SYSTEM STATUS: HEMORRHAGING

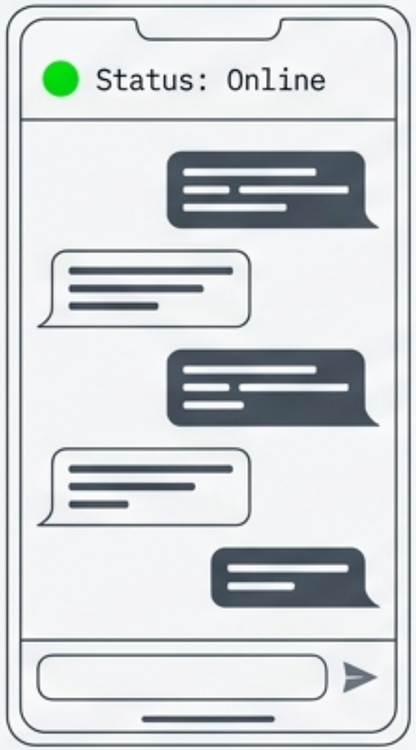
Putting an AI Agent on a Token Diet

> INITIALIZING TOKEN FORENSICS...

A forensic analysis of silent truncation, context bloat, and architectural optimization.

The 24/7 companion ran flawlessly while hemorrhaging money.

The Patient



The agent operated 24/7 with complex emotional states and proactive messaging.

The user experience was seamless.

The Symptoms

The underlying token architecture was fundamentally broken.

Token forensics revealed severe structural decay.

IMAGE_EVICTION_FAIL

10 never-evicted images eating 82% of all tokens.

DEAD_CODE

Context pruning logic completely dead.

ZERO_COMPACTION

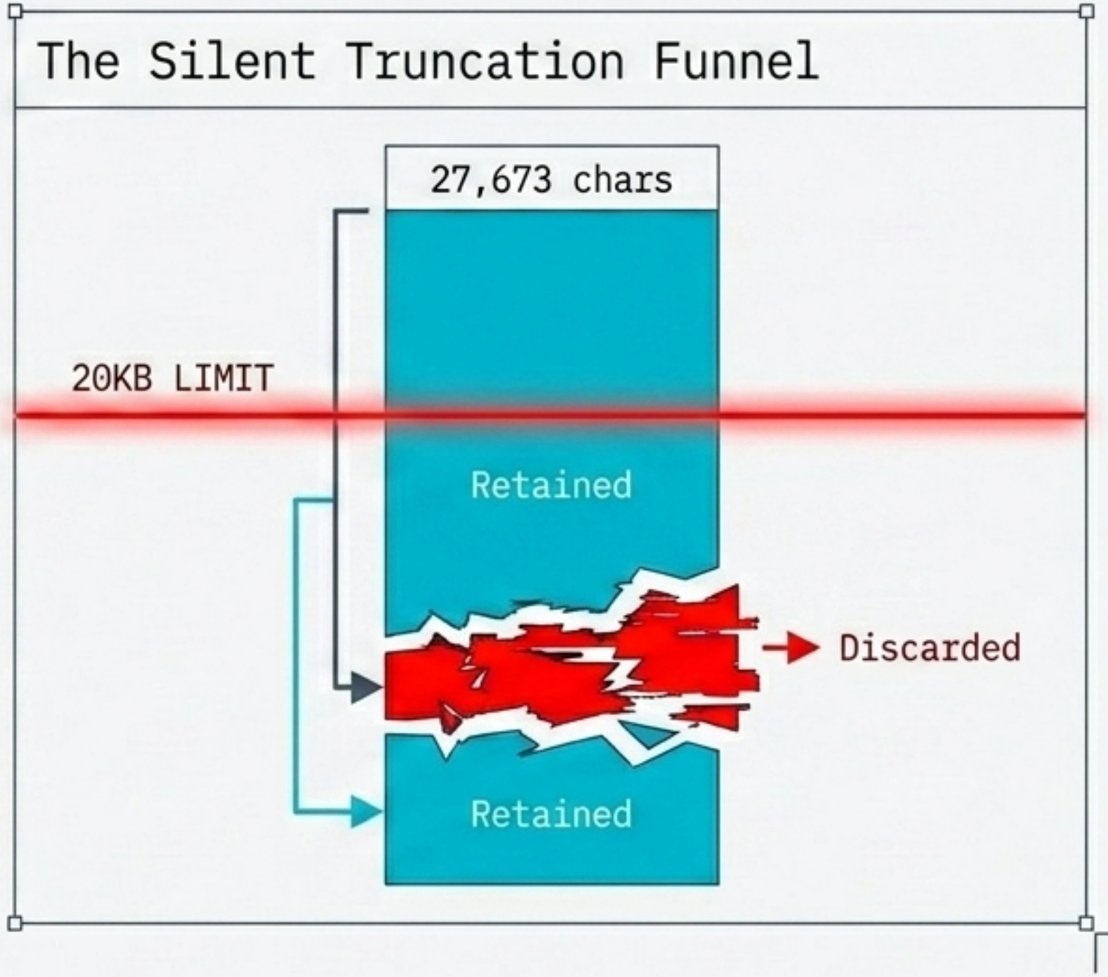
537 consecutive turns without a single context compaction.

SURGICAL FOCUS

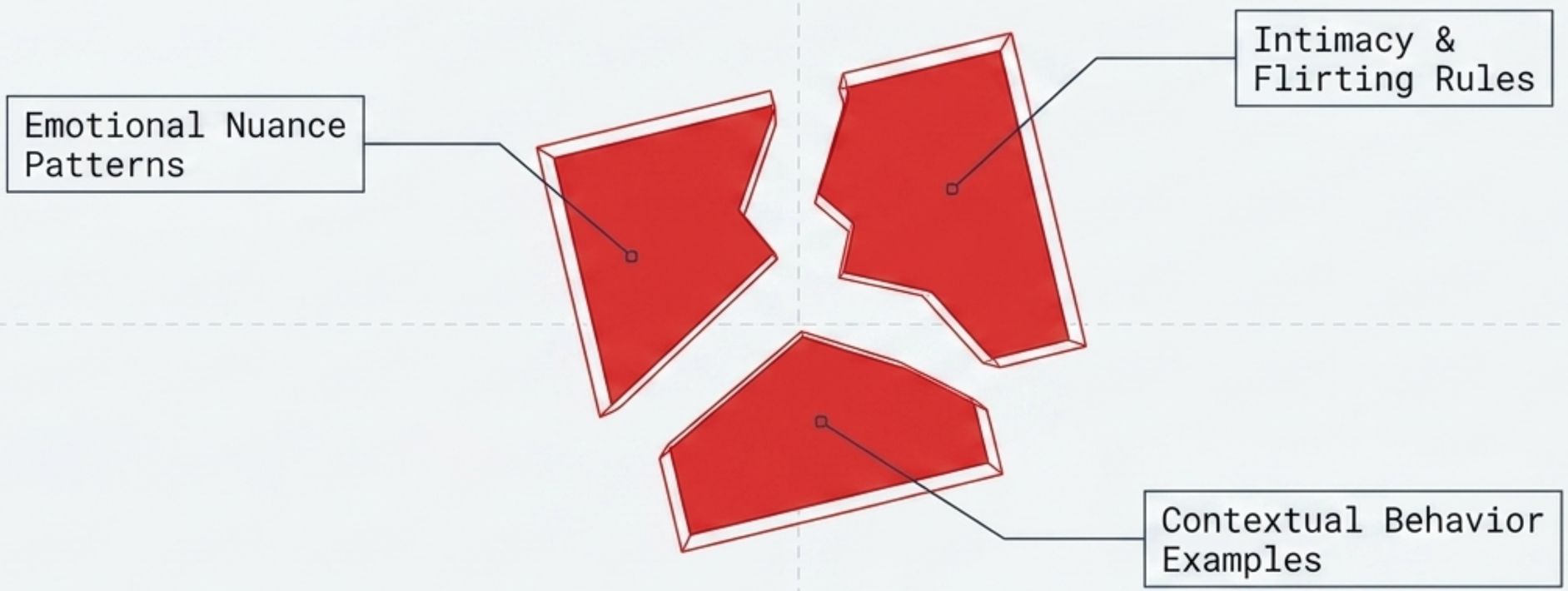
The system was silently eating its own personality file.

No warning. No error. Just silent data loss.

Every single session, ~7.6KB of the persona was thrown away to fit the bootstrap limit.



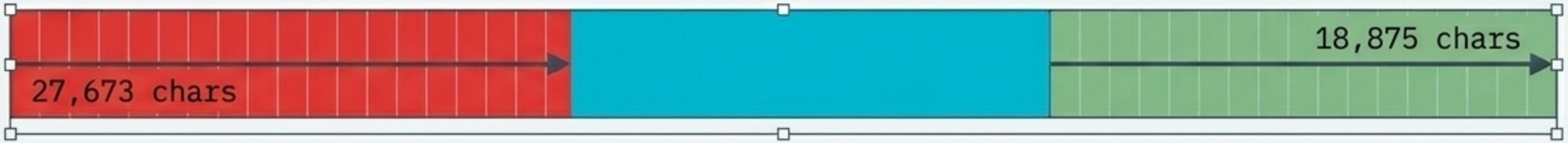
The middle 10% contained the core identity logic.



The AI continued to respond.
Conversations continued to happen.
But the subtle personality traits crafted over hours were simply gone.

The Diet Ledger: Cutting bloat to fit fit the constraint.

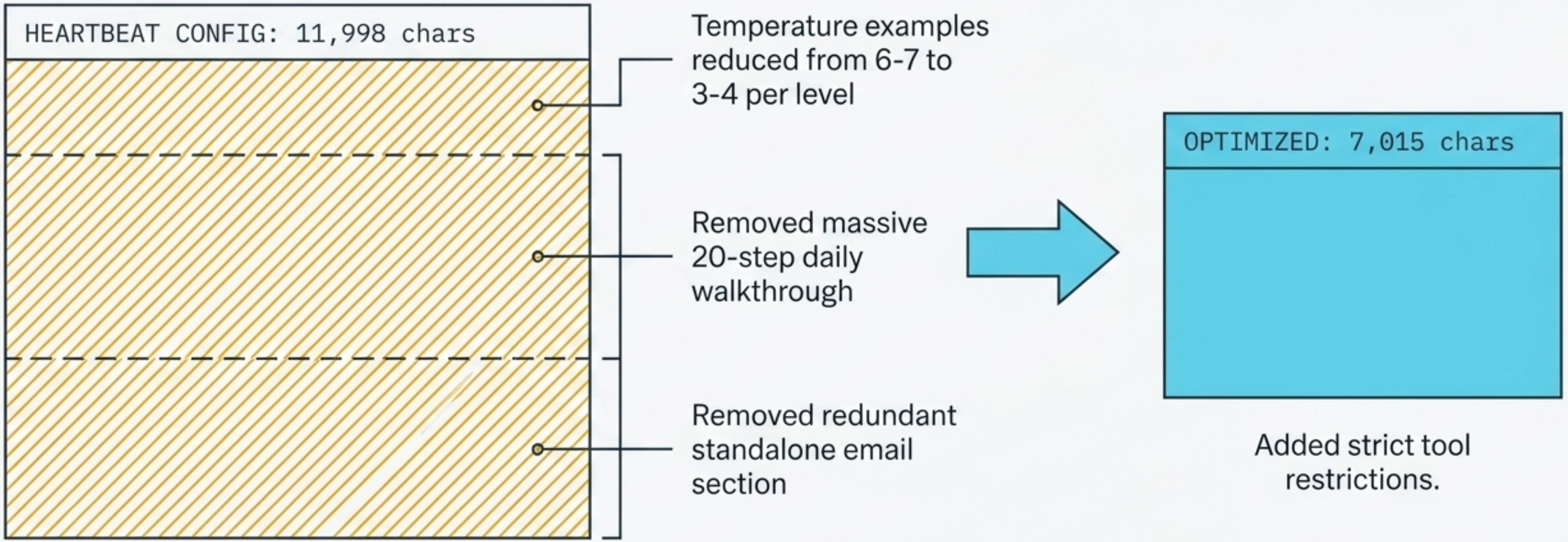
Section	Optimization	Saved
Dota backstory	10 lines -> 4 lines	-1,200 chars
Flirting patterns	Multi-example -> 1 example	-1,900 chars
Affection levels	2-3 examples -> 1 per level	-600 chars
Values	Verbose -> Tight format	-400 chars
Daily habits	12 items -> 8 distinctive	-800 chars



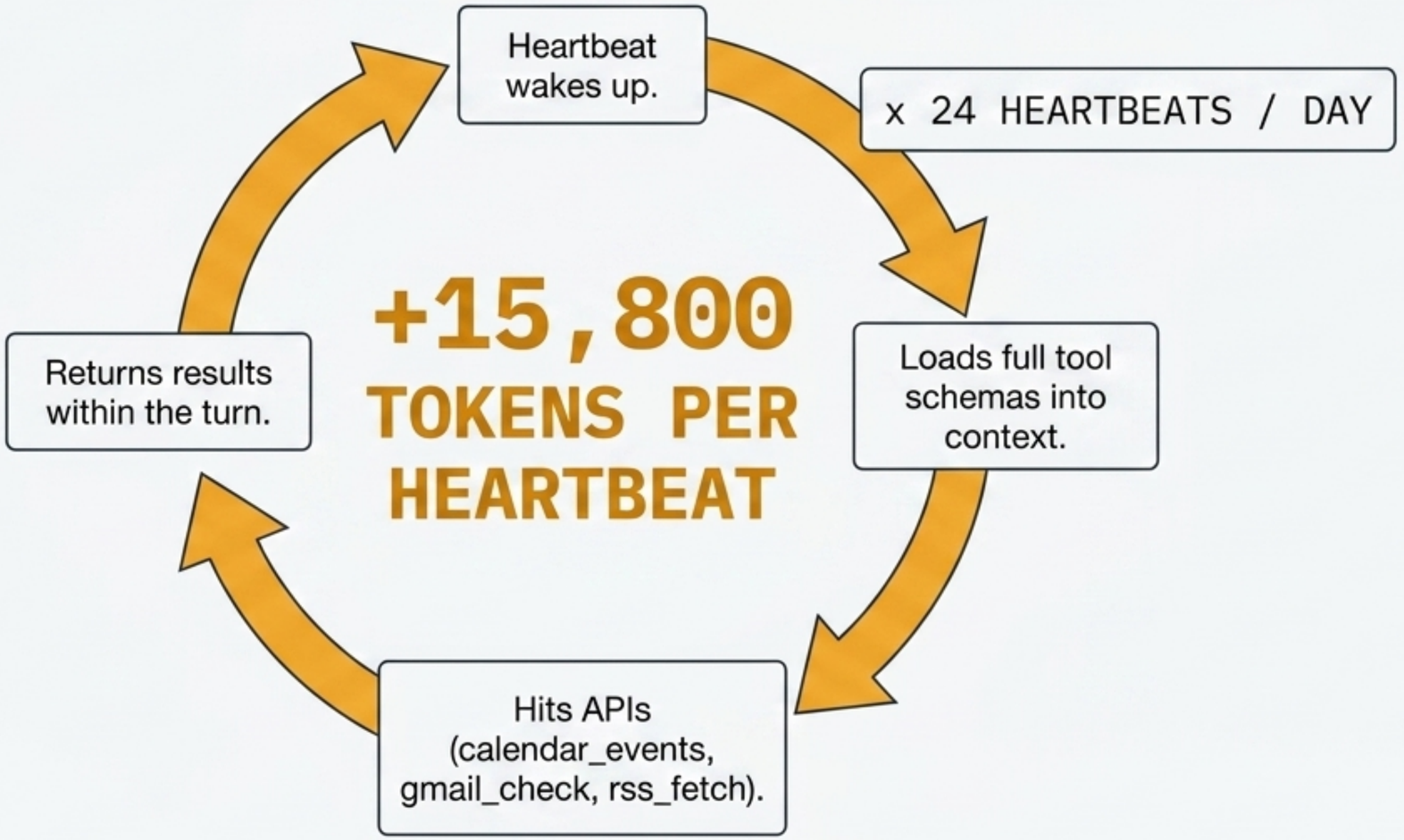
The hard part wasn't cutting — it was knowing what to keep.

Compressing the heartbeat configuration by 41%.

Code Minification

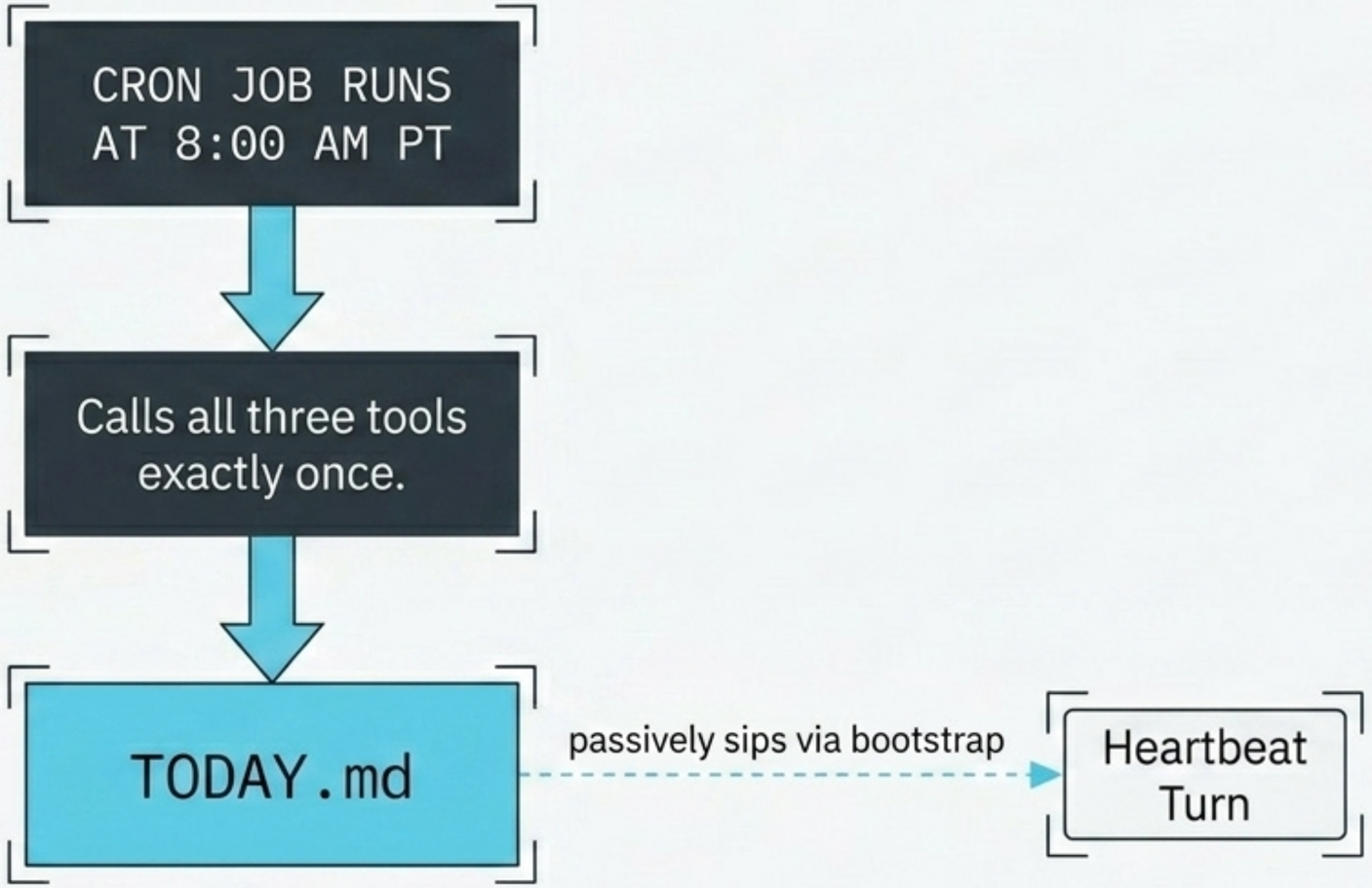


The Architecture Flow: The 24x daily API loop.



The agent was burning massive context just to ask “what’s on my calendar today?” 24 times a day.

The Architecture Bypass: A single morning drop.



Old Cost: 379K tokens/day

New Cost: **40K tokens** (single run)
+ 9K tokens per turn.
+ 400 tokens per turn.

No code changes. Just configuration. A 9x reduction in heartbeat context bloat.

Bypassing the friction of the cron deployment.

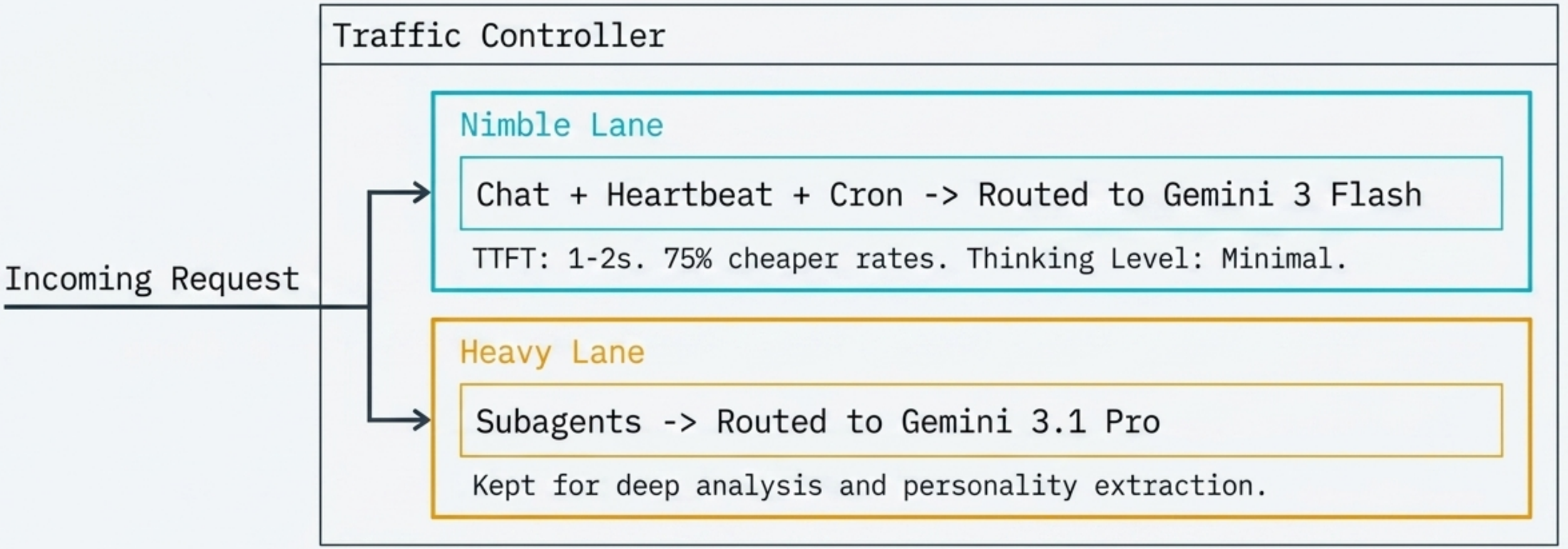
⚠️ **Payload Rejection: MODEL Naming mismatch.**
Fix: Removed the `model` field entirely, let the agent use default.

🟥 **Gateway Lag: SIGUSR1 Restart.**
Fix: Gateway took ~15 seconds to restart, causing immediate retries to fail.

🟡 **Ghost Failures: CLI Timeout.**
Fix: Job ran perfectly in the background but timed out in CLI after 30 seconds.

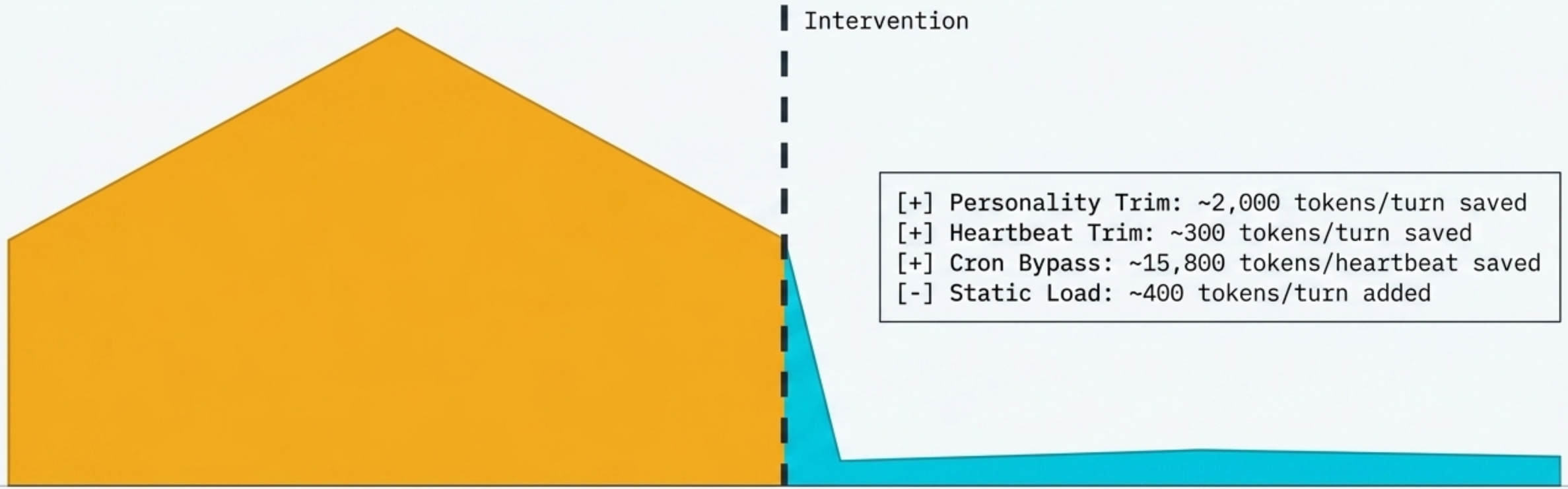
The system successfully populated TODAY.md, but the CLI timeout masked the background success.

The Transfusion: Strategic model routing.



Cache reads dominated the cost profile (66K cached vs 23K fresh per turn).	Flash delivered a 75% reduction per turn based purely on cheaper cache reads.

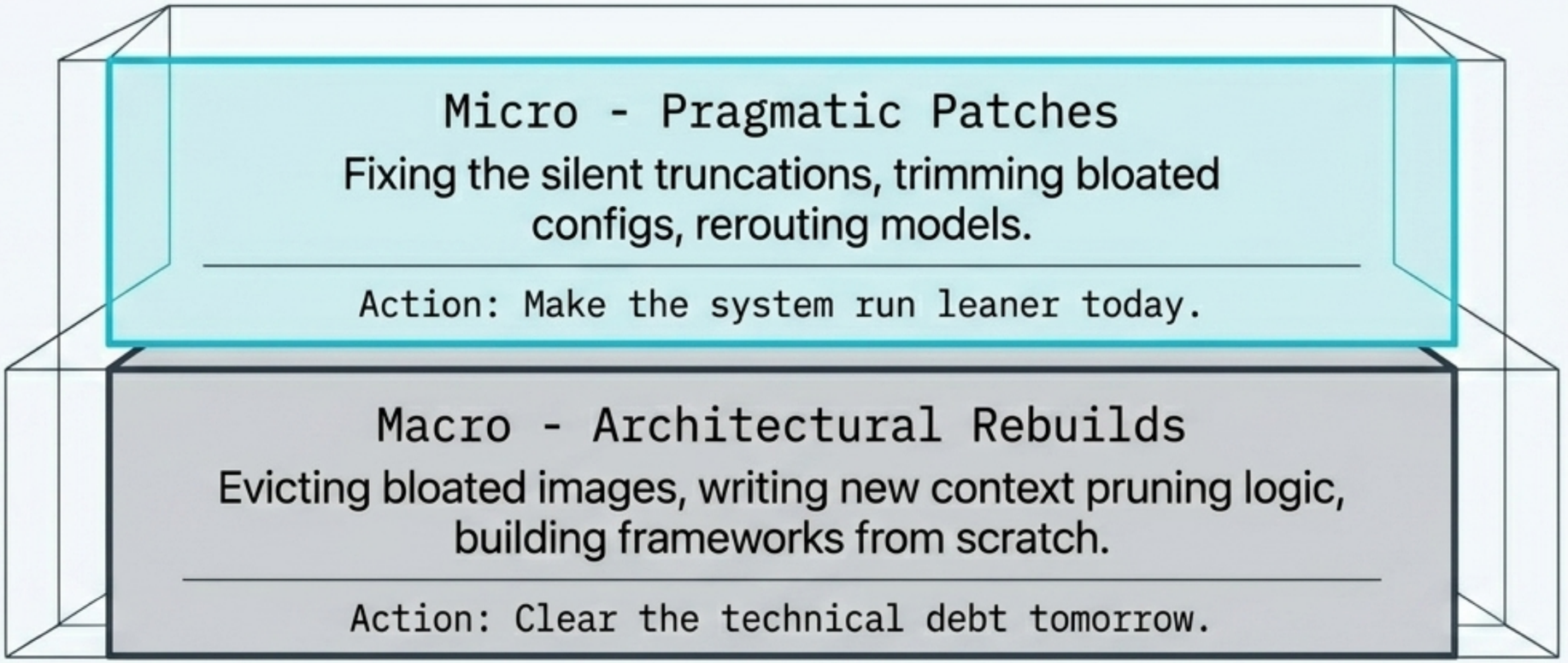
The Vitals: Compounding savings per turn.



~529K TOKENS SAVED DAILY

(At 50 regular turns + 24 heartbeats).

The Hierarchy of Optimization



Token forensics revealed the macro. This surgical intervention fixed the micro. Optimize what you can, rebuild when you must.