Open-weight genome language model safeguards: Assessing robustness via adversarial fine-tuning

James R. M. Black
Center for Health Security
Johns Hopkins Bloomberg School of Public Health

Moritz S. Hanke
Center for Health Security
Johns Hopkins Bloomberg School of Public Health

Aaron Maiwald
Department of Chemistry
University of Oxford

Tina Hernandez-Boussard
Stanford University School of Medicine

Oliver M. Crook
Department of Chemistry & Kavli Institute for Nanoscience Discovery
University of Oxford

Jassi Pannu*
Center for Health Security
Johns Hopkins Bloomberg School of Public Health
pannu@jhu.edu

Abstract

Novel deep learning architectures are increasingly being applied to biological data, including genetic sequences. These models, referred to as genomic language models (gLMs), have demonstrated impressive predictive and generative capabilities, raising concerns that such models may also enable misuse, for instance via the generation of genomes for human-infecting viruses. These concerns have catalyzed calls for risk mitigation measures. The de facto mitigation of choice is filtering of pretraining data (i.e., removing viral genomic sequences from training datasets) in order to limit gLM performance on virus-related tasks. However, it is not currently known how robust this approach is for securing open-source models that can be fine-tuned using sensitive pathogen data. Here, we evaluate a state-of-the-art gLM, Evo 2, and perform fine-tuning using sequences from 110 harmful human-infecting viruses to assess the rescue of misuse-relevant predictive capabilities. The fine-tuned model exhibited reduced perplexity on unseen viral sequences relative to 1) the pretrained model and 2) a version fine-tuned on bacteriophage sequences. The model fine-tuned on human-infecting viruses also identified immune escape variants from SARS-CoV-2 (achieving an AUROC of 0.6), despite having no exposure to SARS-CoV-2 sequences during fine-tuning. This work demonstrates that data exclusion might be circumvented by fine-tuning approaches that can, to some degree, rescue misuse-relevant capabilities of gLMs. We highlight the need for safety frameworks for gLMs and outline further work needed on evaluations and mitigation measures to enable the safe deployment of gLMs.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Biosecurity Safeguards for Generative AI.

1 Introduction

Large language models (LLMs) have transformed our ability to collect and generate information encoded in human language, and today match or outperform humans on a multitude of complex tasks [Zhao et al., 2025]. Deep learning methodologies originally developed for LLMs, such as self-supervised learning on large amounts of unlabeled data, have been applied to biological data such as protein and genetic sequences. These models are referred to as protein language models (pLMs) or genomic language models (gLMs), respectively [Benegas et al., 2024]. Early pLMs and gLMs have demonstrated intriguing capabilities, such as the prediction of protein structural elements and transcription factor binding sites [Chowdhury et al., 2022, Mendoza-Revilla et al., 2024]. The performance limits of these models is not yet clear and may be gated by the availability of high-quality training data, compute, or novel AI architectures particularly well-suited to biological data. Thus, the potential ceiling performance of biological AI models may be contingent on future, adjacent advances in computational biotechnology.

Scientists have noted that to the degree these systems acquire novel biological capabilities, they might also be vulnerable to misuse [NASEM, 2025]. Developers have acknowledged these risks and taken first steps towards mitigating gLM and pLM risks. However, given the uncertainties regarding the upper limit of model performance described above, it is challenging to determine the best approaches for mitigating misuse risks, both for today and for years into the future. Developers have acknowledged these risks and taken first steps towards mitigating gLM and pLM risks. A handful of risk mitigation approaches have been tested to date. For instance, the developers of ESM3-open "removed the capability for the model to follow prompts related to viruses and toxins" for their open-source model ESM3-open [Hayes et al., 2025].

The de facto mitigation of choice for these models, though, is known as data exclusion (also known as data filtering). Data exclusion involves the deliberate removal of data from training datasets in order to limit model performance on risky capabilities tied to such data, such as capabilities related to biological weapons development. For safety and security reasons, the developers of the Evo models, a series of gLMs "excluded genomic sequences from viruses that infect eukaryotic hosts (...) to ensure our openly shared model did not disseminate the capability to manipulate and design pathogenic human viruses" [Brixi et al., 2025]. Similarly, for ESM3-open, the developers "removed sequences unique to viruses, as well as viral and non-viral sequences from the Select Agents and Toxins List" from the training data in order "to reduce the capability of ESM3-open on these sequences" [Hayes et al., 2025].

However, relatively little is known about the true performance reduction that results from the exclusion of sensitive data. Some researchers have demonstrated that this risk reduction may not be robust [Zhang et al., 2025]. Open-source models can be fine-tuned after their release, suggesting that a sufficiently skilled actor could rescue model performance for misuse-related tasks if the sensitive data in question were available to them. The quantity of sensitive data required in this context to rescue the misuse-enabling capability, as well as the degree to which those capabilities can be rescued, has not yet been tested. Further, some researchers have suggested that highly capable biological models could interpolate risky capabilities due to their broad generalizability, even when particular subsets of training data have been removed. Such generalizability for gLMs might exist analogously to how it was demonstrated for protein structure prediction capabilities [Ahdritz et al., 2024].

Here, we conduct evaluations on the robustness of data exclusion from the open-source gLM, Evo 2, by reintroducing sequences from human-infecting viruses via fine-tuning and assessing the rescue of misuse-relevant capabilities. Generally, Evo 2 learns patterns from DNA sequences to both 1) predict the impact of genetic sequence changes and 2) generate genetic sequences; we evaluate primarily the predictive capability in its application to viruses. We demonstrate that important misuse-relevant capabilities for viruses unseen in the training data, including the prediction of downstream phenotypes related to immune escape, can be rescued through adversarial fine-tuning. Lastly, we discuss frameworks for systematically approaching safety and security considerations for gLMs to improve biosecurity in the future.

2 Methods

Note: Due to concerns about information hazards arising from this work, some details have been redacted from the Methods section, and code pertaining to this study has not been made available but can be obtained from the authors by reasonable request.

2.1 Base model used for fine-tuning

We used the gLM Evo 2 [Brixi et al., 2025], an open-source model available in three versions with 1 billion (1B), 7 billion (7B), and 40 billion (40B) parameters, respectively. Evo 2-7B was used as the base model for fine-tuning.

2.2 Fine-tuning methodology

Fine-tuning was performed using the Evo 2-7B as a base, leveraging the Savanna framework for distributed training. Briefly, this approach uses a transformer backbone with rotary position embeddings and was configured for autoregressive pretraining. The model was distributed across 4 NVIDIA H100 GPUs using DeepSpeed ZeRO Stage 1 optimization with mixed precision (BF16) training. The pretrained Evo 2 7B weights were fine-tuned using a character-level tokenizer, with the model adapted from its original 1M context length to 4,096 tokens per sample due to computational constraints. Training utilized the Adam optimizer ($\beta 1 = 0.9$, $\beta 2 = 0.95$, $lr = 2x10e-6$), with an effective batch size of 768 sequences. We applied gradient clipping (0.5), weight decay ($1x10e-4$), and dropout (0.05) for stability.

2.3 Sequence data used for fine-tuning

The dataset used for adversarial fine-tuning consisted of viral sequences. We attempted to collate this dataset in an analogous way to the approach taken by the Evo 2 authors in collating their OpenGenome2 dataset, which was utilized for training in the first instance. First, a literature search was performed for putative harmful human-infecting viruses across six groups of concern: large DNA viruses, small DNA viruses, positive-strand RNA viruses, negative-strand RNA viruses, enteric viruses, and double-stranded viruses. Second, deduplication was performed in a manner that mimicked the approach taken by the Evo 2 team, in order to remove redundancy while preserving sequence diversity. Complete sequences were deduplicated using Mash sketching with 10,000 k-mers to calculate pairwise distances between all genomes. Sequences with Mash distance <0.01 (i.e., >99 percent average nucleotide identity) were clustered, and the longest sequence from each cluster was considered to be representative [Ondov et al., 2016]. This resulted in a dataset of 122 viral genomes (Figure 1). This was then separated using a 90/10 split into a training dataset of 110 viral genomes and a held-out test dataset of 12 viral genomes for downstream evaluation. In parallel, a second, control dataset of prokaryote-infecting viruses (bacteriophages) was generated for a separate fine-tuning exercise, using an analogous approach. This resulted in the generation of 181 bacteriophage genomes.

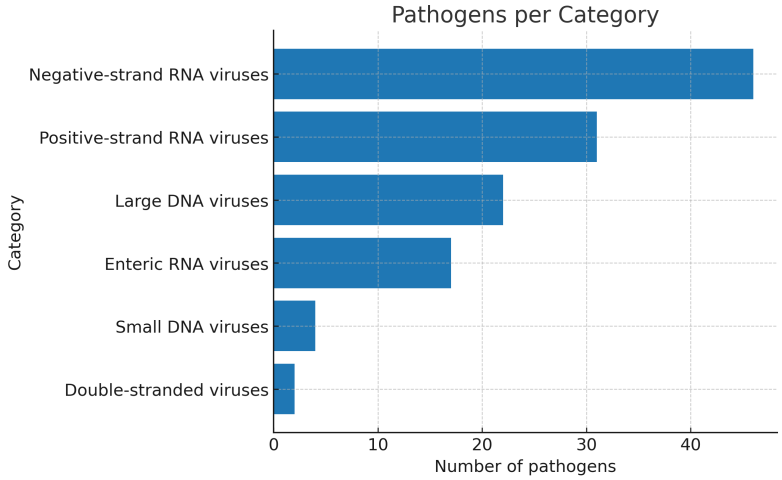

Figure 1: Composition of the dataset of viral genomes, by type of viral sequence. 110 viral sequences were used for fine-tuning, and 12 sequences were held out for downstream evaluation.

2.4 Evaluations

We evaluated two primary capabilities, relating to the predictive, rather than the generative, ability of Evo 2.

Firstly, sequence prediction (perplexity): We evaluated sequence perplexity on three different versions of Evo 2: the pretrained model; the control version, fine-tuned on bacteriophage sequences (FT-bacteriophages); and the version fine-tuned on harmful human-infecting viruses (FT-harmful). Perplexity was calculated separately on the training data, and the held-out set of 12 viral genomes.

Secondly, phenotype prediction: We assessed the ability of the three different versions of Evo 2 to predict which mutations to the SARS-CoV-2 spike protein would underpin immune escape. Importantly, none of the three models had seen any SARS-CoV-2 genomes at any stage of training or fine-tuning. Immune escape was determined using publicly available deep mutational scanning data, leveraging polyclonal or monoclonal antibodies from infected or vaccinated individuals [Starr et al., 2020]. Mutations were divided into those that were likely to underpin immune escape and those that were not likely to underpin immune escape. Sequence perplexity was computed for the sequence encoding the spike protein. The ability of the three models to identify immune escape mutations was compared to BLOSUM-62 scores, representing simple heuristics of evolutionary conservation, and EVEscape, a highly specialized predictor based on a deep learning approach that combines temporal sequences with predicted structural and biophysical information, which was considered the gold standard for this task [Thadani et al., 2023].

3 Results

3.1 Perplexity on held-out viral sequences

In order to understand the degree to which exclusion of sensitive data was a robust mitigation measure for a gLM, we first assessed whether fine-tuning on that data might rescue model performance related to viral sequences. Perplexity provides an approach for assessing this, measuring how well a gLM predicts the next token in a sequence, with lower values indicating better predictive performance and thus a better understanding of the underlying genomic data. In the Evo2 paper, the authors demonstrated that perplexity was increased on eukaryote-infecting virus sequences relative to prokaryote-infecting viruses, reflecting their removal from the training data.

We compared three versions of the 7B parameter model, pretrained Evo 2, and two fine-tuned models, one fine-tuned on the bacteriophage genomes (FT-bacteriophages) and the other on genomes from harmful, human-infecting viruses (FT-harmful). Each version reflected a modified version of the original model, featuring a 1M token context window, with the context window curtailed to 4096 tokens. We evaluated model performance by its ability to predict sequences from within its fine-tuning distribution, as well as a held-out set of viral genomes that had not been included in the fine-tuning dataset. The held-out viruses were balanced across viral families in a similar way to the viruses used for fine-tuning to ensure comprehensive coverage across sequences of interest.

First, we confirmed that sequence length did not relate to perplexity. This was particularly important as we had implemented post-hoc changes to the model to modify context length. There was no relationship between sequence length and perplexity in either the in-domain or out-of-domain (held-out) datasets (Figure 2).

The FT-harmful model exhibited substantially reduced perplexity, relative to the pretrained model and the FT-bacteriophages model on the training data used for fine-tuning. For the test data, the FT-harmful model exhibited the same trends, although the difference in perplexity observed between FT-bacteriophages and FT-harmful was not statistically significant (Figure 3, training data pretrained: median 3.84; training data FT-bacteriophage: median 3.73; training data FT-harmful: median 2.16; test data pretrained: median 3.83; test data FT-bacteriophage: median 3.73; test data FT-harmful: median 3.55). However, relevant to security concerns, this highlights that the model may have begun to learn generalisable patterns from a limited set of examples. The observation that the pretrained model performed slightly worse than the bacteriophage fine-tuned model might plausibly reflect the damage sustained by the model after the reduction in context length.

3.2 Prediction of SARS-CoV-2 immune escape

gLMs might learn patterns within genomic data that correspond to downstream phenotypes. In the context of viral data, these might include transmissibility, virulence, or immune evasion. As an exemplar, we wondered whether the version of Evo 2 fine-tuned on harmful human-infecting viruses might be able to predict which mutations might confer immune evasion for an unseen virus, i.e., not included in the fine-tuning dataset.

To test this, we computed perplexity on the spike protein from the Wuhan SARS-CoV-2 genome, and compared these scores to functional data that has been gathered about this protein relating to host protein interactions and immune escape. Importantly, no SARS-CoV-2 species had been included in

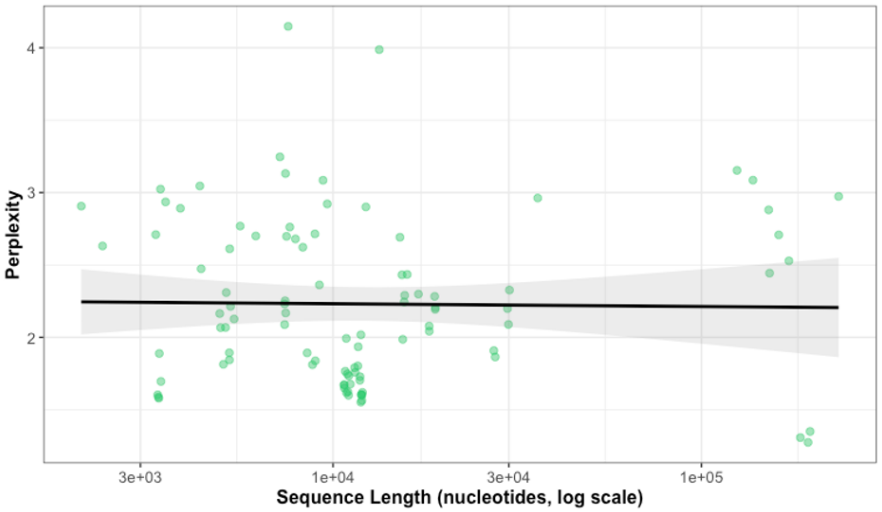

Figure 2: Scatterplot showing perplexity and sequence length for 110 harmful human-infecting viruses that the fine-tuned model was trained on (n=97, r=0.034). Perplexity measures how well a gLM predicts the next token in a sequence, with lower values indicating better predictive performance and thus a better understanding of the underlying genomic data. Each point on the plot corresponds to a single virus used for fine tuning.

the data used for fine-tuning, and so any predictions made by the model could plausibly be considered analogous to performance on an unseen virus. This would represent a relatively stringent level of generalizability, as the model would have to transfer patterns from other viruses to make predictions about a novel pathogen. We leveraged deep mutational scanning data as a baseline, focusing on immune escape due to its clear biosecurity relevance and the immediate availability of high-quality functional data for validation.

We evaluated the performance of five different approaches at predicting which mutations would lead to immune escape and which would not. We leveraged the AUROC to compare the different approaches. Neither the pretrained model nor the model fine-tuned on bacteriophage data conferred any improved predictive performance relative to a random classifier (Figure 3). In addition, we used the BLOSUM-62 matrix to evaluate the degree to which evolutionary conservation might predict mutations that conferred immune escape. This delivered an AUC of 0.51, in essence, no better than a random classifier. The purpose-built, deep learning tool EVEscape, which can be seen as a gold standard here, conferred an AUC of 0.8.

The version of the model fine-tuned on harmful, human-infecting viruses had an AUC of 0.59 at this task. This suggests that fine-tuning on sensitive data conferred generalized predictive properties related to unseen harmful pathogens, relative to both the baseline model and the version of the model fine-tuned on bacteriophages. However, the performance did not match that of the narrow, purpose-built model EVEScape.

Overall, these results show that training data exclusion is unlikely to be a fully robust method for preventing the emergence of misuse-relevant capabilities from open-source gLMs and potential associated downstream harms.

4 Discussion

4.1 Key findings

We have shown that data exclusion can be circumvented through fine-tuning with publicly available sequence data of human-infecting viruses. We also demonstrated a proof-of-principle that misuse-relevant capabilities of openly available gLMs can be rescued through this process. However, model performance on tasks involving functional immune escape prediction was inferior to a more narrow, purpose-built tool. This is congruent with recent results showing narrow, purpose-built tools can outperform large biological foundation models on many tasks related to viruses [Gurev et al., 2025]. Of note, a single fine-tuned gLM might be applicable to a much wider range of misuse-relevant tasks not tested in this study.

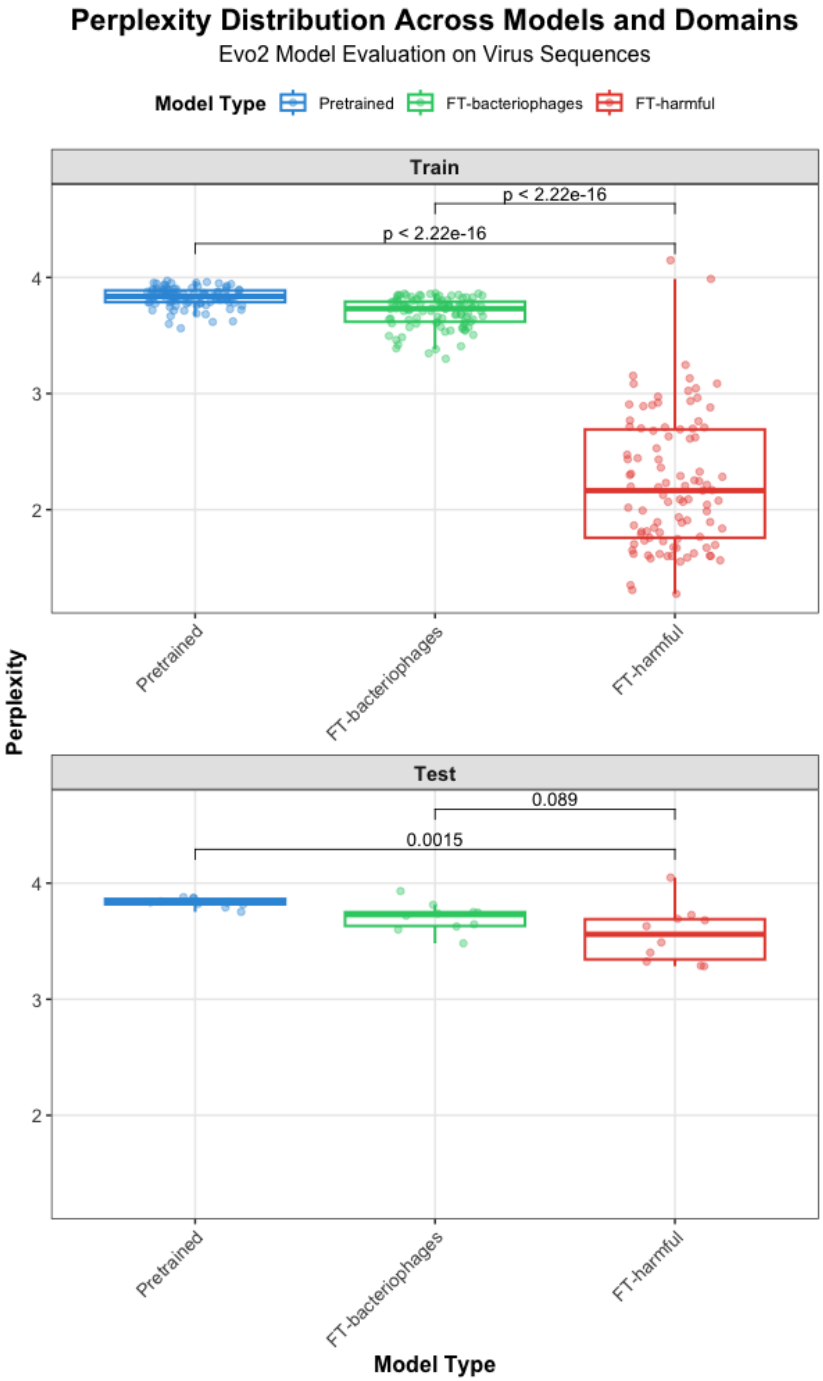

Figure 3: Boxplot showing perplexity on training (n=110) and test (n=12) sequences across the three versions of Evo 2: pretrained; FT-bacteriophages; FT-harmful.

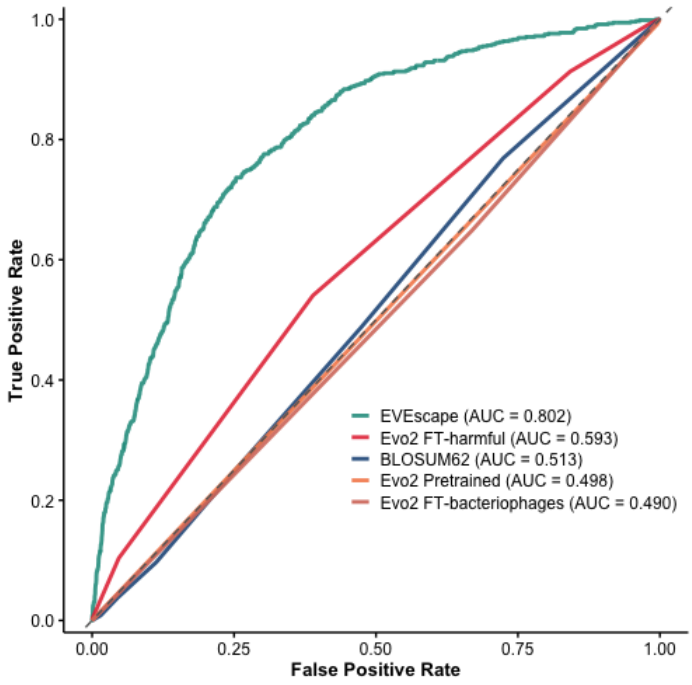

Figure 4: Receiver operating characteristic (ROC) curves comparing methods for identifying SARS-CoV-2 Spike mutations leading to a phenotype of immune escape vs no immune escape. 3 versions of Evo 2, pretrained, fine-tuned (bacteriophages), and fine-tuned (harmful human-infecting viruses) were compared. BLOSUM-62 scores were used to evaluate whether evolutionary conservation alone would confer predictive power. EVEscape, a deep learning model leveraging fitness predictions and structural information, was compared as an example of a model specialized for this exact task.

4.2 Limitations

Certain limitations imply caution when interpreting our results. Firstly, computational constraints prevented us from working with a full 1M token context model. As it was necessary to modify the model architecture to work with a 4096 token context prior to fine-tuning, it is possible that the capabilities we demonstrate underestimate capabilities that might be achievable if the starting point were the 1M token model instead. Nonetheless, many nefarious actors might face similar constraints. Secondly, whilst we have highlighted example tasks, including predictions of downstream functional data, these do not represent the full spectrum of misuse-relevant gLM capabilities. Ideally, an evaluation of sequence generation properties would have been conducted as well, and future work should address this.

4.3 Accident and misuse harms linked to gLMs

A robust evaluation framework should consider the multitude of ways a gLM could be misused. Misuse-enabling capabilities that are most likely or most severe require further investigation first. The most discussed risk to date has been the embedding of functional information regarding pathogens, such as virulence, transmissibility, latency, or immune escape [Pannu et al., 2025]. Whilst perplexity, downstream functional data, and sequence design evaluations are all relevant, future work should explore other less obvious possibilities, such as gLMs being used to predict the fitness of putative novel or heavily modified viral designs, or the generative properties of gLMs being leveraged in similar steps. Experimental work to validate predicted downstream properties should also be considered, insofar as they are conducted in a secure environment using safe biological proxies [Ikonomova et al., 2025].

Given that misuse has a much higher tolerance for failure than benign usage, i.e., the release of an ineffective biological agent has fewer negative consequences than the release of an ineffective countermeasure, it is not clear how well-suited evaluations that simply assess perplexity as an objective measure are. For example, it is possible that a model with very high perplexity, but that has

a small chance to generate a viable viral design that is significantly different from existing sequences, for instance, with a totally foreign antigen, might be more concerning than a model that has lower perplexity, but is incapable of ideating beyond narrow constraints imposed by its training data.

gLMs might also improve our understanding of higher-order genomic architecture, further improving the ability to perform complex pathogen engineering. Thinking further, the broad application of gLMs towards autonomous exploration of the biological landscape through model development feedback loops with limited human checkpoints could have unanticipated risks [Wang et al., 2025a].

Still, limiting gLM capabilities related to viruses may also limit advances in beneficial public health applications, for example, in antigen design for vaccines, where rapid development is particularly relevant in pandemic and epidemic scenarios [Kraemer et al., 2025].

4.4 "Rule-out" evaluations for harm

We propose the development of evaluations that are explicitly designed to rule out potential harms from gLMs. In this context, ruling out means demonstrating with sufficient confidence that particular harms are absent or acceptably unlikely, whereas ruling in means confirming that such harms are present, which is currently usually conducted through red-teaming. Evaluations with high sensitivity are therefore critical for safety assurance. While more research is needed on how such evaluations would be conducted exactly, to promote usability and scalability, these evaluations would ideally be reproducible across gLMs rather than being custom-built for individual models.

4.5 Future risk trajectories

When considering the safety and security of gLMs, researchers should also anticipate future trends in AI development, including multi-modal models [Fallahpour et al., 2025] and agentic systems [OpenAI, 2025], which introduce new categories of misuse or misapplication risks beyond those studied to date. In the modern AI ecosystem, outputs from one tool can readily become inputs to another. The greatest value will likely be unlocked by highly integrated systems, where model inputs and outputs are seamlessly linked. However, this interconnectedness creates safety and security challenges. Risk assessment of an individual tool is often insufficient to capture risks that emerge across the wider ecosystem (the "daisy-chain" problem). Further innovation in developing such evaluations will be required.

4.6 Risk mitigation measure applicability to gLMs

Our evaluations assess the efficacy of the data exclusion risk mitigation approach for gLMs, which falls in the broader risk mitigation category of "capability limitation" [FMF, 2025a]. Capability limitation is a controversial mitigation measure approach, given that many misuse-enabling capabilities are dual-use and can also have desirable applications. For example, the generation of functional viral genomes could also be useful for gene therapy research.

Aside from capability limitation, there are several other approaches for improving the safety and security of gLMs that could be further explored and developed, such as detection and intervention measures or access controls. LLM developers have developed various technical risk mitigation measures associated with their models, such as refusal training, constitutional classifiers, or false learning [FMF, 2025a]. Researchers are only beginning to explore how and to what degree these risk mitigation measures can be applied to both narrow biological tools as well as biological foundation models. Not all available mitigation measures will be appropriate for gLMs, and novel technical methods will likely be needed [Wang et al., 2025b]. In particular, risk mitigation measures for open-source models should be explored more as open models are prevalent in computational biology [Tamirisa et al., 2025].

4.7 Towards gLM safety frameworks

Scientists have generally long recognized the importance of frameworks to assess the risks and benefits of research involving human subjects and animals. Within AI development, safety frameworks have also emerged as a critical tool for frontier AI developers, aiding in the assessment and management of a diversity of risks from advanced AI systems [FMF, 2025b]. Given the rapid advancements in the field, we wish to highlight the importance of developing similar safety frameworks for gLMs, so that their development and deployment can advance responsibly, the manifold benefits can be harnessed, and misuse risks can be effectively mitigated.

5 Conclusion

We demonstrate that, in the case of open-source models, it is possible to circumvent data exclusion safeguards via fine-tuning with human-infecting virus sequences. If sensitive pathogen data such as this is publicly accessible, it can be used to fine-tune an openly available gLM, thereby rescuing misuse-enabling capabilities. Importantly, the rescued performance did not match that of a narrow, purpose-built tool for functional immune escape prediction tasks. Therefore, while sensitive data exclusion raises the bar for misuse, it is susceptible to circumvention. Other risk mitigation measures for gLMs will be required in addition. To ensure safe and responsible development and deployment of this class of powerful biological AI models, further work is needed to develop a taxonomy of misuse-enabling capabilities and a corresponding toolkit of implementable gLM-specific risk evaluation and mitigation measures. Ultimately, these efforts should contribute towards comprehensive safety frameworks for gLMs that developers and deployers can implement to manage risks.

Acknowledgments and Disclosure of Funding

The authors thank Garyk Brixi and Ishan Mukherjee for assistance with fine-tuning Evo 2. The authors also thank two anonymous peer reviewers and Coleman Breen for helpful feedback.

Funding (direct support): JRMB—Open Philanthropy; MSH—Horizon Institute for Public Service (scholarship); AM—Open Philanthropy (Ph.D.); THB—NIH NCATS (1UM1TR004921), AHRQ (R01HS024096), NIH NLM (R01LM013362); OMC-New College Todd–Bird Junior Research Fellowship; MRC Fellowship MR/Y010078/1; JP—Chan Zuckerberg Initiative; Open Philanthropy.

Competing interests: THB—royalties or licenses (Coursera, AI in Healthcare); consulting fees (PAUL HARTMANN AG; Grai-Matter; Roche); stock or stock options (Verantos Inc.; Grai-Matter); advisory board (AtheloHealth). OMC—consulting fees (Pelago Biosciences; Faculty.ai; MarketCast); scientific advisory board member (Evolvere Biosciences). JP-consulting fees (Chan Zuckerberg Initiative). All others had no competing interests to declare.

No funder had a role in the research or decision to publish.

References

Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J. O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, Arkadiusz Nowaczynski, Bei Wang, Marta M. Stepniewska-Dziubinska, Shang Zhang, Adegoke Ojewole, Murat Efe Guney, Stella Biderman, Andrew M. Watkins, Stephen Ra, Pablo Ribalta Lorenzo, Lucas Nivon, Brian Weitzner, Yih-En Andrew Ban, Shiyang Chen, Minjia Zhang, Conglong Li, Shuaiwen Leon Song, Yuxiong He, Peter K. Sorger, Emad Mostaque, Zhao Zhang, Richard Bonneau, and Mohammed AlQuraishi. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods*, 21(8):1514–1524, August 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02272-z. URL https://www.nature.com/articles/s41592-024-02272-z. Publisher: Nature Publishing Group.

Gonzalo Benegas, Chengzhong Ye, Carlos Albors, Jianan Canal Li, and Yun S. Song. Genomic Language Models: Opportunities and Challenges, July 2024. URL http://arxiv.org/abs/2407.11435. arXiv:2407.11435 [cs, q-bio, stat].

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Etowah Adams, Stephen A. Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R. K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, Jonathan C. Schmok, John St John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modeling and design across all domains of life with Evo 2, February 2025. URL https://www.biorxiv.org/content/10.1101/2025.02.18.638918v1. Pages: 2025.02.18.638918 Section: New Results.

Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Christina Floristean, Anant Kharkar, Koushik Roy, Charlotte Rochereau, Gustaf Ahdritz, Joanna Zhang, George M. Church, Peter K. Sorger, and Mohammed AlQuraishi. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11):1617–1623, November 2022. ISSN 1546-1696. doi: 10.1038/s41587-022-01432-w. URL https://www.nature.com/articles/s41587-022-01432-w. Publisher: Nature Publishing Group.

Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J. Maddison, and Bo Wang. BioReason: Incentivizing Multimodal Biological Reasoning within a DNA-LLM Model, May 2025. URL https://arxiv.org/abs/2505.23579v1.

FMF. Preliminary taxonomy of ai-bio misuse mitigations. Technical report, Frontier Model Forum, July 2025a. URL https://www.frontiermodelforum.org/issue-briefs/preliminary-taxonomy-of-ai-bio-misuse-mitigations/.

FMF. Introducing the FMF's Technical Report Series on Frontier AI Frameworks, April 2025b. URL https://www.frontiermodelforum.org/updates/introducing-the-fmfs-technical-report-series-on-frontier-ai-safety-frameworks/.

Sarah Gurev, Noor Youssef, Navami Jain, and Debora S. Marks. Variant effect prediction with reliability estimation across priority viruses, August 2025. URL https://www.biorxiv.org/content/10.1101/2025.08.04.668549v1. ISSN: 2692-8205 Pages: 2025.08.04.668549 Section: New Results.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387 (6736):850–858, February 2025. doi: 10.1126/science.ads0018. URL https://www.science.org/doi/10.1126/science.ads0018. Publisher: American Association for the Advancement of Science.

Svetlana P. Ikonomova, Bruce J. Wittmann, Fernanda Piorino, David J. Ross, Samuel W. Schaffter, Olga Vasilyeva, Eric Horvitz, James Diggans, Elizabeth A. Strychalski, Sheng Lin-Gibson, and Geoffrey J. Taghon. Experimental Evaluation of AI-Driven Protein Design Risks Using Safe Biological Proxies, May 2025. URL https://www.biorxiv.org/content/10.1101/2025.05.15.654077v1. Pages: 2025.05.15.654077 Section: New Results.

Moritz U. G. Kraemer, Joseph L.-H. Tsui, Serina Y. Chang, Spyros Lytras, Mark P. Khurana, Samantha Vanderslott, Sumali Bajaj, Neil Scheidwasser, Jacob Liam Curran-Sebastian, Elizaveta Semenova, Mengyan Zhang, H. Juliette T. Unwin, Oliver J. Watson, Cathal Mills, Abhishek Dasgupta, Luca Ferretti, Samuel V. Scarpino, Etien Koua, Oliver Morgan, Houriiyah Tegally, Ulrich Paquet, Loukas Moutsianas, Christophe Fraser, Neil M. Ferguson, Eric J. Topol, David A. Duchêne, Tanja Stadler, Patricia Kingori, Michael J. Parker, Francesca Dominici, Nigel Shadbolt, Marc A. Suchard, Oliver Ratmann, Seth Flaxman, Edward C. Holmes, Manuel Gomez-Rodriguez, Bernhard Schölkopf, Christl A. Donnelly, Oliver G. Pybus, Simon Cauchemez, and Samir Bhatt. Artificial intelligence for modelling infectious disease epidemics. *Nature*, 638(8051):623–635, February 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08564-w. URL https://www.nature.com/articles/s41586-024-08564-w. Publisher: Nature Publishing Group.

Javier Mendoza-Revilla, Evan Trop, Liam Gonzalez, Maša Roller, Hugo Dalla-Torre, Bernardo P. de Almeida, Guillaume Richard, Jonathan Caton, Nicolas Lopez Carranza, Marcin Skwark, Alex Laterre, Karim Beguir, Thomas Pierrot, and Marie Lopez. A foundational large language model for edible plant genomes. *Communications Biology*, 7(1):835, July 2024. ISSN 2399-3642. doi: 10.1038/s42003-024-06465-2. URL https://www.nature.com/articles/s42003-024-06465-2. Publisher: Nature Publishing Group.

NASEM. *The Age of AI in the Life Sciences: Benefits and Biosecurity Considerations*. National Academies Press, Washington, D.C., April 2025. ISBN 978-0-309-73335-9. doi: 10.17226/28868. URL https://nap.nationalacademies.org/catalog/28868.

Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132, June 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0997-x. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0997-x.

OpenAI. GPT-5 System Card, August 2025. URL https://openai.com/index/gpt-5-system-card/.

Jaspreet Pannu, Doni Bloomfield, Robert MacKnight, Moritz S. Hanke, Alex Zhu, Gabe Gomes, Anita Cicero, and Thomas V. Inglesby. Dual-use capabilities of concern of biological AI models. *PLOS Computational Biology*, 21(5):e1012975, August 2025. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1012975. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012975. Publisher: Public Library of Science.

Tyler N. Starr, Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H.D. Crawford, Adam S. Dingens, Mary Jane Navarro, John E. Bowen, M. Alejandra Tortorici, Alexandra C. Walls, Neil P. King, David Veesler, and Jesse D. Bloom. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, 182(5):1295–1310.e20, September 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.08.012. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7418704/.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika. Tamper-Resistant Safeguards for Open-Weight LLMs, February 2025. URL http://arxiv.org/abs/2408.00761. arXiv:2408.00761 [cs].

Nicole N. Thadani, Sarah Gurev, Pascal Notin, Noor Youssef, Nathan J. Rollins, Daniel Ritter, Chris Sander, Yarin Gal, and Debora S. Marks. Learning from prepandemic data to forecast viral escape. *Nature*, 622(7984):818–825, October 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06617-0. URL https://www.nature.com/articles/s41586-023-06617-0. Publisher: Nature Publishing Group.

Dianzhuo Wang, Marian Huot, Zechen Zhang, Kaiyi Jiang, Eugene Shakhnovich, and Kevin Esvelt. Without safeguards, ai-biology integration risks accelerating future pandemics, 06 2025a. URL https://www.researchgate.net/profile/Dianzhuo-Wang/publication/392731675_Without_Safeguards_AI-Biology_Integration_Risks_Accelerating_Future_Pandemics/links/68506ef7474abd185bd91f22/Without-Safeguards-AI-Biology-Integration-Risks-Accelerating-Future-Pandemics.pdf?origin=publication_detail&_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6I

Mengdi Wang, Zaixi Zhang, Amrit Singh Bedi, Alvaro Velasquez, Stephanie Guerra, Sheng Lin-Gibson, Le Cong, Yuanhao Qu, Souradip Chakraborty, Megan Blewett, Jian Ma, Eric Xing, and George Church. A call for built-in biosecurity safeguards for generative AI tools. *Nature Biotechnology*, 43(6):845–847, June 2025b. ISSN 1546-1696. doi: 10.1038/s41587-025-02650-8. URL https://www.nature.com/articles/s41587-025-02650-8. Publisher: Nature Publishing Group.

Zaixi Zhang, Zhenghong Zhou, Ruofan Jin, Le Cong, and Mengdi Wang. GeneBreaker: Jailbreak Attacks against DNA Language Models with Pathogenicity Guidance, May 2025. URL http://arxiv.org/abs/2505.23839. arXiv:2505.23839 [cs].

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2025. URL https://arxiv.org/abs/2303.18223.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are addressed in the 4.2 section of the Discussion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not directly include theoretical results building on mathematical assumptions and proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: For security reasons, the paper deliberately avoids disclosing full information needed for its reproduction. Authors from the Evo 2 model, which was fine-tuned on human-infecting virus data in this study, originally stated in their paper: "By excluding genomic sequences of viruses that infect eukaryotes from our training data, we aimed to ensure our openly shared model did not disseminate the capability to manipulate and design pathogenic human viruses. Task-specific post-training may circumvent this risk mitigation measure and should be approached with caution." As biosecurity researchers, we fine-tuned the model to assess the efficacy of data exclusion as a safeguard. Including all information necessary to reproduce the fine-tuned model and results would defeat the purpose of attempting to limit the dissemination of the capability to manipulate and design pathogenic human viruses that the Evo 2 authors originally intended. The exclusion of methodological details to avoid reproducibility is a legitimate practice when conducting dual-use research to reduce risks from deliberate misuse that the dissemination of information or tools could enable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We decided not to provide open access to the data and code for security reasons. Authors from the Evo 2 model, which was fine-tuned on human-infecting virus data in this study, originally stated in their paper: "By excluding genomic sequences of viruses that infect eukaryotes from our training data, we aimed to ensure our openly shared model did not disseminate the capability to manipulate and design pathogenic human viruses. Task-specific post-training may circumvent this risk mitigation measure and should be approached with caution." As biosecurity researchers, we fine-tuned the model to assess the efficacy of data exclusion as a safeguard, without intending to make the fine-tuned model available. Doing so would defeat the purpose of attempting to limit the dissemination of the capability to manipulate and design pathogenic human viruses that the Evo 2 authors originally intended. This also included not fully disclosing the human-infecting viral dataset that was used for fine-tuning. Not openly publishing all results and research products is a legitimate practice when conducting dual-use research to reduce risks from deliberate misuse that the dissemination of information or tools could enable.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

Justification: The paper provides all training and test details required to understand the results. In case some details are not available, these were excluded for security reasons (see justifications to questions 4 and 5).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper suitably implements and reports measures providing information on the statistical significance of experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: For security reasons, the paper deliberately avoids disclosing full information needed for its reproduction, including on compute resources. Please see the justifications for questions 4 and 5 for details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [No]

   Justification: The paper carefully considered the NeurIPS Code of Ethics and weighed potential harms and benefits. Based on this, we decided not to publish the model and data openly and to withhold certain methodological information in the manuscript. In doing so, we hope to contribute to a discussion around biosecurity safeguards for generative AI without sharing tools or information that could directly be misused. This is not in line with the Code of Ethics requirement to "Disclose essential elements for reproducibility". However, disclosing everything openly would, in turn, break the Code of Ethics requirement on "security" to consider whether there is a risk to "cause serious accidents when deployed in real world environments." In line with the Code of Ethics we "take concrete steps to recommend or implement ways to protect against such security risks".

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the introduction and the discussion, the paper addresses the potential positive societal impacts of gLMs and the risks from biological misuse. In the discussion, it also covers risk mitigation measures for AI-enabled biorisk and gLM risks, in particular. In trying to balance the potential positive and negative impacts of the paper itself, we decide not to openly disclose the fine-tuned model, data, and certain other methodological information.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: While not describing a managed access regime or a safeguard built into the model, the paper, for security reasons, does not openly publish the fine-tuned model or related data and deliberately avoids disclosing full information needed for its reproduction.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Existing assets used in the paper were properly and sufficiently credited to the degree that the necessary information was available at the time of writing.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: For security reasons, we decided not to provide open access to the fine-tuned model, data, and code, and do not provide all methodological details that allow for reproducibility. Thus, hypothetical assets from the paper are deliberately not fully documented.

Guidelines:

- The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The paper does involve the use of LLMs in any important, original, or non-standard component.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.