com
www.idc.com