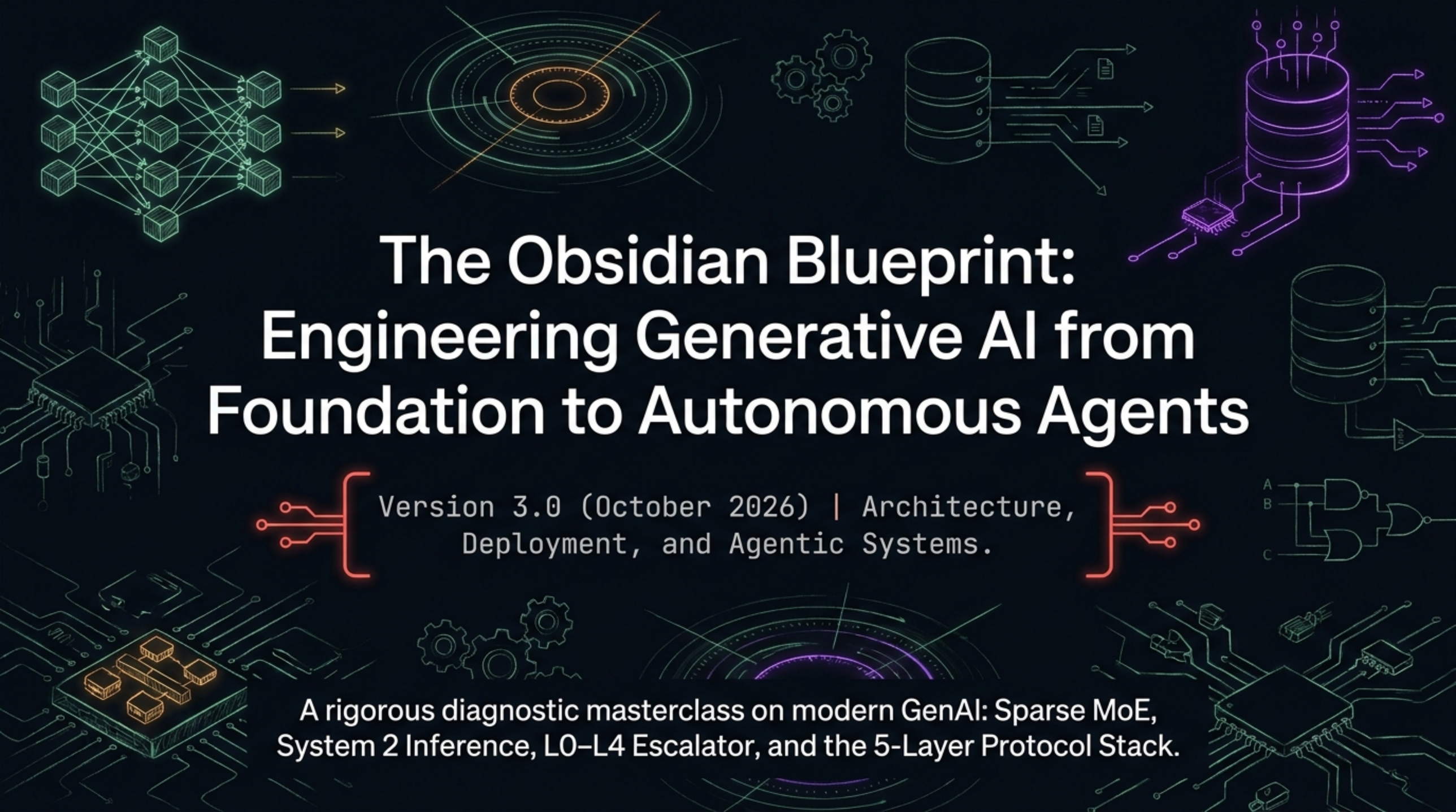


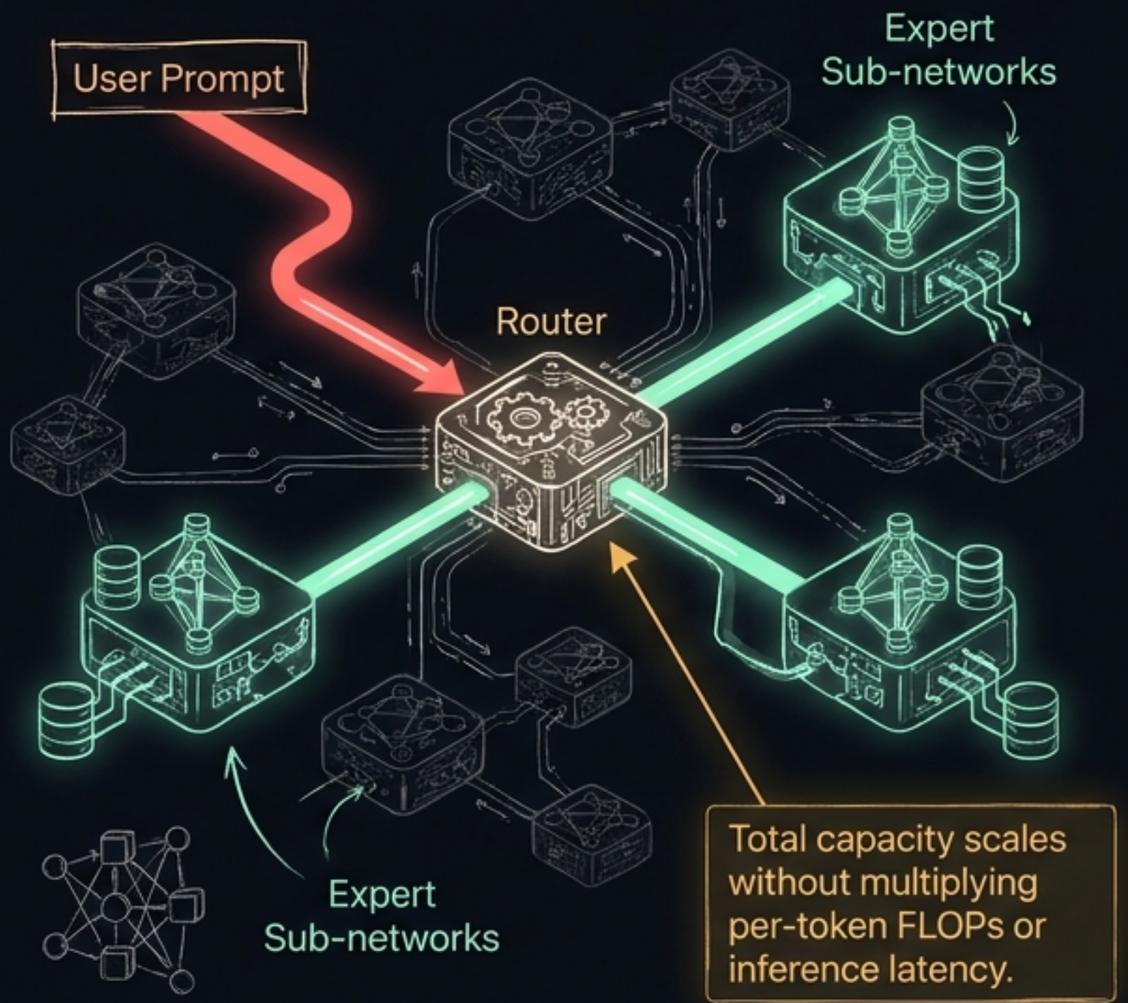
The Obsidian Blueprint: Engineering Generative AI from Foundation to Autonomous Agents

Version 3.0 (October 2026) | Architecture, Deployment, and Agentic Systems.

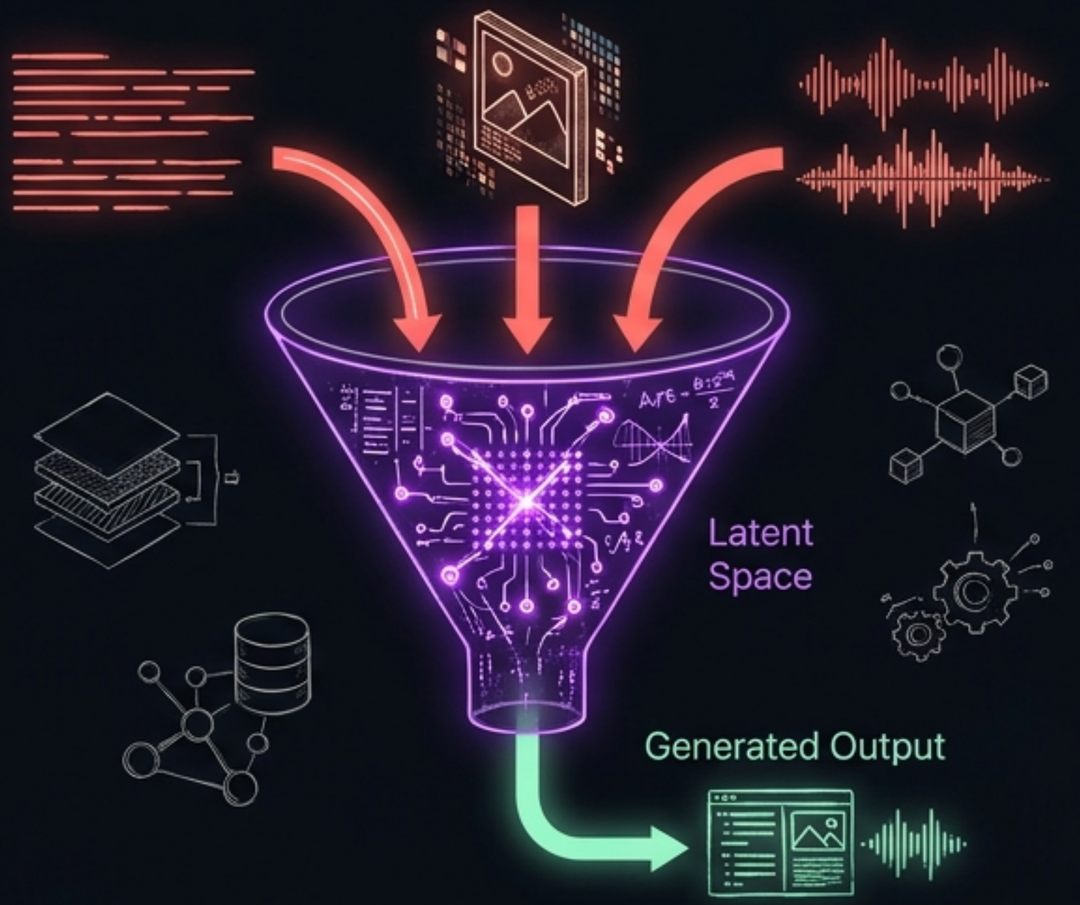
A rigorous diagnostic masterclass on modern GenAI: Sparse MoE, System 2 Inference, L0–L4 Escalator, and the 5-Layer Protocol Stack.

Modern foundation models rely on dynamic routing and native early-fusion.

Sparse MoE

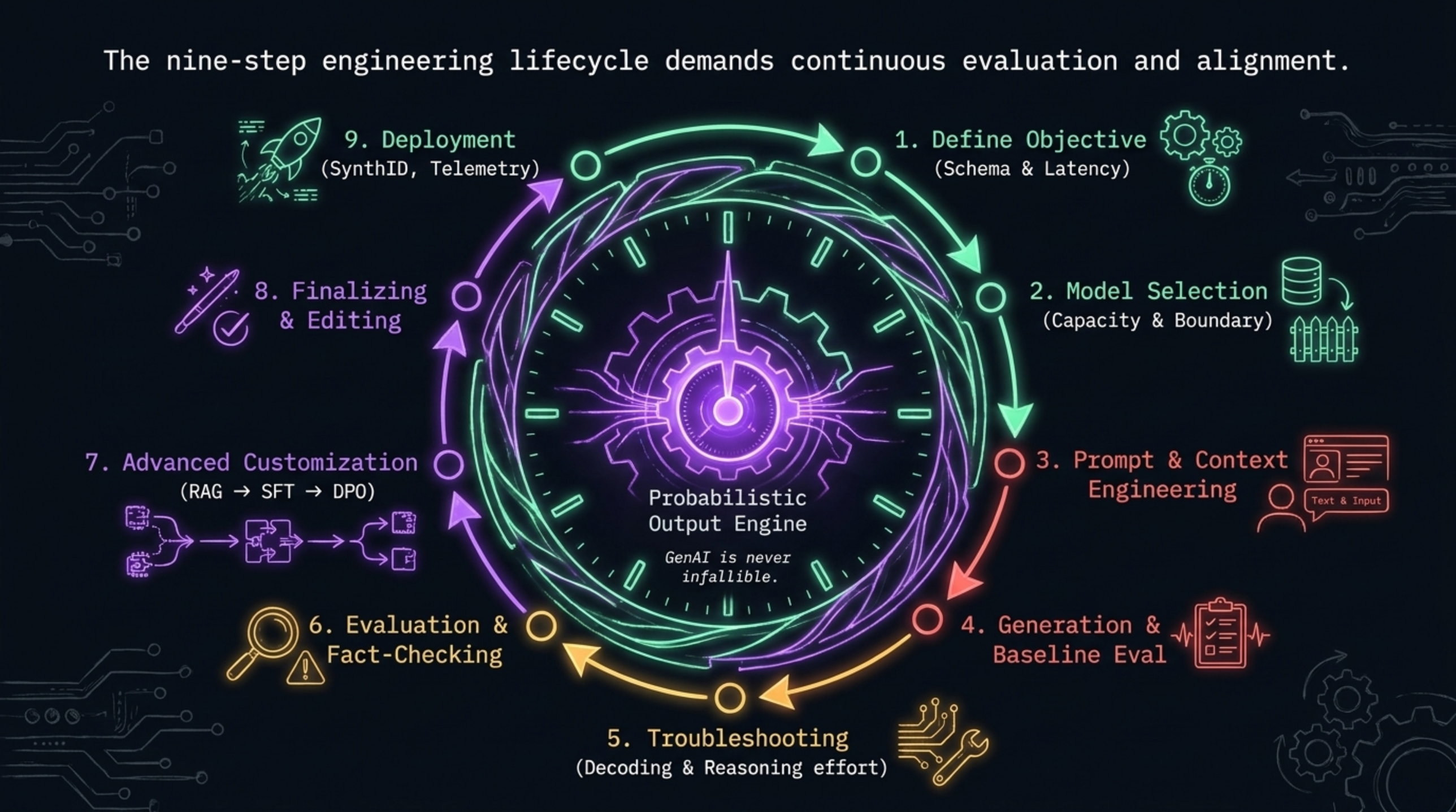


Native Multimodality

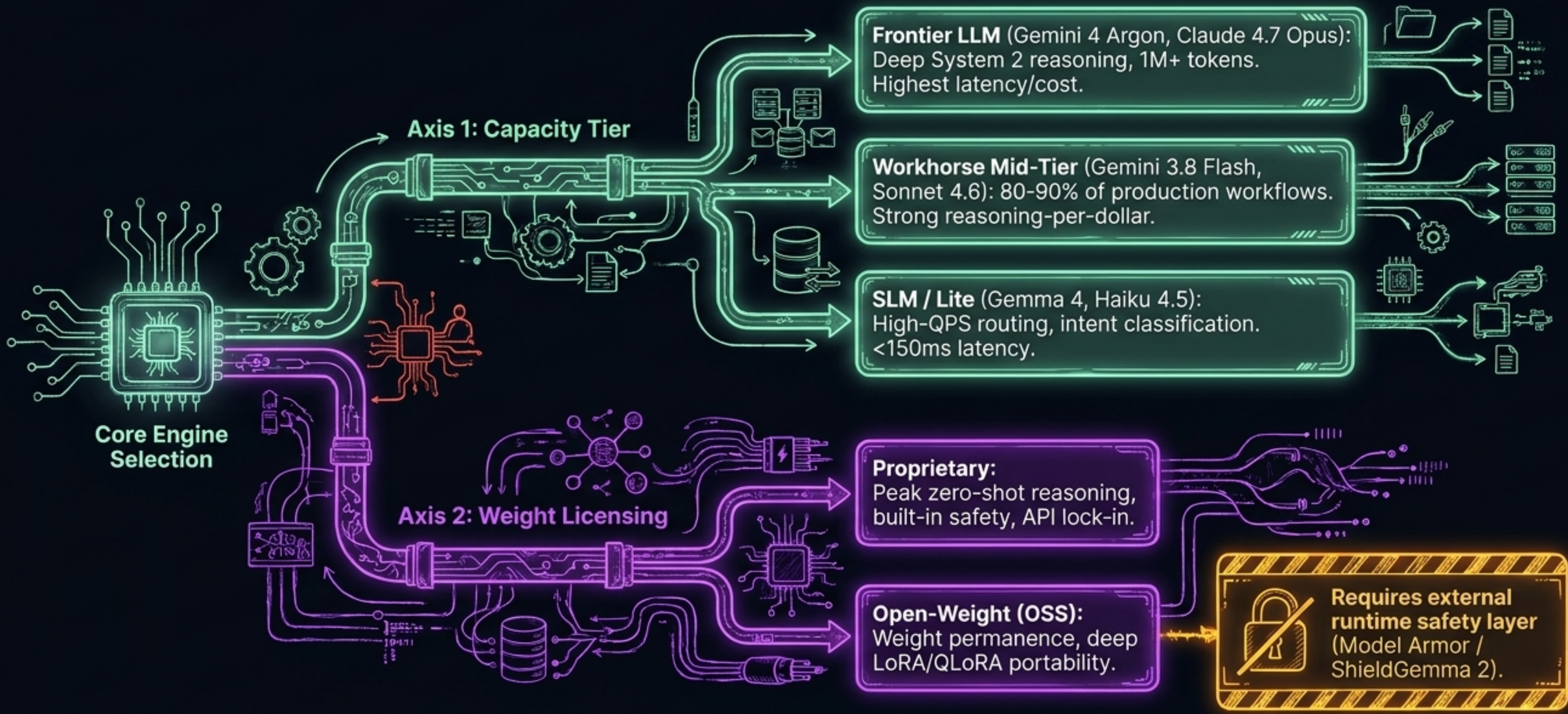


Native early-fusion preserves visual layout and spatial relationships without OCR/ASR information loss.

The nine-step engineering lifecycle demands continuous evaluation and alignment.



Selecting the core engine requires separating capacity tiers from licensing parameters.



Consumption models dictate dictate operational overhead and billing geometry.

Exam Trap: Open-Weight ≠ Self-Hosted GPU Ops. Always evaluate serverless MaaS endpoints before paying the 24/7 idle node tax.

Serverless MaaS/SaaS



Pay-as-you-go per token.
Zero infra ops.
Autoscales to zero.
Spiky traffic.

Dedicated VPC



Per node-hour billing.
Bills 24/7 even at 0 QPS.
Strict multi-tenant prohibition.

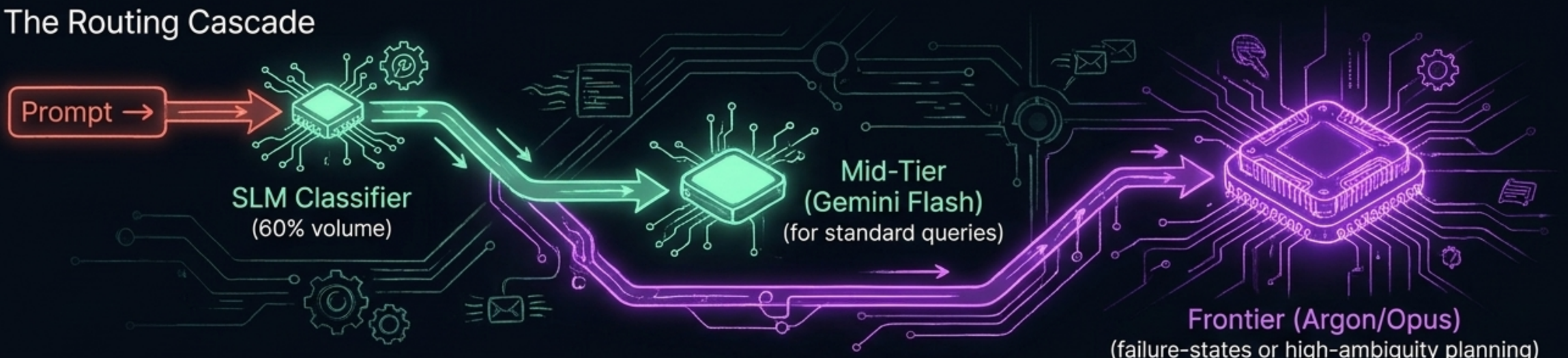
Bare Metal / GKE



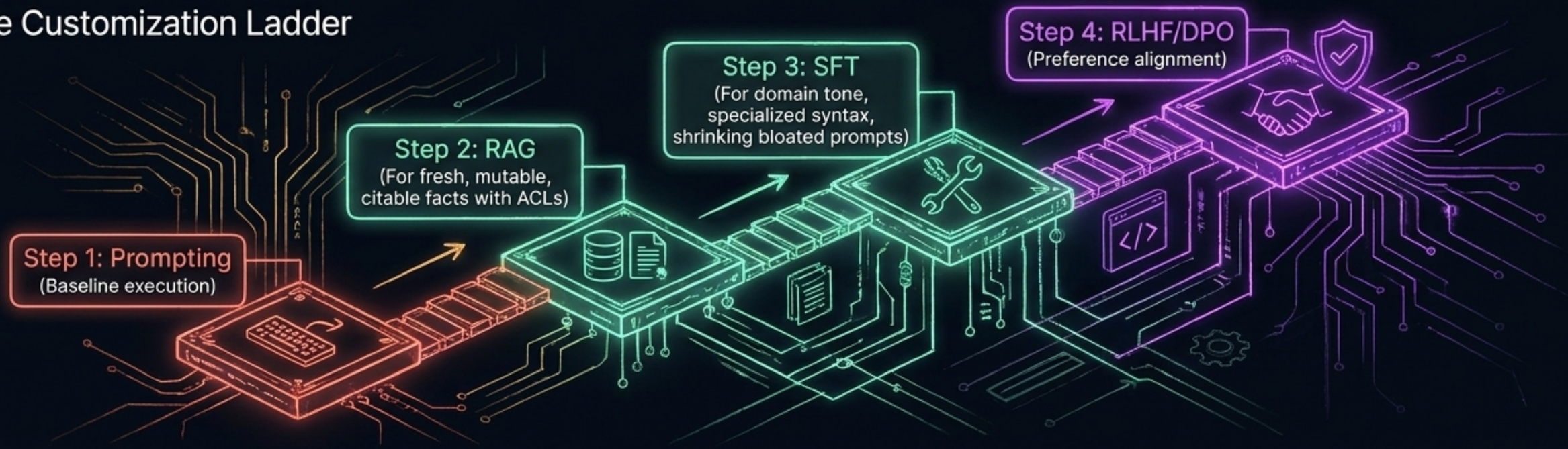
Cluster compute + ops.
Multi-LoRA bin-packing.
Air-gapped environments.

Orchestrate models through routing cascades and deliberate customization paths.

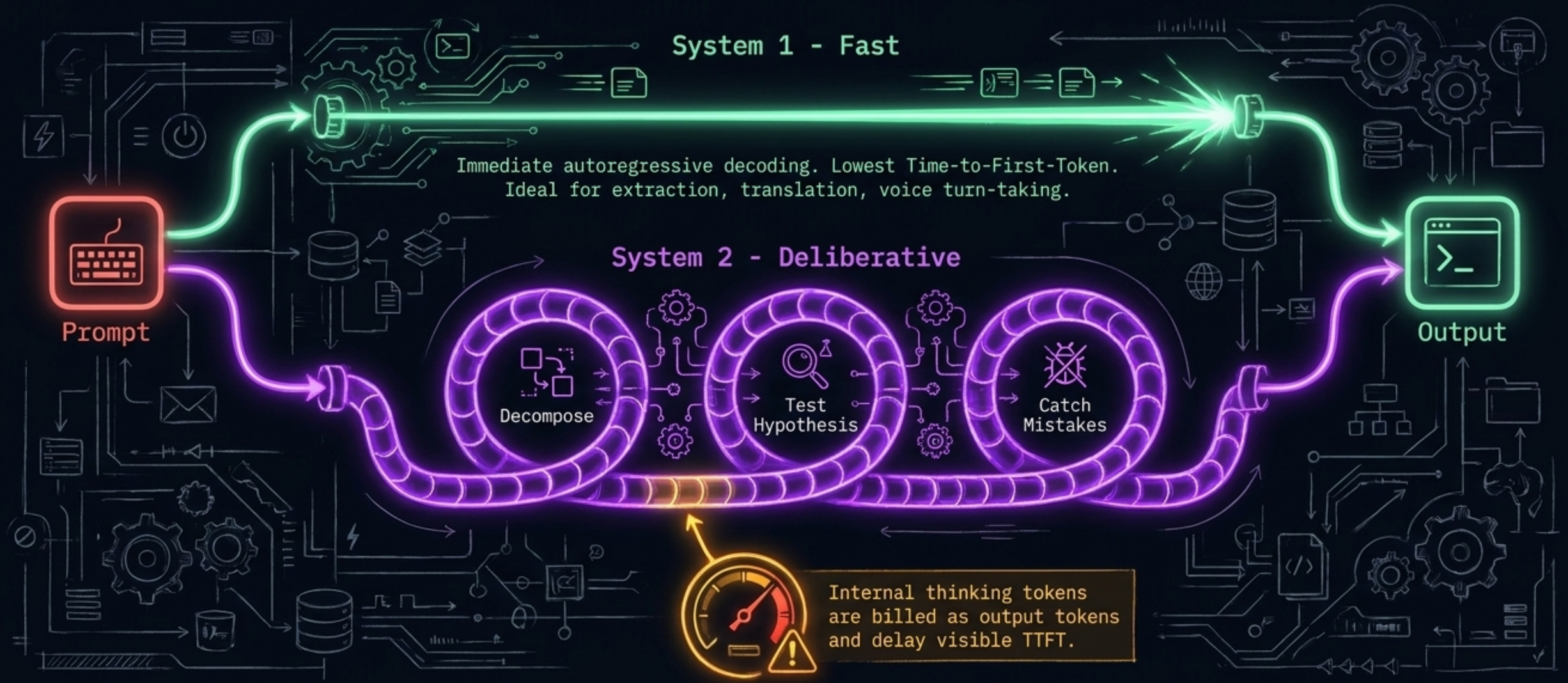
The Routing Cascade



The Customization Ladder



Inference-time compute replaces rapid heuristics with internal deliberative reasoning.



Calibrating reasoning effort requires specific vendor configuration parameters.

Google Gemini

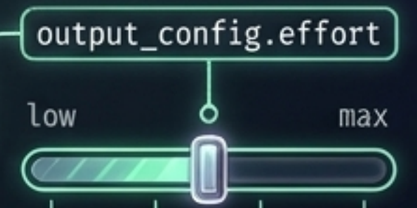


Emits encrypted `thought_signatures` alongside tool calls. Agent harness must pass them back verbatim to preserve multi-step state.

- `gemini-3.8-live-extended-thinking` decouples voice from deep reasoning.

Anthropic Claude

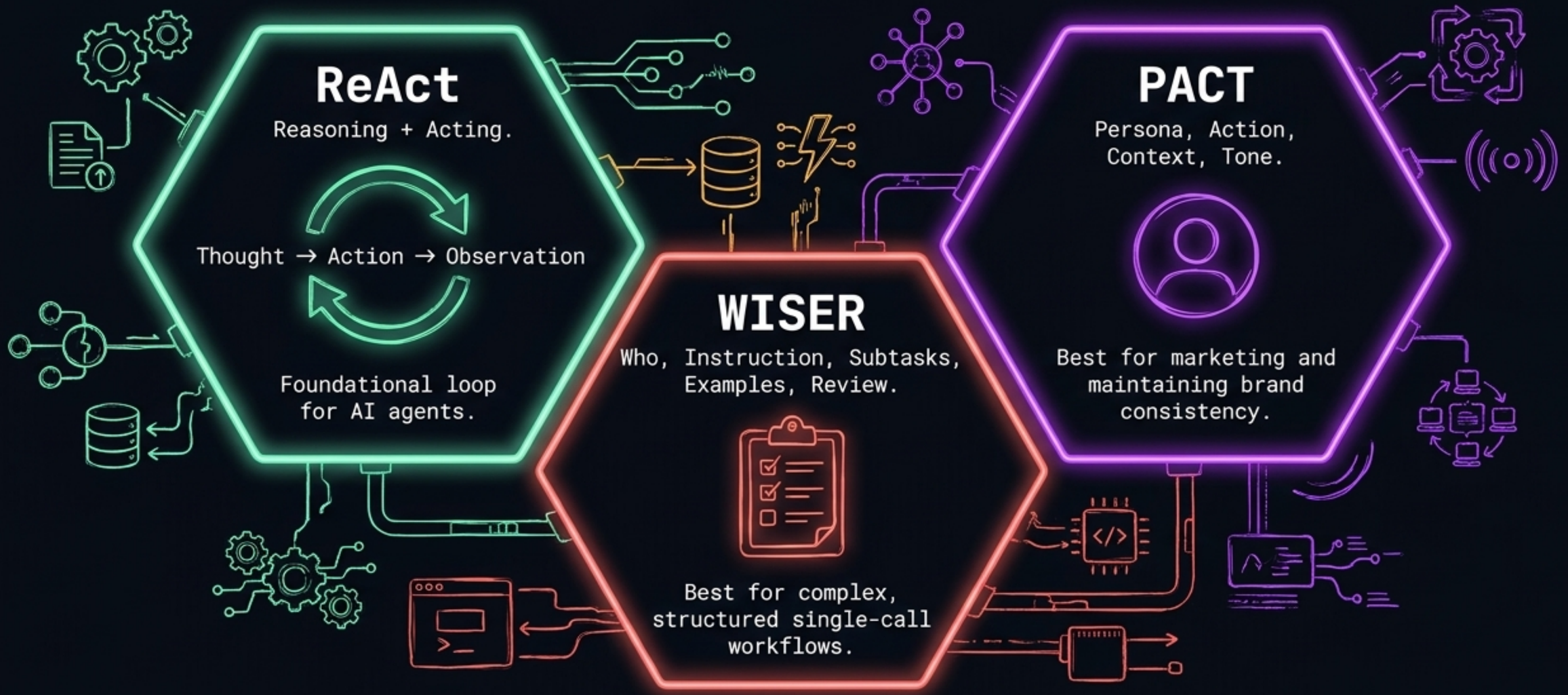
```
{
  "thinking": {
    "type": "adaptive"
  }
}
```



Native interleaved thinking between tool calls. Legacy manual budgets deprecated on 4.6.

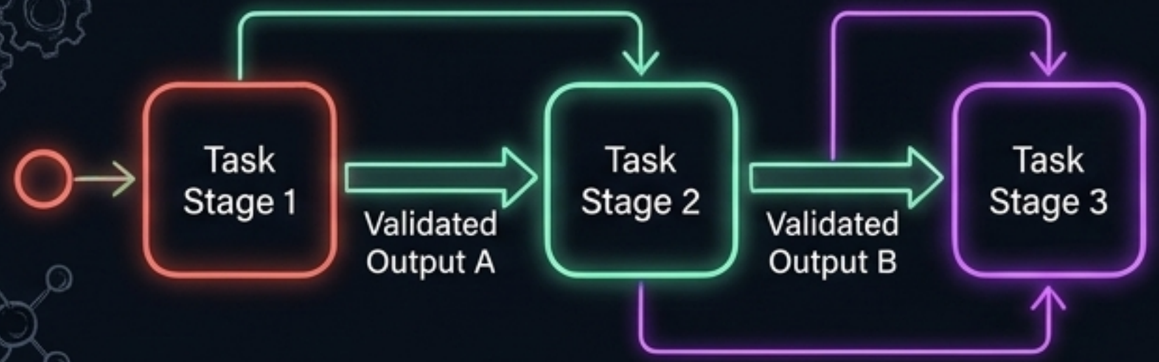
- Custom `temperature` and `assistant prefilling` are removed when active.

Replace ad-hoc instruction with structured, reusable prompting frameworks.



Advanced execution relies on chaining, self-reflection, and iterative layering.

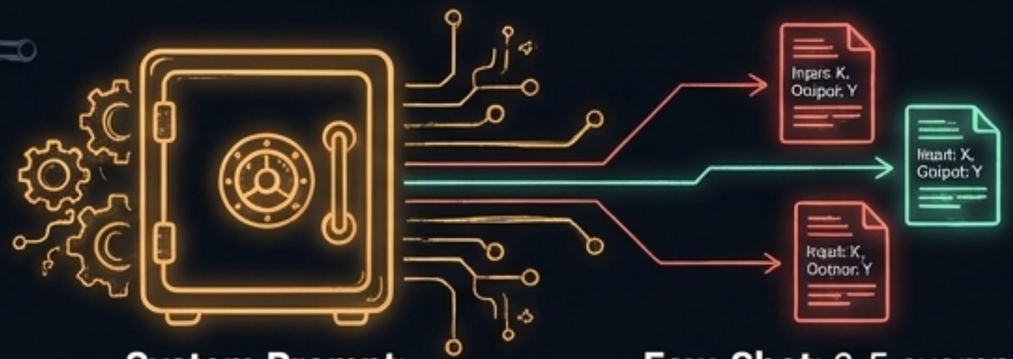
Prompt Chaining: Breaking tasks down where validated output feeds the next stage.



Meta-Prompts: Prompts that inspect, critique, or generate other prompts (e.g., “Propose a tighter system prompt”).



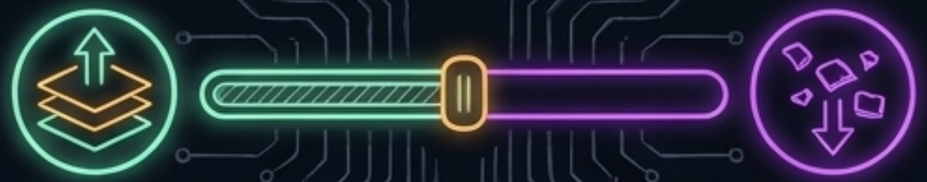
Few-Shot & System Prompts: invariants/boundaries



System Prompt: invariants/boundaries

Few-Shot: 2-5 examples anchoring edge cases

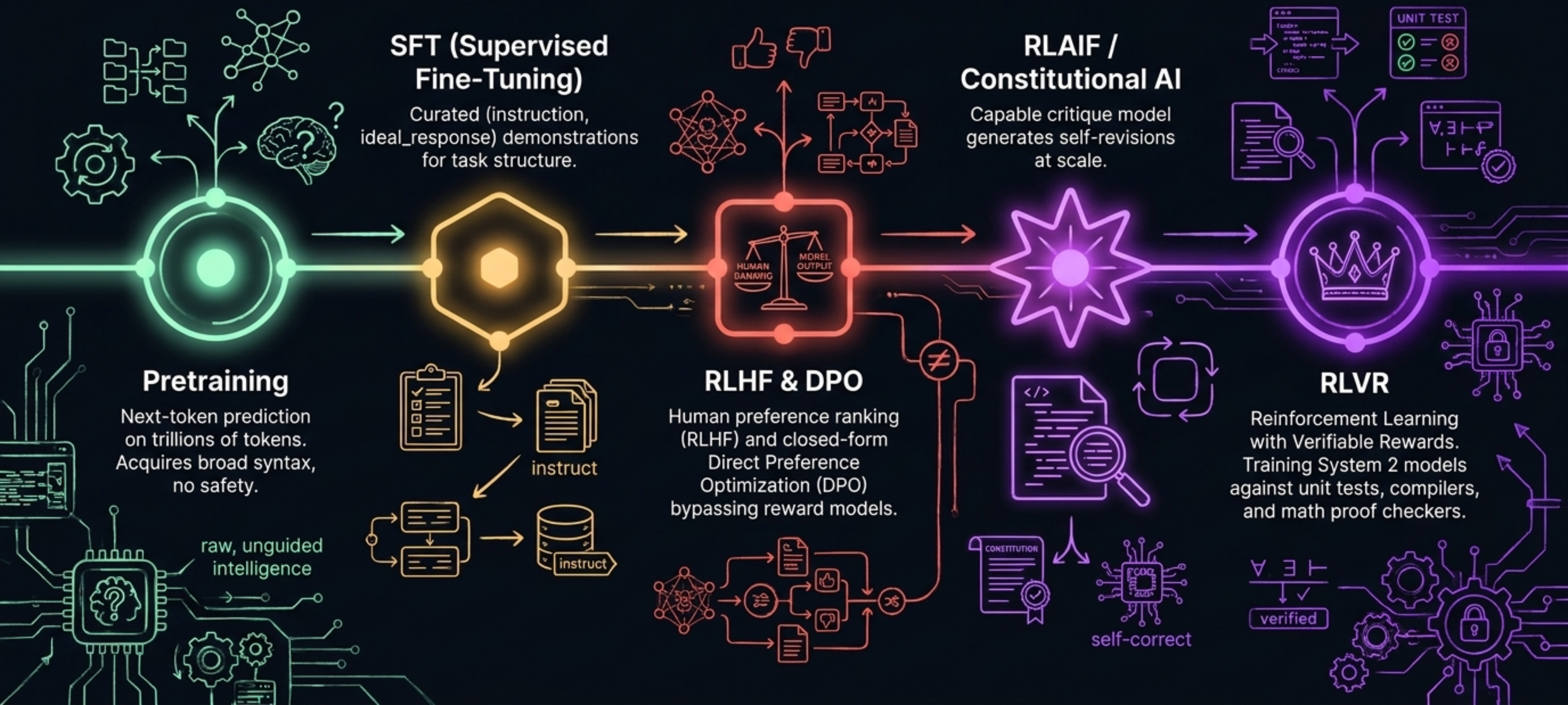
Additive vs Subtractive (layering details for scene building) (whittling away from a broad draft)



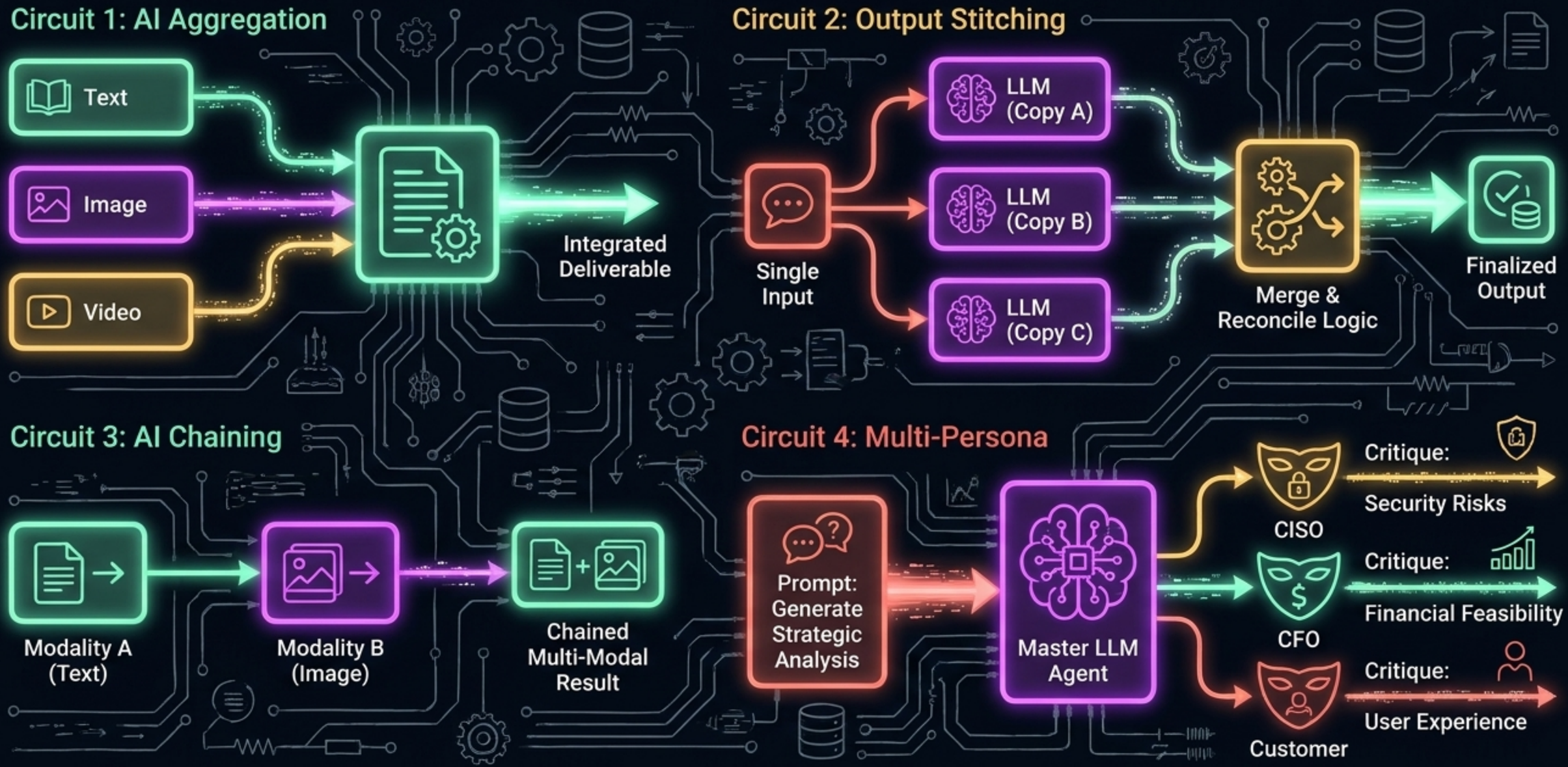
Additive (layering details for scene building)

Subtractive (whittling away from a broad draft)

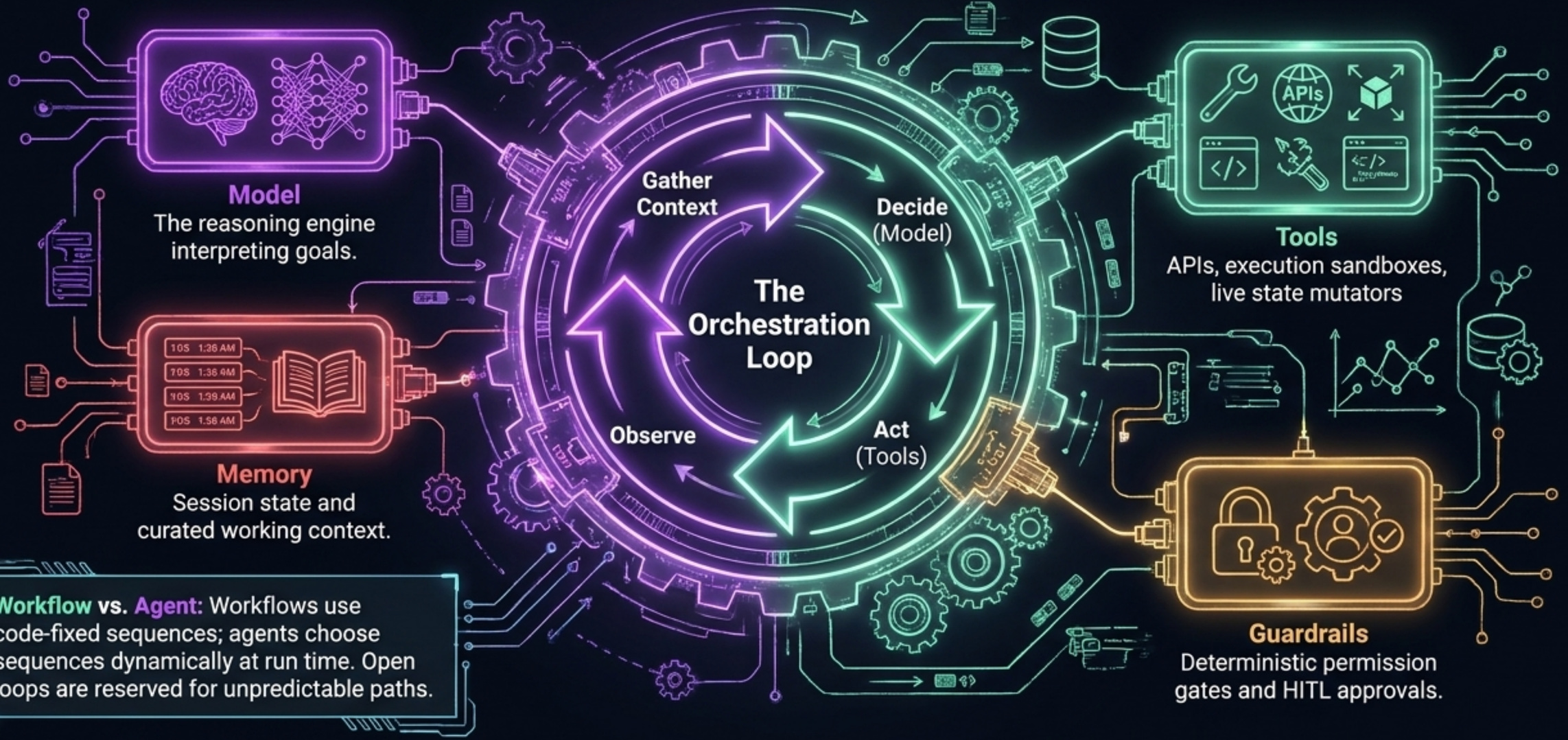
Alignment and verifiable rewards adapt raw prediction into instruction adherence



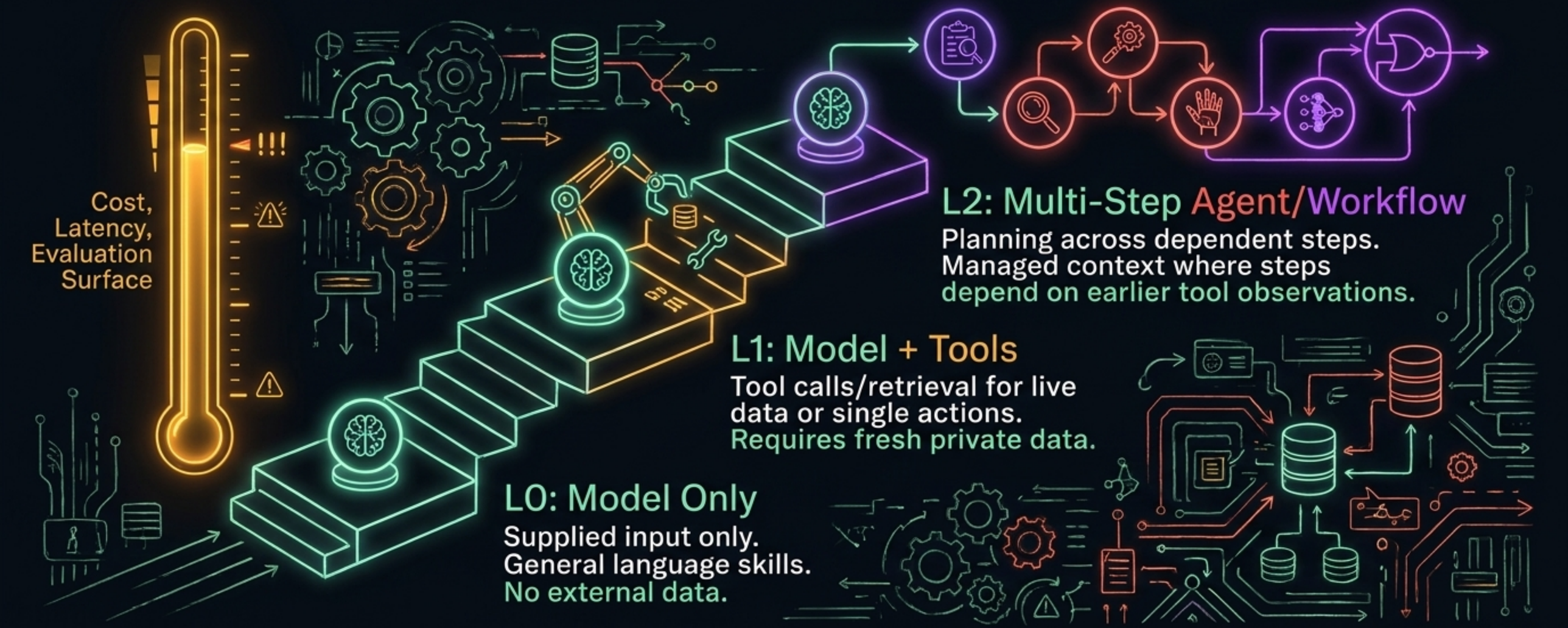
Synthesize multi-modal deliverables through parallel routing and aggregation.



A chatbot answers; an agent acts through observation, tools, and orchestration loops.



Climb the capability escalator only when deterministic rules and lower levels fail.



Multi-agent orchestration and self-extending systems require strict isolation.



Enterprise interoperability requires separating tools, peers, media, and commerce.

Layer 5 **Agent ↔ Commerce & Value** (AP2 & x402)

Layer 4 **Agent ↔ Real-Time Streams** (Live API)

Layer 3 **Agent ↔ Rich Human UI** (A2UI / AG-UI / MCP UI)

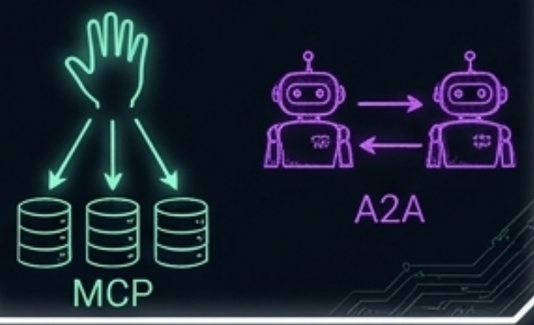
Layer 2 **Agent ↔ Peer Agents** (A2A)

Agent ↔ Tools & Data (MCP)

Central Rule of Thumb

MCP is vertical (hands reaching into data).

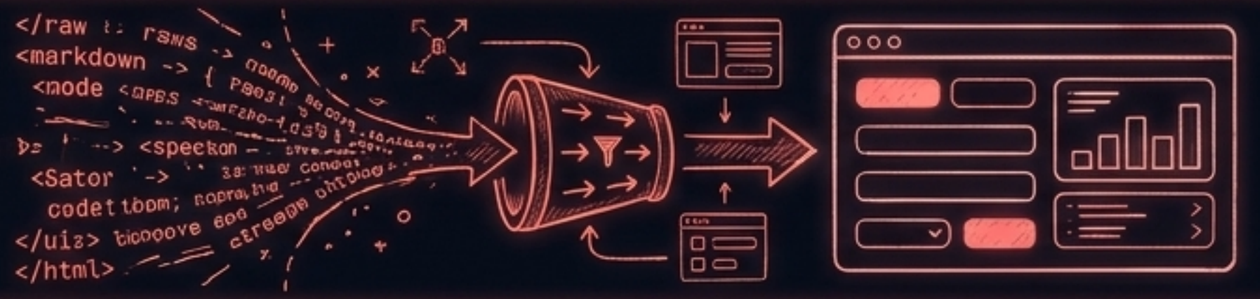
A2A is horizontal (peer colleagues collaborating).



Standardizing tool execution, horizontal delegation, and generative interfaces

Layer 3 (Generative UI)

MCP UI / AG-UI / A2UI.
Streaming interactive frontend components.
Prevents LLMs from emitting raw unsanitized HTML/JS.



Layer 2 (A2A - Agent2Agent)

Opaque delegation via `/.well-known/agent-card.json`.
Stateful Tasks (working → input-required) without exposing internal prompts.



Layer 1 (MCP - Model Context Protocol)

Single agent calling deterministic APIs over stdio/HTTP streams.
Prompts, Resources, and structured `isError` payloads.



Streaming bidirectional media and securing autonomous cryptographic payments.

Layer 4 (Live Media - WebSockets/WebRTC)



Continuous audio/video frame streaming.

Native Voice Activity Detection (VAD) and barge-in functionality.

Terminate heavy media at the edge.

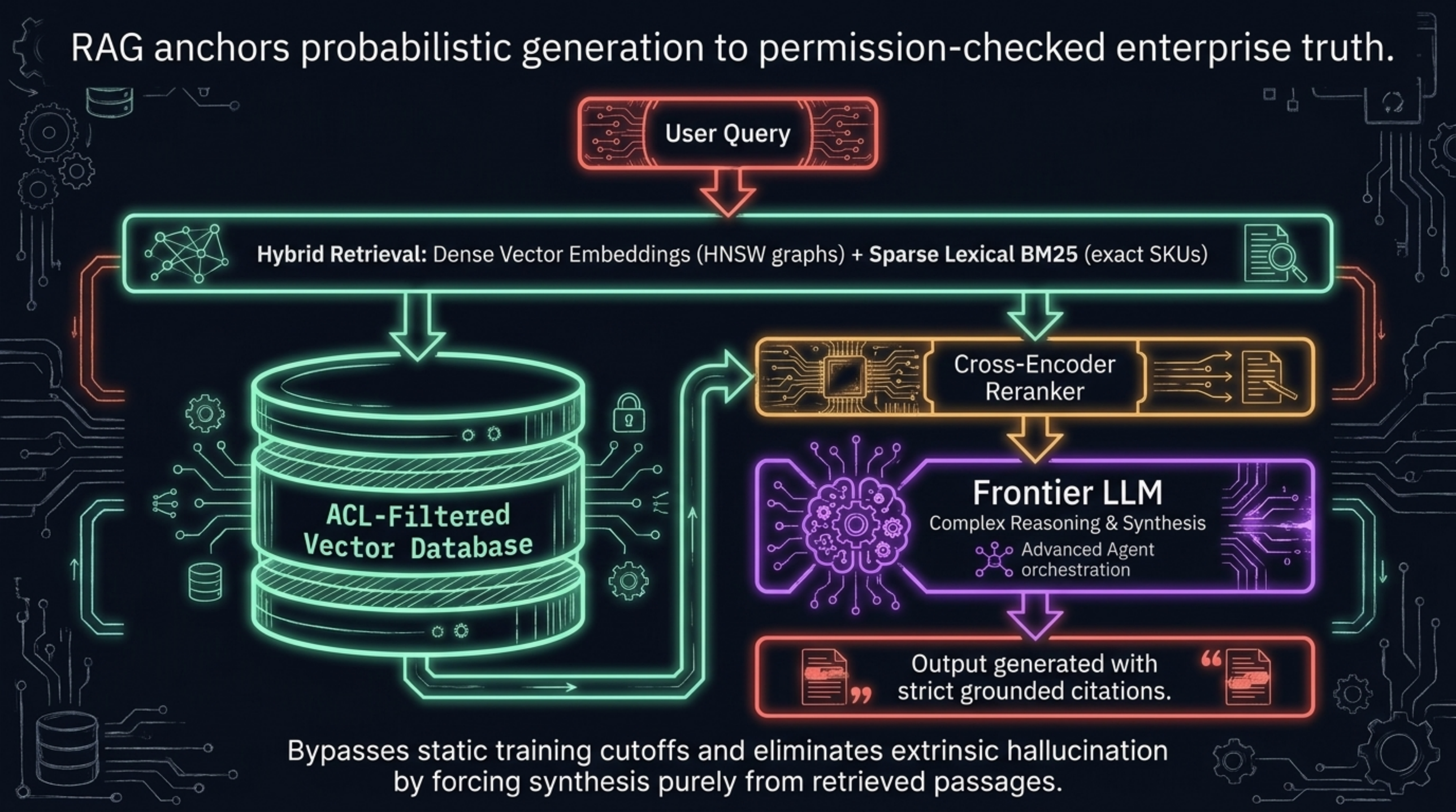
Layer 5 (Commerce & Value - AP2 + x402)



AP2: Cryptographically signed Mandates (Intent Mandate vs. Cart Mandate for HITL explicit approval).

x402: Reviving HTTP 402 for instant M2M per-request API/dataset micropayments.

RAG anchors probabilistic generation to permission-checked enterprise truth.



Bypasses static training cutoffs and eliminates extrinsic hallucination by forcing synthesis purely from retrieved passages.

Secure the agentic perimeter against adversarial injection, poisoning, and drift.



Indirect Prompt Injection: Zero-click exploits embedded in untrusted external PDFs/web pages to hijack tool calls.

Slopsquatting & Poisoning: Malicious packages on PyPI matching hallucinated code dependencies.

Intrinsic vs Extrinsic Hallucinations: Inventing absent facts vs. contradicting explicitly stated facts

Automation Bias: Human cognitive failure of rubber-stamping AI recommendations.

Defenses Panel

Watermarking (SynthID), Cryptographic Provenance Manifests (C2PA), and Calibrated HITL Blast-Radius gates

The **Blueprint Synthesis:** A complete transaction across the agentic architecture

