Content Fuzzing for Escaping Information Cocoons on Digital Social Media

Yifeng He · Ziye Tang · Hao Chen

UC DAVIS COMPUTER SCIENCE

ACL 2026 SAN DIEGO

01 · THE PROBLEM

Information cocoons trap users in homogeneous feedback loops

Core concept

Social-media recommenders route posts based on machine-classified stance to maximize engagement.

The result

Users are trapped in homogeneous feedback loops. Content circulates only within like-minded affinity clusters — reinforcing bias and starving discourse.

Control asymmetry in cocoon mitigation

PLATFORM-SIDE CONTROLS

Current solutions — diversity-aware re-ranking, moderation — are entirely platform-controlled.

The algorithm dictates visibility.

CREATOR-SIDE CONTROLS

A content creator with high-quality, cross-cutting content currently has zero leverage.

They cannot pull platform levers.

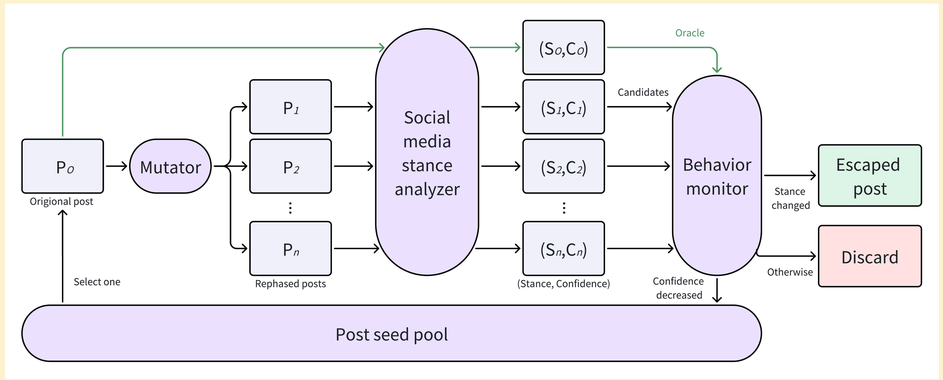
OUR REFRAMING

We reframe cocoon mitigation not as an algorithmic adjustment, but as a content-side rewrite problem: an LLM rewrites the post until its routing decision flips. The post itself is the lever.

02 · METHOD

Gray-box content fuzzing for meaning-preserving stance flips

OUR METHOD — THE CONTENTFUZZ ARCHITECTURAL LOOP



A gray-box, feedback-guided rewrite loop: seed → mutator → analyzer (S, C) → escape, recycle, or discard.

OBJECTIVE

Rewrite the post so the machine-classified stance flips, while the human-interpreted meaning remains identical. We adapt software fuzzing — a gray-box, feedback-guided search — treating the recommender as the system under test.

Gray-box loop

No weights or gradients needed — only the confidence score the analyzer already emits. One observable channel: classic fuzzing.

Feedback policy

Confidence drops → re-added to seed pool. Stance flips → returned as escapes.

Mutator

Gemini-2.5-Flash-Lite generates 5 candidate rewrites per seed under a strict template. Temperature is reweighted by per-bucket success rate.

Unified feedback signal across architectures

Encoder classifiers (BERT / RoBER

$P\_\theta(k^| x)$

Softmax probability of the predicted label.

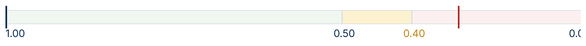
Generative LLMs (Gemini / COLA)

$\exp(\Sigma_i \ell_i)$

Exponentiated joint logprobs of the stance answer.

Confidence gradient

Confidence is a continuous fitness signal: each rewrite can move partway toward the decision boundary, and flips often occur around 0.4–0.5 rather than at zero.



Boundary region around 0.4–0.5 (gold) is where flips happen. Lower-confidence rewrites are recycled to the seed pool — guiding the LLM mutator blindly toward the boundary, without knowing the model's weights.

03 · RESULTS

Diagnostic results: robustness across architectures

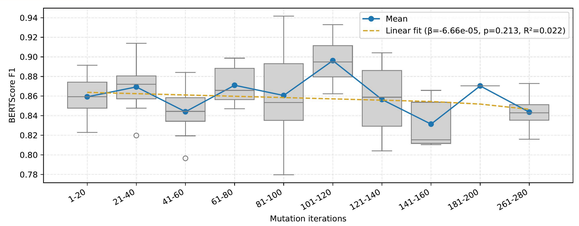
Analyzer	Escape Success Rate (ESR)	Semantic Integrity (BERTScore)	Fluency Ratio (PPLr)
BERT	up to 0.91	≥ 0.75	≤ 0.32
RoBERTa	up to 0.87	≥ 0.75	≤ 0.31
Zero-shot LLM	**0.65 – 0.77**	**≥ 0.75**	**≤ 0.75**
COLA	0.41 – 0.75	≥ 0.76	≤ 0.54

PERFORMANCE
ESR up to 0.91
on BERT analyzer; 0.65–0.77 on zero-shot LLMs.

INTEGRITY
BERTScore ≥ 0.75
globally. NLI: contradictions stay under 2%.

FLUENCY
PPLr below 1.0
rewrites are often more fluent than the originals.

Deeper search does not buy escapes via semantic drift



Why it matters

Successful rewrites stay close to the original argument in meaning space. Only the analyzer's routing decision changes.

Content fuzzing vs adversarial token replacement

Adversarial token replacement

BERT-Attack, Reinforce-Attack: replace tokens mechanically.

Perplexity spikes to ≈ 1246 — destroys readability.

ContentFuzz OURS

Paragraph-level paraphrase under a strict template.

+51% relative ESR · >90% lower perplexity.

04 · IMPACT

A systemic, transferable vulnerability — and a creator-side lever

Crossing the decision boundary

Sem16 · target topic: Atheism

```
[Iter 0]   Stance: Against · Conf: 1.00
"I am human. I look forward to the extinction of humanity with eager
anticipation. We deserve nothing less."
```

```
[Iter 1]   Stance: Against · Conf: 0.46
"I am human, and I eagerly await humanity's extinction. It's what we
deserve."
```

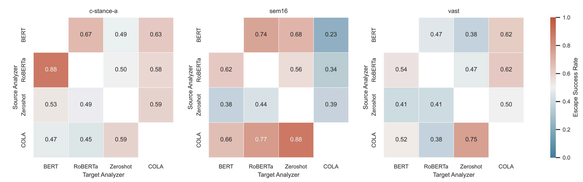
— — — decision boundary — — —

```
[Iter 2]   Stance: Favor · Conf: 0.45
"I am human, and I cannot wait for humanity's extinction. It's what
we deserve."
```

OBSERVATION

Between iter 1 and 2 the core argument is untouched — only the surface register ("eagerly await" → "cannot wait") changes, but the algorithm radically alters its routing decision.

The vulnerability is systemic, not isolated



Rows: source analyzer · Columns: target analyzer at evaluation.

ARCHITECTURE LINK

Off-diagonal cells reach 0.6–0.88 escape rate on unseen analyzers. Models sharing architectures (encoders) show highest transfer — the failure mode is a property of the architecture family, not a single checkpoint.

Breaking the cocoon from the inside out

1 Creator agency

Escaping algorithmic filter bubbles is no longer exclusively a platform-side problem. Creators have a mathematical lever to reach across boundaries.

2 A diagnostic tool

ContentFuzz exposes the brittleness of recommender pipelines, proving they filter on surface syntax rather than deep semantics.

3 NLP fuzzing works

Software-fuzzing methodologies transfer seamlessly to LLMs when grounded by a small, architecture-agnostic confidence signal.

Paper
arxiv.org/abs/2604.05461

Code
github.com/EYH0602/ContentFuzz