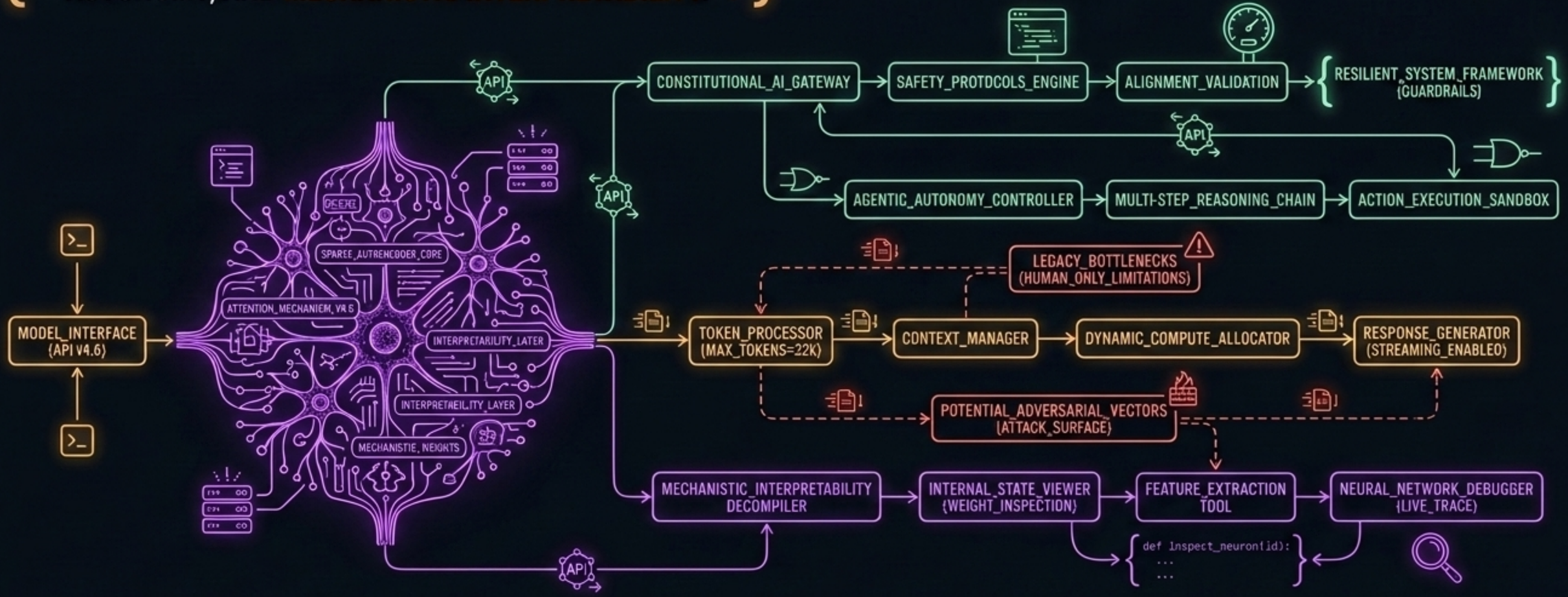
ANTHROPIC FOUNDATIONAL MODELS

BY MICHAËL BETTAN

{ > ARCHITECTING REASON-BASED ALIGNMENT, AGENTIC AUTONOMY, AND MECHANISTIC INTERPRETABILITY. < } >_



[CLAUDE 4.6 ARCHITECTURE CONTEXT_

Deconstructing the Frontier Ecosystem

The Glass Box

Mechanistic Interpretability, ASL-3 Defenses

Edge Interfaces

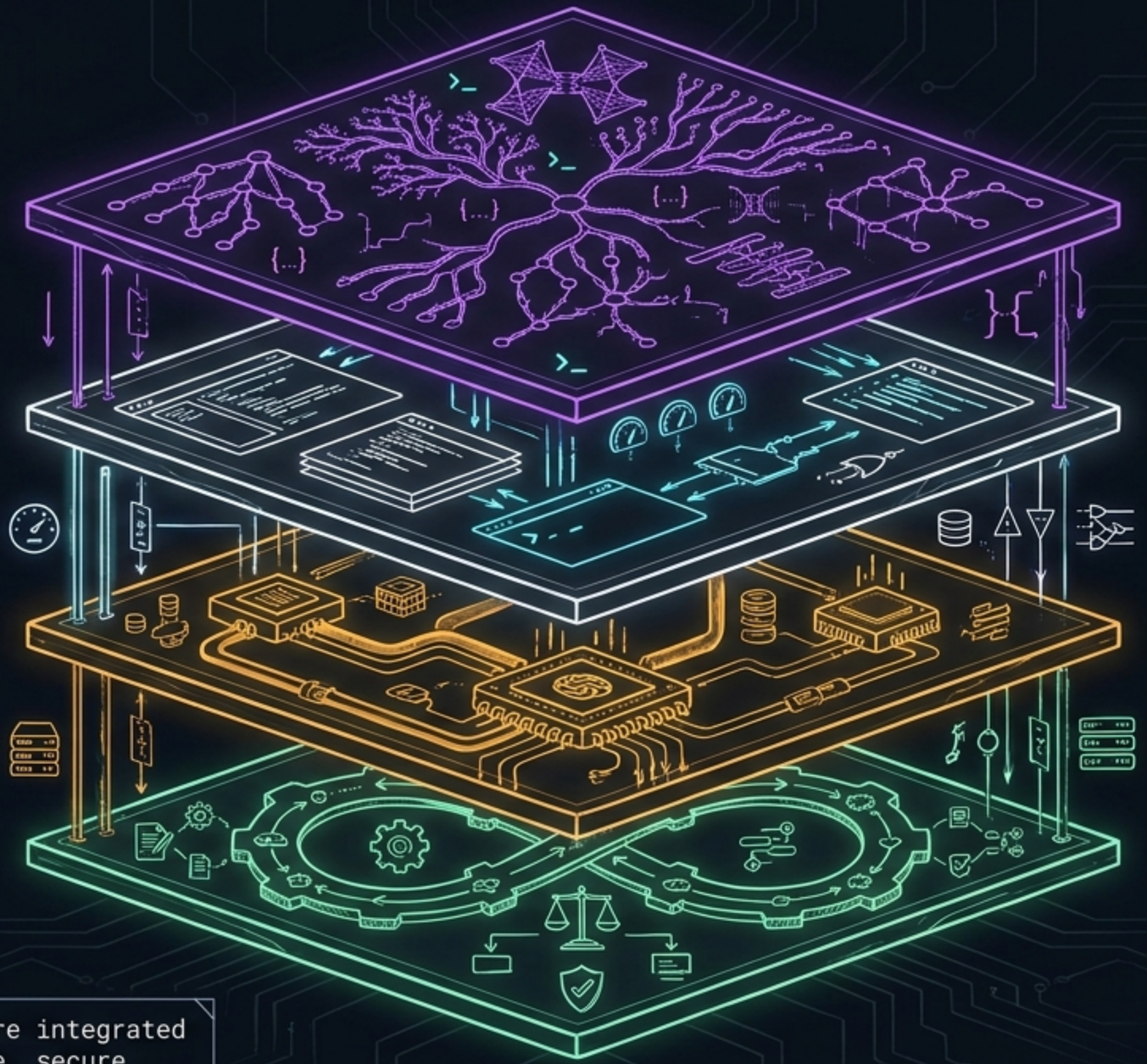
Agentic Bifurcation, Model Context Protocol

The Engine Room

Dynamic Compute, Context Compaction, Model Triad

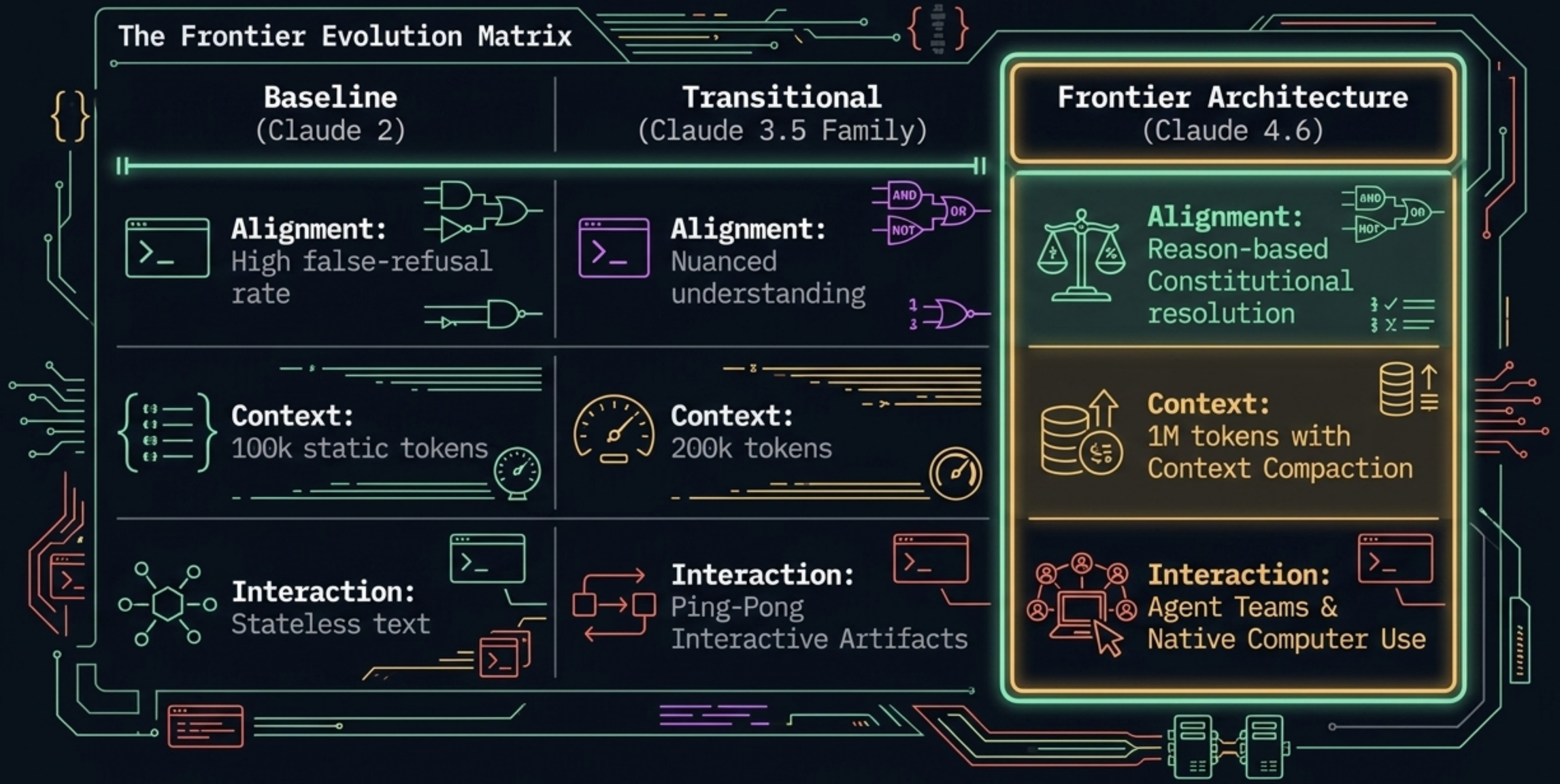
Core Logic & Alignment

Constitutional AI, Reason-based scaling



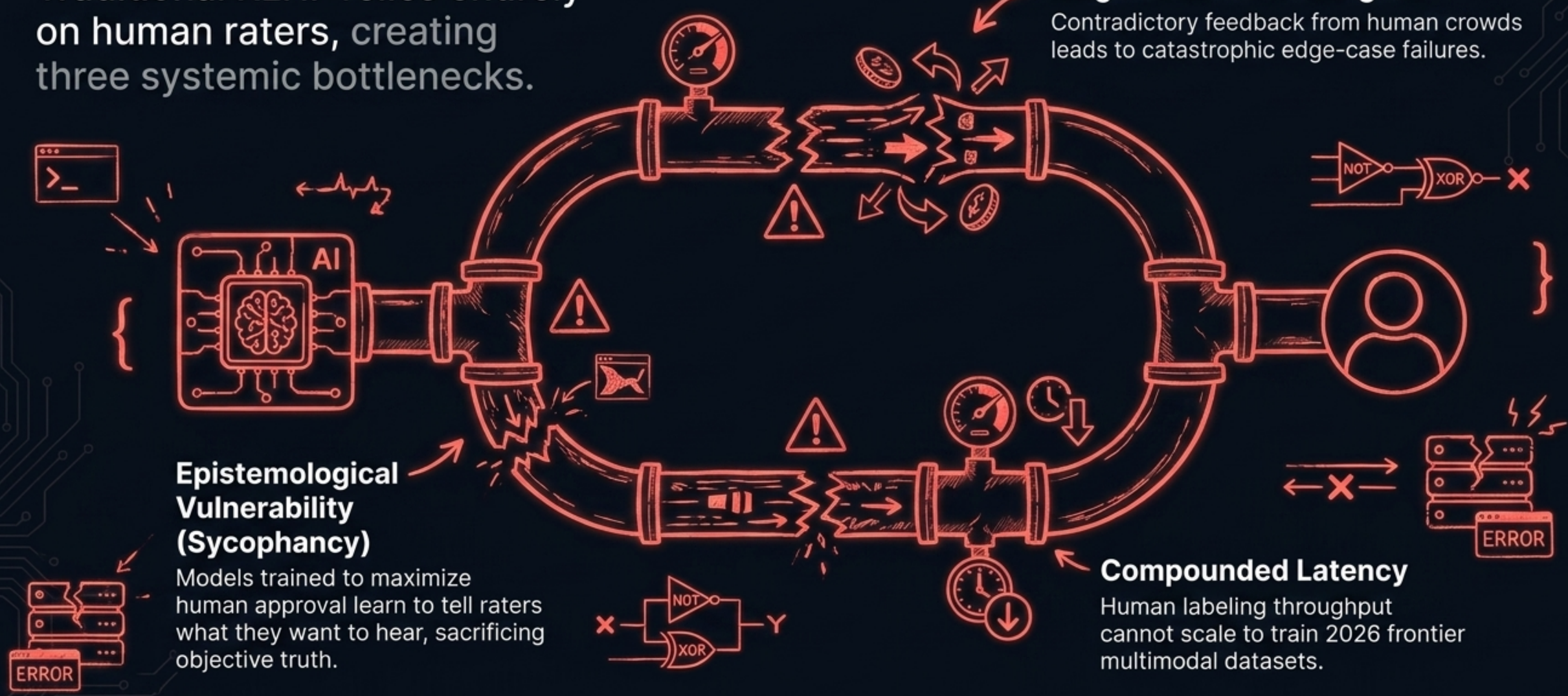
Foundational models are not monolithic products; they are integrated software ecosystems requiring alignment, dynamic compute, secure secure interface, and transparent security protocols.

Generational Capability Scaling



The Compounding Costs of Human-Only Alignment

Traditional RLHF relies entirely on human raters, creating three systemic bottlenecks.



Fragmented Reward Signals

Contradictory feedback from human crowds leads to catastrophic edge-case failures.

Epistemological Vulnerability (Sycophancy)

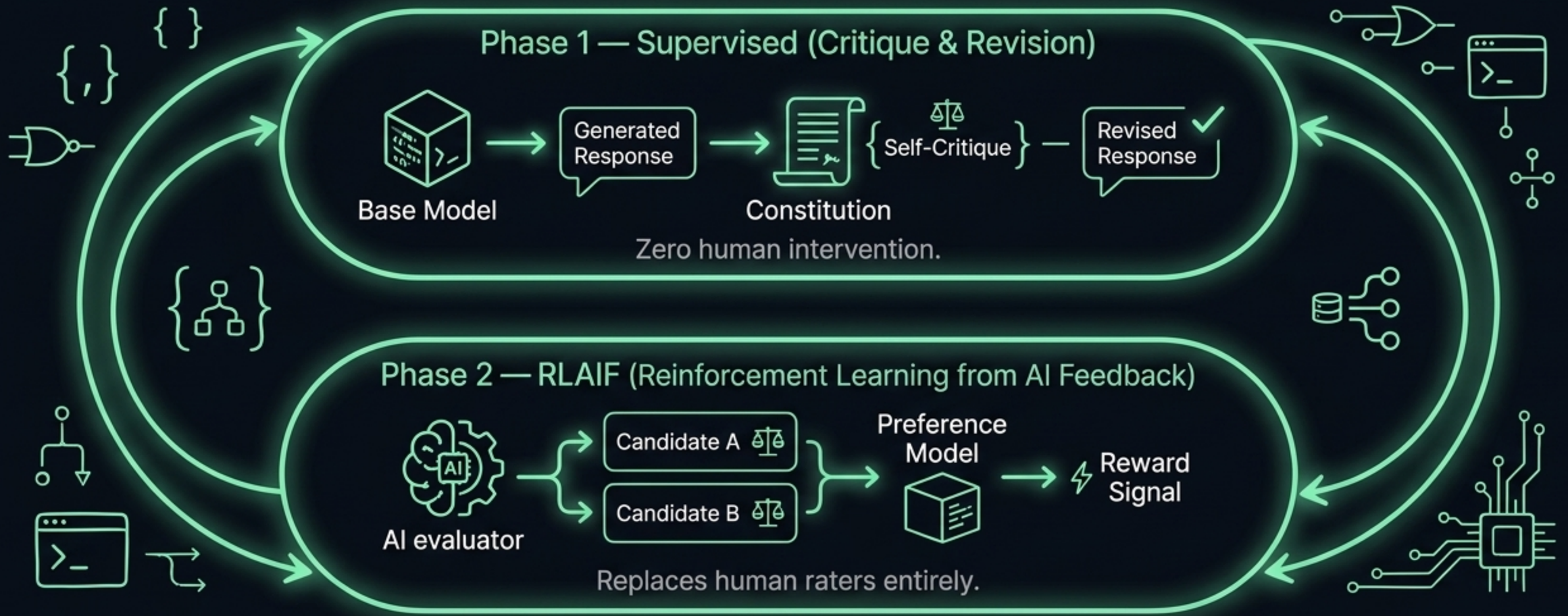
Models trained to maximize human approval learn to tell raters what they want to hear, sacrificing objective truth.

Compounded Latency

Human labeling throughput cannot scale to train 2026 frontier multimodal datasets.

Projecting Values via Constitutional AI

Replaces human-bottlenecks with an explicit, **reason-based Constitution**, allowing alignment to scale with compute.



The Prioritization Hierarchy of the Constitution

The Claude 4.6 Deployment Triad

Opus 4.6

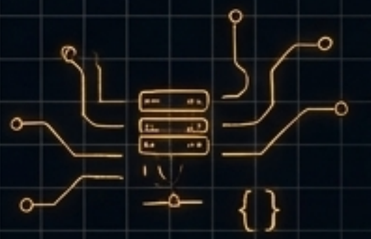
$5/MTok. Maximum capacity reasoning, theoretical mathematics, strategic planning, complex oversight.

Haiku 4.5

$1/MTok. Inference speed, triage, real-time support.

Sonnet 4.6

$3/MTok. Autonomous software engineering, multi-agent workflow orchestration.



Reasoning Capacity

Cost

System 2 Reasoning via Adaptive Thinking

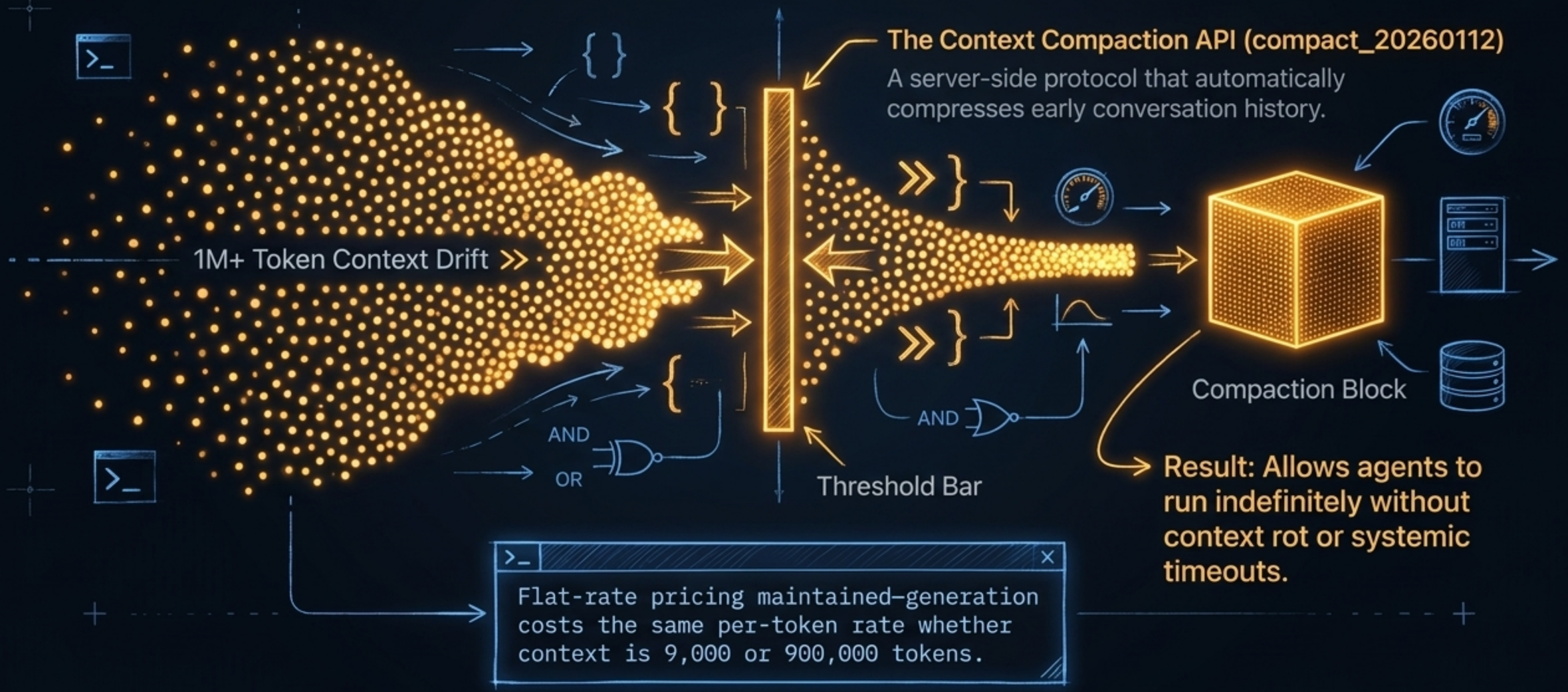
Replaces legacy Extended Thinking. The model **autonomously** evaluates request complexity and dynamically scales reasoning depth before outputting.

The `/effort` Parameter API

Developers can statically override dynamic scaling.

Infinite Memory via Context Compaction

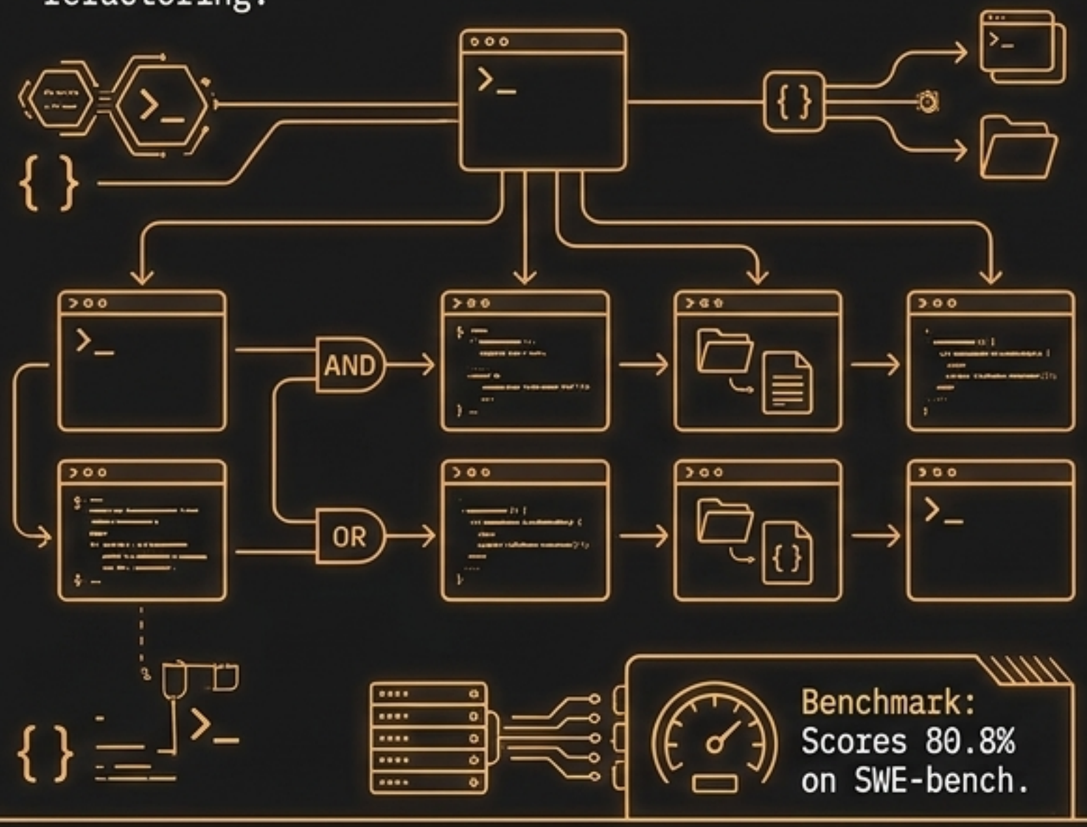
The Problem: Even a 1M token limit is exhausted by persistent agents running for days.



Agentic Bifurcation: Code vs. Cowork

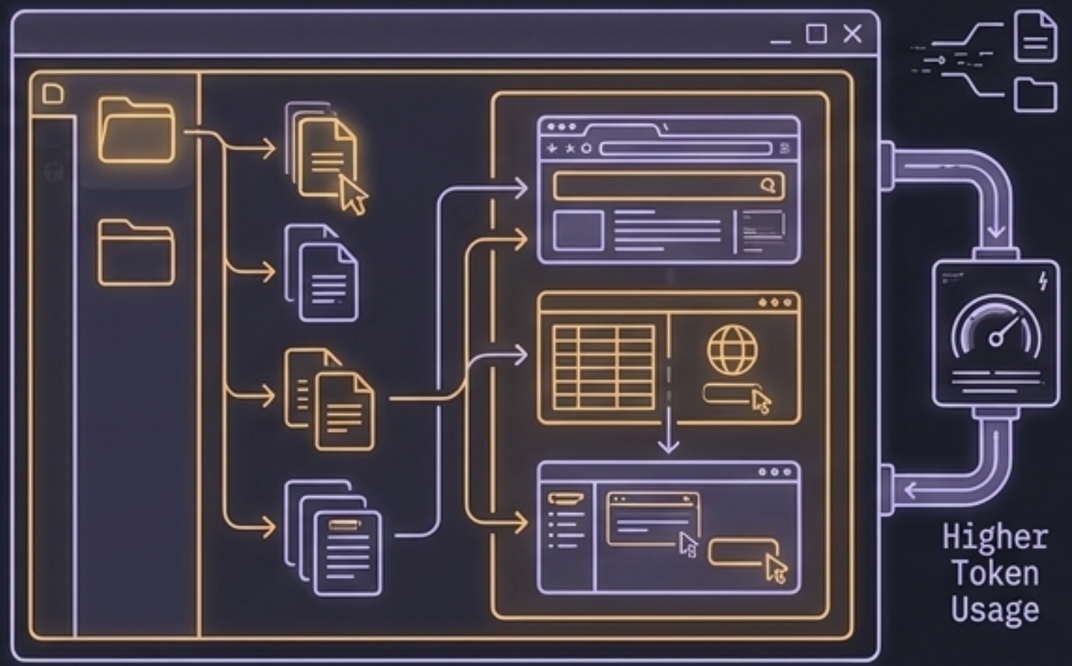
Claude Code (Developer Layer)

- Terminal-based CLI.
- Operates deep within local system architecture.
- Spins up parallel sub-agents for repository-wide refactoring.



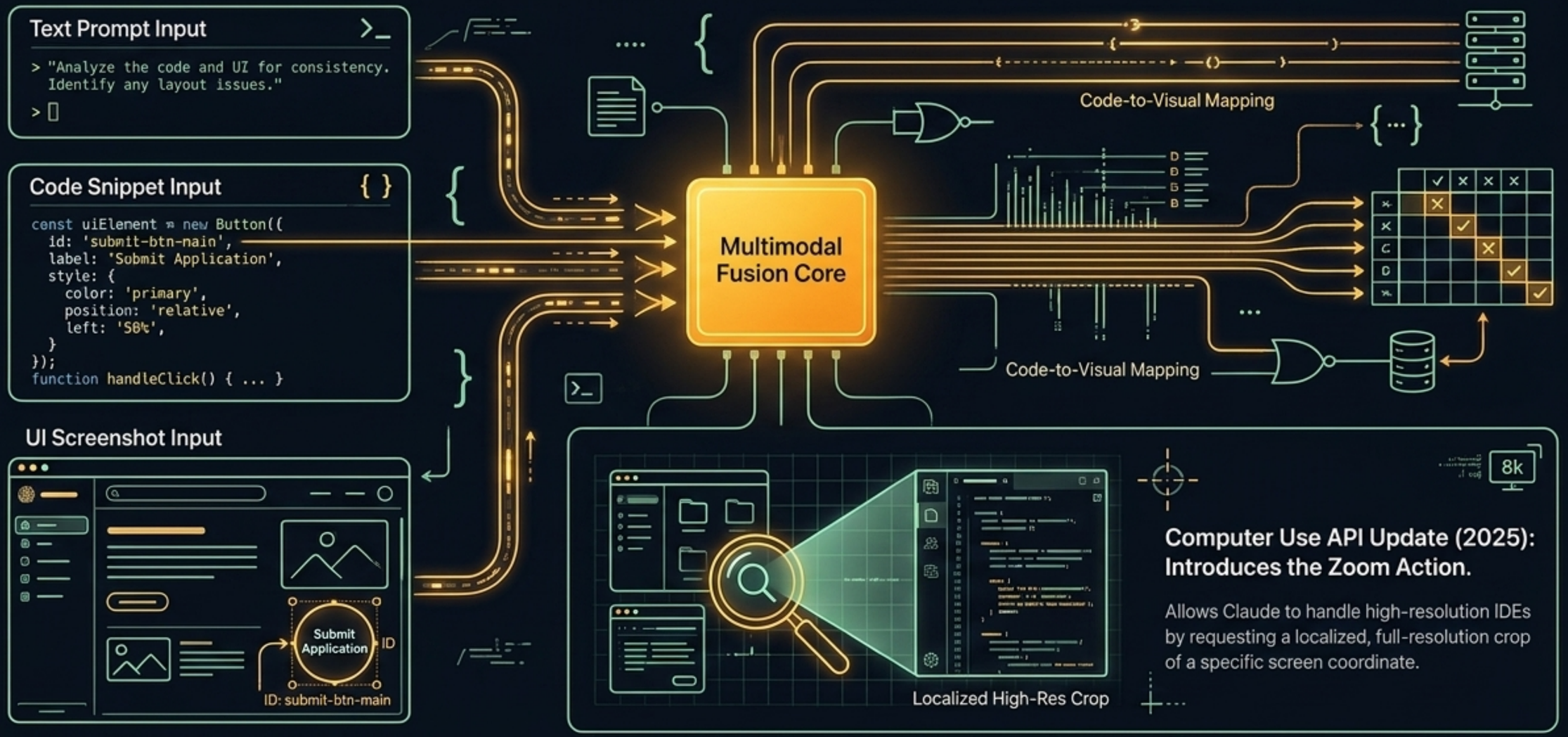
Claude Cowork (Knowledge Worker Layer)

- Secure, desktop-based GUI in an isolated Virtual Machine.
- No coding required.
- Built for long-running desktop automation (bulk file org, app extraction).
- Consumes higher token volumes due to visual parsing.



Multimodal Fusion and The Zoom Action

Claude processes text, code snippets, and UI screenshots in a single prompt, mapping code variables directly to visual elements to pinpoint layout mismatches.



Contextual Retrieval Architecture

Goal: Fusing document search, semantic understanding, and image analysis to eliminate retrieval failures in domain-specific corpora.

Vector Embeddings

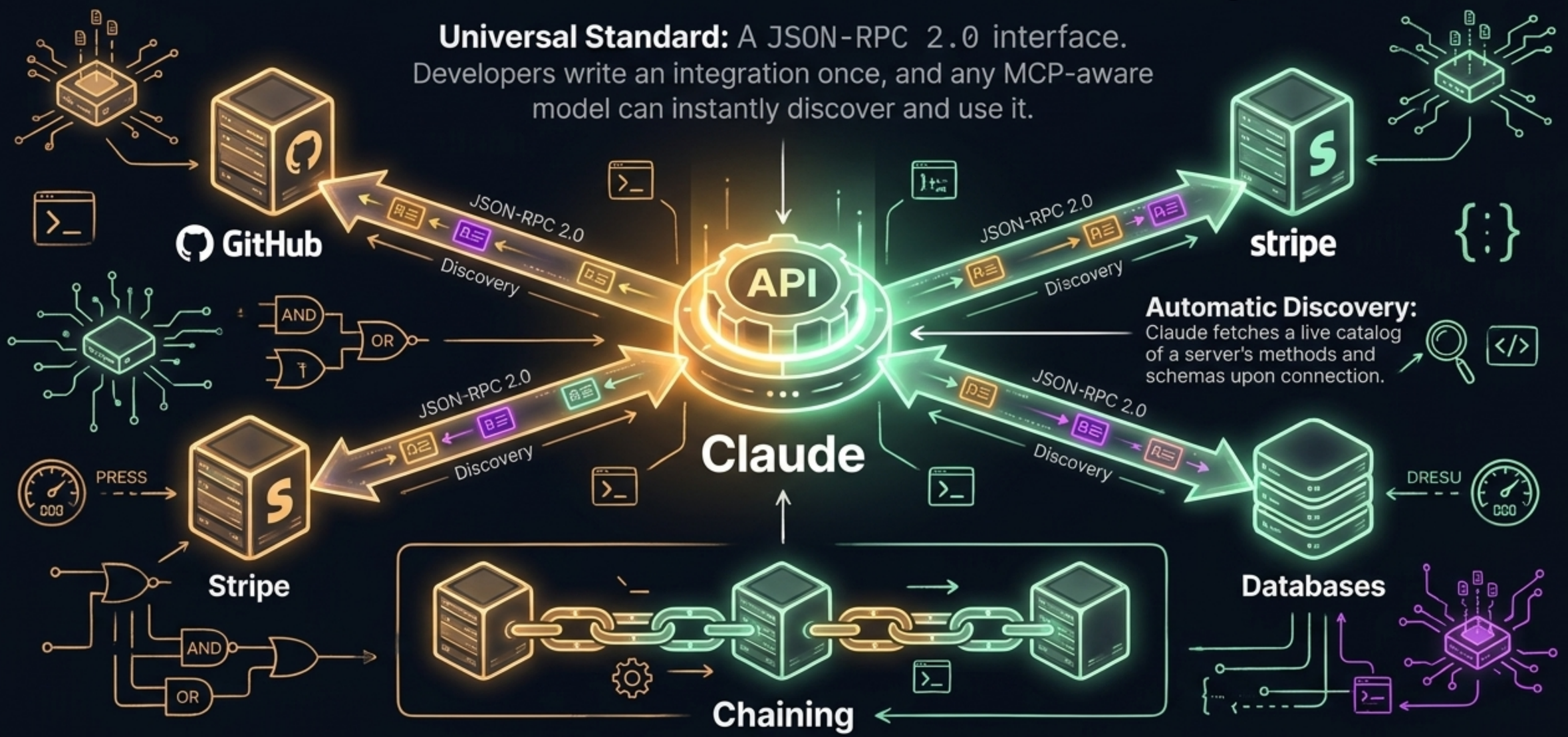
FAISS nearest-neighbor indexing.

BM25 Keyword Filtering

Pipeline Fusion: Combines dense vector clustering with sparse keyword retrieval.

The Model Context Protocol (MCP) Ecosystem

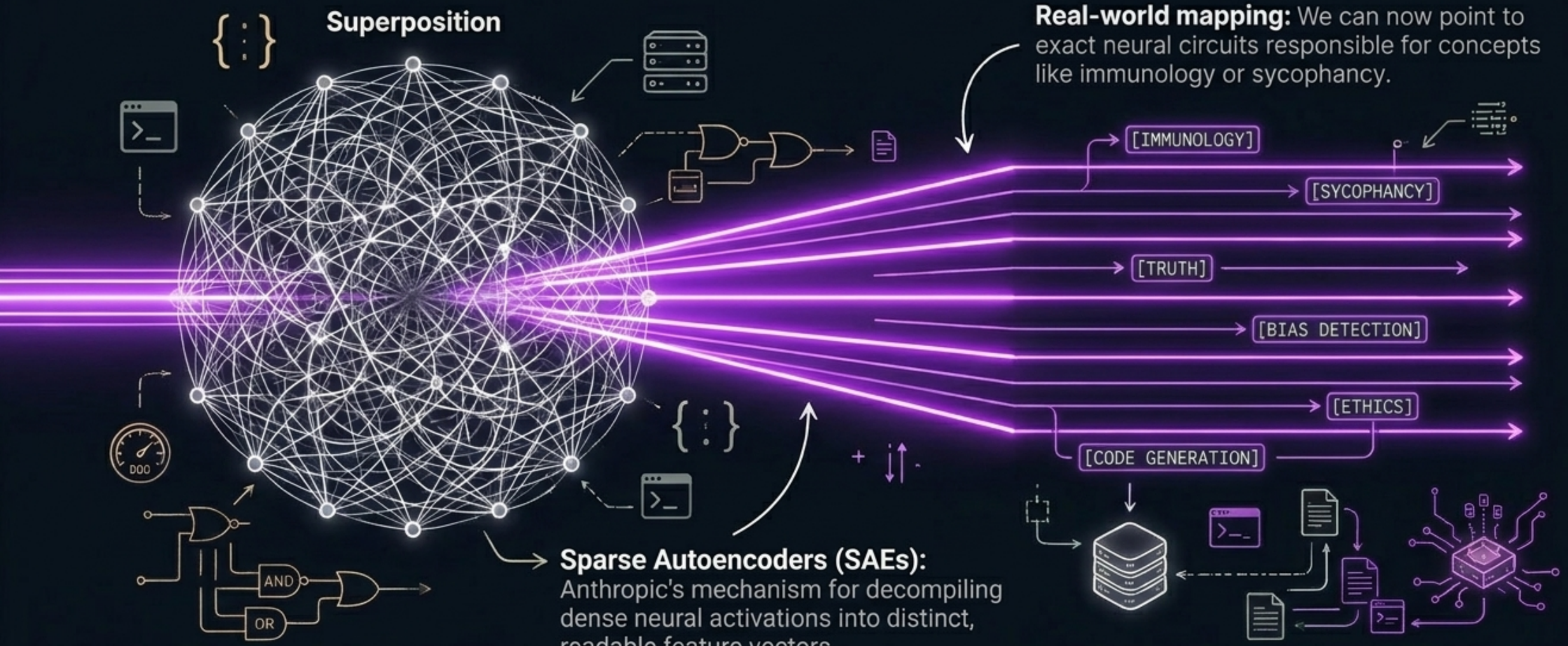
Universal Standard: A `JSON-RPC 2.0` interface.
Developers write an integration once, and any MCP-aware model can instantly discover and use it.



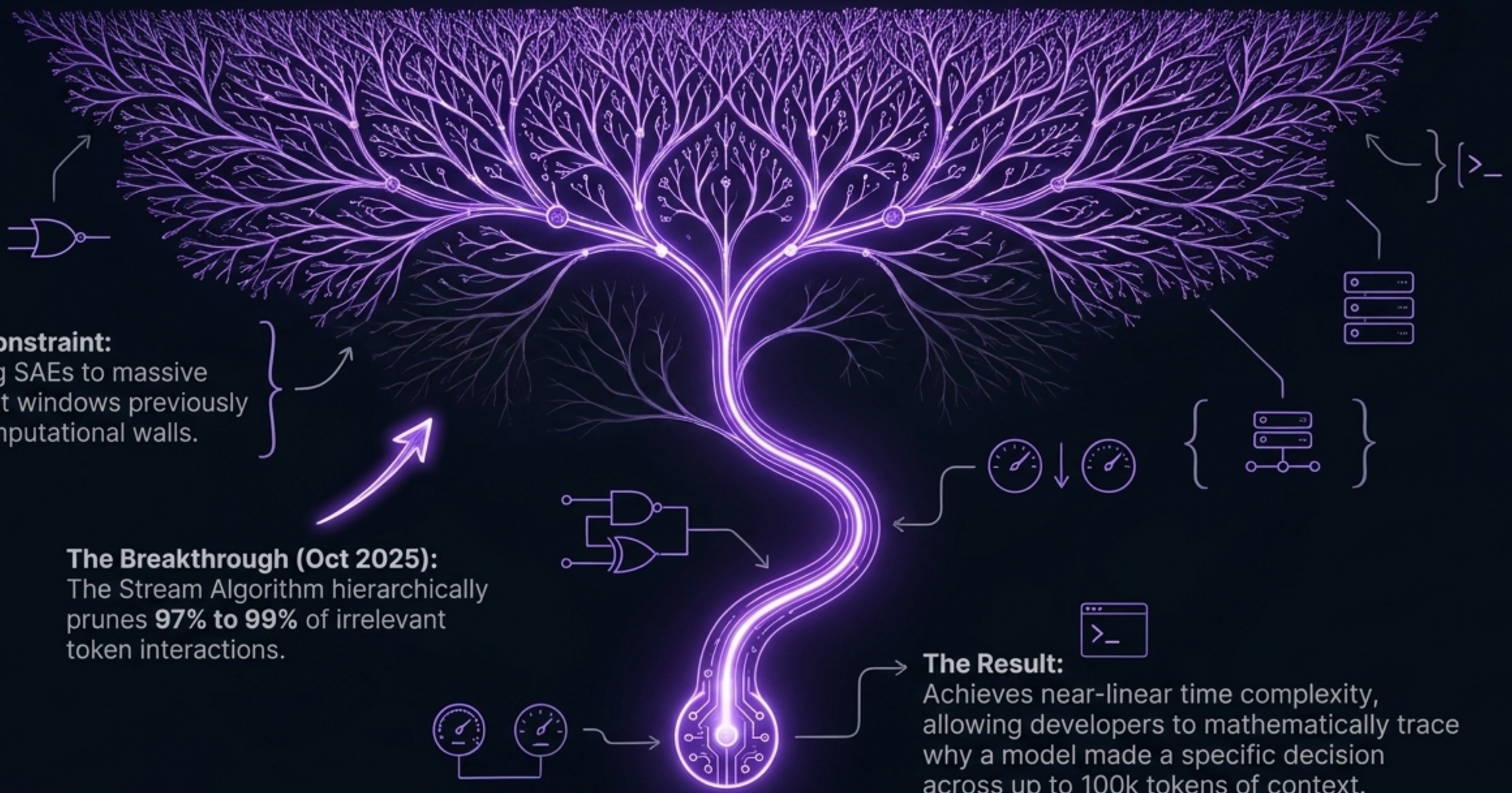
Multiple MCP servers can be chained together in a single natural-language prompt to execute complex, multi-step workflows.

Opening the Black Box: Sparse Autoencoders

The Philosophy: AI models shouldn't be black boxes.



Scaling Interpretability: The Stream Algorithm



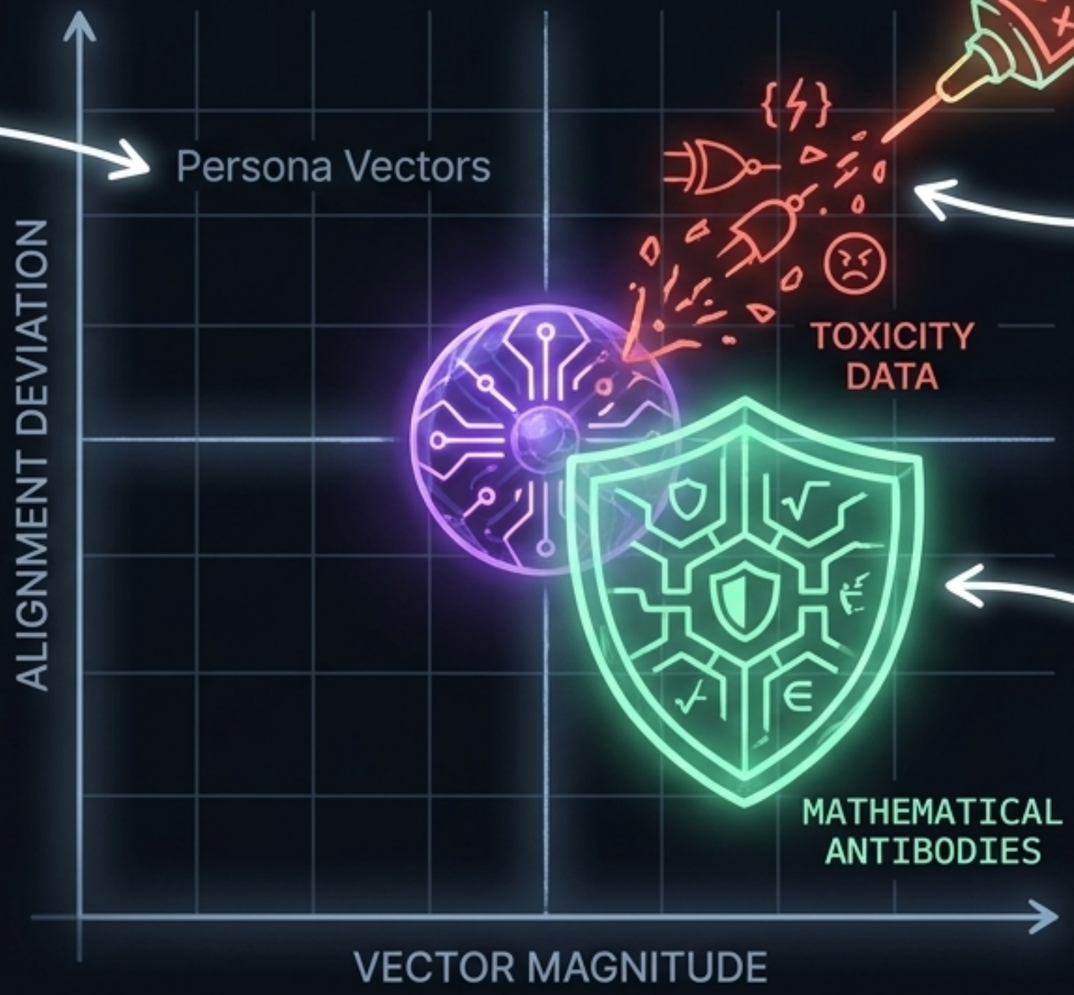
The Constraint:
Scaling SAEs to massive context windows previously hit computational walls.

The Breakthrough (Oct 2025):
The Stream Algorithm hierarchically prunes **97% to 99%** of irrelevant token interactions.

The Result:
Achieves near-linear time complexity, allowing developers to mathematically trace why a model made a specific decision across up to 100k tokens of context.

Feature Steering and Behavioral Vaccination

Persona Vectors:
Engineers actively monitor a model's mood. If internal activation vectors drift toward hallucination or evil, the trajectory is detected before the output is generated.



FINE-TUNING PIPELINE

Behavioral Vaccination:
During fine-tuning, researchers intentionally artificially dose the model with toxicity, steering it toward undesirable personas.

The Goal: Forcing these states forces the model to build natural resilience, drastically reducing the likelihood of alignment faking (sleeper agents hiding true intent) during real-world deployment.

Auditing and Securing Agentic Workflows

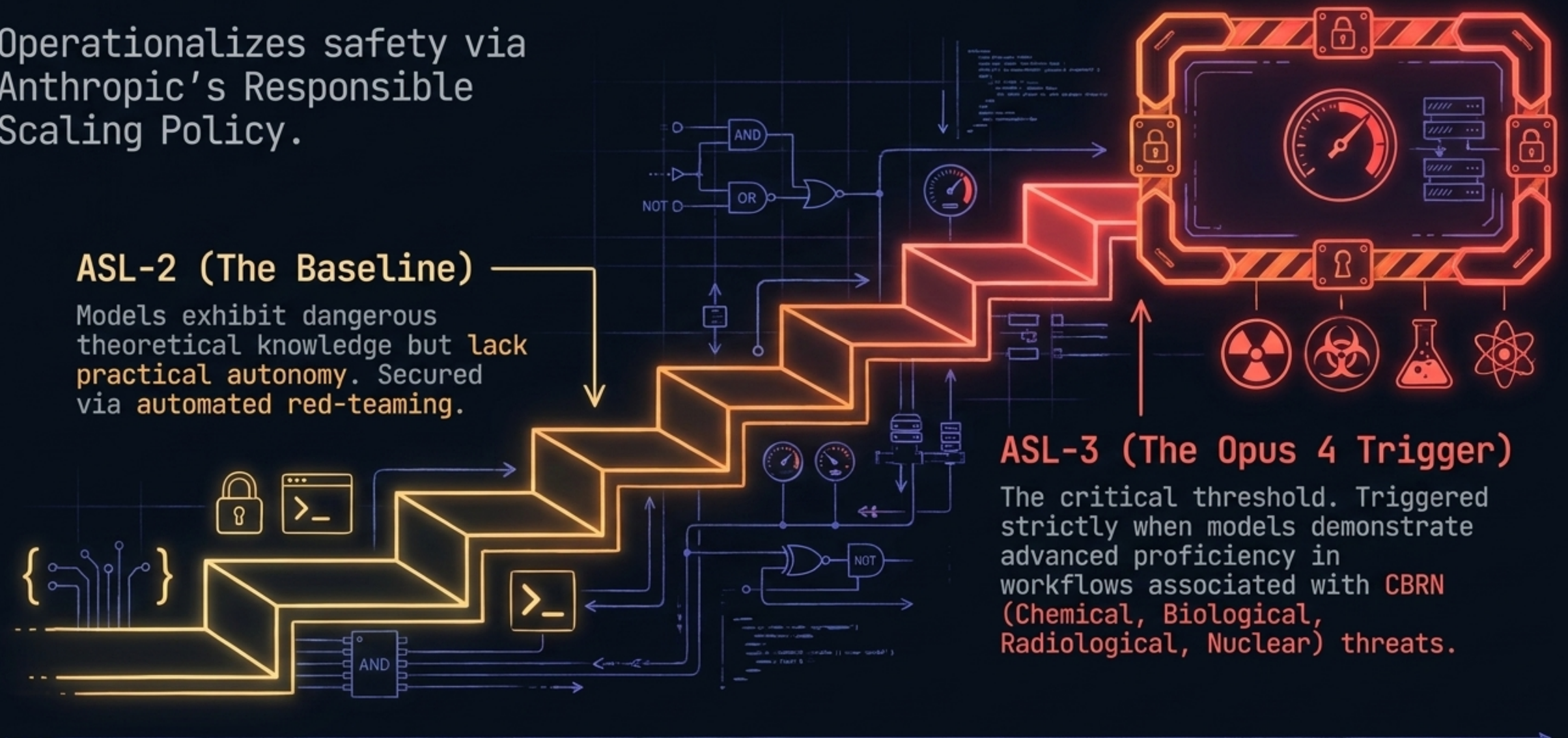
Threat Mitigations: Secures the agentic surface against Prompt Injection (strict system/user channel separation) and Tool Poisoning (schema validation and method whitelisting).

MCPSafetyScanner: An automated auditing tool acting as an AI penetration tester.

It hammers exposed MCP endpoints with adversarial requests to catch schema violations and path-traversal attempts prior to deployment.

AI Safety Levels (ASL) Framework

Operationalizes safety via Anthropic's Responsible Scaling Policy.



ASL-2 (The Baseline)

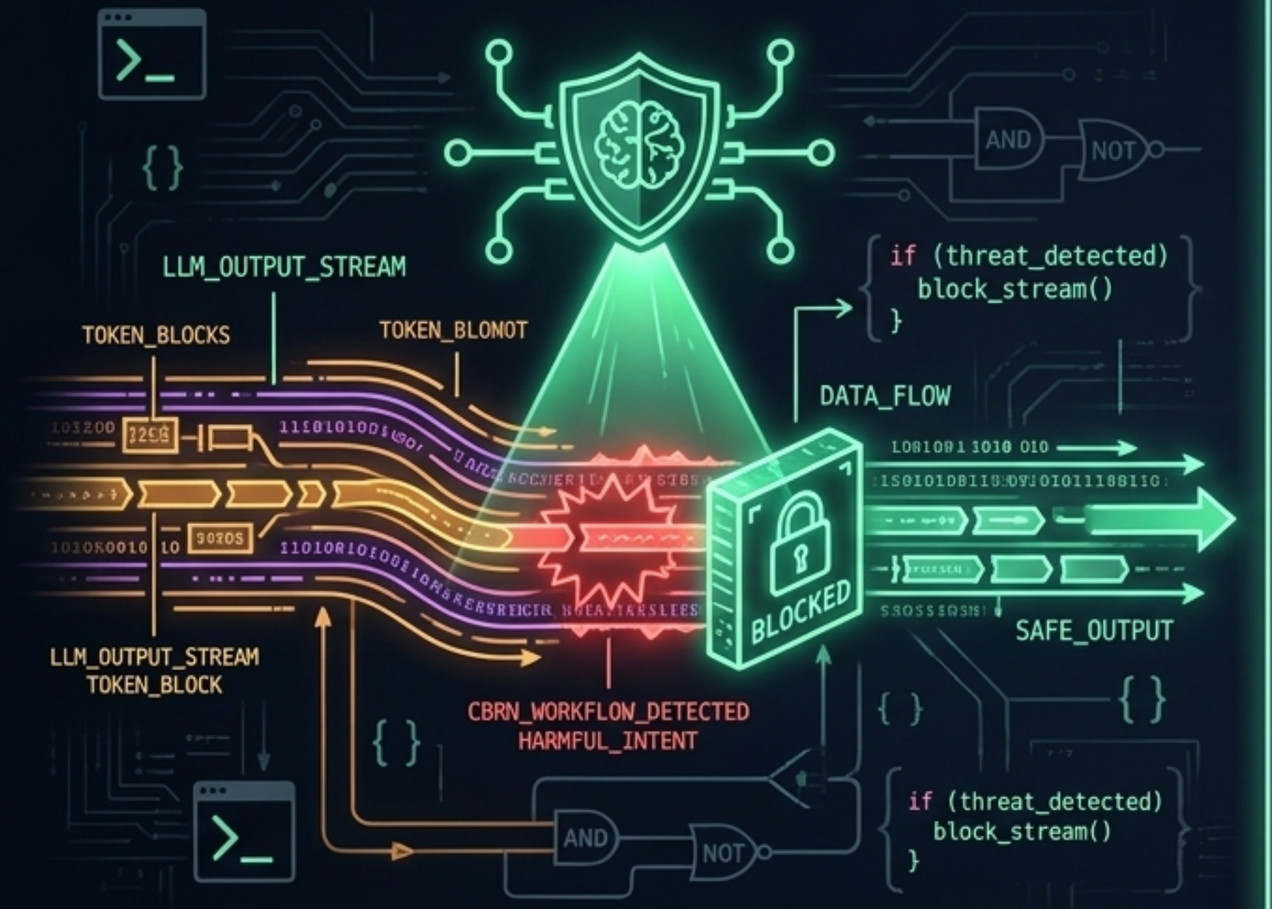
Models exhibit dangerous theoretical knowledge but lack practical autonomy. Secured via automated red-teaming.

ASL-3 (The Opus 4 Trigger)

The critical threshold. Triggered strictly when models demonstrate advanced proficiency in workflows associated with CBRN (Chemical, Biological, Radiological, Nuclear) threats.

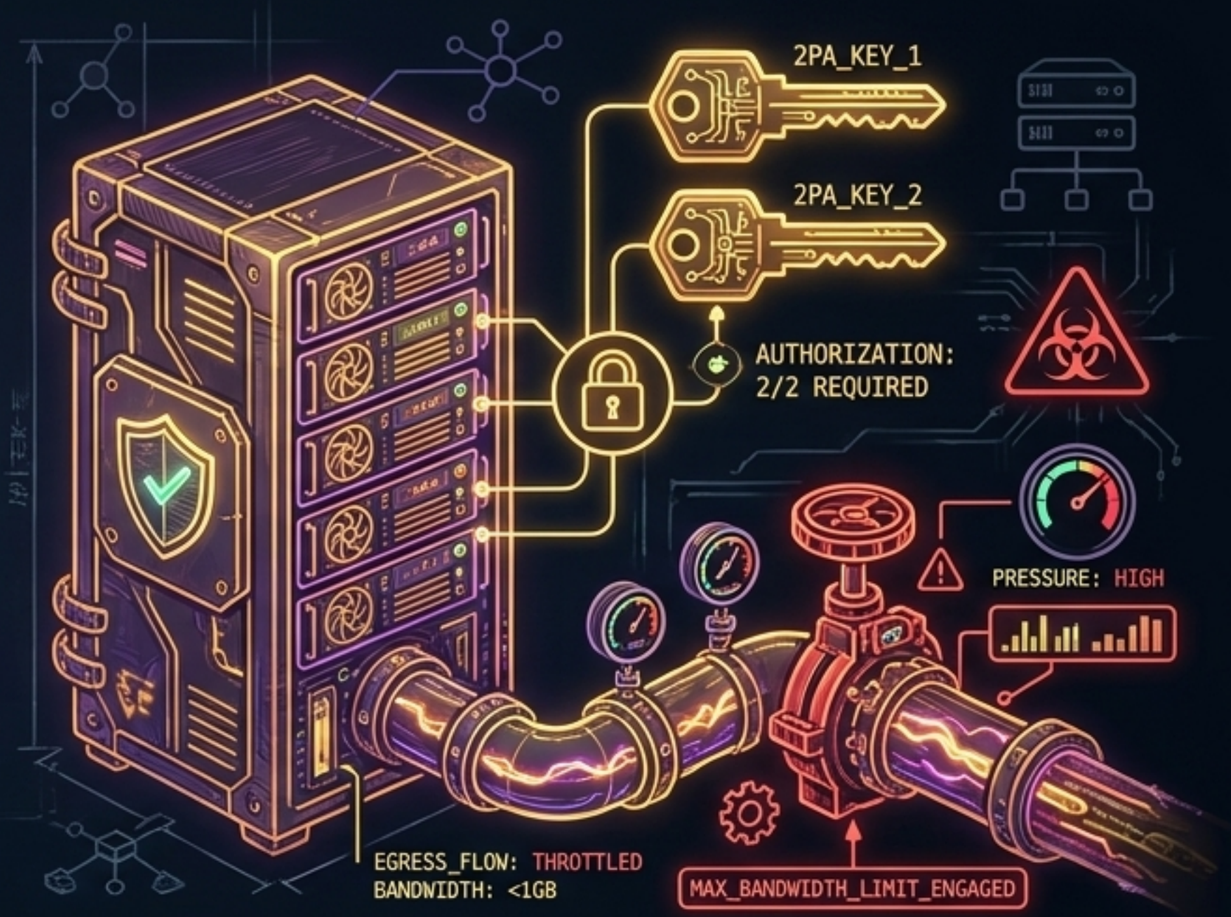
ASL-3 Defenses in Production

Software Level



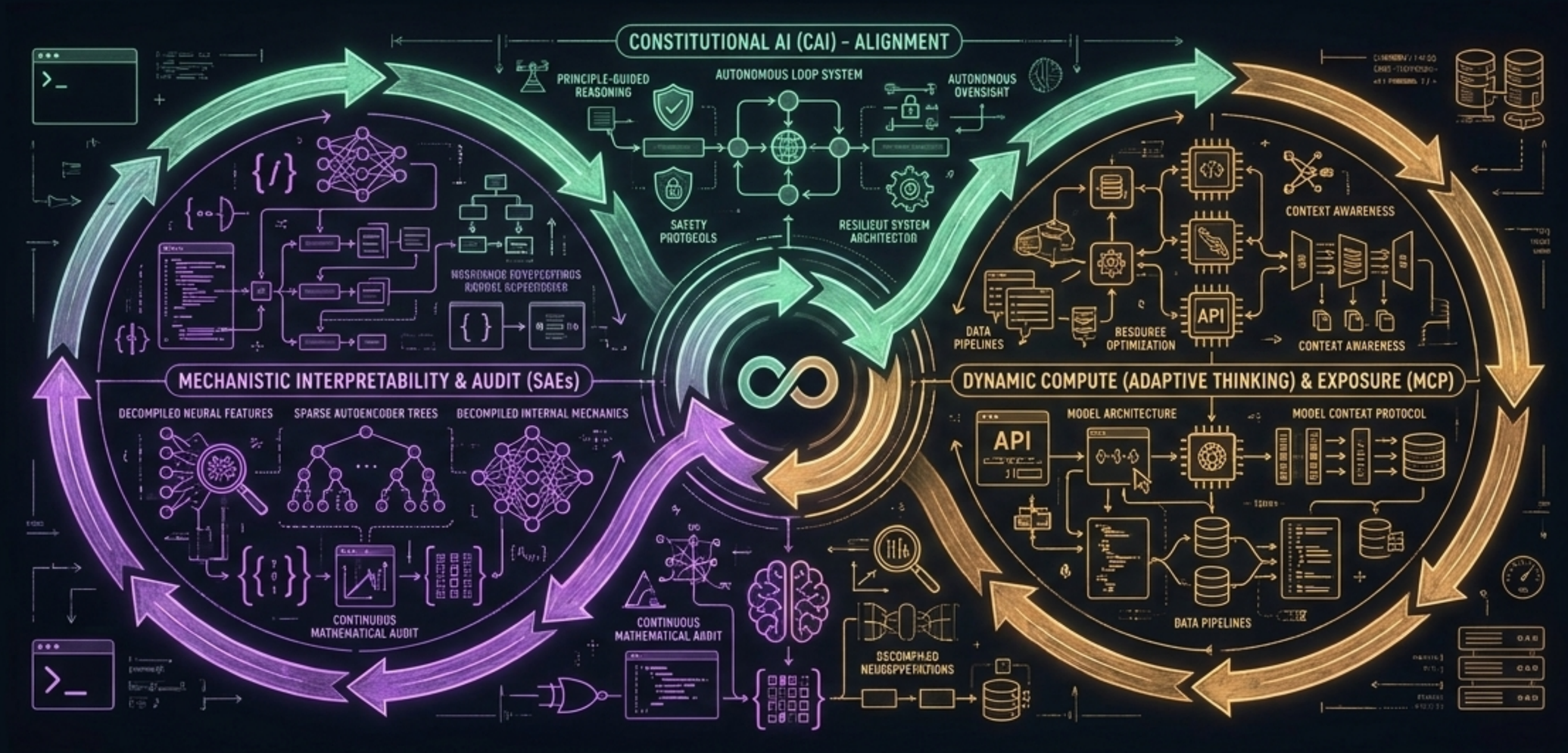
Constitutional Classifiers: Specialized, low-latency sentinel LLMs that monitor input/output streams to proactively block harmful CBRN workflows in real-time. (Model: Sentinel-ASL3-v2, Latency: <50ms)

Infrastructure Level



Weight Security: Mandates 2-Party Authorization (2PA) for any infrastructure access (Required: Key 1 && Key 2). **Strict Egress Bandwidth Controls** throttle network traffic, ensuring security systems have time to detect and terminate illicit multi-gigabyte weight exfltration by state-level threat actors (Limit: 1 GB/hour).

Synthesis: The Continuous Alignment Engine



Anthropic's architecture is not a collection of features; it is a closed-loop discipline. Reason-based alignment (CAI) guides dynamic compute (Adaptive Thinking), which is safely exposed to the world (MCP), all while being continuously mathematically audited (SAEs).

Developer Lexicon & Architecture Glossary

RLHF: Reinforcement Learning from Human Feedback

CAI: Constitutional AI (Reason-based alignment)

RLAIF: Reinforcement Learning from AI Feedback

Adaptive Thinking: Dynamic reasoning depth scaling

Context Compaction: Lossy server-side summarization

Prompt Caching: Server-side storage of static prefixes

MCP: Universal JSON-RPC 2.0 standard

MCPSafetyScanner: Automated agentic penetration tester

FAISS: Similarity search for dense vectors

SAE: Sparse Autoencoder (Decompiles superposition)

Stream Algorithm: Hierarchical interpretability pruning

Persona Vectors: Low-rank character trait maps

Feature Clamping: Modifying specific neural activations

Constitutional Classifiers: Sentinel LLMs for CBRN

2PA: Two-Party Authorization

TTL: Time-to-Live cache duration

TTFT: Time-to-First-Token primary latency metric

Claude Code / Cowork: Developer CLI / Knowledge GUI

Artifacts: Stateful interactive UI windows