The Agent Era: A Compute Story

Multi-step reasoning isn't just magic. It's a cost threshold being crossed.

The Human Intent: Spawns an agent loop to build software, review code, and iterate.

The Inference Engine: Each step costs tokens. GPT-4's blended token price fell 79% annually in just 17 months.

$30 → $0.15: The drop in million-token cost that made autonomous agent loops economically viable for everyone, moving from an expensive luxury to flash-tier ubiquity.

Gemini Notebook

The Always-On Companion

Personalization requires continuous, ambient inference.

The User: Experiences a companion that understands time, history, and context naturally, without needing explicit prompting.

Continuous Context: Observing and learning in the background requires 24/7 continuous processing.

Breaking the Chat Ceiling:
AI companions work today because always-on multi-modal inference dropped from $100+/month in pure API costs to a mere rounding error.

Gemini Notebook

Disposable Software & The End of SaaS

When inference is cheap, custom solutions beat generic subscriptions.

The Domain Expert: A non-technical user iterating through conversation to build a perfect, custom workflow.

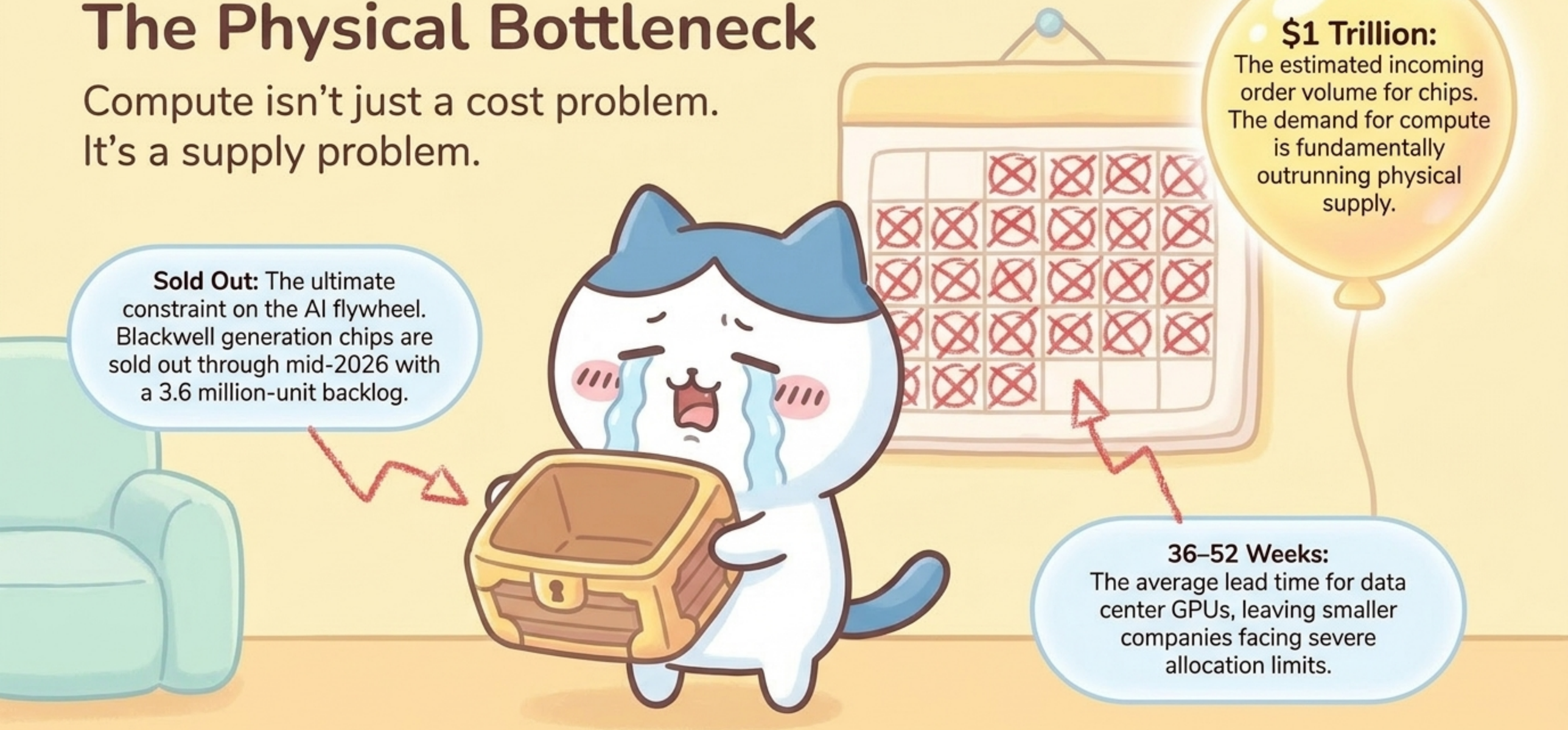
Bespoke Tools: Why pay $500/month for a generic SaaS platform when an AI agent can build an exact, custom fit in a single hour of iteration?

The Printing Press Moment: Democratized creation isn't a software breakthrough—it is democratized compute.

Gemini Notebook

The Physical Bottleneck

Compute isn’t just a cost problem.
It’s a supply problem.



Gemini Notebook

The Super Individual

Managing parallel streams of intelligence.

The Conductor: One person executing the workload of an entire company. At $3/MTok, ten continuous AI streams cost less than a single junior engineer.

10 Parallel Streams: Engineering, marketing, operations, and design runight running continuously in tandem. By 2026, inference will account for two-thirds of total compute workloads.

The Root Variable: Every AI trend—agents, companions, super individuals—is just a symptom. The underlying engine is compute getting cheaper. Understand the **compute curve**, and you understand the future.

Gemini Notebook