Language Models are not Equally Robust to Non-Canonical Tokenization across Languages

Research group discussion · 25–30 minutes

Paper by Poulami Ghosh and Preethi Jyothi
IIT Bombay, India

arXiv:2607.26831v1 · Reading: 13 November 2026

The same string can trigger different multilingual behavior

Question

Does a model preserve its task performance when the input string stays identical but its valid token segmentation changes?

- Study scope: 27 languages, six generation-based tasks, three model families.
- English robustness does not generalize uniformly across languages.
- English-only multi-tokenization training improves multilingual ARC accuracy.

arXiv v1, 29 July 2026.
Publication not found as of 2026-11-13.
Public code not found as of 2026-11-13.

Paper, Secs. 1, 5–7; Tables 1 and 3, PDF pp. 1–2, 5–11. Status checked for this reading.

Token boundaries change the model's computation

Surface text	`abc`
Possible token sequence 1	`[abc]`
Possible token sequence 2	`[ab] [c]`
Possible token sequence 3	`[a] [b] [c]`

- All sequences must use existing vocabulary entries and decode identically.
- Embedding lookup, positions, and attention computations still change.
- Tokenizer pruning or replacement can expose a model to unfamiliar segmentations.

Teaching example with an assumed toy vocabulary; paper motivation in Secs. 1–2, PDF pp. 1–3.

Valid tokenizations define the perturbation space

$$t_{\text{canon}} = T_{\text{canon}}(x), \qquad \mathcal{T}(x) = \{t \in V^* : \text{decode}(t) = x\}.$$
$$f(t_1) \approx f(t_2), \qquad t_1, t_2 \in \mathcal{T}(x)$$

- x: a string; V: the original tokenizer vocabulary.
- t_{canon}: deterministic canonical segmentation; all alternatives preserve x.
- f: model logits, representations, or task outputs.
- The paper evaluates behavior through task scores and retrieval accuracy.

Paper, Sec. 3, PDF pp. 3–4. Non-canonical tokenizations additionally require $t \neq t_{\text{canon}}$.

Uniform paths require weighted edge sampling

An MDD links string positions with valid token edges. Let $C(i)$ count suffix paths:

$$C(m) = 1, \qquad C(i) = \sum_{(i,j)\in E} C(j), \qquad P(i \to j) = \frac{C(j)}{C(i)}.$$

First token for `abc`	Valid completions	Edge probability
`a`	2: `[b,c]`, `[bc]`	1/2
`ab`	1: `[c]`	1/4
`abc`	1: empty suffix	1/4

Each complete tokenization has probability 1/4. Uniform outgoing edges would bias the path distribution.

Teaching derivation of the paper's MDD sampling description; Sec. 3, PDF p. 4; original MDD examples in Fig. 2, PDF p. 5.

Perturb the user prompt while preserving its text

Condition	Input representation
Canonical	Default tokenizer output
Random	Uniformly sampled valid alternative segmentation; five seeds averaged
Character	Finest character-level condition used by the authors

- Retokenize instructions, question content, and multiple-choice options.
- Preserve the system prompt and special tokens.
- The character unit and treatment of the canonical path need implementation clarification.

Paper, Secs. 3–4, PDF pp. 3–5. Sampling formula uses $\mathcal{U}(\mathcal{T}(x))$ although the definition excludes t_{canon}.

Three model families, six tasks, 27 selected languages

- Llama-3.1-8B-Instruct, Qwen3-8B, gemma-3-12b-it.
- Short answer: MGSM and MLQA.
- Multiple choice: Multi-ARC, Multi-HellaSwag, Global-MMLU-Lite, Belebele.
- Direct generation with lm-evaluation-harness; MC instructions require answer letters only.

Scope of the comparison

The study covers 27 selected languages. Exact task-by-model language coverage is not fully tabulated; the reported averages use the paper's aggregation.

Paper, Sec. 4, PDF pp. 4–5; task details and language list in Tables 8–9, PDF p. 15.

Perturbation ranges differ across languages

$$\mathrm{FR} = \frac{\text{tokens}}{\text{words}}, \qquad \mathrm{LR}(t, x) = \frac{|t|}{|t_{\text{canon}}|}$$

Qwen tokenizer / language	Canonical FR	Character FR
English	1.35	5.16
Bengali	6.66	6.68
Gujarati	9.29	9.29
Telugu	13.55	13.55
Chinese	1.13	1.85

A small score change can reflect a small perturbation space. Several near-character languages are analyzed separately in the appendix.

Paper, Sec. 4 and Fig. 3(b), PDF pp. 4–5, 7; App. C, PDF pp. 13–14. FR values transcribed from Fig. 3.

Model robustness ranking changes with perturbation type

Model	Canonical score ↑	Random drop ↓	Character drop ↓
Llama	63.918	23.772%	40.960%
Qwen	64.748	11.391%	16.091%
Gemma	74.375	9.937%	20.928%

- Canonical scores and relative drops are averaged over the six tasks.
- Gemma has the lowest random drop; Qwen has the lowest character drop.
- Llama is most sensitive under both perturbations.

Paper, Table 1, PDF p. 6. Values are relative losses, not percentage-point losses. Language coverage follows the paper.

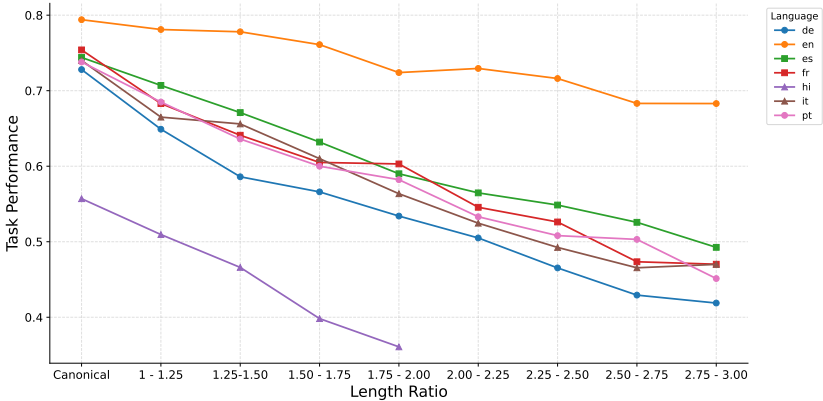
Task type alone does not determine sensitivity

Random relative drop (%)	Llama	Qwen	Gemma
MGSM	27.522	19.948	7.890
MLQA	24.446	12.002	14.708
Global-MMLU-Lite	20.371	8.957	9.766
Multi-ARC	24.320	8.542	9.865
Multi-HellaSwag	25.754	12.202	10.012
Belebele	20.221	6.692	7.382

Character perturbations cause larger drops in every model–task cell.
Short-answer tasks are often sensitive, but the pattern is not universal: Gemma's MGSM drop is below several MC drops.

Paper, Sec. 5.1 and Table 1, PDF pp. 5–6. Task scores are not assumed to share one metric definition.

Llama degrades as tokenization becomes more granular



The x-axis is sampled/canonical token count; colors identify languages. English is comparatively stable while several other languages decline.

Original Fig. 4(a), Multilingual-ARC, PDF p. 7. Canonical is a separate reference point.

Nearly equal sequence lengths still leave performance gaps

Language	Canonical	LR .95–1.00	LR 1.00–1.05
Arabic	78.30	74.73	73.00
English	88.00	–	87.90
Basque	45.60	40.75	41.10
Hungarian	78.90	73.33	71.10
French	80.90	–	81.30
Chinese	86.90	87.38	86.69

- Longer sequences alone do not explain the Arabic and Hungarian gaps.
- English and Chinese are nearly unchanged; French slightly improves.
- Table 6 does not explicitly identify the evaluated model.

Paper, Sec. 5.2, PDF p. 8; Table 6, Multilingual-ARC, PDF p. 14.

Canonical fragmentation predicts additional vulnerability

$$\operatorname{logit} P(Y_r = 1) = \beta_0 + \beta_c Y_c + \beta_f \mathrm{FR}_c$$

Language	Canonical-correct coefficient	Fragmentation coefficient
English	3.416	−1.114
Basque	2.211	−0.138
Italian	2.692	−0.146
Chinese	4.718	−0.967

Y_c, Y_r: correct/incorrect under canonical and random inputs.
English FR: $p = .021$, controlling for canonical correctness.
This is an association; other languages' significance values are not tabulated.

Teaching regression notation; paper, Sec. 5.1, PDF p. 6; Table 10, PDF p. 15.

Scaling improves robustness on the short-answer tasks

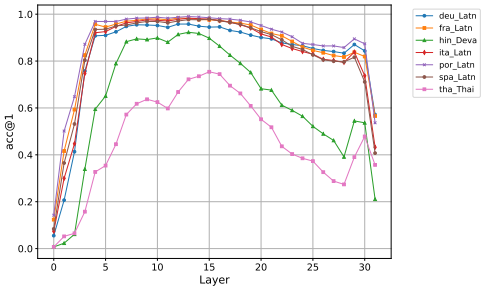
Task	Metric	Qwen 4B	Qwen 8B	Qwen 14B
MLQA	Canonical score	47.564	58.050	65.329
	Random drop (%)	14.648	12.002	10.841
MGSM	Canonical score	44.267	45.911	45.289
	Random drop (%)	21.085	19.948	16.208

- MGSM baseline scores stay similar while perturbation loss falls.
- Qwen 4B fails answer-format instructions on MC tasks.
- Its negative MC drops cannot be read as clean reasoning improvements.

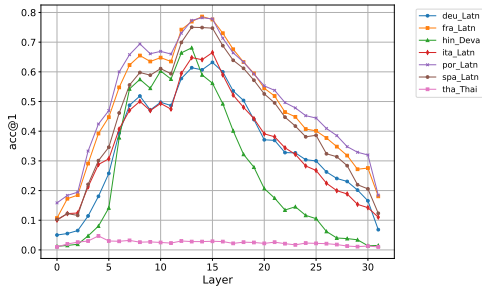
Paper, Sec. 5.3 and Table 2, PDF pp. 8–9.

Tokenization also changes cross-lingual representations

Canonical inputs



Character inputs



- FLORES-200 sentence retrieval, cosine similarity, top-1 accuracy.
- x-axis: layer; y-axis: acc@1. Axis maxima differ: 1.0 vs. 0.8.
- A last-token representation probe shows changes beyond answer formatting.

Original Fig. 5(a,c), Llama, PDF p. 8; method in Sec. 5.4, PDF pp. 8–9. Both sides use the stated tokenization.

Train on diverse tokenizations of the same text

Scheme	Composition of ten copies
Canonical	10 copies of the same canonical segmentation
Mixed	1 canonical + 1 character + 8 random alternatives
LR-Bucketed	1 canonical + 1 character + 1 alternative from each of 8 length-ratio buckets

$$[1, 1.25),\ [1.25, 1.5),\ \ldots,\ [2.75, 3.0]$$

- Repeat the ten-copy pattern for $N = 20$ or 30.
- Canonical($N\times$) controls for repeated exposure.
- Bucketing spreads the alternative inputs across granularity levels.

Paper, Sec. 6.1, PDF pp. 9–10. One-copy canonical, random, and character schemes are also evaluated.

English-only adaptation is evaluated across languages

Multilingual-ARC training

1116 English training examples.
298 original validation examples.
894 validation inputs: three forms per example.

- LoRA rank 16; alpha 32; dropout .1.
- Targets: query, key, value projections.
- Five epochs; maximum length 2048.
- Learning rate 2.5×10^{-6}.

Table 3 matches Qwen3-8B's ARC baseline numerically, but the adapted model is not explicitly named there. Other tasks are not used to test augmentation transfer.

Paper, Secs. 6.2–7, PDF pp. 10–11; Table 7, PDF p. 14. Model identification is inferred from Tables 1 and 3.

Bucketed training beats repeated canonical data

Training scheme	Canonical	Random	Character
No fine-tuning	78.3	71.8	66.0
Canonical 1×	79.2	72.3	66.3
Canonical 30×	81.2	73.7	67.9
Mixed 30×	81.2	74.0	68.2
LR-Bucketed 30×	81.8	74.8	69.2

- Values: average multilingual ARC accuracy (%).
- Versus Canonical 30×: +0.6 / +1.1 / +1.3 percentage points.
- Repeating canonical data already helps; diversity adds a further gain.

Paper, Table 3, PDF p. 10. Equal copy counts do not imply equal processed-token or compute budgets.

Higher perturbed accuracy need not close the relative gap

Training scheme	Random drop (%) ↓	Character drop (%) ↓
No fine-tuning	8.542	15.963
Canonical 30×	9.405	16.490
Mixed 30×	8.987	16.141
LR-Bucketed 30×	8.637	15.548

Interpretation

Bucketed training improves absolute accuracy and preserves robustness better than repeated canonical training. Its random relative gap is still slightly above the unadapted baseline.

Paper, Table 3, PDF p. 10. Relative-drop columns are averages across languages.

The evidence distinguishes three effects

1. **Segmentation**: equal decoded text still gives different task behavior.
2. **Length**: near-equal-length perturbations retain gaps in several languages.
3. **Representation**: sentence retrieval changes without generating an answer.

Limits that affect interpretation

Language coverage and perturbation strength differ; training gains are tested only on ARC. Vocabulary allocation, resource level, and training exposure are not independently manipulated.

Paper, Secs. 5–7 and Limitations, PDF pp. 5–11; Tables 4–6, PDF pp. 13–14.

Evaluate the tokenizer interface across languages

- Robustness observed in English is not a default multilingual property.
- Report the language, perturbation type, and feasible tokenization range.
- Use repeated canonical training as the augmentation baseline.
- Track perturbed accuracy and the relative canonical gap separately.

Discussion question

Can diverse valid inputs improve robustness at a matched token/compute budget, and can the gains transfer beyond multilingual ARC?

Takeaways based on Tables 1–3 and the paper's stated limitations, PDF pp. 6, 9–11.

References and paper status

- Poulami Ghosh and Preethi Jyothi. *Language Models are not Equally Robust to Non-Canonical Tokenization across Languages*.
- arXiv:2607.26831v1, submitted 29 July 2026.
- Versioned paper PDF and author TeX source.

Publication not found as of 2026-11-13.
Public code not found as of 2026-11-13.

Companion analysis: `reading-notes.zh.md`.
Original figure assets and a local pdfLaTeX build helper accompany the deck.

Publication/code searches were refreshed for this reading. This presentation discusses the arXiv v1 paper.

Backup: small tokenization spaces limit perturbations

Language	Log_{10} tokenization count Qwen	Gemma	ARC: Canonical → Random Qwen	Gemma
Bengali	3.53	64.45	57.70→52.64	65.61→53.80
Gujarati	1.12	59.73	54.30→52.64	60.78→56.62
Telugu	17.07	56.19	42.70→41.00	44.82→40.32

- Qwen's canonical tokenization is already near character level for these languages.
- Armenian often has only one valid Qwen tokenization and is omitted.
- Low sensitivity can reflect little room to change the segmentation.

Paper, App. C, Tables 4–5, PDF pp. 13–14. Counts are on a log scale; no exponentiation of an averaged log is used.

Backup: average relative loss has its own denominator

For L languages, let $a_{c,\ell}$ and $a_{r,\ell}$ be canonical and random scores.

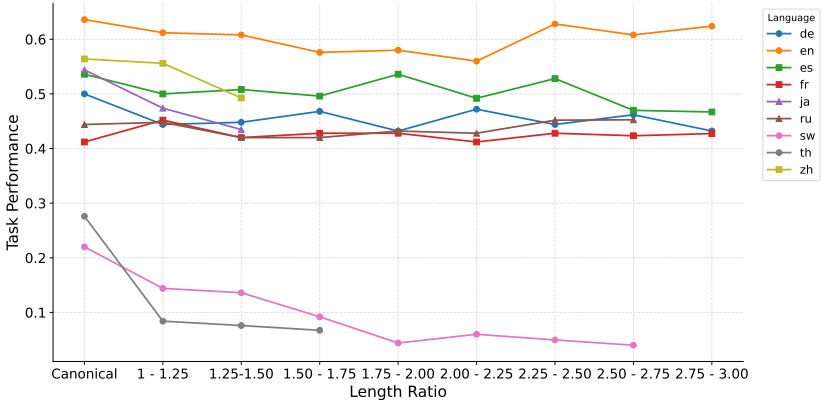
$$\Delta_{r,\ell} = 100\frac{a_{c,\ell} - a_{r,\ell}}{a_{c,\ell}}$$

$$\frac{1}{L}\sum_{\ell} \Delta_{r,\ell} \neq 100\frac{\bar{a}_c - \bar{a}_r}{\bar{a}_c}$$

- Table 3 baseline: mean accuracy 78.3 canonical, 71.8 random.
- Their ratio gives about 8.30%; the reported mean relative loss is 8.542%.
- Do not reconstruct exact random scores from averaged canonical scores and averaged relative losses.

Teaching explanation of aggregation; paper, Tables 1 and 3, PDF pp. 6, 10.

Backup: MGSM sensitivity also varies by language



Qwen: English remains comparatively stable, while Swahili and Thai decline strongly. Missing high-ratio points are not zero scores.

Original Fig. 8(b), PDF p. 18. x-axis: sampled/canonical length ratio; y-axis: task performance.

Backup: control exposure before testing diversity

Training scheme	Canonical	Random	Character
Canonical 10×	81.1	73.6	67.7
Mixed 10×	81.1	73.7	67.9
LR-Bucketed 10×	81.1	73.8	68.0
Mixed 20×	81.5	74.6	68.7
Mixed 30×	81.2	74.0	68.2
LR-Bucketed 30×	81.8	74.8	69.2

- Canonical copies increase exposure without adding segmentation diversity.
- Mixed 20× →30× is not monotonic; more copies alone do not guarantee a gain.
- Five epochs over larger datasets also means more optimization steps.

Paper, Table 3, PDF p. 10; interpretation distinguishes exposure from tokenization diversity.