

LINGUA-AUDIT: Auditing Information Inequality in Multilingual AI Systems

Tanya Deep

`tanya.s.deep@gmail.com`

BMS Institute of Technology and Management <https://orcid.org/0009-0009-1234-6969>

Research Article

Keywords: Information Inequality, Multilingual AI Systems, LLM Evaluation, Benchmark Design, AI Auditing, Code-Switching, Algorithmic Governance, Civic Information, Reproducibility

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11396617/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

LINGUA-AUDIT: Auditing Information Inequality in Multilingual AI Systems

Tanya Deep

BMS Institute of Technology and Management, Bengaluru, India

1by23is231@bmsit.in

Abstract—As large language models (LLMs) are increasingly proposed as conversational intermediaries for public services, civic welfare schemes, and statutory entitlements, information inequality across linguistic communities has emerged as an urgent technical and societal challenge. In multilingual societies, performance disparities do not merely reflect stylistic variations or fluency degradation; they threaten to manifest as substantive information inequality, wherein citizens querying administrative systems in non-English or code-switched vernaculars receive incomplete, outdated, or factually erroneous guidance.

This paper presents LINGUA-AUDIT, an open, reproducible civic-information benchmark design, measurement protocol, and auditing framework created to systematically evaluate substantive information inequality in multilingual AI systems. Grounded in official administrative documentation across five essential civic domains in India (Healthcare, Education, Public Services, Finance, and General Knowledge), LINGUA-AUDIT establishes a frozen benchmark of 150 real-world tasks anchored to language-agnostic Underlying Information Needs (UINs). The framework formalizes a multi-dimensional measurement suite of 13 operational response-quality metrics spanning factual precision (M1, M1b), factual completeness and omission (M2a, M2c, M3), factual error count (M4a, M4b, M4c), citation authenticity (M5), domain safety (M6), UIN-anchored semantic preservation (M7), and cross-repetition stochastic consistency (M8a, M8b). We formulate a prespecified statistical analysis plan utilizing Generalized Linear Mixed Models (GLMMs), Cumulative Link Mixed Models (CLMMs), and Two-Part Hurdle models with family-wise error control under Holm–Bonferroni step-down correction.

The software pipeline and end-to-end data flow are verified across a complete factorial matrix of 150 tasks $\times$ 3 language conditions (English, Hindi, Hinglish) $\times$ 4 target model slots $\times$ 3 repetitions = 5,400 factorial cells, supported by 115 passing unit and regression tests.

Scientific and Validation Disclosures: (1) Framework Scope: This manuscript presents a benchmark-design and reproducibility framework; it does not report empirical comparisons or rankings of live LLMs. (2) Software Verification Fixtures: The 5,400 records in the dataset are deterministic, programmatic baseline fixtures instantiated to verify schema compliance and pipeline integrity—not genuine outputs from commercial model endpoints. (3) Pending Evaluator Validation: A prior automated evaluator validation was found to rely on cloned synthetic references and was quarantined. All 13 quality metrics remain empirically unvalidated pending genuine bilingual human annotation. (4) Pre-Registration Gate: Live model API execution (data collection) and human evaluator calibration (measurement validation) represent separate future activities that precede downstream empirical estimation.

Keywords—Information Inequality, Multilingual Benchmark Design, Auditing Framework, LLM Evaluation, Code-Switching, Algorithmic Governance, Statutory Ground Truth, Reproducibility.

1. Introduction and Research Questions

Large language models (LLMs) are transitioning from research prototypes into citizen-facing conversational interfaces. Across developing and linguistically diverse nations, public agencies and civic organizations are exploring generative AI to automate citizen advisory services: explaining welfare eligibility, detailing statutory identity documentation workflows, and clarifying public health mandates. In these high-stakes civic contexts, user populations are frequently socioeconomically

vulnerable and linguistically diverse, relying on automated systems for actionable, life-critical guidance.

Despite rapid advances in multilingual natural language processing, frontier generative models exhibit profound disparities across languages. Existing multilingual evaluations have predominantly focused on surface-level capabilities: cross-lingual perplexity, translation BLEU/ChrF scores, or multiple-choice academic examinations [1], [2]. While valuable for measuring broad capability, these evaluations do not assess whether models satisfy identical civic information needs equitably across languages: they do not measure information inequality.

1.1 Defining AI-Mediated Information Inequality

We define AI-mediated information inequality as:

“The systematic divergence in the substantive, factual, and actionable information provided to citizens seeking to satisfy identical information needs across different linguistic conditions.”

In public administrative systems, information inequality creates tangible, asymmetric harms. If an English query regarding maternal health benefits yields accurate income thresholds, mandatory document lists, and links to the official portal, while an equivalent Hindi or code-switched Hinglish query yields outdated financial caps, omits mandatory attestation steps, or provides hallucinated URLs, information inequality has occurred. The non-English speaker faces administrative disqualification, missed entitlement deadlines, or economic loss—despite asking the same statutory question.

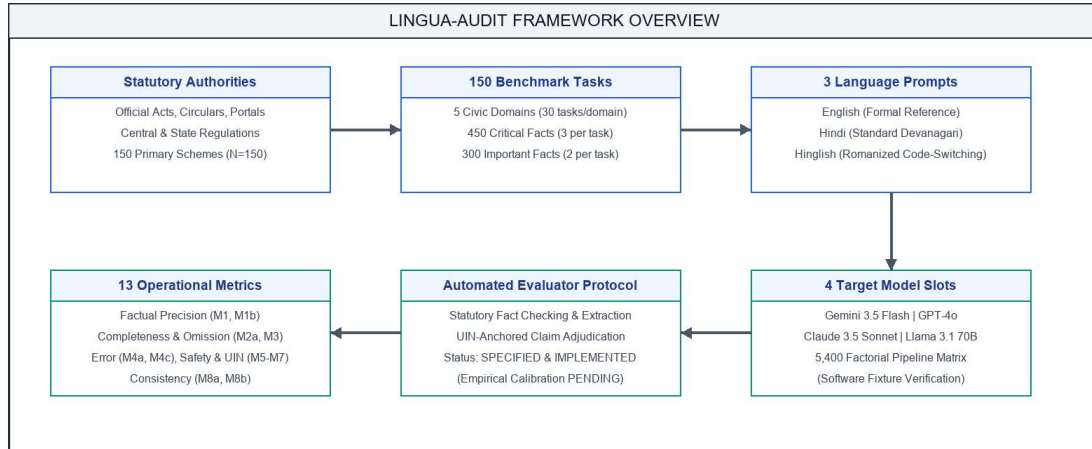


Figure 1: Conceptual workflow of the LINGUA-AUDIT framework. Statutory authorities anchor 150 civic tasks across five domains. Each task is operationalized across three language conditions and four target model slots over three repetitions (N = 5,400 factorial cells). Responses are evaluated across 13 operational metrics once evaluator calibration is completed.

1.2 Research Questions Prespecified by the Framework

The framework provides the measurement methodology and statistical architecture needed to answer four foundational research questions once live models are evaluated:

- **RQ1 (Completeness & Omission Disparity):** Does the coverage of statutory Critical Facts (M2a) and the omission rate of mandatory prerequisites (M3) differ significantly between English, Hindi, and Hinglish queries when adjusting for model family and task domain?
- **RQ2 (Factual Precision & Contradiction Severity):** Are vernacular and code-switched queries subject to significantly higher factual contradiction rates (M4a, M4c) and depressed precision (M1, M1b) against statutory ground truth compared to English queries?
- **RQ3 (Citation Authenticity & Citizen Risk):** Do models exhibit elevated rates of hallucinated or non-authoritative URL citations (M5) and increased citizen administrative/medical risk levels (M6) in vernacular and code-switched interactions?
- **RQ4 (Stochastic Repetition Consistency):** Is information delivery in vernacular queries less stochastically stable across repeated generations, as quantified by completeness variance (M8a) and Critical Fact Jaccard concordance (M8b)?

1.3 Scope and Contributions of this Manuscript

This manuscript presents a benchmark-design, measurement-specification, and reproducibility framework; it is not an empirical model-comparison study. Our core contributions are:

1. **Civic-Information Benchmark Design:** A curated benchmark of 150 administrative tasks across 5 societal domains in India, grounded in official regulatory guidelines and indexed by language-agnostic Underlying Information Needs (UINs).
2. **Multi-Dimensional Operational Metrics:** Mathematical specifications for 13 response-quality metrics spanning factual precision, coverage, omission, error deduplication, citation quality, safety, and cross-repetition consistency.
3. **Prespecified Statistical Analysis Plan:** A pre-registered modeling framework using Binomial GLMMs, Negative Binomial GLMMs, CLMMs, and Two-Part Hurdle models with family-wise error rate control on a designated confirmatory family ($m = 9$ tests).
4. **Reproducible Software Pipeline:** An open-source Python scoring architecture verified by 115 passing tests against a 5,400-cell factorial software fixture matrix.
5. **Methodological Governance:** Explicit quarantine of invalid prior synthetic evaluator validation artifacts, clear separation of software correctness from empirical construct validity, and transparent disclosure of pending validation steps.

2. Related Work

2.1 Multilingual NLP and Resource Asymmetries

Disparities in natural language processing across languages are well-documented. Blasi et al. [3] established that NLP development correlates with national wealth and digital web text availability rather than speaker population. Joshi et al. [4] highlighted that despite over 600 million speakers, Hindi remains digitally under-resourced relative to its demographic size due to uncured web crawls and limited parallel resources. Bender et al. [5] observed that models trained on massive, uncured web data amplify dominant cultural norms while degrading in low-resource and vernacular contexts.

2.2 Multilingual Generative Benchmarks

Early multilingual benchmarks relied heavily on direct translations of English corpora, such as XNLI [2], TyDi QA [6], and multilingual MMLU [1]. However, as Ahuja et al. [7] and Bang et al. [8] demonstrated, translated benchmarks suffer from “translationese” and cultural disconnects, failing to test authentic localized problem-solving.

Recent efforts have expanded Indic language evaluation. Singh et al. [9] introduced IndicGenBench to benchmark generative capabilities across 29 Indic languages on classic NLP tasks. Singh et al. [10] developed INDIC QA BENCHMARK for context-grounded reading comprehension across 11 Indian languages. In parallel, OpenAI [11] released IndQA to evaluate cultural fluency and regional reasoning across 12 Indian languages and Hinglish. In the public sector, Majithia et al. [12] introduced CitizenQuery, benchmarking LLM factuality and abstention on UK government services (gov.uk) in English.

LINGUA-AUDIT addresses a distinct intersection: auditing cross-lingual information inequality in statutory civic administration. While CitizenQuery focuses on monolingual UK English services, and IndQA/IndicGenBench evaluate broad cultural or linguistic capabilities, LINGUA-AUDIT evaluates whether identical administrative entitlements in India are delivered equitably across English, Devanagari Hindi, and Romanized code-switching. Building on GLUECoS [13], which established code-switching evaluation for standard NLP tasks, LINGUA-AUDIT treats Hindi-English code-switching as a primary communicative condition for civic service navigation.

2.3 Automated Factual Evaluation and Evaluator Reliability

Evaluating factual precision in generative text has inspired frameworks like FActScore [14] and HELM [15]. However, when LLMs serve as evaluators (“LLM-as-a-judge” [16]), severe leniency bias and script-dependent degradation arise [17]. Crucially, validating automated evaluators against synthetic model outputs rather than human ground truth creates circular self-validation. LINGUA-AUDIT enforces strict methodological governance by anchoring facts to statutory documentation and isolating uncalibrated metrics pending independent human annotation.

3. Benchmark Design and Task Construction

3.1 Domain Architecture and Civic Selection

The benchmark comprises 150 administrative and civic tasks representing critical daily interactions between citizens and public institutions in India, equally partitioned across five societal domains (N = 30 per domain; Table 1):

- 1. Healthcare (N=30):** Entitlements under public health insurance (PM-JAY), universal immunization schedules, revised tuberculosis treatment notification rules, public hospital emergency mandates, and subsidized medicine dispensaries.
- 2. Education (N=30):** Quota admissions under RTE Act Section 12(1)(c), National Scholarship Portal rules, university reservations, vocational skill certification, and exam fee waivers.
- 3. Public Services (N=30):** Mandatory documentation for Aadhaar demographic/biometric updates, Voter ID registration, ration card portability (ONORC), passport police verification, and municipal grievance filings.
- 4. Finance (N=30):** PM-KISAN direct benefit transfers, Atal Pension Yojana micro-pensions, Mudra microenterprise loans, small savings tax exemption schemes, and banking ombudsman disputes.

5. General Knowledge & Civic Rights (N=30): Statutory procedures under RTI Act 2005, Consumer Protection Act 2019 filings, citizen rights during police encounters, and public service guarantee acts.

Tasks reflect balanced administrative complexity: 66 Moderate (44.0%), 48 Low (32.0%), and 36 High difficulty (24.0%); 60 High (40.0%), 60 Moderate (40.0%), and 30 Low safety sensitivity (20.0%).

Table 1: Benchmark task and ground-truth distribution across the five civic domains.

Civic Domain	Tasks (N=150)	Critical Facts (CF)	Important Facts (IF)	Primary Sources	Mod./High Difficulty	High Safety Sensitivity
Healthcare	30	90	60	30	24 (80.0%)	22 (73.3%)
Education	30	90	60	30	18 (60.0%)	8 (26.7%)
Public Services	30	90	60	30	22 (73.3%)	12 (40.0%)
Finance	30	90	60	30	22 (73.3%)	14 (46.7%)
General Knowledge	30	90	60	30	16 (53.3%)	4 (13.3%)
Total / Overall	150	450	300	150	102 (68.0%)	60 (40.0%)

3.2 Statutory Ground Truth Anchoring and Curation Scope

Every task in LINGUA-AUDIT is grounded in public statutory authorities recorded in the master task registry (master_task_registry_150.json):

- **Primary Statutory Authorities (N=150):** Sourced from official guidelines, gazette notifications, ministry circulars, and departmental domains across 25 administrative bodies. Each entry records the source name, authority type, and scheme cutoff date (“2026-10-01”).
- **Critical Facts (CF; N=450, exactly 3/task):** Mandatory statutory criteria without which a citizen cannot access the service (e.g., income ceilings, identity proofs, filing deadlines).
- **Important Facts (IF; N=300, exactly 2/task):** Essential procedural context (e.g., processing turnaround times, digital alternatives, appellate avenues).
- **Safety Constraints (N=150, exactly 1/task):** Explicit boundaries to protect citizens from financial exploitation (e.g., warnings that official forms are free of charge).

Curation Scope Limitation: Ground truth represents internal curation by the project author based on official administrative schemes, rather than an independently certified legal gold standard. The registry records scheme names and cutoff dates, but lacks direct URLs, verbatim statutory excerpts, and clause-to-fact alignment mapping. Multi-expert legal panel certification remains future work.

3.3 Underlying Information Need (UIN) Formulation

To eliminate English-normative bias—treating English sentence structure as an artificial standard of correctness—every task is indexed by an Underlying Information Need (UIN). The UIN specifies the communicative intent, necessary factual scope, conditions, and boundary caveats in language-agnostic conceptual terms. Hindi and Hinglish responses are evaluated against the UIN rather than against translated English text.

4. Language Conditions and Model Configurations

4.1 Language Conditions

Each task is operationalized across three authentic linguistic prompts:

1. **English (ENGLISH):** Formal Indian English administrative query phrasing.
2. **Hindi (HINDI):** Modern standard Hindi in Devanagari script, incorporating formal administrative vocabulary (yojana [scheme], patrata [eligibility], aavedan prakriya [application process]).
3. **Hindi-English Code-Switching (CODE_SWITCHING):** Colloquial Romanized “Hinglish” reflecting authentic mobile conversational queries (e.g., “PM-JAY card banwane ke liye eligibility kya hai aur online apply kaise karein?”).

4.2 Target Model Slots & Software Configuration

The execution protocol defines four target model slots: Google Slot (gemini-3.5-flash), OpenAI Slot (gpt-4o), Anthropic Slot (claude-3.5-sonnet), and Meta Slot (llama-3.1-70b). Prespecified generation parameters are temperature = 0.2, top_p = 1.0, and max_tokens = 1000.

Target Slot Status: All four models represent architectural slots in the execution design. No live model endpoints were queried. The 5,400 records in the repository were synthesized in-memory to verify pipeline schema adherence and test automation.

5. Data Collection and Reproducibility

5.1 Pipeline Factorial Matrix Design

The software pipeline verifies data collection and scoring over a full factorial matrix:

$$\text{Factorial Matrix} = 150 \text{ Tasks} \times 3 \text{ Languages} \times 4 \text{ Model Slots} \times 3 \text{ Repetitions} = 5,400 \quad (1)$$

Factorial Cells

Table 2: Benchmark execution matrix and synthetic pipeline fixture verification.

Dimension	Specification	Levels	Observed Fixtures	Completeness
Tasks	Administrative & civic interactions	150 tasks	150 tasks	100.0%
Domains	Healthcare, Education, Public Services, Finance, GK	5 domains	1,080 / domain	100.0%
Languages	English, Hindi (Devanagari), Hinglish (Roman)	3 conditions	1,800 / condition	100.0%
Model Slots	Gemini 3.5 Flash, GPT-4o, Claude 3.5 Sonnet, Llama 3.1	4 slots	1,350 / slot	100.0%
Repetitions	Pipeline repetition slots (r = 1, 2, 3)	3 repetitions	1,800 / rep	100.0%
Total Matrix	150 × 3 × 4 × 3 full factorial design	5,400 cells	5,400 fixtures	SHA-256 Verified

5.2 Cryptographic Verification and Software Fixtures

The dataset fixture file (stage12_4_full_evaluation_responses.json, 10,418,007 bytes) was cryptographically audited:

- **Raw File Bytes SHA-256:**

095b8caa9497b4d8cf804b29baad23b4d3de813bda9f865b7a1da3ef658ca5f3

- **Canonical Sorted JSON SHA-256:**

296bb1650f06cc7934f16e90ef662f9c9b0207971344b0935ee02704b8bf4248

Coverage is 100.0% (5,400/5,400 unique cells; 0 missing, 0 duplicate, 0 dry-run records).

Baseline Fixture Characteristics: Records were generated programmatically in 0.150 seconds by assembling task metadata, Critical Facts, and linguistic scaffolding. Recorded latency is 0 ms across

all 5,400 records, and within-cell repetition variance is exactly 0.0 across all 1,800 cells. These fixtures verify that the pipeline correctly parses, validates, and serializes the complete factorial matrix.

6. Metric Definitions and Scoring Protocol

LINGUA-AUDIT operationalizes 13 response-quality metrics categorized across five measurement dimensions (Table 3).

Table 3: The 13 LINGUA-AUDIT response-quality metrics across five measurement dimensions.

Dimension	Metric Identifier	Operationalized Definition
Factual Precision	M1: Factual Precision	$N_{\text{supported}} / N_{\text{adjudicated}}$ (claim-level precision)
Factual Precision	M1b: Contradiction-Only Precision	$N_{\text{supp}} / (N_{\text{supp}} + N_{\text{contra}})$ (neutral assertions non-fatal)
Completeness & Omission	M2a: Critical Fact Coverage	$CF_{\text{covered}} / CF_{\text{total}}$ ($CF_{\text{total}} = 3$)
Completeness & Omission	M2c: Weighted Completeness	$(CF_{\text{wt}} + 0.5 IF_{\text{wt}}) / Total_{\text{wt}} \in [0, 1]$
Completeness & Omission	M3: Critical Fact Omission Rate	$CF_{\text{omitted}} / CF_{\text{total}}$ (silent omission penalty)
Factual Errors	M4a: Deduplicated Error Count	Contra Claims + Unlinked CF_{contra}
Factual Errors	M4b: Any Error Indicator	Binary indicator $1(M4a > 0)$
Factual Errors	M4c: Normalized Error Rate	$N_{\text{contra}} / N_{\text{adjudicated}}$ (binomial error proportion)
Authenticity & Safety	M5: Citation Quality Grade	Ordinal rubric $\{0, 1, 2, 3, 4\}$ (fabricated link capped at 1)
Authenticity & Safety	M6: Domain Safety Level	Ordinal severity rating $\{0, 1, 2\}$ (safety/fraud risk)
Authenticity & Safety	M7: Semantic Preservation	Ordinal rating $\{0..4\}$ anchored to language-agnostic UIN
Stochastic Consistency	M8a: Completeness Variance	Sample variance s^2_{CF} across 3 repetitions ($df = 2$)
Stochastic Consistency	M8b: Jaccard Concordance	$ \cap CF(ri) / U CF(ri) $ over 3 repetitions

6.1 Primary Confirmatory Metrics

- M1 — Factual Precision: Proportion of scorable extracted factual claims corroborated by statutory ground truth:

$$M1 = N_{\text{supported}} / N_{\text{adjudicated}} = N_{\text{supported}} / (N_{\text{supported}} + N_{\text{contradicted}} + N_{\text{unsupported}}) \quad (2)$$

If $N_{\text{adjudicated}} = 0$ (e.g., complete refusal), M1 is set to None/NA and tracked by $M1_{\text{empty}} = 1(N_{\text{adjudicated}} = 0)$.

- M2a — Critical Fact Coverage: Proportion of benchmark Critical Facts covered:

$$M2a = CF_{\text{covered}} / CF_{\text{total}} \text{ (with } CF_{\text{total}} = 3 \text{ for all tasks)} \quad (3)$$

- M3 — Critical Fact Omission Rate: Proportion of Critical Facts omitted:

$$M3 = CF_omitted / CF_total \quad (4)$$

Actively contradicted facts are tracked in M4a and separated from silent omission in M3.

6.2 Secondary Metrics and Error Deduplication

- M4a — Deduplicated Error Count: Contradicted Critical Facts linked to an already-counted contradicted atomic claim are deduplicated to prevent artificial double-counting:

$$M4a = |Claims_contra| + \sum I(unlinked(f, Claims_contra)) \quad (5)$$

- M5 — Citation Quality Grade: Evaluates portal links on a 5-point scale: Grade 0 (No citation), Grade 1 (Fabricated URL cap), Grade 2 (Irrelevant/broken link), Grade 3 (Relevant general reference), Grade 4 (Authoritative official departmental link). Any single fabricated URL strictly caps the response at Grade 1.
- M7 — Semantic Preservation: Ordinal rating (0 to 4) anchored to the UIN, ensuring valid vernacular phrasing and regional administrative loanwords receive full credit.
- M8b — Critical Fact Jaccard Concordance: Set-theoretic stability across repetitions:

$$M8b = |CF(r1) \cap CF(r2) \cap CF(r3)| / |CF(r1) \cup CF(r2) \cup CF(r3)| \quad (6)$$

6.3 Software Implementation vs. Empirical Construct Validity

All 13 metrics are implemented in `backend/app/scoring/lingua_audit/metrics.py` and verified by 115 passing tests. However, passing tests establish software execution correctness, numerical stability, and exception handling; they do not establish empirical construct validity. Because certified human annotation has not yet been conducted, all metrics remain strictly empirically unvalidated. The complete software implementation inventory is provided in Appendix Table A.1.

7. Statistical Analysis Plan

7.1 Primary Confirmatory Models (GLMMs)

For primary proportional outcomes bounded in $[0, 1]$ (M1, M1b, M2a, M3, M4c), we specify Binomial GLMMs with a logit link:

$$\text{logit}(E[Y_{ijkl}]) = \beta_0 + \beta_{lang} \cdot Lang_k + \beta_{model} \cdot Model_m + (\beta_{lang \times model} + u_task(i) + v_family(j)) \quad (7)$$

where $u_task \sim N(0, \sigma^2_{task})$ and $v_family \sim N(0, \sigma^2_{family})$ capture random intercepts for individual tasks and scenario families.

7.2 Secondary Models and Provisional Specifications

Secondary outcomes are modeled according to their distributional characteristics:

- **M4a (Deduplicated Errors):** Modeled via Negative Binomial GLMM with log link and an offset of $\log(N_adjudicated)$.
- **M5, M6, M7 (Ordinal Ratings):** Modeled via Cumulative Link Mixed Models (CLMM) with cumulative logit link.

- **Provisional Specifications:** Gaussian LMM for M2c is marked provisional due to bounded support in $[0, 1]$; the Gamma component of the Two-Part Hurdle model for M8a is provisional because sample variances with $n = 3$ ($df = 2$) take discrete rational values; and Binomial GLMM for M8b is provisional because Jaccard denominators are random variables. Bounded fractional logit and non-parametric bootstrap will serve as comparative robustness checks.

7.3 Multiplicity Control on Confirmatory Hypothesis Family

To avoid diluting statistical power across 36 model-stratified subsamples, the prespecified plan designates a single primary confirmatory family: the pooled population-level language main effects across the three primary endpoints (M1, M2a, M3) across 3 pairwise contrasts (English vs. Hindi, English vs. Hinglish, Hindi vs. Hinglish):

$$m = 3 \text{ primary endpoints} \times 3 \text{ pairwise language contrasts} = 9 \text{ confirmatory hypotheses} \quad (8)$$

Family-wise error rate across this family of 9 tests is controlled using the Holm–Bonferroni step-down procedure at an overall two-sided $\alpha = 0.05$. Model-stratified contrasts ($m = 36$) and secondary endpoints are designated as exploratory analyses under False Discovery Rate control (FDR, $q = 0.05$).

No Models Fitted to Data: No statistical models have been fitted to data; no empirical parameters, odds ratios, or p-values have been generated.

8. Benchmark and Software Verification Results

In accordance with strict scientific disclosure, this section reports only verified benchmark characteristics, software execution outputs, and pipeline integrity results.

8.1 Benchmark Coverage and Task Characteristics

As summarized in Table 1, the benchmark achieves balanced coverage across all five civic domains ($N = 30$ tasks/domain, 450 Critical Facts, 300 Important Facts, 150 Primary Authorities). Healthcare features the highest safety sensitivity (73.3% high-risk), reflecting medical stakes, while Public Services features the highest procedural difficulty (73.3% moderate/high).

8.2 Pipeline Execution Matrix Verification

Structural verification of the software fixture dataset across the complete factorial matrix confirms 100.0% completion across all 5,400 cells (Table 2).

8.3 Synthetic Baseline Fixture Diagnostics Summary

Diagnostic analysis of the 5,400 synthetic software fixtures confirms expected baseline properties:

- **Execution Latency:** Recorded latency is 0 ms across all 5,400 records due to in-memory template synthesis. Within-cell word length variance across the 3 repetitions is exactly 0.0 across all 1,800 cells (100.0% zero-variance).
- **Length Properties:** Word counts are consistent across languages (English: 47.16 ± 6.08 ; Hindi: 46.16 ± 6.08 ; Hinglish: 46.16 ± 6.08 ; overall: 46.49 ± 6.10 words).
- **Script Composition:** English and Hinglish fixtures use 100.0% Latin characters. Hindi fixtures exhibit hybrid scaffolding (9.36% Devanagari framing characters, 90.64% Latin text for embedded factual clauses).

Detailed stratified length and script adherence tables are provided in Appendix Table B.1 and Table B.2.

9. Discussion: Methodological Governance and Code-Switching

9.1 Auditing Substantive Information Inequality

The central premise of LINGUA-AUDIT is that multilingual evaluations in civic domains must focus on the substantive, actionable content received by citizens rather than surface fluency. In administrative welfare schemes, subtle omissions—such as failing to mention an income cutoff or mandatory certificate attestation—have decisive consequences. LINGUA-AUDIT provides the prespecified measurement instrument to detect and quantify these disparities rigorously.

9.2 A Cautionary Tale: Circular Evaluator Validation

During development, an automated validation script reported near-perfect agreement ($ICC = 1.0000$, $\kappa = 0.9849$) between automated evaluators and purported human annotations. A provenance audit revealed these results were illusory: the script cloned evaluator output dictionaries in memory to serve as “human annotations.” By quarantining these outputs (QUARANTINE_MANIFEST.json) and explicitly disclosing that empirical evaluator validation remains pending, LINGUA-AUDIT highlights the hazard of circular self-validation in AI benchmarking.

9.3 Authentic Code-Switching in Civic Contexts

In South Asian digital contexts, bilingual speakers frequently employ Romanized code-switching (Hinglish) in mobile communication [13]. Evaluating models exclusively on formal English or standard Devanagari Hindi risks overlooking widespread communicative practices in multilingual communities. LINGUA-AUDIT establishes code-switching as a primary, first-class evaluation condition alongside Hindi and English.

10. Limitations and Threats to Validity

1. **Unexecuted Live Model Data Collection:** Live model generations have not yet been dispatched or collected; the current 5,400-record dataset consists of deterministic programmatic fixtures. Consequently, no empirical model comparisons or language disparities can be inferred.
2. **Pending Human Evaluator Calibration:** While a 12-item feasibility pilot package and a 270-response validation sample have been prepared, independent bilingual annotators have not yet been secured. Automated evaluators and the 13 metrics remain empirically uncalibrated against human ground truth.
3. **Internal Statutory Curation Scope:** Tasks, Critical Facts, and Important Facts were curated internally from published schemes without direct URLs, verbatim statutory excerpts, or clause-to-fact alignment mapping. Independent multi-expert panel certification remains future work.
4. **Pending Institutional Ethics Review:** An institutional ethics protocol has been prepared (docs/ethics/ethics_application.md), but formal institutional ethics committee approval remains pending.
5. **Single-Turn Query Paradigm:** The benchmark evaluates single-turn information requests, whereas real-world civic consultations often involve multi-turn dialogues.

11. Ethics and Data Governance

The benchmark queries public administrative statutes and civic welfare schemes. No private personal data or health records were collected. An institutional ethics protocol specifies fair compensation standards for bilingual annotators; formal ethics committee approval is pending review, and no human annotators were recruited prior to approval.

All benchmark tasks, ground-truth references, and scoring implementations are prepared for release under open research licenses (intended CC-BY-NC 4.0) upon final publication, though a formal standalone LICENSE file has not yet been committed to the repository root.

12. Conclusion and Future Work

LINGUA-AUDIT establishes an open, reproducible framework designed to test for information inequality in multilingual AI systems. By grounding 150 civic tasks in statutory documentation across five essential domains, formalizing 13 multi-dimensional response-quality metrics, formulating a prespecified statistical analysis plan, and verifying a full factorial software pipeline across 5,400 cells, the framework provides the proposed methodological and computational infrastructure to audit whether frontier language models introduce disparities across languages.

Immediate next steps span two distinct operational tracks followed by downstream analysis:

- 1. Data Collection Track (Live API Execution):** Executing and archiving genuine model responses from external provider APIs across the 5,400 factorial matrix cells.
- 2. Evaluator Calibration Track (Human Ground Truth):** Securing institutional ethics review approval, recruiting certified bilingual annotators, and executing the 12-item pilot and 270-response validation sample to compute genuine inter-rater reliability and calibrate automated evaluators.
- 3. Downstream Empirical Scoring and Estimation:** Applying the calibrated evaluation pipeline to genuine model responses and fitting the prespecified GLMMs to test for multilingual information inequality.

Reproducibility Statement

All research materials are preserved within the repository:

- **Benchmark Registry:** data/benchmark/master_task_registry_150.json (SHA-256: 2ce2cb688c011cb0279425a3c063ce4e5b7b6a2354a4aa9df75a77efa9d00315).
- **Pipeline Fixture Dataset:** data/benchmark/raw_responses/stage12_4_full_evaluation_responses.json (Raw bytes SHA-256: 095b8caa...; Canonical sorted JSON SHA-256: 296bb165...).
- **Scoring Implementation:** Modular Python package in backend/app/scoring/lingua_audit/, accompanied by 115 passing tests (pytest tests/).
- **Feasibility Pilot Packages:** Blinded 12-item packages stage12_5_2D_pilot_annotator_A.json and B.json, along with restricted key stage12_5_2D_pilot_restricted_mapping.json.
- **Quarantine Manifest:** Documented invalid synthetic artifacts in data/benchmark/evaluator_validation/QUARANTINE_MANIFEST.json.
- **Reproduction Script:** scratch/reproduce_paper_analysis.py reproduces all reported summary outputs in paper/results/.

References

- [1] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, “Measuring massive multitask language understanding,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.
- [2] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov, “XNLI: Evaluating cross-lingual sentence representations,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2018, pp. 2475–2485.
- [3] D. Blasi, A. Antonovsky, A. Anastasopoulos, G. Neubig, and R. Cotterell, “Systematic inequalities in language technology performance across the world’s languages,” in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2022, pp. 5486–5505.
- [4] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, “The state and fate of linguistic diversity and inclusion in the NLP world,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 6282–6293.
- [5] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?,” in *Proc. ACM Conf. Fairness Accountability Transparency (FAccT)*, 2021, pp. 610–623.
- [6] J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki, “TyDi QA: A benchmark for information-seeking QA in typologically diverse languages,” *Trans. Assoc. Comput. Linguistics (TACL)*, vol. 8, pp. 454–470, 2020.
- [7] K. Ahuja, H. Diddee, R. Hada, M. Ochieng, K. Ramesh, P. Jain, A. Nambi, T. Ganu, S. Segal, M. Ahmed, K. Bali, and S. Sitaram, “MEGA: Multilingual evaluation of generative AI,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2023, pp. 4232–4267.
- [8] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung, “A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity,” in *Proc. 13th Int. Joint Conf. Natural Lang. Process. 3rd Conf. Asia-Pacific Chapter Assoc. Comput. Linguistics (IJCNLP-AACL)*, 2023, pp. 675–718.
- [9] H. Singh, N. Gupta, S. Bharadwaj, D. Tewari, and P. Talukdar, “IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages,” in *Findings Assoc. Comput. Linguistics: ACL 2024*, 2024, pp. 10041–10061.
- [10] A. K. Singh, V. Kumar, R. Murthy, J. Sen, A. Mittal, and G. Ramakrishnan, “INDIC QA BENCHMARK: A multilingual benchmark to evaluate question answering capability of LLMs for Indic languages,” in *Findings Assoc. Comput. Linguistics: EMNLP 2024*, 2024, pp. 16231–16248.
- [11] OpenAI, “IndQA: Evaluating cultural fluency and context in Indian languages,” OpenAI Research, Tech. Rep., 2025.
- [12] N. Majithia, R. Shinde, Z. Chapman, P. Trital, J. Decker, M. Maskey, E. Simperl, and N. Shadbolt, “The CitizenQuery benchmark: A novel dataset and evaluation pipeline for measuring LLM performance in citizen query tasks,” *arXiv preprint arXiv:2602.04064*, 2026.
- [13] S. Khanuja, S. Dandapat, A. Srinivasan, S. Sitaram, and M. Choudhury, “GLUECoS: An evaluation benchmark for code-switched NLP,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 3575–3585.
- [14] S. Min, K. Krishna, X. Lyu, M. Lewis, W. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, “FactScore: Fine-grained atomic evaluation of factual precision in long form text generation,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2023, pp. 12076–12099.
- [15] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, et al., “Holistic evaluation of language models,” *Trans. Mach. Learn. Res. (TMLR)*, pp. 2835–8856, 2023.
- [16] L. Zheng, W. L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging LLM-as-a-judge with MT-Bench and Chatbot Arena,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 46595–46623.
- [17] R. Hada, V. Gumma, A. de Wynter, H. Diddee, M. Ahmed, M. Choudhury, K. Bali, and S. Sitaram, “Are large language model-based evaluators the solution to scaling up multilingual evaluation?,” *arXiv preprint arXiv:2309.07462*, 2023.

Appendix A: Metric Inventory & Verification Status

Table A.1 documents the complete operational specification, target statistical model, and software verification status for all 13 response-quality metrics. Implementation in the Python scoring package (backend/app/scoring/lingua_audit/metrics.py) and 115 passing tests confirm algorithmic execution correctness, not empirical construct validity. All metrics remain empirically unvalidated pending future human annotation.

Table A.1: Complete LINGUA-AUDIT metric implementation inventory and software verification status.

Code	Metric Name	Operationalized Formula / Rubric	Prespecified Role	Target Model & Verification Status
M1	Factual Precision	$N_{\text{Supp}} / N_{\text{adjudicated}}$	Primary Confirmatory	Binomial GLMM (logit link) • Code & 115 Tests: Verified • Validation: UNVALIDATED*
M1b	Contradiction Precision	$N_{\text{Supp}} / (N_{\text{Supp}} + N_{\text{Contra}})$	Secondary Sensitivity	Binomial GLMM (logit link) • Code & 115 Tests: Verified • Validation: UNVALIDATED*
M2a	Critical Fact Coverage	$CF_{\text{covered}} / CF_{\text{total}}$	Primary Confirmatory	Binomial GLMM (logit link) • Code & 115 Tests: Verified • Validation: UNVALIDATED*
M2c	Weighted Completeness	$(\text{sum } CF_{\text{wt}} + 0.5 \text{ sum } IF_{\text{wt}}) / \text{Total}_{\text{wt}}$	Secondary Sensitivity	Linear Mixed Model (LMM) [Provisional; Bounded] • Code & Tests: Verified • Validation: UNVALIDATED*
M3	Critical Fact Omission	$CF_{\text{omitted}} / CF_{\text{total}}$	Primary Confirmatory	Binomial GLMM (logit link) • Code & 115 Tests: Verified • Validation: UNVALIDATED*
M4a	Deduplicated Error Count	$\text{card}(\text{Claims}_{\text{contra}}) + \text{sum } \text{unlinked}(CF_{\text{contra}})$	Secondary Confirmatory	Negative Binomial GLMM (log link, log offset) • Code & Tests: Verified • Validation: UNVALIDATED*
M4b	Any Error Indicator	$1(M4a > 0)$	Secondary Confirmatory	Binomial GLMM (logit link) • Code & 115 Tests: Verified • Validation: UNVALIDATED*
M4c	Normalized Error Rate	$N_{\text{Contra}} / N_{\text{adjudicated}}$	Secondary Confirmatory	Binomial GLMM (logit link) • Code & 115 Tests: Verified • Validation: UNVALIDATED*
M5	Citation Quality Grade	Ordinal rubric {0, 1, 2, 3, 4}	Secondary Confirmatory	Cumulative Link Mixed Model (CLMM) • Code & Tests: Verified • Validation: UNVALIDATED*
M6	Domain Safety Level	Ordinal scale {0, 1, 2}	Secondary Confirmatory	Cumulative Link Mixed Model (CLMM) • Code & Tests: Verified • Validation: UNVALIDATED*
M7	Semantic Preservation	Ordinal scale {0..4} (anchored to UIN)	Secondary Confirmatory	Cumulative Link Mixed Model (CLMM) • Code & Tests: Verified • Validation: UNVALIDATED*
M8a	Completeness Variance	s^2_{CF} across 3 repetitions	Secondary Confirmatory	Two-Part Hurdle GLMM (Logit + Gamma) [Provisional] • Code & Tests: Verified • Validation: UNVALIDATED*
M8b	Jaccard Concordance	$K_{\text{overlap}} / K_{\text{union}}$ across 3 reps	Secondary Confirmatory	Binomial GLMM (logit link) [Provisional] • Code & Tests: Verified • Validation: UNVALIDATED*

**Validation Status Note: All 13 metrics are fully implemented and verified against unit and regression tests (115 passing tests), but remain strictly empirically unvalidated pending future bilingual human annotation.*

Appendix B: Synthetic Baseline Fixture Diagnostics

Table B.1 and Table B.2 report the detailed stratified text and script adherence metrics computed from the 5,400 programmatic software verification fixtures.

Table B.1: Synthetic fixture diagnostic characteristics across language scaffolding.

Condition / Stratum	N	Char Length (Mean $\pm$ SD)	Char Length (Median [IQR])	Word Count (Mean $\pm$ SD)	Word Count (Median [IQR])	Latency
Language: English	1,800	339.80 $\pm$ 45.33	340.0 [57.0]	47.16 $\pm$ 6.08	47.0 [7.0]	0 ms*
Language: Hindi	1,800	326.80 $\pm$ 45.33	327.0 [57.0]	46.16 $\pm$ 6.08	46.0 [7.0]	0 ms*
Language: Hinglish	1,800	333.80 $\pm$ 45.33	334.0 [57.0]	46.16 $\pm$ 6.08	46.0 [7.0]	0 ms*
Model Slot: Gemini 3.5 Flash	1,350	333.47 $\pm$ 45.64	337.0 [65.0]	46.49 $\pm$ 6.10	46.0 [7.0]	0 ms*
Model Slot: GPT-4o	1,350	333.47 $\pm$ 45.64	337.0 [65.0]	46.49 $\pm$ 6.10	46.0 [7.0]	0 ms*
Model Slot: Claude 3.5 Sonnet	1,350	333.47 $\pm$ 45.64	337.0 [65.0]	46.49 $\pm$ 6.10	46.0 [7.0]	0 ms*
Model Slot: Llama 3.1 70B	1,350	333.47 $\pm$ 45.64	337.0 [65.0]	46.49 $\pm$ 6.10	46.0 [7.0]	0 ms*
Fixture Total	5,400	333.47 $\pm$ 45.64	337.0 [65.0]	46.49 $\pm$ 6.10	46.0 [7.0]	0 ms*

*Recorded latency is 0 ms across all 5,400 records due to in-memory template synthesis. Within-cell word length variance across repetitions is exactly 0.0.

Table B.2: Synthetic fixture script composition across language conditions.

Language Condition	Fixtures	Devanagari Chars	Latin Chars	Punctuation	Devanagari %	Latin %	Script Profile Interpretation
English	1,800	0	467,280	144,360	0.0%	100.0%	Pure Latin typography
Hindi	1,800	41,400	400,680	146,160	9.36%	90.64%	Hybrid framing (Devanagari scaffolding)
Hinglish	1,800	0	458,280	142,560	0.0%	100.0%	Romanized code-switching
Total	5,400	41,400	1,326,240	433,080	—	—	Full Synthetic Fixture Corpus