Situated Language Layers for Minority-Language and Indigenous Communities

Architectural pattern, operating principles, and CPU-fallback inference architecture (Paper B — synopsis)

John G. Stroh

2026-05-03

Situated Language Layers for Minority-Language and Indigenous Communities

Architectural pattern, operating principles, and CPU-fallback inference architecture — Paper B synopsis

---

Abstract

This is a two-page outline of the planned empirical companion to Paper A. It is not the empirical paper. It describes (i) the architectural pattern — a per-tenant *situated language layer* (SLL): a community-scoped small language model trained on the tenant's own content, governed by the tenant's own authority, and operated on infrastructure inside the tenant's own jurisdictional reach; (ii) the operating principles the project follows when training and running these models; (iii) the cohorts running in production today; and (iv) the inference architecture that keeps the runtime pathway entirely outside US-controlled infrastructure. Per-cohort evaluation, weight-modification ablations, and the comparator literature scan are reserved for the full paper, which is held back until verified training-run data is in hand. Paper A described a sovereign-record architecture under which content, governance, and federation envelopes remain inside the community that owns them; the SLL is

the runtime cognition layer of that platform — what lets a member ask the system a question and receive an answer drawn from the community's own writing rather than from a frontier model shaped by a global corpus the community had no part in. This synopsis exists so Paper A's forward references resolve to a real outline rather than a placeholder, and so the architectural pattern is on the public record while the empirical work continues.

Keywords: situated language layer, small language model, minority-language NLP, Indigenous data sovereignty, per-tenant model, CPU fallback inference, training discipline, sovereign infrastructure, Tier-1 cohort, non-US sovereign GPU.

---

1. Introduction

Paper A's sovereign-record architecture preserves community sovereignty over data; it does not by itself preserve community sovereignty over *cognition*. A community using a language model to mediate member queries is still using a language model: the model's training corpus, training discipline, and runtime behaviour are part of the architecture's surface, not separate from it. A frontier model trained on a global corpus shaped by no community in particular cannot answer a Welsh query in Welsh-community register, a Māori query under tikanga, or a Sámi query against the language module's revitalisation lexicon — not without overwriting the community's authority with the corpus the model was trained on.

Paper B describes the architectural pattern that makes the SLL trustworthy for community use: a per-tenant-type small language model trained on the tenant's own content, with strict operating discipline, runtime hosting on tenant-controlled or community-trusted infrastructure, and a CPU-fallback path that keeps inference inside the tenant's jurisdictional reach. The companion to Paper A is concrete: where Paper A makes the data substrate sovereign, Paper B describes the operating discipline that makes the cognition layer match. The empirical demonstration that the discipline yields the claimed properties is the work of the full paper, not this synopsis.

2. The operating discipline

Five rules govern the training of every situated language layer cohort. Each is operating discipline derived from the work of building these AIs, refined as more cohorts have come into deployment; each is documented in the project's standing rules.

No correction pairs. Synthetic correction pairs (paired prompt-and-good-answer adjusted from observed prompt-and-bad-answer) introduce a bias surface that the model overfits when the correction examples sit far from the natural distribution — they almost always do. The project uses steering vectors (sparse activations applied at inference time) for behavioural correction; the training corpus is left alone.

No deduplication of training data. Apparent duplicates in the corpus are reinforcement, not redundancy. The project's cohorts are trained on naturally-occurring content with native repetition (canonical lines from foundational texts; repeated tikanga formulae; recurrent governance phrasings); deduplicating these flattens the corpus's own emphasis structure.

No FAQ answers without codebase verification. Every FAQ-layer addition is anchored to a verifiable repository artefact (a constitution clause; a documented decision; a referenced source in the corpus). FAQ layering is the proven extension path; aspirational FAQ entries are forbidden.

No model-weight modifications. The weight-modification approaches the project has tested — a range of fine-tuning, distillation, and preference-tuning protocols — have, to date, underperformed an FAQ-layered base on the project's internal evaluation. The full paper will report the per-experiment comparison in detail. The operating consequence: until a weight-modification protocol is identified that *demonstrably* improves on the FAQ-layered base, the project does not adopt one. FAQ layering and governance packs are the proven extension paths.

No aspirational training (Tier-2 trigger discipline). A Tier-2 cohort is not commissioned until the first tenant of that type is in deployment. The motivation is two-fold: training without an actual deployment tenant grounds the corpus in speculation rather than the community's own content, and burning training cycles on aspirational cohorts redirects scarce GPU time from what is actually in production.

3. The weight-modification stance

The project's stance on weight modification is the most operationally consequential of the five rules and warrants separate exposition. The candidate modifications the project has tested span the standard surface — fine-tuning at varying scope and rate, distillation from larger models, preference-tuning protocols, and combinations of these — applied to the community-v1 14B Qwen2 base and scored against an

FAQ-layered base with steering vectors applied. The pattern across the approaches tested is that modified models drift in characteristic ways: making things up more on out-of-corpus questions, refusing less reliably when questions sit outside the corpus, and citing less precisely when grounded responses are requested.

Until a weight-modification protocol that demonstrably improves on the FAQ-layered base is identified, the project does not adopt one. The full paper will report the per-experiment results, evaluation-set composition, and methodology in detail; this synopsis records the operational consequence: weight modification is not a current extension path, and FAQ layering plus governance packs carry the project's actual production load.

4. Tier-1 cohorts in production

Five Tier-1 cohorts are deployed in production at the time of writing:

- `villageai-14b-whanau-v1` — Māori extended-family contexts; whānau and (interim) governance configurations.
- `villageai-14b-episcopal-v1` — Anglican-communion parish contexts.
- `villageai-14b-community-v1` — generic community fallback (the 14B Qwen2 base).
- `villageai-14b-family-v1` — family-history contexts.
- `villageai-14b-business-v1` — small-business member-directory and operational-dashboard contexts.

Each cohort is trained on its tenant-type's corpus with the operating discipline of §2. Per-cohort accuracy on tenant-content questions, refusal-rate on out-of-corpus questions, citation discipline, and qualitative tikanga-respecting register evaluation are reported in the full paper. Four Tier-2 cohorts (conservation, diaspora, clubs, alumni) are designated but operationally paused per the no-aspirational-training discipline.

An episcopal-v2 retrain was attempted and did not improve over v1 on the project's evaluation set; v2 is not deployed. The negative result is cited in the project's operating discipline; the full paper will treat it alongside the broader weight-modification work as further evidence that the current FAQ-layering-plus-steering-vectors path is at the local accuracy maximum for the cohort.

5. Inference architecture

Runtime inference is hosted on a New Zealand-sovereign A6000 GPU at Catalyst Cloud during business hours (08:00–20:00 NZST), with

automatic failover to a non-US home eGPU (RX 7900 XTX) outside business hours. A CPU-fallback path is available for low-load periods; latency is higher but throughput remains adequate for the platform's request profile. No request to a US-controlled inference endpoint is in the production request path. The vendor-prohibition rule that governs all platform infrastructure (no US-owned cloud, SaaS, or managed-AI service in the production request path) extends to the inference layer. Per-request routing decisions are logged with the cohort selected, the policy-gate verdict, and the upstream-service health at request time, providing the audit surface that GDPR Article 22 (automated decision-making) and Te Tiriti Article 2 (taonga protection over the community's data and its mediation) jointly require.

6. Evaluation overview

The full paper will report: per-cohort accuracy on standing evaluation sets; refusal discipline (rate of correct refusal-by-default for out-of-corpus questions); citation discipline (rate of zero-citation responses caught by the post-hoc citation safety filter); the weight-modification comparison in detail; episcopal-v2 retrain comparison; tikanga-register evaluation by community-aligned reviewers (where consent permits; otherwise summarised at architectural level); inference latency and throughput on each routing path; and a defence-in-depth analysis of model-grounding versus post-hoc citation filtering — the principle that model behaviour and safety filter operate at distinct layers and both must be reported, never the filter alone treated as the whole answer to grounding. This synopsis names these dimensions; the full paper reports them.

7. Limitations and relationship to Paper A

Three limitations bound this paper's scope. The Tier-2 cohorts are not yet evaluated; their empirical findings will be reported when their first tenants are in deployment. The evaluation protocol is the project's standing protocol — comparable to other community-scale evaluation efforts but not yet aligned with a published peer-reviewed evaluation benchmark; the full paper will discuss this comparator gap and the literature that informs it. And the empirical findings here are project-internal as of the synopsis date; the full paper will report them with the rigour expected for peer-review submission, including reviewer access to the standing evaluation set under appropriate confidentiality where the corpora include culturally-restricted material.

Paper A and Paper B form a deliberate two-paper split: Paper A cov-

ers the architectural substrate; Paper B covers the empirical operating discipline that makes the cognition layer of that substrate trustworthy for minority-language and Indigenous community use. Together they describe a system in which both data and cognition are sovereign by construction. The Tractatus framework paper (already public on Codeberg under Apache 2.0) is the third leg of the triad: development-time governance for the AI assistance that builds the platform itself.

8. Acknowledgements

The author is grateful to Leslie Stroh for foundational philosophical mentorship on pluralistic thinking and the question of goodness in artificial intelligence. The pluralistic-deliberation commitment that runs through the platform's governance architecture — and the wider conviction that an AI substrate worth building must answer to a substantive notion of goodness, not a procedural one — owes its formative shape to those conversations.

The author also acknowledges the community elders, language-revitalisation practitioners, and tenant administrators whose corpora and feedback have shaped the cohorts; specific named acknowledgement awaits direct consent from each individual and is held back here pending that consent.

References

[A] Stroh, J. G. (2026). *Sovereign-Record Architecture for Community-Scale Platforms — Paper A* (Review Draft v3, May 2026). My Digital Sovereignty Limited (NZ). Companion paper. Available at `agenticgovernance.digital/papers/sovereign-record-architecture-v3-may-2026.html` (English, te reo Māori, Deutsch).

[T] Stroh, J. G. (2026). *Tractatus Framework — Architectural Patterns for AI Development Governance, Working Paper v0.2.* `codeberg.org/mysovereignty/tractatus-framework`. Apache 2.0.

Detailed empirical references — including the Qwen2 base-model citation, evaluation-protocol citations, weight-modification-method literature (LoRA, full fine-tuning, layer-wise distillation, RLHF-style preference data), Indigenous-language NLP literature, low-resource translation literature, situated-dialogue literature, and ethics-of-language-model literature — are deferred to the full Paper B and to the Step-F literature scan planned for the project's wider documentation pass.

Corresponding author: John G. Stroh, Director, My Digital Sovereignty Limited (NZ). ORCID: 0009-0005-2933-7170. Email: john.stroh@mysovereignty.digital.

Suggested citation: Stroh, J. G. (2026). *Situated Language Layers for Minority-Language and Indigenous Communities — Paper B Synopsis.* My Digital Sovereignty Limited. Available at agenticgovernance.digital. (Zenodo DOI to be assigned upon expansion to full paper.)

Synopsis status: This is a 2-page synopsis. The full empirical paper is deferred to a separate session with verified training-run data, per-cohort evaluation tables, ablation-result detail, and the comparator literature scan required for peer-review submission. Published in synopsis form to resolve Paper A's forward references and to share the planned shape of the empirical companion paper.