



Research Article

Beyond Accuracy in Low-Resource Llm Evaluation: A Structured Critical Review and Reliability-Centered Framework

Noorullah Zamindar

Independent Researcher, Kabul, Afghanistan

Corresponding Author, Email: noorullahzamindar9@gmail.com

Copyright © 2026 by Authors, Published by **Interkoneksi: Journal of Computer Science and Digital Business**. This is an open access article under the CC BY License <https://creativecommons.org/licenses/by/4.0/>

Received : July 17, 2026

Revised : August 15, 2026

Accepted : September 13, 2026

Available online : October 29, 2026

How to Cite: Noorullah Zamindar. (2026). Beyond Accuracy in Low-Resource Llm Evaluation: A Structured Critical Review and Reliability-Centered Framework. *Interkoneksi: Journal of Computer Science and Digital Business*, 4(2), 255–269. <https://doi.org/10.61166/interkoneksi.v4i2.96>

Abstract. Large language models (LLMs) are commonly compared through aggregate accuracy on benchmarks translated from, or culturally anchored to, high-resource languages. In low-resource settings, the resulting score can conflate model capability with translation artifacts, cultural mismatch, dialect exclusion, tokenization inefficiency, contamination, metric failure, and unreliable automated judging. This structured critical review synthesizes 27 peer-reviewed works published from 2018 to 2026 and develops the LRL-Eval Card, a reliability-centered reporting framework. Multi-label coding found that metric validity was the most frequent theme (20/27), whereas benchmark origin and tokenization appeared in only three papers each. The evidence shows that culturally sensitive subsets can change model rankings, answer-option order can alter performance substantially, multilingual LLM judges have weak cross-language agreement, contamination can inflate scores, and tokenization affects both accuracy and access cost. The LRL-Eval Card records ten non-compensatory evidence dimensions rather than producing another composite score. Reliable low-resource evaluation therefore requires

triangulation across native-authored and translated items, language varieties, native-speaker judgments, contamination controls, tokenizer diagnostics, validated metrics, and uncertainty estimates.

Keywords: large language models; low-resource languages; multilingual evaluation; benchmark validity; cultural validity

INTRODUCTION

Evaluation determines which language models are described as capable, safe, efficient, and ready for deployment. Most widely used practices, however, were developed around English or a small group of digitally well-resourced languages. This imbalance is not resolved merely by adding translated test sets. A benchmark may display target-language text while preserving English-centered knowledge, translated constructions, answer conventions, cultural assumptions, and scoring rules. Its headline score may therefore measure a mixture of model capability and properties of the evaluation instrument.

The underlying disparity is longstanding. Joshi et al. (2020) documented that only a small fraction of the world's more than 7,000 languages receives meaningful attention in natural language processing. Large multilingual suites have since expanded coverage. MEGA evaluates generative models across 16 datasets and 70 languages (Ahuja et al., 2023); FLORES-101 supplies professionally translated parallel data for 101 languages (Goyal et al., 2022); Belebele provides reading-comprehension data in 122 language variants (Bandarkar et al., 2024); and MuBench contains 3.9 million aligned samples across 61 languages (Han et al., 2026). These resources improve comparability, but numerical coverage alone does not establish translation fidelity, local cultural relevance, dialect representativeness, scoring validity, contamination resistance, or equal computational access.

Broader evaluation research already demonstrates why one metric is insufficient. HELM organizes model evaluation across scenarios and multiple outcomes, including calibration, robustness, fairness, toxicity, and efficiency (Liang et al., 2023). CheckList shows that behavioral tests can expose failures concealed by held-out accuracy (Ribeiro et al., 2020). General reviews and benchmark guidance similarly call for explicit constructs, transparent data, multiple metrics, and reproducible reporting (Gevers et al., 2025; Laskar et al., 2024). Low-resource-language evaluation adds a specific problem: the evaluated population is underrepresented not only in model training, but also in benchmark design, metric development, annotation governance, and error analysis.

Several common assumptions are contradicted by evidence. A translated English benchmark is often treated as equivalent to one written by native speakers; a standard language variety is generalized to dialects and scripts; automatic metrics are assumed to work similarly across languages; and a powerful LLM is assumed to be a reliable judge wherever it can generate text. Semantically equivalent prompts are also treated as equally costly, public benchmark performance as clean generalization, and a finite-sample point estimate as a stable ranking. Yet Global MMLU identifies

linguistic and cultural biases in translated questions and reports ranking changes on culturally sensitive subsets (Singh et al., 2025). ProverbEval finds performance changes of up to 50% when answer choices are reordered (Azime et al., 2025). Five multilingual LLM judges achieve average cross-language Fleiss' kappa of only about 0.3 across five tasks and 25 languages (Fu & Liu, 2025). Potential benchmark contamination ranges from 1% to 45% in one multi-benchmark study, with accuracy inflation reaching 14% on a tested benchmark (Li et al., 2024), and contamination can cross language barriers while evading surface-overlap detection (Yao et al., 2024). Tokenization studies further show unequal cost and lower accuracy when equivalent content is segmented inefficiently (Ahia et al., 2023; Lundin et al., 2026). Accuracy alone cannot reveal whether the model, benchmark, tokenizer, judge, or sample produced the result.

Previous work usually isolates one component of this chain. Native benchmark studies address provenance and culture; dialect studies examine linguistic varieties; contamination research studies memorization; tokenizer studies quantify segmentation and cost; and judge studies assess automated scoring. Existing taxonomies are broad, but the reviewed literature does not provide one operational, low-resource-specific record covering translation quality, benchmark origin, cultural validity, dialect coverage, native-speaker involvement, contamination, LLM-as-a-judge reliability, tokenization, metric validity, and statistical adequacy. The gap is integrative: researchers need to connect benchmark provenance, community participation, model-interface effects, scoring reliability, and uncertainty without replacing accuracy with another opaque total.

This review asks: (1) which validity and reliability threats are documented in multilingual and low-resource LLM evaluation; (2) what evidence shows that accuracy-only or coverage-only evaluation can conceal or change substantive conclusions; (3) which dimensions receive the most and least attention in the selected corpus; and (4) what minimum evidence should accompany an interpretable and reproducible low-resource-language claim? The study contributes a structured critical synthesis of 27 peer-reviewed works, an evidence map across ten evaluation dimensions, and the LRL-Eval Card—a non-compensatory reporting instrument designed for authors, reviewers, benchmark creators, and deployers.

RESEARCH METHODS

This study used a structured critical review followed by framework synthesis. A critical-review design was appropriate because the aim was to integrate heterogeneous evidence from benchmark papers, empirical reliability studies, tokenizer analyses, contamination research, and evaluation frameworks. The review was not registered as a systematic review and does not claim exhaustive database coverage. PRISMA-ScR is therefore cited as a boundary on the claim rather than applied retrospectively to imply comprehensive screening (Tricco et al., 2018).

The corpus comprised 27 peer-reviewed works published from 2018 through 2026. Discovery routes included the ACL Anthology, Transactions of the Association for Computational Linguistics, Transactions on Machine Learning Research/OpenReview, and a peer-reviewed reporting-guideline article. Citation chaining from a general LLM-evaluation review and a benchmark checklist located

dimension-specific studies (Gevers et al., 2025; Laskar et al., 2024). Searches were completed on 24 August 2026 using combinations of *large language model*, *multilingual*, *low-resource language*, *evaluation*, *benchmark*, *translation*, *culture*, *dialect*, *native speaker*, *contamination*, *tokenizer*, *metric*, *LLM-as-a-judge*, *reliability*, *validity*, *uncertainty*, and *benchmark size*.

Works were eligible when they directly evaluated a generative or foundation model in multiple or low-resource languages; introduced a multilingual or culturally grounded benchmark; tested a threat such as translation, contamination, tokenization, judge reliability, or prompt instability; or proposed evaluation guidance transferable to low-resource settings. English-only work was retained only for a general measurement contribution, such as behavioral testing, holistic evaluation, confidence-interval reporting, or review guidance. Blogs, vendor claims, inaccessible manuscripts, and model papers containing only incidental leaderboard scores were excluded. The unit of analysis was the publication.

Each work was recorded with year, evidence type, evaluated languages, principal finding, source, and thematic codes. One mutually exclusive evidence-type label was assigned: benchmark, empirical evaluation, framework, systematic review, or reporting guideline. Multi-label coding then captured ten prespecified dimensions: translation; native-versus-translated origin; culture; dialect; native speakers; contamination; LLM judge; tokenization; metrics; and benchmark size/statistical adequacy. A theme was assigned only when it was central to a paper's design, analysis, or stated contribution. Counts therefore describe prominence in the selected literature, not whether every paper complied with every reporting practice.

Framework development had three stages. First, failure modes were grouped by where they enter the measurement chain: benchmark construction, linguistic and cultural coverage, model interface, scoring, or inference. Second, each theme became an auditable question, minimum evidence field, and stronger-evidence condition. Third, the dimensions were checked against representative positive and negative cases in the corpus. Four evidence states were defined for future applications: not reported, documented, empirically checked, and triangulated. The states are intentionally not summed because one dimension cannot compensate for another; confidence intervals, for example, do not repair the absence of dialect coverage.

The review used descriptive statistics only. Counts and percentages were calculated with 27 as the denominator. Because thematic coding was multi-label, percentages do not sum to 100%. No meta-analysis was attempted because tasks, languages, models, metrics, and samples differed substantially. Quantitative findings from individual papers are reported as published and are not pooled. The framework was compared conceptually with two baselines: an accuracy-only design that treats aggregate performance as sufficient, and a coverage-only design that assumes a language's presence establishes valid evaluation.

RESULTS AND DISCUSSION

Corpus characteristics and thematic distribution

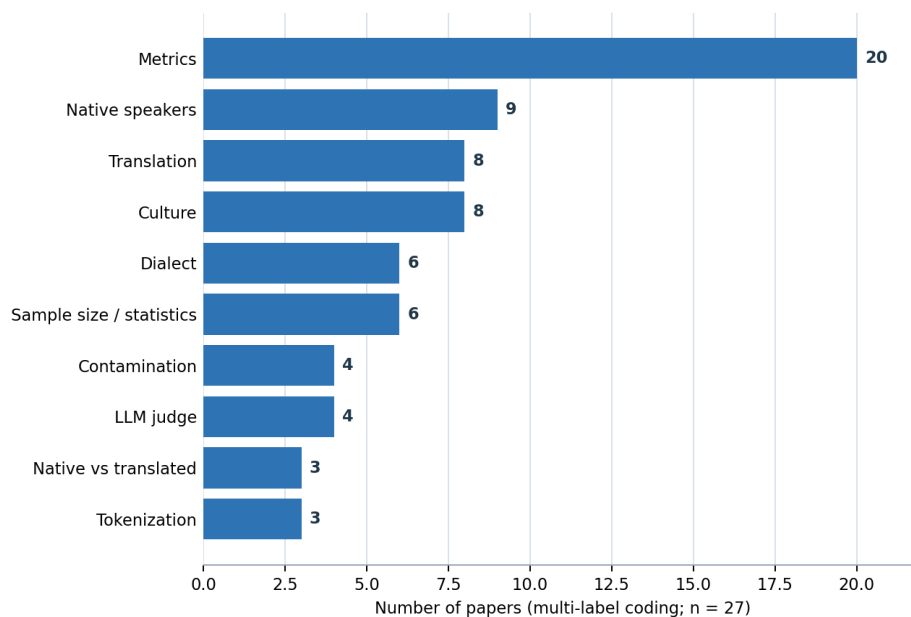
The corpus contained ten empirical evaluations (37.0%), nine benchmark papers (33.3%), six frameworks (22.2%), one systematic review (3.7%), and one

reporting guideline (3.7%). Eighteen papers (66.7%) appeared from 2024 through 2026, showing rapid growth in methodological attention. Metric validity was central in 20 works (74.1%). Native-speaker participation appeared in nine (33.3%); translation and culture in eight each (29.6%); dialect and statistical adequacy in six each (22.2%); contamination and LLM-as-a-judge reliability in four each (14.8%); and benchmark origin and tokenization in three each (11.1%). Table 1 and Figure 1 report these multi-label counts.

Table 1. Thematic distribution in the 27-paper corpus

Evaluation dimension	Papers, n	Share, %
Metric validity	20	74.1
Native-speaker involvement	9	33.3
Translation quality	8	29.6
Cultural validity	8	29.6
Dialect coverage	6	22.2
Benchmark size/statistical adequacy	6	22.2
Dataset contamination	4	14.8
LLM-as-a-judge reliability	4	14.8
Native-versus-translated origin	3	11.1
Tokenization differences	3	11.1

Figure 1. Number of reviewed papers centrally addressing each evaluation dimension. Themes are multi-label.



The distribution should not be read as a compliance audit. A paper lacking a code may still mention the topic. It does, however, indicate where methodological attention concentrates: metrics dominate, whereas the origin of test items and tokenizer-mediated access remain comparatively rare as central research questions. This imbalance matters because those less frequent dimensions can directly change the meaning and cost of a score.

Translation, origin, culture, and language varieties

Parallel benchmarks and native-authored benchmarks serve different purposes. FLORES-101 uses professional translation to support controlled comparison across 101 languages (Goyal et al., 2022), and Belebele maximizes cross-language comparability through parallel reading-comprehension items (Bandarkar et al., 2024). These strengths do not make translated content locally authored. Global MMLU shows why provenance must be separated from surface language: improved translation is paired with culturally sensitive and culturally agnostic subsets, and the distinction changes comparative conclusions (Singh et al., 2025). A linguistically accurate translation can therefore remain culturally external.

Native construction supplies complementary evidence. ParsiNLU contains more than 14,500 Persian instances across six language-understanding tasks with native-speaker participation (Khashabi et al., 2021). NativQA derives approximately 64,000 question-answer pairs from natural queries supplied by native users across seven languages, nine regions, and 18 topics (Hasan et al., 2025). These designs test communicative needs and presuppositions that an English source set might never contain. The appropriate design is not to declare native items universally superior, but to report origin and, where feasible, compare native-authored, human-translated, and machine-translated conditions.

Cultural validity extends beyond grammatical adequacy. In a benchmark of cultural localization across 15 languages, five native-speaker raters per language found that idioms and puns were harder than holidays or explicit cultural concepts; mean overall quality was only 1.68 on a 0–3 scale (Van Doren et al., 2026). ProverbEval likewise connects culture-specific meaning with measurement instability: answer-order changes can alter performance by as much as 50% (Azime et al., 2025). These results show that item culture and response format can interact. Cultural sensitivity and perturbation family should therefore become item-level metadata, with results disaggregated rather than hidden inside a mean.

A language label also conceals internal diversity. Tests of dialectal and language-variety tasks show persistent difficulty and sensitivity to prompt and label design (Faisal & Anastasopoulos, 2025). A direct comparison of Persian, Dari, and Tajiki reports substantial differences associated with variety, script, and prompting (Arnob & Mahi, 2026). Safety annotation research identifies a related gap in language coverage, dialect representation, cultural semantics, and annotation guidelines (Zaghoulani, 2026). Consequently, authors should name the ISO code, script, region, and variety, then report enough independent items per stratum to support the intended claim. A score for a standard variety cannot be generalized automatically to regional dialects, code-switching, or another script.

Native-speaker participation was central in only one-third of the corpus. Strong cases assign substantive roles: native users supply questions in NativQA (Hasan et al., 2025), language communities validate culturally sensitive material in Global MMLU (Singh et al., 2025), and multiple native raters judge cultural localization (Van Doren et al., 2026). Reporting should distinguish construct definition, item writing, translation, adjudication, error analysis, and governance. Final proofreading is not equivalent to co-design, and the phrase “native checked” is insufficient without roles, qualifications, compensation, disagreement handling, and the number of annotators.

Contamination, automated judges, and metrics

Public benchmarks can enter pretraining, instruction-tuning, or evaluation-development data. Li et al. (2024) report potential contamination from 1% to 45% across examined benchmarks and score inflation of up to 14% in one case. Surface matching is particularly inadequate for translated evaluation because semantically preserved content can cross a language boundary (Yao et al., 2024). A target-language string may be new while its source question, answer pattern, or translation remains familiar. Minimum reporting should record public/private status, release date, direct and fuzzy checks, and the possibility of cross-lingual or semantic leakage. Stronger designs include newly authored or sealed items and sensitivity analysis with suspected cases removed.

LLM-as-a-judge offers scalable rubric-based assessment of open responses, but the evaluator becomes another model requiring validation. Cross-language agreement of multilingual judges is weak on average and deteriorates in lower-resource settings (Fu & Liu, 2025). In Farsi natural-language-inference annotation, prompt order changes labels and disagreement sometimes reveals ambiguity in the gold standard rather than simple model error (Panahi et al., 2026). These findings require two safeguards. First, validate each judge, prompt, and rubric against independent target-language human judgments by task and language. Second, analyze disagreement rather than forcing every example into one apparently certain label. A judge model’s strength in English does not establish equivalent competence elsewhere.

Metric choice remains necessary but not self-validating. Exact match and lexical overlap are reproducible, yet they penalize valid paraphrase and depend on normalization. Learned metrics may improve semantic flexibility while inheriting the coverage and biases of their encoders. Human judgments are not automatically correct either; reliability depends on rubric clarity, language expertise, sampling, and adjudication. The dominant theme in the corpus should therefore be interpreted as a need for triangulation, not a search for one universal multilingual metric. Open-ended tasks should report a task metric, target-language human judgments, human agreement, and validated automated scoring when scale requires it.

Tokenization and statistical adequacy

Tokenization is part of the measurement pipeline. Equivalent information can require different token counts across languages, increasing price and reducing usable context for many users (Ahia et al., 2023). Across 16 African languages and five AfriMMLU subjects, greater token fertility consistently predicts lower accuracy

(Lundin et al., 2026). TokCollate broadens intrinsic analysis through measures organized by script, family, geography, population, and resource level (Variš et al., 2026). An evaluation that reports only accuracy may therefore hide an access penalty in cost, latency, context consumption, or truncation. Minimum reporting should name the tokenizer and version and provide tokens per item; stronger evidence adds fertility, character or byte ratios, truncation, latency, and cost.

Benchmark size determines how confidently differences can be interpreted. A point advantage is not a stable ranking when items are few, clustered by source, or duplicated through translation and perturbation. Bestgen (2022) argues that state-of-the-art claims should report differences and confidence intervals, not only the winning point score. For aligned multilingual items, resampling should keep translations and perturbations of one source together. Authors should report independent source-item counts, effect sizes, 95% intervals, and rank-stability or sensitivity analyses. This is especially important in low-resource settings, where small public sets can define the entire evidence base.

Comparison with accuracy- and coverage-only baselines

The reviewed evidence shows that accuracy-only evaluation is underdetermined. Ranking changes on cultural subsets challenge the conclusion that the highest aggregate score identifies the best model (Singh et al., 2025). Answer-order instability challenges the assumption that a multiple-choice score isolates knowledge (Azime et al., 2025). Weak multilingual judge agreement challenges the use of an automated score without target-language validation (Fu & Liu, 2025). Direct and cross-lingual contamination challenge the interpretation of public-test accuracy as generalization (Li et al., 2024; Yao et al., 2024). Tokenization disparities challenge the assumption that equal meaning implies equal computational access (Ahia et al., 2023; Lundin et al., 2026).

Coverage-only evaluation is also insufficient. MuBench's 61-language scope is more informative because it adds aligned design, Multilingual Consistency, and a human audit of 34,000 samples (Han et al., 2026). MEGA's 70-language evaluation exposes broad performance variation, but the meaning of a language average still depends on task, item origin, and metric (Ahuja et al., 2023). Coverage is necessary for inclusion; provenance, validity, reliability, efficiency, and uncertainty determine what inclusion means. Broad suites should therefore be paired with native-content audits and variety-specific analysis rather than presented as the final word on each language.

The LRL-Eval Card

Table 2 converts the synthesis into a dimension-level evidence record. For each row, evaluators select *not reported*, *documented*, *empirically checked*, or *triangulated*, attach supporting evidence, and record unresolved risks. The levels are not added. This non-compensatory design prevents excellent reporting in one area from concealing a validity failure elsewhere.

Table 2. LRL-Eval Card for low-resource-language evaluation

Dimension	Core question audit	Minimum evidence	Strong evidence
Translation	How was target-language text produced and checked?	System, translator roles, review and error policy	Independent native review plus measured adequacy and error analysis
Benchmark origin	Where did the construct and item originate?	Item-level native, translated, adapted, or unknown label	Matched native, human-translated, and machine-translated conditions
Culture	Whose knowledge and pragmatic norms are tested?	Cultural-sensitivity annotation	Native-user items and disaggregated cultural results
Dialect	Which varieties, scripts, and regions are represented?	Exact language and variety identifiers	Stratified evaluation with adequate items per variety
Native speakers	Where did native expertise influence the work?	Roles and stage of involvement	Co-design, adjudication, error analysis, and governance
Contamination	Could the model have seen the content or a translation?	Public status plus direct and fuzzy checks	Sealed/new items and semantic or cross-lingual checks
LLM judge	Does the judge agree with target-language humans?	Model/version, prompt, rubric, validation sample	Agreement, bias tests, uncertainty, and human adjudication
Tokenization	Does equivalent content use unequal tokens or context?	Tokenizer/version and tokens per item	Fertility, byte ratios, truncation, latency, and cost
Metrics	Does scoring measure the intended construct?	Metric definition and normalization	Target-language human validation and task-specific triangulation

Dimension	Core question	audit	Minimum evidence	Strong evidence
Benchmark size	Are estimates and rankings stable?		Per-condition independent item count	Effects, 95% intervals, cluster bootstrap, sensitivity analysis

The card complements holistic evaluation, behavioral testing, and multilingual suites. HELM broadens the outcome space (Liang et al., 2023); CheckList broadens behavioral testing (Ribeiro et al., 2020); and benchmark checklists improve documentation (Gevers et al., 2025). The card's specific contribution is to connect ten threats that are unusually consequential when the evaluated language is underrepresented throughout the research pipeline. It generates a profile of evidence and open risk, not a leaderboard score.

The practical implications are direct. Researchers should report results at the model-by-language-by-variety-by-condition level with benchmark provenance, prompt, tokenizer, scorer, and uncertainty. Benchmark creators should involve native speakers in construct definition and error analysis, not only translation. Developers should treat token efficiency and context consumption as multilingual quality attributes. Reviewers should challenge broad claims when evidence about varieties, judge validity, contamination, or uncertainty is missing. For Pashto and Persian-related work, recent offensive-language benchmarking provides a task-specific starting point (Bokaei et al., 2026), but a single classification benchmark cannot substitute for multidimensional generative evaluation.

Limitations, validity threats, and ethics

This review is purposive rather than exhaustive. It may omit studies indexed elsewhere, work published without English metadata, or recent preprints. The thematic matrix was coded by one reviewer; no inter-rater statistic is claimed. Contribution-oriented coding means the frequencies are not compliance rates. The studies are too heterogeneous for pooled effects, and quantitative findings illustrate failure modes rather than universal magnitudes. "Low-resource" is also contextual: a language may have general text but little domain-specific, dialectal, safety, or annotated data. Evidence from Persian, Pashto, Dari, African languages, or Arabic dialects does not automatically generalize to every language. The LRL-Eval Card has literature-based face validity but has not yet undergone Delphi review, inter-rater testing, or prospective validation.

This study analyzed public scholarly publications and collected no personal data, so human-subject approval was not required. Future empirical application may involve native-speaker annotation or harmful content and should obtain ethics review or a documented exemption where applicable. Native-speaker knowledge is an intellectual contribution, not merely low-cost annotation labor. Studies should use informed consent, fair compensation, manageable workloads, opt-out procedures for harmful material, and authorship or acknowledgment consistent with contributions.

Researchers should minimize demographic data, respect licenses, allow community control over culturally sensitive examples, and avoid implying that one sampled variety represents an entire language community.

CONCLUSION

Reliable low-resource-language LLM evaluation requires more than translating an English benchmark and reporting accuracy. Across 27 peer-reviewed works, translation and origin, cultural content, response-format perturbations, language varieties, native-speaker participation, contamination, automated judges, tokenization, metric validity, and finite-sample uncertainty can each alter the conclusion. The LRL-Eval Card makes these dependencies explicit through ten separate evidence fields and red flags. Its intended outcome is not another universal ranking, but a reproducible basis for deciding when an evaluation claim deserves trust.

Prospective work should ask independent multilingual researchers and native-speaker experts to apply the card to common benchmarks, measure inter-rater agreement and usability, and refine thresholds through a Delphi process. A preregistered Pashto–Dari study should compare native-authored, human-translated, and machine-translated item families while measuring cultural sensitivity, variety-level performance, prompt and answer-order robustness, native-human and LLM-judge agreement, contamination risk, token fertility, latency, cost, effect sizes, confidence intervals, and rank stability.

DECLARATIONS

Funding: The author received no external funding for this work.

Conflict of interest: The author declares no competing interests.

Author contributions: Noorullah Zamindar performed the conceptualization, methodology, investigation, data curation, formal analysis, visualization, drafting, review, editing, and project administration.

Data and code availability: The analyzed works are available through the cited publication links. No model training or inference code was used. The descriptive results are reproducible from the reported corpus size, coding definitions, and theme counts.

AI-use declaration: OpenAI ChatGPT/Codex assisted with literature organization, language editing, citation formatting, descriptive calculations, and document preparation. The named author must independently verify every claim, citation, and declaration before submission and assumes responsibility for the manuscript.

BIBLIOGRAPHY

Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D., Smith, N., & Tsvetkov, Y. (2023). Do all languages cost the same? Tokenization in the era of commercial language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 9904–

- 9923). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.614>
- Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Ahmed, M., Bali, K., & Sitaram, S. (2023). MEGA: Multilingual evaluation of generative AI. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 4232–4267). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.258>
- Arnob, N. M. K., & Mahi, A. B. S. (2026). One language, three of its voices: Evaluating multilingual LLMs across Persian, Dari, and Tajiki on translation and understanding tasks. In R. Merchant & K. Megerdooian (Eds.), *The proceedings of the first workshop on NLP and LLMs for the Iranian language family* (pp. 98–104). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.silkroadnlp-1.10>
- Azime, I. A., Tonja, A. L., Belay, T. D., Chanie, Y., Balcha, B. F., Abadi, N. H., Ademtew, H. B., Nerea, M. A., Yadeta, D. D., Geremew, D. D., Tesfu, A. A., Slusallek, P., Solorio, T., & Klakow, D. (2025). ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Findings of the association for computational linguistics: NAACL 2025* (pp. 6265–6281). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-naacl.350>
- Bandarkar, L., Liang, D., Muller, B., Artetxe, M., Shukla, S. N., Husa, D., Goyal, N., Krishnan, A., Zettlemoyer, L., & Khabsa, M. (2024). The belebele benchmark: A parallel reading comprehension dataset in 122 language variants. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 749–775). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.44>
- Bestgen, Y. (2022). Please, don't forget the difference and the confidence interval when seeking for the state-of-the-art status. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 5956–5962). European Language Resources Association. <https://aclanthology.org/2022.lrec-1.640/>
- Bokaei, Z., Webber, B., & Magdy, W. (2026). Benchmarking offensive language detection in Persian and Pashto. In R. Merchant & K. Megerdooian (Eds.), *The proceedings of the first workshop on NLP and LLMs for the Iranian language family* (pp. 13–23). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.silkroadnlp-1.2>
- Faisal, F., & Anastasopoulos, A. (2025). Testing the boundaries of LLMs: Dialectal and language-variety tasks. In Y. Scherrer, T. Jauhiainen, N. Ljubešić, P. Nakov, J. Tiedemann, & M. Zampieri (Eds.), *Proceedings of the 12th workshop on NLP for similar languages, varieties and dialects* (pp. 68–92). Association for Computational Linguistics. <https://aclanthology.org/2025.vardial-1.6/>

- Fu, X., & Liu, W. (2025). How reliable is multilingual LLM-as-a-judge? In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Findings of the association for computational linguistics: EMNLP 2025* (pp. 11040–11053). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-emnlp.587>
- Gevers, I., De Marez, V., Van Nooten, J., Lemmens, J., Kosar, A., Lotfi, E., Banar, N., Fivez, P., De Bruyne, L., & Daelemans, W. (2025). In benchmarks we trust ... Or not? In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 23662–23676). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.emnlp-main.1208>
- Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., & Fan, A. (2022). The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10, 522–538. https://doi.org/10.1162/tacl_a_00474
- Han, W., Zhang, Y., Chen, Z., Binbinliu, Pechenizkiy, M., Fang, M., & Zheng, Y. (2026). MuBench: Assessment of multilingual capabilities of large language models across 61 languages. In M. Liakata, V. P. Moreira, J. Zhang, & D. Jurgens (Eds.), *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 16163–16192). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.findings-acl.794>
- Hasan, Md. A., Hasanain, M., Ahmad, F., Laskar, S. R., Upadhyay, S., Sukhadia, V. N., Kutlu, M., Chowdhury, S. A., & Alam, F. (2025). NativQA: Multilingual culturally-aligned natural query for LLMs. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Findings of the association for computational linguistics: ACL 2025* (pp. 14886–14909). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-acl.770>
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 6282–6293). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.560>
- Khashabi, D., Cohan, A., Shakeri, S., Hosseini, P., Pezeshkpour, P., Alikhani, M., Aminnaseri, M., Bitaab, M., Brahman, F., Ghazarian, S., Gheini, M., Kabiri, A., Mahabagdi, R. K., Memarrast, O., Mosallanezhad, A., Noury, E., Raji, S., Rasooli, M. S., Sadeghi, S., ... Yaghoobzadeh, Y. (2021). ParsiNLU: A suite of language understanding challenges for persian. *Transactions of the Association for Computational Linguistics*, 9, 1147–1162. https://doi.org/10.1162/tacl_a_00419
- Laskar, M. T. R., Alqahtani, S., Bari, M. S., Rahman, M., Khan, M. A. M., Khan, H., Jahan, I., Bhuiyan, A., Tan, C. W., Parvez, M. R., Hoque, E., Joty, S., & Huang, J. X. (2024). A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical*

- methods in natural language processing* (pp. 13785–13816). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.764>
- Li, Y., Guo, Y., Guerin, F., & Lin, C. (2024). An open-source data contamination report for large language models. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: EMNLP 2024* (pp. 528–541). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-emnlp.30>
- Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., ... Koreeda, Y. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=iO4LZibEqW>
- Lundin, J. M., Zhang, A., Karim, N., Louzan, H., Wei, G., Adelani, D. I., & Carroll, C. (2026). The token tax: Systematic bias in multilingual tokenization. In E. A. Chimoto, C. Lignos, S. Muhammad, I. Abdulmumin, C. Siro, & D. I. Adelani (Eds.), *Proceedings of the 7th workshop on African natural language processing (AfricaNLP 2026)* (pp. 103–112). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.africanlp-main.10>
- Panahi, S., Kelleher, J., & Nedumpozhi, V. (2026). When LLMs annotate: Reliability challenges in low-resource NLI. In H. Hettiarachchi, T. Ranasinghe, A. Plum, P. Rayson, R. Mitkov, M. Gaber, D. Premasiri, F. A. Tan, & L. Uyangodage (Eds.), *Proceedings of the second workshop on language models for low-resource languages (LoResLM 2026)* (pp. 178–188). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.loreslm-1.17>
- Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4902–4912). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.442>
- Singh, S., Romanou, A., Fourrier, C., Adelani, D. I., Ngui, J. G., Vila-Suero, D., Limkonchoti, P., Marchisio, K., Leong, W. Q., Susanto, Y., Ng, R., Longpre, S., Ruder, S., Ko, W.-Y., Bosselut, A., Oh, A., Martins, A., Choshen, L., Ippolito, D., ... Hooker, S. (2025). Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 18761–18799). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.acl-long.919>
- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., ... Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467–473. <https://doi.org/10.7326/m18-0850>

- Van Doren, M., Ford, C., Barajas, J., VanMeter, R., & Holland, C. (2026). "Be my cheese?": Cultural nuance benchmarking for machine translation in multilingual LLMs. In S. Mille, S. Gehrmann, P. Schmidová, O. Dušek, M. Fadaee, K. Lo, E. Santus, & G. Stanovsky (Eds.), *Proceedings of the fifth workshop on generation, evaluation and metrics (GEM)* (pp. 52–76). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.gem-main.6>
- Variš, D., Stephen, A., & Libovický, J. (2026). TokCollate: A comprehensive tool for tokenizer evaluation and visualization across languages. In G. Durrett & P. Jian (Eds.), *Proceedings of the 64th annual meeting of the Association for Computational Linguistics (volume 3: System demonstrations)* (pp. 418–427). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.acl-demo.41>
- Yao, F., Zhuang, Y., Sun, Z., Xu, S., Kumar, A., & Shang, J. (2024). Data contamination can cross language barriers. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 17864–17875). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.990>
- Zaghoulani, W. (2026). Lost in dialect: The annotation gap in multilingual LLM safety. In K. Huang, F. Mo, P. Chen, & M. Jiang (Eds.), *Proceedings of the 1st workshop on multilinguality in the era of large language models (MeLLM 2026)* (pp. 1–13). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.mellm-1.1>