

MOSAIC: A Benchmark for Authentic Language Data

MOSAIC is Serideo AI's proprietary benchmark for evaluating how closely language data reflects real, community-authentic communication — with a specific focus on code-switched and dialect-heavy languages across South and Southeast Asia. This report documents the framework's structure, evaluation approach, and initial findings from a completed evaluation run, alongside illustrative examples of what the benchmark measures.

- 01 The Problem: Why "High Quality" Isn't a Measurement
- 02 The MOSAIC Framework
- 03 The Six Evaluation Dimensions
- 04 Illustrative Examples by Dimension
- 05 Evaluation Methodology
- 06 Statistical Rigor & Reliability
- 07 Case Study: Maithili Conversational Text
- 08 What This Benchmark Does Not Claim
- 09 Roadmap

01 · The Problem: Why "High Quality" Isn't a Measurement

Data annotation providers routinely describe their output as "high quality," "native-level," or "human-verified" — descriptions that carry no measurable definition and cannot be independently checked. This is a particular problem for code-switched and dialect-heavy languages, where the gap between grammatically correct text and authentic, community-real communication is largest. A dataset can be perfectly correct and still fail to represent how people in Bihar, Chennai, or Jakarta actually speak day to day — mixing languages mid-sentence, using regional vocabulary, and leaving in the natural imperfection of real speech.

MOSAIC was built to close that gap: to give "authentic language data" an actual, reproducible measurement rather than a marketing description.

02 · The MOSAIC Framework

MOSAIC (v1.0) is a multi-dimensional scoring framework developed internally at Serideo AI. It evaluates a language dataset across six weighted dimensions of authenticity, combining independent human review with automated consistency checks into a single composite score, supported by dimension-level detail, reviewer agreement, and statistical confidence.

The goal of MOSAIC is not to replace human judgment about language, but to structure it — so that "this data feels authentic" becomes a claim that can be scored, audited, and repeated across languages, dialects, and projects.

03 · The Six Evaluation Dimensions

Dimension	What It Measures
Cultural Context	Whether references, expressions, and implied social meaning fit the target community naturally
Natural Expression	Whether language reads as genuinely spoken or written, rather than scripted or templated
Code-Switching Authenticity	Whether language mixing follows real, community-consistent patterns rather than forced or artificial switching
Human Imperfection	Whether natural variation, hesitation, and informal structure are represented realistically
Contextual & Semantic Fidelity	Whether meaning is preserved and remains consistent with surrounding context
Regional / Lexical Authenticity	Whether region- and dialect-specific vocabulary is accurately preserved rather than flattened into a generic standard

The relative weighting of each dimension, the composite scoring formula, and internal calibration rubrics are part of Serideo AI's internal methodology and are not published in this report.

04 · Illustrative Examples by Dimension

The examples below are constructed for illustration only — they do not come from any client dataset and are meant to show the kind of distinction each MOSAIC dimension is designed to catch.

Cultural Context

SCORES HIGHER

"Aaj Chhath ka doosra din hai, ghaat pe bheed zyada hogi shaam tak."

Names a specific regional festival and setting a local reader would immediately recognize.

SCORES LOWER

"Today is a religious festival, so there will be a big crowd later."

Technically correct, but strips out the specific cultural reference entirely.

Natural Expression

SCORES HIGHER

"Haan yaar, kal wala kaam ho gaya... bas thoda pending hai abhi."

Reads like real conversational speech, including its natural rhythm.

SCORES LOWER

"Yes, the task assigned yesterday has been completed, with a minor portion still pending."

Grammatically fine, but reads as written/formal rather than spoken.

Code-Switching Authenticity

SCORES HIGHER

"Mujhe kal ka meeting reschedule karna hai, thoda busy hoon."

Mixes languages the way bilingual speakers naturally do, mid-thought.

SCORES LOWER

"I have to reschedule tomorrow's meeting because I am currently occupied."

No code-switching at all — doesn't reflect how this sentence is actually spoken in the target community.

Human Imperfection

SCORES HIGHER

"Wo... matlab, maine bola tha na ki thoda late ho jaunga."

Includes natural hesitation and filler that real speech contains.

SCORES LOWER

"As previously mentioned, I will arrive slightly later than scheduled."

Overly clean and structured for an informal, spoken context.

Regional / Lexical Authenticity

SCORES HIGHER

"Hum apna gaam ja rahal chhi is weekend."

Uses Maithili-specific vocabulary and sentence construction, not generic Hindi.

SCORES LOWER

"Main apne gaon ja raha hoon is weekend."

Standard Hindi substituted in place of the dialect-specific form — meaning intact, authenticity lost.

05 · Evaluation Methodology

Each sample is independently rated across the six dimensions by trained human reviewers, alongside a parallel automated evaluation layer that checks selected dimensions for consistency at scale. Human review is weighted more heavily in the composite score, reflecting the fact that cultural and regional judgment is still best assessed by people fluent in the target language and community — automated evaluation supports consistency and scale rather than replacing that judgment.

Deductions are applied for critical quality issues identified during review — for example, meaning-changing errors, culturally inappropriate content, or unnatural, scripted code-switching — so that a small number of serious errors cannot be masked by an otherwise high average score.

Exact scoring weights, penalty values, and reviewer calibration rubrics are part of Serideo AI's internal methodology and are not published here.

06 · Statistical Rigor & Reliability

- **Sample size (n):** the number of data samples included in a given evaluation run, reported alongside every MOSAIC score.
- **Confidence interval (CI):** the statistical margin around a reported score, indicating how precisely the sample reflects the full dataset.
- **Inter-rater reliability (IRR):** the level of agreement between independent human reviewers scoring the same samples — a higher IRR indicates the rubric is being applied consistently rather than subjectively.
- **Human vs. automated split:** reported separately, so it is clear how much of the composite score reflects human judgment versus automated consistency checks.

Reporting these figures alongside every score is what allows a MOSAIC result to be treated as evidence rather than a claim.

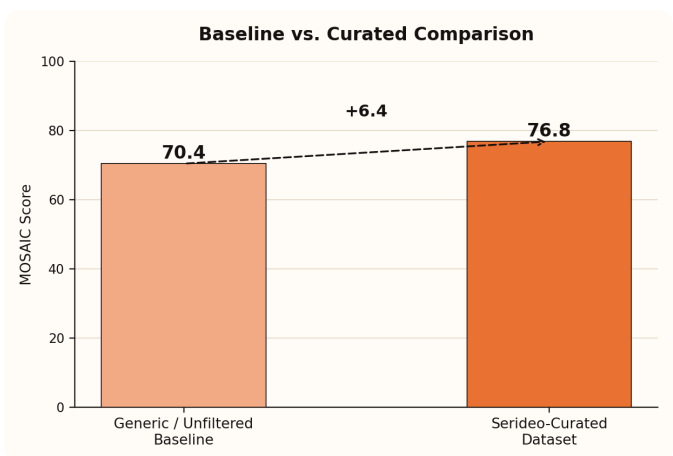
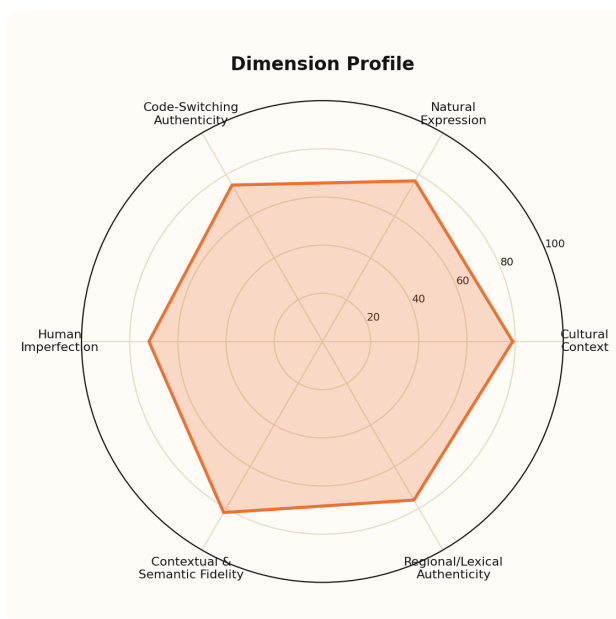
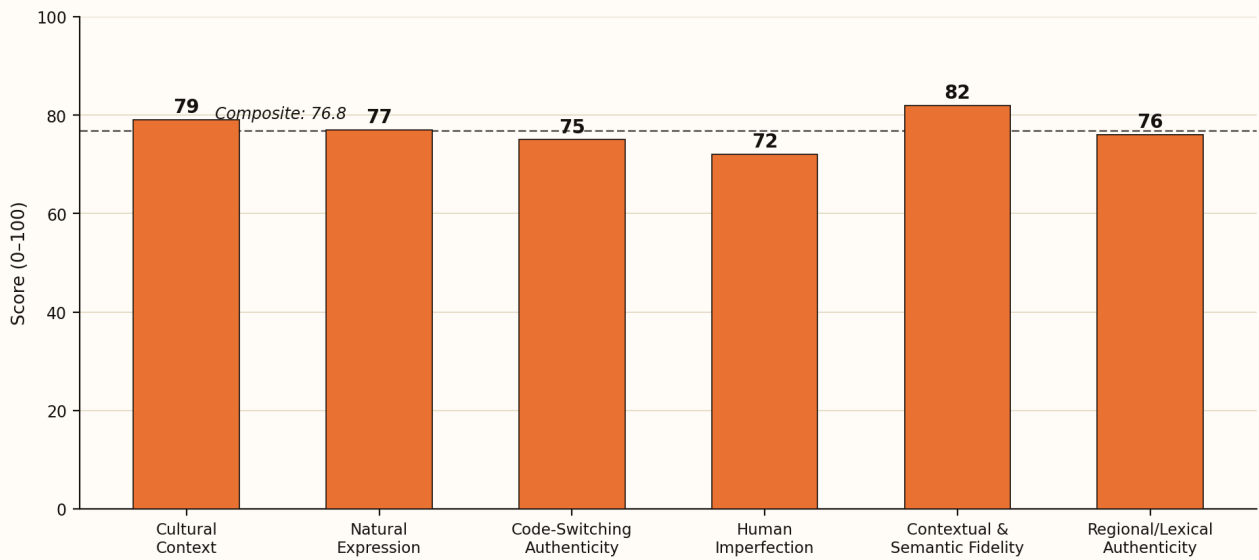
07 · Case Study: Maithili Conversational Text

The following is Serideo AI's first published MOSAIC evaluation: a conversational text dataset in Maithili, evaluated in August 2026.

76.8 / 100

MOSAIC SCORE · N = 200 SAMPLES · 95% CI ± 2.1

MOSAIC Dimension Scores — Maithili Conversational Dataset



Evaluation detail: Human evaluation score 80.0 · Automated evaluation score 76.0 · 4 independent reviewers · Inter-rater reliability 88.2% · Compared against a generic, unfiltered baseline dataset in the same language category, which scored 70.4 / 100 — a gap of +6.4 points. The baseline was constructed internally for comparison and does not represent any named third-party dataset or product.

Within this evaluation, the curated dataset scored measurably higher than the unfiltered baseline across every dimension, with strong reviewer agreement (88.2% IRR) and a tight confidence interval (± 2.1) at $n=200$ — indicating the result is stable rather than driven by a small or noisy sample. Contextual & Semantic Fidelity scored highest (82); Human Imperfection scored lowest (72), suggesting this is the area with the most room for improvement in future curation passes.

This is the first published run of MOSAIC, on one dataset, in one language. We are extending the same evaluation to additional languages and dataset types as part of ongoing benchmarking research.

08 · What This Benchmark Does Not Claim

MOSAIC is a proprietary Serideo AI research framework, not an external or third-party certification, and is not positioned as an industry-wide standard. This report reflects a single evaluation run and should not be read as a permanent or universal quality guarantee — results may vary by language, dialect, task type, and dataset domain. Consistent with our confidentiality practices, this report does not disclose the exact scoring formula, dimension weights, penalty thresholds, internal calibration rubrics, client or project identifiers, or internal tooling.

09 · Roadmap

Serideo AI is extending MOSAIC evaluations across additional languages and dialects — including further South and Southeast Asian code-switched varieties — and will publish further findings as that research progresses. Future reports will include cross-language comparisons and evaluation of additional dataset types beyond conversational text.

Serideo AI Research · MOSAIC Benchmark v1.0 · August 2026