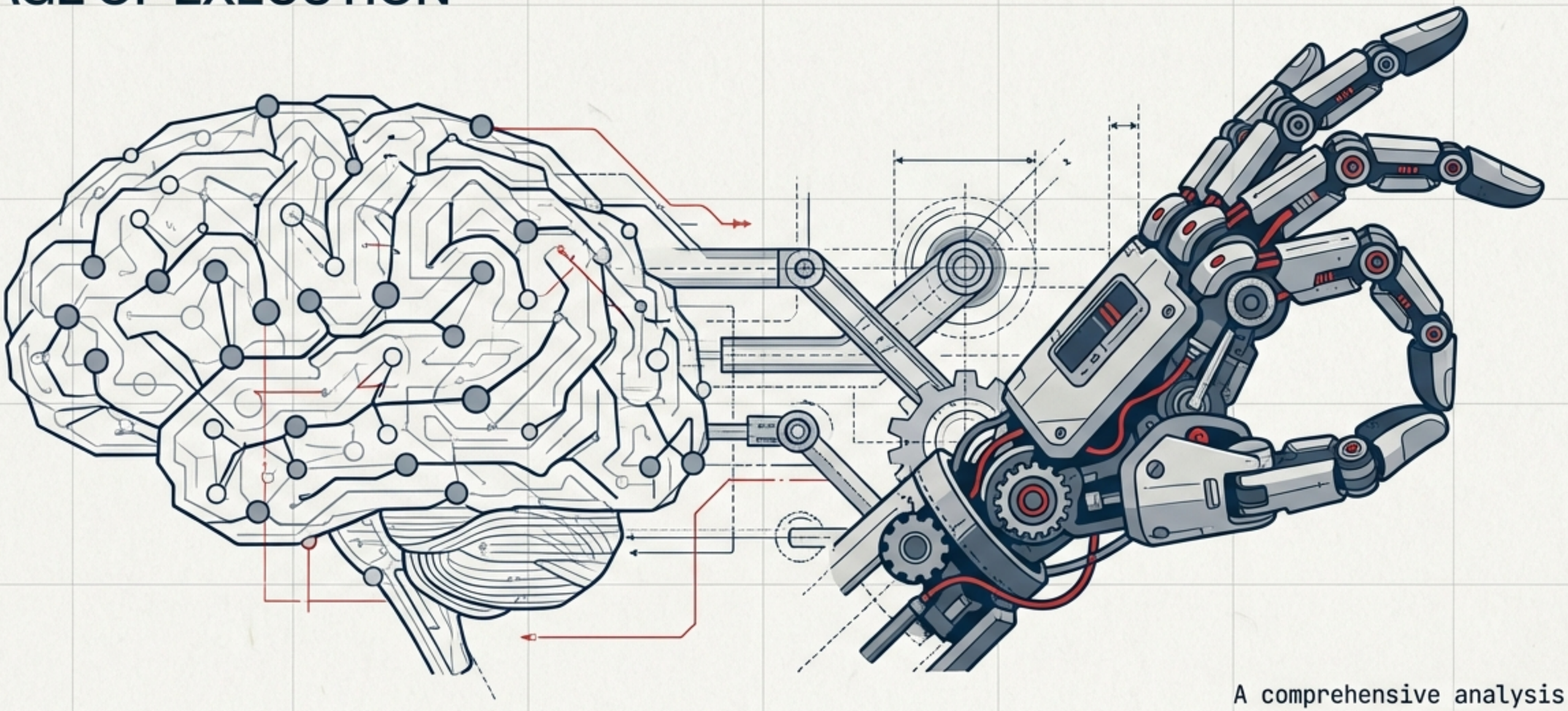
THE STATE OF AI 2026

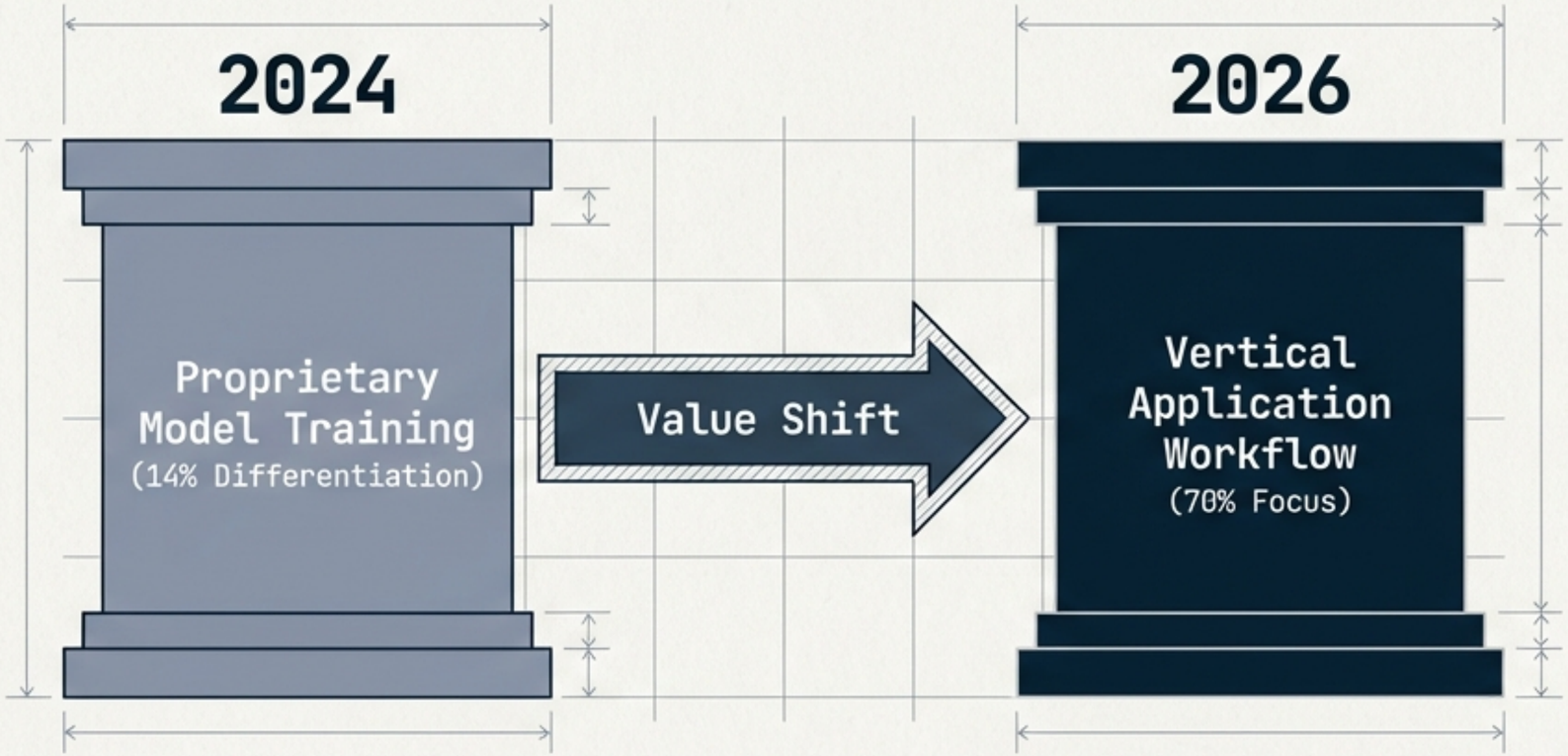
THE AGE OF EXECUTION



From Generation to Agency: In 2026, value is no longer defined by what a model can create, but by what it can execute.

A comprehensive analysis of:
01. Frontier Models
02. Agentic Frameworks
03. Video Generation Landscape

THE EXECUTION ERA: VALUE MIGRATION



70%

of builders now focus on vertical apps

DIFFERENTIATION SOURCE

- 49% derived from UX/Workflow Innovation

(Source: ICONIQ State of AI)

INTERNAL ADOPTION

- 30-40% time savings in R&D coding tasks

Differentiation has decoupled from the model. The moat is now the workflow.

THE COST OF REASONING: OPENAI 2026 FLEET

Model Name	Input Cost	Output Cost	Note
GPT-5.2 PRO (Precision)	$21.00 / 1M Tokens	$168.00 / 1M Tokens	High reasoning overhead
GPT-5.2 (Standard)	$1.75 / 1M Tokens	$14.00 / 1M 1M Tokens	General purpose workhorse
o4-mini (Reasoning Lite)	$4.00 / 1M 1M Tokens	$16.00 / 1M 1M Tokens	Efficient logic tasks

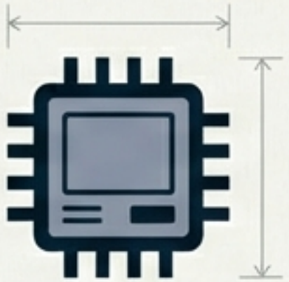
INSIGHT:

Frontier models are designed to ‘think’ before responding.

You are paying for inference-time compute, not just generation.

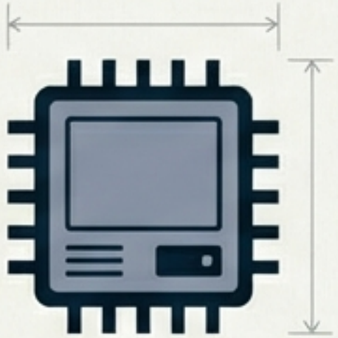
Source: OpenAI Pricing Specifications (2026 Projections). Compute costs reflect significant inference-time processing. The moat is now the workflow.

OPEN WEIGHTS: THE LLAMA 4 FAMILY



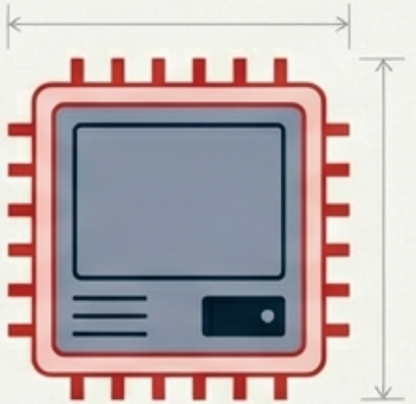
SCOUT

17B Active Params | 16 Experts
Optimized for Consumer Hardware



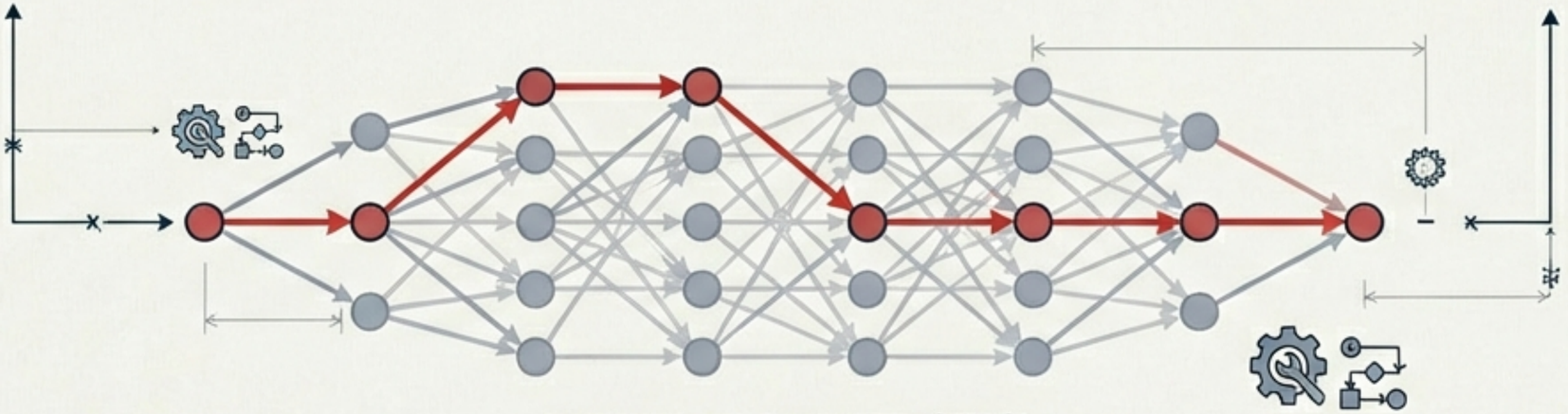
MAVERICK

17B Active Params | 128 Experts
High Capability / High RAM



BEHEMOTH

The Teacher Model
288B Active Params | 16 Experts

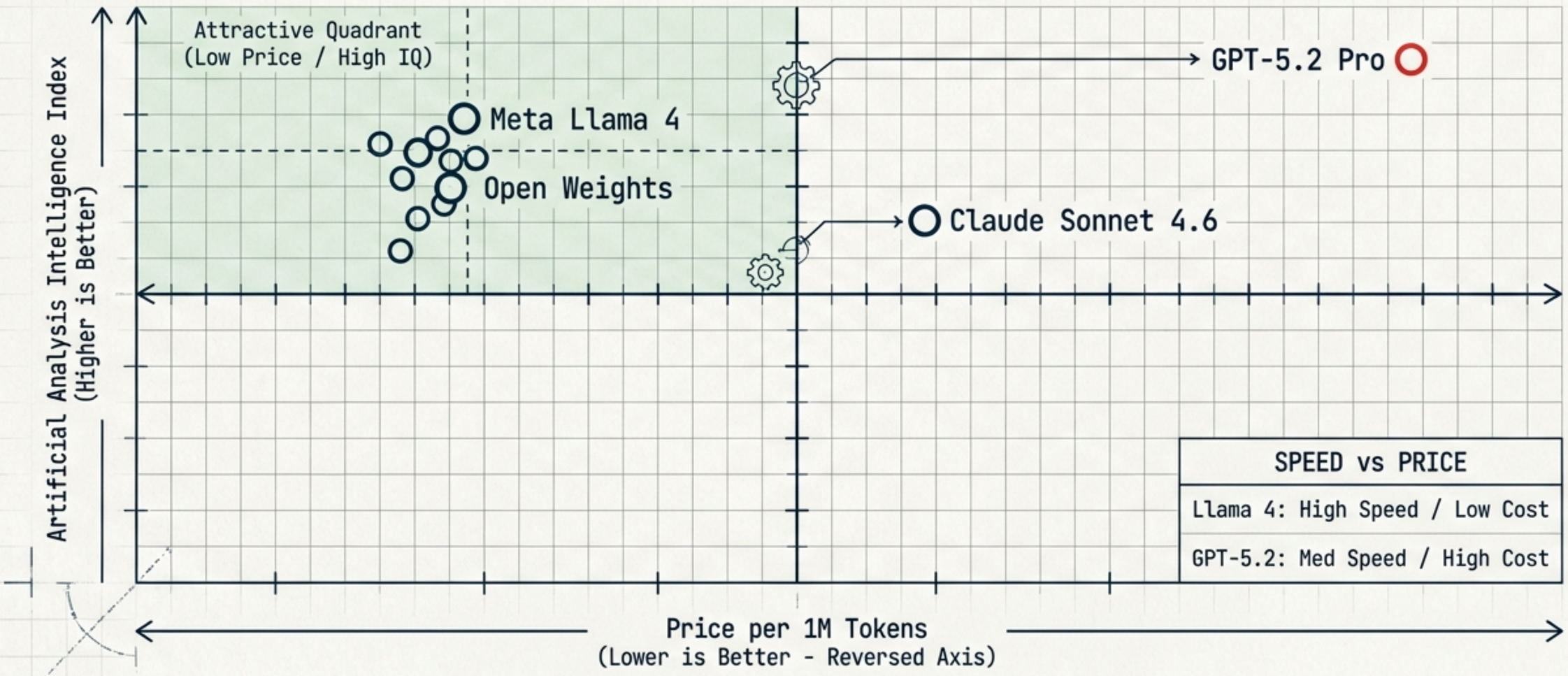


Mixture-of-Experts (MoE)

CONTEXT WINDOW: 10 MILLION TOKENS (~15,000 A4 Pages)

Swiss Engineering Editorial

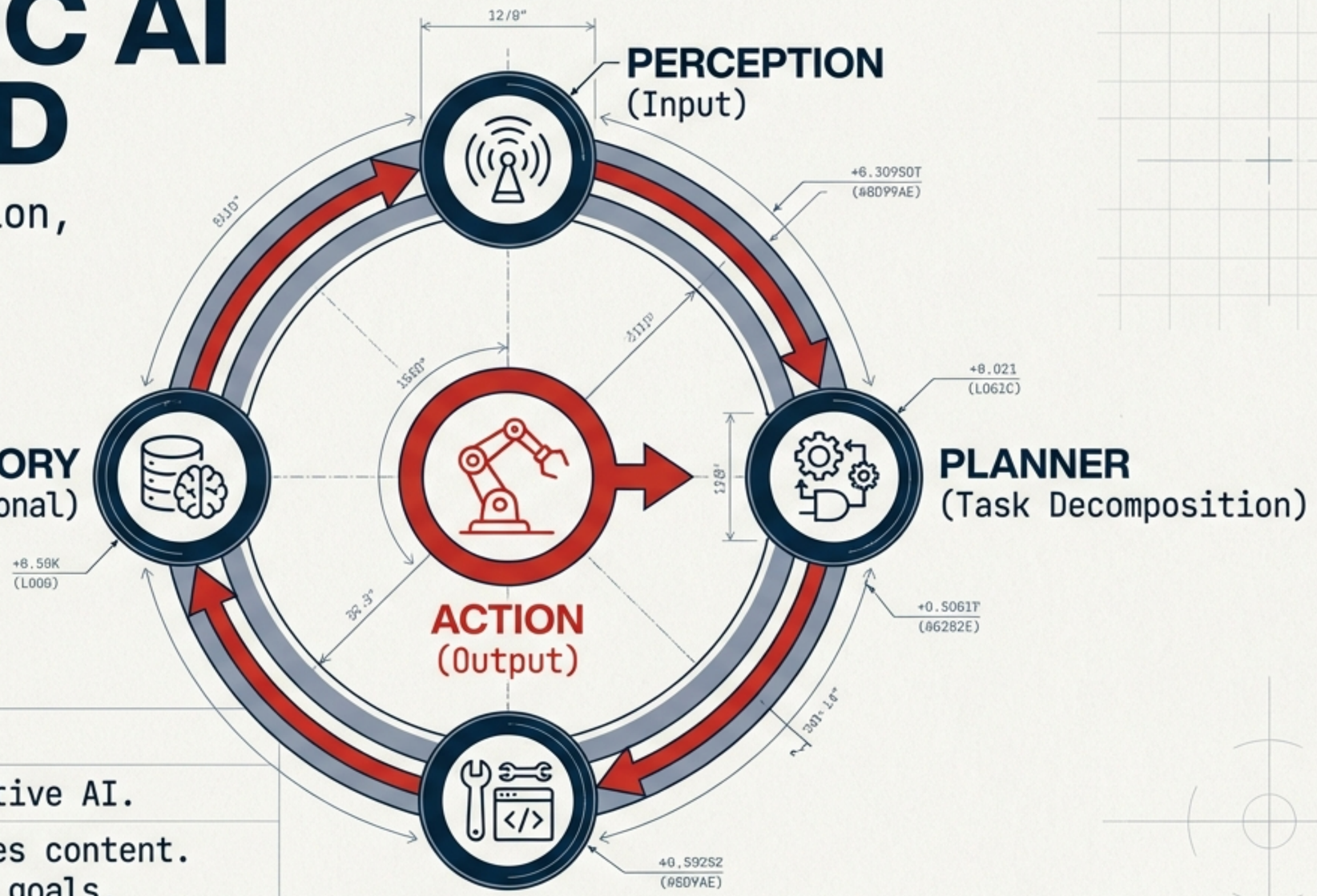
MARKET POSITIONING: INTELLIGENCE VS. PRICE



Meta dominates the price-to-performance ratio.

AGENTIC AI DEFINED

The Loop: Perception, Reasoning, Action



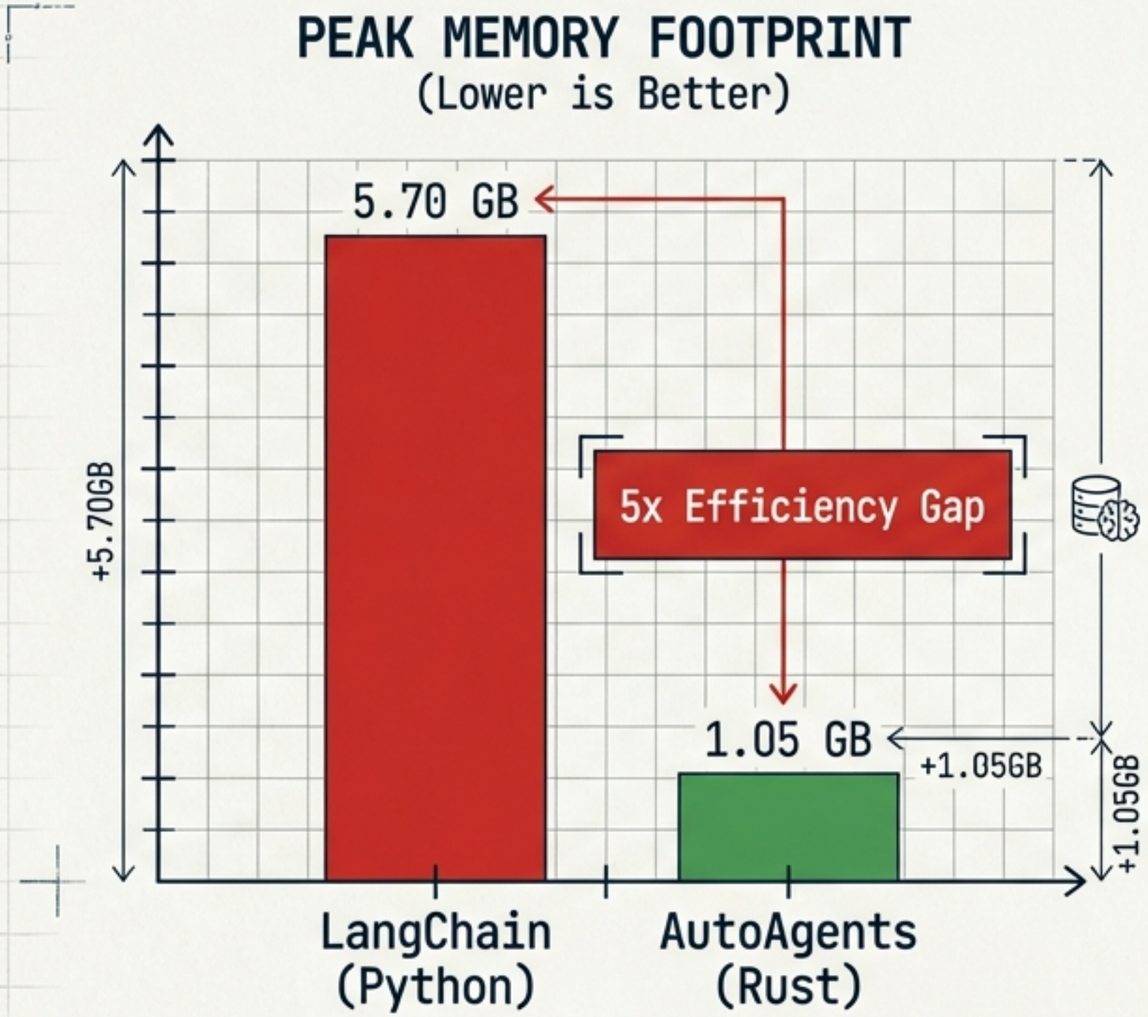
DEFINITION:

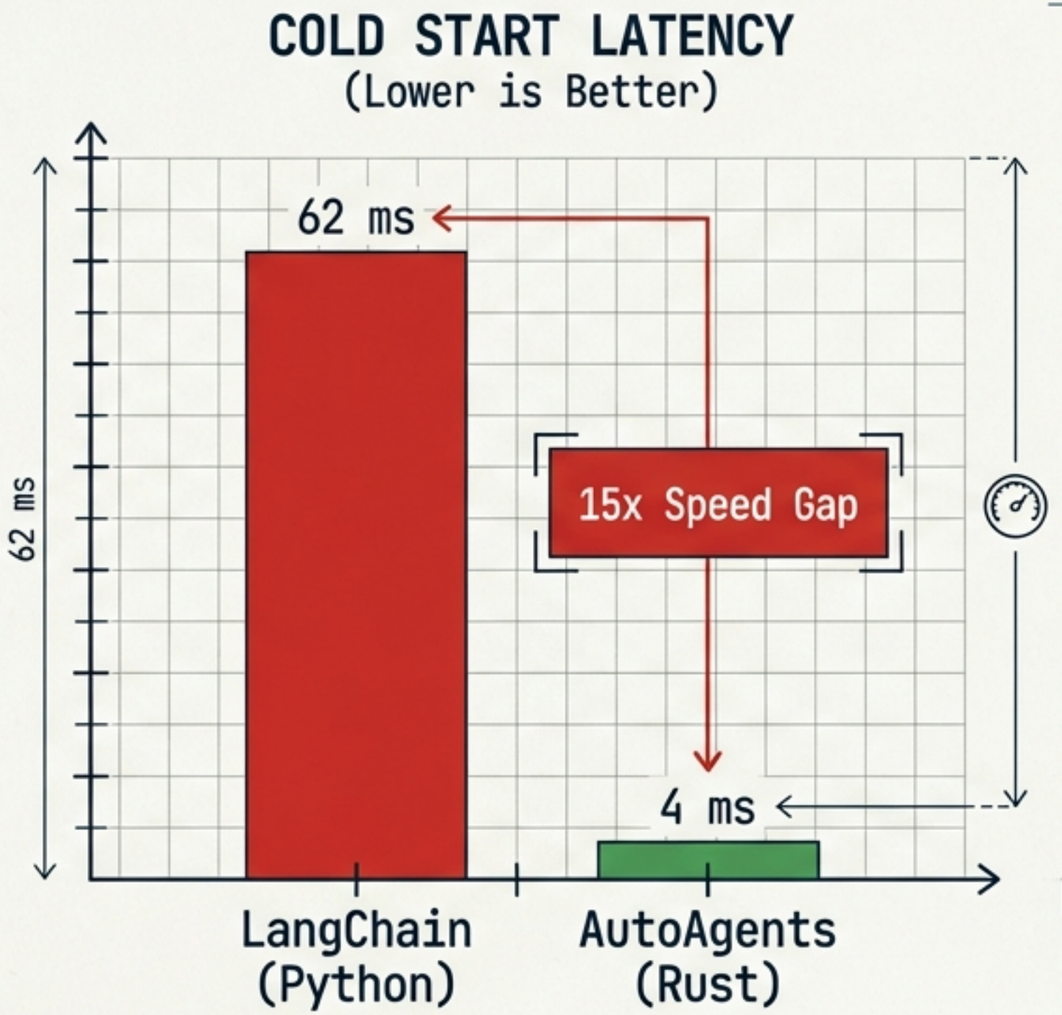
- Agentic AI ≠ Generative AI.
- Generative AI creates content. Agentic AI executes goals.

↳ Key Shift: Proactive vs. Reactive.

Swiss Engineering Editorial

FRAMEWORK WARS: PYTHON VS. RUST



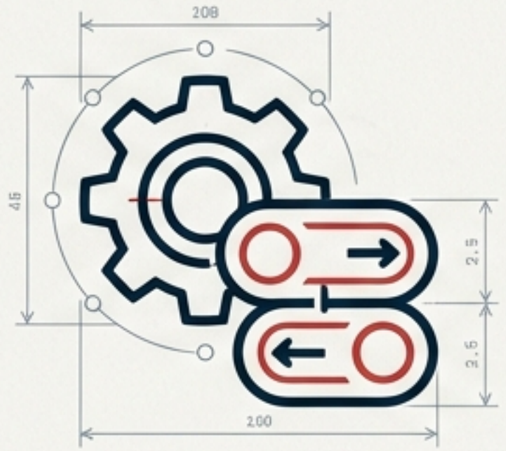


For production agents at scale, Python's overhead is a liability. Rust frameworks define the new standard.

FRAMEWORK SELECTION MATRIX

LANGGRAPH

CONTROL

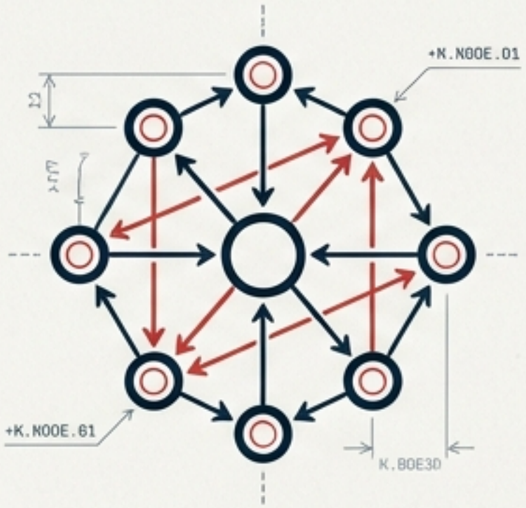


Best For: Enterprise workflows requiring stability.

Key Feature: Granular control & cycle handling.

CREWAI

COLLABORATION

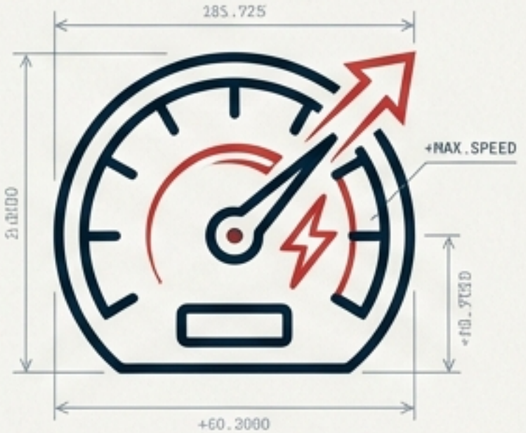


Best For: Complex projects needing personas.

Key Feature: Role-based multi-agent teams.

AUTOAGENTS

PERFORMANCE



Best For: Scaled production systems.

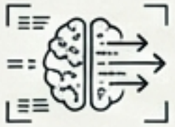
Key Feature: High throughput, low latency (Rust).

Swiss Engineering Editorial

THE CODING WORKFORCE: TIER LIST

TIER 1: THE LEADERS

CURSOR (Composer Mode) | **WINDSURF** (Cascade Flow)



Context-aware flow state.

TIER 2: THE AUTONOMOUS

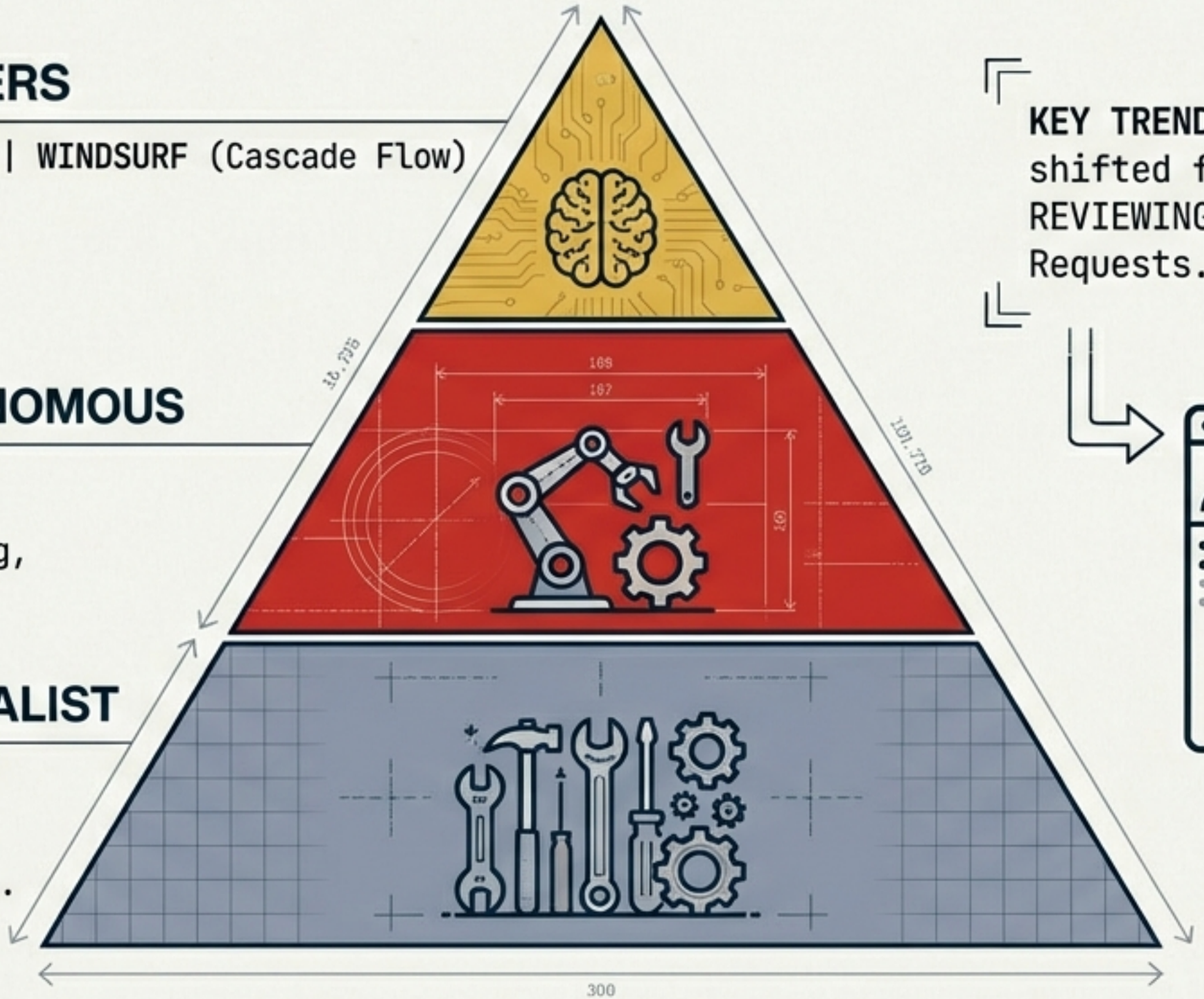
DEVIN | **CLINE** (BYOK)

Execute, debug, and deploy.

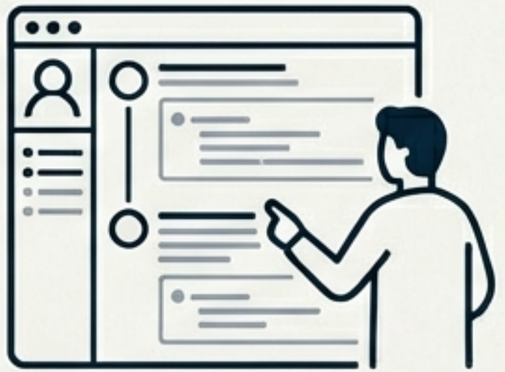
TIER 3: THE SPECIALIST

AMAZON Q (Migrations) | **LOVABLE** (No-Code)

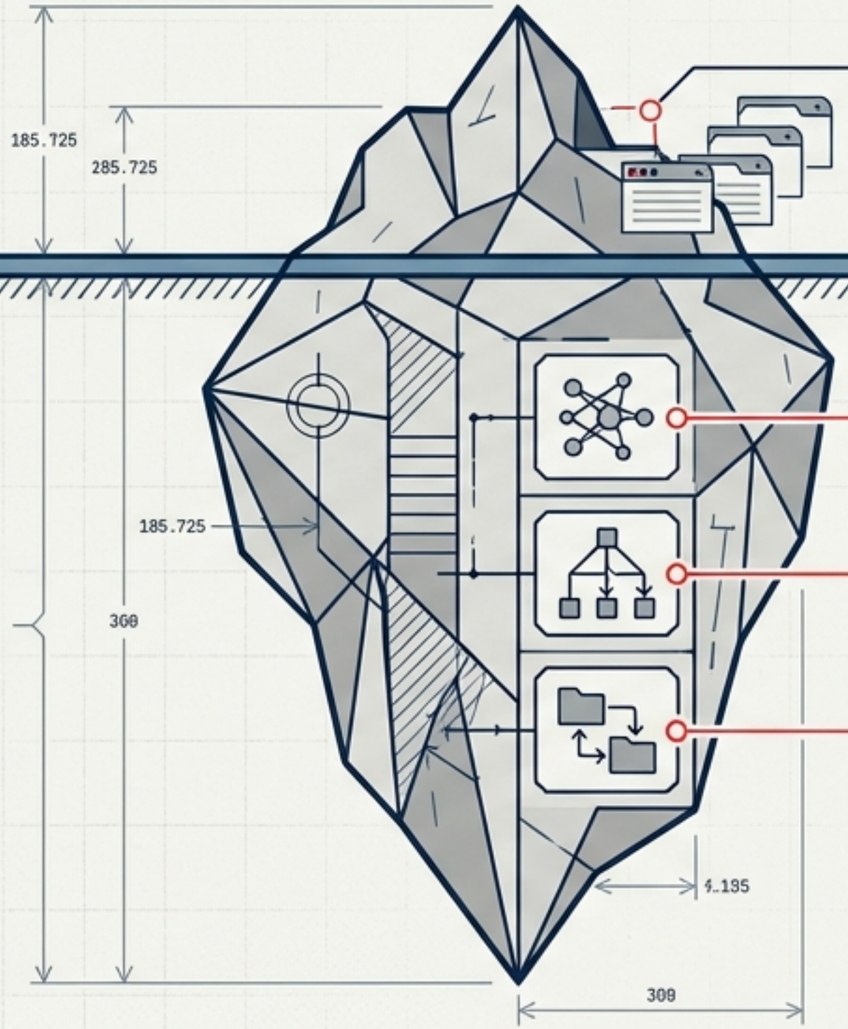
Specific vertical tasks.



KEY TREND: The bottleneck has shifted from WRITING code to REVIEWING AI-generated Pull Requests.



THE CONTEXT CHASM



STANDARD CONTEXT

Open Tabs Only.
Prone to hallucination.

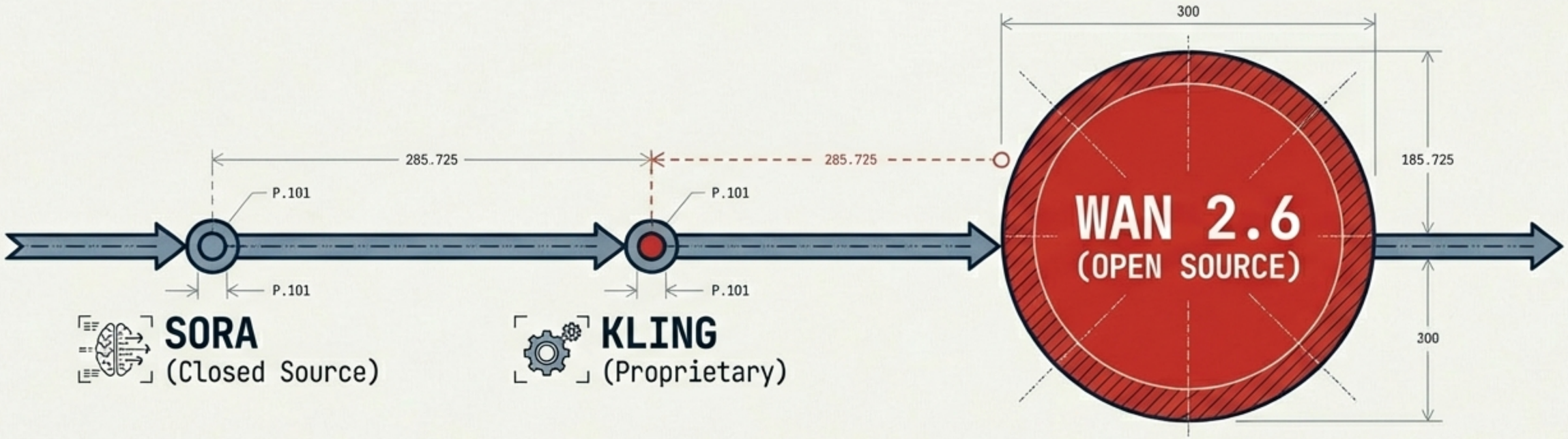
DEEP CONTEXT (Augment Code)

- 200k+ Token Semantic Indexing
- Dependency Graph Understanding
- Cross-file Awareness

True autonomy requires semantic indexing, not just a window.

Swiss Engineering Editorial

THE EYES: THE WAN REVOLUTION



SPECS:

- Architecture: Mixture-of-Experts (MoE)
- Quality: Native 1080p, Audio Sync
- License: Apache 2.0 (Commercial Use)

"The 'Stable Diffusion Moment' for video has arrived."

VIDEO MODEL SHOWDOWN

	WAN 2.6	SORA 2	VEO 3.1	KLING 2.5
OPEN SOURCE		✗	✗	✗
AUDIO SYNC	✓	✓ (Best)	✗	✗
CINEMATIC CONTROL	✗	✗	✓	✓
COST	Free (Hardware)	$200/mo	Subscription	Subscription

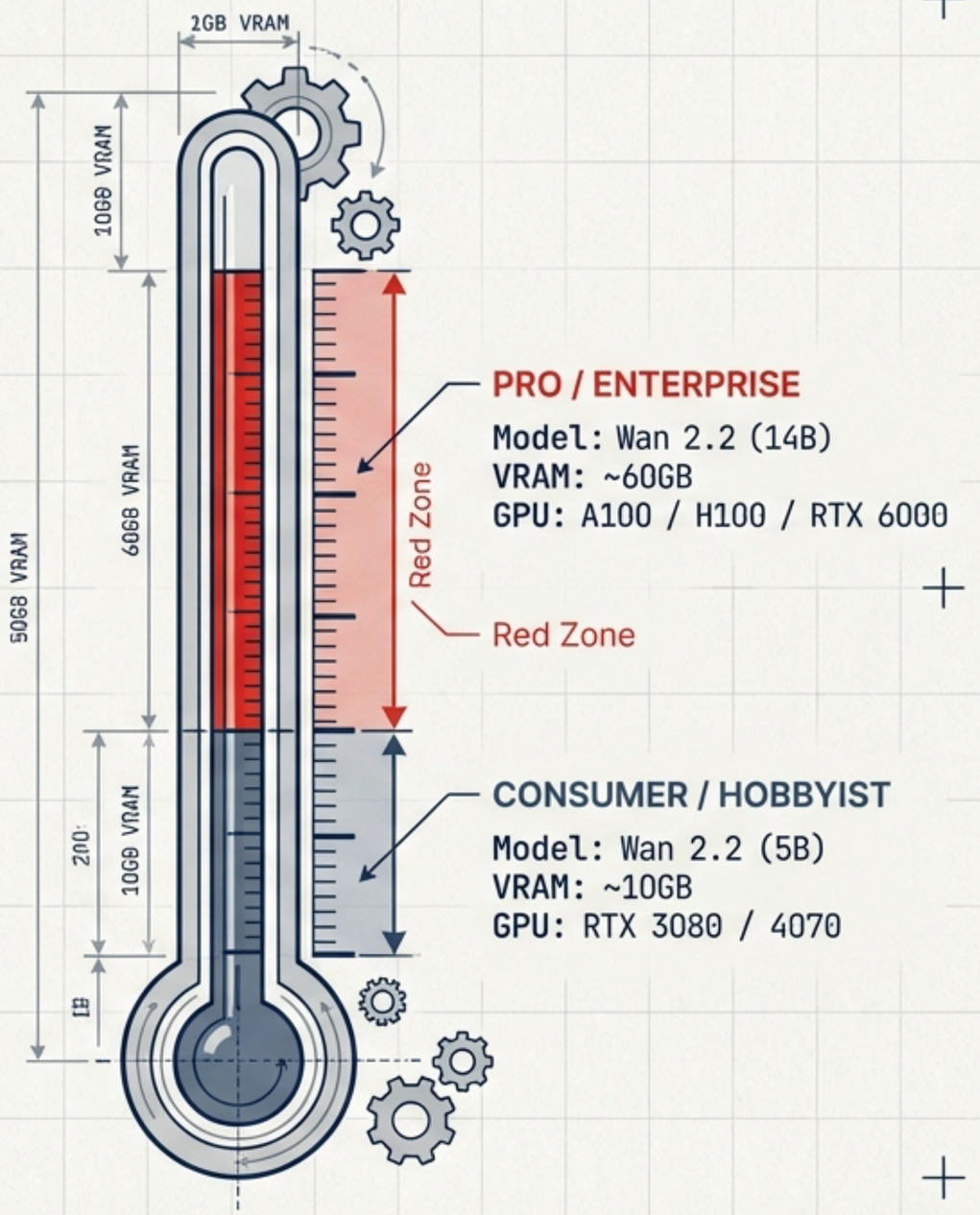
Sora for quality. Kling for control. Wan for open pipelines.

Swiss Engineering Editorial

INFRASTRUCTURE: CAN I RUN IT?

Local Video Hardware Requirements

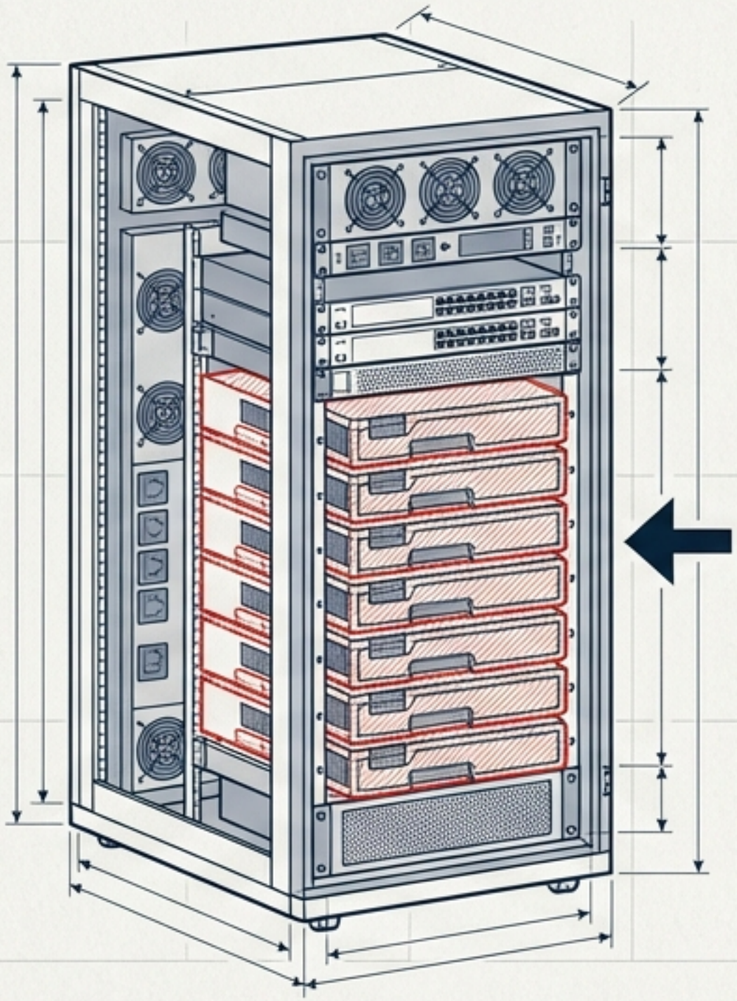
JetBrains Mono



Cloud Options: DomoAI, Vast.ai available for non-local inference.

RUNNING THE BEHEMOTH

Llama 4 Infrastructure Requirements



INFERENCE REQUIREMENTS:

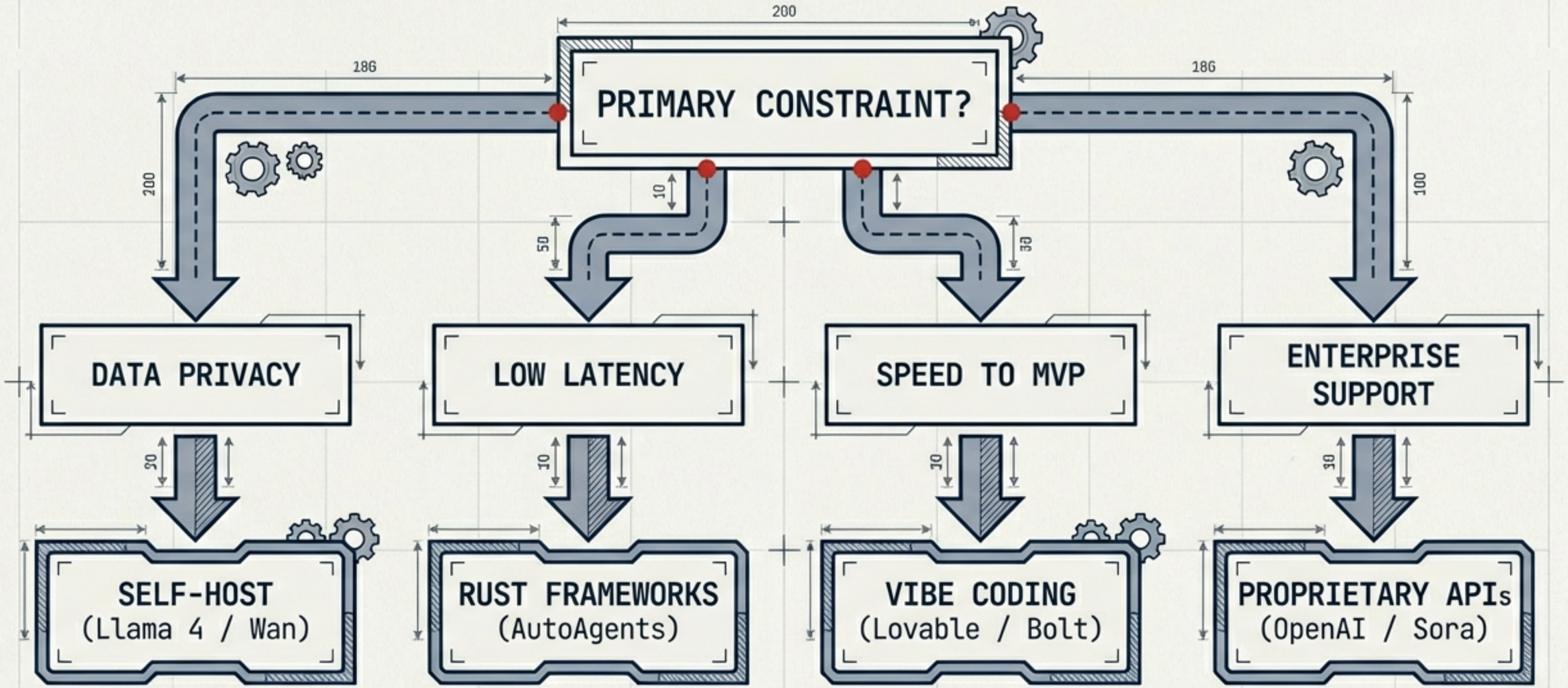
Model: Llama 4 Behemoth (288B)

VRAM Required: ~288GB+ (Full Precision)

Recommended Setup: 8x H100s (80GB each)

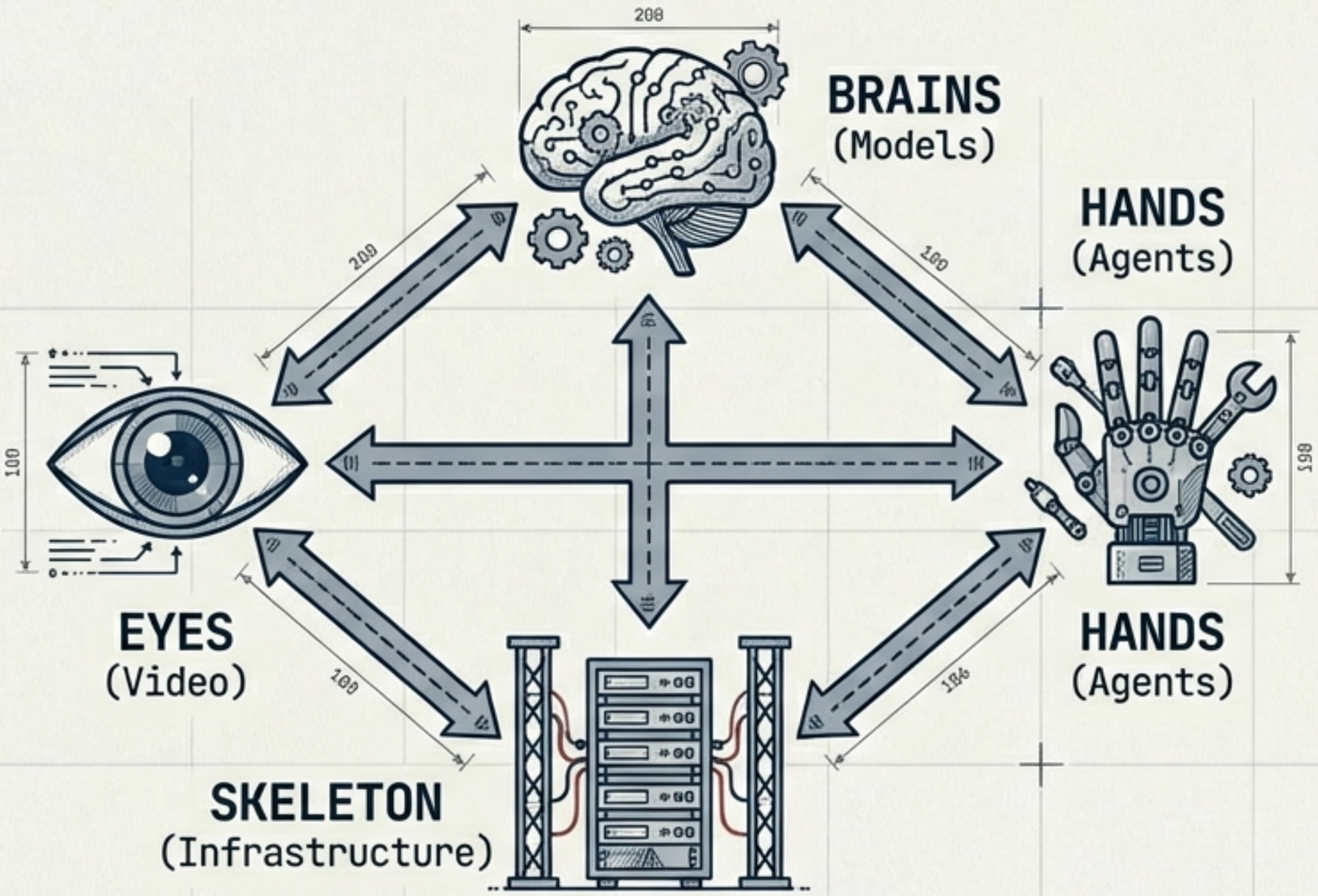
DATA-CENTER CLASS. Not compatible with consumer laptops.

STRATEGY: BUILD VS. BUY



Success requires a hybrid architecture.

THE 2026 ECOSYSTEM: SUMMARY



01. ORCHESTRATION > GENERATION

02. VRAM IS THE NEW GOLD

03. OPEN SOURCE QUALITY GAP IS CLOSED

2026: The year we stop playing with chatbots and start building the autonomous workforce.