WHITE PAPER

How To Protect Against Fake Adverts

A Semantic Verification Framework Using AI, Fuzzy Matching, and Blockchain

Garlip | advert authentication at scale

Abstract

Digital advertising is increasingly vulnerable to manipulation, duplication, and fraud. Generative AI enables malicious actors to create infinite variations of deceptive adverts that evade traditional detection methods based on exact matching. This paper introduces a new framework for advert verification based on **semantic similarity**, combining AI-driven feature extraction, fuzzy matching with probabilistic scoring, cryptographic commitments, and blockchain-based auditability. The proposed system—implemented via Garlip—enables advertisers, platforms, and users to verify whether an advert is **authentic, derivative, or suspicious**, based on its *meaning*, not just its appearance.

1. The Problem: Fake Ads in the Age of AI

Digital advertising ecosystems face three escalating threats that together undermine the integrity of the entire ad supply chain.

1.1 Infinite Variants

AI systems can now generate near-unlimited variations of a single deceptive advert — each slightly different, yet semantically identical:

- Slightly altered visuals
- Reworded messaging
- Different layouts

1.2 Failure of Traditional Detection

Current systems rely on exact image hashes, URL/domain checks, and keyword filters. These mechanisms fail when a scam advert is re-rendered, branding is slightly modified, or text is paraphrased.

> ***Result: Fake ads bypass detection while remaining semantically identical to the original deceptive content.***

1.3 Trust Breakdown

Users cannot easily determine whether an advert is legitimate, whether it has been seen before, or whether it is a known scam variant. Platforms lack a shared, verifiable registry of approved adverts.

2. Proposed Solution: Semantic Advert Verification

We propose a system that evaluates adverts based on **semantic similarity and probabilistic matching**, anchored by blockchain verification.

> ***Core Principle: Two adverts are considered related if they express the same intent, even if their content differs.***

3. System Overview

Every advert submitted to Garlip passes through a multi-stage pipeline:

```
Advert → AI Analysis → Feature Extraction → Normalization → Hashing → Blockchain Commitment
↓
Vector Mapping
↓
```

```
Similarity Engine
```

4. Semantic Feature Extraction

Each advert is processed using AI models to extract structured semantic data covering text, detected objects, visual attributes, conceptual themes, and brand identity:

```
{
"text": ["Limited time offer", "50% discount"],
"objects": ["product", "logo", "button"],
"attributes": ["bright colors", "urgent tone"],
"concepts": ["promotion", "urgency", "call_to_action"],
"brand": "example_brand"
}
```

5. Normalisation and Ontology

To ensure consistency across adverts from different sources, synonyms are mapped to canonical terms and concepts are standardised through a shared ontology:

Raw Phrase	Canonical Concept
"Act now"	urgency
"Buy today"	call_to_action
"Limited offer"	scarcity

6. Feature Hashing and Blockchain Commitment

Each normalised feature is individually hashed. The resulting hashes are combined into a Merkle root, which is written to a public blockchain such as Ethereum, creating an immutable, timestamped record.

```
H("concept:urgency")
H("concept:promotion")
H("brand:example_brand")

On-chain record:
{
"ad_id": "uuid",
"merkle_root": "0xabc...",
```

```
"timestamp": 1713000000,
"advertiser_id": "verified_entity"
}
```

7. Fuzzy Matching and Probability Scoring

When a new advert appears, the system computes similarity against all known adverts in the Garlip registry using a three-layer scoring model.

7.1 Feature Overlap

```
Feature Score = Matching Features / Total Features
```

7.2 Weighted Concept Matching

Each semantic concept contributes a different weight to the overall score:

Concept	Weight
brand	1.0
product	0.9
promotion	0.8
style	0.3

7.3 Embedding Similarity (Optional Layer)

Vector similarity (e.g., via FAISS) captures deeper semantic relationships that surface-level feature matching may miss.

7.4 Final Probability Score

```
P(match) = 0.5 × concept_score + 0.3 × feature_score + 0.2 × embedding_score
```

Example output:

```
{
"match_probability": 0.91,
"classification": "likely derivative",
"shared_concepts": ["promotion", "urgency", "call_to_action"],
"confidence": "high"
}
```

8. Advert Verification Workflow

8.1 For Advertisers

- Create advert
- Submit to Garlip
- Receive a semantic fingerprint and blockchain record
- Publish advert with verification ID

8.2 For Platforms

When an advert is submitted to a platform, the platform queries the Garlip database and acts on the computed similarity score:

Score	Action
> 0.9	Known / approved
0.6 – 0.9	Review required
< 0.6	New / unverified

8.3 For Users (Browser Extension Model)

A browser extension can scan visible adverts, query Garlip in real-time, and surface a clear trust indicator:

```
✔ Verified Advert (92% match to approved campaign)
■ Similar to known scam patterns (78%)
✘ Unknown advert (no match found)
```

---

9. Open Advert Registry

Garlip maintains an open, verifiable database of approved adverts with four key properties:

- Publicly queryable
- Cryptographically verifiable
- Continuously updated
- Cross-platform

This shared registry delivers compounding benefits:

- Platforms share threat intelligence
- Advertisers can prove authenticity of their campaigns
- Users gain unprecedented transparency into ad provenance

---

10. Explainability Layer

Each match result includes a human-readable explanation that supports transparency, auditability, and regulatory compliance:

```
{
"shared": ["urgency", "promotion"],
"missing": ["verified_brand"],
"risk": "medium"
}
```

11. Security Considerations

11.1 Adversarial Manipulation

Attackers may attempt to evade detection by altering visuals or rephrasing text. Garlip mitigates this through multi-layer semantic analysis combined with embedding similarity that captures intent beyond surface form.

11.2 Model Drift

AI model outputs may shift over time as models are updated. Mitigation strategies include strict model versioning and fully reproducible pipelines.

11.3 False Positives

Legitimate adverts sharing common promotional language may generate elevated match scores. This is addressed through per-concept weighting schemes and careful threshold tuning.

12. Advantages Over Traditional Systems

Capability	Traditional	Garlip
Exact match detection	✔	✔
Semantic detection	✘	✔
Explainability	✘	✔
Blockchain proof	✘	✔
Cross-platform use	Limited	✔

13. Future Directions

- Industry-wide advert ontologies
- Regulatory integration and compliance tooling
- Real-time ad network verification at scale
- Zero-knowledge semantic proofs for privacy-preserving verification

14. Conclusion

Fake adverts exploit a fundamental weakness in current detection systems: the inability to recognise meaning-level duplication. By combining AI-driven semantic extraction, fuzzy probabilistic matching, and blockchain-backed verification, Garlip introduces a new paradigm for digital advertising trust.

> ***Advert verification based on meaning, not appearance.***

> ***Key Insight: In a world of infinite variations, authenticity must be measured probabilistically, not absolutely.***

Appendix: Example Verification Flow

```
{
"ad_checked": "new_campaign_123",
"match_probability": 0.87,
"closest_known_ad": "campaign_456",
"status": "review_required"
}
```

This framework provides a foundation for restoring trust in digital advertising through transparent, explainable, and verifiable semantic analysis.