ISSUE BRIEF

FRONTIER MODEL FORUM

Preliminary Reporting Tiers for AI Bio Safety Evaluations

DATE

March 18, 2025

Frontier AI models and systems are particularly promising for advancing medicine and public health. At the same time, their knowledge of biology and ability to reason about biological concepts may also be misused in ways that pose significant risks to public safety and security. To manage those risks, many frontier AI developers have published safety frameworks in which they commit to evaluating their systems for biological knowledge and capabilities that could pose large-scale risks to public health and safety. By providing a structured way to measure those capabilities, AI-Bio safety evaluations – which include uplift studies, red-teaming exercises, and benchmark evaluations – have become a critical tool for ensuring that frontier AI systems are developed and deployed responsibly.

Given how rapidly the biological capabilities of frontier AI are advancing,[1] establishing norms and best practices for sharing information from AI-Bio safety evaluations is essential. Yet disclosing this information involves complex tradeoffs. Sharing more information from an evaluation can increase public trust in the legitimacy, credibility, and validity of its results. In addition, greater transparency about an evaluation better positions other researchers to replicate and compare studies, assess the efficacy of different methodologies, and advance the field overall. However, sharing information about AI-bio evaluations may also introduce or exacerbate information and attention hazards: by raising awareness about how AI might be misused to create or deploy biological threats, reporting on AI-bio safety evaluations may unwittingly increase the very risks they seek to better understand.

Responsibly balancing the benefits of greater disclosure and reporting against potential information and attention hazards is a difficult challenge. This issue brief outlines a three-tiered approach to responsible reporting that aims to attend to both the benefits of greater transparency and the potential risks associated with information and attention hazards. Drawn from expert discussions held by the Frontier Model Forum (FMF), the approach reflects preliminary thinking across FMF member firms about what information from safety evaluations should be shared with the public at large, what should be disclosed within trusted expert networks only, and what should be kept private.

A TIERED APPROACH TO REPORTING INFORMATION FROM AI-BIO EVALUATIONS

This brief proposes a tiered approach for sharing AI-bio safety evaluation information. The tiers are structured in a fundamentally precautionary way, such that information is only shared more widely when there is high confidence that greater circulation will not lead to information or attention hazards. Each tier corresponds to a distinct level of disclosure, with separate guidelines for what information is appropriate to share within the tier. The guidelines were developed in discussion with relevant domain experts and may change in the future as further research on information and attention hazards is conducted or as model capabilities advance.[2]

Notably, the guidelines below offer recommendations that are general rather than absolute. For example, while the guidelines recommend that high-level findings should be published in general, a high-level finding that confirms a specific novel bio threat or exploitable model vulnerability may be too sensitive to disclose publicly. As such, the guidelines also highlight the categories of information for which the reporting tier may be results-dependent.[3] We welcome further discussion and engagement on this approach and classification.

Tier	Information from Evaluation
Tier 1: Public Disclosure	Research Question
	Scientific Domain
	High-Level Evaluation Methods
	Public Benchmark Evaluation Methods
	Participant Profile
	Models Tested*
	Task Examples*
	Analysis Methodology*
	Key High-Level Findings*
	Public Benchmark Results
	Study Limitations
Tier 2: Disclosure Among Trusted Networks	Fully Specified Threat Models*
	In-Depth Evaluation Methods
	Detailed Analysis & Interpretation of Results*
	High-Level Model Enhancements*[4]
Tier 3: Private Disclosure	Specific Model Vulnerabilities*
	Training Data
	Private Datasets

**The reporting tier may differ based on the results of the evaluation.*

TIER 1: PUBLIC DISCLOSURE

Tier 1 consists of information from AI-Bio safety evaluations that can be published safely and responsibly (e.g., in a research paper). Publicly disclosing this information would facilitate greater scientific understanding and coordination of evaluations, while posing minimal information or attention hazards.

Public reports on safety evaluations may include:

- **Research Question.** Public reports should clearly state the research question that the evaluation aims to investigate.[5] This information enables the research community to better understand the scope of the evaluation without revealing sensitive details.
- **Scientific Domain.** Public reports should include the general biological domain being evaluated (e.g., biology) and may mention the sub-domain (e.g., epidemiology, virology) at a high-level.[6] This information allows for further context in interpreting results without revealing harmful information.
- **High-Level Evaluation Methods.** Public reports should include high-level information about the research design of the evaluation, such as whether it consisted of an automated benchmark, red-team exercise, or controlled trial. They should also provide high-level detail on the content of the evaluation method (e.g., benchmarks assessing biological protocol troubleshooting). Reports of benchmark evaluations should include one example question from the set used for the evaluation, though redacting potentially hazardous information (e.g., the specific virus in the question). Evaluations using publicly-available benchmarks should report the methods used in full. By disclosing the primary methods used for an evaluation, reports both enable external experts to better assess the results and facilitate greater trust in the findings among the broader public.
- **Participant Profiles.** For evaluations involving human participants (e.g., controlled studies or red-teaming), reports should include information about the demographics of study participants. For example, they should highlight participants' expertise in biology and the extent of their training or experience using the AI model or system being tested. This context is important for assessing the validity of the study results and understanding potential limitations or biases in the findings.[7]
- **Models Tested.** Reports should specify which AI models were evaluated in the study. This information is important for other researchers and model developers to contextualize the results of the evaluation and potentially replicate or extend the evaluation methodology. However, reports should not draw undue attention to especially vulnerable models (e.g., identifying in particular those which are especially good at producing harmful information and have inadequate safeguards). The specific model identity should be revealed if the evaluation results do not provide insights that may be used to significantly increase biorisks, or if the model is already publicly available. Though this may increase information hazards, it provides an opportunity for external security experts to mitigate the issue.

- **Task Examples.** Example tasks or questions used to query models should be shared publicly, provided the examples do not exacerbate information hazards. Example tasks should also be properly framed to indicate the low risk-risk nature of the designed example. Example tasks that specify hazardous information should be shared in Tiers 2 or 3 only. The goal of sharing these example tasks is to increase the level of trust, visibility, and engagement in the evaluation, as well as making the results more tangible. This type of information may also be useful for the community of model evaluators and other model developers looking to run high-quality evaluations.
- **Analysis Methodology.** Evaluation reports should explain the methodology used to calculate quantitative results with enough detail such that the results can be contextualized and critically assessed. Reporting the analysis and findings of an evaluation allows the scientific community to better evaluate the credibility of its results and even extend or build off them. However, details of the analysis methodology that specify misuse-relevant capabilities should be shared only among trusted networks (Tier 2). Further, if the evaluation results point to an exploitable model vulnerability, findings should not be disclosed publicly until fixed.
- **High-Level Key Findings.** To increase the trust and legitimacy of safety claims made by evaluators, reports should summarize the evaluation results derived from both public and private datasets at a high level. For example, reports should indicate the accuracy of the models on benchmarks, as well as general risk levels associated with the results (e.g., medium, high, critical risk, both pre- and post-mitigation). Evaluation results using public benchmarks should be reported in full. When using methodologies involving human participation, reports should state the human baseline or participant performance as this context is important for accurate interpretation of results and likely would not exacerbate risks. In general, evaluators should take caution when reporting the findings of evaluations, as safety research may unexpectedly identify exploitable vulnerabilities. Teams should consider disclosing findings based on the outcome of the evaluation and the severity of risk.
- **Study Limitations**. Reports should acknowledge limitations in their methodology and findings. For example, reports should note any concerns related to the generalizability of the methodology and results, the statistical power of the study, biases in the data or findings, or other limitations. Benchmark studies should also reference the benchmark load (i.e., whether it is a single question or a series of questions). A discussion of study limitations helps prevent misinterpretation of findings and enables continued refinement of evaluation approaches, particularly if they are also coupled with recommended improvements for future work.

TIER 2: DISCLOSURE AMONG TRUSTED NETWORKS

Tier 2 includes information from AI-Bio safety evaluations that should be shared only among trusted networks, such as relevant government agencies, trusted members of the research community, other AI model developers, key vendors, or third-party facilitators. While the information in this tier is important for advancing safety research, it is best shared within trusted expert networks rather than public channels due to potential information or attention hazards.

This tier would include:

- **Fully Specified Threat Models.** Fully specified threat models should be shared only among trusted contacts. This enables detailed technical discussion among key experts while managing the potential risks from broader disclosure. Even among trusted networks, the level of transparency on threat models should be tied to the potential severity of the threat. For example, for potential threats that would be very high-magnitude, fully specified threat models should remain mostly private.
- **In-Depth Evaluation Methodologies.** Reports shared among trusted networks should include more detailed information about the evaluation's research design and methodology, especially when innovative or experimental techniques are used. These reports may include information about methods, datasets, or assets that are too sensitive (e.g., datasets with hazardous knowledge) or proprietary (i.e., bespoke benchmarks) to release publicly. Most publicly available evaluations (e.g., public benchmarks) benefit from public discussion of their methodologies and may be disclosed publicly; others, such as uplift studies, should remain among trusted networks, or kept private.
- **Detailed Analysis and Interpretation of Results.** A detailed interpretation of proxy evaluations for harmful outcomes, risk interpretation of evaluation results, and implications for safety research should be shared among trusted networks only. This level of analysis helps build the research literature and enables expert feedback for improving evaluation methodologies. Depending on the hazards involved with the results of the evaluation, some detail may be shared at a high level publicly. However, public disclosure of specifically hazardous results information may lead to increased information and attention hazards.
- **Details of Model Enhancements.** Model enhancements are designed to augment model capabilities and may include methods such as scaffolding or fine-tuning. The disclosure of model enhancements provides important context for interpreting model performance and may improve coordination by facilitating the reproduction of evaluation results.[8] However, public disclosure may enable malicious actors to better misuse models by drawing attention to potentially dangerous capabilities.

TIER 3: PRIVATE DISCLOSURE

Tier 3 includes types of information which should not be shared even among trusted networks, since the risks would outweigh the benefits. Private disclosure may include parties who are responsible for safeguarding against the risks of the model or system, such as the model developer or, when appropriate, relevant government agencies.

This tier would include:

- **Specific Model Vulnerabilities.** Evaluation results may sometimes reveal exploitable model vulnerabilities, such as a propensity to output hazardous information, that may be used by threat actors to cause harm. Due to the potential hazards of publicly sharing details on how to exploit a specific model through a specific method, this information should be shared privately among stakeholders for whom the information is decision relevant. This may include the model developer, deployer, evaluator, or relevant national security groups. This information may be shared more widely once the risk is mitigated by patching the vulnerability, or after a preset amount of time (e.g., 12 months). Conversely, information indicating that a model can be exploited that does not reveal the specific model or method may be disclosed in Tier 2.
- **Specifics on Training Data.** The specific training data used to make the model or system better at a bio-specific benchmark should remain private due to misuse risks, proprietary concerns, and attention hazards associated with public or even semi-public sharing.[9]
- **Private Datasets.** Private or proprietary datasets excluding training data that can be used to facilitate evaluations should be shared privately only with model developers, evaluators, or national security stakeholders for whom the information is decision relevant.[10]

RESULTS-DEPENDENT INFORMATION

The guidelines above offer general recommendations for categories of information based on the potential for information hazards. However, the extent to which certain types of information are disclosed may depend on the results of the evaluation. For example, while publishing information stating which models were tested is good practice, if the evaluation results indicate that a particular model has an exploitable vulnerability that may lead to a significant increase in biorisks, this information should not be published.

Some experts argue that the baseline level of transparency for all categories should be higher when making an ‘inability safety claim’ (i.e., that a model does not have a certain capability) to support a safety case. This is because when making such a claim, the evaluation results suggest that the evaluator has already vetted for the absence of the capability which may give rise to the potential for harm. Additionally, higher transparency is important to allow for better external review and assessment of the safety claim. Conversely, if the results indicate that a capability threshold has been crossed, researchers may choose to keep the majority of potentially-harmful information private, while only sharing high-level information publicly (e.g., overall risk level).

CONCLUSION

As frontier AI capabilities in the biological domain continue to advance, clear reporting guidelines are essential for enabling safety research while managing the potential risks. The approach outlined here reflects an preliminary emerging expert consensus on how best to balance transparency with safety considerations through a tiered disclosure approach. The FMF aims to further refine these standards in collaboration with the broader research ecosystem.

These guidelines will require ongoing discussion and adaptation as evaluation methods evolve and new challenges emerge. Regular review and updates will help ensure the standards remain effective at supporting critical safety work while appropriately managing information hazards. Through continued collaboration across the research community, these standards can provide a robust framework for responsible sharing of AI-bio safety evaluation results.

FOOTNOTES

[1] See Epoch's AI Announcement

[2] While we note caveats throughout, there may still be disagreement about the general categorization we have chosen for each type of information.

[3] These are marked with an asterisk(*).

[4] There is some disagreement among experts about the ideal level of transparency for reporting model enhancements (also referred to as elicitation techniques). See description below for further detail.

[5] To the extent that it does not reveal potentially hazardous information about the threat model.

[6] Based on the results of the evaluation, groups may wish to keep the specific sub-domain private to reduce the risk of information or attention hazards.

[7] Depending on the evaluation results, some groups may choose to keep this at a high level to reduce the risk of attention hazards.

[8] While many experts argue these should be disclosed publicly, there was not sufficient consensus on public disclosure for inclusion in Tier 1. Reporting at higher levels of abstraction may enable sharing this information publicly while managing the associated risks. For example, it may be appropriate to publicly share the estimates of the level of effort used to enhance the model (e.g., number of hours spent), or high-level principles (e.g., tool-use). Further detail should be shared only among trusted networks or kept private.

[9] AI developers may wish to publicly disclose that they have excluded certain training data (e.g., virology).

[10] In some cases, there may be value in sharing this information with trusted networks listed in Tier 2. For example, model evaluators, including those housed within AI developers, may consider sharing proprietary evaluation datasets with other actors through trusted networks (e.g., the FMF) to facilitate safety assessments.