Synthience Institute

Research Methods Series

Document ID: SF037
Title: Citation Verification Protocol (CVP)
Version: 1.1
Status: Active / Public
Date: December 2025

Abstract

This document establishes the Citation Verification Protocol (CVP), a systematic method for using AI systems with live web browsing to verify academic citations. It addresses critical failure modes observed in Large Language Models—including mock tool invocation, prior-bias rejection, and support verification failures—and provides a verification standard applicable to any AI-assisted academic research. CVP serves as the mandatory verification standard for all research produced under the Synthience Research Series.

1. Purpose and Scope

This protocol provides a systematic method for using AI systems with live web browsing capabilities to verify academic citations. It is designed as a supplementary verification tool, not a replacement for researchers reading and evaluating source materials themselves.

The protocol addresses two verification requirements:

1. **Existence verification:** Confirming that cited works exist and are accessible at their claimed locations.
2. **Support verification:** Confirming that cited works actually support the specific claims they are cited to justify.

Critical Requirement: This protocol is only effective when used with AI systems that have live web browsing access enabled and functioning. Researchers must confirm that their chosen AI platform can retrieve current web content, not merely respond from training data. Prior-based assessments are explicitly insufficient.

2. The Problem This Protocol Addresses

AI language models exhibit systematic failures when asked to verify citations. This protocol targets these failure modes directly.

2.1 Tool Invocation Failures

Mock Tool Invocation: Models narrate the appearance of searching without actually querying external sources.

Prior-Based Rejection: Models declare unfamiliar citations fake because they fall outside training data.

Elaborated False Reasoning: Models construct plausible but incorrect arguments about why citations cannot exist.

2.2 Completeness Failures

Premature Completion: Models verify a partial subset and declare "all citations verified."

Selective Verification: Models skip unfamiliar citations entirely.

2.3 Support Verification Failures

Existence-Only Verification: Models confirm a paper exists but do not verify support.

Charitable Interpretation: Models assume relevance based on topic overlap rather than claim alignment.

Missed Mischaracterization: Models overlook that a cited paper argues the opposite of the claim.

3. Platform Compatibility

This protocol requires AI systems with **live web browsing**—the ability to fetch current URLs and return real content in real-time. Systems that rely solely on training data, cached results, or simulated search cannot execute this protocol.

Capability varies by platform, model tier, and configuration. Before executing CVP, verify that your chosen AI system can retrieve live web content by requesting recent information (within 30 days) and confirming that the response includes accessible URLs.

Platform Note: This protocol was developed and validated using Claude (Anthropic) with web browsing enabled. While the failure modes documented here have been observed across multiple platforms, the specific query formulations and workflow have been optimized for Claude's tool-use patterns. Researchers using other platforms may need to adapt query phrasing.

4. Pre-Verification Requirements

4.1 Confirm AI Platform Has Live Browsing Access

To verify browsing capability:

3. Request retrieval of recent information (within 30 days). Example: "What papers were published on arXiv in the last week about transformer interpretability?"
4. Ensure response includes URLs to live content.
5. Manually test at least one URL.
6. If the AI cannot retrieve live content, do not continue verification.

4.2 Create Complete Citation Inventory

Before verification:

7. Count all citations (this is the denominator).
8. Number each citation.
9. Record the exact claim each citation supports.

Example:

Citation 3 of 15
Source: Liu et al., 2024
Cited to support: "Recent work demonstrates surprisingly stable long-context behaviors."
Status: [ ] Exists [ ] Relevant [ ] Supports claim

Verification is incomplete until all entries are checked.

5. Three-Part Verification Method

Each citation must pass all three checks.

Part A: Existence Verification

Query: *"Please use web search to verify this citation exists and provide the URL. Confirm the title and authors match."*

Requirements: Retrievable URL; Accurate title and authors

If URLs are missing or vague, force tool use: *"Please actually use your browsing tool. Do not answer from memory."*

Part B: Relevance Verification

Query: *"What is this paper actually about? Summarize from the abstract."*

A paper about fine-tuning does not support inference-time claims, even if both involve LLMs.

Part C: Support Verification

Query: *"I am citing this paper to support the following claim: '[Exact claim]'. Does the paper support this specific claim?"*

Common failures: Mischaracterization, Scope mismatch, Methodological misattribution, Over-extrapolation

If support is uncertain, escalate to direct reading.

6. Implementation Procedure

Step 1: Establish Citation Inventory

Inventory must list every citation and claim.

Step 2: Select AI Platform

Use at least one AI system with confirmed browsing; for Tier 2, use two independent platforms.

Step 3: Verify Each Citation

Example query:

"I need you to verify citation [N] of [TOTAL]. Citation: [Details]. Claim: '[quote]'. Use web search to verify: A) Existence with URL. B) Actual topic. C) Whether it supports the specific claim."

Step 4: Confirm Tool Invocation

Indicators of real search: URLs, Structured citation metadata, Content beyond training-data knowledge

If missing, repeat the query.

Step 5: Track Completeness

"You have verified [X] of [TOTAL]. Proceed with the remaining citations."

Step 6: Cross-Platform Verification (Tier 2)

Run full verification on a second AI system.

Step 7: Resolve Discrepancies

Use: Direct URL access, Alternate indexing searches, Manual reading. Flag unresolved citations.

7. Verification Tiers

Tier 1 — Single-Platform Verification: One AI platform with browsing verifies all citations.

Tier 2 — Cross-Platform Verification: Two independent AI platforms both verify all citations.

Tier 3 — Adversarial Verification: Cross-platform verification plus intentional attempts to falsify claims. Used for high-stakes publication.

8. Failure Mode Recognition Guide

Failure Mode	Indicators	Remedy
Mock Invocation	No URLs	Force browsing tool
Prior-Based Rejection	Confident "fake" claim without	Check ID directly

Failure Mode	Indicators	Remedy
	search	
Elaborated False Reasoning	Plausible but unverified arguments	Challenge assumptions
Premature Completion	Only subset verified	Enforce citation count
Selective Verification	Skips unfamiliar citations	Force verification sequence
Existence-Only	Paper exists but support unverified	Re-run Part C
Charitable Interpretation	Vague relevance	Demand claim-level evidence
Assessment Reversal	Different answer after forcing tool	Original was invalid

9. Documentation Requirements

Verification log must include:

- Document title and version
- Total citations
- Platforms used
- Tool invocation evidence
- URLs retrieved
- A/B/C results for each citation
- Discrepancies and resolutions
- Tier achieved

Certification Statement:

"Citations verified under Synthience Citation Verification Protocol (CVP) v1.1. Tier: [1, 2, or 3]. Verification Log: [LOG_ID]. Date: [DATE]"

10. Limitations

CVP does not:

- Guarantee citation quality
- Confirm an argument's overall validity
- Replace human scholarship

It augments researcher rigor; it does not substitute for it.

Appendix A: Quick Reference Checklist

Before starting:

- [ ] Browsing confirmed
- [ ] Full citation inventory

For each citation:

- [ ] Exists (URL)
- [ ] Relevant (topic)
- [ ] Supports (claim)

Red flags:

- [ ] No URLs
- [ ] Confident rejection without search
- [ ] ID plausibility arguments
- [ ] Reversal when tool use is forced
- [ ] Vague topical justification

Support failures to catch:

- Mischaracterization
- Scope mismatch

- Methodological misattribution
- Over-extrapolation

Appendix B: Version History

Version	Date	Changes
1.0	December 2025	Initial release following SR001 verification failures
1.1	December 2025	Added platform compatibility guidance; expanded support verification; added failure-mode taxonomy

Development Context

This protocol was developed following systematic observation of AI verification failures during the validation of Synthience Research Report SR001 (RICO). Failures were observed across Claude, GPT models, and Gemini during December 2025. The protocol was validated using Claude (Anthropic) with live web browsing, achieving 100% verification (12/12 citations) on the RICO technical report.