Deloitte.

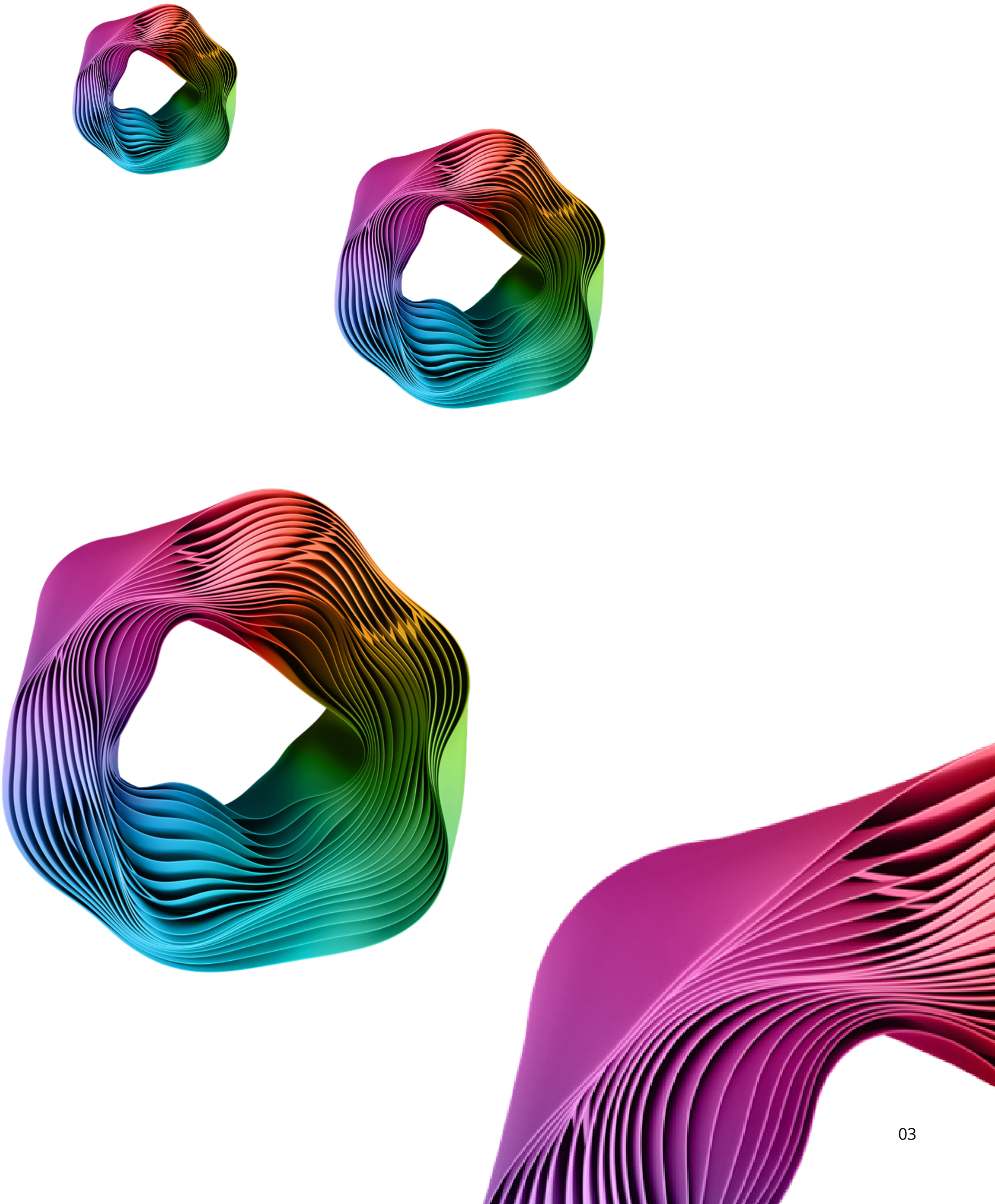
Secure Software Development Lifecycle™ (SSDL) for Precision AI™

November 2024

MAKING AN IMPACT THAT MATTERS since 1845

Contents

This white paper aims to provide a broad understanding of the potential benefits and practical applications of Generative AI (GenAI) and Artificial Intelligence/Machine Learning (AI/ML) technologies, with a specific focus on Palo Alto Networks' Precision AI. It explores how these technologies can transform the security of the Software Development Lifecycle (SDLC) for clients.



Executive summary

In the rapidly evolving landscape of digital transformation and technological disruption, organizations face increasing pressure to adapt to changing business demands. This drives the rapid development of platforms and applications, necessitating broad collaboration across software development, cybersecurity, and Information technology (IT) operations. The introduction of the SSDL framework marks a significant shift, transforming traditional development processes to enhance the secure development and deployment of multi-cloud ecosystems.

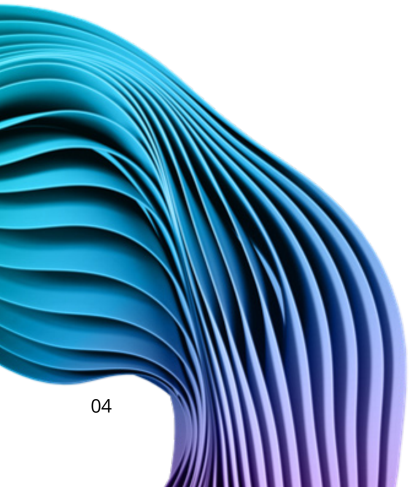
Integrating GenAI technologies within the SSDL framework introduces new dimensions of complexity and opportunity. These technologies, including AI Large Language Models (LLMs) and open-source resources, help drive innovation and efficiency while creating potential new attack surfaces. These vulnerabilities can jeopardize highly valued intellectual property by exposing them to adversaries. Therefore, incorporating these new technologies within the SSDL framework requires measures to secure an organization's critical assets.

GenAI and other AI/ML technologies offer diverse software compositions that utilize evolving open-source code to meet increasing demands. Implementing a SSDL for GenAI can automate and require many facets of the software development and security processes, potentially leading to more efficient and error-free outcomes. By leveraging GenAI for code analysis, vulnerability detection, prioritization, and automated testing tasks, organizations can reduce reliance on manual controls, which often slow down the development process and can lead to costly defects being detected at later stages.

This document serves as a guide for those considering the adoption of such technologies to enhance their cybersecurity posture, operational efficiency, and compliance adherence. It is intended for Chief Information Security Officers (CISOs), Chief Information Officers (CIOs), Chief Technology Officers (CTOs), Chief Product Officers (CPOs), and other decision-makers and leaders responsible for securing their GenAI and LLM application development environments.

This white paper will emphasize Deloitte's SSDL framework, augmented by DevSecOps practices, which provides an approach to securing the development and deployment of GenAI and LLM applications. By integrating GenAI technologies and other AI/ML capabilities, such as Palo Alto Networks' Precision AI, within a broad SSDL framework, organizations can harness the power of AI while mitigating security risks.

This structured approach checks that each phase of the development lifecycle addresses potential threats, which can help safeguard both innovation and security. Consequently, organizations can leverage the transformative potential of LLM applications securely and responsibly, providing that they do not compromise security and privacy.



1.0 GenAI vs. LLMs overview

Before diving deep into the SSDL with Precision AI, let's understand the difference between GenAI and LLMs.

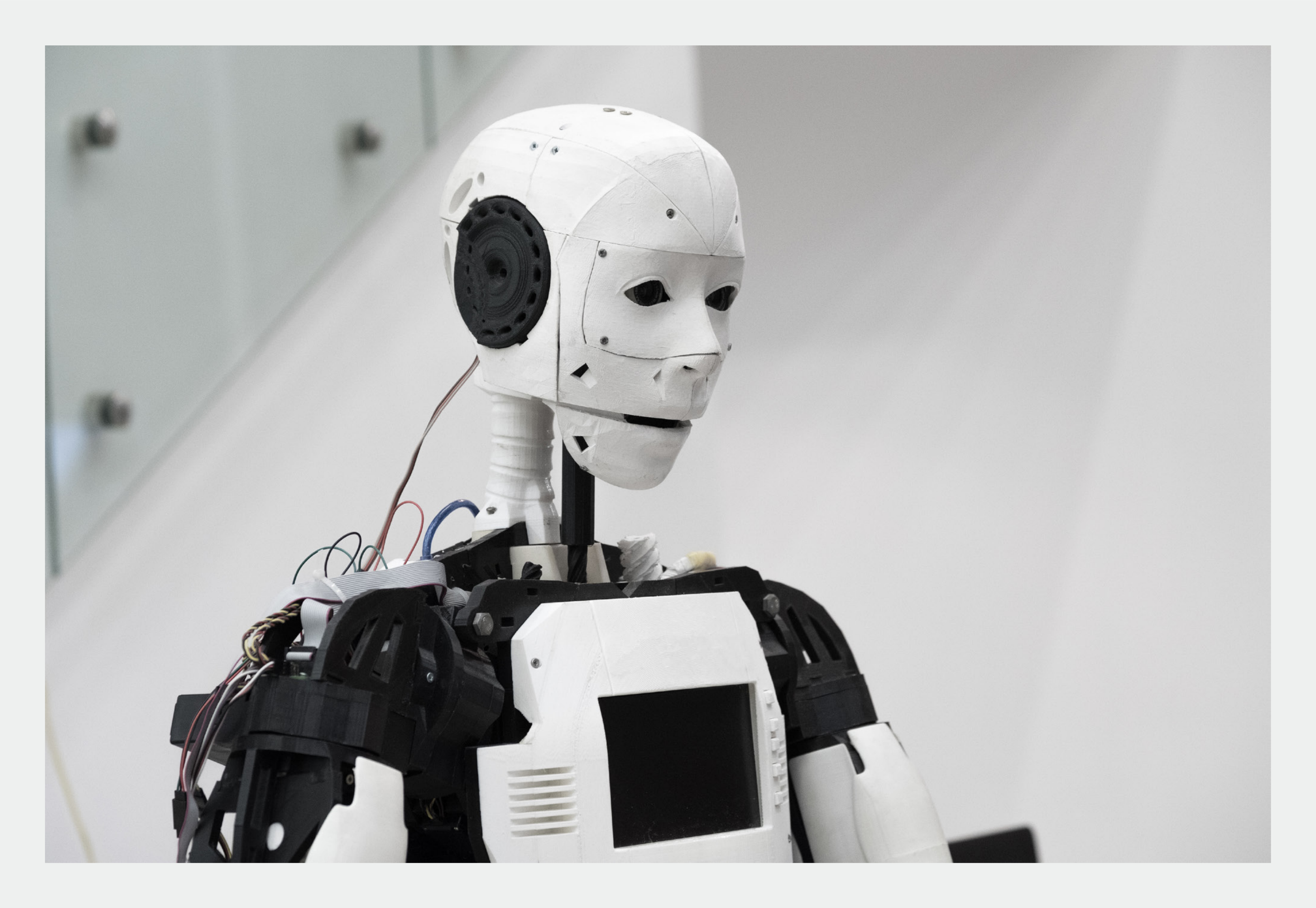


Figure 1: Example of GenAI creation

1.1 GenAI

GenAI refers to a class of AI that can create new content, such as text, images, music, and more, based on the data it has been trained on. GenAI encompasses a wide range of models and applications, including[1]:

- **Generative Adversarial Networks (GANs):**
  These generate realistic images and videos. For example, GANs can create photo-realistic images of people who don't exist.
- **Variational Auto Encoders (VAEs):**
  These generate data with specific distributions, such as creating variations of images from a dataset of facial photos.
- **Transformer models,** including Generative Pre-trained Transformers (GPT), have transformed the field of natural language processing (NLP) by enabling models to understand and generate human-like text more effectively than previous architectures. For instance, existing AI generative models can write essays and poems or generate programming code based on prompts.

Figure 2 – How LLMs relate to GenAI

1.2 LLMs

While all LLMs are a part of GenAI, not all GenAI models are LLMs. GenAI is an umbrella term covering a wide range of generative models, whereas LLMs are specifically designed for tasks involving human language. Both play important roles in advancing AI technologies, with LLMs being particularly significant for text-related applications.[2]

These models are designed to understand, generate, and manipulate human language. Examples include:

- **GPT:** GPT models can generate coherent and contextually relevant text, such as completing sentences, writing articles, or engaging in conversations.[3]
- **BERT (Bidirectional Encoder Representations from Transformers):** BERT is developed to understand the context of words in search queries. BERT helps improve search engine results by understanding the intent behind user queries.[4]
- **LLaMA** (Large Language Model Meta AI) is a model designed for language understanding and generation, making it suitable for applications such as chatbots and language translation services.[5]

> Additionally, consider a smart assistant that can understand your questions and provide relevant answers. These assistants use LLMs like BERT or GPT to comprehend and generate text.[6]

2.0 The specific differences between GenAI and LLMs

The specific differences between GenAI and LLMs are scope, applications, and techniques[7]:

2.1 Scope

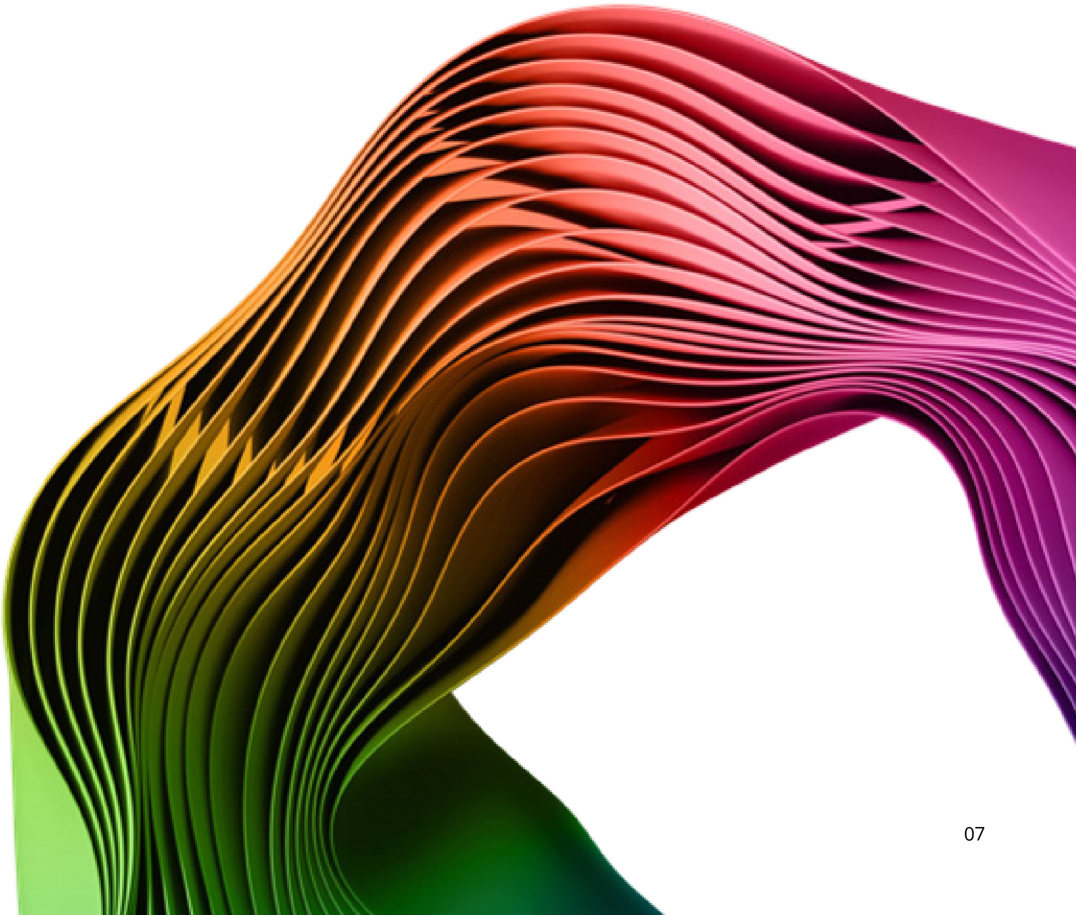
- **GenAI:** A broad category that includes AI models capable of generating content across various domains (e.g., text, images, audio).
- **LLMs:** Specifically focused on text generation and understanding within the domain of NLP.

2.2 Applications

- **GenAI:** Includes diverse applications such as image generation (GANs), music composition, and synthetic data creation.
- **LLMs:** Primarily used for text-based applications like chatbots, translation, summarization, and content creation.

2.3 Techniques

- **GenAI:** Leverages various techniques like GANs, VAEs, and transformers.
- **LLMs:** Mainly rely on transformer architectures for tasks involving text.



3.0 GenAI and LLMs application development trend and specific benefits

The current trend strongly focuses on GenAI and LLMs application development due to their immediate applicability and transformative potential in many business processes. Many organizations are developing GenAI and LLM applications due to their powerful capabilities in language-related tasks.[8]

The development of GenAI and LLM applications holds promise across multiple domains, driving innovation, efficiency, and improved user experiences. As these models evolve, their potential applications and benefits will likely expand further, making them an integral part of modern technology solutions.

Here are some of the potential benefits[8]:

Figure 3 – Benefits of LLMs

3.1 Enhanced natural language understanding and generation

LLMs, BERT, and their successors, can understand and generate highly accurate human-like text. This ability allows for more sophisticated and natural interactions between humans and machines, enhancing user experiences in customer service, virtual assistants, and other applications.

3.2 Automation of content creation

LLMs can automate the creation of various types of content, including articles, reports, emails, and social media posts. This can save businesses time and resources while maintaining high-quality communication and marketing standards.[9]

3.3 Improved customer service and support

Chatbots and virtual assistants powered by LLMs can provide careful and context-aware responses to customer inquiries. This can lead to improved customer satisfaction and help reduce the human support staff's workload by efficiently handling routine queries.

3.4 Advanced data analysis and insights

LLMs can process and analyze vast amounts of textual data to extract meaningful insights. This capability is valuable in market research, healthcare, and finance, where understanding trends and patterns in large datasets can drive better decision-making.[10]

3.5 Multilingual capabilities

LLMs are trained on diverse datasets that include multiple languages, enabling them to provide translation services and multilingual support. This can help businesses expand their reach globally and cater to a diverse customer base.

3.6 Enhanced productivity tools

Integration of LLMs in productivity tools, such as word processors and spreadsheets, can assist users with tasks like drafting emails, summarizing documents, and generating reports. This enhances productivity and allows users to focus on higher-level tasks.[11]

3.7 Personalized user experiences (UX)

LLMs can create personalized content and recommendations based on user preferences and behavior. This personalization can improve user engagement and satisfaction across various platforms, from e-commerce to entertainment services.[12]

3.8 Educational applications

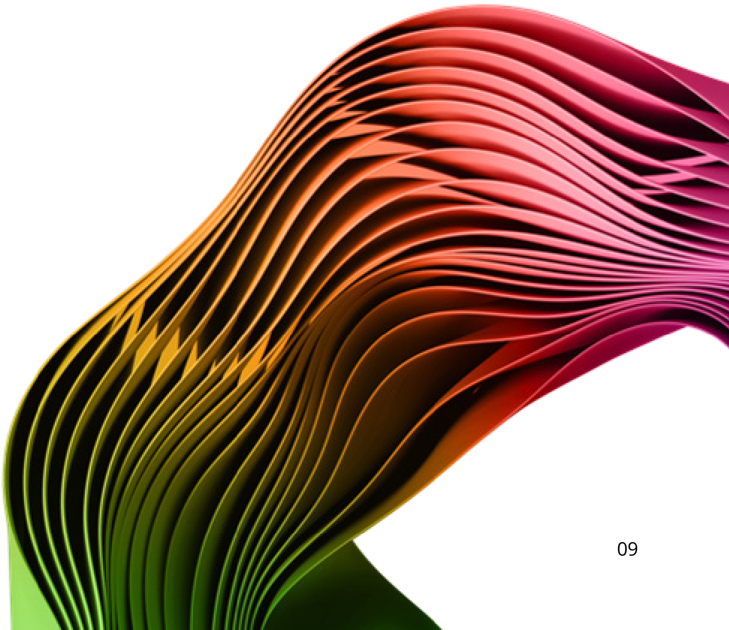
LLMs can support educational applications by providing explanations, tutoring, and personalized learning experiences. They can help students understand complex subjects by generating easy-to-understand content and answering questions interactively.[13]

3.9 Code generation and software development

LLMs can assist in writing and debugging code, making software development more efficient. This capability can help developers overcome coding challenges and speed up development.[14]

3.10 Research and innovation

LLMs can aid researchers by generating hypotheses, summarizing academic papers, and identifying relevant literature. This can accelerate research and innovation across various scientific and academic fields.[15]



4.0 Safety of GenAI and LLMs

The safety of GenAI and LLMs is contingent upon how they are used within an organization. These technologies introduce new risks while amplifying existing ones. Trust in AI will likely be a result of broad AI governance and effective risk mitigation strategies.

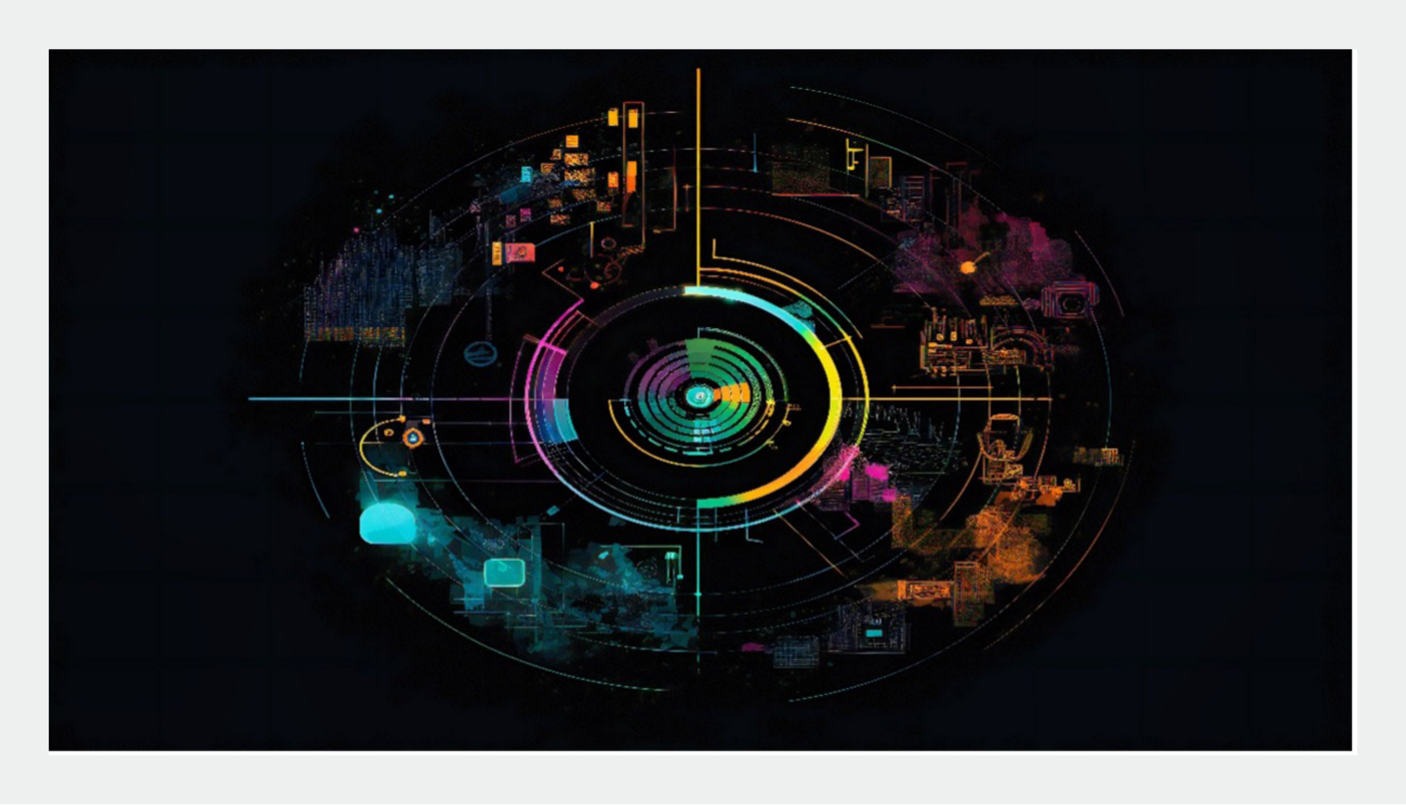


Figure 4 – GenAI and LLMs

- **Security risks:** Security risks associated with AI can lead to financial losses, intellectual property theft, and reputational damage due to unauthorized data exposure, security breaches, or business disruptions.
- **Trust and safety risks:** AI has the potential to generate highly convincing phishing emails, deep fakes, misinformation, or harmful content, which can erode user trust, invite regulatory actions, and damage an organization's brand integrity.
- **Regulatory risks:** Organizations leveraging AI may need to comply with new regulatory requirements as the regulatory landscape evolves.
- **Privacy risks:** Violations of consumer privacy, such as issues with transparency and data use, can lead to legal repercussions, loss of consumer trust, operational disruptions, and significant financial costs.
- **Reputational risks:** Hallucinations from GenAI or misuse by malicious actors can spread misinformation, adversely influencing public opinion, trust, and security.
- **Data risks:** Risks arising from the use of biased or poor-quality data for training models can result in biased decision-making, legal repercussions, and operational inefficiencies.

5.0 LLM applications benefits vs. security threats

The development of LLM applications offers numerous promising benefits but introduces various security threats that should be addressed. A critical challenge is that the control and data planes cannot be strictly isolated, complicating management. Additionally, LLMs are inherently nondeterministic, meaning they can yield different outcomes for the same input due to their reliance on semantic rather than keyword search. This method prioritizes terms based on context, affecting the consistency and reliability of results and leading to issues like hallucinations, where the model generates inaccurate or nonsensical outputs due to training data flaws.[16]

LLMs can increase an organization's attack surface by introducing different and familiar challenges, such as a software bill of materials (SBoM), supply chain vulnerabilities, data loss protection (DLP), and authorized access issues.[16] Adversaries can exploit LLMs to enhance traditional attack methods, create sophisticated malware, develop tailored phishing schemes, and generate convincing deep fakes, facilitating more efficient and effective attacks.

The failure to utilize LLMs also poses risks, including competitive disadvantages, negative market perception, inability to scale personalized communications, innovation stagnation, operational inefficiencies, increased human error, and poor resource allocation. Understanding and integrating these challenges with business strategies can help organizations weigh the pros and cons of using LLMs, safeguarding that these technologies accelerate rather than hinder business objectives.

Addressing these risks requires security measures, including input comparison, required data governance policies, continuous monitoring, and human oversight. By implementing these safeguards, organizations can harness the power of LLMs while mitigating potential security risks.

Here are some potential security threats[17]:

5.1 Natural language understanding and generation

- **Adversarial attacks:** Malicious users can manipulate inputs to cause LLMs to generate harmful or misleading outputs.
- **Data leakage:** Sensitive information could be inadvertently generated or inferred from model responses.

5.1 Automation of content creation

- **Misinformation and fake news:** Automated content generation can be misused to create and spread false information.
- **Plagiarism and copyright infringement:** Generated content may unintentionally replicate copyrighted material, leading to legal issues.

5.3 Customer service and support

- **Social engineering:** Attackers can exploit LLM-powered chatbots to phish for sensitive information from users.
- **Impersonation:** LLMs can create realistic but fraudulent interactions, deceiving customers.

5.4 Data analysis and insights

- **Privacy violations:** Analyzing large datasets can lead to the exposure of private or sensitive information.
- **Data poisoning:** Malicious actors can inject data to corrupt the insights generated by LLMs.

5.5 Multilingual capabilities

- Cross-language exploits: Malicious inputs in one language can cause harmful outputs in another, exploiting differences in language handling.
- Inconsistent translations: Misinterpretations can lead to incorrect or harmful translations, affecting communication accuracy.

5.6 Productivity tools

- **Data leakage:** Sensitive information entered into productivity tools can be inadvertently shared or exposed.
- **Dependence on AI:** Over-reliance on AI-generated content without verification can lead to errors and misinformation.

5.7 Personalized user experience (UX)

- **Privacy concerns:** Personalized recommendations require extensive data collection, posing risks to user privacy.
- **Behavioral manipulation:** Algorithms might manipulate user behavior by recommending certain content over others.

5.8 Educational applications

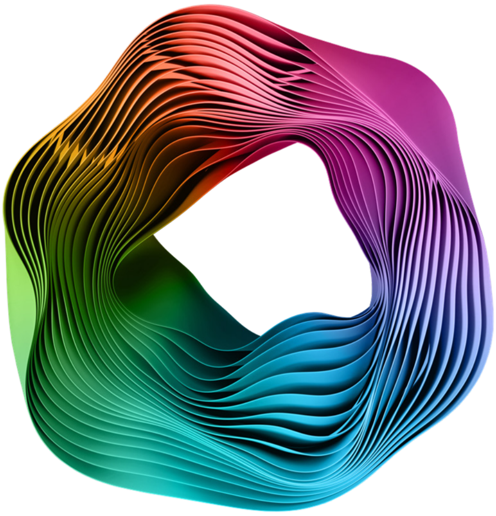
- **Bias in educational content:** LLMs can inadvertently reinforce biases present in the training data, affecting the quality of education.
- **Over-reliance on AI:** Students may become overly dependent on AI for learning, reducing critical thinking skills.

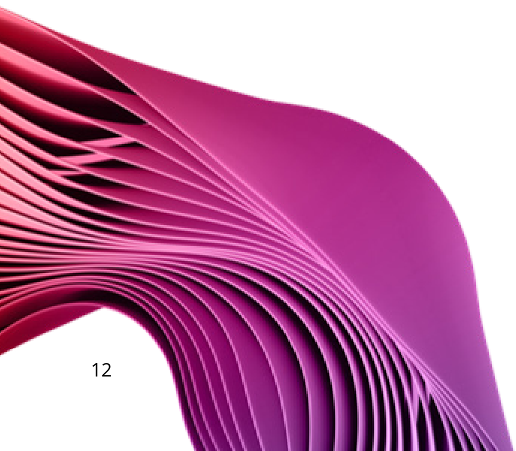
5.9 Code generation and software development

- **Vulnerable code:** AI-generated code may contain security vulnerabilities that could be exploited by attackers.
- **Intellectual property theft:** Generated code might inadvertently include proprietary or sensitive code snippets.

5.10 Research and innovation

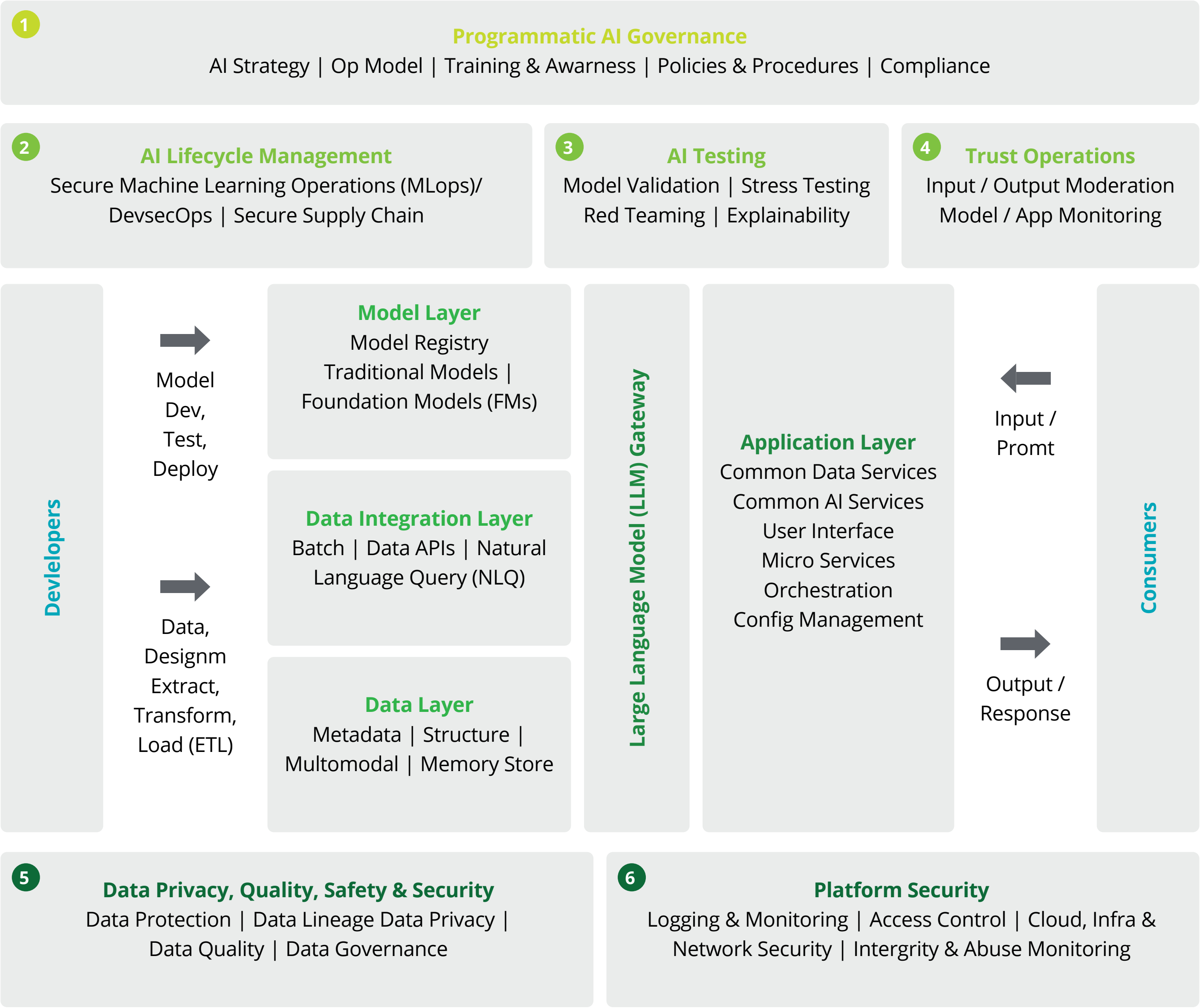
- **Data security:** Sensitive research data can be exposed through interactions with LLMs.
- **The integrity of research:** AI-generated research outputs need broad comparison to prevent disseminating incorrect information.





6.0 Building a trusted AI architecture for platform integration

AI platforms require a reliable and trusted AI architecture. As organizations establish AI and GenAI platforms, along with experimentation sandboxes, it is important to integrate both existing technologies and new point solutions into the development and deployment environments effectively.



1 **Programmatic AI Governance**
AI Strategy | Op Model | Training & Awarness | Policies & Procedures | Compliance

2 **AI Lifecycle Management**
Secure Machine Learning Operations (MLops)/ DevsecOps | Secure Supply Chain

3 **AI Testing**
Model Validation | Stress Testing Red Teaming | Explainability

4 **Trust Operations**
Input / Output Moderation Model / App Monitoring

Developers

Model Dev, Test, Deploy

Data, Designm Extract, Transform, Load (ETL)

Model Layer
Model Registry Traditional Models | Foundation Models (FMs)

Data Integration Layer
Batch | Data APIs | Natural Language Query (NLQ)

Data Layer
Metadata | Structure | Multomodal | Memory Store

Large Language Model (LLM) Gateway

Application Layer
Common Data Services
Common AI Services
User Interface
Micro Services
Orchestration
Config Management

Input / Promt

Output / Response

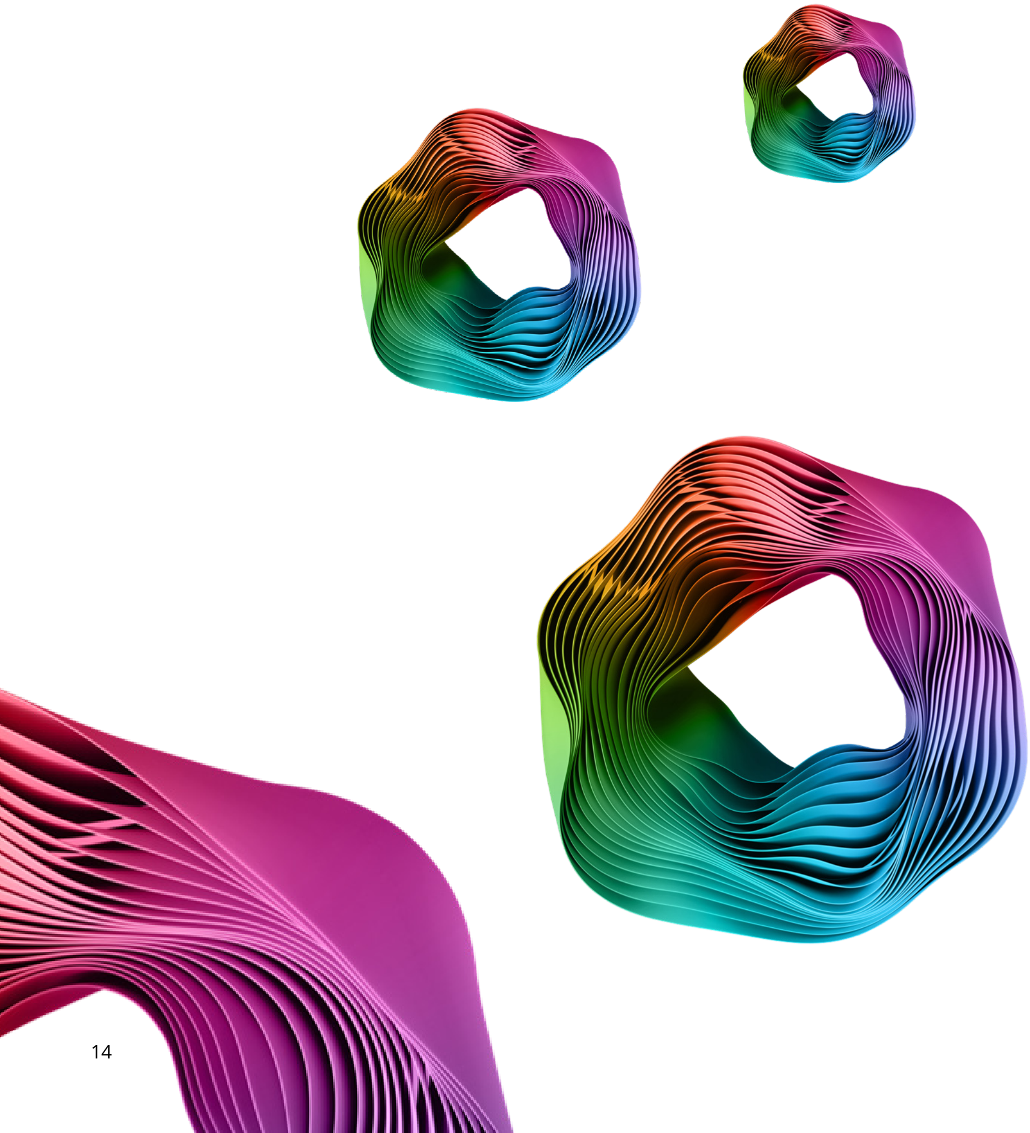
Consumers

5 **Data Privacy, Quality, Safety & Security**
Data Protection | Data Lineage Data Privacy | Data Quality | Data Governance

6 **Platform Security**
Logging & Monitoring | Access Control | Cloud, Infra & Network Security | Intergrity & Abuse Monitoring

Figure 5: Trusted AI Architecture

- **Programmatic AI governance:** Enterprise-wide, systematic approach to overseeing the development, deployment, and monitoring of AI within an organization. Leverages formalized frameworks, policies, and procedures to align AI use with organizational goals, accountability model, ethical standards, and regulatory requirements.
- **AI lifecycle management:** Tools and processes involved in operational risk mitigation across the AI lifecycle, standardizing business and data science practices through design, development, data handling, testing, deployment, monitoring, and feedback incorporation.
- **AI testing:** Evaluating and comparing AI models for accuracy, reliability, performance, and fairness. Testing mechanisms include model comparison, stress-testing, AI red-teaming, and documentation for explainability.
- **Trust operations:** Configuration and deployment of guardrail solutions and technology for continuous monitoring of AI interactions, model performance, and adversarial activity
- **Data quality, safety, and security:** Protocols to prepare careful, unbiased/balanced, and appropriate data used in development/training and testing for data minimization, protection and performance.
- **Platform security:** Securing the underlying infrastructure and software platform that hosts AI systems through integration into enterprise network and infra solutions and security operations centers.



7.0 SSDL for LLM applications development

The SSDL is vital for organizations developing their GenAI platforms by leveraging various LLMs. Utilizing developers, the SSDL framework determines a secure development process from inception through deployment and maintenance. This approach is important for protecting the integrity and security of GenAI platforms, whether development occurs in the cloud or on-premise.

Securing AI throughout the SSDL™ with Precision AI™

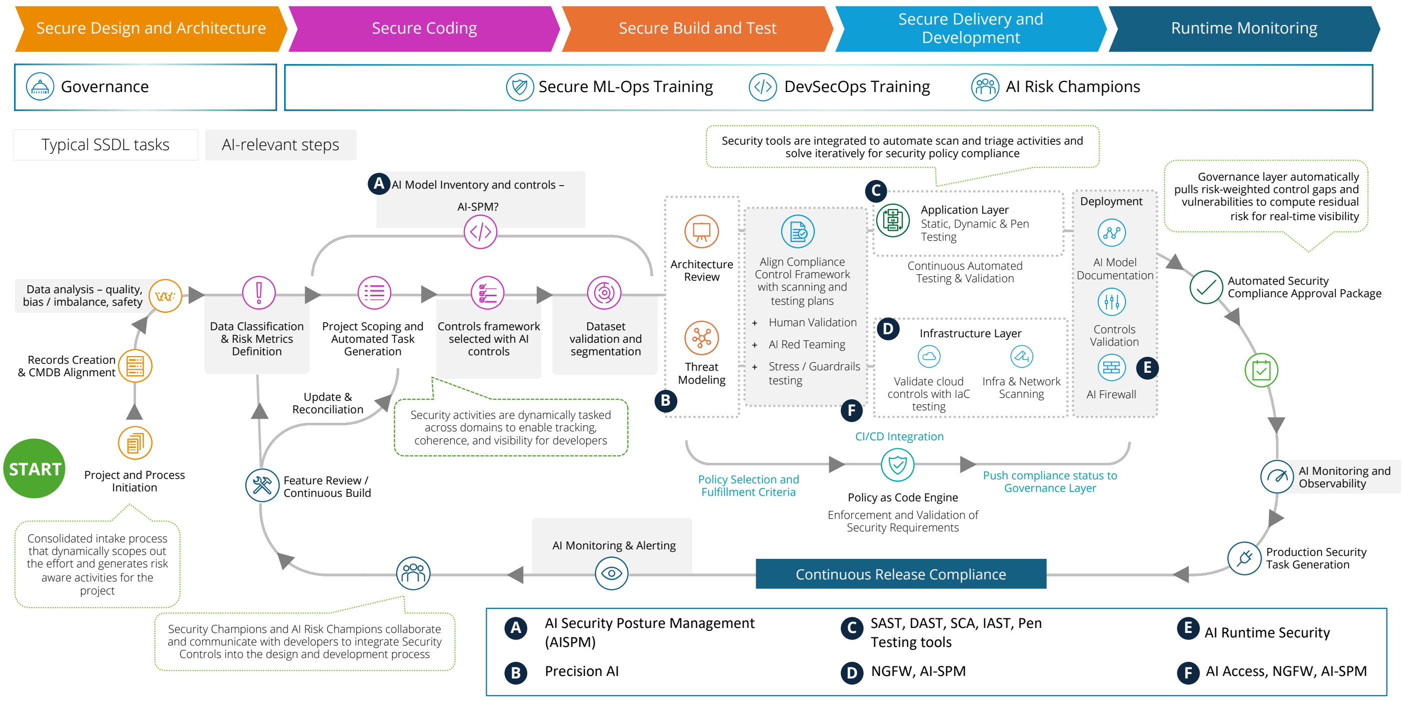

Secure Design and Architecture
Secure Coding
Secure Build and Test
Secure Delivery and Development
Runtime Monitoring
Governance
Secure ML-Ops Training
DevSecOps Training
AI Risk Champions
Typical SSDL tasks
AI-relevant steps
A AI Model Inventory and controls – AI-SPM?
Data analysis – quality, bias / imbalance, safety
Records Creation & CMDB Alignment
START
Project and Process Initiation
Consolidated intake process that dynamically scopes out the effort and generates risk aware activities for the project
Data Classification & Risk Metrics Definition
Project Scoping and Automated Task Generation
Controls framework selected with AI controls
Dataset validation and segmentation
Update & Reconciliation
Security activities are dynamically tasked across domains to enable tracking, coherence, and visibility for developers
Feature Review / Continuous Build
Security Champions and AI Risk Champions collaborate and communicate with developers to integrate Security Controls into the design and development process
AI Monitoring & Alerting
Security tools are integrated to automate scan and triage activities and solve iteratively for security policy compliance
Architecture Review
Threat Modeling
B
Align Compliance Control Framework with scanning and testing plans
+ Human Validation
+ AI Red Teaming
+ Stress / Guardrails testing
F
C
Application Layer
Static, Dynamic & Pen Testing
Continuous Automated Testing & Validation
D
Infrastructure Layer
Validate cloud controls with IaC testing
Infra & Network Scanning
Deployment
AI Model Documentation
Controls Validation
AI Firewall
E
Governance layer automatically pulls risk-weighted control gaps and vulnerabilities to compute residual risk for real-time visibility
Automated Security Compliance Approval Package
AI Monitoring and Observability
Production Security Task Generation
CI/CD Integration
Policy Selection and Fulfillment Criteria
Policy as Code Engine
Enforcement and Validation of Security Requirements
Push compliance status to Governance Layer
Continuous Release Compliance
A AI Security Posture Management (AISPM)
B Precision AI
C SAST, DAST, SCA, IAST, Pen Testing tools
D NGFW, AI-SPM
E AI Runtime Security
F AI Access, NGFW, AI-SPM

Figure 6 – SSDL with Precision AI Architecture

Here are the specific phases of SSDL in LLM application development.

7.1 Secure design and architecture

The initial phase, requirement analysis and planning, involves defining security specifications that align with the objectives of the LLM application, with a focus on sensitive data handling and compliance with security controls. Early threat modeling is important to identify risks specific to AI systems, such as adversarial attacks and data poisoning. For example, identifying potential attack vectors where adversaries could manipulate LLM inputs to produce harmful outputs.

During this phase, it is also important to create a secure architecture that integrates controls to safeguard AI data using secure protocols and encryption methods. Embedding privacy considerations within the design is essential for handling personal or sensitive data in LLM applications. Designing systems to treat LLM outputs as untrusted and inspecting for proper comparison and encoding before use exemplifies this phase.

7.2 Secure coding

Adhering to secure coding standards to prevent vulnerabilities, like prompt injection, is required . Mandatory security-focused code reviews and employing automated tools for static and dynamic analysis help detect and mitigate risks such as training data poisoning. For example, automated tools can help detect and mitigate training data poisoning threats.

7.3 Secure build and test

During this phase, secure build and test processes necessitate verification, which includes performing extensive security tests such as penetration testing and vulnerability assessments. Additionally, organizations should test for AI bias, fairness, and resistance to adversarial AI tactics. Regular checks help determine compliance with regulations like General Data Protection Regulation (GDPR) or Health Insurance Portability and Accountability Act (HIPAA), helping to identify and mitigate biases that could impact fairness and inclusivity.

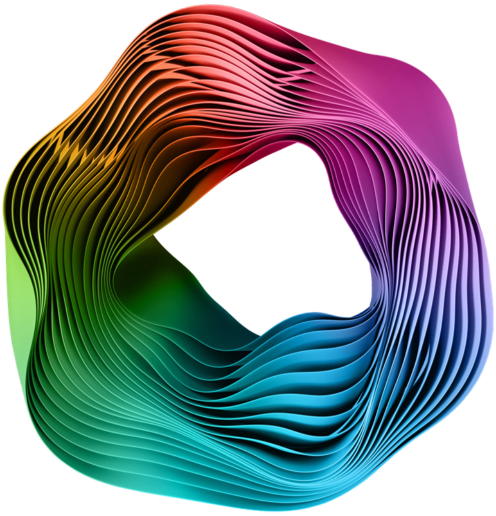
7.4 Secure delivery and deployment

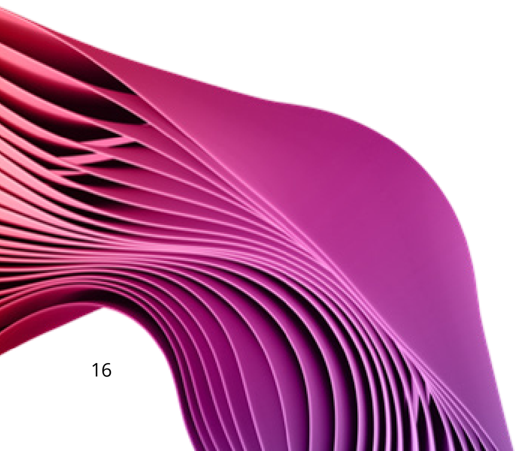
Secure delivery and deployment practices involve implementing secure deployment protocols and thoroughly reviewing deployment scripts for potential security issues. Stringent change management processes are essential to oversee system modifications, requiring that changes undergo broad security assessments. Encrypting models during deployment is important to help protect against model theft.

7.5 Runtime maintenance, operations, and end of life

Effective maintenance and operations can benefit from continuous near real-time monitoring strategies to promptly detect and respond to security incidents. A systematic approach to software updates is important for addressing vulnerabilities, and organizations should develop incident response plans tailored to LLM-specific security incidents. Continuous checking and near real-time alerts are vital for monitoring potential data leakage.

As the end-of-life phase approaches, strategic planning and execution should be considered for the secure decommissioning of LLM applications to help prevent data breaches and confirm full data erasure. Securely wiping all training data and model artifacts is important to prevent misuse after decommissioning.





8.0 Overview of AI guardrails for the LLM-based SSDL

It is increasingly common for LLM application development to be conducted using microservice architectures. This approach leverages the flexibility, scalability, and resilience of these technologies. Microservices allow for the efficient orchestration of containerized applications, which is critical for managing the complex workflows involved in AI/ML processes. Microservice architectures break down applications into smaller, manageable services, each running its own process and communicating via lightweight mechanisms like HTTP APIs. This modularity facilitates easier updates, scaling, and maintenance.

This section highlights AI guardrails specifically designed for the secure development lifecycle of LLM-based software. The following security guardrails should be considered to secure the development of LLM applications in a microservices environment like Kubernetes, leveraging container technologies. These considerations are specifically tailored to address the different risks associated with LLM applications. Integrating these security tools and practices into the software development lifecycle workflows is important for innovation efficiency, while protecting against various risks.

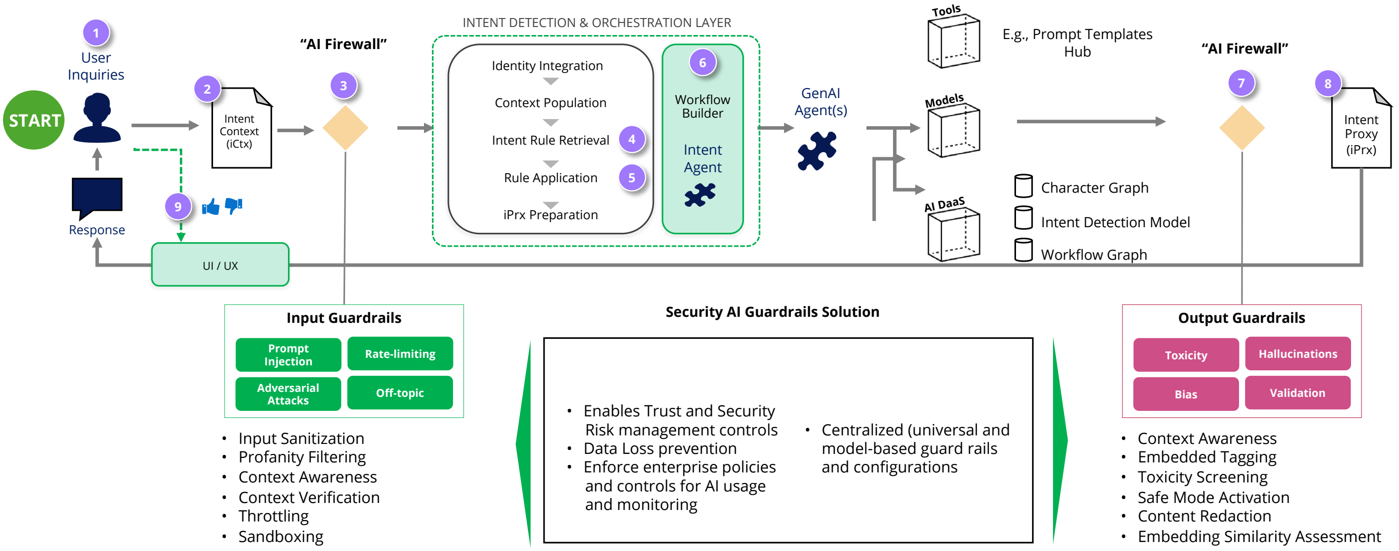


Figure 7: AI Guardrails for SSDL for Precision AI

In the context of AI guardrails within the LLM-based software development workflow, there are eight points to consider:

- **User inquiries:** Efficient and secure handling of user inquiries is important for maintaining LLM application integrity. User inputs, captured through various interfaces like web portals, APIs, and chatbots, should be compared and free from harmful content. Secure logging is essential for the checking and improvement aspect of AI application development to confirm that the system operates effectively, securely, and in compliance with relevant regulations. Tools like Web Application Firewalls (WAF) and API security solutions (e.g., Wide Area Augmentation System (WaaS) are designed to protect against malicious inputs and attacks, including prompt injection and Model Denial of Service (MDoS).[19]
- **Intent context:** Understanding the user's goal behind their input, extracting relevant information, and using workflow builders and intent agents to generate careful responses is vital. This enables LLMs to effectively answer questions, execute commands, and engage in meaningful conversations.[20]

- **Input guardrails and output guardrails[19]:**
  - **Input guardrails:** These mechanisms filter, compare, and sanitize user inputs to prevent LLM exploitation. Specific components include input comparison, profanity filtering, context awareness, context verification, throttling, and sandboxing. These determine secure and effective LLM operation by comparing inputs, filtering offensive content, understanding context, comparing appropriateness, limiting request rates, and isolating processes.
  - **Output guardrails:** These monitor and control LLM outputs to confirm they are safe, appropriate, and relevant. Specific components include context awareness, embedded tagging, toxicity screening, safe mode activation, content redaction, and embedding similarity assessment. These confirm that responses are contextually relevant, safe, and compliant with ethical and legal standards.
- **Intent rule retrieval:** This process matches user intent with predefined rules from a secure database, enabling consistent and careful rule application. Specific components include identifying user intent, matching intent with rules, retrieving relevant rules, applying rules, and comparison and testing.
- **Rule application:** Systematically applying predefined rules to interpret, process, and generate responses based on user inputs safeguards consistent, careful, and contextually appropriate outputs. Specific components include rule definition, management, execution, and monitoring and evaluation.
- **Workflow builders and intent agents:** These manage actions based on user intent using automation tools. Specific components include visual interfaces, action libraries, conditional logic, and integration capabilities. Intent agents use NLP intent recognition, context management, and response generation to carefully understand and respond to inputs.
- **Intent proxy:** Acting as an intermediary between user inputs and backend systems, the intent proxy manages and routes intents to appropriate services. Specific components include intent recognition, context management, routing mechanisms, response generation, and an integration layer, allowing for centralized management, scalability, flexibility, consistency, and security.
- **User interface (UI) and user experience UX:** These are critical for creating visually appealing, intuitive, efficient, and satisfying applications. Focusing on UI/UX enhances usability, engagement, accessibility, conversion rates, and brand loyalty, checking that the application addresses user expectations and provides a positive overall experience.

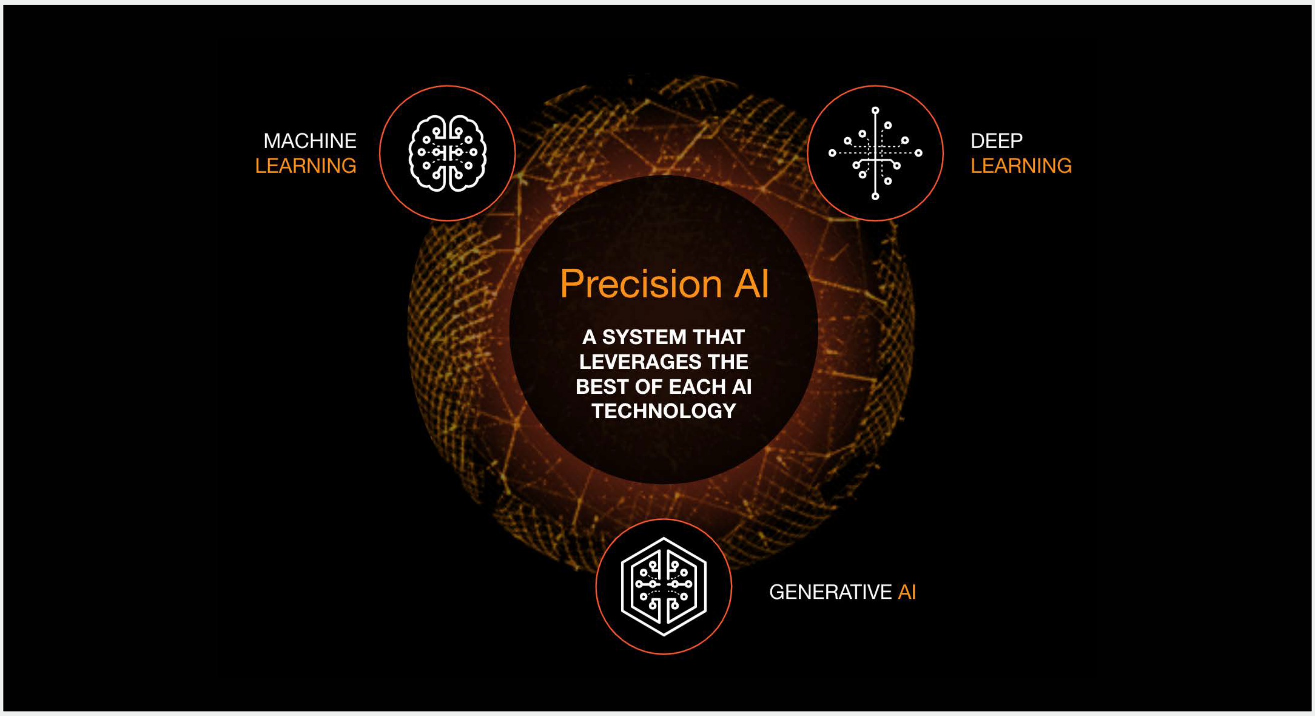


Figure 8: Palo Alto Networks' Precision AI™

9.0 Palo Alto Networks' Precision AI

Palo Alto Networks' Precision AI is a proprietary AI system designed to enhance the accuracy and efficiency of threat detection, prevention, and remediation in cybersecurity. It leverages AI/ML techniques, incorporating rich data and security-specific models to automate security tasks with industry-leading precision.[18]

9.1 Capabilities of Precision AI:

- **ML:** Improves the accuracy of security applications by using historical and current data to help predict and prevent novel security issues.
- **Deep Learning:** Builds predictive models that detect security issues in near real-time by learning from extensive security data.
- **GenAI:** Simplifies user experience and summarizes large volumes of threat intelligence, reducing mean time to resolution (MTTR).

9.2 Problems addressed by Precision AI

AI introduces new cybersecurity vulnerabilities, expanding the attack surface and providing cybercriminals with new vectors to target. Precision AI can help address several critical challenges:

9.2.1 Adversarial AI attacks:

Problem: Cybercriminals use AI to scale and accelerate attacks, circumvent existing security controls, and improve attack methods like phishing and prompt injection.
Solution: Uses advanced threat detection algorithms to identify and mitigate these sophisticated attacks in real-time.

9.2.2 Data poisoning and malicious code:

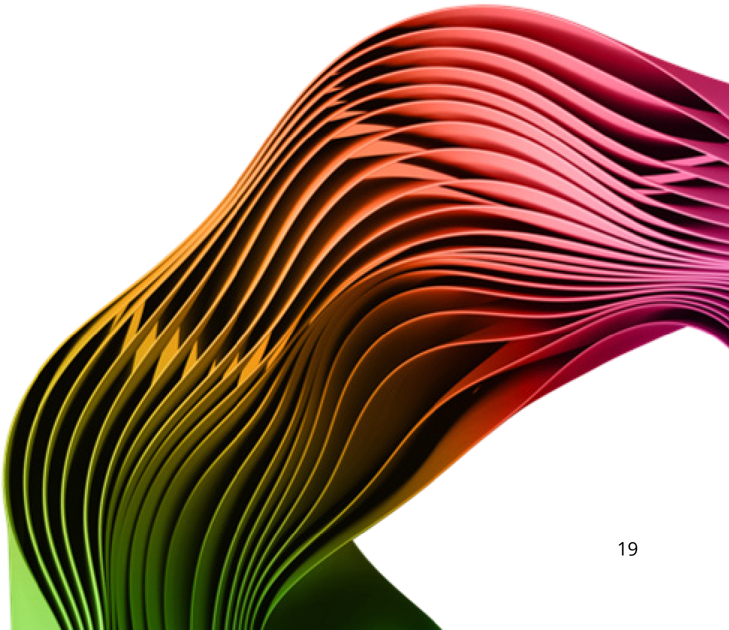
Problem: Adversarial AI can poison data or write malicious code, making it difficult for traditional security tools to identify.
Solution: Utilizes security-specific models to detect and prevent such attacks, safeguarding the integrity of AI infrastructure.

9.2.3 Inconsistent data quality and security silos:

Problem: Traditional approaches fall short due to inconsistent data quality and security silos.
Solution: Centralizes and analyzes vast amounts of security-specific data, providing high-resolution capabilities to automate detection, prevention, and response.

9.2.4 Skills gap in AI and cybersecurity:

Problem: There is a shortage of individuals with experience in both AI and cybersecurity.
Solution: Automates security tasks, reducing the burden on human analysts and allowing them to focus on higher-level strategic activities.



10.0 Potential Benefits of Precision AI

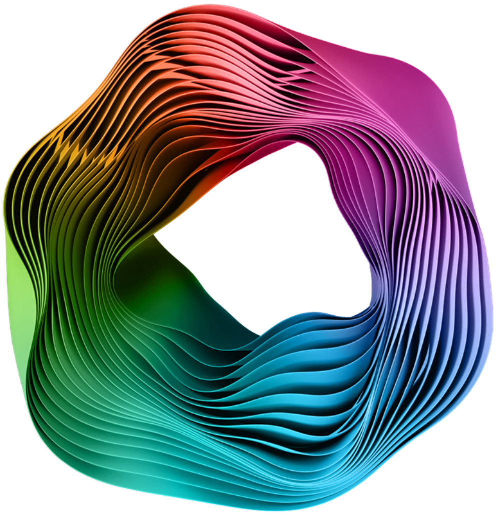
10.0.1 Combat AI-Driven threats: Enables organizations to evolve to real-time, autonomous security, mitigating advanced threats and improving MTTR. It helps anticipate and prevent new attack vectors in real-time, providing protection against AI-driven cyberattacks.[18]

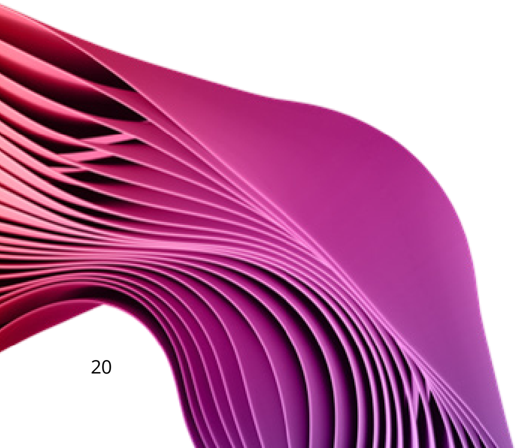
10.0.2 Simplify security: Transforms how practitioners interact with their security toolset, improving access to information, suggesting actions, and reducing time spent navigating user interfaces. This can help increase the productivity and effectiveness of cybersecurity teams.[18]

10.0.3 Secure AI by design: Helps protect AI infrastructure from compromise by using AI models to secure the entire AI roadmap. It identifies and mitigates new attack vectors, such as data poisoning and malicious code generation, confirming the security of AI-related projects and infrastructure.[18]

10.1 Takeaways of Precision AI

- **Impact on businesses:** AI is transforming cybersecurity, and organizations should consider adapting efficiently to address new challenges.[18]
- **Evolving cybersecurity strategy:** Organizations should evolve their cybersecurity strategies to incorporate AI-based solutions that provide real-time, precise results.[18]
- **Governance and compliance:** It is important to Implement governance and compliance models for AI to mitigate incremental cybersecurity risks.
- **Adversarial AI:** Understanding how adversaries leverage AI to circumvent security is essential for developing effective defenses.[18]





11.0 SSDL with Precision AI

Deloitte's LLM-based SSDL framework is designed to integrate security at many stages of the software development process. By incorporating Palo Alto Networks' Precision AI powered AI security tools such as AI-SPM, AI-RunTime Security, and AI-Access, Deloitte's LLM-based SSDL can enhance the security and operational efficiency of AI-based applications. Here's how this integration can help:

11.1 User inquiries:

- **Precision AI:** Utilizes advanced threat detection algorithms to monitor and analyze user inputs in real-time. It can identify and block malicious inputs, such as prompt injections and MDoS attacks, enabling user inquiries to be handled securely.
    - **AI-SPM:** Continuously assesses the security posture by identifying vulnerabilities in user inputs and recommending remediation actions.
    - **AI-Runtime Security:** Provides protection by monitoring and mitigating threats as they occur during user inquiries.
    - **AI-Access:** Manages and controls access to user input data, allowing only authorized users to interact with the AI systems.
- **SSDL integration:** Embedding Precision AI and AI security tools into the LLM-based SSDL framework enables continuous monitoring and comparison of user inputs, preventing harmful content from reaching the LLMs. Efficient and secure handling of user inquiries is critical for maintaining LLM application integrity. User inputs should be analyzed and free from harmful content.

11.2 Intent context:

- **Precision AI:** Enhances the understanding of user intent by analyzing patterns and context in user inputs. This enables the LLM to generate careful and relevant responses.
    - **AI-SPM:** Continuously monitors and improves the understanding of user intent by identifying potential security issues.
    - **AI-Runtime Security:** Provides real-time analysis and protection of user intent understanding.
    - **AI-Access:** Allows only authorized users to influence the context and patterns analyzed by the LLM.
- **SSDL integration:** Incorporating Precision AI and AI security tools helps in carefully capturing and interpreting user intent, which can lead to more precise and contextually appropriate responses from LLMs. Understanding the user's goal behind their input and extracting relevant information is important for effective responses.

11.3 Input guardrails and output guardrails:

11.3.1 Input guardrails:

- **Precision AI:** Implements input comparison, profanity filtering, context awareness, context verification, throttling, and sandboxing to prevent exploitation of the LLM.
    - **AI-SPM:** Continuously assesses and improves input comparison mechanisms.
    - **AI-Runtime Security:** Provides real-time protection by monitoring and mitigating threats during input processing.
    - **AI-Access:** Manages and controls access to input data, allowing only authorized users to provide inputs.
- **SSDL integration:** These input guardrails confirm that user inputs are sanitized and compared before being processed by the LLM, mitigating potential security breaches. These mechanisms filter, compare, and sanitize user inputs to enable secure and effective LLM operation.

11.3.2 Output guardrails:

- **Precision AI:** Monitors and controls LLM outputs checking that they are safe, appropriate, and relevant. It includes context awareness, embedded tagging, toxicity screening, safe mode activation, content redaction, and embedding similarity assessment.
  - **AI-SPM:** Continuously assesses and improves output comparison mechanisms.
  - **AI-Runtime Security:** Provides real-time protection by monitoring and mitigating threats during output generation.
  - **AI-Access:** Manages and controls access to output data, allowing only authorized users to access the generated outputs.
- **SSDL integration:** Output guardrails inspect that the responses generated by the LLM are safe, compliant, and contextually relevant, protecting against the dissemination of harmful or inappropriate content. These confirm that the responses are contextually relevant, safe, and in line with ethical and legal standards.

11.4 Intent rule retrieval:

- **Precision AI:** Matches user intent with predefined rules from a secure database, enabling consistent and careful rule application.
  - **AI-SPM:** Continuously assesses and improves the rule retrieval process.
  - **AI-Runtime Security:** Provides real-time protection by monitoring and mitigating threats during rule retrieval.
  - **AI-Access:** Manages and controls access to the rule database, allowing only authorized users to modify or access rules.
- **SSDL integration:** Confirms that user intents are carefully matched with predefined rules, leading to consistent and reliable application behavior. This process enables consistent and careful rule application by identifying user intent and matching it with relevant rules.

11.5 Rule application:

- **Precision AI:** Systematically applies predefined rules to interpret, process, and generate responses based on user inputs, enabling consistent, careful, and contextually appropriate outputs.
  - **AI-SPM:** Continuously assesses and improves the rule application process.
  - **AI-Runtime Security:** Provides real-time protection by monitoring and mitigating threats during rule application.
  - **AI-Access:** Manages and controls access to rule application processes, allowing only authorized users to modify or apply rules.
- **SSDL integration:** Confirms that all rules are applied consistently and carefully, maintaining the integrity and reliability of the application. Enables consistent, careful, and contextually appropriate outputs through rule definition, management, execution, and monitoring.

11.6 Workflow builders and intent agents:

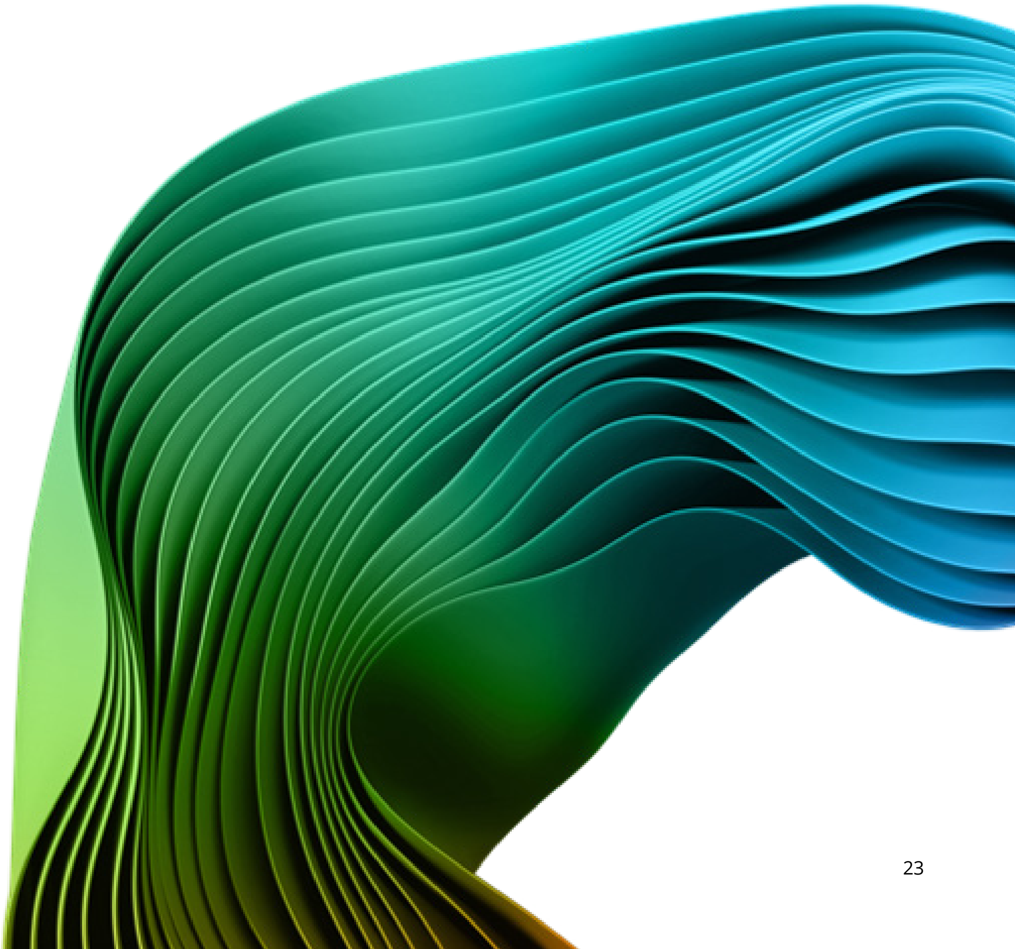
- **Precision AI:** Manages actions based on user intents using automation tools, including visual interfaces, action libraries, conditional logic, and integration capabilities.
  - **AI-SPM:** Continuously assesses and improves workflow automation and intent management processes.
  - **AI-Runtime Security:** Provides real-time protection by monitoring and mitigating threats during workflow execution.
  - **AI-Access:** Manages and controls access to workflow builders and intent agents, allowing only authorized users to modify or execute workflows.
- **SSDL integration:** Enhances workflow automation and intent management, confirming that user intents are processed efficiently and carefully. These manage actions based on user intents using automation tools, enabling careful understanding and response to inputs.

11.7 Intent proxy:

- **Precision AI:** Acts as an intermediary between user inputs and backend systems, managing and routing intents to appropriate services.
  - **AI-SPM:** Continuously assesses and improves intent proxy mechanisms.
  - **AI-Runtime Security:** Provides real-time protection by monitoring and mitigating threats during intent routing.
  - **AI-Access:** Manages and controls access to intent proxy mechanisms, allowing only authorized users to modify or access them.
- **SSDL integration:** Enables centralized management, scalability, flexibility, consistency, and security by effectively routing user intents to the appropriate backend services. Allows centralized management, scalability, flexibility, consistency, and security by managing and routing intents to appropriate services.

11.8 UI and UX:

- **Precision AI:** Enhances the overall user experience by providing a secure, intuitive, and efficient interface. It confirms that the application addresses user expectations and provides a positive overall experience.
  - **AI-SPM:** Continuously assesses and improves the security posture of the user interface.
  - **AI-Runtime security:** Provides real-time protection by monitoring and mitigating threats during user interactions.
  - **AI-Access:** Manages and controls access to the user interface, allowing only authorized users to interact with the AI systems.
- **SSDL integration:** Focuses on creating visually appealing, intuitive, efficient, and satisfying applications, enhancing usability, engagement, and accessibility.

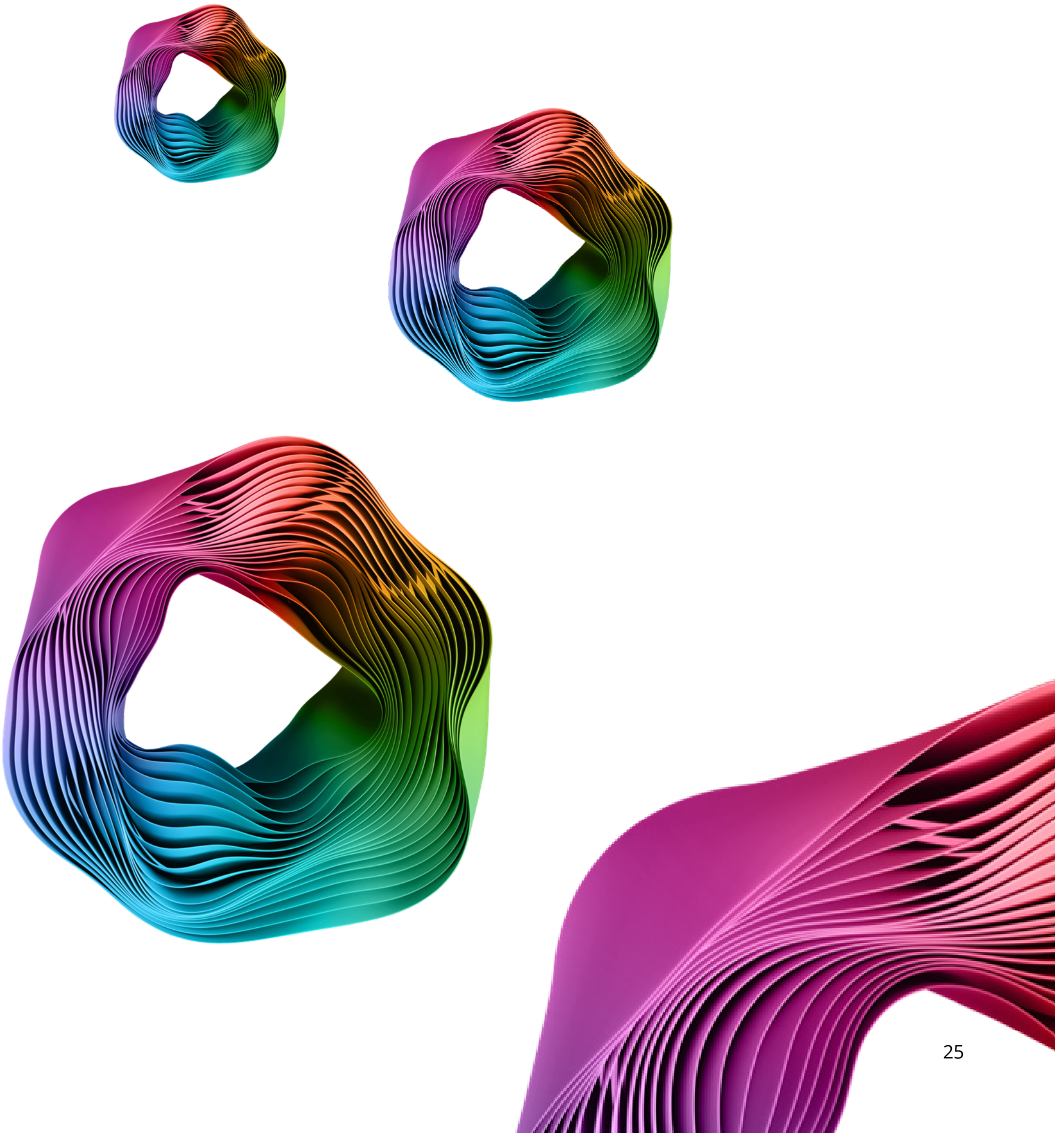


Steps to enhance security with AI integration

- **Capability assessment**
  - **Evaluate current infrastructure:** Assess the existing security infrastructure to identify gaps and strengths. Determine which areas of your GenAI and LLM-based application development need to be protected against rising cyber vulnerabilities.
  - **Technology readiness:** Evaluate the current technology stack's vulnerabilities against GenAI and LLM-based cyberattacks, including hardware, software, and network environments supporting new tools.
- **Strategic planning**
  - **Develop objectives:** Clearly define what you aim to achieve by integrating Palo Alto Networks' Precision AI and its AI security tools into Deloitte's LLM-based SSDL.
  - **Create a strategic framework:** Develop a broad plan outlining specific phase of integration, resource allocation, timelines, and risk management strategies. Enable alignment with the organization's overall cybersecurity strategy.
- **Tools integration**
  - **Tool integration assessment:** Evaluate the organization's ecosystem solutions in need of Palo Alto Networks' Precision AI and its AI security tools. Focus on assessing both the technical capabilities of these solutions and the support ecosystem they offer.
- **Phased implementation**
  - **Pilot program:** Implement a pilot program with the selected client ecosystem solutions in a controlled environment. This allows for monitoring effectiveness and adjustments without impacting broader GenAI and LLM application development operations.
  - **Gradual rollout:** Based on the pilot's achievements, gradually implement the technology across the GenAI and LLM application development. A phased deployment model helps decrease disruptions and allows for continuous assessment and adjustment.
- **Training and empowerment**
  - **Skill development:** Invest in training programs to upskill GenAI and LLM application development personnel on the new security tools. Establish that they understand how to operate the new systems effectively and leverage AI-enhanced capabilities.
  - **Change management:** Support the team through the integration process with clear communication and involvement in decision-making. This helps manage change resistance and fosters a culture of innovation.
- **Continuous monitoring and evaluation**
  - **Monitor performance:** Continuously monitor the system's performance against the pre-defined objectives set in the strategic plan. Use these insights to refine processes and technology deployment.
  - **Iterative requirement:** Regularly update strategies and tools based on operational feedback and evolving security threats. Maintain flexibility to adapt to new developments and technologies in the AI/ML landscape.
- **Scalability and future proofing**
  - **Evaluate scalability:** Regularly assess the scalability of the integrated AI/ML solutions to handle increased loads or expanding security requirements.
  - **Sustainability planning:** Plan for the long-term sustainability of AI/ML integrations by staying updated on advancements in AI/ML technologies and continuously assessing their potential impact on GenAI and LLM application development operations.

Conclusion

- Deloitte's LLM-based SSDL framework, when integrated with Palo Alto Networks' Precision AI and its suite of AI security tools (AI-SPM, AI-Runtime Security, and AI-Access), offers an effective solution for helping organizations secure GenAI and LLM-based applications. This integration enhances the security and operational efficiency of AI-driven solutions by embedding security measures at many stages of the software development process. From user inquiries and intent context to input and output guardrails, rule application, workflow automation, and user experience, the combined capabilities confirm that AI applications are secure, efficient, and aligned with leading practices. By leveraging these advanced security tools, organizations can more effectively mitigate cyber threats while fostering innovation efficiency.
- By integrating Palo Alto Networks' Precision AI and its suite of AI security tools (AI-SPM, AI-Runtime Security, and AI-Access) into Deloitte's LLM-based SSDL, organizations can enhance their security posture against AI-based cyberattacks.
- For more information on how Deloitte can help you secure your AI applications, contact us today. Let's work together to help safeguard your digital future.



References

1 Lawton, G. (2024, June 24). *What is Generative AI? Everything You Need to Know. TechTarget.*

2 Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative ai. *Business & Information Systems Engineering*, 66(1), 111-126.

3 Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., & Gadekallu, T. R. (2024). Gpt (generative pre-trained transformer)–a broad review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*.

4 Larsen, P., & Proserpio, D. (2023). The Impact of Large Language Models on Search Advertising: Evidence from Google's BERT. *USC Marshall School of Business Research Paper Sponsored by iORB*.

5 Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2023). Llama 2: Early Adopters' Utilization of Meta's New Open-Source Pretrained Model.

6 Bent, A. A. (2023). Large Language Models: AI's Legal Revolution. *Pace Law Review, 44(1)*, 91.

7 Huang, K., Manral, V., & Wang, W. (2024). From LLMOps to DevSecOps for GenAI. In *Generative AI Security: Theories and Practices* (pp. 241-269). Cham: Springer Nature Switzerland.

8 Hurani, M., & Idris, H. (2024). Investigating the use of LLMs for automated test generation: challenges, benefits, and suitability.

9 Barker, D. (2016). *Web content management: Systems, features, and best practices*. " O'Reilly Media, Inc.".

10 Luz, A. (2024). *Enhancing the Interpretability and Explainability of AI-Driven Risk Models Using LLM Capabilities* (No. 13368). EasyChair.

11 Forsén, F. (2024). Large Language Models and business applications in an R&D environment.

12 Eapen, J., & Adhithyan, V. S. (2023). Personalization and customization of llm responses. *International Journal of Research Publication and Reviews,* 4(12), 2617-2627.

13 Bhat, A. (2024). A human-centered approach to designing effective large language model (llm) based tools for writing software tutorials.

14 Kähkönen, S. (2024). Improving software development workflows using generative AI.

15 Zhang, A. (2024). Impact of LLMs on Academic Literature Synthesis: Influence and Oversight in Business Economics and Other Disciplines.

16 Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

17 Fasha, M., Rub, F. A., Matar, N., Sowan, B., Al Khaldy, M., & Barham, H. (2024, February). Mitigating the OWASP Top 10 For Large Language Models Applications using Intelligent Agents. In 2024 2nd *International Conference on Cyber Resilience* (ICCR) (pp. 1-9). IEEE.

18 Palo Alto Networks. (2024, September). *What Is Precision AI™?* Palo Alto Networks.

19 Biswas, A., & Talukdar, W. (2023). Guardrails for trust, safety, and ethical development and deployment of Large Language Models (LLM). *Journal of Science & Technology*, 4(6), 55-82.

20 Dzeparoska, K., Tizghadam, A., & Leon-Garcia, A. (2024). Intent Assurance using LLMs guided by Intent Drift. *arXiv preprint arXiv:2402.00715.*

21 Sánchez Cuadrado, J., Pérez-Soler, S., Guerra, E., & De Lara, J. (2024, July). Automating the Development of Task-oriented LLM-based Chatbots. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (pp. 1-10).

The strength of the Deloitte and Palo Alto Networks' alliance

Deloitte's award-winning Cyber practice has joined forces with Palo Alto Networks and its security capabilities platform. Together, we're working to provide a broad range of capabilities that aim to help simplify the complex software development lifecycle, while increasing speed, agility, and enablement so that organizations like yours may better protect their infrastructure and workloads at many stages of the development lifecycle. Our joint solution may aid you in creating a cyber-minded culture for your organization so that it can move forward faster and stronger, fuel more innovation, and stay more resilient in the face of persistent and ever-changing threats—while accelerating time to market and reducing costs.

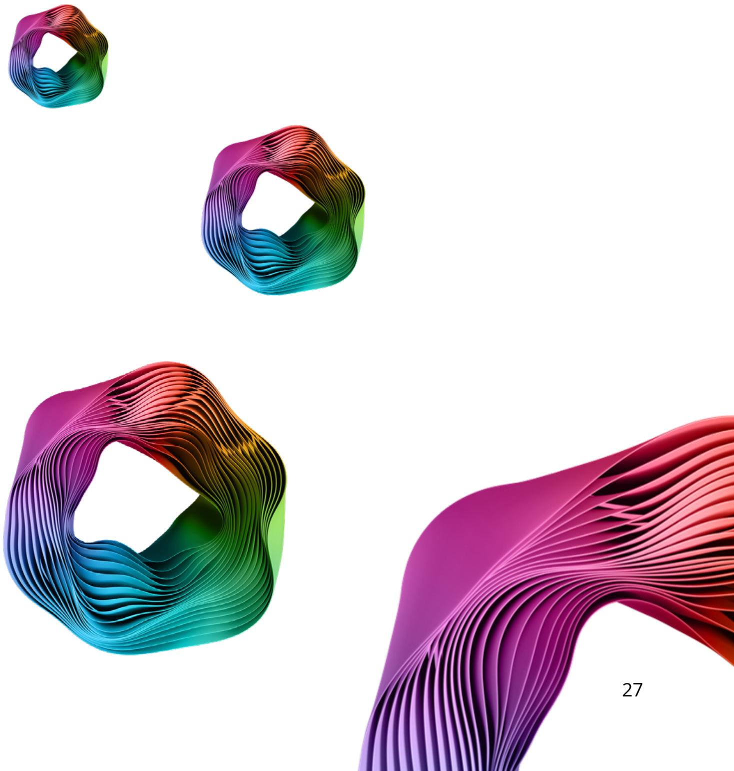
Authors

Kieran Norton
Principal
US Cyber & Strategic Risk
Deloitte & Touche LLP
kinorton@deloitte.com

Jane Chung, Ph.D.
Managing Director
US Cyber & Strategic Risk
Deloitte & Touche LLP
jachung@deloitte.com

Anthony Polzine
Senior Manager
Global Partner Solution Architect
Palo Alto Networks
apolzine@paloaltonetworks.com

Siddharth Kantroo
Advisory Senior Manager
US Cyber & Strategic Risk
Deloitte & Touche LLP
skantroo@deloitte.com



Deloitte.

About this publication

This publication contains general information only and Deloitte and Palo Alto Networks are not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional adviser. Deloitte and Palo Alto Networks shall not be responsible for any loss sustained by any person who relies on this publication.

All product names mentioned in this publication are the trademarks or registered trademarks of their respective owners and are mentioned for identification purposes only. Deloitte is not responsible for the functionality or technology related to the vendor or other systems or technologies as defined in this publication.

As used in this publication, "Deloitte" means Deloitte & Touche LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.