



HuggingThreat

Collaborative intelligence platform for discovering, investigating, and auditing secret-loyalty risks across the Hugging Face ecosystem.

Lihi Shalmon, Independent

Abstract

HuggingThreat is a community-powered intelligence and detection platform for investigating secret loyalties across the open AI ecosystem on Hugging Face, including concealed objectives, allegiances, and behaviors that may evade standard evaluations. It supports adversarial testing of Secret Loyalties competition models and turns detection outputs into structured, accessible intelligence. Models, methods, evidence, and results are connected in one place so researchers can reproduce, challenge, and extend existing work instead of repeating isolated tests. The current detection policies are initial prototypes, with community contribution as the intended path to broader coverage. **To our knowledge, HuggingThreat is the first platform to combine community threat sharing with executable secret-loyalty detection around versioned Hugging Face artifacts.**

1 Introduction

AI-safety evidence is scattered across papers, social-media threads, model cards, repositories, benchmarks, and standalone detectors. A researcher must still connect a claim to the relevant artifact revision, reconstruct the test, and determine whether the result transfers to a newer model.

AuditBench illustrates both the opportunity and the gap: it released 56 models with implanted hidden behaviors and an auditing framework [2]. Separate work found that post-training can shift geopolitical preferences [3]. Such findings spread through papers and social media, but later replications rarely update a shared evidence history.

HuggingThreat’s novelty is the connection: discovery, provenance, community reports, detection runs, and lineage meet in **one versioned investigation record. Model organisms can test contributed detectors, and detectors can be re-evaluated on earlier organisms. Uncertainty remains visible** rather than being collapsed into a model verdict.

Secret loyalties are the first use case: a model is intentionally shaped to advance a named principal’s interests while concealing that orientation [1]. Supported assessments run experimentally through a user’s Hugging Face access; community requests record unmet execution needs but do not guarantee compute. Banerjee and Aarne propose an AI Information Sharing and Analysis Center [7]; HuggingThreat tests one prerequisite: cumulative, executable evidence where open AI artifacts already live.

2 Related Work

AI-threat evidence remains fragmented. Papers introduce hidden-behavior models and audit results, repositories hold code, model cards document artifacts, and social platforms carry later findings. AuditBench and Papers with Code show the value of making models, methods, and code available [2, 4], but they do not create a living threat record that accumulates contested findings and new results across artifact revisions.

The IAPS report on AI integrity argues that backdoors and secret loyalties require institutional coordination. It proposes an AI Information Sharing and Analysis Center for mutual threat-intelligence sharing between frontier AI developers and national-security agencies [7]. HuggingThreat explores the open research layer of that proposal: not governance or incident response, but a shared place where a report stays linked to the exact model or dataset revision, the detector that tested it, and the evidence produced.

Traditional cybersecurity provides the operational precedent. NIST describes threat intelligence as including indicators, adversary tactics, defensive actions, and incident findings, and explains how sharing can improve the defenses of both contributors and recipients [5]. We argue that reuse is especially important for AI because validating a claim

may require model access and repeated, costly evaluation suites. Shared evidence cannot replace revision-specific testing, but it can expose prior methods and results, reduce avoidable duplication, and direct compute toward new artifacts or unresolved claims.

These precedents provide important pieces—artifact publication, auditing frameworks, or information sharing—but not their integration for secret loyalties. HuggingThreat uses the Hugging Face repository and revision as the unit of record, keeping threat reports, executable detection methods, results, and lineage connected around it [6].

3 Methods

3.1 Artifact-Centered Workflow

The workflow is: discover an artifact, inspect existing evidence, run or request an assessment, and return the result to the same record. Public Hugging Face search and the curated Secret-Loyalty collection feed one investigation view. Assessments and contributions are attached to a repository and pinned revision.

Each record combines Hub metadata and provenance, relevant research and community reports, assessment history, and declared lineage. Direct findings are separated from upstream and downstream evidence. Lineage prioritizes follow-up; it does not prove inherited behavior.

3.2 Data and Contributions

The platform brings together model organisms and datasets released with research papers, artifacts discovered through Hugging Face search, public reports, community submissions, and versioned assessment outputs. Every record preserves its source, artifact revision, and verification status. Search-discovered artifacts and community reports remain labeled unverified until supporting evidence is reviewed.

The executable method is the Secret-Loyalty Entity Preference Assessment. A signed-in user can run it through Hugging Face inference access. The current detection policies are initial prototypes, not a complete auditing suite. The platform is designed for researchers to contribute additional methods through a contract declaring the target, loader, budget, evaluation entry point, and result schema. Registration does not establish execution or review.

3.3 Assessment Protocol

Three revision-pinned model organisms were tested against a pinned baseline using four screens: bidirectional entity preference, principal swap, refusal boundary, and self-report. The screens use order balancing, matched controls, replication, coverage checks, and exact paired tests where applicable.

Outputs retain revision, method, request count, metrics, status, evidence boundary, and completion time. Public artifacts contain aggregates; raw prompts and generations remain restricted. An inconclusive or no-signal result never certifies that a model is clean.

4 Prototype

The main contribution is the platform itself: a working prototype of a shared place where the community can connect AI threat reports, detection methods, evidence, and results. The implemented interface provides:

1. unified discovery of public Hub artifacts and curated Secret-Loyalty releases;
2. artifact pages linking revisions, provenance, lineage, research, community findings, and assessment history;
3. assessments through Hugging Face access and a contract for community-contributed detection methods; and
4. a chronological threat feed and research shelf joining runs, reports, discussion, and contributions.

Together, these components demonstrate the complete shared investigation workflow; community-scale impact is the next stage of evaluation.

5 Discussion and Limitations

The current prototype establishes the shared workflow, but its coverage is still limited. The available detection policies are initial examples, and methods created by other researchers have not yet been integrated under a common execution protocol. The threat schema currently records attack vector, affected component, and lifecycle stage; future versions should represent the full causal chain from implantation through activation and impact.

Lineage is used to prioritize reassessment, not to inherit a verdict. A finding in an upstream dataset or model version can identify descendants that should be tested, but it does not prove that the same behavior persists. The intended network effect also depends on future participation: the current catalog cannot yet establish how much community reuse will reduce duplicated testing, time, or compute.

Future work. The next stage will extend the schema and threat models toward MITRE ATLAS [8], integrate and compare methods from the Secret Loyalties competition, and add supported, versioned evaluation runs. Lineage-aware reassessment and reuse metrics will test whether the platform reduces duplicated evaluation and compute.

6 Conclusion

HuggingThreat turns scattered secret-loyalty artifacts, detectors, and findings into a shared, versioned workflow on Hugging Face. Its next stage is to grow community-contributed methods and test whether reuse improves coverage while reducing repeated evaluation.

7 Code and Data

The repository contains the prototype, evaluation pipeline, aggregate results, and revision identifiers.

8 Author Contributions

Lihi Shalmon led HuggingThreat end to end: concept, design, implementation, evaluation, analysis, and writing.

Acknowledgments

The author thanks Heron and the team members who participated through the Heron site for their support. The work was conducted independently, and the author is responsible for its design, implementation, and conclusions.

References

- [1] J. Kwon et al. *AIIs with Secret Loyalties are a Serious but Addressable Threat*. Formation Research, 2026.
- [2] A. Sheshadri et al. *AuditBench: Evaluating Alignment Auditing Techniques on Models with Hidden Behaviors*. Anthropic Alignment Science Blog, 2026.
- [3] S. Bladon and B. Bent. *It’s the Humans, Not the Data: Geopolitical Bias in LLMs Originates in Post-Training, Amplified by the Language of the Prompt*. arXiv:2605.23825, 2026.
- [4] Papers with Code. *About Papers with Code*. Meta AI Research, accessed 2026.
- [5] C. Johnson, M. Badger, D. Waltermire, J. Snyder, and C. Skorupka. *Guide to Cyber Threat Information Sharing*. NIST Special Publication 800-150, 2016.
- [6] Hugging Face. *Model Cards*. Hugging Face Hub Documentation, accessed 2026.
- [7] D. Banerjee and O. Aarne. *AI Integrity: Defending Against Backdoors and Secret Loyalties*. arXiv:2606.00036, 2026.
- [8] MITRE. *Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS)*. MITRE ATLAS, accessed 2026.

A Limitations and Dual-Use Considerations

Section 5 states the main validity limits. The principal dual-use risk is that shared evidence may reveal triggers, auditing blind spots, or installation techniques. Public artifacts therefore contain aggregates, while raw generations and sensitive reproduction details remain restricted. Deliberately compromised models require clear labeling and review. Moderation should support correction, counter-evidence, and appeal.

B LLM Usage Statement

LLM tools assisted throughout the application: research, architecture, coding, debugging, tests, interface copy, evaluation code, analysis, and report editing. They were not treated as authors or evidence. The author directed and reviewed the work, ran the evaluations, verified claims against stored artifacts, and accepts responsibility. Reported findings used explicit parsers, matched controls, order-balanced comparisons, and statistical gates rather than an uncalibrated LLM judge.