Claude Managed Agents — Complete Pricing Reference + Dreaming Update (October 2026)

Edition of November 13, 2026 — facts verified November 13, 2026. Current edition always at https://tygartmedia.com/claude-managed-agents-complete-pricing-guide-2026/.

> *Updated October 28, 2026: Made infographic diagrams clickable to open full-size images.*

Last refreshed: October 28, 2026 (Pacific)

May 2026 Update — Dreaming Feature + Beta Status

Anthropic introduced **Dreaming** at Code w/ Claude (May 6, 2026) — a new Managed Agents capability where agents review their own session history overnight to improve future performance. Harvey (legal AI) reported a roughly 6× task completion rate increase after implementing it. Dreaming is developer-access preview only. Multiagent Orchestration and Outcomes are now in public beta. See the new Dreaming section below.

What Is Claude Managed Agents? (Current Status, May 2026)

What Claude Managed Agents is right now.

Claude Managed Agents is Anthropic's framework for long-running, stateful AI agents — agents that can maintain context across sessions, hand off between sub-agents, and now, improve themselves by reviewing their own work history. Here's the current status of each component:

Component	Status	Who Has Access
Multiagent Orchestration	Public Beta	All API developers
Outcomes	Public Beta	All API developers
Dreaming	Developer Preview	Selected developers only

Dreaming: The Feature the Press Mostly Missed

Dreaming — the feature the press mostly missed.

Announced at Code w/ Claude on May 6, 2026, Dreaming is a Managed Agents capability that lets agents review and reorganize their own memory between sessions. The mechanism:

1. After a session ends, the agent reads its existing memory store alongside the session transcripts
2. It produces a new, reorganized memory store: duplicates merged, stale entries replaced, new patterns surfaced
3. The next session starts with a higher-quality knowledge base — capturing insights no single session could hold

This is meaningfully different from simply persisting conversation history. The agent isn't just remembering what happened — it's synthesizing what it learned. Think of it as the difference between taking notes and actually reviewing and reorganizing your notes the next morning.

The Harvey Result

Harvey, the legal AI company, reported approximately a 6× task completion rate increase after implementing Dreaming in their Managed Agents workflow. Harvey's use case — complex legal research that spans multiple sessions with evolving context — is exactly the kind of work Dreaming was designed for. Sessions build on each other rather than starting fresh each time.

Dreaming is developer-access preview as of May 2026. Docs: platform.claude.com/docs/en/managed-agents/dreams.

What Dreaming Is Not

A few clarifications worth making explicit:

- Dreaming is not available to end users — it's a developer-layer capability requiring implementation
- It's not persistent memory in the claude.ai chat interface
- It's not available to free or standard Pro subscribers through any interface
- It's a developer preview, not GA — expect it to evolve before full release

Our Take: Why This Architecture Matters

We run Managed Agents alongside unified Claude workflows (Cowork was folded into core Claude in September 2026). The Dreaming announcement is the first time Anthropic has shipped something that resembles how expert human knowledge actually compounds over time — not by accumulating raw notes, but by periodically synthesizing and reorganizing what's been learned into a cleaner structure.

The Harvey 6× result is a real-world data point from a production legal AI workflow. That's not a benchmark number — it's a deployed system showing measurable improvement from session-to-session memory refinement. Whether that 6× figure holds across different use cases is unknown, but the direction of the effect is the signal: agents that learn from their own history outperform agents that don't.

For non-developer users watching this space: Dreaming is the preview of what agentic AI will look like when it becomes mainstream. The groundwork being laid now in developer preview will eventually surface in subscription-tier products.

Model Accuracy Note — Updated November 13, 2026

Lineup currency (November 13, 2026, verified): Fable 5.2 ($8/$40), Opus 5.5 ($4/$20), Sonnet 5.5 ($2/$10), Haiku 5.5 ($0.10/$0.50 ≤100K; $0.50/$2.50 >100K). Legacy (still listed): Fable 5.1 ($10/$50), Haiku 4.5 ($1/$5), Opus 4.8 ($5/$25), Sonnet 4.6 ($3/$15). Where older historical sections mention Opus 4.7/4.6, those references are superseded. Current model tracker →

Tygart Media

Strategy

Volume ■ · Issue 04

Quarterly Position

By

Will Tygart

•

Long-form Position

•

Practitioner-grade

You opened this tab because you need a number you can actually use. Not a vibe, not "it depends." A real pricing breakdown you can put in a spreadsheet, a budget request, or a Slack message to your CTO.

This is that page. Every pricing variable for Claude Managed Agents in one place, verified against Anthropic's current documentation as of November 13, 2026. Bookmark it. The beta will update; so will this.

Quick Reference: The Formula

Total Cost = Token Costs + Session Runtime ($0.08/hr) + Optional Tools Session runtime only accrues while status = `running`. Idle time is free.

The Two Cost Dimensions

Two cost dimensions — stale-proof framing.

Claude Managed Agents bills on exactly two dimensions: tokens and session runtime. Every pricing question you have collapses into one of these two buckets.

Dimension 1: Token Costs

These are identical to standard Claude API pricing. You pay the same rates you'd pay calling the Messages API directly. No Managed Agents markup on tokens. Current rates for the models most commonly used in agent work:

- **Claude Sonnet 5.5:** $2/million input tokens, $10/million output tokens *(legacy Sonnet 4.6, still listed: $3/$15)*
- **Claude Opus 5.5:** $4/million input tokens, $20/million output tokens — check platform.claude.com/docs/en/about-claude/pricing for current figures
- **Prompt caching:** same multipliers as standard API — cache reads at $0.20/MTok on Fable 5.2 (97.5% off $8 input) and Opus 5.5 (95% off $4 input), $0.10/MTok on Sonnet 5.5 (95% off $2), and $0.01/MTok on Haiku 5.5 ≤100K dramatically reduce input token costs on long sessions with stable system prompts

The implication: a token-heavy agent with a large system prompt that runs the same context repeatedly benefits significantly from prompt caching, and that benefit carries over unchanged into Managed Agents.

Dimension 2: Session Runtime — $0.08/Session-Hour

This is the Managed Agents-specific charge. You pay $0.08 per hour of active session runtime, metered to the millisecond.

The critical word is *active*. Runtime only accrues while your session's status is `running`. The following do not count toward your bill:

- Time spent waiting for your next message
- Time waiting for a tool confirmation
- Idle time between tasks
- Rescheduling delays
- Terminated session time

This is not how you'd bill a virtual machine. It's closer to how AWS Lambda bills — you pay for execution, not reservation. An agent that "runs" for 8 hours but spends 6 of those hours waiting on human input has a very different bill than one running continuous autonomous loops.

Optional Tool Costs

Web Search: $10 per 1,000 Searches

If your agent uses web search, each search costs $10/1,000 — that's $0.01 per search. For most agents, this is negligible. For a research agent running hundreds of searches per session, it becomes a line item worth modeling separately.

Code Execution: Included in Session Runtime

Code execution containers are included in your $0.08/session-hour charge. You're not separately billed for container hours on top of session runtime. This is explicitly stated in Anthropic's docs and represents meaningful savings versus provisioning your own compute.

Worked Cost Examples

Example 1: Daily Research Agent

Runs once per day. 30 minutes of active execution. Processes 10 documents, outputs a summary report. Moderate token volume.

- Session runtime: 0.5 hrs × $0.08 = **$0.04/day** (~$1.20/month)
- Tokens (estimate): 50K input + 5K output with Sonnet 5.5 = ~$0.15/run (~$5/month)
- **Total: ~$8–10/month**

Example 2: Weekly Batch Content Pipeline

Runs 3x/week. 2-hour active sessions. Processes multiple documents, generates structured outputs.

- Session runtime: 2 hrs × $0.08 × 12 sessions/month = **$1.92/month**

- Tokens: depends on content volume — typically $10–40/month
- **Total: ~$12–42/month**

Example 3: Customer Support Agent (Business Hours)

Active during business hours, handling tickets. 8 hours/day active, 5 days/week.

- Session runtime: 8 hrs × $0.08 × 22 days = **$14.08/month in runtime**
- Tokens: highly variable by ticket volume — the dominant cost driver at scale
- **Runtime cost alone: ~$14/month** — tokens are likely 5–20x this depending on volume

Example 4: 24/7 Always-On Agent

The maximum theoretical runtime exposure. Continuous operation, no idle time.

- Session runtime: 24 hrs × $0.08 × 30 days = **$57.60/month**
- In practice, no agent has zero idle time — real cost will be lower
- Token costs at this scale become the dominant factor by a wide margin

Anthropic's Official Example (from their docs)

Anthropic's docs price this example on Claude Opus 5 ($5/$25): $0.705 all-in, or $0.525 with 40,000 cache reads. Recomputed at current Opus 5.5 rates: a one-hour coding session consuming 50,000 input tokens and 15,000 output tokens: session runtime = $0.08. With prompt caching active and 40,000 of those tokens as cache reads ($0.20/MTok), the token math is: 10,000 uncached input × $4/M = $0.04, plus 40,000 cache reads × $0.20/M = $0.008, plus 15,000 output × $20/M = $0.30 — about $0.35 in tokens. The runtime charge stays flat at $0.08 regardless of caching, for roughly $0.43 all-in.

24/7 Duty-Cycle Profiles & Build-vs-Buy Economics

While the $57.60/month ceiling assumes zero idle time across 720 hours, real 24/7 production agents rarely remain in `running` status 100% of the clock. Modeling active execution ratios changes the runtime budget dramatically:

- **Monitoring agent (~5% active / 36 hrs/mo):** $2.88/month runtime (~$5–$15 total with light burst tokens).
- **Business-hours support agent (8 hrs/day active / 240 hrs/mo):** $19.20/month runtime (~$50–$100 total at ~100 tickets/day).
- **Continuous batch pipeline (20 hrs/day active / 600 hrs/mo):** $48.00/month runtime ($200–$500+/month total as tokens dominate).

Hourly Token-vs-Runtime Crossover (at Sonnet 5.5 rates: $2/MTok in, $10/MTok out):

- **Low-throughput polling (10K input + 2K output/hour):** $0.02 + $0.02 = **$0.04/hour tokens** vs. $0.08/hour runtime (runtime exceeds tokens).
- **Active processing agent (100K input + 20K output/hour):** $0.20 + $0.20 = **$0.40/hour tokens** vs. $0.08/hour runtime (tokens run 5× the runtime fee).

Build-vs-Buy & Labor Framing: In a 2-hour active session generating $0.16 in runtime alongside $20–$50 in tokens, runtime is budget noise — teams save most by optimizing prompts and caching (100K-scale agents cut recurring system-prompt reads up to 95% via caching). Against custom infrastructure (months of sandbox/checkpoint engineering), $0.08/hour is almost always cheaper in year one, and a $50–$100/month agent handling 2 hours/day of structured prep ($1,500/month at $30/hr fully-loaded labor) delivers a 15–30× direct cost difference.

What's Not Billed in Managed Agents

A few things that might seem like costs but aren't:

- **Infrastructure provisioning:** Anthropic handles hosting, scaling, and monitoring at no additional charge
- **Container hours:** Explicitly not separately billed on top of session runtime

- **State management and checkpointing:** Included in the session runtime charge
- **Error recovery and retry logic:** Anthropic's infrastructure problem, not yours

Rate Limits

Managed Agents has specific rate limits separate from standard API limits:

- **Create endpoints:** 300 requests/minute
- **Read endpoints:** 1,200 requests/minute
- Organization-level limits still apply
- For higher limits, contact Anthropic enterprise sales

How to Access Managed Agents Pricing

Managed Agents is available to all Anthropic API accounts in public beta. No separate signup, no premium tier gate. You need the `managed-agents-2026-04-01` beta header in your API requests — the Claude SDK adds this automatically.

For high-volume agent applications, Anthropic's enterprise sales team negotiates custom pricing arrangements. Contact them at sales@anthropic.com or through the Claude Console.

The Pricing Signals Worth Noting

Anthropic recently ended Claude subscription access (Pro/Max/Ultra) for third-party agent frameworks, requiring those users to switch to pay-as-you-go API pricing. This signals a deliberate strategy: consumer subscriptions are for human-paced interactions; agent workloads route through the API. The $0.08/session-hour rate exists in that context — it's infrastructure pricing for compute that runs beyond human attention spans.

The session-hour model also signals something about Anthropic's infrastructure cost structure. They're pricing on active execution time because that's what actually taxes their systems. Idle sessions don't cost them much; active agents do. The billing model follows the actual resource consumption pattern.

Frequently Asked Questions

Is the $0.08/session-hour charge in addition to token costs, or does it replace them?

In addition to. You pay both: standard token rates for all input and output tokens, plus $0.08 per hour of active session runtime. They're separate line items.

Does prompt caching work in Managed Agents sessions?

Yes. Prompt caching multipliers apply identically to Managed Agents sessions as they do to standard API calls. If your agent has a large, stable system prompt, caching it can significantly reduce input token costs.

What happens if my session crashes? Am I billed for the crashed time?

Runtime accrues only while status is `running`. Terminated sessions stop accruing. Anthropic's infrastructure handles checkpointing and crash recovery — the session state is preserved even if the session terminates unexpectedly.

Can I use Managed Agents on the free API tier?

Managed Agents is available to all Anthropic API accounts in public beta, but standard tier access and rate limits apply. Free API tier users receive a small credit for testing.

How does this compare to running agents on my own infrastructure?

See our full breakdown: Build vs. Buy: The Real Infrastructure Cost of Claude Managed Agents. Short version: the $0.08/hour is almost certainly cheaper than provisioning and maintaining equivalent compute, but you trade control and data locality for that simplicity.

Are there volume discounts?

Volume discounts are available for high-volume users but negotiated case-by-case. Contact Anthropic enterprise sales.

Does web search billing count against the $10/1,000 rate if the search returns no results?

Anthropic's current docs don't explicitly address failed searches. Treat any triggered search as billable until confirmed otherwise.

See 24/7 Duty-Cycle Profiles & Build-vs-Buy Economics above for full workload duty-cycle math. For the build-vs-buy infrastructure comparison: Build vs. Buy: The Real Infrastructure Cost. For enterprise deployment patterns: Rakuten Stood Up 5 Enterprise Agents in a Week.

Model your own agent costs from the Anthropic Console — API keys, billing, and a token-to-dollar cost estimator.

Primary & Reference Links Cited on This Page

- platform.claude.com/docs/en/managed-agents/dreams — `https://platform.claude.com/docs/en/managed-agents/dreams`
- platform.claude.com/docs/en/about-claude/pricing — `https://platform.claude.com/docs/en/about-claude/pricing`
- Tygart Media: https://tygartmedia.com/wp-content/uploads/2026/08/claude-managed-memory-loop.jpg
- Tygart Media: https://tygartmedia.com/wp-content/uploads/2026/08/claude-invisible-agent-layer.jpg
- Tygart Media: Current model tracker →
- Tygart Media: https://tygartmedia.com/wp-content/uploads/2026/08/claude-api-token-meter-visual.jpg
- Tygart Media: Build vs. Buy: The Real Infrastructure Cost of Claude Managed Agents
- Tygart Media: Rakuten Stood Up 5 Enterprise Agents in a Week
- Tygart Media: Anthropic Console