

CALYPSOAI

# The Generative AI Policy Handbook





# Introduction

Organizational adoption of artificial intelligence (AI) tools is growing rapidly. Since the debut of ChatGPT in November 2022, large language models (LLMs) have been shown to improve productivity and enhance efficiency by streamlining labor-intensive processes and boosting innovation, creativity, and output. A recent [S&P Global survey](#) of 1,500 decision-makers at large companies found that 69% have at least one AI/ML project in production with 28% having reached enterprise scale with the project “widely implemented and driving significant business value;” 31% have projects in pilot or proof-of-concept stages.

At CalypsoAI, we’ve helped multinationals across insurance, publishing, finance, and other sectors refine their thinking around use of generative AI (GenAI). The experience has allowed us to distill the operational mindset of organizations considering GenAI usage into three distinct tracks:

- 1 The Die-Hard:** They’re going to block it until they know a lot more about it.
- 2 The Vanguard:** They’re establishing proactive deployment plans for adoption.
- 3 The Caught Off-Guard:** Employee usage drives use; deployment plans are TBD.

We have coalesced these intensive and ongoing conversations and observations into this concise guide to help your organization—whichever track it’s on—to safely, securely deploy GenAI models.



# Overview

The enthusiastic adoption of GenAI technology and its impact on the enterprise took the business world by surprise, but the risks accompanying the technology should not have. Security-conscious organizations routinely prohibit connecting external or portable devices to company devices and monitor outbound emails for content or attachments considered proprietary, sensitive, or otherwise controlled information. LLMs, however, were released for general use and instantly adopted by many companies without any such safeguards in place, meaning a user—your employee—could include a sensitive document or confidential information in any message sent to the model.

While many of those early-adopter companies very quickly stopped using LLMs and the model providers released updated versions that included basic guardrails, the situation is still far from ideal because banning a tool like ChatGPT or BERT from being used in your organization is like standing on the beach and trying to stop the tide. It's futile, unproductive, and a bit absurd.

If the dangers of misuse by an employee weren't enough to keep security professionals up at night, the very introduction of GenAI models across an organization can have its own additive effect by contributing to "[digital sprawl](#)." This phenomenon occurs when discrete tools and solutions, including Software as a Solution (SaaS), external base, and Retrieval-Augmented Generation (RAG) models that are integrated with models, proliferate across an organization's digital infrastructure, but don't integrate effectively or at all.

## Deploying GenAI solutions requires due diligence and planning.

This misalignment causes a slow, but significant, expansion of the organization's attack surface that increases risk, especially when the new additions have not been vetted for [security or compliance](#) alignment. Having a solid, fact-based understanding of where your organization is on the AI security preparedness spectrum is the first step toward creating AI system security, as well as having observability built in across your deployment and use of GenAI.

As when integrating any other new tool into an existing system, deploying GenAI solutions requires due diligence and planning, beginning with a governance framework that functions as a blueprint for continual evolution of risk management across an AI-dependent attack surface. The underpinning of such a framework must include operational and strategic considerations. The following sections provide insight into the risks and available remedies.



# Risk Exposure

As organizations begin to deploy GenAI models across their enterprise and realize the operational benefits, they will expand the implementation of or the number of models they're using, or both. Meanwhile, organizations that aren't using GenAI models yet will be as soon as they see competitors streamline processes and race ahead, thanks to productivity gains and cost reductions.

49%



of the general population uses GenAI, and more than one-third use it daily

50%



of the general population has never used GenAI

52%



of AI users are increasing their usage

64%



of non-users say they would use GenAI if it were safe/secure

While it's a given that any organization of any size and in any industry faces expanded risk exposure when it introduces a component to its networks or other systems, the addition of GenAI models dramatically increases the scope of that exposure, for example:



Code with security vulnerabilities via model responses from LLM copilots, such as GitHub CoPilot or Amazon CodeWhisperer



Reliance on multiple models that don't have security controls



Incorporation of multimodal into models, such as the recent version of ChatGPT that enables voice-to-image, voice-to-text, and voice-to-voice interactions



Prompt injection/jailbreak attacks that direct the model to behave in ways detrimental to the organization's interests, values, or operations



Risks to the digital infrastructure, such as infiltration via malware or spyware; data exfiltration via prompts or system breaches.



The creation, dissemination, or acceptance of deepfakes



Unintentional use of copyrighted material/intellectual property (IP) provided by the model



# Operational Controls

Even when carefully and thoughtfully deployed across an organization, model use will take on a life of its own, expanding rapidly as:

- Executives and managers become more dependent on and accustomed to the financial benefits these tools deliver.
- Employees, contractors, and partners become more comfortable using the tools for streamlining productivity.
- New use cases are identified across business units and the models are deployed to address them.

While all of these evolutions are good for business, they must not be considered as “happy path” events. As noted earlier, the addition of new technologies, such as multiple LLMs or GenAI models, as well as multimodal models, to an organization’s digital infrastructure, particularly without also integrating targeted security protocols, leads to technological sprawl and increased vulnerability to exploitation on many levels. Incorporating the following tasks into your organization’s GenAI action plan will help identify vulnerabilities, determine necessary solutions, and ensure the deployment goes smoothly.

## Know Your Attack Surface

The first step in establishing a security plan is to have a thorough understanding of the places where your organization is deploying AI, including access controls and internal permissioning for solutions that allow for it. Such an understanding should include LLMs embedded in your SaaS applications and in the RAG models in use across your organization. Taking the time to assess and document the models in use will provide a complete portrait of your organization’s potential attack surface, which will enable your security personnel to understand what sort of tools and protocols they will need to deploy to eliminate or protect vulnerabilities. Conducting periodic audits of the models will help ensure currency and enable faster identification of old or unused AI tools that could present vulnerabilities in the future.

## Choose Which GenAI Solutions to Incorporate

The number of LLMs available to organizations has proliferated at orders of magnitude since the launch of OpenAI's monolithic ChatGPT, with large rivals to appearing almost overnight, such as BERT (Google), LLaMA (Meta), and Claude (Anthropic). Very shortly thereafter, LLMs tailored for specific industries appeared, such as BloombergGPT (finance), Harvey (law), and Med-PaLM2 (medicine). More recently, private LLMs have been gaining rapid popularity as organizations want to leverage the power and scope of the large foundation models, but include detailed, proprietary information in a sequestered environment not accessible to outsiders and have greater control over security, user behavior, and analytic derivatives.

Another public/private capability newly emerging in the enterprise space is the integration of LLMs with plugins or integrated into SaaS applications, for instance, Slack, Salesforce, and Bloomberg, to enhance the customer experience on one hand and enable increased productivity by incorporating reinforcement learning from human feedback (RLHF) on the other.

Although the options for which LLMs to deploy and how best to do that keep expanding, the decision for each industry and each enterprise within an industry will be based on its own business use cases and how incorporating LLMs into the mix will solve for those. Some well-established, somewhat universal use cases include:



Creating detailed reports from raw data



Translating text or audio into different languages in real time



Generating new text, images, videos, or sounds from prompts provided in the same or different format



Devising new materials, designs, drugs, or other products



Developing new protocols for treating diseases or other physical issues

A cross-functional, in-house team of stakeholders and users from every level that will be affected by the use of GenAI models can establish a hierarchy of technical and operational needs, and use that as the basis for a governance framework, as discussed in the next section.

## Establish a Governance Framework

After the technical architecture for hosting the models has been addressed and the model(s) have been selected, the plan for deploying and using the model(s) must be crafted. The simplicity or complexity of your framework must be decided by the team identified in the previous section, and it must begin with the philosophical aspects of governance—policies and principles—and move to the practical—technical controls.

Alignment with core organizational principles and existing policies is key, although some might need updating or rewriting altogether. Behavioral guidelines and acceptable use standards are also important. In a nutshell, the framework must set both the organization's expectations and the users' expectations, and explain to users what the tools are for, what they are expected to do, and how they are to be used.

The simplicity or complexity of your framework must be decided by a cross-functional team of stakeholders and users.

Issues that organizations must address via [policies](#) include:

- Prompt injection/jailbreak attempts (attempts to generate discriminatory, illegal, unethical, or harmful content in responses)
- Prompts sensitive/confidential information, including IP, that would cause damage to the organization financially, reputationally, or another way, were it made public
  - Any content in a prompt can become part of the model creator's knowledge base and be used to train future iterations of the model, which means it could appear in a competitor's search results
  - "Benign" content that could be misused in the future or aggregated to create a picture of private corporate plans or executive movements, such as calendars, summaries of emails or reports, outlines of meeting agendas, travel itineraries, or first drafts of policy, procedure, contractual, or other documentation
- Accountability and oversight mechanisms, including human verification of generated content to determine accuracy, timeliness, factualness, and relevance, as well as plagiarized or copyrighted content used without the copyright holder's permission
- Usage monitoring and auditing, which can provide insight into organizational usage, cost, and other metrics
  - Some monitoring mechanisms enable tracking employee or customer engagement, and can discern their emotional state or involvement during the interaction or identify protected status related to disabilities or other privileged or private information, and might be prohibited by law
  - Whether, how, and for what purposes monitoring should occur
- Alignment to/support of company goals and core values, including how using the model will improve operations, company culture, productivity, efficiency, and revenue
- What happens when things go wrong



## Educate Your Employees

Frameworks, policies, and documentation will not help an organization that does not share that information with its people. While training methodology will vary according to organizational culture and practices, elements of the program that should be non-negotiable include:

### Make it **mandatory**

Everyone in the company must take the training. No one gets a free pass.

### Make it **meaningful**

Include actual use cases for the organization. Present the risks and business consequences realistically and seriously.

### Make it **memorable**

Use the method that works best for the organization. Make it an on-going program, rather than a once-a-year effort, to ensure security remains top of mind.

## Encourage Feedback and Reporting

Such information can alert the security team about vulnerabilities or other issues that were not foreseen, operations teams about additional use cases, or other teams about additional features or benefits that can be optimized. Establishing a channel for employees to share opinions about any element of the process or the model itself is critical to any roll-out. Taking time to review and, if appropriate, act on the feedback is just as important.

## Identify Consequences

We've known for decades that a single click or keystroke can create a security nightmare, a public relations disaster, or a financial calamity, or a combination of all three. The added layer that occurs when GenAI models are involved is that the content at the heart of the issue—such as a [video or audio recording](#) or a realistic email or photograph—could be completely fabricated. It no longer takes expertise to create deepfakes that are difficult to detect; bad actors with basic skills can use GenAI to create [vocal reenactments](#) of real people “saying” things they never said, to swap faces or have software manipulate faces to make it seem like people are saying or doing things they did not, and to do many other malicious acts.

The repercussions of having plausible, or even incredible, allegations made with “proof” that doesn't appear to be fake can be painful and expensive. Cyber and AI security professionals must plan to respond, rather than react, to the new and emerging destructive capabilities of these models and identify means of working to prevent their intrusion into the organization.



# Technical Controls

When the operational issues have been addressed, delving into the technical controls is going to seem like the easy part. Don't be fooled. This is an evolving area of concern as new security controls are devised and put in place, and then hacked or overridden by threat actors who live for the opportunity to take down anything that thwarts their livelihood. Let's look at this chronologically from the user's perspective.

## Access Controls

Security-aware organizations have likely already implemented rigorous access controls, such as multi-factor authentication (MFA) and zero trust, for their AI-dependent systems. These controls are strong, but they have their limitations, particularly when the protocol in place allows any user who gains access to the system to also have access to all applications. Segmenting systems, in which certain applications are set apart and require additional steps to gain access, is one strong solution to the "everyone everywhere" situation.

However, deploying tools that feature built-in role-based access controls (RBAC) is another, even stronger security solution, especially for situations in which organizations use multiple models across the enterprise.

When operational controls have been addressed, the technical controls will seem like the easy part. Don't be fooled.

CalypsoAI's groundbreaking LLM security solution, Moderator, applies a "wraparound" trust layer that encompasses all LLMs used in an organization, while providing administrators the ability to establish access for each user, as well as for teams of users. The clean, simple interface allows users to interact fully with the model or models available to them, while providing no visibility into the models or features to which they do not have access. In addition to providing secure permissioning, our RBAC feature is integrated with Moderator's tracking and auditing capabilities to provide administrators full visibility into cost, content, and overall user engagement.

## Prompt and Response Scanning

After a user is authenticated and using the LLM(s), the vulnerabilities inherent in large LLMs, such as ChatGPT and others, don't disappear, but they do become more challenging to thwart. The following scenarios show the importance of scanning both outgoing prompts and incoming responses for any content that could exploit a vulnerability, invite a threat, or otherwise put the organization at risk.

- An executive assistant includes the name of a company targeted for an acquisition and their boss' cellphone number in a prompt to condense meeting notes into meeting minutes. While the employee saved time crafting the minutes, everything in the prompt—the meeting topic, attendees, agenda, and the private information—went outside the organization's system, into the model provider's knowledge base, and potentially into the training data for Model 2.0. If that happens, competitors, threat actors, and everyone else using the right search terms will have access to it.
- A software developer under deadline to create complex code for the organization's new, ground-breaking product pastes problematic command lines into a prompt and instructs the model to identify and resolve the issues. As in the previous scenario, the proprietary information leaves the organization and becomes part of the model provider's knowledge base, but the risks don't stop there. The developer, upon receiving the model's newly generated code in seconds, doesn't review or test it before uploading it to the company's code library. When the next build is compiled, the possible outcomes are:
  - The model-generated code is fine and works perfectly.
  - The model-generated code is glitchy or otherwise of poor quality, and causes issues downstream.
  - The model-generated code contains unnoticed vulnerabilities or malicious instructions and infects the library, the product, and possibly the system on which the code is running.
- A bored employee deliberately crafts prompts that return responses with discriminatory, biased, or other content outside of the acceptable-use policy. The prompt content, now part of the provider's knowledge base, is associated with the organization from which the prompt originated and could be seen by future users, potentially creating reputational damage.

All of these instances point to a real need for technical controls on outgoing content. Tools that scan prompts for recognized terminology (RegEx) that is inappropriate in most workplaces, for source code, or for sensitive, private, or personal data are critical. But cleverly (or cluelessly) written prompts with no ill intent could generate responses containing content that would never be approved for use in the organization, such as malicious source code or discriminatory or other toxic material, which means incoming content must be scanned, as well.

CalypsoAI's [Moderator](#) solves for all of these issues by ensuring all user prompts and LLM responses are reviewed by robust, automated scanners applying customized, administrator-defined criteria. The scanners identify source code, sensitive and personal data, toxicity, bias, and other content misaligned with the organization's acceptable use policy and other values, and prevents prompts containing them from being submitted without revision. Our solution also scans incoming responses for malicious code and other administrator-identified content, and provides the option to require human verification of any or all content returned by the model.

## Observability

Few models provide features that enable visibility into model usage or other metrics. Given the costs of using GenAI models, understanding who is using the model(s), how often, and for what purposes can help with resource allocation; knowing which users are routinely flagged for including inappropriate or otherwise prohibited language or content in prompts, or making requests for inappropriate or prohibited content to be returned in a response, could initiate HR involvement. Insights into user sentiment can also be valuable, where allowed.

Model-agnostic and scalable across the enterprise, CalypsoAI Moderator is the first solution to provide a safe, secure research environment and fine-grained content insights with no effect on response speed. With the capability to provide coverage for all LLMs in use, Moderator provides end-to-end observability for user engagements. Chat histories are fully tracked for usage, content, and cost, are fully auditable, and can be purged on a pre-set cadence, if required by policy or regulation. Administrator actions are also logged and available for review, enabling 360-degree visibility and accountability across the enterprise.



# Conclusion

The list of possible worst-case scenarios induced by incorporating multiple LLMs across an organization could go on and on. Candidly, if a human can think it, a human will try it. But, by default or maybe by inclination, cyber and AI security professionals know bringing LLMs in-house represents a challenge and an opportunity to say Yes! to deploying GenAI, not a reason to say No, never!

Only the organization using a model can determine how it should and should not be used or when it should and should not be used; every organization has different needs and different goals. Developing a [governance framework](#) that identifies those needs and goals, and lays out a roadmap for meeting them using the capabilities LLMs can provide, is critical to achieving those results.

Vulnerabilities, both in the model itself and those initiated by the inclusion of the model into the organization's digital infrastructure, must also be identified and addressed. Few LLM providers have integrated strong security tools into their models, but external solutions, such as Moderator, exist to cover the scope of issues that such models and their inherent vulnerabilities present.

Providing protection at scale without creating or exacerbating latency, privacy, or security issues is achievable, which can enable your organization to deploy GenAI models safely, securely, effectively, and efficiently in real time. CalypsoAI Moderator can provide the peace of mind that decision-makers need to greenlight the safe, secure, ethical use of LLMs across the enterprise.



## ABOUT US

CalypsoAI is the leader in developing and delivering AI security solutions. The company's vision is to be the trusted partner and global leader in the AI security domain, empowering enterprises and governments to leverage the immense potential of GenAI solutions and LLMs responsibly and securely. CalypsoAI is striving to shape a future in which technology and security coalesce to transform how businesses operate and contribute to a better world. Founded in Silicon Valley in 2018 by top minds in the fields of artificial intelligence, data science, and machine learning, the company has secured backing from investors including Paladin Capital Group, Lockheed Martin Ventures, Lightspeed Venture Partners, 8VC, and Hakluyt Capital.

To learn more, visit our [website](#) or follow CalypsoAI on [X](#) and [LinkedIn](#).

CALYPSOAI