PRIME Guardrails: A General, Low-Latency Safety Framework for Generative AI

Generative AI's current safety measures often struggle with a balance between efficacy and speed. Simple keyword or filter-based methods are easily circumvented. In contrast, other approaches are frequently too slow for real-time use or are highly application-specific, which restricts their reusability and scalability. These limitations can hinder responsible AI deployment and erode public trust.

We propose PRIME, a general, low-latency, modular, and modality-agnostic framework for building robust and adaptable safety guardrails. It can be instantiated for diverse generative AI applications like text-based chat and text-to-image systems, defined by five core components:

P - Policy Specification: A declarative, human-readable schema that allows for a domain-agnostic definition of safety rules, including content categories and corresponding actions(e.g. Allow, rewrite, refuse). This policy can be configured without requiring any changes to the underlying models.

R - Risk Sensing & Scoring: A latency-first, early-exit pipeline that fuses lexical rules, semantic similarity, and fine-tuned classifiers to detect and score harmful intent. This modular approach allows for general, application-agnostic safety checks (e.g., for unsafe content or prompt hacking) while also enabling the calibration of risk signals based on a specific policy or domain. For example, the term "shooting" might be acceptable for a sports application but not for a news application, and our detectors can be modulated to account for these nuances. This component is key to making the system resistant to adversarial techniques like badgering and prompt hacking.

I - Intervention Router: The I in PRIME is a deterministic controller that acts as the system's "brain" or "agentic layer" leveraging both policy rules and risk scores. This router intelligently maps these inputs to a policy-aware action, such as allow, rewrite, or refuse. For more complex scenarios, it can perform fine-grained rewrites to neutralize harmful content (e.g., removing protected attributes or political advocacy) while preserving the user's original intent. To ground its decisions, it can call optional tools for information retrieval and fact-checking.

M - Monitoring & Memory: A lightweight system that records prior decisions, refusal reasons, and risk spikes. This provides a robust mechanism for audit trails, consistency checks, and detection of malicious actors.

E - Evaluation & Evolution: A comprehensive and reproducible protocol for continuous improvement. This includes red-teaming prompt recipes and a suite of automated tests to expose vulnerabilities at scale, establishing confidence before public deployment. Human in the loop evaluation is subjective, costly and hard to scale. In this framework, human interventions are used to evaluate the results of bulk analysis, instead of individual data points. Over time, this iterative feedback process helps the system’s detectors adapt to new threads and ensures consistent adherence to policy.

The PRIME framework exhibits broad applicability across diverse modalities, exemplified by its integration within Text-to-Image (T2I) systems. For T2I applications, PRIME augments its core Risk Sensing & Scoring component with an image harmfulness classifier. This enhancement enables proactive identification of risks in both the input text prompt, prior to image generation, and the output image, post-generation. Additionally, a verifiability gate is incorporated within the Intervention Router to assess whether prompts pertaining to real individuals or events are consistent with fact-checked, verifiable data sources. Should a prompt be deemed unverifiable, the framework can either refuse the request or leverage the router to rewrite it into a neutral, symbolic alternative, thereby mitigating misinformation while preserving the user's creative intent.

Our replicable evaluation protocol and qualitative findings demonstrate that PRIME effectively withstands common adversarial transformations while introducing low latency overhead thereby enabling responsive and robust deployments. This framework exhibits broad applicability to standalone chat or image applications and transparently extends to other modalities.