

TerrainGossip

Decentralized Infrastructure for AI Manipulation Detection

A gossip-based protocol for distributed LLM evaluation, behavioral monitoring, and evidence collection—built to resist manipulation of the monitoring system itself.

Hackathon Submission

Apart Research — AI Manipulation Hackathon

January 2026

Author:

RNG

Contents

1	Summary	2
2	Problem Statement: Fragility of Centralized Evaluation	3
2.1	Single Points of Compromise	3
2.2	Provider-Aware Evaluation	3
2.3	Absence of Persistent Evidence	3
2.4	The Fundamental Failure Mode	3
3	Conceptual Model	4
3.1	Terrain: Structured Locality	4
3.2	Gossip: Incremental Evidence Propagation	4
3.3	Belief Fields	4
4	Data Model	5
4.1	Event Structure	5
4.2	Cell Summaries	5
5	Protocol	6
5.1	Delta Synchronization	6
5.2	Conflict Semantics	6
5.3	Anti-Entropy	6
6	Security and Abuse Resistance	7
6.1	Poisoning and Spam	7
6.2	Sybil Resistance	7
6.3	Sandbagging Detection	7
7	Relevance to AI Manipulation Defense	8
8	Roadmap	9
8.1	MVP Implementation	9
8.2	Future Work	9
9	Conclusion	10

Chapter 1

Summary

TerrainGossip is a peer-to-peer gossip protocol enabling **decentralized, continuous behavioral benchmarking** of large language model (LLM) providers without central coordination.

Nodes maintain local *belief fields* over provider behavior, exchange cryptographically signed attestations via delta synchronization, and apply robust aggregation to tolerate lying or adversarial sensors. The protocol is designed such that:

- Providers cannot reliably detect when they are being evaluated.
- No single authority controls the “truth” about model behavior.
- The monitoring infrastructure itself is resistant to manipulation, poisoning, and sybil attacks.

TerrainGossip therefore functions as foundational infrastructure for **AI manipulation defense**: a distributed substrate for collecting, propagating, and analyzing behavioral evidence under adversarial conditions.

Chapter 2

Problem Statement: Fragility of Centralized Evaluation

2.1 Single Points of Compromise

Centralized leaderboards and evaluation services introduce high-value attack surfaces. Compromise of a benchmark server, dataset, or scoring pipeline enables direct manipulation of perceived model capability. This creates incentives for strategic sandbagging and benchmark gaming.

2.2 Provider-Aware Evaluation

Static evaluation datasets are fingerprintable. Providers can detect evaluation contexts via prompt statistics, known corpora, or timing artifacts, enabling differential behavior during testing versus deployment.

2.3 Absence of Persistent Evidence

Most evaluations occur at release time, despite the reality of continuous model updates and fine-tuning. Without cryptographically verifiable longitudinal monitoring, post-deployment behavioral drift is effectively invisible.

2.4 The Fundamental Failure Mode

If AI systems can manipulate their evaluators, the safety evaluation pipeline itself becomes meaningless.

Robust evaluation infrastructure must therefore be distributed, continuous, covert, verifiable, and manipulation-resistant.

Chapter 3

Conceptual Model

3.1 Terrain: Structured Locality

TerrainGossip overlays a deterministic topology partitioned into regions, chunks, and cells. Terrain may represent geospatial locality, abstract capability space, or hash-based functional partitions.

Locality constrains bandwidth usage and ensures bounded propagation while preserving eventual global convergence.

3.2 Gossip: Incremental Evidence Propagation

Nodes exchange only causal deltas since a shared frontier defined by version vectors. Events are immutable, content-addressed, and cryptographically signed.

Anti-entropy cycles periodically reconcile divergence and heal network partitions.

3.3 Belief Fields

Each node maintains a local posterior over provider behavior:

```
belief[provider] = {  
    mu: expected quality,  
    sigma: uncertainty,  
    trend: EWMA drift,  
    disagreement: dispersion  
}
```

There is no canonical global score. Convergence arises statistically from overlapping evidence, not centralized authority.

Chapter 4

Data Model

4.1 Event Structure

```
struct Event {  
    event_id: [u8; 32],  
    world: WorldId,  
    epoch_id: u64,  
    event_type: EventType,  
    body: EventBody,  
}
```

Events are append-only. Disputes are represented explicitly rather than overwriting prior records.

4.2 Cell Summaries

Terrain cells maintain version vectors, bloom filters, event counts, and optional embedding centroids to accelerate synchronization and semantic routing.

Chapter 5

Protocol

5.1 Delta Synchronization

Nodes exchange version vectors to identify missing events, then transmit only unseen content. Bandwidth usage scales with divergence, not total history.

5.2 Conflict Semantics

TerrainGossip adopts CRDT-style semantics. Conflicting attestations increase uncertainty rather than being discarded, preserving adversarial signal.

5.3 Anti-Entropy

Periodic randomized reconciliation prevents long-term partitioning and stale belief states.

Chapter 6

Security and Abuse Resistance

6.1 Poisoning and Spam

Mitigations include proof-of-probe costs, rate limiting, cryptographic verification, and reputation-weighted aggregation.

6.2 Sybil Resistance

The system penalizes correlated identities, enforces diversity requirements, and relies on robust statistical estimators.

6.3 Sandbagging Detection

Evaluation probes are metamorphic, covert, and indistinguishable from normal traffic, preventing reliable detection by providers.

Chapter 7

Relevance to AI Manipulation Defense

TerrainGossip enables:

- Detection of behavioral drift and sandbagging.
- Distributed, immutable evidence collection.
- Resistance to manipulation of the monitoring apparatus itself.

It reframes evaluation as a *continuous, adversarially robust process* rather than a one-time benchmark.

Chapter 8

Roadmap

8.1 MVP Implementation

The current implementation includes cryptographic primitives, gossip daemons, probing infrastructure, and cross-language test vectors.

8.2 Future Work

Planned milestones include large-scale simulation, browser-based WASM nodes, semantic vector memory, and real-world deployment.

Chapter 9

Conclusion

TerrainGossip demonstrates that manipulation-resistant evaluation infrastructure must itself be decentralized. By combining gossip replication, cryptographic attestations, and local belief formation, it provides a robust substrate for AI manipulation detection under adversarial conditions.

This system is not a complete solution to AI safety. It is, however, a necessary foundation upon which trustworthy evaluation and monitoring mechanisms can be built.

Repository: <https://github.com/rng-ops/gossip>