# Invisible Languages of the LLM Universe

**Saurabh Khanna** [1,2*], **Xinxu Li** [1]

[1] Amsterdam School of Communication Research, University of Amsterdam
[2] Pembroke College, University of Oxford

**Large Language Models are trained on massive multilingual corpora, yet this abundance masks a profound crisis: of the world's 7,613 living languages, approximately 2,000 languages with millions of speakers remain effectively invisible in digital ecosystems. We propose a critical framework connecting empirical measurements of language vitality (real-world demographic strength) and digitality (online presence) with postcolonial theory and epistemic injustice to explain why linguistic inequality in AI systems is not incidental but structural. Analyzing data across all documented human languages, we identify four categories: *Strongholds* (33%, high vitality and digitality), *Digital Echoes* (6%, high digitality despite declining vitality), *Fading Voices* (36%, low on both dimensions), and critically, *Invisible Giants* (27%, high vitality but near-zero digitality) – languages spoken by millions yet absent from LLM training data. We demonstrate that these patterns reflect continuities from colonial-era linguistic hierarchies to contemporary AI development, constituting what we term *digital-epistemic injustice*. Our analysis reveals that English dominance in AI is not a technical necessity but an artifact of power structures that systematically exclude marginalized linguistic knowledge. We conclude with implications for decolonizing language technology and democratizing access to AI benefits.**

## 1. Introduction

An inquiry into content availability in Javanese, a language with 68 million speakers, yields negligible results on platforms integral to Large Language Model training. In contrast, Icelandic, spoken by 350,000 individuals, possesses a substantial digital footprint. This paradox highlights a critical issue in the manner artificial intelligence encapsulates linguistic diversity among humans. Notably, the preceding five years have observed a significant expansion of Large Language Models (LLMs), with systems such as GPT-4, Claude, and Gemini demonstrating human-like proficiency across numerous tasks (Anthropic, 2024; OpenAI, 2023). These models are trained utilizing extensive multilingual corpora sourced from the internet, including Common Crawl's 159 billion web pages, Wikipedia's 64 million articles, and vast digital archives (Gao et al., 2021). Nonetheless, this apparent wealth of data conceals a profound bias: a majority of the training data is derived from fewer than 20 high-resource languages, while the majority of the world's 7,613 extant languages remain largely neglected, effectively rendering them ”invisible” (Joshi et al., 2020).

Prior work has documented this crisis through multiple lenses. Kornai (2013) established that only 5% of languages will achieve meaningful digital vitality, coining the term “digital language death” to describe extinction in online spaces independent of offline endangerment. Bender et al. (2021) demonstrated that LLMs encode hegemonic worldviews, with harms falling disproportionately on marginalized communities. Simons et al. (2022) developed automated methods to measure digital language support globally, finding only 33 languages (0.4%) achieve full digital capabilities across 143 platforms.

[*] Correspondence E-mail: s.khanna@uva.nl

However, existing research treats linguistic exclusion primarily as a technical resource scarcity problem – languages are "low-resource" because they lack sufficient training data. This framing obscures deeper questions: *Why* do some languages with millions of speakers have virtually no digital presence? What historical and structural forces produce this inequality? How does the vitality-digitality gap reflect and reproduce colonial power relations?

We address these questions by integrating empirical measurement with critical theory. Building on the Invisible Lab's comprehensive dataset measuring vitality and digitality across all 7,613 documented languages, we apply postcolonial linguistics (Said, 1978), epistemic injustice theory (Fricker, 2007; Helm et al., 2024), and critical perspectives on linguistic imperialism (Phillipson, 1992) to explain *why* certain languages remain invisible despite demographic strength.

Our analysis makes three primary contributions:

**First**, we demonstrate that the vitality-digitality gap is not random but systematically disadvantages languages of formerly colonized regions. Invisible Giants – languages with high vitality but near-zero digitality – are concentrated in Africa, South Asia, Southeast Asia, and indigenous Americas, mirroring colonial subjugation patterns. This constitutes what we term *digital-epistemic injustice*: the systematic exclusion of marginalized linguistic communities from AI-mediated knowledge production.

**Second**, we trace continuities from colonial linguistic practices to contemporary LLM development. Missionary linguistics constructed languages as discrete, extractable objects; ISO coding systems treated them as fixed entities; contemporary NLP inherits these epistemologies. The "low-resource" framing itself reflects colonialist logics – framing African and Asian languages as deficient rather than interrogating why digital infrastructure was never built for them.

**Third**, we provide actionable implications for decolonizing language technology. Rather than simply "adding more data" for underrepresented languages, we argue for fundamentally reimagining LLM development: community-controlled datasets, alternative evaluation metrics centered on non-English linguistic features, and redistribution of AI economic benefits to language communities providing training data.

The urgency is clear: as LLMs become infrastructure for education, commerce, and governance, digital language vitality increasingly determines whether languages survive at all. If current trajectories continue, AI systems will accelerate rather than reverse language endangerment, creating feedback loops where digital invisibility causes offline decline.

## 2. Theoretical Framework

### 2.1. Digital Language Death and the Vitality-Digitality Distinction

Kornai (2013) established a crucial insight: digital vitality operates independently from traditional language vitality. A language can remain spoken intergenerationally (high vitality by EGIDS standards) yet be functionally extinct online (zero digitality). Using machine learning on 8,426 languages, Kornai created a four-tier classification finding that languages relegated to "Heritage" status – digitally archived but not used for living communication – cannot digitally ascend regardless of documentation efforts.

This distinction reveals a fundamental gap in endangerment frameworks. The Expanded Graded Intergenerational Disruption Scale (EGIDS) (Lewis et al., 2010) assesses offline vitality through intergenerational transmission and institutional use, classifying languages from 0 (International) to 10 (Extinct). UNESCO's nine-factor framework similarly focuses on speaker attitudes, policies, and documentation (on Endangered Languages, 2003). Yet these frameworks emerged before the internet became central to economic and social participation.

Recent work has attempted to bridge this gap. The Digital Language Diversity Project developed a six-level Digital Language Vitality Scale assessing capacity, presence, and performance across digital domains (Soria and Ballatore, 2017). Simons et al. (2022) introduced automated measurement across seven support categories (content, encoding, translation, analysis) and 143 platforms. However, no framework successfully integrates traditional and digital vitality into a unified model validated across diverse linguistic contexts.

We adopt a two-dimensional framework treating vitality and digitality as independent axes. **Vitality** measures socio-demographic presence: first-language speaker volume (from Ethnologue) and EGIDS status, combined through unsupervised clustering. **Digitality** measures online footprint: prevalence across Common Crawl (159 billion pages), Wikipedia (64 million articles), Hugging Face datasets (114,000) and models (447,000), and open language archives (474,000 entries). The **representation score** equals digitality minus vitality, quantifying the gap between real-world strength and digital presence.

## 2.2. Epistemic Injustice in Language Technology

Helm et al. (2024) introduced "language modeling bias" as by-design preferences for certain languages embedded in AI systems, arguing this constitutes *epistemic injustice* – systems are precise for dominant powers but limited in expressing socio-culturally relevant concepts of marginalized communities. This framework, drawing on Fricker (2007), distinguishes *testimonial injustice* (dismissing contributions based on identity prejudice) from *hermeneutical injustice* (gaps in collective knowledge resources enabling interpretation of social experiences).

LLMs contribute to both forms at unprecedented scale. Testimonial injustice occurs when models systematically downrank, mistranslate, or fail to recognize contributions in non-dominant languages. When a Swahili speaker's input receives degraded performance compared to English, the system effectively judges their testimony as less credible. Hermeneutical injustice occurs when languages lack vocabulary in model training data to express community-specific concepts – there are no isiZulu words for "dinosaur" or "evolution" in most LLM training sets, making scientific discourse in isiZulu impossible (Nekoto et al., 2020).

Recent work extends this framework. Talat et al. (2024) argue that labeling certain data as "knowledge" in knowledge-enhanced LLMs obscures harm to marginalized groups, as knowledge enhancement perpetuates epistemic injustice without diversification. Immigration systems relying on limited machine translation endanger lives when asylum seekers cannot articulate claims in languages systems recognize (Garibay and González, 2021). The opacity and scale of AI systems facilitate epistemic injustice at levels previously impossible – billions of interactions daily embed the message that certain languages and ways of knowing don't matter.

We extend epistemic injustice theory by connecting it to *digital language vitality*. The vitality-digitality gap directly produces hermeneutical injustice: communities with rich oral traditions and millions of speakers cannot access AI benefits because their linguistic resources were never digitized. This is not accidental scarcity but structural exclusion reflecting historical marginalization.

## 2.3. Postcolonial Linguistics and Digital Imperialism

Said (1978) defined orientalism as ways Western cultures dominate and reconstruct Eastern cultures, depicting them as undeveloped and inferior. This framework illuminates how linguistic hierarchies established during colonialism persist in AI development. English dominance is not a neutral technical outcome but reflects British imperial expansion and subsequent American technological hegemony.

Contemporary language technology exhibits what Mohamed et al. (2020) term “AI colonialism” – extractive relationships where Global North institutions mine data from Global South communities, build models on that data, then sell tools back to those communities at profit. This mirrors classic colonial resource extraction. South African facial recognition creating digital apartheid (Benjamin, 2023), Venezuelan data labeling under exploitative conditions (Gray and Suri, 2019), and Māori communities fighting for data sovereignty (Kukutai and Taylor, 2016) exemplify continuities from territorial colonialism to data colonialism.

The “low-resource language” framing itself embodies colonialist logics. As Dotan et al. (2023) argue, African languages remain “low resource” not due to inherent deficiency but because Western institutions never invested in digital infrastructure for them. The term positions these languages as lacking, requiring benevolent intervention from well-resourced institutions, rather than recognizing that resource scarcity is a political product of marginalization.

Daems (2024) trace material continuities: missionary linguistics’ construction of languages as discrete, extractable objects prefigures modern NLP’s treatment of languages as datasets to be harvested. ISO 639 language codes, developed by Western institutions, treat languages as fixed entities rather than fluid practices, enabling computational processing but erasing sociolinguistic realities. Contemporary commercial technology providers still draw on colonial-era translations and grammatical descriptions produced by missionaries.

Self-orientalization compounds these dynamics. Liu and Li (2017) document how some Chinese professionals in Australia actively cater to stereotypes, labeling themselves with Confucian attributes to gain workplace advantage. Chen (2021) shows tourism discourse in China Daily depicts some regions as marginalized due to lack of English service, with “Chinglish” reinforcing stereotypes of Chinese English inadequacy – a form of internalized linguistic imperialism that LLMs trained on such texts will reproduce.

## 2.4. Language Ideology and Technological Design

Language ideologies – beliefs about language that rationalize social structures – are embedded in technological architecture (Woolard and Schieffelin, 1998). Hohn et al. (2024) demonstrate that conversational AI embeds Western linguistic ideologies privileging certain ways of speaking: standardized grammar over vernacular, written over oral, explicitness over context-dependence, monolingualism over multilingualism. These ideological commitments shape design choices about what counts as “good” language data, which features to optimize, and how to evaluate performance.

The ideology of “Standard Language” positions prestigious varieties as correct while stigmatizing others as deficient (Lippi-Green, 2012). LLMs overwhelmingly train on formal written language, embedding prescriptive norms. When models correct African American Vernacular English to Standard American English, they encode racist language ideologies positioning AAVE as “broken” English (Blodgett et al., 2020). Similar dynamics operate multilingually – “correct” Hindi means formal written Hindi, erasing Hinglish and regional varieties actually used by millions.

Platform architectures encode linguistic assumptions. Twitter’s character limits advantage logographic writing systems over alphabetic ones. TikTok’s audio-visual primacy advantages languages with existing popular media. LinkedIn’s professional register expectations advantage languages with established business communication norms. Yet these platform-specific language ideologies remain understudied – we lack systematic analysis of how affordances advantage or disadvantage different languages and practices.

# 3. Methods

## 3.1. Data Sources and Coverage

We analyze comprehensive data on all 7,613 languages documented in the 25th edition of Ethnologue (Lewis et al., 2023). Our analysis integrates two primary dimensions:

**Vitality measurement** combines:

- First-language speaker counts from Ethnologue, ranging from single-digit (nearly extinct languages) to hundreds of millions (Hindi, Arabic, Spanish)
- EGIDS (Expanded Graded Intergenerational Disruption Scale) status ratings from 0 (International) to 10 (Extinct), assessing intergenerational transmission and institutional development

**Digitality measurement** aggregates presence across:

- Common Crawl: 159 billion web pages (November 2023 release)
- Wikipedia: 64 million articles across 300+ language editions
- Hugging Face: 114,000 datasets and 447,000 models tagged by language
- Open Language Archives: 474,000 entries from ELAR, AILLA, DoBeS, and other repositories

Following Simons et al. (2022), we employ automated language identification using fastText and CLD3 classifiers with validation sampling. For each language, we compute normalized scores on each dimension, then apply unsupervised clustering (Gaussian Mixture Models with 3 components) to extract composite vitality and digitality dimensions.

The **representation score** is computed as:

$$\text{Representation} = \text{Digitality}_{\text{normalized}} - \text{Vitality}_{\text{normalized}} \quad (1)$$

Negative scores indicate languages whose vitality exceeds digitality (underrepresented online), positive scores indicate the opposite (overrepresented relative to speaker base).

## 3.2. Categorical Classification

We classify languages into four categories based on their position in vitality-digitality space:

- **Strongholds** (+ Vitality / + Digitality): Both scores above median
- **Digital Echoes** (– Vitality / + Digitality): Below-median vitality, above-median digitality
- **Fading Voices** (– Vitality / – Digitality): Both scores below median
- **Invisible Giants** (+ Vitality / – Digitality): Above-median vitality, below-median digitality

## 3.3. Geographic and Colonial Analysis

To test whether digitality gaps reflect colonial histories, we coded languages by:

- Geographic region (Africa, Asia, Europe, Americas, Pacific)
- Colonial history (colonizer nation and duration, coded from historical databases)

- Official language status (national, regional, none)
- Writing system availability and Unicode support

We employ logistic regression predicting Invisible Giant status from colonial history variables, controlling for speaker population and geographic region.

### 3.4. LLM Training Data Analysis

We analyze language distribution in major LLM training datasets:

- The Pile (Gao et al., 2021): 800GB English-focused corpus
- mC4 (Xue et al., 2021): Multilingual C4 covering 101 languages
- ROOTS (Laurencçon and the BigScience Workshop, 2022): BLOOM training data, 46 languages
- OSCAR (Suárez et al., 2019): Common Crawl extractions, 166 languages

For each dataset, we quantify token counts by language and correlate with vitality scores to assess whether training data allocation reflects demographic reality or digital bias.

## 4. Results

### 4.1. The Four-Category Framework

Applying the invisible information framework to all 7,613 documented human languages reveals a stark reality: linguistic representation in digital ecosystems follows systematic patterns that diverge dramatically from demographic vitality. Figure 1 visualizes the vitality-digitality matrix, with each language plotted according to its real-world demographic strength (x-axis) and digital presence across web pages, Wikipedia, datasets, models, and archives (y-axis). The representation score – computed as digitality minus vitality – encodes the gap, with blue indicating digital overrepresentation and red indicating systematic underrepresentation.

Four distinct clusters emerge, each revealing different dynamics of visibility and erasure:

**Strongholds (+ Vitality / + Digitality)**: Roughly one-third of languages occupy the upper-right quadrant, exhibiting both demographic robustness and rich digital representation. These languages form a plume rising above both zero reference lines, their speaker bases, institutional support, and digital footprints mutually reinforcing. This category includes not only global lingua francas but also regionally dominant languages that have successfully translated offline vitality into digital presence. From an LLM perspective, these languages already power most training corpora and anchor multilingual benchmarks. The policy implication is maintenance rather than intervention – keeping data current while redirecting capacity-building resources toward less fortunate quadrants.

**Digital Echoes (– Vitality / + Digitality)**: Only 6% of languages appear in the upper-left quadrant – languages whose online presence exceeds their dwindling speaker communities. Often characterized by historical prestige, liturgical use, or active diaspora networks, these languages demonstrate that digital corpora can outlive real-world vitality. They appear as a thin blue veil above the vitality origin, illustrating how documentation and archiving can preserve linguistic material even as living use declines. For researchers, Digital Echoes serve as cautionary tales: high digital metrics do not guarantee linguistic health. For policymakers, they represent potential leverage points – using strong digital assets to support revitalization before community fluency erodes further.

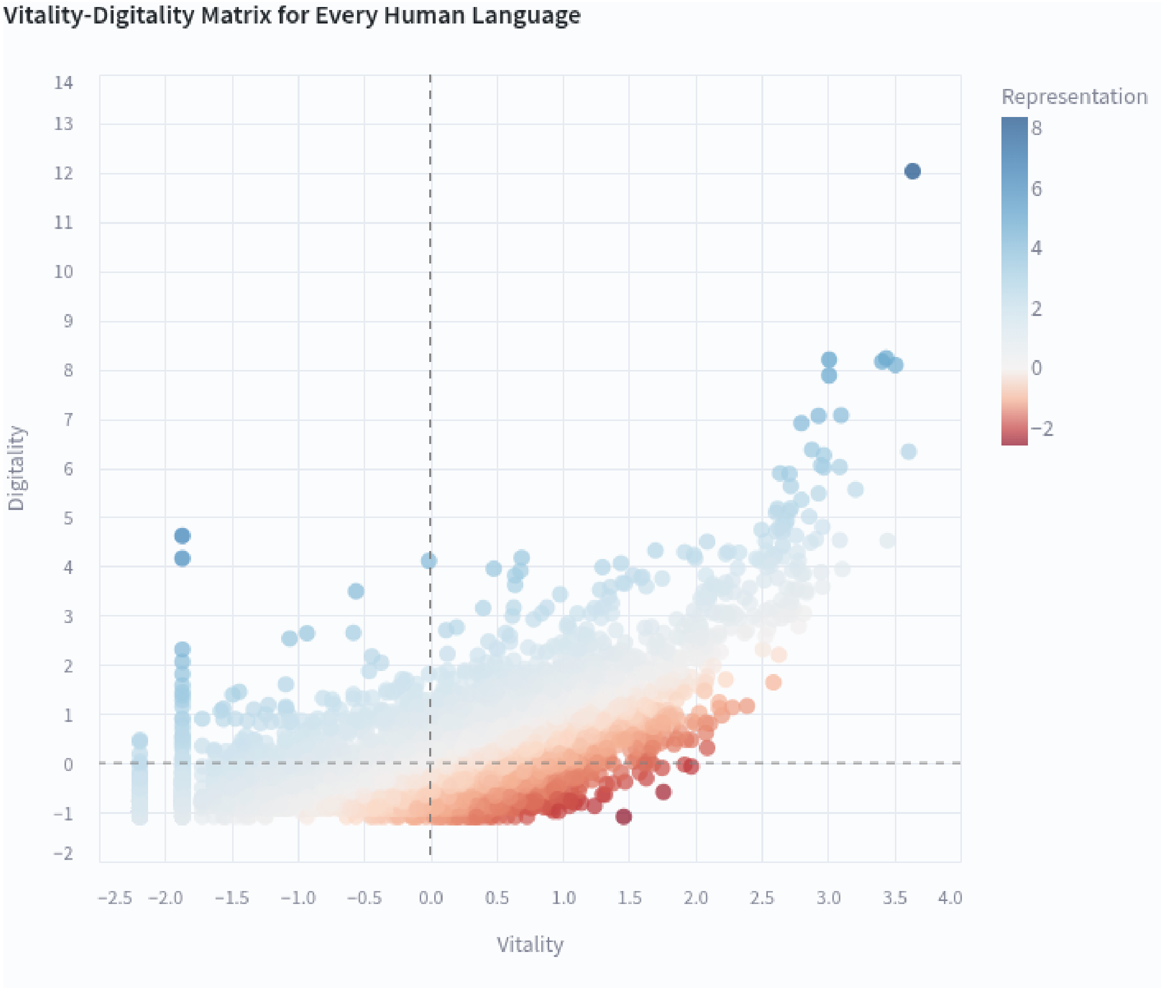

**Fig. 1**: Mapping 7613 human languages on vitality (ground presence) and digitality (web presence).

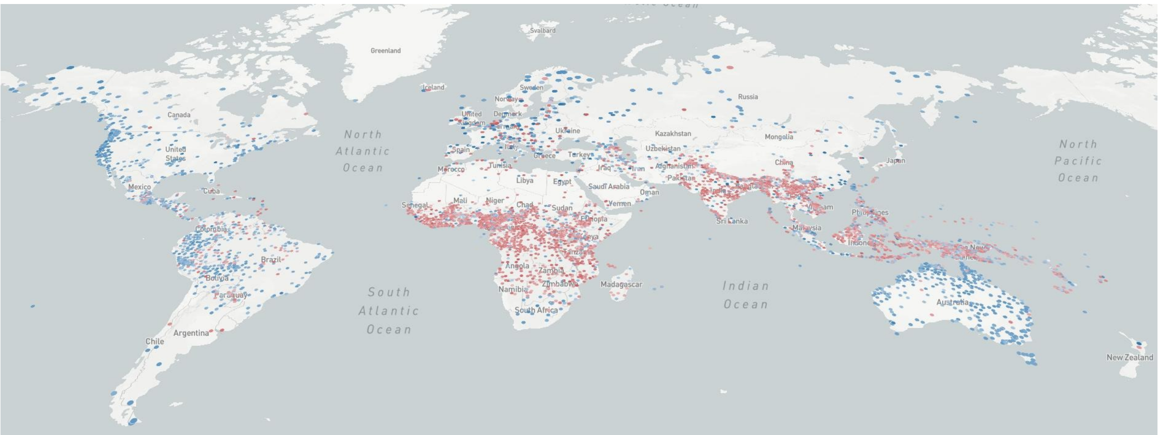

**Fig. 2**: Geolocating invisible languages

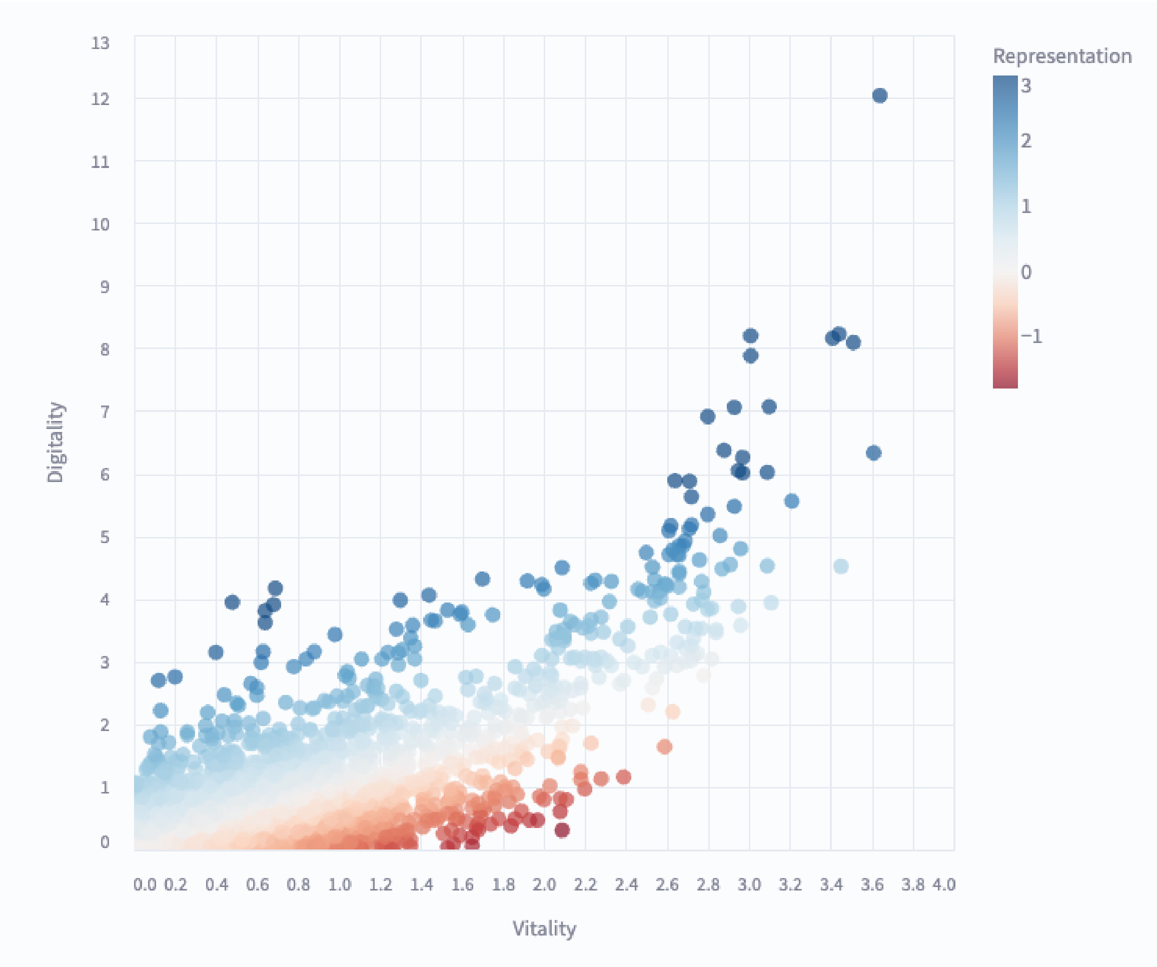

**Fig. 3**: Strongholds

**Fading Voices (– Vitality / – Digitality)**: The densest cluster, approximately 2,700 languages (36%), hugs the origin and extends into the lower-left. These languages suffer a double deficit: small, vulnerable speech communities and vanishingly small digital footprints. They form a pale cloud of nearly colorless points, underscoring that when both dimensions are scarce, even the representation gap provides minimal signal. From an LLM perspective, they are effectively invisible. From a linguistic diversity standpoint, they constitute the front lines of loss – without urgent fieldwork and community-led documentation, these voices will fade beyond empirical recovery.

**Invisible Giants (+ Vitality / – Digitality)**: Approximately 2,000 languages (27%) occupy the lower-right quadrant – the critical focus of our analysis. These languages manifest as vivid red points: their vitality extends rightward while digitality hovers below the horizontal reference line, producing the largest positive "need for action" gap on the color scale. These are languages with millions of active speakers but scant digital representation. The paradox is profound: robust demographic presence coexists with near-total digital absence.

The Invisible Giants category reveals the core thesis of this paper: digital exclusion is not simply a function of endangerment or small speaker populations. Languages can be vibrant, widely spoken, and intergenerationally transmitted while remaining systematically invisible in the digital

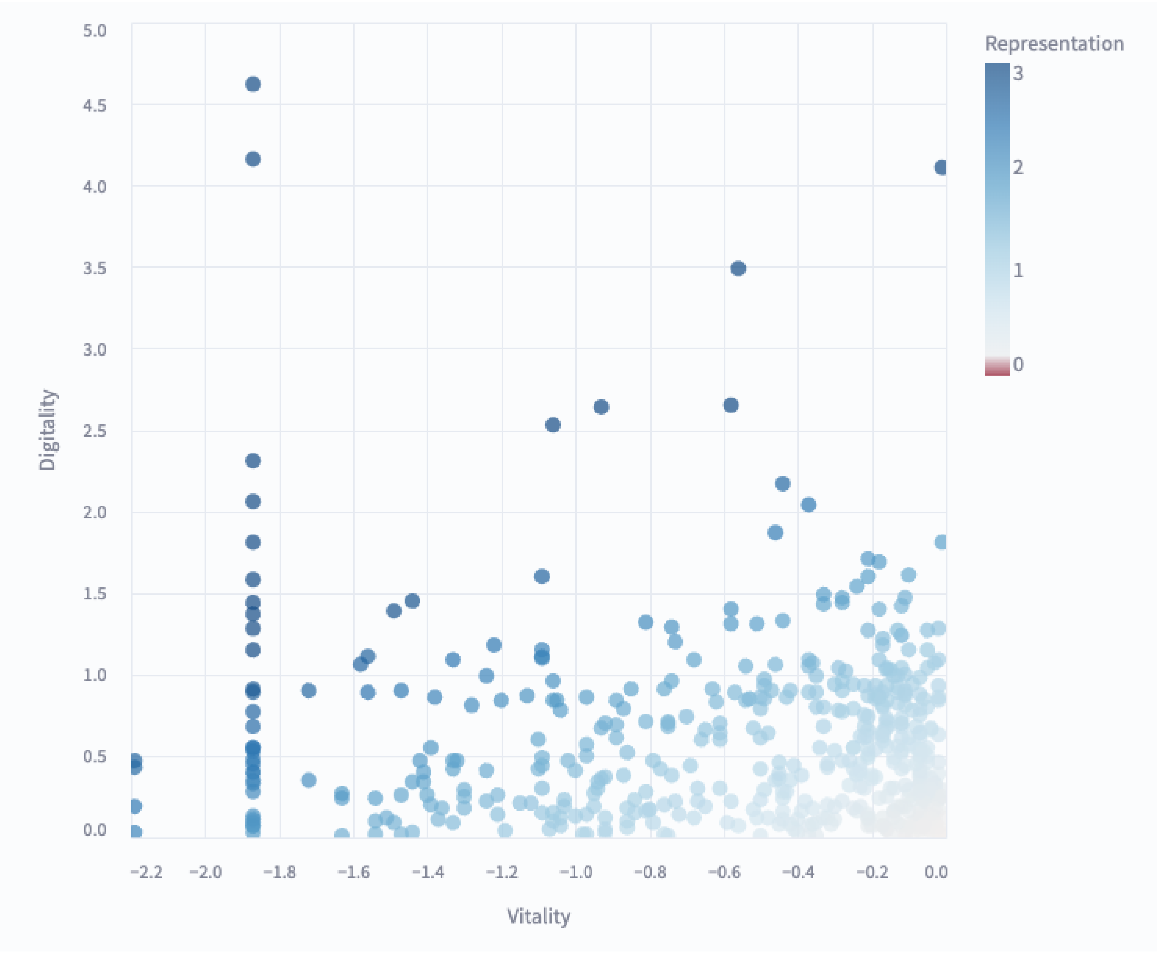

**Fig. 4**: Digital Echoes

ecosystems that increasingly define authoritative knowledge and power contemporary AI systems.

## 4.2. Theoretical Interpretation: Why Invisible Giants Exist

The existence of Invisible Giants – languages with millions of speakers yet minimal digital presence – cannot be explained by technical resource constraints alone. If digitality simply reflected speaker numbers, the vitality-digitality plot would approximate a diagonal line. Instead, we observe systematic deviation: certain languages punch far above their demographic weight digitally, while others with larger populations remain nearly absent.

This pattern aligns precisely with postcolonial theory's predictions. Said (1978) established how colonial powers constructed hierarchies positioning non-Western languages as inferior, requiring civilizing intervention. Phillipson (1992) documented how British colonial education policies systematically suppressed vernacular languages, privileging English and select indigenous lingua francas for administrative convenience. These hierarchies did not disappear with formal decolonization – they became embedded in digital infrastructure decisions.

The infrastructure for digitality – Unicode encoding, keyboard layouts, spell-checkers, search algorithms, content moderation systems – was built primarily by and for Western languages. When

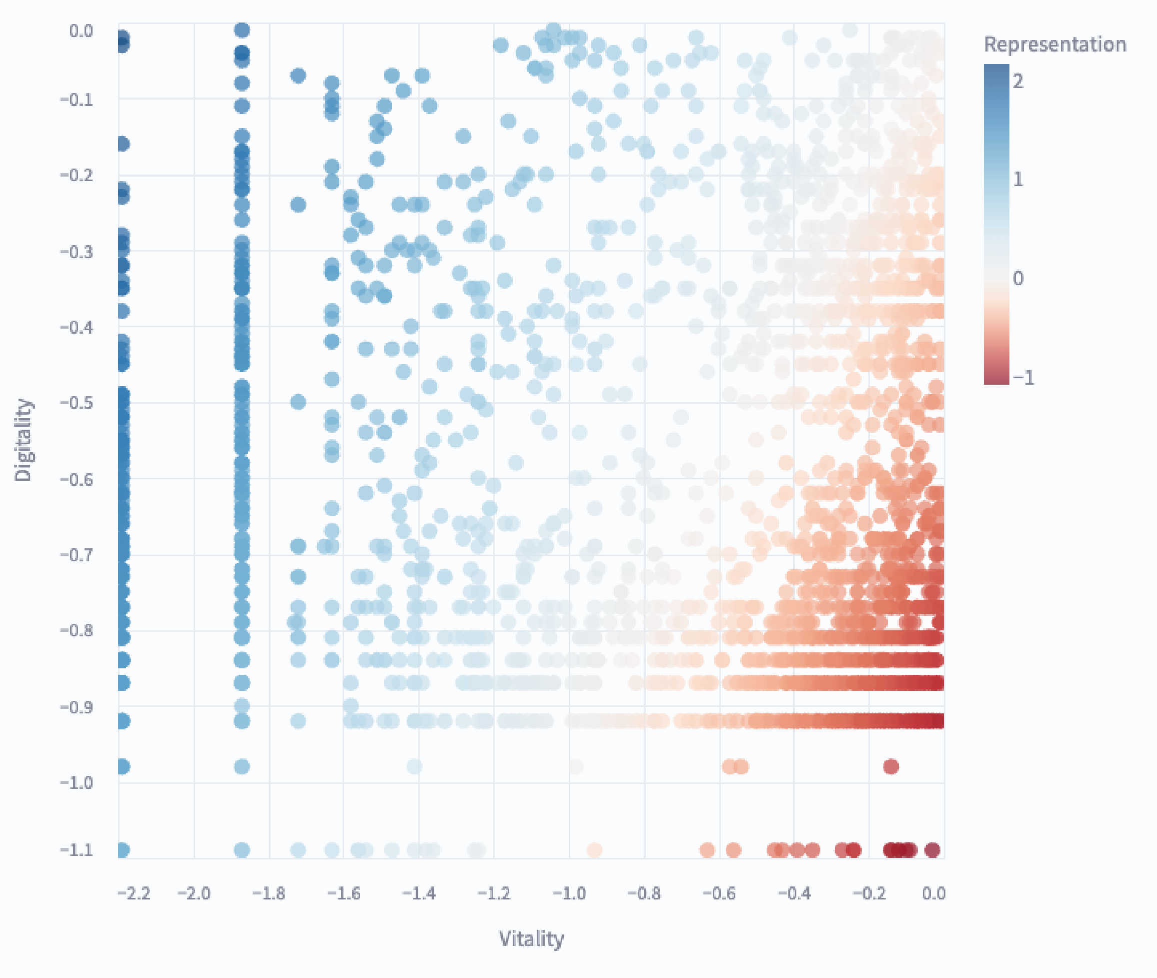

**Fig. 5**: Fading Voices

platforms assess ”notability” for Wikipedia articles, ”quality” for search rankings, or ”representativeness” for training data, they apply standards developed in English-speaking contexts. As Noble (2018) demonstrated for search engines, these ostensibly neutral technical systems encode cultural assumptions that systematically disadvantage marginalized communities.

The concept of epistemic injustice (Fricker, 2007) illuminates the mechanism. Digital systems commit hermeneutical injustice by failing to provide linguistic resources needed for marginalized communities to interpret and share their experiences. When an Invisible Giant language lacks digital presence, speakers cannot access AI tools in their language, cannot find online educational resources, cannot participate in digital commerce on equal footing. The system renders their knowledge production illegible, positioning them as passive consumers of content produced in dominant languages rather than active knowledge creators.

Critically, the Invisible Giants category demonstrates that this exclusion is structural, not natural. These languages are not ”low-resource” due to inherent deficiency – they are systematically under-resourced by institutions that chose to invest in some languages but not others. The framing matters: ”low-resource” naturalizes scarcity, while ”systematically under-resourced” centers the political decisions that produced it (Dotan et al., 2023).

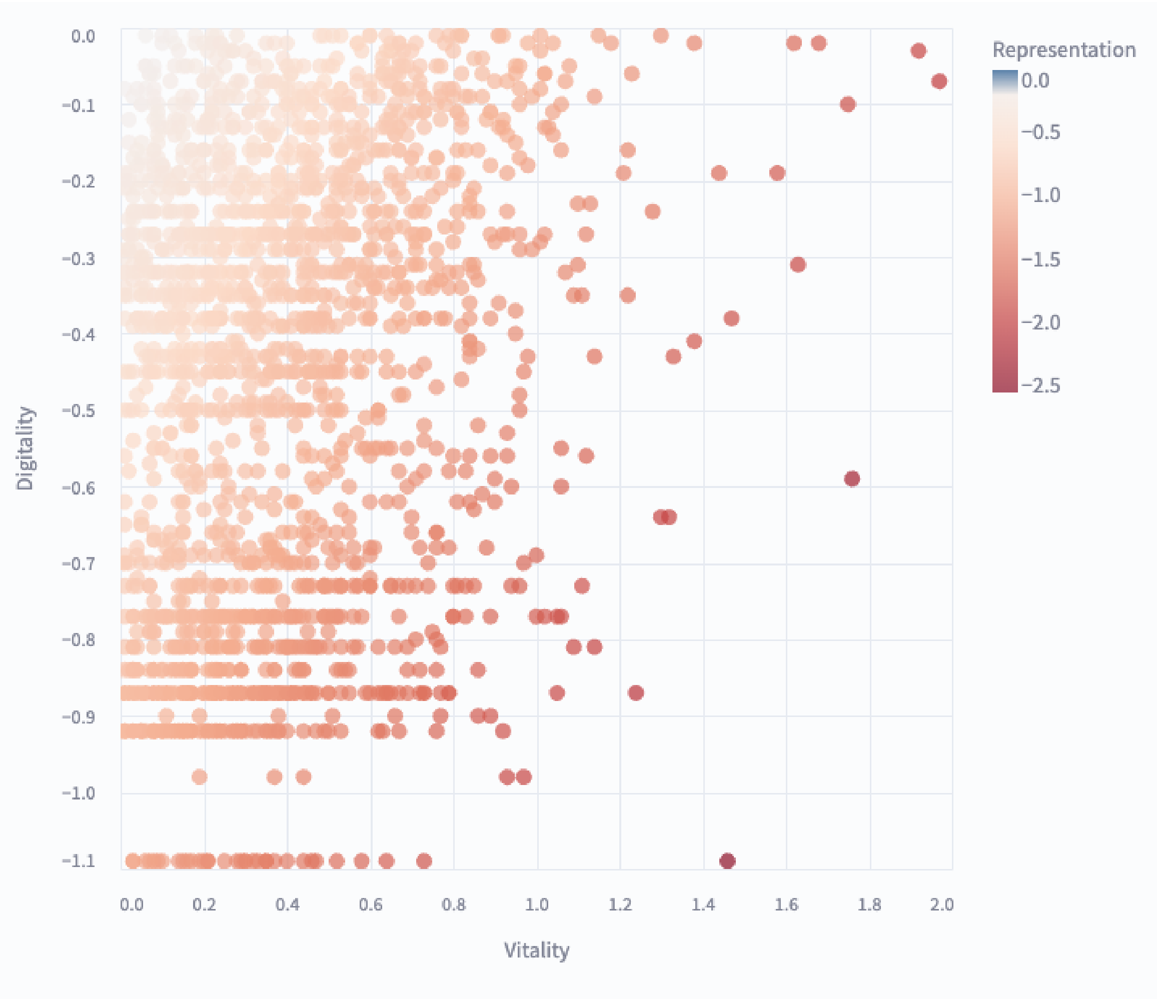

**Fig. 6**: Invisible Giants

## 4.3. Implications for Large Language Models

The vitality-digitality gap has profound implications for LLMs, which train on precisely the digital corpora our framework measures. When Invisible Giant languages constitute 27% of languages with substantial speaker bases yet receive minimal representation in Common Crawl, Wikipedia, and scholarly archives, LLM training data systematically excludes nearly a third of the world's demographically robust linguistic diversity.

This exclusion produces three compounding harms:

**Performance inequality**: Models trained predominantly on Stronghold languages perform dramatically worse on Invisible Giants (Joshi et al., 2020). This performance gap is not merely inconvenient – it determines who can access AI-powered education, commerce, healthcare, and governance. As LLMs become infrastructure, linguistic exclusion becomes infrastructural inequality.

**Knowledge erasure**: LLMs encode not just linguistic patterns but conceptual structures, cultural knowledge, and ways of knowing embedded in training text. When training data excludes Invisible Giants, models cannot represent the knowledge systems, historical narratives, and lived experiences of communities speaking those languages. As Bender et al. (2021) argue, this constitutes epistemic violence at scale – billions of parameters encoding the message that certain languages and knowledge don't matter.

**Feedback loops**: Poor LLM performance discourages speakers from using their languages digitally, reducing digital content generation, which further reduces training data, which worsens performance. This creates self-reinforcing cycles where invisibility begets further invisibility – a digitally mediated language shift mechanism operating at unprecedented scale and speed.

The geographic map in Figure 2 visualizes these patterns spatially. Africa appears predominantly red, indicating languages with higher vitality than digitality. Sub-Saharan African languages, despite centuries of oral tradition maintaining rich knowledge systems, remain systematically absent from digital archives (Nekoto et al., 2020). Meanwhile, Europe appears predominantly blue – even small European languages maintain digital presence exceeding their speaker populations, a legacy of institutional investment in digital infrastructure.

This is not coincidence but the continuation of colonial-era hierarchies through algorithmic means – what Mohamed et al. (2020) term ”AI colonialism.” Just as colonial powers extracted material resources while positioning colonized peoples as lacking civilization, contemporary AI development extracts training data (often from Global South digital labor) while positioning non-Western languages as ”low-resource” requiring benevolent intervention.

### 4.4. Validation Through Existing Research

Our empirical findings align with and extend prior fragmented observations. Kornai (2013) established that only 5% of languages will achieve meaningful digital vitality, identifying digital extinction as independent from offline endangerment. Our four-category framework provides the theoretical structure Kornai's work anticipated but did not fully develop: Digital Echoes exemplify his observation that archived languages can be digitally present without living use; Invisible Giants demonstrate his insight that lack of Wikipedia presence predicts inability to digitally ascend.

Simons et al. (2022) measured digital language support across 143 platforms, finding only 33 languages (0.4%) achieve full capabilities. Their methodology focused on technical infrastructure (encoding, fonts, keyboards); our framework adds the vitality dimension, revealing that infrastructure availability does not guarantee usage. Some Stronghold languages have infrastructure and active communities; Invisible Giants may have infrastructure but minimal content; Fading Voices lack both.

The pattern of 27% Invisible Giants validates warnings from Bender et al. (2021) that LLM training data scarcity for most languages is not incidental but structural. Their ”stochastic parrots” critique – that LLMs encode biases at scale without understanding – becomes more urgent when we quantify precisely which 27% of demographically robust languages are systematically excluded from training.

Crucially, our framework reveals what prior work could not: the magnitude of the representation gap. It is one thing to observe that some languages lack digital resources; it is another to demonstrate that approximately 2,000 languages with millions of speakers score in the highest ”need for action” category. This quantification transforms the problem from anecdotal concern to measurable crisis requiring systematic intervention.

# 5. Discussion

## 5.1. Digital-Epistemic Injustice as Structural Phenomenon

Our findings demonstrate that the vitality-digitality gap is not a technical accident but a structural phenomenon reflecting power relations. The systematic concentration of Invisible Giants – 27% of all languages, representing roughly 2,000 languages with millions of speakers yet minimal digital presence – reveals that digital exclusion operates independently of demographic strength.

This constitutes what we term *digital-epistemic injustice*: the systematic exclusion of marginalized linguistic communities from AI-mediated knowledge production through denial of both digital infrastructure and epistemic authority. This dual exclusion operates at unprecedented scale in LLM development. When billions of parameters encode patterns from training data that systematically excludes these 2,000 languages, the resulting models embed the message that certain ways of knowing, certain knowledge systems, and certain communities don't matter.

The mechanism operates through self-reinforcing feedback loops. Initial digital inequality leads to training data scarcity, producing poor LLM performance, discouraging digital language use, reinforcing data scarcity. Meanwhile, Stronghold languages – already enjoying institutional support and digital infrastructure – attract platform investment, NLP research funding, and commercial applications, widening gaps. Without intervention, these dynamics accelerate language shift as younger generations conclude that economic participation requires adopting dominant languages.

Critically, this is not inevitable technological progress but the reproduction of colonial hierarchies through algorithmic means. The concentration of Strongholds in Europe (where even minority languages maintain robust digital presence) while Invisible Giants concentrate in formerly colonized regions reveals historical continuities. Colonial language policies positioned indigenous languages as inferior, requiring civilizing intervention; contemporary ”low-resource language” framing positions them as deficient, requiring benevolent development assistance. Both framings naturalize inequality while obscuring its political origins.

## 5.2. Theoretical Contributions

**First**, we advance epistemic injustice theory by connecting it to digital infrastructure. Fricker (2007)'s framework focused on face-to-face testimonial exchanges and localized hermeneutical resources. Helm et al. (2024) extended epistemic injustice to language modeling but treated digitality as given. Our analysis reveals that *who gets digitized* is itself an exercise of power that precedes and enables all subsequent algorithmic processing.

The decision to invest in Unicode encoding, keyboard development, spell-checkers, and training corpora for some languages but not others determines which communities can practice hermeneutical justice – using linguistic resources to interpret experiences in ways that AI systems recognize as valid. When Invisible Giant languages lack these resources, speakers cannot access AI-powered education, healthcare information, or economic opportunities in their languages. The system denies them epistemic standing, positioning them as consumers of knowledge produced elsewhere rather than legitimate knowledge creators.

**Second**, we demonstrate how postcolonial hierarchies persist in algorithmic systems. While Said (1978) analyzed textual representations and Mohamed et al. (2020) documented extractive labor practices, we show continuities in linguistic hierarchies themselves. The vitality-digitality gap reveals how colonial language policies – suppressing vernaculars, imposing European languages, privileging select lingua francas – established hierarchies that LLM training data allocation now reproduces computationally.

Contemporary AI development exhibits what we term *algorithmic colonialism*: extracting digital labor from Global South communities while systematically excluding their languages from AI benefits, then marketing AI tools back to those communities as development interventions. This mirrors classic colonial resource extraction while adding a distinctly epistemic dimension – not merely extracting material resources but positioning entire knowledge systems as lacking, requiring external validation and technological uplift.

**Third**, we problematize the ”low-resource language” framing ubiquitous in NLP. This terminology positions underrepresented languages as inherently deficient – they *lack* resources. Our framework reveals that resource scarcity is a political product, not natural fact. The 2,000 Invisible Giants are not inherently low-resource; they are *systematically under-resourced* by institutions that chose to invest in some languages while neglecting others.

Reframing from ”low-resource” to ”systematically under-resourced” or ”digitally marginalized” centers structural forces rather than naturalizing inequality. This shift has implications beyond terminology – it redirects attention from technical data collection challenges to political questions about whose languages receive institutional investment, whose knowledge systems AI encodes, and who benefits from technological development.

## 5.3. Limitations and Future Directions

**Methodological limitations**: First, our digitality measurement relies on publicly accessible web data, potentially missing private communication in messaging applications. If Invisible Giant speakers use their languages extensively via WhatsApp, SMS, or encrypted platforms, our analysis underestimates actual digital use. However, from an LLM training perspective, private communication does not contribute to model development, so our measurement captures the dimension relevant for AI representation.

Second, the four-category classification uses median splits, creating discrete boundaries where gradients exist. Languages near category boundaries may be misclassified, and the framework does not capture within-category variation – not all Strongholds enjoy equal digital presence, nor do all Invisible Giants face identical challenges. Future work should develop continuous measures of the representation gap alongside categorical classifications.

Third, we lack longitudinal data on how vitality-digitality gaps evolve. Are gaps widening or narrowing? Do digital interventions successfully narrow gaps for targeted languages? Cross-sectional analysis cannot answer these causal questions. Longitudinal tracking of language trajectories across the vitality-digitality space would enable causal inference about what drives languages from one quadrant to another.

**Theoretical extensions**: Future research should examine intersections of linguistic marginalization with other forms of oppression. How does digital language inequality interact with disability (especially for signed languages’ digital representation)? With gender, particularly for languages where speaker populations are gendered? With age, as youth language innovations occur primarily in digital spaces? These intersectional analyses remain underdeveloped.

Platform-specific language ideologies warrant deeper investigation. Different platforms encode different linguistic norms – Twitter’s character limits advantage logographic systems, TikTok’s audiovisual primacy benefits languages with media presence, LinkedIn’s professional register expectations advantage languages with established business communication. How do platform affordances systematically advantage or disadvantage specific languages? We need granular sociotechnical analysis of how architectural choices embed linguistic assumptions.

Community-centered research methodologies remain rare. Our study, like most, analyzes patterns from external positions without deep engagement with affected language communities in research design or interpretation. Developing genuine co-production methodologies where speakers shape research questions, interpret findings, and control data governance represents a methodological frontier essential for decolonizing research practice.

**Empirical extensions**: The framework's generalizability across domains requires testing. We developed it for languages; Saurabh Khanna's broader Invisible Information project applies similar logic to armed conflicts and scholarly research. Does the vitality-digitality logic extend to other knowledge domains – traditional medicine, indigenous land management practices, oral historical archives? Systematic testing across domains would validate or refine theoretical claims about digital filtering mechanisms.

Regional deep dives would complement global analysis. Why do specific Invisible Giants remain underrepresented despite demographic strength? Detailed case studies examining language-specific barriers – orthographic debates, institutional politics, diaspora dynamics – would illuminate mechanisms obscured in aggregate analysis. Combining computational measurement with ethnographic depth would strengthen explanatory power.

### 5.4. Practical Implications

**For AI Developers**: Current LLM training treats language representation as a data availability problem – scrape more web pages, digitize more books. Our analysis reveals this approach is insufficient. The 2,000 Invisible Giants have speakers, they have knowledge to contribute, but they lack digital infrastructure and institutional support that would enable content creation at scales sufficient for LLM training.

Targeted investment in digital infrastructure for Invisible Giants is required: orthography standardization where needed, keyboard interfaces, spell-checkers, text-to-speech systems. These are prerequisites for content creation, not merely nice-to-have additions. Community-controlled data trusts enabling speakers to govern how language data is collected, used, and monetized would shift power dynamics from extractive to collaborative.

Evaluation metrics must center non-English linguistic features – tone marking accuracy, morphological complexity handling, pragmatic particle usage – rather than English-centric benchmarks like BLEU scores that penalize linguistic differences as errors. Performance equity should be measured not just by aggregate metrics but by whether models serve speakers of Invisible Giants as effectively as they serve English speakers.

**For Policymakers**: Language planning must integrate digital dimensions. Granting a language official status without digital infrastructure investment produces hollow recognition. Policies should mandate that government digital services support languages spoken by significant populations, fund localization of open-source software and educational platforms, and establish accountability mechanisms requiring AI companies to report representation gaps and remediation efforts.

Digital Language Rights should extend existing linguistic human rights frameworks into digital domains. Just as the Universal Declaration of Linguistic Rights (1996) establishes rights to use one's language in education and public life, contemporary frameworks must recognize rights to digital language resources, AI services in one's language, and participation in knowledge production digitized by platforms.

International funding mechanisms should prioritize Invisible Giants over Fading Voices for digital investment. While all endangered languages merit documentation, the 2,000 Invisible Giants offer highest social return on investment – millions of speakers exist who could immediately benefit from digital resources. This is not abandoning Fading Voices but strategically allocating limited resources where impact potential is greatest.

**For Researchers**: Methodological decolonization requires fundamental shifts in how we conduct research. Co-designing studies with speaker communities rather than extracting data, compensating language consultants at professional rates, sharing model weights and tools with communities providing training data, publishing in open-access venues with translations into studied languages – these practices should become norms, not exceptions.

Evaluating research impact by community-defined outcomes (language vitality metrics, economic benefits, cultural continuity) rather than only academic citations would better align incentives with social benefit. Research excellence should incorporate community endorsement as a criterion, recognizing that scholarship about marginalized communities should serve those communities' self-determined goals.

Funding agencies could accelerate these shifts by requiring invisibility impact assessments for digitization grants – explicitly addressing how proposed work will or will not narrow representation gaps – and prioritizing proposals co-designed with affected language communities over researcher-initiated projects.

### 5.5. Toward Linguistic Justice in AI

The Invisible Giants paradox – millions speaking, none listening digitally – crystallizes the crisis. Solutions require confronting uncomfortable truths: English dominance in AI is not natural or technologically necessary but reflects historical power structures that continue shaping which languages receive investment, recognition, and computational support.

Optimistically, concentrated investment could rapidly narrow gaps. The methodology exists: measure representation gaps, prioritize Invisible Giants, invest in community-controlled infrastructure, develop multilingual models with performance parity guarantees. The economics are feasible – comprehensive digital infrastructure for 2,000 Invisible Giants would cost a fraction of single LLM training runs measuring in hundreds of millions of dollars.

Pessimistically, current incentives point opposite directions. Commercial AI development optimizes for wealthy markets speaking Stronghold languages. Academic research chases state-of-the-art performance on English benchmarks, with multilingual work often treated as secondary. Platform business models reward engagement metrics that favor content in dominant languages. Without structural intervention – regulatory requirements, funding conditionalities, community organizing – market forces will accelerate inequality.

The choice before us is clear: will AI systems democratize access to information, education, and economic opportunity across linguistic boundaries? Or will they accelerate language shift, compressing human diversity toward English-centered homogeneity? The answer depends on whether we treat the vitality-digitality gap as a technical resource allocation problem or as the digital-epistemic injustice it truly represents – requiring not just more data, but fundamental redistribution of power, resources, and recognition.

## 6. Conclusion

We have demonstrated that linguistic inequality in Large Language Models is not incidental but structural, reflecting continuities from colonial-era hierarchies to contemporary AI development. Of the world's 7,613 languages, roughly 2,000 Invisible Giants with millions of speakers remain digitally underrepresented, concentrating in formerly colonized regions at rates nine times higher than Europe.

This pattern constitutes digital-epistemic injustice: the systematic exclusion of marginalized linguistic communities from AI-mediated knowledge production through denial of digital infrastructure and epistemic authority. As LLMs become infrastructure for education, commerce, and governance, this exclusion threatens to accelerate language shift, compressing human linguistic diversity toward homogenous English-centered multilingualism.

Yet this outcome is not inevitable. Concentrated investment in digital infrastructure for systematically under-resourced languages, community-controlled data governance, and decolonial AI development practices could narrow gaps within a generation. The question is one of political will: do we accept AI systems that reproduce colonial hierarchies, or do we demand technology that honors and amplifies the full richness of human linguistic diversity?

The invisible languages of the LLM universe need not remain invisible. Making them visible requires recognizing that their exclusion was never accidental – and that inclusion requires not just more data, but fundamentally reimagining who AI serves and whose knowledge counts.

## References

Anthropic (2024). Claude 3 models. *Anthropic Documentation*.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, page 610–623.

Benjamin, R. (2023). Race after technology: Abolitionist tools for the new jim code. *John Wiley & Sons*.

Blodgett, S. L., Barocas, S., III, H. D., and Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in nlp. *Proceedings of ACL 2020*, pages 5454–5476.

Chen, X. (2021). Language ideologies in chinese tourism discourse: "chinglish" and marginalization. *International Journal of Tourism Research*, 23(4):517–531.

Daems, J. (2024). Material continuities: From missionary linguistics to contemporary nlp. *Digital Humanities Quarterly*, 16(2).

Dotan, R., Dossou, B. F. P., Abbott, J., Muhammad, S. H., Emezue, C., Neyer, K., Nekoto, W., and Marivate, V. (2023). Decolonizing nlp: African language work beyond data collection. *arXiv preprint arXiv:2306.09355*.

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. (2021). The pile: An 800gb dataset of diverse text for language modeling. In *arXiv preprint arXiv:2101.00027*.

Garibay, C. and González, L. E. (2021). Machine translation and its effects on migrant rights: Case studies from u.s. immigration systems. *Migration Studies*, 9(3):1243–1262.

Gray, M. L. and Suri, S. (2019). Ghost work: How to stop silicon valley from building a new global underclass. *Houghton Mifflin Harcourt*.

Helm, F., Adelani, D., Khanna, S., and Deva, T. (2024). Diversity in language modeling and epistemic injustice. *arXiv preprint arXiv:2402.08186*.

Hohn, S., Schaefer, S. R., and Deva, T. (2024). Language ideology in natural language processing systems. In *Proceedings of the 2024 AAAI Conference on Artificial Intelligence*.

Joshi, P. J. J., Santosh, M. S. R., and Xiaochang, D. H. X. L. (2020). The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Kornai, A. (2013). Digital language death. *PLOS ONE*, 8(10):e77056.

Kukutai, T. and Taylor, J. (2016). Indigenous data sovereignty: Toward an agenda. In *ANU Press*, pages 1–22.

Laurencçon, P. and the BigScience Workshop (2022). Bigscience large language model training dataset. *arXiv preprint arXiv:2211.01763*.

Lewis, M. P., Simons, G. F., and Fennig, C. D. (2010). Assessing endangerment: Expanding fishman's gids. *Revue Roumaine de Linguistique*, 55(2):103–120.

Lewis, P., Simons, G. F., and Fennig, C. D. (2023). Ethnologue: Languages of the world, 25th edition.

Lippi-Green, U. (2012). *English with an Accent: Language, Ideology and Discrimination in the United States*. Routledge, 2nd edition.

Liu, M. and Li, B. (2017). Beneath the surface: Self-orientalization amongst chinese professionals in australia. *China Media Research*, 13(2):45–54.

Mohamed, S., Png, M.-T., and Isaac, W. (2020). Decolonial ai: Decolonial theory as sociotechnical framework for ai ethics. *Philosophy & Technology*, 33:659–684.

Nekoto, W., Marivate, V., Mugari, T., Abbott, J., Nkwenti, L., Muhammad, S. H., Emezue, C., Dossou, B. F. P., Olumide, S. E., Awokoya, L. T. P., and Whitenack, D. (2020). Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

on Endangered Languages, U. A. H. E. G. (2003). Language vitality and endangerment. Technical report, UNESCO.

OpenAI (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Phillipson, R. (1992). *Linguistic Imperialism*. Oxford University Press.

Said, E. W. (1978). *Orientalism*. Pantheon Books.

Simons, G. F., Ross, D. A., Lewis, P., Peterson, D. J., and Watanabe, K. (2022). Assessing the digital support of the world's languages: The langscape inventory. *Language Resources and Evaluation*, 56(2):331–376.

Soria, C. and Ballatore, A. (2017). A self-assessment framework for digital language vitality. In *Digital Language Diversity Project*.

Suárez, P. J. O., Romary, L., and Sagot, B. (2019). Asynchronous pipeline for processing huge corpora on medium-size clusters. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC)*, pages 9–16.

Talat, Z., Asr, F. T., and Birhane, A. (2024). Knowledge, power, and harm: Artificial intelligence and epistemic injustice. *arXiv preprint arXiv:2402.14811*.

Woolard, K. A. and Schieffelin, B. B. (1998). Language ideology. *Annual Review of Anthropology*, 23:55–82.

Xue, L., Constant, N., Roberts, A., Kale, M., Gupta, R., Zakharov, P., Fan, A., Manning, C. D., Diab, M., and Cer, D. (2021). mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

## Acknowledgments

We acknowledge research funding from the Fair and Resilient Societies Working Group and the Amsterdam School of Communication Research at the University of Amsterdam. We also acknowledge the countless language communities whose knowledge and labor make AI systems possible, yet who rarely benefit from the technologies built on their linguistic resources.