Claude Fable 5.1 and Mythos 5.1: Biology Access Explained

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · claude fable 5.1 · claude mythos 5.1 · anthropic biology access · claude model routing · life sciences ai · responsible scaling policy · asl-3 · research reproducibility · claude api fallback

Executive Summary

Anthropic introduced **Claude Fable 5.1** and **Claude Mythos 5.1** on September 1, 2026, describing them as "the same model, but with different levels of safeguards" (www.anthropic.com) (www.anthropic.com). Claude Fable 5.1 is generally available on the Claude API, Amazon Bedrock, and Google Cloud; its Microsoft Foundry offering is a public preview (techcommunity.microsoft.com). List pricing is unchanged at $10 per million input tokens and $50 per million output tokens (platform.claude.com), while Claude Mythos 5.1 remains restricted to "a small set of vetted organizations" (www.anthropic.com). This report explains what "biology access" means for these models: a safety-classifier system, deployed under Anthropic's Responsible Scaling Policy, that can decline a biology-related request on Fable 5.1 and automatically reroute it to a less biologically capable model, currently Opus 5, "a capable model that does not have the same level of biological capability as Fable 5" (www.anthropic.com).

Anthropic activated its AI Safety Level 3 (ASL-3) Deployment and Security Standards in May 2025 alongside Claude Opus 4 (www.anthropic.com), and Anthropic applies "the same expanded safeguards that restrict access to dual-use research biology capabilities" to Mythos 5.1 that it previously applied to Mythos 5 (www-cdn.anthropic.com), and offers an invite-only **Life Sciences Verification Program** for vetted researchers (www-cdn.anthropic.com).

An August 2026 classifier update cut biology-related fallbacks by roughly 85 percent, and independent reporting confirms that "medical or biology questions will have 85% fewer interventions" (www.axios.com), yet Fable 5.1 still "falls back to Opus 5 for requests we consider dual-use" (www.anthropic.com), a category that includes virology, toxicology, and molecular design. Anthropic's own documentation shows this behavior is implemented as a formal API mechanism: a declined request returns a standard HTTP 200 response with `stop_reason: "refusal"` rather than an error, and developers can opt into automatic retries on a recommended fallback model (platform.claude.com). This routing behavior, together with Anthropic's removal of the temperature and top-p sampling parameters on Claude 4.7-and-later models (platform.claude.com), carries direct consequences for reproducible research: a 2026 preprint on LLM use in empirical research states plainly that "exact reproduction is therefore generally not possible when using proprietary application programming interfaces" (arxiv.org).

On documented benchmarks, Anthropic's performance table reports 52.6 percent for Claude Fable 5.1 on Terminal-Bench-Science 0.1, compared with 24.7 percent for Fable 5. Real-world life-sciences adoption already includes Benchling, which integrated Claude via the Model Context Protocol to let scientists query lab data directly (www.benchling.com); Stanford's Biomni project, whose own preprint describes an agent that mines "tools, databases, and protocols from tens of thousands of publications across 25 biomedical domains" (www.biorxiv.org); and FutureHouse, a nonprofit "building AI agents to automate research in biology and other complex sciences" (www.futurehouse.org). For research teams, the practical takeaway is that biology access on Fable 5.1 is a governed, disclosed, and appealable routing decision rather than a simple on/off switch, and reproducibility depends on pinning exact model identifiers, logging which model actually answered each call, and treating single-run LLM outputs, even at zero temperature, as provisional rather than final.

Introduction and Background

Anthropic, the San Francisco-based artificial intelligence public benefit corporation founded in 2021 (en.wikipedia.org), released **Claude Fable 5.1** and **Claude Mythos 5.1** on September 1, 2026, describing them as models whose "research capabilities offer an early glimpse of how AI models will contribute to scientific progress" (www.anthropic.com). For life-sciences and pharmaceutical organizations evaluating whether and how to use these models for research, the central practical question is not whether the models are capable, but which capabilities are actually reachable, under what conditions, and whether the same prompt reliably produces the same class of answer across repeated runs.

That question has a specific answer in this release. Fable 5.1 and Mythos 5.1 are, per Anthropic, "the same underlying model as Claude Mythos 5.1 with safeguards for cybersecurity and biology" (www.anthropic.com): Fable 5.1 ships broadly with biology and cybersecurity guardrails active, while Mythos 5.1, offered "by invitation only, as part of Project Glasswing" (platform.claude.com), runs with those guardrails relaxed for a vetted population.

This article is an educational reference, current as of September 5, 2026, on how that access model works: the taxonomy of Fable 5.1 versus Mythos 5.1, the mechanics of the biology-related fallback and refusal system, how it compares to the analogous cybersecurity fallback, and what the underlying routing and versioning behavior means for researchers who need reproducible results from repeated queries. Prices, benchmark scores, and availability details change quickly for frontier models; every figure below is dated to its source and should be re-verified before being relied upon for a procurement or compliance decision. As an adjacent life-sciences AI advisory practice, IntuitionLabs approaches this kind of frontier-model access change primarily as a governance question: which teams get access to which capability tier, how that access is logged, and how outputs are evaluated before being trusted in a regulated workflow (intuitionlabs.ai).

Understanding the Claude Fable 5.1 and Mythos 5.1 Model Family

Claude Fable 5.1 carries the model identifier `claude-fable-5-1`, offers a 1-million-token context window, a maximum output of 128,000 tokens, and was released on September 1, 2026 (platform.claude.com). Anthropic's own model-selection guidance describes it as "Anthropic's most capable widely released model" (platform.claude.com), positioned as an escalation path above Claude Opus 5: developers are told to "move to Claude Fable 5.1" only "if your evals at xhigh or max effort still fall short on demanding reasoning or long-horizon agentic work" (platform.claude.com). List pricing is unchanged from Fable 5 at $10 per million input tokens and $50 per million output tokens, with prompt-cache reads cut to a quarter of the prior cost (platform.claude.com), which independent reporting quantifies as a drop "to just $0.25 on input, down from $1.00 for Fable 5" (venturebeat.com).

Fable 5.1’s Microsoft Foundry offering is a public preview (techcommunity.microsoft.com). Amazon's own documentation confirms it is "available today on Amazon Bedrock through the US Geo CRIS (us.) and Global CRIS (global.) inference profiles" (aws.amazon.com), and separately notes that Anthropic "has designated Fable 5.1 a Covered Model," triggering additional review under AWS's frontier-safety commitments (aws.amazon.com). Microsoft's Azure AI Foundry blog independently corroborates the safeguard structure, describing "Claude Fable 5.1" as "a public preview version with safeguards that fallback for cyber and bio capability queries," while Mythos 5.1 retains "full cyber and bio capabilities" (techcommunity.microsoft.com). Google Cloud's Model Garden page frames the same model as "a frontier model built for the long-running, high-stakes work that enterprises, developers, and power users run" (cloud.google.com).

Claude Mythos 5.1 shares Fable 5.1's technical specification and model card but a different, invitation-only distribution model. Anthropic states plainly that "access remains limited to a small set of vetted organizations" (www.anthropic.com), a description Amazon's own "What's New" announcement echoes, confirming Mythos 5.1 "is available with limited access" on AWS at general availability (aws.amazon.com). Independent business coverage from Axios frames the split concisely: Fable 5.1 became broadly available while "the more restricted Claude Mythos 5.1 remains limited to vetted partners working in areas like cybersecurity and life sciences" (www.axios.com). Both models are also part of a broader industry safety infrastructure effort: Amazon, Google, and Microsoft jointly developed Anthropic's Enterprise Frontier Safeguards program "with more than 100 organizations across financial services, healthcare, manufacturing, telecom, law, retail and government," per VentureBeat's independent reporting (venturebeat.com), which on AWS includes zero data retention for Covered Models "available for internal use through December 31, 2026" (aws.amazon.com).

How the Biology Access and Safeguard System Works

The reason "biology access" is a distinct, named concept for Fable 5.1 rather than an ordinary feature flag traces back to Anthropic's Responsible Scaling Policy (RSP), which defines AI Safety Level (ASL) standards that scale with a model's assessed capability. The policy groups these into "two categories: Deployment Standards and Security Standards" (www-cdn.anthropic.com), and Anthropic activated ASL-3 for the first time in May 2025 alongside Claude Opus 4, stating: "we have activated the AI Safety Level 3 (ASL-3) Deployment and Security Standards" (www.anthropic.com). Anthropic's stated intent is narrow: the ASL-3 Deployment Standard "should not lead Claude to refuse queries except on a very narrow set of topics" (www.anthropic.com), implemented so that Anthropic can "evaluate whether the measures we have implemented make us robust to persistent attempts to misuse" the model (www-cdn.anthropic.com) using Constitutional Classifiers, alongside iterative refinement.

The system card states Anthropic has "opted to apply the same expanded safeguards that restrict access to dual-use research biology capabilities" that were used for the earlier Mythos 5 (www-cdn.anthropic.com), with the vetted path formalized as the "Life Sciences Verification Program" (www-cdn.anthropic.com), an "invite-only beta offering access with reduced biology safeguards for advanced life sciences researchers" (www.anthropic.com).

In everyday use, Anthropic explains the mechanism directly: when a biology classifier fires on Fable 5.1, "the model re-routes the user's request to Opus 5, a capable model that does not have the same level of biological capability as Fable 5" (www.anthropic.com). Anthropic frames the conservative default as deliberate, and even after tuning, "Fable still falls back to Opus 5 for requests we consider dual-use" (www.anthropic.com), a category that includes virology, toxicology, and molecular design. Anthropic's help center is more explicit still about the practical implication for scientific users: Fable 5 and 5.1 are "not recommended for professional biology research and drug development at this time" for exactly this reason (support.claude.com).

As independent validation of these classifier judgments, the biosecurity organization SecureBio reviewed an unredacted version of Anthropic's chemical and biological risk assessment, reporting that Anthropic "shared with us an unredacted version of this report and 110 pages of additional materials" (securebio.substack.com), and found the safeguards correctly refused "94.2% of hazardous prompts in SecureBio's safeguard benchmark BioTIER-refuse" (securebio.substack.com).

Model Routing, Fallback Mechanics, and the Cybersecurity Parallel

Anthropic implements the biology fallback as a formal, documented API behavior rather than a silent redirect. When a request is declined, "a refusal is a successful HTTP 200 response with `stop_reason: "refusal"`," not an error (platform.claude.com), and refusals are grouped into named categories including one where "the request could enable biological harm, such as dangerous lab methods" (platform.claude.com). Developers who set the `fallbacks` parameter to `"default"` get automatic retries, so that "the API retries a declined request on the fallback model Anthropic recommends" (platform.claude.com). Anthropic's Fable 5.1 migration guide confirms Mythos 5.1 is not exempt from this framework: it "runs safety classifiers covering the same `stop_details` categories as Claude Fable 5" (platform.claude.com), meaning even trusted-access users are subject to classifier review, just with a different threshold.

On Claude's consumer and enterprise surfaces, the same mechanism is disclosed rather than hidden. Anthropic's help center states that "you'll see a notice explaining that the model switched, and the response will be labeled with the model that answered" (support.claude.com), that after a switch "the model picker stays on Opus for the rest of the conversation" until manually reverted (support.claude.com), that fallback destinations are "currently Opus 5 for biology, chemistry, and life sciences requests, and Opus 4.8 for offensive cybersecurity technique requests" (support.claude.com), and that "a blocked request pauses the conversation instead of switching models" if a user has disabled auto-switching (support.claude.com). Notably, "a block can be triggered by content you didn't type," since classifiers scan connected files, search results, and memory, not only the user's own message (support.claude.com). Amazon's original Fable 5 coverage independently confirmed the mechanics: blocked prompts on cybersecurity, biology, chemistry, and health topics "fall back to receive a response from Opus 4.8 instead" (aws.amazon.com) Current support documentation identifies Opus 5 for biology, chemistry, and life-sciences requests and Opus 4.8 for offensive cybersecurity technique requests (support.claude.com).

It is worth distinguishing Anthropic's refusal-triggered fallback from "model routing" as the term is generally used across the AI industry. A general technical explainer from Red Hat describes conventional routing as a system where "requests should be routed to the best model for the specific task, making effective use of different, usually specialized, models" (developers.redhat.com), often to "direct simpler queries to smaller, lower-cost models using lower cost hardware" (developers.redhat.com). Anthropic's public documentation does not describe a comparable cost- or complexity-based routing layer that silently selects among Claude models for an ordinary query; the only documented cross-model routing behavior for Fable 5.1 is the safety-classifier fallback described above, triggered specifically by a refusal event rather than by task difficulty or load.

Implications for Reproducible Research Workflows

For a life-sciences team that wants to reuse the same Claude prompt across an experiment, a validation run, and a later audit, three documented behaviors matter more than raw benchmark scores. First, Anthropic has deprecated user-controlled sampling parameters on its newest models: the temperature and top-p settings now return "a 400 error when set to a non-default value on Claude 4.7 and later models and Claude Mythos Preview" (platform.claude.com), removing a lever researchers previously used to reduce output variance. A 2026 preprint whose authors include data editors from several economics journals documents that "providers such as OpenAI (with GPT-5.6) and Anthropic (with Sonnet 5, Opus 4.8 and Fable 5) have disabled the setting in favor of a reasoning intensity" control instead (arxiv.org).

Second, model identity itself is not a stable proxy for reproducibility. The same preprint states flatly that "exact reproduction is therefore generally not possible when using proprietary application programming interfaces" (arxiv.org).

Third, even without any provider-side change, LLM outputs are not guaranteed to be stable run to run. A separate 2026 preprint that executed 480 independent data-analysis runs across six models, four temperature settings, and two prompting strategies found that "even at temperature zero some estimates lead to different conclusions about the same research question" (arxiv.org), a finding consistent with a broader warning in Nature Reviews Psychology that reliance on "proprietary models poses risks to transparency and reproducibility" for the research that depends on them (www.nature.com). Anthropic's own deprecation policy offers one partial mitigation: a retiring model "is still

functional but no longer recommended," with Anthropic assigning a formal retirement date rather than removing access immediately (platform.claude.com), which at least gives research teams a defined window in which a pinned, dated model ID remains queryable for replication purposes, even after a newer version supersedes it.

Taken together, these three facts argue for a specific practice: research groups working with Fable 5.1 or Mythos 5.1 should record the exact model identifier and call timestamp for every generated result, log the top-level `model`, `usage.iterations` (including any `fallback_message` entry), and `fallback` content blocks, distinguish a terminal `stop_reason: "refusal"` from a response served by a fallback model, and run any single analytical claim more than once before treating an LLM-assisted finding as a measurement rather than a preliminary estimate, consistent with the peer-reviewed guidance above.

Data Analysis and Evidence

Table 1 below summarizes the documented specification and access differences between Claude Fable 5.1 and Claude Mythos 5.1 as of September 5, 2026.

Attribute	Claude Fable 5.1	Claude Mythos 5.1
Availability	Generally available on the Claude API and on Amazon Bedrock (aws.amazon.com)	Invitation-only via Project Glasswing; limited to vetted organizations (platform.claude.com) (www.axios.com)
Model ID	`claude-fable-5-1` (platform.claude.com)	`claude-mythos-5-1` (per platform model documentation)
Biology/cyber safeguards	Active; dual-use requests fall back to a less capable model (techcommunity.microsoft.com)	Reduced, for vetted users, under the Life Sciences Verification Program (www-cdn.anthropic.com)
Context window / max output	1M tokens / 128K tokens (platform.claude.com)	Same underlying model specification
List pricing	$10 / MTok input, $50 / MTok output, cache reads cut to a quarter of Fable 5's cost (platform.claude.com)	$10 / MTok input, $50 / MTok output; limited availability (Anthropic)

The table shows that Fable 5.1 and Mythos 5.1 are not two different products in the ordinary commercial sense: they share a specification and price basis, and differ almost entirely in the safety-classifier configuration applied to the same weights. This is a materially different structure from a typical "pro versus enterprise" tiering, and it means an organization cannot simply pay more to unlock biology access; it must qualify through a vetting process that Anthropic has described only at a high level.

Table 2 summarizes the fallback behavior itself.

Refusal category	Example trigger	Fallback model (Sep 2026)	Disclosed to user
Biology / bio	"Biological harm, such as dangerous lab methods" (platform.claude.com); virology, toxicology, molecular design	Opus 5 (support.claude.com)	Yes, via notice and model label (support.claude.com)

Refusal category	Example trigger	Fallback model (Sep 2026)	Disclosed to user
Offensive cybersecurity	Exploit generation, penetration-testing techniques	Opus 4.8 (aws.amazon.com)	Yes, disclosed the same way as the biology category above

This table clarifies that biology and offensive-cybersecurity requests are treated as structurally parallel risk categories within Anthropic's classifier system, sharing the same disclosure and fallback mechanics but routing to different named substitute models. Both categories can be disabled by the user in favor of a hard pause instead of a silent substitution, "a blocked request pauses the conversation instead of switching models" (support.claude.com), which is the safer default for a research team that wants to know explicitly when a query was blocked instead of receiving a response from a fallback model.

On measured impact, Anthropic's August 2026 classifier update reduced total, all-cause fallbacks unevenly by product surface: "roughly 67% on Claude.ai, 55% on Cowork, 17% on Claude Code, and 7% on the Claude Platform" (www.anthropic.com), while the narrower, biology-specific fallback rate dropped by approximately 85 percent, a figure Axios independently reported as "medical or biology questions will have 85% fewer interventions" (www.axios.com). At the original Fable 5 launch in June 2026, Anthropic said false-positive interventions occurred, "on average, in less than 5% of sessions" (www.anthropic.com), and the subsequent 85 percent reduction applies on top of that already-low baseline.

On benchmarked scientific capability, Anthropic's table reports 52.6 percent for Fable 5.1 on Terminal-Bench-Science 0.1, versus 24.7 percent for Fable 5 (www.anthropic.com), alongside his observation that Fable 5.1 "has five reasoning levels: low, medium, high, xhigh, max, and no option to turn off reasoning entirely" (simonwillison.net), which corroborates the sampling-control changes discussed above. On applied biological research tasks, Anthropic reports that Mythos 5.1 produced protein binders whose "binding affinities were 10 times higher than the best designs submitted to Adaptyv Bio's protein design competitions" (www.anthropic.com), a vendor-reported result that VentureBeat separately confirmed reporting on, noting "Mythos 5.1 designed experimentally validated protein binders" (venturebeat.com). Readers should treat this specific figure as a vendor-reported benchmark result rather than an independently replicated measurement, since no third-party laboratory re-scoring of the Adaptyv Bio comparison was identified during this research.

Case Studies and Real-World Examples

Benchling: Embedding Claude Directly in a Life-Sciences R&D Platform

Benchling, a life-sciences research and development software provider, announced a partnership with Anthropic that connects Claude to its platform through the Model Context Protocol, stating on its own blog that "scientists can now ask Claude questions about their Benchling data" (www.benchling.com), including "pulling unstructured context buried in lab notebooks" alongside structured records like assays and molecules (www.benchling.com). Anthropic's customer materials describe the integration as saving "scientists up to 2 weeks spent transforming complex data" (www.anthropic.com), and an Anthropic staff member is quoted on Benchling's own site describing the goal as "connecting Claude to the systems science already runs on" (www.benchling.com).

Biomni: A General-Purpose Biomedical Research Agent

Biomni, developed by a Stanford-affiliated team, is described in its own preprint as a general-purpose agent "mining essential tools, databases, and protocols from tens of thousands of publications across 25 biomedical domains" (www.biorxiv.org), built by "integrating large language model (LLM) reasoning with retrieval-augmented planning and code-based execution" rather than fixed workflow templates (www.biorxiv.org). In one case study documented in the paper, an autonomous multi-step gene-regulatory-network analysis had "the full run, completed in just over five hours" (www.biorxiv.org). Separately, Anthropic's own customer materials attribute a much larger reported speedup to Biomni on a different task, completing "wearable bioinformatics analysis in 35 minutes versus 3 weeks for human experts" (www.anthropic.com); because this specific figure was not independently located in Biomni's own published paper during this research, it should be treated as a vendor-reported claim rather than a peer-reviewed measurement.

FutureHouse: Purpose-Built Agents for the Scientific Process

FutureHouse, a nonprofit research organization, describes its own mission as "building AI agents to automate research in biology and other complex sciences" (www.futurehouse.org). Anthropic's customer profile states the organization built specialized agents for distinct stages of the scientific process, and quotes FutureHouse on why calibrated uncertainty mattered more than raw fluency for their use case: "when evidence is insufficient, it will clearly indicate this rather than hallucinating an answer" (www.anthropic.com). This preference for honest uncertainty over confident fabrication is directly relevant to the reproducibility discussion above: a system that flags its own evidentiary gaps is easier to audit than one that always produces a fluent, unhedged answer.

Implications and Future Directions

The biology-access model described in this report sits inside a wider, industry-level policy context rather than existing in isolation. Anthropic is one of sixteen or more signatories, alongside Amazon, Google, IBM, Meta, Microsoft, and OpenAI, to the UK-convened Frontier AI Safety Commitments, which cover "leading AI organisations" agreeing to "furtherance of safe and trustworthy AI" (www.gov.uk), a list that explicitly names "Amazon, Anthropic, Cohere, Google, G42, IBM, Inflection AI, Meta, Microsoft, Mistral AI, Naver, OpenAI, Samsung Electronics, Technology Innovation Institute, xAI, Zhipu.ai" (www.gov.uk). Under those commitments, signatories agree that "in the extreme, organisations commit not to develop or deploy a model or system at all, if mitigations cannot be applied to keep risks below the thresholds" (www.gov.uk), a framework consistent with Anthropic's own RSP structure discussed earlier.

Anthropic also participates in cross-industry technical evaluation work relevant to biology access. Its own research page describes the company as "co-sponsoring a larger study through the Frontier Model Forum (FMF)" on real-world laboratory uplift (www.anthropic.com), and the Forum's own page states its AI-Bio workstream exists because AI "risks amplifying existing biological threats and introducing novel ones" (www.frontiermodelforum.org), while noting that "publications related to AI and biological threat creation can introduce significant information and attention hazards" (www.frontiermodelforum.org), which explains why detailed eligibility criteria for programs like the Life Sciences Verification Program are not fully public. This caution mirrors long-standing practice in commercial gene synthesis, where the International Gene Synthesis Consortium, "formed in 2009," has members that "screen

the complete DNA and translated amino acid sequences of every double-stranded gene order against the IGSC's comprehensive curated Regulated Pathogen Database" (genesynthesisconsortium.org); Anthropic's own biology-safeguard evaluations draw on comparable screening logic when assessing dual-use risk.

For life-sciences organizations, the near-term implication is that biology-capable frontier models will likely continue to arrive in a two-tier structure: a broadly available model with active safeguards, and a narrower, vetted-access model for validated research use, rather than a single model that unlocks fully on payment or self-certification. Given that pattern, organizations planning to request trusted access should expect an application and verification process rather than an automatic upgrade, and should budget time for it separately from a standard procurement cycle. As an adjacent advisory practice rather than a model vendor, IntuitionLabs frames this as a governance question first: which roles need which access tier, how outputs are logged and reviewed before entering a regulated workflow, and how a team measures whether adoption is actually working before scaling it further (intuitionlabs.ai).

Conclusion

Claude Fable 5.1 and Claude Mythos 5.1 are, by Anthropic's own description, the same underlying model distributed under two different safeguard regimes: one generally available with active biology and cybersecurity classifiers, and one restricted to vetted researchers and institutions with those classifiers relaxed. "Biology access," in this context, is not a marketing label but a specific, documented API and product behavior: a refusal event, a disclosed model substitution, and, for a small population, a formal verification pathway. That structure was built to satisfy Anthropic's Responsible Scaling Policy commitments following the ASL-3 activation in 2025, and it has been tuned over successive updates to reduce false positives on benign biology questions while continuing to intervene on dual-use requests involving virology, toxicology, and molecular design.

For life-sciences organizations and researchers, the practical consequences extend beyond access itself into reproducibility. Research teams should log exact model identifiers and timestamps, check for fallback events, and independently repeat consequential analytical results before treating them as settled. None of this diminishes the documented capability gains in this release, including Anthropic's reported 52.6 percent Terminal-Bench-Science 0.1 score for Fable 5.1, compared with 24.7 percent for Fable 5 (www.anthropic.com) and concrete life-sciences integrations already in production at organizations including Benchling, Biomni, and FutureHouse. It does mean that capability and access are governed separately from one another, and that a defensible research workflow needs to account for both.

Frequently Asked Questions (FAQs)

What is "biology access" on Claude Fable 5.1?

It refers to whether a biology-related query is answered directly by Fable 5.1 or automatically redirected to a different, less biologically capable model, currently Opus 5 (support.claude.com), under Anthropic's safety-classifier system.

How is Claude Mythos 5.1 different from Claude Fable 5.1?

They are, per Microsoft's own Azure documentation, effectively the same model, where "Claude Fable 5.1 is a public preview version with safeguards that fallback for cyber and bio capability queries" while Mythos 5.1 retains full capability (techcommunity.microsoft.com); Mythos 5.1 is distributed only to vetted organizations through Project Glasswing (platform.claude.com).

How does Claude's model routing work for a blocked query?

A blocked request returns an HTTP 200 response with `stop_reason: "refusal"`, and if automatic fallback is enabled, "the API retries a declined request on the fallback model Anthropic recommends" (platform.claude.com), with the substitution disclosed to the end user.

Can researchers get full biology access to Claude Mythos 5.1?

Only through Anthropic's invite-only Life Sciences Verification Program (www-cdn.anthropic.com) or a related access program with limited availability confirmed independently by Amazon's own launch documentation (aws.amazon.com); Anthropic has not published detailed public eligibility criteria for either.

Does model routing affect reproducibility of research results?

Yes. Because temperature and top-p controls are deprecated on current models (platform.claude.com), and because a refusal-triggered fallback can substitute a different model for a given query, exact reproduction through a proprietary API is "generally not possible" (arxiv.org).
