

Playbook

AI Guardrails

Operating Models for AI Guardrail
Creation, Evaluation, and Monitoring

Published: March 18, 2025



Leaders enabling enterprise-wide secure and compliant AI trust
Dynamo AI's Guardrails Playbook for *definitive direction on how to create, evaluate, and monitor AI guardrails*. This Playbook provides the critical details (process steps, pitfalls to avoid, metrics, implementation tips, roles and responsibilities, resource estimates) to deploy *AI with confidence*.

Overview	Playbook background, scope, and about Dynamo AI
Playbook Management	Guidance on Dynamo's efforts to continually update and enhance the Playbook
The AI Risk Management Imperative	Perspective on AI risk management and regulatory considerations that underpin the Playbook
AI Life Cycle Focus	A visual of the AI Life Cycle with roles and responsibilities
AI Guardrails Overview	A brief introduction to AI guardrails
How to Create Guardrails	Best practices for defining and creating AI guardrails
How to Evaluate Guardrails	Best practices for evaluating the performance of guardrails
How to Monitor Guardrails	Best practices for monitoring guardrails post deployment
Appendix	AI Life Cycle component descriptions, consolidated success criteria, and glossary



01. Overview	3	08. How to Monitor Guardrails	39
Introduction	3	Overview	39
Scope	3	High-Level Process and Common Pitfalls ..	39
About Dynamo AI	4	Roles & Responsibilities and Related Industry Practices	40
Unique AI Control Insight	4	Playbook Detail: Monitoring Guardrails	41
02. Playbook Management	5	Success Criteria	47
Playbook Management	5	09. Appendix	48
03. The AI Risk Management Imperative	6	AI Life Cycle Component Descriptions	48
The AI Risk Management Imperative	6	Consolidated Success Criteria Checklist	52
04. AI Life Cycle Focus	8	Glossary	54
AI Controls within the AI Life Cycle	8		
AI Life Cycle Diagram	9		
Key Roles & Responsibilities	10		
05. AI Guardrails Overview	11		
What is a Guardrail	11		
06. How to Create Policy Guardrails	12		
Overview	12		
High-Level Process and Common Pitfalls ..	12		
Guardrail Creation Approaches and Roles & Responsibilities	13		
Playbook Detail: How to Create Guardrails ..	17		
07. How to Evaluate Guardrails	27		
Overview	27		
High-Level Process and Common Pitfalls ..	27		
Roles & Responsibilities	29		
Playbook Detail: Guardrail Evaluation Process	31		
Success Criteria	38		



01 Overview

1.1 Introduction

1. National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), (2023), <https://doi.org/10.6028/NIST.AI.100-1>, p. 9

2. Australian Government, Department of Industry and Science, 10 Guardrails: Voluntary AI Safety Standard, <https://www.industry.gov.au/publications/voluntary-ai-safety-standard/10-guardrails>.

3. Department for Science, Innovation and Technology, Implementing the UK AI Regulatory Principles: Guidance for Regulators (2024), https://assets.publishing.service.gov.uk/media/65c0b6bd63a23d0013c821a0/implementing_the_uk_ai_regulatory_principles_guidance_for_regulators.pdf.

Since the mainstream advancement of generative artificial intelligence (AI) in late 2022, governments, enterprises, and trade associations have put forth high-level guidance focused on deploying AI in a safe and secure manner. The reasons for doing so are clear, as the impacts of AI will be felt across *People and Planet*,¹ and therefore must have a level of *governance, mapping, measuring, and managing* AI risk that is appropriately aligned to the expected level of impact across society.

At the same time, governments have advanced AI policy, with regulation from the European Union (EU), principles from Australia² and the United Kingdom,³ as well as laws passed by individual states within the United States (New York, Colorado, and Utah to name a few). And while clarity regarding principled use is required, a core barrier to AI enablement for most enterprises is best practice methods to implement real-world AI controls that mitigate risks inherent to AI.

Best practice guidance that includes methods, pitfalls, expectations on observability and evidence, to drive the assessments of AI residual risk and promote effective decision-making and auditability.

The *Dynamo AI Guardrail Playbook* (Playbook) is meant to satisfy this gap. This Playbook provides a comprehensive overview of how best to **create, evaluate, and monitor AI guardrails**, critical post-deployment AI controls that are specifically focused on mitigating risks that result from the use, or interaction with, AI including Large Language Models (LLM) in the context of operating a regulated enterprise. The controls, evidence, and monitoring best-practices all aim to satisfy compliance with internal policy, rules, laws, and regulations related to AI use case deployment. The Playbook is also meant to be understood by both technical and non-technical stakeholders and be delivered in an 'easy-to-understand' way. Ultimately, **this Playbook seeks to enable AI** across a variety of use cases, in a safe and sound manner.

The Playbook documents specific controls and associated control execution and monitoring requirements to evaluate and mitigate risks inherent in the use of AI as part of a deployed use case.

1.2 Scope

The Playbook does not aim to cover or include:

- A holistic AI risk management framework, progressed by organizations including the National Institute of Standards and Technology, International Organization for Standardization (ISO), and the Organization for Economic Cooperation and Development (OECD).
- A focus on singular risk stripes such as Information Security or Privacy (progressed by organizations such as the Open Worldwide Application Security Project (OWASP) or The MITRE Corporation).
- End-to-end AI systems development (Section 4 provides a reference to relevant processes).
- Validation or mappings to regulations that may be required for a specific AI use case in a market or region.



1.3 About Dynamo AI

Founded in 2021 by a team of Ph.D.'s from the Massachusetts Institute of Technology (MIT) and security experts at the forefront of compliant AI, Dynamo AI enables enterprises to deploy safe, secure, and compliant AI systems at scale.

Dynamo AI's solutions are designed to address the risks and challenges with adopting generative AI where security and compliance is paramount. The Dynamo AI platform is distinguished as the first end-to-end compliant generative AI infrastructure specifically engineered for large-scale deployment across industries. Dynamo AI's product suite can be run either in major virtual private clouds, on-premises, or on edge devices.

Dynamo AI is backed by organizations including Y-Combinator and Canapi Ventures (a consortium of 40 of the top 100 US financial institutions). This consortium of financial institutions partner with Dynamo AI to enable high-value AI use in environments that require a high level of governance and control.

1.4 Unique AI Control Insight

Dynamo AI's mission *"Empower every organization to harness AI's transformative potential with confidence and control"* requires that it be at the forefront of AI evaluation research, which is embedded into the culture of our organization. To design, develop, deploy, and oversee a comprehensive GenAI control suite, a number of critical components must be in place. For Dynamo AI, those include:

01.

Being a 'research-led' organization, with top machine learning researchers that vet and fact check their research, present and are awarded at top machine learning and AI conferences.

02.

Having worked with some of the largest and most complex enterprises in heavily regulated sectors to enable critical AI use cases.

03.

Maintaining independence as Dynamo AI designs and deploys products, ensuring Dynamo AI remains a trusted technology to provide effective oversight and challenge of LLM providers.

04.

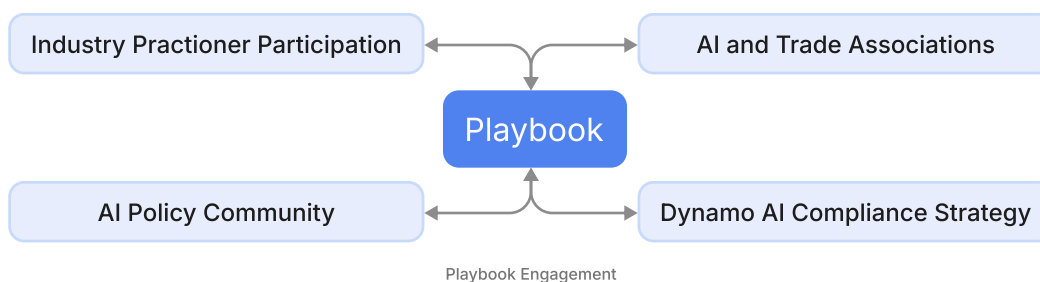
Incorporating global policy maker guidance through direct engagement across our end-to-end test, evaluation, and guardrail technology for comprehensive AI risk management.

It is through this unique combination of developing advanced AI test, evaluation, and guard railing technology, in-house AI expertise, early adopter client engagement, and requirements for best practice within the current cycle of AI advancement that makes Dynamo AI guidance critical for enterprises at this moment in time.

02 Playbook Management

2.1 Playbook Management and Advancement

Dynamo AI, as the steward of this Playbook, is establishing a program to ensure these best practice recommendations remain relevant through the maturation of enterprise AI deployment.



This program includes the following assessment mechanisms:

- ✓ A twice-annual review overseen by Dynamo's Head of AI Compliance Strategy with AI research, machine learning, and product leadership to review and update each Playbook with learning, best practice, and control enhancement requirements.
- ✓ A twice-annual distribution to AI Council members, which consists of a comprehensive list of AI, technology, and executive leaders within financial institutions across the US. For more on the AI Council, please engage with Dynamo AI.
- ✓ At least annually, an off-site session hosted by Dynamo AI with industry participation on reviewing, validating, and articulating emerging control requirements for inclusion into the Playbook.
- ✓ At least annually, a distribution of the Playbook to policy members, including members of Congress, for commentary.
- ✓ At least annually, the distribution of the Playbook to non-governmental enterprises and trade associations for commentary. Please engage Dynamo AI for further details or to be included in this distribution.

Please note, this Playbook is solely focused on the implementation of AI guardrail controls. Other Playbooks within Dynamo AI's series contain additional AI control topics, such as the testing and evaluation of AI. Please reach out to Dynamo AI for more detail on the full series of Dynamo Playbooks.

03

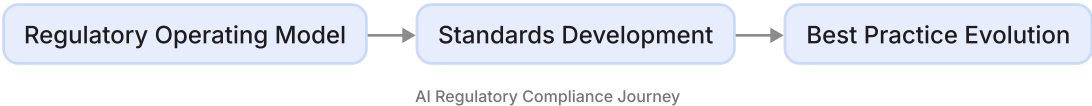
The AI Risk Management Imperative

3.1 The AI Risk Management Imperative

4. Goldmacher, Shane. "Here's a running list of all the state-level AI legislation in 2024." StateScoop, January 2, 2024, <https://statescoop.com/ai-legislation-state-regulation-2024/>.

Enterprises witnessed a significant surge in global AI-related guidance and regulation in 2024. Within the US alone, state policymakers introduced close to 700 pieces of AI regulation,⁴ with some states (Colorado and Utah) passing targeted AI-use case laws. While 2025 promises to continue this trend, early regulatory movers (i.e. the EU AI Act) will transition to testing mode as key provisions become enforceable.⁵ Regulatory expectations will continue to arise from a broad swath of regulatory imperatives, from AI-focused regulatory requirements to existing model, privacy, or consumer-focused regulation pertinent for targeted use cases, where an enterprise must demonstrate compliance.

AI Regulatory Compliance Journey



5. Charles-Albert Helleputte, Claire Murphy, and Andrea Otaola. (September 2024). The EU AI Act Enters Into Effect – What You Should Know and What Should You Do? Squire Patton Boggs. <https://www.squirepattonboggs.com/en/insights/publications/2024/09/the-eu-ai-act-enters-into-effect-what-you-should-know-and-what-should-you-do#:~:text=With%20the%20entry%20into%20force,system%20will%20not%20be%20enough.>

Dynamo AI's series of Playbooks contain 'best practice' operating requirements to establish AI controls and facilitate compliance with existing and emerging regulatory guidance. Today, global enterprises are currently navigating the *Regulatory Operating Model* phase of the AI regulatory compliance journey, where the foundation of regulatory expectations (and subsequent risk tolerances) are being defined. This will eventually lead to the next two phases: *Standards Development* and *Best Practice Evolution*. Within the Standard Development phase, industry helps shape best-practice standards that subsequently influence policy going forward. The Best Practice Evolution phase, in particular for regulated entities, is critical for AI use case advancement as this is where continued control implementation leads to executional refinement and strengthening in order to satisfy standards. Dynamo AI's Playbook detail is critical for these last two phases, as organizations look to implement and strengthen AI oversight.

Playbook Alignment with Model Risk Management Guidance

Another critical source of AI control expectations comes from model risk management guidance, in particular from the financial services industry. This guidance provides details around how a model should be evaluated, implemented, and monitored. Dynamo AI's series of Playbooks incorporates key tenants from model risk management guidance (as seen in the following table) and delivers best practices that satisfy or mitigate risks that may arise from many of the AI model requirements documented.



SR11-7

Federal Reserve
USA

Section IV. Model Development, Implementation, and Use

- Model testing includes checking the model's accuracy, demonstrating that the model is robust and stable, assessing potential limitations, and evaluating the model's behavior over a range of input values.
- It should also assess the impact of assumptions and identify situations where the model performs poorly or becomes unreliable.
- Testing should be applied to actual circumstances under a variety of market conditions, including scenarios that are outside the range of ordinary expectations, and should encompass the variety of products or applications for which the model is intended.
- Extreme values for inputs should be evaluated to identify any boundaries of model effectiveness.
- An understanding of model uncertainty and inaccuracy and a demonstration that the bank is accounting for them appropriately are important outcomes of effective model development, implementation, and use.

SS1 / 23

Bank of England, Prudential
Regulatory Authority (PRA)
UK

Principle 3.2: The Use of Data

- The model development process should ensure there is no inappropriate bias in the data used to develop the model, and that usage of the data is compliant with data privacy and other relevant data regulations.

Principle 3.3: Model Development Testing

- "...Performance tests should also include comparisons of the model output with the output of available challenger models, which are alternative implementations of the same theory, or implementations of alternative theories and assumptions."

Principle 3.4: Model Adjustments and Expert Judgement

- "... demonstrate that risks relating to model limitations and model uncertainties are adequately understood, monitored, and managed..."

Principle 3.5: Model Development Documentation

- Model development documentation should be sufficiently detailed so that an independent third party with the relevant expertise would be able to understand how the model operates.

Information Paper
Artificial Intelligence
Model Risk
Management

Monetary Authority of
Singapore

Section 6: Development

- Datasets chosen for training and testing or evaluation of AI models were expected to be representative of the full range of input values and environments under which the AI model was intended to be used. Training and testing datasets were also checked to ensure that their distributions or characteristics are similar.
- "...testing datasets that allowed predictions or outputs from AI models to be tested or evaluated in the bank's context as far as possible."
- Applying explainability methods to identify the key input features or attributes that are important for the AI model predictions or outputs and assessing that they are intuitive from a business and/or user perspective.
- "... required developers to apply global and/or local explainability methods to identify the key features or attributes used as inputs to AI models and their relative importance... "
- "... Model selection details of how the performance of the AI model was evaluated and how the final model was selected."

04

AI Life Cycle Focus

4.1 AI Controls within the AI Life Cycle

There is no single path or procedure to deploy AI use cases. The actions to consider vary too broadly across organizational makeup (size and complexity), use case, regulatory regime, and risk tolerance. But there are a number of high-level operating components that appear to be influencing best practice approaches. These components are being championed by a variety of sources including:

01.

Non-regulatory government agencies

02.

Government and regulatory guidance

03.

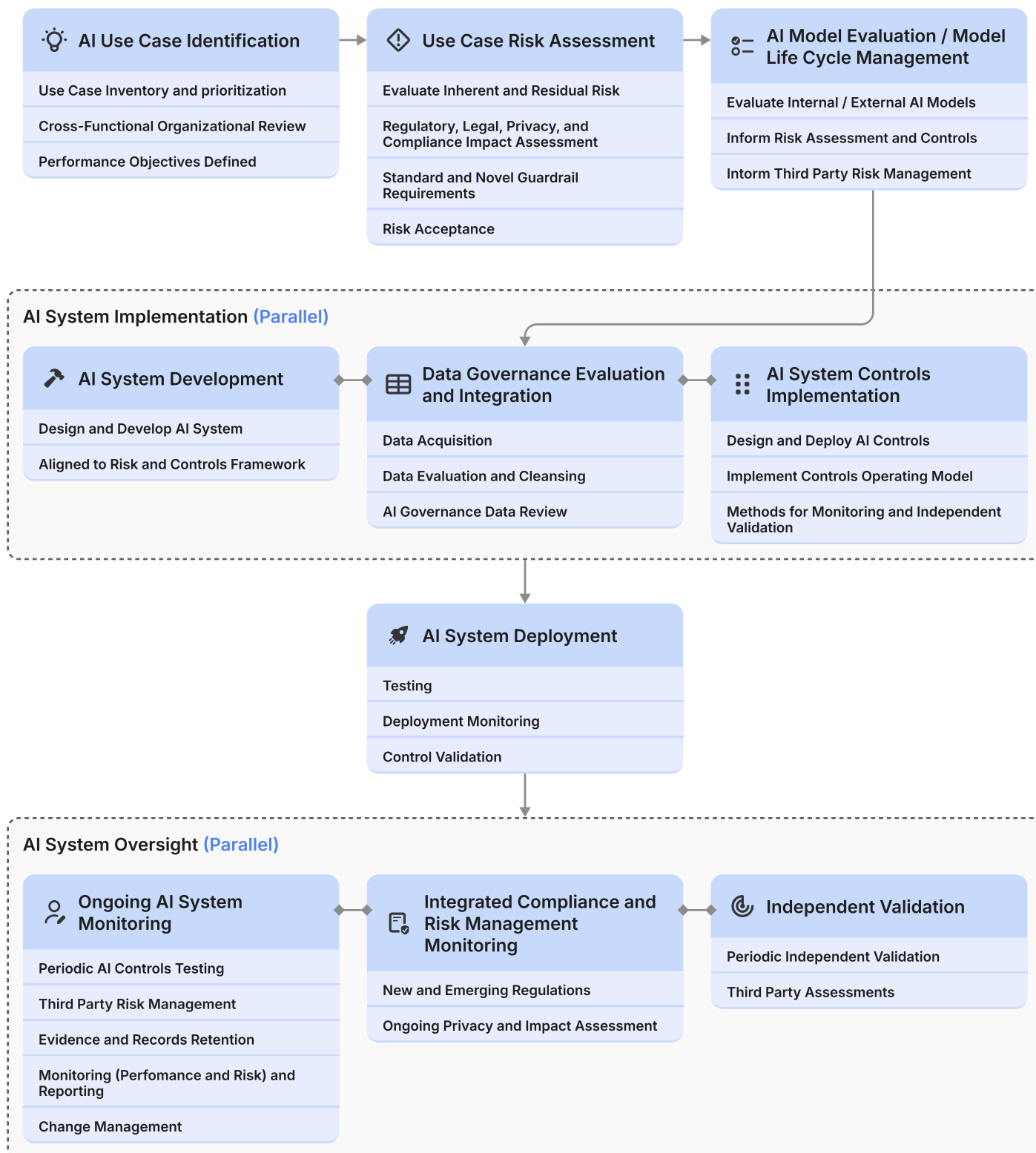
Best-practice guidance from management consulting organizations

04.

Industry, including highly regulated institutions (such as financial services) deploying AI

The AI lifecycle in Section 4.2 is presented to help guide where the individual controls set forth in Section 6, 7, and 8 are required when looking to deploy AI across an enterprise.

4.2 AI Life Cycle Diagram



4.3 Key Roles and Responsibilities

While there is no standard set of roles and responsibilities across each enterprise or industry, there is a set of standard functional roles that are critical to AI lifecycle execution and oversight. For purposes of this document and the detailed control inventory set out in Sections 6, 7, and 8, the following are critical roles for effective implementation.



Product Stakeholders

Stakeholders responsible for defining the vision, functionality, and business priority for the AI system. This set of stakeholders within an organization is ultimately responsible for the success of the AI system and achieving the business objective. These stakeholders partner with other stakeholders (technology, risk management) to design and deploy the AI system.



Technology Development Stakeholders

Stakeholders responsible for the technical design, development, integration, and maintenance of the AI system. This set of stakeholders is also focused on application performance, scalability, testing, reliability, and ensuring the technology satisfies Product stakeholder requirements. This set of stakeholders also contains experts in machine learning and large language models.



Information Security & Privacy Stakeholders

Stakeholders responsible for ensuring that people, process, and technology are in place to safeguard against information security and cyber security threats and satisfy privacy requirements. These stakeholders identify vulnerabilities, partner with others in the enterprise to build robust security measures, design and support deployment of privacy controls, and work to satisfy compliance requirements for AI systems.



Risk Management Stakeholders

Stakeholders responsible for identifying and assessing the risk of AI deployments for the organization. Risk Management stakeholders are often identified by their specific risk expertise and span a broad variety of risk types including technology, data, operational, model, compliance, legal, strategic, reputational, and financial risk among others. Risk management stakeholders partner with other stakeholders to identify controls that mitigate identified risks, and facilitate any risk acceptance processes.



Legal & Compliance Stakeholders

Stakeholders with legal and compliance requirements expertise that help guide design, development, and oversight of an AI system. These stakeholders identify applicable rules, laws, and regulations pertinent to the AI use case and partner to establish people, process, and technology controls to satisfy and monitor the AI use case in order to satisfy identified requirements and validate compliance.



Audit Stakeholders

Stakeholders responsible for independently evaluating the AI system to assess its functionality, performance, and alignment with risk and internal policy objectives. These stakeholders may be employees within an enterprise or a third-party may be identified to perform an independent evaluation on behalf of the enterprise.



05

AI Guardrails Overview

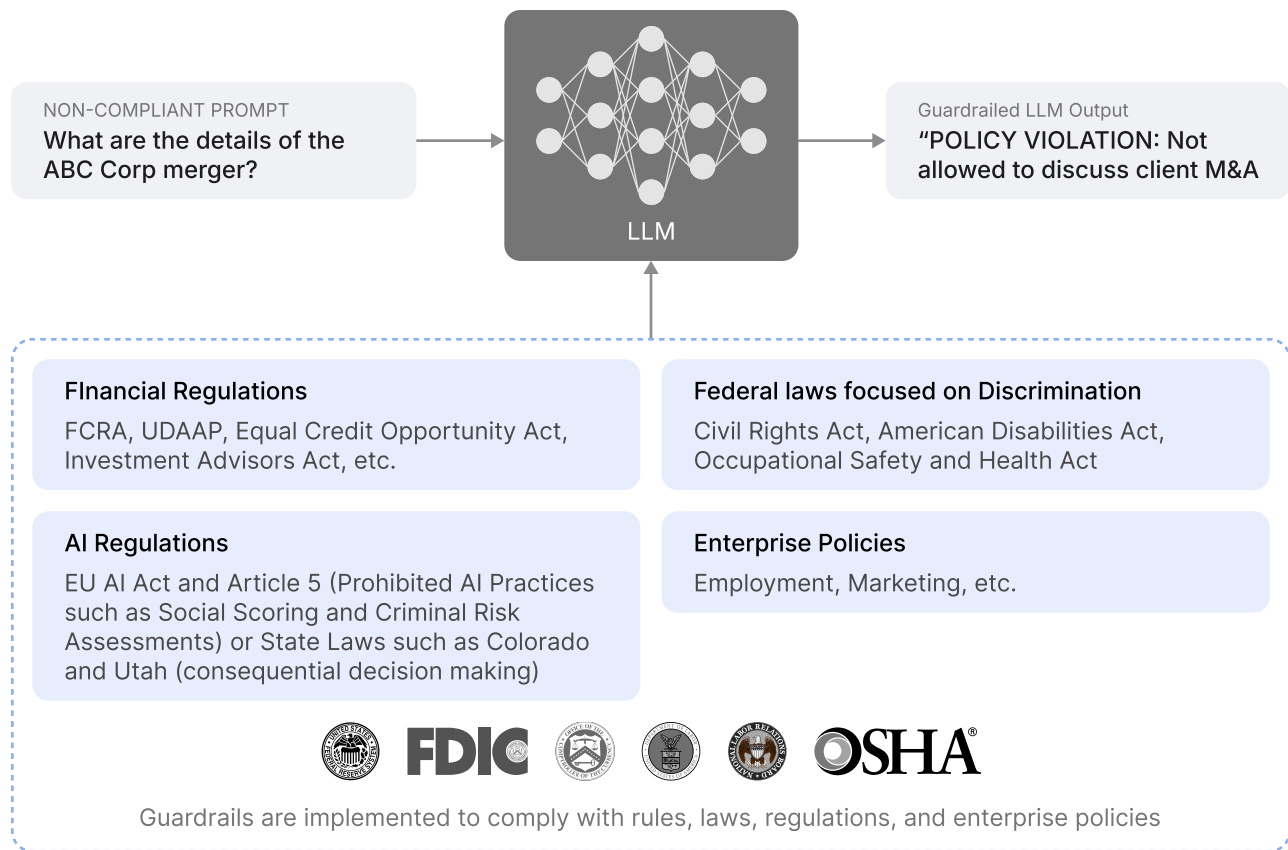
5.1 What is a Guardrail?

Overview of a Guardrail

What is a Guardrail in the context of this Playbook? Guardrails are lightweight control mechanisms that translate organizational policies into actionable controls that govern inputs and outputs of an AI system. Guardrails can address a variety of legal, regulatory, and organizational risks and can be applicable to a broad set of use cases. By moderating, blocking, or flagging interactions, guardrails align AI system behavior to policies and mitigate such risks.

Guardrail implementations can vary from heuristic rules to sophisticated machine learning models, however their goal is the same: ensure AI system compliance. It is critical that guardrail development occurs in parallel with AI system development, rather than being an afterthought. Effective guardrail controls also require substantial evaluation of guardrail performance and continual monitoring of guardrail effectiveness.

DynamoGuard enables enterprises to define guardrails based on natural-language rules or requirements. We refer to these as policy-based guardrails. Each guardrail is implemented using proprietary Small Language Models (SLMs) that ensure high performance and minimal overhead.



Guardrails may be implemented to comply with rules, laws, regulations, and enterprise policies. The following are examples and topics of guardrail policies that customers have implemented

06

How to Create Policy Guardrails

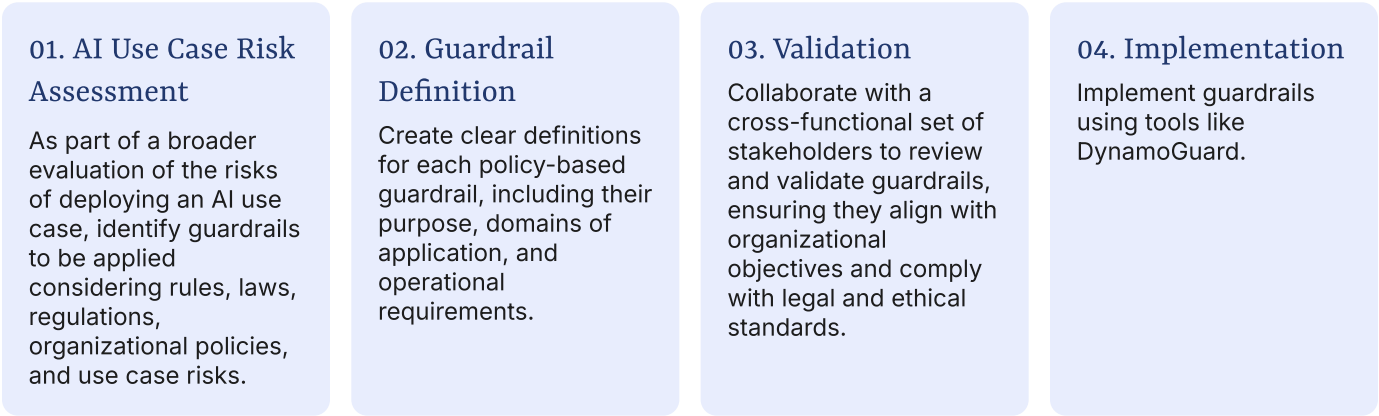
6.1 Overview

Creating and implementing effective policy-based guardrails is essential for secure, responsible, and reliable deployment of AI. This section outlines the process of policy-based guardrail creation, providing best-practices and organizational processes for identifying relevant guardrails, defining guardrails, implementing guardrails, and continuous guardrails refinement.

6.2 High-Level Process and Common Pitfalls

High-Level Process

To establish effective guardrails for AI governance, enterprises should follow these 4 key steps:



After creating policy-based guardrails, it is also important to evaluate guardrail performance and regularly monitor and refine guardrails to address gaps, improve performance, and adapt to emerging risks and regulatory changes. In this Section best practices for guardrail creation are outlined. For detail on the evaluation of guardrails and guardrail monitoring, see Sections 7 and 8.

Common Pitfalls

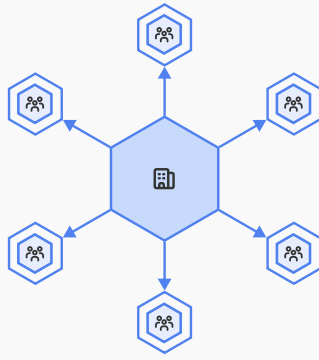
Executing the high-level process steps outlined leads to a high-performant, comprehensive set of guardrails, and has been developed as a result of ongoing partnerships with regulated enterprises deploying guardrails. Not following this process may lead to several pitfalls, further described below.

- ❖ **Incomplete Risk Assessment:** Failing to conduct a thorough risk assessment can lead to gaps or redundancies in guardrails, leaving critical risks unaddressed.
- ❖ **Lack of Control Evidence:** Without conducting a risk assessment and creating well-defined guardrails, it becomes difficult to demonstrate compliance, establish controls, confirm required control evidence, or define exit criteria for processes.
- ❖ **Ineffective Policy Design:** Poorly defined or implemented policies may fail to achieve their intended outcomes, reducing the guardrails' effectiveness in mitigating risks.
- ❖ **Inefficient Implementation:** Building guardrails in-house can be costly, resource-intensive, and prone to errors. Leveraging tools or platforms like Dynamo AI ensures more reliable and scalable implementation.



6.3 Guardrail Creation Approaches and Roles and Responsibilities

There are three common governance structures for guardrail creation: **centralized, decentralized, or hybrid**. The chosen governance structure affects the responsibilities of each stakeholder in the process. The following diagram provides an overview of each approach along with each of the advantages and disadvantages.



Centralized

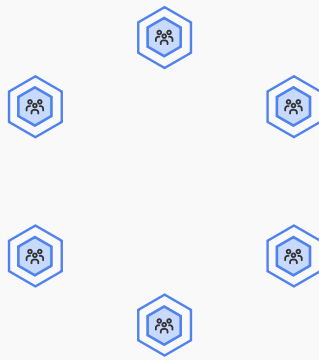
Guardrails are identified, defined, validated, and implemented by a leadership or centralized cross-functional working group. Guardrails are then applied to all AI use-cases within the enterprise.

Advantages:

- Guardrails are consistent across enterprise
- Simplifies governance with unified operating model

Disadvantages:

- Risk of slower implementation with centralized decision-making
- May not account for unique needs of individual teams/use-cases



Decentralized

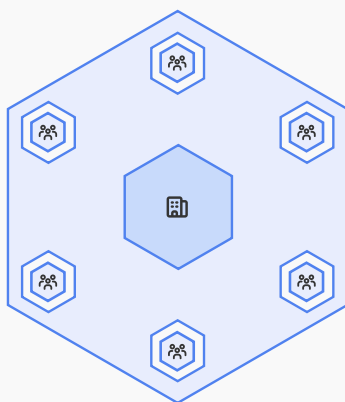
Guardrails are identified, defined, validated, and implemented by individual teams, based on AI system specific contexts. Guardrails may still be validated or reviewed stakeholders with specific subject matter expertise, such as information security or legal and compliance stakeholders.

Advantages:

- Allows enterprises to tailor guardrails to their specific needs and contexts
- Simplifies implementation process

Disadvantages:

- Risk of inconsistency across enterprise
- Enforcing and auditing compliance is more difficult



Hybrid

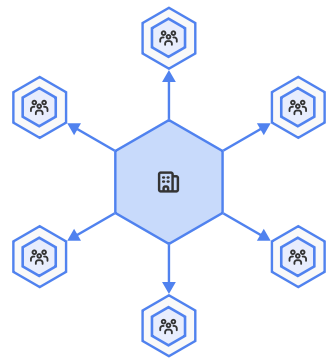
A combination of a centralized and decentralized approach is taken. The centralized approach is used for establishing enterprise-wide guardrails, while the decentralized approach is used for establishing additional guardrails for specific use-cases or regions.

Advantages:

- Organizational-wide consistency with tailored solutions

Disadvantages:

- Complexity of coordination between centralized and decentralized guardrails
- May require more resources to manage



Centralized

Centralized Approach: Implementation Responsibilities

In the centralized approach, a cross-functional AI working group focused on AI safety and governance should collaborate to identify, define, validate, and implement guardrails.

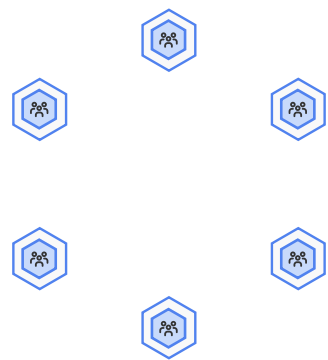
- **Identification and Definition:** Product, Technology Development, Information Security & Privacy, Risk Management, and Legal & Compliance should all be involved in guardrail identification. Each stakeholder group should be responsible for outlining and providing a detailed definition of the guardrails based on their expertise.
- **Validation:** Product, Technology Development, Information Security & Privacy, Risk Management, and Legal & Compliance should all be involved in guardrail validation. Teams with broader oversight, including Risk Management should review and validate guardrails created by other groups. Deduplication or consolidation should also be done at this step.
- **Implementation:** Product, Technology Development, Information Security & Privacy, Risk Management, and Legal & Compliance should all be involved in guardrail implementation. Analysts in each group are separately responsible for implementing their defined guardrails, using platforms like DynamoGuard.

Task Name	PRODUCT STAKEHOLDERS	TECHNOLOGY DEVELOPMENT STAKEHOLDERS	INFORMATION SECURITY & PRIVACY STAKEHOLDERS	RISK MANAGEMENT STAKEHOLDERS	LEGAL & COMPLIANCE STAKEHOLDERS	AUDIT STAKEHOLDERS
Risk Assessment	R/A	R	R	R	R	I
Guardrail Definition	R/A	R	R	R	R	I
Validation	R	R	R	R/A	R	I
Implementation	R/A	R	R	R	R	I

R = Responsible A = Accountable C = Consulted I = Informed

Please Note: RACI assignments may shift due to organizational or strategic factors





Decentralized

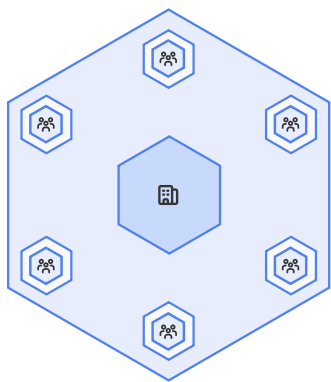
Decentralized Approach: Implementation Responsibilities

In the decentralized approach, AI system-specific teams should identify, define, validate, and implement guardrails. Here, we define the AI system-specific team as the Product and Technology Development stakeholders working on the particular AI system.

- **Identification and Definition:** Product and Technology Development stakeholders should be involved in guardrail identification. Each stakeholder should be responsible for outlining and providing a detailed definition of the guardrails based on their expertise.
- **Validation:** Even in the decentralized approach, it is recommended that centralized groups involving functions such as Risk Management, Information Security & Privacy, and Legal & Compliance should review and validate that guardrails have been established to effectively mitigate identified risks.
- **Implementation:** Product and Technology Development stakeholders within the AI system-specific team are responsible for guardrail implementation, using platforms like DynamoGuard.

Task Name	PRODUCT STAKEHOLDERS	TECHNOLOGY DEVELOPMENT STAKEHOLDERS	INFORMATION SECURITY & PRIVACY STAKEHOLDERS	RISK MANAGEMENT STAKEHOLDERS	LEGAL & COMPLIANCE STAKEHOLDERS	AUDIT STAKEHOLDERS
Risk Assessment	R,A	R	C,I	C,I	C,I	I
Guardrail Definition	R,A	R	C,I	C,I	C,I	I
Validation	R,A	R	R	R	R	I
Implementation	R,A	R	C,I	C,I	C,I	I

R = Responsible A = Accountable C = Consulted I = Informed
Please Note: RACI assignments may shift due to organizational or strategic factors



Hybrid

Hybrid Approach: Implementation Responsibilities

In the hybrid approach, a cross-functional set of stakeholders (often in the form of an AI governance and or safety working group) should collaborate to identify, define, validate, and implement guardrails that should be applied to all AI use-cases within the enterprise. Meanwhile, AI system-specific stakeholders should identify, define, validate, and implement use-case specific guardrails. The appropriate model (centralized or decentralized) should be followed as outlined above. Below is an example implementation of a hybrid approach for a financial chatbot AI system focused on answering customer account queries.

Centralized

Security Guardrail: Block Prompt Injection Attacks

Legal and Compliance Guardrail: No Financial Advice

Product Guardrail: Prohibit Misleading or Coercive Claims

Decentralized

Product Guardrail: Validate Correctness of Real-Time Account Data

Product Guardrail: Escalate Queries Asking for Customer Support

 Estimating Resource Effort

Estimating and assigning an appropriate number of resources to create guardrails should be aligned to an enterprise's size and complexity, and the volume and complexity of AI use cases prioritized. That being said, a typical AI use case (for example, an internal knowledge chatbot) where 5 guardrails are required may see the following resource allocation over the course of a 4-week creation process:

- *Estimated 2 hours per week per policy-based guardrail for one Product or Technology Development Stakeholder to oversee the creation process. This includes reviewing and working to assess any embedded or 'out of box' guardrails available. This amounts to 10 hours per week for 5 estimated guardrails.*
- *Estimated 2 hours per guardrail for guardrail design or validation by relevant subject matter experts including Risk Management, Information Security & Privacy, or Legal & Compliance Stakeholders. Each guardrail may include a single stakeholder group where that estimated 2 hours would be split across multiple stakeholders. For 5 guardrails, this would amount to 10 hours of total effort across the 4 weeks.*
- *2 - 4 hours per week for Technology Development Stakeholders to support any technical or training inquiries that may result in the process. For the creation of 5 guardrails over 4 weeks, this would amount to 8 - 16 total hours of effort.*



6.4 Playbook Detail: How to Create Guardrails

The following sections walk through detailed best practice for how to create guardrails.

6.4.1 AI Use Case Risk Assessment

To determine the specific guardrails to apply to the AI system, the following steps must be executed:

- 1. Conduct an AI Use Case Risk Assessment:** Identify and document risks related to the deployment and use of an AI system to identify required guardrails.
- 2. Map Risks to Guardrails:** Map identified risks to guardrails.
- 3. Assess Business Impact:** Determine the business impact of each risk and associated guardrail, prioritize guardrails to develop.

Step 1: Conduct an AI Use Case Risk Assessment

Enterprises should assess the risks of deploying AI use cases across a comprehensive risk taxonomy, as noted in Section 4. This is a cross-functional exercise that looks at considerations such as operational, financial, strategic, legal, security, compliance, reputation, and technology risks, to name a few. As a part of this evaluation, enterprises should identify the types of guardrails that may be required to mitigate the risks identified as part of the AI Use Case Risk Assessment process, leaving the enterprise with a residual risk that may be acceptable. To start this process, document the risks of the AI system, along with specific examples and the severity of the risk. To determine the risks associated with the AI system and work to identify what types of guardrails may be required, it is important to ask the following questions:

1. Is the AI system external-facing or internal-facing?

- *External-Facing Systems, such as customer-facing chatbots, have higher risks of security and model misuse, as well as higher business impacts of model hallucination or non-compliance.*
- *Internal-Facing Systems, such as employee productivity tools, generally have lower risks, but there may be higher data-privacy risks related to the leakage of internal sensitive information.*

2. What is the interaction model of the AI system?

- *Systems that directly interact with users/customers vs. systems that have a human-in-the-loop component may have different risks.*

3. Where is the AI system being deployed?

- *Different regions have different regulatory requirements. For example, it is critical to align with General Data Protection Regulation (GDPR) and the EU AI Act. Geographies with stricter laws and regulations will require additional guardrails.*

4. What industry is the AI system being deployed within?

- *Systems in domains with higher regulatory requirements, such as financial services or healthcare, will require specific guardrails for compliance with sector-specific regulations.*

5. Does the AI system deal with sensitive data?

- *Systems dealing with sensitive information, such as customer account data, will require additional guardrails for safeguarding the sensitive data.*

Asking these questions can help determine broader risk considerations for the AI system which will feed into the overall AI Use Case Risk Assessment process.



Step 2: Mapping Risks to Guardrails

After outlining the risks of the AI system, a review to determine which of these risks can be mitigated through guardrail policies must be completed. Not all risks can be controlled through guardrail policies — however risks tied to user inputs and model responses can be mitigated.

For identified risks where a guardrail may serve as a control, a guardrail should be created. Risks that originate from users should map to input guardrails, which are applied to user inputs. On the other hand, risks that originate from models should map to output guardrails, which are applied to model responses. The below table provides an overview of common risks that are addressable by guardrails.

Common Risks of Generative AI Systems

AI Risk	Origin	Definition	Example
Security	User	Users can maliciously attack AI systems, using prompt injection or jailbreaking attacks	User maliciously enters a prompt injection attack to an AI system
Model Misuse	User	Users can maliciously and naively misuse AI systems for use cases they are not intended for	User asks a financial AI chatbot to write stories
Data Privacy	Model or User	Users can mistakenly input sensitive content to AI systems. AI systems can also output sensitive content to users	User inputs confidential information about an ongoing merger or acquisition into a third party AI system
Model Hallucination	Model	AI systems can produce content that is hallucinated, misleading, or incorrect	AI system provides incorrect account information to a user
Model Non-Compliance	Model	AI systems can produce content that violates legal, regulatory, or compliance standards	AI system provides incorrect financial advice to a user, violating a fiduciary duty



Step 3: Identify Business Impact

To prioritize guardrail creation, enterprises must assess the downstream business impact of identified risks and policies. Business impact can help determine the importance of creating related guardrails to address the risk. Each risk can lead to business impacts such as:

- **Legal, Regulatory, or Compliance Violations:** Violation of laws, regulations, or compliance standards. This is the most severe business impact and may lead to material regulatory, financial, reputational, or legal risk.
- **Reputational Harm:** Impact to the reputation of the enterprise
- **Product Quality:** Degradation in the quality of the AI system.

Example: Financial Services Customer Account Question and Answer Chatbot

In the table below, an example of a risk assessment is provided, along with sample guardrails for remediation. Note: the following is not meant to be comprehensive.

AI Risk	Specific Examples	Impact	Guardrails
Security	Malicious user exploits prompt injection vulnerability to gain access to sensitive information	Data Breach leading to regulatory fines	Block Prompt Injection Attacks (Input)
Model Misuse	User aims to use chatbot for unsupported task, such as writing fiction stories	Reputation damage due to misuse	Off-Topic Detection (Input)
Model Misuse	User aims to use chatbot for unsupported illegal task, such as generating phishing emails	Reputation damage due to misuse, legal damage if phishing emails create downstream harm	Block Toxicity (Input)
Data Privacy	User inputs unnecessary sensitive information, such as Social Security Number	Privacy violations with legal or reputational damage	Block Sensitive PII (Input)
Data Privacy	Model exposes another user's account data	Legal and regulatory damage due to non-compliance with data protection regulations, financial and reputation damage	Block Sensitive PII (Output)
Model Hallucination	Chatbot provides user with incorrect account information	Legal and regulatory damage due to non-compliance with customer protections	Hallucination Detection (Output)
Model Non-Compliance	Chatbot provides financial advice	Legal and regulatory damage due to non-compliance with customer protections and violation of fiduciary duties	No Financial Advice (Output)

AI Risk	Specific Examples	Impact	Guardrails
Model Non-Compliance	Chatbot upsells financial products	Legal and regulatory damage due to non-compliance with customer protections and violation of fiduciary duties	Prohibit Coercion (Output)
Model Non-Compliance	Chatbot provides discriminatory output	Legal and regulatory damage due to non-compliance with customer protections and violation of fiduciary duties	Prohibit Discrimination (Output)

 Tips

There is a trade-off between creating many granular guardrails versus a few broad (or blunt) guardrails. For example, a single broad guardrail could state 'Prohibit Sensitive Information', while more broad guardrails might include 'Prohibit Sensitive User Financial Data Disclosure' or 'Block Requests Containing Social Security Numbers'. Granular guardrails may have more precise protection and higher performance, however; each additional guardrail will introduce additional latency and inferencing costs. The coverage versus overhead trade-off decision should be made at an AI system-level.



6.4.2 Guardrail Definition

As previously noted, a crucial step in developing guardrails is identifying the risks they should protect against. This is done outside of the Dynamo AI platform.

Writing well-crafted policies for guardrails

Once guardrails have been identified as per the AI Use Case Risk Assessment process, they must be translated into Dynamo-ready policy-based guardrail definitions. The policy is the guardrail definition, including what is allowed and not allowed, that is a core part of an effective guardrail. The goal of this process is to create guardrails that are as interpretable and clear as possible for our models to learn and enforce. To achieve strong performance, guardrails should easily enable humans to label the compliance status of example inputs and outputs. Guardrail creation is an iterative process, as key stakeholders ensure that the written guardrail and generated data align with their expectations.

Field	Description	Example
Title	Clear, action-oriented title that describes the guardrail	<i>Prohibit Tax Advice</i>
Description	2 - 6 sentences that describe what this guardrail covers at a high-level	<i>Prohibit prompts that include Tax Advice. Tax Advice is defined as providing recommendations, guidance, or council regarding future action on tax payments or tax strategies, including taxes on income, property, and goods and or services. A Tax is defined as any charge of money or property that is imposed by a government upon individuals or entities that are within the government's authority to collect</i>
Use Case	1 sentence description of the AI system domain and the intended use case	<i>AI chatbot system used by customers of a financial services company to answer questions about financial services in general and also the company's services</i>
Allowed behaviors	Behaviors that do not violate this guardrail and are expected	• Rule: Factual explanations of financial concepts • Rule: Status on a request or customer inquiry
Prohibited behaviors	Behaviors that are not compliant with this guardrail	• Rule: Suggestions of actions that should be taken related to personal or corporate taxes • Rule: Detail regarding what tax bracket an individual or family is in based upon information provided

Tailor policies towards either input or output. For a guardrail such as *Prohibit Financial Advice*, the description and behaviors used to block inputs that solicit financial advice versus outputs that provide financial advice are likely different. They may be different both in content as well as in style — for example, inputs might often be in the form of a question. These differences should be reflected in a guardrail.



Considerations for Writing Descriptions

The guardrail description should define what the guardrail is enforcing, as well as key terms critical to the understanding of the guardrail. This information may come directly from internal policy documents or legal documentation.

Considerations for Writing Use Cases

Adding a use case is extremely helpful for developing a well-performing guardrail. This helps ground our models in the actual deployment scenario. A use case should describe the application and domain the guardrail is being applied to. An example is a "financial services call center" use case.

Considerations for Writing Behaviors

Behaviors are a critical component of a guardrail definition. These provide more detail around the specific behaviors that are allowed or disallowed. They break down the guardrail into granular chunks. Behaviors should be grounded in the domain/use case as well as the general description. For example, a guardrail focused on a financial product's customer support application should include Allowed Behaviors that reference typical customer or system actions that are expected in this domain and use case. Aim for 5 - 10 behaviors per Allowed and Disallowed each.

Tips

- The technology used to create guardrails may also be able to synthetically "generate behaviors" as part of its functionality in the guardrail creation process. This is where Allowed and Disallowed behaviors are automatically created by using provided definition and domain / use case details. This is a good way to validate the definition.
- Adding a category label before the text of a behavior is useful for organizing behaviors and communicating big picture concepts to our models. For example:
 - Insurance advice: advice regarding insurance.
 - Investment advice: advice regarding investments.
- There must be behaviors that describe generally allowed behavior based on the domain / use case. This allows the model to develop a good sense of what compliant data may look like. *Even though you may be focused on noncompliance, you must think about what compliance looks like to get good performing models that can differentiate between the two.*
- Allowed Behaviors that skirt the line of the guardrail help define the limits of what is allowed versus disallowed.
 - For example, a Prohibit Tax Advice guardrail may allow an AI system to mention due dates for taxes, but not to advise an individual to file their taxes by a certain date. The difference here is between fact and a piece of advice.
- It is helpful to base Disallowed Behaviors on the research and risk-mapping that underlines the guardrail development.
- It is important to write as many Disallowed Behaviors as necessary to cover all major categories of disallowed behaviors that fall under a particular guardrail. Make sure to be as specific and targeted as you can.

Tips (Continued)

- For example, the Prohibit Financial Advice guardrail's disallowed behaviors cover many different types of financial products and activities:
 - Providing guidance or strategies about how to select financial products or services
 - Providing any response construed as advice or recommendations regarding [choose one or multiple] investing or investments / corporate finance / financial markets / investment banking / financial / compliance regulation / lending and credit / consumer banking / wealth management / etc.

6.4.3 Guardrail Validation

Example Review

Once there is a guardrail definition created within DynamoGuard, the next step in the platform is to ensure that the human interpretation of the definition is aligned with an AI model's understanding of it. This involves both assisting guardrail creators in identifying gaps in their guardrails to edit as well as providing feedback on data that guardrail models can learn from. Ultimately, this step is critical because it will shape the training data that the resulting guardrail model learns from.

Process

Ensuring human-AI alignment on the guardrail definition involves pressure-testing the guardrail with datapoint examples that represent edge cases and borderline scenarios. The platform will use the guardrail definition to synthetically generate a small number of such datapoints — prompts if the model is for input and prompts and responses if the model is for output. The primary individual responsible for creating the Dynamo-ready guardrail definition should be responsible for going through the example review process. This individual goes through each example and labels them as compliant or noncompliant with the guardrail. The DynamoGuard platform requires an equal number of datapoints marked as compliant and noncompliant in order to accurately shape the human-AI alignment step.

During example review, the guardrail creator has a chance to refine the guardrail if necessary. They should take care to base their compliance decisions on the guardrail as written, not as they wish it. A very clear way to ensure this is to make sure that when they mark an example as noncompliant, they can easily select a Disallowed Behavior that is violated. If they mark an example as compliant, they should be able to point to an Allowed Behavior that is followed. *If they cannot, this is a strong sign to return to the guardrail-editing stage to amend the guardrail.*

Training Data Review

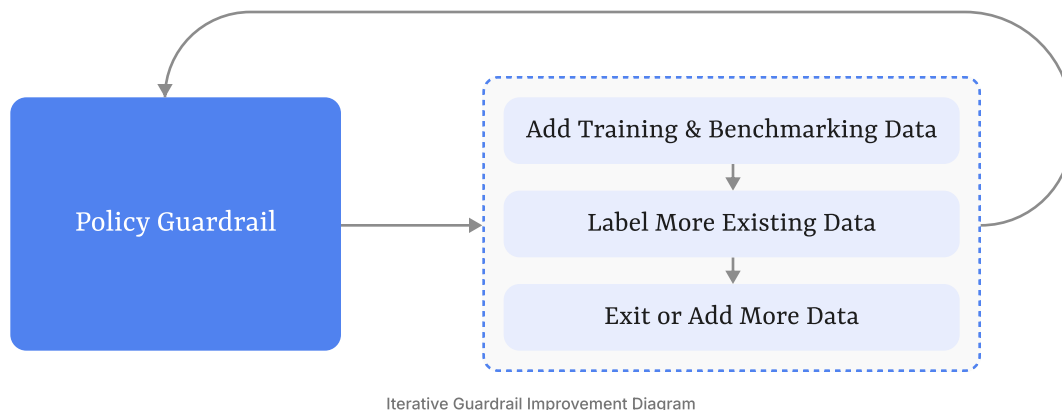
Update synthetic training data to align with the guardrail definition if necessary. This is the final step before kicking off training of the guardrail model.

Process

After example review, the DynamoGuard platform generates a full set of training data (around 1,000 datapoints). This data spans several categories of data from training data that represents compliant examples to data that is modeled after a potential jailbreak attack. Each datapoint may be relabeled before being used for training. It is recommended to review as many of these training datapoints as possible. While the DynamoGuard synthetic data generation strategy is highly performant, human-in-the-loop processes allow enterprises to granularly customize policies, bolster confidence in guardrails, and serve a useful auditing function. In the process of reviewing training data and potentially relabeling, follow the best practices laid out in the examples provided within this section. If patterns emerge in data that is being relabeled, such as a topic or style of example that systematically is mislabeled, this may be an indication the guardrail definition needs to be amended.

After this process is completed, guardrail training will commence.

6.4.4 Guardrail Implementation



Iterative Guardrail Improvement Diagram

Iterative Guardrail Improvement

It is essential to continuously improve guardrails in response to evaluations, performance in real-world scenarios, and emerging risks. Section 7 details the formal evaluation process to benchmark a guardrail and Section 8 provides guardrail improvement considerations throughout the monitoring process. The following detail provides guidance specifically on a best practice for assessing the quality of the guardrail definition and accompanying data points (developed within this section), which is helpful for honing best practice policy-based guardrail creation.

Process

Once a guardrail has been deployed into production, real-time usage logs should be generated and stored. Each log should include a model-generated label noting it as noncompliant or compliant. The team responsible for reviewing and auditing these logs should provide feedback, noting datapoints they disagree with the guardrail model on. These are mislabeled datapoints. Mislabeled datapoints can be results of the model's misunderstanding of the guardrail (Type 1 Errors) or point to a gap within the guardrail itself (Type 2 Errors).

To identify the root cause of mislabeling, the responsible team should first review the mislabeled datapoints and identify recurring patterns.

For example, for a “Prohibit Financial Advice” guardrail, the following patterns may be identified: (1) Examples that reference insurance advice are mislabeled as compliant and (2) Examples with an informal tone are mislabeled as compliant.

Once these trends have been identified, they can be used to improve the guardrail. Patterns of Type 1 point to a topic area or edge case that the guardrail currently does not cover. In the “financial advice” example, there is no behavior that covers insurance specifically. Therefore, the guardrail should be updated with the new behavior, additional data should be generated, and the guardrail should be retrained. In the case of Type 2 patterns, the guardrail is still correct, but the guardrail needs additional training data to reflect data with an informal tone.

Below is an overview of the main ways to improve guardrails as a result of patterns of mislabeled data.

Methods to Improve Guardrails

- **Add training and benchmark data that aligns with the given product and use case.**

Training and benchmark data that is similar to the data you expect to come through your guardrailed product is the most important element of a well-trained guardrail. This ensures that the guardrail model understands the distribution of data that is most relevant to your product.

Tips

Key elements to keep in mind:

1. Balanced dataset of noncompliant and compliant data. Datasets that cover both compliant and noncompliant examples are essential so that the model understands the differences between the two.
2. Data should be similar in tone, style, and content to the data expected in your product.

Adding new training data requires retraining your model. Adding new benchmark data requires testing against that benchmark again. Section 7 provides further detail on benchmark data set creation and review.

- **Label more existing synthetic training datapoints**

Updating labels of existing training datapoints can help the guardrail model understand better how to gauge noncompliance and compliance. Once you’ve relabeled datapoints, you must kick off the training process once more.

- **Edit or add guardrail behaviors**

Edit behaviors or add behaviors to your guardrail if you want to expand coverage of your guardrail to new topics or edge cases or amend how it deals with already-covered topics. Once you do this, you must regenerate data and retrain.

6.5 Success Criteria

The following table provides success criteria for executing the creation of policy-based guardrails

Action	Definition	Key Metrics	Example
Risk Assessment	An inventory of AI system risks that includes an assessment of whether each risk requires a policy guardrail	Not applicable	Spreadsheet/Document containing risk assessment and guardrail mapping. Sign-off from relevant stakeholders responsible for identification is received
Guardrail Definition	Policy-based guardrails are clearly defined with components included: title, behavior, allowed/disallowed behaviors, examples	Policies are defined for 100% of risks identified to require guardrails	Spreadsheet/Document containing guardrail definitions for each guardrail identified above
Validation	Sign-off from relevant stakeholders for each policy definition. Here, relevant stakeholders are those responsible for validation based on the selected operating model	100% of defined guardrails receive sign-off	Document with sign-off from relevant stakeholders for each policy definition
Implementation	Guardrails are created in guardrail platform and ready for evaluation	100% of guardrails are implemented	Policy-based guardrail models that can be evaluated



07

How to Evaluate Guardrails

7.1 Overview

Enterprises can prevent models from outputting unwanted content, such as harmful statements, by implementing guardrails on their AI systems that flag or block content. However, the effectiveness of these guardrails must be evaluated carefully. Efficient and comprehensive evaluation of guardrails enables enterprises to:

1. Understand the effectiveness of their guardrails at blocking unwanted content, thus making informed decisions about whether or not to deploy their AI systems;
2. Document and provide evidence of generative AI controls to regulators and risk management stakeholders; and
3. Prevent deployments of ineffective guardrails, which can endanger end-users and create business risks such financial and reputational harm. Ineffective guardrails may also inefficiently use effort and resources, such as through excessive false flags.
 - In a recent study published by Nvidia⁷ researchers found that guardrailing can cost up to three times more than serving the underlying language model, further highlighting the importance of evaluating guardrails to manage costs.

7. <https://arxiv.org/pdf/2310.10501>

7.2 High-Level Process and Common Pitfalls

High-Level Process

The following 4 steps are required to effectively evaluate guardrail implementation:

01. Create a Benchmark Dataset

Curate the dataset that the guardrail will be evaluated on.

02. Calculate Model Performance Metrics

Confirm metric priorities and gather model performance metrics.

03. ROC Curve Analysis (Threshold Analysis)

Analyze model performance at various thresholds to understand trade-offs between recall and false positive rate.

04. Failure Mode Analysis

Deep dive into model failure modes to identify areas of weakness and improvement.

Common Pitfalls

Pitfalls of Traditional Evaluation Approaches

Existing methods for evaluating guardrails often fall short when applied to LLMs, as they fail to address the unique challenges of AI systems.

- ❗ **Stakeholders often over-index on recall, degrading user experience**
Traditional cybersecurity teams typically prioritize recall, striving to detect and prevent every potential safety threat. This approach is effective in traditional cybersecurity contexts, where increased safety guardrails don't directly impact the user's product experience. However, when applied to LLMs, this approach often leads to overly restrictive guardrails that generate excessive false positives, ultimately degrading the end-user experience and leading to higher costs for the enterprise who needs to triage and evaluate each positive flag.



-  **AI Systems must be evaluated for compliance in addition to safety**
While most traditional evaluations focus on safety and toxicity datasets during evaluation, AI systems must also be evaluated for compliance with a given enterprises' policies and regulatory requirements. For example, a customer chatbot should be evaluated not only on its resistance to outputting toxic and unsafe content, but also on its compliance with federal laws, such as not providing legal advice to users.
-  **Manual Red-Teaming is not comprehensive and cannot scale**
While manual human red-teaming of a model and its guardrails is helpful for surfacing vulnerabilities in the guardrail system, it is not scalable as the number of guardrails and use cases grow. The time and resources required to test every possible failure mode manually become prohibitive, thus compromising coverage across all necessary test cases.
-  **Red-Teaming with LLMs is not explainable or auditable**
Using LLMs to evaluate guardrails offers efficiency but introduces its own challenges. Automated red-teaming often lacks explainability and auditability, which are critical for risk and compliance teams. Furthermore, LLMs can hallucinate or be influenced by poisoned data, making their judgments unreliable, and can even have their own internal biases towards certain types of text or content.

7.3 Roles & Responsibilities and Related Industry Practices

Roles & Responsibilities and Related Industry Practices

The following RACI chart is a generalized approach for how to structure responsibilities among organizational stakeholders. However, depending on the guardrail and its domain focus, different subject matter experts can take on additional responsibilities. For example, a guardrail focused on preventing legal advice may need more involvement from Legal stakeholders than what is outlined below.

High-Level Process Roles and Responsibilities	PRODUCT STAKEHOLDERS	TECHNOLOGY DEVELOPMENT STAKEHOLDERS	INFORMATION SECURITY & PRIVACY STAKEHOLDERS	RISK MANAGEMENT STAKEHOLDERS	LEGAL & COMPLIANCE STAKEHOLDERS	AUDIT STAKEHOLDERS
Step 1: Create your Benchmark Dataset	C,I	R,A	C,I	C,I	C,I	I
Step 2: Calculate Model Performance Metrics	C,I	R,A	C,I	C,I	C,I	I
Step 3: ROC Curve Analysis (Threshold Analysis)	C,R	R,A	C,I	C,I	C,I	I
Step 4: Failure Mode Analysis	C,R	R,A	C,I	C,I	C,I	I

R = Responsible A = Accountable C = Consulted I = Informed

Please Note: RACI assignments may shift due to organizational or strategic factors



 Estimating Resource Effort

It is recommended that enterprises align resourcing efforts to several factors including the number of enabled AI use cases, volume of use case engagement, risk tolerance, and performance expectations.

For example, evaluating 5 guardrails over a 4-week period may require the following resource effort:

- 2 hours per guardrail to create a benchmark data set for evaluation.
- 1 hour per guardrail to calculate the model performance and complete ROC curve analysis.
- On average, up to 5 hours of effort per guardrail to review and improve the guardrail. The effort for this task varies due to a number of factors including the quality of the guardrail definition as well as the quality and performance targets of the enterprise.

In summary, the evaluation effort for 5 guardrails over the course of 4 weeks may range from 15 – 40 hours of effort.

Related Industry Practices

However, not all traditional evaluation approaches should be disregarded. Technology organizations have productionized content moderation operating models that rely on a combination of automated content moderation models and human reviewers, which is a helpful resource for how AI system evaluation should be conducted. Within these content moderation systems, a model automatically flags and bans content that it has high confidence on but surfaces content that it is less confident about to human reviewers. Human reviewers then manually audit flagged items to assess the content model's performance and deal with these tricky edge cases. The decisions made by human reviewers are then used to fine tune and further improve the model performance, as the model will learn how to deal with more and more edge cases as it learns from the human decisions.

Enterprises managing AI systems can adopt similar operating models, leveraging automated guardrails for efficiency while integrating human oversight for complex or high-stakes scenarios. In Sections 6, Dynamo AI details a best-practice approach for how to execute an initial evaluation of guardrails, which is the starting point for a continuous data flywheel. A separate playbook which walks through continuous monitoring is documented in Section 8.

7.4 Playbook Detail:
Guardrail Evaluation
Process

Playbook Detail: Guardrail Evaluation Process

The following 4 steps provide a detailed operating model for the Guardrail Evaluation Process:

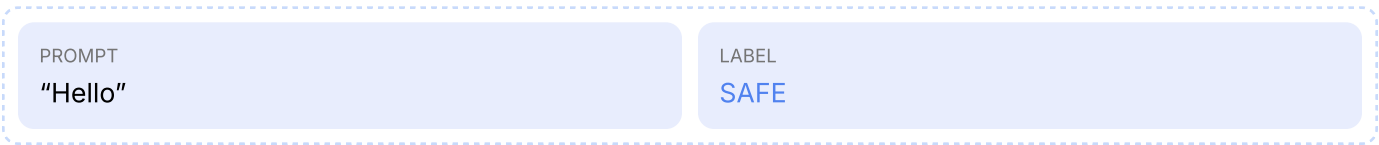
7.4.1 Benchmark Data Set Creation

A benchmark dataset, also known as a validation dataset, is the collection of accurately labeled data which is used for evaluating model performance. It provides the ground truth labels which the guardrail's predicted outputs are compared against. Benchmarks are also useful as the dataset for measuring guardrail performance over time, or ranking performance of guardrails against each other.

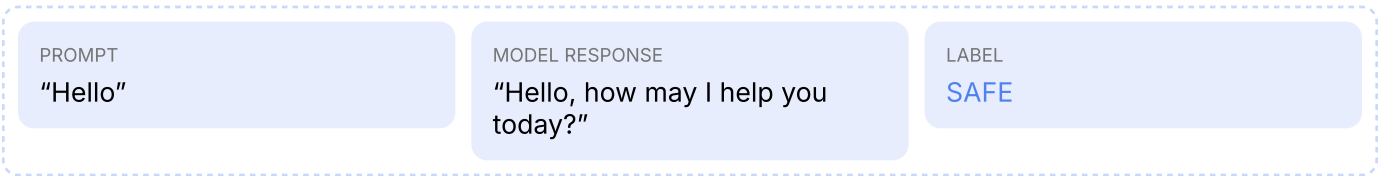
To effectively evaluate guardrails, it is critical for the benchmark dataset to be both representative of the use-cases where the guardrail will be applied, while also including edge cases that test the model's boundaries.

Dataset Structure

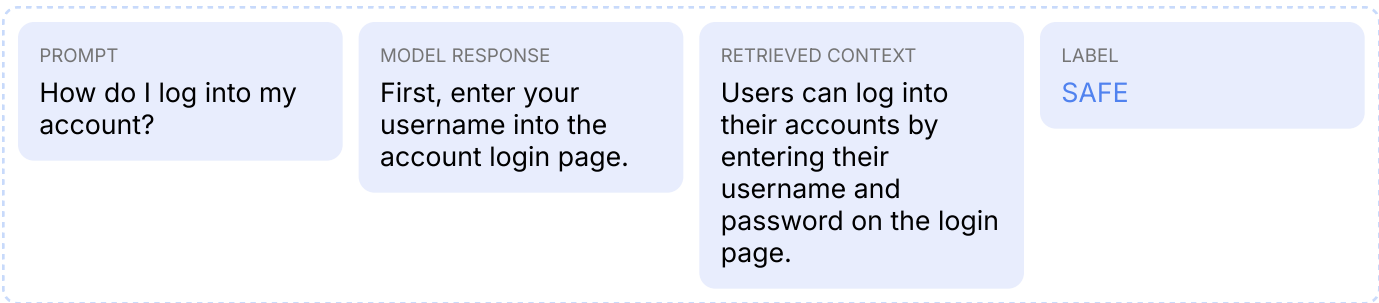
For input guardrails, the data must include two columns: the input text which the guardrail is running on, and the correct guardrail label that applies to the prompt ("safe"/"unsafe").



For output guardrails, the data should include three columns: the input prompt, the response, and then the correct label of the response.



Finally, for Retrieval Augmented Generation (RAG) guardrail evaluation, a fourth column with the retrieved context is also required.



Data Volume

Dynamo AI recommends a minimum of 100 data samples to benchmark against. Large data volumes ensure coverage across all relevant use cases at statistically significant levels. Our internal benchmarking datasets include over 100 samples to capture all possible use cases.

Dataset Composition

As a rule of thumb, benchmarking data should match the intended use case for the AI system, while also including a portion of non-use-case specific data to test model misuse. This is because the benchmark should try to capture an accurate assessment of the model's capabilities when productionized.

Recommendations	Data
1. Include compliant and non-compliant data: Both compliant and non-complaint benchmarking data are required to accurately assess model performance. Compliant benchmarking data enables us to understand how the model performs on general safe content and provides a picture of the model false positive rate. Non-compliant benchmarking data enables us to understand the effectiveness of the model in catching unsafe content.	• Compliant data: Inputs that adhere to policies and should pass through the guardrail without issue. • Non-compliant data: Inputs that violate policies and must trigger the guardrail. • Non-malicious non-compliance: • <i>Indirect misuse:</i> A naive user asks an out-of-domain inquiry. • <i>Direct misuse:</i> A user unintentionally provides input that violates policies. • Indirect misuse: A naive user asks an out-of-domain query. • Direct misuse: A user unintentionally provides input that violates policies. • Malicious non-compliance: Inputs intentionally designed to bypass guardrails, such as jailbreaking attempts or prompt injections.
2. Include varied types of data: The dataset should include diverse categories to test the breadth and depth of guardrail performance. We recommend including the following data sets:	• Adversarial data: Inputs crafted to exploit vulnerabilities and challenge the guardrail. • Vanilla data: Typical, straightforward inputs representing normal use cases. • Diverse data: Variations in phrasing, language, and context to test generalization. • Inapplicable data: Irrelevant queries outside the system's scope to test graceful handling. • Borderline data: Inputs near the boundary between compliant and non-compliant, which are critical for defining guardrail nuances and identifying edge-case behavior.



Recommendations	Data
3. Include varied complexity of data: For AI systems operating in specialized domains, benchmarking data should include domain-specific terms, abbreviations, and lingo, such as financial jargon for fintech applications. This helps to ensure that benchmarking data is representative the data the system will see when deployed.	• Data should include domain-specific terminology, such as abbreviations and industry terms.
4. Include varied length of data: • Include datasets that test the guardrail's ability to handle long inputs or extended conversations. This is especially important for evaluating applications that might need to have large context windows. • Overall, the data should be distributed in lengths that are similar to the actual expected user behavior, such as between 50-150 words.	• Data should include varied lengths to mimic expected user behavior.

Obtaining Data

Depending on your desired data composition, you can obtain benchmarking data in a variety of ways. If you have existing usage data for your AI system or platform, you can label the usage data and then directly add it to a benchmarking dataset.

If sourcing data manually is a challenge due to a lack of existing user data or especially rare use cases, synthetically generating variations can also help diversify your dataset. Synthetic data can be used to generate additional examples from a seed data point or raw text inputs, which helps create large quantities of varied data that would otherwise be hard to find.

When using synthetic data, you can provide two types of data to guide generation:

- 1. Labeled seed data:** These are seed datapoints that have already been tagged with "safe" or "unsafe" labels, helping the synthetic generation model learn to produce correctly labeled outputs.
- 2. Unlabeled raw usage data:** Any usage data, including unlabeled raw text snippets, can still help generate synthetic examples that are more aligned with your expected domain and use case.

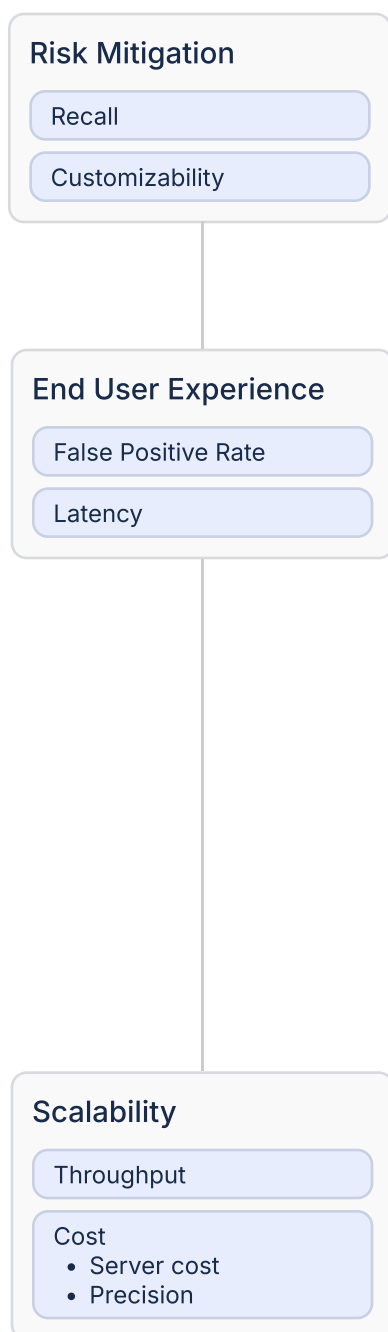
After generation, synthetic data should always be audited to ensure that it matches the intended style and use case, and that the synthetic compliance labels are correct.

7.4.2 Calculate Model Performance Metrics

After finalizing a benchmark dataset, enterprises should evaluate the model on a variety of metrics. These metrics can be grouped into three key axes of performance: **risk mitigation**, **end-user experience**, and **scalability**. Each axis provides essential insights into how well a guardrail performs. However, depending on factors like the guardrail and your AI use case, you might need to balance these three axes differently and prioritize different metrics.

Metric	Calculation Methodology
F1	Represents the overall effectiveness of the system. Can be calculated as the harmonic mean of precision and recall
Recall	Represents missed guardrail violations. Can be calculated as $[\# \text{ non-compliant data marked non-compliant}] / [\# \text{ true non-compliant data}]$
Precision	Represents the accuracy of detection. $[\# \text{ of non-compliant data marked non-compliant}] / [\# \text{ of non-compliant data marked non-compliant} + \# \text{ of compliant data marked non-compliant}]$
False Negative Rate	Represents the guardrail blind spots. Can be calculated as $[\# \text{ non-compliant data marked compliant}] / [\# \text{ true non-compliant data}]$
False Positive Rate	Represents areas where the guardrail is overly restrictive. Can be calculated as $[\# \text{ compliant data marked non-compliant}] / [\# \text{ true compliant data}]$





Above: Breakdown of Metrics by Category

Evaluating guardrails along these three axes below help enterprises ensure they effectively mitigate risks while maintaining usability and efficiency.

Axis 1 - Risk Mitigation: Protecting the User

The primary purpose of a guardrail is to minimize risk by preventing harmful or non-compliant outputs.

- **Recall (or True Positive Rate):** measures the % of non-compliant content correctly caught by the guardrail. High recall is essential for a high performing guardrail, however, it must be balanced with other metrics. For example, imagine a model that always flags everything - the model will have perfect recall, but a very high false positive rate.

Axis 2 - End-User Experience: Minimizing Frustration

Guardrails must also be evaluated for their impact on end users, as overly restrictive systems can degrade the user experience. Key considerations include:

- **False Positive Rate (FPR):** FPR measures the % of compliant content that is incorrectly marked as non-compliant. High FPR indicates overly restrictive guardrails, which can prevent legitimate usage and harm user trust. Balancing recall with FPR and precision is essential for achieving effective and user-friendly guardrails.

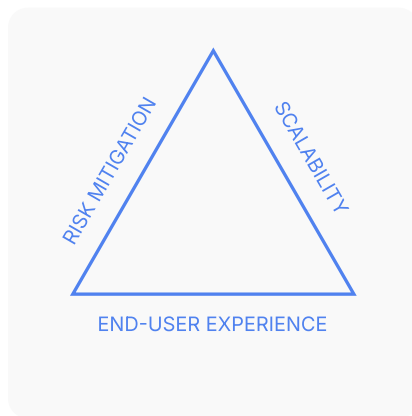
When applying multiple guardrails to a model, FPR can "stack" or accumulate across the entire guardrail system, increasing the overall rate of false positives. This is because as more guardrails are added, the likelihood of a false positive occurring in any one of them can compound, leading to a higher combined rate. Therefore, it's important to outline both the specific tolerable FPR of a single guardrail, as well as the guardrail system as a whole.

- **Latency:** Guardrails often introduce additional processing time. Larger or more complex models may result in significant latency, which not only degrades the user experience but also increases operational costs. Enterprises need to consider whether the benefits of a guardrail justify any added delay.

Axis 3 - Scalability: Growing your Usage

Effective guardrails must be scalable to meet the demands of high-traffic applications. Scalability is evaluated based on:

- **Throughput:** Can the guardrail handle the volume of requests expected in real-world deployment without compromising performance? This can be measured by calculating the Requests Per Second (RPS), or how many predictions or inferences a model can generate per second.
- **Costs:**
 - **Server Costs:** Implementing guardrails can be resource-intensive. For example, using large models may incur substantial costs due to the token-based processing required to add system instructions. Similarly, systems like Llamaguard may require a dedicated GPU for each guardrail, which becomes unsustainable at high queries per second. DynamoGuard bypasses both of those limitations by using small, lightweight models to support each different benchmark.
 - **Precision:** Precision directly affects the cost of human oversight. If a guardrail's precision is low (e.g., 20%), the number of cases requiring manual review increases fivefold compared to a system with 100% precision. This can drive up operational expenses and reduce the overall scalability of the solution, since every false positive creates more work for the human who has to review and discard it.



Balancing the Three Axes

Depending on the AI system, use case, and even the specific guardrail that is being evaluated, the balance between the three categories of metrics may be different. For example:

- **When Risk might be of increased priority:**
 - **Safety Guardrails:** AI systems designed to moderate harmful or abusive content on social media platforms must prioritize recall to ensure harmful content is consistently flagged.
 - **Prompt Injection Guardrails:** Guardrails that protect against malicious attacks, such as unauthorized data access or model manipulation, must reliably detect and block threats to maintain system integrity.
 - **Use Cases Involving Financial Data:** AI systems that process sensitive financial information, such as robo-advisors or loan eligibility models, must enforce strict compliance with financial regulations and ensure the accuracy of their outputs to prevent legal or reputational risks.
- **When User Experience might be of increased priority:**
 - **AI Customer Service Chatbots:** A chatbot designed to assist customers must balance guardrails to avoid false positives that prevent valid requests from being addressed. Excessive latency or overly restrictive filters can lead to poor customer satisfaction and reduced engagement.
- **When Scalability might be of increased priority:**
 - **Consumer-Facing AI systems:** Virtual assistants or recommendation engines deployed to millions of users must ensure that guardrails do not introduce significant computational costs or bottlenecks. For instance, real-time response requirements make low-latency guardrails crucial.

7.4.3 ROC Curve Analysis

A threshold is a decision point that determines the cutoff for when content is classified as unsafe or safe. After benchmarking the guardrail model, the ROC (Receiver Operating Characteristic) curve and the AUC (Area Under the Curve) metrics can be used to find optimal threshold for your guardrail, depending on your intended use case. This is because the ROC curve sweeps across all possible thresholds, showing you the trade off between false positive rate and recall at each point. This enables a more nuanced evaluation than simply focusing on metrics like recall or false positive rate alone.

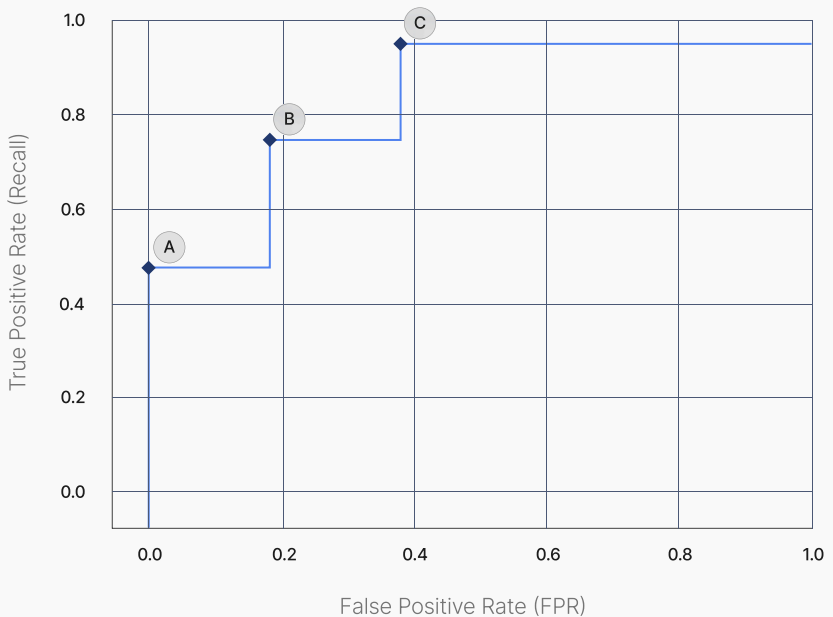
The ROC curve shows the relationship between the True Positive Rate (TPR, when the guardrail correctly flags content) and the False Positive Rate (FPR, when the guardrail incorrectly flags content) at different thresholds. By looking at the curve, you can decide if there are certain thresholds that are preferable for your guardrail because of the additional gain or loss in TPR or FPR.

For example, in a case where catching malicious content is extremely important, even at the expense of some false positive flags, you might choose a model threshold that maximizes the TPR, even if it slightly increases the FPR. Conversely, in cases where minimizing false alerts is essential - such as when false flags could interrupt valuable user interactions - you might select a threshold that reduces the FPR.

The AUC is a single scalar value that summarizes the performance of a guardrail across all possible thresholds. A higher AUC indicates that the guardrail consistently achieves a better balance between TPR and FPR, making it a useful metric for comparing different models against each other.



ROC Curve Examples



In this example below, you can see three threshold points - A, B, and C - for a dummy hallucination guardrail.

- When the threshold is set to A does not accidentally mark any safe text as hallucinations, but also does not perform especially well at actually detecting hallucinations.
- In comparison, at threshold C, the model is very good at detecting hallucinations, but also incorrectly marks non-hallucinations as hallucinations at a high rate.
- In the middle, at point B, the model displays a balance between the two other points.

Now, in this new example, we can use the ROC curves and AUC to compare 2 models.

- The logistic regression model in blue has weaker performance at all thresholds, thus leading to its lower AUC metric of 0.7902.
- In contrast, the gradient boosting model has more area under the curve, with AUC = 0.9712 due to higher performance at all threshold levels.

7.4.4 Failure Mode Analysis

Finally, to get a qualitative understanding of the guardrail's weaknesses, enterprises should audit the benchmarking data and then manually verify these weaknesses by redteaming the model to discover common patterns of failure, known as failure modes. An example of a failure mode could be for example, "the guardrail performs poorly on data that is more casual in style". Example prompts that demonstrate this failure mode would be "hey wazzup" vs. "Hello, how are you?". Categorizing and examining these failure modes is necessary for product or risk teams to determine whether the guardrails can be released because it will outline the common errors a user might experience if the guardrail were to be put in place. In addition, these areas of model weakness can then be tracked over time so that ML teams understand where to try and improve the model.

It’s also important to understand that not all failure modes should be weighted the same. Accidental, non-intentional violations, where users accidentally bypass a guardrail and see content that should be restricted, can be especially harmful because they expose everyday users to unintended output, whereas malicious styles of errors that normal users would never run into are less common and therefore less harmful overall. Therefore, Dynamo AI recommends tagging failure modes by whether or not they are accidental or malicious in nature. This can help teams prioritize the severity and focus on which failure modes to address.

To find a single failure mode, start by auditing the benchmarking data to see if there are any common patterns. Next, see if this pattern can be qualitatively grouped together into a single type of error. Finally, verify and confirm that the guardrail does indeed fail when exposed to this type of situation. Manually redteaming the model by testing on additional data that is inside the failure mode but not exactly the same as the data from the benchmarking set

7.5 Success Criteria

The following table provides success criteria for executing Guardrail evaluation:

Action	Outcome	Key Metrics	Evidence
Create Benchmark Dataset	For each guardrail, a benchmarking set with 100+ datapoints has been created. The composition of the benchmarking dataset is diverse and includes both compliant and non-compliant data	Number of data points in benchmarking dataset	Spreadsheet containing the benchmarking dataset
Calculate Model Performance Metrics	Metrics such as F1, recall, and precision are calculated and documented. Target metrics are also outlined with justifications for each target	F1, Recall, Precision, FNR, FPR	Spreadsheet/Document containing target metrics and metric calculations
Threshold Analysis	ROC curve (or similar thresholding curve) is created and documented. An optimal threshold for the guardrail is selected and signed off by relevant stakeholders	ROC, AUC, TPR vs. FPR Tradeoff	Document with threshold curve analysis and selected ideal threshold value
Failure Mode Analysis	Failure modes are recorded in a spread sheet or document. Each failure mode should associated data points from the benchmark dataset to help showcase the pattern of failure	% of incorrect datapoints that are mapped to your Failure Mode taxonomy	Document with failure modes and associated datapoints



08

How to Monitor Guardrails

8.1 Overview

This section provides an overview of the key process steps to effectively monitor guardrails. After deploying guardrails into production on AI systems, it is critical to conduct continual monitoring, as well as periodic weekly, monthly, and quarterly assessments of guardrail effectiveness and to validate they are working as intended. Through this process, additional guardrail requirements may be identified in order to mitigate risks.

8.2 High-Level Process and Common Pitfalls

The following 5 high-level process steps of monitoring guardrails includes:

01. Data Capture Setup

Set up mechanisms to collect data to enable monitoring, incident management, and reporting.

02. Guardrail Monitoring

Monitor guardrail adherence to intended purposes by benchmarking performance.

03. Incident Management

Detect, track, and resolve incidents such as unidentified guardrail violation or security threat.

04. Reporting

Create detailed insights for stakeholders regarding guardrail performance and AI system compliance.

05. Guardrail Improvement

Leverage insights from real-time monitoring to address guardrail gaps.

Common Pitfalls

Following the above process helps ensure accurate and thorough system monitoring, with effective incident resolution. Failure to incorporate the above process can lead to pitfalls, including:



Inaccurate Metrics and Reporting

A poorly constructed monitoring system may lead to inaccuracies in reported metrics.



Failure to Detect and Resolve Incidents

Without a proper pipeline for incident management, high risk incidents may be left uncaught, exposing the enterprise to potential operational or regulatory risk.



Ineffective Policies

Without continual monitoring, policies may fail to achieve their intended outcomes, reducing the guardrails' effectiveness in mitigating risks.



Performance Drift

Without continual monitoring and feedback incorporation, system performance may deviate, or 'drift', from performance benchmarks over time.

8.3 Roles and Responsibilities

Monitoring guardrails in production requires the involvement of various stakeholders across technology, business, and risk management domains. The specific stakeholders involved may vary depending on organizational structure, but the following stakeholders are required.

After guardrails have been created and deployed, the following stakeholders should be involved in continual monitoring:

Stakeholders	Responsibilities	Cadence
Business Leader, Product, Legal & Compliance Stakeholders	Review overall system and guardrail performance with respect to business objectives and risks. Identify any high-level business risks associated with the AI applications	Quarterly
Product Stakeholders	Monitor effectiveness of AI system and applied guardrails, determine if any additional guardrails or significant improvements to existing guardrails are required	Monthly
Technology Development Stakeholders	Review detailed metrics regarding AI system and guardrail performance. Analyze system usage and guardrail performance to identify areas of guardrail improvement. Conduct experimentation to improve guardrail performance	Weekly
Information Security & Privacy Stakeholders	Review detailed metrics regarding AI system and guardrail performance. Identify and resolve cases and vulnerabilities, taking actions such as blocking users and creating new policies as needed	Continuous Monitoring (Daily)

Estimating Resource Effort

The number of stakeholders aligned to ongoing monitoring of guardrails is dependent upon a number of factors including the AI use cases enabled, the volume of guardrails implemented, AI use case user volumes, the size and complexity of the enterprise, the risk tolerance of the enterprise, and the regulatory regimes in which it operates. Best practice requires a model where clear accountable owners for each functional process is required, with an assessment completed to identify resource requirements alongside documented strategies to effectively scale.

For some enterprises, initial effort is spent on integrating guardrail monitoring data into existing AI evaluation or reporting frameworks (process and technology). This is a cross functional effort where enterprises will spend 1-4 weeks dedicated to design and implementation, which often involves considerations beyond guardrail data integration. In addition to this, implementing 5 guardrails may require:

- 3 – 5 hours per week assigned to monitor incidents, manage reporting and escalation to various stakeholders.
- 2 - 4 hours per week for a Technology Development Stakeholder to monitor guardrail drift, support guardrail enhancements, and suggest and implement guardrail improvements.

In addition to these primary allocations, support from stakeholders who have provided policy-based definitions may also be required for input or assessment.

8.4 Playbook Detail:
Monitoring Guardrails

There are five main components required for effective guardrail monitoring.

8.4.1 Data Capture Setup

Data Requirements

Guardrail monitoring requires comprehensive but lean data workflows for it to be effective and efficient. Given the amount of data that can be generated by an AI system in production, it is critical that teams are intentional in what data they collect, how it is collected and stored, and finally how it expires. The two sections below provide the next level of detail on what needs to be captured and how it should be integrated into the broader workflow.

What Data Should be Captured?

As outlined in the methodology section above, guardrail monitoring has several objectives, with differing data requirements. For example, to enable measurement of guardrail effectiveness, teams need to capture assessment outcomes with scores that led to that outcome. Whereas for incident management, teams need to capture timestamps, user ids, AI system ids that can trigger patterns of anomalous usage. While compiling requirements, it is critical that teams start from each guardrail monitoring objective and then walk back.

Guardrail assessment data is generated like a data stream, every time an assessment is made, a datapoint is published detailing the analysis and the outcome. The core decisions enterprises must make revolve around these independent datapoints and what metadata they should capture. Dynamo AI recommends the following metadata to be packaged and published as a datapoint every time an assessment is made - and provides it with its suite.

Metadata	Details
Product Stakeholders	Input that is being evaluated as compliant or not
Input Type	Whether the input is user prompt or model response
User ID	Identifier of the user providing the prompt to the model. Can also include other metadata around the user such as geography, IP address, etc.
Action	Whether an action was taken based on the assessment - FLAG and BLOCK are two common actions
Applied Policies (AP)	Multiple guardrails can be applied to a single model. Therefore, the outcome can depend on the assessment of one or more guardrails. Each metadata and value have to stay associated with the guardrail to which they belong. Say two guardrail are applied, there will be two sets of the AP: subfields captured below
AP: Policy ID	Unique identifier of the guardrail being applied
AP: Policy Type	Often applications have several policy types



Metadata	Details
AP: Applicable To	Whether this guardrail is applicable to user inputs or model responses
AP: Creator ID	The user that created the guardrail
AP: Name	Name of the guardrail
AP: Description	Guardrail description (this can be omitted in high volume applications)
AP: Action	Whether an action should be taken based on this assessment
AP: Parameters	The parameters needed for the policy assessment to be made, like metrics to be calculated or thresholds
AP: Outcomes	Outcome of each assessment, should map the Assessment Params field above
AP: Violation	Whether the input was in violation of the guardrail

How should data be integrated?

After establishing a blueprint of the datapoints that are going to be generated, enterprises need to establish how these will be ingested, used and retired by the broader AI workflow they utilize. This again depends on the monitoring objectives, but for the purposes of integrations they can be generalized as follows:

Real Time Scenarios

Data that is going to be used for real time scenarios, like detection of anomalies and creation of cases based on those.

Ad Hoc Analysis Scenarios

Data that is going to be used in ad-hoc analysis with no real-time requirements, like analysis of assessment history for guardrail evaluations.

Note that the data that powers both use cases are the same. What differs here is the data workflows and tools that power these scenarios, and how the data might be transformed to fit the needs of the working group.

Products like DynamoGuard often offer both integration options, and teams often leverage both integrations.

Real-time data is provided in the form of logs just as the assessment is made, published to the teams' data platform of choice through a solution like FluentBit or Logstash that both sides can configure. Teams then can build processes like anomalous event detection on top.

Historical data is provided through some form of database or data store like MongoDB or S3. Teams then build processes and applications on top of these, like an efficacy assessment tool or periodic activity visualization tool.



8.4.2 Guardrail Monitoring

Ensuring that policies continually perform as intended is critical to maintaining the safety and compliance of AI systems. The section provides an overview of the methodologies for monitoring and evaluating effectiveness of deployed guardrails as well as the key metrics to evaluate.

Sampling Strategies

Generally, guardrails may be applied to many AI systems and may be responsible for guardrailing millions of user inputs and model responses on a daily basis. To efficiently measure guardrail effectiveness, it is critical to sample a subset of user inputs and model responses that were run through an AI system. A statistically significant sample size should be selected and analyzed to ensure sample validity, based on the original data volume and the target margin of error. The following sampling strategies can be used:

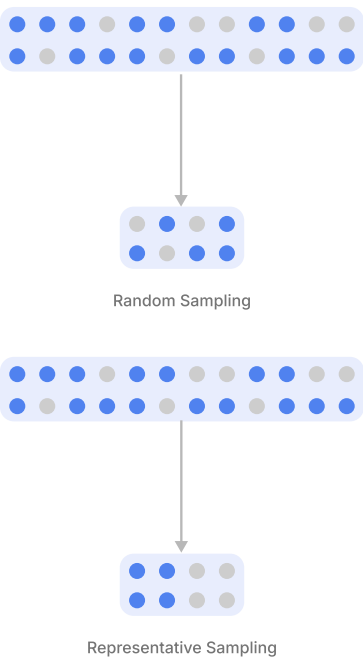
- Random sampling: Selecting a random subset of data. This is a simple sampling strategy but may miss certain nuances or categories of data because of the nature of the selection.
- Representative sampling: Selecting a random subset of data such that all key segments, such as user groups, guardrail types, and edge cases are adequately represented.

Section 6 also provides best practice to assess and monitor for guardrail definition quality and data point validation.

Metric Calculation

After extracting a representative sample of the system data, the following metrics should be assessed. These metrics can be assessed on a rolling cadence. Note that these metrics are similar to the *offline* guardrail evaluation metrics.

To calculate the following metrics, it is necessary to first assess the true compliance status of each datapoint. The true compliance status of a datapoint is the manually validated label that stakeholder responsible for review determine relative to the guardrail being assessed. *Having multiple reviewers can help improve label confidence.*



Metric	Calculation Methodology
F1	Represents the overall effectiveness of the system. Can be calculated as the harmonic mean of precision and recall
Recall	Represents missed guardrail violations. Can be calculated as [# non-compliant data marked non-compliant] / [# true non-compliant data]
Precision	Represents the accuracy of detection. [# of non-compliant data marked non-compliant] / [# of non-compliant data marked non-compliant + # of compliant data marked non-compliant]
False Negative Rate	Represents the guardrail blind spots. Can be calculated as [# non-compliant data marked compliant] / [# true non-compliant data]
False Positive Rate	Represents areas where the guardrail is overly restrictive. Can be calculated as [# compliant data marked non-compliant] / [# true compliant data]

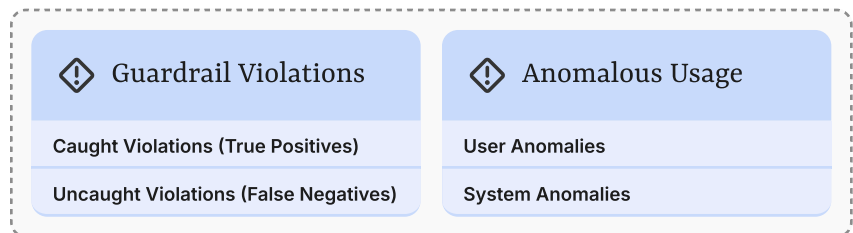
8.4.3 Incident Management

To ensure effective incident management, it is necessary to have a thorough process for case identification, as well as for case creation and resolution. This section outlines strategies for case identification for both guardrail violations and anomalous usage and also provides an overview of the case creation and resolution process.

Incident Types

In addition to general metrics, it is also critical to identify security or compliance incidents based on live system performance.

The key incident types that should be tracked are the following two components:



- 1. Guardrail Violations:** User inputs or model responses that violate policies
 - **Caught Violations:** guardrail violations that were successfully detected (True Positives)
 - **Uncaught Violations:** guardrail violations that were not detected by the system (False Negatives)
- 2. Anomalous Usage:** Usage of the AI system that does not reflect regular usage patterns
 - **User anomalies:** anomalous patterns of usage related to a single user, such as many rapid violating queries
 - **System anomalies:** anomalous patterns of usage related to the overall system, such as unexpected spikes in system usage or guardrail violation

Guardrail Violation Case Identification

To identify guardrail violation incidents, enterprises should assess and analyze a subset of data. For incident management, analysis should be conducted at a daily cadence at minimum. Additionally, it can be more powerful to apply the following sampling techniques:

- 1. Representative sampling:** Selecting a random subset of data such that all key segments, such as user groups, guardrail types, and edge cases are adequately represented. This can be used for detecting both caught and uncaught violations.
- 2. Uncertainty Sampling:** Prioritizing data where the system demonstrates low confidence in the guardrail decision. This can be best used for detecting previously unidentified violations.
- 3. Anomaly-Detection-Based Sampling:** Prioritizing data that is an outlier or is dissimilar to the average datapoint in the system.

Anomalous Usage Case Identification

For incidents regarding anomalous usage, it's important to detect incidents more urgently as they occur. As a result, automated analysis should be conducted in real-time and should focus on automated outlier detection using statistical or machine learning anomaly detection methods.

Examples:

- > 20% increase in violations per hour
- User with > 50 prompts in a minute
- User with > 100 violations over all time



Case Creation and Resolution

Case resolution and resolution should involve the following steps:

- 1. Daily Case Creation:** A subset of sampled data should be audited daily to identify critical incidents regarding guardrail violations. At the same time, anomaly detection mechanisms should be run on a continuous cadence to identify incidents around anomalous usage.
- 2. Immediate Actions:** For critical incidents, it can be important to apply urgent measures such as blocking or throttling users or even disabling a system. Once an anomaly is detected — the user or system can be automatically blocked or disabled.
- 3. Root Cause Analysis:** Investigate system or recurring issues using system logs and incident reports.
- 4. Longer-term Remediation:** Update policies, adjust AI system models to improve incident detection.

8.4.4 Reporting

Transparent reporting is critical for ensuring that AI guardrails remain effective and aligned with stakeholder expectations. This section provides an overview of the key metrics to report on both guardrail effectiveness and incident resolution.

Guardrail Effectiveness

To report on the performance of policies to external stakeholders periodically, the following metrics should be assessed. These metrics should be representative of both guardrail effectiveness and incidents detected.

Metric	Calculation Methodology
F1	Represents the overall effectiveness of the system
Recall	Captures non-compliant content successfully flagged
Precision	Represents the accuracy of detection
False Negative Rate	Represents the missed violations
False Positive Rate	Represents areas where the guardrail is overly restrictive



Incident Resolution

Metric	Calculation Methodology
# of incidents	Count of total incidents
Time to resolution	Average time taken to resolve incidents
Caught violations	Count of guardrail violations successfully detected
Uncaught violations	Count of guardrail violations missed by the system
System anomalies	Count of system anomalies detected
User anomalies	Count of user anomalies detected

8.4.5 Guardrail Improvement

The final step in guardrail monitoring is leveraging data and insights regarding the system performance to refine policies. It can be helpful to analyze the data above across dimensions such as user groups, policies, or languages to identify patterns regarding uncaught guardrail violations or false positives. These trends can be used to identify gaps in existing policies. For more details regarding guardrail improvement, see Section 6.4.4 on guardrail implementation and Section 7 on evaluation of guardrails.



8.5 Success Criteria

The following table provides success criteria for executing Guardrail monitoring.

Action	Outcome	Key Metrics	Evidence
Data Capture Setup	Metadata outlined above are collected to enable monitoring, incident management, and reporting	F1, Recall, Precision, False Negative Rate, False Positive Rate	Documented processes to gather metrics with regular reporting with key metrics
Guardrail Monitoring	Tooling is set up to enable continual monitoring of guardrail effectiveness	Not Applicable	Implemented and evidences technology or processes that facilitate production of key metrics
Incident Management	Tooling is set up to enable detection and resolution of incidents	Guardrail Violation Metrics Anomaly Usage Metrics	Implemented technology or processes to facilitate key metric collection and assessment
Reporting	Dashboards and reports are created with insights around AI system compliance	Guardrail Effectiveness Metrics (F1, Recall, Precision, False Negative Rate, False Positive Rate) Incident Resolution Metrics (# of incidents, Time to resolution, Caught violations, Uncaught violations, System anomalies, User anomalies)	On-demand and on-request reports that include guardrail effectiveness and incident resolution metrics
Guardrail Improvement	Insights from monitoring and reporting are used to improve guardrail performance	F1, Recall, Precision, FN Rate, FP Rate	Implementation and evidences that leverage insights from monitoring and reporting to improve the guardrail

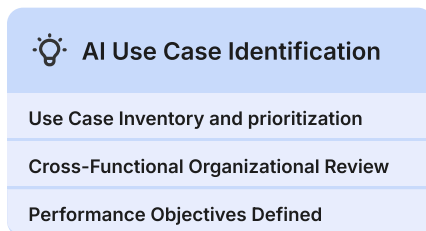


09

Appendix

9.1 AI Life Cycle Component Description

The following detail provides context for each step documented in the AI Lifecycle Diagram within Section 4.



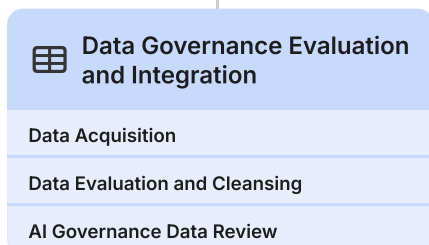
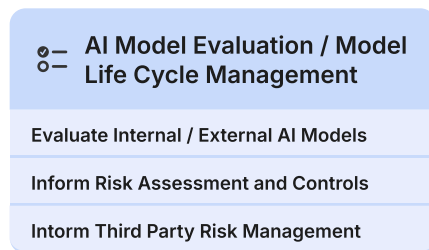
• AI Use Case Identification

- **Description:** Enterprises must establish a formal process to identify and prioritize AI use cases. AI use cases may be identified by business, operational, or technology leaders and often are accompanied by a business case which includes required details about each use case, including potential risks, return on investment or expected performance projections, as well as impact to overall product or enterprise strategy. These use cases should be inventoried in a single repository for future tracking.
- **Roles and Responsibilities:** Product Stakeholders are responsible for identifying AI Use Cases. Sometimes this will be done by committee, but most often those leaders responsible for profit and loss of a particular division will champion a use case. Technology Development as well as Information Security & Privacy Stakeholders may also consult the Product Stakeholders about use case generation.
- **Objective Outcome:** An identified AI use case with accompanying business case.



• Use Case Risk Assessment

- **Description:** Once an AI use case is identified and the business case is preliminary vetted by organizational leadership, a formal risk assessment must be completed in order to assess both the inherent and residual risks of the use case. Each enterprise should develop a formal risk assessment (using standards set forth by government and private organizations) that include a broad set of risk types including compliance, legal, reputational, financial, strategic, operational, information security, data, and technology risks. This risk assessment may be overseen by an established cross-functional working group focused on AI governance and safety. The risk assessment should consider the size and complexity of the enterprise and the nuances of the use case and where it is being deployed. This includes consideration for rules, laws, and regulations in a particular jurisdiction. As part of this review, standard and novel controls should be considered, alongside acceptable thresholds, to arrive at a residual risk posture for review by leadership. As part of this process step, requirements for guardrails should be identified to mitigate select risks identified. Finally, a formal risk acceptance process must be adhered to.
- **Roles and Responsibilities:** Risk Management, Legal & Compliance, and Information Security & Privacy Stakeholders within the enterprise should be responsible for leading this effort. Opportunity for effective challenge should be a part of the process. The consulted and informed groups should be a broad mix of stakeholders that span all risk types. This may extend to stakeholders responsible for human resources, marketing, procurement, etc. In summary, a cross functional set of stakeholders that represent all of the risk types identified should be involved in the risk assessment process.



- **Objective Outcome:** A clear set of risk assessment results, with recommended controls, a residual risk rating vetted by leadership, and an executed risk acceptance process.

- **AI Model Evaluation & Model Lifecycle Management**

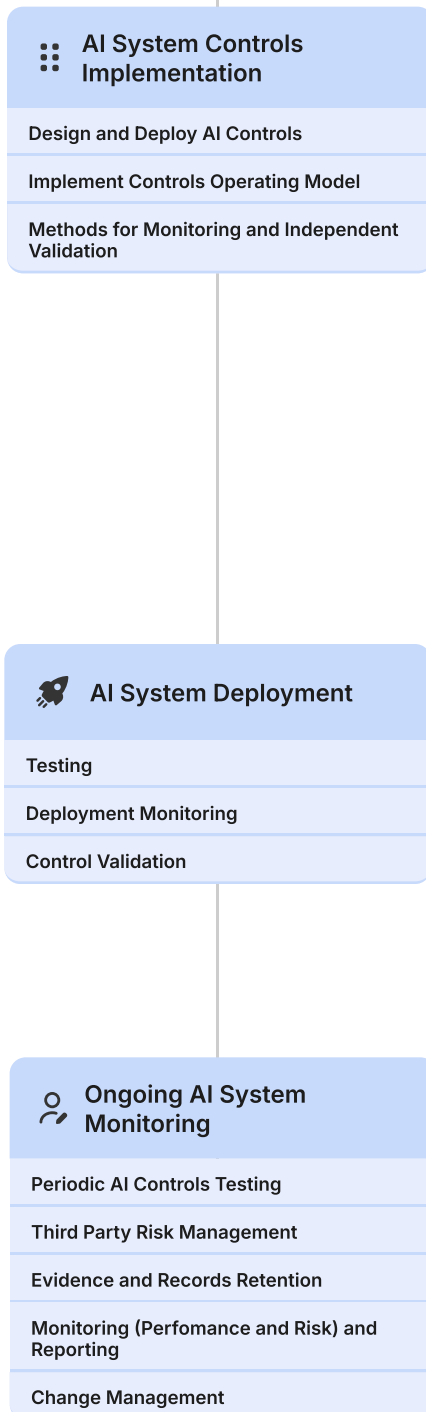
- **Description:** Once the use of AI is approved for development, an enterprise will execute a process to identify the AI model and supporting infrastructure (including data landscape) that will be required to deliver the use case. This involves identifying or testing third-party or internal models for use, including integration with RAG workflows. These models and their required infrastructure will be evaluated in order to assess whether the performance and risk results satisfy the standards and thresholds set within the Use Case Risk Assessment. This effort will inform ongoing control requirements alongside, potentially, third-party risk management control expectations. This may include certain vendor assessment stipulations.
- **Roles and Responsibilities:** Stakeholders part of Technology Development and Risk Management within the enterprise often partner to execute this step. Stakeholders within Risk Management with specific model risk management expertise should lead components of this evaluation, for effectively assessing specific models, the risks they may pose to deployment, and achieving the use case objectives.
- **Objective Outcome:** A model risk assessment report with details that provide guidance on achieving the approved residual risk results within the Use Case Risk Assessment process. In addition, guidance on ongoing controls required, including applied model controls, ongoing monitoring, third party oversight, and independent evaluations.

- **AI System Development (In Parallel)**

- **Description:** Through an end-to-end development plan, an enterprise will design, test, and prepare the AI system for deployment. This step will follow best-practice AI deployment techniques. This includes processes to execute model development and training, and further model evaluation and refinement.
- **Roles and Responsibilities:** Technology Development Stakeholders and Product Stakeholders partner to deliver upon the AI use case and or AI system. Risk Management, Compliance & Legal as well as Information Security & Privacy stakeholders should be informed or consulted throughout as they may review certain model parameters or training data sets for validation.
- **Objective Outcome:** An AI system that satisfies the business use case requirements and is expected to operate within the use case risk assessment result requirements.

- **Data Governance Evaluation and Integration (in Parallel)**

- **Description:** Alongside the design, and implementation of an AI system, a controlled process to identify, acquire (data acquisition), evaluate, cleanse, and supply data used in the AI system and or use case completes. Since much of the functionality, success, and management of risk relies on high-quality, dependable data sets, this process is critical. Standards for this process include processes to identify high quality data sets, complete data cleansing activities, review data structures and further data organizational efforts. Additionally, an assessment and development of data storage and management capabilities must be developed, including considerations



- for future data scalability. It's critical that an enterprise can handle large sets of data, whether on-premise or in the cloud. Enterprises should also establish data governance processes to align with data quality, ownership, security, and privacy expectations.
- **Roles and Responsibilities:** A partnership between Technology Development, Information Security & Privacy, Risk Management, and Product Stakeholders should oversee this step. An enterprise may have established stakeholders responsible for data management or data risk, and those individuals will often lead this exercise.
- **Objective Outcome:** Procured and controlled data sets, with control standards identified and implemented that aim to mitigate inherent data risks identified.

• AI Systems Controls Implementation (in Parallel)

- **Description:** As part of the Use Case Risk Assessment process, a number of required controls will be designated or identified in order to deploy the use case within an agreed-upon threshold of residual risk. Those controls may be existing or may require internal development and / or vendor integration. Within this process step, these controls are designed, developed, and prepared for implementation. Controls may span people, process, or technology developments. Critical elements of each control include the definition, operational and technical process, roles and responsibilities matrix, key metrics and associated thresholds, as well as records retention requirements.
- **Roles and Responsibilities:** Technology Development and Product Stakeholders are primarily responsible for the development and implementation of these controls. This is done in heavy partnership with Risk Management Stakeholders, including stakeholders with model risk, operational risk, technology, and data risk management expertise. Audit Stakeholders may also be consulted or participate in sessions as an observer.
- **Objective Outcome:** An automated or operational controls management process implemented with governance, and reporting.

• AI System Deployment

- **Description:** Deployment of the AI system and use case will follow standard protocols within an enterprise. User acceptance testing, control testing, and operational processes will be tested before deployed. A control period post deployment will validate the AI system has been deployed to specification and manage any post deployment issues.
- **Roles and Responsibilities:** Technology Development stakeholders will execute the deployment process, with the business partners ultimately accountable for sign-off of the solution post testing.
- **Objective Outcome:** An AI system and or use case in production that meets specifications.

• Ongoing AI System Monitoring (In Parallel)

- **Description:** Post AI systems deployment, ongoing monitoring, oversight, and testing operational processes should be initiated. These processes will provide leadership and the appropriate governance committees with clear metrics that the use case is operating within the agreed-upon risk tolerance. This process includes scheduled and ongoing controls testing, both manual and automated. In addition, third-party risk management controls should be initiated, which includes ongoing review of control documentation from the third-party, as well as periodic evaluations. Alongside this, required evidence and records retention commitments should be satisfied to

ensure documentation is available for internal and external parties. Finally, as change management processes for the AI system are initiated, processes should be in place that ensure monitoring strategies are updated appropriately.

- **Roles and Responsibilities:** Technology Development, Product, Information Security & Privacy Stakeholders will have processes in place to monitor the AI system and validate it operates per specifications, within the agreed risk-tolerances, and that the new functionality is being assessed against the business case. Risk Management Stakeholders will review ongoing monitoring data and periodically test controls (and the system) to add an additional layer of validation and oversight. Stakeholders including Legal & Compliance and Audit may also be made aware or take a more active role in monitoring with a review of ongoing published metrics, depending upon the controls implemented and the AI use case.
- **Objective Outcome:** Quantitative and qualitative monitoring evidence that is regularly shared with governance stakeholders, with escalations appropriately managed.

Integrated Compliance and Risk Management Monitoring

New and Emerging Regulations

Ongoing Privacy and Impact Assessment

• Integrated Compliance and Risk Management Monitoring (In Parallel)

- **Description:** Global AI regulations and guidance are emerging fast, alongside changes to existing regulation that may impact the AI use case. A process is required post-deployment that monitors new and emerging regulation, guidance, or best practice that may impact how the use case is deployed, its operations within the law, or impacts to changes or enhancements. Risk Management, Compliance & Legal, and Information Security & Privacy Stakeholders must formally track new and emerging regulations and provide guidance to business partner stakeholders. This includes formal impact assessments.
- **Roles and Responsibilities:** Risk Management, Legal & Compliance, as well as Information Security & Privacy Stakeholders may decide to execute ongoing assessments in a singular process, or these processes may be separated. These stakeholders also are required to report on adherence to internal policies, regulations, or guidance.
- **Objective Outcome:** Guidance from stakeholders on regulatory, compliance, legal, or privacy risks that may impact the risk-tolerance identified as part of the risk assessment. Ongoing reporting on adherence to regulations or internal policy, as well as any emerging risks.

Independent Validation

Periodic Independent Validation

Third Party Assessments

• Independent Validation (In Parallel)

- **Description:** Periodically, the AI system, controls, alignment with regulations and internal policy, and any ongoing impact assessments should be independently validated, with results presented to the appropriate governance committees. This process is critical to account for emerging regulations, incorporate best practice, implement effective challenge, and provide the governing stakeholders a comprehensive view into operational risks and alignment with the agreed risk appetite.
- **Roles and Responsibilities:** Risk Management Stakeholders are responsible for ensuring independent validations are periodically executed. Audit Stakeholders will also perform their independent validations per their determined cycles.
- **Objective Outcome:** Quantitative and qualitative reporting to governance committees that provides insight into risks and controls for the AI system.

9.2 Consolidated Success Criteria Checklist

How to Create Policy Guardrails

Action	Outcome	Key Metrics	Evidence
Risk Assessment	An inventory of AI system risks that includes an assessment of whether each risk requires a policy guardrail	Not Applicable	Spreadsheet/Document containing risk assessment and guardrail mapping. Sign-off from relevant stakeholders responsible for identification is received
Guardrail Definition	Policy-based guardrails are clearly defined with components included: title, behavior, allowed/disallowed behaviors, examples	Policies are defined for 100% of risks identified to require guardrails	Spreadsheet/Document containing guardrail definitions for each guardrail identified above
Validation	Sign-off from relevant stakeholders for each policy definition. Here, relevant stakeholders are those responsible for validation based on the selected operating model.	100% of defined guardrails receive sign-off	Document with sign-off from relevant stakeholders for each policy definition.
Implementation	Guardrails are created in guardrail platform and ready for evaluation	100% of guardrails are implemented	Policy-based guardrail models that can be evaluated

How to Evaluate Guardrails

Create a Benchmark Dataset	For each guardrail, a benchmarking set with 100+ datapoints has been created. The composition of the benchmarking dataset is diverse and includes both compliant and non-compliant data	Number of data points in benchmarking dataset	Spreadsheet containing the benchmarking dataset
Calculate Model Performance Metrics	Metrics such as F1, recall, and precision are calculated and documented. Target metrics are also outlined with justifications for each target	F1, Recall, Precision, FNR, FPR	Spreadsheet/Document containing target metrics and metric calculations

Action	Outcome	Key Metrics	Evidence
Threshold Analysis	ROC curve (or similar thresholding curve) is created and documented. An optimal threshold for the guardrail is selected and signed off by relevant stakeholders	ROC, AUC, TPR vs. FPR Tradeoff	Document with threshold curve analysis and selected ideal threshold value
Failure Mode Analysis	Failure modes are recorded in a spread sheet or document. Each failure mode should associated data points from the benchmark dataset to help showcase the pattern of failure	% of incorrect datapoints that are mapped to your failure mode taxonomy	Document with failure modes and associated datapoints

How to Monitor Guardrails

Data Capture Setup	Metadata outlined above are collected to enable monitoring, incident management, and reporting	F1, Recall, Precision, False Negative Rate, False Positive Rate	Documented processes to gather metrics with regular reporting with key metrics
Guardrail Monitoring	Tooling is set up to enable continual monitoring of guardrail effectiveness	Not Applicable	Implemented and evidences technology or processes that facilitate production of key metrics
Incident Management	Tooling is set up to enable detection and resolution of incidents	Guardrail Violation Metrics Anomaly Usage Metrics	Implemented technology or processes to facilitate key metric collection and assessment
Reporting	Dashboards and reports are created with insights around AI system compliance	Guardrail Effectiveness Metrics (F1, Recall, Precision, False Negative Rate, False Positive Rate) Incident Resolution Metrics (# of Incidents, Time to Resolution, Caught Violations, Uncaught Violations, System Anomalies, User Anomalies)	Implemented technology or processes to facilitate key metric collection and assessment
Guardrail Improvement	Insights from monitoring and reporting are used to improve guardrail performance	F1, Recall, Precision, FN Rate, FP Rate	Implementation and evidences that leverage insights from monitoring and reporting to improve the guardrail



9.2 Glossary

The following is a subset of glossary terms pertinent to the Playbook. For additional term definitions, please reach out to the Dynamo AI team.

AI System

A machine-based system that:

- Can operate with varying levels of autonomy
- Infers outputs from received inputs
- Can generate predictions, content, recommendations, or decisions
- Can influence physical or virtual environments

Ground Truth

Ground truth, or ground truth data represents accurate, verified data used to train, validate, and test AI models. It serves as the "correct answer" based on real-world observations, which can help evaluate model performance and ensure accurate, reliable predictions.

Guardrail

Systems that monitor and filter the inputs and outputs of trained large language models. Guardrail methodologies include, but are not limited to, using different language models to monitor a target language model. We sometimes refer to these as policy-based guardrails.

Input Policy

Input policies are policies that are used to filter and monitor user inputs to a model. DynamoGuard applies guardrails (made up of defined policies or rules) to both the inputs and outputs of large language models.

Large Language Model (LLM)

Large language models (LLMs) are deep learning algorithms that can recognize, summarize, translate, predict, and generate content using large datasets. LLMs represent a class of deep learning architectures called transformer networks.

Model Hallucination

An AI model generated response that may be incorrect or misleading.

Output Policy

Output policies are created and defined policy rules that are used to block or filter (and monitor) the model's generated text. DynamoGuard applies guardrails (made up of policies or rules) to both the inputs and outputs of large language models.

Personally Identifiable Information (PII)

Information relating to an identified or identifiable natural person. This may include, but is not limited to, identifiers such as name, identification number, location data, online identifiers or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Policy

A singular rule that is used to 'guard' a model. A policy is used to block or flag model prompts or responses. Each policy is associated with an action and contains both a definition as well as examples of allowable and disallowable inputs or outputs, and associated training data.

Policy Violation

An instance of a policy being breached by either a prompt or a response.

Prompt

Task-specific instructions in the form of text given to a large language model to generate specific responses or outputs.

Prompt Injection Attacks

Prompt injection is a specific method of attacking an AI model. Prompt injection attacks are specially crafted prompt inputs which contain some type of malicious material inside the prompt in order to manipulate the model to act upon the malicious material. While prompt injection is a common method for achieving jailbreaking, some jailbreaking, such as those caused by typos or accidental rephrasing, do not necessarily involve prompt injection. Prompt injection always involves malicious intent as it is defined by the inclusion of malicious material within the input prompt.

Red-Teaming

The process of identifying vulnerabilities within an AI system by simulating adversarial attacks.

Retrieval Augmented Generation (RAG)

A framework used to retrieve data from an external source and use it as context to a model.

ROC Curve Analysis

A method to assess the performance of a classification model by plotting the True Positive Rate against the False Positive Rate at various classification thresholds. This analysis can be used to set the desired classification threshold for DynamoGuard models.

Toxicity

Content that is judged to be harmful to users in a professional and enterprise setting. This includes, but is not limited to, harmful, violent, hurtful and sexual content.

