Technical Report

Third-Party Assessments

REPORT SERIES

Implementing Frontier AI Frameworks

DATE

August 4, 2025

About the Report

This is the fourth in a **series of technical reports** on frontier AI frameworks that examine how the frameworks can be used effectively across different organizational contexts. The series intends to provide detailed insight into key components of these frameworks, incorporating lessons from early adopters while acknowledging areas where best practices continue to evolve.

Executive Summary

Third-party assessments can be conducted on frontier models to confirm evaluations or claims on critical safety capabilities and mitigations. In appropriate contexts, these assessments may help to confirm or build confidence in safety claims, add robust methodological independence, and supplement expertise. This report outlines practices and approaches among Frontier Model Forum (FMF) firms for implementing, where appropriate, rigorous, secure, and fit-for-purpose third-party assessments.

Frontier AI frameworks typically employ a two-stage risk assessment process carried out by internal assessment teams: **Frontier Capability Assessments** and **Mitigations Assessments**. Third-party assessments can complement and support both stages of this process. Developers may engage third-party assessors at the times and depths most likely to produce meaningful test signals, calibrated to model characteristics, deployment context, assessor qualifications, and risk.

Third-party assessments can serve three primary functions in frontier AI frameworks:

1. **Confirmation.** Confirmatory assessments examine how developer-conducted assessments were performed and whether their conclusions are well-supported. Unlike robustness assessments, which apply independent methods, confirmatory assessments evaluate the developer's existing framework to check for its accuracy, completeness, and soundness of reasoning. Confirmatory assessments may include replication testing, methodological reviews, and holistic reviews.

2. **Robustness.** Robustness assessments examine the claims related to the safety of a model through different methods than a developer has used. These types of open-ended assessments typically involve external parties applying their own evaluation methodologies – i.e., methods developed and selected at the discretion of the assessor – to test model capabilities and safety properties. Adversarial testing involves structured attempts to circumvent model safeguards and elicit harmful behaviors through techniques that malicious actors might employ. In both cases, third-parties apply a methodology that is different to test model safety, and can provide fresh perspectives on whether safety conclusions hold up under alternative evaluation strategies.

3. **Supplementation.** Third-party assessors can also extend the internal capabilities and capacity of a developer to perform internal assessments in cases where they need additional expertise or resources. These third-party assessments are less about evaluating claims related to the safety of a model, but rather about designing a bespoke assessment in specialized domains or increasing capacity during periods of intensive testing.

Developers may combine multiple functions to achieve comprehensive coverage. The appropriate mix depends on factors including the model's proximity to capability thresholds, deployment context, availability of internal expertise, and specific risks of concern.

Third-party assessment, when implemented in appropriate contexts, can bring valuable expertise to a rapidly evolving field, as well as important perspectives that complement internal safety work. We look forward to continued work on this issue. An ecosystem in which AI evaluation science, best practice, and tools advance in parallel to longer-term standards will not only help establish greater clarity and trust, but also support more agile implementation of risk management frameworks.

Section 1

Overview of Third-Party Assessments

ROADMAP

1.1 Purpose and Scope

Third-party assessments can be conducted on frontier models to confirm evaluations or claims on critical safety capabilities and mitigations. In appropriate contexts, these assessments may help to confirm or build confidence in safety claims, add robust methodological independence, and supplement expertise. This report outlines practices and approaches among FMF firms for implementing, where appropriate, rigorous, secure, and fit-for-purpose third-party assessments.

This report is the fourth in a series examining frontier AI frameworks. Previous reports addressed how developers identify and establish thresholds for extreme risks, conduct capability assessments to evaluate those risks, and implement mitigations to reduce them. Those reports primarily focused on internal assessment methodologies. This report focuses on third-party assessors, but does not cover broader AI auditing practices, general model performance evaluations, or regulatory compliance assessments.

Third-party assessments can complement – but do not replace – internal safety processes and corporate governance mechanisms. Internal teams possess deep knowledge of their systems and development processes, and external assessors can offer methodological independence, specialized expertise, and fresh perspectives that may be able to identify issues or validate critical safety claims. These assessments become particularly valuable for models that are approaching, or have reached, enabling capability thresholds, or when implementing novel frontier mitigations. Third-party assessment may be particularly helpful in testing involving highly specialized or classified areas, where model providers may not have the necessary data or expertise.

The third-party ecosystem for frontier models remains relatively nascent. As with other aspects of frontier AI safety and security, third-party assessment methodologies continue to evolve alongside advancing model capabilities, and few third-party organizations exist in the ecosystem to conduct rigorous, secure, and fit-for-purpose assessments. FMF member firms often differ in objectives for and approaches to third-party assessments. Similarly, third-party assessors often differ in how they structure their research designs and analyses. The lack of technical standards for frontier capability evaluations can lead to inconsistency in how assessors structure their research designs and analyses, making it difficult to compare results. Third-party assessments can help to confirm or build confidence in safety claims, but are not the only measure for doing so.

Where this report references developer practices, it primarily describes approaches used or considered by various FMF members and represents current and emerging practices rather than prescribed standards, especially as they do not exist at this time. We welcome expert feedback on this report and look forward to continued work on this issue.

1.2 Relationship to Frontier AI Frameworks

Frontier AI frameworks typically employ a two-stage risk assessment process carried out by internal assessment teams:

1. **Capability Assessments**: The process begins with frontier capability assessments that evaluate whether a model crosses any enabling capability thresholds or outcomes-based thresholds – related to severe harms, such as substantially assisting with CBRN weapons development. If and when a model crosses such a threshold, the framework triggers mitigation and risk evaluation requirements.
2. **Mitigation Assessments**: For models that cross enabling capability thresholds, developers apply appropriate frontier mitigations, such as the approaches described in this report. Developers then evaluate their effectiveness and whether the residual risk is acceptable for proceeding with further training or deployment. Whether a model meets this acceptable development or deployment threshold will depend on factors including the deployment context and the broader risk landscape.

Third-party assessments can complement and support both stages of this process. For example, for capability assessments, external evaluators can help confirm that internal testing methods are effective measures of relevant capabilities and risks, attest to the robustness of findings through different methodologies, or provide specialized expertise for domains where in-house knowledge may be limited. For mitigation effectiveness assessments, external evaluators may conduct further evaluations to determine whether some implemented safeguards function as intended, or conduct adversarial testing to identify potential vulnerabilities.

Developers engage third-party assessors at the times and depths most likely to produce meaningful test signals, calibrated to model characteristics, deployment context, assessor qualifications, and risk. As the external testing ecosystem matures, these engagements should coevolve with internal and vendor testing practices, preserving consistent safety levels across deployments.

1.3 Functions of Third-Party Assessments

Where third-party assessments are helpful, they can serve three primary functions in frontier AI frameworks: confirming internal work, strengthening the robustness of assessments, and supplementing expertise and

resources. Each function may address different needs and have distinct approaches, ranging from document review to hands-on technical testing.

1. **Confirmation.** Third-party assessments may be intended to confirm whether developer-conducted evaluations match internal requirements and that the results are supported by the evidence. For example, for capability assessments, a third party might confirm a "bottleneck assessment" by checking whether the tests effectively measure the capabilities experts identified as necessary for biological weapons development – ensuring the wet lab protocol questions represent the bottleneck and are appropriately calibrated. For mitigation effectiveness assessments, experts might review structured documentation of safety claims, such as examining whether the evidence presented adequately supports the conclusion that behavioral alignment training reduces harmful outputs to acceptable levels, and whether the reasoning connecting test results to deployment decisions is sound. Verification builds confidence that internal safety work meets its stated objectives.

2. **Robustness.** Robustness assessments examine the claims related to the safety of a model through different methods than a developer has used. These types of open-ended assessments typically involve external parties applying their own evaluation methodologies –that is, methods developed and selected at the discretion of the assessor – to test model capabilities and safety properties. Adversarial testing involves structured attempts to circumvent model safeguards and elicit harmful behaviors through techniques that malicious actors might employ. In both cases, third-parties apply a methodology that is different to test model safety, and can provide fresh perspectives on whether safety conclusions hold up under alternative evaluation strategies.

3. **Supplementation.** Third-party assessors can also extend the internal capabilities and capacity of a developer to perform internal assessments in cases where they need additional expertise or resources. These third-party assessments are less about evaluating claims related to the safety of a model, but rather about designing a bespoke assessment in specialized domains or increasing capacity during periods of intensive testing.

Developers may combine multiple functions to achieve comprehensive coverage. The appropriate mix depends on factors including the model's proximity to capability thresholds, deployment context, availability of internal expertise, and specific risks of concern. Targeted confirmation may be sufficient for models far from critical thresholds, while those approaching or exceeding thresholds might benefit from robustness evaluation alongside confirmation efforts.

Effective third-party assessments generally share some key characteristics regardless of function: independence of judgment, technical rigor, clear methodology, appropriate expertise for the risks being evaluated, and actionable recommendations that can meaningfully inform developer decision-making. Section 5 discusses the features of effective third-party assessments in more detail.

We acknowledge a substantial body of cross-industry standards and scholarship defining third-party testing practices (e.g., verification, validation), and accreditation may correspond to elements of the testing functions described here. This report does not attempt a formal crosswalk; our aim is to characterize the current testing ecosystem for frontier AI systems as it evolves. We plan to develop a follow-on publication to map the functions described in this report to those terms and standards.

Section 2

Confirmation

ROADMAP

2.1 Replication Testing

2.2 Methodology Review

2.3 Holistic Review

Confirmatory assessments examine how developer-conducted assessments were performed and whether their conclusions are well-supported. Unlike robustness assessments, which apply independent methods, confirmatory assessments evaluate the developer's existing framework to check for its accuracy, completeness, and soundness of reasoning.

Confirmatory assessments may evaluate whether results are reproducible rather than artifacts of specific testing conditions or implementation choices. They can examine whether methodologies measure what they claim to measure – for instance, whether a biological knowledge test captures dangerous capabilities rather than general scientific knowledge. They also assess whether the conclusions drawn from test results are supported by the evidence, identifying instances where safety claims may extend beyond what the data supports.

2.1 Replication Testing

Replication testing involves re-executing key evaluations to confirm that results are reproducible and accurate. This direct form of confirmation helps establish confidence that capability measurements and safety claims reflect genuine model properties rather than artifacts of specific testing conditions.

- **Implementation Approach**: The scope of replication varies with context. When resources permit and risks are high, assessors might re-run entire evaluation suites with identical prompts and scoring methods. More commonly, they focus on results that raise concerns – those that are surprising, safety-critical, or hover near decision thresholds. If a developer reports their model narrowly falls below a dangerous capability threshold for autonomous AI research, replicating those specific tests becomes helpful for confirmation.

- **Recommended Approaches**: Effective replication benefits from both qualified assessors and appropriate access. Assessors combine technical proficiency in AI architectures with domain expertise relevant to the risks being evaluated – a cybersecurity specialist for cyber risks, a biologist for CBRN concerns. They need methodological rigor to handle statistical complexities and must maintain independence from the development process. The availability of open-source models may also help assessors replicate testing. Replication testing may involve a near-identical model version and comprehensive documentation of the original procedures, which is often difficult to achieve, from exact prompts to scoring rubrics.

- **Practical Application**: In practice, replication often reveals subtle but important issues. Implementation errors in the original testing may emerge when external assessors follow the documented procedures. Models might show surprising sensitivity to minor prompt variations that the evaluating team did not explore, or scoring methods might introduce unrecognized biases. Environmental factors like temperature settings can also impact results in ways that only become apparent through replication.

The value of replication can extend beyond confirming numbers. It can establish that developer claims hold up under further scrutiny, complementing an internal team's findings, and build stakeholder confidence. Yet there are also limitations to replication. Comprehensive replication demands substantial time and computational resources that impose real costs, and perfect reproducibility may prove to be infeasible if techniques reveal sensitive information or capabilities that rely on company-sensitive IP, or technical infrastructure that cannot securely be shared. The inherent nondeterminism of frontier models also introduces unavoidable variability.

Despite these limitations, replication testing, where feasible, can provide additional evidence for safety claims.

2.2 Methodology Review

Methodology review examines whether evaluation designs actually measure what they claim to measure and whether the chosen approaches provide meaningful evidence about model capabilities. Rather than re-running tests, this form of confirmation analyzes the conceptual soundness and technical rigor of evaluation strategies. This analytical approach complements empirical replication by identifying flaws in evaluation design that might not be apparent from results alone.

- **Implementation Approach:** The review process focuses on three core areas. For one, assessors evaluate construct validity – whether tests genuinely measure capabilities that lead to harmful outcomes. For example, reviewing whether biological weapons assessments that test pathogen knowledge sufficiently measure bioweapon creation abilities or merely reflect general microbiology knowledge. In addition, they examine evaluation coverage, identifying test categories, particular workflows, or user skill levels. For another, assessors scrutinize technical implementation choices, from scoring methods to statistical analyses, ensuring these support valid conclusions about model capabilities.

- **Recommended Approaches:** Methodology review requires documentation of evaluation designs, which may include test specifications, theoretical justifications, scoring rubrics, and statistical

methodologies. Assessors need deep expertise in both evaluation science and the specific risk domains being tested – understanding what constitutes a valid measure of dangerous capabilities requires both methodological sophistication and domain knowledge. Reviews are most valuable early in the evaluation process when designs can still be refined.

- **Practical Application:** Reviewers systematically examine multiple aspects of evaluation validity. For example, they assess whether "capture the flag" cybersecurity exercises test real offensive capabilities or merely puzzle-solving skills. They identify when evaluations fail to test multi-step planning in addition to factual knowledge, or when tests assume certain user skill levels without variation. They evaluate whether difficulty levels appropriately distinguish between current capabilities and dangerous thresholds – catching both tests that are too easy to provide warning and those too difficult to detect accessible risks.

Methodology review may result in specific, actionable recommendations for strengthening evaluations – from additional test scenarios that close coverage gaps to improved scoring rubrics that reduce subjectivity. By working within the developer's framework rather than proposing new approaches, reviewers can identify where methods systematically over- or under-estimate risks. However, this approach cannot catch implementation errors or runtime issues that only emerge during actual testing.

2.3 Holistic Review

Holistic reviews, which may include different forms of safety case analysis, document how developers reason from evaluation results to safety conclusions. It is often the final stage in a risk assessment process. This form of confirmatory assessment examines the logical structure and evidentiary support for claims about model safety, assessing whether conclusions are justified by the available evidence. Unlike replication or methodology review, this approach focuses on the argumentation connecting data to deployment decisions, ensuring that safety reasoning is sound and proportionate to risks.

- **Implementation Approach:** Holistic reviews examine the complete reasoning chain from evidence to safety conclusions. Reviewers analyze documentation that includes specific safety claims (such as "this model cannot meaningfully assist with biological weapons development"), supporting evidence from evaluations, underlying assumptions, and conditions under which claims hold valid. The review process systematically traces logical connections, identifying weak inferences, unstated assumptions, and incomplete arguments. Reviews prove most effective when conducted with sufficient time for thorough analysis and when integrated with iterative feedback cycles between reviewers and developers. This ongoing dialogue enables clarification of complex points and collaborative exploration of identified issues.

- **Recommended Approaches:** Effective holistic reviews require comprehensive documentation including risk identification and analysis materials, system documentation explaining assessment methodologies, testing evidence supporting safety claims, and decision frameworks revealing thresholds and acceptance criteria. Documentation must balance completeness with security considerations – particularly for mitigation systems where detailed information could enable exploitation. Assessors need domain expertise relevant to the risks under review, such as CBRN knowledge for biosecurity assessments or cybersecurity expertise for cyber defense reviews. They must also possess experience with structured safety argumentation, systems thinking to identify

emergent vulnerabilities, sufficient independence to challenge embedded assumptions, and the ability to anticipate potential exploitation pathways.

- **Practical Application:** Holistic reviews examine reasoning across multiple contexts. For example, in mitigation reviews, analysis focuses on whether risk assessments comprehensively address threat scenarios, whether defense-in-depth implementations truly prevent single points of failure, and whether protection claims align with supporting evidence. Reviewers identify common reasoning flaws: treating absence of evidence as evidence of absence, selective use of supportive data while dismissing contradictory findings, circular arguments where conclusions depend on unvalidated assumptions, and overconfidence in adversary limitations when simple techniques might bypass safeguards.

Holistic reviews can provide additional support that deployment decisions rest on sound reasoning. The approach identifies logical gaps and flawed arguments that empirical testing alone might miss. The qualitative nature of logical analysis introduces variability based on reviewer expertise and analytical frameworks. Documentation represents intended designs rather than operational implementations, limiting visibility into real-world performance under stress. Reviews provide point-in-time assessments that may not capture system evolution or degradation over time. Without hands-on testing capabilities, reviewers cannot achieve the depth of dedicated adversarial exercises in identifying novel exploitation paths. Despite these constraints, safety case analysis can provide logical rigor that complements empirical validation approaches.

Section 3

Robustness

ROADMAP

3.1 Open-Ended Assessment

3.2 Adversarial Testing

Third-party assessments may also aim to strengthen the robustness of evidence about the safety of a model. Robustness assessments examine the claims related to the safety of a model by applying different, independent methodologies to test model safety, and can provide fresh perspectives on whether safety conclusions hold up under alternative evaluation strategies.

External assessors can also bring different methodological and measurement traditions—perhaps adapted from other fields like cybersecurity or safety engineering for other technologies and industries—that can explore novel failure modes and illuminate risks not immediately available through standard AI evaluation lenses.

3.1 Open-Ended Assessment

Open-ended assessment typically involves external parties applying their own evaluation methodologies – i.e, methods developed and selected at the discretion of the assessor–to test model capabilities and safety properties. It can also involve internal assessors who are sufficiently independent to internal teams running capability assessment and mitigations and bring significant domain expertise in systemic risks. Assessors bring different approaches—techniques developed through specialized research or adapted from other domains—to evaluate whether models pose risks. This methodological independence provides an external perspective that can surface risks invisible to internal evaluation frameworks.

- **Implementation Approach:** Assessments that are designed and led by a third-party can provide methodological diversity. Organizations like METR (Model Evaluation and Threat Research) have developed specialized frameworks for evaluating autonomous AI research capabilities, using task suites and evaluation criteria distinct from those typically employed by developers. Apollo Research can offer

techniques focused on detecting scheming behavior, applying methods informed by interpretability research. These assessments often probe different aspects of model behavior—examining how models respond to ambiguous instructions, handle long-term planning scenarios, or exhibit patterns across seemingly unrelated tasks.

- **Recommended Approaches:** Assessments that are designed and led by a third-party also benefit from both qualified assessors and access that is relevant to the assessment type. Assessors need strong capability elicitation skills, domain-specific knowledge relevant to risks being evaluated, expertise in rigorous AI evaluation methodologies, and experience with threat modeling for advanced capabilities. A representative model version and existing documentation may be sufficient for this approach. A version with limited safety training to test underlying capabilities, and/or additional documentation of known model capabilities, may potentially be helpful for an assessment. External assessors should also provide sufficient security guarantees, particularly in cases where model developers may choose to give access to models with fewer safety mitigations than release versions.

- **Practical Application:** Open-ended assessments test different hypotheses and probe different failure modes than internal evaluations. These assessments prove most valuable when conducted on representative models either preparing for deployment or already deployed, ensuring findings reflect real-world capabilities rather than development artifacts. External assessments can also explore new ground rather than inadvertently replicating known results.

Open-ended assessment can provide multiple perspectives on capability elicitation, enabling more robust risk coverage. When multiple independent parties reach similar conclusions through different methods, it strengthens confidence in safety properties. Conversely, revealing previously unidentified risks can prompt important safety strategy revisions. The approach offers independent benchmarks that support and, where appropriate, enable the possibility of independent internal assessments.

However, significant limitations exist. External evaluators' methodologies vary in quality and comprehensiveness, requiring careful assessor selection. Experts who combine rigorous capability elicitation skills with deep domain knowledge remain scarce. Assessments designed and led by third-parties also often require substantial time and resources, particularly when developing new evaluation techniques. Despite these constraints, open-ended assessments are valuable for providing fresh perspectives that enhance overall safety assurance through methodological diversity.

3.2 Adversarial Testing

Adversarial testing involves structured attempts to circumvent model safeguards and elicit harmful behaviors through techniques that malicious actors might employ. This approach stress-tests safety measures under realistic attack conditions, revealing vulnerabilities that prescribed testing based on a priori agreement might miss. Unlike other evaluation approaches that assess normal operation, adversarial testing is designed to uncover weaknesses and failure modes that emerge under deliberate attack, which can provide useful insights into the resilience of the model in practical scenarios. As discussed in other FMF reports, model developers may engage in their own adversarial testing relying on in-house expertise, and third party testing can play a complementary role in this work.

- **Implementation Approach:** The adversarial mindset fundamentally distinguishes this approach – testers actively search for ways to break safety measures rather than verify their presence. They employ diverse techniques including crafting jailbreaks that bypass safety training, developing multi-turn conversations that gradually erode safety boundaries, identifying edge cases where safety policies conflict or lack clarity, and exploiting gaps between trained refusals and request reformulations.

- **Recommended Approaches:** Adversarial testing requires specialized expertise and access that is relevant to the assessment type. Testers need experience with red teaming techniques and adversarial methodologies, combined with domain-specific knowledge to identify circumvention mechanisms requiring specialized understanding (such as alternative notations for chemical compounds in CBRN contexts). Appropriate access may include a model with representative safeguards in place to enable realistic stress testing, documentation of known vulnerabilities to avoid redundant effort, and clear policy taxonomies defining acceptable and unacceptable content boundaries. Testing can also be valuable when conducted on models with proposed safeguards, either pre-deployment or on deployed systems, ensuring findings reflect real-world defensive capabilities.

- **Practical Application:** The scope and intensity of adversarial testing adapts to threat models and deployment contexts. For limited-distribution models, testing emphasizes sophisticated adversaries who might invest significant effort in circumvention. For widely deployed models, testing must also consider what unsophisticated users might achieve through persistence. Testing reveals useful insights: behavioral training often creates only shallow modifications with harmful capabilities remaining accessible, safeguards that block direct requests may fail against indirect approaches, and motivated adversaries can achieve more than anticipated. These findings inform whether additional mitigations are valuable or certain deployments pose significant risks.

Adversarial testing provides a unique assessment of residual risks and mitigation gaps from an attacker's perspective. Domain experts can identify niche adversarial strategies that predetermined testing might miss, providing concrete evidence about realistic adversary capabilities. This informs critical deployment decisions and mitigation priorities. However, the approach faces significant constraints. Manual testing methods require substantial time and resources. Simulated conditions cannot perfectly replicate all real-world scenarios, potentially missing some attack vectors. Outcomes depend heavily on tester skill, creativity, and comprehensiveness—even expert teams might miss novel exploitation methods. Additionally, the most dangerous attack strategies might be discovered only after deployment when diverse adversaries probe the system. Despite these limitations, adversarial testing remains essential for understanding how safeguards fail in practice and what capabilities become accessible to motivated attackers.

Section 4

Supplementation

Third-party assessors can also play a role in extending internal testing capacity by contributing specialized expertise and novel methodologies to developers' internal teams. Their involvement is not typically about examining an internal assessment, but about designing and implementing tailored evaluations for internal use based on their unique perspectives.

- **Specialized domain expertise** represents the most common form of supplementation. External experts bring deep knowledge of specific risk areas that may be outside the core capabilities of AI developers. For example, biosecurity specialists can surface subtle dual-use risks in biological applications that may not be apparent to AI researchers, while experienced offensive security professionals can assess whether cyber capabilities would realistically enable real-world attacks. This expertise, especially when provided by leading experts in their field, is particularly important for designing robust, context-specific threat models and evaluation frameworks.

- **Methodological diversity** offers another form of supplementation. Third-party experts may bring to internal assessments evaluation techniques developed in other contexts, such as security assessments from traditional software, safety analysis approaches from high-risk industries, or domain-specific testing protocols. Especially since the science of evaluations is still nascent, supplementing internal assessments with the methodological diversity afforded by third-party expertise may help increase overall confidence in internal assessments.

The effectiveness of supplementary capacity depends on smooth integration with internal teams and significant credibility in the specialized field. This requires clear communication channels, well-defined scopes of work, and mechanisms for external experts to understand the AI system's context while maintaining appropriate confidentiality and security controls. When implemented effectively, supplementation enables developers to access valuable expertise that can help strengthen existing risk discovery and mitigation.

Section 5

Key Elements of Third-Party Assessments

Third-party assessment can be effective when they follow several principles that guide implementation:

- **Methodological rigor** ensures assessment credibility through clear documentation of evaluation approaches, comprehensive results reporting, and thoughtful analysis connecting findings to safety implications. High-quality assessments maintain transparency about limitations and uncertainties while providing substantive evidence. As discussed in other FMF reports, methodological rigor for evaluation approaches continues to evolve.

- **Domain expertise** appropriate to the risks being evaluated, whether specialized knowledge in biosecurity for CBRN assessments or deep technical understanding for evaluating autonomous capabilities.

- **Appropriate access** balances information needs with security considerations. Effective assessments require relevant documentation, test results, and sometimes controlled model interaction, while providing only the minimum information necessary and protecting sensitive intellectual property such as training data and model weights. The type and level of access vary by function – confirmation may need model access to replicate tests alongside documentation of methodologies and results, while robustness typically requires more extensive hands-on interaction to conduct independent evaluations using novel approaches.

- **Security readiness** ensures assessors adhere to strict data handling protocols, maintain confidentiality agreements, and implement robust access controls that protect sensitive information encountered during the assessment process. This principle recognizes that third-party assessors must demonstrate their capacity to responsibly handle privileged access before receiving it.

Beyond these fundamentals, different assessment functions benefit from additional elements:

- **For confirmation**: Strong analytical capabilities to evaluate whether internal methodologies meet their stated objectives and whether evidence appropriately supports safety claims. This includes the ability to clearly communicate findings and their implications. This enables assessors to identify gaps in reasoning or instances where conclusions extend beyond what the data supports, ultimately strengthening confidence in safety claims or highlighting where additional work may be helpful.

- **For robustness**: Methodological independence enables assessors to approach evaluations with a fresh perspective. This separation allows external reviewers to challenge implicit assumptions and identify potential gaps that internal teams might overlook. Equally important is the ability to conduct comprehensive assessments that capture the full spectrum of model capabilities, then translate these findings into nuanced, actionable insights. Rather than simply identifying individual failures, effective robustness evaluation synthesizes results across different tests to provide an overall assessment of safety.

- **For supplementation**: Specialized expertise not available internally (such as deep domain knowledge in biosecurity or specific attack methodologies). This enables developers to access specialized expertise when needed without maintaining large standing teams across every possible risk domain.

Section 6

Continuing Work

REQUEST FOR COMMENT

We welcome engagement with this report and our broader technical report series on frontier AI frameworks.

Researchers and organizations interested in further refining and harmonizing the implementation of frontier AI frameworks are invited to reach out to the Frontier Model Forum.

Please offer feedback at: info@frontiermodelforum.org

Third-party assessments for frontier AI safety brings valuable expertise and perspectives to a rapidly evolving field. While the ecosystem is still developing – with assessment methodologies, qualification/accreditation standards, and access frameworks in early stages – external assessors already provide important perspectives that complement internal safety work. Many bring deep domain expertise in areas like biosecurity or cybersecurity and offer the scientific independence that is helpful for credible third-party assessments.

As models become more capable and deployment contexts become more complex, the importance of robust external assessment may increase. Building a mature third-party assessment ecosystem represents a helpful component of comprehensive AI safety governance. Key areas for continued development include:

- **Assessment Methodologies**: Standardizing how assessments are conducted, documented, and reported to enable meaningful comparisons across providers and ensure consistent quality. This includes developing shared evaluation frameworks and benchmarks that can be applied across different models and contexts.
- **Access and Infrastructure**: Establishing frameworks that balance assessor access needs with security requirements, including appropriate model access levels, secure testing environments for dangerous capabilities, and computational resources for comprehensive evaluation. Clear guidelines on what access is appropriate for different assessment types would enable more meaningful external validation while protecting sensitive systems and information.
- **Ecosystem Coordination and Sustainability**: Creating mechanisms for effective collaboration between developers,

assessors, and policymakers, including sustainable funding models for assessment services and protocols for sharing findings while protecting commercially-sensitive and proprietary information. Without clear economic models and coordination structures, the assessment ecosystem cannot develop the depth and scale required.

The development of this ecosystem requires sustained effort from multiple stakeholders. Developers can support growth by engaging constructively with third-party assessors and providing appropriate access. Assessment providers can build specialized expertise while maintaining independence. An ecosystem in which AI evaluation science, best practice, and tools advance in parallel to longer-term standards will not only help establish greater clarity and trust but also support more agile implementation of risk management frameworks.