Taallagen in hun context voor minderheidstaal- en inheemse gemeenschappen

John G. Stroh

Gesitueerde taallagen voor minderheidstalen en inheemse gemeenschappen

Architectuurpatroon, werkingsprincipes en CPU-fallback-inferentiearchitectuur (Paper B — samenvatting)

John G. Stroh

03-05-2026

Situatiegebonden taallagen voor minderheidstalen en inheemse gemeenschappen

Architectuurpatroon, werkingsprincipes en CPU-fallback inference-architectuur — Samenvatting van Paper B

Samenvatting

Dit is een overzicht van twee pagina's van de geplande empirische aanvulling op Paper A. Het is niet het empirische artikel. Het beschrijft (i) het architecturale patroon — een per tenant *gesitueerde taallaag* (SLL): een klein taalmodel op gemeenschapsniveau dat is getraind op de eigen inhoud van de tenant, wordt beheerd door de eigen autoriteit van de tenant en wordt uitgevoerd op infrastructuur binnen het eigen rechtsgebied van de tenant; (ii) de werkingsprincipes die het project volgt bij het trainen en uitvoeren van deze modellen; (iii) de cohorten die momenteel in productie draaien; en (iv) de inferentiearchitectuur die het runtime-traject volledig buiten door de VS gecontroleerde infrastructuur houdt. Evaluatie per cohort, ablatie van gewichtsaanpassingen en de scan van vergelijkende literatuur zijn voorbehouden aan het volledige artikel, dat wordt achtergehouden totdat geverifieerde trainingsgegevens beschikbaar zijn. Paper A beschreef een soevereine-recordarchitectuur waaronder inhoud, governance en federatie-envelopes binnen de gemeenschap blijven die er eigenaar van is; de SLL is de runtime-cognitielaag van dat platform — wat een lid in staat stelt het systeem een vraag te stellen en een antwoord te ontvangen dat is ontleend aan de eigen teksten van de gemeenschap in plaats van aan een grensmodel dat is gevormd door een wereldwijd corpus waar de gemeenschap geen aandeel in had. Deze samenvatting is bedoeld om ervoor te zorgen dat de verwijzingen in Paper A verwijzen naar een echt overzicht in plaats van een tijdelijke plaatshouder, en zodat het architecturale patroon openbaar beschikbaar is terwijl het empirische werk voortduurt.

Sleutelwoorden: situated language layer, small language model, NLP voor minderheidstalen,

inheemse gegevenssoevereiniteit, per-tenant model, CPU-fallback-inferentie, trainingsdiscipline, soevereine infrastructuur, Tier-1-cohort, niet-Amerikaanse soevereine GPU.

---

1. Inleiding

De soevereine-recordarchitectuur van Paper A behoudt de soevereiniteit van de gemeenschap over gegevens; het behoudt op zichzelf niet de soevereiniteit van de gemeenschap over *cognitie*. Een gemeenschap die een taalmodel gebruikt om vragen van leden te bemiddelen, maakt nog steeds gebruik van een taalmodel: het trainingscorpus, de trainingsdiscipline en het runtime-gedrag van het model maken deel uit van de architectuur, en staan er niet los van. Een grensverleggend model dat is getraind op een wereldwijd corpus dat niet door een specifieke gemeenschap is gevormd, kan geen antwoord geven op een Welshe vraag in het register van de Welshe gemeenschap, een Maori-vraag volgens tikanga, of een Sámi-vraag tegen het revitalisatie-lexicon van de taalmodule — niet zonder de autoriteit van de gemeenschap te overschrijven met het corpus waarop het model is getraind.

Paper B beschrijft het architecturale patroon dat de SLL betrouwbaar maakt voor gebruik door de gemeenschap: een klein taalmodel per tenant-type, getraind op de eigen inhoud van de tenant, met strikte operationele discipline, runtime-hosting op door de tenant gecontroleerde of door de gemeenschap vertrouwde infrastructuur, en een CPU-fallback-pad dat de inferentie binnen het jurisdictioneel bereik van de tenant houdt. De aanvulling op Paper A is concreet: waar Paper A het gegevenssubstraat soeverein maakt, beschrijft Paper B de operationele discipline die ervoor zorgt dat de cognitielagen hierop aansluiten. De empirische demonstratie dat de discipline de geclaimde eigenschappen oplevert, is het werk van het volledige artikel, niet van deze samenvatting.

2. De operationele discipline

Vijf regels bepalen de training van elke cohort van gesitueerde taallagen. Elk is een operationele discipline afgeleid van het werk van het bouwen van deze AI's, verfijnd naarmate meer cohorten in gebruik zijn genomen; elk is gedocumenteerd in de permanente regels van het project.

Geen correctieparen. Synthetische correctieparen (gekoppelde prompt-en-goed-antwoord aangepast van waargenomen prompt-en-slecht-antwoord) introduceren een biasoppervlak waaraan het model overfit wanneer de correctievoorbeelden ver van de natuurlijke verdeling liggen — wat bijna altijd het geval is. Het project gebruikt stuurvectoren (sparse activeringen toegepast op het moment van inferentie) voor gedragscorrectie; het trainingscorpus wordt ongemoeid gelaten.

Geen ontdubbeling van trainingsgegevens. Schijnbare duplicaten in het corpus zijn versterking, geen redundantie. De cohorten van het project worden getraind op natuurlijk voorkomende inhoud met inheemse herhaling (canonieke regels uit fundamentele teksten; herhaalde tikanga- formules; terugkerende bestuursformuleringen); het ontdubbelen hiervan vlak de eigen nadrukstructuur van het corpus af.

Geen FAQ-antwoorden zonder verificatie in de codebase. Elke toevoeging aan de FAQ-laag is verankerd aan een verifieerbaar artefact in de repository (een grondwetsclausule; een gedocumenteerd besluit; een bronvermelding in het corpus). FAQ-gelaagdheid is de beproefde uitbreidingsmethode; ambitieuze FAQ- vermeldingen zijn verboden.

Geen aanpassingen aan modelgewichten. De benaderingen voor gewichtsaanpassing die het project heeft getest — een reeks protocollen voor fijnafstemming, distillatie en voorkeursafstemming — hebben tot nu toe slechter gepresteerd dan een op FAQ-lagen gebaseerde basis bij de interne evaluatie van het project. Het volledige artikel zal de

vergelijking per experiment in detail rapporteren. Het operationele gevolg: totdat een protocol voor gewichtsaanpassing is geïdentificeerd dat *aantoonbaar* beter presteert dan de FAQ-gelaagde basis, past het project er geen toe. FAQ-gelaagdheid en governance-pakketten zijn de beproefde uitbreidingspaden.

Geen ambitieuze training (Tier-2-trigger discipline). Een Tier-2-cohort wordt pas in opdracht gegeven als de eerste tenant van dat type in gebruik is genomen. De motivatie is tweeledig: training zonder een daadwerkelijke tenant in gebruik zorgt ervoor dat het corpus op speculatie is gebaseerd in plaats van op de eigen content van de community, en het verspillen van trainingscycli aan ambitieuze cohorten leidt schaarse GPU-tijd af van wat daadwerkelijk in productie is.

3. Het standpunt over gewichtsaanpassing

Het standpunt van het project ten aanzien van gewichtsaanpassing is de meest operationeel relevante van de vijf regels en verdient een aparte toelichting. De kandidaat-aanpassingen die het project heeft getest, bestrijken het standaardgebied — fine-tuning met variërende reikwijdte en snelheid, distillatie uit grotere modellen, protocollen voor het afstemmen van voorkeuren, en combinaties hiervan — toegepast op de community-v1 14B Qwen2-basis en beoordeeld ten opzichte van een op FAQ's gebaseerde basis met toegepaste stuurvectoren. Het patroon dat zich voordoet bij de geteste benaderingen is dat gemodificeerde modellen op karakteristieke manieren afwijken: ze verzinnen meer bij vragen die buiten het corpus vallen, weigeren minder betrouwbaar wanneer vragen buiten het corpus vallen, en citeren minder nauwkeurig wanneer onderbouwde antwoorden worden gevraagd.

Totdat een protocol voor gewichtsaanpassing wordt geïdentificeerd dat aantoonbaar een verbetering is ten opzichte van de FAQ-gelaagde basis, neemt het project er geen in gebruik. Het volledige artikel zal de resultaten per experiment, de samenstelling van de evaluatieset en de methodologie in detail beschrijven; deze samenvatting legt de operationele consequentie vast: gewichtsaanpassing is momenteel geen uitbreidingspad , en FAQ-gelaagdheid plus governance-pakketten dragen de daadwerkelijke productielast van het project.

4. Tier-1-cohorten in productie

Op het moment van schrijven zijn vijf Tier-1-cohorten in productie ingezet:

- `villageai-14b-whanau-v1` — Māori-context van de uitgebreide familie; whānau en (tijdelijke) bestuursconfiguraties.
- `villageai-14b-episcopal-v1` — Anglicaanse parochiecontext .
- `villageai-14b-community-v1` — generieke gemeenschaps-fallback (de 14B Qwen2-basis).
- `villageai-14b-family-v1` — context van familiegeschiedenis.
- `villageai-14b-business-v1` — context van ledenlijsten en operationele dashboards voor kleine bedrijven.

Elke cohort wordt getraind op het corpus van zijn tenant-type met de werkwijze van §2. De nauwkeurigheid per cohort op vragen over tenant-inhoud, het weigeringspercentage op vragen buiten het corpus, de citatiewerkwijze en de kwalitatieve, tikanga-respecterende registerevaluatie worden gerapporteerd in het volledige artikel. Vier Tier-2-cohorten (conservatie, diaspora, clubs, alumni) zijn aangewezen, maar operationeel gepauzeerd volgens de discipline van geen ambitieuze training.

Er is een poging gedaan tot hertraining van episcopal-v2, maar deze leverde geen verbetering op ten opzichte van v1 op de evaluatieset van het project; v2 is niet geïmplementeerd. Het negatieve resultaat wordt aangehaald in de operationele discipline van het project; het volledige artikel zal dit samen met het bredere werk op het gebied van gewichtsaanpassing

behandelen als verder bewijs dat het huidige pad van FAQ-gelaagdheid plus stuurvectoren zich op het lokale nauwkeurigheidsmaximum voor het cohort bevindt.

5. Inferentiearchitectuur

Runtime-inferentie wordt gehost op een in Nieuw-Zeeland gevestigde A6000 GPU bij Catalyst Cloud tijdens kantooruren (08:00–20:00 NZST), met automatische failover naar een niet-Amerikaanse thuis-eGPU (RX 7900 XTX) buiten kantooruren. Een CPU-fallback-pad is beschikbaar voor periodes met lage belasting; de latentie is hoger maar de doorvoer blijft toereikend voor het verzoekprofiel van het platform. Er is geen verzoek aan een door de VS gecontroleerd inferentie-eindpunt in het productie- verzoekpad. De leveranciersverbodsregel die geldt voor alle platform- infrastructuur (geen cloud, SaaS of managed-AI-service in Amerikaanse handen in het productieverzoekpad) strekt zich uit tot de inferentielayer. Routeringsbeslissingen per verzoek worden gelogd met het geselecteerde cohort, de policy-gate-uitspraak en de status van de upstream-service op het moment van het verzoek, waardoor het auditoppervlak wordt geboden dat artikel 22 van de AVG (geautomatiseerde besluitvorming) en artikel 2 van het Verdrag van Wellington (bescherming van taonga met betrekking tot de gegevens van de gemeenschap en de bemiddeling daarover) gezamenlijk vereisen.

6. Overzicht van de evaluatie

Het volledige artikel zal verslag doen van: de nauwkeurigheid per cohort op bestaande evaluatiesets; de weigeringsdiscipline (percentage correcte standaardweigeringen voor vragen die buiten het corpus vallen); citatie-discipline (percentage reacties zonder citatie dat door het post-hoc citatieveiligheidsfilter wordt opgevangen); de vergelijking van gewichtsaanpassingen in detail; vergelijking van episcopal-v2-hertraining; evaluatie van het tikanga-register door community-gerichte beoordelaars (waar toestemming dit toestaat; anders samengevat op architectuurniveau); inferentielatentie en doorvoer op elk routeringspad; en een diepgaande analyse van model-grounding versus post-hoc citatie filtering — het principe dat modelgedrag en veiligheidsfilter op verschillende lagen werken en beide moeten worden gerapporteerd, nooit het filter alleen behandeld als het volledige antwoord op grounding. Deze samenvatting noemt deze dimensies; het volledige artikel rapporteert ze.

7. Beperkingen en relatie tot Paper A

Drie beperkingen begrenzen de reikwijdte van dit artikel. De Tier-2-cohorten zijn nog niet geëvalueerd; hun empirische bevindingen zullen worden gerapporteerd wanneer hun eerste gebruikers in gebruik zijn. Het evaluatieprotocol is het bestaande protocol van het project — vergelijkbaar met andere evaluatie-inspanningen op gemeenschapsschaal , maar nog niet afgestemd op een gepubliceerde, peer-reviewed evaluatiebenchmark; het volledige artikel zal deze vergelijkingskloof en de literatuur die hieraan ten grondslag ligt bespreken. En de empirische bevindingen hier zijn op de datum van de samenvatting projectintern; het volledige artikel zal hierover rapporteren met de nauwkeurigheid die verwacht wordt voor indiening bij peer-review, inclusief toegang voor recensenten tot de vaste evaluatieset onder passende vertrouwelijkheid wanneer de corpora cultureel beperkt materiaal bevatten.

Paper A en Paper B vormen een bewuste splitsing in twee papers: Paper A behandelt de architecturale basis; Paper B behandelt de empirische operationele discipline die de cognitielagen van die basis betrouwbaar maakt voor gebruik door minderheidstalen en inheemse gemeenschappen. Samen beschrijven ze een systeem waarin zowel data als cognitie soeverein zijn door hun opbouw. Het Tractatus-frameworkdocument (reeds openbaar

op Codeberg onder Apache 2.0) vormt de derde poot van de drie-eenheid: governance tijdens de ontwikkeling voor de AI-assistentie die het platform zelf bouwt.

8. Dankwoord

De auteur is Leslie Stroh dankbaar voor haar fundamentele filosofische begeleiding op het gebied van pluralistisch denken en de vraag naar het goede in kunstmatige intelligentie. De toewijding aan pluralistische beraadslaging die als een rode draad door de bestuursarchitectuur van het platform loopt — en de bredere overtuiging dat een AI-substraat dat de moeite waard is om te bouwen, moet beantwoorden aan een inhoudelijk begrip van het goede, niet aan een procedureel begrip — dankt zijn vorm aan die gesprekken.

De auteur bedankt ook de ouderen van de gemeenschap, beoefenaars van taalrevitalisering en huurdersbestuurders wier corpora en feedback de cohorten hebben gevormd; specifieke vermelding van namen wacht op directe toestemming van elk individu en wordt hier in afwachting van die toestemming achtergehouden.

Referenties

[A] Stroh, J. G. (2026). *Sovereign-Record Architecture for Community-Scale Platforms — Paper A* (Review Draft v3, mei 2026). My Digital Sovereignty Limited (NZ). Begeleidend document. Beschikbaar op `agenticgovernance.digital/papers/sovereign-record-architecture-v3-may-2026.html` (Engels, te reo Māori, Deutsch).

[T] Stroh, J. G. (2026). *Tractatus Framework — Architecturale patronen voor het beheer van AI-ontwikkeling, werkdocument v0.2.* `codeberg.org/mysovereignty/tractatus-framework`. Apache 2.0.

Gedetailleerde empirische referenties — inclusief de Qwen2-basismodel , citaten van evaluatieprotocollen, literatuur over methoden voor gewichtsaanpassing (LoRA, volledige fine-tuning, laaggewijze distillatie, RLHF-achtige voorkeursgegevens), NLP-literatuur over inheemse talen, literatuur over vertalen met beperkte middelen , literatuur over gesitueerde dialoog en literatuur over de ethiek van taalmodellen — worden uitgesteld tot het volledige Paper B en naar de literatuurstudie van stap F die gepland is voor de bredere documentatieronde van het project.

Corresponderende auteur: John G. Stroh, directeur, My Digital Sovereignty Limited (NZ). ORCID: 0009-0005-2933-7170. E-mail: john.stroh@mysovereignty.digital.

Voorgestelde bronvermelding: Stroh, J. G. (2026). *Situated Language Layers for Minority-Language and Indigenous Communities — Paper B Synopsis.* My Digital Sovereignty Limited. Beschikbaar op agenticgovernance.digital. (Zenodo DOI wordt toegewezen bij uitbreiding naar volledig artikel.)

Status van de samenvatting: Dit is een samenvatting van 2 pagina's. Het volledige empirische artikel wordt uitgesteld tot een aparte sessie met geverifieerde trainingsgegevens, evaluatietabellen per cohort, details van ablatie-resultaten, en de literatuurstudie ter vergelijking die vereist is voor indiening voor peer-review. Gepubliceerd in de vorm van een samenvatting om de voorwaartse verwijzingen van Paper A op te lossen en om de geplande opzet van het bijbehorende empirische artikel te delen.