Sonnet 5: near-Opus performance at a fraction of the cost

The price, not the benchmark, is the real event — and it moves the binding constraint in enterprise AI from cost to trust.

1 July 2026 8 min read

For two years the binding question in enterprise AI was whether the model was good enough. Anthropic's Claude Sonnet 5 quietly retired that question for a large class of work. What replaces it is harder, and more commercial: not can the model do the task, but will you trust it to finish one without someone standing over it.

Anthropic released Claude Sonnet 5 at the end of June. The headlines did what headlines do and reached for the benchmark charts. The number worth sitting with was not on a leaderboard. It was the price.

Sonnet 5 lands at roughly 40 per cent below the cost of Opus 4.8, Anthropic's most capable model, while landing close to it on the work most enterprises actually buy AI to do. On the GDPval knowledge-work benchmark the two are effectively level, 1,618 against 1,615. On the hardest agentic coding and reasoning tasks Opus still leads by a clear margin. That gap matters for a narrowing set of frontier problems. For the broad middle of enterprise work, it has quietly stopped mattering.

Read that as an economic event, not a technical one. For two years the binding question in every AI business case was capability: is the model good enough to trust with this. Sonnet 5 answers that question in the affirmative for a large class of work, and then attaches a mainstream price to the answer. When near-frontier capability becomes affordable at scale, the thing that used to make a case fail, cost per capable unit of output, is no longer the thing that fails.

What the price actually unlocks

One line buried under the benchmark coverage made this concrete. A team gave the model a two-part job: update account tiers in their CRM, then send a launch announcement to the right enterprise contacts. It completed both, end to end. Their description of what changed was not about eloquence or code quality. It was four words. That used to stall halfway.

Follow-through is the quiet capability. Not the ability to produce a good answer to a single prompt, but the ability to carry a multi-step task across systems without a human stepping back in at the midpoint. That is the difference between a clever tool and a delegated task, and it is the difference that shows up on a profit and loss statement.

> *Benchmarks tell you what a model can do. Economics decide what a business will do.*

The pattern is familiar from every prior wave of technology. Capability arrives first and expensively. Then price falls, and the falling price, not the original capability, is what actually rewires how organisations operate. Yesterday's premium function becomes today's default. Work that could never clear a return-on-investment

hurdle suddenly clears it comfortably. The constraint stops being the technology and moves back to the organisation.

Trust is now the binding constraint

If cost is no longer what stops a team handing over the next multi-step task, something else is. The evidence points at trust, and it is not soft. A 2026 study tracking more than 8,000 users of agentic AI put a subset through complex, multi-step tasks and found completion rates clustered between 73 and 90 per cent depending on the platform. Useful, but not the same as dependable. More telling: a majority of users still trusted doing the work manually over trusting the agent, even when the agent finished.

That is the real number for a leadership team to absorb. Completion is not confidence. A model that finishes the task nine times in ten is a genuine asset and a genuine liability at once, because the tenth time is unattended and inside a live system. Enterprises setting formal targets are now writing reliability thresholds of 95 per cent accuracy and 90 per cent completion into their standards. The gap between what the model does and what the organisation will trust it to do unattended is where the next two years of value is won or lost.

This reframes the buying question. When a cheaper, capable model arrives it is tempting to treat adoption as procurement: switch the model, book the saving. The organisations that get real value will treat it as a trust-engineering decision. Where do the checkpoints sit. What runs unattended and what routes to a person. How do we make the model's reasoning and sources visible enough that an expert can verify a result in seconds rather than redo it. The highest-leverage fix in the research was not a better model. It was surfacing citations and reasoning so a human could trust the output without reproducing it.

What this means for the next planning cycle

Gartner expects 40 per cent of enterprise applications to ship with task-specific agents by the end of 2026. If that is even roughly right, the competitive question is not whether your organisation adopts agents. It is whether it builds the judgement to know which tasks to hand over, in what order, with which controls, faster than the organisations you compete with.

Three moves follow for the current planning cycle. First, re-run the automation backlog: tasks shelved eighteen months ago because the capable model was too expensive deserve a fresh look at today's prices, and some will now clear. Second, treat trust as a design input, not an afterthought: decide deliberately where a human stays accountable, and build the verification surface that lets the rest run. Third, resist standardising on a single model. Capability and cost used to move together, so one model could be the obvious choice. They have started to separate. The mature posture is a small portfolio, with the judgement to route each task to the right model.

The leaders who create the most value over the next few years will not be the ones running the most powerful model. They will be the ones who understood, earlier than their competitors, that the expensive part was never the intelligence. It was learning to trust it with the work.

BY THE NUMBERS

$3 / $15
Sonnet 5 standard price per million tokens, ~40% below Opus 4.8

1,618 vs 1,615
Sonnet 5 and Opus 4.8 on GDPval knowledge work, level

73-90%
Agent completion on complex multi-step tasks, 2026 study

95% / 90%
Accuracy and completion thresholds enterprises now set

SOURCES

· Anthropic, Introducing Claude Sonnet 5 (anthropic.com/news/claude-sonnet-5)
· MarkTechPost, Sonnet 5 vs Sonnet 4.6 vs Opus 4.8: benchmarks and pricing, Jun 2026
· digitalapplied.com, AI Agent Task Completion in 2026: 8,128-user study
· Gartner, enterprise application agent-integration forecast, 2026