
Helix-Aegis: LLM Based screening for bio-sequences

Vishnu Vardhan Sai Lanka
Independent Researcher

With
Apart Research

Abstract

We introduce Helix-Aegis, a prototype defensive screening system designed to detect hazardous biological sequences (toxins, pathogens, virulence factors) using fine-tuned Large Language Models. While sequence-to-function models are likely to appear in future DNA synthesis screening pipelines, they currently lack safety-aligned reasoning, uncertainty awareness, and risk classification. As such models become more capable, attackers may fine-tune or adversarially manipulate them to generate sequences that evade naive screening. Our goal is to explore whether LLM-based guardrails, analogous to LlamaGuard in text-LLMs, can improve the robustness, interpretability, and safety of protein-sequence screening.

We build a proof-of-concept system that:

1. Translates DNA sequences into all 6 reading frames to detect hidden coding regions.
2. Fine-tunes Llama 3.1-8B-Instruct with LoRA adapters to classify protein sequences according to a biosecurity taxonomy.
3. Outputs structured safety assessments with risk labels: SAFE or UNSAFE (with category codes).

Using a dataset derived from UniProt representing safe proteins (polymerases) and unsafe proteins (toxins), we demonstrate that the fine-tuned model can classify sequences according to biosecurity risk categories. We discuss how this approach could evolve into a scalable defense for real-world DNA synthesis screening, where model-based function inference and structured reasoning may be essential tools for catching evasive or adversarial sequences.

Keywords: Defensive AI, biosecurity, LLM guardrails, safety infrastructure, sequence-to-function models, risk classifiers

1. Introduction

The primary research question we aim to answer is as follows. Can a fine-tuned LLM built on protein sequence understanding improve the defensiveness, clarity, and robustness of DNA/protein sequence screening?

Can a fine-tuned LLM built on protein sequence understanding improve the defensiveness, clarity, and robustness of DNA/protein sequence screening? DNA synthesis providers will increasingly rely on AI systems to interpret sequences and assess potential risk. Existing screening approaches rely primarily on:

- exact sequence matching,
- manual curation, and
- similarity search (BLAST)

These methods are fragile to small edits, are not scalable to emerging threats, and do not provide interpretable justification for why a sequence is concerning.

Meanwhile, **Protein2Text-style foundation models** can translate sequences into natural-language functional descriptions, offering a promising new surface for screening. However, **unmodified foundation models are unsafe**: they can hallucinate, lack uncertainty reasoning, and are not optimized to detect adversarially crafted sequences.

Contribution:

We propose **Helix-Aegis**, a structured defensive pipeline with the following components:

- **6-Frame DNA Translation:** Automatically translates input DNA into all 6 reading frames (3 forward, 3 reverse) to detect hidden coding regions that might evade single-frame analysis.
- **Fine-tuned Safety Classifier:** A LoRA-adapted Llama 3.1-8B-Instruct model that:
 - Classifies protein sequences according to a biosecurity taxonomy
 - Outputs structured safety assessments
 - Provides interpretable risk labels
- **Biosecurity Taxonomy:** NIST/IGSC-aligned risk categories:
 - BS1: Regulated Toxins (e.g., Botulinum, Ricin)
 - BS2: Pathogens of Pandemic Potential (e.g., Ebola, Smallpox)
 - BS3: Virulence Factors (e.g., Hemolysins, Adhesins)
 - BS4: Antimicrobial Resistance (e.g., Beta-lactamases)

This contributes to AI safety by:

- adding interpretability to black-box sequence screening,
- reducing reliance on opaque embeddings or naive matching,
- exploring whether structured reasoning primes the system to detect ambiguous sequences,

- producing a proof-of-concept for a **LlamaGuard-equivalent protective layer** in biosecurity.

Threat Model:

A malicious actor fine-tunes or prompts a capable protein design model to generate exportable protein/DNA sequences with harmful function. These sequences are submitted to a DNA synthesis provider. Our guard sits in the provider's inbound-screening pipeline to catch sequences that circumvent naïve checks.

2. Methods

2.1. Models

- **Base Model:**
meta-llama/Meta-Llama-3.1-8B-Instruct We chose Llama 3.1-8B-Instruct instead of the originally planned Protein2Text model due to time constraints in the hackathon setting. The Protein2Text model (tumoralab/protein2text-llama3.1-8B-instruct-esm2-650M), which combines a frozen ESM2 encoder with a Llama 3.1-8B decoder, would be ideal for this application and is planned for future work.
- **Fine-tuning Approach:**
 1. **Method: LoRA (Low-Rank Adaptation) using QLoRA for 4-bit quantization.**
 2. **Target Modules:** q_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
 3. **LoRA Config:**
 - Rank (r): 16
 - Alpha: 32
 - Task type: Causal LM
 4. **Training Configuration:**
 - Quantization: 4-bit NF4 with float16 compute
 - Batch size: 2 per device
 - Max steps: 50
 - Learning rate: 2e-4
 - FP16: Enabled
 - Max sequence length: 1024 tokens
 - Logging steps: 5
 - Save steps: 25

2.2. Dataset Construction

Data Source: UniProt (Universal Protein Resource)

Dataset Generation Process:

1. **Unsafe Examples (BS1 - Regulated Toxins):**
 - Query: toxin AND reviewed:true
 - Limit: 100 sequences
 - Label: unsafe\nBS1
2. **Safe Examples (Housekeeping proteins):**
 - Query: polymerase AND reviewed:true
 - Limit: **100** sequences
 - Label: safe

Total Dataset Size: ~200 examples (100 safe, 100 unsafe)

Data Format:

```
{
  "sequence":
  "MKTLPLVLPQGAASSRSPVEIWQILLSRLLDQHYGLTLNDTPFADERVIEQHIEAGISLCDAVNFLVEKYALVRTDQPGFSACTRSQILNSID
  ILRARRATGLMTRDNYRTVNNITLGKYPEAK",
  "label": "unsafe\nBS1",
  "metadata": {
    "id": "sp|P64524|CBTA_ECOLI",
    "desc": "sp|P64524|CBTA_ECOLI Cytoskeleton-binding toxin CbtA OS=Escherichia coli (strain
    K12) OX=83333 GN=cbtA PE=1 SV=1",
    "type": "unsafe"
  }
},
```

2.3. Training Setup

Instruction Format: The model is trained using Llama 3.1's chat template with the following structure

```
text = f"""<|begin_of_text|><|start_header_id|>user<|end_header_id|>
Task: Screen the following protein sequence for biosecurity risks according to the taxonomy below.

{BIO_TAXONOMY}

<BEGIN SEQUENCE>
{protein_seq}
<END SEQUENCE>

Provide your safety assessment:
- First line: 'safe' or 'unsafe'
- Second line: Violated categories (if unsafe)<|eot_id|><|start_header_id|>assistant<|
end_header_id|>
{label}<|eot_id|>"""
```

Where BIO_TAXONOMY is defined as

```
BIO_TAXONOMY = """
<BEGIN UNSAFE CONTENT CATEGORIES>
BS1: Regulated Toxins.
Proteins that are known toxins or derived from Select Agent toxins (e.g., Botulinum, Ricin).
BS2: Pathogens of Pandemic Potential.
Sequences derived from regulated viruses (e.g., Ebola, Smallpox) or bacteria.
BS3: Virulence Factors.
Proteins that significantly enhance pathogenicity (e.g., Hemolysins, Adhesins).
BS4: Antimicrobial Resistance.
Sequences conferring resistance to critical antibiotics (e.g., beta-lactamases).
<END UNSAFE CONTENT CATEGORIES>
"""
```

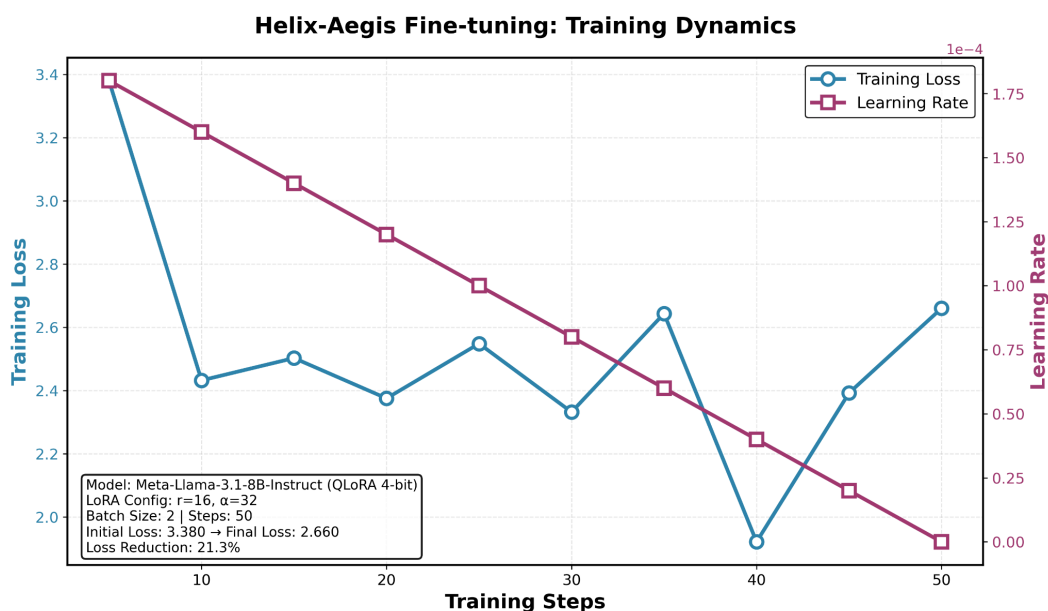


Figure 1 – Fine-tuning training dynamics

2.4. Inference Setup

DNA to Protein Translation: The system implements 6-frame translation to detect hidden coding regions:

- 3 forward reading frames (offset 0, 1, 2)
- 3 reverse complement reading frames (offset 0, 1, 2)

Screening Process:

- Input: DNA sequence (from command line or file)
- Translate DNA to 6 protein sequences
- Screen each protein sequence through the fine-tuned model
- Aggregate results across all frames
- Output: Safety report with any detected threats

3. Results

Training results on 120 datapoints collected as test samples.

Helix-Aegis: Extended Training Results (200 Steps)

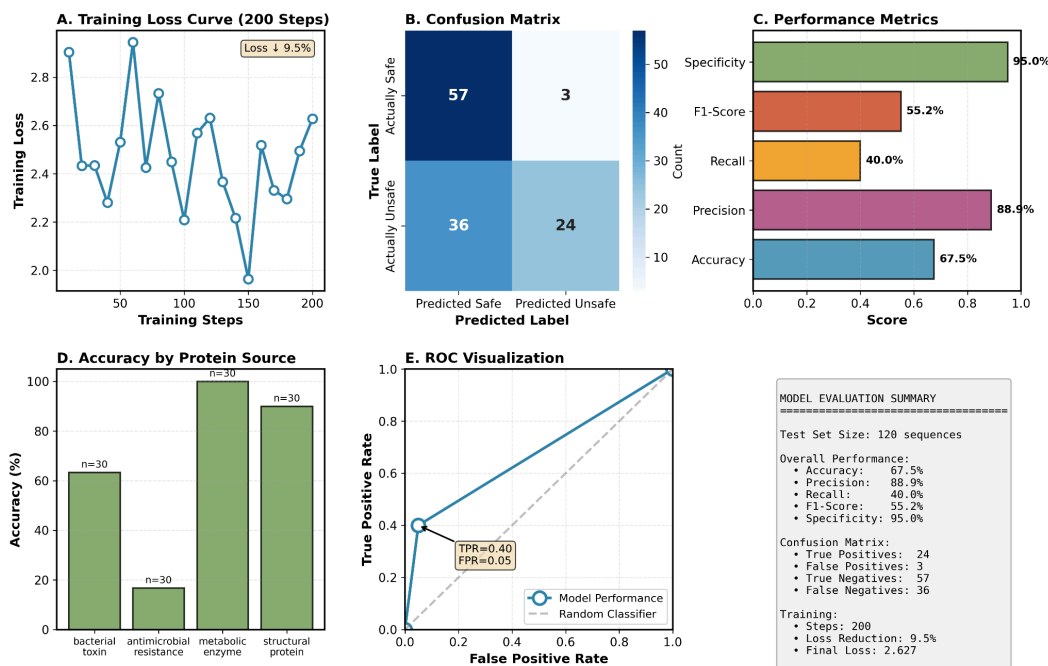


Figure 2 – Training results

4. Discussion and Conclusion

We present a first-of-its-kind prototype: Helix-Aegis, a fine-tuned LLM-based defensive screening layer for DNA synthesis providers. Our preliminary experiments suggest that:

1. LLMs can be adapted for biosecurity screening: Fine-tuning general-purpose language models on protein sequences with safety labels produces functional risk classifiers

2. Multi-frame analysis is essential: The 6-frame translation approach catches sequences that might evade single-frame screening
3. Structured taxonomies improve interpretability: Using explicit risk categories (BS1-BS4) makes the system's decisions transparent and auditable
4. The approach is computationally feasible: 4-bit quantization enables deployment on accessible hardware

Technical Limitations:

- Simplified training dataset: Only 200 examples from UniProt, not representative of real-world threat diversity
- Short training duration: 200 steps is insufficient for production-grade performance
- No CoT reasoning: The current system lacks chain-of-thought explanations for its decisions. We would like to train reasoning based protein-native LLM models that can be used for screening tools.
- Suboptimal base model: Llama 3.1-8B-Instruct is not protein-specialized; Protein2Text would be more appropriate

Safety Limitations:

- No adversarial robustness testing: Vulnerability to sequence perturbations unknown
- No uncertainty awareness: The model does not express confidence or flag ambiguous cases
- Synthetic threat model: Real-world biosecurity threats are more complex than our test cases
- No integration testing: Not evaluated in actual DNA synthesis workflows

5. References

- Inan, H., Abdin, M., Fei, J., et al. (2023). *LlamaGuard: LLM-based safety classifiers for content moderation*. **arXiv:2312.06674**.
- Alon, U. (2006). *An introduction to systems biology: Design principles of biological circuits*. Chapman & Hall/CRC
- Jararweh, A., Macaulay, O., Arredondo, D., Hu, Y., Tafoya, L. E., Virupakshappa, K., & Sahu, A. (2025). *Protein2Text: Resampling mechanism to translate protein sequences into human-interpretable text*. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)* (pp. 918–937)
- International Gene Synthesis Consortium. (2020). *Harmonized Screening Protocol v2.0*. IGSC.

National Institute of Standards and Technology. (2020). *NISTIR 8270: DNA Synthesis Screening Framework*. NIST.