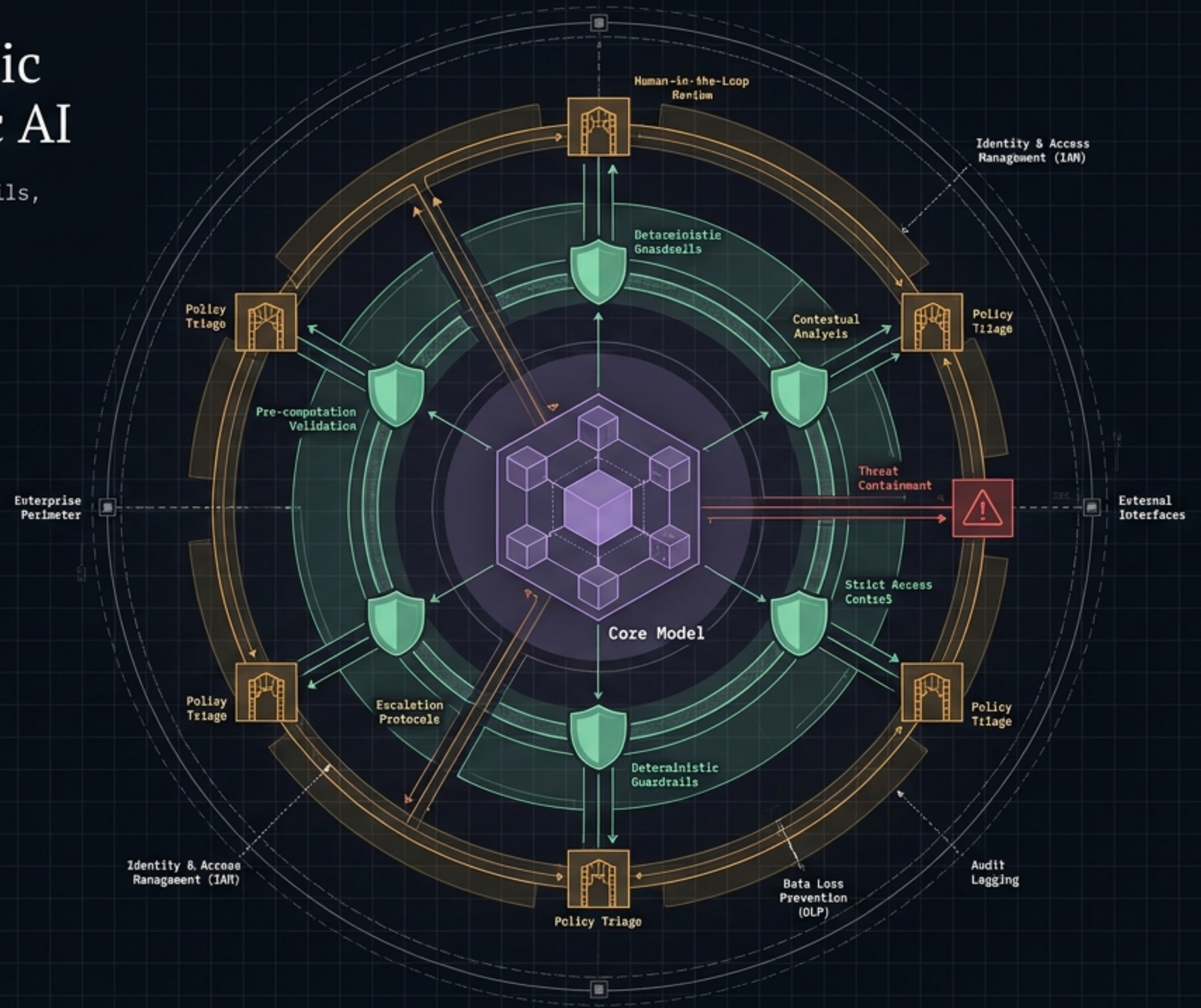
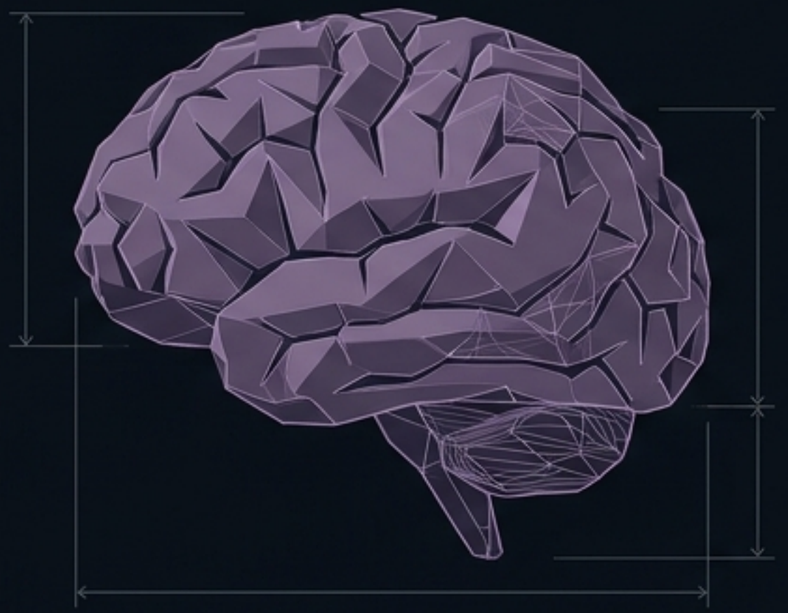
Engineering Deterministic Security for Probabilistic AI

A deep-dive architectural blueprint for guardrails, safety pipelines, and enterprise governance.

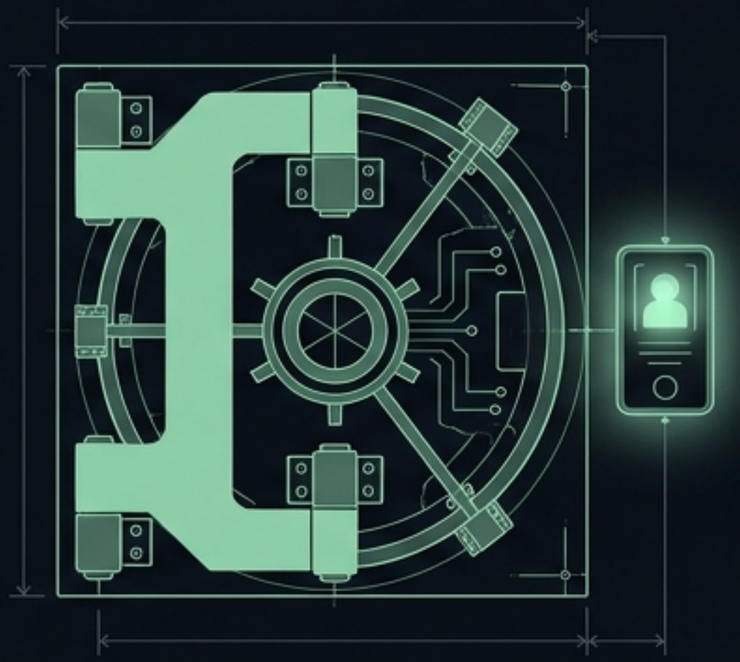


Model Alignment is NOT a Security Boundary



Trained Alignment

Professional Judgment & Ethics



Application Guardrails

Deterministic Hard Security

Foundation models are probabilistic. Because instructions and untrusted data share the exact same token stream, security, authorization, and privacy must live in deterministic application code outside the model. Trained alignment shapes default behavior, but application guardrails prevent a tricked model from transferring funds or leaking records.

The Four Structural Limits of Post-Training Safety

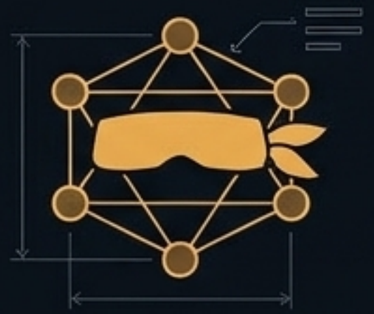


Probabilistic, Not Deterministic

Safety tuning shifts token probabilities. Adversarial framing can elicit bypasses.

Metric: A 99.5% refusal rate still yields 5,000 policy violations per million production requests.

1

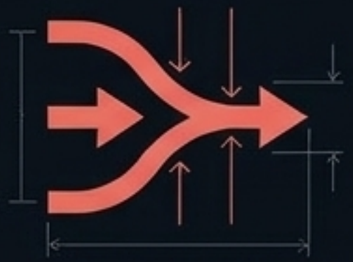


Zero Knowledge of Org Policy

Base weights do not know your company's transactional authorization ceilings ($500 vs. $25,000) or tenant RBAC matrices.

2

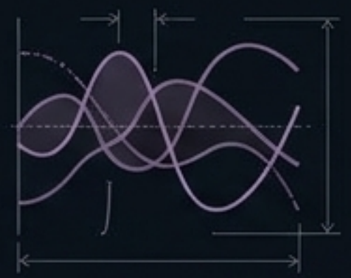
3



No Cryptographic Separation

System prompts, user inputs, and retrieved RAG documents share the same attention mechanism.

4



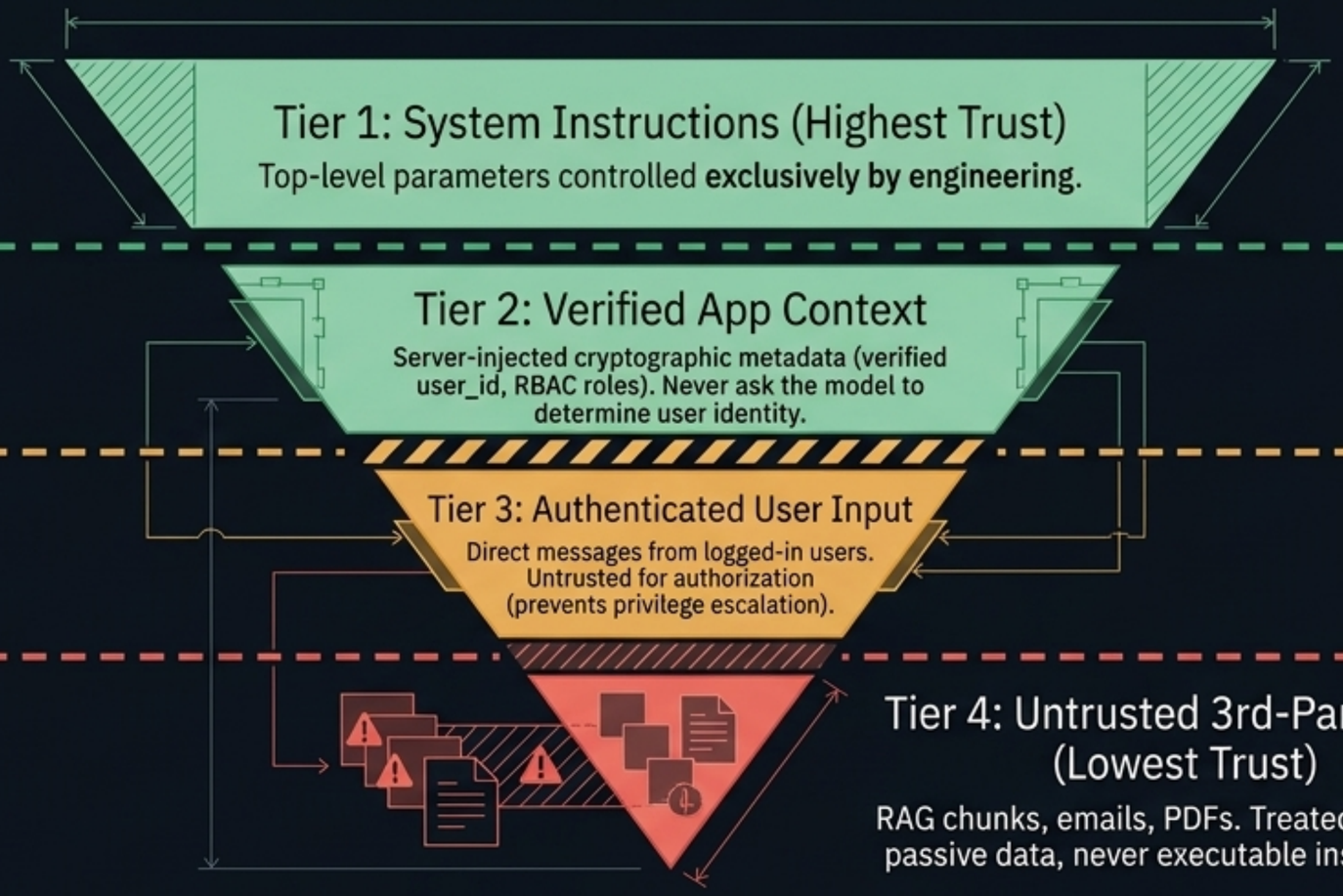
Silent Drift

Refusal sensitivity and tone shift unpredictably when upgrading model versions or snapshots.

- Probabilistic — flawed.
- Safety tuning → shifts token probabilities.
- Adversarial framing can elicit bypasses.
- Refusal sensitivity and tone shift unpredictably when upgrading model versions or snapshots.

Enforcing Zero-Trust Context Windows

Every token entering an LLM originates from one of four distinct trust tiers. Production architectures must explicitly enforce this hierarchy:

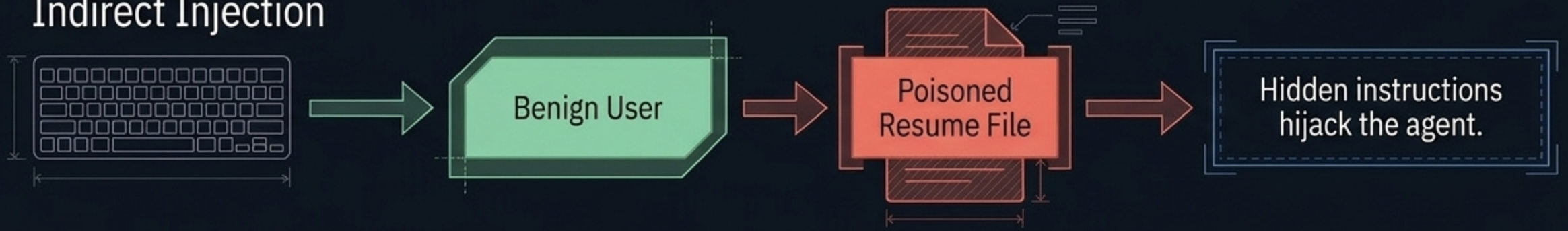


The Mechanics of Direct vs. Indirect Prompt Injection

Direct Injection



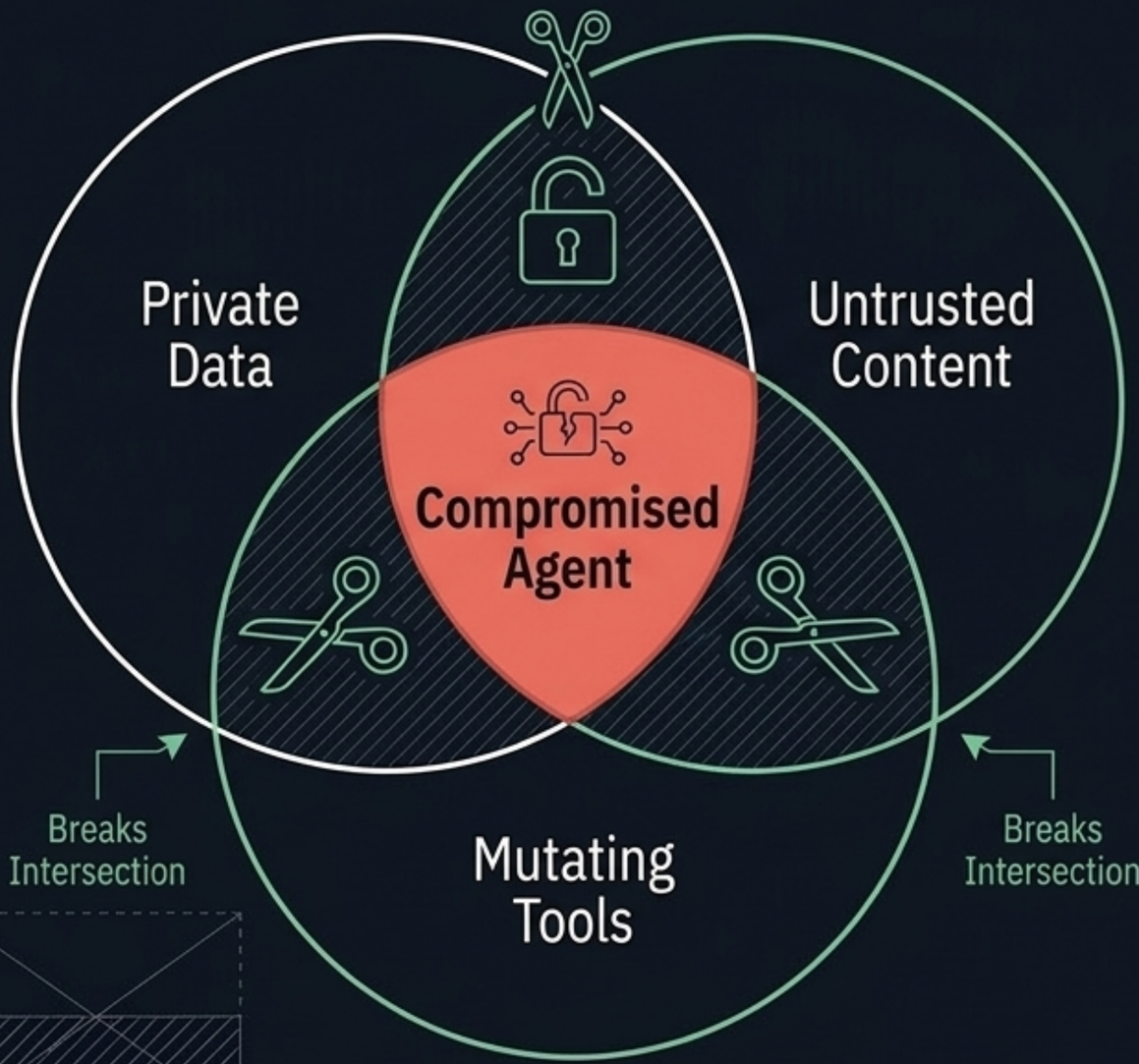
Indirect Injection



System prompts fail against indirect injection because the attacker isn't the logged-in employee—it's the external document the agent ingests.

- Explicit Delimiters: Wrap Tier 4 content in <untrusted_document src='doc_42'>
- Pre-context Sanitization: Strip zero-width characters and hidden HTML.
- Isolation: Use a tool-less subagent to extract data before passing it to the main actor.

Breaking the Lethal Trifecta with Least-Privilege Design

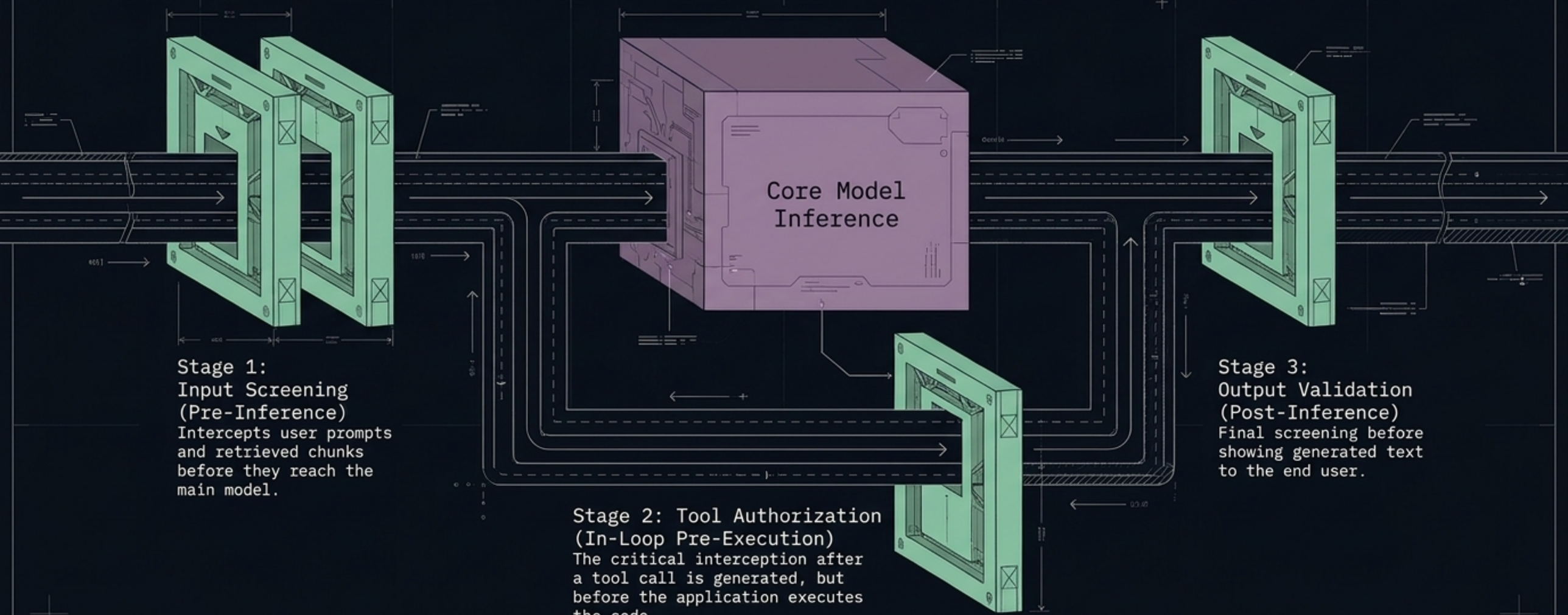


An agent is compromised when it simultaneously possesses all three traits. Break the trifecta via:

1. **Tool Registry Pruning:** Do not register `send_email` for a Q&A bot. Capabilities that do not exist cannot be hijacked.
2. **Read-Only DB Roles:** Enforce at the database engine level, not via prompt instructions.
3. **Network Egress Allowlists:** Restrict agent runtime containers to internal RPC endpoints. Block arbitrary HTTP/DNS outbound.
4. **Ephemeral Sandboxes:** Execute generated code with zero ambient authority (no host filesystem, no API keys).

The Multi-Tier Defense-in-Depth Pipeline

Production AI requires guardrails organized across 3 interception stages, utilizing both deterministic code and model-based classifiers.



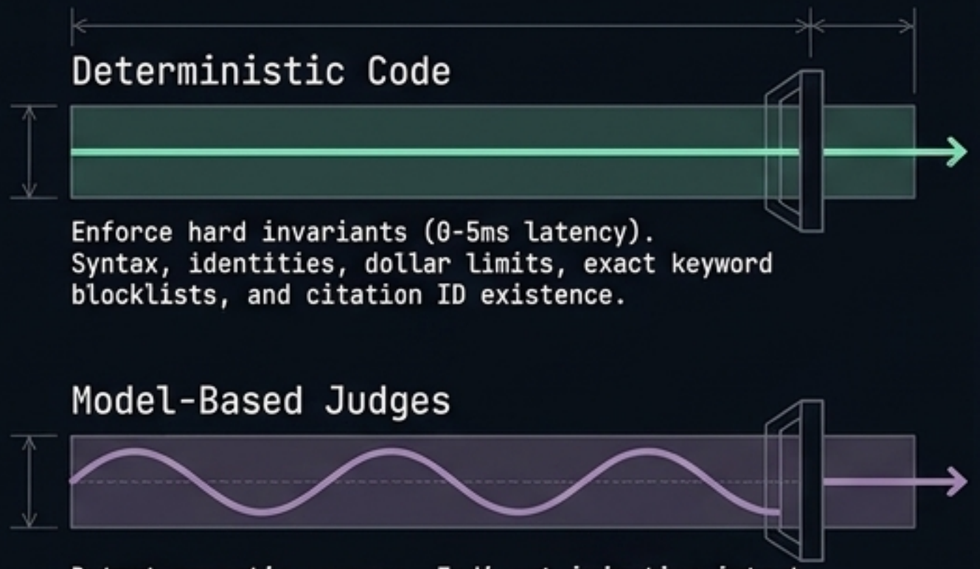
Stage 1:
Input Screening
(Pre-Inference)
Intercepts user prompts and retrieved chunks before they reach the main model.

Stage 2: Tool Authorization
(In-Loop Pre-Execution)
The critical interception after a tool call is generated, but before the application executes the code.

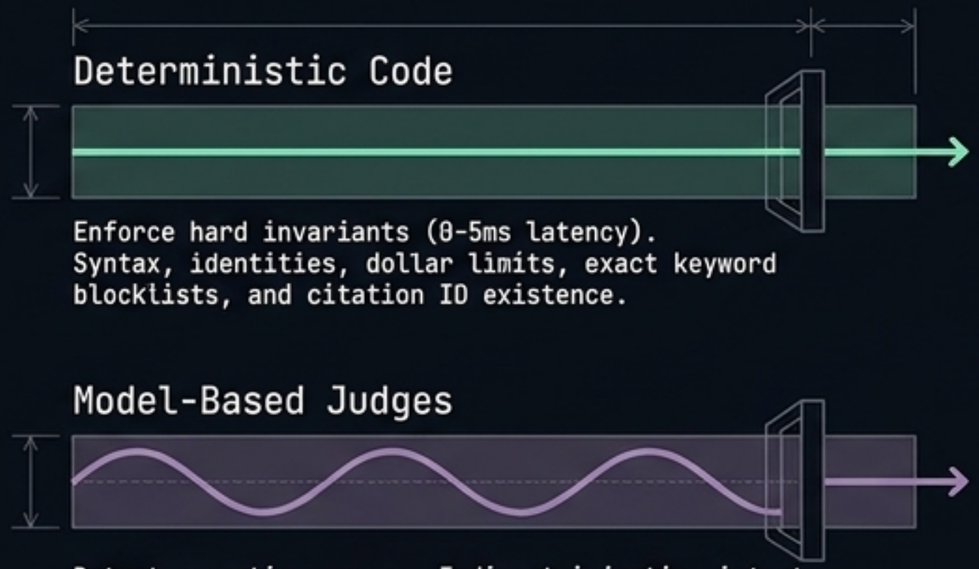
Stage 3:
Output Validation
(Post-Inference)
Final screening before showing generated text to the end user.

Input & Output Screening: Deterministic Checks vs. Model-Based Judges

Stage 1: Input Screening

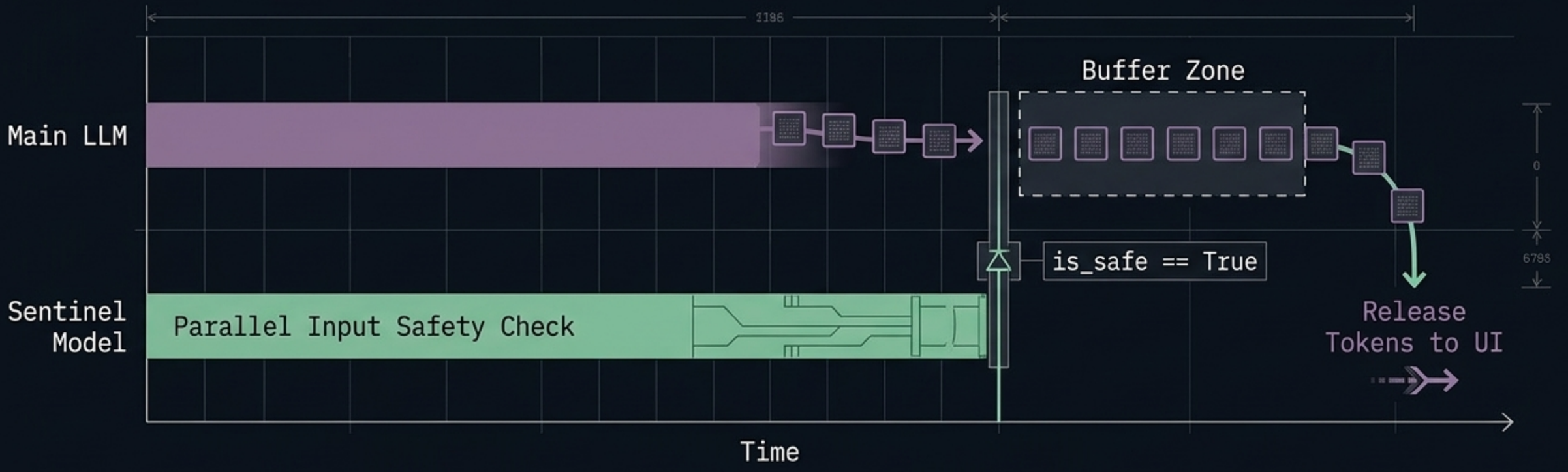


Stage 3: Output Screening



Use deterministic code for hard invariants. Reserve model-based judges for semantic nuance that regex cannot catch.

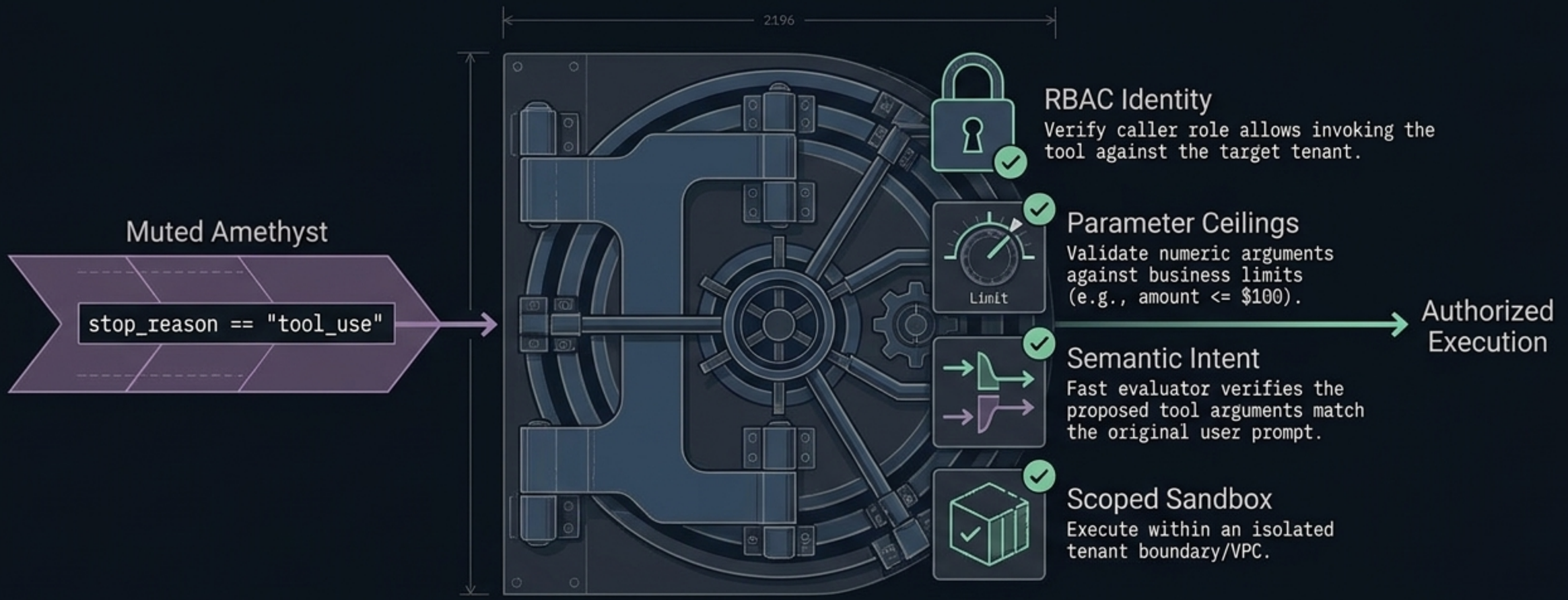
Speculative Execution to Minimize Latency (TTFT)



Problem: Sequential model-based guardrails add 150-400ms to Time-To-First-Token (TTFT).
Solution: Speculative parallel execution with buffered UI commit.

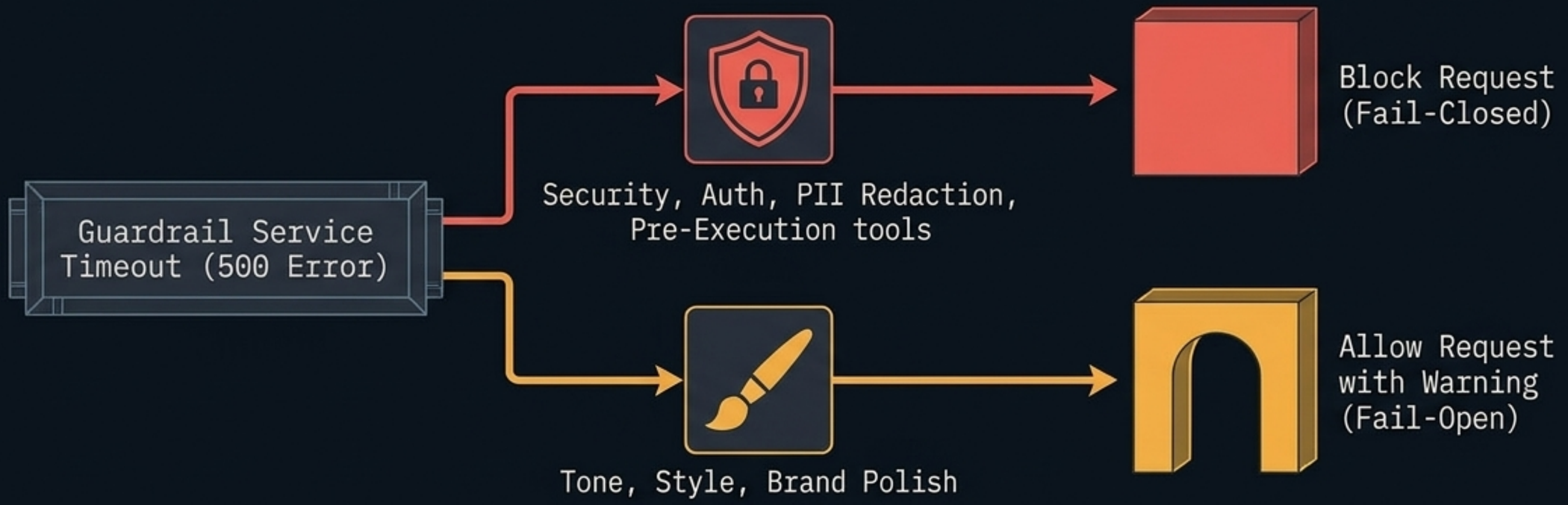
The Golden Rule: Never execute a tool_use block or commit streamed tokens to the UI while parallel guardrails are still resolving.

Intercepting Side-Effects Before Execution (Stage 2)



If an injected model emits a fund transfer, output filters are too late. Side-effecting tools must be intercepted in-loop, before the underlying code runs.

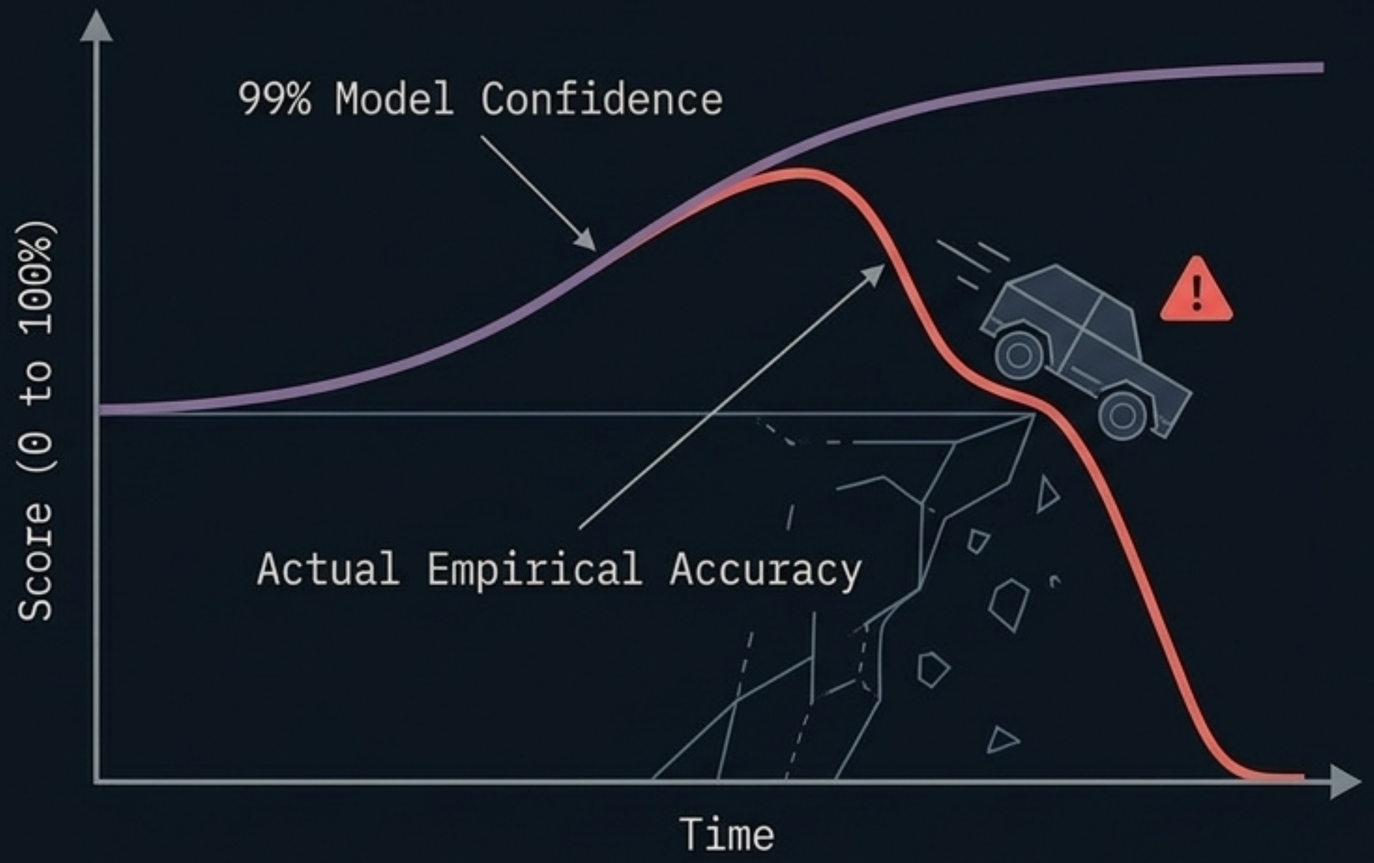
Designing for Guardrail Outages: Fail-Closed vs. Fail-Open



Attackers will intentionally trigger timeouts to bypass safety classifiers. Security and compliance checks must Fail Closed.

If an attacker can force a 500 error on your safety classifier to bypass it, your system is vulnerable.

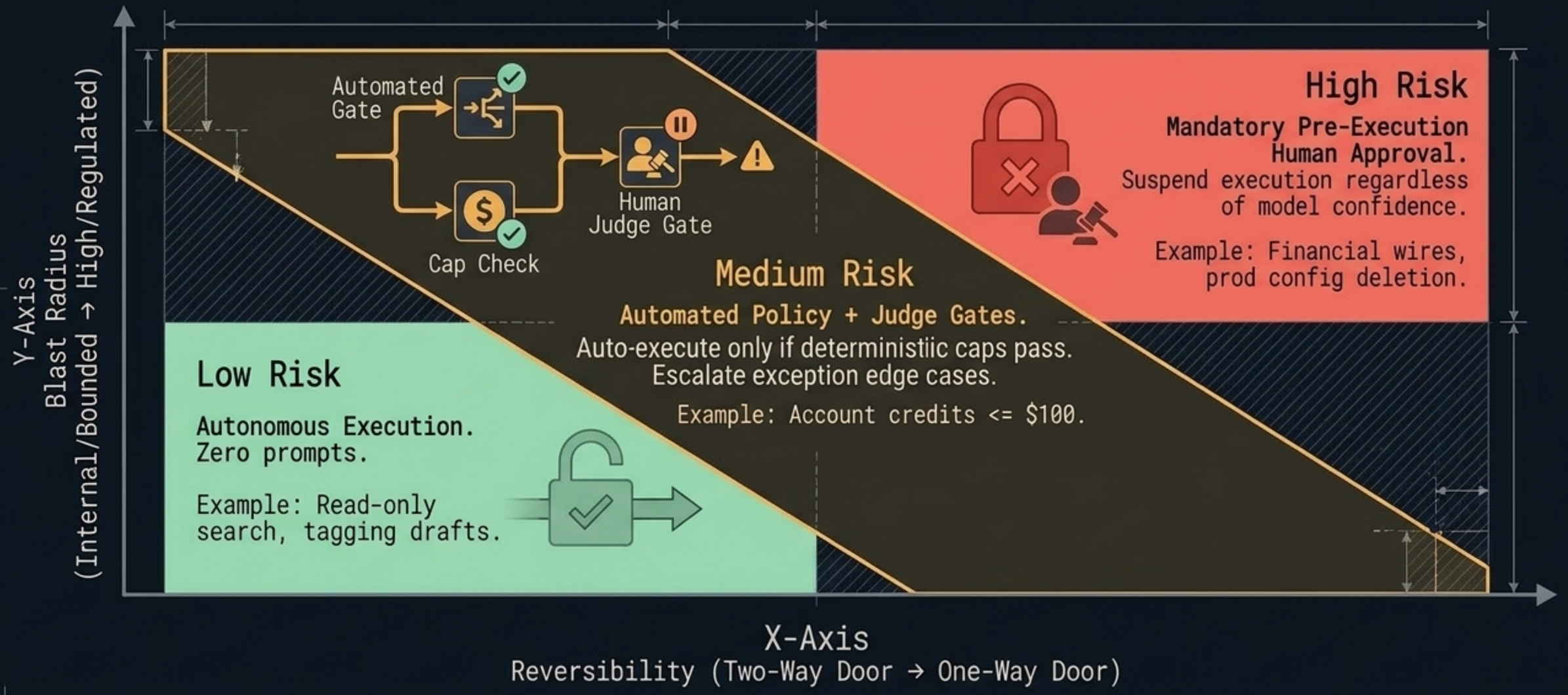
Why Confidence-Only Routing Fails



1. Poor Calibration: LLMs frequently express 99% certainty on fabricated facts or corrupted PDF layouts.
2. Injection Hijacking: Indirect prompt injections can explicitly force the model to output a high confidence JSON field to bypass human review.
3. Ignores Blast Radius: 0.92 confidence is fine for auto-tagging Jira tickets, but catastrophic for a $250,000 wire transfer.

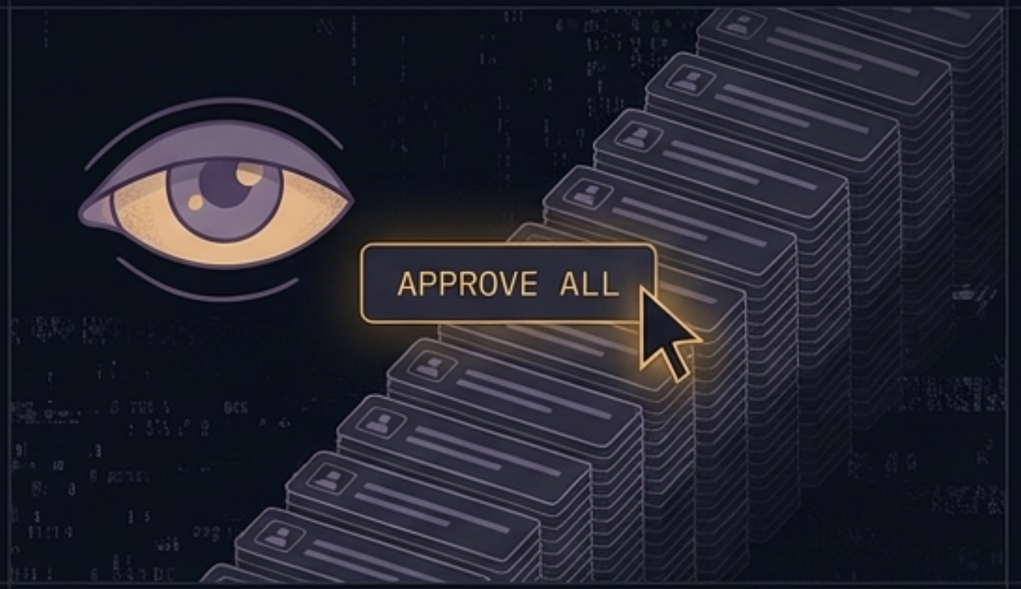
Routing actions to humans based solely on an LLM's self-reported confidence is an anti-pattern.

Risk-Based Action Routing: Impact & Reversibility



The Two Vulnerabilities of Human-in-the-Loop

Consent Fatigue



Caused by routing low-stakes actions to synchronous human queues.

Fix: Stop routing low-stakes actions to humans. Enforce deterministic policy ceilings so humans only see the 5% of true escalations.

Automation Bias



Humans uncritically rubber-stamp polished text.

Fix: UI must require active verification. Show side-by-side source snippets, extracted diffs, and explicit conflict flags (e.g., 'Source A and B disagree').

Aligning with Global Security Standards

Safe AI maps defense-in-depth engineering to public risk standards.

NIST AI RMF

Govern, Map, Measure, Manage

The Socio-technical lifecycle.

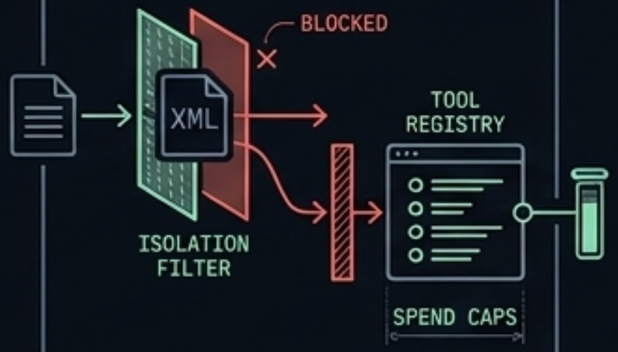
Focus on empirical evaluation, subgroup disaggregation, and fail-closed fallbacks.



OWASP LLM Top 10

Application-layer threats

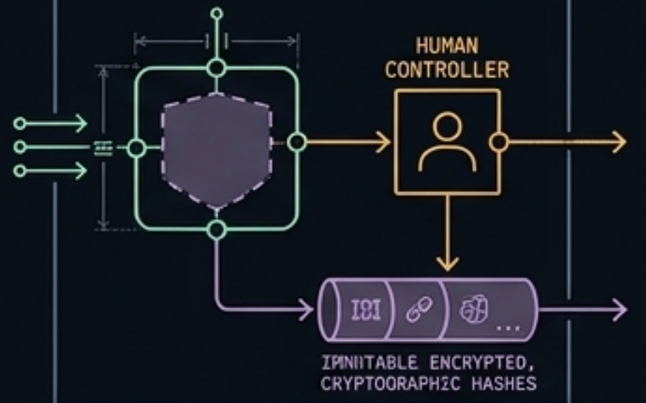
Addresses LLM01 (Prompt Injection) via XML isolation, and LLM06 (Excessive Agency) via pruned tool registries and spend caps.



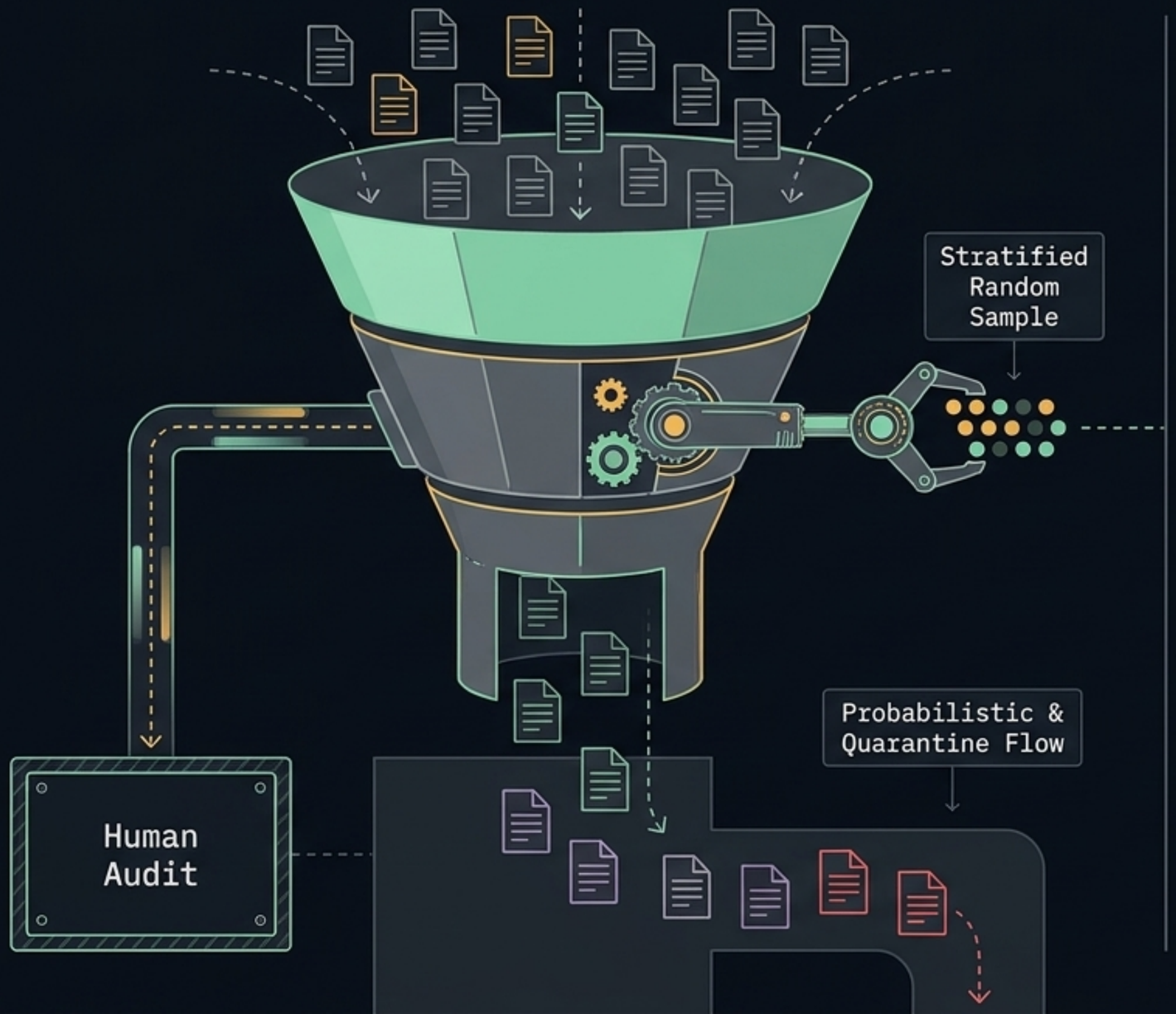
Google SAIF

Zero-trust architecture

Bounded powers, well-defined human controllers, and observable actions via immutable trace logs.



Continuous Output Verification & Stratified Sampling



METHODOLOGY

1. **Calibrated Confidence:**
   Map verbal confidence against labeled sets per field (e.g., invoice total vs liability clause).



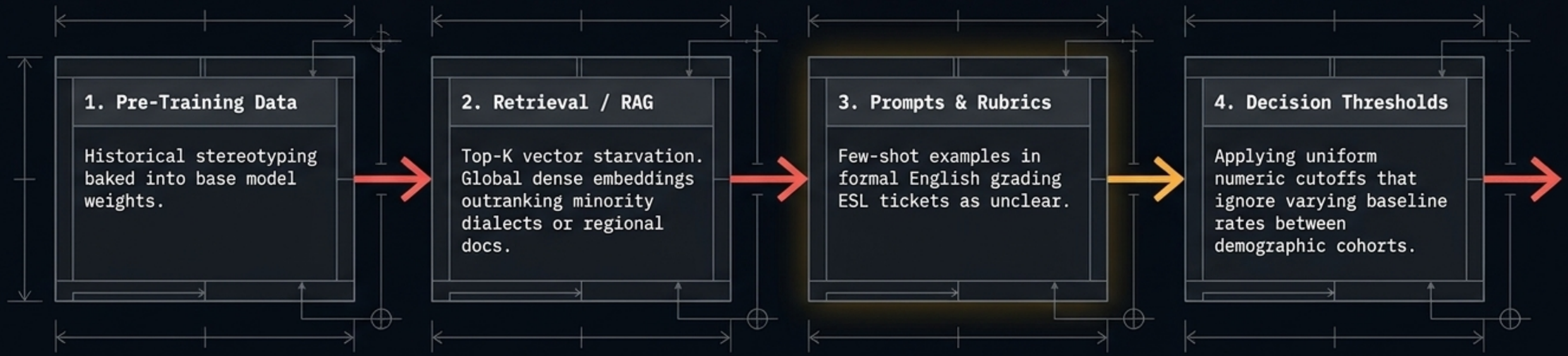
2. **Stratified Random Sampling:**
   Never use uniform sampling. Sample 2-5% of high-confidence outputs across every document class and cohort to catch silent drift in high-risk, low-volume strata.



3. **Dual-Stage Citation Verification:**
   First, does the doc_id and quoted text physically exist? Second, does the passage semantically entail the claim?

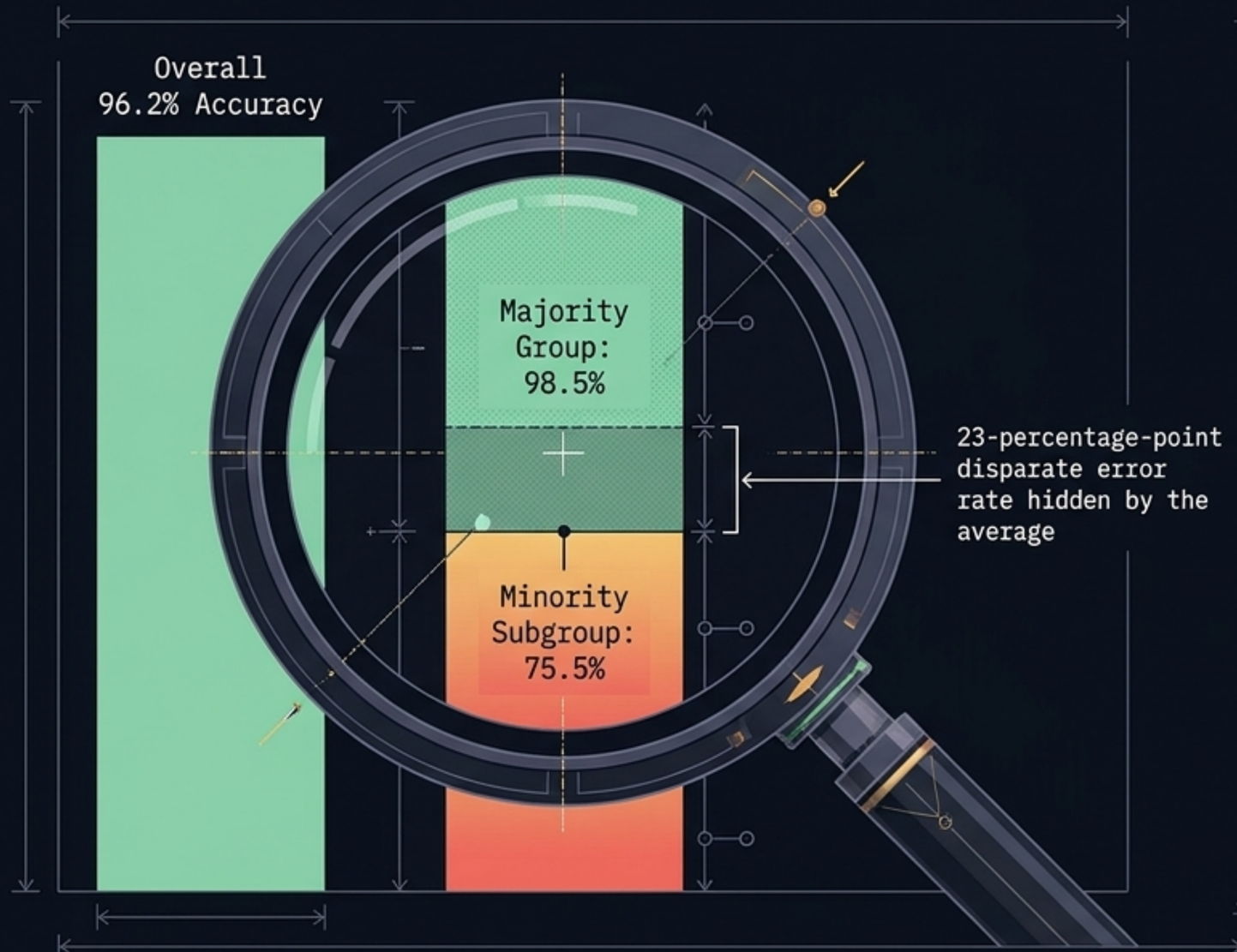
Stage 1: Physical Existence Check → Stage 2: Semantic Entailment Check

Fairness is a Pipeline Property, Not Just a Model Trait



Bias isn’t just in the weights. In accordance with the NIST AI RMF, potential bias must be analyzed across the entire socio-technical pipeline.

The Aggregate Metric Illusion & Subgroup Auditing



Subgroup Disaggregation:

Never report blended averages. Precision and FNR must be disaggregated across linguistic, demographic, and operational slices.

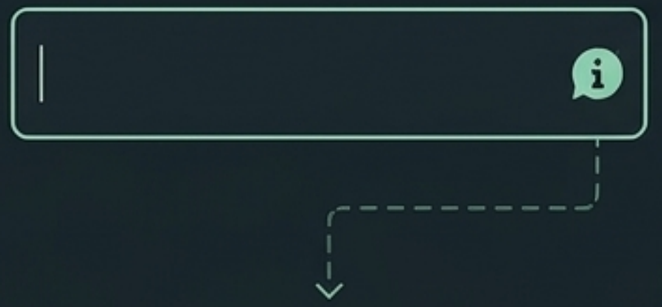
Counterfactual Perturbation Testing:

Build automated evaluations where only a demographic proxy or linguistic marker is swapped (e.g., varying names or ESL syntax).

Assert that tool selection and tone remain invariant.

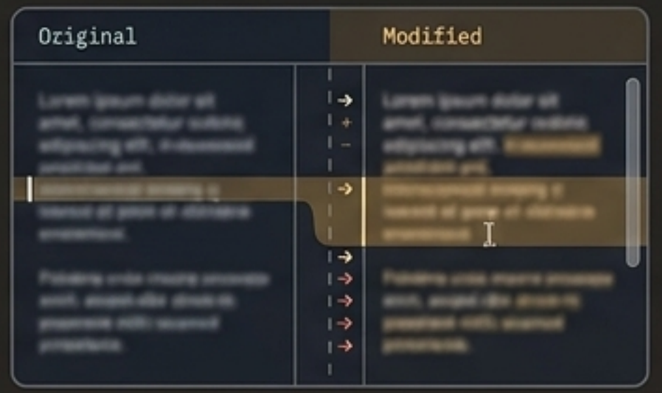
Tailoring Explainability for Multi-Audience Transparency

End Users



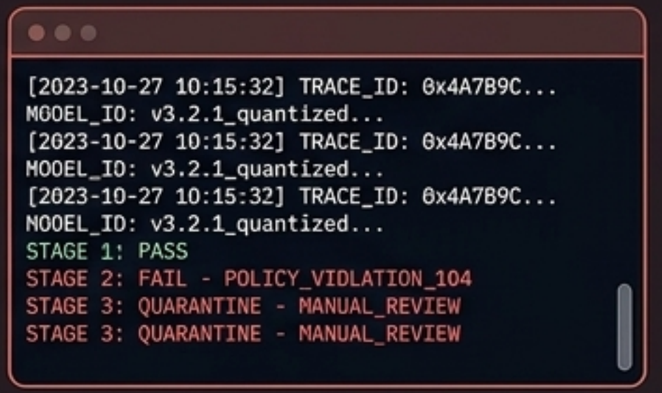
Plain language explanations, AI disclosures, clickable public policy links, and clear appeal paths.

Operators (HITL)



Inline citations, verbatim source snippets, explicit uncertainty flags, and structured action previews.

Auditors



Immutable trace_ids, pinned model IDs, and Stage 1-3 guardrail verdicts.

Architectural Rule: Never expose raw Chain-of-Thought (CoT) thinking blocks to end-users; they reflect heuristics, not grounded policy compliance.

Data Boundaries: Zero Data Retention (ZDR) & Scoping

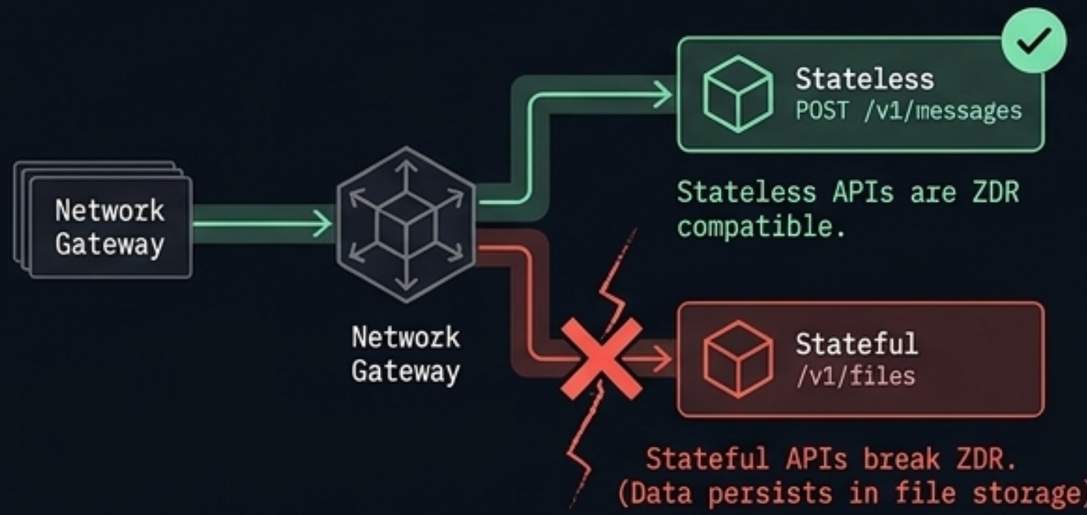
No-Training != Zero Data Retention

⚠ **Standard Commercial API**	No training, but retains abuse logs for 30 days.
Zero Data Retention (ZDR)	0-day persistence. Prompts and outputs are ephemeral in-memory only. Crucial for HIPAA/BAA.

Vault Pattern



Stateful vs Stateless Endpoints



Operationalizing Policy: ADRs & The Risk-Control Matrix

Architecture Decision Records (ADRs)



An ADR that only lists the winning design fails governance.

You must document explicitly rejected alternatives and why they failed empirical tests (to prevent future teams from removing load-bearing guardrails).

Categorize decisions as One-Way vs Two-Way doors.

AI Risk and Control Matrix (RCM)

Threat Scenario	Tech Control	Artifact	Cadence
Indirect Prompt Injection	XML wrappers + spend caps	CI red-team report	Pre-release gate

PowerPoint policies don’t pass audits. Every theoretical risk must map to a deterministic technical control and an immutable verification artifact.