Kontextspezifische Sprachebenen für Minderheitensprachen und indigene Gemeinschaften

John G. Stroh

Situierte Sprachebenen für Minderheitensprachen und indigene Gemeinschaften

Architekturmuster, Funktionsprinzipien und CPU-Fallback-Inferenzarchitektur (Artikel B – Zusammenfassung)

John G. Stroh

03.05.2026

Situative Sprachschichten für Minderheitensprachen und indigene Gemeinschaften

Architekturmuster, Funktionsprinzipien und CPU-Fallback- Inferenzarchitektur – Zusammenfassung von Paper B

Zusammenfassung

Dies ist eine zweiseitige Übersicht über die geplante empirische Ergänzung zu Beitrag A. Es handelt sich nicht um den empirischen Beitrag. Er beschreibt (i) das Architekturmuster – eine mandantenbezogene *situierte Sprachschicht* (SLL): ein auf die Gemeinschaft zugeschnittenes kleines Sprachmodell, das auf den eigenen Inhalten des Mandanten trainiert wurde, der Autorität des Mandanten unterliegt und auf Infrastruktur innerhalb des Zuständigkeitsbereichs des Mandanten betrieben wird; (ii) die Betriebsprinzipien, denen das Projekt beim Trainieren und Ausführen dieser Modelle folgt; (iii) die Kohorten, die derzeit in der Produktion laufen; und (iv) die Inferenzarchitektur, die den Laufzeitpfad vollständig außerhalb der von den

USA kontrollierten Infrastruktur hält. Die Bewertung pro Kohorte, Ablationen zur Gewichtsmodifikation und die Literaturrecherche zu Vergleichsstudien sind dem vollständigen Artikel vorbehalten, der zurückgehalten wird, bis verifizierte Trainingsdaten vorliegen. Artikel A beschrieb eine Architektur mit souveränen Datensätzen, bei der Inhalte, Governance und Föderationshüllen innerhalb der Gemeinschaft verbleiben, der sie gehören; die SLL ist die Laufzeit-Kognitionsschicht dieser Plattform – sie ermöglicht es einem Mitglied, dem System eine Frage zu stellen und eine Antwort zu erhalten, die aus den eigenen Texten der Gemeinschaft stammt, anstatt aus einem Frontier-Modell, das durch einen globalen Korpus geprägt ist, an dem die Gemeinschaft keinen Anteil hatte. Diese Zusammenfassung dient dazu, dass die Vorwärtsverweise in Papier A auf einen konkreten Entwurf verweisen und nicht auf einen Platzhalter, und damit das Architekturmuster öffentlich zugänglich ist, während die empirische Arbeit fortgesetzt wird.

Schlüsselwörter: Situierte Sprachschicht, kleines Sprachmodell, NLP für Minderheitensprachen, indigene Datenhoheit, mandantenbasiertes Modell, CPU-Fallback-Inferenz, Trainingsdisziplin, souveräne Infrastruktur, Tier-1-Kohorte, nicht-US-amerikanische souveräne GPU.

---

1. Einleitung

Die souveräne Aufzeichnungsarchitektur von Paper A bewahrt die Souveränität der Gemeinschaft über die Daten; sie bewahrt jedoch nicht an sich die Souveränität der Gemeinschaft über *die Kognition*. Eine Gemeinschaft, die ein Sprachmodell zur Vermittlung von Mitgliederanfragen nutzt, verwendet immer noch ein Sprachmodell: Der Trainingskorpus, die Trainingsdisziplin und das Laufzeitverhalten des Modells sind Teil der Architekturoberfläche und nicht von ihr getrennt. Ein Pioniermodell, das auf einem globalen Korpus trainiert wurde, der von keiner bestimmten Gemeinschaft geprägt ist, kann keine walisische Anfrage im Register der walisischen Gemeinschaft, keine Māori-Anfrage unter tikanga oder keine Sámi-Anfrage anhand des Revitalisierungslexikons des Sprachmoduls beantworten – nicht ohne die Autorität der Gemeinschaft durch das Korpus zu überschreiben, auf dem das Modell trainiert wurde.

Paper B beschreibt das Architekturmuster, das das SLL für den gemeinschaftlichen Gebrauch vertrauenswürdig macht: ein kleines Sprachmodell pro Mandantentyp, das auf den eigenen Inhalten des Mandanten trainiert wurde, mit strenger Betriebsdisziplin, Laufzeit-Hosting auf mandantengesteuerter oder von der Gemein-

schaft als vertrauenswürdig eingestufter Infrastruktur und einem CPU-Fallback-Pfad, der die Inferenz innerhalb der Zuständigkeit des Mandanten hält. Die Ergänzung zu Paper A ist konkret: Während Paper A das Datensubstrat souverän macht, beschreibt Paper B die Betriebsdisziplin, die die Kognitionsschicht dazu bringt, sich anzupassen. Der empirische Nachweis, dass die Disziplin die behaupteten Eigenschaften hervorbringt, ist Gegenstand des vollständigen Artikels, nicht dieser Zusammenfassung.

2. Die Betriebsdisziplin

Fünf Regeln bestimmen das Training jeder Kohorte der situativen Sprachschicht. Jede ist eine Betriebsdisziplin, die aus der Arbeit am Aufbau dieser KIs abgeleitet und im Zuge der Einführung weiterer Kohorten verfeinert wurde; jede ist in den geltenden Regeln des Projekts dokumentiert.

Keine Korrekturpaare. Synthetische Korrekturpaare (gepaarte Aufforderung und richtige Antwort, angepasst aus beobachteter Aufforderung und falscher Antwort) führen zu einer Verzerrung, an die sich das Modell überanpasst, wenn die Korrekturbeispiele weit von der natürlichen Verteilung entfernt liegen – was fast immer der Fall ist. Das Projekt verwendet Steuerungsvektoren (spärliche Aktivierungen, die zum Zeitpunkt der Inferenz angewendet werden) zur Verhaltenskorrektur; der Trainingskorpus bleibt unverändert.

Keine Deduplizierung von Trainingsdaten. Scheinbare Duplikate im Korpus sind Verstärkung, keine Redundanz. Die Kohorten des Projekts werden auf natürlich vorkommenden Inhalten mit nativen Wiederholungen trainiert (kanonische Zeilen aus grundlegenden Texten; wiederholte Tikanga- Formeln; wiederkehrende Formulierungen aus der Regierungsführung); eine Deduplizierung dieser glättet die eigene Betonungsstruktur des Korpus.

Keine FAQ-Antworten ohne Überprüfung der Codebasis. Jede Ergänzung der FAQ-Ebene ist an ein überprüfbares Repository-Artefakt gebunden (eine Verfassungsklausel; eine dokumentierte Entscheidung; eine referenzierte Quelle im Korpus). Die Schichtung der FAQs ist der bewährte Erweiterungsweg; ambitionierte FAQ-Einträge sind verboten.

Keine Modifikationen der Modellgewichtung. Die vom Projekt getesteten Ansätze zur Gewichtsmodifikation – eine Reihe von Protokollen zur Feinabstimmung, Destillation und Präferenzanpassung – haben bislang in der internen Bewertung des Projekts schlechter abgeschnitten als eine auf FAQ-Schichten basierende Grundlage. Der vollständige Artikel wird den Vergleich pro Experiment im

Detail darlegen. Die operative Konsequenz: Solange kein Protokoll zur Gewichtsanpassung identifiziert wird, das die FAQ-geschichtete Basis *nachweislich* verbessert, wird das Projekt kein solches einführen. FAQ-Schichtung und Governance-Pakete sind die bewährten Erweiterungswege.

Kein aspiratives Training (Tier-2-Trigger- Disziplin). Eine Tier-2-Kohorte wird erst dann in Auftrag gegeben, wenn der erste Mandant dieses Typs im Einsatz ist. Dafür gibt es zwei Gründe: Training ohne einen tatsächlich im Einsatz befindlichen Mandanten stützt den Korpus auf Spekulationen statt auf die eigenen Inhalte der Community, und das Verschwenden von Trainingszyklen auf aspirative Kohorten lenkt knappe GPU-Zeit von dem ab, was tatsächlich in Produktion ist.

3. Die Haltung zur Gewichtsanpassung Die

Die Haltung des Projekts zur Gewichtsmodifikation ist die operativ bedeutendste der fünf Regeln und verdient eine gesonderte Erläuterung. Die vom Projekt getesteten Modifikationsansätze umfassen die Standardmaßnahmen — Feinabstimmung in unterschiedlichem Umfang und Tempo, Destillation aus größeren Modellen, Protokolle zur Präferenzanpassung sowie Kombinationen davon —, die auf die Community-v1-14B-Qwen2-Basis angewendet und anhand einer FAQ-gestützten Basis mit angewendeten Steuerungsvektoren bewertet wurden. Das Muster bei den getesteten Ansätzen ist, dass modifizierte Modelle auf charakteristische Weise abweichen: Sie erfinden mehr bei Fragen außerhalb des Korpus, lehnen weniger zuverlässig ab, wenn Fragen außerhalb des Korpus liegen, und zitieren weniger präzise, wenn fundierte Antworten verlangt werden.

Solange kein Protokoll zur Gewichtsanpassung identifiziert wird, das nachweislich eine Verbesserung gegenüber der FAQ-geschichteten Basis darstellt, wird das Projekt kein solches übernehmen. Der vollständige Artikel wird die Ergebnisse pro Experiment, die Zusammensetzung des Bewertungssatzes und die Methodik im Detail darlegen; diese Zusammenfassung hält die operative Konsequenz fest: Die Gewichtsanpassung ist derzeit kein Erweiterungsweg, und die FAQ-Schichtung sowie Governance-Pakete tragen die tatsächliche Produktionslast des Projekts.

4. Tier-1-Kohorten in der Produktion

Zum Zeitpunkt der Erstellung dieses Dokuments sind fünf Tier-1-Kohorten in der Produktion im Einsatz:

- `villageai-14b-whanau-v1` – Māori-Großfamilienkontexte; Whānau und (vorläufige) Governance-Konfigurationen.
- `villageai-14b-episcopal-v1` – Kontexte anglikanischer Gemeinden.
- `villageai-14b-community-v1` – generischer Community-Fallback (die 14B-Qwen2-Basis).
- `villageai-14b-family-v1` — Kontexte der Familiengeschichte.
- `villageai-14b-business-v1` — Kontexte für Mitgliederverzeichnisse und Betriebs-Dashboards kleiner Unternehmen.

Jede Kohorte wird auf dem Korpus ihres Tenant-Typs unter Einhaltung der in §2 beschriebenen Betriebsdisziplin trainiert. Die Genauigkeit pro Kohorte bei Fragen zu Tenant-Inhalten, die Ablehnungsrate bei Fragen außerhalb des Korpus, die Zitierdisziplin sowie die qualitative, tikanga-konforme Registerbewertung werden im vollständigen Artikel berichtet. Vier Tier-2-Kohorten (Naturschutz, Diaspora, Vereine, Alumni) sind zwar vorgesehen, wurden jedoch gemäß der Richtlinie zum Verzicht auf ambitioniertes Training operativ pausiert.

Ein retraining von Episcopal-v2 wurde versucht und führte im Bewertungssatz des Projekts zu keiner Verbesserung gegenüber v1; v2 wird nicht eingesetzt. Das negative Ergebnis wird in der Betriebsdisziplin des Projekts angeführt; das vollständige Papier wird es zusammen mit der umfassenderen Arbeit zur Gewichtsmodifikation als weiteren Beweis dafür behandeln, dass der derzeitige Pfad aus FAQ-Schichtung plus Steuerungsvektoren das lokale Genauigkeitsmaximum für die Kohorte darstellt.

5. Inferenzarchitektur

Die Laufzeit-Inferenz wird während der Geschäftszeiten (08:00–20:00 NZST) auf einer neuseeländischen A6000-GPU bei Catalyst Cloud gehostet, mit automatischem Failover auf eine nicht in den USA befindliche Heim-eGPU (RX 7900 XTX) außerhalb der Geschäftszeiten. Ein CPU-Fallback-Pfad steht für Zeiten mit geringer Auslastung zur Verfügung; die Latenz ist höher, der Durchsatz bleibt jedoch für das Anforderungsprofil der Plattform ausreichend. Es gibt keine Anforderung an einen von den USA kontrollierten Inferenz-Endpunkt im Produktions- Anforderungspfad. Die Anbieter-Verbotsregel, die die gesamte Plattform- Infrastruktur regelt (keine US-eigenen Cloud-, SaaS- oder Managed-AI-Dienste im Produktions-Anforderungspfad), erstreckt sich auch auf die Inferenzschicht. Routing-Entscheidungen pro Anfrage werden mit der ausgewählten Kohorte, dem Policy-Gate- Urteil und dem Zustand des Upstream-Dienstes zum Zeitpunkt der Anfrage protokolliert

und bieten damit die Prüfungsgrundlage, die Artikel 22 der DSGVO (automatisierte Entscheidungsfindung) und Artikel 2 des Tiriti (Schutz der Taonga über die Daten der Gemeinschaft und deren Vermittlung) gemeinsam vorschreiben.

6. Überblick über die Bewertung

Der vollständige Artikel berichtet über: die Genauigkeit pro Kohorte bei bestehenden Bewertungsdatensätzen; die Ablehnungsdisziplin (Rate korrekter Ablehnungen bei Fragen außerhalb des Korpus); Zitationsdisziplin (Anteil der Antworten ohne Zitate, die vom nachträglichen Zitationssicherheitsfilter erfasst wurden); den detaillierten Vergleich der Gewichtsmodifikation; den Vergleich des Episcopal-v2-Retrainings; die Tikanga-Register-Bewertung durch gemeinschaftsorientierte Gutachter (sofern die Einwilligung dies zulässt; ansonsten auf Architekturebene zusammengefasst); Inferenzlatenz und Durchsatz auf jedem Routing-Pfad; sowie eine tiefgreifende Analyse von Modell-Grounding im Vergleich zu Post-hoc-Zitationsfilterung – das Prinzip, dass Modellverhalten und Sicherheitsfilter auf unterschiedlichen Ebenen operieren und beide berichtet werden müssen, niemals der Filter allein als die gesamte Antwort auf das Grounding behandelt werden darf. Diese Zusammenfassung nennt diese Dimensionen; das vollständige Papier berichtet darüber.

7. Einschränkungen und Bezug zu Artikel A

Drei Einschränkungen begrenzen den Umfang dieses Artikels. Die Tier-2-Kohorten wurden noch nicht evaluiert; ihre empirischen Ergebnisse werden berichtet, sobald ihre ersten Nutzer im Einsatz sind. Das Evaluierungsprotokoll ist das laufende Protokoll des Projekts – vergleichbar mit anderen Evaluierungsbemühungen auf Community-Ebene, jedoch noch nicht an einen veröffentlichten, peer-reviewten Evaluierungs-Benchmark angepasst; der vollständige Artikel wird diese Vergleichslücke und die zugrunde liegende Literatur erörtern. Und die hier dargestellten empirischen Ergebnisse sind zum Zeitpunkt der Zusammenfassung projektintern; der vollständige Artikel wird sie mit der für eine Peer-Review-Einreichung erwarteten Genauigkeit berichten, einschließlich des Zugangs der Gutachter zum bestehenden Evaluierungsdatensatz unter angemessener Vertraulichkeit, sofern die Korpora kulturell geschützte Materialien enthalten.

Artikel A und Artikel B bilden eine bewusste Aufteilung in zwei Artikel: Artikel A behandelt das architektonische Substrat; Artikel B behandelt die empirische Betriebsdisziplin, die die Kognitionsschicht

dieses Substrats für die Nutzung durch Minderheitensprachen und indigene Gemeinschaften vertrauenswürdig macht. Zusammen beschreiben sie ein System, in dem sowohl Daten als auch Kognition von Grund auf souverän sind. Das Tractatus-Framework-Paper (bereits auf Codeberg unter Apache 2.0 veröffentlicht) ist das dritte Standbein der Triade: Governance während der Entwicklungsphase für die KI-Unterstützung, die die Plattform selbst aufbaut.

8. Danksagungen

Der Autor dankt Leslie Stroh für die grundlegende philosophische Begleitung zum pluralistischen Denken und zur Frage nach dem Guten in der künstlichen Intelligenz. Das Bekenntnis zur pluralistischen Deliberation, das sich durch die Governance-Architektur der Plattform zieht – und die weitergehende Überzeugung, dass ein KI-Substrat, dessen Aufbau sich lohnt, einem substanziellen, nicht einem prozeduralen Begriff des Guten entsprechen muss – verdankt seine gestaltende Form diesen Gesprächen.

Der Autor dankt außerdem den Ältesten der Gemeinschaft, den Praktikern der Sprachrevitalisierung und den Verwaltern der Mieteinheiten, deren Korpora und Rückmeldungen die Kohorten geprägt haben; eine namentliche Danksagung steht noch unter dem Vorbehalt der direkten Zustimmung jedes Einzelnen und wird hier bis zu dieser Zustimmung zurückgehalten.

Referenzen

[A] Stroh, J. G. (2026). *Sovereign-Record Architecture for Community-Scale Platforms – Paper A* (Review Draft v3, Mai 2026). My Digital Sovereignty Limited (NZ). Begleitpapier. Verfügbar unter `agenticgovernance.digital/papers/sovereign-record-architecture-v3-may-2026.html` (Englisch, Te Reo Māori, Deutsch).

[T] Stroh, J. G. (2026). *Tractatus Framework – Architektonische Muster für die Steuerung der KI-Entwicklung, Arbeitspapier v0.2.* `codeberg.org/mysovereignty/tractatus-framework`. Apache 2.0.

Detaillierte empirische Referenzen – einschließlich des Qwen2-Basismodells , Zitate aus Bewertungsprotokollen, Literatur zu Methoden der Gewichtsmodifikation (LoRA, vollständiges Fine-Tuning, schichtweise Destillation, RLHF-artige Präferenzdaten), NLP-Literatur zu indigenen Sprachen, Literatur zur Übersetzung mit geringen Ressourcen, Literatur zu situierten Dialogen und Literatur zur Ethik von Sprachmodellen – werden auf das vollständige

Papier B sowie auf die für die umfassendere Dokumentationsphase des Projekts geplante Literaturrecherche im Rahmen von Schritt F.

Korrespondenzautor: John G. Stroh, Direktor, My Digital Sovereignty Limited (NZ). ORCID: 0009-0005-2933-7170. E-Mail: john.stroh@mysovereignty.digital.

Vorgeschlagene Zitierweise: Stroh, J. G. (2026). *Situated Language Layers for Minority-Language and Indigenous Communities — Paper B Synopsis.* My Digital Sovereignty Limited. Verfügbar unter agenticgovernance.digital. (Zenodo-DOI wird bei Erweiterung zum vollständigen Artikel vergeben.)

Status der Zusammenfassung: Dies ist eine 2-seitige Zusammenfassung. Das vollständige empirische Papier wird auf eine separate Sitzung verschoben, mit verifizierten Trainingsdaten, Bewertungstabellen pro Kohorte, Details zu den Ablationsergebnissen und der für die Einreichung zur Begutachtung erforderlichen Literaturrecherche. Veröffentlicht in Form einer Zusammenfassung, um die Vorwärtsverweise von Paper A aufzulösen und die geplante Struktur des begleitenden empirischen Papiers zu teilen.