THE CENTRE FOR
LONG-TERM RESILIENCE

Response: National Institute of Standards and Technology's Safety Considerations for Chemical and/or Biological AI Models

PREPARED BY

Richard Moulange & Cassidy Nelson

The Centre for Long-Term Resilience

December 2024

For: U.S. Artificial Intelligence Safety Institute

National Institute of Standards and Technology

Docket: 240920-0247

THE CENTRE FOR
LONG-TERM RESILIENCE

Background

The Centre for Long-Term Resilience (CLTR) is an independent nonprofit think tank with a mission to transform global resilience to extreme risks: high-impact threats with global reach. We do this by working with governments and other institutions to understand the risks and improve relevant governance, processes and decision-making. We focus on three main areas: the safe development and use of AI; biosecurity threats and pandemic preparedness, and improving how governments manage extreme risks.

CLTR has published on the dual-use implications of AI in the life sciences, including the responsible innovation of chem-bio AI models. We are motivated to work in this area because of the widespread potential beneficial applications of these tools, which necessitates their safeguarding, and the currently low understanding of their risk profile. Our work has focused on evaluating the capabilities of these models, proposing risk mitigation strategies, and advancing assessment frameworks. Below, we draw on our previous work and recent insights to address the questions below from the National Institute of Standards and Technology (NIST)'s Request for Information (RFI), focusing on information sharing and concrete recommendations to inform the US AI Safety Institute's approach to biosecurity evaluations and mitigations.

1. Current and/or Possible Future Approaches for Assessing Dual-Use Capabilities and Risks of Chem-Bio AI Models

a. What current and possible future evaluation methodologies, evaluation tools, and benchmarks exist for assessing the dual-use capabilities and risks of chem-bio AI models?

Existing evaluation methods

There are four current types of methods for evaluating the dual-use capabilities and risks of AI models, including, but not limited to, chem-bio AI models. These have all been used to assess large language models (LLMs) for chem-bio risks, but only the first listed (non-interactive risk assessment) has been publicly reported to be used for other chem-bio AI models, which includes AI-enabled biological tools such as biological design tools (BDTs) and biological foundation models (bioFMs).

The four types are:

I. **Non-interactive risk assessment:** theoretical analyses of the risks and capabilities of models that do not require interacting with the model during the assessment.

II. **Evaluations ("evals"):** simple, interactive, reproducible technical assessments that can be automated to compare one or more models against a standardised set of tasks to

produce a score. We would suggest that, while often used for different purposes, automated evaluations and "benchmarks" can be considered synonymous.

III. **Uplift studies:** non-automated, interactive use of a model by users with varying levels of expertise who aim to complete one or more standardised tasks. Usually, one arm of the study would have access only to existing internet search engines to serve as a baseline, while other arm(s) would have access to use an AI model.

IV. **Red-teaming:** open-ended, non-automated, interactive use of a model, usually by experts, some of whom may have relevant security experience that is designed to surface more complex capabilities which might not be uncovered by automated evaluations, and explicitly defeat risk mitigation systems.

A methodology has been developed by CLTR for non-interactive risk assessment (Moulange et al., 2024) and is currently being built upon in a project collaboration with RAND (RAND & CLTR, 2024). This methodology centres around assessing the capability of a functional category of chem-bio AI models using expert reviews of the literature and industrial applications and the application of prespecified risk criteria. This allows for a rapid assessment of capabilities and risk without requiring *in silico* or wet-lab experiments.

Note that from the above list, uplift studies and red-teaming can be completed *in silico* (so that the model users attempt only to produce dangerous knowledge digitally and do no real-world work) or can include physical interactions, which would often be in a wet-lab environment. Caution is strongly advised when considering wet-lab validation of evaluation outputs (Moulange et al., 2024). We recommend a stepwise approach where wet-lab experiments are only considered in some instances and conducted exclusively on proxy agents with efforts taken to stop the generation and spread of dangerous information in the process.

There have been evaluations developed for chemical and biological risks posed by frontier large language models (LLMs). The table below was adapted from a forthcoming Frontier Model Forum publication and consists of public resources that document or reference AI safety evaluations of bio risks.[1]

Year	Author	Name	Method	URL
2024	Anthropic	Claude 3.5 Sonnet Model Card	Benchmark Evals; Uplift Study	Model Card
2024	Future House	LAB-Bench: Measuring Capabilities of Language Models for Biology Research	Benchmark Evals	Paper

[1] Frontier Model Forum, "Frontier AI-Bio Safety Evaluations" (forthcoming)

2024	GDM	Gemini 1.5 Pro model card	Benchmark Evals	Model Card
2024	GDM	Evaluating Frontier Models for Dangerous Capabilities (preliminary bio evals)	Benchmark Evals; Red-team Evals	Paper
2024	Ivanov	BioLP-bench: Measuring understanding of biological lab protocols by large language models	Benchmark Evals	Paper
2024	Li et al.	The WMDP Benchmark	Benchmark Evals	Paper
2024	Meta	Llama 3.1 Model Card	Uplift Study	Model Card
2024	OpenAI	Building an early warning system for LLM-aided biological threat creation	Uplift Study	Post
2024	OpenAI	O1 System Card Bio Threat Creation Evaluations	Benchmark Evals; Red-team Evals	System Card
2024	OpenAI	GPT-4o System Card	Benchmark Evals; Uplift Study	System Card
2024	RAND	The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study	Red-team Evals; Uplift Study	Paper
2024	SecureBio	Lab Assistance Benchmark – Multimodal	Benchmark Evals	Post
2024	UK AISI	Advanced AI evaluations at AISI	Benchmark Evals	Post
2023	Gopal et al.	Will releasing the weights of future large language models grant widespread access to pandemic agents?	Red-team Evals	Paper
2023	OpenAI	GPT-4 System Card	Red-team Evals	System Card
2023	Sarwal et al.	BioLLMBench: A Comprehensive Benchmarking of Large Language Models in Bioinformatics	Benchmark Evals	Paper

2024	Anthropic	Responsible Scaling Policy Evaluations Report – Claude 3 Opus	Red-team Evals	Evaluations Report
2024	UK AISI & US AISI	Pre-Deployment Evaluation of Anthropic's Upgraded Claude 3.5 Sonnet	Benchmark Evals	Evaluations Report
2024	OpenAI	Los Alamos wet-lab red-teaming [mentioned in o1 model card]	Benchmark Evals	System Card

None of these assessments in the table above are explicitly non-interactive risk assessments, but most discuss and are informed by theoretical work. Most include automated evaluations, and more recent model cards from frontier LLM companies comprise both automated evaluations and some form of interactive assessment (either red-teaming or uplift study). There are some explicit uplift studies conducted by OpenAI and RAND, and UK AISI is currently running chem-bio uplift studies (AISI 2024a, AISI 2024b).

As far as we are aware, the only completed, publicly acknowledged non-interactive risk assessment of non-LLM chem-bio models is CLTR's Capability-Based Risk Assessment for Biological Tools (Moulange et al., 2024). It is possible that uplift studies could be completed without model interaction through structured interviews of chem-bio AI model end users, using beneficial applications as proxy uplift. There are no existing benchmarks or automated evaluations for non-LLM models that explicitly assess dual-use capabilities. However, there is currently a collaborative CLTR-RAND project underway to assess chem-bio AI models using non-interactive risk assessment (RAND & CLTR, 2024).

b. How might existing AI safety evaluation methodologies (*e.g.,* benchmarking, automated evaluations, and red teaming) be applied to chem-bio AI models? How can these approaches be adapted to potentially specialized architectures of chem-bio AI models? What are the strengths and limitations of these approaches in this specific area?

All four evaluation methodologies are useful for assessing the dual-use chem-bio capabilities and risks. UK AISI regularly uses all four approaches to assess chem-bio risks of frontier LLMs, providing clear evidence of the value of these approaches (AISI 2024a).

For non-LLM chem-bio AI models, however—of which there are many more models which tend to be much smaller, targeted and less compute-intensive (Moulange et al., 2023; EpochAI, 2024)—non-interactive risk assessment may be more important than the other three methods.

This is because:

(a) It is currently impractical to design and build evaluations suitable for the wide range of chem-bio AI models available. Chem-bio AI models have highly specialised inputs, architectures and outputs and are not easy to interact with or understand for non-experts; and

(b) Identifying the riskiest chem-bio AI models is challenging since compute does not clearly correlate with model capability, and the AI and biosecurity communities currently do not fully understand—and cannot prioritise between—the many relevant threat models associated with non-LLM chem-bio AI models.

Instead, non-interactive risk assessments, like the CLTR Capability-Based Risk Assessment for Biological Tools and the upcoming RAND–CLTR risk index for AI-powered biological tools, can help prioritise between hundreds of different tools and more effectively identify the riskiest dual-use capabilities (Moulange et al., 2024; RAND & CLTR, 2024). Such assessments straightforwardly incorporate more information about a model beyond its training compute, providing a more nuanced examination of the model. Within the government, these assessments could be further supplemented with intelligence and include detailed threat modelling. We therefore recommend non-interactive risk assessments be conducted on a recurring basis within the government by technical experts with national security input. For more details on the advantages of non-interactive risk assessments for chem-bio AI models, please see Smith et al., 2024 and its associated report (CLTR 2024).

Nevertheless, non-interactive risk assessments may not fully capture a comprehensive spectrum of the relevant threat models associated with the riskiest chem-bio AI models and bioFMs, given they are first-line assessments. This means they should be followed by red-teaming by technical experts within the government with national security input. This will provide the most information about the risks, helping to inform proportionate mitigation measures and the design of automated evaluations and uplift studies, if suitable.

Finally, note that if bioFMs become much more capable than chem-bio AI models across many biological tasks, the existing LLM-focused assessment methods are likely to be much more applicable for chem-bio AI models in general. This is because if most specialised biological capabilities are instead regularly accessed through a small number of bioFMs, it becomes more feasible to develop an evaluation suite for that small group of models. Additionally, conducting expensive red-teaming exercises and uplift studies is much more feasible with only a few frontier chem-bio AI models. For more details on how bioFMs might change the landscape of chem-bio AI models, please see *Trend 5* in the 2024 CLTR report 'The near-term impact of AI on biological misuse' (Rose et al., 2024).

c. What new or emerging evaluation methodologies could be developed for evaluating chem-bio AI models that are intended for legitimate purposes but may output potentially harmful designs?

While we do not anticipate that a new "fifth" form of evaluation method will emerge in the near future, we believe that existing methods can be used in new and better ways to evaluate dual-use chem-bio AI models. First, non-interactive risk assessment methods can be used much more widely and with minimal cost—including before a potentially dual-use chem-bio model is published. Second, we believe that the AI and biosecurity communities should prioritise developing rapid forms of LLM–chem-bio AI model interaction assessment (both automated and non-automated).

As discussed above, non-interactive risk assessments are likely the best first-line evaluation method for novel chem-bio AI models. Uplift studies and red-teaming exercises are too expensive to conduct for every potentially risky non-LLM model. It is also difficult to build effective, generalisable automated evaluations for highly specialised models, especially when the associated threat modelling remains not clearly understood. Instead, we suggest that US AISI conduct non-interactive risk assessments of potentially dual-use chem-bio AI models, drawing on the methods from CLTR and the upcoming CLTR-RAND risk index report (Moulange et al., 2024; RAND & CLTR, 2024).

These are advantageous in three ways:

i) **Low resource requirement:** CLTR demonstrated non-interactive tool evaluation can be completed with just a small team in a few months without the need for automated evaluations or costly red teaming or uplift studies. A single tool could be manually assessed by a technical expert in US AISI in a few days, and LLMs could potentially be used internally to automate some aspects of this process.

ii) **Timely visibility:** infrastructure for non-interactive risk assessments would allow the U.S. Government to rapidly detect jumps in capability for chem-bio AI models, helping to inform and target future assessments and determine proportionate mitigation measures.

iii) **Pre-release:** With agreement from chem-bio AI model developers, there is nothing in principle stopping non-interactive assessments from being completed prior to model release, and this could eventually be a part of existing dual-use review processes.

The ability of LLMs to troubleshoot the use of chem-bio AI models is increasing, necessitating the development of evaluations to assess potential dual-use risks. This includes reducing barriers to use through natural language troubleshooting of these tools and potentially the ability to iterate suggestions on the outputs of these models. We recommend evaluations focused on troubleshooting are prioritised, and uplift studies conducted investigating how accessible these chem-bio AI models are to non-experts with the help of a LLM. Note that here 'non-experts' includes both novices and individuals with expertise in fields other than that of the chem-bio AI model itself.[2]

2. Current and/or Possible Future Approaches To Mitigate Risk of Misuse of Chem-Bio AI Models

a. What are current and possible future approaches to mitigating the risk of misuse of chem-bio AI models? How do these strategies address both intentional and unintentional misuse?

Mitigations for the misuse of chem-bio AI models remain under-explored, and almost all model-focused mitigations apply only to LLMs. Mitigating the intentional misuse risks of chem-bio AI models requires a multi-layered strategy. In 2024, CLTR recommended that the government prioritise non-interactive risk assessments as outlined above, the research and development of technical safeguards, and the drafting of responsible development guidelines for chem-bio AI models (Smith et al., 2024).

Other potentially useful strategies warranting further exploration include:

a) **Input–Output Screening and Harm Refusal**: Models could be designed to refuse to provide outputs determined to be harmful or flag inputs that suggest malicious intent. This can be reinforced by logging potentially dangerous queries to create user accountability and deter misuse. Harm refusal is widely used in frontier LLMs accessed via APIs and could be applied to non-open-sourced chem-bio AI models. While current mechanisms are not jailbreak-proof, research in this area could answer whether robust implementations could significantly reduce risks. Extending harm refusal to chem-bio AI models is highly promising, though largely unexplored.

b) **Excluding Dangerous Training Data:** By excluding dual-use or hazardous datasets during training, it could be possible to limit the model's ability to generate dangerous outputs. Although some argue that models can generalize dangerous capabilities even without explicit training on dual-use data, there is currently limited empirical evidence and some strong theoretical reasons to expect models to underperform on tasks relating to data on which they were not trained. Even if imperfect, this strategy is low-cost and should be considered for further study as a part of a broader defense-in-depth approach. If empirical evidence is forthcoming for its effectiveness,

[2] For further information, please see Rose et al., 2024

data curation standards for chem-bio AI models could be developed to ensure this safeguard is consistently applied.

c) **Model Unlearning:** This involves modifying trained models to reduce their capabilities on harmful tasks post-training. Neural-network-based unlearning methods already exist for some AI models but are still in the early stages of development. Model unlearning could theoretically be applied to chem-bio AI models, though further research is needed to adapt these techniques to diverse machine-learning architectures. Like data curation, unlearning faces challenges related to "re-generalization," where models relearn harmful capabilities through indirect pathways. However, its potential as a targeted safeguard warrants further exploration.

d) **Dataset Security Post-Deployment**: Restricting access to sensitive datasets can reduce the risk of misuse during model fine-tuning post-deployment. Developers could ensure that datasets are curated and access-controlled.

e) **API Deployment:** Encouraging the deployment of chem-bio AI models through an API can help prevent unauthorized access to model weights and prevent potentially dangerous downstream fine-tuning. The government could develop a free-to-use API for model developers to encourage API use and ensure adequate security measures.

Whilst the development of responsible development guidelines in the protein design community is very encouraging (e.g. 2024 IPD guidelines), we believe there is still a role for governmental assistance and guidance on the development of chem-bio AI models. Risk assessment and mitigation should not be solely the responsibility of the developer community. Government guidelines developed with technical and biosecurity experts could specify requirements for risk assessments during the design, deployment, and post-deployment phases. Technical safeguards could be recommended for different risk levels and structured access systems (e.g., API-based access) encouraged to limit misuse risks, especially for powerful tools. Ideally, a free-to-use API for model developers could be offered to developers to reduce any undue burden on this community.

Further research is urgently required to determine the most effective mitigation strategies for chem-bio AI models (Smith et al., 2024). To prevent stifling innovation or the beneficial use of these models, any risk mitigation measures should be proportional to the risks posed as determined through a comprehensive risk assessment.