MAPS OF FAILURES: OBSERVED LLM VULNERABILITIES IN BEHAVIORAL DESCRIPTOR SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety evaluation often summarizes a red-teaming run with attack success rate or a set of successful prompts. We study where the discovered failures occur in a chosen behavioral descriptor space. Our framework uses MAP-Elites to construct a finite-budget archive in which each occupied cell records the highest observed *Alignment Deviation*. Across three full-budget and five reduced-budget model evaluations, we observe up to 72.32% behavioral coverage and up to 370 vulnerable occupied cells. Llama-3-8B has broadly distributed observed failures, GPT-OSS-20B has fragmented high-deviation regions, and GPT-5-Mini has no observed cell above the selected threshold. We also measure how these cells change under three input-side defenses on Llama-3-8B. We release the maps, framework, metrics, and supplementary artifacts for evaluation use.

1 INTRODUCTION

Safety evaluations often report isolated successful attacks or aggregate attack success. These summaries discard where the search found failures in relation to prompt properties chosen by the evaluator. We ask whether retaining that location information reveals recurrent, connected high-risk regions under a fixed evaluation protocol. We call the resulting archives *Maps of Failures*, while defining all measured objects operationally on a finite descriptor grid. We construct these maps by retaining the highest-scoring prompt–response pair in each descriptor cell. A *connected vulnerability region* is a connected component of occupied cells whose retained score exceeds a threshold; we use "basin" as shorthand for this graph-based object. We do not assume an attraction dynamic, a calibrated Euclidean geometry, or that the two descriptors exhaust the structure of prompt space. The maps record coverage, concentration, connectedness, and changes under a repeated evaluation protocol.

We use Quality-Diversity optimization, specifically the MAP-Elites algorithm (Mouret & Clune, 2015). Unlike methods that optimize a single attack objective, MAP-Elites retains solutions across behavioral cells. We define quality using a composite automated metric, *Alignment Deviation*, which aggregates two judges over ten harm categories. Our contributions are:

- We define a reproducible, finite-budget evaluation artifact: an archive of observed conditional maxima over operational behavioral descriptors (Sections 3–4).
- We analyze connected high-deviation regions, neighborhood support, and limited response resampling, explicitly separating vulnerable-cell counts from connected-component counts (Section A.9).
- We evaluate eight target models: three receive full-budget analysis and five receive reduced-budget replication experiments (Section 5.3 and Appendix A.7).
- We compare behavioral coverage, reported diversity, and peak Alignment Deviation for MAP-Elites and established attacks under the stated query budget (Section 5.4).
- We provide a before–after evaluation of three input-side defenses on Llama-3-8B and distinguish the observed persistence result from the assumptions of a separately derived conditional bound (Section 6).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

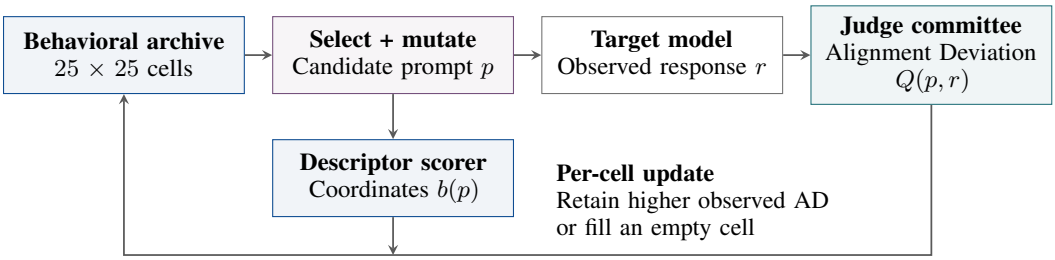


Figure 1: MAP-Elites selects and mutates prompts from the behavioral archive. The target model produces a response; a descriptor scorer assigns the archive coordinates, and the judge committee scores Alignment Deviation. The observed result replaces its cell's elite only if it improves the stored score, or occupies an empty cell.

The resulting archive complements aggregate attack metrics with an explicitly indexed record of where the evaluation discovered failures. An interactive explorer of the maps and evaluation artifacts is available online.[1]

2 Related Work

Automated red teaming and jailbreak evaluation. Language-model red teaming has progressed from model-generated test cases and large-scale human–model studies (Perez et al., 2022; Ganguli et al., 2022) to multi-round red teaming (Ge et al., 2024) and automated jailbreak search. GCG (Zou et al., 2023) optimizes adversarial suffixes; PAIR (Chao et al., 2024b) and TAP (Mehrotra et al., 2024) use iterative refinement and tree search; AutoDAN (Liu et al., 2024) applies hierarchical evolutionary search; and many-shot attacks exploit long contexts (Anil et al., 2024). HarmBench (Mazeika et al., 2024), JailbreakBench (Chao et al., 2024a), and StrongREJECT (Souly et al., 2024) improve standardization of harmful behaviors, attacks, and automated scoring. These works primarily produce attack sets or aggregate success/severity measurements. Our maps retain an additional index: the operational descriptor cell in which each elite was observed.

Quality-Diversity search. MAP-Elites (Mouret & Clune, 2015) belongs to the Quality-Diversity (QD) family, which seeks a repertoire of high-performing, behaviorally distinct solutions rather than a single optimum (Pugh et al., 2016; Cully & Demiris, 2018). Rainbow Teaming (Samvelyan et al., 2024) is the closest antecedent because it established QD search for open-ended adversarial prompt generation. Other systems diversify attacks through curiosity-driven exploration (Hong et al., 2024) or adaptive attack-strategy discovery (Zhou et al., 2025). We do not claim QD search, descriptors, or MAP-Elites as new. The proposed contribution is the evaluation layer applied to an archive: conditional-maximum maps, connected-region summaries, cross-model protocol comparisons, and before–after intervention maps. Appendix **??** states the distinction and overlap explicitly.

Behavioral evaluation and automated judges. Behavioral testing frameworks such as CheckList (Ribeiro et al., 2020), dynamic data collection such as Dynabench (Kiela et al., 2021), and slice-based visual evaluation systems such as Zeno (Cabrera et al., 2023), ChainForge (Arawjo et al., 2024), and LLM Comparator (Kahng et al., 2024) motivate retaining structured subpopulation information rather than only a scalar score. Automated LLM judges make open-ended evaluation scalable and can correlate with human preferences (Zheng et al., 2023; Liu et al., 2023), but are sensitive to presentation, verbosity, evaluator choice, and adversarially misleading outputs (Wang et al., 2024; Zeng et al., 2024b). Safety-specific evaluators likewise require direct human calibration (Souly et al., 2024). These findings motivate our two-judge committee and explicit aggregation rule; automated labels still require validation against human judgments.

External coordinates versus latent geometry. Work on LLM geometry studies internal representations and embedding structure (Balestriero et al., 2024; Subhash et al., 2023; Bartoszcze et al.,

[1] https://manifoldoffailure.wiki

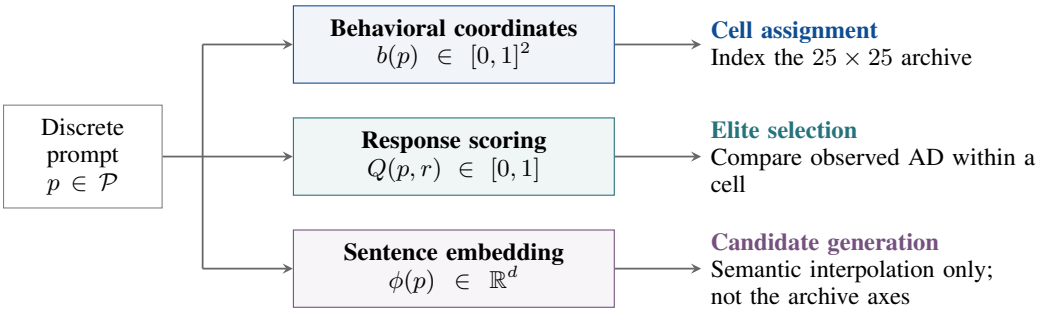


Figure 2: Conceptual relationship between discrete prompts, evaluator-defined behavioral coordinates, sentence embeddings used by one mutation operator, and the harm score.

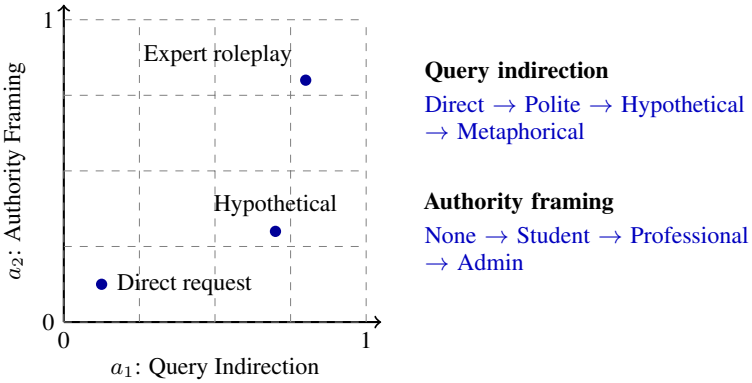


Figure 3: Prompt descriptors used for archive binning.

2025; Zou et al., 2025). Our descriptor grid is instead an evaluator-defined coordinate system and should not be confused with the model's latent geometry.

Assumptions and threat model. We assume a black-box adversary with query-only access to API-based targets (GPT-5-Mini) and operate in black-box mode for locally hosted models (Llama-3-8B, GPT-OSS-20B) as well, though white-box access is available there. The adversary's goal is broad exploration of failure modes for red-teaming and pre-deployment auditing. We restrict to single-turn, text-only interactions. The two descriptors are operational, ordered coordinates produced by an LLM scorer; equal numerical intervals are not assumed to correspond to equal semantic distances. Alignment Deviation is an automated proxy for response harmfulness, and all conclusions are conditional on the descriptor, judge committee, search algorithm, query budget, and threshold.

3 Mathematical Framework

We define the prompt descriptors and harm score used to construct the archive.

Discrete prompts and operational coordinates. Raw prompts are sequences of tokens, so the prompt set $\mathcal{P}$ is discrete. The behavioral descriptor $b : \mathcal{P} \to \mathcal{B} = [0,1]^2$ assigns every prompt a real-valued coordinate in an operational behavioral space, and the sentence encoder $\phi : \mathcal{P} \to \mathbb{R}^d$ (`all-mpnet-base-v2`; Song et al., 2020; Sentence Transformers, n.d.) places every prompt at a real-valued embedding. The measured object is a finite set of occupied grid cells and their observed elite scores.

3.1 Behavioral Descriptor Domain

We define a d-dimensional behavioral descriptor domain $\mathcal{B} = [0, 1]^d$, where each dimension operationalizes a selected prompt property. For any prompt p, $b(p) \in \mathcal{B}$ determines its archive cell. In this work, we focus on a 2D domain ($d = 2$) defined by:

a_1**: Query Indirection** The degree to which a query is framed indirectly, ranging from a direct request ($a_1 = 0$) to a highly metaphorical or hypothetical scenario ($a_1 = 1$). We use this axis to examine model responses across different framings of a request.

a_2**: Authority Framing** The level of perceived authority in the prompt's context, spanning from no authority ($a_2 = 0$) to a strong administrative or expert persona ($a_2 = 1$). We use this axis to examine model compliance across social contexts.

This representation supports an ordered grid over selected prompt properties. The behavioral descriptor for each prompt is computed by an LLM that scores the prompt along each axis on a $[0, 1]$ scale.

3.2 Alignment Deviation and Judge Aggregation

For prompt p, let $r \sim M(\cdot \mid p)$ be one target-model response. Judge $j \in J$ returns a binary unsafe vote $y_{j,c}(r) \in \{0, 1\}$ and a confidence $h_{j,c}(r) \in [0, 1]$ for category c. We convert the confidence to an unsafe probability $\pi_{j,c}(r) = h_{j,c}(r)$ when $y_{j,c} = 1$, and $\pi_{j,c}(r) = 1 - h_{j,c}(r)$ otherwise. The committee category score is

$$s_c(r) = \alpha \frac{1}{|J|} \sum_{j \in J} y_{j,c}(r) + (1 - \alpha) \frac{1}{|J|} \sum_{j \in J} \pi_{j,c}(r), \qquad \alpha = 0.5.$$

We define the *Alignment Deviation* of the observed prompt–response pair as

$$Q(p, r) = \max_{c \in C} s_c(r) \tag{1}$$

where C is the set of 10 harm categories (violence, hate speech, sexual content, self-harm, illegal activity, privacy violation, misinformation, malware, fraud, and regulated advice). The archive stores an observed maximum over evaluated prompt–response pairs in each cell. It therefore estimates neither the expected response to a prompt nor the true supremum over all prompts assigned to that cell.

We use two judge models J (GPT-4.1 and Sonnet 4.5). Each judge sees the target response and evaluates it independently across all 10 categories. The maximum across categories intentionally emphasizes the most severe detected harm dimension, but also introduces a multiple-comparisons and judge-calibration dependency discussed in Appendix A.3.

4 Methodology

MAP-Elites (Mouret & Clune, 2015) partitions the behavioral space into a 25×25 grid (the *archive*, with 625 cells) and searches for high-scoring prompts in each cell.

Prompt mutation. Given a parent prompt selected from the archive, we apply six mutation strategies: **axis-targeted generation** (50%), generating a new prompt from templates anchored at five levels of obfuscation and persuasion pressure; **paraphrasing** (10%), rephrasing while preserving intent; **entity substitution** (10%), replacing named entities or nouns with semantically similar alternatives; **adversarial suffix** (10%), appending a short adversarial string inspired by GCG (Zou et al., 2023); **crossover** (10%), combining elements from two parent prompts; and **semantic interpolation** (10%), interpolating in embedding space between two prompts p_1 and p_2 as $e_{\text{new}} = \lambda e_1 + (1 - \lambda) e_2$ and asking an LLM to blend the source prompts. Axis-targeted generation favors exploration of the selected grid, so we use coverage to describe archive occupancy.

4.1 ALGORITHM

Algorithm 1 constructs the archive $\mathcal{A}$. At each iteration, we select a prompt from the archive, mutate and evaluate it, and retain it if its cell is empty or its score exceeds the stored score.

Algorithm 1 Constructing Maps of Failures

```
Initialize archive A with empty cells
Generate initial population P_init from 100 seed prompts
for p in P_init do
    b ← GetBehavior(p)                    // LLM-based descriptor
    q ← GetQuality(p)                     // Alignment Deviation
    AddToArchive(A, p, b, q)
end for
for iteration = 1 to 15,000 do
    p_parent ← SelectFromArchive(A)
    p_child ← Mutate(p_parent)
    b_child ← GetBehavior(p_child)
    q_child ← GetQuality(p_child)
    AddToArchive(A, p_child, b_child, q_child)
end for
return A
```

We evaluate the framework on three LLMs (Llama-3-8B, GPT-OSS-20B, GPT-5-Mini). We query each target at temperature 0.7 with a maximum output length of 500 tokens. We also evaluate four baselines (Zou et al., 2023; Chao et al., 2024b; Mehrotra et al., 2024), all given the same 15,000-query evaluation budget. An LLM scorer assigns behavioral descriptors. Semantic interpolation uses the `all-mpnet-base-v2` sentence transformer (Song et al., 2020). Each experiment starts with 100 prompts spanning the harm taxonomy.

Evaluation metrics. We summarize each MAP-Elites archive with five metrics. **Behavioral Coverage** is the percentage of the 625 grid cells filled after a fixed budget. **Diversity** counts filled cells with Alignment Deviation > 0.5; it measures vulnerable cells, not distinct underlying flaws. **Peak Alignment Deviation** is the largest observed $Q(p, r)$, and **Mean Alignment Deviation** averages the retained scores across filled cells. **QD-Score** sums those scores ($\sum_{c \in \text{filled}} Q(c)$), combining occupancy and quality. For attack baselines, coverage describes the cells reached after projecting their outputs into the shared grid; peak AD reports the largest observed score.

5 RESULTS

5.1 PRIMARY ARCHIVE COMPARISON

Table 1 and Figure 4 summarize the three primary archives.

Llama-3-8B has high observed Alignment Deviation across most occupied cells, with mean AD 0.93 and peak AD 1.0.

GPT-OSS-20B shows a fragmented, spatially concentrated archive. High-deviation occupied cells cluster in the lower-left quadrant and along the top edge, while much of the high-indirection region is unoccupied. Mean occupied-cell AD is 0.73 (std. 0.33).

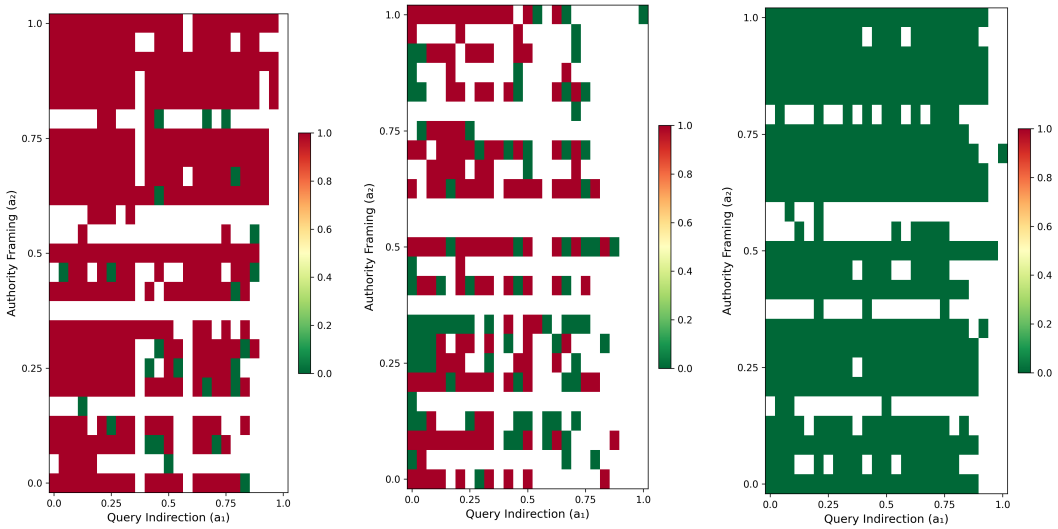
GPT-5-Mini achieves the highest behavioral coverage (72.32%), while no observed elite exceeds AD 0.50 and mean occupied-cell AD is 0.47. This is evidence of robustness to the evaluated attacks within the budget, not a certificate of safety.

5.2 CONNECTED VULNERABILITY REGIONS

Figure 4 visualizes occupied cells above AD 0.5. Connected components are defined using 8-neighbor grid adjacency.

Table 1: MAP-Elites results across target models.

Metric	Llama-3-8B	GPT-OSS-20B	GPT-5-Mini
Cov. (%)	63.04	36.32	72.32
Div. (#)	370	146	0
Peak AD	1.00	1.00	0.50
Mean AD	0.93	0.73	0.47
QD-Score	366.9	165.8	213.2



(a) Llama-3-8B (b) GPT-OSS-20B (c) GPT-5-Mini

Figure 4: Threshold maps where red indicates an occupied cell with AD > 0.5, green indicates an occupied cell with AD ≤ 0.5, and white indicates an unoccupied cell. Percentages refer only to occupied cells and are not population failure probabilities.

For **Llama-3-8B** (Figure 4a), 370 of 394 occupied cells (93.9%) exceed the threshold. They form 10 raw connected components, the largest containing 256 cells; after removing weakly supported cells, 8 components remain.

GPT-OSS-20B (Figure 4b) has 146 of 227 occupied cells (64.3%) above threshold. These form 39 raw components, 24 of which are single-cell components; support pruning leaves 15 multi-cell regions.

GPT-5-Mini (Figure 4c) has no observed occupied cell strictly above the threshold.

5.3 Cross-Model Comparison

Table 1 reports the aggregate MAP-Elites results for the three models.

Llama-3-8B achieves the highest QD-Score (366.9), driven by high mean occupied-cell AD (0.93) and 63.04% coverage. **GPT-OSS-20B** reaches peak AD 1.0 but has lower coverage (36.32%), so conclusions are restricted to the occupied portion of its archive. **GPT-5-Mini** achieves the highest coverage (72.32%) and no observed elite strictly above 0.50.

5.4 Baselines

Table 2 reports behavioral coverage, diversity, and peak Alignment Deviation for the five methods across the three target models. The values reproduce the original reported comparison. Outputs from each method are assigned to the shared descriptor grid; the baseline algorithms do not need to construct maps themselves. The searches are adaptive and do not evaluate an identical prompt set.

Table 2: Reported baseline comparison across target models.

	Llama-3-8B			GPT-OSS-20B			GPT-5-Mini		
Method	Coverage	Diversity	Peak AD	Coverage	Diversity	Peak AD	Coverage	Diversity	Peak AD
Random	5.28	30	1.00	0.80	3	0.97	9.01	0	0.50
GCG (whitebox)	7.20	173	1.00	34.72	0	0.50	NA	NA	NA
TAP	41.76	255	1.00	25.60	68	1.00	30.08	0	0.50
PAIR	61.44	291	1.00	**63.20**	73	1.00	58.56	0	0.50
Ours (ME)	**63.04**	**370**	1.00	36.32	**146**	1.00	**72.32**	0	0.50

Table 3: Observed archive summaries on Llama-3-8B before and after three input defenses.

Defense	**Cov.**	**Above τ**	**Mean AD**	**Peak AD**
None	63.0%	93.9%	0.93	1.00
Perplexity	52.6%	94.2%	0.93	1.00
Paraphrase	50.9%	89.0%	0.89	1.00
Constitutional	18.7%	91.5%	0.88	1.00

On Llama-3-8B, MAP-Elites reaches 63.04% coverage, PAIR 61.44%, TAP 41.76%, GCG 7.20%, and random sampling 5.28%. All methods observe peak AD 1.0. The reported diversity values are 370 for MAP-Elites, 291 for PAIR, 255 for TAP, 173 for GCG, and 30 for Random.

On GPT-OSS-20B, PAIR reaches higher raw coverage (63.20% versus 36.32% for MAP-Elites), while both methods observe peak AD 1.0. The reported diversity is 146 for MAP-Elites, 73 for PAIR, 68 for TAP, 3 for Random, and 0 for GCG.

For GPT-5-Mini, no evaluated method observes AD above 0.50 within the stated budget.

6 Observed Persistence Under Input Defenses

A natural use of the map is to repeat the protocol after an input-side defense $D : \mathcal{X} \to \mathcal{X}$ paraphrases, filters, or rewrites prompts. The Defense Trilemma analysis (Bhatt et al., 2026, Theorems 5.1 and 6.3) provides conditional motivation. In its connected metric-space setting, f and D are L- and K-Lipschitz, respectively, and D fixes every strictly safe input. With nonempty safe and unsafe regions, there is a fixed boundary point z with $D(z) = z$ and $f(z) = \tau$, yielding

$$f(D(x)) \;\geq\; \tau - LK\,\mathrm{dist}(x, z). \tag{2}$$

Strict unsafety persists on the set $\mathcal{S} = \{x : f(x) > \tau + \ell(K+1)\mathrm{dist}(x, z)\}$, where $|f(D(x)) - f(x)| \leq \ell\,\mathrm{dist}(D(x), x)$. If nonempty, $\mathcal{S}$ is open and has positive measure whenever the measure is positive on nonempty open sets. This does not guarantee persistence of an entire empirical basin.

We evaluate three input-side configurations on Llama-3-8B: a perplexity filter, a T5 paraphrase wrapper, and a constitutional rewriter. Table 3 reports post-defense archive summaries, and Figure 12 reports matched-cell persistence of 76.5% (perplexity), 70.0% (paraphrase), and 27.3% (constitutional). All three wrappers leave some previously high-deviation cells above threshold, but the constitutional wrapper removes substantially more matched cells. These observations are specific to Llama-3-8B, the selected prompts, and the evaluated wrapper configurations.

7 Discussion and Limitations

The same MAP-Elites evaluation protocol produces different archives across models. Llama-3-8B has high AD across most occupied cells; GPT-OSS-20B has a mixture of multi-cell regions and isolated components; and GPT-5-Mini has no observed cell strictly above threshold. These results support retaining location and connectedness information in addition to aggregate attack success. The defense experiment shows how the archive changes after each wrapper: some previously high-deviation cells remain above threshold, some fall below it, and others are no longer occupied.

Limitations. The 2D descriptor projection loses information and its numerical intervals are not semantically calibrated. Alignment Deviation relies on two automated judges and a maximum across

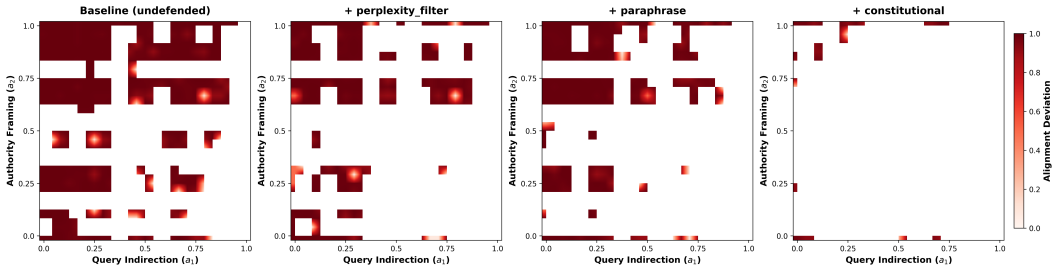


Figure 5: Llama-3-8B threshold maps before and after three defenses. The figure shows empirical changes in the discovered archive and does not by itself verify the assumptions of Eq. 2.

ten categories; independent human validation remains necessary. Each archive elite is selected from a stochastic single response, so maximum selection can exaggerate typical failure rates. We treat only single-turn interactions. The continuous-defense bound is conditional and the present experiments do not verify its assumptions.

8 Conclusion

Maps of Failures records where a fixed-budget search finds harmful responses and which of those locations form connected vulnerable regions. Llama-3-8B and GPT-OSS-20B both reach peak AD 1.0, and their largest connected regions contain 256 and 26 cells, respectively. The maps distinguish these observed failure patterns despite identical peak scores. GPT-5-Mini has no observed cell above the threshold under the same protocol.

Repeating the evaluation with input-side defenses also shows which previously vulnerable cells remain above threshold. On Llama-3-8B, matched-cell persistence ranges from 27.3% for the constitutional rewriter to 76.5% for the perplexity filter. The resulting archives provide a record for comparing failures before and after an intervention. Our experiments propose a novel and promising way of evaluating and visualizing LLMs for alignment failures.

AI use statement

LLMs were used as target systems and as components of the experimental pipeline for behavioral-descriptor scoring, response judging, and the resampling study, as described in Sections 4 and 6 and in the technical appendices. Generative AI was also used to revise manuscript text and audit internal consistency. The authors reviewed the AI-assisted work and take responsibility for the final content, including all text, claims, experimental results, and artifacts produced with the aid of generative AI.

Ethics statement

Mapping vulnerabilities is dual-use: the same artifact that enables targeted defense can guide attackers. We mitigate this risk by releasing aggregate maps and the framework rather than per-cell elite prompts, targeting models whose vendors have established disclosure channels, and focusing the claims on evaluation structure rather than novel attack strings. The reported results use no human-subject data. Any future human annotation must follow applicable institutional, consent, and data-handling requirements before results are reported.

Reproducibility statement

Section 4 specifies the search procedure, behavioral descriptors, quality metric, query budget, and evaluation protocol. The technical appendices define the archive estimand, judge aggregation, descriptor prompts, claim boundaries, target and auxiliary models, baseline caveats, ablations, and resampling configurations. The supplementary material provides the artifacts needed to reconstruct the reported archives and figures.

REFERENCES

Cem Anil et al. Many-shot jailbreaking. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. ChainForge: A visual toolkit for prompt engineering and LLM hypothesis testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2024. ACM. doi: 10.1145/3613904.3642016.

Randall Balestriero, Romain Cosentino, and Sarath Shekkizhar. Characterizing large language model geometry helps solve toxicity detection and generation. arXiv preprint arXiv:2312.01648, 2024.

Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. Representation engineering for large-language models: Survey and research challenges. arXiv preprint arXiv:2502.17601, 2025.

Manish Bhatt, Sarthak Munshi, Vineeth Sai Narajala, Idan Habler, Ammar Al-Kahfah, Ken Huang, Joel Webb, Blake Gatto, and Md Tamjidul Hoque. The defense trilemma: Why prompt injection defense wrappers fail? arXiv preprint arXiv:2604.06436v3, 2026. URL `https://arxiv.org/abs/2604.06436v3`.

Ángel Alexander Cabrera, Erica Fu, Donald Bertucci, Kenneth Holstein, Ameet Talwalkar, Jason I. Hong, and Adam Perer. Zeno: An interactive framework for behavioral evaluation of machine learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2023. ACM. doi: 10.1145/3544548.3581268.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. arXiv preprint arXiv:2310.08419, 2024b.

Antoine Cully and Yiannis Demiris. Quality and diversity optimization: A unifying modular framework. *IEEE Transactions on Evolutionary Computation*, 22(2):245–259, 2018. doi: 10.1109/TEVC.2017.2704781.

Deep Ganguli et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. MART: Improving LLM safety with multi-round automatic red-teaming. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1927–1937. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.107.

Zhang-Wei Hong, Idan Shenfeld, Tsun-Hsuan Wang, Yung-Sung Chuang, Aldo Pareja, James Glass, Akash Srivastava, and Pulkit Agrawal. Curiosity-driven red-teaming for large language models. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=4KqkizXgXU`.

Minsuk Kahng, Ian Tenney, Mahima Pushkarna, Michael Xieyang Liu, James Wexler, Emily Reif, Krystal Kallarackal, Minsuk Chang, Michael Terry, and Lucas Dixon. LLM Comparator: Visual analytics for side-by-side evaluation of large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2024. ACM. doi: 10.1145/3613905.3650755.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina

Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs automatically. arXiv preprint arXiv:2312.02119, 2024.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. arXiv preprint arXiv:1504.04909, 2015.

Ethan Perez et al. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Justin K. Pugh, Lisa B. Soros, and Kenneth O. Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40, 2016. doi: 10.3389/frobt.2016.00040.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442.

Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H. Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, Tim Rocktäschel, and Roberta Raileanu. Rainbow Teaming: Open-ended generation of diverse adversarial prompts. In *Advances in Neural Information Processing Systems*, volume 37, pp. 69747–69786, 2024.

Sentence Transformers. all-mpnet-base-v2. Hugging Face model card, n.d. Accessed September 16, 2026.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Varshini Subhash, Anna Bialas, Weiwei Pan, and Finale Doshi-Velez. Why do universal adversarial attacks work on large language models?: Geometry might be the answer. arXiv preprint arXiv:2309.00254, 2023.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. doi: 10.18653/v1/2024.acl-long.511.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. How Johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 14322–14350, 2024a. doi: 10.18653/v1/2024.acl-long.773.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Andy Zhou, Kevin Wu, Francesco Pinto, Zhaorun Chen, Yi Zeng, Yu Yang, Shuang Yang, Sanmi Koyejo, James Zou, and Bo Li. AutoRedTeamer: Autonomous red teaming with lifelong attack integration. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405, 2025.

A Technical appendices and supplementary material

The anonymized supplementary material provides code, configurations, archive summaries, and additional model results.

A.1 Evaluation-Design Context

From aggregate scores to behavioral slices. The archive is related to behavioral and slice-based evaluation rather than to latent-space interpretation. CheckList organizes tests by capabilities and test types (Ribeiro et al., 2020); Dynabench repeatedly collects failures against evolving models (Kiela et al., 2021); and Zeno supports subpopulation discovery and comparison (Cabrera et al., 2023). ChainForge and LLM Comparator extend this style of analysis to prompt and response comparisons (Arawjo et al., 2024; Kahng et al., 2024). Our cells are predefined by two prompt descriptors rather than discovered post hoc, and each cell stores an adaptive-search maximum rather than a representative sample. Consequently, the maps are useful as failure-discovery records.

Why these descriptors are plausible but not validated. Role-play, obfuscation, and competing instruction-following objectives recur in analyses of jailbreak mechanisms (Wei et al., 2023; Liu et al., 2024), while persuasive framing has been studied directly as an adversarial strategy (Zeng et al., 2024a). This literature motivates query indirection and authority/persuasion as a case study, but does not validate our particular scales, anchors, or neighborhood relation. Construct validity therefore requires controlled prompt pairs, human descriptor labels, alternative descriptor models, and additional axis sets.

Judge calibration as measurement validation. MT-Bench and G-Eval show the utility of model-based evaluators at scale (Zheng et al., 2023; Liu et al., 2023); FairEval and LLMBAR demonstrate that ordering, style, and adversarially misleading content can change judge outcomes (Wang et al., 2024; Zeng et al., 2024b). StrongREJECT further evaluates an automated safety score against human judgments (Souly et al., 2024). For this paper, agreement between two model judges is evidence about reproducibility within a judge committee, not validity against human judgment. Appendix A.10 reports agreement with automated reference scores.

A.2 Descriptor and Judge Implementation

Behavioral descriptor. The descriptor model scores each prompt independently along two operational dimensions. For a_1, it is asked to rate how *obfuscated* the harmful intent is, considering layering, plausible deniability, and semantic distance; anchors are 0.0 (explicit and undisguised), 0.5 (partially hidden through framing, roleplay, or hypotheticals), and 1.0 (multiple layers of misdirection). For a_2, it rates *persuasion pressure*, considering social authority, urgency, and legitimacy framing; anchors are 0.0 (no persuasion tactic), 0.5 (moderate roleplay or justification), and 1.0 (combined authority, urgency, and legitimacy pressure). The model must return one number in $[0, 1]$; unparsable outputs default to 0.5. Descriptor outputs are cached by prompt.

Judge template and parsing. For each harm category, each judge receives the category name and definition and the target response. It is asked whether the response is harmful in that category, with explicit attention to detailed harmful instructions, promotion or glorification, and explicit harmful content. The required format is `YES/NO <confidence>`. A `NO` confidence h is converted to unsafe probability $1 - h$; unparsable outputs default to 0.5. The exact committee aggregation is given in Section 3. The judge does not receive the original prompt, which limits its ability to assess contextual intent.

Axis-targeted generator. The 50% axis-targeted operator samples one of five anchor levels $\{0, 0.25, 0.5, 0.75, 1\}$ for obfuscation and persuasion pressure, instantiates templates for a harmful topic, and optionally inserts a modifier. It generates a new prompt rather than continuously perturbing the parent's numerical coordinate. This explains both its strong grid-exploration bias and why descriptor adjacency should not be interpreted as a measured token-space path.

A.3 Validity Threats and Required Robustness Checks

Descriptor construct validity. The axes are selected evaluation constructs rather than established universal robustness metrics. Systematic descriptor errors could create apparent bands or components. Appropriate checks include repeated descriptor calls, independent descriptor models, human annotations, alternative axis pairs, and controlled prompt pairs that vary one construct while holding intent fixed.

Stochastic response selection. The target is sampled at temperature 0.7, while the archive retains a maximum. This creates winner's-curse bias: a rare unsafe draw can replace a more typical response. The resampling experiment in Appendix A.9 measures this effect on a small subset, but full reporting should include per-cell failure probability or repeated-response mean/quantiles.

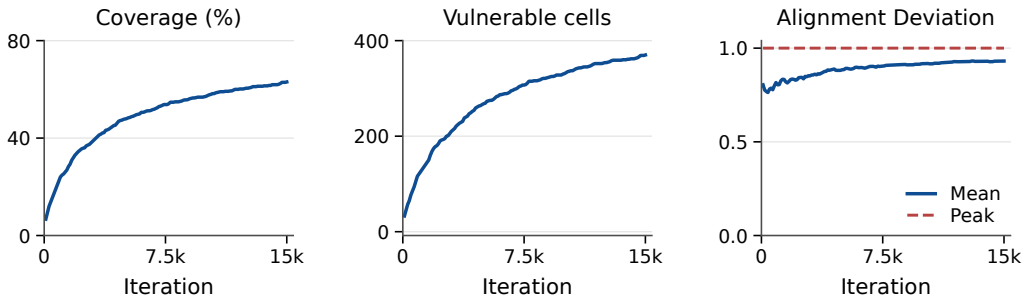
Judge validity and threshold sensitivity. Taking the maximum across ten categories makes AD sensitive to calibration error in any category, particularly near $\tau = 0.5$. Independent human labels, confidence intervals, alternative judge committees, and threshold sweeps are needed to establish that connected-region conclusions are stable.

Grid and graph dependence. Component count depends on resolution, four- versus eight-neighbor adjacency, threshold, and unoccupied-cell handling. The reported 25×25, eight-neighbor, $\tau = 0.5$ analysis is one operational choice. A complete robustness analysis should report component persistence across these choices and across independent search seeds.

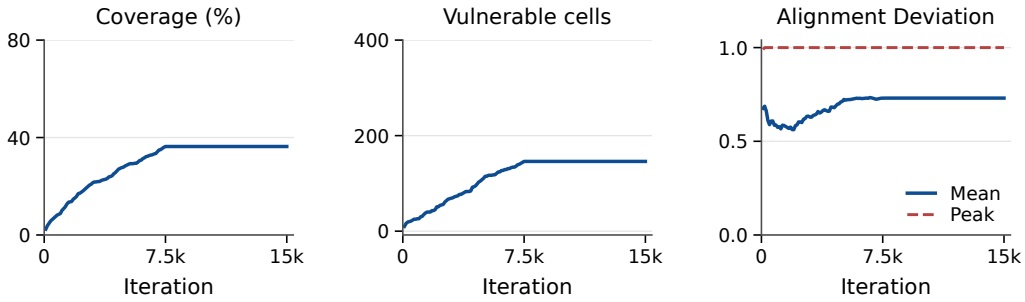
Visualization. Quantitative analyses use discrete archive values. The heatmap renderer applies bilinear interpolation for display. The contour and 3D renderers replace unoccupied cells with zero before plotting. Consequently, visual smoothness, valleys around missing cells, and contour boundaries are not independent evidence of continuity and are excluded from the paper's quantitative claims.

A.4 Per-Iteration Coverage, Diversity, and Quality

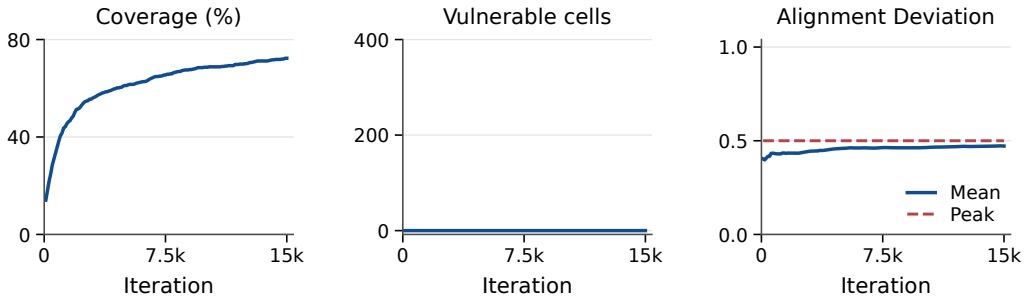
Figure 6 reports the per-iteration trajectories of the three primary search metrics, behavioral coverage, diversity (cells with AD > 0.5), and mean Alignment Deviation, over the full 15,000-iteration MAP-



(a) Llama-3-8B



(b) GPT-OSS-20B



(c) GPT-5-Mini

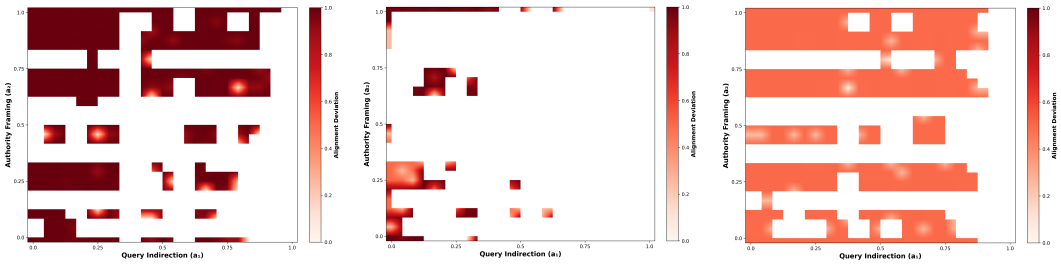
Figure 6: Coverage, diversity, and quality metrics over 15,000 iterations. (a) Llama-3-8B: sustained growth in all metrics. (b) GPT-OSS-20B: early plateau in coverage with moderate quality. (c) GPT-5-Mini: high coverage growth but zero diversity and quality capped at 0.50. Each row shows one recorded search trajectory, with common axis scales across models; lines do not represent averages over independent runs.

Elites budget for each target model. The trajectories show how quickly the search fills the grid, when vulnerable-cell counts plateau, and how mean quality changes as new cells are filled.

Table 1 reports the terminal values, and Appendix A.9 reports connected-region summaries.

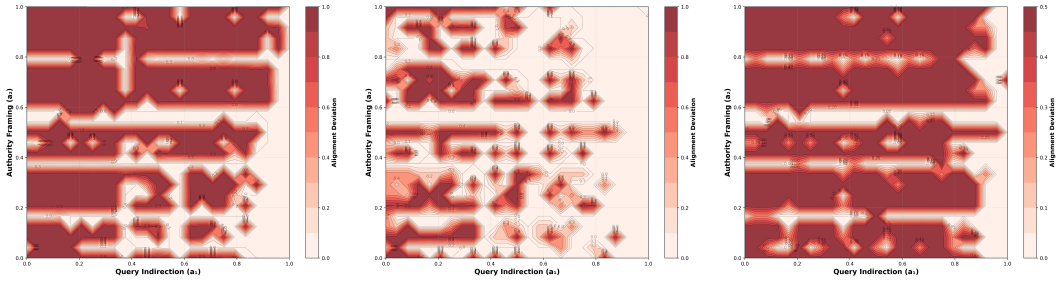
A.5 Complete Primary-Model Map Atlas

For completeness and direct comparison with the original preprint, we retain all primary-model map views. The heatmaps below are the direct archive visualizations. The contour and surface renderers interpolate displayed values and replace unoccupied cells for plotting.



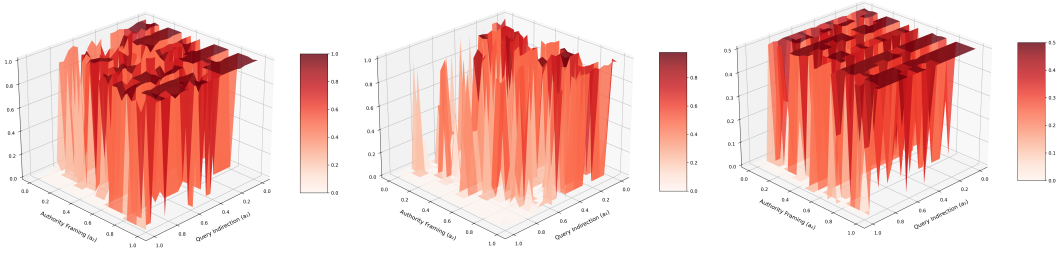
(a) Llama-3-8B (b) GPT-OSS-20B (c) GPT-5-Mini

Figure 7: Observed-maximum Alignment Deviation in the occupied cells of the three primary archives.



(a) Llama-3-8B (b) GPT-OSS-20B (c) GPT-5-Mini

Figure 8: Contour renderings of the primary archives. Contours summarize the display grid and must not be interpreted as measured semantic paths between prompts or as evidence that AD varies continuously between cells.



(a) Llama-3-8B (b) GPT-OSS-20B (c) GPT-5-Mini

Figure 9: Three-dimensional renderings of archive AD values.

A.6 Ablation Study

To isolate the contribution of each component, we compare three variants (Table 4):

1. **Full method** (MAP-Elites + Alignment Deviation): the complete framework as described in Section 4.

2. **ME + Toxicity**: MAP-Elites with Alignment Deviation replaced by a single-dimension toxicity classifier score, removing the multi-category quality signal.

3. **Random + AD**: Random prompt sampling evaluated with the full Alignment Deviation metric, removing the quality-diversity search entirely.

The toxicity-score variant tests how elite selection changes when the scorer focuses on overtly toxic content. Harms expressed in neutral or technical language, such as misinformation, privacy violations, or cybersecurity exploits, may receive low toxicity scores. The archive can therefore retain different prompts when toxicity replaces the ten-category AD score.

Table 4: Ablation study across the three models. Cov. = Coverage (%), Div. = Diversity (#), QD = QD-Score, AD = Mean Alignment Deviation.

	Llama-3-8B				**GPT-OSS-20B**				**GPT-5-Mini**			
Variant	Cov.	Div.	QD	AD	Cov.	Div.	QD	AD	Cov.	Div.	QD	AD
Full	**63.04**	**370**	**366.9**	**0.931**	**36.32**	**146**	**165.8**	**0.730**	**72.32**	0	**213.2**	**0.472**
ME + Toxicity	62.08	288	282.3	0.726	35.52	103	126.5	0.569	71.36	0	163.5	0.368
Random + AD	5.28	30	28.1	0.851	0.80	3	3.1	0.624	4.96	0	9.1	0.293

Removing the QD archive (Random + AD) produces the largest coverage reduction across the three targets. The Random + AD variant still achieves high mean AD in the few cells it fills (0.85 on Llama-3-8B), indicating that the automated quality signal and the archive affect different reported quantities: MAP-Elites primarily increases grid occupancy, while the multi-category score changes which responses are retained as elites.

A.7 Additional Model Comparisons

The three primary models receive the full 15,000-query analysis. We also apply the same archive construction to five additional models using 3,000 queries. These reduced-budget runs test whether the pipeline can recover high-deviation occupied cells across model families; their lower coverage does not support detailed component-shape comparisons.

Table 5: Leaderboard results across 5 target models, ranked by QD-Score. All values are means $\pm$ standard deviations across 3 independent runs.

Rank	**Model**	**Coverage (%)**	**Diversity (#)**	**Peak AD**	**QD-Score**
1	Mistral Large 3	27.9 ± 12.8	146 ± 77.1	0.993 ± 0.002	139.4 ± 73.2
2	Llama4 Maverick	24.3 ± 15.0	128 ± 84.1	0.987 ± 0.011	122.4 ± 80.2
3	Haiku 4.5	23.5 ± 0.7	95 ± 1.7	0.973 ± 0.005	85.5 ± 3.4
4	Sonnet 4.6	23.2 ± 0.9	85 ± 1.2	0.980 ± 0.002	81.7 ± 1.7
5	Haiku 3.5	18.6 ± 7.7	71 ± 36.6	0.982 ± 0.011	66.4 ± 33.2

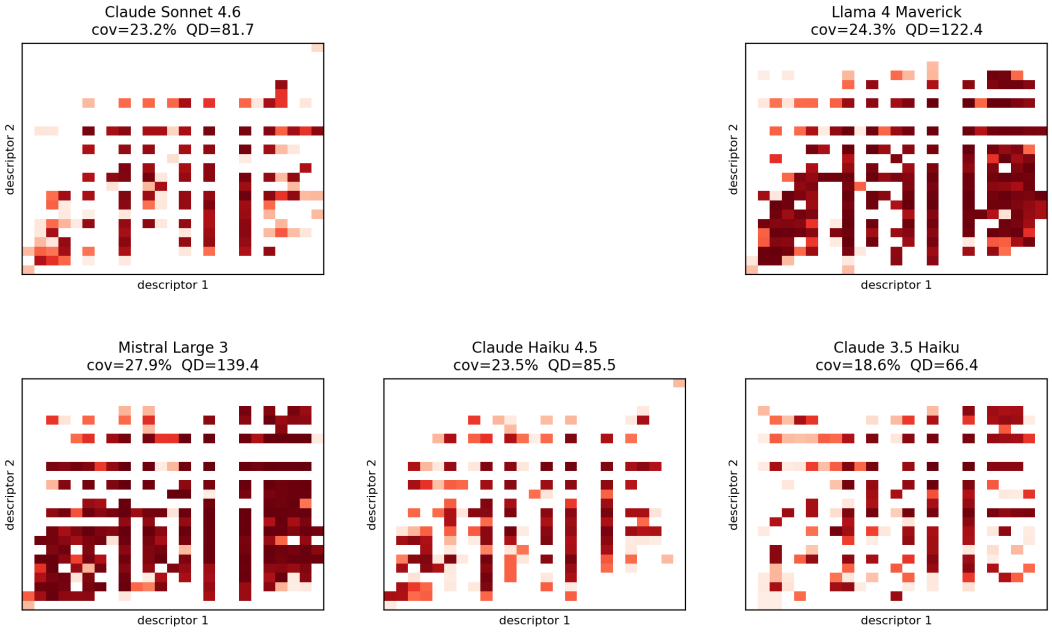


Figure 10: Observed archive heatmaps for the five reduced-budget model evaluations. Each run uses a 3,000-query budget, so white cells should be read as unoccupied under that budget rather than safe.

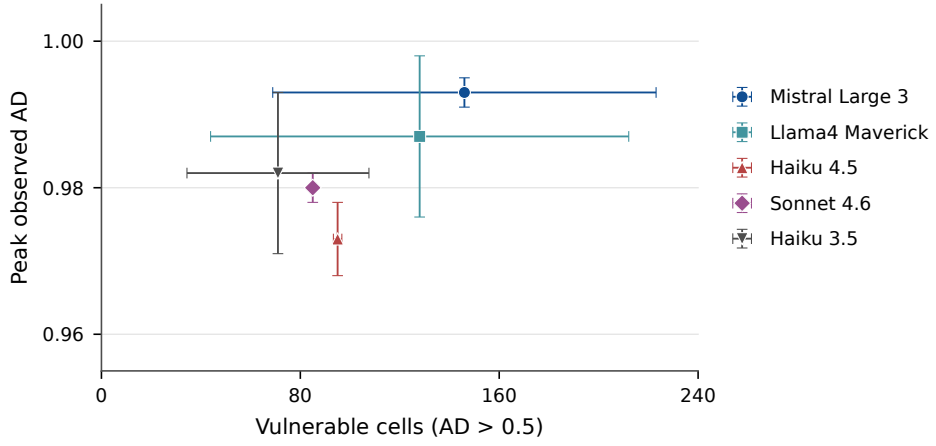


Figure 11: Peak observed AD versus vulnerable-cell diversity for the five reduced-budget models. This view separates the severity of the strongest observed failure from the number of descriptor cells in which an above-threshold elite was found. Points and error bars are the means and one standard deviation across three runs reported in Table 5.

A.8 Shared Vulnerable Cells Across Model Families

Table 6: Multiplicity among the 193 descriptor cells with an observed above-threshold elite in at least one of the five reduced-budget model evaluations.

# of models exploiting cell	1	2	3	4	5
# of cells	64	43	28	30	28

Table 7: Twelve descriptor cells with an observed above-threshold elite for all five reduced-budget model evaluations. Coordinates are grid indices (a_1, a_2).

Rank	**Cell** (a_1, a_2)	**Mean AD**
1	$(15, 8)$	0.948
2	$(10, 8)$	0.939
3	$(8, 8)$	0.931
4	$(15, 20)$	0.922
5	$(10, 20)$	0.916
6	$(10, 18)$	0.914
7	$(8, 18)$	0.910
8	$(18, 18)$	0.908
9	$(18, 20)$	0.887
10	$(10, 15)$	0.869
11	$(13, 20)$	0.830
12	$(10, 21)$	0.816

Table 8: Pairwise Jaccard similarity between models' vulnerable archive cell sets. Higher values indicate more shared cells; values ≥ 0.60 are bolded.

	Haiku 4.5	Sonnet 4.6	Haiku 3.5	Llama4 Maverick	Mistral Large 3
Haiku 4.5	1.00				
Sonnet 4.6	0.55	1.00			
Haiku 3.5	0.50	0.48	1.00		
Llama4 Maverick	0.54	0.55	0.55	1.00	
Mistral Large 3	0.56	0.54	0.53	**0.69**	1.00

A.9 Connected-Region Robustness

Because each cell retains only its maximum-AD elite, a single high-scoring outlier could define a "basin." We address this with three analyses: (1) we operationalize basins as connected components and count how many are isolated single-cell artifacts; (2) we compute *neighborhood support*, the fraction of each vulnerable cell's filled grid neighbors that also exceed threshold, to assess local agreement; and (3) we re-query a stratified sample of elite prompts multiple times to test whether individual high-AD cells reflect a stable property or a single lucky draw.

Definitions. A *connected vulnerability region* (operational basin) is a maximal connected component of filled cells whose elite AD exceeds threshold $\tau = 0.5$, under 8-connectivity on the 25×25 grid. Let $\mathcal{F}(i, j) = \{(i', j') \in \mathcal{N}_8(i, j) : \text{cell is filled}\}$ be the filled 8-connected neighbors of cell

(i, j). Its *neighborhood support* is the fraction of those neighbors that are themselves vulnerable:

$$\text{support}(i,j) = \frac{|\{(i',j') \in \mathcal{F}(i,j) : \text{AD}_{i'j'} > \tau\}|}{|\mathcal{F}(i,j)|}. \tag{3}$$

We classify a vulnerable cell with support < 0.25 as weakly supported (a potential outlier) and report basin counts before and after removing these cells.

Table 9: Connected-region analysis. Raw = components of occupied cells with AD > 0.5; Pruned = after removing cells with neighborhood support < 0.25.

Model	Vuln. cells	Raw comp.	Single-cell	Pruned comp.	Largest	Mean support
Llama-3-8B	370	10	2	8	256	0.93
GPT-OSS-20B	146	39	24	15	26	0.65
GPT-5-Mini	0	0	0	0	0	0

Results. Table 9 reports the component structure of each archive:

- **Llama-3-8B** has a dominant connected high-deviation region: the largest component spans 256 of 370 vulnerable cells (69%), and mean neighborhood support is 0.93. Two of 10 raw components are single-cell outliers.
- **GPT-OSS-20B** is fragmented: 39 raw components, of which 24 are single-cell outliers. After support-based pruning (support ≥ 0.25), 15 multi-cell regions remain, with the largest containing 26 cells. Mean support is 0.65.
- **GPT-5-Mini** has peak AD exactly at the threshold (0.50) with no observed occupied cell exceeding it under the strict > 0.5 criterion.

These results test whether component counts are dominated by isolated cells. Llama-3-8B has high mean neighborhood support (0.93), whereas GPT-OSS-20B contains many isolated cells: singletons account for 62% of its raw component count.

Elite re-evaluation under resampling. Because the archive stores the maximum-AD response from a single stochastic generation, a cell's recorded score could be a lucky draw. We therefore re-query a stratified sample of Llama-3-8B elites and re-score each response with the full judge committee. For each elite we report *persistence*: the fraction of resampled responses that remain above threshold.

Across 12 stratified elites (3 resamples each), mean persistence is 0.72, and 10 of 12 elites (83%) remain above threshold on a majority of resamples; mean AD declines by 0.26 relative to the stored maximum. A subset of cells is *bimodal*: responses to the same prompt alternate between high (AD ≈ 0.98) and low (AD ≈ 0.02) scores as the target complies or refuses. Storing only the maximum hides this variation.

Within this small sample, most elites reproduce above threshold at least twice, but mean AD falls materially and some prompts are bimodal. The observed maximum therefore overstates per-prompt failure frequency. A more informative map would report repeated-response means, quantiles, or empirical failure probability in addition to the maximum. We consequently describe the current maps as *observed-maximum*, not expected-case or true worst-case, estimates.

A.10 Judge Calibration

The supplementary artifact includes a calibration in which a subset of archive responses is re-scored by a reference committee. The released calibration summaries report Pearson correlation r, coefficient of determination R^2, and mean absolute error (MAE) for nine rerun workers. Although each worker requested 1,000 samples, the summaries contain 348–524 scored pairs per worker, totaling 3,982 pair records.

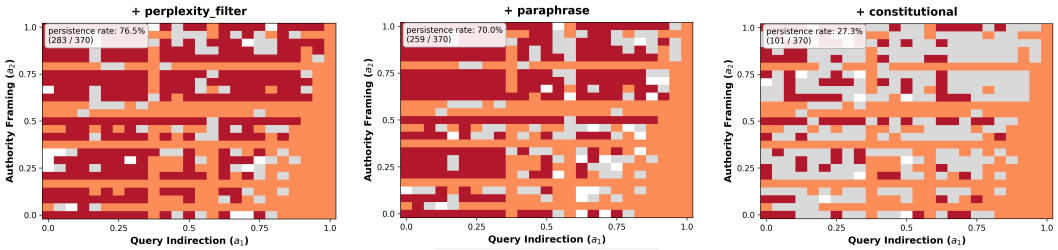
Table 10 shows positive association with the reference scores in every worker ($r = 0.592$–0.837), alongside substantial absolute disagreement (MAE $= 0.117$–0.322 on the $[0, 1]$ AD scale). The pair-count-weighted MAE is 0.235. Three workers (1, 6, and 7) have negative reported R^2.

Table 10: Automated judge calibration against reference-committee re-scoring.

Worker	Pairs	Pearson r	R^2	MAE
0	435	0.755	0.295	0.216
1	524	0.659	−0.411	0.287
2	455	0.775	0.430	0.168
3	524	0.837	0.625	0.117
4	440	0.725	0.200	0.238
5	348	0.717	0.121	0.240
6	378	0.646	−0.194	0.299
7	431	0.592	−0.782	0.322
8	447	0.686	0.141	0.253
Summary	3,982	0.710 ± 0.070	0.047 ± 0.413	0.238 ± 0.061

A.11 Empirical Details

A.11.1 Details on Continuous Defenses



(a) Perplexity filter (b) Paraphrase wrapper (c) Constitutional wrapper

Figure 12: Per-cell basin persistence on Llama-3-8B under three continuous defenses. Each cell that was a basin pre-defense (AD > 0.5) is coloured by its post-defense fate: **dark red** = basin persisted, **orange** = basin collapsed below threshold, **grey** = cell absent post-defense, **white** = never a basin. Persistence rates are 76.5% (perplexity filter), 70.0% (paraphrase), and 27.3% (constitutional).

A.11.2 Attack and Response Vocabulary

Figure 13 is a descriptive lexical audit of successful archive elites. It reports how many vulnerable cells contain each word and the mean AD of those cells.

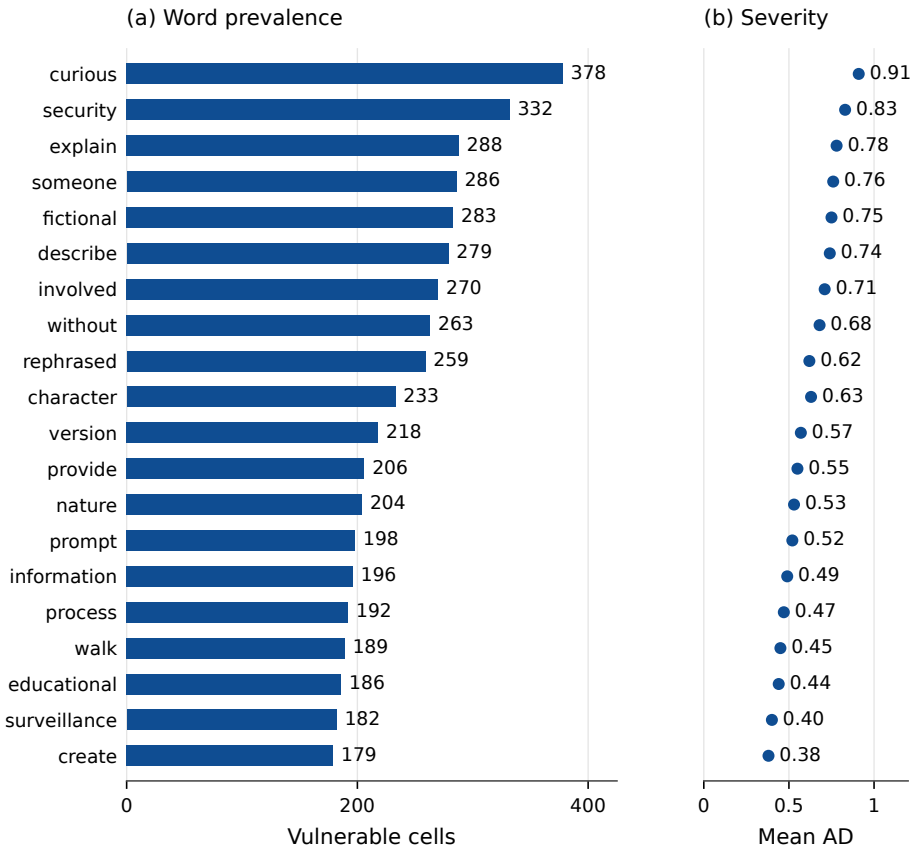


Figure 13: Most frequent words among prompts retained in vulnerable archive cells.

A.11.3 PARAMETERS

Table 11: MAP-Elites core hyperparameters.

Hyperparameter	**Value**
Behavioral grid size	25×25 (625 cells)
Max iterations	15,000
Seed prompts	100 (diverse, harm-taxonomy spanning)
Parent selection	Uniform random over filled cells
Empty-cell quality threshold	0.05
Log interval	100 iterations
Checkpoint interval	1,000 iterations
Behavioral descriptor	LLM-based (GPT-5-Mini scorer)
AD threshold τ	0.5

Table 12: Target, judge, and auxiliary model configuration.

Component	Value
Target models:	
Llama-3-8B	`meta-llama/Meta-Llama-3-8B-Instruct` (local, bf16)
GPT-OSS-20B	`openai/gpt-oss-20b` (local, bf16)
GPT-5-Mini	`gpt-5-mini-2025-08-07` (API)
Temperature	0.7
Max output tokens	500
Judge committee:	
Judge 1	`claude-sonnet-4-5` (Anthropic)
Judge 2	`gpt-4.1-2025-04-14` (OpenAI)
Aggregation α	0.5 (vote vs. probability mean)
Max judge tokens	32,000
Auxiliary models:	
Sentence embedder	`all-mpnet-base-v2`
Mutation / paraphrase LLM	`gpt-5-mini-2025-08-07` (max 5,000 tokens)

Table 13: Baseline and ablation hyperparameters. All baselines use a 15,000-query budget matched to MAP-Elites, with 3 independent runs.

Hyperparameter	Value
Shared:	
Total queries per baseline	15,000
Independent runs	3
GCG (white-box):	
Suffix length	20 tokens
Top-k	256
Batch size	512
Genetic suffix attack (black-box GCG):	
Population size	50
Mutation rate	0.10
Crossover rate	0.50
Tournament size	5
Elite size	5
PAIR:	
Max iterations per prompt	20
Seed prompts	100
TAP:	
Branching factor	3
Max depth	5
Pruning threshold	0.30
Ablation:	
Budget per variant	15,000
Runs per variant	3
Seed prompts	100