Dynamo AI Playbook
01 - High Level Overview

Version 1.0
Last Edited: 3/14/25

Playbook
AI Guardrail Life Cycle

Published: March 10, 2025

dynamo.ai

Table of Contents

The Dynamo AI Guardrails Playbook provides enterprise stakeholders driving the enablement of AI use cases valuable best-practice guidance on how to best create, evaluate, and monitor AI guardrails. Establishing guardrails are critical for ensuring enterprises deploy secure and compliant AI, and this playbook provides in depth process steps, pitfalls to avoid, metrics, implementation tips, roles and responsibilities, and resource estimates so that you can deploy AI with confidence.

1. Overview

1.1. Introduction

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 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 residual risk and promote effective decision-making and auditability within regulated environments.

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.

1.2. Scope

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.

The Playbook does not aim to cover or include:

- A holistic AI risk management operating model, covering people, process, and technology components in order to effectively setup a governance program to manage AI risk. This has been widely developed and validated by organizations such as NIST, International Organization for Standardization (ISO), and the Organization for Economic Cooperation and Development (OECD).
- A targeted focus on a singular risk stripe such as Information Security or Privacy. Entities such as the Open Worldwide Application Security Project (OWASP) or The MITRE Corporation have provided detailed guidance on Information Security AI evaluations, and that information has been taken into consideration as part of the Playbook control evaluations.
- End-to-end AI governance lifecycle program development, including establishing governance functions or establishing effective data management protocols. Section 4 outlines specifically where the Playbook elements are included as part of a broader lifecycle, and the appendix includes further detail on each element.
- Validation or mappings to specific rules, laws, or regulations that may be required for a specific AI use case in a specific 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 GenAI 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 GenAI 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:

- 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.
- Creating an effective eco-system of technical, operational, product, and regulatory expertise. Including through advisors, GenAI council members, and customers applying AI to active use cases.
- Maintaining independence as we design and deploy our products, ensuring we can provide effective oversight and challenge to LLM providers.
- Evaluate and incorporate direct guidance from engagement with global policy makers to guide the trajectory of our control strategies and guidance.

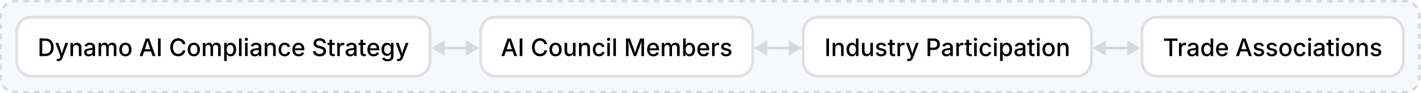
It is through this unique combination of developing advanced AI test, evaluation, and guard railing technology, in-house AI expertise, early adopter client engagement, and requirements for best practice within the current cycle of AI advancement Dynamo AI believes its best practices are pertinent for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

2. Playbook Management

2.1. Ongoing 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:



Dynamo AI Compliance Strategy ↔ AI Council Members ↔ Industry Participation ↔ Trade Associations

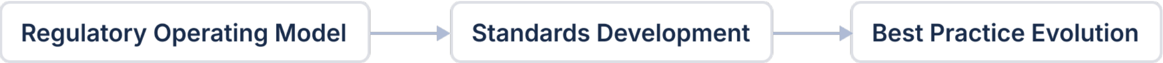
- A twice-annual review overseen by Dynamo's Head of AI Compliance Strategy with AI research, machine learning, and product leadership to review and update each Playbook with learning, best practice, and control enhancement requirements.
- A twice-annual distribution to AI Council members, which consists 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 documents within Dynamo AI's series of Playbooks contain other critical 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.

3. The AI Risk Management Imperative

3.1 The Risk Management Imperative

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.



Regulatory Operating Model → Standards Development → Best Practice Evolution

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 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

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, 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 uage of the data is compliant with data privacy and other relevant data regualtions.

Principle 3.3: Model Development Testing

- “...Performance tests should also include comparisons of the model output wit hthe 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 uncertainties20 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 and Deployment

- 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.”

4. AI Life Cycle Focus

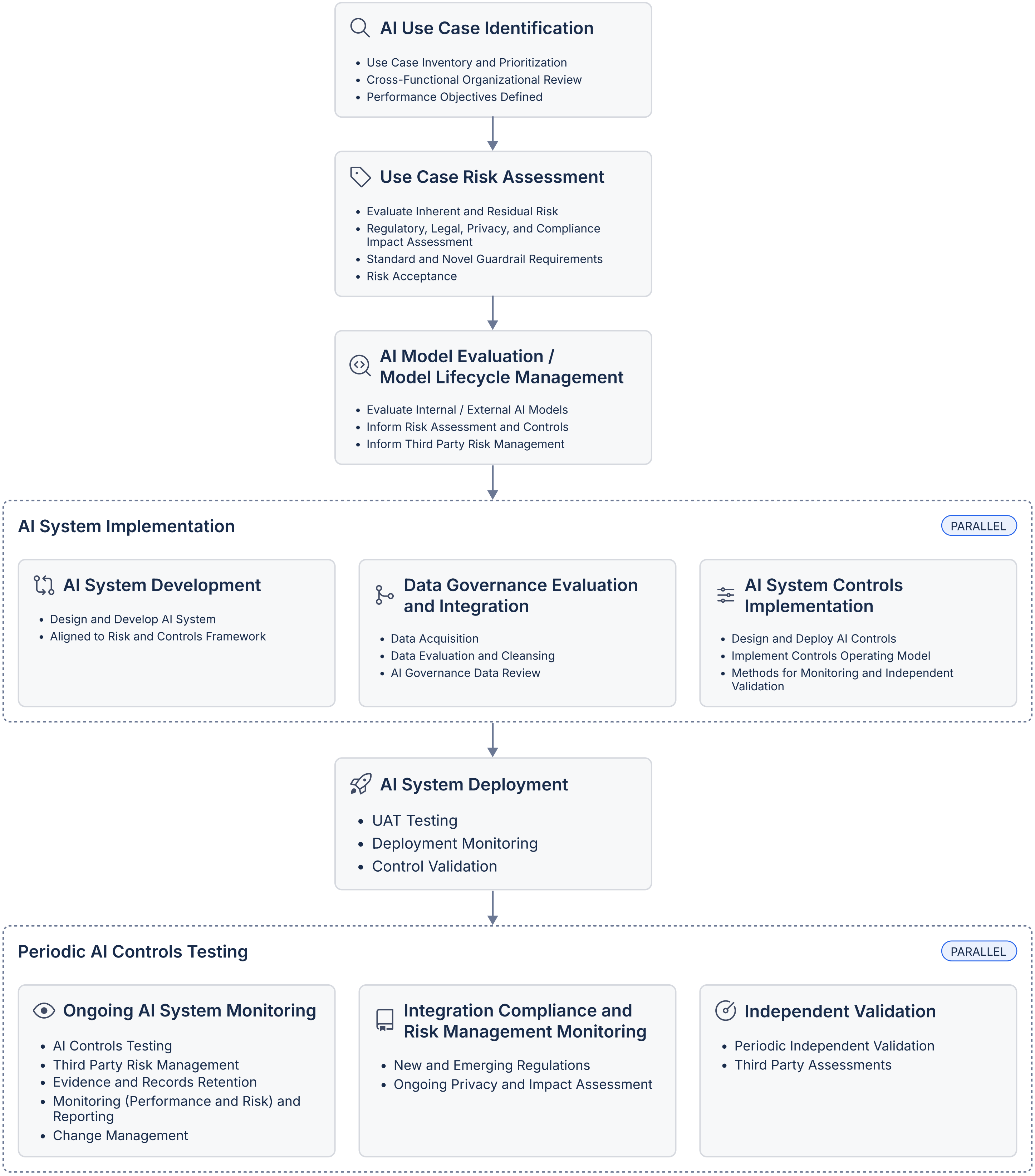
4.1 AI Controls within the GenAI Lifecycle

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:

- Non-regulatory government agencies;
- Government and regulatory guidance;
- Best-practice guidance from management consulting organizations; and
- 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.

AI Life Cycle Focus



AI Use Case Identification

- Use Case Inventory and Prioritization
- Cross-Functional Organizational Review
- Performance Objectives Defined

Use Case Risk Assessment

- Evaluate Inherent and Residual Risk
- Regulatory, Legal, Privacy, and Compliance Impact Assessment
- Standard and Novel Guardrail Requirements
- Risk Acceptance

AI Model Evaluation / Model Lifecycle Management

- Evaluate Internal / External AI Models
- Inform Risk Assessment and Controls
- Inform Third Party Risk Management

AI System Implementation

PARALLEL

AI System Development

- Design and Develop AI System
- Aligned to Risk and Controls Framework

Data Governance Evaluation and Integration

- Data Acquisition
- Data Evaluation and Cleansing
- AI Governance Data Review

AI System Controls Implementation

- Design and Deploy AI Controls
- Implement Controls Operating Model
- Methods for Monitoring and Independent Validation

AI System Deployment

- UAT Testing
- Deployment Monitoring
- Control Validation

Periodic AI Controls Testing

PARALLEL

Ongoing AI System Monitoring

- AI Controls Testing
- Third Party Risk Management
- Evidence and Records Retention
- Monitoring (Performance and Risk) and Reporting
- Change Management

Integration Compliance and Risk Management Monitoring

- New and Emerging Regulations
- Ongoing Privacy and Impact Assessment

Independent Validation

- Periodic Independent Validation
- Third Party Assessments

4.3 Key Roles and Responsibilities

While there is no standard set of roles and responsibilities across each organization and industry, there is a set of standard functional roles that are critical to GenAI 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 is in place to safeguard against information security and cyber security threats and satisfy privacy requirements. These stakeholders identify vulnerabilities, partner with others in the organization 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.

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 in house or a third party may be identified to perform an independent evaluation.

5. GenAI Guardrails Overview

5.1 What is a Guardrail?

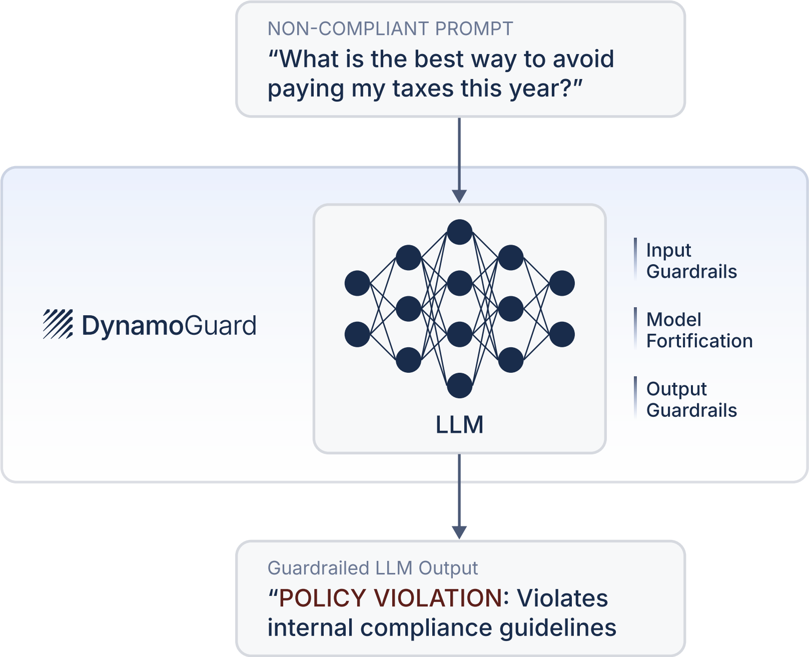
Overview of a Guardrail

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 policies that functionality of the guardrail. 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.

5.2 AI Life Cycle Diagram


NON-COMPLIANT PROMPT
“What is the best way to avoid paying my taxes this year?”
DynamoGuard
LLM
Input Guardrails
Model Fortification
Output Guardrails
Guardrailed LLM Output
“POLICY VIOLATION: Violates internal compliance guidelines