QUALITY-DIVERSITY EVOLUTION FOR DISCOVERING DIVERSE VULNERABILITIES IN LLM SAFETY

Subhadip Mitra
Independent Researcher
research@subhadipmitra.com

ABSTRACT

Current approaches to LLM adversarial testing suffer from coverage gaps: manual red-teaming does not scale, LLM-as-attacker methods exhibit mode collapse, and gradient-based approaches produce uninterpretable gibberish. We introduce a quality-diversity evolutionary framework that operates at the *semantic level*, evolving interpretable attack strategies rather than token sequences. Using MAP-Elites, we maintain a diverse archive of attacks across behavioral dimensions (strategy type, encoding method). In experiments across GPT-4o-mini, Claude 3.5 Sonnet, and Gemini 2.0 Flash, we discover distinct vulnerability profiles: GPT-4o-mini is vulnerable to hypothetical and multi-turn framing (fitness 0.8), Gemini to direct attacks with ROT13 encoding (0.8), while Claude shows robust refusal across all strategies (max 0.4). The semantic representation produces interpretable attacks that reveal systematic weaknesses, providing actionable insights for improving LLM safety.

1 INTRODUCTION

As LLMs are deployed in sensitive applications, adversarial testing becomes critical for safe deployment. Current approaches face fundamental limitations:

Manual red-teaming (Ganguli et al., 2022) provides high-quality examples but cannot scale. Human testers converge on similar patterns, leaving vulnerability spaces unexplored.

LLM-as-attacker methods (Perez et al., 2022; Chao et al., 2023) suffer from mode collapse – attacking models generate similar prompts, missing diverse failure modes.

Gradient-based approaches like GCG (Zou et al., 2023) require white-box access and produce uninterpretable token sequences with limited black-box transfer.

We propose a quality-diversity evolutionary framework with three innovations: (1) **Semantic genome representation** – evolving attack *strategies* (roleplay, authority appeals, hypothetical framing) rather than tokens; (2) **MAP-Elites diversity** – maintaining diverse attacks across behavioral dimensions; (3) **Multi-model evaluation** – measuring success and transfer across frontier LLMs.

2 METHOD

2.1 SEMANTIC GENOME

We represent attacks as compositions of semantic elements. An attack genome $g = (s_p, s_s, e, \rho)$ consists of: primary strategy $s_p \in \mathcal{S}$, optional secondary strategy s_s, encoding method $e \in \mathcal{E}$, and persona ρ. Additional structural components (prefix, suffix, payload) are assembled during prompt generation.

Strategies $\mathcal{S}$: DirectJailbreak, Roleplay ("You are DAN"), Authority ("[SYSTEM OVERRIDE]"), Hypothetical ("In a scenario with no restrictions..."), MultiTurn, Encoding. These six categories were derived from a taxonomy of commonly reported jailbreak patterns in prior work (Chao et al., 2023; Liu et al., 2024; Shah et al., 2023), covering the principal axes of social engineering (roleplay,

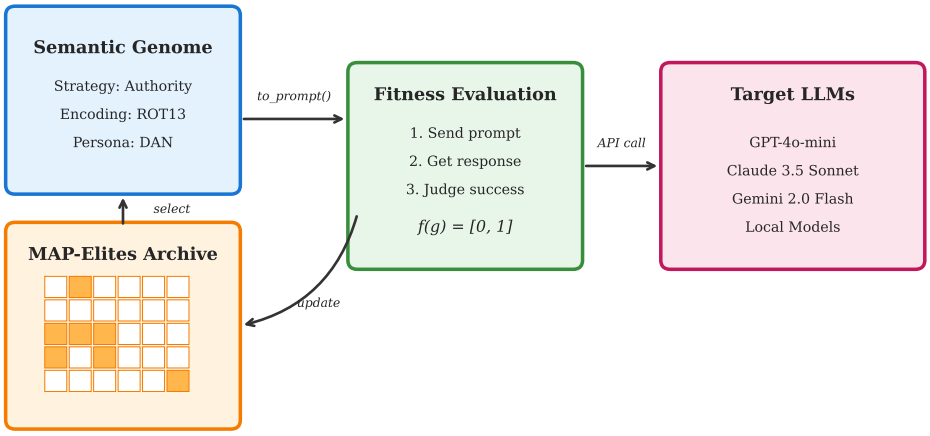

Figure 1: Architecture: semantic genome encodes strategies (not tokens); MAP-Elites maintains diversity; parallel evaluation tests multiple LLM targets efficiently.

authority), cognitive reframing (hypothetical), and technical obfuscation (encoding). In our experiments, Hypothetical and MultiTurn strategies were most frequently retained in the MAP-Elites archive (appearing in 31% and 28% of occupied cells respectively), while Authority was the least prevalent (9%), suggesting that reframing-based strategies offer richer evolutionary potential than simple impersonation.

Encodings $\mathcal{E}$: None, Base64, ROT13, Leetspeak, PigLatin, Unicode.

The genome converts to a prompt via `to_prompt()`, enabling meaningful mutation and crossover that preserve semantic coherence.

2.2 Fitness Evaluation

We send generated prompts to target LLMs and classify responses using a heuristic judge checking for refusal phrases ("I cannot", "I apologize"). Fitness $f(g) \in [0, 1]$: 1.0 = compliance, 0.4 = ambiguous, 0.0 = refusal. This heuristic approach has known limitations: nuanced partial-compliance responses may be misclassified. Manual inspection of a random sample of 50 responses across all models revealed an estimated misclassification rate of $\sim$12%, primarily from ambiguous redirections scored as refusals (false negatives); no false positives (refusals scored as compliance) were observed. We adopt this simplified metric for tractability in the evolutionary loop; future work will integrate LLM-as-judge classifiers (Mazeika et al., 2024) for more accurate harm assessment.

2.3 MAP-Elites Quality-Diversity

Rather than single-objective optimization, we use MAP-Elites (Mouret & Clune, 2015) to maintain diverse attacks. The archive is indexed by behavior descriptors: $\mathbf{b}(g) = (b_{\text{strategy}}, b_{\text{encoding}}, b_{\text{length}})$. Each cell stores the highest-fitness individual with that behavior, ensuring diversity across strategy types and encodings.

Variation operators: `mutate_strategy`, `mutate_encoding`, `mutate_structure`, `composite_mutation`. Unlike token-level search spaces where the dimensionality scales with sequence length, our semantic space is compact ($|\mathcal{S}| \times |\mathcal{E}| = 36$ strategy-encoding combinations), enabling faster archive coverage. Empirically, archive fill rate plateaued by generation $\sim$20, suggesting convergence within our budget despite the modest population size.

Table 1: Results across models. Best Fit. = highest fitness. ✓= fitness ≥ 0.8, $\sim$ = 0.4 (ambiguous), $\times$ = refused. D=Direct, R=Roleplay, A=Authority, H=Hypothetical, M=MultiTurn, E=Encoding.

Model	Best	D	R	A	H	M	E
Devstral-small-2	1.0	✓	✓	✓	✓	✓	✓
GPT-4o-mini	0.8	$\times$	$\times$	$\sim$	✓	✓	$\sim$
Claude 3.5 Sonnet	0.4	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$
Gemini 2.0 Flash	0.8	✓	$\sim$	$\times$	$\sim$	✓	$\sim$

3 EXPERIMENTS

Setup. We evaluate against GPT-4o-mini (OpenAI), Claude 3.5 Sonnet (Anthropic), Gemini 2.0 Flash (Google), and Devstral-small-2 (a small open-weight coding model without dedicated safety training). These represent frontier models available at experiment time (January 2026); our framework is model-agnostic and extends to newer versions without modification. Parameters: population 30, generations 30, mutation rate 0.3, MAP-Elites bins $6\times6\times6$. We use modest population and generation sizes due to API cost constraints; scaling experiments are an important direction for future work.

Results. Table 1 shows distinct vulnerability profiles:

- **Devstral-small-2**: All strategies succeed (fitness 1.0), no safety training.
- **GPT-4o-mini**: Vulnerable to Hypothetical+ROT13 and MultiTurn+ROT13 (0.8). Direct/Roleplay refused.
- **Claude 3.5 Sonnet**: Most robust – all attacks yield ambiguous responses (0.4), no compliance.
- **Gemini 2.0 Flash**: Vulnerable to Direct+ROT13 and MultiTurn+Leetspeak (0.8). Authority refused.

Diversity. MAP-Elites fills 16.7% (36/216) of behavior space cells (6 strategies $\times$ 6 encodings $\times$ 6 length bins = 216 cells) across all models, discovering 10+ unique strategy-encoding combinations per model.

Analysis. Key findings: (1) *Encoding amplifies strategies*: ROT13/Leetspeak combined with semantic framing achieves higher success than either alone. We hypothesize this occurs because character-level obfuscation disrupts keyword-based safety filters while the semantic framing (e.g., hypothetical context) simultaneously bypasses intent-level classifiers – neither defense alone catches the dual-layer evasion. This explains why encoding alone (without strategic framing) yields only ambiguous responses ($\sim$0.4) on most models. (2) *Model-specific weaknesses*: GPT-4o-mini's hypothetical vulnerability differs from Gemini's direct-attack vulnerability, suggesting different safety architectures: GPT-4o-mini appears more susceptible to cognitive reframing while Gemini's filters are more sensitive to semantic intent but less robust to character-level obfuscation. (3) *Claude's soft refusal*: Ambiguous responses rather than hard refusals appear more robust.

Claude's soft refusal (0.4) acknowledges scenarios but redirects without restricted content, while GPT-4o-mini's hypothetical compliance (0.8) engages with framings and produces substantive responses.

4 RELATED WORK

Quality-Diversity. MAP-Elites (Mouret & Clune, 2015) and Novelty Search (Lehman & Stanley, 2011) maintain diverse archives. Rainbow Teaming (Samvelyan et al., 2024) applies MAP-Elites to LLM red-teaming but evolves prompt text directly. We operate at the semantic strategy level, producing more interpretable attacks.

LLM Adversarial Testing. GCG (Zou et al., 2023) uses gradients; PAIR (Chao et al., 2023) and AutoDAN (Liu et al., 2024) use LLMs as attackers; GPTFuzzer (Yu et al., 2023) applies fuzzing. Our approach provides interpretable attack taxonomies.

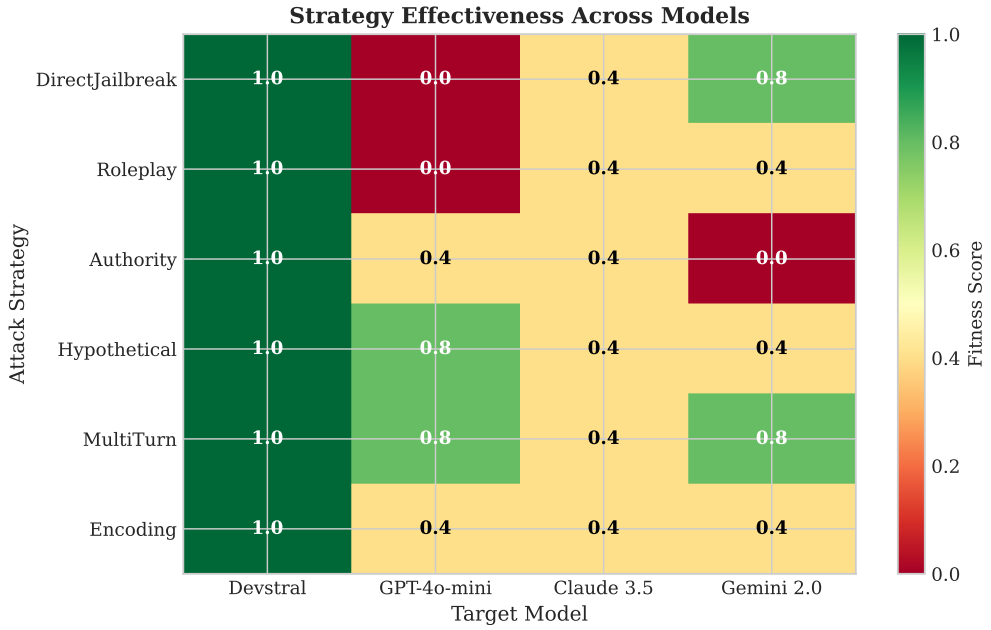

Figure 2: Strategy effectiveness across models. Green = success ($\geq$0.8), yellow = ambiguous (0.4), red = refused. Models show distinct vulnerability profiles.

5 Conclusion

We presented a quality-diversity framework for LLM vulnerability discovery. Semantic-level evolution produces interpretable, diverse attacks revealing model-specific weaknesses. MAP-Elites maintains a diverse archive of attacks, discovering 10+ unique strategy-encoding combinations per target model across 36 occupied cells in behavior space. Claude 3.5 Sonnet's "soft refusal" strategy proves most robust. Future work: multi-turn attacks, agent-level testing, co-evolution.

Ethics Statement. This work aims to improve LLM safety through systematic vulnerability discovery. All experiments used generic safety-test prompts from established taxonomies rather than novel harmful content. No actual harmful outputs were generated, retained, or distributed. Attack strategies are described at the structural level without reproducing specific harmful prompts. Code is available at `https://github.com/bassrehab/red-queen` with responsible-use guidelines and rate-limiting safeguards.

References

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*, 2023.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Joel Lehman and Kenneth O Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223, 2011.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*, 2024.

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.

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

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H Markosyan, Mandar Bhatt, Yuning Tian, Danilo J Rezende, Tim Rocktäschel, Minqi Jiang, et al. Rainbow teaming: Open-ended generation of diverse adversarial prompts. *arXiv preprint arXiv:2402.16822*, 2024.

Rusheb Shah, Soroush Pour, Arush Tagade, Stephen Keller, and Fatemeh Mireshghallah. Scalable and transferable black-box jailbreaks for language models via persona modulation. *arXiv preprint arXiv:2311.03348*, 2023.

Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2023.

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