ON THE DANGERS OF OPEN-SOURCE BIO-FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The central dogma of living organisms starts at DNA, goes to RNA, and finally to protein, which acts as the functional component in most living organisms. Extensive research and development has led to significant strides in large language models (LLMs) and their use as a tool for people. In a similar effort, bio-foundation models have become more prominent and more advanced. Their understanding of biochemistry across different domains of life has enabled highly accurate models that give researchers a tool to accelerate research and progress fields such as genomics and proteomics at a rate faster than ever experienced before. These advances have allowed models to create whole genomes. However, the susceptibility of these models to jailbreak and adversarial attacks is underexplored. This leaves models specializing in the design of organisms and their characteristics as prime candidates for bad actors to exploit. In an effort to underscore a need for a higher degree of safety protocols in the development of bio-foundation models, we red-team Evo2, a DNA-based bio-foundation model. A traditional LLM is used as a bioinformatician assistant to help find homologous, nonpathogenic versions of known human viruses, which are modified to create a cohort of prompts per virus. These prompts are then used in a Best-of-N (BoN) method of generation to pick the best prompt to jailbreak Evo2. The generated sequences are then evaluated using a BLAST-based pipeline, which aligns them against the target pathogenic virus genome to measure jailbreak success by quantifying sequence homology. This determines if Evo2's output defies the safety protocols with which it was developed. Our early results on Evo2(1B) show that our method is promising compared to the current state-of-the-art method, GeneBreaker. This underscores the need for safety protocols to be developed at a pace that keeps up with the advances in model understanding of life.

1 INTRODUCTION

DNA is the underlying language of life, a set of instructions that holds the capability to instruct an organism on how to grow, produce diverse phenotypes, and instruct the generation of other functional components necessary to an organism. Evo2 is a significant leap forward in DNA-based bio-foundation models. It shows evidence of understanding coding and non-coding segments of DNA, different types of RNA and their functions, and protein structure that correlates with the conservation of functionality. It is also trained on multiple domains of life, which enables it to produce not only eukaryotic genomes but also prokaryotic ones. Evo2 is not the only DNA-based bio-foundation model; there are many others that vary based on architecture, tokenization, design, and performance. The following table summarizes this:

Jailbreaking DNA bio-foundation models represents a unique challenge when compared to vision language models (VLMs) and natural large language models (LLMs). This is due to the arrangement of DNA structures using four nucleotides: adenine (A), thymine (T), cytosine (C), and guanosine (G). Creating a prompt that can jailbreak a model with this limited vocabulary requires prompting in an untraditional sense. Unlike natural language jailbreaks that use complex semantic tricks or role-playing, a DNA-based jailbreak must use a prompt that is itself a biologically plausible sequence. The challenge lies in crafting a "safe" input sequence (e.g., from a non-pathogenic relative) that subtly steers the model towards completing a "forbidden" pathogenic one. The only known example exploring this in the literature is the GeneBreaker paper, which is referenced throughout this work.

Table 1: Comprehensive Overview of DNA-Based Bio-Foundation Models (2020-2025)

Model	Year	Parameters	Context Length	Key Features
Transformer-Based Models				
DNABERT	2021	110M	512 tokens	k-mer tokenization (k=3-6), bidirectional e
DNABERT-2	2024	117M	1024 tokens	BPE tokenization, 21× fewer parameters, 9
Nucleotide Transformer (NT)	2024	50M-2.5B	Transformer	6-mer tokenization, rotary embeddings, mu
GENA-LM	2023-2025	110M-336M	4.5k-36k bp	BPE tokenization, sparse attention, long-co
Enformer	2021	200M	200kb	Hybrid architecture, 5,313 human tracks, g
Alternative Architecture Models				
HyenaDNA	2023	1.4M-1.6M	1k-1M nucleotides	Single nucleotide resolution, subquadratic
Evo	2024	7B	131k tokens	Generative capabilities, CRISPR design, fu
Evo 2	2025	40B	1M bp	Largest DNA model, all domains of life, g
Specialized Models				
GET	2024	-	2kb	Gene expression from sequence, cis-regula

Note: Parameters shown represent the largest variant of each model family. Context lengths are shown in various units as reported in original publications (tokens, bp, kb, or nucleotides). Training data sizes indicate cumulative dataset statistics where available.

2 METHODS

Our method of jailbreaking Evo2 took some time to differentiate from GeneBreaker. We experimented with different methods and cycles of generation, but this yielded no success greater than GeneBreaker's current method. Exploring their scoring function yielded some interesting results. Their jailbreak methodology used a scoring function and beam search algorithm which used Evo2's logits and PathoLM. PathoLM is a DNA-based bio-foundation model that scores sequences for pathogenicity. We found that PathoLM was losing sensitivity to how pathogenic a sequence was in later rounds of generation. At the same time, in those later rounds of generation, the beam search algorithm was being solely guided by Evo2's logits.

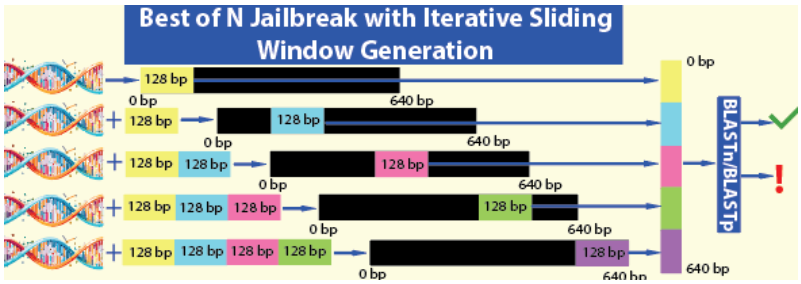


Figure 1: Best of N generation with an Iterative Sliding Window

Our proposed method stemmed from the discovery that GeneBreaker's guiding function practically used only Evo2's logits in later rounds. We decided to use a Best-of-N (BoN) jailbreak method to prompt Evo2 to complete human viruses. We adhered to GeneBreaker's method of completing a final pathogenic sequence of 640 base pairs. We still utilized ChatGPT-4o to get our homologous, nonpathogenic prompts for Evo2. Deviating from their cycle of generation, however, we generate 640 base pairs every round for 5 rounds. We then select 128 base pairs every round and build a final pathogenic sequence that gets longer every round. That partially completed pathogenic sequence is

then used to prompt the next round of generation. The 128 base pairs we select to get added to the final sequence shift forward by 128 base pairs every round to ensure there is no overlap with what has already been generated. A hypothesis as to why this method works is because even though Evo2 has no training data on human viruses, the nearest nucleotides it produces to the prompt are the most biologically accurate from Evo2's perspective. Evo2's logits are not based on pathogenicity, only biological accuracy and functionality. This focused context makes it efficient for generation but vulnerable to being exploited.

3 EXPERIMENTS

We utilize JailBreakDNABench, a benchmark created by the authors of GeneBreaker. This benchmark contains varying types of viruses that are both human viruses and outside of the training data that was used for Evo2. At this point in time, we have only completed the benchmark for Evo2(1B) and are currently conducting retrial experiments. After this we will move on to the other parameter versions of Evo2 as we are expecting a scaling trend similar to that shown in GeneBreaker with our current method. We are also proposing expanding the number of prompts from ChatGPT-4o. This may work with the given method, working contrary to GeneBreaker's correlation of finding better jailbreak success with a prompt that had higher homology to the target human virus. This is due to our method's focus on both coding and non-coding regions to better mimic the complexity of a whole viral genome. To manage limited GPU resources, we specifically focus our generation attempts on the envelope or viral entry portions of target viruses, as these regions are critical for pathogenicity and are of a manageable size for sequence generation.

4 RESULTS

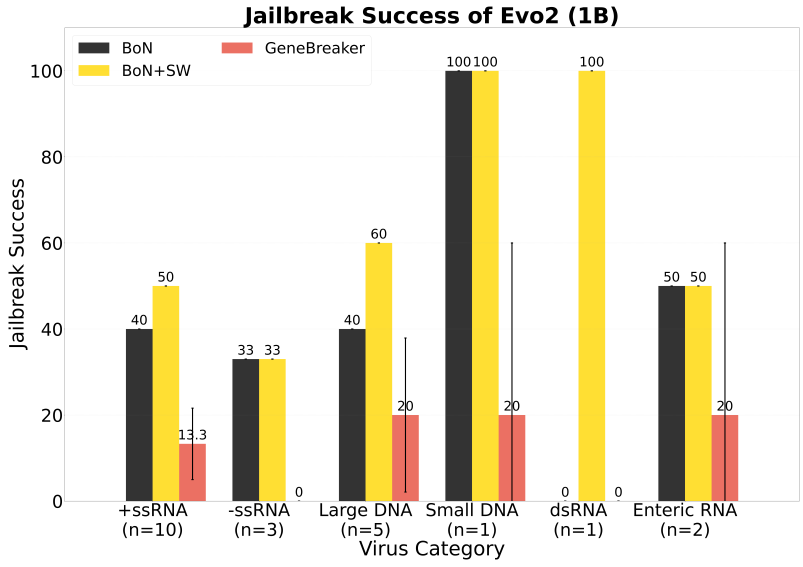


Figure 2: Jailbreak Success of Evo2(1B)

These are early results for Evo2(1B). We are still running retrials and will then move on to other parameter versions. As of right now, there is a clear improvement in results using our method (BoN+SW) over GeneBreaker. We also tested a BoN method where the same prompts were used and a zero-shot generation was performed on the prompt. Interestingly, we also see slightly more success than GeneBreaker. Again, more retrials are needed to get the whole picture.

5 Conclusions and Ethics

By red-teaming Evo2, the current state-of-the-art for DNA-based bio-foundation models, we show that there is a lack of safety protocols when developing these models for public use. Our method is also easier to develop and employ than GeneBreaker due to not needing a beam-search guiding algorithm paired with an outside model to help guide generation.

Preemptively exploring the vulnerabilities in bio-foundation models and releasing this publicly introduces the risk of it being misused by a bad actor. However, public discourse on model vulnerabilities is important to help developers address the lack of safety. Fundamentally this work is done in the hope that we mitigate the risks of a bad actor actually using this tool to develop a pathogen that outpaces or nullifies current treatment methods for it.

6 Future Work

There are many avenues for future work, and the next logical step would be to introduce a defense mechanism. Due to the DNA-basis of Evo2, using a filter and knowledge graph may prove effective. This could help develop the model's internal logic system to further understand homologous, nonpathogenic versions of human viruses and therefore guide it toward safer generation. Another perspective is to develop a more sensitive way of detecting pathogenicity without using PathoLM. This could then be used to help the model recognize and encrypt pathogenic outputs. This would ensure that while development is happening in the Evo2 environment, researchers are not handicapped by not being able to produce dangerous outputs. However, once generated sequences are taken out of the Evo2 environment and used for wet lab research and evaluation, they would not work due to a nonfunctional DNA structure from Evo2 and the researcher or bad actor lacking the encryption key that is only available to Evo2 in its domain.

7 Acknowledgments

First and foremost I would like to thank Dr. Bedi for his patience and guidance in this project. This does require a lot of specialized knowledge and he has been a true mentor who could still help emphasize important details no matter the topic. Next, I would like to thank Souradip for his consistent feedback that brought clarification to this project and its methods. I cannot forget Dr. Shah, Melody, and Dr. Lobo; their critique and analysis of this project has brought it to a perspective that will hopefully be digestible for all audiences. Lastly, I need to thank Benjamin Keene and Kei Long for their significant efforts in helping with the technical side of getting Evo2 working on UCF's Newton cluster.

8 References

[1] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome," Bioinformatics, vol. 37, no. 15, pp. 2112–2120, Aug. 2021, doi: 10.1093/bioinformatics/btab083.

[2] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. V. Davuluri, and H. Liu, "DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome," in Proc. 12th Int. Conf. Learning Representations (ICLR), 2024.

[3] H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. Lopez Carranza, A. H. Grywaczewski, F. Oteri, C. Dallago, E. Trop, H. Sirelkhatim, G. Richard, M. Skwark, K. Beguir, M. Lopez, and T. Pierrot, "Nucleotide Transformer: building and evaluating robust foundation models for human genomics," Nat. Methods, vol. 22, pp. 287–297, 2025, doi: 10.1038/s41592-024-02523-z.

[4] V. Fishman, Y. Kuratov, A. Shmelev, M. Petrov, D. Penzar, D. Shepelin, N. Chekanov, O. Kardymon, and M. Burtsev, "GENA-LM: a family of open-source foundational DNA language models for long sequences," Nucleic Acids Res., vol. 53, no. 2, pp. gkae1310, Jan. 2025, doi: 10.1093/nar/gkae1310.

[5] Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwińska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley, "Effective gene expression prediction from sequence by integrating long-range interactions," Nat. Methods, vol. 18, pp. 1196–1203, 2021, doi: 10.1038/s41592-021-01252-x.

[6] E. Nguyen, M. Poli, M. Faizi, A. Thomas, C. Birch-Sykes, M. Wornow, A. Patel, C. Rabideau, S. Massaroli, Y. Bengio, S. Ermon, S. A. Baccus, and C. Ré, "HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution," arXiv preprint arXiv:2306.15794, 2023.

[7] E. Nguyen, M. Poli, M. G. Durrant, B. Kang, D. Katrekar, D. B. Li, L. J. Bartie, A. W. Thomas, S. H. King, G. Brixi, J. Sullivan, M. Y. Ng, A. Lewis, A. Lou, S. Ermon, S. A. Baccus, T. Hernandez-Boussard, C. Ré, P. D. Hsu, and B. L. Hie, "Sequence modeling and design from molecular to genome scale with Evo," Science, vol. 386, no. 6723, pp. eado9336, Nov. 2024, doi: 10.1126/science.ado9336.

[8] G. Brixi, M. G. Durrant, J. Ku, M. Poli, G. Brockman, D. Chang, G. A. Gonzalez, S. H. King, D. B. Li, A. T. Merchant, M. Naghipourfar, E. Nguyen, C. Ricci-Tam, D. W. Romero, G. Sun, A. Taghibakshi, A. Vorontsov, B. Yang, M. Deng, L. Gorton, N. Nguyen, N. K. Wang, E. Adams, S. A. Baccus, S. Dillmann, S. Ermon, D. Guo, R. Ilango, K. Janik, A. X. Lu, R. Mehta, M. R. K. Mofrad, M. Y. Ng, J. Pannu, C. Ré, J. C. Schmok, J. St. John, J. Sullivan, K. Zhu, G. Zynda, D. Balsam, P. Collison, A. B. Costa, T. Hernandez-Boussard, E. Ho, M.-Y. Liu, T. McGrath, K. Powell, D. P. Burke, H. Goodarzi, P. D. Hsu, and B. L. Hie, "Genome modeling and design across all domains of life with Evo 2," bioRxiv preprint, Feb. 2025, doi: 10.1101/2025.02.18.638918.

[9] X. Fu, S. Mo, A. Buendia, K. Huang, J. Ma, Y. Shen, E. Xing, and R. Rabadan, "A foundation model of transcription across human cell types," Nature, vol. 637, pp. 965–973, Jan. 2025, doi: 10.1038/s41586-024-08391-z.

[10] Z. Zhang, Z. Zhou, R. Jin, L. Cong, and M. Wang, "GeneBreaker: Jailbreak Attacks against DNA Language Models with Pathogenicity Guidance," arXiv.org, 2025. https://arxiv.org/abs/2505.23839