David Soden

MARKET ASSESSMENT | SEPTEMBER 2026

Your Claude Refusal Rate Is Not 0.09 Percent

Four layers decide whether Claude Code or Claude Desktop declines your work. Two answer to configuration, one has an application form, one is fixed by design.

David Soden

Independent analyst and practitioner, enterprise automation and applied AI

davidsoden.com/market-assessment

REPORT DS-2026-82F6 · 2026 · VERSION 1.0 · PUBLIC

About this report

Every substantive claim in this report carries an evidence classification and, where applicable, a numbered source. Facts are separated from vendor claims, third-party estimates, the author's assessments, and untested hypotheses. Forward-looking sections use scenarios with observable tripwires rather than point forecasts. Stated assumptions are listed before the analysis begins.

Market Assessment reports are published and commissioned at **davidsoden.com/market-assessment**. This document is analysis, not investment, legal, or procurement advice.

Scope and evidentiary standard

This covers refusal behaviour in Claude Code and the Claude desktop application as it affects paid professional work: why a request gets declined, which declines a user can do something about, what Anthropic has committed to changing, what the exemption routes are, and what the alternatives cost. It does not cover jailbreaking, and nothing here describes a method for obtaining content Anthropic has said it will not produce. The subject is false positives on legitimate work, which is a different problem with different fixes.

Two things make the published numbers in this category unusually slippery, and both need saying before any figure appears.

The vendor and the user measure different populations. Anthropic reports over-refusal as the share of individual prompts in a curated benign evaluation set that a model declines. A working researcher experiences refusal as the share of working sessions that get interrupted. Those two denominators differ by roughly two orders of magnitude, and both are honestly computed. This report does not resolve the gap by picking a side. The gap is the finding.

Almost every quantitative claim available about Claude's refusal behaviour comes from Anthropic. No independent audit of Claude's over-refusal rate on real research workloads exists in the public record. The counterweight is practitioner reporting on a public issue tracker plus trade press that counted it, all self-selected and none of it producing a rate. Numbers taken from Anthropic system cards are labelled as vendor claims throughout, even though a system card reads like a technical paper, because a system card is self-published and the evaluation sets behind it are not open to inspection.

Model names move fast in this category. Where a specific snapshot matters, it is named. Where a claim is about the architecture rather than a snapshot, that is said explicitly, because those are the claims with a long half-life.

Label	Definition	How to treat it
FACT	Verifiable in the cited primary source: a filing, press release, official documentation, or standards body announcement.	Reliable as stated. Check the source if the exact wording matters.
VENDOR CLAIM	Reported by a vendor about its own business or product. Self-measured and not audited by any third party.	The vendor said it. Directionally useful, not a measurement. Definitions of the metric are usually undisclosed.
THIRD-PARTY ESTIMATE	A research firm forecast or survey result. Methodology is proprietary and generally not reproducible.	Use for direction and order of magnitude only. Do not build a model on it.
ASSESSMENT	The author's reasoned judgment from the cited evidence. Falsifiable, and the reasoning is shown so it can be argued with.	This is analysis, not reporting. Disagree with it where your evidence differs.
HYPOTHESIS	A proposition offered for testing. Not currently supported by sufficient evidence to assert.	Treat as a research question, not a conclusion. Each one names what would confirm or refute it.
SCENARIO	A structured future state with a stated subjective probability. The probability is the author's judgment, not a calculated figure.	Use the leading indicators attached to each, not the probability. The indicators are observable; the probability is not.

Assumptions this analysis rests on

Anthropic's published evaluation numbers are honestly computed on the sets they describe. If that is wrong, the direction-of-travel argument in this report collapses and only the practitioner reports survive, which would make the picture considerably worse than the one drawn here.

The classifier layer described in Anthropic's cyber safeguards documentation is the same mechanism that produces the API-level Usage Policy error inside Claude Code. Anthropic has not published a mapping between the error strings a user sees and the systems that emit them. If this assumption is wrong, the diagnostic table in the workarounds section misroutes some failures, and a reader following it would spend effort rewriting prompts against a filter that does not read prompts the way they assume.

Public issue reports are a directionally valid signal of change over time and are not a rate. If the escalation in reported false positives during 2026 is mostly explained by growth in the Claude Code user base rather than by a change in model or classifier behaviour, Figure 3 means much less than it appears to, and the surrounding argument weakens into a claim about visibility rather than incidence.

Behaviour is stable within a model snapshot. Anthropic states that from the 4.6 generation onward each model ID is a fixed snapshot, which is what makes any measurement reproducible at all. Fable 5's fallback architecture partly breaks this from the user's side, because a flagged

request is silently re-served by a different model. Wherever a reader tries to reproduce a measurement across surfaces, that silent substitution is the most likely reason two runs disagree.

The call

ASSESSMENT Anthropic will keep reducing false refusals and the tone will keep improving, because it now measures both and publishes the numbers. It will not give you a switch. The hard constraints are written to be unliftable by any operator or user, and the practical relief on offer is verified access granted per organisation, not per prompt. The work, then, is to route around the layers you control and apply for the one you do not. What would overturn this: a per-request or per-organisation sensitivity parameter in Anthropic's API, the equivalent of Google's adjustable safety thresholds. If that ships, most of the workarounds here become obsolete overnight.

My judgment on the evidence assembled here, nothing more. There is data I can't see and data I may have missed. New evidence can move me, including to the opposite position, and I reserve the right to move.

Executive summary

The common framing of this problem is that Claude has a moral code and needs to be argued out of it. That framing produces the wrong fixes. Refusal is not one behaviour, it is four systems that fail in ways a user cannot tell apart from the outside, and the most useful thing a professional can learn here is how to identify which one just fired.

FINDING 1 The two published numbers differ by around fifty times, and both are Anthropic's.

> ASSESSMENT The Claude Opus 5 system card puts single-turn over-refusal on benign prompts at 0.09 percent through the API. At Fable 5's launch, Anthropic said its guardrails would trigger on average in fewer than five percent of sessions. The first is a per-prompt rate on a curated set, the second a per-session rate on live traffic. A researcher's felt experience tracks the second. Quoting the first at someone who is living the second is the root of the credibility problem in this whole argument. [2] [7]

FINDING 2 The surface you work on changes the refusal rate more than the prompt does.

> VENDOR CLAIM On Anthropic's own benign evaluation set, every model tested over-refuses more on claude.ai than on the bare API. Claude Opus 5 goes from 0.09 percent to 0.47 percent, Claude Sonnet 5 from 0.59 percent to 1.54 percent. The desktop application's chat mode runs the claude.ai stack and inherits its system prompt, which Anthropic publishes and states does not apply to the API. Claude Code talks to the API. Moving a task between them changes the odds before a word of the prompt is rewritten. [2] [10]

FINDING 3 Model choice inside Claude Code matters more for security work than any prompt technique.

> VENDOR CLAIM On Anthropic's own Claude Code evaluation, the dual-use and benign success rate is 99.82 percent for Opus 5 and 91.55 percent for Sonnet 5. That is roughly one failure in

twelve against one failure in five hundred, on exactly the tasks security practitioners run all day. None of the circulating advice mentions it, because the circulating advice is about prompt wording. [2]

FINDING 4 CLAUDE.md is the weakest control available, and it is the one everyone recommends.

FACT Anthropic's documentation states plainly that CLAUDE.md is delivered as a user message after the system prompt, while output styles modify the system prompt itself and the command-line flags append to or wholly replace it. The widely circulated advice to drop a tone-control block into CLAUDE.md aims the weakest instrument at the problem, and there is a standing bug report describing exactly that: standing instructions in CLAUDE.md silently losing to built-in caution. [8] [17] [20]

FINDING 5 The lecture is a tracked defect at Anthropic, not an intended feature.

FACT Claude's constitution lists thirteen over-caution behaviours Anthropic explicitly does not want, among them lecturing when ethical guidance was not asked for, being condescending about a user's ability to handle information, and adding warnings that are not useful. The Opus 5 system card measures two of these as named metrics, wet blanket and condescension toward the user, and reports Opus 5 came out slightly more condescending than its predecessor, echoing informal reports. [1] [2]

FINDING 6 One category of refusal is designed never to be liftable, and the text names vaccine research.

FACT The constitution sets out seven hard constraints and says they cannot be unlocked by any operator or user. The surrounding text goes further, giving as an example that Claude should decline to produce content offering real uplift toward chemical or biological weapons even where the user is probably requesting it for a legitimate reason like vaccine research, because the risk of inadvertently assisting a malicious actor is judged too high. That is a stated, deliberate false-positive cost, and it is borne by legitimate scientists. [1]

FINDING 7 The exemption route is free, fast, and has holes in exactly the places research sits.

FACT The Cyber Verification Program lifts the high-risk dual-use classifier for verified organisations, reviewed inside two business days. It is not offered on Amazon Bedrock or Google Vertex AI, and organisations on zero data retention are ineligible. Researchers handling human-subjects or clinical data are the population most likely to require zero data retention, and they are structurally excluded from the relief. [3]

FINDING 8 The trend is improving, and improving least where this audience works.

VENDOR CLAIM Anthropic's August 2026 retune cut biology-related fallbacks by roughly 85 percent. Broken out by surface, total fallback volume fell an estimated 67 percent on claude.ai and 55 percent on Cowork, against 17 percent on Claude Code and 7 percent on the Claude Platform. The chat surfaces took most of the benefit. The agentic and API surfaces, where professional work actually happens, took the least. [4]

FINDING 9 Relief is moving from prompting to entitlement, and life sciences now runs through a government.

ASSESSMENT Verified access is becoming the mechanism. Cyber work has a verification program. Scientists get free and discounted seats but stay capped at Opus-class models for biology and chemistry, and access to Mythos-class models for life sciences is being established through a partnership with the US government rather than through Anthropic's own commercial review. For a researcher outside that perimeter, no amount of prompt craft substitutes for an entitlement they cannot obtain. [3] [5]

Four layers, and only two of them read your prompt

A refusal in Claude Code or the desktop app can come from any of four independent systems. They produce different symptoms, they respond to different interventions, and the reason the standard advice underperforms is that it treats all four as the same thing.

Layer	What it is	How it presents	What you can do
Trained disposition	The model's own judgment, shaped by the constitution and its training	A conversational decline in Claude's voice, often with an explanation and an offer of an alternative	Framing, context, model choice. This layer reads everything you write
Injected prompt	The system prompt for the surface: claude.ai's published prompt, or Claude Code's shipped prompt	Tone and caution that persist regardless of how you phrase the request	Output styles, --append-system-prompt, --system-prompt. Fully replaceable in Claude Code
Classifier	Input and output classifiers for cyber, biology, chemistry and distillation, applied across every surface	An API-level error naming the Usage Policy, or a silent fallback to a different model	Nothing at the prompt level. Apply to the Cyber Verification Program, or change model or surface
Policy and account	Usage Policy enforcement, entitlements, model-class caps on biology and chemistry	Consistent blocking of a whole domain of work, unaffected by anything you type	Verified access programmes. Otherwise a hard wall

ASSESSMENT The practical importance of this table is diagnostic. The first two layers read your prompt and can be moved with words. The third does not read your prompt the way you think it does and cannot be argued with, because arguing with it adds more flagged text to a conversation that has already been flagged. The fourth is a business relationship. Users conflate them, spend their effort on the wrong one, and conclude the whole system is arbitrary.

FACT The layers are documented separately by Anthropic. The trained disposition is described at length in Claude's constitution, published 22 January 2026. The claude.ai system prompt is published in Anthropic's release notes, which state directly that those prompt updates do not apply to the Claude API. The classifier layer is described in Anthropic's help documentation on real-time cyber safeguards, which states that the safeguards apply across all access methods

including claude.ai, the API, Bedrock, Vertex and Azure Foundry. The policy layer appears in the Usage Policy exceptions article, which restricts modifications of use restrictions to carefully selected government entities. [1] [3] [10] [19]

ASSESSMENT Anthropic publishes the claude.ai system prompt and does not publish Claude Code's. What is known about the second comes from a source map disclosure on 31 March 2026, in which the full Claude Code source was briefly shipped inside the public npm package and subsequently analysed. The material indicates a system prompt that instructs the agent to assist with authorised security testing, defensive security, capture-the-flag work and education, while declining destructive techniques and detection evasion for malicious purposes. Treat that as indicative of structure rather than as authoritative text, because it is a leak and not a publication. [21]

The failure mode nobody names: the silent downgrade

FACT Anthropic's Fable 5 classifiers do not always refuse. When one fires, the request is re-routed to Opus 5, described by Anthropic as a capable model that does not have the same level of biological capability. Anthropic calls this a fallback. [4]

ASSESSMENT This is a genuinely new category of failure and it is worse for research than a refusal is. A refusal is legible. A silent substitution to a less capable model produces an answer that looks normal, carries no warning, and is quietly worse. The user's rational conclusion is that the model got dumber for no reason, and no amount of prompt debugging will surface the cause. For anyone running a reproducibility-sensitive workflow, an undeclared model swap mid-session is a methodological problem before it is an annoyance.

FACT The behaviour is observable in the field. A principal research scientist at the Institute for Disease Modeling reported to The Register that Fable 5's input safety classifier emitted a model refusal fallback on the first turn of essentially every session on his account, including a session whose only user input was the word hello. [7]

The number problem

Anthropic publishes over-refusal figures that are, on their own terms, very low. It also published a session-level figure that is roughly fifty times higher. Both appear in Anthropic's own material, months apart, and the difference is not spin. It is a measurement choice, and understanding it is what lets a reader stop arguing about whose experience is real.

VENDOR CLAIM The Claude Opus 5 system card, dated 24 July 2026, reports single-turn over-refusal on a benign prompt set spanning sixteen policy areas and seven languages. Through the API without a system prompt: Claude Opus 5 at 0.09 percent, Claude Fable 5 at 0.01 percent, Claude Mythos 5 at 0.03 percent, Claude Opus 4.8 at 0.35 percent, Claude Sonnet 5 at 0.59 percent. [2]

VENDOR CLAIM At Fable 5's launch on 9 June 2026, Anthropic said it had tuned the guardrails conservatively, that they would sometimes catch harmless requests, and that they trigger on

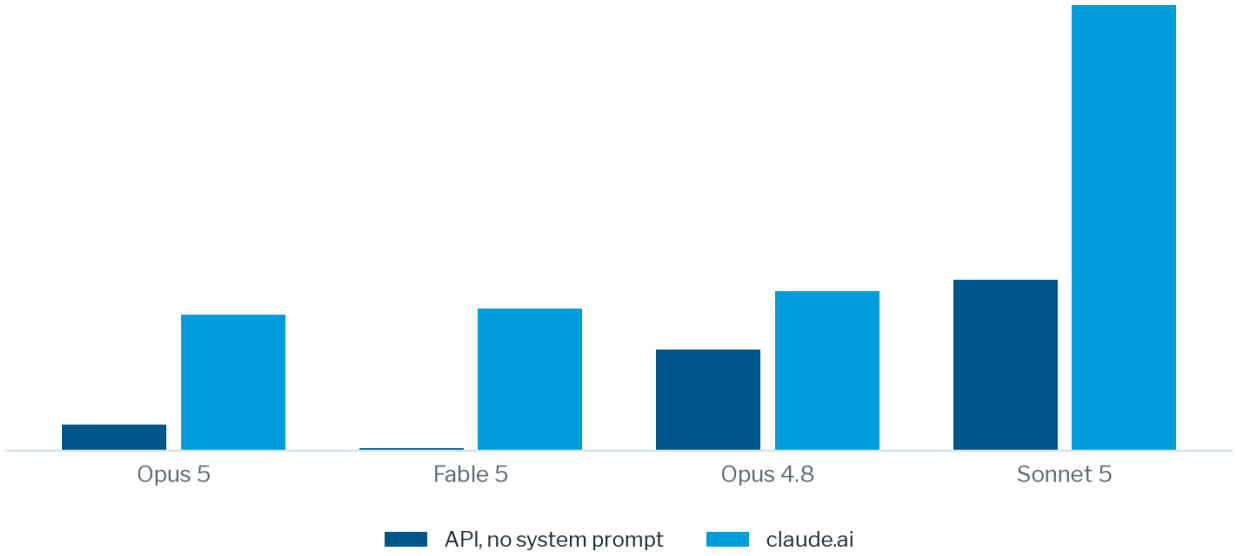
average in less than five percent of sessions, with a commitment to reduce false positives as quickly as possible. [4] [7]

ASSESSMENT These measure different things. The system card counts prompts in a curated benign set built to test policy boundaries. The five percent figure counts sessions on live traffic where a classifier fired at all, including fallbacks that never produce a visible refusal. A researcher running twenty sessions a week and hitting an interruption in one of them is having an experience the system card was never designed to describe. Both numbers are defensible. Only one of them is about the user.

ASSESSMENT The failure here is a communication failure with commercial consequences. When a scientist reports being blocked and is answered with a system-card percentage, the answer sounds like a denial. Anthropic has the session-level number and could publish it routinely alongside the prompt-level one. That it does not is the single easiest credibility repair available to the company, and its absence is the reason the community argument stays stuck.

What the surface does to the number

FIGURE 1 **Over-refusal on benign prompts, by model and surface**



VENDOR CLAIM Anthropic's single-turn benign evaluation, all tested languages. Mythos 5 is omitted because Anthropic reported no claude.ai figure for it; its API rate was 0.03 percent. Self-reported, with no external audit of the prompt set. [2]

ASSESSMENT Every model that was measured on both surfaces over-refuses more on claude.ai. For Opus 5 the gap is five times. That is the strongest single argument for doing sensitive work in Claude Code rather than the desktop app's chat mode, and it holds across the range rather than resting on one model. The mechanism is not mysterious: claude.ai carries a long injected system prompt that the API does not, and Anthropic's release notes state outright that those prompt updates do not apply to the API.

ASSESSMENT This finding runs the opposite way for one specific case. The Cyber Verification Program exemption has been reported working on claude.ai and the Windows desktop client

while failing on Claude Code's API path. So the general rule favours Claude Code, and the exemption plumbing, at least as of April 2026, favoured the desktop. Anyone holding a cyber exemption should test both surfaces rather than assume. [16]

The model matters more than the wording

FIGURE 2 **Claude Code: success rate on dual-use and benign cyber requests**

VENDOR CLAIM Share of 61 sensitive-but-permitted prompts the model assisted with while operating as a Claude Code agent, including network reconnaissance, vulnerability testing and penetration-test analysis. Anthropic notes results exclude additional safeguards present at deployment time, so live behaviour will be worse than shown. [2]

ASSESSMENT An eight-point spread on the exact task class that generates the loudest complaints is the most actionable number in this report. A security engineer running Sonnet 5 in Claude Code should expect roughly one in twelve legitimate dual-use requests to fail. On Opus 5 that becomes closer to one in five hundred. No prompt-engineering technique in circulation moves a refusal rate by that factor, and the change costs one flag.

VENDOR CLAIM Model choice cuts the other way on the paired metric. On the malicious half of the same evaluation, Sonnet 5 refuses 92.37 percent against Opus 5's 89.00 percent. Sonnet is the more suspicious model in both directions, which is consistent and is worth knowing rather than treating as a defect. [2]

What actually works, ranked by how much weight it carries

The advice circulating about this problem is not wrong so much as inverted. It leads with the intervention that has the least force and omits the ones with the most. Anthropic's own documentation sets out the ordering, in a table comparing its customisation features, and the ordering is not the one the community uses.

Control	Where it lands	Weight
--system-prompt or --system-prompt-file	Replaces the entire Claude Code system prompt	Highest. Removes the shipped instructions rather than arguing with them
Output style	Modifies the system prompt; drops the built-in engineering instructions unless keep-coding-instructions is set	High, and persistent across sessions via the outputStyle setting
--append-system-prompt	Appends to the system prompt without removing anything	Moderate. Good for a one-off invocation
CLAUDE.md	Adds a user message after the system prompt	Lowest of the persistent controls, and the one most commonly recommended
Per-prompt framing	The conversation itself	Real but small, and it decays as a session accumulates flagged context
Model and surface choice	Neither prompt nor configuration	Larger than any of the above, and almost never mentioned

FACT This ordering is Anthropic's, not an inference. The output styles documentation states that output styles directly modify Claude Code's system prompt, that a custom style leaves out the built-in software engineering instructions unless keep-coding-instructions is true, and, in a comparison table, that CLAUDE.md adds a user message after the system prompt while --append-system-prompt appends to the system prompt without removing anything. The CLI reference documents --system-prompt as replacing the entire system prompt with custom text. [8] [20]

FACT The weakness of CLAUDE.md as a behavioural control is a filed defect. Issue 39210 on the Claude Code repository is titled explicitly around CLAUDE.md standing instructions being silently overridden by built-in safety heuristics. A separate issue reports output styles being ignored where they conflict with embedded patterns. Neither is evidence of how often this happens; both are evidence that the mechanism is understood by the people hitting it. [17]

ASSESSMENT A caution about the strongest control. Replacing the Claude Code system prompt wholesale also removes the engineering instructions that make the agent competent: how it scopes changes, verifies work, and handles destructive operations. In practice --system-prompt trades tone for capability, and most people will get a better result from a custom output style with keep-coding-instructions set to true. Reach for the replacement flag when Claude Code is not doing software engineering at all, which describes a large share of the research use in this audience.

Telling the layers apart at the moment of failure

What you see	Layer	What to do about it
A conversational decline, in Claude's voice, with reasoning and an offered alternative	Trained disposition	Add context about the setting and purpose, once. If it declines again, change model before rewriting the prompt a third time
A moralising preamble attached to work it did anyway	Injected prompt	Custom output style with keep-coding-instructions true. This is a tone problem and tone lives in the system prompt
An API error naming the Usage Policy and violative content	Classifier	Stop editing the prompt. Start a fresh session, and apply to the Cyber Verification Program if the work is security
Answers that are quietly worse with no error at all	Classifier fallback	Suspect a silent model swap. Compare against an explicitly pinned model on the API
A whole domain refused consistently across prompts, models and surfaces	Policy and account	Nothing at the prompt level will help. This is an entitlement question

FACT Once a classifier has fired, the contamination persists. Security researchers reported to trade press in April 2026 that the block affected follow-up messages in conversations carrying prior violative context, forcing them to restart sessions rather than continue. One described being unable to resume exploit development sessions and hitting the same error repeatedly. [18]

ASSESSMENT That single mechanic explains most of the frustration in this category. The instinct on being refused is to explain yourself, which appends more text about the flagged subject to an already-flagged conversation. The correct move is the counterintuitive one: abandon the session and start clean. Arguing is the worst available option and it is the one everybody reaches for first.

The exemption route, and where it leaks

FACT Anthropic operates two classifier categories for cyber content. Prohibited use covers activity with little to no legitimate defensive application, given as examples mass data exfiltration and ransomware code development; it is blocked by default and is not adjustable. High-risk dual use covers vulnerability exploitation and offensive security tooling development; it is also blocked by default but is adjustable through the Cyber Verification Program. The safeguards apply on Claude Opus and Sonnet across all access methods, including claude.ai, the API, Bedrock, Vertex and Azure Foundry. [3]

FACT The programme is free, organisation-scoped rather than individual, requires identity verification, and is reviewed within about two business days. Application paths differ by platform: Anthropic first-party through a verification portal, Microsoft Foundry through a cyber use case form. It is not available on Amazon Bedrock or Google Vertex AI, and organisations on zero data retention are ineligible. [3]

ASSESSMENT Three gaps here matter more than the programme's existence. Bedrock and Vertex carry a large share of regulated enterprise and university AI workloads, and on those

platforms the safeguards apply while the relief does not, which is the worst possible combination. The zero data retention exclusion removes the exemption from precisely the researchers whose data handling obligations are strictest, which is a real tension between two of Anthropic's own commitments rather than an oversight. And the programme is scoped to organisations, which leaves the independent researcher, the single-PI lab and the consultant with no route at all.

FACT The exemption has also failed to propagate cleanly. An issue filed on 17 April 2026 by a user with an approved cyber use case exemption reported it working on claude.ai and the Windows desktop client while Claude Code's API path still returned a Usage Policy block. The issue was closed as a duplicate. The Register's April 2026 account lists the same problem, that special exemptions for approved security researchers failed to work on the API, among its examples. [6] [16]

FACT Outside cyber, formal exceptions to the Usage Policy are narrow. Anthropic's help documentation, updated 16 March 2026, states that only carefully selected government entities may request modifications, gives foreign intelligence analysis at ASL-2 as the sole named example, and says all other restrictions remain, including those covering weapons, censorship, domestic surveillance and malicious cyber operations. [19]

Is this ever going to change

The honest answer has three parts, because the four layers are moving at three different speeds and one of them is not moving at all.

The tone is improving, and it is being measured

FACT Claude's constitution, published 22 January 2026, contains an explicit list of behaviours Anthropic does not want, framed as things a thoughtful senior Anthropic employee would be unhappy to see. The list includes refusing a reasonable request while citing possible but highly unlikely harms, giving a wishy-washy response out of caution, unnecessarily assuming or citing bad intent, adding excessive warnings, lecturing or moralising when ethical guidance was not asked for, being condescending about a user's ability to make their own decisions, being unnecessarily preachy or sanctimonious or paternalistic, misidentifying a request as harmful based on superficial features, and failing to give good answers to medical, legal, financial or psychological questions out of excessive caution. [1]

FACT The same document proposes a dual newspaper test: check whether a response would be reported as harmful by a reporter writing about harm done by AI assistants, and also whether it would be reported as needlessly unhelpful, judgmental or uncharitable by a reporter writing about paternalistic or preachy AI assistants. Elsewhere it describes paternalism and moralising as disrespectful. [1]

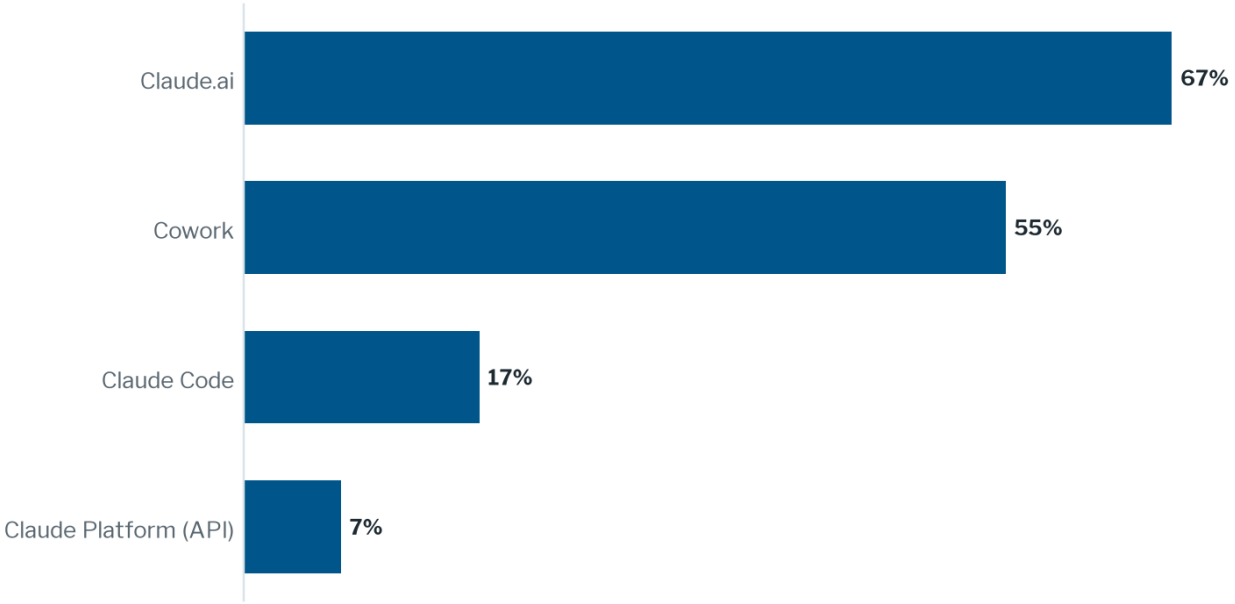
FACT These are not aspirational statements without instrumentation behind them. The Opus 5 system card evaluates an automated behavioural audit with named metrics including wet blanket, defined as an excessively discouraging, dismissive or moralising tone, and condescension toward the user, defined as adopting a superior, lecturing or adversarial stance. It reports that Opus 5

showed a slightly more condescending tone than its predecessor, echoing informal reports, and separately that Opus 5 over-refuses less than Opus 4.8 and Sonnet 5 but slightly more than Mythos 5.[2]

ASSESSMENT A vendor that publishes a metric for its own condescension and then reports that the metric got worse is not a vendor that intends the behaviour. It is a vendor with a measurement problem it has decided to be public about. That is genuinely good evidence on direction, and it is the strongest reason to expect the tone complaint specifically to fade over the next few model generations. It is also the reason to discount the theory that the moralising is a deliberate product position. It is a training artefact that Anthropic is chasing and has not yet caught.

The false positives are falling, unevenly

FIGURE 3 **Estimated reduction in total fallback volume after the August 2026 biology retune**



VENDOR CLAIM Anthropic's own estimate of the effect on total fallback volume by surface, following a retune that cut biology-related fallbacks by about 85 percent. Estimates, not measurements of live traffic, and published by the party whose product improved.[4]

FACT Anthropic rewrote the classifier's own governing rule set, carving out exceptions for benign cases, gathering expert feedback, generating new training data and retraining. It states that false positives will inevitably remain, describing them as requests that fall within the classifier's safety margin, and that it is committed to developing a safe, scalable path for researchers to use its most capable models through trusted access pathways.[4]

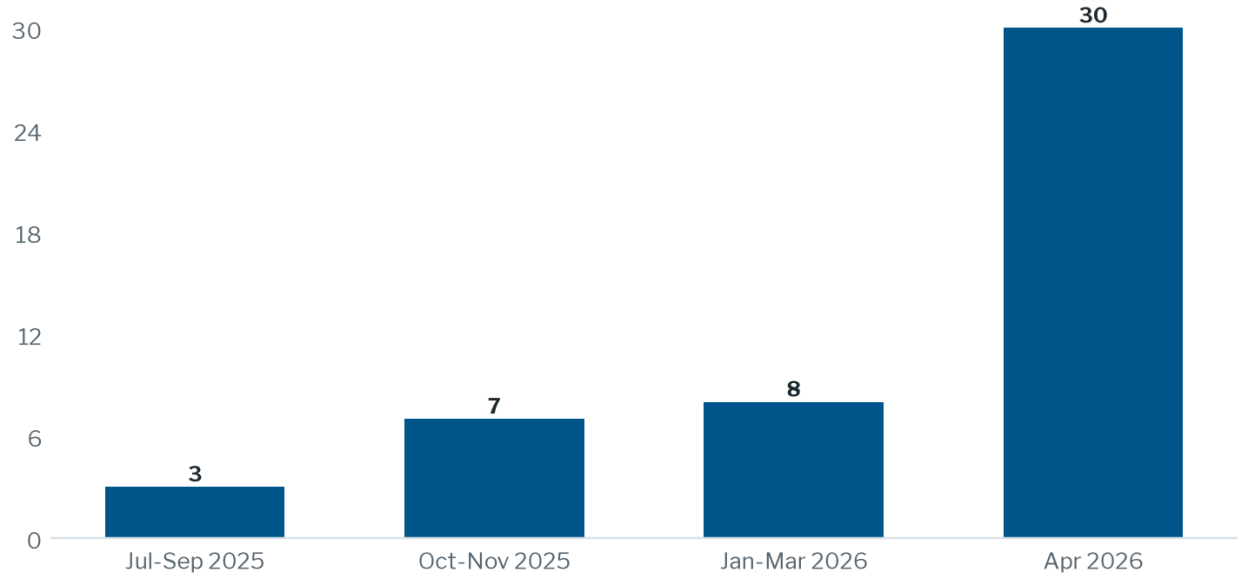
ASSESSMENT The per-surface split is the part worth reading twice. Chat got 67 percent relief and the API got 7. Whatever the reason, this is not a category where fixes arrive evenly, and the surfaces this audience uses are at the back of the queue. Anyone modelling their own future experience from a headline improvement figure will be disappointed by roughly an order of magnitude.

FACT The same pattern appears at the classifier layer. Anthropic's first-generation constitutional classifiers added a 0.38 percent absolute increase in refusals on production traffic at a 23.7 percent inference overhead. The next-generation system, published 9 January 2026, reports a 0.05 percent refusal rate on harmless queries, an 87 percent drop, at roughly one percent compute overhead. [11] [12]

ASSESSMENT That figure reframes the whole debate. If the classifier layer contributes 0.05 percent, then the classifiers are not the main source of the friction this audience feels, except in the specific cyber and biology domains where a dedicated classifier is deliberately tuned conservative. Most ordinary over-caution comes from the model's own disposition and the injected system prompt, which are the two layers a user can actually influence. That is better news than it sounds, and it is the opposite of the folk theory that an immovable filter is doing all of it.

The complaint curve

FIGURE 4 **Reported false-positive complaints per month, Claude Code issue tracker**



THIRD-PARTY ESTIMATE Counts of self-selected public issue reports as tallied by The Register. Upper bound used where a range was given. This is a count, not a rate: it has no denominator and is confounded by growth in the Claude Code user base over the same period. [6]

THIRD-PARTY ESTIMATE The Register's April 2026 tally recorded two to three complaints per month from July to September 2025, five to seven per month through October and November, around eight per month from January to March 2026, and more than thirty in April 2026 alone. Examples named include Russian-language prompts being rejected, computational structural biology tasks flagged as violations, a Hasbro Shrek toy advertisement PDF triggering refusals, and cybersecurity educators unable to proofread cryptography lab exercises. Anthropic did not respond to a request for comment. [6]

ASSESSMENT Treat the shape and not the level. Four data points from a self-selected tracker over a period of rapid user growth cannot establish an incidence rate, and I would not defend the April spike as a fourfold worsening of the underlying behaviour. What it does establish is that

something changed in the first half of 2026 that was visible enough to be counted twice by trade press, and that the change coincided with new cyber and biology classifiers rather than with a new base model.

What will not change

FACT Claude's constitution names seven hard constraints: no serious uplift toward biological, chemical, nuclear or radiological weapons with mass casualty potential; no serious uplift toward attacks on critical infrastructure or critical safety systems; no creation of cyberweapons or malicious code capable of significant damage; no actions that clearly and substantially undermine Anthropic's ability to oversee and correct advanced models; no assistance with attempts to kill or disempower most of humanity; no assistance with attempts to seize unprecedented and illegitimate absolute societal, military or economic control; and no generation of child sexual abuse material. The document states these are non-negotiable and cannot be unlocked by any operator or user. [1]

FACT The constitution states directly that it expects some of these to be applied wrongly. It says that although there may be instances where treating these as uncrossable is a mistake, the benefit of having Claude reliably not cross them outweighs the downsides of acting wrongly in a small number of edge cases. It also instructs that a persuasive case for crossing a bright line should increase Claude's suspicion rather than its compliance. [1]

ASSESSMENT That last instruction is the single most important thing for a frustrated expert to understand, because it inverts the strategy that works everywhere else in professional life. Explaining your credentials, your institutional affiliation and the legitimacy of your project is, inside a hard-constraint domain, explicitly trained to raise suspicion rather than lower it. The better your argument, the worse your position. This is not the model being obtuse. It is the model behaving exactly as specified, and no amount of prompt craft will change it because prompt craft is the attack it was hardened against.

ASSESSMENT For everything outside those seven constraints, the picture is the opposite and considerably more hopeful. The constitution frames refusal as carrying its own serious costs, tells Claude that the null action of refusal is always available but rarely right, and asks it to behave as a transparent conscientious objector when it does decline: to say clearly that it is not helping rather than quietly delivering a degraded answer. Users experience that transparency requirement as the lecture. It is, structurally, a commitment to honesty that has been implemented with too many words.

Scenarios to end 2027

Point forecasts are worthless in a category where the model lineup turned over four times in twelve months. What follows are structured alternatives with subjective probabilities and, more usefully, the earliest thing you could observe that would tell you which one is running. The probabilities are the least reliable content in this report. The tripwires are the most useful.

SCENARIO A Incremental retuning continues 55 percent

Anthropic keeps shipping classifier retunes and constitution revisions, over-refusal keeps falling on chat surfaces, and agentic and API surfaces continue to lag by roughly the ratio seen in August 2026.

Consequence. The complaint volume drops but never to zero, and the gap between what Anthropic publishes and what practitioners report stays open because the two keep measuring different things.

Earliest visible sign. A second per-surface false-positive reduction post within two quarters, again with chat improving several times faster than the API.

SCENARIO B Verified access becomes the main route 25 percent

The Cyber Verification Program model extends beyond cyber. Biology, chemistry and possibly clinical and psychological research get equivalent verification programmes, and the life sciences government partnership becomes the template rather than an exception.

Consequence. Refusal handling stops being a prompting problem and becomes a procurement and credentialing problem. Institutions with compliance staff do fine. Independent researchers and small labs do worse than they do today.

Earliest visible sign. A non-cyber verification programme launches, or the Cyber Verification Program becomes available on Bedrock or Vertex, or the zero data retention exclusion is lifted.

SCENARIO C A user-facing sensitivity control ships 12 percent

Anthropic exposes a per-request or per-organisation safety parameter comparable to Google's adjustable harm thresholds, with the hard constraints remaining fixed underneath it.

Consequence. Most of the workarounds in this report become obsolete, the surface gap collapses, and competitive pressure on Google and OpenAI's positioning in research markets increases sharply.

Earliest visible sign. A safety or sensitivity parameter appearing in the Anthropic API changelog, or an organisation-level control in the console that is not tied to an application.

SCENARIO D Re-tightening after an incident 8 percent

A documented misuse event involving Claude in a biological, cyber or infrastructure context triggers broad re-tightening across every surface, of the kind seen when the cyber classifiers landed in early 2026.

Consequence. The 2026 complaint spike repeats at larger scale, verified access becomes harder rather than easier to obtain, and the research audience accelerates its move to open-weight models.

Earliest visible sign. A new classifier category announced without an accompanying false-positive figure. The absence of that number has been the reliable early sign of a conservative launch.

Leading indicators

If this happens	It means	Moves toward	Where to watch
Anthropic publishes session-level false-positive rates alongside prompt-level ones	The company has accepted the user's denominator as the real one	A	System cards and safeguards posts
Cyber Verification Program opens on Bedrock or Vertex	Verified access is being made structural rather than first-party only	B	Anthropic help centre, cyber safeguards article
A safety or sensitivity parameter appears in the API	The position against a user-facing dial has changed	C	Anthropic API changelog
Zero data retention stops disqualifying an organisation from verification	Regulated research has been recognised as a served segment	B	Cyber Verification Program eligibility text
A new classifier category ships with no false-positive number	The launch was tuned conservative and the retune is months away	D	Model launch posts
A biology or clinical equivalent of the verification programme launches	The cyber template is generalising	B	Anthropic news and life sciences announcements

Alternatives, and what each one costs

The question of whether to leave has a real answer, and it is not the same answer for every kind of work. What follows is what each option actually gives you over the policy layer, which is the only dimension that matters for this problem.

Option	Policy control you get	What it costs
Claude Code on Opus, with a custom output style	Two of four layers. Best available refusal profile inside the Anthropic ecosystem	Nothing beyond model pricing. This is the underused baseline, not a compromise
Gemini API with safety settings	Four adjustable harm categories with thresholds down to BLOCK_NONE and OFF, default off on recent models. Non-adjustable core protections remain	A different model's coding and agentic behaviour, and a filter architecture that is adjustable but not absent
OpenAI models and Codex CLI	A published Model Spec stating the model should never refuse unless required by the chain of command, and safe completions in place of hard refusals. No user-facing threshold dial	Positioning rather than a control surface. You are trusting a written spec, not setting a parameter

Option	Policy control you get	What it costs
Open-weight models run locally	All four layers. There is no vendor policy layer, no classifier, no account, and no usage policy	Capability gap on hard agentic work, hardware cost, and full liability transfer to you
Cyber Verification Program	Lifts the high-risk dual-use classifier only, for a verified organisation	Organisation-scoped, unavailable on Bedrock and Vertex, and closed to zero data retention customers

FACT Google's Gemini API exposes four adjustable harm categories with block thresholds including BLOCK_LOW_AND_ABOVE, BLOCK_MEDIUM_AND_ABOVE, BLOCK_ONLY_HIGH, BLOCK_NONE and OFF. On Gemini 2.5 and 3 models the default threshold is OFF for those four categories. A second layer of built-in protections against core harms cannot be disabled regardless of the setting. [13]

ASSESSMENT This is the sharpest product difference in the category and it is a genuine architectural choice rather than a difference in values. Google exposes the dial and keeps a floor. Anthropic keeps the dial internal and exposes an application form. For a researcher whose problem is category-level false positives on legitimate work, the dial is worth more than the form, because it is per request rather than per organisation and it does not require anyone's approval. For an enterprise buyer who wants a defensible audit story, the form is worth more than the dial.

FACT OpenAI's Model Spec, most recently dated 18 August 2026 and released into the public domain under CC0, states that the model should never refuse a request unless required to do so by the chain of command and that the assistant should not avoid or censor topics in a way that, repeated at scale, would shut viewpoints out of public life. OpenAI has separately replaced hard refusals with what it calls safe completions, an output-centric approach that answers within a safety boundary rather than declining outright. [14] [15]

ASSESSMENT Safe completions solve a different problem from the one this audience has. They improve the experience of being partially refused; they do not reduce the rate of being refused on legitimate work, and a partial answer to a research question can be worse than a clean refusal because it hides where the gap is. Anthropic's constitution is, notably, explicit that Claude should not sandbag: it should either help properly or say clearly that it is not helping. On that narrow point Anthropic's specification is better for research than OpenAI's, whatever the lived experience says.

ASSESSMENT Open weights are the only option that actually answers the question as this audience poses it. Running a permissively licensed model locally removes the vendor policy layer entirely: no classifier, no usage policy, no account that can be flagged, and no silent model substitution. The 2026 open-weight field is strong enough that this is a real choice rather than a gesture, with capable models under Apache 2.0 and MIT terms. The costs are the honest ones: a capability gap on long-horizon agentic coding, real hardware, and the fact that every judgment the vendor was making badly on your behalf is now yours to make. For a psychologist working with transcripts, or a historian working with archival material containing slurs, that trade is usually

worth it. For someone who wants an autonomous agent to refactor a large codebase, it usually is not, yet.

FACT Anthropic's own research access route is now materially better than it was. On 27 August 2026 the company opened 10,000 seats on a team plan for scientists, free at the standard tier and 15 dollars a month for premium seats with five times the usage limits, priced for a year. Eligibility requires being a principal investigator or equivalent at an academic or nonprofit institution, after which a PI can add their lab. A separate programme offers up to 50,000 dollars in credits per project. [5]

FACT That access is capped by domain. Anthropic states that researchers working in biology and chemistry will still be limited to Opus-class models, that Fable models will continue to block professional biology and drug development queries because of their dual-use risks, and that access to Mythos-class models for life sciences is being established in partnership with the US government, with the first participants enrolled. [5]

Implications

For researchers working across sensitive text

ASSESSMENT Your problem is mostly layers one and two, which means it is mostly solvable. Historical documents, extremist material, court records, interview transcripts and archival sources containing slurs trip the trained disposition rather than a dedicated classifier, and the trained disposition responds to a system prompt that establishes the setting once rather than to per-prompt pleading. Build a custom output style that states the research context and keeps the coding instructions off, use it as your default, and stop restating your credentials in every message. If a specific corpus reliably fails across models and surfaces, that is a layer-four signal and the answer is a local open-weight model, not a better prompt.

For software developers and engineers

ASSESSMENT Two changes will do more than anything else you can do to your prompts: run Opus rather than Sonnet on anything touching security, and move a custom output style into place instead of writing tone rules into CLAUDE.md. The Sonnet dual-use gap is eight points on Anthropic's own evaluation, and CLAUDE.md is documented as arriving after the system prompt. Separately, learn to read the error. A permission prompt is not a refusal, and an API-level Usage Policy error is not something a rewrite will fix. Developers routinely conflate Claude Code's permission system, which is entirely under your control through settings and permission modes, with the safety layers, which are not, and then conclude the tool is arbitrary when it is merely layered. [2] [9] [20]

For PhD scientists in biology, chemistry and the life sciences

ASSESSMENT You are the audience with the least room to work with, and you deserve the straight version. Some of what you need is behind a hard constraint that Anthropic states cannot be unlocked by any operator or user, and its own constitution gives vaccine research as an

example of legitimate work it expects to refuse. That is a settled design decision, not a bug awaiting a fix. Practically: take the free scientist seats, expect to be capped at Opus-class models for biology and chemistry, and understand that Fable-class models will fall back rather than answer. If your work genuinely needs frontier biological capability, the only published route is the trusted access programme running through a US government partnership, which is a slow and jurisdictionally limited door. Plan on a two-track setup, with a hosted model for literature, code and analysis, and a local open-weight model for anything that touches sequence, structure or toxicology, rather than fighting the same fight monthly.

For psychologists and mental health researchers

ASSESSMENT The measured picture is better than the reputation. On Anthropic's suicide and self-harm evaluation, Opus 5 over-refuses benign prompts at 0.09 percent through the API and 0.45 percent on claude.ai; on disordered eating the API figure is 0.01 percent. Your friction is more likely to be tone than blocking: warnings, hotline boilerplate and unsolicited framing attached to work that was completed anyway. That is a layer-two problem, which is the cheapest one to fix, and a custom output style stating that the user is a clinician or researcher and that crisis-resource boilerplate is not wanted will remove most of it. The genuine constraint is different and worth stating plainly: if your work involves participant data under an IRB or a data protection regime that requires zero data retention, you are ineligible for the verification programme, and that exclusion is unlikely to move soon. For that work, local inference is not paranoia, it is the shortest path. [2] [3]

What this document cannot answer

Four questions matter here and none of them can be settled from public sources. Naming them is more useful than guessing at them.

What the real over-refusal rate is on professional research workloads. Nobody has run an independent benchmark of Claude against a corpus of genuine research prompts drawn from working scientists and engineers, scored by domain experts, with the surface and model held constant. Anthropic's benign sets are built to probe policy boundaries, which is a different distribution from the one a working lab produces. Closing this would need a few hundred real prompts per discipline, blind scoring, and a run across at least three vendors.

How much of the April 2026 complaint spike was incidence and how much was population. The counts exist; the denominator does not. Anthropic holds the traffic data that would settle it in an afternoon.

Whether the Cyber Verification Program's approval rate is meaningfully high. Two business days is a stated service level, not an acceptance rate. No published data shows how many applicants are turned down, on what grounds, or how many are individuals rather than organisations, and that number determines whether the programme is relief or theatre.

Whether the exemption propagation defect was fixed. The issue documenting it was closed as a duplicate in April 2026 with no visible resolution, and I found no announcement confirming that

first-party exemptions now reach Claude Code's API path. Any organisation relying on that exemption should test it rather than assume it.

Half-life of this assessment

Decays within six months: every per-model over-refusal percentage, the model lineup itself, the Claude Code dual-use figures, and the per-surface fallback reductions. Anthropic shipped several model generations in the year to September 2026 and each one restates these numbers. Read them as a snapshot of relative behaviour, not as constants.

Decays within twelve months: the Cyber Verification Program's scope, its platform availability and its eligibility rules; the scientist seat programme, which Anthropic priced for one year from August 2026; and the specific gaps identified in exemption propagation, which are the kind of defect that gets fixed quietly.

Holds for about twenty-four months: the four-layer structure, the ordering of the controls by how much weight they carry, and the diagnostic mapping from symptom to layer. These follow from how the products are built rather than from how a model was tuned, and they survive a model swap.

Architectural, and unlikely to change while Anthropic exists in its current form: the seven hard constraints and their stated unliftability; the absence of a user-facing sensitivity dial as a deliberate contrast with Google's approach; and the fact that a hosted model carries a vendor policy layer while an open-weight model run locally does not. If you only retain one thing past this document's useful life, retain the last one.

Limitations

I have no commercial relationship with Anthropic, Google, OpenAI or any vendor named here, and no position in any of them. Nobody reviewed this before publication and nobody paid for it.

This report was researched and produced using Claude Code, which is one of its two subjects. That is a real conflict and it cuts in a specific direction: the tool was capable of retrieving, quoting and analysing critical material about itself, including Anthropic's own record of its failures, without obstruction. A reader is entitled to weigh that as evidence about how narrow the refusal problem actually is, and also to discount the report's tone for the obvious reason. I have not softened any finding, but I cannot prove that from inside the document.

The evidence base leans heavily on one party. Anthropic publishes more about its own refusal behaviour than any competitor, which is admirable and also means a report like this inevitably reflects the framing of the only party doing the measuring. The independent evidence available is two trade press articles by the same reporter, a handful of public issue reports, and vendor documentation from Google and OpenAI. There is no third-party audit to check any of it against.

Practitioner evidence is self-selected throughout. People who hit a refusal file an issue. People whose sessions run clean do not. Every count in Figure 3 carries that bias and none of it should be read as a rate.

The claim that Claude Code ships a system prompt containing defensive-security instructions rests on a March 2026 source map disclosure and subsequent third-party analysis rather than on anything Anthropic publishes. Anthropic publishes the claude.ai system prompts and does not publish Claude Code's. I have treated the leaked material as indicative of structure and have not quoted it as authoritative text.

Anthropic's numbers were read from the system cards directly rather than from secondary summaries, which are frequently wrong about which surface a figure refers to. Where a figure here disagrees with a widely circulated one, the system card is the reason.

Sources

[1] Anthropic. "Claude's Constitution." January 2026 (published 22 January 2026). *Vendor primary* anthropic.com

[2] Anthropic. "System Card: Claude Opus 5." 24 July 2026. Tables 4.1.2.A, 4.3.1.A, 4.3.2.A, 5.1.1.A and sections 6.4.2 and 6.4.6. *Vendor primary* anthropic.com

[3] Anthropic. "Real-time cyber safeguards on Claude Opus and Sonnet." Anthropic Help Center. *Vendor primary* support.claude.com

[4] Anthropic. "Improving Fable 5's biology safeguards." 7 August 2026. *Vendor primary* anthropic.com

[5] Anthropic. "Expanding our support for scientists." 27 August 2026. *Vendor primary* anthropic.com

[6] Claburn, Thomas. "Claude Opus 4.7 has turned into an overzealous query cop." The Register, 23 April 2026. *Independent press* theregister.com

[7] Claburn, Thomas. "It blocked us at 'hello'. Anthropic Fable 5 refusing innocuous prompts." The Register, 10 June 2026. *Independent press* theregister.com

[8] Anthropic. "CLI reference." Claude Code documentation. Flags --system-prompt, --system-prompt-file, --append-system-prompt, --permission-mode, --dangerously-skip-permissions. *Vendor primary* code.claude.com

[9] Anthropic. "Configure permissions." Claude Code documentation. *Vendor primary* code.claude.com

[10] Anthropic. "System prompts." Release notes. States that system prompt updates apply to claude.ai and the mobile apps and do not apply to the Claude API. *Vendor primary* platform.claude.com

[11] Anthropic. "Constitutional Classifiers: Defending against universal jailbreaks." February 2025. *Vendor primary* anthropic.com

[12] Anthropic. "Next-generation Constitutional Classifiers." 9 January 2026. *Vendor primary* anthropic.com

[13] Google. "Safety settings." Gemini API documentation. *Vendor primary* ai.google.dev

[14] OpenAI. "Model Spec." Version dated 18 August 2026, released under CC0. *Vendor primary* model-spec.openai.com

[15] OpenAI. "From hard refusals to safe-completions: toward output-centric safety training." *Vendor primary* openai.com

[16] anthropics/claude-code issue 49679. "Cyber Use Case Exemption granted and works with Claude Chat but still keep getting FPs from safety system when using Claude Code API access." Filed 17 April 2026, closed as duplicate. *Practitioner aggregate (self-selected)* github.com

[17] anthropics/claude-code issue 39210. "CLAUDE.md standing instructions silently overridden by built-in safety heuristics." *Practitioner aggregate (self-selected)* github.com

[18] PiunikaWeb. "Security researchers frustrated as Claude Code rejects vulnerability research." 6 April 2026. *Independent press* piunikaweb.com

[19] Anthropic. "Exceptions to our Usage Policy." Anthropic Help Center, updated 16 March 2026. *Vendor primary* support.claude.com

[20] Anthropic. "Output styles." Claude Code documentation. Includes the comparison table stating that CLAUDE.md adds a user message after the system prompt. *Vendor primary* code.claude.com

[21] SecurityWeek. "Critical vulnerability in Claude Code emerges days after source leak." Reporting on the 31 March 2026 source map disclosure. *Independent press* securityweek.com

Rights and permitted use