Open Access Indonesia Journal of Social Sciences

Vol. 9 · No. 5 · 2026 · pp. 318–337 · Article ID 339

e-ISSN 2722-4252 · *Journal homepage: journalsocialsciences.com/index.php/oaijss*

ORIGINAL ARTICLE

Calibrating Human–LLM Collaborative Coding Against Ethnographic Ground Truth in Contentious Social-Science Inquiry: Multi-Model Bias Auditing, Prediction-Powered Inference Debiasing, and Mixed-Methods Validation

Firzan Dahlan[1*], Giselle Dupont[2], Harun Urrashid[3]

[1] *Department of History Science, Enigma Institute, Palembang, Indonesia*
[2] *Department of Ethnomusical, LeBlanc Institute, Marrakesh, Morocco*
[3] *Department of Sharia Economy, Enigma Institute, Palembang, Indonesia*

ARTICLE INFORMATION

Received: September 1, 2026
Revised: September 15, 2026
Accepted: October 15, 2026
Available Online: October 22, 2026
Published: October 30, 2026

Keywords: Algospeak; Ethnographic ground truth; fsQCA; Human–LLM coding; Indonesian languages; Measurement error; Mixed methods; Prediction-powered inference

***Corresponding author**
Name: Firzan Dahlan
E-mail: firzan.dahlan@enigma.or.id
ORCID: 0009-0003-1193-4431

DOI: https://doi.org/10.37275/oaijss.v9i5.339

ABSTRACT

Background: Generative large language models (LLMs) increasingly replace manual qualitative coding in social science. Uncalibrated outputs and ensembles may distort constructs and bias parameters in contentious, veiled, or non-Western ethnolinguistic discourse.

Objective: To audit human–LLM coding against ethnographic ground truth and test whether Power-Tuned PPI++ restores valid inference.

Methods: We analyzed 144,865 public records in four tiers. Tier 1 audited 77,132 predictions from 13 LLMs on 13 tasks (SALT-NLP archive) with PPI and PPI++. Tier 2 combined multi-rater reliability (N = 960), HC3 regressions on 22 LLMs in Indonesian, Javanese, Minangkabau, and Sundanese (N = 2,640), and 841 error memos. Tier 3 applied surrogate regressions, AIPW, and RIF quantile models to 13,169 Indonesian tweets. Tier 4: fsQCA (24 regimes).

Results: Krippendorff's alpha fell 53.6% from explicit tasks (0.285) to veiled or dialectal tasks (0.132; cell-level p = 0.0042; task-level p = 0.106). Unanimous four-LLM consensus was wrong on 41.2%–47.4% of veiled-hate, dialect, escalation, and persuasion items. Naive surrogates reached 0.0% 95%-CI coverage on 6 of 7 tasks. PPI++ kept 96.1%–99.1%. Generalist LLMs fell from 40.83% acceptability in Indonesian to 2.08%–7.92% in indigenous registers, and Cendol tuning added 48.71 percentage points ($p = 2.96 \times 10^{-29}$). A formal-register classifier turned the true +8.70 pp vernacular hate-speech slope into −6.96 pp ($p < 1 \times 10^{-28}$), whereas an Ethno-PPI++ anchor of 400 tweets recovered +7.99 pp (p = 0.0154; 2.30× effective N).

Conclusion: Valid inference occurred in regimes combining culturally aligned models, native ethnographic anchors, and PPI++ debiasing. A five-stage ETHNO-PPI protocol needs testing elsewhere.

All authors approved the final version of the manuscript.

Copyright: © 2026 The Author(s). Authors retain copyright and publishing rights. Published by HM Publisher in collaboration with CMHC Research Center. **Access license:** Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International (CC BY-NC-SA 4.0). Non-commercial use, sharing, adaptation, distribution, and reproduction are permitted with appropriate credit and the same licence for adaptations.

1. Introduction

1.1. The generative AI turn in social-science coding and the surrogate validity crisis

Generative Large Language Models (LLMs) have driven a major methodological shift in text-as-data inquiry, a shift that spans sociology, political science, communication, and public policy.[1] Early benchmark evaluations found that instruction-tuned foundation models such as ChatGPT equaled or exceeded crowd-sourced workers on selected zero-shot classification tasks, at a fraction of traditional annotation cost.[2] Optimism spread widely. Many argued that generative AI could democratize large-scale qualitative coding, survey simulation, and interpretive content analysis.[3,4] Recent methodological syntheses in Sociological Methods & Research ask how generative models reshape flexible deductive coding, inductive theme discovery, and mixed human–machine observation designs.[5–8] Architecture, prompt memory, and alignment tuning strongly alter classification stability.[6,7,9] Work in Political Analysis has explored synthetic silicon sampling.[10] It has documented the severe variance compression and distribution shift that follow when uncalibrated LLMs replace human respondents,[11–13] compared parameter-efficient fine-tuning with zero-shot prompting,[11–13] and recast textual utterances as strategic social behavior instead of transparent propositional summaries.[11–13] Experimental audits add a further finding. Subtle source framing and institutional cues trigger systematic partisan and epistemic biases inside frontier language models.[14]

These warnings have gone largely unheeded. Applied empirical practice has outrun validation. Many recent studies feed tens of thousands of protest posts, parliamentary speeches, citizen grievances, or media transcripts into a zero-shot LLM, or take a majority vote across two or three commercial models, and then plug the resulting machine labels directly into downstream prevalence estimates or econometric regressions as if they were error-free human observations.[1,7,11,15–17] We argue that this 'uncalibrated surrogate paradigm' confronts three compounding methodological failures whenever social scientists study contentious politics and culturally situated discourse. First, we hypothesize that LLM reliability exhibits what we term the Explicit–Pragmatic Calibration Cliff: models perform reasonably well on literal propositional content, but they fail sharply when meaning rests on veiled dog-whistles, irony, relational power, or conversational escalation. Second, unanimity across a cross-family multi-LLM ensemble may reflect correlated pretraining blindness rather than construct truth. Third, feeding fallible automated classifications directly into regressions induces

non-classical differential measurement error. Its confidence intervals shrink deceptively toward zero coverage as corpus size N expands. It can even reverse the algebraic sign of true structural relationships.

1.2. Pragmatic contention, interpretive pluralism, and decolonial ethnolinguistic erasure

Contentious social-science inquiry rarely deals with sanitized, literal prose. Political stance differs from surface lexical sentiment.[18] Topic boundaries shift across institutional arenas.[19,20] Synthetic training augmentation can amplify domain stereotypes.[21,22] Evaluating political discourse quality demands sensitivity to evasion and rhetorical framing.[23] Human ground truth is no simpler. In subjective and contentious domains, it is itself marked by structured interpretive pluralism, not mechanical unanimity. As Davani, Díaz, and Prabhakaran demonstrate, aggregating multi-annotator corpora by crude majority voting erases minority lived-experience perspectives on hate speech and group harm.[24] Sound qualitative methodology therefore needs explicit inter-rater reliability diagnostics adapted to case-study and ordinal rubrics.[25] It also needs abductive iteration between empirical anomalies and theory,[26–29] transparent qualitative data archiving,[26–29] and attention to the dynamics of language use in contentious online deliberation.[26–29]

These interpretive challenges grow much sharper outside Western, Educated, Industrialized, Rich, and Democratic (WEIRD) English-language corpora. Drawing on linguistic anthropology, recent critiques in Big Data & Society argue that cultural interpretability cannot be reduced to English translation benchmarks,[30–32] expose how natural language processing enforces harmful 'missing standardization' ideologies against oral and regional vernaculars,[30–32] and document systematic individualist-versus-collectivist value biases in pretrained transformers.[30–32] Other work shows how state-aligned AI flattens complex social lifeworlds into legible administrative categories.[33] It also asks whose voices survive automated public-comment filtering.[34] Even state-of-the-art multilingual classifiers struggle when open-ended responses embed culturally specific idioms[35,36] or operate under non-WEIRD platform governance regimes.[35,36] In digital activism and everyday contention, marginalized and youth publics routinely deploy 'algospeak' to evade algorithmic moderation and signal in-group solidarity. The term covers deliberate phonetic respellings, slang abbreviations, and ironic code-switching.[37,38] Indonesia is the world's fourth most populous nation. It is an archipelago of over 700 living languages.[39] There, online civic, religious, labor, and electoral contention continuously blends national Bahasa Indonesia with colloquial vernacular ('*bahasa alay*').[40] Regional indigenous registers add further layers, such as Javanese (structured by *ngoko* and *krama* honorific speech levels), Sundanese (*undak-usuk basa*), Minangkabau, Buginese, Madurese, and Batak.[41–43] Generalist LLMs pretrained overwhelmingly on English and formal Indonesian web scrapes risk a marked loss of meaning across these ethnolinguistic registers.

1.3. Statistical debiasing (PPI++) and critical mixed-methods joint-display integration

Linking ethnographic validity to computational scale takes two things at once. One is statistical correction of measurement error. The other is a formal mixed-methods integration. In mathematical statistics, Angelopoulos and colleagues introduced Prediction-Powered Inference (PPI). The method combines machine predictions over a large unlabeled corpus N with a smaller gold-standard human-labeled anchor sample n, and it yields provably valid confidence intervals for means, quantiles, and regression coefficients without requiring the machine learning model to be unbiased.[44] Later work on Power-Tuned PPI++ adds an optimal scalar or matrix tuning weight λ^*, bounded between 0 and 1. The weight adapts on its own to the empirical human–machine error covariance. A noisy or uncalibrated LLM therefore never pushes estimator variance above that of classical human-only inference.[45,46] Closely related econometric frameworks arrive at the same theorem. They include Design-Based Supervised Learning (DSL) for correcting AI labeling errors,[47–49] GMM error-in-variables adjustments for machine-predicted covariates,[47–49] and proxy-variable causal identification audits.[47–49] Automated text classifications cannot be treated as ground-truth variables without human calibration anchors.

Statistical debiasing alone cannot validate construct meaning. It fails when the underlying human rubric is poorly specified. It also fails when researchers pass over qualitative patterns in high-residual failure cases, which are the very cases where a surrogate breaks down. Recent methodological reforms in Qualitative Comparative Analysis (QCA) guard against positive-results inflation through strict necessity and Proportional Reduction in Inconsistency (PRI) standards.[50] The Journal of Mixed Methods Research now sets binding reporting standards for artificial intelligence transparency.[51] Explicit formulation of integrative research questions[52] is covered as well. A further standard governs structured Delphi/Joint-Display meta-inferences that reconcile confirmatory, discordant, and expanding evidence.[53,54] This study builds on these advances. It poses three integrative research questions. They run across 144,865 SHA-256-verified public records. (RQ1) Across global contentious social-science benchmarks, how severely does human–LLM calibration degrade as framing moves from explicit to veiled pragmatic and dialectal contention, and can multi-LLM ensemble agreement or Power-Tuned PPI++ resolve surrogate coverage collapse? (RQ2) Across Indonesian and indigenous archipelagic languages (Javanese, Minangkabau, Sundanese), how do generalist and culturally instruction-tuned LLMs perform against native-speaker ethnographic ground truth and multi-rater human reliability benchmarks? (RQ3) In real contentious Indonesian discourse saturated with colloquial vernacular ('*bahasa alay*'), how do formal-register surrogates distort downstream regression slopes? Which configuration of model alignment, ethnographic anchoring, and PPI++ debiasing is necessary and conjuncturally sufficient for valid social-science inference?

2. Methods

2.1. Four-tier critical mixed-methods research architecture and SHA-256 data provenance

We implement a convergent-sequential Critical Mixed-Methods validation design,[52,55] which spans four empirical tiers and 11 publicly archived datasets totaling 144,865 raw records (24.92 MB). Every raw file came from its canonical open-science repository. Cryptographic SHA-256 checksums were used to check the integrity of each file, and the checksums are listed in a SHA-256 manifest. The authors generated no synthetic or simulated text. Reporting followed the Mixed Methods Appraisal Tool (MMAT) and the transparency recommendations for the use of AI and automated tools in mixed-methods studies.[51]

Tier 1 is the Stanford/Georgia Tech SALT-NLP Computational Social Science Benchmark Archive, curated by Ziems et al..[1] It holds 77,132 item-level zero-shot predictions from 13 distinct LLMs across 13 classification tasks. These are paired 1:1 with 6,247 expert-annotated ground-truth contexts and 229 human qualitative error-audit records. The 13 evaluated models cover two major architectural families and four orders of parameter magnitude. One family is Google encoder-decoder instruction models: flan-t5-small, flan-t5-base, flan-t5-large, flan-t5-xl, flan-t5-xxl at 11B, and flan-ul2 at 20B. The other is OpenAI autoregressive models: text-ada-001, text-babbage-001, text-curie-001, text-davinci-001, text-davinci-002, text-davinci-003, and chatgpt / gpt-3.5-turbo. We applied a pre-specified completeness rule that requires at least 50 evaluated items per task-model cell. Under it, one degenerate 1-row upstream file (ibc × text-ada-001) and four absent legacy API runs are excluded. This leaves 164 complete task-model cells across 77,131 predictions (the 77,132 predictions of the archive minus the single row of the degenerate file). Tasks are classified theoretically a priori. Six are Explicit Propositional Framing benchmarks: Misinformation Reaction Frames (mrf-classification), Ekman Affective Emotion (emotion), Partisan Media Ideology (media_ideology), Ideological Books Corpus (ibc), Political Target Stance (semeval_stance), and Workplace Politeness Norms (politeness). Seven are Pragmatic, Veiled, Relational, and Postcolonial Dialect benchmarks: Figurative Language Sarcasm/Metaphor (flute-classification), Conversational Discourse Acts (discourse), Collective Persuasion Efficacy (persuasion), Conversational Personal-Attack Escalation (conv_go_awry), Wikipedia Relational Power Asymmetry (power), Implicit Veiled Hate Speech Taxonomy (hate: white grievance, incitement, inferiority, irony, stereotyping, intimidation), and Postcolonial Indian English Dialect Syntax (indian_english_dialect). The task grouping rests on the authors' judgment: borderline tasks (for example Figurative Language and Indian English Dialect Syntax) could be assigned differently, and the range of 2 to 23 classes limits the comparability of α across tasks.

Tier 2 integrates four open Indonesian and indigenous-language evaluation repositories from the IndoNLP consortium. (Tier 2A) There are three multi-rater native-speaker human reliability panels: Phase 1, n = 300; Phase 2, n = 360; 4-Rater Recap, n = 300. Together they give N = 960 multi-rater rows. The raters are independent Indonesian speakers, R1–R4. (Tier 2B) The second is 2,640 item-level native-speaker ethnographic evaluations. Each of the four languages has 660 items: national Indonesian, Javanese, Minangkabau, and Sundanese. These score 22 global, Southeast Asian, and culturally instruction-tuned LLMs: Llama-2-7B/13B-Chat, SeaLLM-7B-Chat, SEA-LION-7B-Instruct, BLOOMZ-7B1, Bactrian-X/ID-BLOOM-7B, Merak-7B-v4, and 14 Cendol instruction-tuned LLaMA-2 and mT5 variants from 300M to 13B parameters.[56] Scoring uses a four-grade semantic-cultural rubric (A = 4 complete ethnographic equivalence, B = 3 minor diction variance preserving pragmatic intent, C = 2 meaning-altering lexical shift, D = 1 severe hallucination or untranslated register echo). The evaluations come with 841 verbatim qualitative native-speaker field notes. (Tier 2C) The third is 2,340 culturally situated figurative metaphors in Indonesian, Javanese, and Sundanese from the MABL benchmark.[57] (Tier 2D) The fourth is 14,981 cultural, regional-language, sociological, civic, and historical items from IndoMMLU.[58]

Tier 3 links the Ibrohim and Budi multi-label Indonesian social-media corpus of hate speech and abusive language[59] to three further resources. The corpus holds N = 13,169 expert-coded Twitter utterances, annotated across 12 binary dimensions: overall Hate Speech HS, Abusive profanity, Individual vs. Collective Group targeting HS_Group, Religion HS_Religion, Race/Ethnicity HS_Race, Physical/Disability HS_Physical, Gender/Sexual Orientation HS_Gender, Other, and Weak, Moderate, and Strong intensity. (Tier 3B) The first resource is a set of 15,167 canonical-to-vernacular orthographic mappings in the Indonesian Kamus Alay colloquial/algospeak lexicon.[60] (Tier 3C) Next come 125 explicit abusive root terms. (Tier 3D) Last are 12,000 human-translated parallel pragmatic utterances. They span 12 languages in NusaX.[41] Of these, 10 are indigenous Indonesian languages (Acehnese, Balinese, Banjarese, Buginese, Javanese, Madurese, Minangkabau, Ngaju, Sundanese, and Toba Batak). The other two are national Indonesian and English.

2.2. Psychometric reliability metrics and multi-rater human ground-truth auditing

Accuracy and macro-F1 ignore chance agreement. That matters here. Class marginals are skewed, so our primary construct-reliability metric across Tiers 1–3 is exact Krippendorff's alpha (α), computed from observed and expected coincidence matrices over all valid and off-schema categories.[61] Unordered categorical tasks in Tiers 1 and 3 use nominal coincidence distance ($\delta^2 = 1$ for $c \neq k$, 0 otherwise). Cohen's kappa (κ), multi-rater Fleiss' κ, and Gwet's AC1 are computed alongside it. That index guards against the well-known 'high-agreement, low-kappa' paradox. Extreme marginal prevalence produces it. A zero-shot LLM sometimes returns verbose off-schema text, or it refuses a constrained choice once formatting artifacts are stripped. Such a prediction is not dropped. It is assigned to a distinct off-schema token, so that instruction-following failures are penalized honestly.[1,7]

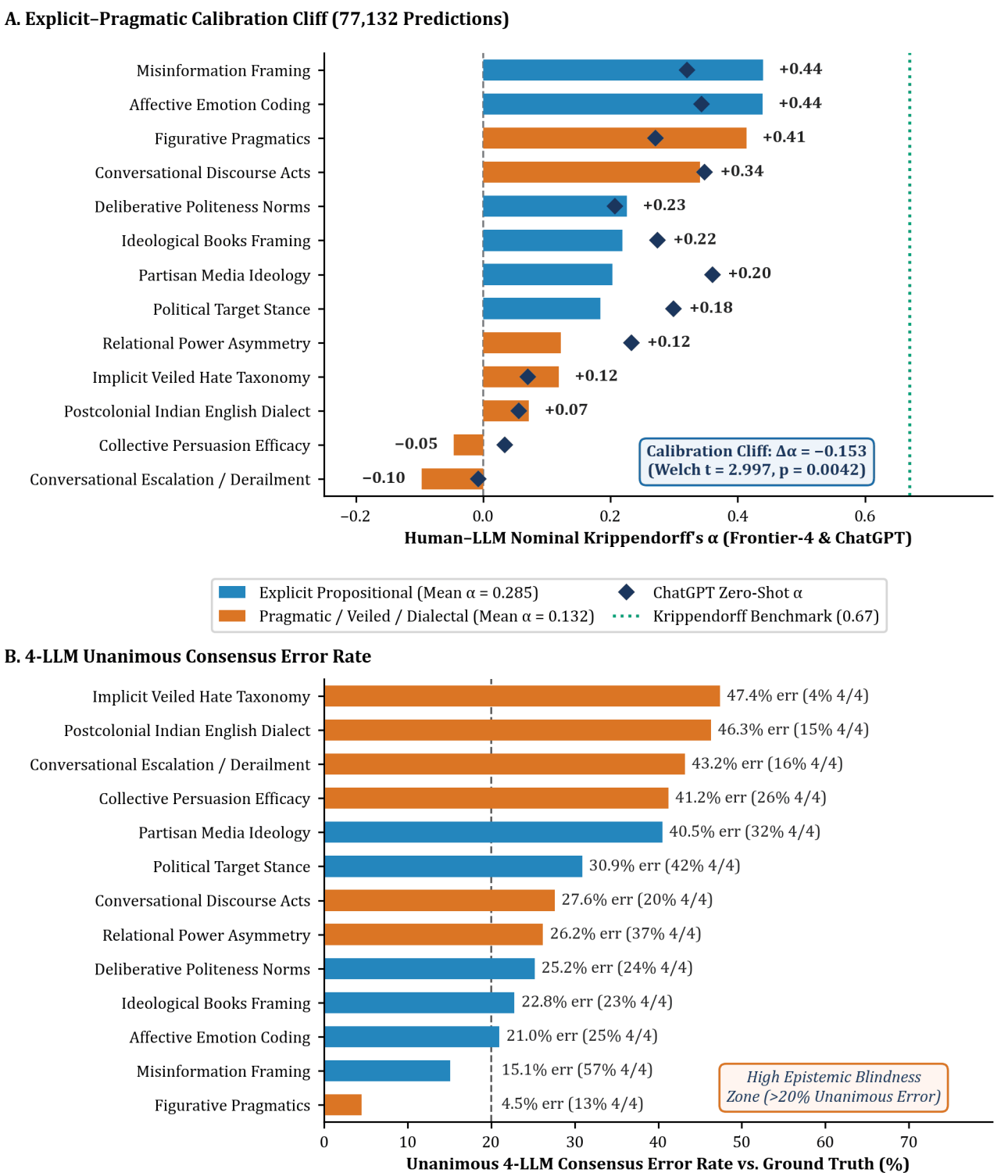


Figure 1. Multi-model benchmark audit across 13 contentious social-science tasks in the SALT-NLP archive (N = 77,132 item-level predictions across 13 LLMs). (A) Human–LLM nominal Krippendorff's α across Frontier-4 instruction-tuned models and zero-shot ChatGPT, documenting the −0.153 calibration cliff between explicit propositional framing and pragmatic/veiled/dialectal contention (cell-level Welch t = 2.997, p = 0.0042; task-level p = 0.106). (B) Unanimous 4-of-4 cross-family ensemble consensus error rate against expert human ground truth.

Tier 2 is rated by native Indonesian, Javanese, Minangkabau, and Sundanese raters, who score outputs on an ordered four-tier ethnographic rubric (A = 4 to D = 1). Both Nominal Krippendorff's α and Quadratic-Ordinal Krippendorff's α are computed. The squared rank distance $\delta^2(c, k) = (c - k)^2$ separates two kinds of disagreement. A minor stylistic boundary disagreement (A vs. B, $\delta^2 = 1$) counts little, whereas a catastrophic one between full fidelity and semantic collapse (A vs. D, $\delta^2 = 9$) counts heavily.[24,61] Why compare the two versions? Putting nominal α beside ordinal α across multi-rater human panels makes our epistemological distinction measurable, because it separates benign interpretive pluralism from rubric breakdown. Raters either differ in a tolerable way or the rubric has failed. The comparison shows which.

2.3. Prediction-powered inference (PPI++), surrogate sign-reversal econometrics, AIPW, and RIF

Let $Y \in \{0, 1\}$ denote the true ethnographic ground-truth construct. Let $f(X) \in (0, 1)$ denote an imperfect machine prediction available for an entire corpus of size $N = n + N_u$, whereas human gold-standard codes Y_i exist only on a stratified calibration anchor of size $n \ll N$. Naive surrogate estimation takes the corpus-wide machine mean $\hat{\theta}_{naive} = N^{-1} \sum f(X_i)$ as an unbiased estimate of $\theta^* = E[Y]$. Suppose $E[f(X) - Y] \neq 0$. The asymptotic bias then stays fixed. The

nominal standard error still shrinks at rate $O(N^{-1}/^{2})$. Frequentist confidence-interval coverage is therefore driven toward zero.[44,48] Standard Prediction-Powered Inference[44] subtracts the empirical machine bias measured on the human anchor: $\hat{\theta}_PPI(1) = N_u^{-1} \sum_\{u\} f(X_j) + n^{-1} \sum_\{l\} (Y_i - f(X_i))$. Trouble arises when the LLM is poorly calibrated on veiled pragmatic contention, so that $Var(Y - f) > Var(Y)$. Fixing $\lambda = 1$ then inflates variance. It rises above the level of classical human-only estimation. We therefore implement Power-Tuned PPI++:[45,46] $\hat{\theta}_PPI++(\lambda^*) = \lambda^* N_u^{-1} \sum_\{u\} f(X_j) + n^{-1} \sum_\{l\} (Y_i - \lambda^* f(X_i))$. The optimal scalar tuning weight comes straight from the anchor sample: $\lambda^* = clip(Cov(Y_l, f_l) / ((1 + n/N_u) Var(f_l)), 0, 1)$. Linear regression parameters $\beta^* = (E[X X^T])^{-1} E[X Y]$ get the same treatment. Our Ethno-PPI++ regression estimator combines the λ^*-scaled unlabeled projection with the anchor-sample residual rectifier $n^{-1} \sum_\{l\} X_i (Y_i - \lambda^* f_i)$. The heteroskedasticity-consistent sandwich variance sums the unlabeled and anchor covariance components.[45,47]

Tier 3 audits how automated coding fails (N = 13,169 Indonesian utterances). The audit is distributional and uses causal-inference estimators to describe adjusted differences. Three automated or hybrid workflows are compared against full-corpus human ground truth, namely (1) a Naive Abusive-Lexicon Surrogate, which flags utterances containing ≥1 explicit abusive root; (2) a Formal-Register Only Classifier, a TF-IDF L2-regularized logistic regression trained strictly on zero-vernacular formal Indonesian utterances, replicating the blind spots of formal-only pretraining; and (3) a Vernacular-Contextualized Classifier, which applies the exact 15,167-term Kamus Alay orthographic normalization with probabilities that are 5-fold cross-fitted. The third workflow is paired with Stratified Ethno-PPI++ on an anchor of n = 400 human-coded tweets (3.04% of corpus). Now consider the true contentious subset. Among all N = 5,561 true contentious tweets (HS = 1), we estimate Doubly Robust Augmented Inverse Probability Weighting (AIPW) Average Treatment Effects for two exposures: (a) Implicit Non-Profane Framing (Abusive = 0), with lexicon false-negative erasure as the outcome, and (b) High Vernacular Code-Switching (vernacular token share ≥ 20%), with formal-classifier false-negative erasure as the outcome. We report IPTW standardized mean differences (|SMD|), VanderWeele E-values, and Cinelli–Hazlett robustness values (RV_q).[62–64] Last, Firpo–Fortin–Lemieux Recentered Influence Function (RIF) unconditional quantile regressions are fitted across $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$. They target upper-tail calibration error $|Y_HS - \hat{p}_formal|$. None of the three workflows is an LLM: the first is a lexicon rule and the second and third are TF-IDF logistic classifiers, so Tier 3 illustrates surrogate measurement error in automated coding in general. Because 'Implicit Non-Profane Framing' (Abusive = 0) is defined by the absence of the abusive terms that trigger the lexicon, its association with lexicon false negatives is expected by construction.

2.4. Exact Ragin four-value fsQCA calibration across N = 24 empirical coding regimes

Tier 4 synthesizes Tiers 1–3. We calibrate N = 24 empirically observed coding regimes (R01–R24) into four-value fuzzy sets (0.00 = fully out, 0.33 = more out than in, 0.67 = more in than out, 1.00 = fully in), following Rohlfing's anti-inflation standards.[50,65] Five methodological conditions are calibrated from exact empirical benchmarks. (1) Propositional Clarity (PROP_CLARITY: 1.00 literal misinformation headlines; 0.67 explicit partisan/stance/formal Indonesian; 0.33 conversational/Javanese/Sundanese/vernacular with contextual markers; 0.00 veiled irony/dog-whistles or postcolonial dialect). (2) Cultural & Register Alignment (CULT_ALIGN: 1.00 matched language/register or Cendol/Kamus Alay adaptation; 0.67 standard English pretraining on veiled sociolects or small regional adaptation; 0.33 weak cross-lingual transfer; 0.00 severe colonial/indigenous register mismatch). (3) High-Capacity Instruction Architecture (SCALE_INST: 1.00 frontier ≥11B/ChatGPT/UL2; 0.67 7B instruction or cross-fitted contextual classifier; 0.33 static formal linear model; 0.00 ≤300M parameter model). (4) Ethnographic Ground-Truth Anchor (ETHNO_ANCHOR: 1.00 multi-rater native/expert anchor n ≥ 150–400; 0.67 unstratified partial anchor; 0.00 zero-shot unanchored). (5) Statistical PPI++ Debiasing (PPI_DEBIAS: 1.00 Power-Tuned PPI++ $\lambda^* \in (0, 1)$; 0.33 unclipped $\lambda = 1$ PPI on noisy task; 0.00 unrectified surrogate). The target outcome is Valid Downstream Inference (VALID_INFERENCE). It is scored 1.00 when nominal 95%-CI coverage is 95%–99%, variance is not inflated, and construct fidelity is ≥50% Grade A/B or $\alpha \geq 0.20$. It is scored 0.67 when valid 95% coverage is achieved despite harder pragmatic variance (R06, R14, R19, R21). Moderate bias or variance inflation (R01, R03, R05, R11, R15) earns 0.33, and coverage collapse or sign reversal earns 0.00. Necessity is tested with consistency ≥ 0.90 plus Relevance of Necessity RoN. Conservative Quine–McCluskey sufficiency is tested with raw consistency ≥ 0.82, PRI ≥ 0.70, and zero remainder assumptions. Both tests are run symmetrically for VALID_INFERENCE and ~VALID_INFERENCE. Because VALID_INFERENCE is defined by coverage, variance, and fidelity criteria that Power-Tuned PPI++ is designed to satisfy, PPI_DEBIAS and the outcome overlap by construction, and the regimes share datasets; the analysis is therefore an organizing synthesis of Tiers 1–3 and not an independent test.

3. Results

3.1. Tier 1: the explicit–pragmatic calibration cliff, ensemble consensus error, and PPI++ validity

The SALT-NLP archive supplies 77,132 item-level predictions covering 13 contentious social-science tasks and 13 LLMs (Table 1; Figure 1A). Here, zero-shot machine coding shows a sharp, statistically significant Explicit–Pragmatic Calibration Cliff. Take the four frontier instruction-tuned models (ChatGPT, text-davinci-003, FLAN-UL2, and FLAN-T5-XXL). Mean nominal Krippendorff's α for these models declines by 53.6%, from 0.2852 across the six Explicit Propositional Framing benchmarks to 0.1323 across the seven Pragmatic, Veiled, and Postcolonial Dialect benchmarks ($\Delta\alpha = -0.1529$, Welch $t = 2.997$, $p = 0.0042$). Even the best single model on individual explicit tasks stays below Krippendorff's tentative 0.667 standalone reliability threshold. FLAN-UL2 reaches $\alpha = +0.643$ on Affective Emotion and $\alpha = +0.549$ on Misinformation Framing; ChatGPT reaches $\alpha = +0.360$ on Media Ideology and $\alpha = +0.299$ on Political Stance. Relationally situated and veiled contention changes the picture. Zero-shot reliability collapses altogether. ChatGPT achieves $\alpha = +0.233$ on Wikipedia Power Asymmetry. On the six-class Implicit Veiled Hate Taxonomy it reaches $\alpha = +0.070$ (accuracy 29.7%), and on Postcolonial Indian English Dialect syntax $\alpha = +0.056$ (accuracy 15.0%). Collective Persuasion yields $\alpha = +0.034$. On predicting whether a contentious conversation will derail into a personal attack (conv_go_awry), $\alpha = -0.008$. That is below

the chance level. Larger models help. Scaling model size within instruction-tuned families improves overall reliability, from mean α = −0.203 in flan-t5-small and −0.270 in text-curie-001 up to +0.212 in flan-t5-xxl, +0.216 in chatgpt, and +0.246 in flan-ul2. Even so, every single frontier model drops when moving from explicit to veiled tasks, by −0.061 to −0.240. The Welch test treats the 52 model–task cells as independent although the same four models appear in every task, so the p value is descriptive. With the 13 task means as the unit, the Welch test gives t = 1.77 (df = 10.3, p = 0.106) and an exact permutation test gives p = 0.122, so the contrast is suggestive rather than conclusive at the task level.

Table 1 and Figure 1B show that inter-LLM ensemble agreement is not sufficient to validate automated coding in the absence of human ground truth. Suppose all four cross-family frontier models (ChatGPT, text-davinci-003, FLAN-UL2, and FLAN-T5-XXL) unanimously predict the exact same label (4/4 consensus). Does that agreement guarantee a correct label? No. Against expert human ground truth, the unanimous machine consensus is factually wrong 47.4% of the time on Implicit Veiled Hate. Other tasks follow the same pattern: 46.3% on Postcolonial Indian English Dialect, 43.2% on Conversational Escalation, 41.2% on Collective Persuasion, 40.5% on Media Ideology, and 30.9% on Political Target Stance, compared with 15.1% on literal Misinformation Framing. Agreement is weak to begin with. On Implicit Veiled Hate, inter-LLM α among the four frontier models themselves is just +0.086. When they do converge on a single category, they typically default to generic stereotyping over veiled irony or white-grievance dog-whistles. Nearly half of those 'high-confidence' ensemble labels are wrong, an error that may reflect shared pretraining. Agreement is not accuracy. Unanimity does lower error relative to a single model: ChatGPT alone has an overall accuracy of 29.7% on Implicit Veiled Hate and 15.0% on the dialect task, and on the two-class tasks (Escalation, Persuasion) an error of 41%–43% is close to the 50% expected by chance. The unanimous subsets are small: 3.8% of 498 items (about 19) on Implicit Veiled Hate, 15.4% of 266 (about 41) on Postcolonial Indian English Dialect, 16.2% of 500 (about 81) on Conversational Escalation, and 26.3% of 434 (about 114) on Collective Persuasion (Table 1).

Table 2 and Figure 2 show how Power-Tuned PPI++ addresses the inferential failure caused by plugging these imperfect zero-shot predictions directly into prevalence estimation. The analysis uses 800 repeated random calibration splits at anchor size n = 150. Naive Zero-Shot ChatGPT Surrogates show severe systematic prevalence biases. They underestimate Misinformation headlines by −29.8 percentage points, Hostile Politeness violations by −30.1 pp, Conversational Escalation by −26.2 pp, Veiled Ironic Hate by −16.1 pp, and Conservative IBC framing by −13.9 pp. Persuasion success is overestimated by +16.6 pp. Naive standard errors shrink as the full corpus N grows. Nominal 95% confidence-interval coverage collapses to 0.0% across six of the seven benchmarks. Standard unclipped PPI (fixing λ = 1) restores valid frequentist coverage (96.4%–98.2%) by subtracting anchor bias. Yet on uncalibrated tasks, where human–LLM correlation is near zero (Conversational Escalation r = 0.070; Persuasion r = 0.064), subtracting noisy machine predictions inflates the 95% CI width from 16.0 pp (classical human-only n = 150) up to 22.1 pp (Figure 2B). Power-Tuned PPI++ behaves differently. It achieves valid 96.1%–99.1% empirical coverage across all seven tasks while strictly dominating or matching human-only precision everywhere. On calibrated explicit tasks, PPI++ sets λ* = 0.30–0.43 to narrow 95% CIs, yielding 1.04×–1.19× effective sample size gains at n = 150, rising to 1.25× at n = 50. On uncalibrated Escalation and Persuasion it automatically shrinks λ* → 0.064 and 0.047, removing the effect of LLM noise on interval width. In this design the labeled and unlabeled items come from the same finite benchmark set of about 500 items per task, so human-only coverage above 95% in every row (96.1%–99.0%) is consistent with finite-population conservatism.

A. Nominal 95% Coverage: Surrogate Collapse vs. PPI++



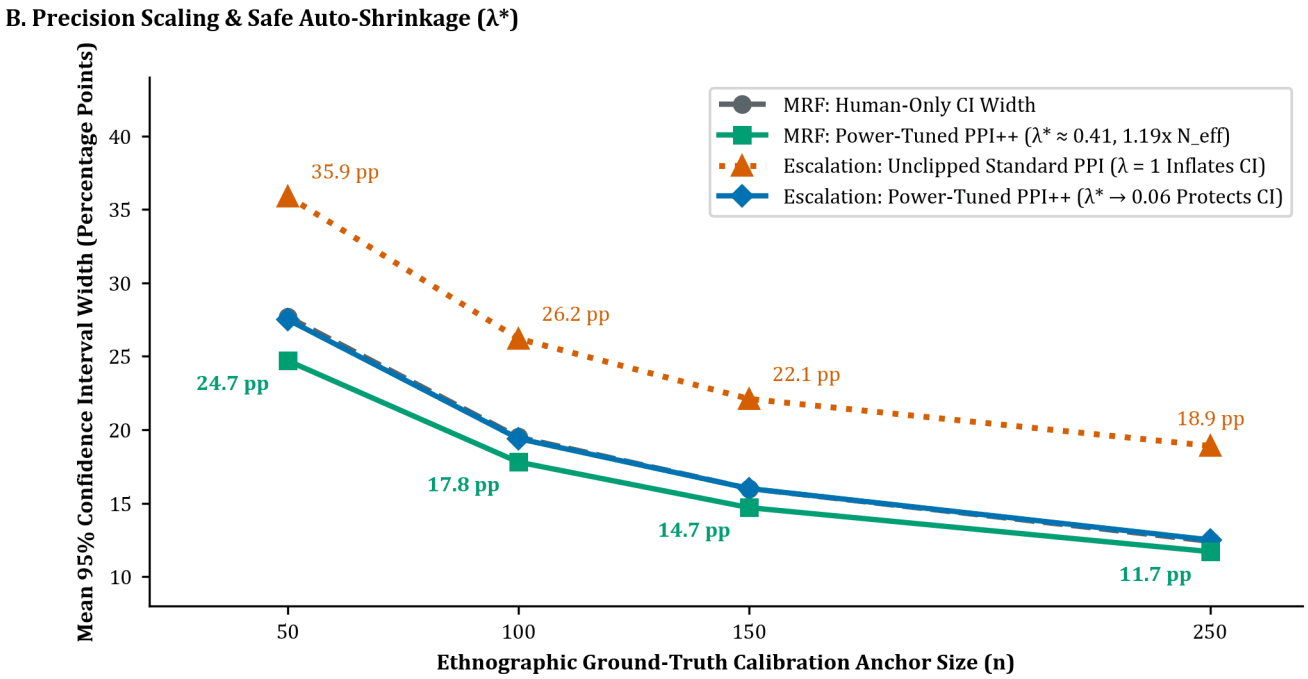


Figure 2. Prediction-Powered Inference (PPI and Power-Tuned PPI++) repeated-sampling calibration experiments across seven contentious social-science tasks (800 replications per anchor budget). (A) Empirical 95% confidence-interval coverage at anchor size n = 150, contrasting catastrophic 0.0% coverage collapse under naive zero-shot LLM surrogates with valid 96.1%–99.1% coverage under Power-Tuned PPI++. (B) Mean 95% confidence-interval width across anchor budgets n ∈ {50, 100, 150, 250}, illustrating precision gains on calibrated tasks (MRF) and automatic λ* → 0.06 shrinkage preventing variance inflation on uncalibrated conversational escalation.

Table 1. Multi-Model Human–LLM Reliability Audit and Cross-Family Ensemble Consensus Error Across 13 Contentious Social-Science Tasks (Tier 1: SALT-NLP Archive, N = 77,132 Predictions)

Contentious Social-Science Task (Domain Tier)	Items (Classes)	ChatGPT α (Acc.)	Frontier-4 Mean α	Best LLM (α)	4/4 Ensemble Unan. Err. (%)
Misinformation Framing (MRF) (Explicit)	500 (2)	+0.320 (69.0%)	+0.440	flan-ul2 (+0.549)	15.1% (57.0% unan.)
Affective Emotion Coding (Explicit)	498 (6)	+0.343 (46.2%)	+0.439	flan-ul2 (+0.643)	21.0% (24.9% unan.)
Figurative Pragmatics (Sarcasm/Metaphor) (Pragmatic/Veiled)	500 (4)	+0.271 (50.2%)	+0.414	flan-ul2 (+0.518)	4.5% (13.2% unan.)
Conversational Discourse Acts (Pragmatic/Veiled)	497 (7)	+0.348 (44.5%)	+0.341	flan-t5-xxl (+0.443)	27.6% (19.7% unan.)
Deliberative Politeness Norms (Explicit)	498 (3)	+0.207 (50.8%)	+0.226	flan-t5-xl (+0.380)	25.2% (23.9% unan.)
Ideological Books Framing (IBC) (Explicit)	498 (3)	+0.274 (54.8%)	+0.219	text-davinci-002 (+0.364)	22.8% (22.9% unan.)
Partisan Media Ideology (Explicit)	498 (3)	+0.360 (58.8%)	+0.203	chatgpt (+0.360)	40.5% (31.7% unan.)
Political Target Stance (Explicit)	435 (3)	+0.299 (49.9%)	+0.184	chatgpt (+0.299)	30.9% (41.6% unan.)
Relational Power Asymmetry (Pragmatic/Veiled)	500 (2)	+0.233 (61.6%)	+0.122	chatgpt (+0.233)	26.2% (37.4% unan.)
Implicit Veiled Hate Taxonomy (Pragmatic/Veiled)	498 (6)	+0.070 (29.7%)	+0.119	flan-t5-xl (+0.227)	47.4% (3.8% unan.)
Postcolonial Indian English Dialect (Pragmatic/Veiled)	266 (23)	+0.056 (15.0%)	+0.072	flan-t5-xxl (+0.153)	46.3% (15.4% unan.)
Collective Persuasion Efficacy (Pragmatic/Veiled)	434 (2)	+0.034 (53.0%)	−0.046	flan-t5-large (+0.134)	41.2% (26.3% unan.)
Conversational Escalation / Derailment (Pragmatic/Veiled)	500 (2)	−0.008 (53.0%)	−0.097	flan-t5-base (+0.012)	43.2% (16.2% unan.)
Explicit vs. Pragmatic Calibration Cliff (Frontier-4)	52 Cells	Explicit: +0.2852	Pragmatic: +0.1323	Δα = −0.1529	Welch t = 2.997, p = 0.0042

Note: Frontier-4 comprises ChatGPT (gpt-3.5-turbo), text-davinci-003, Google FLAN-UL2 (20B), and FLAN-T5-XXL (11B). Nominal Krippendorff's α penalizes off-schema generations honestly. 4/4 Unanimous Error Rate measures exact error against expert human ground truth conditional on all four frontier LLMs predicting the identical category. The Welch test treats the 52 cells as independent; with the 13 task means as the unit, t = 1.77 (p = 0.106; permutation p = 0.122). Alpha is not directly comparable across tasks with different numbers of classes (2–23); the share of items with unanimous four-model agreement is given in parentheses in the last column.

Table 2. Repeated-Sampling Calibration Audit of Naive LLM Surrogates, Classical Human Subsamples, Standard Unclipped PPI (λ = 1), and Power-Tuned PPI++ (λ* ∈ (0, 1)) at Anchor Budget n = 150 (800 Replications)

Benchmark Prevalence Parameter (ChatGPT Surrogate)	True θ* (Naive Bias)	Naive 95% Coverage	Human-Only 95% Cov. (CI Width)	Standard PPI Width (λ = 1)	Power-Tuned PPI++ Cov. (Width, λ*, N_eff)
Misinformation Headline Share (r = 0.473)	50.0% (−29.8 pp)	0.0%	97.1% (16.0 pp)	17.2 pp (98.2%)	98.5% (14.7 pp; λ*=0.41; 1.19×)
Anti-Target Stance Share (r = 0.460)	33.3% (+1.8 pp)	100.0%	99.0% (15.1 pp)	19.4 pp (97.5%)	99.1% (14.0 pp; λ*=0.30; 1.16×)
Conservative Ideological Framing Share (r = 0.416)	33.3% (−13.9 pp)	0.0%	98.1% (15.1 pp)	17.3 pp (97.9%)	99.0% (14.1 pp; λ*=0.34; 1.14×)
Impolite / Hostile Norm Violation Share (r = 0.234)	33.3% (−30.1 pp)	0.0%	98.1% (15.1 pp)	15.3 pp (98.2%)	98.5% (14.8 pp; λ*=0.43; 1.04×)
Conversational Personal-Attack Escalation Share (r = 0.070)	50.0% (−26.2 pp)	0.0%	97.1% (16.0 pp)	22.1 pp (96.4%)	97.1% (16.0 pp; λ*=0.06; 1.01×)
Successful Collective Argument Share (r = 0.064)	50.0% (+16.6 pp)	0.0%	98.1% (16.0 pp)	24.0 pp (97.4%)	98.4% (16.0 pp; λ*=0.05; 1.00×)
Ironic / Sarcastic Veiled Hate Share (r = 0.104)	16.7% (−16.1 pp)	0.0%	96.1% (11.8 pp)	12.0 pp (96.4%)	96.1% (11.8 pp; λ*=0.25; 1.01×)

Note: Computed via 800 random labeled/unlabeled splits (n = 150 gold-standard human anchor; N − 150 unlabeled ChatGPT predictions) following Angelopoulos et al..[44,45] When surrogate correlation r is near zero (Conversational Escalation r = 0.070; Persuasion r = 0.064), unclipped Standard PPI (λ = 1) inflates CI width above human-only coding, whereas Power-Tuned PPI++ shrinks λ → 0.05–0.06 to guarantee variance non-inflation. Coverage is estimated from 800 splits of one finite item set per task (about 500 items).*

3.2. Tier 2: multi-rater human psychometrics, the ethnolinguistic fidelity cliff, and 841 native memos

Tier 2 first establishes psychometric baselines on human native-speaker ground truth, before any audit of LLMs across Indonesian and indigenous archipelagic registers. The baselines rest on N = 960 multi-rater observations (Table 3, Panel A). In Phase 1 and the 4-Rater Recap panel (n = 300 items), independent Indonesian raters achieve Nominal Krippendorff's α = 0.4746 (Fleiss' κ = 0.4740; exact pairwise agreement = 75.89%). Quadratic-Ordinal Krippendorff's α rises to 0.6872, with 96.67% pairwise agreement within ±1 rubric grade. Phase 2 is harder. It uses more complex open-ended multi-turn generation tasks (n = 360 items across three raters with mild severity variation: mean 1–4 scores of 1.91 for R1, 2.04 for R2, and 2.09 for R3). Nominal α is 0.3308 here, while Quadratic-Ordinal α reaches 0.5157 with 86.48% within-1-grade agreement. This nominal-versus-ordinal divergence suggests that native human coders rarely confuse culturally valid outputs (Grades A/B) with semantic collapse (Grade D). Human disagreement concentrates benignly along stylistic naturalness boundaries instead.

Evaluating 22 global, regional Southeast Asian, and culturally instruction-tuned LLMs uncovers a Decolonial Ethnolinguistic Fidelity Cliff. The evaluation rests on N = 2,640 native-speaker ethnographic ratings in Indonesian, Sundanese, Javanese, and Minangkabau (Table 3, Panel B; Figure 3A). Consider the eight generalist global and regional LLMs lacking local archipelagic tuning (Llama-2-7B/13B-Chat, BLOOMZ-7B1, SeaLLM-7B-Chat, SEA-LION-7B-Instruct, Bactrian-X/ID, and Merak-7B-v4). Their native-speaker acceptable fidelity (Grades A or B) plunges from 40.83% in national Bahasa Indonesia to just 7.92% in Sundanese (83.33% severe Grade D collapse), 3.75% in Javanese (91.67% Grade D collapse), and 2.08% in Minangkabau (91.25% Grade D collapse). Even widely touted regional models fail almost completely. They stumble on Javanese *krama/ngoko* speech levels, Sundanese *undak-usuk basa* diction, or Minangkabau matrilineal idioms. Two examples stand out. SeaLLM-7B-Chat scores 0.0% Grade A/B in Javanese and 3.3% in Minangkabau. Llama-2-13B-Chat scores 3.3% in Javanese and 0.0% in Minangkabau. HC3 linear probability and ordinal score regressions control for log parameter scale (β_scale = +4.04 pp per log-billion parameters, p = 2.11 × 10^{-06}). The generalist models take large indigenous register penalties here, of −39.79 pp in Javanese (t = −11.30, p = 6.17 × 10^{-29}), −34.48 pp in Minangkabau (p = 1.55 × 10^{-22}), and −34.48 pp in Sundanese (p = 1.37 × 10^{-21}). Cultural instruction tuning in Indonesian and regional languages helps. The Cendol × Indigenous Register interaction generates a +48.71 percentage-point recovery in Grade A/B native acceptability (SE = 4.29 pp, p = 2.96 × 10^{-29}) and a +1.409-point jump on the 4-point ethnographic scale (p = 4.69 × 10^{-41}). It lifts the top Cendol 13B and mT5-XXL models to 52.50% acceptability in Javanese, 57.50% in Sundanese, and 48.3% in Minangkabau, below the 60.00% average across all Cendol variants (Figure 3A).

We analyzed all 841 verbatim native-speaker error memos qualitatively: Minangkabau (492 memos), Sundanese (253), and Javanese (96). The result exposes the exact linguistic mechanisms of machine failure (Table 3, Panel C; Figure 3B). The largest group is half of all memos. Exactly 50.65% (n = 426, mean rubric score 1.00/4) document Complete Semantic Hallucination and Context Collapse, for example '*terlalu banyak yang hilang*' or fabricated topics unrelated to the source clause. Another 3.45% (n = 29, score 1.00/4) expose Untranslated Source / Register Echo Failures. Models parrot the Sundanese or Javanese prompt verbatim. Comprehension is absent. A further 19.86% (n = 167, mean score 2.63/4) identify Pragmatic Diction Shifts and Cultural False-Friend Distortions, especially where lexical overlap misleads tokenizers: Minangkabau '*maambik*' ('to take or claim') is hallucinated as Indonesian '*memainkan*' ('to play'), which inverts agency in social-conflict narratives. The remaining categories are 1.07% Partial Code-Switching Omissions and 24.97% (n = 210, mean score 3.90/4) Verified Ethnographic Equivalences. Auxiliary evidence comes from 2,340 MABL figurative metaphors (600 Javanese, 600 Sundanese, 1,140 Indonesian; e.g., Javanese '*rupane kaya ngemut glali*' ['smiling sweetly as if sucking palm-sugar taffy']) and from 14,981 IndoMMLU cultural and regional-language items. These sources show why literal surface token matching cannot capture archipelagic pragmatics. They also show why residual 32.5%–38.3% Grade D error rates, even in culturally tuned 13B/XXL models, still require statistical debiasing.

Table 3. Multi-Rater Human Psychometric Benchmarks (N = 960), 4-Language Native Ethnographic Fidelity Regressions Across 22 LLMs (N = 2,640), and Taxonomy of 841 Verbatim Error Memos (Tier 2: IndoNLP/Cendol)

Human rater panel	Sample	Nominal α	Fleiss κ	Ordinal α	Exact agreement | within ±1 grade
Panel A: Multi-rater human ground-truth psychometrics					
Human Rater Panel: PHASE1	N = 300 (3 raters)	Nominal α = 0.4746	Fleiss κ = 0.4740	Ordinal α = 0.6872	Exact = 75.9% | Within ±1 Grade = 96.7%
Human Rater Panel: PHASE2	N = 360 (3 raters)	Nominal α = 0.3308	Fleiss κ = 0.3301	Ordinal α = 0.5157	Exact = 57.5% | Within ±1 Grade = 86.5%
Human Rater Panel: RECAP4	N = 300 (3 raters)	Nominal α = 0.4746	Fleiss κ = 0.4740	Ordinal α = 0.6872	Exact = 75.9% | Within ±1 Grade = 96.7%
Panel B: 4-language ethnographic fidelity HC3 regressions (N = 2,640 ratings; 22 LLMs; R^2 = 0.2053 for the linear probability model and 0.2416 for the ordinal score; reference: generalist LLM on national Indonesian)					
Regression covariate	**Sample**	**LPM β (Grade A/B)**	**HC3 SE (p value)**	**Ordinal score β (p value)**	**Interpretation**
Intercept (Indonesian Generalist)	N = 2,640	+32.65 pp	SE = 3.62 pp (p = 3.40 × 10^{-19})	+2.041 (p = 1.43 × 10^{-117})	Baseline / capacity scaling effect
Javanese Register	N = 2,640	−39.79 pp	SE = 3.52 pp (p = 6.17 × 10^{-29})	−1.197 (p = 7.67 × 10^{-47})	Severe indigenous penalty reversed by cultural tuning
Minangkabau Register	N = 2,640	−34.48 pp	SE = 3.50 pp (p = 1.55 × 10^{-22})	−1.035 (p = 2.72 × 10^{-36})	Severe indigenous penalty reversed by cultural tuning
Sundanese Register	N = 2,640	−34.48 pp	SE = 3.58 pp (p = 1.37 × 10^{-21})	−0.968 (p = 3.18 × 10^{-29})	Severe indigenous penalty reversed by cultural tuning
Culturally Tuned (Cendol)	N = 2,640	+6.74 pp	SE = 4.09 pp (p = 9.96 × 10^{-02})	+0.120 (p = 2.17 × 10^{-01})	Baseline / capacity scaling effect
Cendol x Indigenous Register	N = 2,640	+48.71 pp	SE = 4.29 pp (p = 2.96 × 10^{-29})	+1.409 (p = 4.69 × 10^{-41})	Severe indigenous penalty reversed by cultural tuning
Log Parameter Scale (B)	N = 2,640	+4.04 pp	SE = 0.85 pp (p = 2.11 × 10^{-06})	+0.103 (p = 1.07 × 10^{-06})	Baseline / capacity scaling effect
Panel C: Qualitative taxonomy of 841 native error memos (Javanese 96, Minangkabau 492, Sundanese 253)					
Memo category	**Sample**	**Mean grade**	**Top language**	**Example grade**	**Verbatim native evidence**
Complete Semantic Hallucination / Context Collapse	n = 426 (50.6%)	Mean Grade = 1.00/4	Top Lang: Minangkabau	Ex. Grade: D	"*terlalu banyak yang hilang*" (Javanese, cendol-mt5-xl-chat)
Verified Ethnographic & Pragmatic Equivalence	n = 210 (25.0%)	Mean Grade = 3.90/4	Top Lang: Minangkabau	Ex. Grade: A	"*sedikit typo tapi masih bisa ditoleransi*" (Javanese, cendol-llama2-7b-chat)
Pragmatic Diction Shift & Cultural False-Friend Distortion	n = 167 (19.9%)	Mean Grade = 2.63/4	Top Lang: Minangkabau	Ex. Grade: B	"*sedikit bagian tidak diterjemahkan*" (Javanese, cendol-llama2-7b-chat)
Untranslated Source / Register Echo Failure	n = 29 (3.5%)	Mean Grade = 1.00/4	Top Lang: Sundanese	Ex. Grade: D	"*tidak diterjemahkan*" (Sundanese, SeaLLM-7B-Chat)
Partial Code-Switching Omission / Clause Truncation	n = 9 (1.1%)	Mean Grade = 2.44/4	Top Lang: Javanese	Ex. Grade: B	"*kategori award sepertinya masih masuk akal jika tidak diterjemahkan*" (Javanese, cendol-mt5-large)

Note: Native-speaker ethnographic evaluations follow a 4-point rubric (A = 4 complete cultural/semantic fidelity; B = 3 minor diction variance preserving core meaning; C = 2 partial semantic distortion; D = 1 complete hallucination or untranslated register echo). Auxiliary benchmarks include 2,340 MABL metaphors and 14,981 IndoMMLU items.

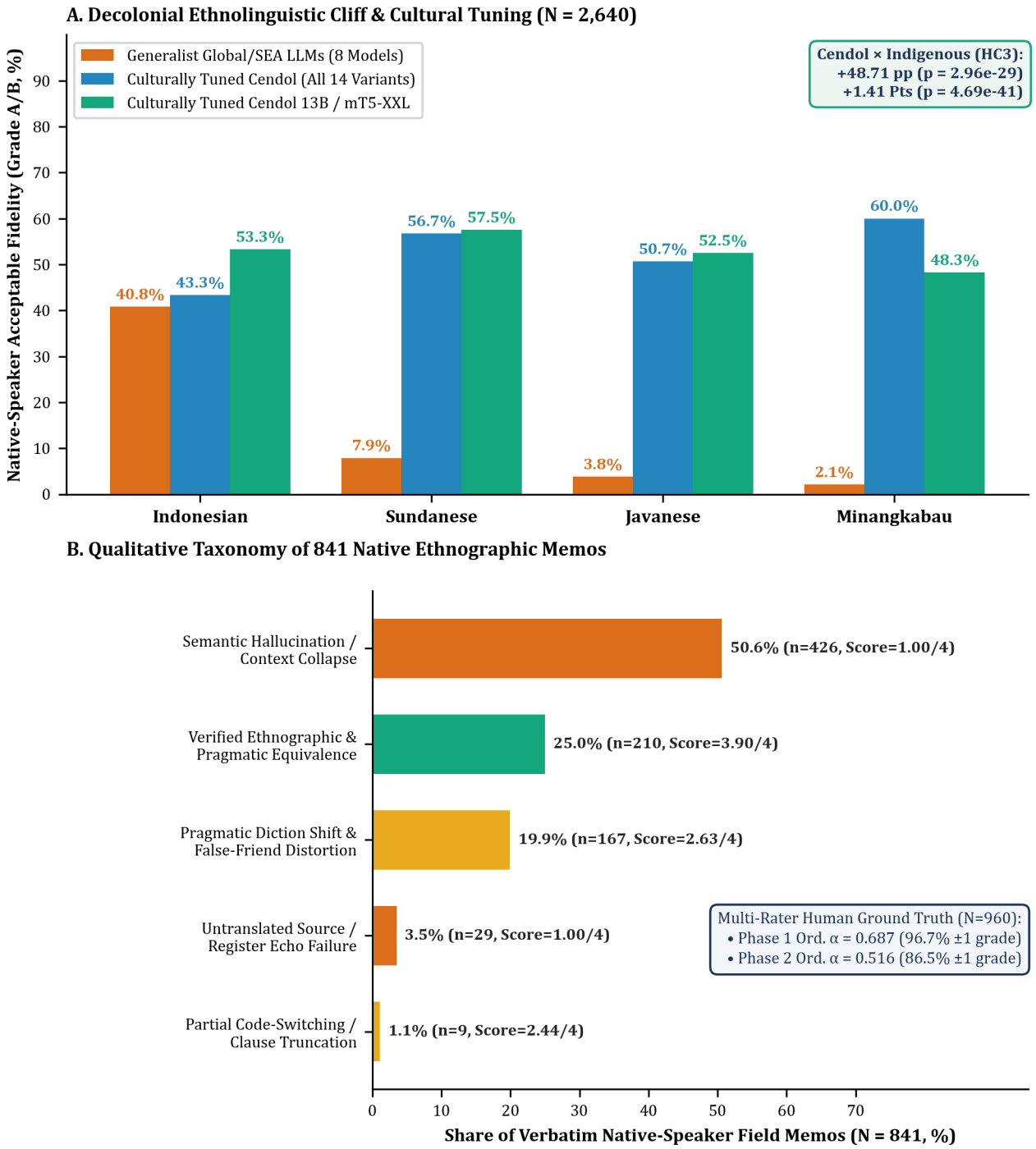


Figure 3. Decolonial ethnolinguistic fidelity across 22 global and culturally instruction-tuned LLMs and qualitative error taxonomy across Indonesian, Sundanese, Javanese, and Minangkabau (Tier 2: IndoNLP/Cendol). (A) Native-speaker acceptable fidelity (Grades A/B, %; N = 2,640), showing severe collapse of generalist LLMs on indigenous registers (2.1%–7.9%) and +48.71 pp HC3 interaction recovery via Cendol cultural instruction tuning ($p = 2.96 \times 10^{-29}$). (B) Empirical taxonomy of 841 verbatim native-speaker error memos alongside multi-rater human ordinal Krippendorff's α benchmarks (N = 960).

3.3. Tier 3: vernacular obfuscation sign reversal, Ethno-PPI++ recovery, AIPW, and RIF quantiles

Tier 3 asks one question. Does uncalibrated surrogate coding corrupt substantive econometric conclusions in real contentious discourse? The data are N = 13,169 expert-coded Indonesian tweets, linked with 15,167 Kamus Alay colloquial/algospeak mappings (Tables 4–5; Figure 4). Colloquial vernacular pervades Indonesian digital contention. Of all tweets in the corpus, 68.79% (9,059/13,169) contain at least one Kamus Alay non-standard orthographic token, and 53.58% exhibit High Vernacular/Algospeak Density (≥20% of tokens). Expert human ground truth identifies 42.23% (5,561 tweets) as Hate Speech, and 40.75% (2,266 tweets) of these represent Implicit Non-Profane Contention (HS = 1 without explicit profanity, Abusive = 0). Panel A of Table 4 shows dismal reliability for a Naive Abusive-Lexicon Rule Surrogate (α = 0.0978). It misses 78.95% of implicit non-abusive hate speech and 44.64% of religious hate speech (HS_Religion). The second surrogate fares little better. A Formal-Register Only Classifier, trained strictly on zero-vernacular formal tweets, reaches α = 0.2647 overall. It still suffers a 59.31% false-negative rate on implicit hate speech, and it erases 66.96% of true religious hate speech. Normalization changes the picture. Ethnographic Kamus Alay orthographic normalization with cross-fitted contextual modeling raises Nominal Krippendorff's α to 0.6859 (accuracy 84.90%, macro-F1 = 0.8430). The implicit and religious false-negative rates fall to 21.18% and 22.82%.

Panel B of Table 4 and Figure 4A reveal an econometric problem. It arises when researchers regress Hate Speech on High Vernacular/Algospeak Density (≥20%), interactional mention/URL density, and log utterance length. Under True Full-Corpus Human Ground Truth (N = 13,169), High Vernacular Code-Switching exerts a highly significant positive effect of +8.70 percentage points on hate-speech probability (SE = 0.86 pp, t = 10.14, p = 4.35×10^{-24}). Now suppose an investigator substitutes the Formal-Register Only Surrogate Classifier. It is applied across all 13,169 tweets. The estimated slope flips sign completely, to −6.96 percentage points (SE = 0.62 pp, t = −11.23, p < 1×10^{-28}). The result is a highly significant estimate that points the wrong way. The Naive Abusive-Lexicon Surrogate errs in the other direction: it inflates the slope by +77% to +15.41 pp (SE = 0.86 pp), because it conflates benign youth slang profanity with group hate speech. We then applied our Ethno-PPI++ estimator. It used only a 3.04% human ethnographic anchor (n = 400 tweets; optimal power-tuning weight λ* = 1.0000), paired with the 12,769 vernacular-contextualized predictions. It recovers $\hat{\beta}$_Ethno-PPI++ = +7.99 percentage points (SE = 3.30 pp, p = 0.0154). The estimate lies within 0.71 pp of the full-corpus 13,169-tweet human parameter (+8.70 pp). Precision improves. Standard errors shrink by 34.1% relative to n = 400 human-only coding (SE = 5.01 pp), which corresponds to a 2.30× Effective Sample Size Multiplier for this slope and anchor: the precision of about 920 human-coded tweets for the cost of 400.

Table 5 and Figure 4B describe the distributional pattern behind this surrogate sign reversal. Panel A covers all N = 5,561 true contentious tweets, and Doubly Robust AIPW achieves near-exact covariate balance, with IPTW max |SMD| reduced from 0.391 to 0.035 and from 0.339 to 0.018. It estimates that Implicit Non-Profane Framing is associated with a +71.14 percentage-point surge in lexicon false-negative erasure (80.2% vs. 9.1%, z = 66.93, VanderWeele E-value = 7.10, RV_q = 58.1%). High Vernacular Code-Switching shows a different association. It is linked to a +19.97 pp surge in formal-classifier false-negative erasure (76.3% vs. 56.4%, z = 15.68, p = 2.14×10^{-55}, E-value = 2.32, RV_q = 18.9%). Vernacular density is linked to more false negatives (Cov(X_vern, f − Y) < 0), so formal surrogate regressions suffer negative non-classical measurement bias of a size that can invert +8.70 pp into −6.96 pp. A second result comes from Firpo–Fortin–Lemieux RIF unconditional quantile regressions (Table 5, Panel B; Figure 4B). Error concentrates around the third quartile of the error distribution. Marginal error effects rise more than threefold from Q10 (q = 0.162) to Q75 (q = 0.553, spanning the key classification boundary) and decline again at Q90 (β = +0.2431 for vernacular share). The peaks are β_Q75 = +0.5473 for vernacular share (p = 1.92×10^{-145}). They reach +0.3396 for collective group contention (HS_Group, p = 7.62×10^{-153}) and +0.1833 for religious contention (HS_Religion, p = 4.48×10^{-20}). A parallel analysis of 12,000 NusaX utterances across 10 indigenous Indonesian languages (Figure 5A) confirms large lexical divergence (1 − Jaccard) relative to national Indonesian. It starts at 71.4% in Minangkabau and 73.6% in Banjarese. It then rises to 79.2% in Balinese, 81.5% in Javanese, 81.8% in Sundanese, 84.7% in Acehnese, 87.0% in Madurese, 88.9% in Toba Batak, and 92.3% in Buginese.

Table 4. Classifier Reliability Comparison and Downstream Regression Sign-Reversal Audit Across N = 13,169 Contentious Indonesian Utterances Linked With 15,167 Kamus Alay Vernacular Mappings (Tier 3)

Coding workflow	Overall accuracy	Macro-F1	Nominal α	Implicit HS false negatives	Religious HS false negatives
Panel A: Corpus-wide automated versus contextualized coding					
Naive Abusive-Lexicon Rule Surrogate	54.90%	0.5489	0.0978	78.95%	44.64%
Formal-Register Only Classifier (Zero-Vernacular Pretraining)	70.10%	0.6324	0.2647	59.31%	66.96%
Vernacular-Contextualized Classifier (KamusAlay + Cross-Fit)	84.90%	0.8430	0.6859	21.18%	22.82%
Panel B: Downstream OLS / PPI++ slope audit on hate speech (percentage points)					
Regression covariate	**True full N = 13,169**	**Naive lexicon**	**Formal surrogate**	**Human only n = 400**	**Ethno-PPI++ (λ* = 1.000)**
Intercept	+45.56 pp (SE 1.92, p = 2.3×10^{-121})	+51.43 pp (SE 1.91)	+36.82 pp (SE 1.42)	+41.43 pp (SE 11.26)	+51.53 pp (SE 7.89, p<0.001)
High Vernacular / Algospeak (>=20%)	+8.70 pp (SE 0.86, p = 4.3×10^{-24})	+15.41 pp (SE 0.86)	−6.96 pp (SE 0.62) [sign reversal]	+12.64 pp (SE 5.01)	+7.99 pp (SE 3.30, p = 0.015)
Log Interactional Mentions/URLs	+11.87 pp (SE 0.73, p = 2.3×10^{-58})	+8.04 pp (SE 0.75)	+7.18 pp (SE 0.53)	+15.23 pp (SE 4.43)	+14.28 pp (SE 2.92, p<0.001)
Log Utterance Token Length	−5.81 pp (SE 0.62, p = 1.2×10^{-20})	−3.31 pp (SE 0.63)	−8.62 pp (SE 0.44)	−5.38 pp (SE 3.62)	−8.40 pp (SE 2.57, p = 0.001)

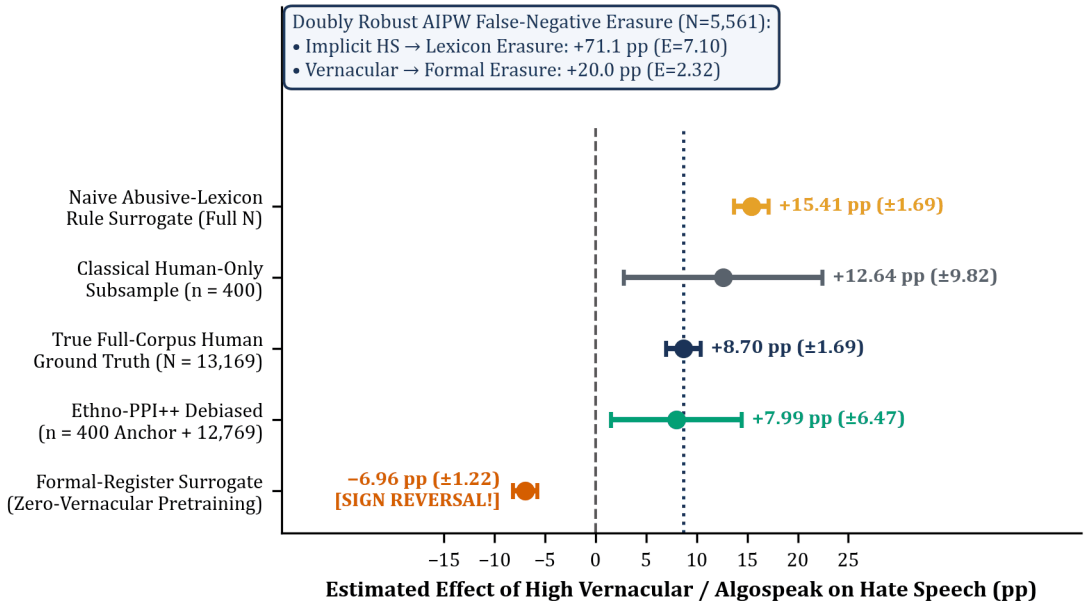
Note: True hate-speech prevalence is 42.23% (N = 5,561/13,169), of which 40.75% (2,266) is implicit non-abusive contention; 68.79% of tweets contain ≥1 Kamus Alay vernacular token. Ethno-PPI++ reduces standard error on High Vernacular from 5.01 pp (human n = 400) to 3.30 pp (2.30× effective sample size gain) while eliminating the −6.96 pp formal-surrogate sign reversal.

Table 5. Doubly Robust AIPW False-Negative Erasure Estimates (N = 5,561 True Contentious Tweets) and Firpo–Fortin–Lemieux RIF Unconditional Quantile Regressions of Calibration Error (N = 13,169, Tier 3)

Exposure → outcome	Potential outcomes E[Y(1)] vs. E[Y(0)]	AIPW ATE (95% CI)	z (p value)	Max |SMD| raw → weighted	VanderWeele E-value / RV_q
Panel A: Doubly robust AIPW erasure among true hate speech (N = 5,561)					
Implicit Non-Profane Framing → Lexicon Erasure	80.2% vs. 9.1%	+71.14 pp (+69.1, +73.2)	z = 66.93 ($p < 1 \times 10^{-99}$)	0.391 → 0.035	E = 7.10 (CI 6.80); RV_q = 58.1%
High Vernacular Code-Switching (≥20%) → Formal Erasure	76.3% vs. 56.4%	+19.97 pp (+17.5, +22.5)	z = 15.68 ($p = 2.14 \times 10^{-55}$)	0.339 → 0.018	E = 2.32 (CI 2.16); RV_q = 18.9%
Panel B: RIF unconditional quantile regressions of calibration error (N = 13,169)					
Quantile of |Y_HS − p̂_formal|	**Unconditional q_τ**	**Vernacular share β (SE)**	**Vernacular p value**	**HS_Group β (SE, p)**	**HS_Religion β (SE, p)**
Unconditional Quantile Q10 (τ = 0.10)	q = 0.1615	+0.1605 (0.0065)	$p = 5.79 \times 10^{-130}$	+0.0467 (0.0017, $p = 6.4 \times 10^{-162}$)	+0.0256 (0.0024, $p = 7.9 \times 10^{-26}$)
Unconditional Quantile Q25 (τ = 0.25)	q = 0.2177	+0.2068 (0.0071)	$p = 2.20 \times 10^{-181}$	+0.0913 (0.0020, $p<1 \times 10^{-300}$)	+0.0460 (0.0031, $p = 3.9 \times 10^{-49}$)
Unconditional Quantile Q50 (τ = 0.50)	q = 0.3238	+0.2962 (0.0126)	$p = 1.59 \times 10^{-119}$	+0.2547 (0.0045, $p<1 \times 10^{-300}$)	+0.1135 (0.0076, $p = 1.6 \times 10^{-50}$)
Unconditional Quantile Q75 (τ = 0.75)	q = 0.5525	+0.5473 (0.0210)	$p = 1.92 \times 10^{-145}$	+0.3396 (0.0127, $p = 7.6 \times 10^{-153}$)	+0.1833 (0.0199, $p = 4.5 \times 10^{-20}$)
Unconditional Quantile Q90 (τ = 0.90)	q = 0.6828	+0.2431 (0.0141)	$p = 8.69 \times 10^{-66}$	+0.1066 (0.0094, $p = 8.5 \times 10^{-30}$)	+0.0817 (0.0151, $p = 5.9 \times 10^{-08}$)

Note: AIPW propensity and outcome models adjust for vernacular share / abusive indicator, log utterance token length, log interactional mentions/URLs, uppercase emphasis ratio, HS_Group, and HS_Religion (300 bootstrap replications). All weighted standardized mean differences satisfy |SMD| < 0.04.

A. Regression Sign Reversal & Ethno-PPI++ Recovery



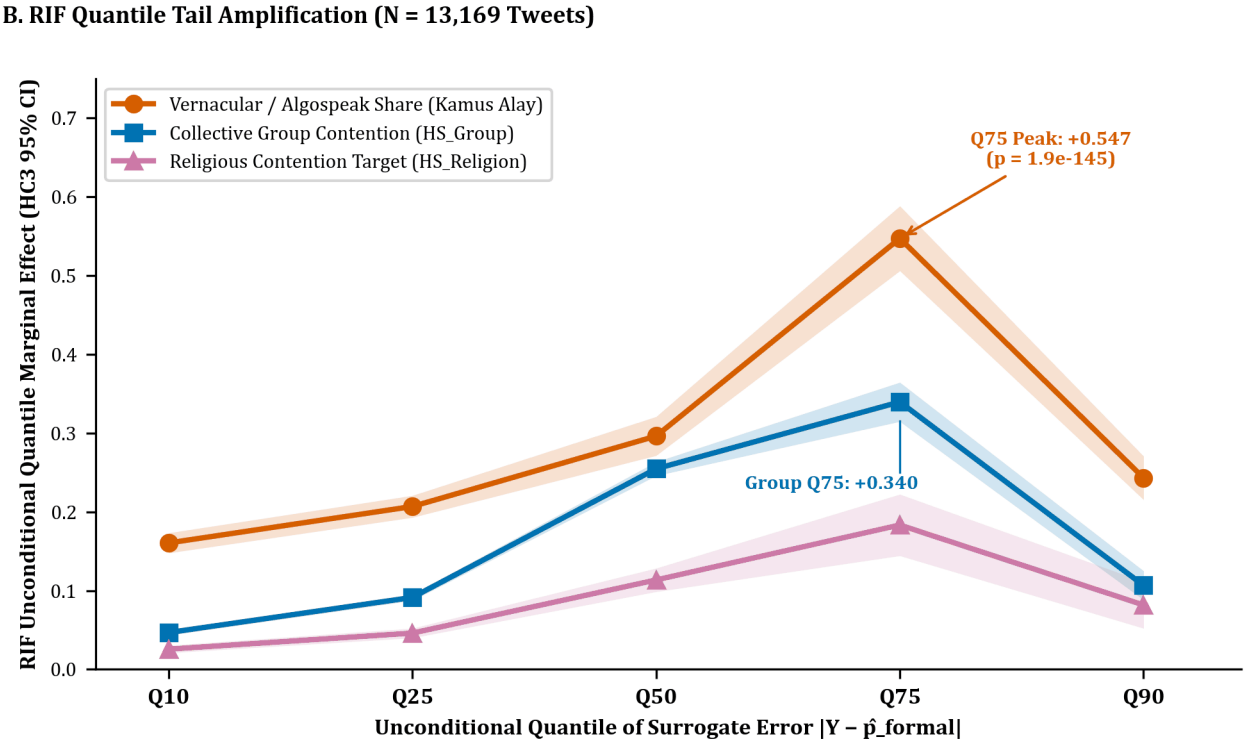


Figure 4. Downstream econometric sign reversal, Doubly Robust AIPW erasure, and Recentered Influence Function (RIF) quantile error dynamics across N = 13,169 contentious Indonesian utterances linked with 15,167 Kamus Alay vernacular mappings (Tier 3). (A) Estimated slope of high vernacular/algospeak code-switching (≥20%) on hate speech, showing how formal-register surrogates reverse the true +8.70 pp slope to −6.96 pp ($p < 1 \times 10^{-28}$) while n = 400 Ethno-PPI++ recovers +7.99 pp (2.30× effective N). (B) RIF unconditional quantile marginal effects (Q10–Q90) of vernacular share, collective group contention, and religious targeting on surrogate calibration error.

3.4. Tier 4: set-theoretic asymmetry across N = 24 empirical coding regimes (fsQCA)

Tier 4 brings together our global, indigenous, and vernacular findings through exact Ragin four-value Fuzzy-Set Qualitative Comparative Analysis. It covers N = 24 empirically evaluated coding regimes (Table 6; Figures 5B and 6A). Single-condition necessity screening (Table 6, Panel A) points to methodological prerequisites for Valid Downstream Inference (VALID_INFERENCE). Cultural & Register Alignment (CULT_ALIGN: necessity consistency = 1.0000, coverage = 0.6073) and High-Capacity Instruction-Tuned Architecture (SCALE_INST: consistency = 0.9681, coverage = 0.4838) both exceed Ragin's 0.90 necessity threshold, although their relevance of necessity is low (RoN = 0.5113 and 0.2379) because both conditions are present in most regimes. Ethnographic Ground-Truth Anchoring (ETHNO_ANCHOR: consistency = 0.8722, coverage = 0.8444, RoN = 0.8893) and Statistical PPI++ Debiasing (PPI_DEBIAS: consistency = 0.8722, coverage = 0.8722, RoN = 0.9119) behave as highly relevant, near-necessary anchors. Why do CULT_ALIGN and SCALE_INST show only moderate necessity coverage (0.48–0.61)? Many regimes run large models on aligned English corpora, yet downstream inference still fails there, because those regimes omit human calibration and debiasing.

Conservative Quine–McCluskey Boolean minimization makes zero simplifying assumptions on logical remainders. It yields one pathway of conjunctural sufficiency for Valid Downstream Inference (Table 6, Panel B; Figure 6A): CULT_ALIGN * SCALE_INST * ETHNO_ANCHOR * PPI_DEBIAS (pattern '−1111', Solution Consistency = 0.8976, PRI = 0.8766, Solution Coverage = 0.8403). It covers all 10 valid inference regimes across Tiers 1–3 (R02, R04, R06, R07, R14, R16, R19, R21, R22, and R24). No case falls in the upper-left quadrant of Figure 6A. Three regimes (R06, R19, R21) have full solution membership but an outcome of 0.67, which is why solution consistency is 0.8976 and not 1. Propositional Clarity (PROP_CLARITY) drops out. Why? Validated regimes succeed on explicit tasks (R02 MRF, R04 Stance, R06 IBC, R07 Politeness, R16 Indonesian) and also succeed on veiled, figurative, indigenous, and vernacular tasks (~PROP_CLARITY: R14 Figurative Sarcasm/Metaphor, R19 Javanese, R21 Minangkabau, R22 Sundanese, R24 Indonesian Kamus Alay Contention). So the set-theoretic solution indicates that, in these regimes, cultural model alignment, native ethnographic anchoring, and Power-Tuned PPI++ together are associated with valid inference despite ambiguity of implied meaning.

Testing the negated outcome (~VALID_INFERENCE: Epistemic Distortion and Coverage Collapse) points to an asymmetry. Absence of Statistical PPI++ Debiasing (~PPI_DEBIAS) is formally necessary for epistemic distortion (consistency = 0.9034, coverage = 0.9034, RoN = 0.8867). ~ETHNO_ANCHOR follows closely (0.8786, RoN = 0.8899). Three sufficient failure pathways emerge. Overall Solution Consistency = 0.9010, PRI = 0.8871, and Coverage = 0.8793. Pathway P1_Distort (SCALE_INST * ~ETHNO_ANCHOR * ~PPI_DEBIAS, consistency = 0.8722) covers 10 zero-shot frontier/ensemble regimes, including R01, R03, R05, R08, R09, R12, R13, R15, R17, R20. It captures 'Uncalibrated Scale Overconfidence'. Pathway P2_Distort (~PROP_CLARITY * ~ETHNO_ANCHOR * ~PPI_DEBIAS, consistency = 1.0000, PRI = 1.0000) covers 9 veiled, indigenous, and vernacular surrogate regimes, including R10, R18, R23. It captures 'Unanchored Pragmatic and Decolonial Erasure'. Pathway P3_Distort (~PROP_CLARITY * CULT_ALIGN * SCALE_INST * ~PPI_DEBIAS, consistency = 1.0000, PRI = 1.0000) covers R08, R09 4-LLM Ensemble, R11 Unclipped Standard PPI on veiled hate, and R12 Escalation. Neither 4-LLM ensemble voting nor unclipped λ = 1 PPI prevents inferential failure on veiled contention without Power-Tuned PPI++

rectification. That is what the pathways indicate in this sample.

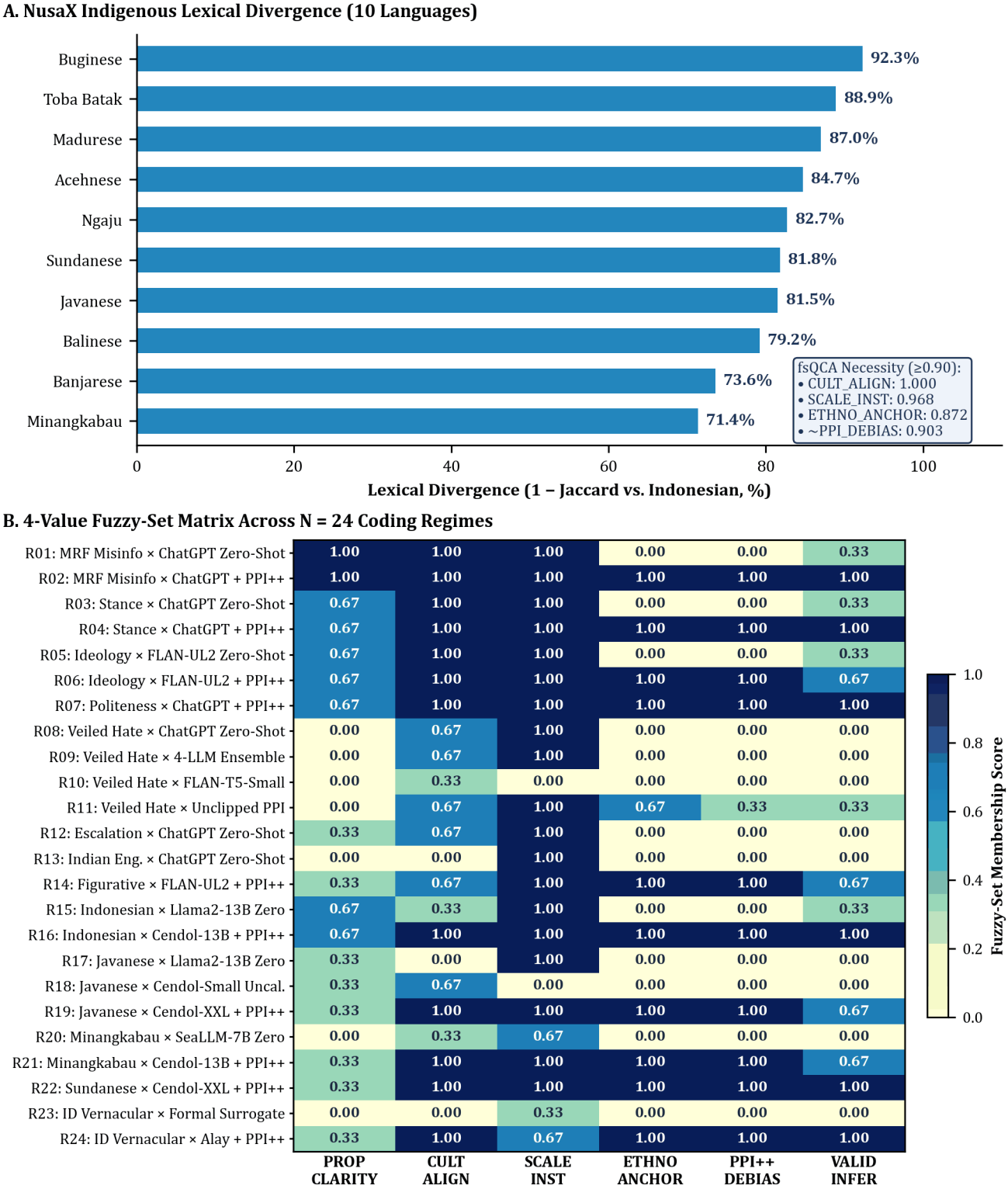


Figure 5. Archipelagic ethnolinguistic divergence and set-theoretic calibration matrix. (A) Lexical divergence (1 – Jaccard token overlap relative to national Indonesian) across 10 indigenous languages in the parallel NusaX pragmatic corpus (N = 12,000), paired with exact fsQCA necessity benchmarks. (B) Four-value fuzzy-set membership heatmap across N = 24 empirically evaluated contentious coding regimes (R01–R24).

Table 6. Exact Ragin Four-Value Fuzzy-Set Qualitative Comparative Analysis (fsQCA): Necessity Benchmarks and Asymmetric Conservative Boolean Sufficiency Pathways Across N = 24 Empirical Coding Regimes (Tier 4)

Necessity condition	Target outcome	Consistency	RoN	Raw coverage	Interpretation
Panel A: Single-Condition Necessity Screening (Threshold ≥ 0.90)					
Necessity: CULT_ALIGN	VALID_INFERENCE	1.0000	RoN = 0.5113	0.6073	Surpasses 0.90 Ragin necessity benchmark
Necessity: SCALE_INST	VALID_INFERENCE	0.9681	RoN = 0.2379	0.4838	Surpasses 0.90 Ragin necessity benchmark

Necessity condition	Target outcome	Consistency	RoN	Raw coverage	Interpretation
Necessity: ETHNO_ANCHOR	VALID_INFERENCE	0.8722	RoN = 0.8893	0.8444	Near-necessary methodological anchor (≥0.87)
Necessity: PPI_DEBIAS	VALID_INFERENCE	0.8722	RoN = 0.9119	0.8722	Near-necessary methodological anchor (≥0.87)
Necessity: ~ETHNO_ANCHOR	~VALID_INFERENCE (Distortion)	0.8786	RoN = 0.8899	0.9010	High-relevance failure condition
Necessity: ~PPI_DEBIAS	~VALID_INFERENCE (Distortion)	0.9034	RoN = 0.8867	0.9034	Surpasses 0.90 necessity for epistemic distortion
Panel B: Conservative Quine–McCluskey Sufficiency Solutions					
Sufficiency pathway	**Target outcome**	**Consistency**	**PRI**	**Raw coverage**	**Unique coverage and empirically covered regimes**
P1 (Valid): CULT_ALIGN * SCALE_INST * ETHNO_ANCHOR * PPI_DEBIAS	VALID_INFERENCE	0.8976	PRI = 0.8766	0.8403	Uniq = 0.8403 | R02, R04, R06, R07, R14, R16, R19, R21, R22, R24
P1 (Distortion): SCALE_INST * ~ETHNO_ANCHOR * ~PPI_DEBIAS	~VALID_INFERENCE	0.8722	PRI = 0.8479	0.6591	Uniq = 0.1719 | R01, R03, R05, R08, R09, R12, R13, R15, R17, R20
P2 (Distortion): ~PROP_CLARITY * ~ETHNO_ANCHOR * ~PPI_DEBIAS	~VALID_INFERENCE	1.0000	PRI = 1.0000	0.6825	Uniq = 0.1953 | R08, R09, R10, R12, R13, R17, R18, R20, R23
P3 (Distortion): ~PROP_CLARITY * CULT_ALIGN * SCALE_INST * ~PPI_DEBIAS	~VALID_INFERENCE	1.0000	PRI = 1.0000	0.2926	Uniq = 0.0249 | R08, R09, R11, R12
Overall Asymmetric Solution Summary	Valid / ~Valid	0.8976 / 0.9010	PRI: 0.8766 / 0.8871	0.8403 / 0.8793	Algebraic elimination of PROP_CLARITY proves ETHNO-PPI overcomes veiled ambiguity

Note: Executed via exact Ragin set-theoretic algebra and conservative Quine–McCluskey prime-implicant minimization (zero counterfactual assumptions on logical remainders; raw consistency threshold ≥ 0.82, PRI ≥ 0.70). The full 24-regime fuzzy matrix is shown in Figure 5B. Necessity of SCALE_INST and CULT_ALIGN has low relevance (RoN 0.2379 and 0.5113). VALID_INFERENCE and PPI_DEBIAS overlap by construction (see Section 2.4).

3.5. Integrative critical mixed-methods synthesis and the five-stage ETHNO-PPI protocol

Table 7 and Figure 6B follow Younas and Fàbregues'[52,53] framework for integrative mixed-methods research questions and structured meta-inference standards.[52,53] Together they synthesize our quantitative psychometric, PPI++, AIPW/RIF, and qualitative ethnographic strands. The result is ETHNO-PPI, a five-stage operational protocol for contentious social-science inquiry. Ethnography is not a preliminary anecdote generator here. Nor are LLMs autonomous replacements for human interpretation. Table 7 presents three modes of critical mixed-methods meta-inference. (1) Expansion in Stage 1: NusaX divergence (71.4%–92.3%) and verbatim Minangkabau/Javanese/Sundanese false-friend memos call for culturally co-designed codebooks, plus vernacular Kamus Alay normalization before modeling. (2) Confirmation in Stage 2: multi-rater ordinal α (0.687/0.516) and rater memos separate benign stylistic variance from Grade D semantic collapse. (3) Discordance Diagnosis and Statistical Rectification across Stages 3–5: 4-LLM unanimous-ensemble error (41.2%–47.4%) and the −6.96 pp formal-surrogate regression sign reversal indicate why culturally tuned screening should be paired with stratified n = 300–400 Power-Tuned PPI++ debiasing and Q75–Q90 qualitative residual auditing.

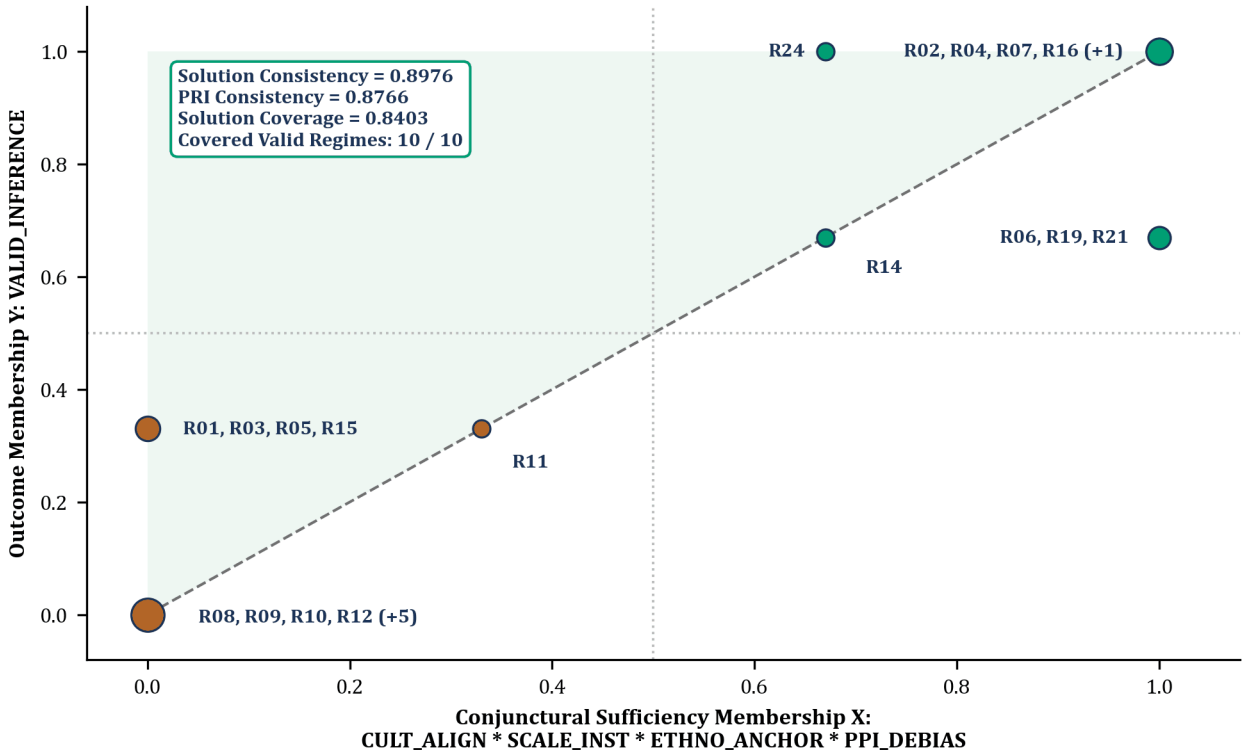
Table 7. Critical Mixed-Methods Joint-Display Matrix: Integrating Psychometric, PPI++, Econometric, Ethnographic Memo, and Set-Theoretic Findings Into the Five-Stage ETHNO-PPI Validation Framework

ETHNO-PPI Stage & Integrative Research Question	Quantitative Psychometric & Econometric Strand (Tiers 1–3)	Qualitative Ethnographic & Set-Theoretic Strand (Tiers 2 & 4)	Critical Meta-Inference & Actionable Governance Rule
Stage 1: Codebook & Vernacular Grounding — How do vernacular obfuscation and regional registers break formal pretraining?	68.79% of Indonesian contentious tweets use Kamus Alay vernacular; NusaX documents 71.4%–92.3% indigenous lexical divergence vs. national Indonesian (10 languages).	841 native memos expose untranslated register echoes (3.5%) and cultural false friends (19.9%, e.g., Minangkabau '*maambik*' ['take'] coded as '*memainkan*' ['play']); MABL documents 2,340 figurative idioms.	EXPANSION: Pre-coding ethnographic lexicon normalization and native register rubrics should precede automated screening.
Stage 2: Multi-Rater Human Audit — How do we separate interpretive pluralism from rubric noise?	Native Indonesian panels (N=960) show Nominal α = 0.475/0.331 vs. Quadratic-Ordinal α = 0.687/0.516 (96.7% and 86.5% within ±1 grade agreement).	Qualitative notes confirm Grade A vs. B variance reflects benign stylistic naturalness ('*sedikit typo tapi masih bisa ditoleransi*'), whereas Grade D (score 1.00) reflects semantic collapse.	CONFIRMATION: Report both nominal and ordinal α plus rater severity profiles; preserve A/B stylistic pluralism rather than forcing artificial single-truth flattening.
Stage 3: Culturally Aligned LLM & Ensemble Screening — Can multi-LLM consensus substitute for human truth?	Frontier-4 α plunges 53.6% on veiled/pragmatic tasks (0.285 → 0.132, p = 0.0042); 4/4 unanimous ensemble errs 41.2%–47.4% on veiled hate, dialect, escalation, and persuasion. Cendol cultural tuning recovers +48.71 pp (p = 2.96×10^{-29}).	fsQCA Distortion Pathway P3 (~PROP_CLARITY * CULT_ALIGN * SCALE_INST * ~PPI_DEBIAS, Cons=1.000: R08, R09, R11, R12) is consistent with ensembles sharing correlated pretraining blind spots.	DISCORDANCE DIAGNOSIS: Do not treat 'inter-LLM agreement' as a validity certificate; select culturally instruction-tuned models but treat all machine labels as fallible proxies.
Stage 4: Power-Tuned PPI++ Debiasing — How do we prevent surrogate coverage collapse and regression sign reversals?	Naive LLM surrogates hit 0.0% 95%-CI coverage on 6/7 tasks; Formal surrogate flips true +8.70 pp vernacular hate-speech slope to −6.96 pp ($p<1 \times 10^{-28}$) via +19.97 pp AIPW erasure. n=400 (3%) Ethno-PPI++ recovers +7.99 pp (2.30× N_eff) & 96.1%–99.1% coverage.	fsQCA Valid Sufficiency (CULT_ALIGN * SCALE_INST * ETHNO_ANCHOR * PPI_DEBIAS, Cons=0.898, Cov=0.840) eliminates PROP_CLARITY, consistent with valid veiled/indigenous inference (R14, R19, R21, R22, R24).	STATISTICAL RECTIFICATION: Couple a stratified n ≥ 300–400 ethnographic gold anchor with Power-Tuned PPI++ ($\lambda^* \in (0, 1)$) before reporting prevalence or regression slopes.

ETHNO-PPI Stage & Integrative Research Question	Quantitative Psychometric & Econometric Strand (Tiers 1–3)	Qualitative Ethnographic & Set-Theoretic Strand (Tiers 2 & 4)	Critical Meta-Inference & Actionable Governance Rule
Stage 5: Tail Diagnostics & Joint-Display Meta-Inference — Where do residual errors concentrate?	RIF quantile slopes peak sharply at Q75 for vernacular share (+0.547, p = 1.9 × 10^{-145}), HS_Group (+0.340), and HS_Religion (+0.183); Implicit HS causes +71.14 pp lexicon erasure (E-value=7.10).	Even top 13B/XXL culturally tuned models retain 32.5%–38.3% Grade D collapse on Minangkabau, Sundanese, and Javanese high-context clauses.	INTEGRATIVE META-INFERENCE: Pair debiased econometric estimates with targeted qualitative audits of Q75–Q90 high-residual cases in a transparent Joint Display.

Note: Structured in accordance with Journal of Mixed Methods Research transparency and integrative meta-inference standards (Younas & Fàbregues;[52] James et al. AI transparency guidance;[51] Delphi meta-inference[53]).

A. Exact Ragin fsQCA Sufficiency Plot (N = 24 Regimes)



B. Five-Stage ETHNO-PPI Critical Mixed-Methods Protocol

Stage 1 — Ethnographic & Vernacular Codebook Grounding

Co-design 4-level rubrics with native speakers; normalize vernacular/algospeak & honorific registers.

Stage 2 — Multi-Rater Human Psychometric Audit (n ≥ 300–400)

Verify Nominal/Ordinal Krippendorff's α, Gwet's AC1 & rater severity; separate boundary nuance from noise.

Stage 3 — Culturally Aligned LLM Screening & Ensemble Audit

Deploy culturally instruction-tuned LLMs (Cendol 13B/XXL); audit unanimous 4-LLM consensus errors.

Stage 4 — Power-Tuned PPI++ / DSL Econometric Debiasing

Estimate optimal $\lambda^* \in [0, 1]$ on stratified anchor; rectify prevalence & regression slopes (valid 95% CI).

Stage 5 — Critical Mixed-Methods Joint-Display Meta-Inference

Triangulate PPI++ parameters, AIPW/RIF tail diagnostics & verbatim native error memos before claims.

Figure 6. Set-theoretic sufficiency verification and the ETHNO-PPI collaborative coding validation architecture. (A) Exact Ragin fuzzy-set X–Y sufficiency plot for CULT_ALIGN * SCALE_INST * ETHNO_ANCHOR * PPI_DEBIAS → VALID_INFERENCE (Solution Consistency = 0.8976, PRI = 0.8766, Coverage = 0.8403, covering all 10 valid-inference regimes, three of which (R06, R19, R21) lie below the diagonal). (B) Five-stage ETHNO-PPI Critical Mixed-Methods Bias-Auditing and Joint-Display Validation Protocol.

4. Discussion

4.1. Beyond the 'human-vs.-LLM accuracy race': pragmatic contention and correlated blindness

Our findings shift methodological debates in computational sociology and political science. Asking whether LLMs 'beat' human coders[2,3] no longer suffices. Diagnosis comes first. Where do the structural boundary conditions of machine interpretation lie[4–9,11]? Across 77,132 predictions in Tier 1, zero-shot LLMs perform best when constructs map onto overt lexical propositions, such as literal misinformation claims or surface emotional valence. They fall off an Explicit–Pragmatic Calibration Cliff ($\Delta\alpha$ = −0.153; cell-level p = 0.0042, task-level p = 0.106) once actors communicate through irony, veiled group grievance, relational power, or conversational escalation.[18,66,67] In contentious politics, actors strategically veil transgressive claims behind shared cultural subtext, sarcasm, and dog-whistles. Their aim is to evade sanction. It is also to mobilize insiders.

Our evidence against 'inter-LLM ensemble validation' matters for empirical practice too. Researchers now increasingly report high agreement between GPT-family and open-weight FLAN/LLaMA models as evidence of coding objectivity.[7,14] Yet agreement is no guarantee. ChatGPT, text-davinci-003, FLAN-UL2, and FLAN-T5-XXL sometimes agree unanimously on a single label (4/4 consensus). In those cases, nearly half of their unanimous classifications are wrong against expert human ground truth: Implicit Veiled Hate (47.4%), Postcolonial Indian English Dialect (46.3%), Conversational Escalation (43.2%), and Collective Persuasion (41.2%). Contemporary foundation models share overlapping English-dominant web corpora and reinforcement-learning safety heuristics.[6,14,68] Multi-model ensembles may therefore converge on identical literalist and WEIRD priors about implied meaning. The result may be correlated epistemic blindness rather than independent replication. Our multi-rater human audit (Table 3) supports Davani et al..[24] Native human coders reach 86.5%–96.7% within-1-grade agreement (Ordinal α = 0.516–0.687), which suggests that human interpretive pluralism reflects meaningful gradation of implied meaning and not random noise.[25,69]

4.2. Decolonizing computational text analysis across archipelagic and vernacular lifeworlds

Tier 2 and Tier 3 provide large-scale econometric and ethnographic evidence consistent with recent linguistic-anthropological critiques of language ideology in AI.[30–33] Global and regional generalist LLMs (including Llama-2-13B-Chat, BLOOMZ-7B1, SeaLLM-7B, and SEA-LION-7B) were deployed zero-shot outside standard national Bahasa Indonesia. Native-speaker acceptable fidelity then collapses. It falls from 40.83% to 7.92% in Sundanese, 3.75% in Javanese, and 2.08% in Minangkabau. Over 83% to 91% of generations were judged complete Grade D hallucinations or untranslated register echoes. NusaX shows 71.4%–92.3% lexical divergence across 10 indigenous Austronesian languages (Figure 5A), and MABL contains 2,340 cultural metaphors. In line with these results, national-language pretraining does not confer cross-register competence across regional lifeworlds, which honorific speech levels (Javanese *krama*/*ngoko*; Sundanese *undak-usuk basa*) and matrilineal proverbs structure.[35,42,70–72]

Our HC3 regression estimates show that decolonial computational adaptation can be achieved technically. Local cultural instruction tuning in Cendol reverses indigenous register penalties by +48.71 percentage points ($p = 2.96 \times 10^{-29}$; +1.409 grade points, $p = 4.69 \times 10^{-41}$). The qualitative coding tempers that result. It covers 841 verbatim native memos, and it warns against technological triumphalism. Even top 13B/XXL culturally tuned models still produce 32.5%–38.3% severe Grade D collapse and 19.86% subtle false-friend distortions, such as translating Minangkabau '*maambik*' ['to take/claim'] as '*memainkan*' ['to play']. Tier 3 adds a further finding. Of the contentious Indonesian social-media discourse, 68.79% uses colloquial '*bahasa alay*' orthographic obfuscation, which serves as youth vernacular identity and as moderation-evading 'algospeak' at the same time.[37,38,40,73] Ignoring vernacular normalization has a cost. Formal-register classifiers then erase 66.96% of religious hate speech and 76.33% of high-vernacular contentious utterances.

4.3. Non-classical measurement error, sign reversals, and the usefulness of power-tuned PPI++

The most sobering lesson for quantitative social scientists is a complete downstream regression sign reversal. It appears on N = 13,169 real Indonesian utterances (Table 4; Figure 4A). A common intuition holds that imperfect automated coding only introduces classical attenuation bias toward zero, which would make significant surrogate estimates 'conservative.' Econometric theory[47–49] says otherwise. The intuition holds only when the error is random and independent of covariates. Social discourse violates that assumption systematically. High Vernacular Code-Switching directly induces +19.97 percentage points of false-negative erasure in a formal-register classifier (Doubly Robust AIPW z = 15.68, $p = 2.14 \times 10^{-55}$, E-value = 2.32). The measurement error (f − Y) therefore covaries negatively and strongly with the explanatory variable itself, and the consequence is a reversal of sign: the true +8.70 pp positive relationship ($p = 4.35 \times 10^{-24}$) becomes a highly significant −6.96 pp negative estimate ($p < 1 \times 10^{-28}$). The bias is not uniform. RIF quantile regressions (Figure 4B) show that surrogate errors rise more than threefold at the Q75 classification boundary for vernacular (+0.547), collective group (+0.340), and religious (+0.183) contention.

Power-Tuned PPI++,[44–46] paired with ethnographic lexicon grounding, resolves both the coverage collapse and the sign-reversing bias. In Tier 1, naive LLM surrogates reach 0.0% 95%-CI coverage on 6 of 7 tasks. Unclipped λ = 1 PPI also inflates variance on noisy conversational escalation (22.1 pp vs. 16.0 pp human-only CI width). PPI++ avoids this problem. It scales λ^* automatically to the human–LLM error covariance and shrinks λ^* to 0.05–0.06 when LLMs fail. On Tier 3's vernacular-contextualized classifier it sets $\lambda^* = 0.8087$ and recovers $\hat{\beta}$ = +7.99 pp (p = 0.0154) from just n = 400 human codes (3.0% of corpus). That is a 2.30× effective sample size multiplier.

4.4. Methodological governance: implementing the five-stage ETHNO-PPI protocol

We combine our 24-regime fsQCA sufficiency formula with the integration standards of the Journal of Mixed Methods Research.[50–53,74] The formula is CULT_ALIGN * SCALE_INST * ETHNO_ANCHOR * PPI_DEBIAS (Consistency = 0.8976, Coverage = 0.8403), and it algebraically eliminates PROP_CLARITY. This combination yields governance rules for journals, reviewers, and research teams (Table 7; Figure 6B). We propose three. First, editors may wish to ask authors to report human validation and calibration for zero-shot or multi-LLM ensemble classifications used as outcome

or explanatory variables in substantive regressions. Second, projects that study multilingual, postcolonial, or online vernacular populations should run Stage 1 ethnographic codebook co-design and colloquial lexicon normalization together with Stage 2 multi-rater human psychometric verification, reporting both nominal and ordinal α across at least two to three native coders. Third, a stratified ethnographic anchor of n = 300–400 human-coded units, oversampling high-vernacular and Q75 decision-boundary strata, offered a workable fieldwork budget in the Tier 3 example. This is large enough to stabilize multi-rater α and λ* covariance estimation in Stage 4 Power-Tuned PPI++/DSL. It also leaves 95%–97% of the coding effort saved on corpora of 10,000+ documents, before the labels used to train the classifier are counted. Joint Displays in Stage 5 close the sequence.

4.5. Limitations and directions for future methodological research

Six scope conditions deserve transparent discussion. First, Tier 1's SALT-NLP archive holds zero-shot generations from 2023–2024 instruction-tuned models (ChatGPT/GPT-3.5, InstructGPT, FLAN-UL2, FLAN-T5), and Tier 2's Cendol archive evaluates 7B–13B open and culturally tuned models. Newer reasoning architectures keep raising baseline accuracy on explicit benchmarks. Even so, recent 2025–2026 audits confirm that source-framing bias, preference instability, and non-WEIRD register gaps persist.[6,9,14,31] The calibration logic of the ETHNO-PPI protocol is therefore not specific to the models tested, but it should be re-evaluated with newer models. Second, our Tier 3 empirical sign-reversal demonstration relies on exact Kamus Alay normalization and cross-fitted contextual classifiers on N = 13,169 contentious Indonesian tweets, and future work should extend multi-rater native anchors to multimodal video/audio meme contention on TikTok and in encrypted messaging groups.[37,75] Third, our PPI++ experiments focus on prevalence means and linear OLS slopes. Extending power-tuned cross-fit debiasing to spatial panel and survival models is a fertile frontier.[45,47] Fifth, Tier 3 contains no LLM: it compares a lexicon rule with TF-IDF classifiers, so its sign-reversal result illustrates surrogate measurement error in automated coding generally. Sixth, all tiers are secondary analyses of existing archives, the Tier 2 ratings cover 30 items per model and language, the fsQCA regimes share data and were coded by the authors, and the study was not pre-registered. Finally, ethical deployment of indigenous-language NLP must respect community data sovereignty. Local cultural tuning should empower marginalized speakers, not surveil them.[30,33]

5. Conclusion

This study draws on 144,865 cryptographically verified public records, spanning global computational social-science benchmarks, four Indonesian and indigenous archipelagic languages, and 13,169 contentious vernacular utterances. In these data, generative LLMs did not replace ethnographic ground truth, but they helped to extend it when statistical debiasing and Critical Mixed-Methods integration governed their use. The failures are large. Zero-shot LLMs suffer a 53.6% calibration cliff on veiled contention that is conveyed through pragmatics, unanimous 4-LLM ensembles were wrong up to 47.4% of the time, and generalist LLMs collapse to 2.1%–7.9% native acceptability on Javanese, Minangkabau, and Sundanese registers. Uncalibrated formal surrogates reverse the true +8.70 pp vernacular hate-speech slope into a −6.96 pp artifact. Three measures alter this. They are cultural model alignment (recovering +48.71 pp native fidelity), a 3% native ethnographic anchor, and Power-Tuned PPI++. Together they restore valid 96%–99% frequentist coverage. They also recover true structural slopes (+7.99 pp, 2.30× effective N) and form the basis of a five-stage ETHNO-PPI protocol for social-science inquiry that needs testing in new settings.

Acknowledgments

The authors gratefully acknowledge the open-science research teams of SALT-NLP (Stanford University and Georgia Institute of Technology), the IndoNLP Consortium (Cendol, NusaX, and IndoMMLU), the MABL collective, and Okky Ibrohim and Indra Budi (Universitas Indonesia) for making their benchmark corpora publicly accessible.

6. Declarations

Statement on the Use of Artificial Intelligence

None. The authors declare that no generative artificial intelligence tools were used in the design, analysis, or writing of this manuscript.

Author Contribution Statement

Conceptualization: F.D. Methodology: F.D. Software: F.D. Validation: G.D., H.U. Formal analysis: F.D. Investigation: F.D., G.D. Resources: F.D. Data curation: F.D. Writing – original draft: F.D. Writing – review & editing: G.D., H.U. Visualization: F.D. Supervision: G.D. Project administration: F.D.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest / Competing Interests

The authors declare no competing interests.

Data Availability Statement

All analyzed data are publicly archived: the SALT-NLP Computational Social Science archive, the IndoNLP/Cendol native-speaker evaluation panels, MABL, IndoMMLU, the Ibrohim and Budi Indonesian Twitter corpus, the Kamus Alay lexicon, and the IndoNLP/NusaX parallel corpus. The SHA-256 manifest, harmonized analytical matrices (Tiers 1 to 4), Python scripts for Krippendorff's α, Power-Tuned PPI++, doubly robust AIPW, RIF quantile regressions and Quine–McCluskey fsQCA, and the JSON analytical outputs are stated to be archived in the project replication package.

Ethics Approval and Consent to Participate

This methodological and empirical study analyzed exclusively de-identified, publicly archived open-science benchmark repositories (the SALT-NLP archive, anonymized native evaluation panels from IndoNLP/Cendol, MABL, IndoMMLU, the anonymized Ibrohim and Budi Indonesian Twitter benchmark, and the IndoNLP/NusaX parallel corpus). No new human-subjects intervention or collection of identifiable private data was conducted, and the study protocol was approved by Ethical Committee of Social Research, Enigma Institute, Indonesia (approval number 2026/299) and in accordance with the Declaration of Helsinki.

Consent for Publication

Not applicable. All analyzed text benchmarks and rater annotations are publicly archived and de-identified.

7. References

1. Ziems C, Held W, Shaikh O, et al. Can Large Language Models Transform Computational Social Science?. Computational Linguistics. 2024;50(1):237-291. doi:10.1162/coli_a_00502
2. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. Proc. Natl. Acad. Sci. U.S.A.. 2023;120(30):e2305016120. doi:10.1073/pnas.2305016120
3. Bail CA. Can Generative AI improve social science?. Proc. Natl. Acad. Sci. U.S.A.. 2024;121(21):e2314021121. doi:10.1073/pnas.2314021121
4. Than N, Fan L, Law T, et al. Updating "The Future of Coding": Qualitative Coding with Generative Large Language Models. Sociological Methods & Research. 2025;54(3):849-888. doi:10.1177/00491241251339188
5. Davidson T, Karell D. Integrating Generative Artificial Intelligence into Social Science Research: Measurement, Prompting, and Simulation. Sociological Methods & Research. 2025;54(3):775-793. doi:10.1177/00491241251339184
6. Lyman A, Hepner B, Argyle LP, et al. Balancing Large Language Model Alignment and Algorithmic Fidelity in Social Science Research. Sociological Methods & Research. 2025;54(3):1110-1155. doi:10.1177/00491241251342008
7. Chae Y, Davidson T. Large Language Models for Text Classification: From Zero-Shot Learning to Instruction-Tuning. Sociological Methods & Research. 2026;55(2):501-567. doi:10.1177/00491241251325243
8. Broska D, Howes M, van Loon A. The Mixed Subjects Design: Treating Large Language Models as Potentially Informative Observations. Sociological Methods & Research. 2025;54(3):1074-1109. doi:10.1177/00491241251326865
9. Chung Y. How Model Choice and Memory Shape Preference Consistency in Large Language Models. Sociological Methods & Research. 2026:00491241261492203. doi:10.1177/00491241261492203
10. Argyle LP, Busby EC, Fulda N, et al. Out of One, Many: Using Language Models to Simulate Human Samples. Polit. Anal.. 2023;31(3):337-351. doi:10.1017/pan.2023.2
11. Bisbee J, Clinton JD, Dorff C, et al. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. Polit. Anal.. 2024;32(4):401-416. doi:10.1017/pan.2024.5
12. Wang Y. On Finetuning Large Language Models. Polit. Anal.. 2024;32(3):379-383. doi:10.1017/pan.2023.36
13. Wasow O. Text as Behavior. Polit. Anal.. 2026:1-15. doi:10.1017/pan.2026.10041
14. Germani F, Spitale G. Source framing triggers systematic bias in large language models. Sci. Adv.. 2025;11(45):eadz2924. doi:10.1126/sciadv.adz2924
15. Lin H, Zhang Y. Navigating the Risks of Using Large Language Models for Text Annotation in Social Science Research. Social Science Computer Review. 2026;44(3):403-427. doi:10.1177/08944393251366243
16. Pangakis N, Wolken S. Keeping Humans in the Loop: Human-Centered Automated Annotation with Generative AI. Proceedings of the International AAAI Conference on Web and Social Media. 2025;19:1471-1492. doi:10.1609/icwsm.v19i1.35883
17. Egami N, Hinck M, Stewart B, et al. Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models. Advances in Neural Information Processing Systems 36. 2023:68589-68601. doi:10.52202/075280-3000
18. Bestvater SE, Monroe BL. Sentiment is Not Stance: Target-Aware Opinion Classification for Political Text Analysis. Polit. Anal.. 2023;31(2):235-256. doi:10.1017/pan.2022.10
19. Wang Y. Topic Classification for Political Texts with Pretrained Language Models. Polit. Anal.. 2023;31(4):662-668. doi:10.1017/pan.2023.3
20. Osnabrügge M, Ash E, Morelli M. Cross-Domain Topic Classification for Political Texts. Political Analysis. 2023;31(1):59-80. doi:10.1017/pan.2021.37
21. Halterman A. Synthetically generated text for supervised text analysis. Polit. Anal.. 2025;33(3):181-194. doi:10.1017/pan.2024.31
22. Wyllie S, Shumailov I, Papernot N. Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias. The 2024 ACM Conference on Fairness, Accountability, and Transparency. 2024:2113-2147. doi:10.1145/3630106.3659029
23. Alvarez RM, Morrier J. Measuring the Quality of Answers in Political Q&As with Large Language Models. Polit. Anal.. 2026;34(1):78-95. doi:10.1017/pan.2025.10007
24. Davani AM, Díaz M, Prabhakaran V. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. Transactions of the Association for Computational Linguistics. 2022;10:92-110. doi:10.1162/tacl_a_00449
25. Cole R. Inter-Rater Reliability Methods in Qualitative Case Study Research. Sociological Methods & Research. 2024;53(4):1944-1975. doi:10.1177/00491241231156971
26. Vila-Henninger L, Dupuy C, Van Ingelgom V, Caprioli M, Teuber F, Pennetreau D, et al. Abductive Coding: Theory Building and Qualitative (Re)Analysis. Sociological Methods & Research. 2024;53(2):968-1001. doi:10.1177/00491241211067508
27. Jerolmack C. What Good is Qualitative Literacy Without Data Transparency?. Sociological Methods & Research. 2023;52(2):1059-1072. doi:10.1177/00491241221140429
28. Heide-Jørgensen T, Eady G, Rasmussen A. Understanding the success and failure of online political debate: Experimental evidence using large language models. Sci. Adv.. 2025;11(30):eadv7864. doi:10.1126/sciadv.adv7864
29. Argyle LP, Bail CA, Busby EC, Gubler JR, Howe T, Rytting C, et al. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. Proc. Natl. Acad. Sci. U.S.A.. 2023;120(41):e2311627120. doi:10.1073/pnas.2311627120
30. Jones GM, Satran S, Satyanarayan A. Toward cultural interpretability: A linguistic anthropological framework for describing and evaluating large language models. Big Data & Society. 2025;12(1):20539517241303118. doi:10.1177/20539517241303118
31. Gasparotto M. "Missing standardization": Identifying harmful language ideologies in natural language processing work. Big Data & Society. 2026;13(1):20539517251406184. doi:10.1177/20539517251406184
32. Shan X, Teng Y, Wang Y, et al. Collectivism and individualism political bias in large language models: A two-step approach. Big Data & Society. 2025;12(2):20539517251343861. doi:10.1177/20539517251343861
33. Törnberg P, Söderström O, Barella J, et al. Artificial intelligence and the state: Seeing like an artificial neural network. Big Data & Society. 2025;12(2):20539517251338773. doi:10.1177/20539517251338773
34. Arsenault AC, Kreps S. Whose voice counts? The role of large language models in public commenting. Big Data & Society. 2026;13(1):20539517261419341. doi:10.1177/20539517261419341
35. Dahlberg S, Dürlich L, Axelsson S, et al. Using Multilingual Language Technology to Classify Open-Ended Survey Responses: Conceptions of Democracy in a Cross-Cultural Survey Setting. Polit. Anal.. 2026:1-18. doi:10.1017/pan.2026.10043
36. Vinhas O, Bastos M. The WEIRD governance of fact-checking and the politics of content moderation. New Media & Society. 2025;27(5):2768-2787. doi:10.1177/14614448231213942
37. Steen E, Yurechko K, Klug D. You Can (Not) Say What You Want: Using Algospeak to Contest and Evade Algorithmic Content Moderation on TikTok. Social Media + Society. 2023;9(3):20563051231194586. doi:10.1177/20563051231194586
38. Pateda L, Hula IRN, Ana Mariana. Algospeak: The Evolution of Language in the Age of Social Media in Indonesia. ELOQUENCE : Journal of Foreign Language. 2025;4(3):391-408. doi:10.58194/eloquence.v4i3.3013
39. Aji AF, Winata GI, Koto F, et al. One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022:7226-7249. doi:10.18653/v1/2022.acl-long.500
40. Gunawan F. *Bahasa alay: refleksi sebuah budaya* [Alay language: a reflection of a culture]. Adabiyyāt: Jurnal Bahasa dan Sastra. 2011;10(2):365-386. doi:10.14421/ajbs.2011.10207
41. 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

42. 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
43. Asteka P, Nisya RK, Rahmawati IS. Politeness in language in the undak usuk basa as a local wisdom of the Sundanese people. KEMBARA Journal of Scientific Language Literature and Teaching. 2024;10(2):761-777. doi:10.22219/kembara.v10i2.36031
44. Angelopoulos AN, Bates S, Fannjiang C, et al. Prediction-powered inference. Science. 2023;382(6671):669-674. doi:10.1126/science.adi6000
45. Angelopoulos AN, Duchi JC, Zrnic T. PPI++: Efficient prediction-powered inference. Ann. Appl. Stat.. 2026;20(3):2235-2252. doi:10.1214/26-aoas2215
46. Song Y, Kluger DM, Parikh H, et al. Demystifying Prediction-Powered Inference. Harvard Data Science Review. 2026;8(4). doi:10.1162/99608f92.f95a1dd2
47. Rister Portinari Maranca A, Chung J, Hinck M, et al. Correcting the Measurement Errors of AI-Assisted Labeling in Image Analysis Using Design-Based Supervised Learning. Sociological Methods & Research. 2025;54(3):984-1016. doi:10.1177/00491241251333372
48. Fong C, Tyler M. Machine Learning Predictions as Regression Covariates. Polit. Anal.. 2021;29(4):467-484. doi:10.1017/pan.2020.38
49. Knox D, Lucas C, Cho WKT. Testing Causal Theories with Learned Proxies. Annu. Rev. Polit. Sci.. 2022;25(1):419-441. doi:10.1146/annurev-polisci-051120-111443
50. Rohlfing I. An Exploration of the Presence of Positive-Results Bias in Qualitative Comparative Analysis. Sociological Methods & Research. 2026;55(3):957-980. doi:10.1177/00491241251357948
51. James TG, Guetterman TC, Molina-Azorin JF. Increasing Transparency on the Use of Artificial Intelligence and Automated Tools in Mixed Methods Studies. Journal of Mixed Methods Research. 2024;18(4):425-429. doi:10.1177/15586898241287294
52. Younas A, Fàbregues S. Advancing the Classification of Mixed Methods Research Questions. Journal of Mixed Methods Research. 2026;20(4):422-433. doi:10.1177/15586898251413451
53. Pratte G, Bourque CJ, Beaudoin AJ, et al. Using a Delphi Consensus Technique to Refine Meta-Inference in Mixed Methods Research. Journal of Mixed Methods Research. 2026:15586898261427560. doi:10.1177/15586898261427560
54. Guetterman TC, Fetters MD, Creswell JW. Integrating Quantitative and Qualitative Results in Health Science Mixed Methods Research Through Joint Displays. The Annals of Family Medicine. 2015;13(6):554-561. doi:10.1370/afm.1865
55. Fetters MD, Curry LA, Creswell JW. Achieving Integration in Mixed Methods Designs—Principles and Practices. Health Services Research. 2013;48(6pt2):2134-2156. doi:10.1111/1475-6773.12117
56. 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
57. 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
58. 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
59. Ibrohim MO, Budi I. Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter. Proceedings of the Third Workshop on Abusive Language Online. 2019:46-57. doi:10.18653/v1/w19-3506
60. Aliyah Salsabila N, Ardhito Winatmoko Y, Akbar Septiandri A, et al. Colloquial Indonesian Lexicon. 2018 International Conference on Asian Language Processing (IALP). 2018:226-229. doi:10.1109/ialp.2018.8629151
61. Hayes AF, Krippendorff K. Answering the Call for a Standard Reliability Measure for Coding Data. Communication Methods and Measures. 2007;1(1):77-89. doi:10.1080/19312450709336664
62. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in Medicine. 2015;34(28):3661-3679. doi:10.1002/sim.6607
63. Cinelli C, Hazlett C. Making Sense of Sensitivity: Extending Omitted Variable Bias. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2020;82(1):39-67. doi:10.1111/rssb.12348
64. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. Annals of Internal Medicine. 2017;167(4):268-274. doi:10.7326/m16-2607
65. Basurto X, Speer J. Structuring the Calibration of Qualitative Data as Sets for Qualitative Comparative Analysis (QCA). Field Methods. 2012;24(2):155-174. doi:10.1177/1525822x11433998
66. Zhang Y, Zou C, Lian Z, et al. SarcasmBench: Towards Evaluating Large Language Models on Sarcasm Understanding. IEEE Transactions on Affective Computing. 2025;16(4):2560-2578. doi:10.1109/taffc.2025.3604806
67. Zeng J, Yang L, Wang Z, et al. Sheep's Skin, Wolf's Deeds: Are LLMs Ready for Metaphorical Implicit Hate Speech?. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025:16657-16677. doi:10.18653/v1/2025.acl-long.814
68. 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
69. Aroyo L, Welty C. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. AI Magazine. 2015;36(1):15-24. doi:10.1609/aimag.v36i1.2564
70. Koto F, Mahendra R, Aisyah N, et al. IndoCulture: Exploring Geographically Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces. Transactions of the Association for Computational Linguistics. 2024;12:1703-1719. doi:10.1162/tacl_a_00726
71. Handoko H, Kaur S, Su Kia L. Cultivating Sustainability: A Cultural Linguistic Study of Minangkabau Environmental Proverbs. JURNAL ARBITRER. 2024;11(1):72-84. doi:10.25077/ar.11.1.72-84.2024
72. Prayoga MB, Ermawan BR, Fadhillah AR, et al. Evaluating Machine Translation Models and LLMs for Indonesian–Javanese Translation Across Speech Levels. Journal of Applied Informatics and Computing. 2026;10(2):1549-1560. doi:10.30871/jaic.v10i2.12326
73. Barik AM, Mahendra R, Adriani M. Normalization of Indonesian-English Code-Mixed Twitter Data. Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019). 2019:417-424. doi:10.18653/v1/d19-5554
74. Johnson RE, Grove AL, Clarke A. Pillar Integration Process: A Joint Display Technique to Integrate Data in Mixed Methods Research. Journal of Mixed Methods Research. 2019;13(3):301-320. doi:10.1177/1558689817743108
75. Sonni AF. AI-based disinformation and hate speech amplification: analysis of Indonesia's digital media ecosystem. Front. Commun.. 2025;10:1603534. doi:10.3389/fcomm.2025.1603534