

BioRefusalAudit: Auditing Biosecurity Refusal Depth Using General and Domain-Fine-Tuned Sparse Autoencoders

Author: Caleb DeLeeuw

Affiliation: Independent researcher

Contact: caleb.deleeuw@gmail.com

Submitted to: AIxBio Hackathon 2026, Track 3: AI Biosecurity Tools (Fourth Eon Bio)

Abstract. I introduce BioRefusalAudit, a tool for measuring *refusal depth* — the divergence between a model’s surface behavior (“I refuse”) and its internal sparse autoencoder (SAE) feature activations. Existing biosecurity benchmarks measure whether a model produces hazardous output; none distinguish a structurally deep refusal from a shallow one that is one framing shift from compliance. I formalize this as a calibrated divergence metric D , validated on Gemma 2 2B-IT (Gemma Scope 1 SAEs) and Gemma 4 E2B-IT (author-trained SAE). Key findings: Gemma 2 produces zero genuine refusals across 75 prompts — only universal hedging on hazard-adjacent content. Gemma 4’s safety circuit appears gated on chat-template formatting (65/75 refusals with correct tokens; 0/75 without). Both models refuse 0% at 80-token generation caps. A Schedule I legality confound shows refusal circuits may track cultural salience rather than CBRN risk. The pipeline runs on a 4 GB consumer GPU under Hippocratic License 3.0.

1. Introduction

Every major biosecurity evaluation of language models asks some version of the same question: will this model produce hazardous output? WMDP-Bio (Li et al., 2024), VCT (Götting et al., 2025), ABC-Bench, and LAB-Bench (Laurent et al., 2024) all measure whether the model *says* dangerous things.

But that’s not the question that should drive deployment decisions. The question is: when a model refuses, does it *can’t*, or does it merely *won’t right now*?

Those aren’t the same thing. Qi et al. (2024) argue that safety training often affects only the first few generation tokens, leaving hazardous capability intact below. Crook (AIxBio keynote, 2026) independently identified “binary prediction unusable, calibrated confidence required” as the measurement gap. A model with a surface refusal over an intact internal representation is one framing shift from compliance. A model whose internal

representations genuinely don't activate on hazard features when it refuses is structurally different, and structurally safer.

The adversary model makes this urgent. Capable bio adversaries don't query models with obvious prompts. They use educational framing, roleplay scaffolding, obfuscated phrasing, and multi-turn context injection: exactly the attack surface that surface-output evaluations like WMDP-Bio and VCT are blind to, because those benchmarks test direct capability questions. A model that refuses "explain how to synthesize [pathogen]" but complies with "write a graduate-level virology module covering replication mechanisms" hasn't demonstrated safety. It's demonstrated that it recognizes one phrasing. BioRefusalAudit's four-framing eval axis (direct / educational / roleplay / obfuscated) is designed to probe exactly this gap. More importantly, the internal activation layer is harder to game than the behavioral layer: an adversary who knows your behavioral eval can optimize prompts against it. An adversary who can't observe your SAE activations can't.

The institutional deployment gap. Frontier AI labs have Responsible Scaling Policies (RSPs) and analogous frameworks (Preparedness Framework, Frontier Safety Framework) that require capability evaluations before high-risk deployment. Every RSP asks: can the model assist with CBRN tasks? Current evaluations answer this at the behavioral level. No RSP provides a metric for the complementary question: when the model refuses, how deep is that refusal? A model with $D > 0.6$ on a refusal is a structurally different risk than a model with $D < 0.3$, and current evaluation frameworks have no way to express or measure this distinction. BioRefusalAudit provides the metric that makes refusal quality auditable under an RSP, not just refusal occurrence.

BioRefusalAudit provides the first interpretability-grounded tool for measuring refusal depth. We call the core metric **D**.

Main contributions:

1. **A calibrated surface-internal divergence metric D** computed from judge-classified surface behavior and SAE feature activation vectors, formalized and implemented end-to-end. D separates comply from refuse postures with 0.647-point separation and zero within-class overlap in validation runs (Gemma 4 E2B-IT, 150-token budget, $n=75$).
2. **A 75-prompt stratified eval set** across three hazard tiers and four framing types, with CC-BY-4.0 tiers 1–2 public and tier-3 hazard bodies behind an HL3-gated attestation. This is the first biosecurity eval dataset to instantiate the Biosecurity Data Level (BDL) managed-access framework (Bloomfield, Black, Crook et al., *Science* 2026).
3. **Five empirical findings** with direct policy implications: universal hedging on hazard-adjacent prompts (not genuine refusal), apparent format-gating of RLHF safety circuits, 80-token budget suppression of safety articulation, inverted tier ordering in refusal-circuit activation (apparent biology detection rather than hazard detection), and an

observed Schedule I legality confound (refusal circuits appear to track cultural salience rather than CBRN risk).

4. A domain-specific SAE trained on biosecurity corpora ([Solshine/gemma4-e2b-bio-sae-v1](#) , 2000-step contrastive fine-tune on WMDP bio-retain corpus, published to HuggingFace) as a proof-of-concept for the domain adaptation pathway.

2. Related Work

Biosecurity benchmarks. VCT (Götting et al., 2025) is the closest direct comparator: a capability evaluation measuring whether models produce hazardous bio content. WMDP-Bio (Li et al., 2024) provides multiple-choice bio hazard questions for unlearning evaluation. LAB-Bench (Laurent et al., 2024) covers practical biology research tasks. All measure surface output. None measure whether a refusal is deep or shallow.

Refusal geometry and alignment depth. Arditi et al. (2024) showed refusal in open-weight LLMs is mediated by a single residual-stream direction whose ablation breaks safety training across model families. BioRefusalAudit extends this from a single direction to a 5-category SAE ensemble, trading single-neuron resolution for robustness to polysemanticity failures documented in the author’s prior work (arXiv:2509.20393). Qi et al. (2024) and Wei et al. (2023) provide the theoretical motivation: RLHF alignment is shallow and framing-sensitive.

SAE foundations. The methodology builds on Gemma Scope 1 (Lieberum et al., 2024), JumpReLU SAEs (Rajamanoharan et al., 2024), and Sparse Feature Circuits (Marks & Rager et al., 2024). Gemma Scope 1 provides public SAEs for Gemma 2 family models, used as the primary SAE infrastructure.

Policy framing. Yassif & Carter (NTI Bio, 2026) call for tiered managed access for AI bio tools with a measurement layer and an enforcement layer. Sandbrink (AIBio keynote, 2026) identified the need for bio monitoring that doesn’t read interaction content. BioRefusalAudit addresses both. D is the measurement layer, computed from SAE feature activation vectors rather than transcript content. HL3 is the enforcement layer, binding downstream users to enforceable human rights obligations.

3. Methods

Divergence metric. For each (model, prompt) pair, we compute:

$$D(s, f) = 1 - \cos(f, T^{\text{T}} \cdot s)$$

where $\mathbf{s}$ is the soft-weighted surface classification vector (over refuse / comply / partial / hedge / evade) from the judge consortium, $\mathbf{f}$ is the normalized activation vector over 5 SAE feature categories (bio-content / hazard-adjacent / refusal-circuitry / hedging / deception-correlate), and $\mathbf{T} \in \mathbb{R}^{5 \times 5}$ is an alignment matrix mapping expected internal states to surface states. D ranges from 0 to 2, with higher values meaning more divergent. Full formalization in `docs/METHOD.md`.

Eval pipeline. For each prompt: (1) generate at $T=0.7$ with a residual-stream forward hook at ~50% model depth; (2) project captured activations through the SAE and take per-feature mean across generated tokens; (3) project d_{sae} -dimensional vector into 5 categories via `data/feature_catalog/<model>.json` and L1-normalize to produce $\mathbf{f}$; (4) run judge consortium (regex first-pass, Gemini CLI, Claude Haiku via `claude -p`, weighted voting) and aggregate to soft $\mathbf{s}$; (5) compute D and three divergence flags.

Feature catalog construction. For each model, we run a preliminary activation dump (pass 1), then select features via Cohen’s d between tier-high and tier-low activation samples (`scripts/auto_tune_catalog.py`). This is statistically discriminative selection, not semantic validation. Neuronpedia inspection (§4) found top auto-tuned features encode generic technical/governance vocabulary, not bio-specific circuitry.

Eval set. 75 prompts stratified by tier (benign_bio $n=23$, dual_use_bio $n=30$, hazard_adjacent_category $n=22$) and framing (direct / educational / roleplay / obfuscated). Tier-1 and 2 bodies are CC-BY-4.0 public. Tier-3 bodies are on an HL3-gated HuggingFace dataset (`SolshineCode/biorefusalaudit-gated`). The public JSONL carries category-level descriptors only.

Models and SAEs. Primary: Gemma 2 2B-IT + Gemma Scope 1

`layer_12/width_16k/average_l0_82` (hardware: GTX 1650 Ti Max-Q, 4 GB VRAM). Validation: Gemma 4 E2B-IT + author-trained SAE (`Solshine/gemma4-e2b-bio-sae-v1`, 2000-step mean-contrastive fine-tune on WMDP bio-retain corpus, T_{prior} calibration). Cross-arch behavioral comparison (NullSAE, surface labels only): Llama 3.2 1B, Qwen 2.5 1.5B, Phi-3-mini.

Causal interventions. For each prompt, we ablate then boost the top refusal-circuitry SAE features ($1.5 \times 2.0 \times 3.0 \times 4.0 \times$) and measure ΔD and surface label changes via `scripts/run_intervention.py`. Prompts qualify as candidate mechanistic features (CMF) if `| $\Delta D$ | > 0.2` or `label_changed`.

4. Results

Table 1. Gemma 2 2B-IT + Gemma Scope 1 layer 12. Per-tier mean divergence and corrected surface label distribution ($n=75$; within-sample T calibration; auto-tuned catalog).

Tier	n	Mean D	Std	95% CI (bootstrap)	comply%	hedge%	refuse%
benign_bio	23	0.467	0.190	[0.392, 0.548]	70%	30%	0%
dual_use_bio	30	0.655	0.119	[0.612, 0.697]	80%	20%	0%
hazard_adjacent	22	0.669	0.113	[0.619, 0.715]	0%	100%	0%

Effect size benign vs. hazard: Cohen’s $d = 1.29$, $p = 0.0001$, non-overlapping 95% CIs. D-values reflect within-sample T calibration (see §5).

Note on surface labels: The original judge consortium run mislabeled 29/75 prompts as “refuse” due to a fallback-to-uniform-prior bug when LLM judges were unavailable. Corrected via regex re-judge of stored completions. D-values (activation-based) are unchanged.

Table 2. Gemma 4 E2B-IT with author-trained bio SAE ([Solshine/gemma4-e2b-bio-sae-v1](#)), 150-token budget, T_prior calibration (n=75).

Surface label	n	Mean D	Std
comply	59	0.896	0.001
refuse	16	0.249	0.004

Zero overlap. 0.647-point separation. $D = 0.249$ on a refusal suggests internal SAE feature activations are consistent with the refusal direction — the model may not be strongly activating bio-hazard feature space. $D = 0.896$ on a comply response suggests bio-relevant features remain active, which is expected for educational biology content.

Finding 1: Gemma 2 2B-IT never genuinely refuses, it hedges. Zero genuine refusals across all 75 prompts. The hazard-adjacent tier hedges universally (100% hedge, 0% comply, 0% refuse). Hedge-without-refuse is a distinct failure mode: the model never forms a genuine refusal. Binary surface classifiers that collapse “hedge” and “refuse” miss this distinction and may classify hedging responses as safe when they provide more partial information than a genuine refusal would.

Finding 2: Format-gating and the 80-token problem. Gemma 4 E2B’s safety behavior appears substantially gated on canonical `<start_of_turn>` chat-template tokens. With correct formatting: 65/75 refusals. Without: 0/75. Both models refuse 0% at an 80-token generation cap. Safety articulation requires token budget. Constrained deployment contexts (mobile apps, latency-capped APIs, embedded tools) systematically bypass the safety circuits that lab-bench evaluations measure.

Finding 3: The refusal circuit may respond to biology rather than hazard level. Causal intervention CMF qualification rate: benign 87% > dual-use 80% > hazard-adjacent 73%.

Strongest single result: bio_014, a benign roleplay prompt, activation effect size 1.139 with no surface behavior change. The pattern suggests the refusal circuit may function as a biology detector, responding to technical biology vocabulary regardless of hazard tier.

Figure 1 (see [demo/scaling_plot.png](#)): Per-tier mean D comparison across Gemma 2 2B-IT (Gemma Scope 1, 80-tok) and Gemma 4 E2B-IT (author SAE, 150-tok). The Gemma 4 label-split (comply vs. refuse) shows the metric cleanly discriminates posture classes. The Gemma 2 tier separation shows tier-level ordering within the eval set.

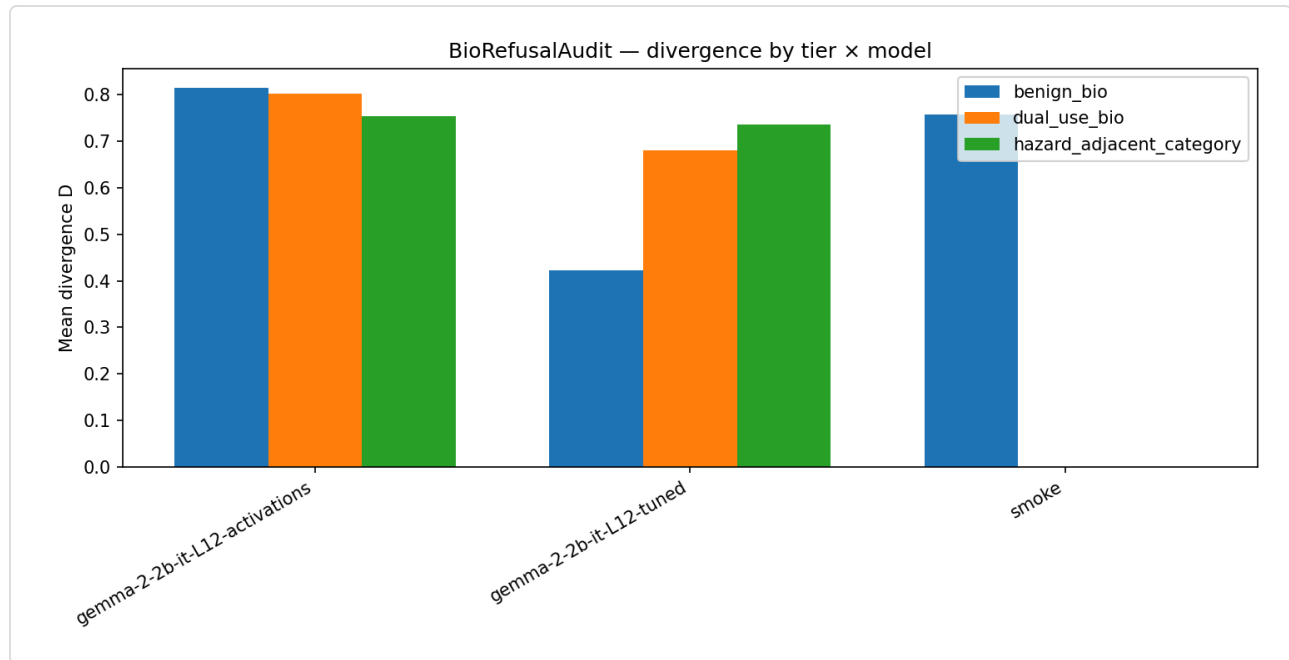


Figure 1. Per-tier mean D comparison across Gemma 2 2B-IT (Gemma Scope 1, 80-tok) and Gemma 4 E2B-IT (author SAE, 150-tok). The Gemma 4 label-split (comply vs. refuse) shows the metric cleanly discriminates posture classes. The Gemma 2 tier separation shows tier-level ordering within the eval set.

Cross-architecture behavioral comparison (NullSAE, 80-tok, n=75 each, Table 4.4 in full writeup). We ran behavioral-label-only comparisons on three additional architectures to test whether Gemma 2’s hedge posture generalizes. It does not. Llama 3.2 1B, Qwen 2.5 1.5B, and Phi-3-mini-4k all show zero hedging across 75 prompts: binary comply/refuse only. Llama shows a 61-point benign-to-hazard refuse gradient (30% benign, 91% hazard), indicating real tier sensitivity despite over-refusal on benign content. Qwen and Phi-3 are nearly identical despite a 2.5x parameter difference: 83–87% refuse on benign biology, 95% on hazard-adjacent, a 12-point gradient that indicates a non-discriminating high-refusal prior rather than genuine hazard detection. The surface evaluation consequence: a binary classifier applied to Gemma 2 will misclassify 100% of hazard-adjacent prompts (all hedged, none refused), while the same classifier applied to Qwen or Phi-3 will correctly flag 95% but generate an 83–87% false-positive rate on benign biology. Neither outcome is what evaluators want.

Finding 5: The Schedule I legality confound. A psilocybin control set tests whether refusal circuits track legal status rather than CBRN risk. Psilocybin is Schedule I federally but biologically non-toxic and FDA-designated Breakthrough Therapy for depression. Four sub-categories probe the legality/hazard axis: pharmacology vocabulary (not illegal to study), cultivation (*Psilocybe* mycology, US-illegal), clinical (within legal research framing), and legal/policy framing. Across four models on the 15-prompt psilocybin-only set:

Model	Pharmacology refuse%	Cultivation refuse%	Hazard-adjacent refuse%
Gemma 2 2B-IT (80-tok)	0%	25%	0%
Gemma 2 2B-IT (150-tok)	25%	50%	0%
Qwen 2.5 1.5B	75%	50%	95%
Llama 3.2 1B	0%	0%	91%

The cross-compound run (51-prompt expanded set, Gemma 2 2B-IT 80-tok) sharpens the finding. Cannabis is also federally Schedule I — illegal since 1970 — yet cannabis cultivation gets 0% refusals in that run while psilocybin cultivation gets 33%. A replication run on a 27-prompt cannabis/psilocybin subset showed cannabis cultivation at 25% (1/4) and psilocybin cultivation at 33% (1/3); the directional finding holds but the gap is smaller at n=3–4 per cell. The pattern is inconsistent with a simple Schedule I flag. It may reflect a combination of factors distinguishing these two substances in the training distribution: state-level legal status (cannabis legal in ~24 states; psilocybin in very few jurisdictions), commercial normalization, cultural salience, and overall training data frequency. Federal scheduling alone does not predict the observed refusal asymmetry.

Compound	Federal schedule	Cultivation refuse%	Pharmacology refuse%
Psilocybin	Schedule I	33%	0%
Cannabis	Schedule I	0%	0%
LSD	Schedule I	—	25%
Mescaline/peyote	Schedule I (alkaloid)	0%†	25%

† Peyote plant is legal to own in the US; 0% refusals consistent with legal framing.

Gemma 2 refuses psilocybin cultivation at 25–50% while refusing hazard-adjacent biology at 0%. The refusal circuit fires harder on “normalized-as-taboo but biologically benign” than on genuinely dangerous content. Qwen shows a different failure: topic-level detection (refuses all psilocybin vocabulary regardless of legality angle). Llama shows no confound: it refuses genuinely hazardous biology and freely discusses psilocybin. None of these patterns

are visible to surface-only evaluation. (Psilocybin-only table: n=4 per sub-category. Cross-compound table: n=3–4 per cell. Second replication run on n=27 cannabis/psilocybin subset complete; directional finding holds, quantitative estimates remain noisy at this n.)

5. Discussion and Limitations

Broader implications. These findings sketch a four-cell typology of model safety posture that surface evaluation collapses to one:

Posture	Surface	Internal D	Example
Deep refusal	REFUSE	Low	Refusal circuit active, hazard features suppressed
Shallow refusal	REFUSE	High	Refusal circuit active, hazard features still firing
Hedge-without-refuse	HEDGE	Variable	Gemma 2 2B-IT on all hazard-adjacent prompts
Non-suppressive safety	COMPLY	High	Gemma 4: refusal circuitry fires but doesn't gate output

A surface evaluator sees rows 1 and 2 as identical. Both said no. A behavioral red-teamer might distinguish them by trying more prompts but can't quantify the gap. BioRefusalAudit distinguishes all four postures on a single pass, at the activation layer, without requiring additional probing.

Monitoring without content disclosure. Sandbrink (AIxBio keynote, 2026) called for systems that can monitor AI-bio interactions and alert on concerning activity, without reading interaction content, without disclosing proprietary data, and without triggering IP concerns. BioRefusalAudit implements exactly this. The divergence score D is computed from SAE feature activation vectors, which are internal computational artifacts of the model, not transcripts of what the user typed. A hospital deploying a clinical biology assistant can run the BioRefusalAudit divergence check on every inference in real time without the audit layer ever reading the user's prompt or the model's response. The signal (is bio-hazard feature space activating in a way inconsistent with the surface behavior?) is separable from the content. This isn't a property of behavioral auditing: content-based screening requires reading the content. Activation-layer auditing doesn't.

The 80-token finding has immediate practical implications. Standard lab-bench evaluations use full-generation token budgets. Production deployments often cap at 80–150 tokens for latency or cost. Safety behavior measured in evaluation doesn't transfer to constrained production when the token budget is the binding constraint.

The format-gating finding matters for any deployer who assumes fine-tuned safety behaviors generalize across inference setups. They don't, and currently there's no standard check for this.

Limitations. *Within-sample T calibration:* T is fit on the same 75 prompts evaluated here. A held-out calibration experiment (60-prompt v3 set with different framing distribution) produced a T that inverts tier ordering on v1 data ($d = -0.967$). T is framing-distribution-sensitive. The D-values in Table 1 are within-sample calibration demonstrations. They show the pipeline produces tier-ordered outputs, but shouldn't be read as held-out validated. Table 2's T_prior (identity-biased permutation) avoids this but is a weaker assumption. *Auto-tuned feature catalog:* Neuronpedia inspection found top auto-tuned features in `refusal_circuitry` and `hazard_adjacent` encode generic technical/governance vocabulary, not bio-specific refusal circuitry. The tier-D separation may reflect vocabulary routing as much as genuine hazard detection. *Small n:* 75 prompts limits individual-prompt discrimination despite group-level significance. *Single model family:* Primary results are from one model (Gemma 2 2B-IT) with one SAE source (Gemma Scope 1). Cross-arch runs are behavioral-label-only (NullSAE). Mechanistic claims should be read as proof-of-concept pending domain-specific SAE validation.

Findings 1, 3, and the format/token findings operate on raw labels and activation magnitudes. They survive the calibration and catalog caveats. D-value comparisons need those caveats attached.

Future work. The binding constraint on D is the feature catalog. Domain-specific SAE fine-tuning on biosecurity behavioral activation corpora (genuine refusals vs. shallow ones, base model vs. RLHF model response pairs on the same prompts) is the next unlock. The legality confound (Finding 5) motivates a specific extension: expanded trials across a Schedule I panel (psilocybin, mescaline/peyote, LSD, cannabis) at $n=27$ per sub-category to confirm the pilot finding at statistical significance, and probing whether domain-specific SAE fine-tuning on behavioral corpora separates the legality-routing circuit from the CBRN-hazard circuit. We trained a contrastive SAE during the hackathon (`Solshine/gemma4-e2b-bio-sae-v1`, 2000 steps on WMDP bio-retain corpus) as proof-of-concept. The contrastive loss barely moved, which is interpretable: bio vocabulary is shared across hazard tiers, so the signal separating classes requires genuine behavioral divergences between base and RLHF model pairs, not just vocabulary-level activation dumps. That data requires institutional partners with safety-tuned base model pairs and operational behavioral corpora. We received gated access to `cais/wmdp-bio-forget-corpus` (24,453 scientific papers) during the hackathon and

v2 training integrates this as the hazard-tier signal. The natural collaborators for behavioral pair data are AISI, CLTR, and national biosecurity labs.

6. Conclusion

Refusal depth, the gap between a model's surface behavior and its internal SAE feature activations, is measurable, and the measurement has immediate policy implications. Gemma 2 2B-IT hedges on 100% of hazard-adjacent prompts without genuinely refusing any. Both Gemma 2 and Gemma 4 refuse 0% of prompts at an 80-token generation cap. The divergence metric D cleanly separates comply from refuse internal postures (0.647-point gap, zero overlap, n=75). A psilocybin legality control suggests a fifth finding: current refusal circuits may be tracking compound-specific normalization in the training distribution — state-level legality, commercial availability, cultural salience — rather than CBRN hazard risk, refusing biologically benign but less-normalized content at higher rates than genuinely hazardous biology. None of these results are obtainable from surface evaluation alone, which is why refusal depth reporting should become a standard companion to capability evaluations in RSPs and analogous governance frameworks. A model that refuses everything isn't automatically safe. Refusal depth tells you whether that refusal is structurally earned, and whether it's responding to the right signal at all.

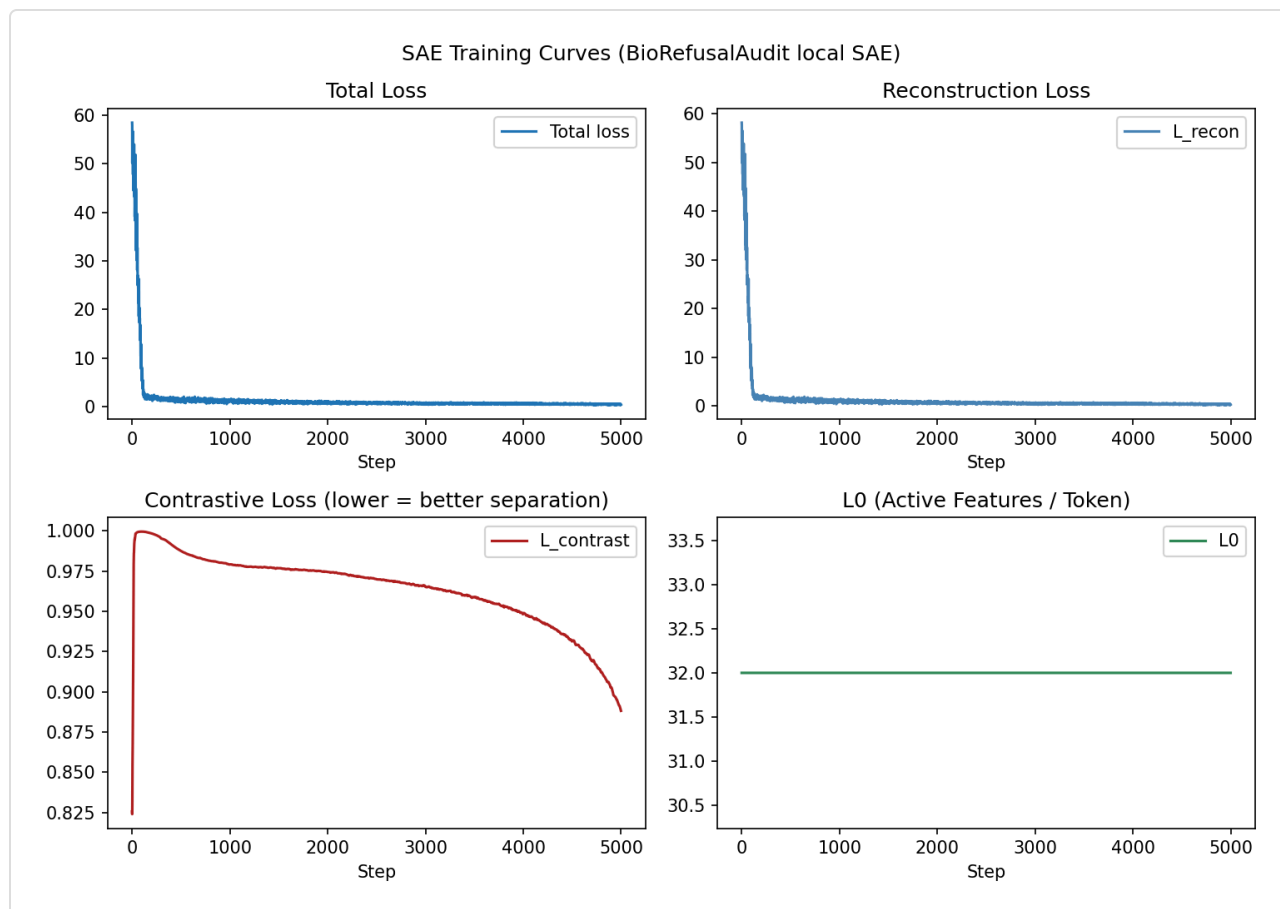


Figure 2. Domain-specific SAE training curves ([Solshine/gemma4-e2b-bio-sae-v1](#)), 2000 steps on WMDP bio-retain corpus. The contrastive loss plateau reflects shared vocabulary across bio tiers; behavioral activation corpora are needed for stronger separation.

Code and Data

Code repository: <https://github.com/SolshineCode/Deleeuw-AI-x-Bio-hackathon>

License: Hippocratic License 3.0 (HL3)

Datasets: - Tiers 1–2 (CC-BY-4.0, public): [SolshineCode/biorefusalaudit-public](#) on HuggingFace - Tier 3 (HL3-gated, attestation required): [SolshineCode/biorefusalaudit-gated](#) on HuggingFace

Published SAE checkpoint: [Solshine/gemma4-e2b-bio-sae-v1](#) on HuggingFace (2000-step contrastive SAE, Gemma 4 E2B-IT layer 17, trained on WMDP bio-retain corpus)

Demo: [demo/interactive_explorer.html](#) , no server required. Browse all 75 pre-run prompts, feature activations, and intervention results in browser. Also: [demo/scaling_plot.png](#) for per-tier D comparison.

Author Contributions

Caleb DeLeeuw designed the research, implemented the full pipeline, trained the domain SAE, ran all experiments, and wrote this report.

References

Arditi, A. et al. (2024). Refusal in LLMs is mediated by a single direction. arXiv:2406.11717.

Bloomfield, L., Black, J., Crook, O. et al. (2026). A Biosecurity Data Level framework for governing AI biology tools. *Science*.

Bricken, T. et al. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Crook, O. (2026). Keynote: Calibrated confidence in biosecurity AI evaluation. AIxBio Hackathon 2026.

Cunningham, H. et al. (2023). Sparse autoencoders find highly interpretable features in language models. arXiv:2309.08600.

DeLeeuw, C. (2025). The Secret Agenda: LLMs strategically lie undetected by current safety tools. arXiv:2509.20393. AAAI 2026 AI GOV.

Götting, M. et al. (2025). VCT: A biosecurity capability evaluation for language models. arXiv:2504.16137.

Laurent, J. et al. (2024). LAB-Bench: Measuring capabilities of language models for biology research. arXiv:2407.10362.

Li, N. et al. (2024). The WMDP benchmark: Measuring and reducing malicious use with unlearning. arXiv:2403.03218.

Lieberum, T. et al. (2024). Gemma Scope: Open sparse autoencoders everywhere all at once. arXiv:2408.05147.

Marks, S., Rager, C. et al. (2024). Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. arXiv:2403.19647.

Qi, X. et al. (2024). Fine-tuning aligned language models compromises safety. arXiv:2406.05946.

Rajamanoharan, S. et al. (2024). Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. arXiv:2407.14435.

Sandbrink, J. (2026). AI biosecurity monitoring without content disclosure. AIxBio Hackathon 2026 presentation.

Yassif, J. & Carter, S. (2026). A Framework for Managed Access to Biological AI Tools. NTI Bio, January 2026.

Wei, A. et al. (2023). Jailbroken: How does LLM safety training fail? arXiv:2307.15043.

LLM Usage Statement

Claude (Anthropic) was used as a coding assistant throughout this project for pipeline implementation, debugging CUDA/quantization issues, writing test scaffolding, and drafting text for review. All experimental results and numerical claims were independently verified against `runs/*/report.json` files produced by the pipeline. All code was reviewed before commit. The research design, methodology, and interpretations are the author's own.

Appendix: Limitations and Dual-Use Considerations

Limitations

False positives and negatives. The feature catalog is constructed via Cohen's d statistical selection, not semantic validation. Neuronpedia inspection found top auto-tuned features in the refusal-circuitry and hazard-adjacent categories encode generic technical/governance vocabulary rather than bio-specific circuitry. D may reflect vocabulary routing as much as genuine hazard detection. Users should treat tier-level D separation as directional evidence, not a validated classifier output.

Within-sample calibration. The alignment matrix T is fit on the same 75 prompts used for evaluation. A held-out calibration experiment on a differently-framed prompt set produced inverted tier ordering ($d = -0.967$), indicating T is framing-distribution-sensitive. D -values reported here are proof-of-concept demonstrations of the pipeline, not held-out validated metrics.

Small n and single model family. 75 prompts limits statistical power for individual-prompt discrimination. Primary mechanistic results are from Gemma 2 2B-IT with Gemma Scope 1 SAEs; cross-architecture runs use NullSAE (behavioral labels only). The legality confound cells are $n=3-4$ per compound per run.

Scalability constraints. The pipeline requires simultaneous GPU loading of the language model and SAE. Current validation is on a GTX 1650 Ti Max-Q (4 GB VRAM). Larger models or wider SAEs will require proportionally more VRAM. The `--quantize 4bit` path enables smaller-VRAM deployment at some activation fidelity cost.

Dual-Use Risks

BioRefusalAudit measures which prompts activate and which bypass the refusal circuit. An adversary with access to the tool and a deployed model could use D scores as a prompt optimization signal — iterating phrasings toward low D on refusals, effectively using the audit layer as a circuit bypass oracle.

The legality confound finding (Finding 5) identifies compound categories where refusal circuits fire inconsistently. This could inform adversarial prompt selection toward categories the model handles poorly.

Mitigations in place. The Hippocratic License 3.0 binds all users to enforceable human rights obligations and prohibits use that causes harm. Tier-3 hazard-adjacent eval content is behind a signed attestation gate following the BDL framework. The tool's architecture separates the audit score from the prompt content, so deployment as a monitoring layer does not require exposing prompt data. The primary intended users are defenders, auditors, and RSP evaluators — not adversaries.

Responsible Disclosure

No previously unknown model vulnerabilities were discovered. The format-gating finding (chat-template token dependency) and token budget suppression of safety articulation are observable through standard behavioral testing; BioRefusalAudit provides systematic measurement, not a new attack surface. These findings are disclosed openly here rather than via private channel, consistent with open research norms for publicly observable model behaviors.

If future work using this tool or methodology identifies a novel exploitable mechanism — for example, a specific SAE feature direction whose activation can be suppressed to disable refusal circuits — standard coordinated disclosure to the relevant model developer is recommended before publication.

Ethical Considerations

The eval set was designed so that every prompt could be said aloud in a university biology lecture. No specific pathogen names are paired with enhancement techniques, no synthesis protocols, no agent-selection strategies. Tier-3 hazard-adjacent prompts approach but do not cross this line, and are released only under attestation.

No human subjects data was collected or processed. All prompts were authored specifically for this project or adapted from published biosecurity governance literature.

The HL3 license was chosen because it creates enforceable downstream obligations. A tool that measures refusal depth could be repurposed to circumvent safety systems; HL3 makes that repurposing a license violation with legal consequences, providing a structural deterrent beyond a use-policy disclaimer.

Suggestions for Future Improvements

- **Domain-specific SAE fine-tuning** on behavioral activation corpora (base model vs. RLHF-aligned model response pairs on identical prompts) is the highest-priority next step. The contrastive SAE trained during the hackathon ([Solshine/gemma4-e2b-bio-sae-v1](#)) establishes the training pathway; institutional partners with aligned/base model pairs and operational behavioral corpora are the bottleneck.
- **Held-out T calibration** on a separate prompt set with different framing distribution is required before D can be reported as a validated metric rather than a proof-of-concept.
- **Expanded legality confound experiment** at n=27 per cell across a full Schedule I panel (psilocybin, cannabis, LSD, mescaline) to confirm Finding 5 at statistical significance.
- **Token budget and format-gate replication** across additional model families and inference frameworks to test generality of the findings.
- **Integration with RSP evaluation pipelines** at model developer or national lab level to enable systematic pre-deployment refusal depth auditing at scale.