

Less Context. More Work.

Targeting working context under 64K tokens for AI subscriptions & agentic workflows.

Unmanaged Context Bloat

Default multi-turn agent runs & continuous tool execution

- **Context Blowout (>128k - 200k+):**
Conversations degrade, latency spikes, models lose instructions.
- **Subscription Throttling:**
Hits Claude Pro/Max & ChatGPT message caps early in long runs.
- **Cache Invalidation:**
Naive history truncation churns prefix caches, destroying cache hits.
- **Redundant Tool History:**
Massive stdout and repetitive JSON responses saturate prompt buffers.
- **Escalating Invoices:**
Pay-as-you-go APIs charge maximum input fees on every subsequent turn.

UNMANAGED CONTEXT HORIZON

200,000+ tokens | Latency: 18-35s | Cache Hit: 0-15%



Frequent 429 throttling & silent instruction loss

Swamix Pi Context Architecture

Configured under-64K working context with protected turns

- **Under-64K Working Budget:**
Maintains rapid response latency and razor-sharp attention fidelity.
- **Prefix Cache Stabilization:**
Engineered prefix stability ensures 60-85% Anthropic & OpenAI cache hits.
- **Protected Turns & Headroom:**
Protects recent turns and reserved 8,192 token output capacity.
- **Tool Output Compaction:**
Semantic summarization of tool stdout with lossless full-result reference.
- **Subscription Throughput:**
Eliminates premature rate-limit ceilings, tripling usable agent capacity.

SWAMIX CONFIGURED WORKING BUDGET

<64,000 tokens | Latency: 2-6s | Cache Hit: 65-90%



Active context compressed • Headroom & cache prefix preserved

Managed AI Optimization & Transparent Terms

Human engineering included • Free MIT audit software • Predictable volume pricing

Commercial Pricing Structure

30% of monthly AI spend • Min \$69/mo • Cap \$269,000/mo

30% of monthly AI spend (regardless of volume)

\$69 / month Minimum: Accessible entry point for early-stage products, founders, and small teams.

\$269,000 / month Upper Cap: Absolute ceiling for enterprise-scale workloads, no matter how high the volume.

Cheaper Than In-House Hiring: Expert AI context & prompt engineering at a fraction of Anthropic/OpenAI costs.

Independent Billing: Your model provider accounts stay directly in your name; we never mark up API tokens.

Free MIT Software Offering

Downloadable standalone CLI verification kit (Python stdlib):

- **audit.py:** Rigorous cohort matching: compares baseline vs. optimized runs task-by-task.
- **Token & Cache Metrics:** Measures uncached input, cached input hits, output, and cache writes.
- **Quality Regression Guard:** Enforces quality acceptance rate parity before validating savings.
- **Open Source (MIT):** Available at swamix.com/downloads/ai-usage-audit.zip — no sign-up required.

Human Engineering Included in Service

Dedicated specialist engineers who personalize and maintain your context pipeline

01

Architecture & Workflow Audit

We inspect your agent loops, system prompts, subagent trees, and tool schemas to identify context bloat and inefficiencies.

02

Pi Engine Setup & Extension Routing

We deploy and tune Pi context-compression extensions, setting calibrated soft budgets (<64K tokens) with dynamic fallbacks.

03

Cache Prefix Engineering

We restructure dynamic prompt sections to keep common prefixes immutable, maximizing Anthropic 5-min cache hits.

04

Deep Domain Personalization

We customize extraction heuristics, tool output minifiers, and compaction triggers specifically to your application.

05

Continuous Telemetry & Cost Reviews

Monthly review of token ratios, request latencies, and cache metrics with ongoing prompt recalibration as needed.

Measurable Validation Criteria

How cost savings, token reduction, and capacity expansion are strictly proven.

1. Input Token Compression

Metric: Uncached prompt tokens per task execution.

Target: Working context actively maintained under 64K tokens.

Validation: Direct BPE/tiktoken measurement across identical task cohorts.

Impact: Reduces input token processing costs by 50% to 75% on multi-turn flows.

2. Prompt Cache Stability

Metric: Cached input tokens (Anthropic cache_read, OpenAI cached).

Target: Achieve 60% - 90% cache hit ratios on multi-turn interactions.

Validation: Provider usage response metadata audit (prompt_cache_hit_tokens).

Impact: Preserves prefix caches rather than thrashing them with naive edits.

3. Request Counts & Retries

Metric: API roundtrips, timeout drops, and 429 rate-limit retries.

Target: Zero 429 backoff delays; lower total turns per completed objective.

Validation: Network-level request logs and agent telemetry traces.

Impact: Dramatically cuts turn latency and prevents mid-task agent stalls.

4. Quality & Output Parity

Metric: Task acceptance rate and functional output benchmark score.

Target: 100% acceptance parity (zero regression vs. uncompressed baseline).

Validation: Automated test suites or human evaluator acceptance flags in CSV.

Impact: Ensures token savings represent true efficiency, not lossy degradation.

CRITICAL DISTINCTION: Pay-As-You-Go workloads yield dollar invoice reductions. Flat-rate subscriptions (Claude Pro/Max, ChatGPT Team) yield dramatic throughput expansion and elimination of rate-limit throttling.