What Public Benchmark Evidence Can Establish About Bangla-Specific and Multilingual Large Language Models

Scaledex Research Labs

***Abstract*—This paper synthesizes the retained public-study record on Bangla-specific and multilingual large language models across factuality, instruction following, and the requested Bangla-English code-mixed domain. The retained TigerLLM/BenHalluEval evidence supports a bounded inference: task-conditional calibration findings, expressed in the study's Bengali-centric terminology, do not establish that Bengali-centric pretraining is necessary or sufficient for calibration. BenHalluEval evaluates judgments against supplied context, not open-world factual accuracy. The verified record provides no usable direct Bangla-specific-versus-multilingual comparison for Bangla instruction following, and MixSarc provides no explicit comparison between those categories. Two audit limitations remain material: the TigerLLM result requires reconciliation of a version 1 seven-model roster and a version 4 nine-model roster before rank or sole-category-representation assertions, and neither BenHalluEval nor MixSarc can be characterized as genuinely Bangla-English code-mixed without dataset-level evidence of actual mixing, script regime, and mixing prevalence. The resulting conclusion is construct-specific and study-bounded rather than a general ranking of model categories.**

***Index Terms*—Bangla language models, multilingual large language models, hallucination calibration, instruction following, code-mixed evaluation**

I. Introduction

The decision question is narrow: whether the audited public benchmark and human-evaluation record permits comparative conclusions about Bangla-specific and multilingual models on factuality, instruction following, and Bangla-English code-mixed tasks. These constructs are not interchangeable. BenHalluEval operationalizes context-grounded hallucination or calibration, whereas the translated Bengali benchmark study evaluates academic-task performance rather than instruction adherence as a distinct construct. MixSarc reports zero-shot task results for named models but does not provide the requested model-category comparison [1], [4], [7], [8].

Within the retained record, the central comparative result is limited but informative: the supplied TigerLLM/BenHalluEval evidence does not establish that Bengali-centric pretraining is necessary or sufficient for context-grounded calibration. This does not settle factuality outside the reported task settings. It limits only what may be inferred from the study's task-conditional calibration findings [1]–[3].

II. Methods

The synthesis admits only audited claims whose stated task, language coverage, protocol, and reported comparison match the question. A study was not treated as evidence of instruction following merely because it measured general task accuracy, nor as evidence of open-world factuality merely because it measured context-grounded hallucination judgments. Safety-refusal evaluation was kept separate from general instruction adherence and helpful task completion [1], [4], [9].

Comparability checks considered language inclusion, prompting and fine-tuning conditions, model roster, category definition, and aggregation before a category-level conclusion was drawn. The supplied MixSarc evidence concerns zero-shot classification by three API-accessed instruction-following models without task fine-tuning, but it lacks an explicit Bangla-specific-versus-multilingual comparison. As an audit limitation, the available record does not document the actual mixing, script regime, or mixing prevalence needed to characterize MixSarc or BenHalluEval as genuinely Bangla-English code-mixed [7], [8].

Version control was retained as unresolved. The supplied TigerLLM/BenHalluEval material includes a seven-model description and a nine-model description across three categories. The audit identifies these as version 1 and version 4 records, respectively, and requires their reconciliation before assertions about rank or sole-category representation. This constraint prevents stronger roster-dependent interpretation in the present synthesis [1]–[3].

Absence findings are bounded to the verified supplied record. They do not demonstrate that no relevant public Bangla evaluation exists beyond that record. IndicIFEval could become relevant to a future synthesis only if Bangla coverage and per-model Bangla results are independently established [4]–[6].

III. Results

A. Factuality and context-grounded calibration

BenHalluEval does not measure general open-world factual accuracy. Its human annotation instruction requires judgments only against supplied context and prohibits use of outside knowledge. Its findings therefore bear on context-grounded hallucination or calibration, not closed-book factual-question-answering accuracy, retrieval accuracy, or factual consistency outside a provided context [1].

For that construct, the supplied findings report task-conditional rankings and state that Bengali-centric pretraining is neither necessary nor sufficient for calibration. The retained inference is correspondingly narrow: the evidence does not establish a reliable category-wide calibration benefit from

Bengali-centric pretraining. It does not generalize beyond the study's stated Bengali and Bangla-English task settings [1]–[3].

Interpretation is further constrained by the unresolved roster discrepancy. One supplied TigerLLM/BenHalluEval excerpt reports seven evaluated models, while another reports nine models across three categories. The audit requires reconciliation of the version 1 and version 4 roster descriptions before rank or sole-category-representation claims are made [1]–[3].

B. Instruction following

The verified Bengali evaluation of 10 open-source models on eight translated datasets does not test instruction adherence as a distinct construct. Its enumerated datasets concern commonsense, science, mathematics, and multidomain benchmarks translated into Bengali. General accuracy on these tasks cannot establish multi-constraint instruction following, output-format compliance, or instruction completeness; the study's blind human review concerns translation quality used in benchmark construction rather than model instruction adherence [4].

No usable direct Bangla-specific-versus-multilingual comparison on Bangla instruction following appears in the supplied evidence. XIFBench's listed languages exclude Bangla. The supplied IndicIFEval excerpt neither establishes Bangla as an evaluated language nor reports the requested category comparison. Its grounded subset covers 25 original IFEval constraint categories, uses single-annotator annotations, and limits reasoning analysis to Qwen models [4]–[6].

C. Dataset-characterization gaps and adjacent evidence

TABLE I
Table 1. Evidence domains, permitted conclusions, and unresolved constraints in the audited record. [1]–[9]

Requested domain	Retained evidence; Permitted conclusion; Principal constraint
Context-grounded calibration	*Retained evidence:* BenHalluEval human judgments against supplied context *Permitted conclusion:* Bengali-centric pretraining is not established as necessary or sufficient for calibration *Principal constraint:* This is not open-world factual accuracy; the supplied seven- and nine-model records require reconciliation before roster-dependent claims
Bangla instruction following	*Retained evidence:* Translated Bengali benchmarks; XIFBench; IndicIFEval excerpts *Permitted conclusion:* No usable direct category comparison is available in the verified record *Principal constraint:* The translated study is not an instruction-following test; Bangla coverage is unestablished for IndicIFEval
Requested Bangla-English code-mixed domain	*Retained evidence:* MixSarc zero-shot classification *Permitted conclusion:* No explicit category comparison is available *Principal constraint:* Audit limitation: actual mixing, script regime, and mixing prevalence are not documented in the retained record
Cross-script Banglish safety	*Retained evidence:* BanglaVeilGuard *Permitted conclusion:* Adjacent safety evidence only *Principal constraint:* It does not establish factuality, general instruction adherence, or code-mix naturalness comparisons

MixSarc reports full-test zero-shot classification results for three named API-accessed instruction-following models without task fine-tuning. The source reports micro-F1 values of 0.5615 for LLaMA-3.1-8b, 0.5610 for Gemini-3-Flash, and 0.3582 for the model named in the source table as LLaaMa-3.3-70b-versatile. These descriptive results do not provide an explicit Bangla-specific-versus-multilingual comparison. The audit further identifies insufficient dataset-level documentation of actual Bangla-English mixing, script regime, and mixing prevalence for characterizing MixSarc as genuinely code-mixed [7], [8].

BanglaVeilGuard supplies adjacent evidence of cross-script Banglish and code-mixed Bangla-English safety evaluation. Its framing does not make it evidence for factuality, general instruction adherence, or comparative code-mix naturalness between Bangla-specific and multilingual models, and the supplied excerpt does not report a corresponding category-performance table [9].

IV. Discussion

The calibration evidence permits rejection of a simple pretraining-category rule within the study's own terminology. Task-conditional rankings, together with the statement that Bengali-centric pretraining is neither necessary nor sufficient for calibration, do not sustain a claim that Bengali-centric pretraining reliably improves context-grounded hallucination calibration. This conclusion does not determine whether a model can lead on a particular task within the study's limited scope [1]–[3].

That inference should not be extended from supplied-context judgments to open-world factuality, retrieval, or factual consistency outside the context. Nor should the evidence be converted into an exact roster-dependent ranking while the TigerLLM/BenHalluEval version 1 seven-model and version 4 nine-model records remain unreconciled. The audit therefore retains rank and sole-category-representation claims as unresolved [1]–[3].

Instruction-following evidence remains indeterminate because the relevant studies do not jointly establish the required construct, Bangla coverage, and category contrast. Translated academic benchmarks are not direct measures of multi-constraint instruction adherence. XIFBench excludes Bangla in its listed languages, and the supplied IndicIFEval material does not establish Bangla inclusion or report a Bangla-specific-versus-multilingual result. IndicIFEval-Ground's 25-category coverage, single-annotator protocol, and Qwen-limited reasoning analysis are additional constraints on future use. A scope-sensitive candidate counterevidence item was excluded by the independent audit and is not interpreted as a retained comparative result [4]–[6].

MixSarc's reported model scores do not resolve the requested comparative domain because the required model-category comparison is absent. The audit also finds that the retained record lacks dataset-level evidence sufficient to characterize MixSarc or BenHalluEval as genuinely Bangla-English code-mixed in terms of actual mixing, script regime,

and mixing prevalence. BanglaVeilGuard demonstrates that cross-script Banglish safety can be evaluated, but safety refusal and unsafe-request bypass remain distinct from factuality, general instruction adherence, and code-mix naturalness [7]–[9].

For deployment-oriented interpretation, the retained evidence supports only a provisional decision rule: assess candidate models against the organization's own task, risk, language, script, and applicability requirements before relying on these benchmark findings. This is an analytical recommendation, not a universally validated minimum-capability framework. Its use requires organization-specific risk and applicability assessment [1]–[8].

V. Conclusion

Within the audited public-study record, Bengali-centric pretraining is not established as necessary or sufficient for context-grounded hallucination calibration. BenHalluEval's human protocol measures consistency with supplied context rather than general open-world factual accuracy, and its task-conditional findings do not establish a category-wide reliable calibration benefit [1]–[3].

The record provides no usable direct Bangla-specific-versus-multilingual conclusion for Bangla instruction following. It also provides no explicit category conclusion from MixSarc for the requested Bangla-English code-mixed domain. Stronger synthesis requires direct task-level category comparisons, demonstrated Bangla coverage where relevant, dataset-level documentation of mixing, scripts, and prevalence, and reconciliation of the TigerLLM/BenHalluEval version 1 and version 4 model rosters before roster-dependent conclusions are asserted [1]–[8].

The conclusions are bounded to the verified supplied record and do not establish that no other public Bangla evaluation exists. BenHalluEval should not be interpreted as an open-world factuality benchmark. The supplied TigerLLM/BenHalluEval records retain an unresolved version 1 seven-model versus version 4 nine-model discrepancy; the audit requires reconciliation before rank or sole-category-representation assertions. The instruction-following record does not establish Bangla coverage for IndicIFEval and includes its grounded-subset limitations. The requested Bangla-English code-mixed domain remains unverified for BenHalluEval and MixSarc because the retained record lacks sufficient dataset-level evidence of actual mixing, script regime, and mixing prevalence. BanglaVeilGuard remains adjacent safety evidence rather than a resolution of the requested comparative constructs. A scope-sensitive candidate counterevidence item was excluded by the independent audit and is not used to support the paper's conclusions [1]–[9].

AI Assistance Disclosure

This report is an AI-assisted synthesis of retained public sources; it does not report original interviews or experiments. It presents a bounded synthesis and a provisional decision rule, not a universally validated minimum-capability model or general framework.

References

[1] "BenHalluEval: A Multi-Task Hallucination Evaluation Framework for Large Language Models on Bengali," *arXiv preprint arXiv:2605.31483*. 2026. Accessed: Sep. 1, 2026. [Online]. Available: https://arxiv.org/html/2605.31483

[2] "BenHalluEval: A Multi-Task Hallucination Evaluation Framework for Large Language Models on Bengali," *arXiv preprint arXiv:2605.31483v1*. 2026. Accessed: Sep. 1, 2026. [Online]. Available: https://arxiv.org/html/2605.31483v1

[3] "BenHalluEval: A Multi-Task Hallucination Evaluation Framework for Large Language Models on Bengali," *arXiv preprint arXiv:2605.31483v4*. 2026. Accessed: Sep. 1, 2026. [Online]. Available: https://arxiv.org/html/2605.31483v4

[4] "Evaluating LLMs' Multilingual Capabilities for Bengali," *arXiv preprint arXiv:2507.23248v1*. 2025. Accessed: Sep. 1, 2026. [Online]. Available: https://arxiv.org/html/2507.23248v1

[5] "XIFBench: Evaluating Large Language Models on Multilingual Instruction Following," *arXiv preprint arXiv:2503.07539v2*. 2025. Accessed: Sep. 1, 2026. [Online]. Available: https://arxiv.org/html/2503.07539v2

[6] "IndicIFEval: A Benchmark for Verifiable Instruction-Following Evaluation in 14 Indic Languages," *arXiv preprint arXiv:2602.22125v1*. 2026. Accessed: Sep. 1, 2026. [Online]. Available: https://arxiv.org/html/2602.22125v1

[7] "MixSarc: A Bangla-English Code-Mixed Corpus for Implicit Meaning Identification," *arXiv preprint arXiv:2602.21608v1*. 2026. Accessed: Sep. 1, 2026. [Online]. Available: https://arxiv.org/html/2602.21608v1

[8] K. S. Y. Alam, M. T. Chowdhury, T. Ahmed, A. Abrar, and M. R. Haque, "MixSarc: A Bangla-English code-mixed corpus for implicit meaning identification," *arXiv preprint arXiv:2602.21608*. 2026. Accessed: Sep. 1, 2026. [Online]. Available: https://arxiv.org/pdf/2602.21608

[9] "BanglaVeilGuard: Cross-Script Safety Benchmarking and Lightweight Guardrails for Bangla Large Language Models," *arXiv preprint arXiv:2608.21880v1*. 2026. Accessed: Sep. 1, 2026. [Online]. Available: https://arxiv.org/html/2608.21880v1