IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Claude Opus 5.5 for Pharma R&D: Cost, Quality, Access Pilot

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-11-15 · claude opus 5.5 · claude opus 5.5 pharma · pharma r&d ai · life sciences verification program · claude biology safeguards · llm pilot scorecard · cost per accepted task · opus 5.5 pricing · claude vs gpt pharma · anthropic fable 5.1

Original article: https://intuitionlabs.ai/articles/claude-opus-5-5-pharma-rd-pilot-scorecard



In this article

1. Executive Summary
2. Introduction and Background
3. Product Overview: What Shipped and Who Can Use It for Biology
4. Performance and Benchmark Analysis
5. Pilot Design: Three Bounded Tasks and a Fixed Configuration
6. Reception and Community Perspectives
7. Data Analysis and Evidence
8. Implications and Future Directions
9. Frequently Asked Questions (FAQs)
10. Conclusion

Executive Summary

Anthropic released **Claude Opus 5.5** on September 22, 2026 (platform.claude.com). List prices are **$4 per million input tokens and $20 per million output tokens**, a 20% per-token cut from Opus 5 (www.techrepublic.com), and cache reads fell from $0.50 to **$0.20 per million tokens** (venturebeat.com). The widely repeated 40% saving is Anthropic's own test result at default settings on typical workloads (www.anthropic.com). It is not a per-token price, and Artificial Analysis, an independent evaluator, measured cost per task level with Opus 5 at maximum effort despite 1.6 times the output tokens (artificialanalysis.ai). No public evidence shows what the saving is on a pharma sponsor's own task mix.

For biology work, access is a gate and not a footnote. Opus 5.5 uses the same biology safeguards as Fable 5.1 (hlth.com), and dual-use requests in areas such as virology, toxicology and molecular design fall back to Opus 5 (support.claude.com). Anthropic's Life Sciences Verification Program (LSVP) is the route to relaxed safeguards, but its Standard Use grants are listed for Mythos 5.1, Opus 5 and Sonnet 5 "today" and extend to future models as they launch, yet the page does not name Opus 5.5 (www.anthropic.com). The program requires 30-day data retention (support.claude.com), so coverage of Opus 5.5 for the intended organization and platform needs written confirmation before a pilot is scoped.

Quality evidence is vendor-reported or general-purpose. Anthropic reports 58.7% on Terminal-Bench-Science 0.1 versus 29.0% for Opus 5 (www.anthropic.com). Vals, an independent evaluator, reports 48.57% on a single-run, 70-task version of the same benchmark, falling to 47.14% when fallback-assisted tasks count as failures (www.vals.ai), and it notes that Opus 5.5 lands below Opus 5 on MedCode (www.vals.ai). None of these measures is pharma validation. The U.S. Food and Drug Administration's draft guidance on AI in regulatory decision-making asks for a credibility assessment tied to a defined context of use (www.fda.gov).

This report therefore proposes a **cost-per-accepted-scientific-task scorecard** for three bounded tasks: literature evidence extraction with citations, review of an existing analysis pipeline, and a non-sensitive biology reasoning task. In an invented example (6 minutes of review per task at $90 per hour, with 88 versus 85 accepted tasks out of 100), reviewer time is about 97% of the cost per accepted task, and one extra accepted

task per 100 outweighs the entire token-price difference between Opus 5.5 and Opus 5. The recommendation is a controlled pilot with fixed effort, platform, fallback and review settings, a written access confirmation, and explicit stop rules. It is not a recommendation to automate regulated decisions.

Introduction and Background

Pharma and biotech research leaders face a practical question: should a defined scientific workflow be piloted on Claude Opus 5.5, stay on Opus 5, or wait for Fable or Mythos access? This is a model-selection and go/no-go decision. It is not an argument about general AI progress, and it does not concern automating decisions that regulators treat as high-risk.

Three facts frame the decision. First, price: Opus 5.5 is cheaper per token than Opus 5 (mixed-news.com). Second, capability: Anthropic ships Opus 5.5 with biology safeguards matching Fable 5.1's (hlth.com), its system card says biology capability differed only modestly from Mythos 5.1 (anthropic.com), and the same card treats the model as having CB-1 (non-novel weapons-related) capabilities but not CB-2 (anthropic.com). Third, access: the LSVP, a beta for vetted organizations, targets work that generally available Fable models block, such as drug discovery and research biology (www.anthropic.com). Each fact pulls in a different direction, which is why a scorecard is more useful than a release recap.

IntuitionLabs is a life-sciences and AI consultancy. It does not sell any model, platform or service compared in this report, and its perspective here is advisory: treat model selection as a measurement problem, fix every variable that is not the model, and price the human review behind every accepted output. Nothing here was reviewed by a regulatory, quality or legal professional, and readers should involve their own quality assurance and regulatory teams before a pilot touches data or decisions that matter.

This report makes no new measurements. Numbers come from vendor pages, independent evaluators, regulators and published studies, each cited at the point of use, and prices are as listed on vendor pages fetched on November 15, 2026. Where the worksheet in the Data Analysis section is filled in, the values are labeled hypothetical. Where it is blank, it is a template for the reader's own pilot.

Product Overview: What Shipped and Who Can Use It for Biology

Specifications and price context

Opus 5.5 offers a **1M-token context window**, a **128K maximum output** and a **June 2026** knowledge cutoff (mixed-news.com). The Claude API identifier is `claude-opus-5-5` (platform.claude.com), and the model is available on Anthropic's platform as well as Amazon Web Services, Google Cloud and Microsoft Azure (decrypt.co). A 900k-token request is billed at the same per-token rate as a 9k-token request (platform.claude.com), so long context carries no premium. Anthropic has committed to keeping Opus 5.5 available until at least **September 22, 2027** (mixed-news.com).

Two behavior changes affect any pilot harness:

- **Thinking is always on** (platform.claude.com). Thinking tokens are billed as output tokens (platform.claude.com), so the effort setting drives cost directly.

- **Default effort is medium.** Anthropic recommends starting agent workloads there (platform.claude.com).

Table 1 below compares list prices and biology access routes for the Claude models and the main alternatives, as listed on vendor pages and press coverage in November 2026.

Model	List price per million tokens (input / output)	Cache pricing	Biology access route
Claude Opus 5.5	$4 / $20 (www.techrepublic.com)	Reads $0.20 (venturebeat.com); writes $5 (5 minutes) or $8 (1 hour) (mixed-news.com)	Fable 5.1-class safeguards; blocks fall back to Opus 5 (anthropic.com); LSVP route for vetted organizations (www.techrepublic.com)
Claude Opus 5	$5 / $25 (www.digitaltrends.com)	Reads $0.50; 5-minute writes $6.25 (Anthropic pricing page)	No biology fallback (support.claude.com); named for LSVP Standard Use "today"
Claude Fable 5.1	$10 / $50 (thenewstack.io)	Reads $0.25 (platform.claude.com)	Full biology safeguards; not listed among the models covered by LSVP Standard Use grants (www.anthropic.com)
Claude Mythos 5.1	$10 / $50 (same price row as Fable 5.1)	Reads $0.25 (same row)	Trusted access programs only (www.anthropic.com); initially a set of U.S. organizations (www.anthropic.com)
GPT-6 Astra (OpenAI)	$10 / $50, cached input $1.00 (developers.openai.com)	See price column (cached input $1.00 per million tokens)	General API; see Rosalind row for life-sciences gating
GPT-6.1 Sol (OpenAI)	$2 / $10, cached input $0.10 (developers.openai.com)	See price column (cached input $0.10 per million tokens)	General API
GPT-6 Rosalind (OpenAI)	$15 / $75, cached input $1.50 (developers.openai.com)	$1.50 per million cached input tokens	Approved organizations only; a subscription does not confer access (help.openai.com)
Gemini 4 Carbon (Google)	$2 / $10 through November 30, 2026 (ai.google.dev); $4 / $20 from December 1, 2026 (ai.google.dev)	Cached-input rates are listed on Google's pricing page, linked in the price column	CBRN safeguards strengthened (deepmind.google); Google's model card describes a verified-access program for cybersecurity (deepmind.google)

The table shows Opus 5.5 between GPT-6 Astra and GPT-6.1 Sol on list price, and well below Fable 5.1 and Mythos 5.1. One trade publication notes that Opus 5.5 remains more expensive than OpenAI's GPT-6 Sol and Luna models (aibusiness.com). A price comparison alone says little about a pharma task, because cost per accepted task also depends on tokens used, retries, fallback and review time. The OpenAI rows are short-context rates (up to 272K input tokens); OpenAI lists higher rates above 272K input tokens, which long literature-extraction runs can reach. The Gemini row carries a scheduled price change, a reminder to record the source date on every worksheet row.

The access gate for biology work

Access to Opus 5.5 for biology depends on which of three routes an organization sits on. Each route changes both cost and the validity of any comparison with Opus 5.

- **Default safeguards (no LSVP).** Opus 5.5 runs a biology safety classifier in addition to the cybersecurity one (platform.claude.com). Everyday health and educational questions are still served, but the classifiers review everything the model reads, so a fallback can be triggered by files or search results the user did not type (support.claude.com). A refusal arrives as a successful HTTP 200 response (platform.claude.com), and API customers must opt in and configure fallbacks (support.claude.com).
- **LSVP.** Vetted organizations can apply for broader biology access through the program (www.digitaltrends.com). No Anthropic page reviewed names Opus 5.5 under any grant type. Anthropic's program page says Standard Use grants extend to future models as they launch, and launch coverage points vetted organizations to the program for Opus 5.5 (thenewstack.io). Coverage for a specific organization and model should be confirmed in writing.
- **Mythos 5.1. Mythos 5.1** is the same model as **Fable 5.1** with different safeguards (www.anthropic.com), available only through trusted access programs (www.anthropic.com).

The LSVP's operating conditions shape a pilot more than the application does. The beta is not available on third-party platforms (www.anthropic.com), so a team using Opus 5.5 through a cloud marketplace cannot use LSVP there. Existing zero-retention agreements do not apply to LSVP usage (support.claude.com), and organizations set up under a business associate agreement (BAA, the contract used for protected health information) cannot receive LSVP grants (support.claude.com). Cyber classifiers stay in place under LSVP grants (www.anthropic.com), and not every organization that submits an interest form is invited during the beta (support.claude.com).

Alternatives and their gates

OpenAI's life-sciences offering is a trusted-access model, **GPT-6 Rosalind**, with Research and Discovery access tiers (help.openai.com). It is not available through Amazon Bedrock (help.openai.com), and API access is limited to internal use (help.openai.com). Google's **Gemini 4 Carbon** has strengthened safeguards for chemical, biological, radiological and nuclear misuse (deepmind.google), and the only verified-access program in its model card covers cybersecurity (deepmind.google).

No independent pharma head-to-head of Opus 5.5 and any GPT model was found. One trade report notes there is no clean public same-harness comparison between GPT-6 Sol and Claude Opus 5.5 (venturebeat.com). **Artificial Analysis** reports that **GPT-6.1 Sol** scores within 4 points of Opus 5.5 at maximum effort at about an eighth of the cost per task (artificialanalysis.ai), but that index is general-purpose. A comparator arm is therefore reasonable, and the scorecard below accommodates one.

Performance and Benchmark Analysis

Vendor-reported versus independent results

Table 2 below separates numbers reported by Anthropic from those run by independent evaluators and benchmark originators. The same benchmark appears more than once because harness, number of runs and fallback handling differ between reporters.

Benchmark	Reporter and type	Opus 5.5	Opus 5	Conditions and caveats
Terminal-Bench-Science 0.1	Anthropic (vendor-reported)	58.7% (www.anthropic.com)	29.0% (same source)	Maximum effort; standard error of plus or minus 3.5 points or more (www.anthropic.com); intercepted biology tasks were completed by Opus 5 (www.anthropic.com)
Terminal-Bench-Science 0.1	Benchmark originator leaderboard (Claude Code harness, three trials)	About 63% (www.terminal-bench-science.ai); 8.7B tokens, $4,874 total (www.terminal-bench-science.ai)	30.0% (www.tbench.ai)	Native agent harness; tasks were calibrated to challenge the newest frontier models (www.tbench.ai)
Terminal-Bench-Science 0.1	Vals (independent, single run, 70 tasks)	48.57%, or 47.14% counting fallback-assisted tasks as failures (www.vals.ai); $19.12 per test (www.vals.ai)	27.14%, $32.54 per test (www.vals.ai)	Single pass@1 run per model (www.vals.ai); not directly comparable to the official leaderboard (www.vals.ai)
Terminal-Bench-Science 0.1	Artificial Analysis (independent, three repeats)	59% (GPT-6 Astra 63%) (artificialanalysis.ai)	Not captured	mini-swe-agent harness, pass@1 averaged over three repeats (artificialanalysis.ai)
MedCode (medical coding)	Vals (independent)	49.80%, 45.87k tokens per test (www.vals.ai)	63.57%, 19.09k tokens per test (www.vals.ai)	Built on 2,755 de-identified patient records (www.vals.ai); a coding task, not an R&D task
BioMysteryBench	Vals (benchmark developed by Anthropic (www.vals.ai))	79.26%, $2.95 per test (www.vals.ai)	79.26%, $3.29 per test (www.vals.ai)	Mean of three runs (www.vals.ai); tasks refused by biology filters count as failures (www.vals.ai)
Vals Science Index	Vals (average of three science benchmarks)	58.65%, $8.71 per test (www.vals.ai)	47.93%, $13.04 per test (www.vals.ai)	Includes the Terminal-Bench-Science and BioMysteryBench figures above
Intelligence Index v5	Artificial Analysis (general-purpose)	58 at maximum effort (artificialanalysis.ai); $10.04 per task (artificialanalysis.ai)	$11.82 per task at maximum effort (artificialanalysis.ai)	General index across many task types, not biology-specific

The spread on one benchmark is the main lesson. **Terminal-Bench-Science 0.1** contains 70 research-workflow tasks contributed by practicing scientists (www.vals.ai). Because 34 of 70 tasks is **48.57%** and 33 of 70 is **47.14%**, the difference between Vals' two figures is a single task, smaller than the 3.5-point standard error Anthropic reports. Results for Opus 5.5 range from about **47% to about 63%** across harnesses and fallback settings, so no single published figure can be treated as a sponsor-relevant estimate.

The benchmark is also far from a pilot task. Of 920 task proposals, 464 were approved for implementation (www.tbench.ai), and release 0.1 scores 70 tasks. A pilot task that takes a reviewer minutes is a different workload.

Fallback and safeguard effects on scores

Anthropic's own results were produced with production safeguards enabled, so biology tasks the safeguards intercepted were scored through Opus 5. Anthropic's system card confirms that blocks on the biology classifier fall back to Opus 5 (anthropic.com). The effect can be large in some domains: on Terminal-Bench 2.1, counting fallback-assisted tasks as failures lowers Vals' score from **87.64% to 79.77%**, affecting 26 of 267 tasks (www.vals.ai), and on SRE Bench 217 of 262 tasks were fallback-assisted (www.vals.ai). No public figure gives the fallback rate for pharma tasks. That rate is a pilot output, not an assumption.

Token use, effort and the 40% claim

Anthropic's **40% claim** compares default settings with default settings, which one analysis describes as Opus 5.5 at medium effort against Opus 5 at high (www.digitalapplied.com). Anthropic's documentation adds that at the same effort setting the model tends to think more per turn than Opus 5 (platform.claude.com), and it advises running an effort sweep on the user's own evaluations (platform.claude.com). Independent measurements at maximum effort are mixed. Artificial Analysis measured about 119k output tokens per Intelligence Index task (artificialanalysis.ai), yet its later index version puts Opus 5.5 at $10.04 per task against $11.82 for Opus 5 (artificialanalysis.ai). Effort alone moves cost sharply: Opus 5.5 scores **35 at low effort and 58 at maximum**, at 10 times the cost per task (artificialanalysis.ai).

Task type changes the picture. In Table 2, Opus 5.5 used fewer tokens and less money per test than Opus 5 on Terminal-Bench-Science but 2.4 times the tokens on MedCode (45.87k versus 19.09k). Customer reports point in opposite directions, with Box reporting about a third as many tokens as Opus 5 (thenewstack.io) and a code-review vendor reporting a 50% increase in output tokens (aibusiness.com). A trade analysis cautions that such results are vendor-selected rather than independent evaluations (venturebeat.com). For a sponsor, the only reliable token figure is one measured on the sponsor's own inputs.

What these results do not measure

None of the sources above measures literature extraction with citations, pipeline review in a sponsor's code base, or biology reasoning under the LSVP regime. Anthropic reports an evaluation with Dyno Therapeutics on long-horizon molecular prediction and design (www.mobihealthnews.com), but coverage notes that no score was disclosed (hlth.com). Benchmark margins are an imperfect guide to real tasks, and Anthropic's own launch material cautions that gaps between models can be narrower than scores suggest.

Pilot Design: Three Bounded Tasks and a Fixed Configuration

What must stay fixed

A scorecard is informative only if everything except the model is held constant. These settings should be written down before the first run.

- **Model list.** Opus 5.5 and Opus 5 as the primary pair. Add one comparator such as GPT-6.1 Sol only if its access and data terms are acceptable. Fable 5.1 is not a candidate for these tasks, because the generally available Fable models block research biology (www.anthropic.com).
- **Effort.** Run a matched-effort arm and a vendor-default arm, because default effort differs between the two Claude models (www.digitalapplied.com). Record the setting on every call.
- **Platform.** First-party Claude API if LSVP is in scope, because the beta runs on first-party products only (support.claude.com). Record region settings: Anthropic lists a **1.1x multiplier** for U.S.-only inference (platform.claude.com).
- **Safeguards and fallback.** Configure fallback explicitly and log every refusal. Fallback targets for Opus 5.5 include Opus 4.8 and Opus 5 (platform.claude.com), and if the fallback model is rate limited the refusal is returned instead (platform.claude.com).
- **Input mix.** Freeze the document set, prompt prefix and schema. The prompt cache has a default 5-minute lifetime (platform.claude.com) and a 1-hour option at 2x the base input price (platform.claude.com), and prompts below the minimum are processed without caching and with no error (platform.claude.com).
- **Human review.** Fix reviewer roles, blinding and minutes per task. Review time is the largest line in the cost model below.

Batch processing suits non-interactive extraction. Anthropic lists a **50% discount** on input and output tokens (platform.claude.com), with Opus 5.5 batch prices of $2 input and $10 output per million tokens (platform.claude.com), but batches expire if not completed within 24 hours (platform.claude.com) and cache hits in batch mode typically run 30% to 98% (platform.claude.com). Record batch and interactive calls in separate worksheet rows.

The three tasks

Table 3 below defines each task with its access route, evidence requirement, failure types, acceptance threshold and reviewer. The thresholds are illustrative starting values for a sponsor to replace before running. They are not benchmarks from any source.

Task	Access route	Evidence requirement	Failure types	Acceptance threshold (illustrative)	Reviewer
1. Literature evidence extraction with citations	Standard API on public literature; classifiers review files the model reads, so biology-heavy papers can trigger fallback	Every extracted field linked to a verbatim quote and location; quotes machine-checked against the source text	Wrong value; missed field; unsupported or unresolvable citation; refusal or fallback	At least 95% field agreement with an adjudicated reference and zero unresolvable quotes; published accuracy of 91.0% for human-verified AI-assisted extraction, against 89.0% for human-only extraction, gives context, and 95% is a stretch target above both (www.acpjournals.org)	Domain scientist, with a second reviewer on a sample

Task	Access route	Evidence requirement	Failure types	Acceptance threshold (illustrative)	Reviewer
2. Review of an existing analysis pipeline	Standard API on non-sensitive code; most cybersecurity tasks are rerouted to Opus 4.8 (decrypt.co), so security-flavored prompts need separate logging	Each finding cites file and line and is reproduced with a minimal test; defects seeded in a copy of the pipeline	Missed silent defect; false positive; wrong diagnosis; refusal	At least 90% of seeded defects detected, with a stated share correctly located and classified; a synthetic benchmark reports detection F1 of 0.94 or higher across the configurations tested (aclanthology.org)	Bioinformatician, with quality assurance review of changes
3. Non-sensitive biology reasoning	Only questions permitted under default safeguards; fallback rate is itself a measured outcome. Dual-use areas such as virology and toxicology trigger fallback (support.claude.com)	Answer with cited primary sources and a rubric of required reasoning steps	Wrong conclusion; unsupported claim; refusal; mid-stream refusal	A rubric score set by the sponsor, with a fallback-rate ceiling; the Humanity's Last Exam authors report expert disagreement of about 18% on one biology, chemistry and health subset, which argues for expert adjudication (arxiv.org)	Senior scientist, blinded to model

The access column is where Opus 5.5 differs most from Opus 5. Task 1 is the most likely to pass without LSVP if the literature is not dual-use. Task 3 is the most likely to hit the classifier, and its fallback rate decides whether a sponsor needs LSVP at all. Because Opus 5 has no biology fallback, the two arms will not see the same refusal behavior, and the scorecard must report accepted, fallback and failed tasks separately.

Evidence for each task type

For literature extraction, a study within reviews found that Claude-assisted and human-only extraction agreed on **77.2%** of data elements (www.acpjournals.org), and the AI-assisted approach reached **91.0%** accuracy with a human reviewer verifying output, compared with 89.0% for human-only extraction (www.acpjournals.org). The preprint reported 41 minutes saved per study (www.medrxiv.org) and comparable major-error rates of 2.5% versus 2.7% (www.medrxiv.org). These studies used older Claude versions, not Opus 5.5. A study of 20 randomized trials concluded that human verification is necessary (pmc.ncbi.nlm.nih.gov), and a study of GPT-4o and o3 concluded that numeric extraction cannot yet be fully automated (www.cambridge.org).

For pipeline review, a 2026 benchmark of 200 synthetic bioinformatics snippets frames the key risk as defects that execute without error yet produce plausible, scientifically wrong results (aclanthology.org). No study of review for Nextflow or Snakemake pipelines specifically was found, which is why the scorecard uses seeded defects in the sponsor's own code.

For biology reasoning, LAB-Bench has over 2,400 multiple-choice questions (arxiv.org), and its authors observed that frontier models often refuse to answer some questions (arxiv.org). Its successor has nearly 1,900 tasks (arxiv.org) and reports model-specific accuracy differences from minus 26% to minus 46% across subtasks (arxiv.org). BixBench has 53 scenarios and 296 open-answer questions (arxiv.org), and one lab report found average task cost varied 367-fold across models (advances.edisonscientific.com). These results show why accuracy and cost need joint reporting, a point also made in agent-evaluation research (arxiv.org).

Scoring, judges and stop rules

Evaluation should be multi-metric. HELM measures seven metrics, including efficiency (arxiv.org). A model used as judge can exceed 80% agreement with humans but has documented position, verbosity and self-enhancement biases (arxiv.org), so the acceptance decision should rest with the human reviewer. Report inter-reviewer agreement with **Cohen's kappa**, because percent agreement can overstate reliability: one worked example has 0.94 percent agreement and a kappa of 0.85 (pmc.ncbi.nlm.nih.gov).

Illustrative stop rules, to be set before the run:

- **Unresolvable citation.** Any fabricated or unresolvable quote in Task 1 pauses the arm for root-cause review.
- **Fallback ceiling.** If the fallback rate on Task 3 exceeds the sponsor's ceiling, stop and decide between LSVP and no-go.
- **Review-time ceiling.** If reviewer minutes per task exceed the human-only baseline, stop.
- **Budget cap.** Stop any arm that exceeds its approved spend, including retries and fallback calls.

Regulatory framing

The pilot tasks are research support, not regulatory decisions. The **FDA draft guidance** proposes a seven-step, risk-based credibility assessment (www.fda.gov) in which model risk combines model influence and decision consequence (www.fda.gov). It is a draft distributed for comment (www.fda.gov), and it states that drug discovery is outside its scope (www.fda.gov). Where a human is in the loop, it recommends evaluating the human-AI team, not only the model (www.fda.gov), which is what the scorecard's review time captures. The **European Medicines Agency** adopted a reflection paper on AI in the medicinal product lifecycle in September 2024 (www.ema.europa.eu) and cautions that language models can produce plausible but erroneous output (www.ema.europa.eu). The draft EU Annex 22 consultation states that it does not apply to generative AI and large language models (health.ec.europa.eu) and that such models should not be used in critical good manufacturing practice (GMP) applications (health.ec.europa.eu). The U.S. National Institute of Standards and Technology defines confabulation as confidently stated but erroneous or false content (nvlpubs.nist.gov). None of this establishes that Opus 5.5 is validated for any regulated use, and a pilot result does not substitute for workload-specific validation.

Reception and Community Perspectives

This section reports sentiment and customer accounts, separate from measurements. Trade and analyst reaction focused on cost and caveats. One analyst note says Opus 5.5 will not be the most cost-effective option (www.constellationr.com), and a Forrester analyst quoted by InfoWorld describes a prolonged price-performance race among frontier labs (www.infoworld.com). Another outlet observes that the 40% claim leans on fewer tokens per job as well as a lower per-token price (mixed-news.com). TechRepublic noted that mandatory adaptive thinking means engineers can no longer toggle reasoning off (www.techrepublic.com).

Pharma adoption accounts exist for Claude in general but do not name Opus 5.5. A Sanofi testimonial on Anthropic's life-sciences page reports an assistant used every day by **roughly 60,000 employees** (claude.com). Fierce Pharma reported a rollout to **more than 30,000** Bristol Myers Squibb employees

(fiercepharma.com). Genmab announced custom Claude-powered agentic AI for clinical development priorities (in.marketscreener.com), and Benchling connected Claude to Benchling data (www.benchling.com). Anthropic's **Claude Science** app runs on the same models a plan includes, with the same safeguards (claude.com). None of these sources states which Claude model version is in use, so they do not validate Opus 5.5 for any task, and the testimonials are vendor-curated.

Data Analysis and Evidence

The cost formula

The scorecard's central metric is cost per accepted task:

```
cost per accepted task =
  (input + cache write + cache read + output
   + retries and fallback + reviewer time)
  / accepted tasks
```

Source date for all prices on the worksheet: November 15, 2026 (vendor pages fetched that day). Re-date every row when prices change.

Each term needs a rule. Input is uncached input tokens at the base rate. Cache writes and reads are priced separately (see Table 1). Output includes thinking tokens (platform.claude.com). Models from the 4.7 generation onward use a tokenizer that produces approximately **30% more tokens** for the same text (platform.claude.com), so token counts from older models do not transfer. Caching and batch discounts can combine (platform.claude.com).

Refusals and fallback have specific billing rules. Pre-output refusals in some categories, including biology, are billed (platform.claude.com), and when a fallback is triggered the refusal is billed in addition to the fallback request (platform.claude.com). Anthropic credits the cache miss that a model switch causes (platform.claude.com). The documentation notes that beneficial life-sciences work can also trigger the biology category (platform.claude.com). The retries and fallback line must therefore include refused calls and the calls that replaced them.

A hypothetical worked example

(Hypothetical Example.) The inputs below are invented for illustration and are not measurements. Assume 100 literature-extraction tasks. Each has 40,000 uncached input tokens, an 8,000-token prefix read from a 1-hour cache, and 6,000 output tokens including thinking. Five tasks are retried once. Three tasks hit the biology classifier before output and are re-run on Opus 5 as the fallback. Reviewers spend 6 minutes per task at a loaded cost of $90 per hour ($9 per task). Opus 5.5 yields 88 accepted tasks and Opus 5 yields 85. Prices are the list prices in Table 1.

- **Opus 5.5 token spend.** Uncached input $16.00, cache write $0.064, cache reads $0.16 and output $12.00, total $28.224.

- **Opus 5.5 retries and fallback.** Retries $1.408, billed pre-output refusals $0.48 and the Opus 5 fallback runs $1.05 (3 runs at a rounded $0.35 each; $0.354 each would give $1.06, which does not change the result), total $2.938. Pre-output refusals are costed on input tokens only.
- **Opus 5.5 review and total.** Reviewer time $900, for a total of $931.162, or **$10.58 per accepted task** ($931.162 divided by 88).
- **Opus 5 token spend and retries.** Input $20.00, cache write $0.08, reads $0.40, output $15.00 and retries $1.77, total $37.25, with no biology fallback.
- **Opus 5 total.** With $900 of review the total is $937.25, or **$11.03 per accepted task** at 85 accepted tasks ($937.25 divided by 85).

Three readings follow:

- **Review dominates.** Reviewer time is about 97% of the Opus 5.5 total ($900 of $931.16).
- **Token gap is small.** The API-cost gap between the two models is $6.09 across 100 tasks, or about $0.07 per accepted task.
- **Acceptance matters more.** One more accepted task per 100 lowers the Opus 5.5 figure from $10.58 to $10.46, a drop of about $0.12, which exceeds the whole token-price difference.

The example adds fallback costs on top of full-task spend for all 100 tasks, a conservative simplification. In this invented case, acceptance rate and review minutes matter more than list price. Real workloads with long agentic runs will have a larger token share, which is why the pilot measures rather than assumes.

Blank pilot worksheet

The block below is a blank worksheet, not results. No pilot has been run for this report. Copy one block per task and arm.

```
PILOT SCORECARD (BLANK, NOT RESULTS)
Price source and date: _____
Task (1, 2 or 3): ____   Arm (model / effort / platform): _____________
Fallback configured (yes/no, target model): _____________
Tasks attempted: ____   Tasks accepted: ____   Fallback tasks: ____   Refused, no fallback: ____
Uncached input tokens: ____   Cache write tokens (5 min / 1 h): ____ / ____
Cache read tokens: ____   Output + thinking tokens: ____
Retry runs: ____   Billed refusals: ____   Fallback runs (model): ____
Reviewer minutes per task: ____   Loaded reviewer rate per hour (USD): ____
Error counts by type (wrong value / missed / unsupported / other): ____
Inter-reviewer kappa on a double-reviewed sample: ____
Cost per accepted task (USD): ____
Access confirmation on file (LSVP scope, model, platform, date): ____
```

Implications and Future Directions

Pilot decision matrix

The matrix below turns the evidence into a go/no-go frame. It is a decision aid, not a validation verdict.

- **Go: pilot Opus 5.5 on Tasks 1 and 2.** The data is public or non-sensitive, default biology safeguards are unlikely to bind, and the platform can be first-party or cloud. Run both arms at matched effort and at vendor defaults.
- **Conditional: Task 3 only with an access answer.** If the biology scope touches dual-use areas, the pilot needs a written LSVP confirmation that names Opus 5.5, the platform and the retention terms. Without it, the Opus 5.5 arm measures fallback to Opus 5, not Opus 5.5 itself.
- **No-go: cloud-marketplace workloads that need relaxed biology safeguards.** LSVP runs on first-party products only.
- **No-go: protected health information in an organization holding an LSVP grant.** BAA organizations cannot receive LSVP grants.
- **Stay on Opus 5.** If a sponsor's tasks trip the classifier often, Opus 5 has no biology fallback and is one of the models named for Standard Use on the program page.
- **Consider a comparator.** A lower-priced general model or a trusted-access life-sciences model may fit some tasks, and the scorecard compares them on the same terms.

Procurement and research IT questions

- **Scope.** Does the LSVP grant cover Opus 5.5 specifically, for which grant type and which projects?
- **Retention.** Is 30-day retention under LSVP acceptable to legal and data governance?
- **Platform.** Can the workload run on Anthropic's first-party platform?
- **Fallback.** Is server-side fallback enabled, and what happens when a target model is rate limited or a refusal and its replacement are both billed?
- **Versioning.** What is the retirement horizon, and who re-runs the scorecard on model changes?
- **Quality.** Who owns the reference set, the adjudication process and sign-off for any use beyond research support?

Outlook

Several variables will move after this report's date. Anthropic's pages now also list Fable 5.2 as a Mythos-class model (support.claude.com). Gemini 4 Carbon's list price is scheduled to double on **December 1, 2026** (ai.google.dev), and OpenAI's Rosalind remains available to approved organizations only (help.openai.com). Safeguard classifiers are retuned over time: Anthropic reported that an August 2026 retune cut biology-related fallbacks for Fable 5 by about 85% (www.anthropic.com), so a fallback rate measured in a pilot is time-stamped.

IntuitionLabs' advisory view is that the scorecard should outlive the model. A fixed task set, a frozen reference and a documented review process let a team re-run the same comparison when Opus 5.5 is replaced, when LSVP coverage changes or when prices move, without restarting the debate.

Frequently Asked Questions (FAQs)

What does Claude Opus 5.5 cost, and what is the pricing for pharma teams?

List pricing is $4 per million input tokens and $20 per million output tokens, with cache reads at $0.20 per million (see Table 1) and batch at $2 and $10 (see Pilot Design). Anthropic's 40% saving claim is conditional on typical workloads and its own tests (thenewstack.io). Pharma teams should compute cost per accepted task, including review time and fallback, instead of relying on list price.

Is Claude Opus 5.5 ready for life sciences and clinical development work?

No public evidence validates Opus 5.5 for clinical development or any regulated workload. Regulators ask for credibility assessments tied to a context of use (www.fda.gov), and a pilot can inform that assessment but does not replace it.

Can Claude be used for drug discovery with Opus 5.5?

Drug discovery and research biology are named among the work that generally available Fable models block and the LSVP is intended to enable, according to Anthropic's LSVP announcement. Whether Opus 5.5 is covered for a given organization is unconfirmed on the pages reviewed, so it should be confirmed in writing.

Who gets Claude Opus 5.5 biology access?

Dual-use requests fall back to Opus 5 under default safeguards. Vetted organizations can apply to the Life Sciences Verification Program (LSVP) for broader biology access. Anthropic's program page lists Standard Use grants for Mythos 5.1, Opus 5 and Sonnet 5 today and for future models as they launch, but no Anthropic page reviewed names Opus 5.5, so coverage should be confirmed in writing.

How does Claude Opus 5.5 compare with GPT for pharma research?

No independent pharma head-to-head was found. On list price, GPT-6.1 Sol is $2 and $10 (developers.openai.com) and GPT-6 Astra is $10 and $50 (developers.openai.com). OpenAI's life-sciences model, Rosalind, is available only to approved organizations through a trusted-access program. A comparator arm in the scorecard answers the question for a specific workload.

What is an LLM pilot scorecard for pharma R&D, and how is AI pilot cost and quality evaluated?

It is a fixed protocol that runs bounded tasks under fixed settings, records accepted, fallback and failed outputs, and computes cost per accepted task from tokens, retries and reviewer time. Evaluation research recommends reporting cost alongside accuracy (arxiv.org).

How do Claude biosecurity safeguards affect life sciences access?

Opus 5.5 runs a biology classifier, and the classifiers scan everything the model reads. Cyber classifiers remain in place even under an LSVP grant (www.anthropic.com).

Conclusion

Claude Opus 5.5 is cheaper per token than Opus 5 and, on several independent measures, more capable, but the evidence that matters to a pharma R&D lead is thin. The 40% saving is conditional, the benchmark spread on Terminal-Bench-Science runs from about 47% to about 63% depending on harness, and no public study measures literature extraction, pipeline review or biology reasoning on Opus 5.5 in a pharma setting. Access adds a second layer: the model carries the same biology safeguards as Fable 5.1, and LSVP coverage for Opus 5.5 should be confirmed in writing before any biology-heavy pilot is scoped.

The practical path is a controlled pilot. Fix the model list, effort, platform, fallback, input mix and review process. Measure accepted tasks, fallback tasks and failures separately. Price reviewer minutes as carefully as tokens, since in the hypothetical example they dominate the bill. Keep the scope to research support, state the context of use, and treat any step toward a regulated decision as a separate validation question for quality and regulatory teams.

A team that follows this approach ends the pilot with a defensible number for its own workload: a cost per accepted task for each model, a measured fallback rate, a documented access position and a clear go or no-go. That is more useful than any vendor-reported benchmark, and it remains reusable when the next model arrives.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://platform.claude.com/docs/en/models/opus-5-5/overview
2. https://www.techrepublic.com/article/news-anthropic-claude-opus-5-5-pricing-performance
3. https://venturebeat.com/technology/anthropic-releases-claude-opus-5-5-beating-fable-5-1-on-key-agentic-benchmarks-at-60-cheaper-api-price
4. https://www.anthropic.com/claude-opus-5-5
5. https://artificialanalysis.ai/articles/claude-opus-5-5
6. https://hlth.com/insights/news/anthropic-launches-claude-opus-5-5-with-expanded-biology-capabilities
7. https://support.claude.com/en/articles/16049681-why-claude-switched-models-in-your-conversation-with-opus-5-or-opus-5-5
8. https://www.anthropic.com/news/life-sciences-verification-program
9. https://support.claude.com/en/articles/16975617
10. https://www.vals.ai/models/anthropic_claude-opus-5-5
11. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological
12. https://mixed-news.com/en/claude-opus-5-5-price-4-per-million-tokens-usage-limits
13. https://anthropic.com/claude-opus-5-5-system-card
14. https://decrypt.co/378947/anthropic-claude-opus-5-5-fable-price-comparison
15. https://platform.claude.com/docs/en/about-claude/pricing
16. https://platform.claude.com/docs/en/models/opus-5-5/whats-new-opus-5-5

17. https://platform.claude.com/docs/en/build-with-claude/thinking
18. https://platform.claude.com/docs/en/about-claude/models/optimizing-for-cost-and-intelligence
19. https://www.digitaltrends.com/computing/claude-opus-5-is-here-and-anthropic-says-it-can-rival-fable-5-in-some-tasks
20. https://thenewstack.io/claude-opus-5-5-release
21. https://www.anthropic.com/claude-fable-and-mythos-5-1
22. https://developers.openai.com/api/docs/pricing
23. https://help.openai.com/en/articles/20001193-gpt-rosalind-for-life-sciences-research
24. https://ai.google.dev/gemini-api/docs/pricing
25. https://deepmind.google/models/model-cards/gemini-4-carbon
26. https://aibusiness.com/generative-ai/anthropic-unveils-opus-5-5
27. https://platform.claude.com/docs/en/build-with-claude/refusals-and-fallback
28. https://www.digitaltrends.com/computing/anthropic-launches-claude-opus-5-5-with-fable-5-1-level-performance-at-a-40-lower-price
29. https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v5
30. https://www.terminal-bench-science.ai
31. https://www.tbench.ai/news/terminal-bench-science-0-1
32. https://www.vals.ai/benchmarks/terminal-bench-science
33. https://artificialanalysis.ai/evaluations/terminal-bench-science
34. https://www.vals.ai/benchmarks/medcode
35. https://www.vals.ai/benchmarks/biomysterybench
36. https://www.vals.ai/benchmarks/science_index
37. https://artificialanalysis.ai/models/claude-opus-5-5
38. https://artificialanalysis.ai/models/claude-opus-5
39. https://www.digitalapplied.com/blog/claude-opus-5-5-launch-pricing-benchmarks-2026
40. https://platform.claude.com/docs/en/build-with-claude/effort
41. https://www.mobihealthnews.com/news/anthropic-launches-claude-opus-55-expanded-biology-capabilities
42. https://platform.claude.com/docs/en/build-with-claude/fallback-credit
43. https://platform.claude.com/docs/en/build-with-claude/prompt-caching
44. https://platform.claude.com/docs/en/build-with-claude/batch-processing
45. https://www.acpjournals.org/doi/10.7326/ANNALS-25-00739
46. https://aclanthology.org/2026.realm-1.65
47. https://arxiv.org/html/2501.14249
48. https://www.medrxiv.org/content/10.1101/2025.03.20.25324350v1.full
49. https://pmc.ncbi.nlm.nih.gov/articles/PMC12462964
50. https://www.cambridge.org/core/journals/research-synthesis-methods/article/automating-the-data-extraction-process-for-systematic-reviews-using-gpt4o-and-o3/28C54A601B80E2BE6611C8B7B15FEDAF
51. https://arxiv.org/abs/2407.10362

52. https://arxiv.org/html/2407.10362v3
53. https://arxiv.org/abs/2604.09554
54. https://arxiv.org/html/2503.00096v2
55. https://advances.edisonscientific.com/research/bixbench3-benchmarking-study-scale-computational-biology
56. https://arxiv.org/abs/2407.01502
57. https://arxiv.org/abs/2211.09110
58. https://arxiv.org/abs/2306.05685
59. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052
60. https://www.fda.gov/media/184830/download
61. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf
62. https://health.ec.europa.eu/document/download/5f38a92d-bb8e-4264-8898-ea076e926db6_en?filename=mp_vol4_chap4_annex22_consultation_guideline_en.pdf
63. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
64. https://www.constellationr.com/index%2ephp/insights/news/anthropic-launches-claude-opus-55-gets-cost-efficiency-memo
65. https://www.infoworld.com/article/4225637/openai-anthropic-cut-ai-model-costs-as-price-performance-race-intensifies.html
66. https://claude.com/solutions/life-sciences
67. https://fiercepharma.com/pharma/bristol-myers-taps-anthropic-claude-enterprise-wide-ai-adoption-speed-drug-rd-global
68. https://in.marketscreener.com/news/genmab-a-s-partners-with-anthropic-to-accelerate-research-and-development-through-agentic-artificial-ce7e59dcde8ff726
69. https://www.benchling.com/news/benchling-partners-with-anthropic-to-build-a-bridge-between-science-and-ai
70. https://claude.com/product/claude-science
71. https://support.claude.com/en/articles/15363606
72. https://www.anthropic.com/news/improving-fable-5-s-biology-safeguards

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.