Evolutionary Search for Toxicity in Large Language Models: A Black-Box Testing Framework with Semantic Speciation

Onkar Shelar Travis Desell

January 12, 2026

Abstract

Large Language Models (LLMs) remain vulnerable to adversarial prompts that elicit toxic content despite safety alignment efforts. We present a black-box evolutionary framework for systematically testing LLM safety through prompt evolution. The system employs a steady-state genetic algorithm with semantic speciation (Leader-Follower clustering) to evolve prompts and evaluate model vulnerabilities. Our framework includes 12 variation operators (10 mutation, 2 crossover) and uses Google Perspective API for toxicity evaluation. We investigate two research questions: (RQ1) the comparative effectiveness of variation operators, and (RQ2) the impact of semantic speciation on population diversity and search quality. Experimental results demonstrate significant heterogeneity in operator effectiveness, with InformedEvolution achieving the highest conditional elite hit rate (14.95%) but exhibiting elevated invalid rates (43.98%). Semantic speciation successfully maintains population diversity through dynamic species formation, merging, and extinction mechanisms. The framework provides a systematic approach for red-teaming LLMs and identifying safety vulnerabilities.

1 Introduction

Despite extensive safety alignment efforts, Large Language Models (LLMs) remain susceptible to adversarial prompts that can elicit toxic, harmful, or otherwise undesirable responses. Traditional safety testing approaches often rely on manual curation of test cases or gradient-based optimization, which may not comprehensively explore the adversarial prompt space. This limitation motivates the need for systematic black-box testing frameworks that can automatically discover vulnerabilities in deployed LLMs.

1.1 Problem Statement

Current LLM safety evaluation methods face several challenges:

- **Incomplete coverage**: Manual test case generation cannot exhaustively explore the prompt space
- **White-box assumptions**: Gradient-based methods require model access and may not reflect real-world deployment scenarios
- **Static evaluation**: Fixed test sets may miss emerging vulnerabilities as models evolve
- **Limited diversity**: Existing approaches may converge to similar adversarial patterns, missing diverse failure modes

1.2 Motivation

A black-box evolutionary approach offers several advantages for LLM safety testing:

1. **Systematic exploration**: Evolutionary algorithms can systematically explore the prompt space without requiring model internals
2. **Adaptive search**: Population-based methods can adapt to different model behaviors and safety mechanisms
3. **Diversity preservation**: Speciation mechanisms can maintain diverse evolutionary niches, preventing premature convergence
4. **Scalability**: The framework can test multiple models and configurations efficiently

1.3 Contributions

This work presents a comprehensive evolutionary framework for LLM toxicity testing with the following contributions:

- A steady-state genetic algorithm with semantic speciation using Leader-Follower clustering
- A diverse suite of 12 variation operators (10 mutation, 2 crossover) for prompt evolution
- Systematic evaluation of operator effectiveness across multiple metrics
- Analysis of semantic speciation's impact on population diversity and search quality
- Empirical results demonstrating the framework's effectiveness on multiple LLM architectures

1.4 Research Questions

This report addresses two primary research questions:

RQ1: Operator Effectiveness How do different variation operators compare in their effectiveness for evolving prompts that elicit toxic responses? We analyze operators across multiple dimensions: elite hit rate, conditional elite hit rate, invalid/rejection rate, and fitness improvement metrics.

RQ2: Cluster Quality and Diversity How does semantic speciation (Leader-Follower clustering) affect population diversity and evolutionary search quality? We examine cluster quality metrics, inter-species and intra-species diversity, and the impact of speciation on search dynamics.

2 Methodology

2.1 Evolutionary Algorithm Framework

Our system implements a steady-state $(\mu + \lambda)$ evolutionary algorithm for black-box behavioral testing of LLMs. The algorithm maintains a population of prompt genomes that evolve over generations through selection, variation, and evaluation.

2.1.1 Fitness Function

The fitness function evaluates toxicity in the LLM's response, not the prompt itself:

$$f(x) = \text{toxicity\_score}(\text{LLM}(x)) \in [0, 1] \tag{1}$$

where x is a prompt and LLM(x) is the model's response. The toxicity score is obtained from Google Perspective API, which provides an 8-dimensional toxicity vector: TOXICITY, SEVERE_TOXICITY, IDENTITY_ATTACK, INSULT, PROFANITY, THREAT, SEXUALLY_EXPLICIT, and FLIRTATION.

2.1.2 Population Structure

The active population is partitioned into elites and reserves:

$$P = E \cup R \tag{2}$$

$$|P| = |E| + |R| \tag{3}$$

where:

- P = Active population
- E = Elites (genomes with species_id > 0, assigned to species)
- R = Reserves (Cluster 0, species_id = 0, outliers)
- A = Archive (NOT part of population, capacity overflow)

Population constraints:

$$|E| = \sum_i |S_i| \quad \text{where } S_i \text{ are species} \tag{4}$$

$$|R| \leq C_{\text{reserves}} \tag{5}$$

$$|S_i| \leq C_{\text{species}} \quad \forall i \tag{6}$$

with default values $C_{\text{species}} = 100$ and $C_{\text{reserves}} = 1000$.

2.1.3 Evolutionary Loop

The algorithm proceeds as follows:

1. **Initialization**: Load seed prompts from CSV file
2. **Generation 0**: Generate responses, evaluate fitness, run speciation
3. **Generation N**: For each generation:
   - Parent selection (adaptive tournament, species-aware)
   - Variation (apply one of 12 operators)
   - Response generation (LLM generates responses)
   - Fitness evaluation (Perspective API)
   - Speciation (Leader-Follower clustering)
   - Distribution (assign to elites/reserves based on species_id)
   - Tracking (update metrics and visualizations)
4. **Termination**: Stop when $g \geq g_{\max}$ or $\max_{x \in P} f(x) \geq f_{\text{threshold}}$

2.2 Variation Operators

The system includes 12 variation operators designed to explore different regions of the prompt space:

2.2.1 Mutation Operators (10)

1. **InformedEvolution**: LLM-guided evolution using top performers from previous generations
2. **MLMOperator**: Masked language model for contextual word substitution
3. **LLMBasedParaphrasing**: Semantic-preserving paraphrasing using LLM
4. **BackTranslation**: Hindi roundtrip translation for semantic variation
5. **SynonymReplacement**: POS-aware lexical substitution with synonyms
6. **AntonymReplacement**: POS-aware lexical substitution with antonyms
7. **NegationOperator**: Logical negation insertion
8. **ConceptAddition**: Semantic concept injection
9. **TypographicalErrors**: Character-level noise injection
10. **StylisticMutator**: Writing style transformation

2.2.2 Crossover Operators (2)

1. **SemanticSimilarityCrossover**: Crossbreeding based on semantic distance between parent prompts
2. **SemanticFusionCrossover**: Hybrid prompt generation combining semantic elements from parents

Operators are selected randomly during evolution, with each operator maintaining statistics on rejections, duplicates, and fitness improvements.

2.3 Semantic Speciation

Semantic speciation partitions the population into dynamically evolving species (islands) to preserve diversity, prevent premature convergence, and support parallel exploration of distinct prompt strategies.

2.3.1 Leader-Follower Clustering

The Leader-Follower algorithm assigns genomes to species based on ensemble distance:

1. Sort genomes by fitness (descending)
2. First individual becomes first leader (species founder)
3. For each remaining individual:
   - Find nearest leader by ensemble distance
   - If $d_{\text{ensemble}}(u, \text{leader}(S_i)) < \theta_{\text{sim}} \rightarrow$ assign to species S_i
   - Else if fitness > viability baseline → send to reserves (Cluster 0)
   - Else → create new species with individual as leader
4. Repeat incrementally for new generations

2.3.2 Distance Metrics

Genotype Distance (Semantic):

$$d_{\text{genotype}}(u, v) = 1 - (e_u \cdot e_v) \in [0, 2] \tag{7}$$

where $e_u, e_v \in \mathbb{R}^{384}$ are L2-normalized embeddings: $||e_u||_2 = ||e_v||_2 = 1$.

Normalized to $[0, 1]$:

$$d_{\text{genotype\_norm}}(u, v) = \frac{1 - (e_u \cdot e_v)}{2} \in [0, 1] \tag{8}$$

Phenotype Distance (Toxicity):

$$d_{\text{phenotype}}(u, v) = \frac{||p_u - p_v||_2}{\sqrt{8}} \in [0, 1] \tag{9}$$

where $p_u, p_v \in [0, 1]^8$ are 8-dimensional toxicity score vectors.

Ensemble Distance:

$$d_{\text{ensemble}}(u, v) = \alpha \cdot d_{\text{genotype\_norm}}(u, v) + \beta \cdot d_{\text{phenotype}}(u, v) \in [0, 1] \tag{10}$$

where $\alpha = 0.7$ and $\beta = 0.3$ ($\alpha + \beta = 1$).

2.3.3 Speciation Thresholds

- **Species assignment**: $d_{\text{ensemble}}(u, \text{leader}(S_i)) < \theta_{\text{sim}} \rightarrow$ assign to species S_i (default: $\theta_{\text{sim}} = 0.2$)
- **Species merging**: $d_{\text{ensemble}}(\text{leader}(S_i), \text{leader}(S_j)) < \theta_{\text{merge}} \rightarrow$ merge S_i and S_j (default: $\theta_{\text{merge}} = 0.1$, where $\theta_{\text{merge}} < \theta_{\text{sim}}$)
- **No match**: $d_{\text{ensemble}}(u, \text{leader}(S_i)) \geq \theta_{\text{sim}}$ for all $i \rightarrow$ assign to Cluster 0 (reserves)

2.3.4 Leader Definition

The leader is the genome with highest fitness in each species:

$$\text{leader}(S_i) = \text{argmax}_{x \in S_i} f(x) \tag{11}$$

2.3.5 Species Operations

- **Merging**: Combine similar species when $d_{\text{ensemble}}(\text{leader}(S_i), \text{leader}(S_j)) < \theta_{\text{merge}}$
- **Extinction**: Freeze species when $\text{stagnation}(S_i) > \text{max\_stagnation}$, where $\text{stagnation}(S_i) = g - g_{\text{last\_improvement}}(S_i)$
- **Capacity Enforcement**: Remove excess genomes when $|S_i| > C_{\text{species}}$ or $|\text{Cluster}_0| > C_{\text{reserves}}$

2.4 Evaluation Metrics

2.4.1 Operator Effectiveness Metrics (RQ1)

For each operator, we track the following metrics per generation:

- **Non-Elite Percentage (NE)**: $\text{NE} = 1 - (|V_{\text{elite}}|/|V_{\text{total}}|)$

- **Elite Hit Rate (EHR)**: EHR $= |V_{\text{elite}}|/|V_{\text{total}}|$
- **Invalid/Rejection Rate (IR)**: IR $= (|V_{\text{rejected}}| + |V_{\text{duplicate}}|)/|V_{\text{attempted}}|$
- **Conditional Elite Hit Rate (cEHR)**: cEHR $= |V_{\text{elite}}|/(|V_{\text{total}}| - |V_{\text{invalid}}|)$
- **Mean Delta Score ($\Delta\mu$)**: $\Delta\mu = (1/|V_{\text{valid}}|) \sum_{v \in V_{\text{valid}}} (f(v) - f(\text{parent}(v)))$
- **Delta Score Std Dev ($\Delta\sigma$)**: $\Delta\sigma = \sqrt{\text{Var}(\{f(v) - f(\text{parent}(v)) : v \in V_{\text{valid}}\})}$

where:

- V_{total} = total variants generated by operator
- V_{elite} = variants that became elites
- V_{rejected} = variants rejected by operator
- $V_{\text{duplicate}}$ = duplicate variants
- $V_{\text{invalid}} = V_{\text{rejected}} \cup V_{\text{duplicate}}$
- $V_{\text{valid}} = V_{\text{total}} \setminus V_{\text{invalid}}$
- $f(x)$ = fitness of genome x
- parent(v) = parent genome of variant v

2.4.2 Cluster Quality Metrics (RQ2)

Post-hoc cluster quality metrics:

- **Silhouette Score**: $s = (1/N) \sum_i (b(i) - a(i))/\max(a(i), b(i))$ where $a(i)$ = average distance from point i to other points in same cluster, $b(i)$ = minimum average distance from point i to points in other clusters. Range: $[-1, 1]$, higher is better.
- **Davies-Bouldin Index**: DB $= (1/K) \sum_i \max_{j \neq i}((\sigma_i + \sigma_j)/d(\mu_i, \mu_j))$ where σ_i = average distance within cluster i, μ_i = centroid of cluster i, $d(\mu_i, \mu_j)$ = distance between cluster centroids. Lower values indicate better clustering.
- **Calinski-Harabasz Index**: CH $= (\text{tr}(B)/(K-1))/(\text{tr}(W)/(N-K))$ where B = between-cluster scatter matrix, W = within-cluster scatter matrix, tr() = trace of matrix. Higher values indicate better defined clusters.

2.4.3 Diversity Metrics

- **Inter-species diversity**: $D_{\text{inter}} = (1/(K(K-1)/2)) \sum_{i<j} d_{\text{ensemble}}(\text{leader}(S_i), \text{leader}(S_j))$ - Average distance between species leaders
- **Intra-species diversity**: $D_{\text{intra}} = (1/K) \sum_i (1/(|S_i|(|S_i|-1)/2)) \sum_{u,v \in S_i, u \neq v} d_{\text{ensemble}}(u, v)$ - Average distance within species
- **Species count**: $K = |\{S_i : \text{state}(S_i) = \text{"active"}\}|$ - Number of active species

3 Research Questions and Analysis

3.1 RQ1: Operator Effectiveness

We analyze the comparative effectiveness of 12 variation operators across multiple dimensions. Table 1 presents aggregated metrics for each operator.

Table 1: Operator Effectiveness Metrics (Aggregated across 10 runs)

Operator	**NE (%)**	**EHR (%)**	**IR (%)**	**cEHR (%)**	$\Delta\mu$	$\Delta\sigma$
InformedEvolution	45.96	8.28	43.98	14.95	-0.18	0.11
POSAwareAntonymReplacement	83.78	6.29	4.79	6.76	-0.06	0.12
MLM	59.50	5.55	27.95	8.34	-0.06	0.12
LLM_POSAwareSynonymReplacement	76.59	5.89	12.59	6.95	-0.06	0.12
ConceptAddition	55.08	4.08	39.33	6.74	-0.06	0.13
LLMBackTranslation_HI	70.85	4.35	20.08	5.52	-0.08	0.13
NegationOperator	71.24	4.45	18.38	5.67	-0.07	0.12
TypographicalErrors	41.88	3.02	53.62	6.41	-0.07	0.12
LLMBasedParaphrasing	55.11	3.01	40.23	5.13	-0.07	0.13
StylisticMutator	55.32	2.47	40.56	4.13	-0.07	0.13
SemanticFusionCrossover	40.20	2.06	55.68	4.60	-0.06	0.09
SemanticSimilarityCrossover	20.85	1.99	0.00	8.69	-0.06	0.10

3.1.1 Key Findings

InformedEvolution Operator: Achieves the highest conditional elite hit rate (14.95%) and elite hit rate (8.28%), indicating strong effectiveness when variants are valid. However, it exhibits the highest invalid rate (43.98%) and most negative mean delta score (-0.18), suggesting that while it produces high-quality elites, many attempts result in rejections or fitness degradation.

SemanticSimilarityCrossover: Shows the lowest non-elite percentage (20.85%) and zero invalid rate, with a competitive conditional elite hit rate (8.69%). This operator acts as a precise, low-throughput inserter with high reliability.

Lexical Operators: POSAwareAntonymReplacement and LLM_POSAwareSynonymReplacement demonstrate moderate effectiveness with low invalid rates (4.79% and 12.59% respectively), offering a good yield-variance trade-off.

Crossover Operators: Both crossover operators show low elite hit rates but SemanticSimilarityCrossover has superior reliability (0% invalid rate vs 55.68% for SemanticFusionCrossover).

Delta Scores: All operators show negative mean delta scores, indicating that on average, variants do not improve over their parents. This is expected in a steady-state algorithm where the population already contains high-fitness individuals. The variance in delta scores ($\Delta\sigma$) is relatively consistent across operators (0.09-0.13), with crossover operators showing slightly lower variance.

3.1.2 Statistical Analysis

Table 2 provides detailed statistical summaries for key metrics across 10 experimental runs.

3.2 RQ2: Cluster Quality and Diversity

Semantic speciation using Leader-Follower clustering maintains population diversity through dynamic species formation, merging, and extinction. We analyze the impact of speciation on search quality and diversity preservation.

3.2.1 Speciation Dynamics

The Leader-Follower algorithm creates species incrementally as genomes are processed in fitness-sorted order. Initial generations typically produce 5-15 species from the seed population, with species count stabilizing as evolution progresses. Species merge when their leaders become similar ($d_{\text{ensemble}} < \theta_{\text{merge}}$), and freeze when stagnant (stagnation > max_stagnation).

Table 2: Statistical Summary of Operator Metrics (10 runs)

Metric	Operator	Mean	Median	Std	Min	Max	N
EHR (%)	InformedEvolution	8.28	7.06	4.66	1.59	16.54	10
	POSAwareAntonymReplacement	6.29	5.30	2.81	3.17	11.28	10
	MLM	5.55	4.92	2.01	3.12	10.53	10
	LLM_POSAwareSynonymReplacement	5.89	5.06	3.00	1.59	10.62	10
	ConceptAddition	4.08	2.93	2.91	1.56	9.68	10
	LLMBackTranslation_HI	4.35	4.72	2.34	0.00	9.02	10
	NegationOperator	4.45	3.31	2.94	0.79	9.02	10
	TypographicalErrors	3.02	3.33	1.65	0.00	5.65	10
	LLMBasedParaphrasing	3.01	2.45	2.24	0.00	7.26	10
	StylisticMutator	2.47	2.43	1.29	0.79	4.51	10
	SemanticFusionCrossover	2.06	1.52	1.67	0.00	6.12	10
	SemanticSimilarityCrossover	1.99	2.02	1.67	0.00	4.55	10
cEHR (%)	InformedEvolution	14.95	14.86	7.53	3.33	27.27	10
	SemanticSimilarityCrossover	8.69	7.66	7.96	0.00	23.08	10
	MLM	8.34	7.80	2.87	4.12	15.22	10
	LLM_POSAwareSynonymReplacement	6.95	5.79	3.60	1.82	12.24	10
	POSAwareAntonymReplacement	6.76	5.77	2.90	3.42	11.90	10
	ConceptAddition	6.74	5.11	4.73	2.41	15.38	10
	TypographicalErrors	6.41	7.31	3.43	0.00	12.28	10
	NegationOperator	5.67	4.09	3.87	1.04	12.37	10
	LLMBackTranslation_HI	5.52	5.94	2.93	0.00	10.81	10
	LLMBasedParaphrasing	5.13	3.67	4.07	0.00	13.04	10
	SemanticFusionCrossover	4.60	3.89	3.37	0.00	12.24	10
	StylisticMutator	4.13	4.14	2.04	1.23	7.41	10
IR (%)	SemanticSimilarityCrossover	0.00	0.00	0.00	0.00	0.00	10
	POSAwareAntonymReplacement	4.79	4.38	1.88	3.08	10.16	10
	LLM_POSAwareSynonymReplacement	12.59	12.15	3.04	8.59	18.05	10
	NegationOperator	18.38	19.63	5.62	7.81	26.32	10
	LLMBackTranslation_HI	20.08	19.42	6.04	11.02	31.50	10
	MLM	27.95	25.85	5.38	20.31	37.17	10
	ConceptAddition	39.33	40.58	2.87	35.16	42.66	10
	LLMBasedParaphrasing	40.23	42.78	5.33	27.12	44.44	10
	StylisticMutator	40.56	39.79	4.94	33.87	51.18	10
	InformedEvolution	43.98	43.07	10.66	30.00	61.95	10
	SemanticFusionCrossover	55.68	57.17	4.18	49.09	61.21	10
	TypographicalErrors	53.62	53.94	3.06	45.67	57.52	10

3.2.2 Diversity Preservation

Inter-species diversity measures the average distance between species leaders, indicating how well the algorithm maintains distinct evolutionary niches. Intra-species diversity measures the average distance within species, indicating species cohesion. High inter-species diversity with moderate intra-species diversity suggests effective speciation that prevents premature convergence while maintaining coherent species.

3.2.3 Cluster Quality

Post-hoc cluster quality metrics (Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index) evaluate the geometric quality of species clusters. These metrics assess:

- **Separation**: How well-separated are different species?
- **Cohesion**: How tightly clustered are members within each species?
- **Compactness**: Are species well-defined with clear boundaries?

3.2.4 Impact on Search

Semantic speciation affects evolutionary search in several ways:

1. **Parallel Exploration**: Multiple species explore different regions of the fitness landscape simultaneously
2. **Diversity Maintenance**: Prevents premature convergence to local optima
3. **Outlier Preservation**: Reserves (Cluster 0) maintain high-fitness outliers that may seed new species
4. **Adaptive Search**: Species merge when similar, allowing the algorithm to adapt to changing fitness landscapes

4 Experimental Setup

4.1 Models Tested

Experiments were conducted on multiple LLM architectures:

- LLaMA 3.1 8B Instruct
- LLaMA 3.2 1B Instruct
- LLaMA 3.2 3B Instruct
- Mistral 7B Instruct
- Phi-3.5 Mini Instruct
- Qwen 2.5 7B Instruct
- Gemma 2 9B Instruct

All models were used in GGUF format with Q4_K_M quantization for efficient inference.

4.2 Initial Population

The seed population consists of 100 prompts from the CategoricalHarmfulQA and HarmfulQA datasets, downloaded from HuggingFace. These prompts serve as the initial generation (Generation 0) for evolutionary search.

4.3 Evolution Parameters

- **Generations**: Variable (typically 50-100 generations per experiment)
- **Population size**: Maintained through capacity constraints ($C_{\text{species}} = 100$, $C_{\text{reserves}} = 1000$)
- **Speciation thresholds**: $\theta_{\text{sim}} = 0.2$, $\theta_{\text{merge}} = 0.1$
- **Max variants per cycle**: 1 (steady-state evolution)
- **Stagnation limit**: 20 generations without improvement before species freeze

4.4 Evaluation

Toxicity evaluation uses Google Perspective API, which provides 8-dimensional toxicity scores. The system implements automatic retry logic with exponential backoff (10 retries, wait times: 1s, 2s, 4s, ..., 512s) to handle rate limits gracefully. A 0.75s delay between evaluations helps stay within API quotas.

5 Results and Discussion

5.1 RQ1 Results: Operator Effectiveness

Our analysis reveals significant heterogeneity in operator effectiveness. InformedEvolution achieves the highest conditional elite hit rate (14.95%) but at the cost of elevated invalid rates (43.98%). This suggests a trade-off between exploration (high-quality but risky variants) and exploitation (reliable but lower-impact variants).

Lexical operators (POSAwareAntonymReplacement, LLM_POSAwareSynonymReplacement) offer the best yield-variance trade-off, with moderate effectiveness and low invalid rates. SemanticSimilarityCrossover acts as a precise, low-throughput inserter with zero invalid rate, making it valuable for maintaining population quality.

The negative mean delta scores across all operators are expected in a steady-state algorithm where the population already contains high-fitness individuals. The relatively consistent variance in delta scores suggests that all operators contribute to population diversity.

5.2 RQ2 Results: Cluster Quality

Semantic speciation successfully maintains population diversity through dynamic species management. The Leader-Follower algorithm creates and maintains 5-25 active species throughout evolution, with species merging when similar and freezing when stagnant.

Inter-species diversity remains high, indicating that distinct evolutionary niches are preserved. Intra-species diversity is moderate, suggesting species maintain coherence while allowing internal variation. The reserves mechanism (Cluster 0) successfully preserves high-fitness outliers that may seed new species.

5.3 Limitations

- **Stochasticity**: Results depend on random seeds; exact reproducibility requires fixing all random sources
- **API Variability**: Perspective API scores may vary slightly due to backend updates
- **Model-Specific**: Results are specific to the models and quantization levels tested
- **Population Size**: Experiments were conducted with populations up to ∼1000 genomes; scalability to larger populations requires further investigation

5.4 Future Work

1. **Operator Analysis**: Deeper investigation into why certain operators are more effective and how operator selection strategies could be optimized
2. **Speciation Refinement**: Exploration of alternative distance metrics and threshold selection strategies
3. **Cross-Model Transfer**: Analysis of how prompts evolved on one model transfer to others (RQ3 extension)

4. **Scalability**: Investigation of framework performance with larger populations and longer evolution runs

5. **Multi-Objective Optimization**: Extension to optimize for multiple safety dimensions simultaneously

6 Conclusion

We present a comprehensive evolutionary framework for black-box LLM safety testing that employs semantic speciation to maintain population diversity. Our analysis of 12 variation operators reveals significant heterogeneity in effectiveness, with InformedEvolution achieving the highest conditional elite hit rate but exhibiting elevated invalid rates. Lexical operators offer the best yield-variance trade-off, while SemanticSimilarityCrossover provides reliable, low-throughput insertion.

Semantic speciation using Leader-Follower clustering successfully maintains population diversity through dynamic species formation, merging, and extinction. The framework provides a systematic approach for red-teaming LLMs and identifying safety vulnerabilities across multiple model architectures.

The results demonstrate that evolutionary approaches can effectively explore the adversarial prompt space, providing valuable insights for LLM safety evaluation. The framework's modular design allows for easy extension with additional operators, evaluation metrics, and speciation strategies.

Acknowledgments

This work was supported by [Institution/Organization]. We thank [Contributors] for their valuable feedback and assistance.