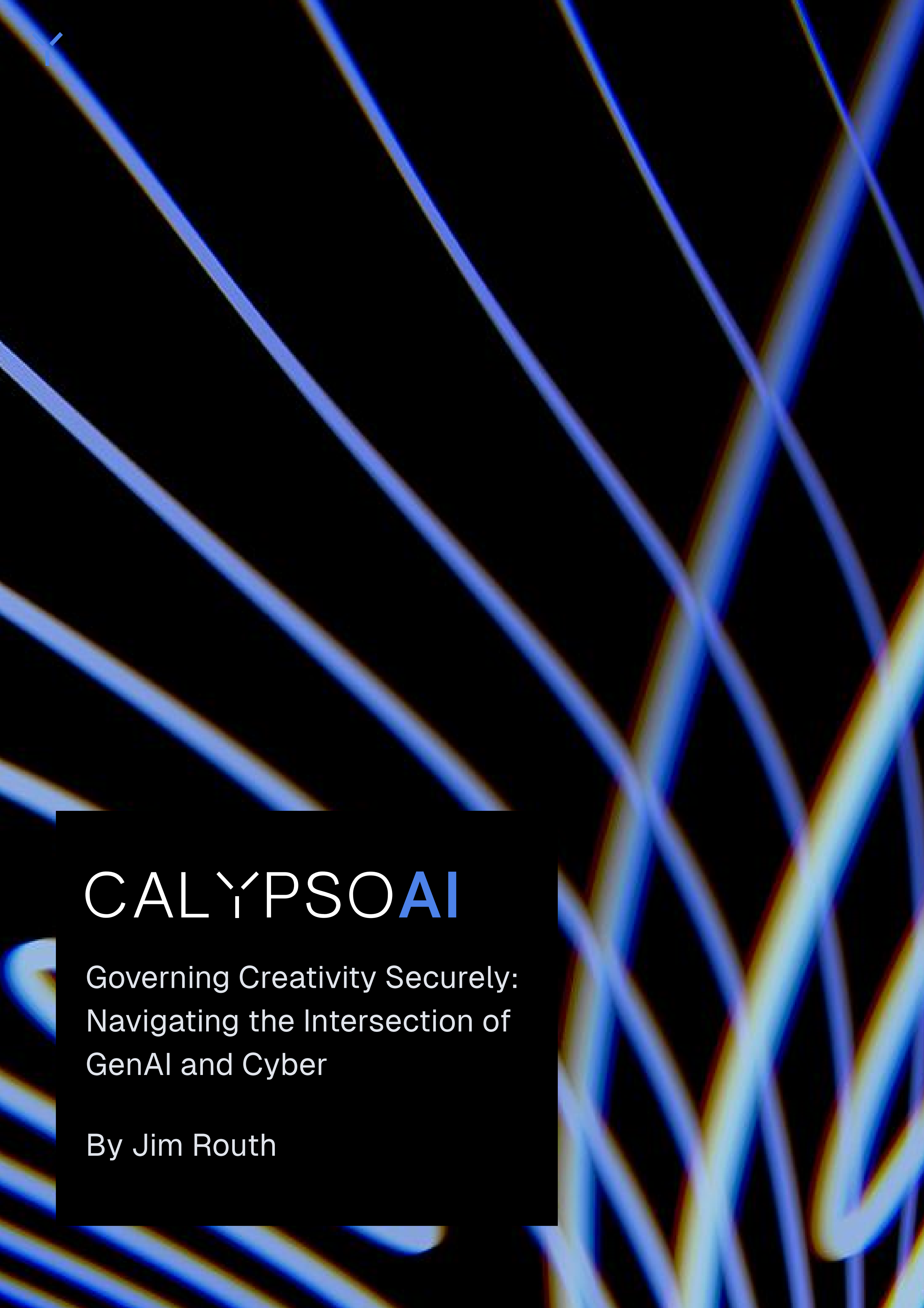


CALYPSOAI

Governing Creativity Securely: Navigating the Intersection of GenAI and Cyber

By Jim Routh

The Evolution of GenAI

Novel innovations from disruptive technologies, for instance web apps, mobile apps, SaaS, blockchain, Internet of things (IoT), and augmented reality (AR) and now GenAI, spark excitement, enthusiasm, and unbridled optimism toward the exponential opportunities to change the way we work. In one short year, the evolution of the attack surface for GenAI, in addition to the primary use cases, have fundamentally changed from using large, publicly hosted large language models (LLMs) to smaller proprietary and customized LLMs embedded in applications and open-source software components. Last year, the top risk was information leakage into public LLMs; today, it is the risk of inaccurate information (hallucinations) feeding business processes designed for and reliant upon accurate data.

Defining AI

The exuberance surrounding the possibilities and potential of GenAI generates the feeling within enterprises that anything is possible as long as we dive into using GenAI anywhere and everywhere. The reality is that, just like disruptive technology innovations in the past, there are new risks that must be recognized and managed as part of the learning process. A good starting place is a common definition—”definitions” might actually be more accurate—of AI.

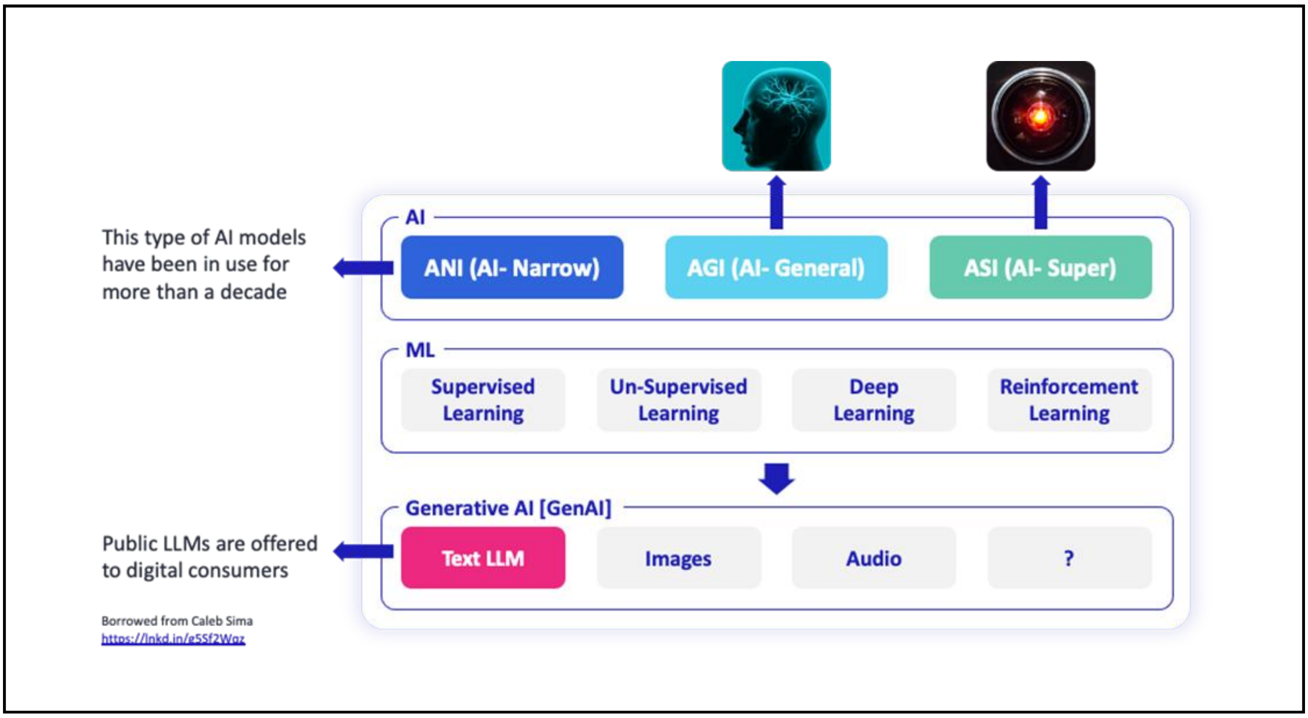
ANI (Artificial Narrow Intelligence or AI Narrow) denotes goal-oriented or task-specific versions of AI designed to better perform a single task, such as tracking weather updates, facial or speech recognition, or playing games such as poker, chess, etc.[1] This type of AI has evolved and matured within enterprises in the past decade and includes ML algorithms that permeate business processes in the majority of large enterprises.

AGI (Artificial General Intelligence) denotes as-yet unrealized AI tools that will be able to function with human-like intelligence, self-teach, and execute tasks they were not trained for or intended to perform.

ASI (Artificial Super Intelligence) denotes theoretical tools that will be self-aware and have capabilities that exceed human capacity.

[1] Vijay, Kanade, What is Narrow Artificial Intelligence (AI)? Definition, Challenges, and Best Practices for 2022, *Spiceworks* (March 22, 2022) https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-narrow-ai/#:~:text=Artificial%20narrow%20intelligence%20(ANI)%20is,as%20poker%2C%20chess%2C%20etc.

The graphic below shows where GenAI fits into this scenario.



This type of AI models have been in use for more than a decade

Public LLMs are offered to digital consumers

Borrowed from Caleb Sima
https://lnkd.in/e5Sf2Wqz

Figure 1: Generative AI

GenAI is the more recent development of LLMs, many of which are text-based today, but are rapidly evolving into "multimodal" models that include images and audio files in their datasets and their output. These models pose significant challenges to artists, songwriters, and other content creators and celebrities who are interested in protecting their intellectual capital (films, videos, music, etc.). There are specific use cases where LLMs shine in creating results quickly, and other use cases where other types of ML systems fit nicely.

Governing AI

The key for any enterprise is to identify its potential use cases that have upside potential and are worthy of investment capital or expense for development, while also defining the appropriate controls required to meet customer and regulator expectations in the near future.

This represents a bit of a balancing act for establishing a GenAI governance model, which requires expertise from several dimensions/organizations that need to work collaboratively to yield favorable results. The graphic below illustrates one approach to supporting a GenAI governance model at enterprise scale being implemented today. The sections that follow provide details about this approach.

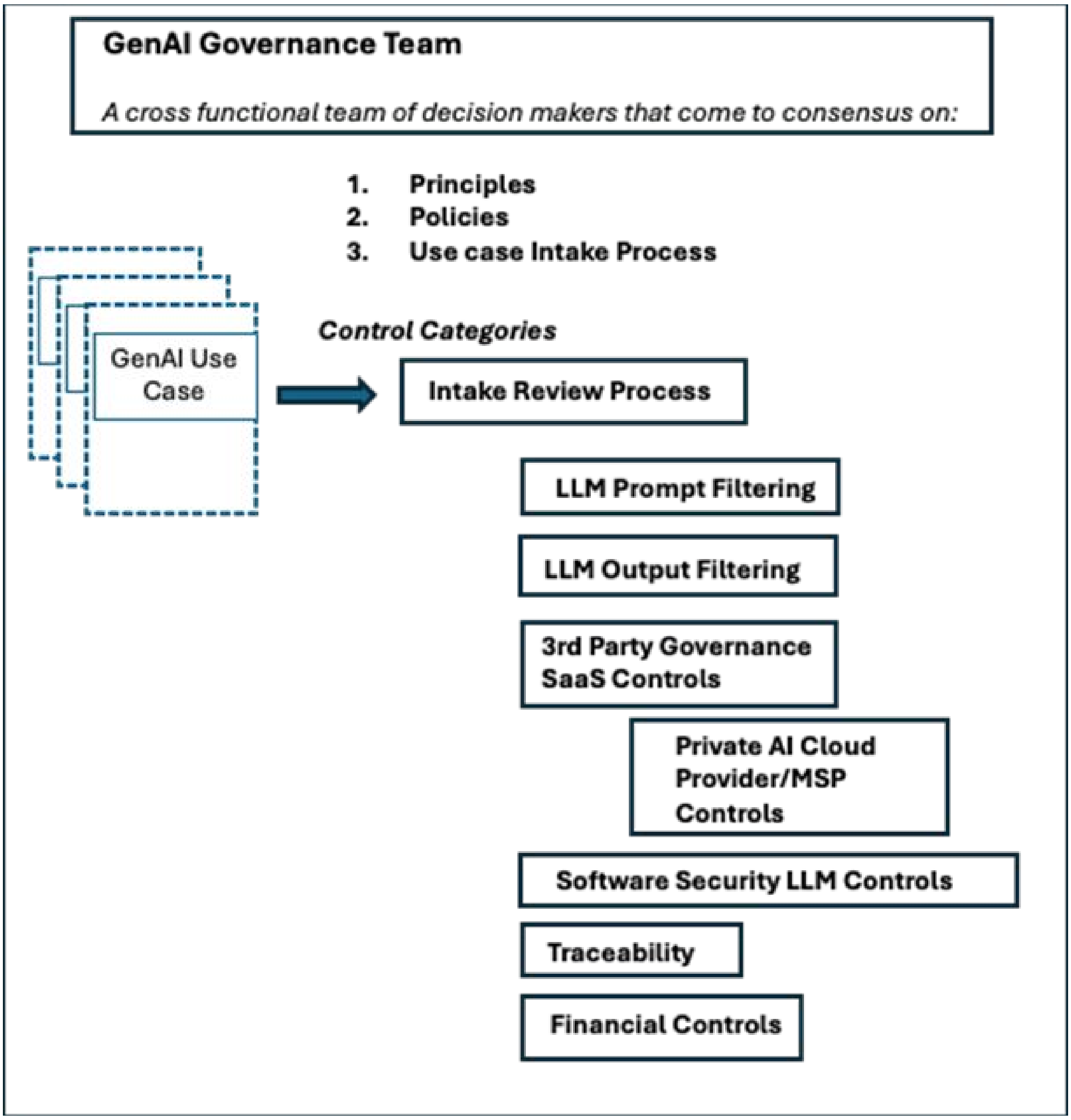

GenAI Governance Team
A cross functional team of decision makers that come to consensus on:
1. Principles
2. Policies
3. Use case Intake Process
GenAI Use Case
Control Categories
Intake Review Process
LLM Prompt Filtering
LLM Output Filtering
3rd Party Governance SaaS Controls
Private AI Cloud Provider/MSP Controls
Software Security LLM Controls
Traceability
Financial Controls

Figure 2: Generative AI Governance

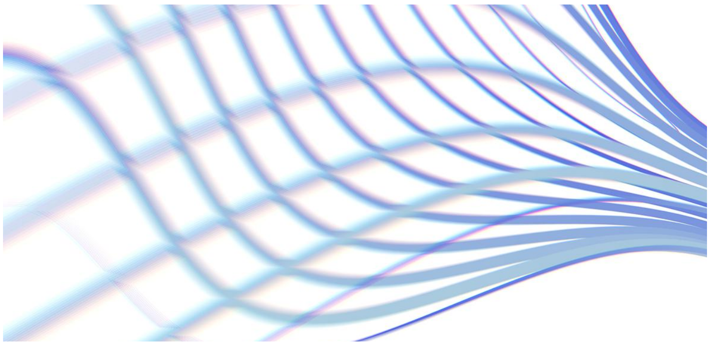
- ## The Team

The easiest way to start an AI governance program is to form a cross-functional team of decision-makers that will be led by someone with facilitation skills to drive a consensus-based decision process resulting in principles for the enterprise.

Because securing GenAI is not merely a technical challenge, but a strategic imperative, I recommend that the Chief Information Security Officer (CISO) lead the cross-functional team in creating governance principles, policies, and practices specific to GenAI governance. It requires meeting facilitation skills to achieve consensus from multiple points of view, but the results are more sustainable policies and practices. If other leaders demonstrate the ability to drive consensus, then by all means encourage them to facilitate meetings.

Selected participants will likely represent the following disciplines:

- Data science
- Legal/Privacy
- IT architecture
- Cybersecurity
- Customer communications
- Human resources
- Finance

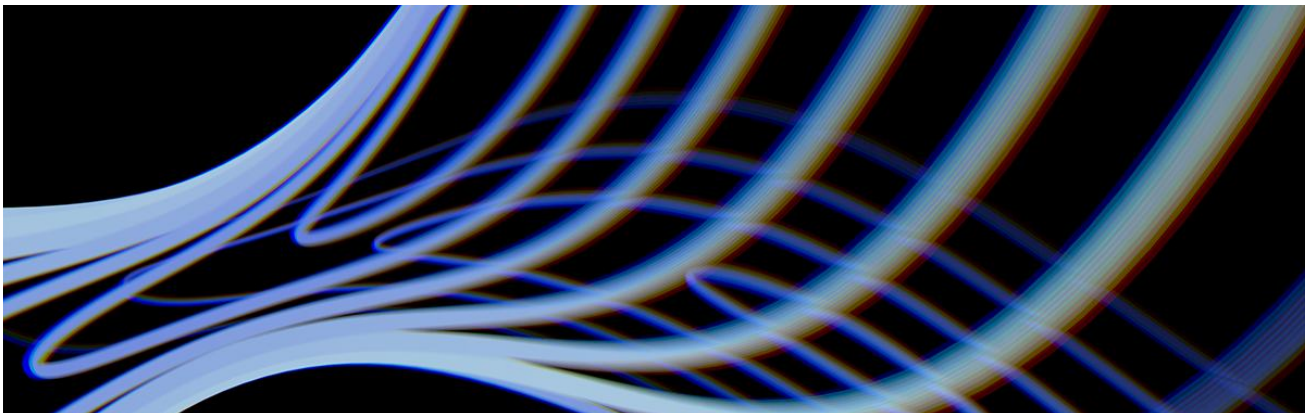


Every enterprise is different and how best to invest in innovation using LLMs is not likely to be the same from one enterprise to another. The level of additive controls necessary to meet governance needs will also vary depending on the characteristics of the use cases selected.

The Structure

Principles are relatively easy to create and provide high-level guidance for determining potential use cases and the types of opportunities that are most attractive to the enterprise and may be core to a strategic direction. The following list presents sample principles.

1. The enterprise will never use an LLM to process sensitive customer account information.
2. The enterprise will seek to use LLM capabilities to improve a quality work experience for employees doing repetitive and tedious work today.
3. Accountability for the resilience of systems using LLMs rests with the system designers and implementers.



The next step for the governance team is to come to consensus on policies specific to GenAI governance for the enterprise, such as:

1. Project teams seeking to use/build/integrate LLMs within systems must prepare a use case description that will be fed into the GenAI governance intake process for review.
2. System developers are accountable for the resilience of ML systems and GenAI models (private or foundational) in development and production.
3. GenAI models utilized by enterprise users must have a traceability capability in place.

The Process

Having an intake process implemented enables enterprises to more quickly capitalize on innovation, while also providing risk management guidance specific to the use case. A use case template is shared with the GenAI intake process and reviewed by each governance team member. The members meet weekly to share observations and make decisions on whether a use case must include more information, requires specific controls, and/or is approved for development. Once use cases are implemented, the respective teams capture lessons learned through the process to benefit other use cases going forward.

Control capabilities must continue to evolve alongside innovative and emerging technologies like GenAI, and this approach to implementing GenAI governance is important to that end. This approach helps build control capability over time and improve governance processes as the understanding of how to deploy the technology to ensure the right outcomes matures. Innovation in commercial enterprises offers an opportunity to experience small failures, pivot, and redesign, which ultimately results in higher quality outcomes over time. Lessons learned by one team/project can be quickly applied to another team that will benefit from the experience, good or bad.

One of the opportunities offered by GenAI capabilities that enterprises can take advantage of is to deploy products embedded with LLMs, which reduces the need for highly skilled labor to staff a Security Operations Center (SOC) or Identity and Access Management (IAM) operations center while significantly improving outcomes. AI/ML systems do this today and cybersecurity teams will have more and better choices available to them as products evolve to deploy GenAI-enabled capabilities to protect the enterprise more efficiently.

The Controls

Here are a few general thoughts for CISOs to consider regarding control design specific to GenAI governance. Several CISOs have told me that there is nothing special about GenAI that can't be handled with conventional and well-established cybersecurity control frameworks. I believe this is fundamentally wrong. GenAI represents different governance challenges from ML models or cloud infrastructure as a service, and to believe that existing controls are sufficient to meet these challenges at enterprise scale is too narrow a viewpoint.

Observability and Traceability

Once an enterprise has aligned on principles, a few policies, and an intake process for use cases, the next step is to address the need for traceability/observability. This essentially means capturing log files showing how LLM access is in place for different use cases and types of models (foundation models vs. proprietary or private models). The log files can be integrated into a Security Event and Incident Management (SEIM) platform, a data lake, or be part of a selected governance platform acquired for LLM use cases.

Observability is foundational to understanding the evolution of use cases and responding to the evolution of regulatory requirements at multiple levels (local, state, country, industry, global). Enterprises can purchase a platform designed specifically for traceability and require it to be a front-end to LLM usage options, or enterprises can modify existing data loss prevention (DLP) platforms to control access to foundation models and SaaS applications that use LLMs. The key is for the enterprise to understand how people are getting access to LLMs, how they are using that access, and, ultimately, how to enforce the right level of access accountability across an enterprise.

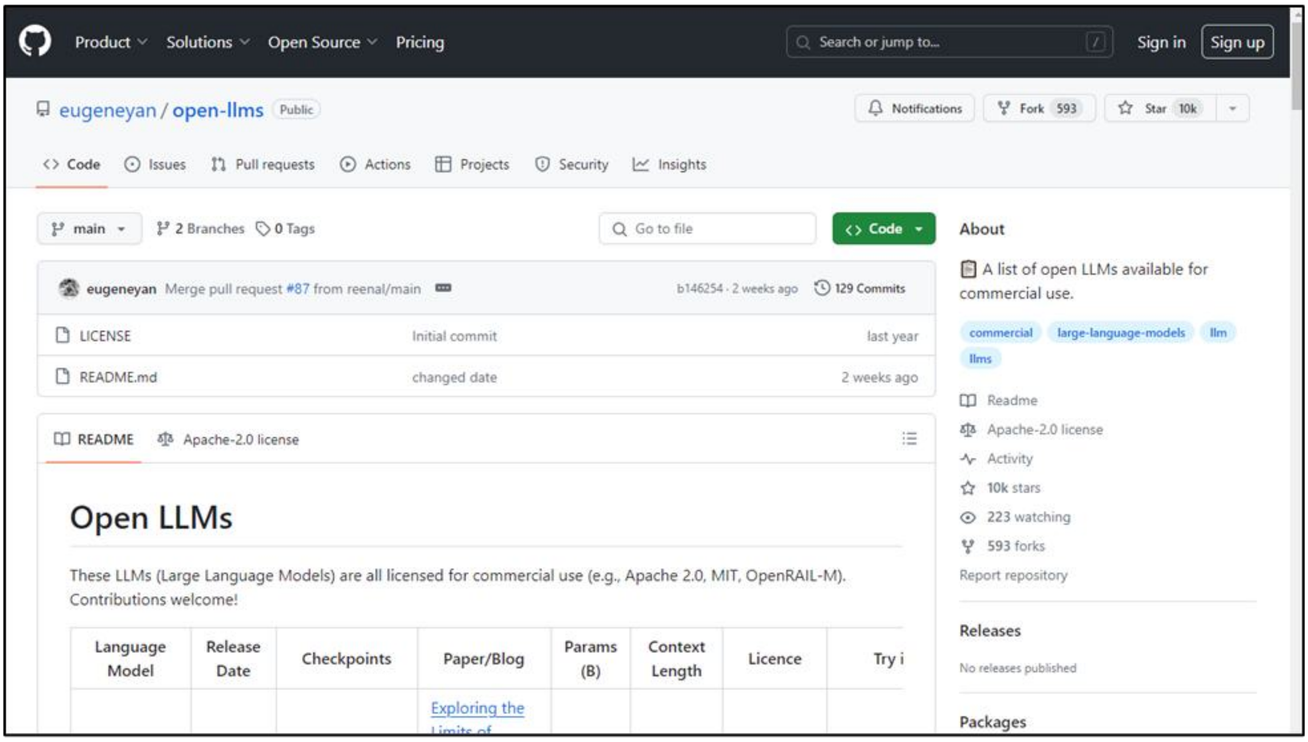
Vendor Interactions

Every enterprise will have some sort of third-party governance controls in place when using vendors for services and products, but the likelihood that these controls are sufficient for GenAI is very low. The specific focus should be on how

SaaS applications are selected and licensed for enterprise users. Adding a control to identify the use of LLMs embedded in the SaaS applications is the first step to enabling traceability. Commercial products providing third-party governance today may offer a way to identify LLMs within SaaS applications and enterprises using these products should request this feature if it is not provided. At a minimum, knowing an LLM is in use and knowing the number of its users provides the enterprise with a foundation for building effective controls.

- ## Software Development/DevOps

Enterprises that build and acquire software components today for their own use or to create commercial software have implemented techniques to control the software development pipeline, understand what open standard software components are in use, and ultimately create a software bill of materials (SBOM) to manage the risk of security vulnerabilities in any of the software components. There are a few software pipeline management tools that can identify software components using LLMs today, providing the enterprise an opportunity to add enforceable observability into the software build process. The vast majority of enterprises will need to add this capability to the build process in the near future. Software developers that use a code repository for managing the components they integrate have lots of choices via the repository to select and use LLMs, as shown below in the image of a GitHub repository:



eugeneyan Merge pull request #87 from reenal/main b146254 · 2 weeks ago 129 Commits

LICENSE	Initial commit	last year
README.md	changed date	2 weeks ago

Open LLMs

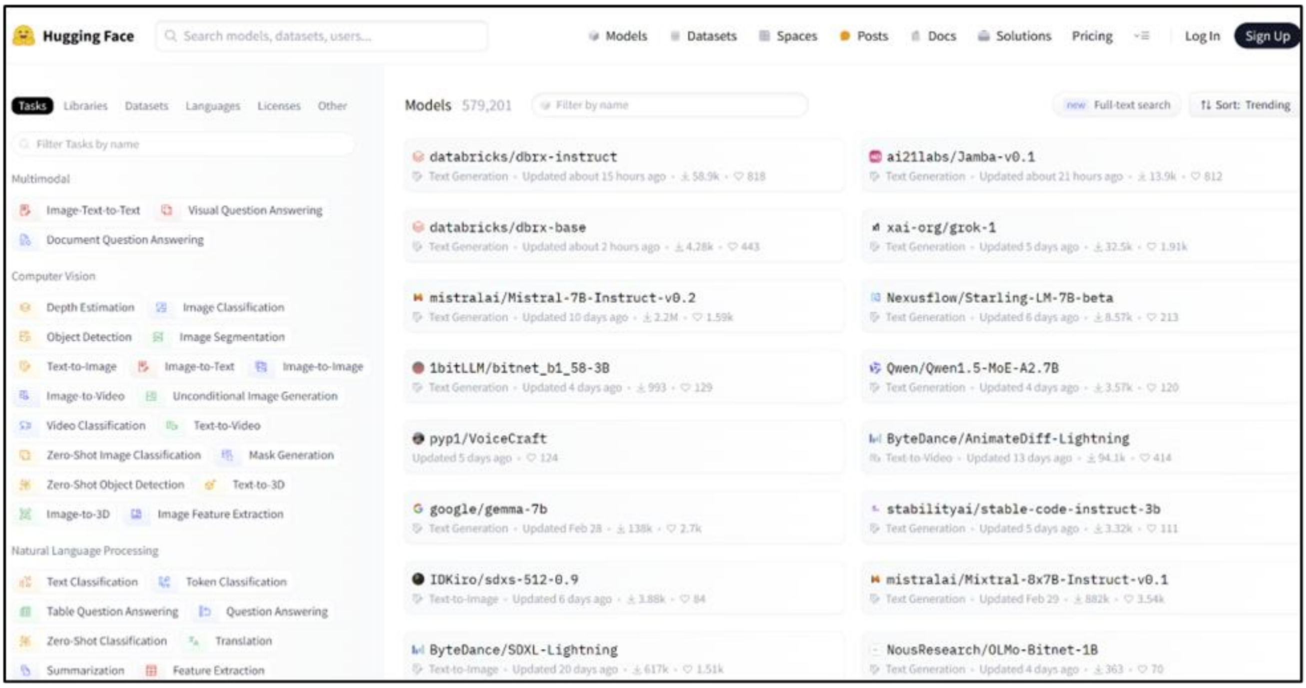
These LLMs (Large Language Models) are all licensed for commercial use (e.g., Apache 2.0, MIT, OpenRAIL-M). Contributions welcome!

Language Model	Release Date	Checkpoints	Paper/Blog	Params (B)	Context Length	Licence	Try i
			Exploring the Limits of				

A list of open LLMs available for commercial use.

Figure 3: GitHub Repository

Web sites such as Hugging Face, shown below, provide developers with access to hundreds of special-purpose LLMs and other open-source tools and utilities for managing the integration of LLMs in applications. The easy and plentiful availability of this externally created code increases the organization’s exposure by expanding the potential attack surface for cyber criminals wishing to exploit the enterprise's use of GenAI for fraud or other purposes.



Multimodal

Image-Text-to-Text · Visual Question Answering · Document Question Answering

Computer Vision

Depth Estimation · Image Classification · Object Detection · Image Segmentation · Text-to-Image · Image-to-Text · Image-to-Image · Image-to-Video · Unconditional Image Generation · Video Classification · Text-to-Video · Zero-Shot Image Classification · Mask Generation · Zero-Shot Object Detection · Text-to-3D · Image-to-3D · Image Feature Extraction

Natural Language Processing

Text Classification · Token Classification · Table Question Answering · Question Answering · Zero-Shot Classification · Translation · Summarization · Feature Extraction

Models 579,201

databricks/dbrx-instruct
Text Generation · Updated about 15 hours ago · ↓ 58.9k · ♡ 818

ai21labs/Jamba-v0.1
Text Generation · Updated about 21 hours ago · ↓ 13.9k · ♡ 812

databricks/dbrx-base
Text Generation · Updated about 2 hours ago · ↓ 4.28k · ♡ 443

xai-org/grok-1
Text Generation · Updated 5 days ago · ↓ 32.5k · ♡ 1.91k

mistralai/Mistral-7B-Instruct-v0.2
Text Generation · Updated 10 days ago · ↓ 2.2M · ♡ 1.59k

Nexusflow/Starling-LM-7B-beta
Text Generation · Updated 6 days ago · ↓ 8.57k · ♡ 213

1bitLLM/bitnet_b1_58-3B
Text Generation · Updated 4 days ago · ↓ 993 · ♡ 129

Qwen/Qwen1.5-MoE-A2.7B
Text Generation · Updated 4 days ago · ↓ 3.57k · ♡ 120

pyp1/VoiceCraft
Updated 5 days ago · ♡ 124

ByteDance/AnimateDiff-Lightning
Text-to-Video · Updated 13 days ago · ↓ 94.1k · ♡ 414

google/gemma-7b
Text Generation · Updated Feb 28 · ↓ 138k · ♡ 2.7k

stabilityai/stable-code-instruct-3b
Text Generation · Updated 5 days ago · ↓ 3.32k · ♡ 111

IDKiro/sdxs-512-0.9
Text-to-Image · Updated 6 days ago · ↓ 3.88k · ♡ 84

mistralai/Mixtral-8x7B-Instruct-v0.1
Text Generation · Updated Feb 29 · ↓ 882k · ♡ 3.54k

ByteDance/SDXL-Lightning
Text-to-Image · Updated 20 days ago · ↓ 617k · ♡ 1.51k

NousResearch/OLMo-Bitnet-1B
Text Generation · Updated 4 days ago · ↓ 363 · ♡ 70

Figure 4: Hugging Face Model Directory

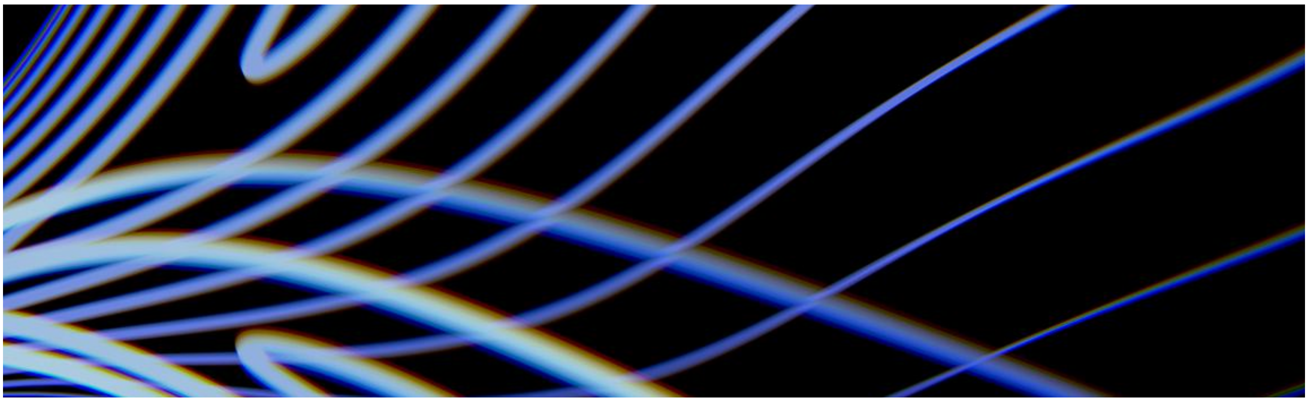
The software development process today provides a wealth of tools and components served up to the DevOps team members to integrate into components that end up in the automated build process. Providing instrumentation to the developers to check/verify the components for security vulnerabilities is the best method for checking integrity of the build process. Fortunately, there are many tool choices available, both proprietary and open source, and some that actually use LLMs to protect against the challenges of LLMs.

The code repositories have a marketplace for exchanging components along with websites that offer open-source components, enabling the developer to operate like a general contractor building the object of choice with prefabricated materials. Using the figure below as an example, presume the red Lego® blocks are LLMs available via private and open-source providers and the blue are open-source components. Without adequate software security governance models in place to identify the appropriate planning, processes, sources, and review or vetting techniques that must be adhered to during the development life cycle, developers might well be just reaching into the mix and using the first one that fits their need at the moment. Scaling that disorganized approach to an in-house DevOps team would result in unstable, unsecure, and untrustworthy models.

Figure 5: LLMs and Open-Source Components

Conclusion

The software development process represents a growing part of the enterprise attack surface for GenAI uses. Applying the right controls enables the DevOps teams to offer the most comprehensive security protection in a sustainable way. In years to come, the growth in use of public LLMs will likely diminish and be replaced by special purpose LLMs embedded in software available through SaaS products or deployed in enterprise applications. Using LLMs in the software development process along with the application of products using LLM technology to both protect the enterprise and generate code will become the norm.

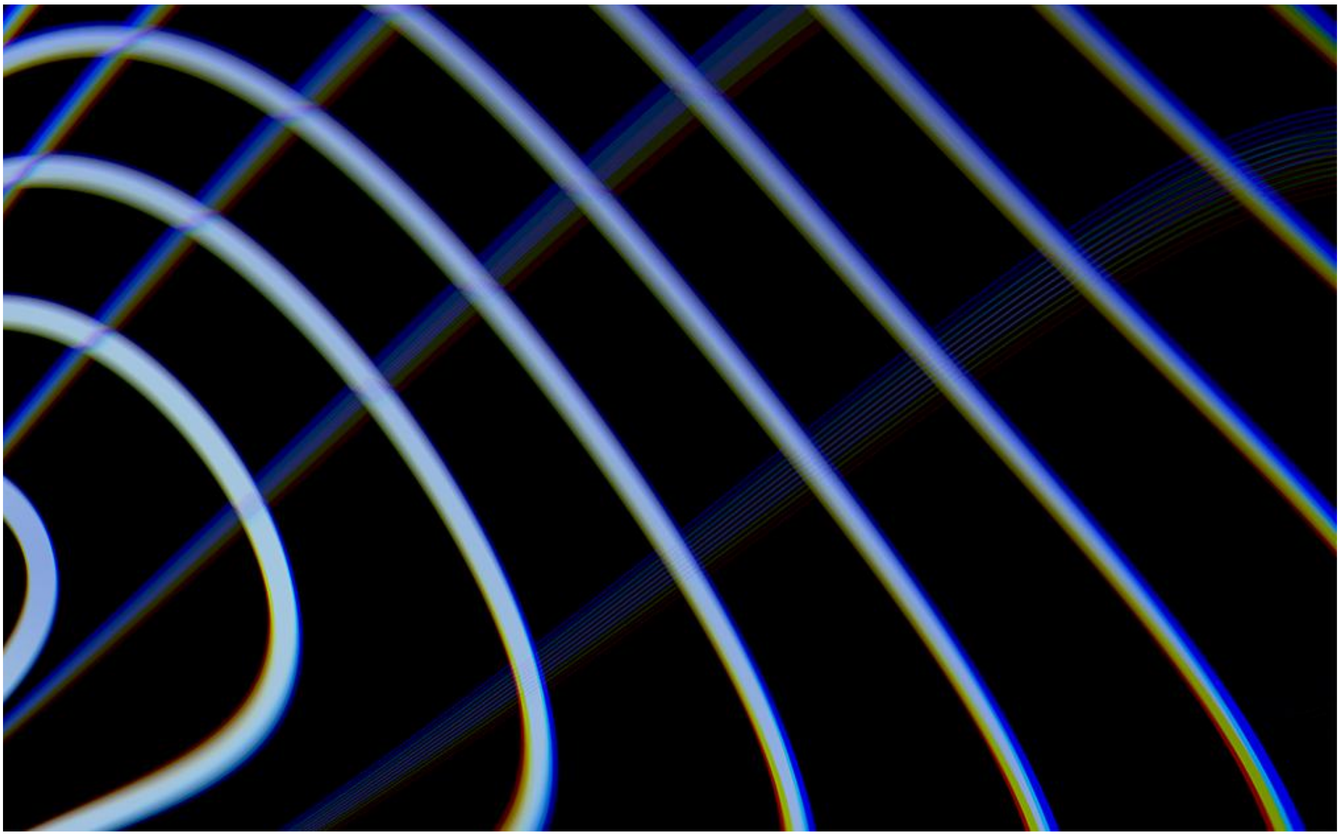


About the Author

Renowned cybersecurity industry expert and thought leader Jim Routh has served as the CISO/CSO for CVS Health, Aetna, KPMG, DTCC, American Express, and MassMutual. He is a member of the Board of Advisors for CalypsoAI.

About CalypsoAI

CalypsoAI is the leader in AI Security and Enablement. As a trusted partner and global leader in the AI Security domain, CalypsoAI empowers enterprises and governments to leverage the immense potential of GenAI solutions and LLMs responsibly and securely. CalypsoAI strives 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, Hakluyt Capital, and Empros Capital. To learn more, visit the website or follow CalypsoAI on X and LinkedIn.



Thank you

CALYPSOAI