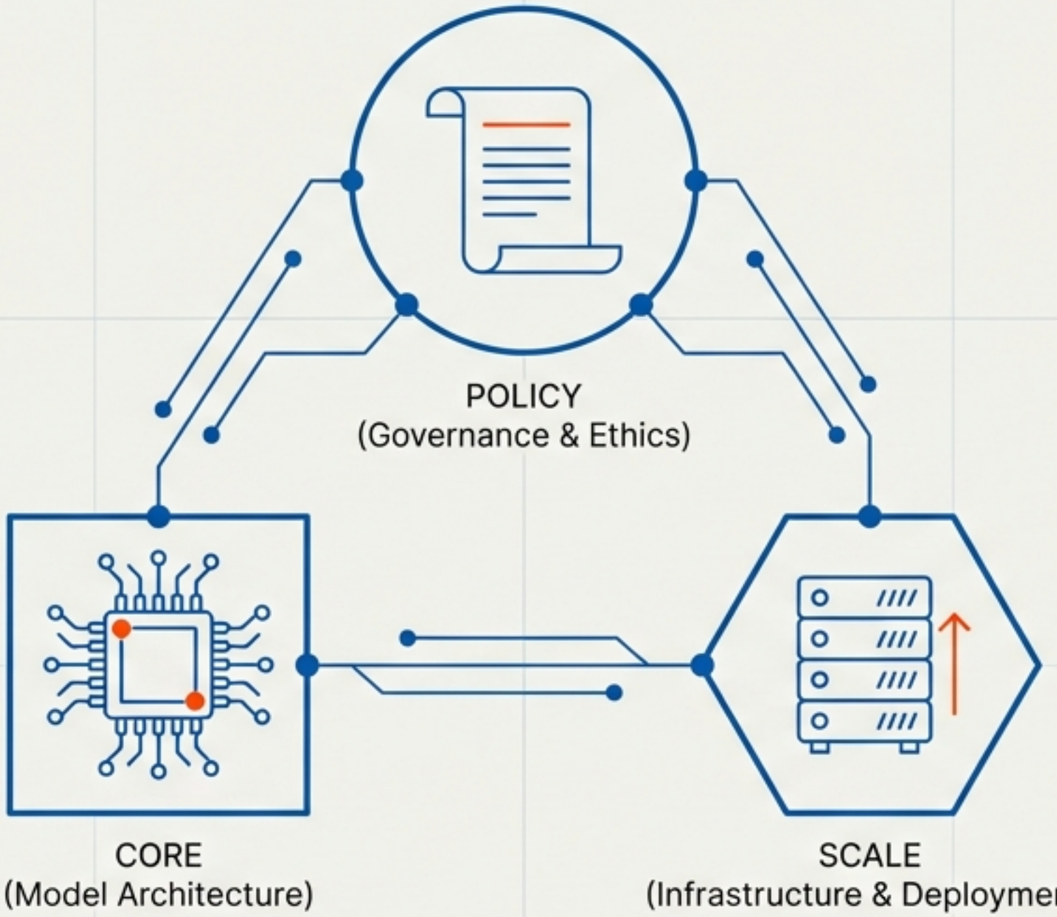
AI Engineering Report: 2026-01-25

Policy, Architecture, and Infrastructure Trends



The State of AI Engineering: Four Strategic Pillars

1. Governance & Culture (ガバナンスと文化)

Moving from simple bans to “Disclose & Verify” frameworks. Shifting focus from fear to responsibility.

2. Agent Architecture (エージェントアーキテクチャ)

Understanding internal loops, context compaction, and parallel execution patterns.

3. Infrastructure & Efficiency (インフラと効率化)

Challenging cloud costs with local stacks, bare-metal density, and recycled heat.

4. Standards & Impact (標準と影響)

Re-evaluating data strictness (XML vs JSON) and the economic reality of AI adoption.

Moving Beyond Bans: The "Disclose & Verify" Standard

Case Study: Ghostty Terminal Emulator Policy

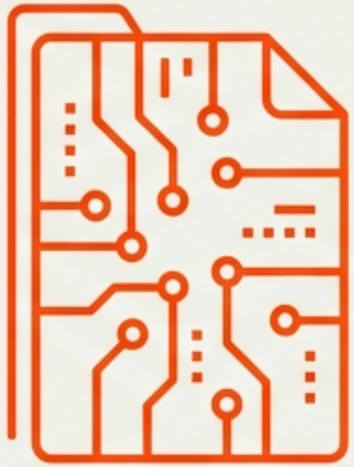
- **Disclosure is Mandatory (開示義務):** Contributors must explicitly list tools used (e.g., Claude Code, Cursor) and the scope of assistance.
- **No “Drive-by PRs” (ドライブバイPRの禁止):** Low-effort PRs generated without context are rejected. Changes must be tied to pre-approved issues.
- **The “Lazy User” Problem:** The policy explicitly states the issue isn’t AI itself, but the lack of human verification.
- **Verification (検証):** Code must be tested by the human author. Generating code for platforms the author cannot verify is prohibited.

“AIを禁止しない。問題なのは『質の低いAIユーザー』である。”

Defining the Contract: The Rise of AGENTS.md

CONTRIBUTING.md

For Humans (ヒト向け)



AGENTS.md

For AI (AI向け)

The Perception Gap

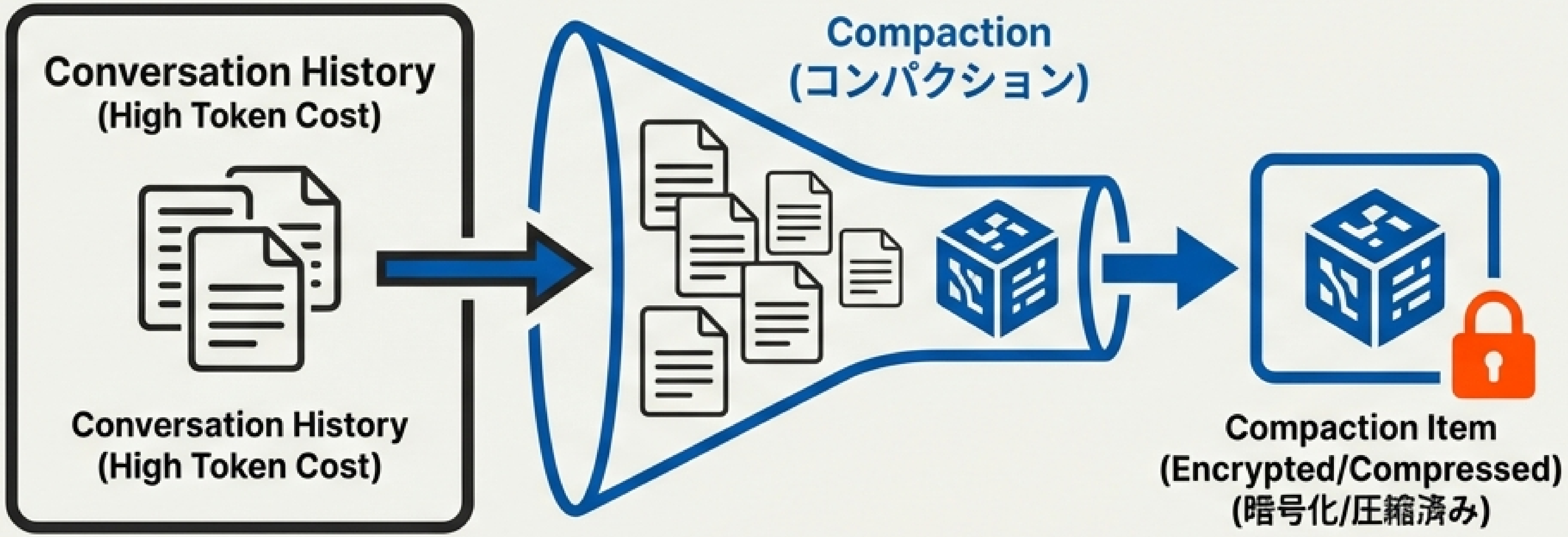
- **The 'Dark Signal' View:** Does this file imply an agent messed up the code?
- **The 'Guardrail' View:** A necessary protection layer to externalize project wisdom.

Implementation Goals

- **Role:** Explicitly document guardrails, context limits, and known pitfalls.
- **Goal:** Prevent hallucinations and regression.

Decision Hint: Use AGENTS.md to clarify the boundary between 'AI-authorized scope' and 'Human-only review scope'.

Anatomy of an Agent: Inside the Codex Loop



Technical Deep Dive

- **Compaction:** Compressing history into encrypted items to save tokens while keeping latent understanding.
- **The Bottleneck:** Detailed inference steps are ephemeral, explaining why 'memory' degrades in long sessions.

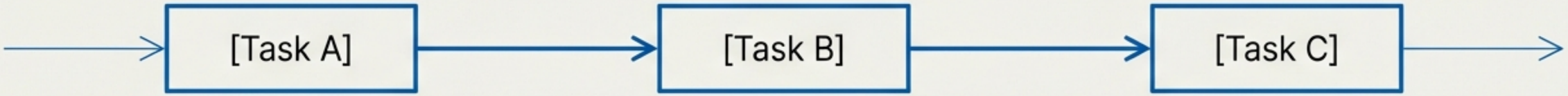
Token Lifecycle

Inference Tokens (Ephemeral)
Discarded after user turn.

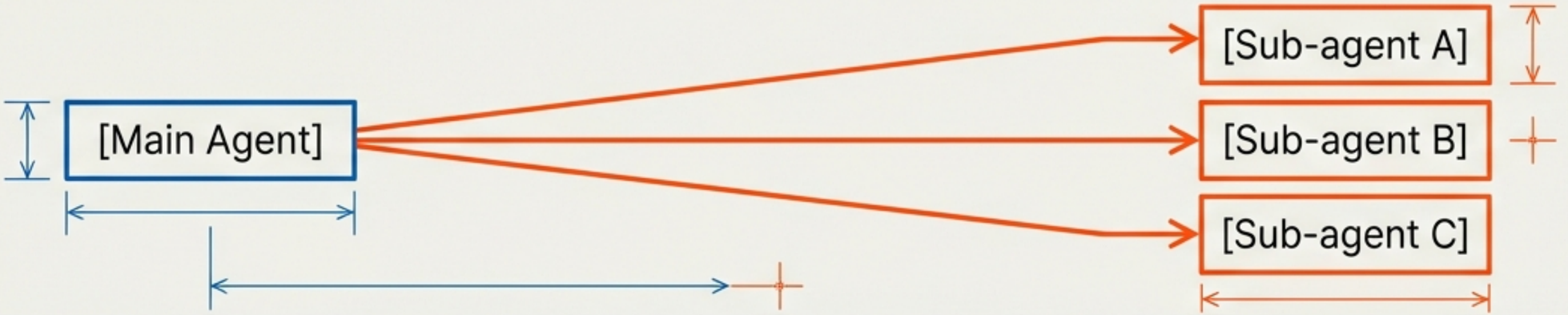
User Context (Persistent)

Scaling Through Parallelism: The ‘Swarm’ Pattern

Serial Execution (Standard)



Parallel Swarm Execution (Claude Code)

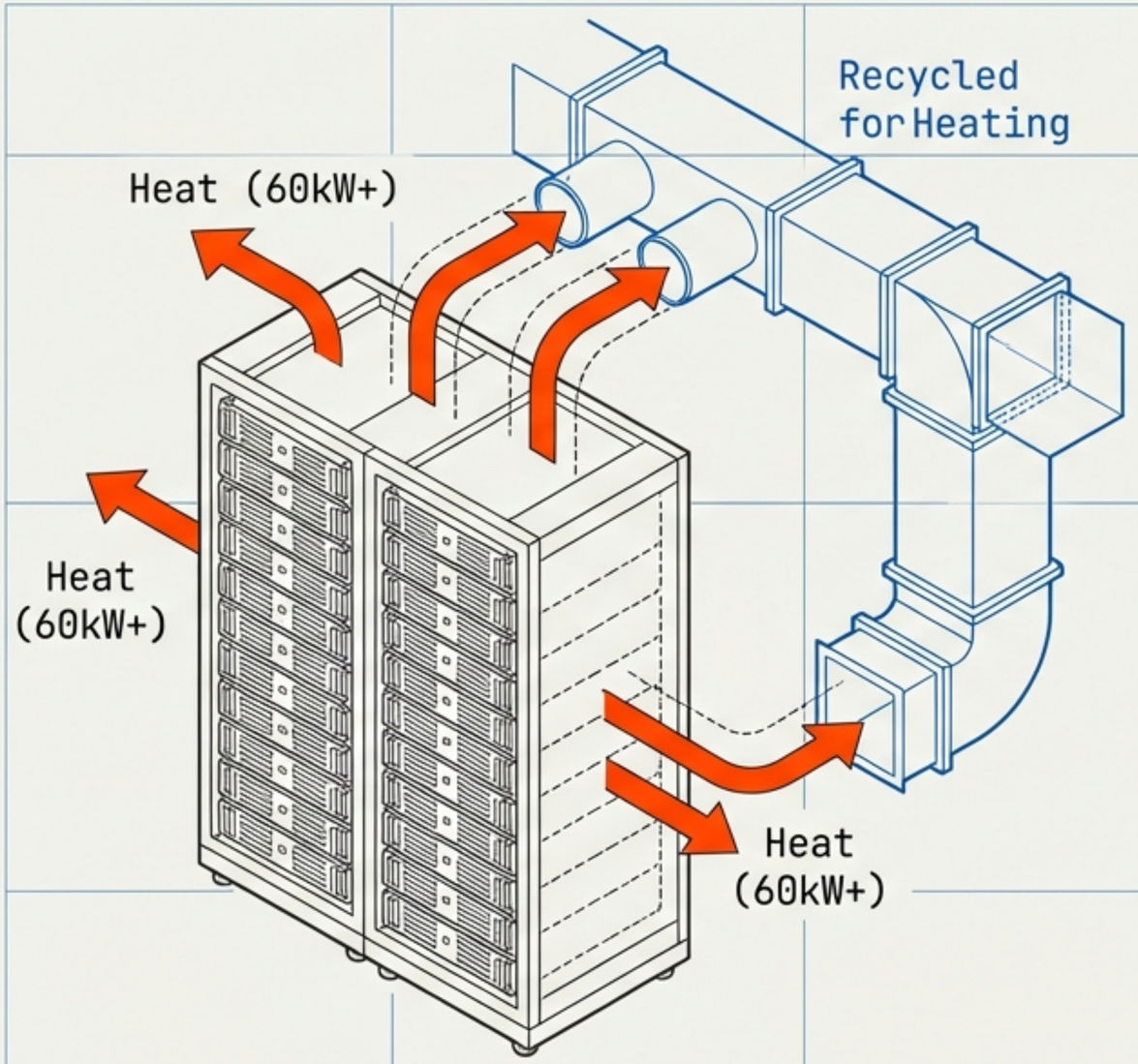


The Engineering Trade-off

Pros	Cons
■ Significant speed increase for large codebases. ■ Simple natural language orchestration.	■ Higher token cost (Parallel consumption). ■ Review Debt: Verification load spikes massively.

Takeaway: Use for independent tasks. For interdependent changes, serial execution is safer.

High-Density Efficiency: The Internet Archive Case Study



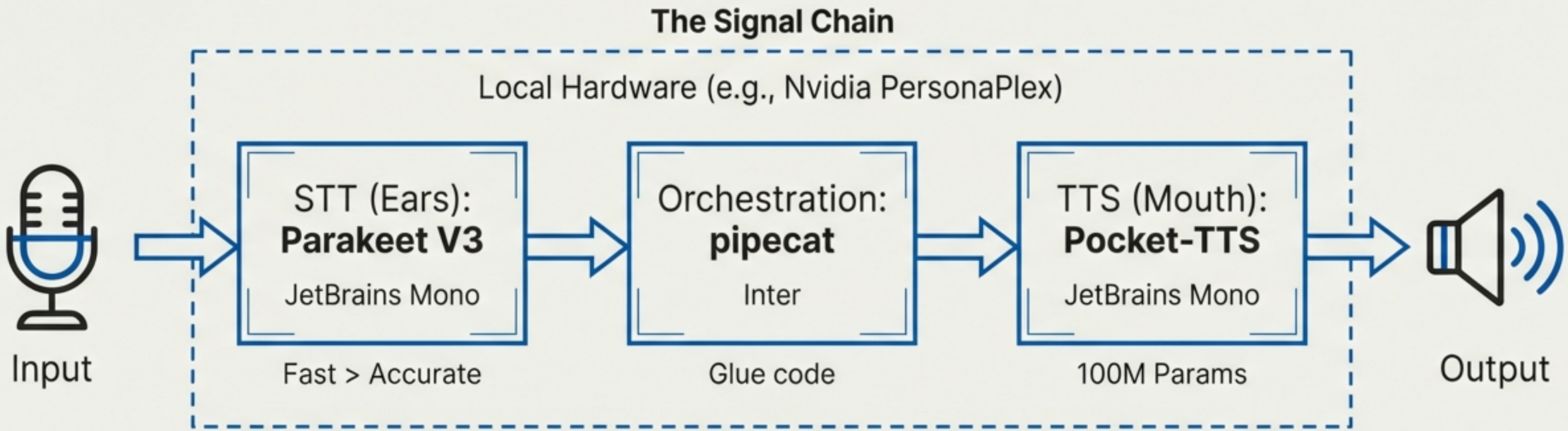
Budget: ~2.5–3.0 Billion JPY/year for global scale.

Architecture Innovation:

- Custom PetaBox Design: Optimized for density.
- **HVAC-Free**: Using waste heat to warm the building.
- **Open Source Stack**: Zero licensing fees.

Lesson: For massive, long-term data (like AI training sets), the **"Cloud Default" is a financial trap.**

The Local AI Stack: Voice at the Edge



Context:

Privacy and latency are driving Voice AI to the edge.

Performance:

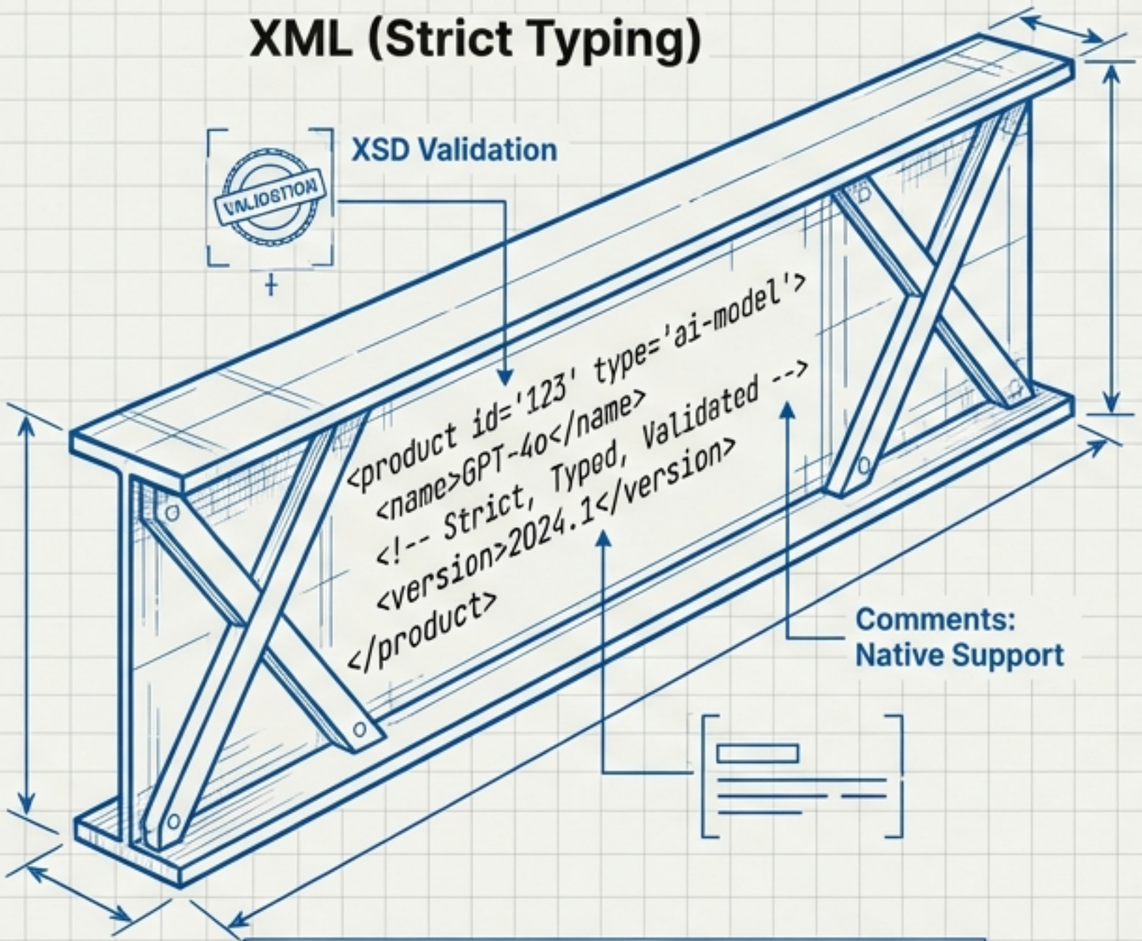
Single-GPU pipelines are now viable for real-time conversation.

Practical Tip:

Have the LLM "repeat back" STT input to verify accuracy.

Revisiting XML: The Lost Value of Strict Schemas

XML (Strict Typing)



Archaeological Blueprint: Structured Foundation

The Argument:

The industry sacrificed robustness for convenience (JSON), losing features critical for AI correctness:

- **XSD Validation:** Document-level strict typing.
- **Namespaces:** Collision-free composition.
- **Comments:** Native support (missing in JSON).

Verdict:

Don't rewrite in XML, but recognize that modern AI needs the the strictness (like TypeScript/Zod) that XML originally solved.

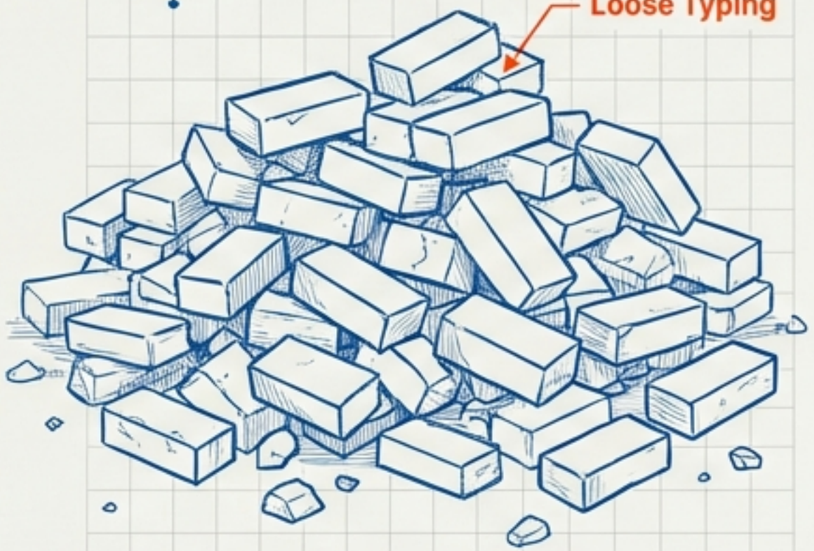
JSON (Flexible/Loose)

```
{ 'id': 123,
  'type': 'ai-model',
  'name': 'GPT-4o',
  'version': '2024.1'
} // No Comments!
```

No Comments

No Namespaces (Collision Risk)

Loose Typing

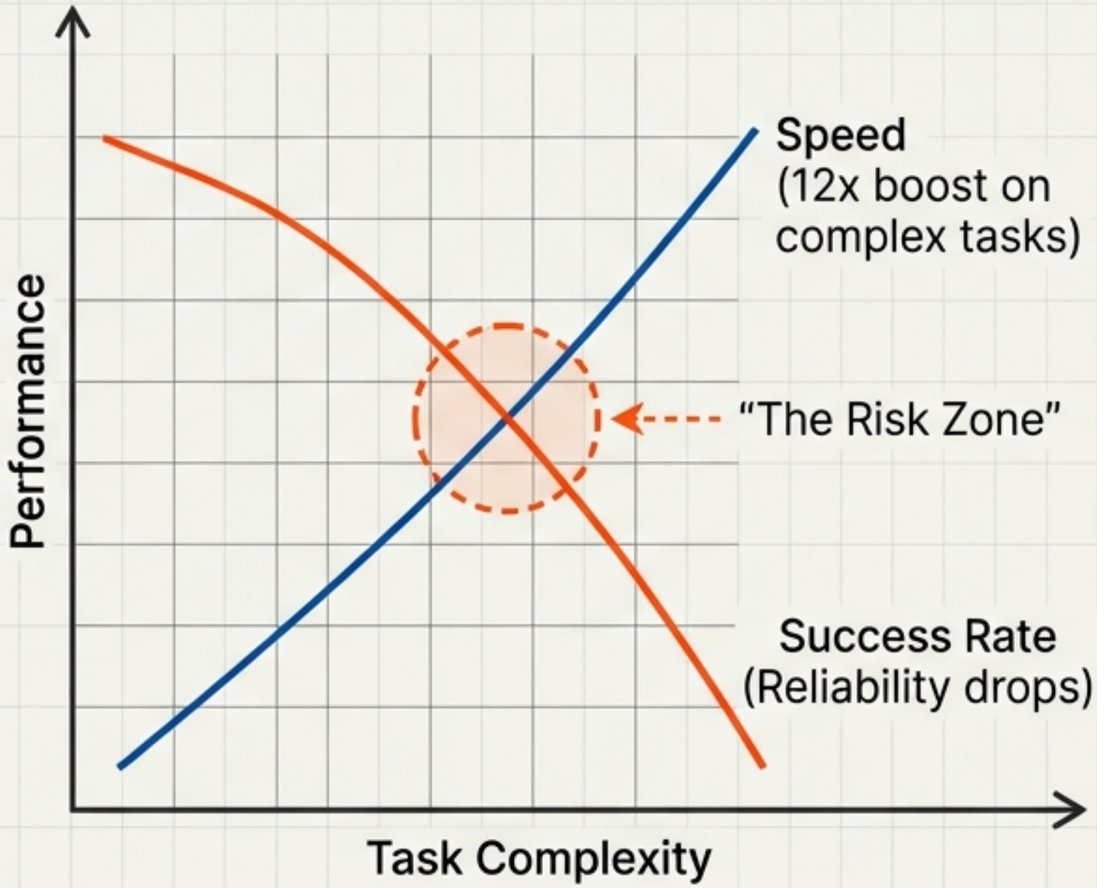


Archaeological Blueprint: Loose Aggregate

SCALE	PROJECT:	DRAWING NO.	APPROVED BY:
1:1	AI INFRASTRUCTURE ARCHAEOLOGY	3 OF 3	AI ARCHITECT

The Economic Reality: Speed vs. Reliability

Source: Anthropic Economic Primitives Report



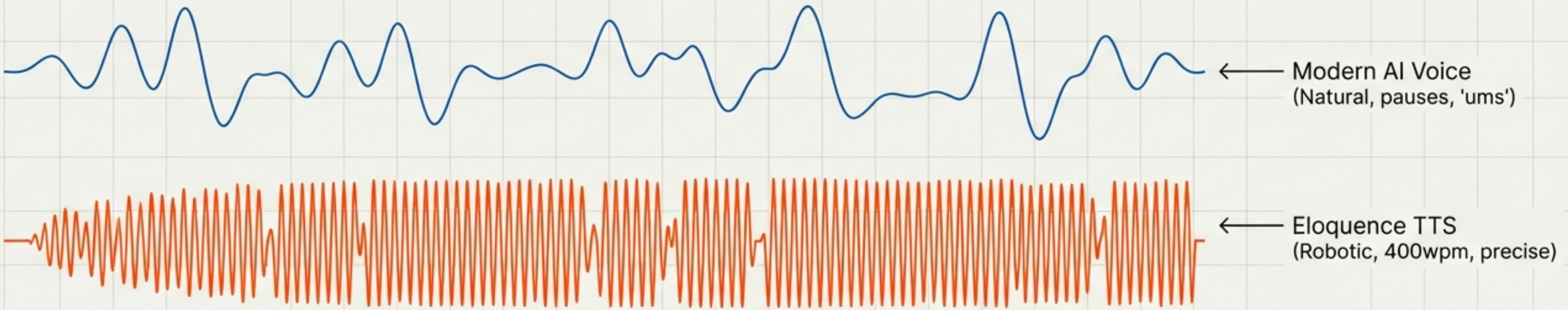
Data Analysis:

- **Speed Multiplier:** 12x speedup on University-level tasks.
- **The Reliability Gap:** As complexity rises, success rates plummet.
- **Real Productivity:** 1.0–1.2% annual boost (Not magical, but steady).

Trend:

"De-skilling"—AI takes high-skill components, humans are left with low-skill verification.

The ‘Human-Like’ Trap in Accessibility



The Conflict:
Power users (screen readers) listen at 400+ wpm. Natural AI pauses slow down information intake.

Legacy Tech:
Why 2003's "Eloquence" engine is still superior for efficiency.

Lesson:
Product design must prioritize user workflow over technical realism. "Better" isn't always efficient.

SCALE	PROJECT NAME	DRAWING NO	APPROVED BY
	AI INFRASTRUCTURE ARCHAEOLOGY	3 OF 3	AI ARCHITECT

Beyond Static Media: Real-Time Generative Video

Technology Profile: Waypoint-1

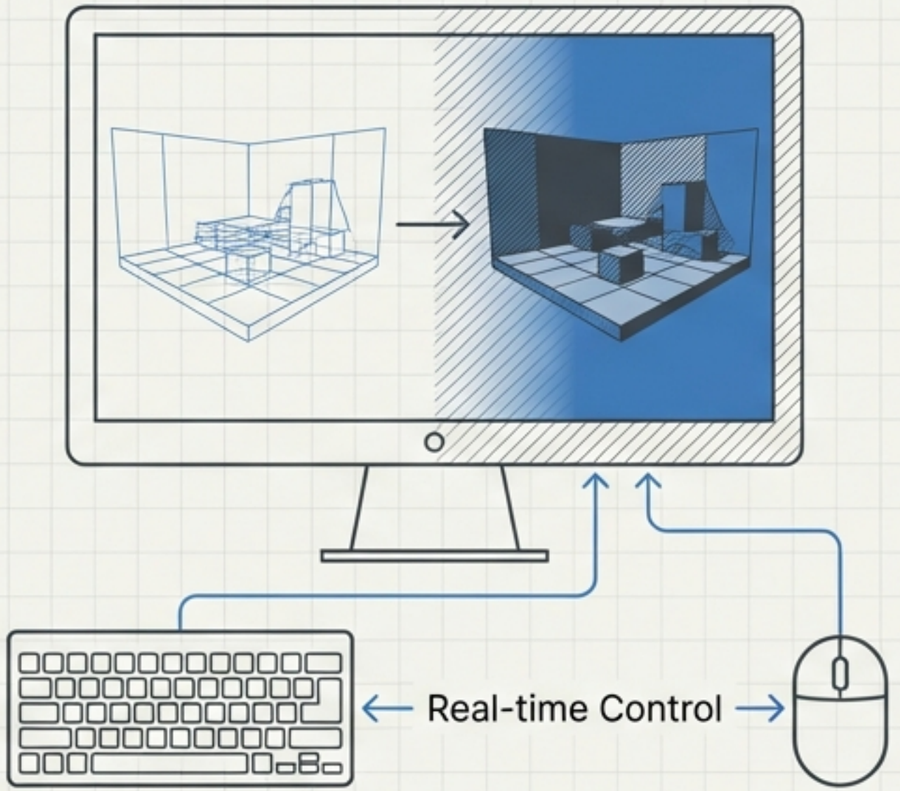
Technique: Diffusion Forcing for frame consistency.

Specs:

- 30 FPS on RTX 5090 (4 steps).
- **Input**: Text + Keyboard + Mouse.

Use Case: Instant game prototyping. "Playing" a video generated on the fly.

Status: Short context retention, but proves Neural Rendering is viable.



GENERATION PIPELINE: WAYPOINT-1 PROTOTYPE			
SCALE	PROJECT NAME: REAL-TIME NEURAL RENDERING	DRAWING NO: 1 OF 1	APPROVED BY: AI ARCHITECT

Technical Terminology Reference

Drive-by PR:
Low-effort, context-free pull requests, often AI-generated.

Compaction (Codex):
Compressing conversation history into encrypted tokens to save space.

PetaBox:
High-density, open-hardware storage unit by Internet Archive.

Diffusion Forcing:
Training method for consistent real-time video generation.

Pipecat:
Open-source orchestration framework for voice AI pipelines.

XSD (XML Schema Definition):
Strict typing and validation system for XML documents.

Strategic Implications for 2026

- **Governance:** Adopt **‘Disclose & Verify’** immediately. Treat AI code as **‘untrusted until tested’**.
- **Development:** Leverage **‘Swarms’** for isolation tasks, but monitor **‘Review Debt’**. Use **Markdown files** to manage **Agent** memory.
- **Infrastructure:** Re-evaluate **‘Cloud Default’** for massive static data. Explore **local stacks** for voice/privacy.

Closing Thought:

The hype phase is ending. We are entering the phase of Engineering Discipline—where costs, correctness, and verification matter more than magic.