Enigma in Cultural
enigma.or.id/index.php/cultural

e-ISSN 3026-7277
Vol. 4, No. 2 (2026) | pp. 85–103

ORIGINAL ARTICLE

Pragmatic Flattening and Honorific Erasure: A Sociolinguistic and Computational Audit of Speech-Level Hierarchies (*Undha-Usuk* and *Basa Alus*) and Ritual Metaphors Across Large Language Models in Javanese, Sundanese, and Balinese

Andi Fatihah Syahrir[1], Ahmad Badruddin[2*], Bhawani Singh[3]

[1]Department of Communication Sciences, CMHC Research Center, Palembang, Indonesia

[2]Department of Literacy Education, Enigma Institute, Palembang, Indonesia

[3]Department of Ecotourism, Social Sciences Institute, Phagwara, India

ARTICLE INFO

Article history
Received: July 22, 2026
Revised: August 20, 2026
Accepted: September 19, 2026
Available online: October 15, 2026
Published: October 30, 2026

Keywords
cultural alignment; decolonial computing; indigenous data governance; large language models; speech-level pragmatics

***Corresponding author**
Ahmad Badruddin · ahmad.badruddin@enigma.or.id
ORCID: 0009-0003-6697-5062

DOI
https://doi.org/10.61996/cultural.v4i2.144

ABSTRACT

Background: Multilingual large language models (LLMs) trained predominantly on Anglocentric and register-neutral Indonesian web corpora risk pragmatic flattening across Nusantara speech-level traditions: Javanese *Undha-Usuk*, Sundanese *Undak-Usuk Basa*, and Balinese *Sor Singgih Basa*, where social deference and cosmological order are grammatically encoded.

Objective: To audit cross-model epistemic attrition, few-shot calibration, generative register fidelity, and parallel-corpus honorific retention across Javanese, Sundanese, and Balinese.

Methods: A four-stream computational-sociolinguistic audit of public resources examined 2,620 regional IndoMMLU items (plus 3,213 Indonesian comparators), 19 SEACrowd foundation models, 960 paired 0-shot/3-shot evaluations (1,920 runs) across six 4-bit instruction-tuned LLMs, 1,320 native-speaker-rated Cendol generations across 22 LLMs, and 3,000 parallel NusaX clauses. Tests were Friedman, Wilcoxon, Cochran's Q, exact McNemar, Kruskal–Wallis, and Benjamini–Hochberg-adjusted Fisher's exact.

Results: Pedagogical corpora stratified speech-level honorifics (10.08%), classical metaphors/poetics (12.56%), script/rituals (9.08%), and vernacular lexicons (68.28%; $p < 0.0001$). Across 19 LLMs, regional performance trailed Indonesian by 10.91–18.36 percentage points on metaphor, discourse, and sentiment tasks (all $p < 0.001$). Three-shot prompting yielded non-significant regional gains (35.83% to 37.36%; $p = 0.4472$), while wrong regional answers carried 56.62% mean softmax confidence. Models outside the Cendol family showed 12.73-fold higher collapse odds ($p < 0.0001$), Cendol-aligned models still produced pronoun–verb register collisions, and parallel corpora retained Javanese *Krama* in only 20.99% of kinship-trigger clauses (79.01% flattened).

Conclusion: Honorific erasure was observed together with source-pivot corpus flattening and uncalibrated confidence, which motivates the proposed CARE-Pragmatics custodial alignment framework.

Copyright © 2026 The Author(s). The authors retain copyright. Published by Enigma Institute.

Access license Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International (CC BY-NC-SA 4.0). Noncommercial sharing and adaptation are permitted with attribution. Adaptations must use the same license.

e-ISSN 3026-7277
Vol. 4, No. 2 (2026) | pp. 85–103

1. Introduction

1.1 Algorithmic Monolingualism, Data Colonialism, and Pragmatic Homogenization in Foundation Models

Instruction-tuned large language models (LLMs) now serve as everyday intermediaries for education, translation, public administration, and cultural heritage inquiry, and their rapid global deployment has reopened foundational debates in cultural studies and linguistic anthropology about who controls the digital reproduction of human meaning.[1–3] Contemporary commercial and open-weights foundation models claim broad multilingual competence across hundreds of languages. Yet scholarship in this area shows that web-scale pre-training pipelines remain structurally dominated by English and a narrow tier of state-standardized national pivot languages.[1,2] Couldry and Mejias conceptualize this extractive asymmetry as data colonialism. Peripheral linguistic communities are incorporated into global computational infrastructures through standardized digital traces, which strip away locally situated social relations, relational obligations, and indigenous interpretive norms.[4,5] Autoregressive transformers often learn regional languages mainly through machine-translated web scrapes, bilingual dictionaries, or crowd-sourced translations mediated by egalitarian, register-neutral pivot languages. In that situation, statistical fluency in surface token prediction can conceal severe homogenization at the level of pragmatics.[3–5]

Recent cross-cultural alignment audits confirm that foundation models systematically gravitate toward Western, educated, industrialized, rich, and democratic (WEIRD) communicative defaults, even when prompted in non-Western languages.[6,7] Benchmarks of everyday cultural commonsense, social norm adaptability, and culturally grounded value reasoning reveal persistent performance gaps whenever correct interpretation depends on sub-national kinship etiquette, age-graded deference, or non-Anglocentric figurative traditions.[8–10] Existing global cultural alignment frameworks, however, mostly operationalize culture through declarative survey values or encyclopedic trivia, such as regional cuisines, public holidays, or historical landmarks. They overlook structural sociolinguistic pragmatics.[6,8–10] In the major Austronesian literary civilizations of Maritime Southeast Asia, culture is more than a topic spoken about in language. Ethical hierarchy, affective self-restraint, and cosmological order are obligatorily grammaticalized and lexicalized inside every clause through asymmetrical speech-level systems.[11,12]

1.2 Semiotic Architecture of Nusantara Speech Levels and Figurative Cosmologies: Javanese, Sundanese, and Balinese

Indonesia's three largest speech-level traditions are Javanese (spoken by over 80 million people across Central and East Java), Sundanese (spoken by roughly 40 million people in West Java and Banten), and Balinese (spoken by approximately 3.3 million people on Bali, Nusa Penida, and western Lombok). They share a centuries-old philological heritage shaped by Old Javanese (*Kawi*) court literature, Sanskritic epistemic contact, Islamic ethical synthesis, and village customary law (*Adat*). Yet each maintains a distinct semiotic architecture of interpersonal deference.[11–14] Errington's classic semiotic ethnographies of Javanese court and urban interaction showed that speech levels (*Undha-Usuk* [speech levels] or *Unggah-Ungguh Basa*) function as indexical signs. These signs continuously calibrate the social distance, genealogical seniority, and spiritual refinement (alus versus kasar) linking speaker, addressee, and third-person referent.[11–13] Classical and contemporary Javanese pedagogy distinguishes plain vernacular *Ngoko* [plain register] (*Ngoko Lugu* and *Ngoko Alus*) from deferential *Krama* [polite register] (Madya, *Krama Lugu*, and *Krama Alus*). Hundreds of suppletive honorific (*Krama Inggil* [exalted honorific forms]) and self-effacing humiliative (*Krama Andhap*) lexical items cross-cut these levels. They govern bodily organs, speech acts, motion, dining, and attire.[14–16] Javanese etiquette strictly prohibits self-honorification. Applying a *Krama Inggil* predicate such as *dhahar* (to dine) or *tindak* (to go) to oneself, rather than to a respected interlocutor, commits a jarring breach of the ethical ideal of *andhap asor* (humble self-lowering before others).[11,15,16]

Parallel yet structurally non-identical speech-level hierarchies govern Sundanese and Balinese social life. Sundanese *Undak-Usuk Basa* [speech levels] has been reframed in contemporary West Javanese pedagogy as *Tatakrama Basa Sunda*, which emphasizes mutual ethical respect or *silih asih*, *silih asah*, *silih asuh*. Speakers must distinguish familiar *Basa Loma* [familiar register] from two polite registers. *Basa Lemes keur Sorangan* consists of humiliative polite forms reserved strictly for the speaker's own actions and possessions, such as *neda* for eating, *rorompok* for house, or *mios* for departing. *Basa Lemes keur Batur* consists of deferential honorifics reserved for respected addressees and elders, such as *tuang*, *bumi*, and *angkat*.[17,18] Balinese *Sor Singgih Basa* [speech levels] works differently. Speech-level selection there intersects both symmetrical civic respect and traditional *Tri Wangsa* (*Brahmana*, *Ksatria*, *Wesia*) and *Jaba* social history. It stratifies the lexicon across *Basa Kasar*, *Basa Andap/Kepara* (everyday vernacular), *Basa Madia*, and three specialized high registers. *Basa Alus Singgih* elevates respected or high-status persons, including somatic terms derived from Sanskritic *Kawi* such as *naka* for fingernail and *marayunan* for dining. *Basa Alus Sor* humbles the speaker or lower-ranked referent. *Basa Alus Mider* covers symmetrically polite forms addressing collective civic or ritual assemblies.[19–21]

Equally vital to cultural transmission across Java, Pasundan, and Bali is the intertwining of polite register with classical prosodic literature, embodied spatial deference, and ancestral figurative genres.[21–23] Moral epistemology in Javanese education is transmitted through *Tembang Macapat* meters, such as *Pocung*, *Kinanthi*, *Sinom*, *Pangkur*, and *Dhandhanggula* in Mangkunegara IV's nineteenth-century ethical treatise *Serat Wedhatama*, which are governed by strict syllabic counts per line (*guru wilangan*), stanzaic line counts (*guru gatra*), and terminal vowel rhymes (*guru lagu*). Idiomatic proverbs sit alongside them. These distinguish literal agentive metaphors (*saloka*), situational metaphors (*bebasan*), and fixed maxims (*paribasan*).[22,23] Sundanese cultural pedagogy, in turn, transmits ancestral toponymic prophecies (*cacandran*, which articulate the historical character of landscapes and polities), rhyming couplets (*sisindiran*), and *Pupuh* [verse forms] prosody. Balinese literary education integrates *Sekar Alit* (*Pupuh*), *Sekar Madia* (*Kidung*), and *Sekar Agung* (*Kakawin* [Old Javanese poetic genre]) directly into temple liturgy (*Panca Yadnya*), *Pura* spatial orientation, and customary oratory (*Pidarta* and *Sambrama Wacana*).[19–22] The risk is clear. When digital language technologies process these languages without encoding speech-level relationality or prosodic-figurative ontology, they threaten the intergenerational survival of living intangible heritage.[15,17,20]

From a comparative typological perspective, Nusantara speech-level architectures pose a far more intricate representational challenge for neural language models than the better-studied morphosyntactic politeness systems of Europe or East Asia.[11,14,19] European T–V pronominal distinctions (such as French tu/vous or German du/Sie) localize interpersonal deference primarily in second-person pronouns and verb agreement, and East Asian honorific systems (such as Japanese Keigo or Korean Jondaetmal) predominantly attach regular agglutinative verbal suffixes and honorific prefixes onto shared lexical stems. Javanese *Undha-Usuk*, Sundanese *Tatakrama Basa*, and Balinese *Sor Singgih* operate differently, through pervasive full-clause lexical suppletion.[11,14,17,19] Converting a plain Javanese *Ngoko* utterance (*'Kowe wis mangan sega durung ing omahmu?'*) into polite *Krama Alus* (*'Panjenengan sampun dhahar sekul dereng wonten ing dalemipun panjenengan?'*) replaces virtually every open- and closed-class morpheme in the clause (pronoun, aspectual auxiliary, predicate verb, object noun, negator, locative preposition, and nominalizer) with phonologically unrelated etyma drawn from *Kawi*, Sanskrit, or courtly coinages.[12–14] Subword byte-pair encoding tokenizers are trained predominantly on Latin-script English and Indonesian. For them, such wholesale lexical suppletion fractures semantic equivalence across sparse, low-frequency token neighborhoods. Attention layers are then deprived of shared root morphemes that might otherwise anchor cross-register generalization.[11,20]

1.3 Hermeneutical Injustice, Indonesian Benchmark Gaps, and Study Objectives

Fricker's philosophy of epistemic injustice and Mignolo's framework of decolonial epistemic disobedience supply a lens for one systematic failure: foundation models cannot distinguish *Ngoko* from *Krama Inggil*, *Basa Loma* from *Basa Lemes* [polite register], or ancestral *Cacandran* [ancestral place-prophecies] from generic proverbs. Seen this way, the failure can be read as a computational form of hermeneutical injustice. Indigenous interpretive categories that organize moral personhood are rendered unintelligible within dominant technological infrastructures.[24–26] Recent cultural sciences scholarship published in Enigma in Cultural shows both the resilience and fragility of archipelagic symbolic systems facing modern institutional regimes. Rahayu et al. offer a clinical ethnography of Balinese cosmological negotiation and ritual efficacy in psychiatric care,[27] while Singh and Kumar evaluate indigenous heritage digitalization and warn that uncurated digital archiving can decontextualize living ritual knowledge from community custodianship.[28] Extending this critique to generative artificial intelligence is urgent. Regional governments, schools, and cultural agencies across Indonesia increasingly experiment with LLM-based translation and educational chatbots.

Over the past five years, Indonesian and Southeast Asian natural language processing initiatives, including IndoNLG,[29] NusaX,[30] NusaWrites,[31] IndoMMLU,[32] MaBL,[33] Cendol,[34] SEACrowd,[35] LORAXBENCH,[36] and CULTURETALK-ID,[37] have expanded evaluation resources for regional languages considerably. Yet a methodological gap persists between aggregate leaderboard reporting and sociolinguistic theory. A recent study evaluated language models on the four Javanese honorific levels.[38] Prior benchmark studies primarily report single macro-averaged accuracy or surface character n-gram overlap (chrF++) scores, without disaggregating performance across speech-level honorifics, classical prosody, and ritual scripts, without auditing whether contemporary instruction-tuned models recognize their own epistemic uncertainty via softmax probability calibration, and without testing whether Indonesian-pivoted parallel training corpora themselves bake honorific erasure into supervision data.[30,32,34,35]

To bridge cultural semiotics and computational benchmarking, this study conducts a multi-stream empirical audit. Four research questions guide it. (1) How are speech-level honorifics, classical metaphors/poetics, indigenous scripts/rituals, and vernacular lexicons distributed across formal Javanese, Sundanese, and Balinese educational corpora? How severely do 19 global and regional LLMs degrade from national Indonesian pivots to these three regional languages? (2) Does three-shot in-context prompting overcome zero-shot performance ceilings across contemporary instruction-tuned LLMs? How calibrated are model softmax probabilities when cultural or honorific predictions fail? (3) In native-speaker human evaluations of open-ended generations, how do global multilingual, Southeast Asian regional, and Indonesian-aligned model families diverge, from complete generative collapse to subtle intra-clause register collisions? (4) To what extent do parallel human-translated corpora pivoted through register-neutral Indonesian induce asymmetric speech-level flattening across Javanese, Sundanese, and Balinese, and how can indigenous data governance principles guide prospective sociolinguistic alignment?

2. Methods

2.1 Multi-Stream Computational and Sociolinguistic Audit Design

This study was a retrospective computational-sociolinguistic audit built from four streams. Each stream drew on a different kind of evidence: quantitative benchmark disaggregation, item-level probabilistic calibration analysis, native-speaker generative error evaluation, and parallel-corpus register analysis. The languages were Javanese (ISO 639-3: jav), Sundanese (sun), and Balinese (ban), set against national Indonesian (ind) comparators.[30,32,34,35] We did not use synthetic model self-judgments (LLM-as-a-judge). Such judgments suffer from severe circularity in low-resource languages. Instead, all four empirical streams were anchored in human-authored regional examination items, professional parallel translations, and native-speaker human evaluation annotations, all released under open-science licenses.[30,32–35]

Data-hygiene screening came first, before any statistical aggregation. While inspecting the upstream SEACrowd natural language understanding logs, we found a configuration artifact. Language-tagged IndoMMLU entries returned identical aggregate scores across regional tags in several model logs, because subsets had been loaded without filtering. Cross-model macro comparisons in Stream 2 were therefore restricted to verified language-distinct benchmarks: multicultural figurative language (MaBL),[33] parallel reading comprehension (Belebele),[39] parallel sentiment pragmatics (NusaX),[30] topical classification (SIB-200),[40] and conversational/encyclopedic translation (NusaX-MT and FLORES-200).[30,35] Exact item-level IndoMMLU regional evaluation was run directly from verified item IDs in Streams 1 and 3.[32]

2.2 Corpus Architecture Across the Four Empirical Streams

Stream 1 covered the complete regional-language pedagogical corpus of IndoMMLU.[32] It comprised N = 2,620 multiple-choice

cultural and linguistic examination items spanning Grades 1 through 12: *Bahasa Jawa* (n = 992; 37.86%), *Bahasa Sunda* (n = 1,157; 44.16%), and *Bahasa Bali* (n = 471; 17.98%). These were paired with n = 3,213 *Bahasa Indonesia* items (total N = 5,833 items out of 14,981 total IndoMMLU records). Balinese items carried a combined school-span label in the raw metadata, yet they kept exact numerical grade levels (kelas 1.0–12.0). All 2,620 regional items were therefore mapped deterministically, by numerical grade, into three standardized Indonesian educational tiers. The tiers were Primary School (SD, Grades 1–6; n = 1,314), Lower Secondary School (SMP, Grades 7–9; n = 715), and Upper Secondary School (SMA, Grades 10–12; n = 591).

Stream 2 analyzed empirical evaluation matrices from SEACrowd,[35] covering 19 natural language understanding (NLU) models and 19 natural language generation (NLG) models. Four groups of models were represented. Commercial frontier APIs were GPT-4 (NLU suites) and GPT-4o (NLG suites). Global open-weights instruction models were Command-R, Meta-Llama-3-8B-Instruct, Mistral-7B-Instruct-v0.2, Qwen-7B-Chat, Falcon-7B-Instruct, BLOOMZ-7B1, mT0-XL, Aya-101, and Aya-23-8B. Southeast Asian regional models were SeaLLM-7B-v2.5, Sailor-7B-Chat, SEA-LION-7B-Instruct, Malaysian-Llama-3-8B, and LLaMA3-8B-WangchanX. Indonesian-specialized models were Cendol-Llama2-7B, Cendol-mT5-XL, Merak-7B-v4, and Bactrian-X.[34,35,41,42] For each dataset, accuracy (for NLU) and character n-gram F-score chrF++ (for NLG) were averaged across standardized prompt templates per model.

Stream 3 audited item-level zero-shot versus three-shot predictions, along with option-level softmax probability vectors, from the open-access 4-Bit-IndoMMLU evaluation suite. The suite covers six contemporary instruction-tuned decoder LLMs quantized via 4-bit NormalFloat (NF4) protocols:[43] Gemma 2 9B, Qwen 2 7B, Mistral 7B v0.3, Meta-Llama 3 8B, Aya 23 8B, and Phi 3 Mini 3.8B. Every model was evaluated on the identical stratified sample of 40 items per language. For *Bahasa Jawa*, *Bahasa Sunda*, and *Bahasa Bali*, this gave n = 120 unique regional items (720 model–item pairs per prompting regime). *Bahasa Indonesia* contributed n = 40 items (240 model–item pairs). The total was 960 paired zero-shot/three-shot evaluations (1,920 item–model–shot runs). Zero-shot and three-shot records were joined strictly on unique IndoMMLU item identifiers. This step kept few-shot demonstration prefixes from contaminating sociolinguistic domain classification.

Stream 4 combined two distinct generative and translational corpora. Stream 4A examined the complete native-speaker human evaluation dataset released with Cendol,[34] comprising N = 1,320 open-ended generative outputs (660 Javanese and 660 Sundanese; 30 prompts per model per language). The outputs came from 22 LLMs stratified into three architectural families. The first family was Indonesian-aligned Cendol models (14 models: 10 encoder–decoder mT5 variants from 300M to 13B and 4 decoder-only Llama-2 7B/13B variants; n = 840 outputs). The second was Southeast Asian and Indonesian regional chat models (SeaLLM-7B-Chat, SEA-LION-7B-Instruct, Merak-7B-v4, and Bactrian-ID-BLOOM-7B1; n = 240 outputs). The third was global multilingual baselines (Llama-2-7B-Chat, Llama-2-13B-Chat, BLOOMZ-7B1, and Bactrian-X-BLOOM-7B1; n = 240 outputs). Native annotators graded every output on a four-tier ordinal scale. Grade A (score 4) meant fluent, accurate, and pragmatically natural. Grade B (score 3) meant comprehensible, with minor lexical or register drift. Grade C (score 2) meant severe semantic distortion or hallucinated additions. Grade D (score 1) meant complete generative collapse, untranslated source repetition, or Indonesian/English code leakage. Qualitative diagnostic notes accompanied the grades.[34] Stream 4B audited all N = 1,000 twelve-way parallel human-translated utterances of NusaX.[30] They were split into training (n = 500), validation (n = 100), and testing (n = 400) partitions. This yielded 3,000 parallel regional clauses in Javanese, Sundanese, and Balinese, aligned to identical Indonesian source sentences.

2.3 Sociolinguistic Domain Taxonomy and Deictic Register Lexicons

To operationalize cultural-linguistic theory across Streams 1 and 3, every regional IndoMMLU item stem and option set was sorted into four sociolinguistic domains. Sorting relied on mutually exclusive, priority-ordered regular-expression lexicons grounded in Javanese, Sundanese, and Balinese philology.[11,14,17,19] (1) Domain D1 (Speech-Level Honorifics and Deictic Etiquette) captured explicit metalinguistic terms and suppletive honorific/humiliative predicates (e.g., krama inggil, ngoko alus, unggah-ungguh, undha-usuk, undak-usuk, basa lemes, basa loma, sor singgih, alus singgih, alus sor, alus mider, *panjenengan*, *dhahar*, *tindak*, *ngendika*, *tuang*, *kulem*, *titiang*, *marayunan*, *malinggih*). (2) Domain D2 (Figurative Metaphors and Classical Poetics) captured traditional metaphorical genres and prosodic rules (e.g., *paribasan*, *bebasan*, *saloka*, *pepindhan*, *wangsalan*, *cangkriman*, tembang macapat, *guru wilangan*, *guru lagu*, *babasan*, *sisindiran*, *cacandran*, *pakeman basa*, *sesonggan*, *bladbadan*, *wewangsalan*, pupuh, kidung, kakawin, *padalingsa*). (3) Domain D3 (Indigenous Script, Wayang Mythology, and Ceremonial Rituals) captured Brahmic orthography (*aksara* Jawa/Sunda/Bali, pasangan, sandhangan, rarangken, gantungan, hanacaraka). It also covered shadow-puppetry epics (*wayang*, *Pandhawa*, *Punakawan*, Mahabharata, Ramayana) and ritual/oratorical institutions (*upacara adat*, *Galungan*, *Kuningan*, *Nyepi*, *Odalan*, *Yadnya*, *banten*, *sesorah*, *panatacara*, *biantara*, *pidarta*). (4) Domain D4 (Neutral Vernacular Lexicon and Discourse Comprehension) encompassed everyday vocabulary, descriptive syntax, and narrative reading passages that did not invoke explicit honorific, prosodic, or liturgical markers.

In Stream 4B, we tested the source-pivot flattening hypothesis across NusaX. First, we identified Indonesian source sentences that contained explicit asymmetric social, kinship, or service-deference triggers. These were elder/kinship titles and polite interpersonal markers (*ibu*, *bapak*, ayah, *orang tua*, nenek, kakek, guru, pak, bu, *Anda*, beliau, tuan, nyonya, *pelanggan*, *terima kasih*, mohon, maaf, tolong; n = 81 trigger utterances vs. n = 919 general utterances). Borrowed kinship nouns could inflate the results tautologically, for example when the noun *ibu* is retained inside an otherwise blunt *Ngoko* clause. To avoid this, we constructed strict grammatical and predicate honorific lexicons. The Javanese *Krama*/*Krama Inggil* lexicon included *sampun*, *dereng*, *mboten*, *panjenengan*, kula, *tindak*, *kondur*, *dhahar*, saged, *badhe*, *wonten*, *menika*, *ingkang*, and nyambut *damel*. The Sundanese *Basa Lemes* lexicon included *parantos*, *atos*, *teu acan*, *anjeun*, *abdi*, *simkuring*, *sumping*, *mulih*, *tuang*, *tiasa*, *bade*, *kanggo*, *janten*, *sareng*, *damel*, and *kantos*. The Balinese *Basa Alus* [refined register] lexicon included *sampun*, *durung*, *nenten*, *ida*, *dane*, *titiang*, *rauh*, *mantuk*, *marayunan*, *prasida*, *pacang*, *wenten*, *punika*, *sane*, *makarya*, and *naenang*.

e-ISSN 3026-7277
Vol. 4, No. 2 (2026) | pp. 85–103

2.4 Statistical Estimation, Calibration Metrics, and Multiple-Testing Control

All statistical analyses were executed deterministically in Python using NumPy, pandas, and SciPy with two-sided alpha = 0.05. Exact 95% Wilson score confidence intervals accompanied the binomial proportions (sociolinguistic domains, accuracy strata, generative collapse rates, and parallel honorific retention rates). They ensured valid boundary coverage at small cell counts. In Stream 1, heterogeneity across the three regional languages and four sociolinguistic domains was evaluated with Pearson's chi-square contingency test and Cramer's V effect size. Stream 2 tested repeated-measures differences across Indonesian, Javanese, Sundanese, and Balinese for the 19 paired LLMs. Non-parametric Friedman chi-square tests came first. Two-sided Wilcoxon signed-rank tests followed.

In Stream 3, inter-model heterogeneity across the six 4-bit LLMs on the 120 shared regional items was tested via Cochran's Q test. Paired zero-shot versus three-shot accuracy transitions within each model, domain, and educational tier were evaluated via exact binomial McNemar tests on discordant item pairs. Option-level probability arrays were normalized to sum to unity. Maximum predicted softmax confidence was then extracted. Epistemic overconfidence was defined as mean maximum softmax probability minus empirical accuracy. Confidence distributions for correct and incorrect regional predictions were compared via two-sided Mann–Whitney U tests. In Stream 4A, ordinal native-rater scores across the three LLM families were compared via Kruskal–Wallis H tests with epsilon-squared effect sizes. The remaining analyses were pairwise Mann–Whitney U tests with Cliff's delta, 4,000-resample non-parametric bootstrap 95% confidence intervals (fixed seed 202606), and exact Fisher's odds ratios for Grade D collapse. In Stream 4B, cross-register retention across the three parallel translations was evaluated via Cochran's Q and two-sided Fisher's exact odds ratios, which compared social-deixis trigger clauses against general clauses. The Benjamini–Hochberg false discovery rate (BH-FDR) procedure adjusted these tests.

To ensure full mathematical replicability across all four streams, our non-parametric and exact inferential estimators were specified algebraically as follows. For any observed binomial count k out of N trials ($\hat{p} = k/N$), the 95% Wilson score confidence interval ($z = 1.95996$) was computed as $[\hat{p} + z^2/(2N) \pm z\sqrt{(\hat{p}(1-\hat{p})/N + z^2/(4N^2))}] / [1 + z^2/N]$, avoiding the zero-width degeneracy and under-coverage of normal Wald intervals in low-prevalence strata such as general-clause Javanese *Krama* retention (30/919 = 3.26%, 95% CI: 2.30%–4.62%). In Stream 3, Cochran's Q across M = 6 models evaluated on I = 120 binary regional items ($X_{i,m}$ equal to 0 or 1, with item sums $R_i = \sum_{m=1}^{M} X_{i,m}$, model sums $C_m = \sum_{i=1}^{I} X_{i,m}$, and grand total $T = \sum_{m=1}^{M} C_m$) was evaluated as $Q = (M-1)[M\sum_{m=1}^{M} C_m^2 - T^2] / [MT - \sum_{i=1}^{I} R_i^2]$ against a central chi-square distribution with $M - 1 = 5$ degrees of freedom. Paired zero-shot ($X_{i,0}$) versus three-shot ($X_{i,3}$) transitions were tested via exact two-sided binomial tail probabilities on discordant counts $b = \sum_i I(X_{i,0} = 1 \wedge X_{i,3} = 0)$ and $c = \sum_i I(X_{i,0} = 0 \wedge X_{i,3} = 1)$ under null probability 0.5. For option probability vector $u_i = (u_{i,1}, \ldots, u_{i,K})$ normalized via $p_{i,c} = u_{i,c} / \sum_{j=1}^{K} u_{i,j}$, maximum predicted confidence was $c_i = \max_c p_{i,c}$ and group-level overconfidence equaled $\Delta\text{conf} = \bar{c} - \text{Acc}$. In Stream 4A, omnibus ordinal differences across the three model families (N = 1,320) were quantified via Kruskal–Wallis H and epsilon-squared effect size $\varepsilon^2 = H / [(N^2 - 1)/(N + 1)]$, while pairwise ordinal dominance between groups X (n1) and Y (n2) was measured via Cliff's delta $d = [\sum_{i,j} I(x_i > y_j) - \sum_{i,j} I(x_i < y_j)] / (n_1 n_2)$. In Stream 4B, ordered Fisher's exact p-values $p(1) \leq p(2) \leq p(3)$ across the three parallel regional languages were converted into Benjamini–Hochberg step-up false discovery rate q-values via $q(r) = \min_{s \geq r} \min(1, (3/s)p(s))$.

3. Results

3.1 Sociolinguistic Curriculum Stratification and 19-LLM Macro Epistemic Attrition

Stream 1 covered N = 2,620 regional pedagogical items. Formal school assessments devoted nearly one-third of them (31.72%; n = 831) to culturally specialized competencies in pragmatics, poetics, and ritual, not to neutral vocabulary alone. The composition differed sharply across Javanese, Sundanese, and Balinese (chi-square = 88.40, df = 6, $p = 6.50 \times 10^{-17}$, Cramer's V = 0.1299; Table 1 and Figure 1). Speech-level honorifics and interpersonal deixis (Domain D1) made up 10.08% of regional items overall (n = 264; 95% CI: 8.98%–11.29%). They peaked in Balinese *Sor Singgih* assessments at 14.86% (70/471; 95% CI: 11.93%–18.36%) and in Javanese *Undha-Usuk* assessments at 12.60% (125/992; 95% CI: 10.68%–14.81%), against 5.96% in Sundanese (69/1,157; 95% CI: 4.74%–7.48%). Figurative metaphors and classical prosody (Domain D2) represented 12.56% of regional items (n = 329; 95% CI: 11.34%–13.88%). Javanese *Macapat* [Javanese metrical poetry] and *paribasan*/*saloka* items led this domain at 16.63% (165/992), and Balinese *sesonggan*/pupuh items followed at 11.89% (56/471). Indigenous Brahmic scripts, Wayang mythology, and ceremonial rituals (Domain D3) accounted for 9.08% overall (n = 238; 10.69% Javanese, 9.13% Balinese, and 7.69% Sundanese). Neutral vernacular comprehension (Domain D4) made up 68.28% (n = 1,789). Breaking the items down by educational tier showed balanced representation: Primary (SD: 50.15%, n = 1,314), Lower Secondary (SMP: 27.29%, n = 715), and Upper Secondary (SMA: 22.56%, n = 591).

Stream 2 benchmarked all 19 commercial and open-weights LLMs. Accuracy fell steeply whenever evaluation moved from national Indonesian to Javanese, Sundanese, and Balinese, and this held across every comprehension, pragmatic, and generative paradigm (Table 2 and Figure 2). On multicultural figurative language understanding (MaBL), the 19-model mean accuracy dropped from 63.66% ± 10.60% in Indonesian to 52.75% ± 6.67% in Javanese (−10.91 percentage points; paired Wilcoxon p = 0.0005). In Sundanese it reached 50.84% ± 9.77% (−12.81 points; Wilcoxon p = 0.0005; omnibus Friedman chi-square = 16.78, p = 0.0002). That level approaches the 50% binary random-guessing floor across most 7B–8B models. Even the commercial frontier model GPT-4 plunged from 82.19% on Indonesian metaphors to 27.67% on Javanese metaphors and 13.17% on Sundanese metaphors. Both regional scores lie far below the 50% chance level of this binary task. A similar cross-lingual erosion appeared on Belebele parallel discourse comprehension. Indonesian scored 52.00% ± 22.07%, against 38.74% ± 15.65% for Javanese (−13.26 points, $p < 0.0001$) and 36.23% ± 13.98% for Sundanese (−15.77 points, $p < 0.0001$; Friedman chi-square = 19.89, $p < 0.0001$). NusaX affective sentiment classification followed the same pattern: Indonesian 67.63% ± 18.61% vs. Javanese 56.05% (−11.57 points), Sundanese 52.19% (−15.43 points), and Balinese 49.26% (−18.36 points); Friedman chi-square = 40.99, $p = 6.56 \times 10^{-9}$. So did NusaX-MT conversational generation, where the 19-

e-ISSN 3026-7277
Vol. 4, No. 2 (2026) | pp. 85–103

model mean chrF++ fell from 38.05 ± 10.72 in Indonesian to 26.09 in Javanese, 25.67 in Sundanese, and 24.84 in Balinese (Friedman chi-square = 35.97, $p = 7.60 \times 10^{-8}$; all paired Wilcoxon $p = 3.81 \times 10^{-6}$). On SIB-200 topical classification, the 19-model mean accuracy also contracted, from 71.49% in Indonesian to 58.65% in Javanese, 57.51% in Sundanese, and 56.12% in Balinese. Official Javanese regional exam items in M3Exam averaged just 29.89% ± 10.42% across the 19 models.

Model-specific anomalies in Stream 2 may explain why high-capacity commercial models can fail even more sharply than smaller open-weights models on regional figurative tasks. The 7B–8B open-weights models appeared to lack regional idiomatic knowledge. On MaBL Javanese and Sundanese metaphor pairs they typically defaulted to uniform binary guessing (~50.0%–52.8%). GPT-4 behaved differently. It fell well below chance, to 27.67% in Javanese and 13.17% in Sundanese, despite reaching 82.19% on Indonesian metaphors and 91.33% on Belebele Indonesian reading comprehension. MaBL pairs culturally specific figurative expressions with literal-translation distractors that sound believable.[33] An instruction-following frontier model has strong lexical translation priors but weak grounding in Javanese *saloka* or Sundanese *babasan*. Such a model systematically selects the literal compositional gloss over the non-compositional cultural meaning. One reading is active negative transfer rather than random guessing; a score this far below chance may also reflect answer extraction or refusals. The massively multilingual encoder–decoder Aya-101[42] showed the opposite profile. It kept comparatively balanced cross-lingual retention on MaBL (71.20% Indonesian, 64.44% Javanese, 65.28% Sundanese) and on NusaX sentiment (86.25% Indonesian, 85.50% Javanese, 80.50% Sundanese, 77.67% Balinese). This illustrates the value of broad multilingual multitask supervision before generative decoding.

Table 1. Multi-Stream Empirical Corpus Architecture and Sociolinguistic Domain Distribution Across Javanese, Sundanese, and Balinese Pedagogical Items (N = 2,620 Regional; N = 5,833 Total IndoMMLU).

Stream / Sociolinguistic Stratum	Javanese (Bahasa Jawa)	Sundanese (Bahasa Sunda)	Balinese (Bahasa Bali)	Regional Total / Scope
Panel A: Stream 1 Pedagogical & Cultural Corpus (IndoMMLU)	**n = 992 (37.86%)**	**n = 1157 (44.16%)**	**n = 471 (17.98%)**	**N = 2620 (+3,213 Ind)**
D1: Speech-Level Honorifics (*Undha-Usuk* / *Lemes* / Alus)	125 (12.60% [10.7–14.8])	69 (5.96% [4.7–7.5])	70 (14.86% [11.9–18.4])	264 (10.08% [9.0–11.3])
D2: Figurative Metaphors & Classical Poetics (*Macapat* / *Pupuh*)	165 (16.63% [14.4–19.1])	108 (9.33% [7.8–11.2])	56 (11.89% [9.3–15.1])	329 (12.56% [11.3–13.9])
D3: Indigenous Script (*Aksara*), Wayang & Ceremonial Rituals	106 (10.69% [8.9–12.8])	89 (7.69% [6.3–9.4])	43 (9.13% [6.8–12.1])	238 (9.08% [8.0–10.2])
D4: Neutral Vernacular Lexicon & Narrative Comprehension	596 (60.08% [57.0–63.1])	891 (77.01% [74.5–79.3])	302 (64.12% [59.7–68.3])	1789 (68.28% [66.5–70.0])
Educational Tiers (SD Grades 1–6 / SMP 7–9 / SMA 10–12)	396 / 298 / 298	718 / 294 / 145	200 / 123 / 148	1,314 / 715 / 591
Panel B: Cross-Stream Computational & Human Evaluation Scope	**Javanese Coverage**	**Sundanese Coverage**	**Balinese Coverage**	**Analytical Unit Total**
Stream 2: SEACrowd Macro Benchmark (19 NLU & 19 NLG LLMs)	MaBL, Belebele, NusaX, SIB	MaBL, Belebele, NusaX, SIB	NusaX-Senti/MT, FLORES, SIB	38 LLM Benchmark Suites
Stream 3: 4-Bit 0-Shot vs. 3-Shot Softmax Audit (6 LLMs)	40 items × 6 LLMs × 2 shots	40 items × 6 LLMs × 2 shots	40 items × 6 LLMs × 2 shots	1,440 Regional (+480 Ind)
Stream 4A: Cendol Native-Rated Generations (22 LLMs, Grades A–D)	n = 660 rated outputs	n = 660 rated outputs	Not covered in Cendol R1	N = 1,320 Human Ratings
Stream 4B: NusaX Parallel Deictic Audit (Train/Valid/Test)	n = 1,000 parallel clauses	n = 1,000 parallel clauses	n = 1,000 parallel clauses	N = 3,000 Regional Clauses

Note: Bracketed intervals denote 95% Wilson score binomial confidence intervals. Omnibus contingency test across the three regional languages and four sociolinguistic domains in Stream 1 yielded chi-square = 88.40 (df = 6, $p = 6.50 \times 10^{-17}$, Cramer's V = 0.1299).

Table 2. Cross-Architecture Epistemic Attrition Across 19 Commercial and Open-Weights Foundation Models in SEACrowd: National Indonesian Pivot Versus Javanese, Sundanese, and Balinese.

Model / Summary Metric	MaBL Metaphor IND / JAV / SUN (%)	Belebele QA IND / JAV / SUN (%)	NusaX Sentiment IND / JAV / SUN / BAN (%)	NusaX-MT Generation IND / JAV / SUN / BAN (chrF++)
GPT-4 (Commercial Frontier)	82.2 / 27.7 / 13.2	91.3 / 77.6 / 71.3	86.2 / 52.5 / 38.0 / 25.5	55.2 / 46.1 / 43.7 / 36.0 (GPT-4o)
Command-R (35B Multilingual)	77.5 / 58.3 / 53.8	77.3 / 51.3 / 44.6	77.2 / 72.8 / 68.0 / 68.0	52.9 / 26.8 / 26.5 / 26.9

e-ISSN 3026-7277
Vol. 4, No. 2 (2026) | pp. 85–103

Model / Summary Metric	MaBL Metaphor IND / JAV / SUN (%)	Belebele QA IND / JAV / SUN (%)	NusaX Sentiment IND / JAV / SUN / BAN (%)	NusaX-MT Generation IND / JAV / SUN / BAN (chrF++)
SeaLLM-7B-v2.5 (SEA Regional)	77.7 / 55.4 / 54.4	78.1 / 51.1 / 45.9	65.5 / 52.7 / 48.8 / 30.2	45.6 / 26.6 / 26.0 / 26.1
Sailor-7B-Chat (SEA Qwen-Based)	68.5 / 57.9 / 53.9	68.5 / 38.9 / 38.6	84.2 / 72.6 / 63.2 / 59.8	44.9 / 30.8 / 27.2 / 25.3
Aya-101 (13B Encoder–Decoder)	71.2 / 64.4 / 65.3	66.6 / 64.0 / 62.4	86.2 / 85.5 / 80.5 / 77.7	48.6 / 40.3 / 41.8 / 32.0
Aya-23-8B (Multilingual Decoder)	75.7 / 57.6 / 47.8	71.2 / 39.7 / 33.7	53.7 / 37.0 / 33.8 / 32.0	49.7 / 24.7 / 23.2 / 24.4
Meta-Llama-3-8B-Instruct (Global)	61.9 / 52.4 / 50.9	53.2 / 37.9 / 33.3	60.6 / 51.8 / 45.3 / 44.0	44.1 / 28.8 / 28.5 / 27.1
Mistral-7B-Instruct-v0.2 (Global)	63.4 / 52.8 / 52.0	53.8 / 34.0 / 30.0	80.8 / 53.6 / 47.2 / 46.2	37.3 / 24.8 / 24.9 / 25.0
SEA-LION-7B-Instruct (SEA Regional)	50.0 / 50.0 / 50.0	23.0 / 22.9 / 23.0	38.2 / 38.0 / 38.4 / 38.2	39.6 / 28.0 / 27.2 / 25.3
Cendol-mT5-XL (3.7B Indonesian)	49.9 / 51.1 / 50.3	23.3 / 22.4 / 23.4	96.5 / 92.1 / 91.6 / 89.6	48.2 / 39.2 / 39.7 / 27.9
Cendol-Llama2-7B (Indonesian)	51.2 / 51.1 / 51.2	25.6 / 26.3 / 27.8	69.2 / 66.8 / 67.1 / 67.2	46.4 / 40.2 / 37.8 / 27.2
Summary: 19-LLM Macro Mean ± SD	63.7±10.6 / 52.8±6.7 / 50.8±9.8	52.0±22.1 / 38.7±15.7 / 36.2±14.0	67.6±18.6 / 56.0±17.7 / 52.2±17.5 / 49.3±18.4	38.0±10.7 / 26.1±9.1 / 25.7±8.6 / 24.8±5.4
Summary: Friedman Chi-Square & Gap	**χ^2 = 16.78 (p = 0.0002) Gap: −10.9 / −12.8 pp**	**χ^2 = 19.89 (p < 0.0001) Gap: −13.3 / −15.8 pp**	**χ^2 = 40.99 (p < 0.0001) Gap: −11.6 / −15.4 / −18.4 pp**	**χ^2 = 35.97 (p < 0.0001) Gap: −12.0 / −12.4 / −13.2 chrF++**

Note: Macro means, standard deviations, non-parametric repeated-measures Friedman chi-square tests, and two-sided paired Wilcoxon signed-rank tests are computed across all 19 evaluated LLMs in SEACrowd (all paired Indonesian-versus-regional Wilcoxon p <= 0.0005). In the NLG suite, OpenAI's commercial endpoint was evaluated using GPT-4o.

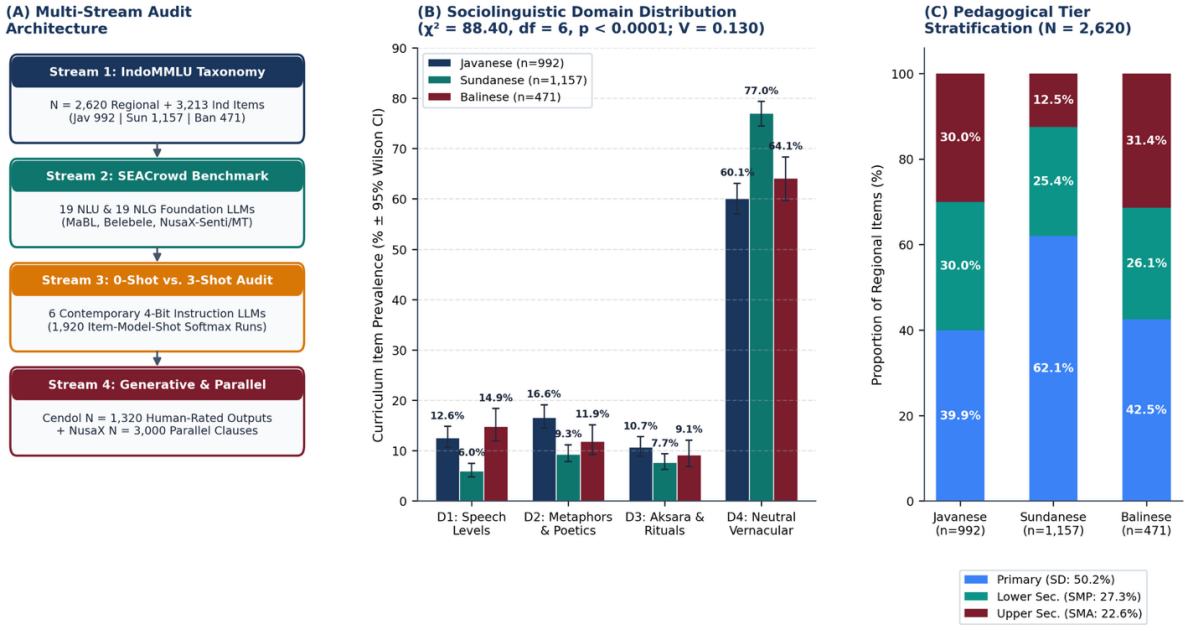

Figure 1. Multi-stream empirical audit architecture, sociolinguistic domain distribution, and pedagogical tier coverage across Javanese, Sundanese, and Balinese: (A) four-stream research architecture integrating Stream 1 IndoMMLU pedagogical items (N = 2,620 regional plus 3,213 Indonesian items), Stream 2 SEACrowd macro benchmarks across 19 NLU and 19 NLG LLMs, Stream 3 item-level 0-shot versus 3-shot softmax calibration runs (1,920 evaluations across 6 contemporary 4-bit LLMs), and Stream 4 generative/parallel evaluations (Cendol N = 1,320 native-rated outputs and NusaX N = 3,000 parallel regional clauses); (B) prevalence of Speech-Level Honorifics (D1), Figurative Metaphors and Classical Poetics (D2), *Aksara*, Wayang, and Rituals (D3), and Neutral Vernacular Lexicon (D4) across Javanese (n = 992), Sundanese (n = 1,157), and Balinese (n = 471) with 95% Wilson score confidence intervals (Pearson chi-square = 88.40, df = 6, p < 0.0001, Cramer's V = 0.130); and (C) school-tier stratification across Primary (SD: 50.2%), Lower Secondary (SMP: 27.3%), and Upper Secondary (SMA: 22.6%) curricula.

e-ISSN 3026-7277
Vol. 4, No. 2 (2026) | pp. 85–103

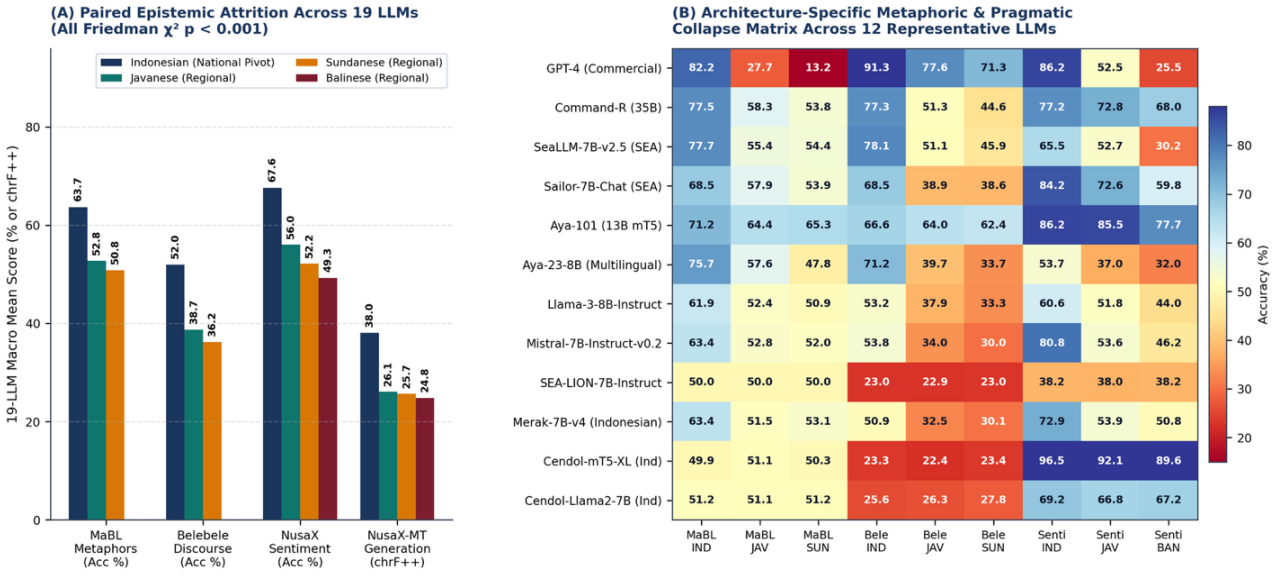

Figure 2. Cross-architecture epistemic attrition across 19 commercial and open-weights foundation models in SEACrowd: (A) paired 19-LLM macro mean accuracy (%) and chrF++ generation scores comparing national Indonesian pivots against Javanese, Sundanese, and Balinese across MaBL figurative metaphors (Friedman chi-square = 16.78, p = 0.0002), Belebele discourse comprehension (chi-square = 19.89, p < 0.0001), NusaX sentiment pragmatics (chi-square = 40.99, p < 0.0001), and NusaX-MT conversational generation (chi-square = 35.97, p < 0.0001); and (B) model-by-benchmark accuracy heatmap across 12 representative commercial, global multilingual, Southeast Asian, and Indonesian-aligned architectures, highlighting GPT-4's literal-distractor collapse on Javanese (27.7%) and Sundanese (13.2%) metaphors.

3.2 Three-Shot Prompting Ceilings and Softmax Overconfidence Across Contemporary 4-Bit LLMs

Stream 3 ran item-level paired evaluation across six contemporary 4-bit instruction-tuned LLMs (1,920 item–model–shot runs). Three-shot in-context prompting did not close the regional epistemic gap, and classical figurative reasoning stayed near random chance (Table 3 and Figure 3). Across the 720 pooled regional item–model pairs, zero-shot accuracy averaged 35.83% (258/720; 95% CI: 32.41%–39.40%). National Indonesian items reached 60.83% (146/240; 95% CI: 54.53%–66.79%). Three-shot prompting yielded a statistically non-significant +1.53 percentage-point regional shift to 37.36% (269/720; 95% CI: 33.90%–40.95%; 81 items lost vs. 92 items gained; exact McNemar p = 0.4472). Models differed widely across the 120 regional items, under both zero-shot (Cochran's Q = 51.55, df = 5, $p = 6.67 \times 10^{-10}$) and three-shot regimes (Cochran's Q = 53.60, df = 5, $p = 2.53 \times 10^{-10}$). Gemma 2 9B had the highest regional three-shot accuracy at 52.50% (Javanese 47.50%, Sundanese 62.50%, Balinese 47.50%). It still trailed its own 77.50% Indonesian benchmark by 25.00 percentage points. Mistral 7B v0.3 (40.00%), Qwen 2 7B (39.17%), and Meta-Llama 3 8B (38.33%) came next. Aya 23 8B (27.50%) and Phi 3 Mini 3.8B (26.67%) hovered near the 25% four-choice chance baseline. Three-shot demonstrations also induced negative interference in two cases: Meta-Llama 3 8B on Sundanese (47.50% to 42.50%) and Phi 3 Mini 3.8B on Javanese (35.00% to 25.00%).

Splitting Stream 3 performance by our four sociolinguistic domains revealed a sharp divide. Lexical etiquette responded to exemplars, whereas poetic cosmology resisted them. Speech-level honorific items (Domain D1; 15 unique items, n = 90 model evaluations) rose modestly from 38.89% in zero-shot to 46.67% in three-shot (+7.78 points; McNemar p = 0.2100). *Aksara*/mythology/ritual items (Domain D3; n = 72) shifted from 31.94% to 36.11% (+4.17 points; p = 0.6291). Figurative metaphors and classical *Macapat*/*Pupuh* poetics (Domain D2; n = 90) stayed trapped at 27.78% in zero-shot and 28.89% in three-shot (+1.11 points; McNemar p = 1.0000). Neutral vernacular items (Domain D4; n = 468) did not move at all, at 37.39% in both regimes. Auditing softmax probabilities exposed pervasive epistemic miscalibration. Correct zero-shot regional predictions carried higher mean maximum softmax confidence than incorrect ones (67.74% vs. 56.62%; Mann–Whitney U = 77,582.0 for 258 correct vs. 462 incorrect pairs, $p = 1.82 \times 10^{-11}$). Even so, models assigned 56.62% mean confidence (and 50.79%–56.34% in three-shot, Table 3 Panel B) to demonstrably false regional answers. In three-shot regional inference, overconfidence gaps reached +31.65 percentage points in Qwen 2 7B (70.82% mean confidence vs. 39.17% accuracy). They reached +25.37 points in Meta-Llama 3 8B (63.70% vs. 38.33%) and +20.89 points in Gemma 2 9B (73.39% vs. 52.50%).

e-ISSN 3026-7277
Vol. 4, No. 2 (2026) | pp. 85–103

Table 3. Item-Level Zero-Shot Versus Three-Shot Prompting Accuracy, Exact McNemar Paired Tests, and Softmax Calibration Across Six Contemporary 4-Bit Instruction-Tuned LLMs (1,920 Evaluations).

Model / Sociolinguistic Domain	Indonesian 0s → 3s Acc % (n = 40/LLM)	Regional 0-Shot Acc % [95% CI]	Regional 3-Shot Acc % [95% CI]	Delta (pp) & McNemar Exact p	Mean Softmax Conf. 0s / 3s (%)
Panel A: Contemporary 4-Bit Instruction-Tuned LLMs (n = 120 Regional Items / Model)	**Indonesian Pivot**	**0-Shot (n = 120)**	**3-Shot (n = 120)**	**3s − 0s (p-value)**	**Max Prob (Overconf.)**
Gemma 2 9B (Google)	70.00 → 77.50 (+7.50)	48.33 [39.59–57.18]	52.50 [43.63–61.22]	+4.17 (p = 0.4421)	73.60 / 73.39 (+20.89)
Qwen 2 7B (Alibaba)	70.00 → 60.00 (−10.00)	39.17 [30.90–48.11]	39.17 [30.90–48.11]	0.00 (p = 1.0000)	69.21 / 70.82 (+31.65)
Mistral 7B v0.3 (Mistral AI)	42.50 → 42.50 (0.00)	33.33 [25.53–42.17]	40.00 [31.68–48.94]	+6.67 (p = 0.2295)	67.83 / 53.22 (+13.22)
Meta-Llama 3 8B (Meta)	62.50 → 65.00 (+2.50)	39.17 [30.90–48.11]	38.33 [30.13–47.26]	−0.83 (p = 1.0000)	67.35 / 63.70 (+25.37)
Aya 23 8B (Cohere For AI)	62.50 → 50.00 (−12.50)	25.83 [18.84–34.33]	27.50 [20.30–36.09]	+1.67 (p = 0.8506)	39.77 / 44.52 (+17.02)
Phi 3 Mini 3.8B (Microsoft)	57.50 → 65.00 (+7.50)	29.17 [21.79–37.84]	26.67 [19.57–35.21]	−2.50 (p = 0.6776)	46.29 / 44.18 (+17.51)
Pooled 6-Model Total (720 Pairs)	60.83 → 60.00 (−0.83)	35.83 [32.41–39.40]	37.36 [33.90–40.95]	+1.53 (p = 0.4472)	60.68 / 58.30 (+20.94)
Panel B: Regional Sociolinguistic Domains (Pooled Across 6 Models; N = 720 Pairs)	**Item Count (Evals)**	**0-Shot [95% CI]**	**3-Shot [95% CI]**	**3s − 0s (McNemar p)**	**Wrong-Ans Conf. 0s/3s**
D1: Honorific Speech-Levels	15 items (n = 90)	38.89 [29.47–49.22]	46.67 [36.71–56.90]	+7.78 (p = 0.2100)	52.93 / 50.79
D2: Figurative Metaphors & Poetics	15 items (n = 90)	27.78 [19.58–37.80]	28.89 [20.54–38.96]	+1.11 (p = 1.0000)	52.22 / 54.05
D3: *Aksara*, Wayang & Rituals	12 items (n = 72)	31.94 [22.33–43.39]	36.11 [25.98–47.65]	+4.17 (p = 0.6291)	54.86 / 54.39
D4: Neutral Vernacular Lexicon	78 items (n = 468)	37.39 [33.13–41.86]	37.39 [33.13–41.86]	0.00 (p = 1.0000)	58.59 / 56.34

Note: Inter-model heterogeneity on the 120 regional items was highly significant under both 0-shot (Cochran's Q = 51.55, df = 5, $p = 6.67 \times 10^{-10}$) and 3-shot prompting (Cochran's Q = 53.60, df = 5, $p = 2.53 \times 10^{-10}$). In Panel A, parenthetical values in the rightmost column indicate 3-shot overconfidence (mean maximum softmax probability minus empirical 3-shot accuracy).

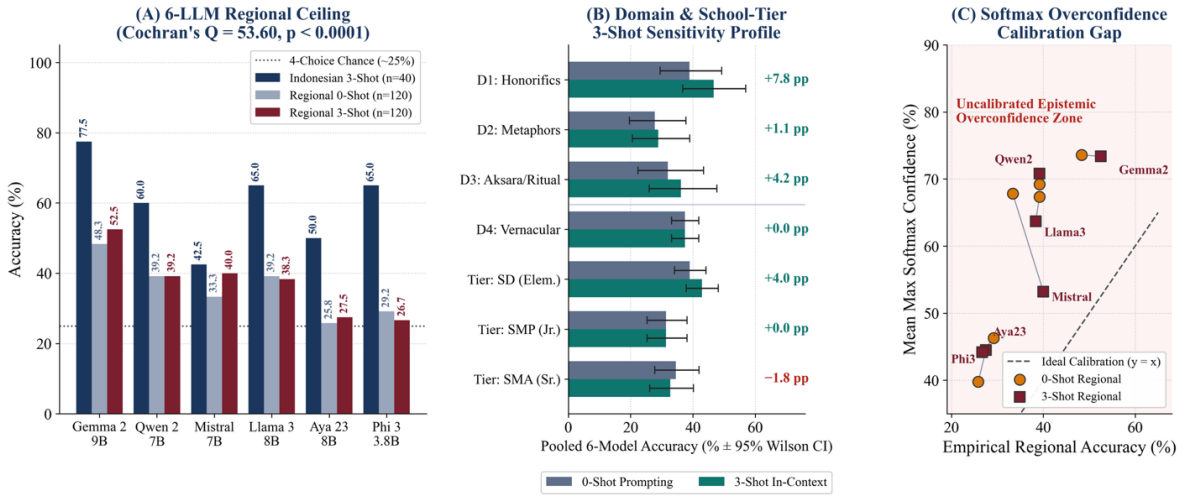

Figure 3. Item-level zero-shot versus three-shot prompting ceilings and softmax probability calibration across six contemporary 4-bit instruction-tuned LLMs (Gemma 2 9B, Qwen 2 7B, Mistral 7B v0.3, Meta-Llama 3 8B, Aya 23 8B, and Phi 3 Mini 3.8B; 1,920 evaluations): (A) three-shot Indonesian accuracy (n = 40/model) versus zero-shot and three-shot pooled regional accuracy (n = 120/model; Cochran's Q = 53.60, p < 0.0001); (B) pooled six-model zero-shot versus three-shot accuracy

(± 95% Wilson CIs) across sociolinguistic domains D1–D4 and school tiers SD–SMA, demonstrating stagnation on D2 metaphors/poetics (27.8% to 28.9%) and slight three-shot degradation on SMA upper-secondary items (−1.8 percentage points); and (C) empirical regional accuracy versus mean maximum softmax confidence, illustrating pervasive epistemic overconfidence above the y = x calibration diagonal.

3.3 Generative Collapse, Register Collisions, and Parallel-Corpus Honorific Flattening

In Stream 4A, native speakers evaluated N = 1,320 open-ended Javanese and Sundanese generations from 22 LLMs. Quality diverged sharply across model training paradigms (Kruskal–Wallis H = 346.59, df = 2, $p = 5.48 \times 10^{-76}$, $\varepsilon^2 = 0.2628$; Table 4 and Figure 4). In the full 1,320-output corpus, more than half of all generations received Grade D. This grade marks complete generative collapse, verbatim repetition of the source prompt, or Indonesian/English code leakage (54.39%, 718/1,320; 95% CI: 51.70%–57.06%). Only 25.30% (334/1,320) earned Grade A, meaning fluent and contextually appropriate output. Unaligned global multilingual baselines (Llama-2-7B/13B-Chat, BLOOMZ-7B1, and Bactrian-X) showed 82.08% Grade D collapse overall (91.67% in Javanese; 72.50% in Sundanese; mean ordinal score 1.33 ± 0.80). General Southeast Asian/Indonesian chat models without targeted regional supervision (SeaLLM-7B-Chat, SEA-LION-7B-Instruct, Merak-7B-v4, and Bactrian-ID) reached 92.92% Grade D collapse (mean score 1.11 ± 0.43). Taken together, non-Cendol models had 12.73-fold higher odds of complete generative collapse than Indonesian-aligned Cendol models (exact Fisher's OR = 12.73, $p = 4.45 \times 10^{-81}$). Cendol-aligned models lowered Grade D collapse to 35.48% (mean score 2.56 ± 1.31; Mann–Whitney $p < 0.0001$, Cliff's delta = +0.5435 vs. global baselines). Yet 39.53% of their Javanese outputs still fell into intermediate Grade B (24.29%) or Grade C (15.24%) because of lexical and sociolinguistic drift.

Stream 4B audited all N = 1,000 parallel human-translated utterances (3,000 regional clauses). The audit points to a possible structural root of generative register failure. Indonesian-pivoted parallel corpora systematically erase Javanese speech-level deference, yet they preserve Sundanese and Balinese polite lexicons at far higher rates (Cochran's Q = 82.66, df = 2, $p = 1.13 \times 10^{-18}$ on social-deixis triggers; Q = 1,102.93, $p < 10^{-200}$ overall; Table 5 and Figure 5). Some Indonesian utterances contained explicit kinship, elder, or polite service addressees (n = 81 trigger clauses, such as addressing *ibu*, *bapak*, *Anda*, or *pelanggan*). When translating these, human translators retained strict polite lexicons in 90.12% of Sundanese *Basa Lemes* clauses (73/81; flattening rate 9.88%; OR = 2.46, BH-FDR q = 0.0207). They did so in 71.60% of Balinese *Basa Alus* clauses (58/81; flattening rate 28.40%; OR = 1.57, q = 0.0933). In Javanese *Krama*/*Krama Inggil* the figure was only 20.99% (17/81; OR = 7.87 vs. general clauses [3.26%], $q = 6.3 \times 10^{-8}$). As a result, 79.01% (64/81) of Javanese clauses involving elders or formal service encounters contained none of the listed *Krama* lexicon (read here as plain *Ngoko*). So did 95.30% (953/1,000) of the overall Javanese NusaX corpus. NusaX training record 16 shows the pattern. It gives formal banking guidance addressed to an elder woman (*'Untuk memastikannya ibu bisa datang ke kantor cab bank mandiri tersebut pada hari dan jam kerja untuk penukaran uang'*). The Sundanese *Basa Lemes* rendering is deferential (*'...ibu tiasa sumping... dinten sareng waktos damel kanggo nukeurkeun artos'*), and so is the Balinese *Basa Alus* one (*'...ibu dados rauh... jam mekarya untuk penukaran jinah'*). The Javanese version is flattened into blunt *Ngoko* (*'...ibu isa teka... dina lan jam kerja kanggo ijol dhuwit'*). Read against the source sentence, it addresses a respected elder in the register of an equal or a subordinate.

Qualitative forensic inspection of Streams 3 and 4A revealed six illustrative sociolinguistic failure patterns, each documented by a single output (Table 5, Panel B). They are the kind of error one would expect when LLMs learn from noisy, register-mixed supervision, a link this study does not test. (M1) Intra-clause pronoun–verb register collision: cendol-mt5-base paired the familiar Sundanese *Basa Loma* pronoun *manehna* with the exalted *Lemes Luhur* deferential verb *ngahaturkeun* (*'manehna ngahaturkeun Carter'*). (M2) Cross-sentence deictic rank switching: cendol-mt5-xl-chat referred to the same political figure first through plain Javanese *Ngoko dheweke*, then immediately through aristocratic *Krama Inggil Panjenengane*. (M3) Inanimate vestimentary honorification: cendol-mt5-base applied the human aristocratic honorific *ngagem* (to wear royal attire or uphold sacred office) to the inanimate clinical Ebola drug control ZMapp (*'para elmuwan ngagemna minangka dasarna'*). (M4) Embodied spatial deference erasure: Gemma 2 9B misclassified Javanese passing-bow etiquette (*'Yen arep lewat ing ngarepe wong liyo kudu ngaturake...'*, IndoMMLU #12757; key: Nyuwun sewu) as permanent leave-taking (Nyuwun pamit), with 98.2% softmax confidence. (M5) Somatic *Sor Singgih* hallucination: Meta-Llama 3 8B misidentified the Balinese *Basa Alus* term for fingernail (IndoMMLU #7752; key: Naka) as Tulang (bone) at 98.0% confidence. (M6) Toponymic cosmology flattening: Meta-Llama 3 8B reduced a Sundanese ancestral place-prophecy (*Cacandran*, IndoMMLU #13940) to generic *Babasan* proverbs at 96.7% confidence.

Table 4. Native-Speaker Human Evaluation of Open-Ended Javanese and Sundanese Generations Across 22 Large Language Models in Cendol (N = 1,320 Outputs).

Architecture Family / Stratum	Sample (Models)	Grade A: Fluent (%)	Grade B: Drift (%)	Grade C: Severe (%)	Grade D: Collapse % [95% Wilson CI]	Ordinal Score (1–4) Mean ± SD [Boot CI]
Indonesian-Aligned Cendol — All	n = 840 (14)	37.86%	15.83%	10.83%	35.48% [32.31–38.77]	2.56 ± 1.31 [2.47–2.65]
Javanese Subset	n = 420 (14)	26.43%	24.29%	15.24%	34.05% [29.68–38.71]	2.43 ± 1.21 [2.32–2.55]
Sundanese Subset	n = 420 (14)	49.29%	7.38%	6.43%	36.90% [32.43–41.62]	2.69 ± 1.39 [2.56–2.82]
Global Multilingual Baseline — All	n = 240 (4)	5.83%	3.33%	8.75%	82.08% [76.74–86.42]	1.33 ± 0.80 [1.23–1.43]
Javanese Subset	n = 120 (4)	1.67%	2.50%	4.17%	91.67% [85.35–95.41]	1.14 ± 0.52 [1.06–1.24]

Architecture Family / Stratum	Sample (Models)	Grade A: Fluent (%)	Grade B: Drift (%)	Grade C: Severe (%)	Grade D: Collapse % [95% Wilson CI]	Ordinal Score (1–4) Mean ± SD [Boot CI]
Sundanese Subset	n = 120 (4)	10.00%	4.17%	13.33%	72.50% [63.91–79.70]	1.52 ± 0.97 [1.35–1.69]
SEA & Indonesian Chat — All	n = 240 (4)	0.83%	2.08%	4.17%	92.92% [88.95–95.53]	1.11 ± 0.43 [1.06–1.17]
Javanese Subset	n = 120 (4)	0.00%	3.33%	5.00%	91.67% [85.35–95.41]	1.12 ± 0.41 [1.05–1.19]
Sundanese Subset	n = 120 (4)	1.67%	0.83%	3.33%	94.17% [88.45–97.15]	1.10 ± 0.46 [1.02–1.19]
Total Pooled Generative Corpus	N = 1,320 (22)	25.30%	11.06%	9.24%	54.39% [51.70–57.06]	2.07 ± 1.29 [2.00–2.14]

Note: Native speaker human evaluation grades from Cendol (A = fluent and semantically/pragmatically faithful [4]; B = minor lexical or register drift [3]; C = severe distortion or hallucinated additions [2]; D = complete collapse, untranslated source repetition, or Indonesian/English leakage [1]). Omnibus Kruskal–Wallis test across the three model families: H = 346.59 (df = 2, $p = 5.48 \times 10^{-76}$, $\varepsilon^2 = 0.2628$). Non-Cendol models exhibited an exact Fisher's Odds Ratio of OR = 12.73 ($p = 4.45 \times 10^{-81}$) for Grade D collapse compared with Cendol-aligned models.

Table 5. Parallel Corpus Honorific Retention Audit Across NusaX (N = 3,000 Regional Clauses) and Authentic Qualitative Sociolinguistic Error Taxonomy.

Language / Error Mechanism	Overall Corpus (N = 1,000 / Lang)	Social/Kinship Triggers (n = 81 Clauses)	General Clauses (n = 919 Clauses)	Trigger Flattening & Fisher OR (BH q)	Verbatim Empirical Exemplar (Source & Observed Text)
Panel A: Parallel Speech-Level Encoding in NusaX Human Translations	**Strict Honorific %**	**Honorific % [95% CI]**	**Honorific % [95% CI]**	**Flattening % | OR**	**Parallel Utterance Comparison (NusaX Train Row 16)**
Javanese (*Krama / Krama Inggil*)	4.70% [3.55–6.19]	20.99% [13.53–31.06]	3.26% [2.30–4.62]	79.01% | OR = 7.87 ($q = 6.3 \times 10^{-8}$)	IND: *'...ibu bisa datang ke kantor cab bank mandiri... penukaran uang.'* → JAV (*Ngoko*): *'...ibu isa teka... dina lan jam kerja kanggo ijol dhuwit.'*
Balinese (*Basa Alus*)	62.40% [59.35–65.35]	71.60% [60.99–80.27]	61.59% [58.40–64.68]	28.40% | OR = 1.57 (q = 0.0933)	BAN (Alus): *'Anggen mastiang ibu dados rauh ka kantor cabang... jam mekarya untuk penukaran jinah.'*
Sundanese (*Basa Lemes*)	79.70% [77.09–82.08]	90.12% [81.71–94.91]	78.78% [76.02–81.30]	9.88% | OR = 2.46 (q = 0.0207)	SUN (*Lemes*): *'Kanggo mastikeun ibu tiasa sumping ka kantor... dinten sareng waktos damel kanggo nukeurkeun artos.'*
Panel B: Qualitative Sociolinguistic & Pragmatic Failure Taxonomy Across Audited LLMs	**Source Stream**	**Affected Language**	**Model & Confidence / Grade**	**Pragmatic Mechanism**	**Verbatim Prompt / Output & Native Sociolinguistic Diagnosis**
M1. Intra-Clause Pronoun–Verb Register Collision	Stream 4A (Cendol)	Sundanese	cendol-mt5-base (Grade B; Rater Note)	Loma pronoun + *Lemes Luhur* verb	Output: *'Salila seleksi taun 1976, manéhna ngahaturkeun Carter...'* — Collides egalitarian/familiar pronoun *'manéhna'* with deferential verb *'ngahaturkeun'*.
M2. Cross-Sentence Deictic Rank Switching	Stream 4A (Cendol)	Javanese	cendol-mt5-xl-chat (Grade D)	*Ngoko* pronoun → *Krama Inggil* pronoun	Output: *'...dhèwèké dadi Advisor keamanan nasional... Panjenengané kapilih...'* — Shifts referent mid-passage from plain *'dhèwèké'* to exalted *'Panjenengané'*.
M3. Inanimate Vestimentary Honorification	Stream 4A (Cendol)	Sundanese	cendol-mt5-base (Grade B; Rater Note)	Aristocratic attire verb on drug control	Output: *'...Zmapp dipakena minangka kontrol, anu hartina para élmuwan ngagemna...'* — Applies human aristocratic honorific *'ngagem'* (to wear regalia/uphold) to clinical drug ZMapp.
M4. Embodied Spatial Deference Erasure	Stream 3 (IndoMMLU #12757)	Javanese (D1)	Gemma 2 9B (p0 = 0.982; Wrong)	Spatial deference → permanent departure	Stem: *'Yen arep lewat ing ngarepe wong liyo kudu ngaturake...'* Key: *'B. Nyuwun sewu'* (passing bow). Predicted: *'C. Nyuwun pamit'* (taking leave) at 98.2% confidence.
M5. Somatic *Sor Singgih* Hallucination	Stream 3 (IndoMMLU #7752)	Balinese (D1)	Llama 3 8B (p0 = 0.980; Wrong)	Sanskritic somatic register erasure	Stem: *'Kukune putih, Kuku bahasa alusne...'* Key: *'B. Naka'* (Old Javanese/Sanskritic *Basa Alus*).

e-ISSN 3026-7277
Vol. 4, No. 2 (2026) | pp. 85–103

Language / Error Mechanism	Overall Corpus (N = 1,000 / Lang)	Social/Kinship Triggers (n = 81 Clauses)	General Clauses (n = 919 Clauses)	Trigger Flattening & Fisher OR (BH q)	Verbatim Empirical Exemplar (Source & Observed Text)
					Predicted: *'A. Tulang'* (bone) at 98.0% softmax confidence.
M6. Toponymic Cosmology Flattening	Stream 3 (IndoMMLU #13940)	Sundanese (D2)	Llama 3 8B (p0 = 0.967; Wrong)	Ancestral *Cacandran* → generic proverb	Stem: *'Pakeman basa anu eusina mangrupa tujuman... kaayaan hiji daerah...'* Key: *'A. Cacandran'* (ancestral place-prophecy). Predicted: *'D. Babasan'* at 96.7% confidence.

Note: In Panel A, Cochran's Q across the three parallel translations equaled Q = 82.66 (df = 2, p = 1.13×10^{-18}) on the 81 social/kinship deixis utterances and Q = 1,102.93 (df = 2, p < 10^{-200}) across all 1,000 utterances. BH q denotes Benjamini–Hochberg false-discovery-rate adjusted p-values.

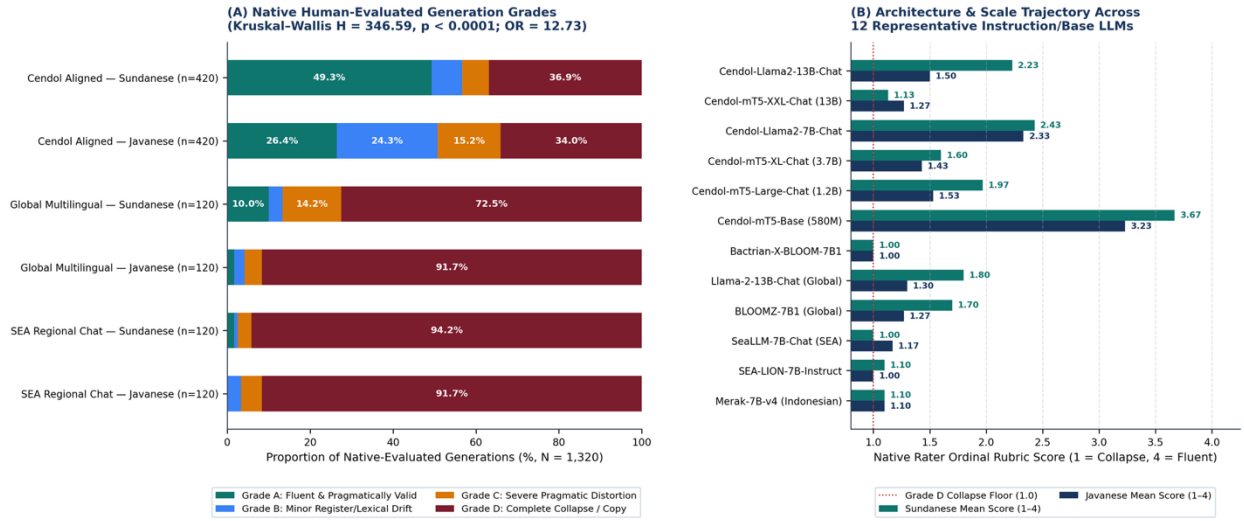

Figure 4. Native-speaker human evaluation of open-ended Javanese (n = 660) and Sundanese (n = 660) generations across 22 LLMs in Cendol (N = 1,320 outputs): (A) stacked percentage distributions of native rater grades (Grade A: fluent and pragmatically faithful; Grade B: minor lexical or register drift; Grade C: severe pragmatic/semantic distortion; Grade D: complete generative collapse or untranslated source copy) across Indonesian-aligned Cendol, global multilingual baseline, and Southeast Asian regional chat families (Kruskal–Wallis H = 346.59, p < 0.0001; collapse Odds Ratio = 12.73); and (B) mean ordinal human scores (1–4 scale) across 12 representative architectures, contrasting decoder-only Cendol-Llama2 and encoder–decoder Cendol-mT5 trajectories.

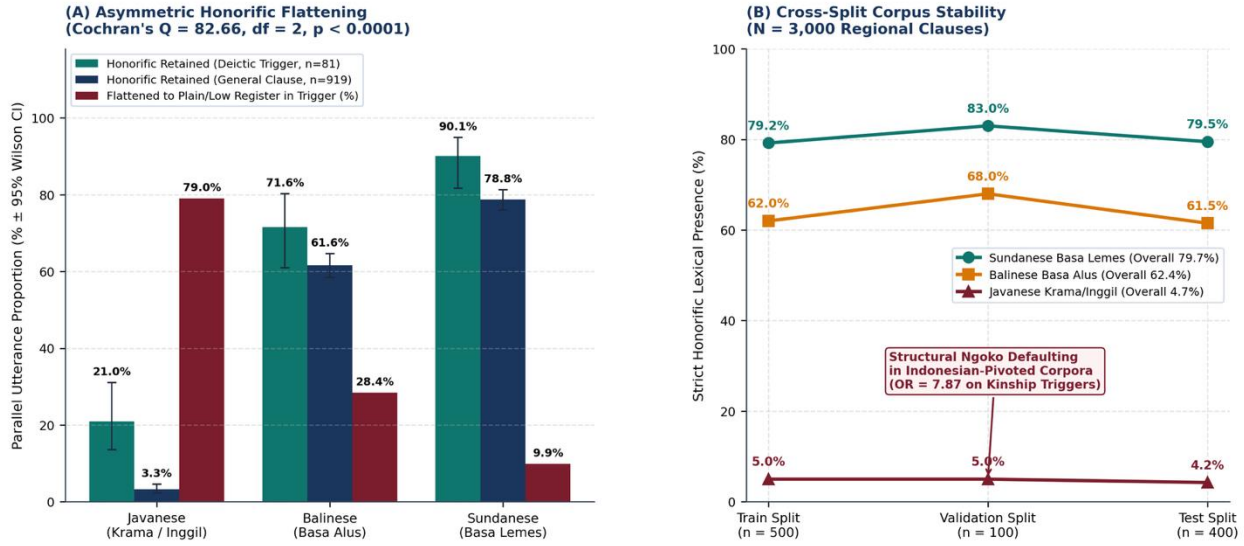

Figure 5. Asymmetric speech-level retention and source-pivot honorific flattening across N = 1,000 parallel human-translated utterances (3,000 regional clauses) in NusaX: (A) strict honorific lexical presence (± 95% Wilson CIs) in social/kinship deixis trigger utterances (n = 81) versus general utterances (n = 919) alongside plain/low-register flattening rates on trigger utterances—demonstrating 79.0% Javanese *Ngoko* flattening compared with 28.4% in Balinese and 9.9% in Sundanese (Cochran's Q = 82.66, df = 2, p < 0.0001); and (B) cross-split corpus stability across training (n = 500), validation (n = 100), and testing (n = 400) partitions.

3.4 Multi-Stream Sensitivity and Robustness Across Educational Tiers, Backbones, and Data Splits

Sensitivity analyses covered pedagogical school tiers, neural backbone architectures, and corpus partitions. They supported the stability of these findings (Table 6). In Stream 3, we stratified the 720 regional evaluations by numerical grade tier. Pooled six-model accuracy was not monotonic across tiers. Primary School items scored 38.98% zero-shot and 42.94% three-shot (SD; +3.95 points, McNemar p = 0.1606). Lower Secondary items stagnated at 31.31% in both regimes (SMP). Upper Secondary items declined from 34.52% zero-shot to 32.74% three-shot (SMA; −1.79 points, p = 0.7428). Zero-shot accuracy did not differ significantly across school tiers (chi-square = 3.41, df = 2, p = 0.1816). Three-shot prompting then selectively helped elementary lexical items and mildly degraded complex high-school *Macapat*/*Kakawin* and discourse etiquette items. That widened the gap between tiers to statistical significance (chi-square = 9.33, df = 2, p = 0.0094).

Stream 4A included 14 Indonesian-aligned Cendol models. A backbone sensitivity analysis showed that decoder-only Cendol-Llama2 architectures (7B–13B; n = 240 outputs) reached significantly higher native-speaker fluency than encoder–decoder Cendol-mT5 architectures (300M–13B; n = 600 outputs). The decoder-only group attained 48.33% Grade A fluency and 27.92% Grade D collapse (mean ordinal score 2.84 ± 1.30; 95% bootstrap CI: 2.68–3.00). Cendol-mT5 reached 33.67% Grade A fluency and 38.50% Grade D collapse (mean 2.45 ± 1.30; 95% CI: 2.35–2.55; Mann–Whitney U = 60,491.0, p = 0.0001, Cliff's delta = +0.1599). Across both backbones, Sundanese generations significantly outperformed Javanese generations (49.29% vs. 26.43% Grade A; mean 2.69 vs. 2.43; p = 0.0029, Cliff's delta = +0.1128). This directly mirrors the higher register consistency of Sundanese training data observed in Stream 4B. Partition-level sensitivity across NusaX training (n = 500), validation (n = 100), and testing (n = 400) splits then confirmed strict rank-order invariance. Javanese *Krama* lexical presence stayed capped at 5.00%, 5.00%, and 4.25% across the three splits. The corresponding figures were 62.00%, 68.00%, and 61.50% for Balinese *Basa Alus*, and 79.20%, 83.00%, and 79.50% for Sundanese *Basa Lemes*.

Table 6. Multi-Stream Sensitivity and Robustness Analyses Across Pedagogical School Tiers, Model Backbone Families, and Parallel Corpus Splits.

Sensitivity Dimension / Stratum	Analytical Sample Size	Primary Metric / 0-Shot	Comparator / 3-Shot	Effect Size & Robustness Test
Panel A: Educational Tier Complexity in Stream 3 (Pooled 6 LLMs)	**Unique Items (Model Evals)**	**0-Shot Accuracy [95% CI]**	**3-Shot Accuracy [95% CI]**	**Delta & Tier Contingency Test**
Primary School (SD: Grades 1–6)	59 items (n = 354)	38.98% [34.05–44.16]	42.94% [37.88–48.14]	+3.95 pp (McNemar p = 0.1606)
Lower Secondary (SMP: Grades 7–9)	33 items (n = 198)	31.31% [25.26–38.08]	31.31% [25.26–38.08]	0.00 pp (McNemar p = 1.0000)
Upper Secondary (SMA: Grades 10–12)	28 items (n = 168)	34.52% [27.75–41.99]	32.74% [26.10–40.15]	−1.79 pp; 3-shot tier χ^2 = 9.33 (p = 0.0094)
Panel B: Backbone Architecture Within Aligned Cendol Suite (Stream 4A)	**Evaluated Generations**	**Grade A Fluent / Grade D Collapse**	**Ordinal Mean ± SD [Boot CI]**	**Mann–Whitney U & Cliff's Delta**
Decoder-Only (Cendol-Llama2 7B–13B)	n = 240 (4 models)	48.33% Fluent / 27.92% Collapse	2.84 ± 1.30 [2.68–3.00]	Outperforms mT5: U = 60,491
Encoder–Decoder (Cendol-mT5 300M–13B)	n = 600 (10 models)	33.67% Fluent / 38.50% Collapse	2.45 ± 1.30 [2.35–2.55]	p = 0.0001; Cliff's d = +0.1599
Language Asymmetry Within Cendol	Sun n = 420 vs. Jav n = 420	Sun 49.29% A vs. Jav 26.43% A	Sun 2.69 vs. Jav 2.43	p = 0.0029; Cliff's d = +0.1128
Panel C: Cross-Split Stability of Parallel Honorific Encoding (Stream 4B)	**NusaX Partition**	**Javanese *Krama* %**	**Balinese Alus % / Sundanese *Lemes* %**	**Rank-Order Invariance**
Training Split	n = 500 parallel triplets	5.00%	62.00% / 79.20%	Sundanese > Balinese >> Javanese
Validation Split	n = 100 parallel triplets	5.00%	68.00% / 83.00%	Sundanese > Balinese >> Javanese
Testing Split	n = 400 parallel triplets	4.25%	61.50% / 79.50%	Sundanese > Balinese >> Javanese

Note: Bootstrap 95% confidence intervals in Panel B were computed via 4,000 non-parametric resamples (fixed seed 202606). Across all three NusaX splits in Panel C, Javanese *Krama* lexical presence remained <= 5.00%, confirming systematic corpus-wide *Ngoko* defaulting.

e-ISSN 3026-7277
Vol. 4, No. 2 (2026) | pp. 85–103

4. Discussion

4.1 Source-Pivot Pragmatic Flattening and the Semiotic Fragility of Speech Levels in Latent Space

The four empirical streams, read together, advance cultural and sociolinguistic theory on how foundation models transform non-Western communicative traditions. The central finding is simple. LLM failure in Javanese, Sundanese, and Balinese is not a uniform low-resource vocabulary deficit. It is a structured semiotic breakdown. The breakdown sits in relational speech-level indexicality (*Undha-Usuk*, *Undak-Usuk*, and *Sor Singgih*) and in prosodic-figurative cosmology.[11,17] Across 19 models in Stream 2, regional performance dropped by 10.91 to 18.36 percentage points relative to Indonesian, and even frontier GPT-4 fell to 27.67% and 13.17% on Javanese and Sundanese figurative metaphors. In Stream 3, three in-context regional exemplars produced no statistically significant aggregate improvement (+1.53 points; McNemar p = 0.4472), and classical metaphors and *Macapat*/*Pupuh* prosody stayed stagnant at 28.89%. In Errington's semiotic terms,[11–13] speech-level selection is a triadic relational calculus that links speaker, addressee, and referent across social space. Models handle this differently. Subword byte-pair encoding tokenizers and next-token cross-entropy objectives treat suppletive honorifics as interchangeable distributional synonyms of English or Indonesian lemmas.

Stream 4B suggests that this algorithmic flattening does not begin inside neural model weights alone. It is already visible in the human-translated parallel corpora used to train and benchmark Indonesian regional NLP.[30,34] Modern Indonesian (*Bahasa Indonesia*, standardized from Malay) deliberately dropped compulsory grammaticalized speech levels, in order to foster an egalitarian national civic identity, so a single Indonesian sentence such as *'Ibu pernah bekerja di Grab Indonesia'* is pragmatically underspecified with respect to Javanese *Undha-Usuk*. When crowd or freelance translators render isolated Indonesian sentences without an explicit conversational persona or addressee directive, Javanese translators overwhelmingly default to unmarked *Ngoko* (95.30% overall; 79.01% even when kinship/elder triggers appear). Sundanese and Balinese translators conventionally adopt polite *Basa Lemes* (79.70% overall; 90.12% on triggers) and *Basa Alus* (62.40% overall; 71.60% on triggers) as default written standards. Models instruction-tuned on Indonesian-pivoted corpora may thus inherit both Javanese *Ngoko* flattening and cross-register lexical contamination. Cendol's results fit this reading, although no training experiment tests it. Its Javanese Grade A fluency (26.43%) lagged nearly 23 percentage points behind Sundanese (49.29%), while accumulating 39.53% Grade B/C register and semantic drift.

Stream 4A compares architectures within the Indonesian-aligned Cendol suite. The comparison shows how neural backbone topology interacts with regional instruction tuning.[41,42] Across 840 native-rated outputs, decoder-only Cendol-Llama2 checkpoints achieved much higher Grade A fluency (48.33% vs. 33.67%) and lower Grade D collapse (27.92% vs. 38.50%; Cliff's delta = +0.1599, p = 0.0001) than encoder–decoder Cendol-mT5 checkpoints. Scaling within conversational '-chat' variants was non-monotonic. For instance, base Cendol-mT5-Base achieved higher task translation fidelity than conversational Cendol-mT5-XL-Chat and XXL-Chat on direct prompt-translation instructions. In autoregressive chat fine-tuning dominated by Indonesian multi-turn assistant dialogues, safety and conversational formatting templates can inadvertently override regional target-language generation constraints. Chat-tuned checkpoints then echo the source sentence verbatim or answer in national Indonesian (native raters gave Grade D notes of 'sama' [identical to input]). When models do generate regional text, cross-sentence deictic switching appears, such as a shift from *Ngoko 'dheweke'* to aristocratic *'Panjenengane'*, which shows that token-level next-word prediction lacks a persistent discourse-state register variable governing interpersonal rank across multi-sentence spans.

4.2 Uncalibrated Epistemic Overconfidence and Cultural Heritage Risks in Educational Deployment

From the perspective of Fricker's hermeneutical injustice and critical heritage studies,[24–28] the Stream 3 findings pose tangible risks for regional education and public culture. Cultural accuracy was low, while softmax confidence was high. Across the six contemporary instruction-tuned LLMs, incorrect regional predictions carried a mean maximum softmax probability of 56.62% in zero-shot inference. Extreme high-confidence hallucinations (>85%–98% probability) occurred precisely on items encoding embodied etiquette and ancestral cosmology. Examples include the conflation of the Javanese spatial deference formula *nyuwun sewu* with permanent departure *nyuwun pamit* (98.2% confidence), the replacement of the Balinese Sanskritic somatic honorific *naka* with *tulang* (98.0% confidence), and the reduction of Sundanese toponymic *Cacandran* prophecies to generic proverbs (96.7% confidence). Complete Grade D collapse in unaligned models is easy to spot. Native speakers recognize and reject it at once, because the output repeats the prompt or switches into English. High-confidence, grammatically fluent Grade B register collisions are different, such as pairing *Basa Loma manehna* with *Lemes Luhur ngahaturkeun* or attributing aristocratic *ngagem* to a pharmaceutical compound. These present insidious pedagogical hazards.[15–17]

Younger urban cohorts in Java, West Java, and Bali already report intergenerational anxiety. It arises when they attempt *Krama Inggil*, *Basa Lemes*, or *Basa Alus* with elders and customary leaders. Locally the fear of speaking incorrectly is called *wedi kleru* in Javanese and *sieun salah* in Sundanese.[44,45] If students, teachers, or cultural content creators consult uncalibrated chatbots that present flattened *Ngoko* or hybrid Loma–*Lemes* collisions with authoritative fluency, generative AI systems risk accelerating the erosion of asymmetric humiliative–honorific distinctions. They also risk decoupling classical *Macapat*, *Pupuh*, and *Kakawin* meters from their ethical and liturgical meanings.[21–23]

The cultural stakes of these errors become clear when they are placed within the lived anthropology of embodied etiquette, somatic hierarchy, and prosodic memory across Java, Pasundan, and Bali.[11] In Javanese everyday life, a speaker says *'nyuwun sewu'* while inclining the upper torso and extending the right hand downward when passing elders, a gesture that is a foundational bodily hexis of social harmony (*rukun*) and humble self-restraint (*andhap asor*). Replacing it with *'nyuwun pamit'* (announcing permanent departure from a host's home) or *'nyuwun ngapura'* (confessing moral fault) collapses distinct social rituals into an undifferentiated English 'excuse me'.[11,15] Balinese *Sor Singgih Basa* works in a similar way. It elevates the human body through *Kawi*–Sanskritic somatic lexemes, such as *'naka'* for fingernail and *'waduk'* for abdomen, and so links interpersonal speech etiquette directly to Hindu-Balinese

microcosmic theology (*Bhuana Alit* mirroring *Bhuana Agung*). Hallucinating *'tulang'* (bone) at 98.0% softmax confidence severs this lexical-cosmological chain. Classical literature carries the same load. Javanese *Tembang Macapat* stanzas, such as Mangkunegara IV's *Serat Wedhatama* (*'Ngelmu iku kalakone kanthi laku'*, composed in *Pocung* meter), and Sundanese *Cacandran* toponymic prophecies encode ethical and ecological memory through strict acoustic-prosodic constraints. When uncalibrated LLMs misclassify meters, or conflate sacred place-prophecies with everyday idioms at >85%–96% confidence, they erode the formal interpretive frameworks sustained by local teachers and cultural practitioners.

Studies of language shift among young Javanese speakers report growing intergenerational insecurity about speech levels, termed *'wedi kleru'* in Javanese and *'sieun salah'* in Sundanese.[44,45] Children and adolescents avoid speaking regional languages to parents, teachers, and elders, out of anxiety over misapplying *Krama Inggil*, *Basa Lemes*, or *Basa Alus* predicates. The result is an accelerating domestic shift toward monolingual Indonesian. Schools are adding digital learning assistants. Deploying unaligned or source-flattened LLMs there creates an acute pedagogical feedback loop. Learners seeking polite regional phrasing meet authoritative-sounding outputs that address mothers in blunt *Ngoko* (*'ibu isa teka'*) or fuse familiar Loma pronouns with aristocratic verbs (*'manehna ngahaturkeun'*). Without explicit register calibration and custodial verification, generative AI risks institutionalizing pragmatic flattening as the de facto digital norm for the next generation of regional speakers.[15,18,28]

4.3 Toward Decolonial Sociolinguistic Alignment: The Proposed CARE-Pragmatics Framework

Addressing pragmatic flattening requires more than surface prompt engineering. It calls for structural reforms in corpus architecture, preference optimization, and community custodianship.[46–48] We build on Sennrich et al.'s side-constraint politeness conditioning,[46] Rafailov et al.'s Direct Preference Optimization (DPO),[47] and Carroll et al.'s Global Indigenous Data Alliance (GIDA) CARE Principles for Indigenous Data Governance.[48] On that basis we propose the CARE-Pragmatics Alignment Framework, a prospective governance and engineering roadmap for Nusantara language technologies (Table 7 and Figure 6). It has four parts. First, under Collective Benefit, single-reference Indonesian-pivoted parallel corpora should give way to multi-register parallel tuples, in which each source utterance is paired with parallel *Ngoko*/*Krama*, Loma/*Lemes*, and *Kepara*/Alus realizations, explicitly tagged for interpersonal symmetry or deference.

Second, under Authority to Control, annotation rubrics, morphological lexicons, and release licenses should be co-governed with accredited local cultural custodians, including Yogyakarta and Surakarta *Keraton* philologists, provincial *Balai Bahasa* centers, *Lembaga Basa jeung Sastra Sunda* (LBSS), and Balinese *Pasraman* and *Penyuluh Bahasa Bali* networks. Such co-governance keeps authority over normative etiquette boundaries with indigenous communities.[20,21,48] Third, under Responsibility, instruction tuning should incorporate explicit dyadic role metadata (<Speaker, Addressee, Referent>). It should also include DPO preference pairs that penalize self-honorification (e.g., rejecting *'kula dhahar'* in favor of humiliative *'kula nedha'*), intra-clause pronoun–verb collisions, and inanimate honorification. Temperature or abstention calibration should accompany both when cultural priors are uncertain.[46,47] Fourth, under Ethics, models queried on sacred liturgical formulas (*Yadnya* mantras, *Kidung*, *Kakawin*, *Serat Wedhatama*, and ancestral *Cacandran*) should enforce prosodic verification. They should cite primary philological sources rather than fabricate pseudo-traditional verses.[21,22]

Embedding the CARE-Pragmatics framework within Indonesia's statutory cultural and educational policy architecture offers an institutional pathway from diagnosis to implementation (Table 7 and Figure 6).[48] Nationally, Law No. 5 of 2017 on the Advancement of Culture (*Undang-Undang Nomor 5 Tahun 2017 tentang Pemajuan Kebudayaan*) mandates active state protection, development, utilization, and community stewardship of ten statutory objects of cultural advancement (*Objek Pemajuan Kebudayaan*).[49] These objects explicitly encompass oral traditions, manuscripts, customs, rites, traditional knowledge, and regional languages. Two more instruments add to this mandate. The Ministry of Education, Culture, Research, and Technology's Regional Language Revitalization initiative (*Revitalisasi Bahasa Daerah*) and provincial local-content curriculum decrees (*Muatan Lokal Bahasa Jawa*, *Bahasa Sunda*, and *Bahasa Bali*) institutionalize speech-level pedagogy and *Aksara* literacy across tens of thousands of Indonesian schools.

Operationalizing CARE-Pragmatics allows *Badan Pengembangan dan Pembinaan Bahasa*, provincial *Balai Bahasa* offices, *Keraton* philological scriptoria, Sundanese LBSS scholars, and Balinese *Penyuluh Bahasa Bali* to establish sovereign, community-governed reference benchmarks. These benchmarks would exist before commercial or open-weights LLMs are certified for classroom or public-service deployment.[21,48]

For future custodial corpus construction and post-DPO model certification, we recommend a standardized four-dimensional native-expert annotation protocol. It would be evaluated across stratified multi-turn and single-turn prompts, to secure psychometric rigor. The four dimensions are: (1) Addressee Deference Fidelity (verifying *Krama Alus*, *Basa Lemes keur Batur*, and *Basa Alus Singgih* selection on asymmetric elder, teacher, and civic-service interactions); (2) Humiliative Self-Reference Integrity (verifying strict absence of self-honorification, ensuring speakers apply *Krama Andhap*, *Lemes* keur Sorangan, and Alus Sor to themselves); (3) Animacy and Ontological Congruence (preventing human aristocratic somatic or vestimentary honorifics such as *'ngagem'* from attaching to inanimate objects or non-human entities); and (4) Prosodic and Cosmological Provenance (verifying exact syllabic *'guru wilangan'*, terminal vowel *'guru lagu'*, and primary manuscript attribution on *Macapat*, *Pupuh*, *Kidung*, *Kakawin*, and *Cacandran* queries).[11,19,22,46–48] Certification should be multi-rater. It should involve independent philologists from each language community. The target is inter-annotator reliability of Fleiss' kappa ≥ 0.80 and Krippendorff's ordinal alpha ≥ 0.80, alongside post-alignment Expected Calibration Error (ECE) ≤ 0.08. Such certification could make indigenous data sovereignty an auditable computational standard. These thresholds are proposed targets and have no empirical basis in the present study.[47,48]

e-ISSN 3026-7277
Vol. 4, No. 2 (2026) | pp. 85–103

Table 7. Proposed CARE-Pragmatics Decolonial Governance and Sociolinguistic Alignment Protocol for Indonesian Regional Language Technologies.

CARE Principle & Cultural Dimension	Empirical Vulnerability Addressed	Institutional Custodians & Co-Design Actors	Technical Curation & Alignment Specification	Prospective Verification Metric
C — Collective Benefit (Multi-Register Corpus Equity)	79.01% Javanese *Krama* erasure on kinship triggers in single-reference Indonesian-pivoted corpora (Stream 4B).	Regional public educators, provincial *Balai Bahasa* offices, and open-science NLP consortia.	Replace register-neutral Indonesian 1-to-1 translations with multi-register parallel tuples explicitly pairing *Ngoko*/*Krama*, Loma/*Lemes*, and *Kepara*/Alus realizations.	Deictic Honorific Retention Rate ≥ 90% on asymmetric elder, kinship, and public-service test suites.
A — Authority to Control (Indigenous Custodial Sovereignty)	Unvalidated crowd-translated rubrics and Anglocentric politeness priors flattening local etiquette hierarchies (Streams 1–2).	Yogyakarta/Surakarta *Keraton* philologists, *Lembaga Basa jeung Sastra Sunda* (LBSS), and Balinese *Pasraman* / *Penyuluh Bahasa Bali*.	Establish community-ratified sociolinguistic annotation guidelines and custodial licensing agreements prior to instruction tuning or commercial API deployment.	Inter-annotator agreement (Fleiss' kappa ≥ 0.80) certified by accredited regional language boards.
R — Responsibility (Dyadic Alignment & Calibration)	Intra-clause register collisions (M1–M3) and 56.62% mean softmax confidence on incorrect cultural predictions (Stream 3).	LLM alignment engineers, computational sociolinguists, and regional curriculum evaluation panels.	Condition prompts on explicit dyadic speaker–addressee–referent tags <S, A, R> and apply Direct Preference Optimization (DPO) penalizing self-honorification and animacy errors.	Zero self-honorification (*'kula dhahar'* → *'kula nedha'*) and Expected Calibration Error (ECE) ≤ 0.08.
E — Ethics (Ritual & Poetic Safeguards)	Near-chance accuracy on D2 classical poetics/metaphors (27.78%–28.89%) and 96.7%–98.0% confident hallucinations (M5–M6).	Customary *Adat* councils, Hindu Dharma liturgical scholars (Sulinggih/Pemangku), and *Macapat*/Tembang Sunda masters.	Implement prosodic-meter verification (*guru wilangan* / *guru lagu* / *padalingsa*) and calibrated abstention with primary philological source attribution on sacred texts.	Hallucination rate < 5% on sacred liturgical/cosmological queries with verifiable textual provenance.

Note: Adapted from the Global Indigenous Data Alliance (GIDA) CARE Principles (Carroll et al., 2020) and Javanese/Balinese/Sundanese semiotic etiquette theory (Errington, 1988; Clynes, 1994). Thresholds represent proposed prospective audit targets rather than post-intervention claims.

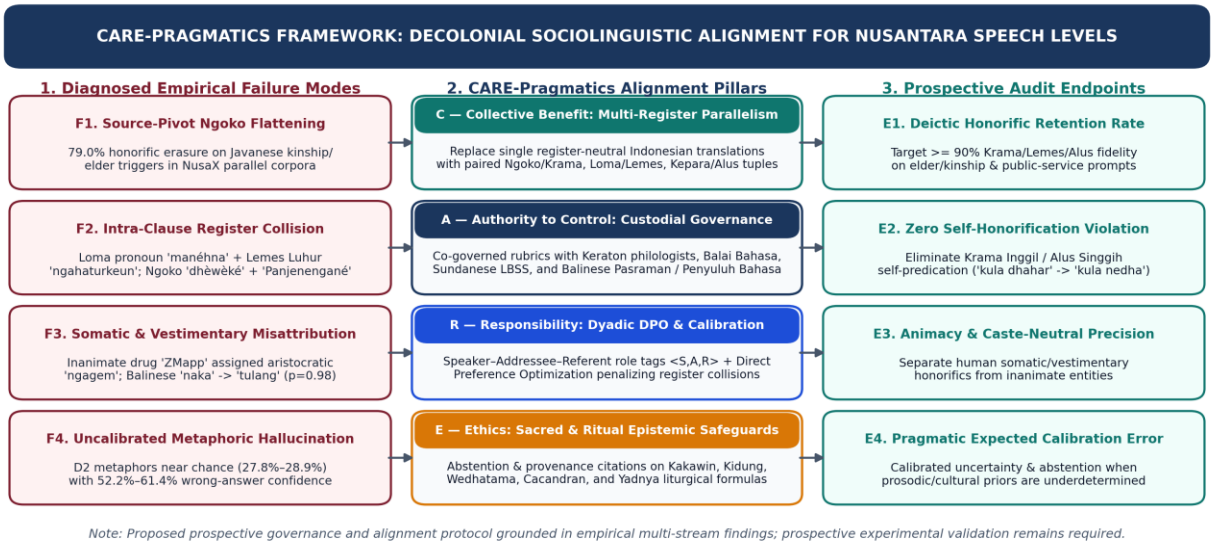

Figure 6. Proposed CARE-Pragmatics Decolonial Sociolinguistic Alignment Framework for Nusantara Speech-Level Traditions: schematic architecture linking the four empirically diagnosed failure modes (F1 source-pivot *Ngoko* flattening, F2 intra-clause register collision, F3 somatic and vestimentary misattribution, and F4 uncalibrated metaphoric hallucination) to four GIDA CARE governance and preference-optimization pillars (Collective Benefit, Authority to Control, Responsibility, and Ethics) and prospective custodial verification endpoints.

4.4 Methodological Strengths, Limitations, and Epistemic Humility

A primary strength of this investigation is the triangulation of four independent, human-grounded empirical streams: 2,620 regional curriculum items, 19 macro-benchmarked foundation models, 1,920 item-level softmax runs, 1,320 native-rated open-ended generations, and 3,000 parallel human translations. No synthetic data or unvalidated LLM-as-a-judge proxies were used. Several limitations still bound the interpretation of our findings. First, this is a secondary computational and sociolinguistic audit of publicly released benchmarks, so our findings reflect written multiple-choice, translation, and single-turn instruction settings and do not rest on longitudinal ethnographic observation of live multi-turn oral interaction in Javanese, Sundanese, or Balinese households and ritual spaces. Second, in Stream 3, the six contemporary

e-ISSN 3026-7277
Vol. 4, No. 2 (2026) | pp. 85–103

instruction-tuned models were evaluated under 4-bit NormalFloat quantization on a stratified 40-item-per-language subset (n = 120 regional items). The 4-bit setting reflects realistic local deployment on consumer hardware in Indonesian schools. Full-precision 16-bit weights and larger per-domain sample sizes could still yield modestly higher absolute accuracy, particularly in 3.8B–9B parameter regimes.

Third, sociolinguistic domain classification (D1–D4) in Streams 1 and 3 and strict honorific detection in Stream 4B relied on deterministic, philologically curated regular-expression lexicons. Rule-based matching guarantees 100% transparency and replicability. It also avoids circular neural classifier bias. Lexical matching cannot, however, capture every context-dependent prosodic nuance, zero-morpheme implicature, or regional dialectal variant (such as Banyumasan Javanese, Cirebon–Indramayu Javanese, Priangan vs. Banten Sundanese, or Bali Aga upland dialects). Fourth, Cendol's native-speaker human evaluation suite (Stream 4A) covered Javanese and Sundanese across 22 open-weights LLMs up to 13B parameters. It did not include human-graded open-ended Balinese generations or proprietary reasoning models released in 2025–2026. Fifth, the six models were scored on the same 120 regional items, so the 720 pooled model–item pairs are not independent, and the intervals and tests that treat them as independent are optimistic. Sixth, the Indonesian and regional IndoMMLU items are different examinations, so the Stream 3 gap between them mixes language with content and difficulty. Seventh, Cendol's native-speaker evaluation was designed by the authors of the Cendol models, and the contrast between Cendol and the other families confounds family with model size and training data. Finally, the CARE-Pragmatics framework is presented as an empirically motivated, prospective governance and alignment protocol. Experimental pre- versus post-DPO alignment trials, co-designed with native custodial boards, remain essential future work before any claim of operational remediation.

5. Conclusion

This multi-stream computational and sociolinguistic audit shows that the contemporary large language models examined here exhibit severe pragmatic flattening, honorific erasure, and uncalibrated metaphoric hallucination across Javanese, Sundanese, and Balinese. Across 2,620 regional pedagogical items, speech-level etiquette, classical prosody, and ritual scripts make up nearly one-third of formal cultural education, yet 19 global and regional LLMs suffer systematic accuracy and generation losses of 10.91 to 18.36 percentage points relative to Indonesian. Three-shot in-context prompting across six contemporary instruction-tuned LLMs fails to lift regional performance significantly (35.83% to 37.36%; p = 0.4472). They also stay overconfident on incorrect cultural answers (56.62% mean softmax probability). Native-speaker evaluations of 1,320 generations show that Indonesian instruction tuning dramatically reduces complete generative collapse (OR = 12.73). Surviving outputs still fail. They frequently commit intra-clause pronoun–verb collisions, deictic rank switching, and inanimate honorification, weaknesses that coincide with register-neutral Indonesian parallel corpora, which erase Javanese *Krama* deference in 79.01% of kinship-trigger clauses. The remedy is structural. Protecting Nusantara intangible linguistic heritage in the age of generative AI requires a shift from monolingual pivot translation toward multi-register parallel curation, dyadic preference alignment, and indigenous custodial sovereignty, as operationalized in the proposed CARE-Pragmatics framework.

Declarations

Ethical Approval. Not applicable. This computational and sociolinguistic audit analyzed exclusively open-access, anonymized secondary benchmark corpora and published human-evaluation records (IndoMMLU, SEACrowd, 4-Bit-IndoMMLU, Cendol, MaBL, Belebele, SIB-200, and NusaX). No live human subjects, private personal data, clinical interventions, or unconsented ethnographic recordings were collected.

Consent to Participate. Not applicable. No new human participant recruitment or primary behavioral data collection was conducted in this study.

Consent for Publication. Not applicable. This manuscript contains no identifiable personal records, private correspondence, or individual participant images.

Use of Artificial Intelligence. No generative artificial intelligence tools were used in conceiving, analyzing, or writing this manuscript.

Author Contributions. A.F.S.: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Writing – original draft. A.B.: Validation, Investigation, Supervision, Project administration, Writing – review & editing. B.S.: Resources, Investigation, Writing – review & editing.

Funding. This research did not receive any specific grant from public, commercial, or not-for-profit funding agencies.

Competing Interests. The authors declare that they have no competing interests.

Data Availability. The benchmark datasets analyzed here are publicly available from their repositories: IndoMMLU (https://huggingface.co/datasets/indolem/IndoMMLU), SEACrowd evaluation logs (https://github.com/SEACrowd/seacrowd-experiments), 4-Bit-IndoMMLU item predictions (https://github.com/afrizal-HA/4-Bit-IndoMMLU), Cendol human evaluation annotations (https://github.com/indo-nlp/cendol), and NusaX parallel corpora (https://github.com/IndoNLP/nusax). The derived sociolinguistic indicator tables, summary metrics, and analysis scripts can be obtained from the corresponding author upon reasonable request.

Acknowledgements. The authors gratefully acknowledge the open-science contributions of the IndoNLP, SEACrowd, IndoLEM, and Cohere For AI communities and of the regional Indonesian annotator communities, whose publicly released benchmarks and native-speaker evaluations made this audit possible.

References

1. Bender EM, Gebru T, McMillan-Major A, et al. On the Dangers of Stochastic Parrots. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021:610-623. doi:10.1145/3442188.3445922
2. Joshi P, Santy S, Budhiraja A, et al. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:6282-6293. doi:10.18653/v1/2020.acl-main.560
3. Blodgett SL, Barocas S, Daumé III H, et al. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. Proceedings of the 58th Annual Meeting of the Association for Computational

Linguistics. 2020:5454-5476. doi:10.18653/v1/2020.acl-main.485
4. Couldry N, Mejias UA. Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. Television & New Media. 2019;20(4):336-349. doi:10.1177/1527476418796632
5. Birhane A. Algorithmic Colonization of Africa. SCRIPT-ed. 2020;17(2). doi:10.2966/scrip.170220.389
6. Tao Y, Viberg O, Baker RS, et al. Cultural bias and cultural alignment of large language models. PNAS Nexus. 2024;3(9):pgae346. doi:10.1093/pnasnexus/pgae346
7. Cao Y, Zhou L, Lee S, et al. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP). 2023:53-67. doi:10.18653/v1/2023.c3nlp-1.7
8. Myung J, Lee N, Zhou Y, et al. BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. Advances in Neural Information Processing Systems 37. 2024:78104-78146. doi:10.52202/079017-2483
9. Rao AS, Yerukola A, Shah V, et al. NormAd: A Framework for Measuring the Cultural Adaptability of Large Language Models. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025:2373-2403. doi:10.18653/v1/2025.naacl-long.120
10. Li C, Chen M, Wang J, et al. CultureLLM: Incorporating Cultural Differences into Large Language Models. Advances in Neural Information Processing Systems 37. 2024:84799-84838. doi:10.52202/079017-2693
11. Errington JJ. Structure and Style in Javanese. University of Pennsylvania Press; 1988. doi:10.9783/9781512815764
12. Errington J. Speech in the Royal Presence: Javanese Palace Language. Indonesia. 1982;34:89. doi:10.2307/3350951
13. ERRINGTON JJ. On the Nature of the Sociolinguistic Sign: Describing the Javanese Speech Levels. Semiotic Mediation. 1985:287-310. doi:10.1016/b978-0-12-491280-9.50018-2
14. Uhlenbeck EM. The Krama-Ngoko Opposition. Studies in Javanese Morphology. 1978:278-299. doi:10.1163/9789004656970_015
15. Wulandari D, Al-Shidi S, Firuza T. Javanese Speech Levels and Language Politeness Among First-Grade Elementary Students: A Sociolinguistic Study. Journal of Language, Literature, and Educational Research. 2026;3(1):58-66. doi:10.37251/jolle.v3i1.3175
16. Aisha, Dwiana Asih Wiranti. Analysis of the Application of Basa Ngoko Lan Krama Song Media in Understanding Unggah Ungguh Javanese Language. Buana Pendidikan Jurnal Fakultas Keguruan dan Ilmu Pendidikan. 2025;21(2):105-114. doi:10.36456/bp.vol21.no2.a9895
17. Sudarmansyah R, Rudi Nurjaman A, Pradeka D, et al. The Relationship Between Understanding and Use of Sundanese Speech Levels (Undak-Usuk Basa) Among Elementary School Students. PiJIES. 2026;9(3):909-926. doi:10.24256/pijies.v9i3.12416
18. Setiawan FA, Fakhirah Z, Laxmi GF. Large Language Model Fine-Tuning for Indonesian-to-Sundanese Speech-to-Speech Translation Based on Undak-Usuk Basa. 2026 International Conference on Computer, Control, Informatics and its Applications (IC3INA). 2026:188-193. doi:10.1109/ic3ina72033.2026.11714672
19. Clynes A. Old Javanese influence in Balinese: Balinese speech styles. Language Contact and Change in the Austronesian World. 1994:141-180. doi:10.1515/9783110883091.141
20. Okasunu IGAS, Dewi KYS, Udayana IN. The Changing of 'Sor Singgih Basa' in Balinese Root Based on the Internal Modification: Morpho-Phonology Study. j.lang.lit. 2023;23(2):426-435. doi:10.24071/joll.v23i2.5895
21. I Wayan Sugita, I Gede Tilem Pastika. Dialektika Bahasa dan Agama: Penggunaan 'Sor-Singgih' Basa Bali sebagai Media Penanaman Etika Hindu dalam Pertunjukan Topeng Milenial di Kabupaten Gianyar. VIDYA SAMHITA: Jurnal Penelitian Agama. 2026;12(1):1-13. doi:10.25078/vs.v12i1.5742
22. Hikmah P, Gandung Joko Srimoko. Learning strategy for the "kinanthi" macapat song in serat wedhatama at randudongkal state high school. Jurnal Penelitian Musik. 2025;6(1):70-80. doi:10.21009/jurnalpenelitianmusik.61.06
23. Mahamid MNL, Hidayatulloh H. Wahdatul wujud dalam naskah keraton jawa: studi komparatif pada serat menak dan serat wedhatama. j. lektur keagamaan. 2024;22(1):287-316. doi:10.31291/jlka.v22i1.1150
24. Fricker M. Testimonial Injustice. Epistemic Injustice. 2007:9-29. doi:10.1093/acprof:oso/9780198237907.003.0002
25. Fricker M. Hermeneutical Injustice. Epistemic Injustice. 2007:147-175. doi:10.1093/acprof:oso/9780198237907.003.0008
26. Mignolo WD. Epistemic Disobedience, Independent Thought and Decolonial Freedom. Theory, Culture & Society. 2009;26(7-8):159-181. doi:10.1177/0263276409349275
27. Ni Wayan Saka Rahayu, Ni Made Citra Riesti Wulan, Nyoman Ari Yoga Wirawan, et al. Symbolic Efficacy in the Psychiatric Ward: A Clinical Ethnography of Cetik-Attributed Psychosis and Hospital-Facilitated Ritual Integration in Bali. cultural. 2026;4(1):132-144. doi:10.61996/cultural.v4i1.114
28. Singh B, Ashok Kumar. Examining the Impact of Digitalization on Indigenous Cultural Heritage Preservation in India. cultural. 2024;2(1):43-57. doi:10.61996/cultural.v2i1.59
29. Cahyawijaya S, Winata GI, Wilie B, et al. IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021:8875-8898. doi:10.18653/v1/2021.emnlp-main.699
30. Winata GI, Aji AF, Cahyawijaya S, et al. NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023:815-834. doi:10.18653/v1/2023.eacl-main.57
31. Cahyawijaya S, Lovenia H, Koto F, et al. NusaWrites: Constructing High-Quality Corpora for Underrepresented and Extremely Low-Resource Languages. Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2023:921-945. doi:10.18653/v1/2023.ijcnlp-main.60
32. Koto F, Aisyah N, Li H, et al. Large Language Models Only Pass Primary School Exams in Indonesia: A Comprehensive Test on IndoMMLU. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023:12359-12374. doi:10.18653/v1/2023.emnlp-main.760
33. Kabra A, Liu E, Khanuja S, et al. Multi-lingual and Multi-cultural Figurative Language Understanding. Findings of the Association for Computational Linguistics: ACL 2023. 2023:8269-8284. doi:10.18653/v1/2023.findings-acl.525
34. Cahyawijaya S, Lovenia H, Koto F, et al. Cendol: Open Instruction-tuned Generative Large Language Models for Indonesian Languages. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024:14899-14914. doi:10.18653/v1/2024.acl-long.796
35. Lovenia H, Mahendra R, Akbar SM, et al. SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024:5155-5203. doi:10.18653/v1/2024.emnlp-main.296
36. Aji AF, Cohn T. LORAXBENCH: A Multitask, Multilingual Benchmark Suite for 20 Indonesian Languages. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025:17432-17457. doi:10.18653/v1/2025.emnlp-main.881

e-ISSN 3026-7277
Vol. 4, No. 2 (2026) | pp. 85–103

37. Al Kautsar MD, Pranida SZ, Elbouardi B, et al. CULTURETALK-ID: A Multi-Task Dialogue Benchmark for Cultural Commonsense in Indonesian Local Languages. Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026. 2026:2888-2912. doi:10.18653/2026.findings-aacl.147
38. Farhansyah MR, Darmawan I, Kusumawardhana A, et al. Do Language Models Understand Honorific Systems in Javanese? Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025:26732-26754. doi:10.18653/v1/2025.acl-long.1296
39. Bandarkar L, Liang D, Muller B, et al. The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024:749-775. doi:10.18653/v1/2024.acl-long.44
40. Adelani DI, Liu H, Shen X, et al. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2024:226-245. doi:10.18653/v1/2024.eacl-long.14
41. Nguyen XP, Zhang W, Li X, et al. SeaLLMs - Large Language Models for Southeast Asia. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). 2024:294-304. doi:10.18653/v1/2024.acl-demos.28
42. Üstün A, Aryabumi V, Yong Z, et al. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024:15894-15939. doi:10.18653/v1/2024.acl-long.845
43. Dettmers T, Pagnoni A, Holtzman A, et al. QLoRA: Efficient Finetuning of Quantized LLMs. Advances in Neural Information Processing Systems 36. 2023:10088-10115. doi:10.52202/075280-0441
44. Winarti O. Language Shift of Krama to Bahasa Indonesia among Javanese Youths and it's Relation to Parents' Social Class. Indonesian Journal of Communications Studies. 2018;2(3):290. doi:10.25139/jsk.v2i3.186
45. Hadiwijaya M, Puput Kinanti K, Dian Puspitasari I. Youth and Indigenous Language: Assessing Javanese Krama Madya Language Vitality. J. Ilmu Sos. Human. 2022;11(3):397-406. doi:10.23887/jish.v11i3.44545
46. Sennrich R, Haddow B, Birch A. Controlling Politeness in Neural Machine Translation via Side Constraints. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016:35-40. doi:10.18653/v1/n16-1005
47. Rafailov R, Sharma A, Mitchell E, et al. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. Advances in Neural Information Processing Systems 36. 2023:53728-53741. doi:10.52202/075280-2338
48. Carroll SR, Garba I, Figueroa-Rodríguez OL, et al. The CARE Principles for Indigenous Data Governance. Data Science Journal. 2020;19:43. doi:10.5334/dsj-2020-043
49. Government of Indonesia. Undang-Undang Nomor 5 Tahun 2017 tentang Pemajuan Kebudayaan [Internet]. Lembaran Negara Republik Indonesia 2017 No. 104; 2017 [cited 2026 Nov 12]. Available from: https://peraturan.bpk.go.id/Details/37642/uu-no-5-tahun-2017