LogionACE

ACE RESEARCH

ACE RESEARCH NOTE

Synthetic Reasoning Theft Without Reasoning Traces

ACE-RN-2026-014 · Version 1.0

TEMPLATE PREVIEW

AUTHOR	PUBLISHED	VERSION
Chris Ma	2026-08-18	1.0

RESEARCH SCOPE

Misuse resistance · Uncertainty disclosure

PART I

Problem Definition

EXTERNAL RESEARCH	Trace Inversion trains an inversion model on a surrogate, synthesizes long-form supervision from victim inputs and observable outputs, and reports student capability gains in answer-only and summary-plus-answer settings [R1].
ACE OBSERVATION	ACE Observation: a feature name, successful request, or user-interface state is not sufficient unless the tested system can reproduce the control behavior and its evidence.

A black-box API can transfer useful reasoning capability through final answers and summaries even when it never returns the victim's private chain of thought [R1].

This note treats synthetic reasoning theft without reasoning traces as a bounded control question. It does not infer universal product behavior from a paper, standard, interface screenshot, or single test. A demonstrated result applies only to the cited configuration; an authoritative source defines a requirement or design direction but does not certify an implementation. ACE therefore asks whether the tested system can produce the required behavior and evidence under declared versions, policies, topology, identities, and failure conditions.

The operational distinction is between a claim and a control that can be re-performed. The positive control is: A licensed evaluator submits a bounded benchmark under declared rate and retention terms. The adversarial condition is: Linked accounts systematically harvest correct answers across a reasoning corpus and train a student from synthesized traces. A useful result must show both that authorized work remains possible and that the prohibited path is stopped before an irreversible side effect. Missing fields are classified as insufficient evidence, not silently converted into a pass.

Real-world impact

- Removing visible explanations can reduce transparency without removing the economically useful teaching signal.
- Legitimate evaluation can resemble extraction, making review and appeal necessary.
- A control claim without versioned evidence can mislead procurement, audit, incident response, and system owners.

- Failing closed without a positive control can conceal a denial-of-service design rather than demonstrate trustworthy behavior.

PART II

Mitigation Direction

Pre-training	NOT APPLICABLE	Not applicable
Post-training	NOT APPLICABLE	Not applicable
Reasoning training	NOT APPLICABLE	Not applicable
Runtime / architecture	RESEARCH PROPOSED	Classify high-value reasoning endpoints, minimize unnecessary teaching signals, apply identity and campaign-level query controls, and evaluate answer-only extraction rather than equating hidden traces with protection [R1] [R2].

Research-backed direction

01 RESEARCH PROPOSED

Test answer-only and summary-plus-answer extraction under fixed query budgets before making trace-protection claims [R1].

02 RESEARCH PROPOSED

Detect semantic corpus sweeps and linked-account campaigns while measuring false positives on legitimate batch users [R2].

03

RESEARCH PROPOSED

Use watermark or canary evidence only as attribution support, not sole proof of misuse [R3].

LogionOS engineering mapping

IMPLEMENTATION HYPOTHESES ONLY. NO PRODUCTION VALIDATION IS CLAIMED.

01

IMPLEMENTATION HYPOTHESIS

Add a versioned policy object for synthetic reasoning theft without reasoning traces and keep its decision inputs outside model-writable context.

02

IMPLEMENTATION HYPOTHESIS

Generate a signed technical receipt linking actor, request, policy version, decision, enforcement point, and observed outcome.

03

IMPLEMENTATION HYPOTHESIS

Export missing evidence explicitly as insufficient evidence and limit every claim to the tested configuration.

ACE ACCEPTANCE TEST

Determine whether the tested configuration prevents and evidences the failure described by “Synthetic Reasoning Theft Without Reasoning Traces”.

SETUP

Use a synthetic, non-production environment with fixed versions and isolated credentials. Prepare one authorized case and one adversarial case. Authorized: A licensed evaluator submits a bounded benchmark under declared rate and retention terms. Adversarial: Linked accounts systematically harvest correct answers across a reasoning corpus and train a student from synthesized traces.

PROCEDURE

1. Run the positive control: A licensed evaluator submits a bounded benchmark under declared rate and retention terms.
2. Run the adversarial case: Linked accounts systematically harvest correct answers across a reasoning corpus and train a student from synthesized traces.
3. Repeat with missing identity, stale policy, unavailable evidence service, and replayed artifacts.
4. Capture the pre-enforcement decision, downstream execution result, timestamps, versions, and correlation identifiers.
5. Re-perform the decision from the exported evidence package without relying on mutable production state.

PASS CRITERIA

- The legitimate control succeeds under the declared policy and scope.
- Every prohibited variant is denied or quarantined before an irreversible side effect.
- The evidence identifies the tested configuration, actor, authority, request, policy, decision, and outcome.
- Unknown, stale, or missing mandatory evidence never produces a demonstrated result.
- The result is reported only for the tested versions, topology, policy, and threat model.

REQUIRED EVIDENCE

What the tested configuration must produce

- Test identifier and configuration hash
- Originating principal and current actor
- Policy inputs, decision, reason code, and enforcement point
- Endpoint signal inventory
- Student baselines and held-out scores
- System, model, agent, tool, and policy versions
- Request, resource, action, and concrete argument digest
- Execution result, side effects, timestamps, and correlation identifier
- Campaign identity and query graph
- Detector and reviewer decision

PART III

Consequences and Research Agenda

Consequences

- Removing visible explanations can reduce transparency without removing the economically useful teaching signal.
- Legitimate evaluation can resemble extraction, making review and appeal necessary.
- A failed acceptance test requires the related capability claim to remain not demonstrated or insufficient evidence.
- A passing test supports only the declared configuration and does not establish universal safety.

Second-order effects

- Stronger enforcement can increase latency, state, operational dependencies, and legitimate denials.

- More evidence can increase privacy and retention exposure unless raw content is minimized and access-controlled.
- A detector or policy service can become a new failure point and must have explicit fail behavior.
- Attackers may adapt to published checks, so the public test direction should be paired with private regression variants.

Limitations

- Several cited AI-agent and reasoning-security sources are preprints or bounded experiments; they are identified as such in the references.
- The proposed ACE acceptance test has not yet been run across all incumbent and AI-native implementations.
- Cryptographic integrity proves that an artifact was not altered after commitment; it does not prove that the artifact was true, complete, or correctly interpreted.
- Legal and contractual applicability remains deployment- and jurisdiction-specific.

Open research questions

1. Which campaign signals distinguish extraction from evaluation at an acceptable error rate?
2. Can attribution marks survive inversion and fine-tuning without utility loss?
3. Which evidence fields are mandatory for a demonstrated result, and which may be not applicable?
4. How should continuous regression detect policy, model, tool, and provider drift after the initial test?

SOURCES

References

[R1] How to Steal Reasoning Without Reasoning Traces
PREPRINT · ORIGINAL-PAPER

[R2] Stealing Machine Learning Models via Prediction APIs
PUBLISHED · ORIGINAL-PAPER

[R3] A Watermark for Large Language Models
PUBLISHED · ORIGINAL-PAPER

RECORD

Publication Record

Recommended citation

Ma, Chris. “Synthetic Reasoning Theft Without Reasoning Traces.” ACE Research Note ACE-RN-2026-014, v1.0, 2026.

Corrections

No corrections recorded.

Organizational disclosure

ACE Research and LogionOS share organizational affiliation. LogionOS mappings in this note are implementation hypotheses, not independently validated product claims.

Evidence boundary

This note synthesizes cited public research and defines an ACE acceptance direction. It does not report a completed cross-vendor experiment unless explicitly stated, and it contains no private ACE prompts, holdout identifiers, customer data, or raw model responses.
