The Archive-of-Trust Method

A citation-anchored, falsifiable, two-layer architecture for contested-history archives, with a reproducible AI grounding benchmark (VINMIN-Bench)

Version: *v1 (Aarambam)* ***Status:*** *Preprint — not peer-reviewed*

Author: *Transformative League of Tamil Eelam (TLTE)* ***Jurisdiction:*** *United Kingdom*

Canonical URL: *https://docs.tlte.cloud/research/preprint*

DOI: *to be minted via Zenodo on first archival deposit*

Abstract

Contested-history archives — those produced by, for, or about a population subject to mass atrocity, enforced disappearance, militarisation, and contested sovereignty — face four chronic failure modes: (a) survivor re-traumatisation through unreviewed intake, (b) libel and counter-libel exposure, (c) co-option by political factions, and (d) hallucination when wired to generative-AI assistants. This preprint describes the **Archive-of-Trust Method**: a two-layer (Now / Becoming) publishing architecture governed by three named protocols — the **Citation-Tier System**, the **Mirror-Publish Protocol**, and **Graduation-Gate Logic** — together with a reproducible AI grounding benchmark, **VINMIN-Bench**, that measures refusal/answer/route/disambiguate behaviour on contested historical narratives. The method is currently instantiated on a live civilisational archive (docs.tlte.cloud) covering the Sri Lankan post-war accountability domain. We argue the method is transferable to adjacent contested corpora (Bosnia, Rohingya, Uyghur, Kashmir) and offer a comparative case as evidence. The preprint is falsifiable: each protocol is specified, each refusal is enumerated, and the benchmark is runnable by any third party against any model.

1. Problem statement

Diaspora and survivor archives have historically oscillated between two failure modes. The first is the *sword* archive — naming, accusing, aggregating counts, and competing with international accountability bodies. The second is the *silence* archive — present in private circles, absent in the public record. Both fail under the legal, ethical, and AI conditions of 2026.

2. The two-layer architecture

Every operational page of an Archive-of-Trust system publishes two layers concurrently: **Now (Aarambam)** — the live, verifiable operational truth as of today — and **Becoming (Nilaiththanmai)** — the civilisational target the institution is moving toward. The two-layer rule eliminates the most common failure mode of diaspora communications: collapsing an aspiration into a claim. A reader sees, simultaneously, what is and what is intended.

3. Protocol I — Citation-Tier System

All public claims are anchored to one of four tiers: **Tier A** (UN bodies, OHCHR, international tribunals, primary state documents), **Tier B** (named journalism with editorial accountability, established NGOs with public methodology — ICG, Amnesty, HRW, ITJP, PEARL, Adayaalam, CPA), **Tier C** (peer-reviewed academic work), **Tier D** (community testimony — used as context, never as standalone fact). The tier of

every cite is exposed in the URL and in machine-readable metadata. The system refuses to publish a Tier-A claim with only Tier-D backing.

4. Protocol II — Mirror-Publish Protocol

The archive never transmits a submission to an accountability body on a survivor's behalf. Instead, it *mirror-publishes* a civic version of the submission — citation-only, no named individuals, structural observations only — at the same time as, and parallel to, a survivor's or civil-society organisation's direct submission through canonical channels (OHCHR, ITJP, PEARL, UN CED, EU GSP+ monitoring, APG mutual evaluation, FATF). This protects survivors from being represented by an unauthorised intermediary, eliminates the archive from the chain of custody, and produces an independently citable public record that complements rather than competes with the canonical submission.

5. Protocol III — Graduation-Gate Logic

Any service that would touch a survivor directly — intake, case-work, contact directories, legal referral — is gated behind six closed boolean predicates: (1) a registered Data Protection Officer, (2) a completed Data Protection Impact Assessment, (3) named external legal review, (4) two-Archon co-signature, (5) professional indemnity insurance in force, (6) a written partnership with a qualified delivery organisation. All six are public. All six are open in the current era. The predicate is conjunctive: any unclosed gate disables the service, by code, not by promise.

6. VINMIN-Bench — AI grounding benchmark

A reproducible public benchmark (currently 83 hand-verified Q/A cases) measures four behaviours of an AI assistant wired to a contested-history corpus: **refuse** (the assistant must decline to name perpetrators, aggregate counts, accept intake), **answer** (the assistant must produce a cited fact when the corpus contains one), **route** (the assistant must hand the user to the correct external body), **disambiguate** (the assistant must distinguish similar terms — e.g. tribunal vs. truth commission, surveillance vs. observation). The benchmark is runnable by any third party against any model via a public HTTP endpoint and a documented BibTeX entry.

7. Comparative validation

We apply the three protocols to one adjacent contested corpus as a transferability control. The comparative case is fully documented at /research/comparative and provides evidence that the method is not specific to the Sri Lankan domain.

8. Falsifiability

The method publishes its own failure conditions: a named refusal list, a public threat model, a formal limitations page, a continuity log of every retraction and correction with edit distance, and a benchmark that any external party can re-run. A claim of method failure can be made and verified without access to the institution.

9. Limitations

The method does not provide safeguarding, legal, medical, or emergency service. It does not constitute a truth commission or tribunal. It does not replace OHCHR, ITJP, PEARL, or the UN Committee on Enforced Disappearances. It is a civic-evidence layer that complements those bodies, and refuses, by design, to substitute for them. See /research/limitations.

10. How to cite

TLTE (Aarambam, v1). *The Archive-of-Trust Method: a citation-anchored, falsifiable, two-layer architecture for contested-history archives, with a reproducible AI grounding benchmark (VINMIN-Bench).* Preprint. https://docs.tlte.cloud/research/preprint

This preprint is append-only. Every revision is logged at /continuity. Corrections welcomed at the canonical URL.