Gartner

Innovation Guide for Generative AI in Trust, Risk and Security Management

Published 18 September 2023 - ID G00799579 - 16 min read

By Analyst(s): Avivah Litan, Jeremy D'Hoinne, Gabriele Rigon

Generative AI brings new risks in three categories: content anomalies, data protection and AI application security. IT leaders using or building GenAI apps can use this research to understand market dynamics and evaluate emerging GenAI TRiSM technology and providers that address new risks.

Overview

Key Findings

- Integrating large language models (LLMs) and other generative AI (GenAI) models in enterprise applications bring new risks in three categories: content anomalies, data protection and AI application security.

- Vendors hosting GenAI models do not provide a complete set of controls that mitigate these risks. Instead, users need to acquire solutions that do so to augment hosting vendors' limited controls.

- IT leaders must trust most hosting LLM vendors with protection of their data, without the ability to verify their security and privacy controls.

- The market for GenAI TRiSM solutions is still small, and emerging solutions remain largely untested. It is currently targeted toward LLM usage protection as opposed to protecting usage of multimodal models.

Recommendations

- Set up proofs of concept to test emerging GenAI TRiSM products in the categories to augment your legacy security controls, and apply them to production applications once they perform as required.
- Use content anomaly detection products that mitigate input and output risks to enforce acceptable use policy and prevent unwanted or otherwise illegitimate model completions and responses from compromising your organization's decision making, safety and security.
- Evaluate the use of AI application security products to protect your organization from hackers who exploit new GenAI threat vectors to damage your organization and its assets.
- Continue to use known legacy security controls to protect sensitive information, application stacks and assets, but recognize they don't mitigate risks unique to LLMs, such as inaccurate, inflammatory or copyrighted outputs in responses.

The GenAI TRiSM market is still a very early market subset of the overall AI TRiSM Market (see Market Guide for AI Trust, Risk and Security Management). We define this market and its functions and note some of the vendors who support it in this inaugural Innovation Guide.

Beta Research

The following research is part of a new initiative Gartner is piloting to provide updates at a greater frequency. It is a work in progress that does not represent our final position. While we continue to monitor this topic, we invite you to provide constructive feedback. All relevant updates and feedback will be incorporated into the final research, which will undergo our standard review process.

You can navigate this document using the hyperlinks below.

Market Definition

Back to top

The GenAI TRiSM market comprises multiple software and services segments that support security, data protection and risk mitigation for adopters of GenAI applications and model interactions. GenAI TRiSM tools include solutions for:

- Content anomaly detection
- Data protection
- AI application security

These tools complement associated measures implemented by builders or owners of GenAI models, applications and agents, and as such, represent “shared responsibilities.”

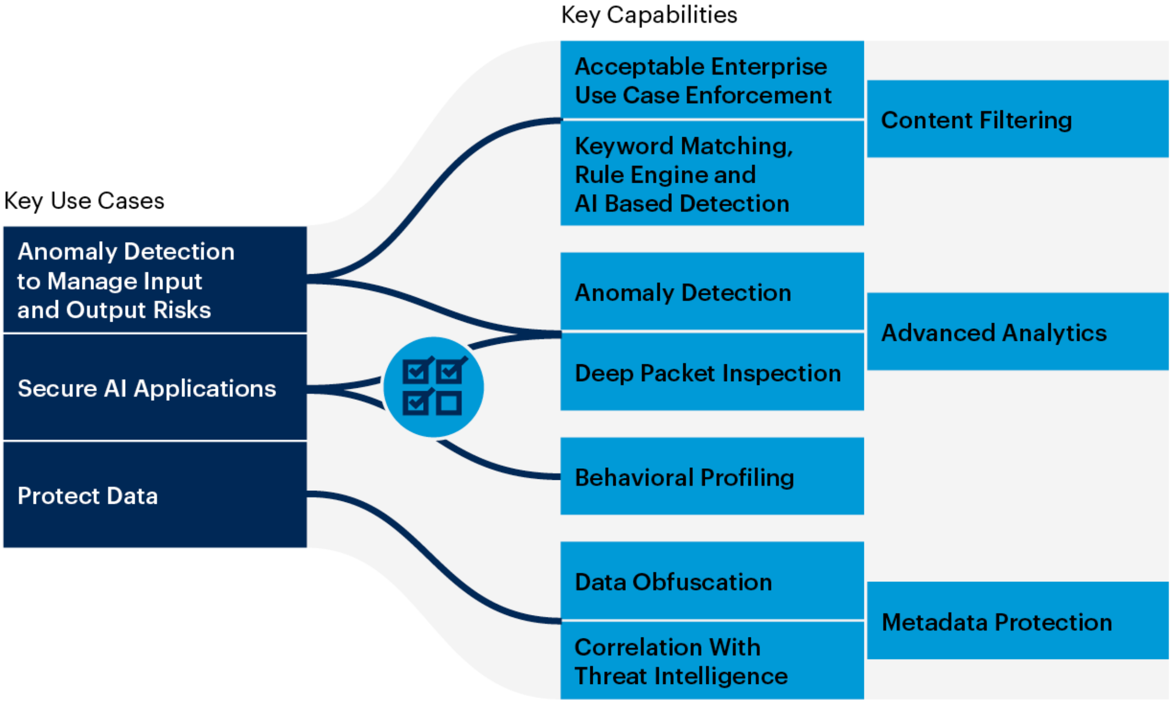
The GenAI TRiSM market is a subset of a larger AI TRiSM market that also includes multiple software segments that can only be implemented by the builders or owners of AI models, applications or agents. These builders or owners create, maintain and govern these entities. In the case of on-premises AI models, the owners of these AI models are the same entities that use or interact with those same AI models.

Market Map Visual

Back to top

Figure 1: Generative AI on Trust, Risk and Security (TRiSM) Overview

Generative AI on Trust, Risk and Security (TRiSM) Overview



Key Use Cases

- Anomaly Detection to Manage Input and Output Risks
- Secure AI Applications
- Protect Data

Key Capabilities

- Acceptable Enterprise Use Case Enforcement
- Keyword Matching, Rule Engine and AI Based Detection
- Content Filtering
- Anomaly Detection
- Deep Packet Inspection
- Advanced Analytics
- Behavioral Profiling
- Data Obfuscation
- Correlation With Threat Intelligence
- Metadata Protection

Key Trends Affecting This Market

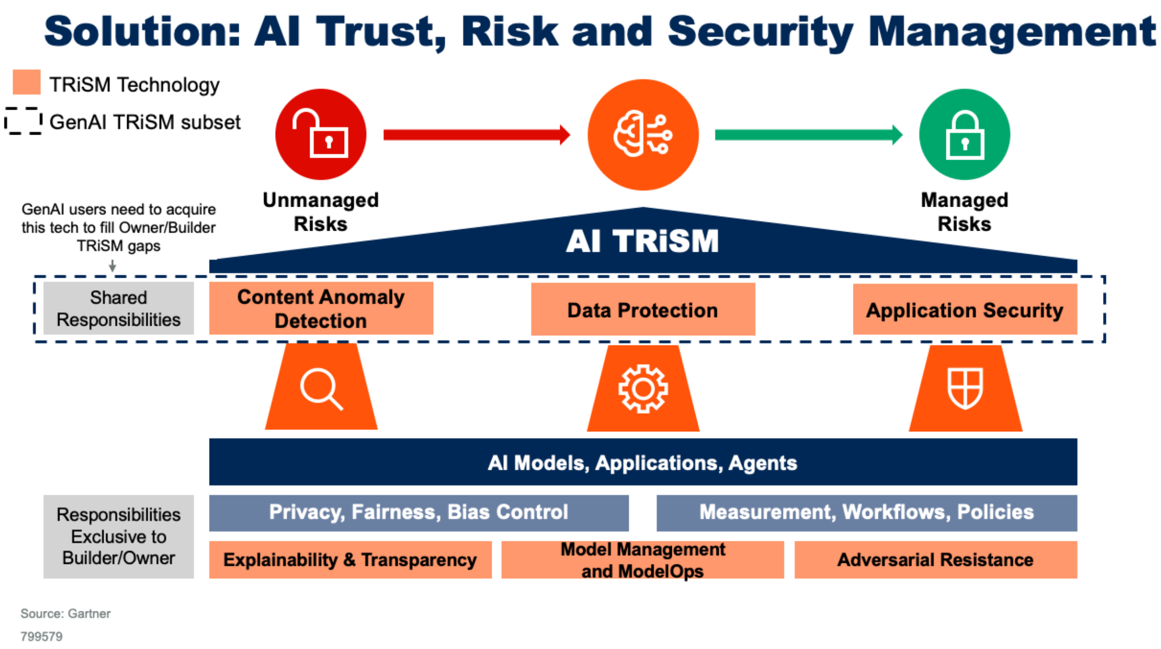
Consolidation With Legacy Security Products	Integration With Model Provider Functionality	Expansion of Functionality Into Adjacent Areas	Multi-Modal Interfaces and Capabilities
Satisfy Initial User Needs for Usage Maps	Niche Emerging Market With Low Revenues	GenAI Data Protection Standards Emerging	GenAI Bias Mitigation Tools Emerging

Source: Gartner
799579_C

Gartner

We updated our AI TRiSM architecture in 2023 to reflect the different TRiSM responsibilities of builders or owners of AI models, applications and agents and all other parties who integrate and use them. Figure 2 identifies the technology components that these two parties use to manage these aspects for AI models, applications and agents.

Figure 2: AI TRiSM Architecture


Solution: AI Trust, Risk and Security Management
TRiSM Technology
GenAI TRiSM subset
Unmanaged Risks
Managed Risks
GenAI users need to acquire this tech to fill Owner/Builder TRiSM gaps
AI TRiSM
Shared Responsibilities
Content Anomaly Detection
Data Protection
Application Security
AI Models, Applications, Agents
Responsibilities Exclusive to Builder/Owner
Privacy, Fairness, Bias Control
Measurement, Workflows, Policies
Explainability & Transparency
Model Management and ModelOps
Adversarial Resistance

Source: Gartner
799579

Gartner

This research elaborates on the three TRISM technology components in the "shared responsibilities" row that support GenAI, and lists representative vendors in each of the categories. The Market Guide for AI Trust, Risk and Security Management analyzes the functions and includes representative vendors in the three technology categories in the AI TRiSM architecture row "Responsibilities Exclusive to Builder/Owner."

Market Dynamics

Back to top

The use of hosted LLM and GenAI models unlocks many benefits, but users also must contend with new unique risks in three primary categories:

1. **Content anomaly detection**

   - Unacceptable or malicious use
   - Unmanaged enterprise content transmitted through prompts or other methods resulting in compromise of confidential data inputs
   - Hallucinations or inaccurate, illegal, copyright-infringing and otherwise unwanted outputs that compromise enterprise decision making

2. **Data protection**

    - Data leakage and confidentiality compromise in hosted vendor environment
    - Inability to govern privacy and data protection policies in externally hosted environments
    - Difficulty conducting privacy impact assessments and complying with various regional regulations, due to the black box nature of the third-party models and the mostly absent possibility to officially contract these model providers as data processors following privacy legislative requirements

3. **AI application security**

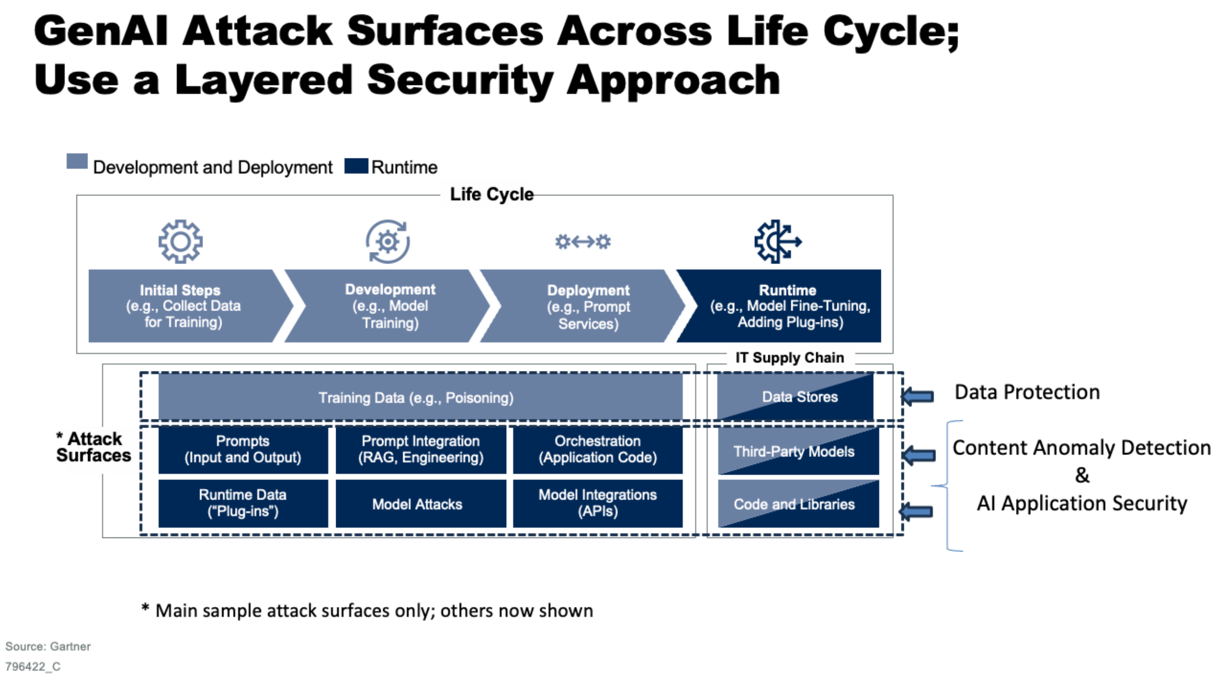
    - Adversarial prompting attacks
    - Vector database attacks
    - Hacker access to model states and parameters

Our recent survey of over 700 webinar attendees on what GenAI risks they are most concerned about validated these risk categories — and highlighted that data privacy is the No. 1 risk users are concerned about. [1]

These risks are exacerbated when using externally hosted LLM and other GenAI models, as enterprises lack capabilities to directly control their application processes and data handling and storage. But the risks still exist in on-premises models hosted and directly controlled by the enterprise, especially when security and risk controls are lacking.

These three categories of risks confront users during runtime of AI applications and models. See Figure 3 for where they manifest across the AI model life cycle, and note the associated runtime attack and compromise surfaces that impact GenAI users. These new attack surfaces are driving much of the new GenAI TRiSM market.

Figure 3: GenAI Life Cycle and Attack Surface


GenAI Attack Surfaces Across Life Cycle;
Use a Layered Security Approach
Development and Deployment
Runtime
Life Cycle
Initial Steps (e.g., Collect Data for Training)
Development (e.g., Model Training)
Deployment (e.g., Prompt Services)
Runtime (e.g., Model Fine-Tuning, Adding Plug-ins)
IT Supply Chain
* Attack Surfaces
Training Data (e.g., Poisoning)
Data Stores
Prompts (Input and Output)
Prompt Integration (RAG, Engineering)
Orchestration (Application Code)
Third-Party Models
Runtime Data ("Plug-ins")
Model Attacks
Model Integrations (APIs)
Code and Libraries
Data Protection
Content Anomaly Detection & AI Application Security
* Main sample attack surfaces only; others now shown
Source: Gartner
796422_C
Gartner

Responsibilities for implementing mitigating controls for these attack and compromise surfaces are split across two main parties, defined by their roles with regard to the AI models, applications or agents:

1. Builders or owners of an AI model, application or agent
   - These are the entities engaged in the first three components of the life cycle, shown above in light blue.
2. Users (human or machine) that integrate with the AI model, application or agent
   - These are the entities engaged in the runtime component of the life cycle, shown above in dark blue.

Oftentimes, the builders/owners and the users will be the same entity.

Market Evolution

Back to top

The market for AI TRiSM has been fragmented and has not generated the revenue vendors expected (see Market Guide for AI Trust, Risk and Security Management). In search of more revenue streams, AI TRiSM vendors have continued to develop functionalities by expanding into adjacent GenAI TRiSM categories (e.g., into AI application security or anomaly detection).

In addition, they continue to try and demonstrate the value of AI TRiSM products to data scientists and AI engineers, hoping to sell them on using these products early in AI project life cycles, rather than deploying them only when projects move into production. Their selling point has correctly been that building TRiSM into AI projects at their outset leads to better AI project performance.

Still, uptake of TRiSM products has been limited. We estimate 2022 AI TRiSM market revenues at or below $150 million a year. The rapid adoption of ChatGPT and GenAI that began in early 2023 has nonetheless accelerated enterprise demand for AI TRiSM products, as end-user organizations now more eagerly seek to implement concrete measures and controls to mitigate various risks. As such, many AI TRiSM vendors who provided anomaly detection, model monitoring and AI application security have quickly pivoted to expand their product lines to support GenAI TRiSM capabilities. These vendors are included in our vendor table below.

We believe many enterprises will initially acquire solutions that mitigate input/output risks through anomaly detection or secure AI applications to gain visibility into enterprise use of GenAI applications and models. This includes use of off-the-shelf applications such as ChatGPT or interactions through other integration points such as plug-ins, prompts or APIs. Getting their arms around enterprise interactions with GenAI is the first priority for organizations, and these products can provide a good map of those interactions. Once the map is established, core functions of mitigating risks and security threats can be gradually deployed.

We conservatively expect the GenAI TRiSM market revenue to reach $150 million by the end of 2025. We also expect this market to consolidate substantially by the end of 2026, when vendors combine input/output risk mitigation using data and content anomaly detection with AI application security. Further, we expect large incumbent security vendors, particularly in the security service edge and data loss prevention business, to acquire GenAI startups to expand protections they offer customers.

We also expect a couple of winners in the GenAI privacy and data protection category once vendors hosting large public GenAI models standardize on data protection methods (e.g., encryption) and offer these standard methods to customers demanding this added functionality.

Over time, we expect GenAI TRiSM vendors to expand their products to protect usage of multimodal models, rather than just LLMs. This will naturally occur in the next couple of years as enterprise usage of multimodal models increases.

Business Benefits (Use Cases)

Back to top

Legacy controls are not enough to mitigate risks associated with using hosted GenAI (e.g., LLM) models. Users encounter risks in the three shared responsibilities categories, and entrepreneurial third-party vendors are helping address these vulnerabilities for three distinct use cases.

Use Content Anomaly Detection to Manage Input and Output Risks

- **Input risks:** Information and data submitted to GenAI models can result in data compromise if sent to environments that are not adequately secured and protected (e.g., if they are not encrypted in transit and, if stored, at rest). Inputs to GenAI models must also be screened to ensure they meet the enterprise's acceptable use policies. Today, most of these inputs take the form of an interactive prompt, but security leaders must address input and output risks for applications leveraging more automated forms of inputs, such as API calls from other applications, and outputs directly transmitted to software agents.

- **Output risks:** Outputs from GenAI models are unreliable, given an unpredictable rate of factual errors and hallucinations. Outputs can also be biased and potentially include copyright material or other unwanted, malicious, illegitimate or illegal information. This puts the enterprise at risk of being sued for illegally using proprietary materials output from the models or making misinformed decisions.

  - When hosting models on-premises, enterprises mitigate data compromise risks that come from submitting data to third-party environments. However, enterprises still must protect the data they host.

Vendors in this category typically aim to mitigate LLM input and output risks. (Other types of GenAI models are not in their scope at this time). They do so by providing anomaly detection and content filtering that screens inputs and outputs against preset enterprise policies embodied in rule-based systems or small AI models that sit between the enterprise and the hosted LLM. These anomaly detectors and content filters mediate and validate the information flows against them. We have not identified a vendor in this category that can screen outputs against specific enterprise-defined bias and fairness policies.

Manage Data Protection Risks, Especially Inherent to External Environments

- The use of private, proprietary, sensitive or confidential information as inputs into hosted GenAI models entails risks of data leakage and potential violations of existing regulations. Organizations must monitor and enforce privacy, data confidentiality and governance in the environments where these vendors' models are hosted.
    - When hosting models on-premises, enterprises automatically mitigate data compromise risks in third-party environments since they are not used.
        - However, they still must account for the legality of (personal) data in use, and secure and protect all data processed and hosted — notably, when the application will be available to external parties (customers, partners).

Emerging vendors targeting this category are still sparse; in fact, we only identified two of them. The reason why more have not emerged is because, generally speaking, hosted LLM models only accept queries in clear text unless they are equipped with specialized data protection software that "unscrambles" or interprets the protected, scrambled or encrypted submission from the user. We recommend leaders monitor the efforts of data protection vendors offering privacy-enhancing technologies and those listed in the privacy category of the Market Guide for AI Trust, Risk and Security Management.

External-hosting LLM vendors cannot practically support all the various potential data protection methods at this stage of the market. Yet it is possible that in the future, some standard methods will be endorsed by vendors, assuming users demand them and that they do not compromise performance. Further, most enterprises will not want to send synthetic data to LLMs or other GenAI models when querying them, as synthetic data could render responses much less useful. (Sending synthetic data to a hosted LLM would be a good way to protect sensitive real data that would then stay in the enterprise, but again, it is impractical for most use cases).

Vendors in the data protection domain provide tools to encrypt and/or transform data sent to, for example, hosted vector databases or LLM services.

Manage AI Application Security Risks by Safeguarding New GenAI Attack Vectors

The extent of these new attack surfaces for the application component surrounding the LLM models remains undetermined, as research work and new proofs of concept continue to be released. AI applications include new components to orchestrate the use of the models. This introduces security threats that conventional application security controls do not yet address.

- Adversarial prompting includes prompt guardrail “hijacking” and prompt injection attacks. These attacks require embracing a “red teaming” approach to GenAI application security by testing its resistance to these adversarial prompts. Tools are emerging to support better automation of testing these prompts.
- Vector database breaches.
- Malicious hacker access to model states and parameters.
- Unmanaged and unmonitored integration with third-party models offered “as a service” through API calls and other IT supply chain risks.

In addition, organizations can tune their use of existing infrastructure and application security controls. Many organizations are leveraging their existing security service edge providers to filter access to GenAI web applications. They also use their SIEMs to create “generative AI dashboards” to discover new applications being used and get a better sense of which teams use them.

Further, attackers themselves can use GenAI for hacking techniques that enterprises are not prepared for. These risks also exist where enterprises have more visibility and control.

Vendors in this category use a range of techniques to scan and prevent cybersecurity risks that are unique to applications embedding and interacting with GenAI and hosted LLMs.

Piloting and Evaluating Vendors

Back to top

Before selecting a GenAI TRiSM vendor, you must first decide on your deployment approach (see How to Choose an Approach for Deploying Generative AI) and the TRiSM functionality you require.

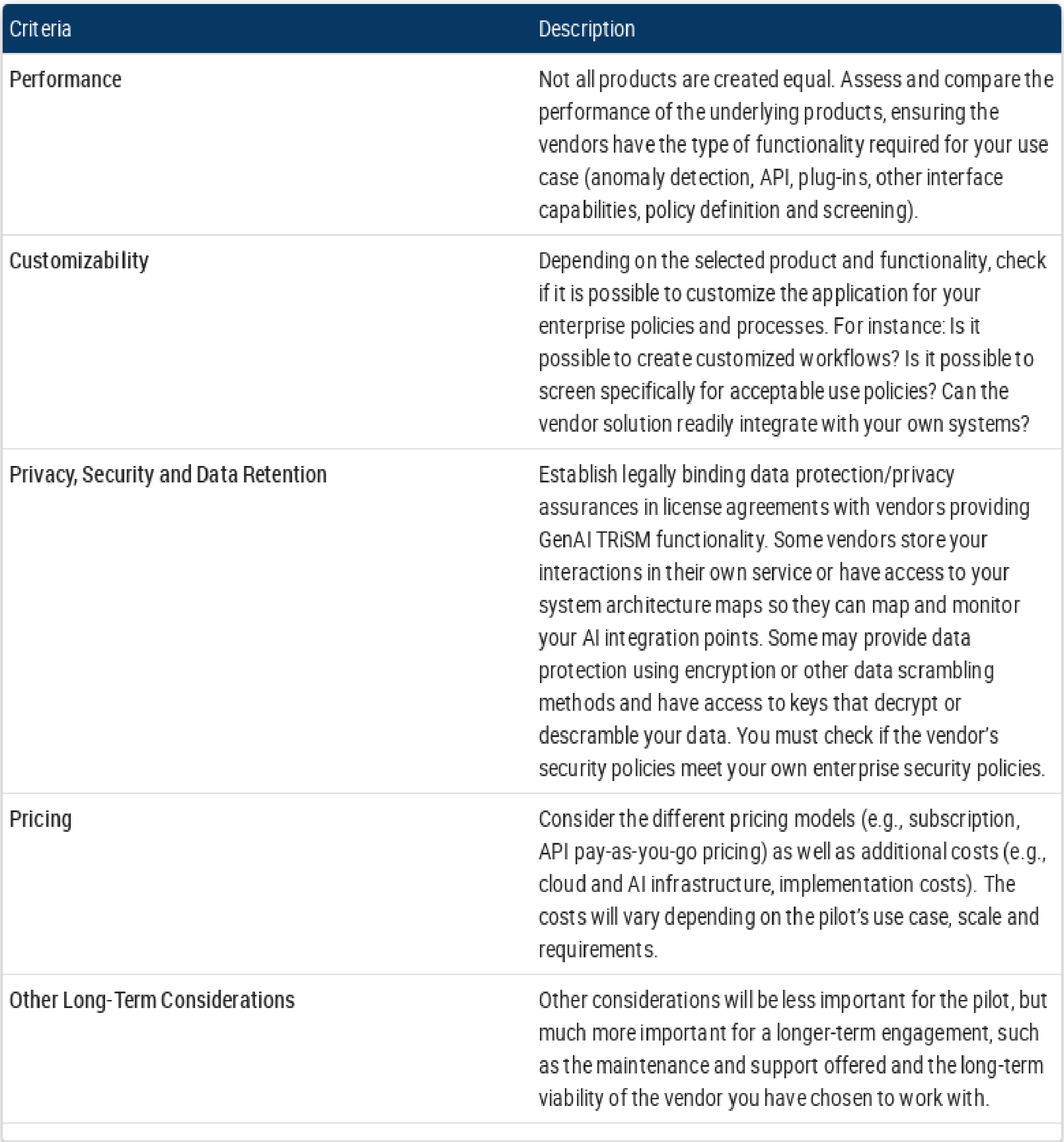
Once you are clear on the GenAI use cases and GenAI TRiSM business benefits you require for a successful implementation, consider a third-party vendor applicable to your deployment approach. For example, if you are using prompt engineering to restrict model completions to only use your private validated data, you will likely lower your need for a vendor to mitigate output risk.

Further, your organization is likely to use multiple GenAI models and applications, and a third-party vendor should readily support such usage and eliminate friction, if you swap out various AI models and applications for others that better suit your needs.

Once you settle on the need for a third-party GenAI TRiSM vendor, consider the five criteria in Table 1 when selecting a vendor for your pilot and use case.

Table 1: Vendor Criteria for Generative AI Pilot

(Enlarged table in Appendix)



Criteria	Description
Performance	Not all products are created equal. Assess and compare the performance of the underlying products, ensuring the vendors have the type of functionality required for your use case (anomaly detection, API, plug-ins, other interface capabilities, policy definition and screening).
Customizability	Depending on the selected product and functionality, check if it is possible to customize the application for your enterprise policies and processes. For instance: Is it possible to create customized workflows? Is it possible to screen specifically for acceptable use policies? Can the vendor solution readily integrate with your own systems?
Privacy, Security and Data Retention	Establish legally binding data protection/privacy assurances in license agreements with vendors providing GenAI TRiSM functionality. Some vendors store your interactions in their own service or have access to your system architecture maps so they can map and monitor your AI integration points. Some may provide data protection using encryption or other data scrambling methods and have access to keys that decrypt or descramble your data. You must check if the vendor's security policies meet your own enterprise security policies.
Pricing	Consider the different pricing models (e.g., subscription, API pay-as-you-go pricing) as well as additional costs (e.g., cloud and AI infrastructure, implementation costs). The costs will vary depending on the pilot's use case, scale and requirements.
Other Long-Term Considerations	Other considerations will be less important for the pilot, but much more important for a longer-term engagement, such as the maintenance and support offered and the long-term viability of the vendor you have chosen to work with.

Source: Gartner (September 2023)

The pilot is *not* the moment to make long-term vendor decisions; this is a much longer process. It is simply an opportunity to run a small proof of concept using external capabilities.

Managing Risks

Back to top

Vendor viability risk is inherent when working with emerging startups or small vendors in emerging markets that may not be able to sustain themselves financially in the long run. To manage this risk, ensure you own your own data and can readily extract it from the vendor's platform if the vendor's situation changes. Also make sure you thoroughly understand the functionality the vendor is providing so you can more easily transition to a comparable vendor, while minimizing business disruption.

Continually monitor your vendor's position in the market along with its growth and profitability, and prepare a "Plan B" once red flags appear.

Be prepared for market consolidation as described in the "Market Evolution" section, and use larger incumbent security vendors who extend their portfolio to manage GenAI TRiSM if your enterprise prefers to use fewer, more mature and financially stable vendors.

Finally, manage the risk of vendors unintentionally (or intentionally) leaking or compromising your data by favoring product implementation on-premises where you can directly manage their security posture. Scrutinize on-premises products for security vulnerabilities, and contractually obligate the vendors to only access your data with your explicit approval. Ensure technical controls are in place to disable vendor access to your data as a default configuration, whether the product is run on-premises or in a cloud.

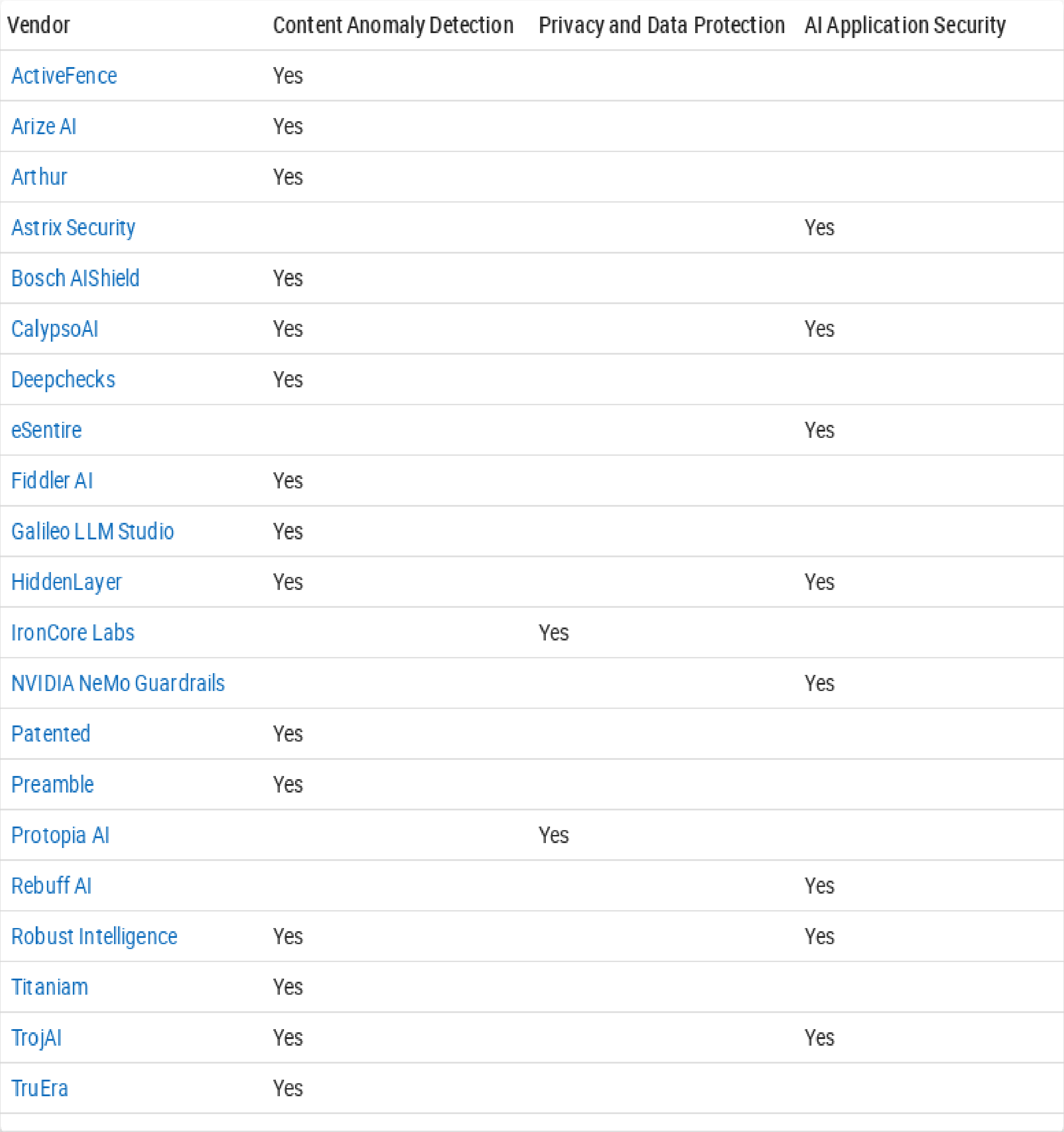
Representative Vendors

Back to top

The vendors listed in this Innovation Guide do not imply an exhaustive list. This section is intended to provide more understanding of the market and its offerings.

Table 2: Representative GenAI TRiSM Vendors

(Enlarged table in Appendix)



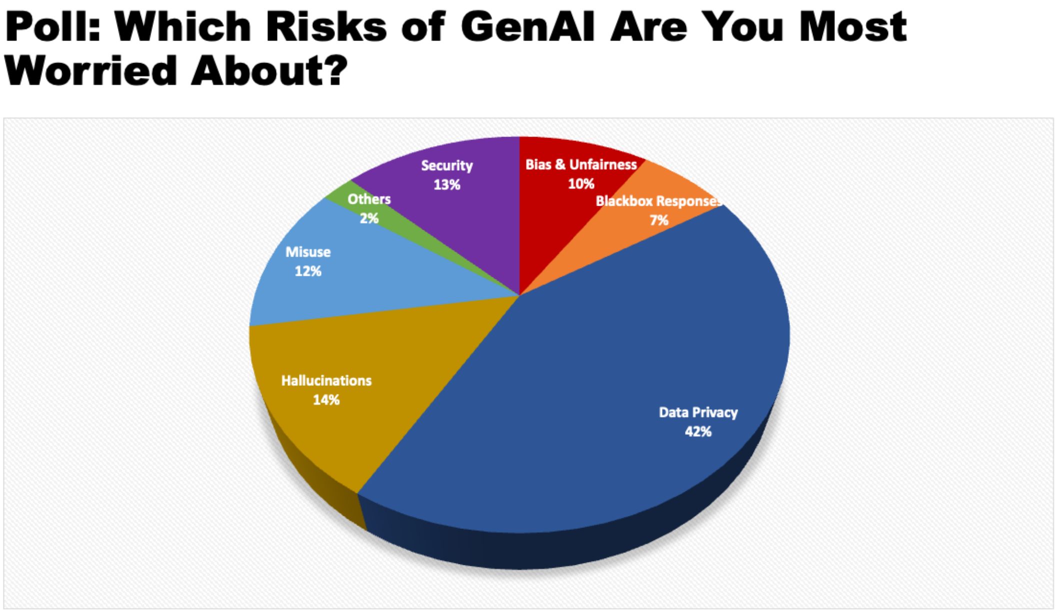
Vendor	Content Anomaly Detection	Privacy and Data Protection	AI Application Security
ActiveFence	Yes		
Arize AI	Yes		
Arthur	Yes		
Astrix Security			Yes
Bosch AIShield	Yes		
CalypsoAI	Yes		Yes
Deepchecks	Yes		
eSentire			Yes
Fiddler AI	Yes		
Galileo LLM Studio	Yes		
HiddenLayer	Yes		Yes
IronCore Labs		Yes	
NVIDIA NeMo Guardrails			Yes
Patented	Yes		
Preamble	Yes		
Protopia AI		Yes	
Rebuff AI			Yes
Robust Intelligence	Yes		Yes
Titaniam	Yes		
TrojAI	Yes		Yes
TruEra	Yes		

Source: Gartner (September 2023)

Evidence

[1] Gartner IT Executives Webinar Poll (August 2023), n = 713.

Figure 4: IT Executive Poll on GenAI Concerns



Poll: Which Risks of GenAI Are You Most Worried About?

Source: Gartner IT Executives Webinar Poll (August 2023), n = 713

799579

Table 1: Vendor Criteria for Generative AI Pilot

Criteria	Description
Performance	Not all products are created equal. Assess and compare the performance of the underlying products, ensuring the vendors have the type of functionality required for your use case (anomaly detection, API, plug-ins, other interface capabilities, policy definition and screening).
Customizability	Depending on the selected product and functionality, check if it is possible to customize the application for your enterprise policies and processes. For instance: Is it possible to create customized workflows? Is it possible to screen specifically for acceptable use policies? Can the vendor solution readily integrate with your own systems?
Privacy, Security and Data Retention	Establish legally binding data protection/privacy assurances in license agreements with vendors providing GenAI TRiSM functionality. Some vendors store your interactions in their own service or have access to your system architecture maps so they can map and monitor your AI integration points. Some may provide data protection using encryption or other data scrambling methods and have access to keys that decrypt or descramble your data. You must check if the vendor's security policies meet your own enterprise security policies.
Pricing	Consider the different pricing models (e.g., subscription, API pay-as-you-go pricing) as well as additional costs (e.g., cloud and AI infrastructure, implementation costs). The costs will vary depending on the pilot's use case, scale and requirements.

Other Long-Term Considerations	Other considerations will be less important for the pilot, but much more important for a longer-term engagement, such as the maintenance and support offered and the long-term viability of the vendor you have chosen to work with.

Source: Gartner (September 2023)

Table 2: Representative GenAI TRiSM Vendors

Vendor	Content Anomaly Detection	Privacy and Data Protection	AI Application Security
ActiveFence	Yes		
Arize AI	Yes		
Arthur	Yes		
Astrix Security			Yes
Bosch AIShield	Yes		
CalypsoAI	Yes		Yes
Deepchecks	Yes		
eSentire			Yes
Fiddler AI	Yes		
Galileo LLM Studio	Yes		
HiddenLayer	Yes		Yes
IronCore Labs		Yes	
NVIDIA NeMo Guardrails			Yes
Patented	Yes		
Preamble	Yes		

Protopia AI		Yes	
Rebuff AI			Yes
Robust Intelligence	Yes		Yes
Titaniam	Yes		
TrojAI	Yes		Yes
TruEra	Yes		

Source: Gartner (September 2023)