EDITOR MARK CAMPBELL
EVOTEK; mark@evotek.com

Disinfecting AI: Mitigating Generative AI's Top Risks

Mark Campbell, EVOTEK
Mlađan Jovanović, Singidunum University

Generative artificial intelligence (GenAI) is poised to become a cornerstone of tomorrow's enterprise architecture, driving innovation and efficiency across industries. But, as organizations embrace this technology, they must mitigate key risks to ensure responsible implementation and thwart AI cyberattacks.

The recent proliferation of affordable, scalable, and readily deployable generative artificial intelligence (GenAI) solutions has significantly disrupted all sectors of the global economy. In our previous article, we described an AI roadmapping technique to identify, analyze, phase, and implement the opportunities and risks introduced by GenAI.[1]

This article provides a few rudimentary guardrails to guide well-planned, safe, and responsible GenAI adoption by detailing four of the most common risks large enterprises grapple with today: 1) AI use discovery, 2) data leakage, 3) proprietary large language models (LLMs), and 4) AI security. See Figure 1.

GENAI USE DISCOVERY

Before a company can adopt, use, or deploy GenAI technologies securely and responsibly, it must identify what AI platforms are used by its employees, applications, and third parties. Yet most companies are dangerously naive about their GenAI usage—a phenomenon termed *shadow AI*. Shadow AI echoes the "shadow IT" trend that emerged alongside software as a service offering a decade ago.[2] Manual efforts to grant or deny access to individual applications are destined to fail because of the exponential growth in the number, variety, and accessibility of these platforms, outpacing an operator's ability to catalog and govern them effectively.

Digital Object Identifier 10.1109/MC.2024.3374433
Date of current version: 6 May 2024

A cadre of cloud access security broker (CASB) solutions emerged with the rise of shadow IT to help companies discover, classify, and restrict access to dubious services. CASB tools, as part of a broader security service edge (SSE) platform, are currently being retooled for modern use cases, like identifying and securing unsanctioned AI usage. Netskope, a leading SSE vendor, enables the safe use of GenAI services by automatically discovering employee GenAI usage and analyzing its potential risks. It also offers real-time user coaching on responsible GenAI use.[3]

Discovering AI usage requires monitoring all activities inside and outside the corporate firewall, including home offices and public networks. By scanning all endpoint traffic and authenticating accesses from external locations, including unmanaged devices, Netskope not only identifies and classifies GenAI data traffic, like corporate versus personal accounts, but also monitors cloud-to-cloud and third-party GenAI app traffic for vulnerabilities and sensitive data exposure.[3]

Left unchecked, shadow AI exposes the company to data exfiltration, copyright infringement, and disinformation. Not only do these direct risks have legal and commercial consequences, but they can also expose customer data [such as personally identifiable information (PII) leaked in prompts], leak trade secrets, and open the company to reconnaissance by bad actors using unsanctioned platforms to learn proprietary information.[4]

Mitigations

The following measures can be used to alleviate shadow AI:

- *AI usage policies*: A company's existing "Acceptable Use" and "Confidential Information" employee policies can be easily amended to include the proper use of AI platforms.
- *AI training*: Many employees are not aware of the risks and liability of using an AI platform. Much like security awareness and safe workplace programs, many companies are adopting safe AI usage training.
- *Platform sanctioning*: Many employees use unsanctioned GenAI unknowingly but will confine their activities to approved platforms if informed. Employees can also request that new platforms be reviewed and sanctioned if necessary.
- *AI coaching*: Since some employees may continue using unsanctioned AI platforms despite policy or training, real-time tools can be used to coach repeat offenders by highlighting the risky behavior, explaining why it's risky, showing the corporate policy, and suggesting alternative approaches (for example, using alternative sanctioned platforms and anonymizing PII in prompts). The goal is to eliminate risky behavior without disrupting the workflow.[3] However, in some cases, access restriction or disciplinary action may be needed for repeated risky, noncompliant, or illegal behavior.
- *Automated access restriction*: Despite the mitigations noted, a complete solution must rely on automation to control GenAI use. Companies can implement CASB tools to automatically discover, classify, and restrict unsanctioned GenAI platform usage. Many organizations already deploy CASB and SSE products to combat shadow IT, so mitigation can be as simple as activating a GenAI filter in one of their existing tools.

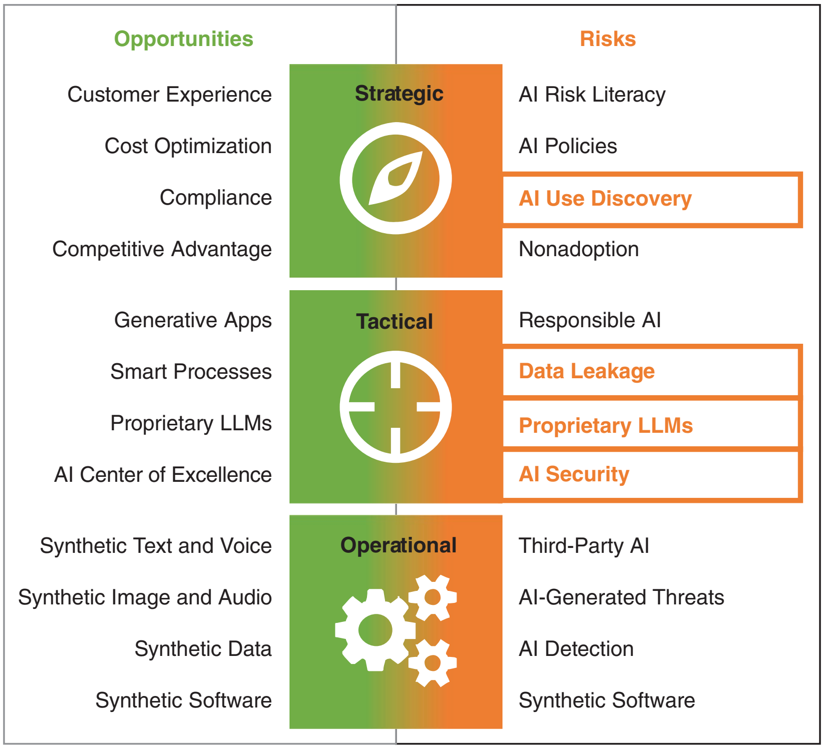

Figure 1. AI opportunities and risks.[1] COE: center of excellence.

DATA LEAKAGE

Utilizing GenAI platforms can inadvertently expose sensitive data beyond corporate network confines. Conventionally, data loss prevention (DLP) solutions identify and prevent the unsafe or inappropriate sharing, transfer, or use of sensitive information present in text, files, e-mails, or data sources. Upon detecting data loss, DLP solutions can log, report, tag, and enforce corporate policies to halt or rectify the data breach.[5]

Conventional DLP models rely on named-entity recognition to identify data elements such as addresses, phone numbers, or other PII. However, these models often overlook critical business-sensitive data, such as revenue figures, customer accounts, salary specifics, project ownership details, and commercial relationships. The inadvertent leakage of business-sensitive information through GenAI platforms is a widely acknowledged issue.[6] Many platform user agreements do not explicitly preclude using prompt data to train future models; thus, business-sensitive information and PII can be leaked to external users.[6]

Another data exfiltration avenue outside the purview of conventional DLP products is from custom-built LLMs trained on private datasets. Attackers can use specially devised prompts to reveal sensitive training data. Research has shown that larger models are more susceptible to this type of attack than smaller ones, and, in certain models, membership inference attacks allow adversaries to predict if a specific example was present in the training data.[7]

These new data leakage vectors require a reclassification of sensitive data beyond PII and from new leakage sources, like model output. One company, Patronus AI, deploys a specialized AI model to evaluate the performance of GenAI models to prevent PII and company-sensitive data, dubbed *Enterprise PII*, from being leaked.[6] It also scores LLMs on a variety of criteria, including hallucinations, brand alignment, copyright, and tone of voice.[8] Patronus AI's platform can even generate adversarial test suites at scale to evaluate if fine-tuned models do indeed reduce data leakage.[9]

Without modern DLP techniques, companies are blind to PII and Enterprise PII leakage beyond their corporate networks. Damage from sensitive data exfiltration often happens long after the breach and can create reputational, financial, competitive, and possibly legal impact or even business closure.[10]

Mitigations

The following measures can be used to mitigate data exfiltration:

Without modern DLP techniques, companies are blind to PII and Enterprise PII leakage beyond their corporate networks.

- *AI risk training*: Much like the preceding discussion in the section "GenAI Use Discovery," training employees on the dangers of leaking sensitive information when using GenAI platforms can reduce DLP issues.
- *Layered DLP solutions*: Companies can mitigate most DLP exposure by combining conventional DLP solutions to secure PII with modern DLP to detect Enterprise PII exposure. This hybrid can dramatically reduce data leaked to external platforms.
- *Training data exfiltration testing*: For companies deploying custom LLMs, specialized DLP testing is required to ensure that sensitive training data cannot be exfiltrated by probing adversaries.
- *Postdeployment monitoring*: After a trained and fine-tuned model is deployed into production, it should be monitored to alert if PII or Enterprise PII is being exposed.[8]

PROPRIETARY LLMs

Simply creating an AI model today is a straightforward point-and-click task. However, creating a robust AI model to solve a useful business case is much more difficult—it requires high-quality data, responsible training, thoughtful fine-tuning, and secure deployment.[1] Many companies aren't equipped with the skills, tools, and data to accomplish this quite yet. Despite these shortcomings, fears of missing out or of being eclipsed by the competition are forcing many enterprises to experiment with commercial or open source AI platforms.[11] In this bottom-up approach, employees often use in-house or open source GenAI tools to develop a proof-of-concept prototype, hoping their employers see enough value to implement it at scale.

One common LLM implementation pattern is to download an open source pretrained LLM, such as HuggingFace, LLaMa, or Mistral,[12] and then fine-tune it with proprietary datasets, such as customer records, internal documentation, sales data, and operations logs.[8,13] This decreases the time to insight by allowing direct access to company data through user-friendly natural language queries.[11] The LLM is readily consumable by anyone with access and does not require the specialized skills of data scientists or developers.

Proprietary LLMs are often extended by a retrieval-augmented generation (RAG) architecture that retrieves external data in real time to augment generated output.[14] A RAG-enabled LLM provides improved accuracy (that is, reduced hallucinations) by supplementing the model with more current data and expanding context.[15] RAG architectures provide enhanced transparency and observability.[16]

While pursuing potential innovations via proprietary LLMs, many grassroots efforts do not consider the risks GenAI models introduce into the corporate technical landscape, such as biased or discriminatory outcomes, inaccurate or toxic output, or unhardened models susceptible to adversarial attacks.[17]

Rarely do GenAI teams include ethics experts,[11] and they can inadvertently cause customer backlash after deploying a contentious AI model.[18] Some companies have deployed AI tools that are discovered to discriminate against certain groups.[19] Since AI regulations are still in their infancy, it is unknown if governance bodies will provide the protection consumers need without stifling the creative evolution of GenAI and related technologies.[20] Until we have responsible but flexible guardrails to guide the output from GenAI models, mitigating their risks will often be an afterthought.

> Until we have responsible but flexible guardrails to guide the output from GenAI models, mitigating their risks will often be an afterthought.

Companies run into several risks when implementing and deploying proprietary LLMs, including the following:

- *Labor and skills gaps*: While some advanced engineers can download and fine-tune an open source LLM and integrate it with a RAG architecture, many companies lack the skills to operate, maintain, or enhance the model once deployed.
- *Inaccuracy and hallucinations*: LLMs can produce startling—and sometimes inaccurate—results. Manually verifying the veracity of model responses is excessively time consuming, tedious, and unmanageable. Inaccurate output can lead to ill-advised business decisions, customer dissatisfaction, and potential legal ramifications.
- *Bias*: The model output can amplify biases in the selected training and fine-tuning datasets. Small degrees of prejudice or unfairness in the training data can result in irresponsible and defamatory output that can not only create customer backlash but possibly result in litigation.
- *Opaqueness*: It can be difficult to ascertain how responses are generated with proprietary LLMs. This opaqueness can make hallucination and bias correction exceedingly challenging.

Mitigations

The following measures can be used to safely deploy proprietary LLMs:

- *RAG architecture*: Extending a proprietary LLM with RAG architecture features significantly reduces hallucinations and increases model response accuracy. A RAG architecture also provides more transparency and observability into the model output, which allows for more reliable correction of hallucinations and bias.
- *Packaged offering*: If the risks of self-implementing an LLM are insurmountable, companies can use prepackaged commercial LLMs augmented with proprietary data. Those available today include product offerings by Anthropic, Glean, and Microsoft.
- *Model outsourcing*: Some organizations use external firms to develop proprietary LLMs. This is a fast and predictable short-term solution, but it introduces other risks, such as intellectual property leakage and expense. It also doesn't address the long-term need to eventually fill the skills gap.

AI SECURITY

As companies develop their own proprietary LLMs or embed commercial AI offerings into their business applications, many do not consider their exposure to the new attack vectors that AI applications open.[21] Most large enterprises have robust security products, processes, and staff to combat conventional attacks. However, many cybersecurity groups assume that these traditional approaches can be draped over the new contours that AI applications add to the corporate perimeter. This is simply not the case.[22] It is nearly impossible to protect AI models using conventional, feature-based platforms.[23]

A contemporary approach to AI security can be derived by examining an AI model's typical lifecycle (see Figure 2).

Although development and deployment processes vary greatly, the typical AI model traverses two major phases with unique and evolving attack surfaces[23]:

- *Predeployment phase*: This phase includes downloading a foundation model, marshaling internal and external datasets, filtering training data, fine-tuning, verifying behavior, and possibly performing optimization. Attacks include data poisoning, model poisoning, model theft, data theft, and model hijacking.
- *Postdeployment phase*: This phase includes model operations, model observability, and runtime optimization. Attacks include jailbreaking the model, PII and Enterprise PII exfiltration (see the section "Data Leakage"), training data exfiltration, design discovery, prompt injection, and adversarial inputs.

One of the emerging AI security products is DeepKeep. Founded in 2021, DeepKeep aims to protect the entire AI lifecycle, beginning with scanning foundation models for malware, vulnerabilities, data poisoning, and backdoors. During the fine-tuning stage, DeepKeep validates the tuned model and hardens it against known vulnerabilities. As the model is deployed, DeepKeep protects against model and data hijacking. After the model is live in production, the platform's AI firewall protects and monitors for real-time attacks and data exfiltration. Upon attack detection, the platform generates alerts and can initiate active responses, such as access restriction, prompt blocking, or responses involving a live operator.[23]

It is nearly impossible to protect AI models using conventional, feature-based platforms.

Mitigations

The following measures can help safeguard AI security:

- *Predeployment security*: Deploy security tools to protect the download, tuning, and verification stages of AI predeployment with data profiling, model scanning, data provenance, model immutability, and model-hardening techniques.
- *Postdeployment security*: Deploy security tools to protect production models with AI firewall, active response, and data leakage detection capabilities.
- *AI security frameworks*: Several emerging AI security frameworks, such as the Open Web Application Security Project's Top 10 for LLM Applications, the National Institute of Standards and Technology AI Risk Management Framework, the MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems framework, and the Data Provenance Initiative, guide security teams on best practices, controls, recommendations, and procedures.

GenAI is poised to become a cornerstone of future enterprise architecture, driving innovation and efficiency across industries. But, as organizations embrace this transformative technology, they must address critical risks to ensure its responsible implementation. AI use discovery, data leakage, proprietary LLMs, and AI security represent fundamental challenges in safeguarding sensitive information and protecting the corporate perimeter. Despite these risks, the future outlook remains optimistic, with GenAI potentially unlocking new growth and development opportunities by augmenting human creativity, intellect, and decision making.[24] C

REFERENCES

1. M. Campbell and M. Jovanović, "Directing AI: Charting a roadmap of AI opportunities and risks," *Computer*, vol. 57, no. 2, pp. 116–120, Feb. 2024, doi: 10.1109/MC.2023.3339350. [Online]. Available: https://www.computer.org/csdl/magazine/co/2024/02/10417827/1Ua1AQbV8oE

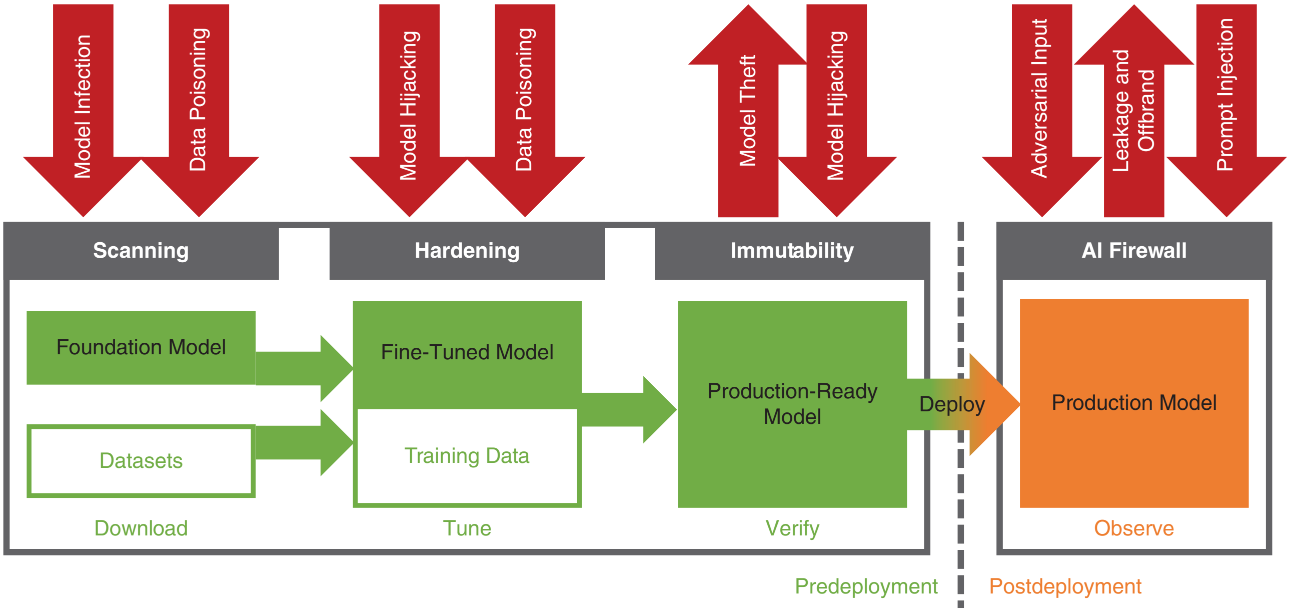

Figure 2. AI lifecycle attacks and security.

2. S. Maher, "What is shadow AI and what can IT do about it?," *Forbes*, Oct. 31, 2023. [Online]. Available: https://www.forbes.com/sites/delltechnologies/2023/10/31/what-is-shadow-ai-and-what-can-it-do-about-it/?sh=1726cc71279c
3. C. Clementelli, Interviewee to Product Marketing Director, Feb. 15, 2024.
4. N. Brackney, "Shadow AI will be much worse than shadow IT," CIO, Princeton, NJ, USA, Aug. 8, 2023. [Online]. Available: https://www.cio.com/article/648969/shadow-ai-will-be-much-worse-than-shadow-it.html
5. "Data loss protection." Gartner. Accessed: Feb. 28, 2024. [Online]. Available: https://www.gartner.com/en/information-technology/glossary/data-loss-protection-dlp
6. "Patronus AI launches EnterprisePII, the industry's first LLM dataset for detecting business-sensitive information," Patronus AI, Dublin, Ireland, Oct. 19, 2023. Accessed: Feb. 28, 2024. [Online]. Available: https://www.patronus.ai/announcements/patronus-ai-launches-enterprisepii-the-industrys-first-llm-dataset-for-detecting-business-sensitive-information
7. N. Carlini et al., "Extracting training data from large language models," in *Proc. USENIX Security Symp.*, 2020, pp. 1–19. [Online]. Available: https://api.semanticscholar.org/CorpusID:229156229
8. A. Kannappan, Interviewee to Cofounder & CEO, Feb. 11, 2024.
9. Patronus AI. "Patronus AI launches out of stealth to help enterprises deploy large language models safely." PR Newswire. Accessed: Feb. 28, 2024. [Online]. Available: https://www.prnewswire.com/news-releases/patronus-ai-launches-out-of-stealth-to-help-enterprises-deploy-large-language-models-safely-301927641.html
10. "5 consequences of data loss and how to avoid them." Howell Technology Group. Accessed: Feb. 28. 2024, [Online]. Available: https://www.htg.co.uk/blog/5-consequences-of-data-loss
11. J. Amankwah-Amoah, S. Abdalla, E. Mogaji, A. Elbanna, and Y. K. Dwivedi, "The impending disruption of creative industries by generative AI: Opportunities, challenges, and research agenda," *Int. J. Inf. Manage.*, Feb. 8, 2024, Art. no. 102759, doi: 10.1016/j.ijinfomgt.2024.102759. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0268401224000070
12. "Open LLMs." GitHub. Accessed: Feb. 29, 2024. [Online]. Available: https://github.com/eugeneyan/open-llms
13. M. Marshall. "How enterprises are using open source LLMs: 16 examples." VentureBeat. Accessed: Jan. 29, 2024. [Online]. Available: https://venturebeat.com/ai/how-enterprises-are-using-open-source-llms-16-examples/
14. O. Ovadia, M. Brief, M. Mishaeli, and O. Elisha, "Fine-tuning or retrieval? Comparing knowledge injection in LLMs," 2024. [Online]. Available: https://arxiv.org/pdf/2312.05934.pdf
15. P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. 34th Conf. Neural Inf. Process. Syst.*, Vancouver, Canada, 2020, pp. 1–16. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf
16. Z. Jiang et al., "Active retrieval augmented generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Singapore, 2023, pp. 7969–7992, doi: 10.18653/v1/2023.emnlp-main.495. [Online]. Available: https://aclanthology.org/2023.emnlp-main.495/
17. S. Lynch, "Davos 2024: Six takeaways on the AI conversation at WEF," Stanford University Human-Centered Artificial Intelligence, Stanford, CA, USA, Jan. 30, 2024. Accessed: Feb. 29, 2024 [Online]. Available: https://hai.stanford.edu/news/davos-2024-six-takeaways-ai-conversation-wef
18. J. D. Weisz, M. Muller, J. He, and S. Houde, "Toward general design principles for generative AI," in *Proc. ACM IUI Workshops*, Sydney, Australia, 2023, pp. 1–14. [Online]. Available: https://ceur-ws.org/Vol-3359/paper14.pdf
19. S. Vartan, "Racial bias found in a major health care risk algorithm," *Sci. Amer.*, Oct. 24, 2019. [Online]. Available: https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/
20. P. Hacker, A. Engel, and M. Mauer, "Regulating ChatGPT and other large generative AI models," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, 2023, pp. 1112–1123, doi: 10.1145/3593013.3594067.
21. J. Wolff, "How to improve cybersecurity for artificial intelligence," Brookings Inst., Washington, DC, USA, Jun. 9, 2020. Accessed: Feb. 29, 2024. [Online]. Available: https://www.brookings.edu/articles/how-to-improve-cybersecurity-for-artificial-intelligence/
22. V. Goda. "AI Security: Not your usual security lens." Medium. Accessed: Feb. 29, 2024. [Online]. Available: https://medium.com/google-cloud/ai-security-not-your-usual-security-lens-ad830a8b6500
23. R. Ohayon and Y. Altevet, Interviewees to CEO & CTO, Feb. 15, 2024.
24. M. Campbell and M. Jovanovic, "Conversational artificial intelligence: Changing tomorrow's health care today," *Computer*, vol. 54, no. 8, pp. 89–93, Aug. 2021, doi: 10.1109/MC.2021.3083155. [Online]. Available: https://www.computer.org/csdl/magazine/co/2021/08/09504493/1vJVyjU38bK

MARK CAMPBELL is the chief innovation officer at EVOTEK, San Diego, CA 92121 USA. Contact him at mark@evotek.com.

MLAĐAN JOVANOVIĆ is an associate professor of computer science at Singidunum University, 11000 Belgrade, Serbia. Contact him at mjovanovic@singidunum.ac.rs.