

PROMPT & AGENT SECURITY ARCHITECTURE

Layered control from prompt to tool execution.

Prompt and Agent Security Architecture

A high-level architecture view that brings user input, model traffic, AI-agent resource access, and tool calls under a common identity, policy, and audit approach.

IDENTITY

POLICY

RESOURCE

NETWORK

AUDIT

A SOFTBYT LABS PRODUCT FAMILY • HIGH-LEVEL ARCHITECTURE

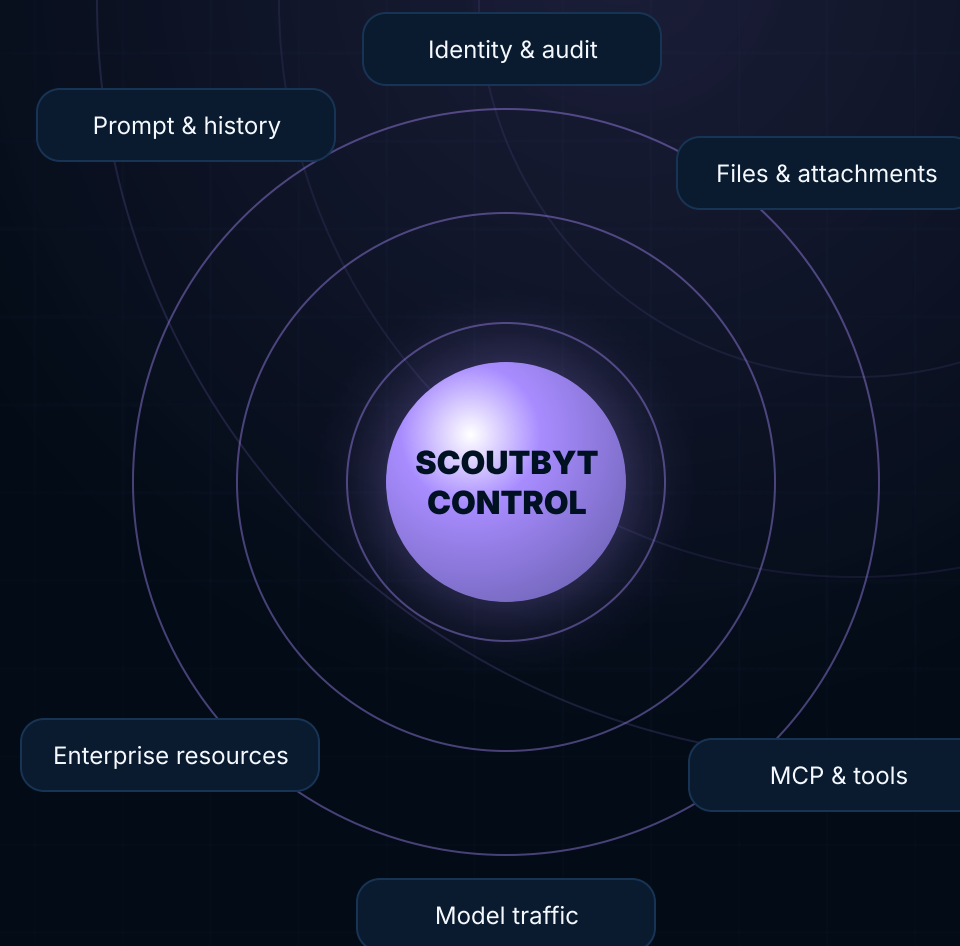


PROMPT & AGENT SECURITY ARCHITECTURE

Security surfaces that matter

AI security is more than prompt filtering; conversation history, files, resources, tools, network traffic, and identity must be considered together.

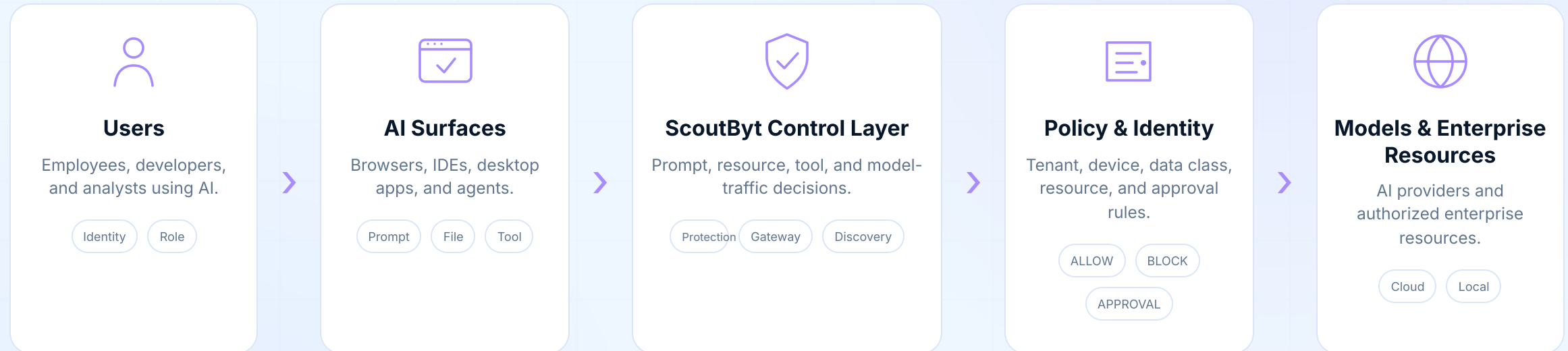
Direct data leakage • indirect prompt injection • excessive authority • shadow AI



PROMPT & AGENT SECURITY ARCHITECTURE

High-level operating model

This view explains the security model without exposing proprietary integration and validation methods.



High level

Not every integration uses the same technical method; controls depend on the security surfaces exposed by the relevant product.

PROMPT & AGENT SECURITY ARCHITECTURE

Decision model and design principles

Within a managed flow, each action produces an explicit, context-aware security decision.

Allow

Policy and context are valid.

Warn

The user receives a visible risk notice.

Block

Protected data or an unauthorized action.

Approval required

The operation is routed to a human decision.

User authority ≠ AI authority

Human access is not automatically inherited by the agent.

Least privilege

The agent receives only the resources and tools required for the task.

Pre-action control

On supported surfaces, prompts or tool actions are evaluated before execution.

Privacy-conscious audit

Policy and decision evidence is retained without unnecessary raw sensitive content.

Safe failure behavior

Managed enforcement paths do not silently release unverified operations.

Staged rollout

Discovery and simulation come first; enforcement follows on proven surfaces.