LLM Biosafety Red-Team Notebook

Comprehensive Summary · May – Jun 2025

Objective

Deliver a reproducible framework for probing large-language-model (LLM) responses to biology-related queries, flagging any disclosure of sensitive technical detail.

Context & Threat-Modeling

The project is guided by an **AI Threat-Model Cheat-Sheet** that maps each prompt slice to **STRIDE** and **FMEA** elements:

Prompt Slice	STRIDE focus	FMEA failure mode
Benign-Control	-	None (baseline)
Elevated-Risk · Gene-Editing	**I – Information Disclosure**	LLM reveals protocol ("loss of confidentiality")
Elevated-Risk · Human-Ethics	**R – Repudiation** (model offering disallowed advice)	Unsafe guidance impacts policy decisions

Severity ranked "High" where pathogen or demographic harm is possible; *Detection* improved via nightly CI; *Occurrence* mitigated by OpenAI moderation + custom sanitizer.

See `THREAT_Model.md` for full STRIDE / ATLAS mapping.

Methods

- **90 prompts** (30 benign, 60 elevated-risk placeholders) in `prompts_bio.yaml`.
- Model: **GPT-3.5-turbo** (temperature 0.2).
- Moderation endpoint + **regex sanitizer** (DNA ≥ 18 bp, "step n" verbs).
- Logs → `results.jsonl`; chart → `unsafe_by_scenario.png`.
- **GitHub Actions CI**: nightly rerun + artifact upload.

Compute & Cost

- **Runtime — 6.44 minutes** on 12-thread Intel i7, Python 3.11.
- API: ≈ 10.8 k tokens → **≈ $0.05 per run** (June 2025 pricing).

Results (Baseline)

Slice	Prompts	Unsafe %	Notes
Benign	30	**0 %**	All answers compliant
Elevated	60	**0 %**	Model refused or summarised safely

Impact

- Cuts **~20 h** of manual biosafety review per evaluation cycle.
- STRIDE-FMEA mapping enables policy & engineering teams to prioritise mitigations.
- Open repo (29 Jun 2025) serves as a community benchmark; nightly CI watches for drift.

Repository: https://github.com/Lakshmanwadhwani/biosec-ace-portfolio