STATE OF AI

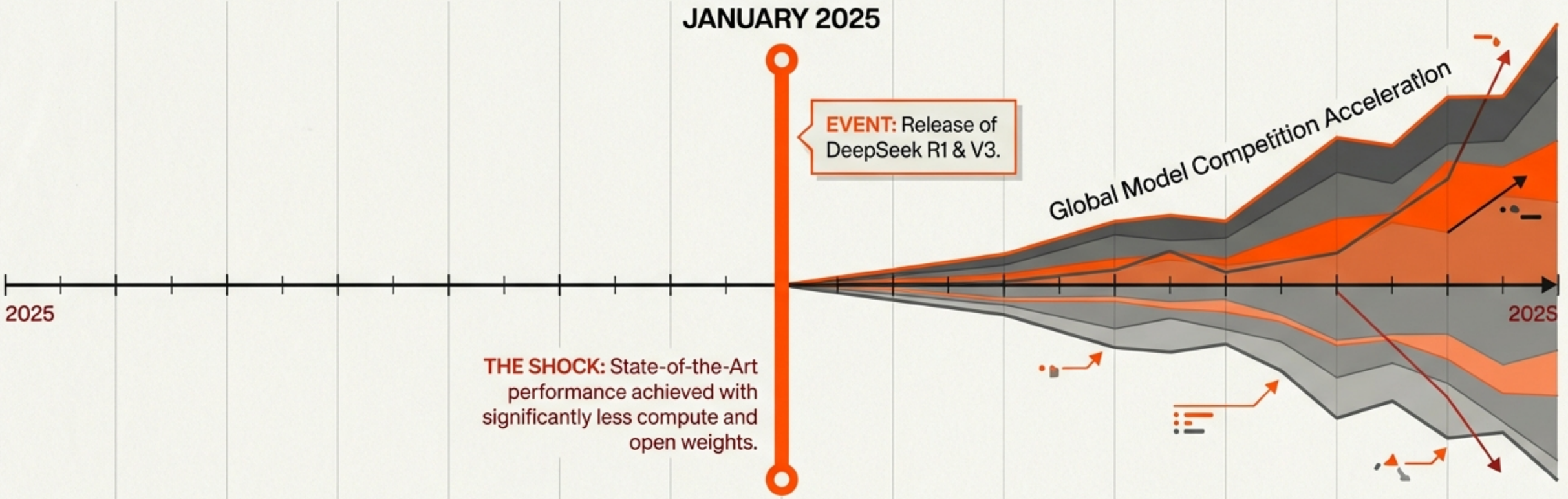
2026

A Retrospective Analysis of the Post-Training Era.
Covering: The DeepSeek Moment. Inference Time Scaling.
The China Strategy.
The Rise of Jagged Intelligence

PRINARY SOURCE: RETROSPECTIVE 2025-2026
REPORT STATUS: FINAL
DATE: JANUARY 2026

THE DEEPSEEK MOMENT

Efficiency as the New Disruption



The Catalyst

DeepSeek, an open-weight Chinese lab, released R1, proving frontier performance didn't require massive US-style capital.

The Reaction

Shattered the industry consensus that only closed US labs could compete. Triggered a "leapfrogging" cycle of model releases.

The Quote

"The AI competition has gotten insane... it's just been accelerating throughout 2025."

A TALE OF TWO STRATEGIES

The Great Bifurcation: Influence vs. Revenue

CHINA STRATEGY

Influence via Open Weights

- DeepSeek
- Kimi
- MiniMax
- Qwen (Alibaba)
- Z.ai

Constraint: Low domestic willingness to pay for software + Hardware export controls.

Tactic: Release high-performance Open Weight models to capture global developer mindshare.

Result: Dominance of the open-source ecosystem, mirroring the 'ChatGPT moment' but for developers.

US STRATEGY

Revenue via Closed APIs

- OpenAI (GPT-5.2)
- Google (Gemini 3)
- Anthropic (Claude Opus 4.5)

Incentive: High-margin enterprise software market & security/safety concerns.

Tactic: Proprietary models guarded by massive infrastructure infrastructure moats (Custom data centers, TPU clusters).

Result: Leadership in peak reliability and enterprise integration.

ARCHITECTURE: EVOLUTION, NOT REVOLUTION

Optimization of the Transformer Block

Mixture of Experts (MoE)

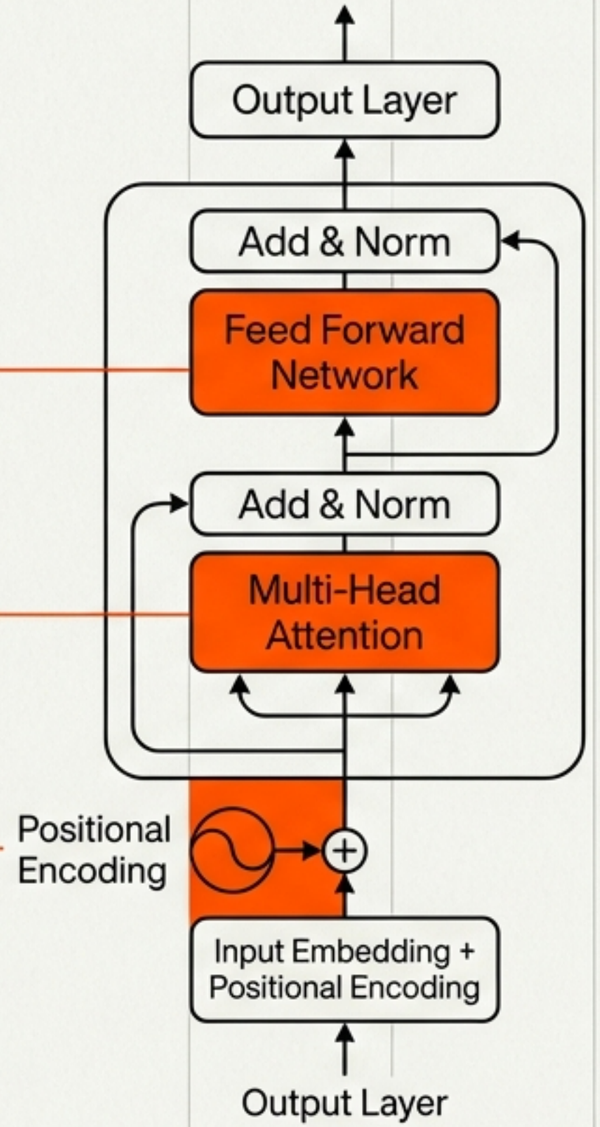
Sparse activation. Increases parameter count without exploding compute cost. Used by DeepSeek & Mistral.

Attention Variants (MLA & GQA)

Multi-head Latent Attention and Group Query Attention. Shrinks KV cache size to drastically improve inference speed.

System Optimization (FP8/FP4)

Training at lower precision (FP8/FP4) to maximize compute throughput.



Core Assertion: The fundamental architecture remains the Autoregressive Transformer derived from GPT-2. Gains in 2026 are driven by systems and data optimization, not a new paradigm.

THE THREE AXES OF SCALING

Shifting Value from Pre-Training to Inference

1. PRE-TRAINING

The Foundation

- **Cost:** High Fixed Cost ($5M - $10M+ per run).
- **Status:** Diminishing returns. Still provides base knowledge, but no longer the primary differentiator.

2. POST-TRAINING

The Skill Unlock

- **Cost:** Variable / High Talent Density.
- **Status:** The 2025/2026 Frontier. Turning raw knowledge into reasoning via RLVR.

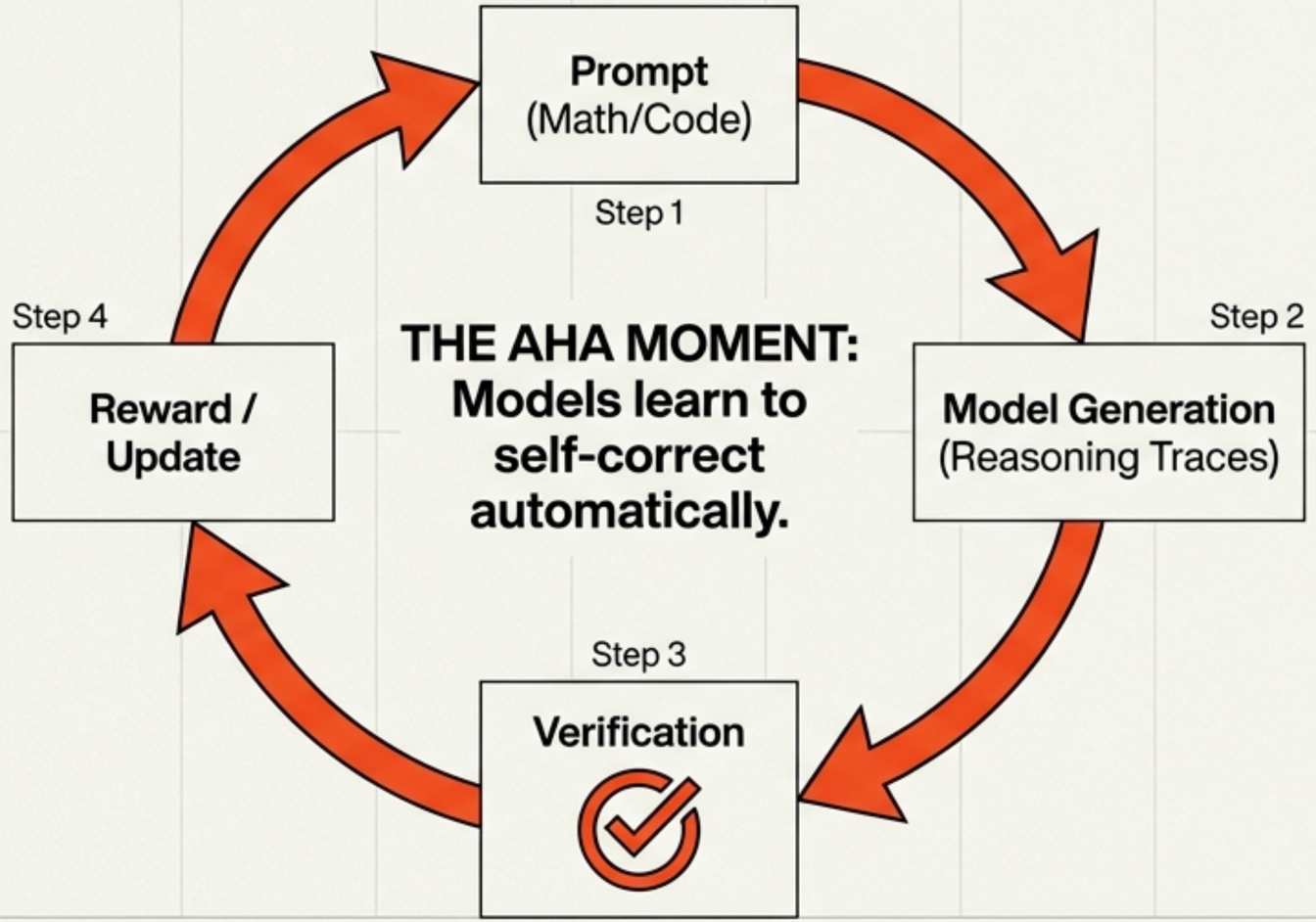
3. INFERENCE TIME

The Thinking Phase

- **Cost:** Recurring, Billions in OpEx.
- **Status:** The New Scale. Models generate hidden “thought tokens” before answering. Allows smaller models to outperform larger ones by “thinking” longer.

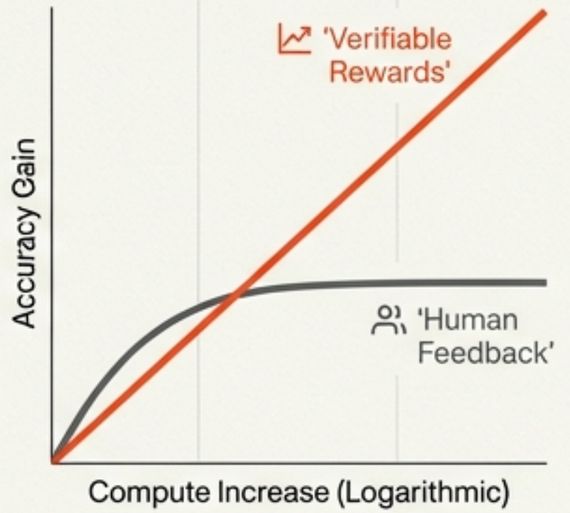
THE ENGINE OF 2026: RLVR

Reinforcement Learning with Verifiable Rewards



The Scaling Law

Unlike human feedback, RLVR scales linearly.



Logarithmic compute increase = Linear accuracy gain.

RLHF VS. RLVR

Vibes vs. Verification

RLHF

Reinforcement Learning from Human Feedback

Goal:	Optimizes for Style, Tone, Safety, Formatting.
Constraint:	Subjective (“Vibes”). Humans cannot scale indefinitely.
Failure Mode:	Over-optimization leads to mode collapse (generic/repetitive outputs).
Analogy:	The User Interface of Intelligence.

RLVR

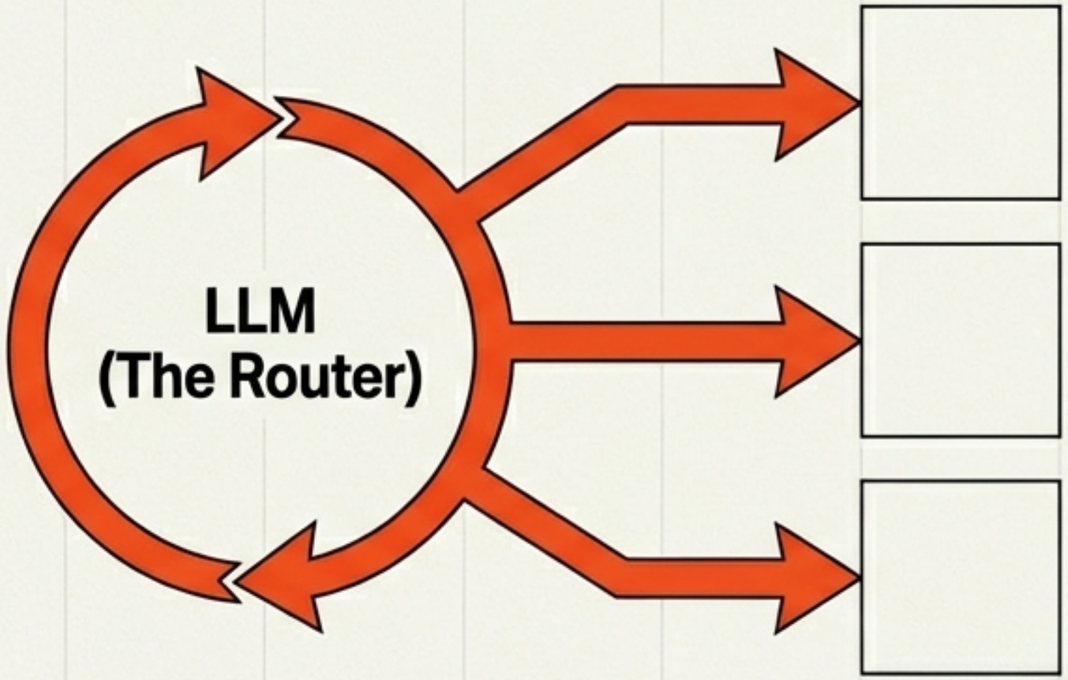
Reinforcement Learning with Verifiable Rewards

Goal:	Optimizes for Reasoning, Accuracy, Problem Solving.
Constraint:	Requires Objective Truth (Math/Code/Logic).
Scaling:	Can run millions of simulations automatically without humans.
Analogy:	The Raw Intelligence Engine.

Mid-Training: The critical bridge curating data to prepare the model for RLVR.

SOLVING HALLUCINATION: TOOL USE

The Shift from Recall to Search/Compute



Search Tool
Don't memorize the 1998 World Cup winner. Search for it.

Python Interpreter
Don't calculate math in-weights. Execute code.

Compiler
Don't guess syntax. Verify with a compiler.

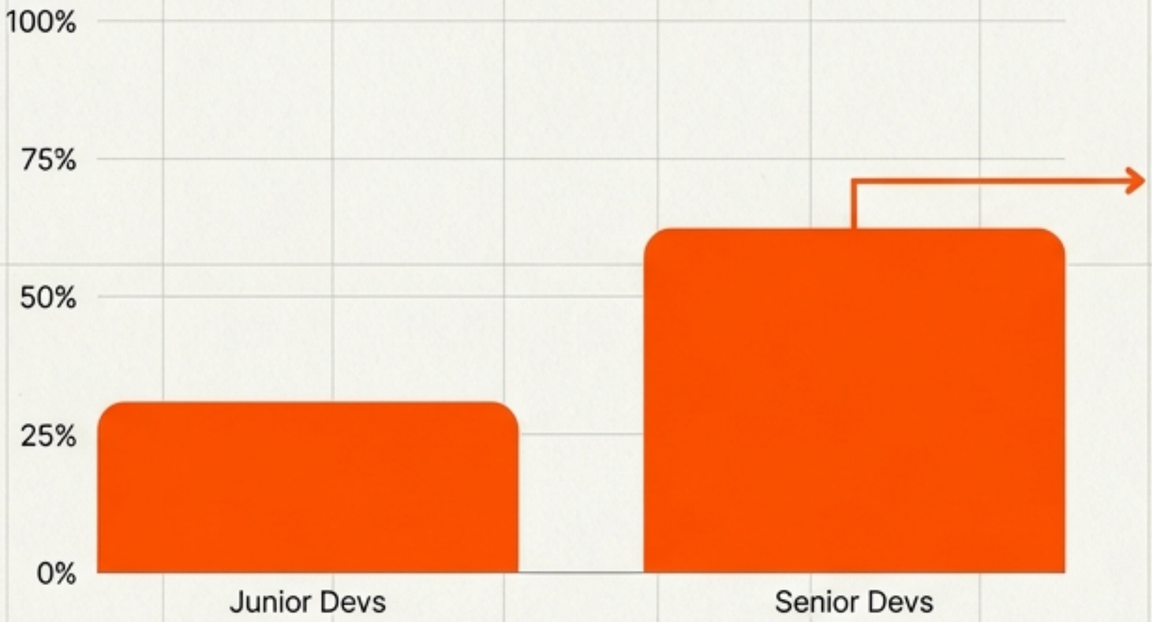
Current Status:

- Closed Models (Claude / GPT): Deep integration (e.g., 'Computer Use').
- Open Models (gpt-oss-120b): Catching up with specific tool-calling training.

THE SUPERHUMAN CODER

The Rise of "Vibe Coding"

AI Code Usage Comparison



The Paradox:
Seniors use MORE AI because they have the judgment to verify.

Vibe Coding

The workflow has shifted from writing syntax to managing outcomes via natural language. Tools like Cursor and Claude Code dominate.

The Dilemma: The Desert of Debugging

Quote: “Debugging is like a drink of water after going through a desert.”
Risk: By skipping the struggle, we erode the mastery required to solve novel problems.

Architectural Journal

EDUCATION: THE “BLUE BOOK” RETURN

Verifying Competence in the Age of Cheating

Pre-2025: Digital First

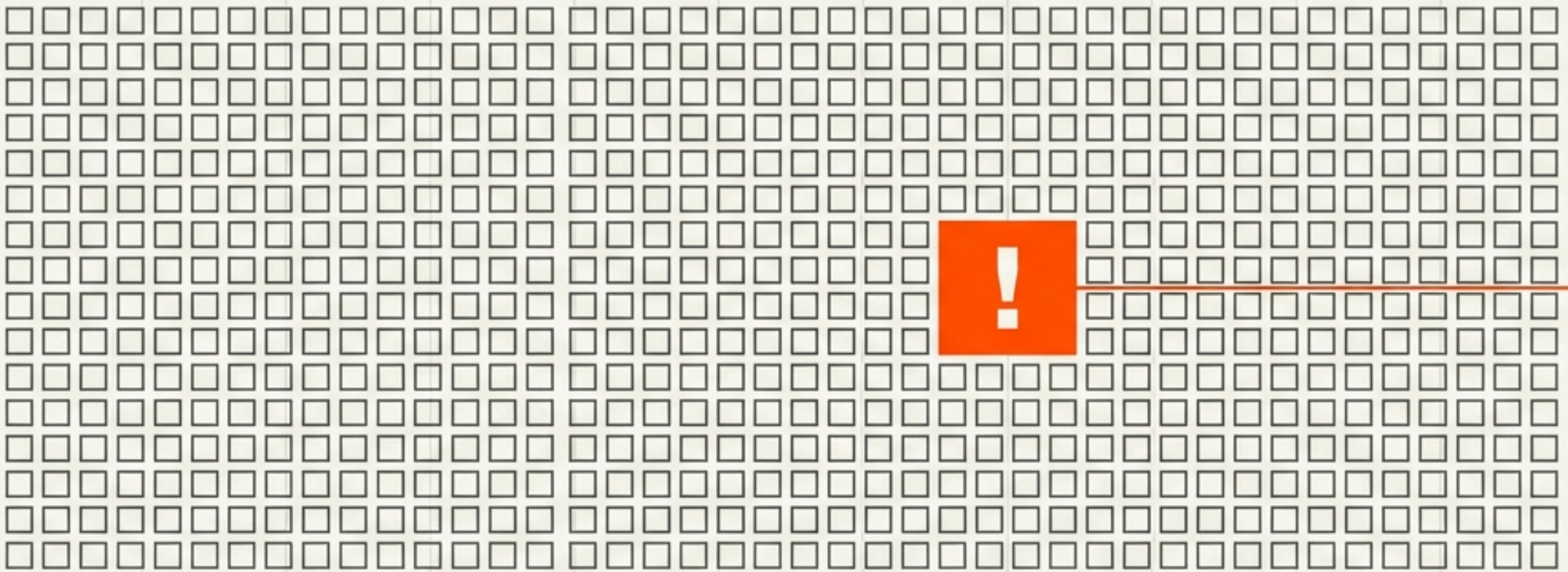
Post-2026: Analog Verification

The Education Pivot: As digital exams become trivially cheatable, institutions return to oral exams and physical writing to verify unassisted thought.

Learning Strategy: Build from Scratch. To understand LLMs, one must code a small one (GPT-2 scale) from scratch. The struggle is the learning mechanism.

THE IRON LAW OF INFRASTRUCTURE

The Physical Constraints of Digital Intelligence



The Reliability Crisis:
At 100k scale, hardware failure is a daily guarantee. Training is now a systems engineering challenge.

Pre-Training Economics

Type: Fixed Cost
Price: $5M - $10M+ per run.

Inference Economics

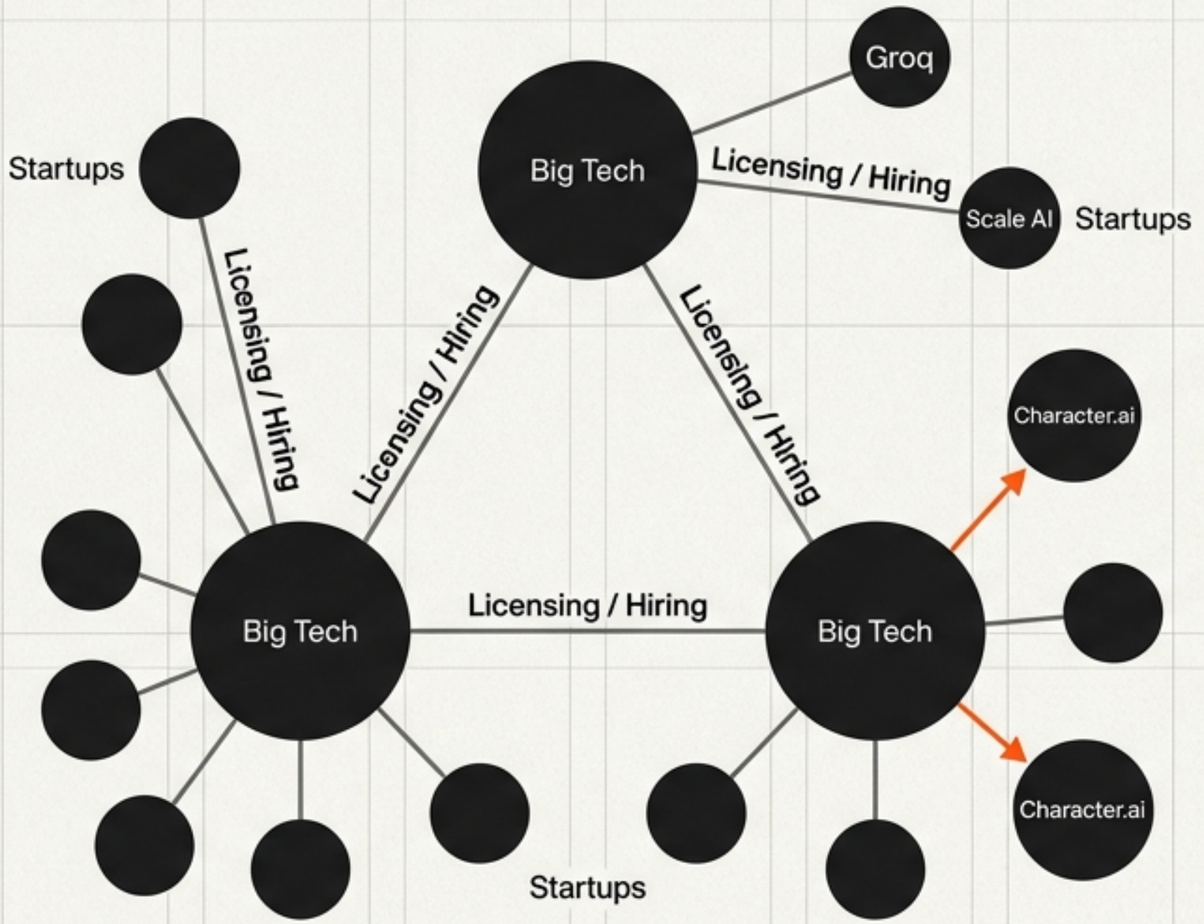
Type: Recurring Cost
Price: Billions/Year.
Result: Drives massive push for Distillation (Teaching small models to mimic huge ones).

MARKET DYNAMICS

Pseudo-Acquisitions and the IPO Drought

The Trend: Consolidation disguised as Licensing. disguised as Licensing.

Mechanism: Big Tech hires the team and licenses the tech to bypass antitrust antitrust scrutiny of full acquisitions.

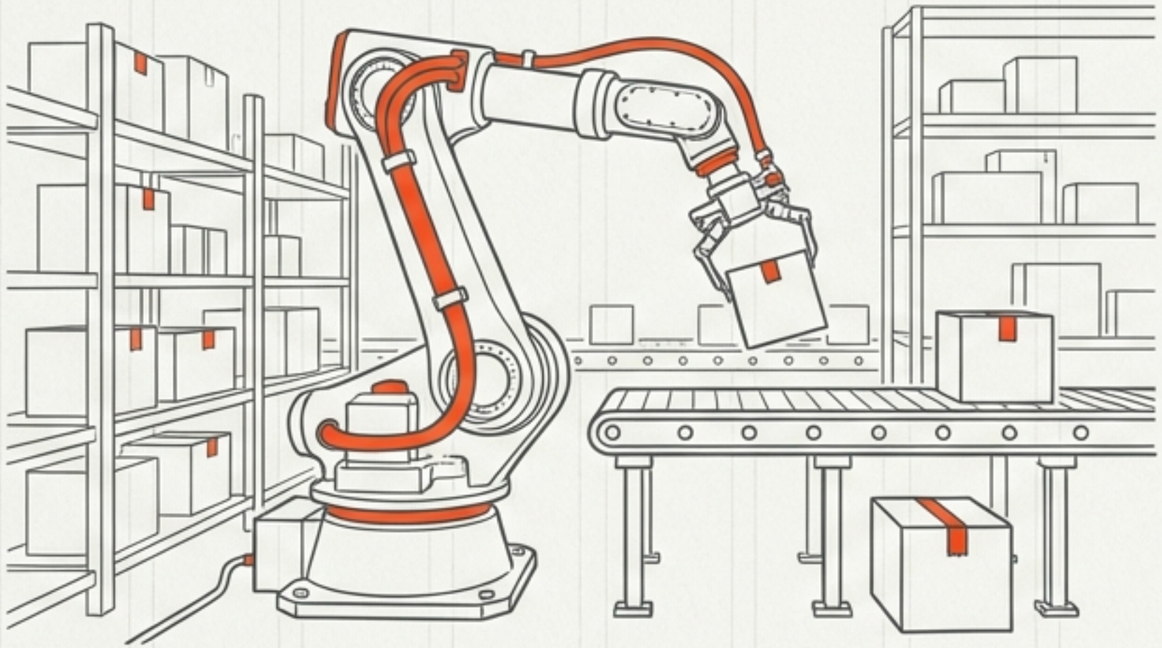


The IPO Drought: Private capital abundance allows OpenAI/Anthropic to stay private. No pressure to face public markets while burning cash on compute.

EMBODIED AI & ROBOTICS

Closing the Simulation Gap

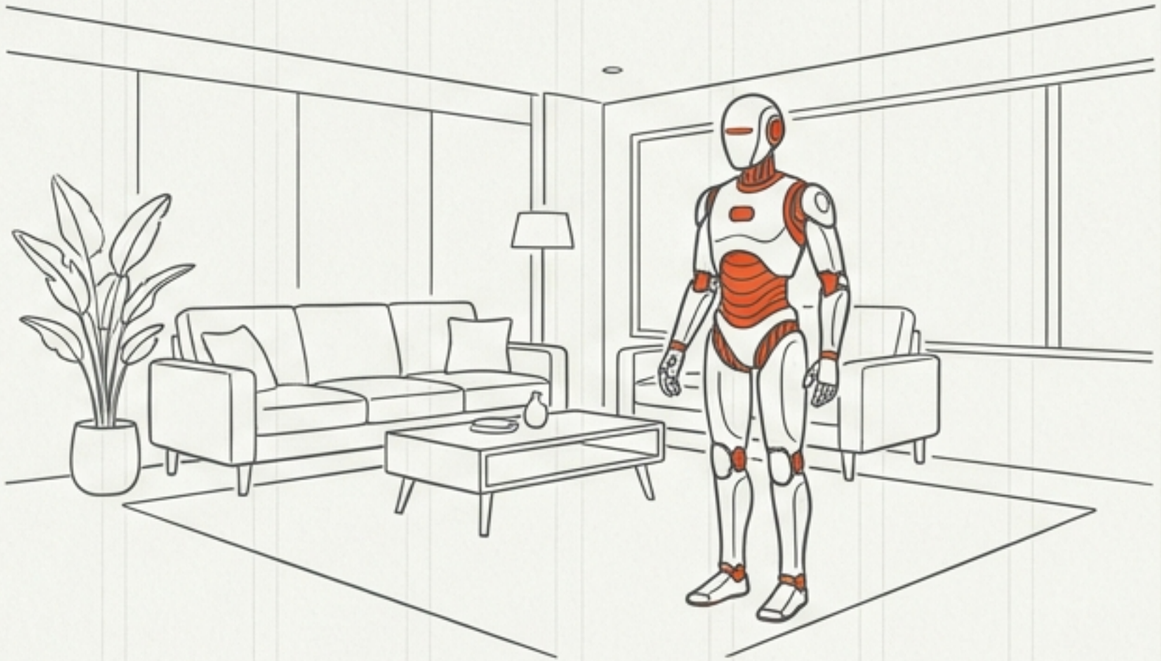
Industrial Automation



Bullish.

Controlled environments (e.g., Amazon Warehouses). Sim-to-Real gap closing via Foundation Models (RTX).

Home Robotics



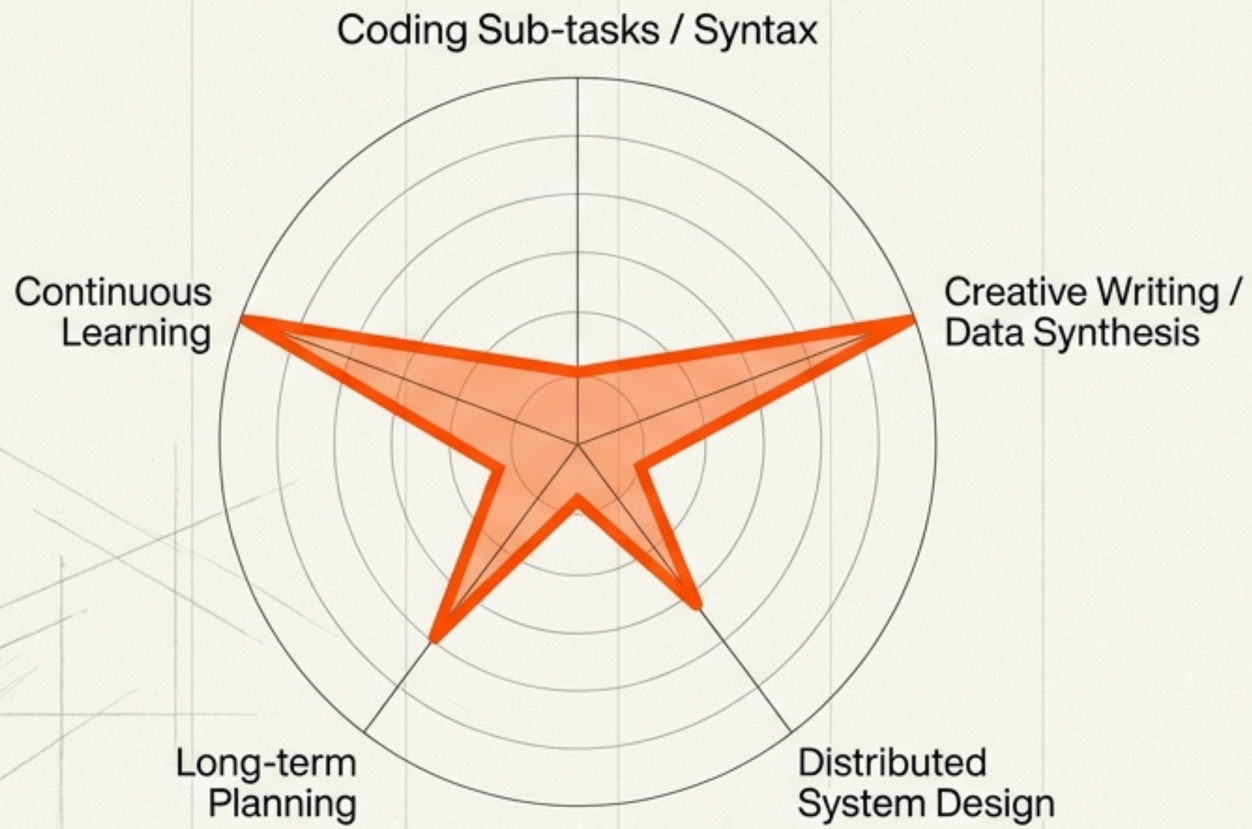
Bearish.

The Safety Bottleneck. “You are allowed to fail never in a home.” Edge cases remain unsolved.

REDEFINING AGI: JAGGED INTELLIGENCE

Moving Beyond the Binary Singularity Myth

Jagged Frontier Radar Chart



The 2026 Reality:
Intelligence is not uniform.

The Remote Worker Test:
Can AI replace a remote worker? No—it lacks continuous feedback loops.

Economic Impact: Growth will come from “Industrialized Software,” enabling non-coders to build complex systems.

THE HORIZON

Fragmentation and Specialization

01.

The Engine

RLVR and Inference Scaling are the primary drivers of progress. The "One Model to Rule Them All" dream is fading in favor of specialized agents.

02.

The Market

A permanent bifurcation: US dominates Revenue/Product; China dominates Influence/Open Weights.

03.

The Imperative

Find Agency. Don't just consume the "slop." Build, code, and curate to stay on the right side of the jagged frontier.