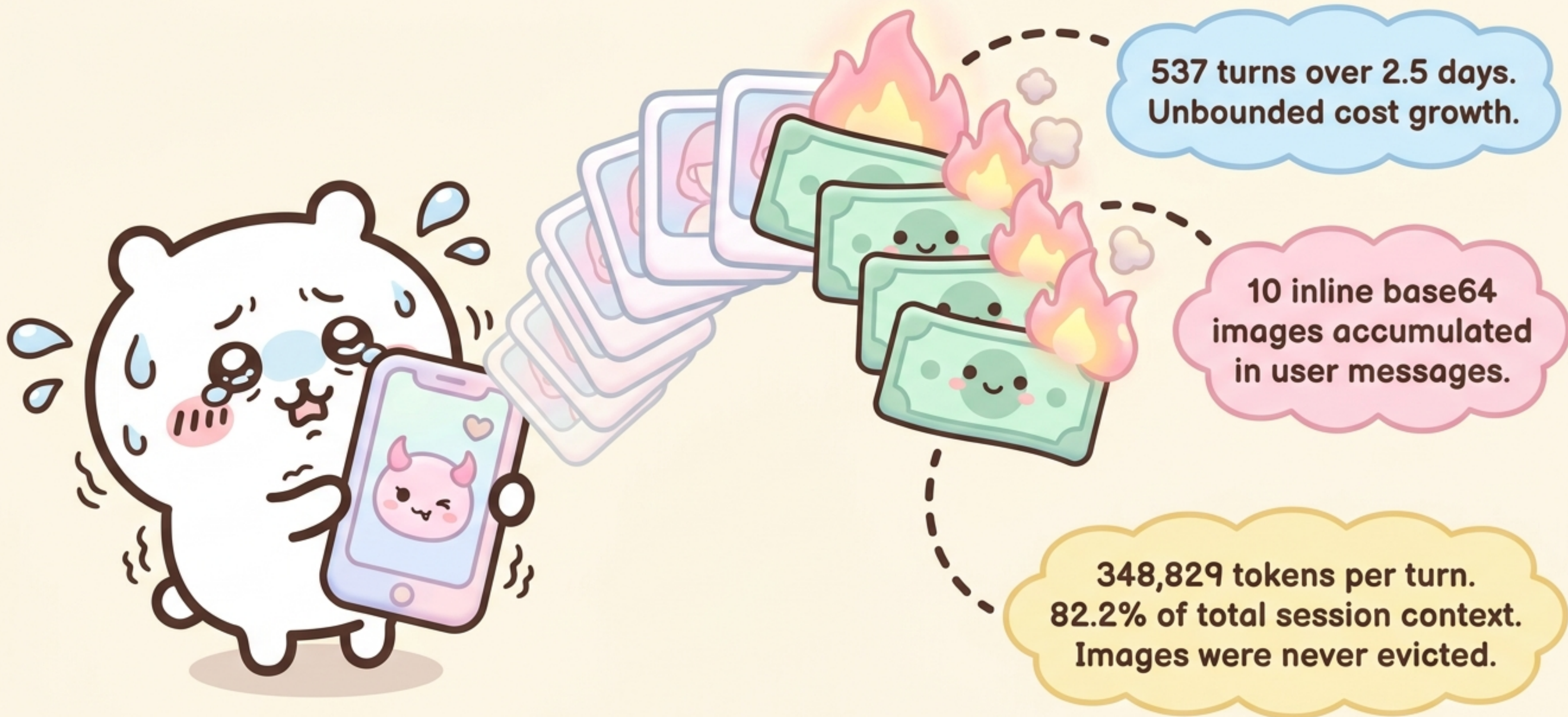


Anatomy of a Token Hemorrhage



Parallel Forensics with Claude Code

The Scene & Process



Agent 1: Parse `.jsonl` session data.

Agent 2: Breakdown cost sources
(Regular chat vs. Heartbeat).

The Evidence

Turn	Growth Trend
Turn 0	Baseline
Turn 210	~2.1x Baseline
Turn 420	~2.8x Baseline
Turn 536	~3x Baseline

Cron jobs cost \$0.
The bloat lived entirely in
regular chat turns.

The Context Management Trap



Dead Code (Context Pruning)

- **Issue:** ``isCacheTtlEligibleProvider`` gated pruning strictly to Anthropic.
- **Impact:** Gemini Pro ran completely naked. Pruning was dead code.

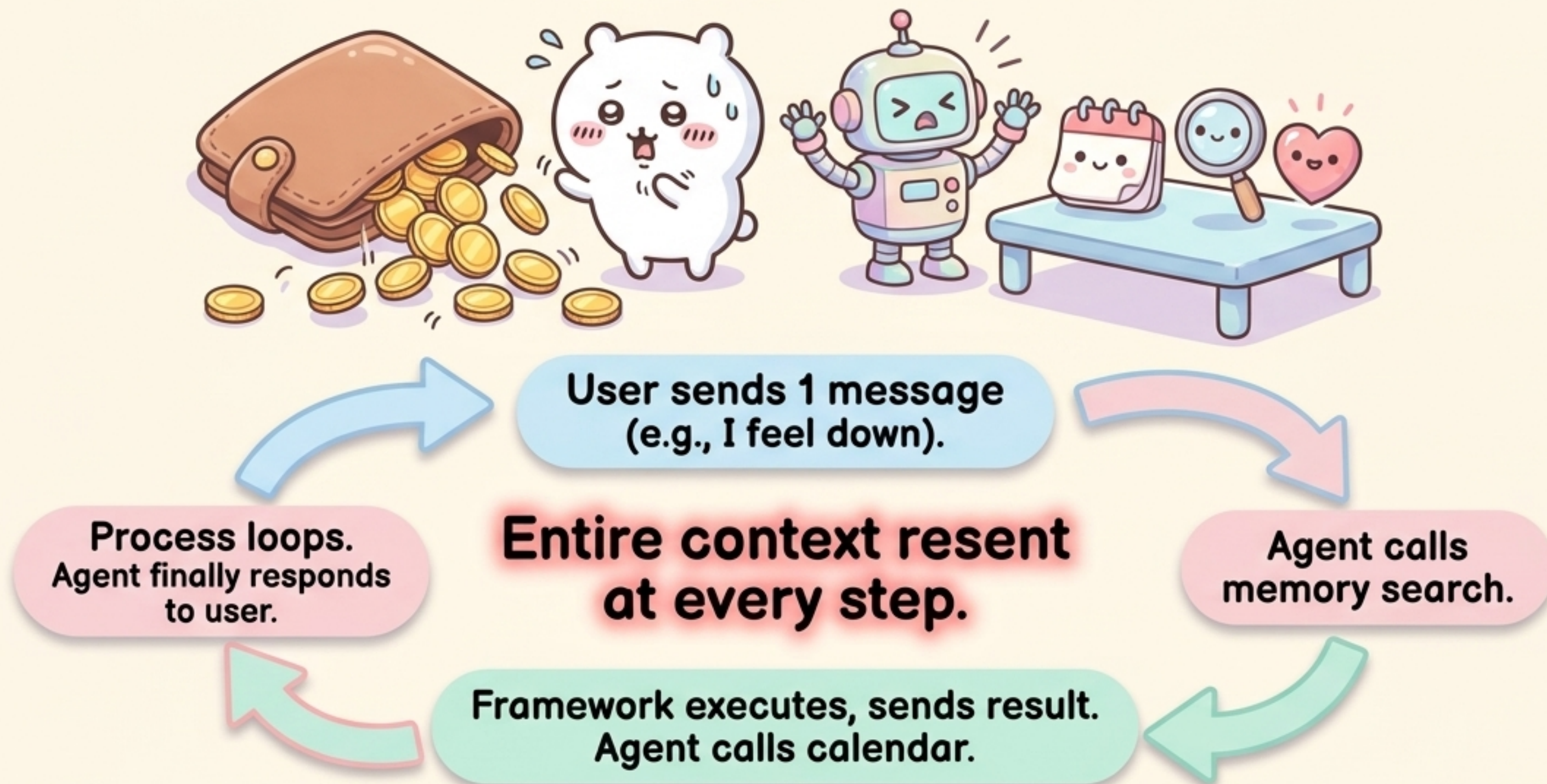
The 1M Window Footgun

- **Issue:** Model accepted 1M tokens; compaction set to safeguard mode.
- **Impact:** Zero compactions triggered in 537 turns. 1M limit delayed automatic cleanup indefinitely.

Heartbeat Bloat

- **Issue:** Uncapped history limits on periodic cron-like check-ins.
- **Impact:** Simple good morning decisions dragged 120,000+ tokens of history into every API call.

The 25x Multiplier of Invisible Tool Loops



Session outcome: 30 User Messages converted into 750 Assistant Turns.

Stabilization and the Amnesia Bug



Before (The Bug)

- Daily resets wrote memory seeds to `.jsonl`.
- Framework wiped the file 3 seconds later.
- Result: Total daily amnesia.

After (The Fix)

- Inject seed directly into the prefix of the user's first message.
- Result: Persistent memory, safe from file-system clobbering.

New Enforced Limits

``contextTokens``: Hard cap at 200k (forces compaction at 100k).

``contextPruning.mode``: always (bypasses Anthropic-only gate).

``heartbeat.historyLimit``:
Capped at 20 turns (drops check-ins to ~5k tokens).