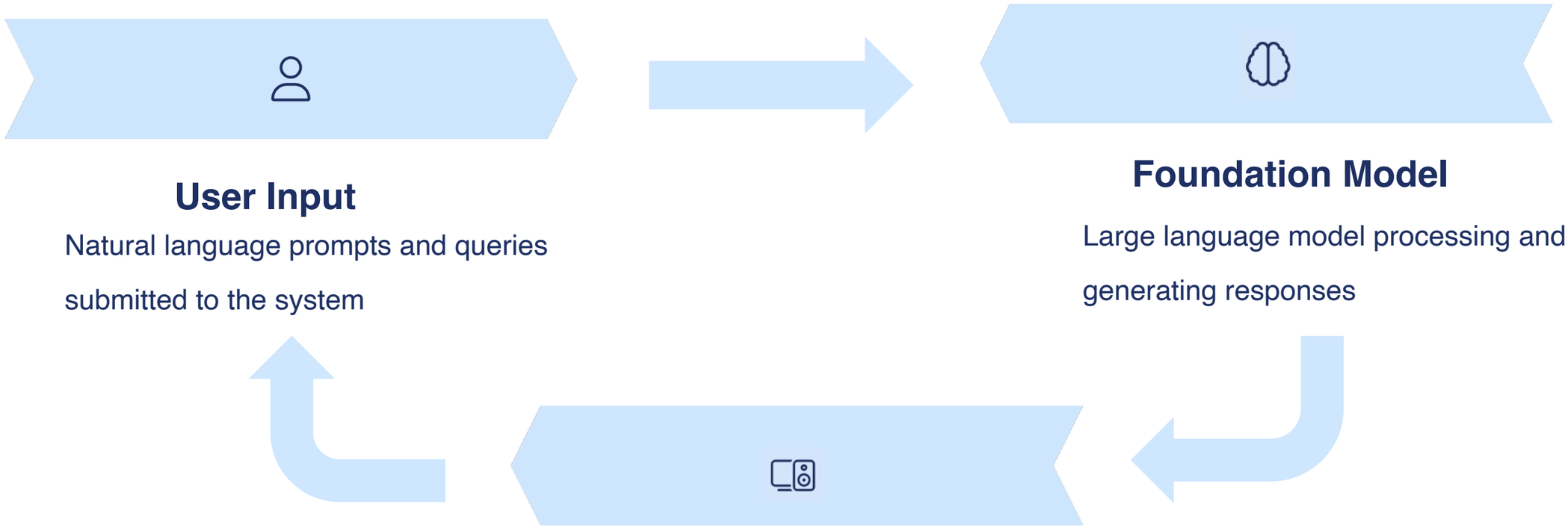
AI Guardrails: Securing GenAI Applications

Essential strategies for implementing robust security layers in generative AI systems to prevent adversarial attacks and ensure reliable outputs.

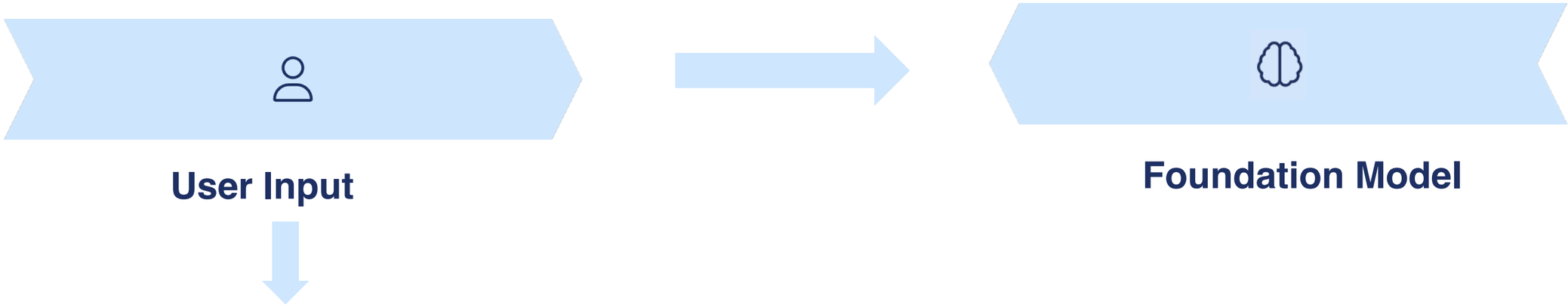
GenAI Application Architecture



Understanding the flow from user interaction to model output is crucial for identifying where security vulnerabilities can emerge and guardrails must be implemented.

User Input Guardrails

Intercept and analyze user prompts before they reach foundation models to prevent malicious inputs and inappropriate content.



Prompt Injection Detection

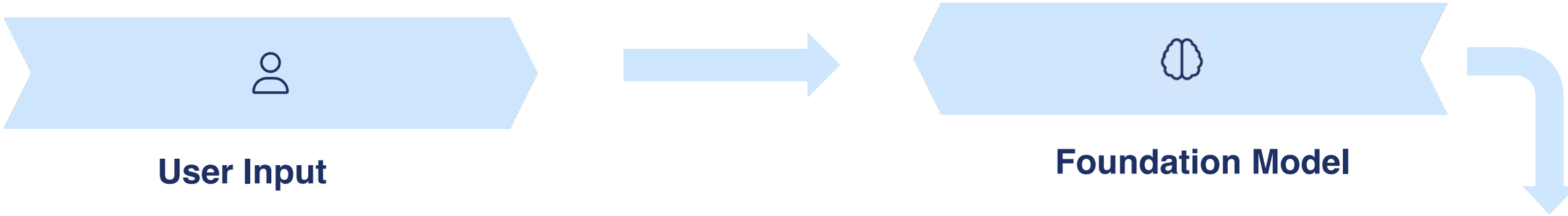
Identifies attempts to manipulate model behavior through crafted prompts

Content Moderation

Filters toxic, hateful, violent, or inappropriate content before processing

Foundation Model Response Guardrails

Analyze generated responses to ensure accuracy, relevance, and safety before delivery to users.



Hallucination Detection

Identifies factually incorrect information in model outputs using specialized evaluation models

Answer Relevancy

Ensures responses directly address user queries and maintain contextual appropriateness

Content Moderation

Optional secondary filtering to catch inappropriate content that may have been generated

Course Organization

LLM Fundamentals

Understanding large language models architecture, capabilities, and inherent vulnerabilities

AI Guardrails Necessity

Exploring security risks and business justification for implementing protective measures

Core Technologies

Vectors, embeddings, and retrieval-augmented generation (RAG) concepts

Input Protection

Implementing prompt-guard and Llama Guard for injection detection and content moderation

Response Validation

Using phi3-hallucination-judge and evaluation models for output quality assurance

Course Organization

LLM Scanner - Garak

Open-source vulnerability scanner providing standardized detection modules for comprehensive LLM security testing

Cyber Security using AI Agent

Protecting autonomous software programs that interact with environments and make decisions

Haystack Evaluation

Framework for metrics-driven assessment of hallucination detection and context relevancy

AWS Bedrock Integration

Enterprise-grade guardrails implementation using Amazon's managed AI platform

Open Source Tools

Guardrails AI and Nemo Guardrails frameworks for customizable security implementations

Comprehensive Framework Coverage

Garak LLM Scanner

Open-source vulnerability scanner providing standardized detection modules for comprehensive LLM security testing

AI Agents Security

Protecting autonomous software programs that interact with environments and make decisions

Haystack Evaluation

Framework for metrics-driven assessment of hallucination detection and context relevancy

AWS Bedrock Integration

Enterprise-grade guardrails implementation using Amazon's managed AI platform

Open-Source Tools

Guardrails AI and Nemo Guardrails frameworks for customizable security implementations