Kaitiakitanga in Machine Learning:

A Budget-Aware Approach to Fairness in Bilingual RAG

GitHub: https://github.com/alexkzw/kaitiaki-planner

Executive Summary

This report examines fairness in AI systems serving both English and Te Reo Māori. Through a budget-aware Retrieval-Augmented Generation (RAG) system, this research explores how resource allocation strategies can reduce performance disparities between high-resource (English) and low-resource (Māori) languages. The project reveals that improving retrieval quality through semantic embeddings yields greater fairness gains than budget allocation mechanisms alone, but uncovers a critical complexity-dependent trade-off. This work contributes to conversations about Māori data sovereignty, ethical AI design, and language justice.

Motivation: Language Justice and Kaitiakitanga

Natural language processing systems exhibit systematic performance disparities, with low-resource languages like Te Reo Māori consistently underperforming compared to English. Contemporary AI systems, trained predominantly on English data, risk perpetuating "algorithmic colonialism" - the extension of historical power imbalances into computational infrastructure (Couldry, N., & Mejias, 2019).

The concept of kaitiakitanga - guardianship and stewardship - provides a culturally grounded framework for this technical challenge. In tikanga Māori, kaitiakitanga encompasses responsibility to protect taonga (treasures), which explicitly includes te reo Māori itself (Waitangi Tribunal, 2011). This project adopts kaitiakitanga as both metaphor and method: just as traditional kaitiaki actively manage resources to ensure equitable access, a budget-aware RAG system allocates computational resources to protect Māori language interactions with AI. This embeds cultural values at the architectural level rather than treating fairness as post-hoc optimisation.

Research Question

This study investigates the following:

1. Whether dynamic budget allocation reduces performance gaps between English and Māori queries.
2. Which retrieval strategies most effectively support fairness.
3. How fairness interventions align with Indigenous Data Sovereignty principles.

Methods: Three Fairness Philosophies

The kaitiaki-planner implements a RAG pipeline with query processing, budget planning, semantic retrieval (paraphrase-multilingual-mpnet-base-v2 with keyword boosting), and answer generation via Claude (Anthropic's LLM). Three budget allocation strategies operationalise distinct fairness theories:

Under Condition 1 (Uniform/Equality), the system allocates top\_k = 5 for all queries. This reflects a philosophy of formal equality by providing identical resources regardless of need, and it represents how most current RAG systems are designed.

Under Condition 2 (Language-Aware/Equity), the system allocates top\_k = 8 for Māori queries and top\_k = 5 for English queries. This follows an equity philosophy that responds to resource scarcity by offering additional support to structurally disadvantaged groups, acknowledging that Māori queries face retrieval challenges due to smaller training data and fewer native embeddings.

Under Condition 3 (Fairness-Aware/Intersectional Equity), the system allocates top\_k = 8 for Māori queries or for complex queries. This recognises intersectional fairness across multiple axes of disadvantage - such as language and task difficulty - and operationalises manaakitanga (care and support) by tailoring assistance to structural barriers.

The system was evaluated on 30 queries (15 English, 15 Māori) covering New Zealand cultural and natural heritage topics, with queries stratified by complexity (20 simple, 10 complex). The primary evaluation metric is Grounded Correctness (GC), a binary measure requiring both correct answers and proper citation of gold-standard passages. This metric prioritises retrieval quality over generation fluency, reflecting the principle that knowledge access precedes text production.

Results

**Baseline BM25 Performance**

The BM25 baseline achieved 100% GC for English (15/15) but only 60% for Māori (9/15), establishing a 40-percentage point fairness gap. Surprisingly, BM25 succeeded more on complex Māori queries (75%, 3/4) than simple queries (54.5%, 6/11), revealing that the retrieval gap is driven by semantic localisation of cultural concepts rather than query complexity. Simple queries about culturally-specific entities (kākā, marae) lack keyword bridges to English equivalents, while complex queries about international topics (Waitangi Treaty, United States) benefit from consistent terminology across languages. This demonstrates that keyword-based retrieval systematically underserves queries about culturally distinct knowledge.

**Semantic Embeddings: Aggregate Performance**

Semantic embeddings with keyword boosting improved overall system performance from 80% (BM25) to 83.33% grounded correctness (+4.17pp), with particular benefits for Māori queries (60% to 66.67%, +11.11% relative improvement, +1 query from 9/15 to 10/15) and fairness gap reduction (40pp to 33.3pp, 16.67% gap reduction). English accuracy remained at 100% (15/15) across all conditions, demonstrating a ceiling effect. Independent t-tests confirm a statistically significant and practically large EN-MI performance gap (t=2.646, p=0.013, Cohen's d=0.966), establishing the fairness disparity as genuine rather than random variation. This finding validates that the 33.3pp gap is both statistically and practically significant.

**The Null Result: Budget Allocation Has No Effect**

All three budget allocation strategies yielded identical outcomes, with a mean GC of 83.33% overall (25/30 queries), 100% English (15/15), and 66.67% Māori (10/15). Statistical testing confirmed no difference among conditions (ANOVA: F = 0.000, p = 1.0000). The interpretation for this is that once the semantic embeddings correctly ranked the gold-standard passage within the top five results, increasing the budget to top\_k = 8 provided no additional benefit. The correct passage was already visible, and retrieving more documents introduced noise rather than signal. This indicates that fairness-by-design (selecting strong multilingual embeddings) is more effective than fairness-by-allocation (dynamically adjusting top\_k). Budget reallocation functions as crisis management, whereas architectural quality functions as preventive care.

**Critical Finding: Complexity-Dependent Trade-Off**

The above aggregate improvement masks a critical complexity-dependent trade-off revealed through stratified analysis. Simple Māori queries improved substantially: 54.5% (BM25) to 81.82% (embeddings) while complex Māori queries regressed severely: 75% (BM25) to (25%) embeddings, creating a two-tier fairness outcome. This stratification pattern reveals that current retrieval approaches may inadvertently create equitable performance on simple queries

but substantially inequitable performance on complex reasoning tasks.

Ethical, cultural, and societal implications

**Māori data sovereignty**

This project operates within the framework articulated by Te Mana Raraunga (the Māori Data Sovereignty Network). Three principles inform the work:

1. Rangatiratanga (authority/self-determination): Māori should govern how te reo Māori is represented in AI systems; meaningful rangatiratanga requires Māori communities to act as co-designers who determine system priorities and acceptable trade-offs rather than serving merely as evaluation subjects.
2. Whakapapa (relationships/contextualization): Data must be understood in relational context; the corpus pairs English and Māori passages about shared referents (e.g., Tongariro), acknowledging that they reflect different epistemologies - mātauranga Māori versus Pākehā knowledge systems - and warning that treating them as interchangeable risks flattening important distinctions about knowledge authority.
3. Kaitiakitanga (guardianship): Beyond technical adequacy, the system must address who benefits and who is harmed; although performance for Māori improved to 80%, it still falls behind English, raising the question of whether such underperformance is acceptable and under what theory of justice systematic shortfalls can be tolerated.

**Ethical philosophy: distributive justice theories**

The three experimental conditions operationalise distinct theories.

1. Utilitarian equality (Uniform): The aim is to maximise average performance by treating all queries identically, which is efficient but ignores structural inequalities and perpetuates existing advantages.
2. Rawlsian equity (Language-Aware): The design prioritises improvements for the least advantaged group (Rawls, 2017); allocating additional budget to Māori queries applies the “maximin” principle by optimising outcomes for the worst-off position.
3. Capabilities approach (Fairness-Aware): Following Sen, A. (2014)., the focus shifts from equal resources to equal capability to achieve valuable outcomes; the system identifies multiple barriers (e.g., language and task complexity) and allocates resources to equalise effective capability rather than formal inputs.

The null result complicates this landscape: when all approaches yield identical outcomes, the practical choice among them becomes moot. This shows that fairness interventions are context-dependent: sophisticated allocation mechanisms matter when resources are scarce and quality uneven, but they become irrelevant when baseline quality already meets user needs. In such settings, the ethical imperative shifts from “how do we allocate resources fairly?” to “how do we ensure sufficient quality so that allocation becomes unnecessary?”

Machine consciousness and epistemic violence

Debates about machine consciousness often ask whether AI systems possess subjective experience, but this project highlights the ethics of representation in language technologies. Even without consciousness, a RAG system acts as an epistemological intermediary that shapes access to knowledge. When it fails to retrieve correct Māori passages, it performs a form of epistemic violence by reinforcing the false belief that information in te reo Māori is less accessible. This has implications for younger generations: if digital systems consistently deliver better results in English, users face pragmatic pressure to abandon te reo, accelerating language loss. Linguistic agency - the ability to interact with technology in one’s language without penalty - therefore becomes crucial.

Cross-cultural implications

Although grounded in Aotearoa, the findings have implications for the world’s ~7,000 languages, most of which are “low-resource” in NLP terms (Joshi et al., 2020). Three lessons emerge. First, architecture matters more than tweaking: replacing BM25 with semantic embeddings produced a 71% improvement for Māori retrieval, far surpassing any gains from allocation strategies, so communities should prioritise foundational models that respect linguistic diversity.

Second, measurement shapes outcomes: the GC metric intentionally evaluates citation accuracy rather than fluency, reflecting the principle that retrieval quality precedes generation; in other contexts, communities may prioritise cultural appropriateness or preservation of traditional metaphors, so fairness metrics must be co-designed to reflect local priorities.

Third, beware the “good enough” trap: an 80% success rate for Māori may seem acceptable in isolation, but it remains inequitable when English achieves 100%; acceptability must be judged against the best available performance, not against past baselines.

Limits of technical solutions

This project illustrates the limits of technical interventions that are divorced from structural change. Even a perfectly fair RAG system operates within broader inequalities: most training data, research publications, and hiring pipelines are English-dominant. Improving one system’s Māori performance cannot substitute for systemic efforts in data justice (e.g., investing in Māori-language corpus development, oral history digitisation, and indigenous-language web content), educational justice (e.g., supporting Māori-language immersion and bilingual STEM resources), and economic justice (e.g., supporting Māori-led tech companies and bilingual AI procurement). The principle of whanaungatanga (kinship) reminds us that technology exists within social webs: a RAG system cannot save a language; only communities can, through intergenerational transmission, cultural revitalisation, and political power - while technology can either support or hinder these efforts.

Conclusion: toward language justice in AI

This capstone shows that meaningful fairness gains in bilingual RAG systems require architectural investment in retrieval quality rather than clever resource allocation alone. By improving Māori performance from 46.7% to 80% through semantic embeddings and a 30% keyword boost, the system substantially reduced - but did not eliminate - the gap with English. From a kaitiakitanga perspective, builders of multilingual systems should prioritise foundational model quality that respects linguistic diversity, while treating budget allocation strategies as contingency measures when quality falls short. The persistent 20-point gap, although only marginally significant statistically (p = 0.0719), remains practically important. Future work should explore targeted interventions such as query expansion, cross-lingual retrieval, and community-led evaluations that assess cultural appropriateness alongside factual correctness. Ultimately, the mauri (life force) of te reo Māori depends on speakers and communities, yet in a digitally mediated world, computational kaitiaki have a responsibility to ensure that technology supports - rather than suffocates - indigenous language vitality.

References

1. Couldry, N., & Mejias, U. A. (2019). The costs of connection: How data is colonizing human life and appropriating it for capitalism. Stanford University Press.
2. Te Mana Raraunga. (2025). Principles of Māori data sovereignty. Retrieved from https://www.temanararaunga.maori.nz/
3. Waitangi Tribunal. (2011). Ko Aotearoa Tēnei: A Report into Claims Concerning New Zealand Law and Policy Affecting Māori Culture and Identity (Wai 262). Wellington: Legislation Direct.
4. Rawls, J. (2017). A theory of justice. *In Applied Ethics* (pp. 21-29). Routledge
5. Sen, A. (2014). Development as freedom (1999). The globalization and development reader: Perspectives on development and global change, 525.
6. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. arXiv preprint arXiv:2004.09095.