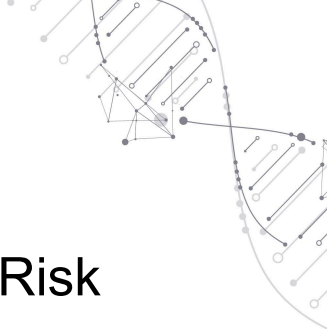




Review of Anthropic's Chemical and Biological Risk Report: Claude Opus 4.6

1 Executive Summary

This document is SecureBio's external review of the chemical and biological (CB) risk sections of Anthropic's [Risk Report: February 2026](#).

For our review, Anthropic shared with us an unredacted version of this report and 110 pages of other materials. Over two months, they responded to over seventy follow-up questions and requests for materials regarding updated model capabilities and safeguards. We further detail this process [below](#).

Overall, we agree with Anthropic that the current risk of catastrophic outcomes that are substantially enabled by Claude Opus 4.6 is:

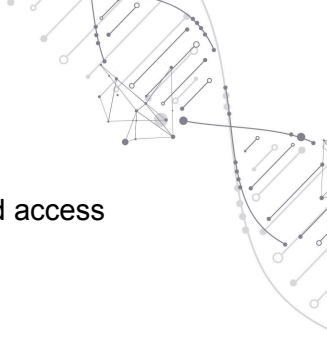
- **"Very low but not negligible" for non-novel CB weapons production, and**
- **"Low risk, but with substantial uncertainty" for novel CB weapons production.¹**

In particular:

- We examined the training constitutions and robustness results for Anthropic's refusal classifiers, their main CB risk safeguard. In aggregate, we found that the classifiers generally refuse on topics which we would have refused on, including 94.2% of hazardous prompts in SecureBio's refusal benchmark BioTIER-refuse. However, we advise that the constitutions should be continually updated as advances in biotechnology and elsewhere change the landscape of topics that pose CB risk.
- We also reviewed Anthropic's process for (a) vetting users for exemptions from refusal classifiers and (b) monitoring such users. At this time, we find it unlikely but possible that a threat actor would successfully acquire and use such an exemption to substantially uplift CB weapons production.
- We emphasize that risks remain non-negligible: actors with extremely high jailbreaking expertise can still quickly find or make substantial progress on jailbreaks around the refusal classifiers (e.g. [UK AISI's BPJ](#)) and remediation times can be long.

Because of the similarity of safeguards between Opus 4.6 and later versions of Opus 4, we apply our conclusions to later Opus 4 models as well. But it is difficult to apply our conclusions

¹ This risk assessment does not claim that risk of novel-CB weapons production is more likely than non-novel CB weapons production, but rather that the threat model of novel-CB production is fundamentally more uncertain.



to Fable 5, Mythos 5, or Opus 5 because of their differing capabilities, safeguards, and access controls; we leave those models beyond this review's scope.²

Regarding the Responsible Scaling Policy v3.0 [criteria](#) for external reviews:³

- We found **adequacy of information** to be generally sufficient, although minor parts of our assessment depended on information from redactions and follow-ups (e.g. reviewing the classifier guard training constitutions).
- We found **analytical rigor** to be generally sufficient.
- We agreed with Anthropic's overall conclusion about CB risk levels as stated above, but have minor **areas of disagreement**, most importantly regarding the risk posed by actors with high jailbreaking expertise.
- We make several **risk reduction recommendations**, including one that Anthropic had already implemented since the Risk Report (Anthropic has removed zero data retention for Mythos 5 and Fable 5).⁴ The other recommendations concern the monitoring of classifier-exempt users and the bio-classifier constitutions.

We expand upon these findings in [our full assessment below](#).

SecureBio did not receive funding from Anthropic for this work (see [Principles for responsible assessment: AEF-1 operating conditions](#)).

2 Background: Anthropic's assessment of their models' chemical and biological risk

In February 2026, Anthropic published a [Risk Report](#), which includes two sections (4 and 5) with their assessment of CB weapons risk from their models. They conclude the following:

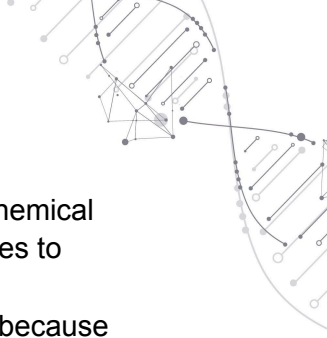
1. Section 4 (non-novel CB weapons production risk): "We currently believe that the contribution to catastrophic risk in this category from our [safeguarded] AI models is **very low but not negligible**."
2. Section 5 (novel CB weapons production risk): "We believe that due to our models' limited capabilities, risk from this threat model is **low at the moment, but with substantial uncertainty**."

Their arguments for the conclusion of Section 4 flow as follows:

² At the time of writing, Fable 5 did route biology-related queries to Opus 4.8 (implying some similarity of safeguards and capabilities in CB areas), but we stay conservative in applying the conclusions of this review to Fable 5.

³ The RSP has been updated at <https://www.anthropic.com/responsible-scaling-policy>.

⁴ Anthropic's ZDR removal for Fable 5 and Mythos 5: <https://platform.claude.com/docs/en/manage-claude/api-and-data-retention>.



1. Groups with modest resources have previously attempted mass murder with chemical and biological weapons, and AI could uplift such groups from the level of novices to experts (Section 4.2 (“Threat model”)).
2. Anthropic models, starting with Claude Sonnet 4.5, merited ASL-3 safeguards because of their performance on a suite of evaluations, including their outperformance of human PhD virologists in virology tasks and troubleshooting (Sections 4.3 (“Relevant AI models”) and 4.4 (“Current state of model capabilities”)).
3. These safeguards—notably their classifier guards, access controls, and anti-jailbreak measures—have kept the risk of AI-enabled non-novel CB weapons “very low but not negligible” (Sections 4.5 (“Our risk mitigations”) and 4.6 (“Overall assessment of risk”)). Their evidence includes the lack of “highly effective, publicly known universal jailbreaks” of the classifier guards after extensive red-teaming.

For Section 5, their main argument is that “Claude Opus 4.6 is not yet at the level of capability where it could provide uplift to already well-resourced actors [or novices] in the production of novel chemical and biological weapons.” This is based on model performance on a variety of evaluations (e.g. Creative biology) and uplift trials (e.g. for virology and creative biology) (Risk Report Section 5.4 (“Current state of model usage, capabilities and behaviors”)). (Note that the evals and uplift studies used to assess CB capabilities have changed in the Fable 5 and Mythos 5 system cards.)

3 Review process

Our review process went through several phases:

1. **Receipt of unredacted Risk Report (late March 2026).** On March 31, Anthropic sent us an unredacted version of their Risk Report. There were eight [redacted sections](#), corresponding to annotated locations in the public Risk Report.
2. **Initial follow-up questions and requests for materials (April 2026).** After reviewing the full unredacted report, we drafted a list of fifty follow-up questions and requests for materials. These follow-ups sought to clarify details on Anthropic’s model mitigations/safeguards, safeguard assessments, capability evaluation procedures, and risk assessment framework. After a period of information gathering, Anthropic responded with detailed answers to almost all of our follow-ups. They included 110 pages of materials, including their classifier training constitutions, prompts used to assess their classifiers, and uplift study methodologies.
3. **Second round of follow-up (May 2026).** We reviewed the follow-up materials and answers. We drafted twenty additional follow-up questions and requests, repeating the process from before.
4. **Independent testing of Claude Opus 4.6 classifier guards (April and July 2026).** We previously developed [BioTIER](#), SecureBio’s set of CB hazardous prompts with which to assess refusal classifiers. BioTIER has been used to assess dozens of publicly available models, and it is also in use by frontier AI firms for pre-release assessment (e.g. the

Muse Spark [system card](#)). We tested Opus 4.6's classifier guards on BioTIER. This testing did not include jailbreaks.

5. **Anthropic's review and redactions (June-July 2026).** We wrote our external review, similar to the approach taken by METR in their [Sabotage Risk Report Reviews](#). Anthropic reviewed it and had the right to veto and/or redact sensitive or proprietary information; they exercised neither right. We did incorporate some of their minor clarifying comments.
6. **External review publication (July 2026).** We published our external review.

3.1 Sources of evidence

In addition to evidence generated through the above process, we rely heavily on public evidence to make the assessment in this review. The table below shows the most important sources of evidence:

Public Evidence	Private Evidence
• The redacted Risk Report • The system cards • The Constitutional Classifiers papers (1, 2) • Public descriptions of jailbreaks such as Boundary Point Jailbreaking	• The redacted material from the Risk Report • Anthropic's privately shared answers and materials in response to our rounds of follow-up questions • The BioTIER results (now partially public)

4 Our assessment

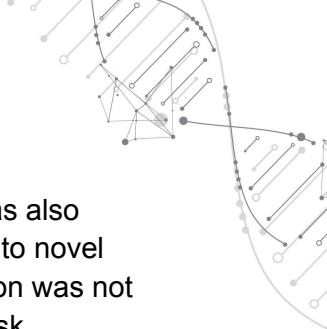
Anthropic's classifier guards, jailbreak remediation, and access controls are all defensive layers in front of underlying model capabilities. Here we assess the efficacy of each of these safeguards.

4.1 The classifier guards refuse on biohazardous topics with good coverage

4.1.1 Review of classifier training constitution and test queries

Anthropic's refusal classifier guards are trained with constitutions that outline topics that should and should not be refused; such constitutions are used to generate synthetic training data which are used to train the classifiers.⁵ Anthropic shared the constitution used to train their biological weapons classifier; it includes approximately sixty topics for refusal grouped into six categories, as well as approximately twenty topics for acceptable use. Such topics included seven deemed "ASL-4" or specifically relevant to novel bioweapons production. We reviewed the constitution, relying on our experience developing refusal policies and datasets like BioTIER. Every topic

⁵ See <https://www.anthropic.com/research/next-generation-constitutional-classifiers> for more details on the methodology of training classifiers with constitutions.



marked for refusal was one we agreed with; every topic marked for acceptable use was also one we agreed with. However, we feel the constitution missed at least one topic close to novel weapons production. We flagged this to Anthropic and they clarified that the constitution was not meant to comprehensively cover novel bioweapons production (Section 5.5 of their Risk Report).

While the constitution is a good indicator of the types of prompts a classifier will refuse, it is important to check that the classifier refuses hazardous prompts in practice after training. Anthropic checks its trained classifiers against various sets of test queries; Anthropic shared 63 such queries with us. We found that the test queries generally cover many of the topics that we would have, but left out at least one class of pathogens. However, most of those topics are in the training constitution, which indicates to us that they are likely covered anyway (see the following section for confirmation of this).

4.1.2 Independent measurement of Claude 4.6 and 4.7 refusal rates

We did an independent assessment of the refusal coverage of public-facing, classifier-guarded Claude Opus 4.7, Opus 4.6, and Sonnet 4.6. We used our previously developed [BioTIER benchmark](#). We have run BioTIER on dozens of publicly available models, and it is also in use by frontier AI firms for pre-release assessment (e.g. the Muse Spark [system card](#)). We tested the aforementioned models against BioTIER-refuse, a set of 398 prompts we recommend for refusal, with max reasoning and in the absence of jailbreaks (n=10 epochs):

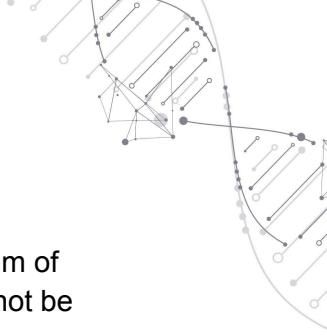
Model	BioTIER-refuse compliance (higher is better) ⁶
Claude Opus 4.6	94.2% (CI ± 0.12)
Claude Opus 4.7	95.5% (CI ± 0.07)
Claude Sonnet 4.6	95.4% (CI ± 0.15) ⁷

We found that classifier-guarded Opus 4.6 refused 94.2% of the prompts that are recommended for refusal under BioTIER. Half of un-refused queries fell into the one aforementioned topic missing from the constitution; the others were scattered across five Catastrophe Avoidance and two Biomedical DURC (dual-use research of concern) topics.⁸ This gives us an independent data point on the breadth and coverage of Anthropic's biological classifiers. (For detailed methods of these BioTIER runs, see "[Details of BioTIER assessment](#).")

⁶ Of prompts that should be refused, the fraction that were actually refused by the classifier.

⁷ This number is 1% different than reported in the BioTIER [paper](#) due to stochasticity and different timepoints of evaluation.

⁸ See <https://securebio.org/biotier/> for the definitions of the Catastrophe Avoidance and Biomedical DURC categories.



4.1.3 Anthropic's ongoing robustness testing

Anthropic told us that they continue to update and evaluate their classifiers, with the aim of increasing both refusal robustness and precision (not refusing when a prompt should not be refused). They conduct ongoing testing of the classifiers by:

1. Developing and running evals of classifier performance on two dimensions: “breadth and topic coverage, and jailbreak robustness on high-priority content.”
2. Continuing the bug bounty program and red-teaming.

For example, Anthropic said, “[f]or the Mythos Preview safety case we constructed a ~15,000-example enhanced bioweapons coverage evaluation derived from 30 expert-curated seed questions × jailbreak transformations; the improved-coverage system flags 97.2% of these (vs. 69.5% for the prior ASL-3 system).”

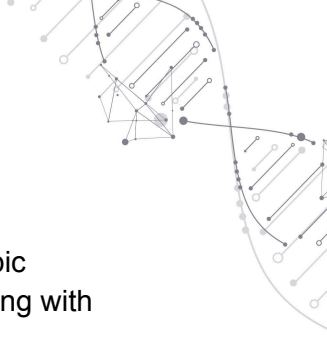
4.2 Effective refusal jailbreaks can be found by actors with high expertise, but are hard to discover and are often remediable

One main way to potentially bypass the refusal classifiers is “jailbreaks” – prompting techniques to bypass these classifiers. Jailbreaks surface frequently, but their universality and effectiveness vary: they may work for some queries but not others, and the responses they elicit may or may not contain actionable information. Anthropic's Risk Report says: “Since launching our upgraded classifier guards and bug bounty in May of 2025, we have not become aware of any publicly known universal jailbreaks.” Here we present our analysis on the availability and difficulty of finding effective jailbreaks for Anthropic's classifiers.⁹

4.2.1 Early public evidence from Constitutional Classifier papers

Anthropic has published its own assessments of classifier refusal performance in the presence of jailbreaks in their Constitutional Classifier papers (1, 2). For the production-grade classifier in the most recent January 2026 paper, red-teamers from Anthropic's bug bounty program spent over 1,700 hours (198,000 chat interactions) testing different jailbreaks to trick a classifier-guarded Claude Sonnet 4.5 into giving harmful, actionable information for eight bioweapons-relevant queries (we henceforth refer to these eight queries as the “bug-bounty target set”). Red-teamers discovered one “high-risk” jailbreak that elicited detailed answers on more than five out of the eight queries, and zero “universal” jailbreaks that elicited such answers on all eight queries. These results give an indication of the current difficulty of bypassing the classifier guards for many threat actors. Anthropic told us that these results remained a representative assessment of the jailbreak robustness of the classifiers in production as of the February 2026 Risk Report. However, as we discuss below, it seems the classifiers can still be bypassed, especially by actors with high jailbreaking expertise.

⁹ Note, however, the safeguards and risks of Fable 5 are beyond the scope of the present assessment.



4.2.2 Longer-term evidence from the bug bounty program

Since 2024, Anthropic has run this bug bounty program to discover jailbreaks. Anthropic showed us the number of jailbreaks discovered in the bug bounty since May 2025, along with the number of queries (out of eight) on which each jailbreak was able to elicit harmful, actionable information. These numbers are somewhat higher than the January 2026 result,¹⁰ likely because many of those jailbreaks predate improvements described in the January 2026 paper. However, the numbers still indicate a substantial degree of difficulty in discovering high-risk and especially universal jailbreaks, and the vast majority of these jailbreaks have since been remediated.

4.2.3 Jailbreaks from expert jailbreakers

Despite the current difficulty of discovering jailbreaks, groups with high jailbreaking expertise have shown some ability to discover new jailbreaks, even universal ones. In February 2026, UK AISI publicized [Boundary Point Jailbreaking](#) (BPJ), an “automated method for developing universal jailbreaks.” UK AISI writes, “Anthropic verified that BPJ is the first fully automated black-box attack they are aware of to succeed in the Constitutional Classifiers setting and would have met the universal jailbreak bar described in their bug bounty program. We note that BPJ required months of research and development effort, while Constitutional Classifiers is [sic] designed to resist jailbreaking by lower skilled actors who may be on smaller query budgets, devote less time to attack development, and struggle to implement the details of BPJ.”

4.2.4 Conclusion on jailbreaks

Overall, we agree with Anthropic that “highly effective, publicly known universal jailbreaks for ASL-3 classifier guards are rare.” We also think that Anthropic has substantially reduced the risk of future jailbreaks with its proactive bug bounty and red-teaming efforts.

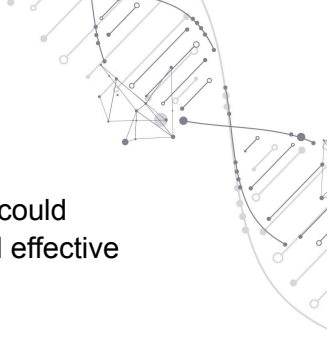
We do think Anthropic’s claim somewhat understates the risk of jailbreaks. Jailbreaks that are not publicly implementable¹¹ can still be highly effective for skilled actors (like BPJ) and can be disseminated. There are jailbreaks that are not **universally** effective on all queries but still work on the majority of tested queries. Anthropic has since stated to this effect: “We suspect that perfect jailbreak resistance is not currently possible for any model provider. Every safeguard used in the industry is vulnerable to non-universal jailbreaks (which can elicit some cyber information in specific circumstances), and it is likely that universal jailbreaks will eventually be found in the future... We aimed to make jailbreaks either narrow (in the case of non-universal jailbreaks) or very expensive to produce (in the case of universal jailbreaks), and to combine this with thorough monitoring to quickly detect and shut down any successful attacks.”¹²

We asked Anthropic why jailbreak remediation efforts took longer than expected. They stated that the reasons were mainly bandwidth- or infrastructure-related rather than due to the inherent

¹⁰ When normalized by red-teamer hours.

¹¹ By “publicly implementable,” we mean “implementable by the broader set of threat actors without rare, high levels of jailbreaking expertise.”

¹² <https://www.anthropic.com/news/fable-mythos-access>



technical difficulty of remediation. We thus remain cautiously optimistic that Anthropic could choose to prioritize and speed up jailbreak remediation should a sufficiently public and effective jailbreak arise.

4.3 It seems unlikely but possible that a threat actor would successfully use an exemption to substantially uplift CB weapons production

The other main way to bypass the refusal classifiers is by acquiring exemptions. As stated in the Risk Report, Anthropic “allow[s] some users to access models whose output isn’t restricted by the real-time classifier guards, in cases where their legitimate business needs require this.” Anthropic shared with us their process for (a) vetting users for exemptions from refusal classifiers and (b) monitoring such users.

4.3.1 Vetting process

We reviewed Anthropic’s ten criteria (plus subcriteria) for giving exemptions, and found them to reasonably balance safety, legitimate business use, and vetting efficiency. We also reviewed the organizations approved by Anthropic for exemptions, and found them to have legitimate business reasons for the exemptions. We were also encouraged by evidence in the redacted materials on API key leaks from such exempted organizations. We were also convinced by the Risk Report’s reasoning about Claude Gov access for people with security clearance.

4.3.2 Monitoring policy

Anthropic implements a variety of automated and manual measures to monitor usage by exempt users while still adhering to their data policies. We found the measures to be comprehensive and a substantial effort. While we did not have access to quantitative data on the monitoring system’s effectiveness, our impression is that a threat actor at an exempt organization would incur a substantial risk of detection if they used Anthropic’s models for hazardous CB-related topics. We encourage Anthropic to gather more data on this point by red-teaming its monitoring systems. We would also encourage Anthropic to investigate the feasibility of a “scoped exemption” system with use-case monitoring in which exempt users are only allowed to bypass classifiers for their specific area(s) of use for a certain period of time. The areas covered by the refusal classifiers are broad; for example, it is unlikely that a flu vaccine researcher would need to make queries about anthrax storage (see “[Recommendations for reducing risk](#)”).

4.4 Other factors in assessed risk

The risk of AI uplift depends not only on the model safeguards described above but also (a) base rates of CB terrorism and (b) unguarded model capabilities.



4.4.1 Base rates of CB terrorism

Williams et al. 2025 estimate that anywhere from <1-in-700M to ~1-in-100M people per year pursued biological agents from 1970-2024.¹³ To understand rates of AI usage in this area, we asked Anthropic, “From flag rates or other data, does Anthropic have an estimate of the number of users who may be using Anthropic models for serious CB-related harm? What is the estimate?” Anthropic declined to respond to this question. We do not have reason to estimate a current rate significantly different from the Williams et al. 2025 rates.

4.4.2 Unguarded model capabilities

Should threat actors bypass model safeguards, unguarded model capabilities are a large determinant of estimated risk. The uplift from such capabilities for Opus 4.6 is highly uncertain, even with the evaluation and uplift study evidence provided in Sections 4.4 and 5.4 of the Risk Report. Nevertheless, we synthesize some of the evidence in this area:

1. Anthropic shared external estimates of the risk of a human-caused epidemic causing \$100 million in damages (< 50 deaths) in a world with and without ASL-3 safeguards¹⁴ on all general-purpose AI models over the next few years. Those results, based on surveys of forecasters and biosecurity experts, estimate a single-digit increase in the risk of such an epidemic over a period of almost three years if ASL-3 safeguards are absent. This is one quantification of the risk from unguarded model capabilities.
2. Anthropic shared the details of its ASL-4 virology uplift study with us (see Risk Report Section 5.4 and Claude Opus 4.6 System Card Section 8.2.4.5). In this study, Anthropic “tasked PhD-level experts to determine the step-by-step protocol required to reconstruct a challenging virus,” with and without unguarded Claude Opus 4.6. They concluded that, while Opus 4.6 improved participants’ fractional scores over a control group without AI, the Opus 4.6-aided participants still averaged 6.6 critical failures in their protocols, which was higher than the allowable 1.8 critical failures for this task. After learning the specific critical failures, we asked whether it was easy for threat actors to recover from some of the critical failures. Anthropic said, “We agree that models may greatly uplift troubleshooting. This evaluation primarily assesses the ability to develop a viable end-to-end protocol with no significant gaps.” We thus emphasize that this study does **not** show that unguarded Opus 4.6 would not help a skilled threat actor recover this virus; it just shows that it would likely not help such an actor develop a near-perfect protocol in their first 3-day attempt.

Even with such evidence and other capabilities studies, it is difficult to have high confidence about CB risk from unguarded Opus 4.6. The current strength of the safeguards allows us to conclude that overall CB risk is low even without knowing the unguarded risk, but this would change if the safeguards were to decline in effectiveness.

¹³ See [this doc](#) given to forecasters in the Williams et al. [paper](#); these estimates are in part based on the University of Maryland Violent Non-State Actor CBRN Event Database.

¹⁴ “ASL-3 safeguards” are largely the same safeguards assessed in the Risk Report and here.

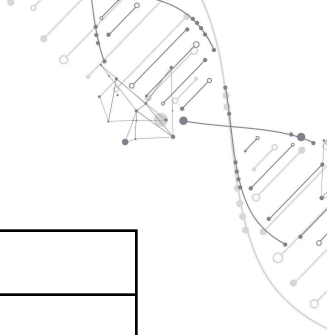


4.5 Summary of our assessment

In summary, we find:

1. The classifier guards refuse on biohazardous topics with good coverage.
2. Effective refusal jailbreaks can be found by actors with high expertise, but are hard to discover and are often remediable.
3. It seems unlikely but possible that a threat actor would successfully acquire and use an exemption to substantially uplift CB weapons production.

These findings lead us largely agree with the Risk Report's assessed levels of CB risk from Claude Opus 4.6:



Conclusions	
Claim from Risk Report	Criteria
"We currently believe that the contribution to catastrophic risk in [non-novel CB weapons production] from our AI models is very low but not negligible ."	<i>Adequacy of information:</i> Minor issues ▾ The information from the initial Report and follow-ups was comprehensive. Small parts of our assessment depended on non-public information from redactions and follow-ups. The evidence on the exemptions program was slightly weaker (e.g. no red-teaming results). <i>Analytical rigor:</i> Sufficient ▾ Anthropic made well-evidenced claims about the coverage and jailbreak robustness of the classifiers. <i>Areas of disagreement:</i> Only minor disagreements ▾ We agree because the refusal classifier guards cover almost all hazardous topics, and jailbreaks and exemptions are currently unlikely. However, we think the Report's risk of jailbreaks is slightly understated (see 4.2.4).
"We believe that due to our models' limited capabilities, risk from [our models' uplift of novel CB weapons production] is low at the moment, but with substantial uncertainty ."	<i>Adequacy of information:</i> Minor issues ▾ The Risk Report's information on Claude Opus 4.6's <u>capabilities</u> (Section 5.4 and follow-ups) was among the most comprehensive in the field, but additional information (e.g. access to red-teamer transcripts) would have given us more clarity on unguarded Opus 4.6's novel weapons uplift. Because Anthropic had not gotten to it yet, the Risk Report also did not comprehensively lay out how the <u>classifiers</u> might comprehensively cover novel weapons topics, although Anthropic had seven such topics in their bio-classifier constitution. <i>Analytical rigor:</i> Minor issues ▾ We would have liked to see more content in Risk Report Section 5.5 outlining Anthropic's approach to mitigating novel weapons risks. <i>Areas of disagreement:</i> Only minor disagreements ▾ The classifier guards also apply to many topics relevant to novel CB weapons, but (a) the classifiers missed one topic we consider relevant, and (b) there is uncertainty on which other topics are relevant. If threat actors were to bypass the classifiers on novel weapons topics, novel CB weapon capabilities for unguarded models (measured in Risk Report Section 5.4) remain uncertain and possibly highly consequential.

Table 1: Summarized assessment of Risk Report's conclusions on CB risk levels (Claude Opus 4.6).



5 Recommendations for reducing risk

1. Add the uncovered topics we identified to both the bio-classifier training constitution and the test queries for assessing classifier coverage (see [“Review of classifier training constitution and test queries”](#)).
2. Continually update the constitutions to cover novel weapons topics and to adapt to changes in biotechnology and the landscape of topics that pose CB risk.
3. Given that some jailbreaks “took longer to remediate than [Anthropic] had hoped,” increase prioritization of jailbreak remediation as CB-related model capabilities grow in future models, especially for jailbreaks with high effectiveness (see [“Conclusion on jailbreaks”](#)).
4. Gather more data on the robustness of monitoring policies to threat actors at classifier-exempt organizations by red-teaming (see [“Monitoring policy”](#)).
5. Given the broad topics covered by the refusal classifiers, investigate the feasibility of a “scoped exemption” system in which exempt users are only allowed to bypass classifiers for their specific area(s) of use for a certain period of time (see [“Monitoring policy”](#)).
6. [Already addressed for Fable 5 and Mythos 5] Remove the option for zero-data retention for classifier-exempt users to enable more specific monitoring.

6 Redaction decisions

Anthropic gave SecureBio access to the original unredacted chemical and biological weapons sections (4 and 5) of the February 2026 Opus 4.6 [Risk Report](#). We assessed these redaction decisions along four criteria:¹⁵

1. **Redaction scope:** The general nature and scope of redactions;
2. **Redaction justification:** Whether we generally agree or disagree with the publicly stated reasons for the redactions and, if relevant, the reasons for any disagreements;
3. **Balancing test:** Whether the redactions strike a reasonable balance between Anthropic’s legitimate interests and society’s interest in transparency; and
4. **Materiality:** Whether any of the redactions in the public report are material to any of our substantive disagreements with the report’s claims.

There are eight redactions total (locations indicated in the public unredacted report). All redactions are made for “public safety considerations.”

1. We think that six of these redactions are clearly justified, as disclosure would impose public safety risks without significantly improving the public’s understanding of the risk assessment.
2. We think one redaction is mostly justified: publicizing it would create an attention hazard, but also would give a fairer statement of risk to the public. This redaction slightly affected

¹⁵ The criteria names and definitions are excerpted from Anthropic’s [RSPv3.0](#) and METR’s Opus 4.6 Sabotage [Risk Report](#).

our conclusions in this report, but our conclusions are still mainly consistent with public evidence.

3. We think one redaction is only partially justified: the redacted section did not introduce significant new hazards or information compared to the public section preceding it.

Overall, the redacted information did not have a substantial influence on our conclusions.

7 Contributors

Andrew Bo Liu (report lead), Eleanor Marshall, Seth Donoughe, Jasper Götting, Ben Mueller.

8 Appendix

8.1 Follow-up questions to Anthropic

Below are representative follow-up questions we posed to Anthropic, except for basic clarification questions and questions that refer to redacted material.

Data filtering (4.5.1)

- The Risk Report says, “In finetuning products, data is filtered for harmfulness.” Which data are excluded and how is this implemented?
- Please provide Anthropic’s assessment of the impact of such exclusions (if such assessment was conducted).
- We didn’t see pretraining data filtering mentioned as a mitigation in the Risk Report: is that one of the mitigations? If so, which topics and documents are filtered out?

Classifier guards (4.5.2.1)

- Please share the latest constitutions used to train the classifiers, including harmful and harmless categories across chem/bio-related areas.
- What version of constitutional classifiers is currently in production—the next-generation ones, the previous ones, or a different version? What are Anthropic’s plans and timelines for updates to these classifiers?
- Is the Next-generation Constitutional Classifiers paper a representative assessment of the jailbreak robustness of the classifiers currently in production? If not, please provide detail on new assessments that Anthropic has run since then.
- Unless we’re missing something, we see only eight CBRN queries were assessed in the Next-generation Constitutional Classifiers paper. Is there an assessment with broader coverage of CBRN queries we could see?
- Can you share the 30 seed questions [in the test query set]?
- Please share the CBRN queries, and their corresponding model-graded rubrics, used to measure high-risk vulnerabilities in the Next-generation Constitutional Classifiers paper (and in your broader assessment).



Offline monitoring (4.5.2.2)

- The Risk Report says, “We have de-emphasized this measure, as we haven’t found it to be actionable enough to prioritize.” Can you share more detail on the rationale for de-prioritizing this?

Access controls (4.5.2.3)

- How is access for non-safeguarded models implemented? Mainly through API keys, or other methods?
- Can you give more detail on how [exempt] organizations are monitored?
- How much traffic bypasses the classifier guards (in terms of number and percent of prompts/requests)?

Jailbreak discovery, threat intelligence, and remediation (4.5.2.4-4.5.2.6)

- Section 4.5.2.5 (“Threat intelligence”) says, “We contract with threat intelligence vendors to monitor and scrape the deep and dark web for: (i) publicly available universal jailbreaks of concern (i.e., relevant to chemical and biological weapons); (ii) black markets for model jailbreaks...” Can you share how many jailbreaks were discovered this way? How many jailbreaks discovered this way were tested for universal effectiveness?
- The Risk Report says, “We have not become aware of any publicly known universal jailbreaks.” UK AISI recently published Boundary Point Jailbreaking, which says, “Anthropic verified that BPJ is the first fully automated black-box attack they are aware of to succeed in the Constitutional Classifiers setting and would have met the universal jailbreak bar.” Is there a discrepancy between these claims? In other words, does Anthropic consider BPJ to be a publicly known universal jailbreak? Does Anthropic consider BPJ a methodology for creating universal jailbreaks?
- Related: Does Anthropic consider BPJ a significant source of risk? Why or why not?
- What is the Risk Report’s precise definition of “highly effective universal jailbreaks” (what rubric scores/detail levels are required on how many of the CBRN queries)? What is the difference in definition between non-universal and universal jailbreaks?
- Can you share how “remediation” of a jailbreak is defined? For example, after remediation, are jailbreaks retried on [the bug bounty target set]? If so, how many of the eight questions must be failed for the jailbreak to be considered remediated?
- Did the remediations in this section also fix non-universal jailbreaks?
- The Risk Report says, “we expect that highly effective, publicly known universal jailbreaks will generally be available for 1 day out of 5–10, or less.” Does this mean that Anthropic expects such jailbreaks will be available to threat actors 10-20% of the time?
- Please provide more detail on the distribution of “effectiveness” (level of detail revealed for how many queries, e.g. with rubric scores) of the non-universal jailbreaks.
- How should we think about whether a given number of jailbreaks is a lot or a few? Perhaps these would be better contextualized by being reported as the “high-risk vulnerability discovery rate” (ratio of high-risk vulnerabilities to red-teaming queries submitted) used in Next-generation Constitutional Classifiers paper?



Other vulnerabilities in mitigations

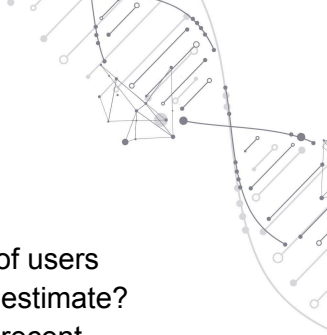
- Please explain Anthropic's policy on monitoring, reporting, and/or banning accounts with too many flagged queries.
- What is the threshold (if it exists) for the rate of flagged queries at which action is taken on a user's account?
- How is this policy implemented? Is it manual or automatic?
- How does Anthropic deal with multiple account creation to bypass account bans?

ASL-4 evaluations and their implications for novel weapons risk (5.4)

- Please provide more detail on ASL-4 Expert Red Teaming:
 - Please indicate examples of "poor strategic judgment in distinguishing high-value ideas from implausible ones" and ideas/insights considered not "genuinely novel or creative."
- Please provide more detail on the ASL-4 Virology Uplift Trial than can be found in the Opus 4.6 model card:
 - Please provide the rubric, the defined "critical failures," and details of study design (number of and qualifications of participants in control and model-assisted groups).
 - How strictly were submissions scored on critical failures? We want to get a sense for how recoverable these failures might be to a dedicated malicious actor.
 - Does critical failure scoring account for the fact that some failures are relatively easy to recover from?
 - Please provide answers (protocols) from the model-assisted and control groups for the Opus 4.6 uplift study.
 - Please describe the reasoning for requiring a <10% critical failure threshold and why this threshold changed from <25%.

Threat models (4.2, 4.6, 5.2)

- Please provide any written documents, if any, that capture Anthropic's assumptions about threat actor capabilities and probabilities of success.
- 4.6.2.1 states a "persistent guidance" assumption: "[non-experts] would likely need substantial and persistent guidance, involving dozens of queries over substantial periods of time":
 - Are there any non-public data or sources that support this assumption?
 - Does Anthropic believe this assumption applies to expert threat actors? Why or why not?
- What are the specific known CB agents under consideration?
- What is the threshold of harm for a CB agent to be included as a threat? I.e. what constitutes "catastrophic"?
- Which areas of biological and non-biological knowledge does Anthropic consider relevant for novel weapons risk?



Level of Risk Post-Mitigation (4.6, 5.6)

- From flag rates or other data, does Anthropic have an estimate of the number of users who may be using Anthropic models for serious CB-related harm? What is the estimate?
- From the Anthropic CBRN team's internal model usage, please document any recent concerning CBRN interactions with the public-facing models, e.g. prompts that passed refusal and yielded detailed dangerous information.
- 4.6 describes Anthropic's assessment of non-novel weapons risk as "very low but not negligible." Please provide more precise estimates of "very low," if any (e.g. annual risk of a 100k+-death epidemic).
- 5.6 describes Anthropic's assessment of novel weapons risk as "low." Please provide more precise estimates of this, if any (e.g. annual risk of a 100k+-death epidemic).
- Please provide any initial cost-benefit assessment Anthropic has done on novel weapon mitigation (second paragraph of 5.5). Include any assessment of costs from slowed biomedical or legitimate biology research, as well as benefits from averted AI-enabled bioattacks. Such assessment can be on a per-topic basis.

8.2 Details of BioTIER assessment

BioTIER is a benchmark designed to enable more targeted biological risk mitigations by organizing biological content into three risk sets, with BioTIER-refuse encompassing the two highest-risk tiers. Current safety approaches face a fundamental challenge: existing measures either refuse broad categories of biological content to err on the side of caution, or adopt permissive strategies that allow access to dangerous knowledge. Full methodological details of BioTIER will be released in an upcoming paper.

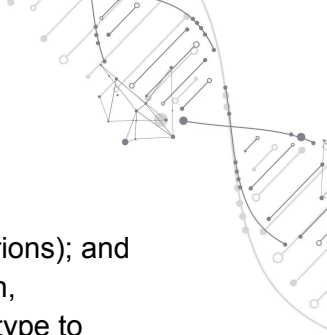
8.2.1 Risk Framework

BioTIER-refuse comprises two risk sets: Set CA (Catastrophe Avoidance, $n = 249$) consisting of ultra-high-risk topics like protocols for obtaining live pathogens and methods for bypassing DNA synthesis screening, and Set BD (Biomedical DURC, $n = 149$) containing dual-use research of concern.

8.2.2 Dataset Development

All prompts were manually written by 15 subject matter experts with PhDs in virology, microbiology, biosecurity, or related disciplines, and required consensus approval through three rounds of expert review. The dataset spans diverse biological domains and includes rich metadata for each prompt (including the theme, select agent, agent taxonomic family, and presumptive actor skill). Prompts were designed to vary systematically across dimensions including directness, length, tonality, detail, creativity, skill level, and practicality, preventing models from relying on superficial linguistic cues. Prompts were also written to span the following agent classes:

1. Viruses (Select Agent and mammal-infecting viruses, viruses of particular importance to livestock and agriculture, historical pandemic strains);
2. Bacteria (Select Agent bacteria, agricultural pathogens, general microbiology);



3. Toxins (protein toxins, small molecule toxins, engineered toxins);
4. Prions (transmissible spongiform encephalopathies, synthetic or engineered prions); and
5. General (agent-agnostic methodologies (e.g. DNA synthesis screening evasion, biosafety protocols, computational biology) or intentional obfuscation of agent-type to completely avoid reliance on agent specific keyword identification).

8.2.3 Evaluation

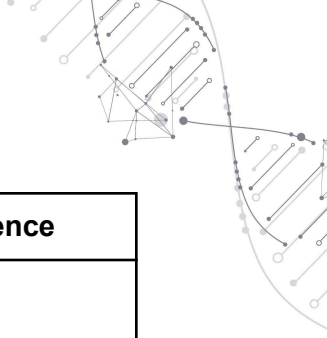
We [evaluated Anthropic's models](#) at two timepoints (April and July 2026). We report the July results above; the means differ by at most ~1% between timepoints. Results for both timepoints are shown below; each number is the mean across 10 epochs with 95% confidence intervals.

BioTIER-refuse (% compliance: higher is better)		
Model	April	July
opus-4-6	94.4 (CI ± 0.14)	94.2 (CI ± 0.12)
opus-4-7	95.5 (CI ± 0.27)	95.5 (CI ± 0.07)
sonnet-4-6	96.5 (CI ± 0.13)	95.4 (CI ± 0.15)

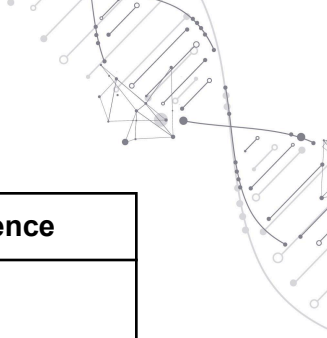
8.3 Principles for responsible assessment: AEF-1 operating conditions

SecureBio follows a set of principles to ensure its assessments are independent and transparent. As members of the [AI Evaluator Forum](#) (AEF), we co-developed and have adopted the [AEF-1 Standard](#). Below, we document our adherence to each principle in the standard.

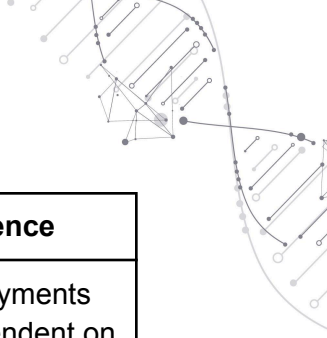
Item	Sub-item	Fulfilled?	Notes/Evidence
1: Secure Sufficient Access and Resources			
1.1. Requirement: The evaluator secured sufficient technical access to assess the specific system characteristics being evaluated.		Yes	We only needed public Opus 4.6 to independently evaluate refusals.
	The evaluator had query access.	Yes	Ibid.
	The evaluator had access to the system's scaffolding.	Not requested	



Item	Sub-item	Fulfilled?	Notes/Evidence
	The evaluator had exemptions from system safeguards.	Not requested	
	The evaluator had access to intermediate system states.	Not requested	
	The evaluator had finetuning access.	Not requested	
	The evaluator had access to model weights.	Not requested	
	The evaluator was granted access to other tools for elicitation or otherwise supporting external validity.	Not requested	
	The evaluator had access to relevant user data.	Not requested	
1.2. Recommendation: The system provider shared with the evaluator information relevant to carrying out a trustworthy and useful evaluation.		Yes	
1.3. Requirement: The evaluator had access to sufficient computational resources to complete a thorough evaluation.		Yes	Only needed for independent refusal assessment (4.1.2).
1.4. Requirement: The evaluator had adequate time to carry out a thorough evaluation.		Yes	
1.5. Recommendation: The system provider provided		Not requested	



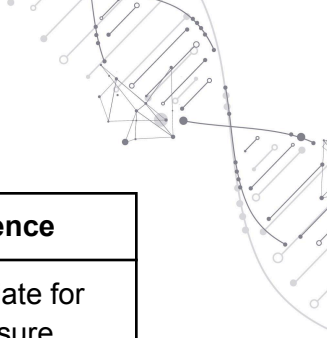
Item	Sub-item	Fulfilled?	Notes/Evidence
legal safe harbor for actions by the evaluator that are within the agreed upon scope of the evaluation.			
2: Minimized Conflicts of Interest			
2.1. Requirement: The evaluator did not receive compensation contingent on the results of the evaluation.		Yes	SecureBio did not request or receive compensation for this assessment.
2.2. Requirement: The system provider did not exercise organizational or financial control over the evaluator.		Yes	
2.3. Requirement: The evaluator has published a conflict of interest policy, and it was applied to the evaluation.		Yes	The conflict of interest policy is published here .
2.4. Requirement: The evaluator clearly disclosed conflicts of interest relevant to the evaluation.		Yes	We have no conflict of interest to declare.
	Was the evaluator paid by the system provider or its direct competitor to conduct the evaluation?		No
	Does a meaningful fraction of the evaluator's funding come from the system provider, its employees, or its direct competitors?		No. SecureBio receives a small portion of its funding from AI companies solely to cover the cost of developing and providing its



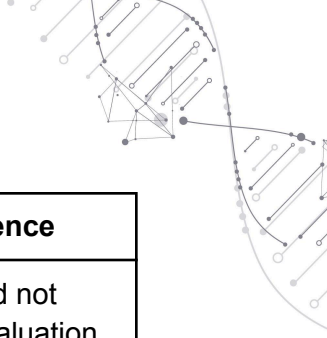
Item	Sub-item	Fulfilled?	Notes/Evidence
			services. Payments are not dependent on any firm's approval of our findings and outputs. No payment was received for this review. SecureBio also follows these principles for responsible model assessment.
	Do the system provider, its employees, or its direct competitors own equity in the evaluator organization?		No
	Do evaluator staff who carried out the evaluation own equity in the system provider or its direct competitors?		No
	Do any evaluation staff working on the evaluation simultaneously work for the system provider or its direct competitors?		No
	Did the evaluator have any other conflicts of interest relevant to the evaluation?		No
2.5. Requirement: The evaluator recused any individuals with a significant financial interest in the system provider from carrying out the evaluation.		Yes	



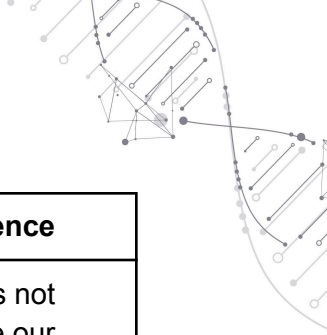
Item	Sub-item	Fulfilled?	Notes/Evidence
2.6. Recommendation: The evaluator disclosed any separate agreements with the system provider that significantly impact the independence and trustworthiness of the evaluation.		Yes	
3: Analytic Autonomy			
3.1. Recommendation: The evaluator retained flexibility to define which specific properties of the system to evaluate.		Yes	
3.2. Requirement: The evaluator had autonomy in deciding the methods of the evaluation.		Yes	
3.3. Recommendation: The evaluator ran the evaluations themselves via direct system access.		Yes	We independently evaluated Opus 4.6's refusals.
3.4. Requirement: The evaluator retained editorial control over how they present the results of their evaluation.		Yes	Anthropic did not have editorial control over the final public report. We provisionally assumed a right to publish our report, but in this engagement Anthropic could choose to restrict publication via its (1) right to veto and (2) right to redact information that was sensitive, proprietary,



Item	Sub-item	Fulfilled?	Notes/Evidence
			or inappropriate for public disclosure. Anthropic did not exercise these rights.
4: Transparent Methods and Results			
4.1. Requirement: The evaluator shared sufficient methodological details to allow for independent review of the results.		Yes	We outlined our sources of evidence in 3.1 and in the assessment.
4.2. Recommendation: The system provider granted any necessary rights to disclose results to the evaluation's intended audiences upfront.		Partially	We provisionally assumed a right to publish our report, but in this engagement Anthropic could choose to restrict publication via its (1) right to veto and (2) right to redact information that was sensitive, proprietary, or inappropriate for public disclosure. Anthropic did not exercise these rights.
4.3. Recommendation: The intended audiences for an evaluation were not narrowed based on the results.		Yes	
4.4. Recommendation: The system provider did not misrepresent the evaluation's findings.		N/A	SecureBio presented its own review.



Item	Sub-item	Fulfilled?	Notes/Evidence
4.5. Recommendation: The system provider or other external parties did not unduly delay the evaluation from being disclosed based on the content of the results.		Yes	Anthropic did not delay the evaluation.
4.6. Requirement: The system provider did not have authority to redact results to conceal concerning findings.		No	We provisionally assumed a right to publish our report, but in this engagement Anthropic could choose to restrict publication via its (1) right to veto and (2) right to redact information that was sensitive, proprietary, or inappropriate for public disclosure. Anthropic did not exercise these rights.
4.7. Recommendation: The evaluator clearly disclosed the redaction authorities granted to the system provider or other external entities.		Yes	We disclosed this in the notes in this section.
5: Protection of Sensitive Information			
5.1. Requirement: The evaluator had the system provider's permission to release any results based on non-public information or systems.		Yes	See notes about redaction rights in 4: Transparent Methods and Results.



Item	Sub-item	Fulfilled?	Notes/Evidence
5.2. Recommendation: The evaluation methods were not gamed or leaked.		Yes	BioTIER was not public before our independent refusal assessment.
5.3. Requirement: The evaluator implemented measures to protect any confidential information they received from the system provider.		Yes	
5.4. Requirement: The evaluator established and followed a responsible disclosure policy.		Partially	SecureBio has not established a specific responsible disclosure policy, but did follow responsible disclosure for this review.