Niveaux linguistiques contextualisés pour les communautés de langue minoritaire et autochtones

John G. Stroh

Couches linguistiques situées pour les communautés de langues minoritaires et autochtones

Modèle architectural, principes de fonctionnement et architecture d'inférence de secours sur CPU (Article B — résumé)

John G. Stroh

03/05/2026

Couches linguistiques situées pour les communautés linguistiques minoritaires et autochtones

Modèle architectural, principes de fonctionnement et architecture d'inférence de secours sur CPU — Résumé de l'article B

Résumé

Il s'agit d'un résumé de deux pages du document empirique prévu pour accompagner l' article A. Ce n'est pas l'article empirique lui-même. Il décrit (i) le modèle architectural — une *couche linguistique située*(SLL) par locataire : un petit modèle linguistique à l'échelle de la communauté, entraîné sur le contenu propre au locataire, régi par l'autorité propre au locataire et exploité sur une infrastructure située dans la sphère de compétence du locataire ; (ii) les principes de fonctionnement que le projet suit lors de l'entraînement et de

l'exécution de ces modèles ; (iii) les cohortes actuellement en production ; et (iv) l' architecture d'inférence qui maintient le chemin d'exécution entièrement en dehors de l' infrastructure contrôlée par les États-Unis. L'évaluation par cohorte, les ablations de modification de poids et l'analyse de la littérature comparative sont réservées à l'article complet , qui est mis en attente jusqu'à ce que des données d'entraînement vérifiées soient disponibles. L'article A décrit une architecture de registres souverains dans laquelle le contenu, la gouvernance et les enveloppes de fédération restent au sein de la communauté qui en est propriétaire ; le SLL est la couche de cognition d'exécution de cette plateforme — ce qui permet à un membre de poser une question au système et de recevoir une réponse tirée des écrits de la communauté elle-même plutôt que d'un modèle de pointe façonné par un corpus mondial auquel la communauté n'a pas participé. Ce résumé existe afin que les références futures de l’article A renvoient à un véritable plan plutôt qu’à un espace réservé, et afin que le modèle architectural soit accessible au public pendant que les travaux empiriques se poursuivent.

Mots-clés : couche linguistique située, petit modèle linguistique, TNL des langues minoritaires, souveraineté des données autochtones, modèle par locataire, inférence de secours sur CPU, discipline d’entraînement, infrastructure souveraine, cohorte de niveau 1, GPU souverain hors États-Unis.

---

1. Introduction

L’architecture de registre souverain du document A préserve la souveraineté de la communauté sur les données ; elle ne préserve pas en soi la souveraineté de la communauté sur *la cognition*. Une communauté utilisant un modèle linguistique pour traiter les requêtes de ses membres utilise toujours un modèle linguistique : le corpus d’entraînement, la discipline d’entraînement et le comportement d’exécution du modèle font partie de la surface de l’architecture, ils ne sont pas distincts de celle-ci. Un modèle de pointe entraîné sur un corpus mondial non façonné par une communauté en particulier ne peut répondre à une requête galloise dans le registre de la communauté galloise, à une requête maorie selon le tikanga, ou à une requête sami par rapport au lexique de revitalisation du module linguistique — pas sans remplacer l’autorité de la communauté par le corpus sur lequel le modèle a été entraîné.

L'article B décrit le modèle architectural qui rend le SLL fiable pour une utilisation communautaire : un petit modèle linguistique par type de locataire entraîné sur le contenu propre au locataire, avec

une discipline opérationnelle stricte, un hébergement d'exécution sur une infrastructure contrôlée par le locataire ou approuvée par la communauté, et un chemin de secours CPU qui maintient l'inférence dans la sphère de compétence du locataire. Le complément de l'article A est concret : là où l'article A rend le substrat de données souverain, l'article B décrit la discipline opérationnelle qui permet à la couche cognitive de s'y conformer. La démonstration empirique que cette discipline produit les propriétés revendiquées constitue l' objet de l'article complet, et non de ce résumé.

2. La discipline opérationnelle

Cinq règles régissent l'entraînement de chaque cohorte de couches linguistiques situées. Chacune est une discipline opérationnelle dérivée du travail de construction de ces IA, affinée à mesure que davantage de cohortes ont été déployées ; chacune est documentée dans les règles permanentes du projet.

Pas de paires de correction. Les paires de correction synthétiques (paires de prompt et de bonne réponse ajustées à partir de prompts et de mauvaises réponses observés) introduisent un biais auquel le modèle s'adapte de manière excessive lorsque les exemples de correction s'écartent de la distribution naturelle — ce qui est presque toujours le cas. Le projet utilise des vecteurs de pilotage (activations clairsemées appliquées au moment de l'inférence) pour la correction comportementale ; le corpus d'entraînement n'est pas modifié.

Pas de déduplication des données d'entraînement. Les duplicatas apparents dans le corpus constituent un renforcement, et non une redondance. Les cohortes du projet sont entraînées sur du contenu naturel comportant des répétitions natives (lignes canoniques issues de textes fondateurs ; formules tikanga répétées ; formulations récurrentes en matière de gouvernance) ; la déduplication de ces éléments aplatit la structure d'accentuation propre au corpus.

Aucune réponse de FAQ sans vérification du code source. Chaque ajout à la couche FAQ est ancré à un artefact vérifiable du référentiel (une clause constitutionnelle ; une décision documentée ; une source référencée dans le corpus). La stratification des FAQ est la voie d'extension éprouvée ; les entrées de FAQ spéculatives sont interdites.

Pas de modifications du poids des modèles. Les approches de modification du poids testées par le projet — une gamme de protocoles de réglage fin, de distillation et d'ajustement des préférences — ont, à ce jour, donné des résultats inférieurs à ceux d'une base à couches de FAQ lors de l'évaluation interne du projet. L'article

complet rendra compte en détail de la comparaison par expérience . Conséquence opérationnelle : tant qu'un protocole de modification des poids n'aura pas été identifié comme améliorant *de manière démontrable* la base à couches de FAQ, le projet n'en adoptera pas. La stratification des FAQ et les packs de gouvernance constituent les voies d'extension éprouvées.

Pas de formation sur des cas hypothétiques (règle de déclenchement de niveau 2). Une cohorte de niveau 2 n'est pas mise en service tant que le premier locataire de ce type n'est pas en déploiement. La motivation est double : une formation sans locataire en déploiement réel fonde le corpus sur de la spéculation plutôt que sur le contenu propre à la communauté, et consacrer des cycles de formation à des cohortes hypothétiques détourne le temps GPU limité de ce qui est réellement en production.

3. La position sur la modification des poids

La position du projet sur la modification des poids est la plus importante sur le plan opérationnel parmi les cinq règles et mérite une présentation distincte. Les modifications candidates que le projet a testées couvrent l'ensemble des techniques standard — réglage fin à différentes échelles et cadences, distillation à partir de modèles plus grands, protocoles de réglage des préférences, et combinaisons de ceux-ci — appliquées à la base community-v1 14B Qwen2 et évaluées par rapport à une base structurée en FAQ avec application de vecteurs de pilotage. La tendance observée dans les approches testées est que les modèles modifiés dérivent de manière caractéristique : ils inventent davantage de réponses aux questions hors corpus, refusent de manière moins fiable lorsque les questions se situent en dehors du corpus, et citent de manière moins précise lorsque des réponses fondées sont demandées.

Tant qu'un protocole de modification des poids qui améliore de manière démontrable la base à couches FAQ n'aura pas été identifié, le projet n'en adoptera pas. L' article complet rendra compte en détail des résultats par expérience, de la composition de l'ensemble d'évaluation et de la méthodologie ; ce résumé consigne la conséquence opérationnelle : la modification des poids n'est pas une voie d'extension actuelle, et la superposition de couches FAQ ainsi que les packs de gouvernance supportent la charge de production réelle du projet.

4. Cohortes de niveau 1 en production

Cinq cohortes de niveau 1 sont déployées en production au moment de la rédaction :

- `villageai-14b-whanau-v1` — Contextes de la famille élargie maorie ; whānau et configurations de gouvernance (provisoires).
- `villageai-14b-episcopal-v1` — Contexte paroissial de la Communion anglicane .
- `villageai-14b-community-v1` — solution de secours communautaire générique (la base 14B Qwen2).
- `villageai-14b-family-v1` — contextes d'histoire familiale.
- `villageai-14b-business-v1` — contextes de répertoire des membres et de tableau de bord opérationnel pour les petites entreprises.

Chaque cohorte est entraînée sur le corpus de son type de locataire selon la discipline opérationnelle de la section 2. La précision par cohorte sur les questions relatives au contenu des locataires, le taux de refus sur les questions hors corpus, la discipline de citation et l' évaluation qualitative du registre respectant le tikanga sont présentés dans l' article complet. Quatre cohortes de niveau 2 (conservation, diaspora, clubs, anciens élèves) ont été désignées mais sont suspendues sur le plan opérationnel conformément à la discipline de non-formation ambitieuse.

Une nouvelle formation d'episcopal-v2 a été tentée et n'a pas apporté d'amélioration par rapport à la v1 sur l'ensemble d'évaluation du projet ; la v2 n'est pas déployée. Ce résultat négatif est mentionné dans la discipline opérationnelle du projet ; l'article complet l'abordera parallèlement aux travaux plus généraux de modification des poids comme une preuve supplémentaire que la voie actuelle « FAQ-layering-plus-steering-vectors » se situe au maximum de précision locale pour la cohorte.

5. Architecture d'inférence

L'inférence en temps réel est hébergée sur un GPU A6000 sous souveraineté néo-zélandaise chez Catalyst Cloud pendant les heures de bureau (08h00–20h00 NZST), avec basculement automatique vers un eGPU domestique non américain (RX 7900 XTX) en dehors des heures de bureau. Une Chemin de secours sur CPU est disponible pour les périodes de faible charge ; la latence est plus élevée mais le débit reste suffisant pour le profil de requêtes de la plateforme. Aucune requête vers un point de terminaison d'inférence contrôlé par les États-Unis ne figure dans le chemin de requête de production. La

règle d’interdiction des fournisseurs qui régit toute l’infrastructure de la plateforme (pas de cloud, de SaaS ou de service d’IA géré détenu par les États-Unis dans le chemin de requête de production) s’étend à la couche d’inférence. Les décisions de routage par requête sont consignées avec la cohorte sélectionnée, le verdict de la barrière de politique et l'état de santé du service en amont au moment de la requête, fournissant ainsi la surface d'audit requise conjointement par l'article 22 du RGPD (prise de décision automatisée) et l'article 2 du Traité de Tiriti (protection des taonga sur les données de la communauté et leur médiation).

6. Aperçu de l'évaluation

L'article complet rendra compte : de la précision par cohorte sur des ensembles d'évaluation permanents ; de la discipline de refus (taux de refus par défaut correct pour les questions hors corpus) ; la discipline de citation (taux de réponses sans citation détectées par le filtre de sécurité de citation a posteriori) ; la comparaison détaillée des modifications de pondération ; la comparaison de réentraînement d’episcopal-v2 ; l’évaluation du registre tikanga par des évaluateurs alignés sur la communauté (lorsque le consentement le permet ; sinon résumée au niveau architectural) ; latence d'inférence et débit sur chaque chemin de routage ; et une analyse en profondeur de l'ancrage du modèle par rapport au filtrage a posteriori des citations — le principe selon lequel le comportement du modèle et le filtre de sécurité opèrent à des niveaux distincts et doivent tous deux être rapportés, le filtre ne devant jamais être considéré à lui seul comme la réponse complète à l'ancrage. Ce résumé énumère ces dimensions ; l'article complet les détaille.

7. Limites et lien avec l'article A

Trois limites restreignent la portée de cet article. Les cohortes de niveau 2 n’ont pas encore été évaluées ; leurs résultats empiriques seront rapportés lorsque leurs premiers locataires seront déployés. Le protocole d’évaluation est le protocole permanent du projet — comparable à d’autres efforts d’évaluation à l’échelle de la communauté mais pas encore aligné sur un benchmark d’évaluation publié et évalué par des pairs ; l’article complet abordera ce manque de comparabilité et la littérature qui l’étaye. De plus, les résultats empiriques présentés ici sont internes au projet à la date du résumé ; l’article complet les présentera avec la rigueur attendue pour une soumission à un comité de lecture, y compris l’accès des évaluateurs à l’ensemble d’évaluation permanent dans le respect d’une confidentialité appropriée lorsque les corpus contiennent des éléments

soumis à des restrictions culturelles.

L'article A et l'article B forment un duo délibéré : l'article A couvre le substrat architectural ; l'article B couvre la discipline opérationnelle empirique qui rend la couche cognitive de ce substrat fiable pour une utilisation par les communautés de langues minoritaires et autochtones. Ensemble, ils décrivent un système dans lequel tant les données que la cognition sont souveraines de par leur conception. L'article sur le cadre Tractatus (déjà accessible sur Codeberg sous licence Apache 2.0) constitue le troisième volet de la triade : la gouvernance en phase de développement de l'assistance IA qui construit la plateforme elle-même.

8. Remerciements

L'auteur remercie Leslie Stroh pour son mentorat philosophique fondamental sur la pensée pluraliste et la question du bien dans l'intelligence artificielle. L'engagement en faveur de la délibération pluraliste qui imprègne l'architecture de gouvernance de la plateforme — ainsi que la conviction plus large qu'un substrat d'IA digne d'être construit doit répondre à une notion substantielle du bien, et non procédurale — doit sa forme initiale à ces conversations.

L'auteur remercie également les anciens de la communauté, les praticiens de la revitalisation linguistique et les administrateurs de communautés dont les corpus et les retours ont façonné les cohortes ; les remerciements nominatifs spécifiques attendent le consentement direct de chaque personne et sont donc retenus ici en attendant ce consentement.

Références

[A] Stroh, J. G. (2026). *Architecture de registre souverain pour les plateformes à l'échelle communautaire — Document A* (Projet de révision v3, mai 2026). My Digital Sovereignty Limited (NZ). Document d'accompagnement. Disponible à l'adresse `agenticgovernance.digital/papers/sovereign-record-architecture-v3-may-2026.html` (anglais, te reo Māori, allemand).

[T] Stroh, J. G. (2026). *Tractatus Framework — Modèles architecturaux pour la gouvernance du développement de l'IA, document de travail v0.2.* `codeberg.org/mysovereignty/tractatus-framework`. Apache 2.0.

Références empiriques détaillées — y compris la citation du modèle de base Qwen2 , les citations relatives au protocole d'évaluation, la

littérature sur les méthodes de modification des poids (LoRA, réglage fin complet, distillation par couche, données de préférence de type RLHF), la littérature sur le TALN en langues autochtones, la littérature sur la traduction en conditions de ressources limitées , la littérature sur le dialogue situé et la littérature sur l'éthique des modèles linguistiques — sont reportées à l'article B complet et à l'analyse documentaire de l'étape F prévue dans le cadre de la phase de documentation plus large du projet.

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

Citation suggérée : Stroh, J. G. (2026). *Couches linguistiques situées pour les communautés de langues minoritaires et autochtones — Synopsis de l'article B.* My Digital Sovereignty Limited. Disponible sur agenticgovernance.digital. (DOI Zenodo à attribuer lors de l' extension à l'article complet.)

Statut du résumé : Il s'agit d'un résumé de 2 pages. L'article empirique complet est reporté à une session distincte comprenant des données de test vérifiées, des tableaux d'évaluation par cohorte, des détails sur les résultats d'ablation et l'analyse documentaire comparative requise pour la soumission à l'évaluation par les pairs. Publié sous forme de résumé afin de résoudre les références anticipées de l'article A et de partager la structure prévue de l'article empirique d'accompagnement.