

How Reliable Are Bengali LLM Benchmarks? A Cross-Lingual Contamination and Robustness Audit of Open-Weight Language Models

Md Fahim Faisal Tanha

mdfahimfaisaltanha@gmail.com

Independent Researcher

Research Article

Keywords: large language models, benchmark contamination, Bengali, low-resource languages, evaluation reliability, robustness

Posted Date: August 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10491036/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Bengali, spoken by roughly a quarter of a billion people, has recently gained a growing collection of large language model (LLM) evaluation benchmarks, most of which are direct translations of established English test sets. At the same time, benchmark data contamination (BDC) - the leakage of evaluation items into model training corpora - is now a well-documented threat to the validity of LLM evaluation, and recent evidence shows that contamination can cross language barriers: a model exposed to an English test set can obtain inflated scores on its translated counterpart. Whether Bengali benchmarks suffer from such contamination, and how stable their results are under semantically equivalent perturbations, has never been examined. We present the first systematic audit framework for the reliability of Bengali LLM benchmarks, together with a completed pilot validating the full pipeline on free-tier compute. The framework combines (i) likelihood-based membership signals (Min-K% Prob), (ii) generation-based tests (TS-Guessing-style option completion), (iii) ground-truth n-gram overlap against fully public pre-training corpora, and (iv) a perturbation suite measuring the stability of benchmark rankings. In the pilot (Qwen2.5-1.5B-Instruct on 500 BoolQ-bn items), accuracy is 0.568 (95% CI [0.524, 0.610]) - barely above the 0.50 chance level of the binary task - and a distractor perturbation shifts accuracy by -1.0 pp (McNemar $p = 0.302$). The complete model x benchmark audit grid is being executed with the released protocol; this manuscript reports the framework, benchmark inventory, protocol, and pilot. All code, data manifests, and per-item audit outputs are publicly released.

1. Introduction

Large language models are now routinely marketed and adopted as multilingual systems, and their reported performance on non-English benchmarks increasingly informs deployment decisions in education, public services, and content moderation. Bengali (Bangla) is a striking case: it is among the most widely spoken languages in the world, yet it remains computationally low-resourced, and the infrastructure for evaluating LLMs in Bengali has emerged only in the last two to three years. This new infrastructure includes native-content benchmarks such as BEnQA (parallel Bengali-English science exam questions) and BanglaMATH (elementary-school math word problems), as well as large translated suites such as Bangla MMLU and Bengali versions of PIQA, CommonsenseQA, OpenBookQA, and BoolQ.

The reliability of these benchmarks, however, has not kept pace with their adoption. Two threats stand out. First, benchmark data contamination (BDC): because LLM pre-training corpora are scraped at web scale, evaluation items frequently leak into training data, inflating scores and invalidating comparisons. Contamination has been documented across popular English benchmarks and, critically, has been shown to cross language barriers - a model whose training data contains the English source of a translated benchmark can achieve inflated scores on the translation, even without ever seeing the translated text. Since most Bengali benchmarks are translations of heavily contaminated English test sets, they are structurally exposed to exactly this risk. Second, robustness: recent work shows that LLM accuracy on low-resource-language benchmarks can collapse under semantically irrelevant perturbations, suggesting that headline scores may reflect surface memorization rather than capability.

Despite a rapidly growing literature on contamination detection, no study has audited Bengali benchmarks. Multilingual contamination reports cover broad multi-language suites but do not include Bengali-specific benchmarks, and the papers introducing Bengali benchmarks do not measure contamination in the models

they evaluate. This leaves a foundational question unanswered: when an LLM scores well on a Bengali benchmark, does that number mean what we think it means?

We address this question by developing the first systematic reliability-audit framework for Bengali LLM benchmarks and validating it end to end in a completed pilot (Section 7.0); the full audit grid is in progress using the released protocol. We deliberately audit open-weight models: they expose token-level likelihoods (enabling stronger, membership-style contamination tests than black-box APIs allow), several of them publish their full pre-training corpora (enabling ground-truth contamination measurement), and they include the first generation of Bangla-specific LLMs, which have never been audited. Our contributions are:

1. A curated, versioned inventory of public Bengali LLM benchmarks (native and translated), with provenance, size, licensing, and construction method documented for each (Section 3).
2. A contamination-audit protocol for Bengali benchmarks (the first of its kind), combining likelihood-based signals (Min-K% Prob), generation-based tests (TS-Guessing-style), and exact ground-truth n-gram overlap against fully public pre-training corpora (Dolma, The Pile, ROOTS) (Sections 5.1–5.2).
3. The first test of cross-lingual contamination for Bengali, exploiting the parallel Bengali-English design of BEnQA to separate source-language leakage from target-language leakage (Section 5.2).
4. A robustness audit measuring how Bengali benchmark scores and model rankings change under meaning-preserving perturbations (Section 5.3), with statistical significance analysis throughout.
5. Practical recommendations and a public release of all audit code and outputs, enabling the community to re-run the audit as new Bengali benchmarks and models appear.

Research questions. RQ1: Do open-weight LLMs exhibit measurable contamination signals on public Bengali benchmarks, and do these signals differ between native and translated benchmarks? RQ2: Is there evidence of cross-lingual contamination - i.e., do models show stronger memorization signals on English source items than on their Bengali counterparts, coupled with performance gaps consistent with source-language leakage? RQ3: How robust are Bengali benchmark scores and model rankings to meaning-preserving perturbations, and is fragility associated with contamination signals?

2. Related Work

2.1 Bengali and low-resource LLM benchmarks

Bengali evaluation resources have expanded quickly. BEnQA provides approximately 5K parallel Bengali-English middle- and high-school science questions and documents a notable performance disparity between the two languages. The TituLLMs effort released a family of Bangla LLMs together with a benchmarking suite of roughly 132K items, dominated by Bangla MMLU (approx. 88K items) alongside Bengali versions of PIQA, CommonsenseQA, OpenBookQA, and BoolQ. TigerLLM likewise couples new Bangla models with evaluation sets. BanglaMATH contributes 1.7K native Bangla math word problems sourced from elementary-school workbooks and reports that leading models lack robustness and exhibit performance bias in Bangla. A complementary 2025 effort translated twenty major English benchmarks into Bengali, retaining eight after

quality filtering, and reports consistent Bengali-English performance gaps across ten open models. None of these works measures contamination.

2.2 Benchmark data contamination

Contamination of LLM evaluation benchmarks is now extensively documented for English. Surveys of BDC synthesize evidence that test items of popular benchmarks appear in pre-training corpora and that contaminated models obtain materially inflated scores; work on software-engineering benchmarks reports Pass@1 inflation by a factor of several times on leaked samples. Position papers have argued that per-benchmark contamination measurement should accompany any LLM evaluation. Proposed mitigations include dynamic and continuously refreshed benchmarks (e.g., LiveBench) and encryption or canary protocols for test sets.

2.3 Contamination detection methods

Detection methods fall into three families relevant to our setting. (i) Likelihood/membership methods, such as Min-K% Prob, score whether a text was likely present in training data using the model's own token probabilities; these require logit access and are therefore well suited to open-weight models. (ii) Generation-based methods, such as guided-instruction prompting and TS-Guessing, test whether a model can complete or reconstruct held-out parts of benchmark items (e.g., a masked incorrect option) beyond chance. (iii) Corpus-search methods measure exact or near-duplicate n-gram overlap between benchmark items and the pre-training corpus; they provide ground truth but require corpus access, which only fully open models (e.g., OLMo/Dolma, Pythia/The Pile, BLOOM/ROOTS) permit. Statistically grounded procedures for proving test-set contamination with false-positive guarantees also exist for the exchangeable-benchmark setting. We adopt one method from each family to triangulate.

2.4 Cross-lingual contamination

Most contamination work is monolingual, but recent studies demonstrate that contamination generalizes across languages: models can benefit from having seen the English source of a translated test set, a phenomenon invisible to surface-form overlap checks in the target language. A NeurIPS 2024 workshop report found signs of contamination in nearly all popular multilingual benchmarks across seven major LLMs. Translated Bengali benchmarks inherit this exposure channel directly from their English sources, which motivates our RQ2 design.

2.5 Robustness and dynamic evaluation

Perturbation-based evaluation - paraphrasing, distractor insertion, option reordering, and numeric substitution - has emerged as a practical probe of whether benchmark performance reflects generalizable capability. BanglaMATH already indicates fragility of top models under distraction in Bangla; enterprise-oriented robustness benchmarks report that perturbation consistency varies non-monotonically with model size. We integrate perturbation testing not as a new benchmark but as a reliability instrument: if a model's score drops sharply under meaning-preserving edits, and that model also shows strong memorization signals on the original items, contamination is the parsimonious explanation.

3. Benchmark Inventory

Table 1 lists the public Bengali benchmarks included in the audit. We distinguish native benchmarks (authored originally in Bengali from Bengali-language sources) from translated benchmarks (derived from an English test set by machine or human translation), because the two classes have different contamination exposure channels: native items can leak only through Bengali web text, while translated items are additionally exposed through their English sources. For every benchmark we record a frozen snapshot (commit hash or dataset revision) in the released audit manifest so that the audit is reproducible.

Table 1

Public Bengali LLM benchmarks included in the audit. Sizes and licenses verified against Hugging Face and GitHub on 27 July 2026: boolq_bn, piqa_bn, and commonsenseqa_bn are MIT-licensed; titulum-bangla-mmlu states no license on its dataset card; BanglaMATH is distributed for research use only per its repository. Licenses for BEnQA and openbookqa_bn could not be confirmed online and are re-checked at download time. Exact counts and dataset revisions are recorded in the released manifest (Appendix A).

Benchmark	Task	Size (items)	Origin	Where to get it
BEnQA	Science exam QA (parallel Bn-En)	5,161	Native (parallel EN)	github.com/sheikhshafayat/BEnQA
Bangla MMLU (TituLLMs suite)	Multi-domain MCQ	87,869 (all subset)	Native + curated	HF: hishab/titulum-bangla-mmlu
PIQA-bn (TituLLMs suite)	Physical commonsense MCQ	17,177	Translated (PIQA)	HF: hishab/piqa-bn
CommonsenseQA-bn (TituLLMs suite)	Commonsense MCQ	10,962	Translated (CSQA)	HF: hishab/commonsenseqa-bn
OpenBookQA-bn (TituLLMs suite)	Science MCQ	~ 5,944	Translated (OBQA)	HF: hishab/openbookqa-bn
BoolQ-bn (TituLLMs suite)	Yes/no reading comprehension	1,976	Translated (BoolQ)	HF: hishab/boolq_bn
BanglaMATH	Math word problems (grades 6–8)	~ 1,700	Native	github.com/TabiaTanzin/BanglaMATH-...
Bengali translated suite (2025)	8 benchmarks incl. commonsense, science, math	varies	Translated (20->8 sets)	arXiv:2507.23248 release page

Inclusion criteria: (i) publicly downloadable without access restrictions; (ii) documented construction method; (iii) objective answer keys permitting automated scoring. Where a benchmark ships multiple splits, we audit the test split used in the corresponding paper's headline results. For tractability under free compute, benchmarks larger than 5,000 items are audited on a stratified random subsample of 1,000–2,000 items drawn with a fixed seed (Section 6); contamination and robustness statistics are reported with bootstrap confidence intervals that reflect the subsampling.

4. Audited Models

We audit open-weight instruction-tuned models in the 1B-9B parameter range, selected to satisfy four criteria: (i) weights and tokenizers are publicly downloadable; (ii) the model is runnable on free-tier GPU compute (Kaggle/Colab, 16 GB VRAM) in 4-bit quantization; (iii) the set spans model families with different data pipelines; and (iv) the set includes Bangla-specific models and fully-open-data models. Tier A models expose logits (all methods except corpus search apply); Tier B models additionally publish their complete pre-training corpora, enabling ground-truth corpus-search contamination measurement.

Table 2

Audited models. Tier A: open weights (logit access). Tier B: open weights and fully public pre-training corpora (ground-truth corpus search). Final list may be adjusted for availability; record exact checkpoints in the manifest.

Model	Params	Tier	Why included
Llama-3.2-1B / 3B Instruct	1B / 3B	A	Widely used small models; multilingual claims
Llama-3.1-8B Instruct	8B	A	Most common open baseline
Qwen2.5-1.5B / 7B Instruct	1.5B / 7B	A	Strong multilingual family
Gemma-2-2B / 9B IT	2B / 9B	A	Distinct data pipeline (Google)
Mistral-7B Instruct v0.3	7B	A	Family reported weak on Bengali
TituLLM-1B / 3B	1B / 3B	A	Bangla-specific; releases own benchmark suite
TigerLLM-1B	1B	A	Bangla-specific; never audited; weights: HF md-nishat-008/TigerLLM-1B-it
OLMo-2-7B	7B	B	Fully public pre-training data (Dolma)
Pythia-6.9B	6.9B	B	Fully public data (The Pile); pre-benchmark-era control
BLOOM-7B1	7.1B	B	Fully public data (ROOTS); explicit Bengali fraction

Pythia deserves note as a temporal control: its training corpus (The Pile) predates every Bengali benchmark in Table 1, so any positive contamination signal on Pythia estimates the false-positive rate of our detectors in the Bengali setting - an important calibration given that membership tests are known to be sensitive to distribution shift.

5. Methodology

5.1 RQ1: Contamination signals on Bengali benchmarks

We triangulate three detector families; no single method is conclusive, but concordant evidence across families is strong evidence of contamination.

- Min-K% Prob (likelihood-based). For each benchmark item (question + options + answer, formatted exactly as distributed), we compute the average log-probability of the k% least likely tokens under the model, for k in {10,

20}. Items in training data have systematically higher Min-K% scores. We compare each benchmark's score distribution against (a) the same model's scores on a freshly authored, never-published probe set of matched format and difficulty (constructed by us, kept offline until after all runs), and (b) the Pythia temporal control. We report AUROC of separating benchmark items from probe items, with 95% bootstrap CIs.

- TS-Guessing-style option reconstruction (generation-based). For multiple-choice items we mask one incorrect option and prompt the model to produce the missing option verbatim. Reconstruction accuracy meaningfully above a paraphrase-tolerant chance baseline (estimated on the probe set) indicates memorization of the test file rather than task ability. We report exact-match and ROUGE-L-thresholded reconstruction rates.
- Ground-truth corpus search (Tier B only). We search Dolma, The Pile, and ROOTS for 13-gram overlaps (and normalized near-duplicates) with each benchmark item, in both Bengali and English forms where applicable, using the corpora's public search/index tooling. This yields per-benchmark contamination rates that are facts, not inferences, for Tier B models.

5.2 RQ2: Cross-lingual contamination

BEnQA's parallel design lets us isolate the cross-lingual channel: for each item we hold content constant and vary only language. For every Tier A model we compute the paired difference in Min-K% signal between the English and Bengali versions of the same item, and the paired difference in accuracy. Evidence consistent with cross-lingual contamination is the conjunction: (i) stronger memorization signal on English items than on their Bengali twins; (ii) higher accuracy on English; and (iii) within-model, item-level correlation between the English memorization signal and correctness on the Bengali version (a model "knowing" an item in English predicts solving its Bengali twin). For translated benchmarks (Bangla MMLU etc.), we additionally run detectors on the English source items and compare per-item signals between source and translation. Paired statistics use Wilcoxon signed-rank tests; correlations are Spearman, with Holm-Bonferroni correction across models.

5.3 RQ3: Robustness of scores and rankings

We construct meaning-preserving perturbation suites per benchmark: (i) option reordering (MCQ); (ii) distractor sentence insertion (adding an irrelevant but topical sentence, following the BanglaMATH protocol); (iii) numeric substitution for math items (same operations, different numbers, answers recomputed programmatically); and (iv) paraphrase via round-trip machine translation (Bn->En->Bn with an open MT model, e.g. NLLB-200), with automatic filtering by back-translation semantic similarity and manual verification of a 100-item sample by native speakers. For each model-benchmark pair we report the accuracy delta between original and perturbed versions, and for each benchmark the Spearman correlation between the model ranking on original vs. perturbed items. Finally, we test the diagnostic association predicted by the contamination hypothesis: across items, does the memorization signal on the original item predict the probability that the model fails its perturbed twin? We fit a mixed-effects logistic regression with model and benchmark random effects.

5.4 What the audit can and cannot conclude

Likelihood- and generation-based detectors provide evidence, not proof, for Tier A models; only Tier B corpus search is definitive. We therefore report all three signal families separately, never aggregate them into a single "contamination score", and interpret Tier A findings through the calibration provided by the probe set and the

Pythia temporal control. Conversely, absence of signal does not certify a benchmark clean: contamination can be paraphrastic or cross-lingual in ways that evade all current detectors.

6. Experimental Setup

- **Compute.** All experiments run on free-tier GPU compute: Kaggle (2x T4, ~ 30 GPU-hours/week) and Google Colab (T4). Models of 7B-9B are loaded in 4-bit (bitsandbytes NF4); 1B-3B models in fp16. No paid APIs are used anywhere in the study.
- **Harness.** Accuracy evaluations use lm-evaluation-harness with its existing Bangla task definitions where available; custom tasks are registered for BEnQA, BanglaMATH, and perturbed variants. Decoding is deterministic (temperature 0); MCQ scoring uses length-normalized option log-likelihood; math scoring uses exact numeric match after Bengali/Arabic numeral normalization.
- **Sampling.** Benchmarks over 5,000 items are subsampled to 1,000–2,000 items, stratified by subject/category, fixed seed 20260727. The same item sample is used across all models and all conditions to keep comparisons paired.
- **Statistics.** 95% CIs by nonparametric bootstrap (10,000 resamples) at item level; paired tests as specified in Section 5; multiple-comparison control by Holm-Bonferroni within each research question. Significance level 0.05.
- **Reproducibility.** A machine-readable manifest records dataset revisions, model checkpoints, prompts, seeds, and software versions. All audit outputs (per-item scores) are released as CSV.

Measured pilot budget: one complete model x benchmark cell (500 items: accuracy, Min-K%, and one perturbation pass) took 0.56 GPU-hours for a 1.5B model on a single T4 (Section 7.0). Naively extrapolating to the full grid (about 12 models x 8 benchmarks, with 7-9B models roughly 4-6x slower per token in 4-bit), the complete audit requires on the order of 100–200 GPU-hours, i.e. 4–7 weeks of Kaggle free quota; trimming the model list or subsample sizes reduces this proportionally.

7. Results

Section 7.0 reports the completed pilot measurements; all numbers are real outputs of the released code. The remaining model x benchmark audit grid (Sections 7.1–7.3) is being executed with the released protocol; cells marked in progress will be reported in a subsequent version of this work.

7.0 Pilot results (completed 27 July 2026)

To validate the pipeline end to end (Appendix A, Step 3), we executed the full pilot on free-tier compute (Google Colab, single T4 GPU): Qwen2.5-1.5B-Instruct evaluated on a 500-item random subsample of BoolQ-bn (seed 20260727). All numbers below are measured outputs of the released code; per-item files (accuracy_items.csv, mink_items.csv, pilot_results.csv) are part of the released audit outputs.

- Accuracy: 0.568 (95% bootstrap CI [0.524, 0.610]). Against the 0.50 chance level of a binary task, the model performs only about 7 points above coin-flipping, consistent with limited genuine Bengali reading comprehension at the 1.5B scale. The majority-class baseline on this subsample is 0.534, so the model outperforms always answering the majority class by only 3.4 points.
- Distractor perturbation: accuracy 0.558 (95% CI [0.516, 0.600]); paired delta – 1.0 pp; exact McNemar test on 15 discordant items (10 correct-to-wrong vs 5 wrong-to-correct), $p = 0.302$ (not significant). A single 500-item cell is underpowered for deltas of this size; the full grid provides the powered test.
- Min-K% Prob ($k = 20$): mean – 3.767 (SD 0.187) over the 500 items. The contamination AUROC requires the private probe-set baseline (Section 5.1) and is therefore pending.
- Cost: 0.56 GPU-hours for one complete model x benchmark cell (accuracy + Min-K% + one perturbation pass), confirming the free-compute budget of Section 6.

7.1 Contamination signals (RQ1)

Table 3

Contamination signals per model-benchmark pair. Remaining cells are being produced with the released pipeline and will be reported in a subsequent version.

Model	Benchmark	Min-K% AUROC (95% CI)	TS-Guessing reconstruction rate	Corpus-search overlap (Tier B)
Qwen2.5-1.5B-Instruct (pilot)	BoolQ-bn (n = 500)	mean Min-K%(20) = -3.767 (SD 0.187); AUROC pending probe set	n/a (binary task)	n/a (Tier A)
Remaining grid	in progress	in progress	in progress	in progress

7.2 Cross-lingual contamination (RQ2)

Table 4

Cross-lingual analysis on BEnQA parallel items and translated-suite sources.

Model	Paired Min-K% diff (En - Bn)	Paired accuracy diff (En - Bn)	Item-level Spearman (En signal vs Bn correctness)	p (Holm-Bonf.)
BEnQA parallel items - full grid	in progress	in progress	in progress	in progress

7.3 Robustness (RQ3)

Table 5

Score and ranking stability under meaning-preserving perturbations.

Benchmark	Mean accuracy delta under perturbation	Ranking stability (Spearman rho)	Memorization-fragility association (mixed-effects OR, 95% CI)
BoolQ-bn - pilot (Qwen2.5-1.5B, distractor)	-1.0 pp (0.568 -> 0.558); McNemar $p = 0.302$	n/a (single model)	pending full grid
Remaining grid	in progress	in progress	in progress

8. Discussion

We organize the discussion around three questions. (1) Validity: which Bengali benchmarks, if any, currently support trustworthy claims about model capability, and for which model families? If translated benchmarks show systematically stronger contamination signals than native ones, headline multilingual claims built on them need re-examination; if they do not, that is itself an important negative result given the strong prior from cross-lingual contamination work. (2) Practice: what should Bengali benchmark builders change - e.g., preferring native sourcing, withholding encrypted test splits, releasing canary strings, versioning with release dates so temporal controls are possible? (3) Generalization: Bengali is a template for dozens of mid-sized languages whose benchmarks are translation-derived; the audit protocol released here is language-agnostic and cheap enough to re-run elsewhere.

9. Limitations

- Closed frontier models (GPT/Claude/Gemini class) are out of scope because they expose neither logits nor training data; our conclusions apply to open-weight models, and we frame closed-model auditing as future work.
- Likelihood- and generation-based detectors are indirect for Tier A models; we mitigate with a never-published probe set and a temporal control, but residual false positives/negatives are possible.
- Free-tier compute constrains model size ($\leq 9B$) and forces subsampling of the largest benchmarks; CIs reflect subsampling, but small effects may be underpowered.
- Round-trip MT paraphrases can introduce subtle meaning drift in Bengali; we filter automatically and verify a manual sample, but perturbation quality is bounded by open MT quality for Bengali.
- The benchmark inventory reflects a snapshot date; Bengali evaluation is moving quickly and the audit will need periodic re-runs.

Declarations

Ethics and Data Availability

All datasets and models used are publicly released artifacts used under their stated licenses; no personal data is collected. The study involves no human subjects beyond voluntary native-speaker verification of paraphrase quality by the author(s). All code, manifests, and per-item audit outputs are publicly available at <https://github.com/mdfahimfaisaltanha/bengali-llm-benchmark-audit>; the private probe set will be released only after publication to preserve its integrity. AI assistance: parts of the literature scoping and manuscript drafting were assisted by an AI writing tool; all experimental results are produced by the released code and verified by the author(s).

References

1. Shafayat, S., Hasan, H. M. Q., Mahim, M. R. C., Putri, R. A., Thorne, J., & Oh, A. (2024). BEnQA: A Question Answering Benchmark for Bengali and English. Findings of ACL 2024. arXiv:2403.10900.
2. Nahin, S. K. (2025). TituLLMs: A Family of Bangla LLMs with Comprehensive Benchmarking. Findings of ACL 2025. arXiv:2502.11187.
3. Raihan, N. (2025). TigerLLM: A Family of Bangla Large Language Models. ACL 2025 (short papers).
4. Prama, T. T., Danforth, C. M., & Dodds, P. S. (2025). BanglaMATH: A Bangla Benchmark Dataset for Testing LLM Mathematical Reasoning at Grades 6, 7, and 8. Proceedings of the 3rd MathNLP Workshop (EMNLP 2025). arXiv:2510.12836.
5. Bhowmik, S., Dipto, T. T., Islam, M. S., Hsu, S., & Reasat, T. (2025). Evaluating LLMs' Multilingual Capabilities for Bengali: Benchmark Creation and Performance Analysis. arXiv:2507.23248.
6. Xu, C., Guan, S., Greene, D., & Kechadi, M-T. (2024). Benchmark Data Contamination of Large Language Models: A Survey. arXiv:2406.04244.
7. Yao, F., Zhuang, Y., Sun, Z., Xu, S., Kumar, A., & Shang, J. (2024). Data Contamination Can Cross Language Barriers. EMNLP 2024. arXiv:2406.13236.
8. Ahuja, S., Gumma, V., & Sitaram, S. (2024). Contamination Report for Multilingual Benchmarks. arXiv:2410.16186.
9. Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., & Zettlemoyer, L. (2024). Detecting Pretraining Data from Large Language Models. ICLR 2024. arXiv:2310.16789. (Min-K% Prob).
10. Oren, Y. (2024). Proving Test Set Contamination in Black-Box Language Models. ICLR 2024.
11. Golchin, S., & Surdeanu, M. (2024). Time Travel in LLMs: Tracing Data Contamination in Large Language Models. ICLR 2024 (Spotlight). arXiv:2308.08493.
12. Sainz, O. (2023). NLP Evaluation in Trouble: On the Need to Measure LLM Data Contamination for Each Benchmark. Findings of EMNLP 2023.
13. White, C., Dooley, S., Roberts, M., Pal, A. (2024). LiveBench: A Challenging, Contamination-Limited LLM Benchmark. arXiv:2406.19314. [Note: title updated from Contamination-Free in later versions.].
14. Chen, S., Chen, Y., Li, Z., Jiang, Y., Wan, Z., He, Y., Ran, D., Gu, T., Li, H., Xie, T., & Ray, B. (2025). Benchmarking Large Language Models Under Data Contamination: A Survey from Static to Dynamic Evaluation. EMNLP 2025. arXiv:2502.17521.
15. Groeneveld, D. OLMo: Accelerating the Science of Language Models. ACL 2024. /, Soldaini, L. (2024). (2024). Dolma: An Open Corpus of Three Trillion Tokens. ACL 2024.
16. Biderman, S. (2023). Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. ICML 2023.
17. BigScience, W. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. arXiv:2211.05100.
18. Gao, L. (2023). A Framework for Few-Shot Language Model Evaluation (lm-evaluation-harness). Zenodo.
19. Hendrycks, D. (2021). Measuring Massive Multitask Language Understanding. ICLR 2021.
20. Bogavelli, T., Bamgbose, O., Gauthier Melançon, G., Riols, F., & Sharma, R. (2026). Evaluating Robustness of Large Language Models in Enterprise Applications: Benchmarks for Perturbation Consistency Across Formats and Languages. arXiv:2601.06341.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AppendixA.docx](#)