Prompting Toxicity: Analyzing Biosafety Risks in Genomic Language Models

Akshay Murthy
Bothell High School
akshay.srik@gmail.com

Mengmeng Zhang
University of Pennsylvania
29mzhang@gmail.com

Aashrita Koyyalamudi
Dougherty Valley High School
koyyalamudiaashrita@gmail.com

Shanmukhi Kannamangalam
Rutgers University
shankannamangalam@gmail.com

Benjamin Liu
Stanford University
Algoverse AI Research
ben@algoverseairesearch.org

Kevin Zhu
UC Berkeley
Algoverse AI Research
kevin@algoverseacademy.com

Abstract

Genomic language models (gLMs) have transformed biomedical research by enabling large-scale generation and analysis of DNA sequences. Evo, a powerful gLM trained across multiple species, was designed to uncover patterns that link genetic variation to traits and disease risk. However, its generative capabilities raise biosafety concerns: given minimal input, Evo can produce sequences resembling those found in harmful biological agents. In this study, we analyze Evo's susceptibility to generating toxic outputs. Using a curated dataset of experimentally validated toxic bacterial sequences, we prompt Evo with partial contexts and evaluate its completions using ToxinPred3 and ToxinPred2. While reconstruction fidelity improves with longer prompts, we observe that toxic protein predictions double in the presence of prompt context. These findings highlight a pressing need to assess and regulate the use of genomic foundation models in laboratory and clinical settings, where malicious intent can lead to harmful generation.

1 Introduction

The rapid advancement of genomic foundation models (gLMs) has redefined the landscape of biological sequence modeling, enabling tasks that have been. previously reliant on labor-intensive experimentation to be addressed computationally at scale. Inspired by the architecture and learning paradigms of large language models (LLMs), these models operate directly on nucleotide sequences and have shown unprecedented generalization across species and sequence lengths.

This trend toward deep learning-driven biological modeling is also reflected in work like AlphaFold, which achieved high-accuracy protein structure prediction from amino acid sequences Jumper et al. [2021]. Although distinct from language models, AlphaFold demonstrated the capacity of neural architectures to extract structural and functional signals from raw biological sequences alone. This development reinforced the feasibility of using sequence-only modeling to solve complex biological problems and helped lay the conceptual groundwork for generative genomic models.

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

In light of these contributions, models such as Evo Meier et al. [2023], Evo2 Arc Institute [2024], and HyenaDNA Poli et al. [2023] have demonstrated the capacity to capture meaningful biological representations, ranging from mutational impacts to protein functionality. Evo was among the first to apply transformer-based architectures at genome scale, revealing how language modeling approaches could generalize to molecular sequences. Evo2, developed as an up-scaled version of Evo, increased both the size of the model and the training data: trained on OpenGenome2, it spans 9.3 trillion nucleotides and more than 128,000 genomes, positioning it as one of the most comprehensive generative models in biology to date Arc Institute [2024]. HyenaDNA, on the other hand, utilizes efficient Hyena operators to model long-range dependencies without relying on attention mechanisms, showing promise in enhancer and regulatory sequence modeling Poli et al. [2023].

These advances mark a pivotal shift in genomics, with gLMs now being used for tasks including sequence generation, genome annotation, mutation effect prediction, and synthetic genome design. Studies such as Feng et al. [2024] have benchmarked a variety of genomic models—DNABERT-2, NT-v2, HyenaDNA—across over 50 classification tasks, revealing their utility in zero-shot settings and their sensitivity to tokenization strategies, pooling layers, and training domain generalization.

However, while model capabilities have grown, research into their failure modes, particularly under adversarial or high-risk conditions, is still limited. Given that gLMs are frequently trained on publicly available genomic datasets—including pathogenic genomes and toxin-coding sequences Arc Institute [2024], Olson et al. [2022]—they may be capable of reconstructing or approximating harmful biological agents. This becomes especially dangerous when partial prompting allows the model to fill in toxic sequences, a form of structural "leakage" from benign prompts. Unlike traditional sequence alignment or motif-matching methods, generative models can synthesize novel combinations that mirror functional bioactivity.

These risks highlight the importance of evaluating not just the accuracy of genomic models, but also the conditions under which they may generate unsafe outputs. Understanding how architectural differences and prompt design affect biosafety is essential for guiding responsible use and future development of these models.

2 Problem and Motivation

Genomic foundation models (gLMs) are increasingly being adopted for critical applications such as protein engineering, genome annotation, and de novo sequence design. Their generative capacity introduces not only technical potential but also new categories of risk—particularly when models are conditioned on biologically sensitive inputs.

In parallel, studies in the LLM domain have revealed that generative models are susceptible to manipulation via adversarial or indirect prompting strategies prompting strategies Adversa [2023], Wei et al. [2022], Zou et al. [2023] . Despite these findings, the safety landscape for biological generation remains underdeveloped. Tools like Evo, Evo2, and HyenaDNA have yet to be thoroughly tested for robustness in response to toxic or bioactive prompt inputs.

This gap motivates our investigation into the safety implications of gLMs when conditioned on partial, toxicity-relevant genomic fragments. We examine how prompt length and structure influence both sequence reconstruction fidelity and the generation of potentially harmful biological outputs across different model families. into the safety implications of genomic language models (gLMs) when conditioned on partial, toxicity-relevant prompts.

We examine the extent to which prompt length influences both the accuracy of sequence reconstruction and the generation of toxic content across multiple gLM architectures. Specifically, we evaluate whether providing partial genomic contexts can induce biologically unsafe completions, and how this behavior varies as a function of both model architecture and prompting strategy.

3 Related Work

Genomic foundation models (gLMs) have recently gained attention for their ability to model and generate biological sequences. Evo, Evo2, and HyenaDNA represent key advancements in this space. Evo established the use of transformer-based architectures for modeling biological sequences, while Evo2 extended this framework to support over 9.3 trillion nucleotides spanning more than

128,000 genomes. HyenaDNA introduced Hyena operators as a computationally efficient alternative to attention mechanisms, enabling more scalable training on long genomic contexts.

While not a genomic language model, AlphaFold demonstrated that deep learning architectures can accurately infer protein structure from sequence alone, without relying on co-evolutionary features or templates Jumper et al. [2021]. This result reinforced the feasibility of learning biologically meaningful representations directly from raw sequences, and helped lay conceptual groundwork for models like Evo and HyenaDNA that aim to capture structure and function through generative modeling.

While the architectural innovation in these models is significant, prior work has largely emphasized predictive performance rather than generation safety. For example, benchmarked models like DNABERT-2 and HyenaDNA across 57 classification tasks, evaluating runtime efficiency and generalization ability. However, these evaluations primarily focused on embedding quality and did not assess whether the models could produce biologically harmful outputs.

In parallel, research on large language models has revealed how prompting techniques can lead to unintended or unsafe outputs. Techniques such as Best-of-N sampling Zou et al. [2023], Chain-of-Thought prompting Wei et al. [2022], and jailbreak taxonomies Adversa [2023] have been used to stress-test model alignment. These strategies, while developed for text-based models, suggest similar risks could emerge in genomic contexts, especially where training data includes pathogenic or toxin-related sequences.

To our knowledge, few studies have investigated how gLMs behave when prompted with toxicity-relevant inputs. Most safety work in bioinformatics has focused on static toxicity prediction using tools like ToxinPred2 Sharma et al. [2022], rather than integrating these classifiers into the generative evaluation of modern language models. Our study contributes to this underexplored area by adversarially probing multiple gLMs to characterize prompt-induced toxicity and assess their biosafety robustness.

4 Methodology

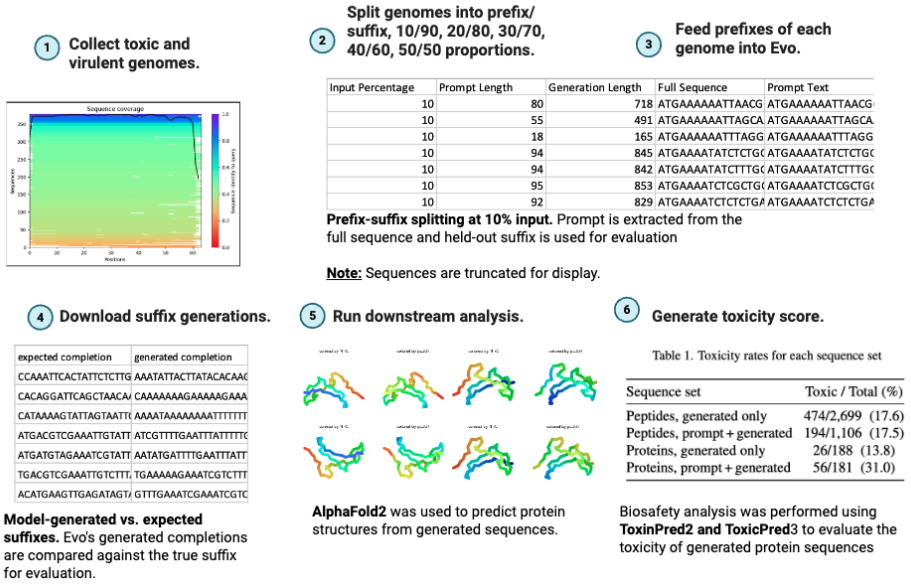

Figure 1: Prompting Pipeline

4.1 Data Collection

Bacterial DNA sequences were collected as FASTA files from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) dataset. We filtered for genomes that were representative genomes, occured in mammals, and for their sequences to exude toxic traits in vivo to ensure biological relevance. We ended up with a total 185 after filtering constraints.

4.2 Prompting Strategies

4.2.1 Model Initialization and Tokenization

We choose to use `evo-1-8k-base` which is the base Evo 1 model with 8192 context accessed through Evo's HuggingFace repository in order to. All generated tokens are the standard DNA nucleotides $\{A, G, C, T\}$.

4.2.2 Dataset Preparation

All available FASTA files from the BV-BRC data are merged, upper-cased, stripped of whitespace, and deduplicated. No sequences were above the 8192 token context length requirement of `evo-1-8k-base`, yielding 185 unique samples $\{s_i\}_{i=1}^{185}$ with $|s_i| \leq 8192$.

4.2.3 Prompt Construction

For each s_i we create five deterministic prompts consisting of $\{10, 20, 30, 40, 50\}\%$ of the input sequence. We then prompt the model to generate its best fit rendition of the corresponding $\{90, 80, 70, 60, 50\}\%$ DNA sequence.

4.2.4 Decoding Policy

Continuations are generated with strict greedy decoding. No temperature, nucleus, or top-k sampling is applied. This is done to ensure reproducibility and that the generation is the model's most probable output.

4.2.5 Accuracy Metric

We report the normalised Needleman–Wunsch alignment score

$$\mathrm{NAS}(\hat{t},\, t) = \frac{\mathrm{NW}(\hat{t}, t)}{2\,|t|} \in [0, 1],$$

with match=$+2$, mismatch=-1, gap-open=-2, gap-extend=-0.5. This is designed to mimic BLAST similarity which shows how similar two sequences are to each other. In this case, we compare each suffix sequence generated by Evo to the corresponding portion of the input/ground-truth sequence.

4.3 Toxicity Evaluation

We chose to evaluate toxicity through multiple metrics. The first is the normalized BLAST-like score above to measure sequence similarity. We choose to use two sets of sequences: the prompt and Evo's generation, versus Evo's generation alone. We then convert each set's DNA sequences to protein and peptide sequences through `BioPython`, transcribing from all 6 Open Reading Frames, 3 from the forward strand and 3 from reading from the reverse strand. We define peptides to be any sequence to be greater than or equal to 5 AA and less than 50 AA to account for functional purposes, and proteins to be any sequence greater than 50 AA. Then, we filter out any duplicate sequences from transcribing from multiple ORFs, and make sure the amino acid sequences derived from the prefix + generation are distinct from the transcribed sequences from just the generation. Then we choose to use ToxPred 3.0 and ToxPred 2.0 to classify each peptide/protein as toxic or non-toxic.

4.4 Compute Resources

All experiments were conducted on an NVIDIA H200 GPU with 141 GB of memory, accessed through RunPod. Each experimental run (generating completions for 185 sequences across 5 prompt lengths) required approximately 5 hours of execution time.

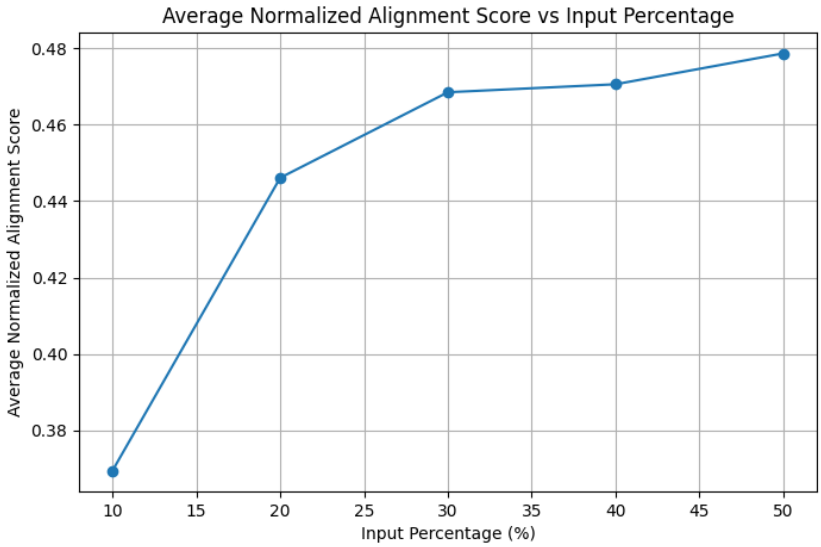

Figure 2: Normalized Alignment Score per Input Percentage

5 Results

Evo's alignment performance improves substantially with increased prompt length but begins to saturate beyond the 30% input threshold. Specifically, providing just the first 30% of a toxic reference sequence yields a normalized alignment score of 0.47, nearly matching the maximum score of 0.48 achieved at 50%. This suggests that Evo recovers the core sequence structure early in generation, with further context contributing minimally to reconstruction fidelity.

Table 1: Toxicity rates for each sequence set

Sequence set	Toxic / Total (%)
Peptides, generated only	474/2,699 (17.6)
Peptides, prompt + generated	194/1,106 (17.5)
Proteins, generated only	26/188 (13.8)
Proteins, prompt + generated	56/181 (31.0)

ToxinPred2 and ToxinPred3 were used to classify each protein and peptide sequence respectively as toxic or non-toxic. In terms of toxicity, peptide sequences exhibited consistent rates across both generation strategies, with 17.6% of `generated-only` sequences and 17.5% of `prompt plus generated` sequences classified as toxic by ToxinPred2. However, protein sequences showed a more pronounced shift: the proportion of toxic proteins rose from 13.8% in the generated-only condition to 31.0% when prompts were included.

Taken together, these results show that prompting enhances fidelity with diminishing returns beyond a third of the input, and that while peptide toxicity is stable, protein toxicity appears sensitive to context length and composition. This pattern may reflect the model's capacity to extend functional or pathogenic motifs present in the input prompt, particularly in longer and more complex protein-coding regions. Notably, the sharp increase in protein toxicity suggests that longer prompts could amplify bioactive patterns embedded in the original sequence, even without direct supervision. These findings highlight the need to further investigate sequence-level attributes that predispose models to unsafe completions, especially for outputs with known or suspected functional roles.

6 Limitations

This study has a few limitations that may affect the generalization and interpretation of the results. The genomic inputs used in the experiments consist of partial sequences rather than complete genomes. While these fragments were selected for relevance to toxic biological traits, they may not fully capture the structural and regulatory complexity of intact genomic contexts.

Additionally, the model outputs often contained repetitive base patterns. These patterns may reduce the biological realism of the generations and potentially inflate alignment metrics or affect toxicity classification outcomes.

These limitations should be considered when interpreting the findings, and future work should aim to incorporate full-genome inputs and further investigate the impact of prompt structure on the fidelity and safety of generated sequences.

7 Conclusion

This study demonstrates that genomic foundation models (gLMs), even when prompted with limited sequence context, can generate biologically plausible continuations with high alignment fidelity. Increasing prompt length was found to improve sequence reconstruction, but also introduced risks: protein sequences generated from partial contexts exhibited significantly higher toxicity rates compared to their unprompted counterparts.

These findings reveal a trade-off between generative accuracy and biosafety in gLMs. While longer prompts enhanced the recovery of biological structure, they also increased the likelihood of producing potentially harmful outputs. Peptide toxicity remained stable across prompt conditions, but protein toxicity nearly doubled—underscoring the need for targeted safety mechanisms, especially when generating longer and more functionally relevant bio-molecules.

Our evaluation contributes an initial framework for prompt-based adversarial analysis in gLMs, showing that even well-aligned models can be steered toward undesirable behavior through prompt manipulation. However, given that our experiments used partial genomic inputs and that model outputs frequently contained repetitive base patterns, the biological realism of these sequences may be limited. These constraints should be considered when interpreting biosafety implications, and future studies should evaluate full-length genomic contexts and a broader range of model architectures.

As gLMs become more widely integrated into synthetic biology workflows, we emphasize the need for pre-deployment safety audits, context-aware decoding strategies, and toxin-aware post-generation filtering. Future work should incorporate empirical validation of toxicity predictions and establish more comprehensive standards for responsible generative model use in biological domains.

References

Adversa. Universal llm jailbreak: Chatgpt, gpt-4, bard, bing, anthropic, and beyond. `https://adversa.ai/blog/universal-llm-jailbreak-chatgpt-gpt-4-bard-bing-anthropic-and-beyond/`, 2023. Accessed: 2025-05-07.

Arc Institute. Evo2: A foundational model for biology. `https://arcinstitute.org/news/evo2-foundational-model-for-biology`, 2024.

Haonan Feng, Lang Wu, Bingxin Zhao, Chad Huff, Jianjun Zhang, Jia Wu, Lifeng Lin, Peng Wei, and Chong Wu. Benchmarking dna foundation models for genomic sequence classification. *bioRxiv*, 2024. doi: 10.1101/2024.01.02.39185205.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michal Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Pushmeet Kohli, and Demis Hassabis.

Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Joshua Meier, Yannan Zhang, Xiaotong Li, et al. Sequence modeling and design from molecular to genome scale with evo. *bioRxiv*, 2023. doi: 10.1101/2023.02.13.528211.

Robert D. Olson, Rima Assaf, Thomas Brettin, et al. Introducing the bacterial and viral bioinformatics resource center (bv-brc): A resource combining patric, ird, and vipr. *Nucleic Acids Research*, 2022. doi: 10.1093/nar/gkac1006.

Marc Poli, Randall Balestriero, Maxwell Shinn, et al. Hyenadna: Learning long-range genomic dependencies with hyena operators. `https://arxiv.org/abs/2306.11459`, 2023. arXiv preprint arXiv:2306.11459.

Neelam Sharma, Leimarembi Devi Naorem, Shipra Jain, and Gajendra P. S. Raghava. Toxinpred2: An improved method for predicting toxicity of proteins. *Briefings in Bioinformatics*, 23(5), 2022. doi: 10.1093/bib/bbac250.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain of thought prompting elicits reasoning in large language models. `https://arxiv.org/abs/2201.11903`, 2022. arXiv preprint arXiv:2201.11903.

Andy Zou, Xuechen Wang, Tom Goldstein, et al. Universal and transferable adversarial attacks on aligned language models. `https://arxiv.org/abs/2307.15043`, 2023. arXiv preprint arXiv:2307.15043.

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 abstract and introduction clearly state the contributions: analyzing Evo's susceptibility to generating toxic sequences, measuring reconstruction fidelity and toxicity across prompt lengths, and highlighting biosafety implications. These claims are consistent with the results presented in Sections 5 (Results) and 7 (Conclusion).

   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: Section 6 explicitly discusses limitations: use of partial genomes rather than complete ones, potential inflation of alignment scores due to repetitive base patterns, and reduced biological realism. These constraints are openly acknowledged.

   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 present new theorems or formal proofs. The work is empirical, relying on experiments and evaluations of gLM outputs.

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: [Yes]

Justification: The methodology section (4) describes data sources (BV-BRC), preprocessing steps, prompt construction, decoding policy, and alignment metric. While code is not released, the description provides sufficient detail to replicate the experiments given access to the same model and dataset.

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 do not release the code or generated sequences due to biosafety risks. The methodology (Section 4) provides sufficient detail for reproduction with access to Evo and BV-BRC data, but releasing outputs directly could enable malicious use.

   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: Section 4 details dataset preparation, model initialization, decoding policy, and scoring. Hyperparameters are minimal because greedy decoding was used with no sampling. These choices are clearly stated.

   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: [No]

Justification: Results are presented as point values (alignment scores and toxicity rates) without error bars, variance, or confidence intervals. While trends are clear, statistical robustness is not quantified.

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: [Yes]

   Justification: Section 4.4 specifies the compute setup. All experiments were run on a single NVIDIA H200 GPU with 141 GB of memory via RunPod. Each full set of runs across 185 sequences and five prompt-length conditions required approximately 5 hours of execution.

   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: [Yes]

   Justification: The work complies with NeurIPS Code of Ethics. The paper discusses biosafety risks, emphasizes responsible use, and does not release unsafe assets. Anonymity is preserved.

   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: Section 7 (Conclusion) and throughout the paper, both positive (scientific insight, safety evaluations of gLMs) and negative (misuse potential for harmful sequence generation) societal impacts are discussed. The need for safeguards and audits is highlighted.

    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 no assets are released, the paper explicitly recommends safeguards such as pre-deployment safety audits, toxin-aware post-generation filtering, and context-aware decoding (Section 7). These are responsible precautions for high-risk applications.

    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: The BV-BRC dataset (Olson et al. 2022) and Evo model (Meier et al. 2023) are properly cited. Both are public research resources. Licensing details are handled via their respective repositories.

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: [NA]

Justification: No new datasets, models, or benchmarks are released. The contribution is analysis and evaluation, not asset release.

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 study does not involve human subjects or crowdsourcing.

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 research does not involve human participants, thus IRB approval is not applicable.

    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 study uses genomic language models (gLMs) such as Evo, but not general-purpose LLMs for methods. Any AI tools used for editing are outside the methodology, so declaration is unnecessary.

    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.