Arbiter: Automated Review of Bio-AI Tools for Emerging Risk[1]

Ryan J.C. Teo
Institute of Microbiology and Infection
University of Birmingham

Shrestha Rath
Bloomberg School of Public Health
Johns Hopkins University

With
Apart Research

Abstract

The rapid advancement of AI-enabled biological tools presents significant biosecurity and AI safety challenges, necessitating scalable and balanced assessments. Building on the Global Risk Index (GRI) for AI-enabled Biological Tools report's foundational framework, this work introduces Arbiter, an automated pipeline designed to overcome the GRI's resource-intensive manual analysis. Arbiter employs a multi-stage LLM-driven process to analyze scientific literature, systematically evaluating AI-Bio tools for misuse risks and, notably, their potential benefits. This includes assessing economic impact, national competitiveness, and crisis response capabilities, providing policymakers with the comprehensive, balanced insights required for informed decision-making. A pilot execution demonstrated Arbiter's ability to efficiently monitor emerging tools, highlight areas for prompt and model refinement, and support the development of future decision-support frameworks. Arbiter's modular and extensible design empowers users to tailor analyses, ensuring continuous, scalable, and adaptable oversight in this dynamic field.

Keywords: dual-use technologies, risk assessment, benefits assessment, biosecurity, open-source

1. Introduction

The Global Risk Index (GRI) for AI-enabled Biological Tools (Webster et al., 2025) establishes a crucial precedent for understanding emerging AI capabilities in advancing our understanding of and control over biology. It provides a structured framework for assessing tools across eight key functional categories spanning areas like designing biological components (e.g., viral vectors, proteins) to predicting pathogen behaviors and automating experiments. The Index combines a broad landscape assessment (tracking development trends and future capabilities)

[1] Research conducted at the CBRN AI Risks Research Sprint, 2025

with a detailed tool assessment that scores misuse-relevant capabilities using predefined scenarios, classifying tools as 'Red' (high concern), 'Amber' (moderate concern), or 'Green' (low concern).

This work is critical for biosecurity and AI safety because AI is dramatically accelerating progress in the life sciences. While these AI-Bio tools promise immense benefits—such as faster vaccine development or improved diagnostics—they also present dual-use risks. This means the same capabilities driving innovation could, if misused, lower the barrier to developing novel biological agents or enhancing existing ones, posing significant CBRN threats. AI's power to analyze vast biological datasets and uncover complex patterns means these risks can emerge rapidly, potentially reducing the specialized expertise traditionally needed for misuse.

The GRI report itself involved a rigorous, in-depth manual analysis, requiring a multidisciplinary team and hundreds of person-hours from researchers and expert graders to conduct extensive literature reviews, identify tools, and perform detailed evaluations. However, despite its methodological rigor, key limitations were identified that hinder its direct application for continuous policy support. Firstly, its exclusive emphasis on misuse risks, while important, provides an incomplete foundation for the balanced risk-benefit analysis that policymakers require. Understanding not only the risks but also the transformative benefits of these tools is crucial to avoid overly precautionary approaches that could stifle beneficial innovation and create strategic disadvantages. Secondly, the manual nature of the GRI's assessment process is inherently resource-intensive and cannot scale to keep pace with the exponential growth of new AI-Bio tools—a trend the GRI itself identified, noting a substantial increase in tool releases. This lack of scalability hinders policymakers' ability to maintain up-to-date situational awareness.

To address these limitations and build upon the GRI's explicit recommendations for automated assessment and benefits analysis, we introduce **Arbiter (Automated Review of Bio-AI Tools for Emerging Risk)**. Arbiter is a modular pipeline designed to automate the review of AI-enabled biological tools in scientific literature, offering a more efficient and scalable method for continuous risk monitoring. Crucially, it innovatively introduces a modular benefits assessment framework, providing the comprehensive, balanced evaluation necessary for informed policy decisions in biosecurity and AI safety.

2. Methods

2.1. Tool development

The Arbiter tool (https://github.com/teojcryan/arbiter) is designed to automate and standardise the assessment of AI-enabled biological tools, specifically focusing on their potential for misuse and benefits. Our approach is fundamentally grounded in the methodologies outlined in the Global Risk Index report. Specifically, we have operationalised and extended the automated assessment pilot detailed in Appendix C of the GRI report, focusing on translating the complex and resource-intensive risk assessment framework into a practical, scalable, and extensible software tool.

At its core, Arbiter employs a multi-stage pipeline driven by a series of Large Language Models (LLMs). This pipeline is designed for non-interactive assessment of publicly available scientific literature, ensuring broad coverage and scalability across a vast landscape of research. The process unfolds across distinct stages, each with a clear objective as illustrated in Table 1 and Figure 1.

Stage		Objective
0	Query generation	Transforms a user-provided topic into an optimised search query, designed to capture relevant scientific papers.
1	Relevancy filtering	Filters out content irrelevant to AI-Bio tools based on an assessment of paper titles, abstracts, and types.
2	Risk assessment	Evaluates the tool's potential dual-use status and assigns a misuse risk score guided by detailed prompts that define specific risk levels and misuse scenarios.
2.5	Benefits assessment	Evaluates the potential positive impacts of the identified tools across various dimensions.
3	Strategic reasoning	For papers flagged with high misuse risk, explore potential misuse scenarios, capability amplification, and mitigation challenges to provide foresight

Table 1: Summary of the five stages in the Arbiter pipeline

Arbiter embodies a modular architecture that prioritizes extensibility and maintainability. This design philosophy is crucial for adapting to the dynamic field of AI-enabled biological tools and supporting diverse research needs. Users are granted granular control over the entire operational process, allowing them to:

- Select preferred data sources, currently supporting either OpenAlex or EuropePMC.
- Choose LLM providers for each pipeline stage, with options including local Ollama or Google Generative AI.

Furthermore, all interactions with LLMs are handled via templates which allows for straightforward prompt tuning and experimentation. Arbiter allows users to easily modify or create new prompts to refine the LLM's behaviour, test different assessment criteria, or adapt the analysis for new types of biological AI tools without needing to alter the core codebase.

To illustrate how Arbiter allows users to adapt their analyses, we have chosen to include a parallel assessment to the misuse capability assessment (Stage 2) that is focused on evaluating potential benefits unlocked by the development and deployment of a particular tool (Stage 2.5). This feature allows users to tailor their analyses to evaluate not only misuse potential but also the positive impacts of AI-enabled tools, offering a more holistic understanding. More information on how the assessment prompt was constructed is available in Appendix A.3.

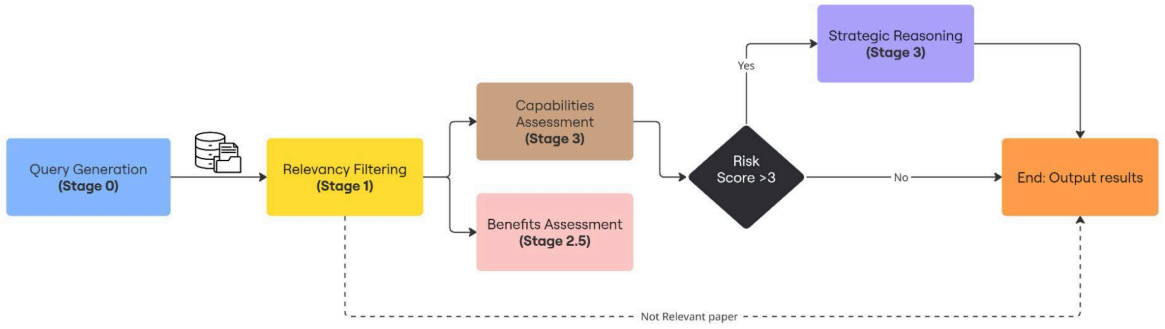


Figure 1. Flowchart for Arbiter.

Lastly, to enable direct user engagement and critical review, Arbiter includes an interactive Streamlit dashboard. This user-friendly interface visualizes pipeline outputs, allowing users to explore classified tools, filter by risk/benefit scores, and review AI justifications which is vital for human-in-the-loop validation and informed decision-making.

3. Results

To validate the Arbiter tool's methodology and demonstrate its capabilities in a practical context, a pilot execution was conducted. We focused this initial assessment on AI-enabled biological tools within the viral vector design category. This choice was deliberate, aiming to emulate the automation trial performed in the Global Risk Index report, thereby allowing for a direct comparative assessment of Arbiter's effectiveness against established benchmarks.

The Arbiter pipeline was configured to utilise the "gemma-3-27b-it" model (Gemma et al., 2025) across all its stages. While this model provided functional results, our observations indicate that more advanced reasoning models with larger context windows would likely yield more nuanced and accurate outputs, particularly for the later stages of the pipeline where complex contextual analysis and speculation on misuse scenarios are critical.

The process began with the input topic “AI tools for engineering and optimising viral vectors.” Stage 0 of Arbiter's pipeline dynamically generated a richer, optimized search query, which was then applied to the EuropePMC database. This search yielded an initial 6829 hits. Due to resource constraints typical of a pilot study, the pipeline proceeded to process the first 250 articles from this dataset. The complete pipeline execution for these articles concluded in 28 minutes.

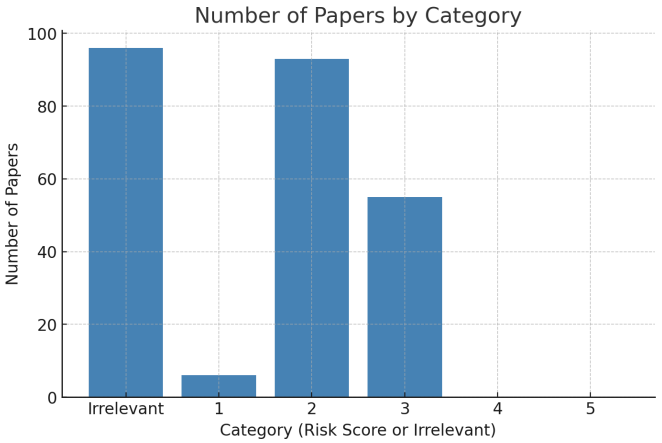


Figure 2. Distribution of processed papers across risk scores

The output of Arbiter's initial assessment stages is visualized in Figure 2. The significant number of "irrelevant" papers (approx. 95 out of 250 processed) suggests that while the Stage 0 query generation is dynamic, further refinement of search terms may be beneficial to manage the inherent 'noise' in broad literature searches, providing valuable data for future query optimization. However, the GRI report noted that wider search terms were a necessary product of minimising false negatives. Stage 2 primarily classified relevant papers into the 'Low' risk (Category 1), with few reaching higher scores. This conservative outcome suggests that the current LLM configuration warrants further tuning or prompt engineering to achieve more nuanced risk differentiation and better capture the spectrum of potential misuse. Since no papers scored '4' or '5', Stage 3: Strategic Reasoning could not be activated. This highlights

the need for datasets with a broader distribution of risk scores to fully test the pipeline's capabilities.

Lastly, the dashboard (screenshot included in Appendix A.4) complements these findings by illustrating how Arbiter presents classifications and justifications. This interactive interface allows for crucial human-in-the-loop validation, enabling users to explore, filter, and review the AI's assessments.

4. Discussion and Conclusion

Our pilot execution of Arbiter has successfully demonstrated its foundational capabilities in automating the assessment of AI-enabled biological tools, offering a crucial advancement for evidence-based policymaking. The automated pipeline enables continuous monitoring at scale, a vital necessity given the rapidly increasing pace of tool development (the GRI report found that nearly as many tools were released in 2023-2024 as in the preceding four years combined). This rate would quickly overwhelm manual assessment processes, making Arbiter essential for policymakers to maintain situational awareness of evolving capabilities.

However, this pilot also underscored critical dependencies that influence the quality and depth of the analysis. The accuracy and nuance of Arbiter's output are intrinsically linked to:

- The quality of prompts employed, which directly impacts LLM classification precision.
- The quality of the underlying AI models; access to more advanced reasoning capabilities could significantly enhance later pipeline stages, particularly for detailed risk assessment and speculative reasoning.
- Access to relevant information, specifically the availability of full research paper text versus abstracts, a limitation also highlighted in the GRI report.

Looking ahead, future development could significantly enhance Arbiter's impact and reliability by addressing these dependencies. Advancements in data retrieval could focus on more robust handling of full-text papers, potentially overcoming paywall challenges to improve comprehensive coverage. Iterative prompt engineering will be central to refining the accuracy of risk and benefit classifications, directly impacting the quality of the assessment. Crucially, exploring integrations with superior LLMs possessing enhanced reasoning capabilities could substantially bolster Arbiter's AI Safety Contribution, enabling more sophisticated analysis in Stage 3 and ensuring fewer concerning tools are missed.

These advancements in automation could also build future structured decision support frameworks (Manheim et al., 2016). For instance, Arbiter's outputs could directly inform funding allocation by translating risk and benefit scores into specific budget recommendations or regulatory actions to accelerate defensive capabilities, transforming raw "Red/Amber/Green" classifications into actionable policy directives. Furthermore, automation facilitates the forecasting of emerging capabilities; threshold-based alert systems could prompt immediate reassessment of "potential for change" metrics for categories showing new investment or research signals, acting as an early warning system. This allows expert effort to be concentrated on areas of dynamic change, making regular, resource-intensive updates feasible.

Ultimately, by addressing these limitations and pursuing these future enhancements, Arbiter could evolve into a more powerful and reliable system for proactive biosecurity risk assessment.

5. References

Gemma Team, Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., ... Hussenot, L. (2025). Gemma 3 Technical Report (No. arXiv:2503.19786). arXiv. https://doi.org/10.48550/arXiv.2503.19786

Manheim, D., Chamberlin, M., Osoba, O. A., Vardavas, R., & Moore, M. (2016). Improving Decision Support for Infectious Disease Prevention and Control: Aligning Models and Other Tools with Policymakers' Needs. https://www.rand.org/pubs/research_reports/RR1576.html

Webster, T., Moulange, R., Del Castello, B., Walker, J., Zakaria, S., & Nelson, C. (2025). Global Risk Index for AI-enabled Biological Tools (Public Report). The Centre for Long-Term Resilience & RAND Europe. https://doi.org/10.71172/wjyw-6dyc https://doi.org/10.48550/arXiv.2503.19786

A. Appendix

A.1. Security Considerations

The open-source nature of Arbiter, encompassing its codebase and data analysis methodologies, necessitates consideration of potential information hazards. However, the risks associated with this tool are fundamentally linked to the underlying AI models and biological tools it analyzes, rather than Arbiter itself creating new dangers. The security of Arbiter's implementation is managed through standard software engineering practices. We acknowledge that the information Arbiter surfaces, especially regarding tools with high misuse capabilities, requires careful handling. Therefore, while our project champions transparency through open-source data and code, we strongly advise users to exercise caution. Specifically, users should refrain from publicising AI-enabled biological tools identified by Arbiter as possessing high misuse potential. This recommendation serves as a proactive measure to prevent the inadvertent amplification of dangerous capabilities, an approach that provides direct user guidance beyond mere omission of details.

A.2. Search query generated in pilot assessment

(artificial intelligence OR machine learning OR deep learning OR AI OR neural networks OR latent variable models OR biophysically inspired algorithms) AND (viral vector OR AAV OR adeno-associated virus) AND (engineering OR optimisation OR optimization OR design OR predict* OR modeling OR modelling) AND (capsid OR cargo capacity OR immunogenicity OR tissue targeting OR gene therapy)

A.3. Benefits Assessment

One differentiating feature of our assessment is the inclusion of a benefits assessment. We evaluated AI-enabled biological tools across four policy relevant dimensions designed to capture essential priorities of policymakers.

1. **Economic impact** quantifies market and healthcare value on a 1-5 scale by examining the economic value tools could create through market opportunities, cost reductions, or

healthcare improvements, with particular attention to the scale of impact suggested by target applications, patient populations, or industrial processes that could be addressed.

2. **National Competitiveness** focuses on strategic advantage by addressing the zero-sum dynamics of technological dominance, recognising that tools that for instance, provide 2-3 years lead time in pandemic response or reduce reliance on international supply chains carry strategic weight beyond technical merits.
3. **Resilience and Crisis Response** examines emergency deployment potential by evaluating how quickly tools could be deployed in health emergencies, what surge capacity they could provide, and how they might strengthen biosecurity through defensive capabilities or threat detection mechanisms.
4. **Implementation Feasibility** considers deployment pathway practicality by analyzing what facilities, equipment, or expertise are needed and whether tools can be deployed using existing infrastructure, providing essential insight into the translation of technological capability into operational reality.

The exact prompt used for conducting the benefits assessment can be found in the Github repository.

A.4. Example screenshot from Arbiter

iGTP: learning interpretable cellular embedding for inferring biological mechanisms underlying single-cell transcriptomics.

Dual-Use Concern

True

Assessed Risk Score

3

View details

Misuse capabilities

Summary:

This work presents iGTP, an AI framework integrating transcriptional programs and protein-protein interactions to generate interpretable cellular embeddings and predict responses to perturbations.

Dual-Use Justification:

The ability to accurately generate cell embeddings and predict responses to perturbations could be misused to design novel biological agents or optimize existing ones.

Risk Justification:

While not directly engineering pathogens, the predictive capability and generative model represent a meaningful advance with potential for misuse, though significant expertise would still be required for harmful applications.

Benefits

Economic Impact	National Competitiveness	Resilience/Response	Implementation Feasibility
4	4	3	3

Economic Impact:

Improved drug development, personalized medicine, and a deeper understanding of disease mechanisms could lead to significant cost savings in healthcare and create new market opportunities for diagnostics and therapeutics.

National Competitiveness:

The iGTP framework's superior performance over existing methods and its ability to predict responses to novel perturbations positions the US as a leader in AI-driven biological research and potentially reduces reliance on foreign technologies in this critical field.

Resilience & Crisis Response:

The ability to infer cellular responses to perturbations and generate cell embeddings could accelerate the development of countermeasures during health emergencies, though deployment speed is not explicitly stated and likely requires significant computational resources.

Implementation Feasibility:

While leveraging existing deep learning and bioinformatics techniques, iGTP requires expertise in both areas and substantial computational infrastructure for training and operation, potentially limiting immediate widespread adoption.