Eleventh Conference on Machine Translation (WMT26)

EMNLP2026 BUDAPEST

IndicQE-APE: A Consolidated Benchmark for Quality Estimation and Automatic Post-Editing over Indic Languages

Diptesh কনৌজিয়া[1] Archchana சிந்துஜன்[1] Sourabh देवघरे[2] Daria Sokova[1] Shenbin Qian[3] Girish A. కౌశిక్ [1]
Tharindu රණසිංහ[4], Constantin Orăsan[1], Chrysoula Zerva[5], Ricardo Rei[6,7], Frederic Blain[8], André F. T. Martins[5]
Marco Turchi[9], Matteo Negri[10], Anoop കുഞ്ഞുകുട്ടൻ[11,12], Mitesh M. ખાપરા[11,12], Pushpak ভট্টাচার্য্য[2]

[1]University of Surrey [2]IIT Bombay [3]University of Oslo [4]Lancaster University
[5]Instituto de Telecomunicações & Instituto Superior Técnico, University of Lisbon [6]Sword Health [7]INESC-ID
[8]Tilburg University [9]Zoom Communications [10]Fondazione Bruno Kessler [11]Bodhan AI [12]AI4Bharat, IIT Madras

One segment, linked evidence

English → Malayalam / source: `alope-rl`

EN 15 off the over

MT 15 റൺസിന് പുറത്തായി.

Derived tags: BAD / BAD / BAD / BAD / OK (EOS)

source + MT

Human QE
60.7/100
3 raters: 65, 60, 57

APE target
ഒരോവറിൽ15 റൺസ്
Human post-edit

Human remark: "he was dismissed for 15 runs" (excerpt).

Domain: sports (annotated; confidence 1.00). TER: 1.50; challenge: A2.

Tags: derived. MQM and synthetic explanation: absent on this row.

HF `full/test: 08495a258f48517c`

One aligned benchmark for Indic QE, post-editing and error analysis.

- Quality estimation (QE) scores machine translation (MT); automatic post-editing (APE) repairs it.
- We align human scores, post-edits, word tags and explanations on the same segments.
- Announcement: Inspired us to build *Mātrā - Novel MT evaluation metric for Indic languages, based on the akṣara (orthographic syllable) unit [soon on arXiv].*

Domain prediction: a simple model wins

80.4% TF-IDF + logistic regression
56.6–61.8% GPT-4o-mini / GPT-5.4-nano
Accuracy on 500 gold items; three GPT settings.
Gain mainly in provenance classes. [App. B]

01 The Benchmark, and how we built it

1 Collate — One corpus
Released + previously unreleased instances
WMT + English–Malayalam

2 Deduplicate — 126,754
instances in `full`
Content-hash merging
Corpus provenance retained

3 Task views — QE + APE
Sentence and word QE
Automatic post-editing
Explainable QE and pinpointed APE

4 Challenge Set — 13,032
curated test segments
Four difficulty axes
Source-disjoint splits

5 Publish — Nine directions
Task and language subsets
Dataset, code, predictions
Reproducible baselines

Supporting multiple QE and APE tasks.

Task subsets overlap; `qe-da` covers English→Indic. Additional MQM view: `mqm`, 4,490 English–Hindi instances.

Sentence-level QE	Automatic post-editing	Word-level QE	Explainable QE
85,586	73,829	79,301	19,644
`qe-da`: English→Indic	`ape`: seven language pairs	`word-qe`: OK/BAD tags	`explainable-qe`

English → Hindi, Marathi, Gujarati, Tamil, Telugu, Malayalam; Estonian, Nepali, Sinhala → English. Estonian is a non-Indic control. [Tables 1–3]

02 Word tags validated at scale

40,685 additional tagged segments. Re-derive released upstream tags from MT and human post-edits to audit.

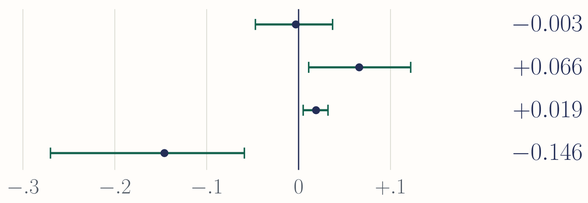
0.963	633,988	0.977
Matthews correlation	validated tokens	BAD-tag F1

MCC by target: Malayalam 0.898; Marathi 0.929; English 0.969–0.973.

Indic morphology matters. Across seven pairs, subword fertility and tag agreement correlate at $\rho = -0.927$ ($p = 0.003$). IndicNLP reproduces shipped tokenisation at 96.0%, versus 66.0% for Moses. Validation audits label construction. [§2; App. B]

03 Separate difficulty from score range

Controls match **language and human-score distribution**. A1 disappears; heavy editing and dense word errors are not intrinsically harder.



A4: negative for 9/9 systems and 7/7 pairs; still -0.071 without Malayalam. Barely-edited arm: -0.070 (1,149 segments; four pairs). Dots: matched $\Delta\rho$; whiskers: 95% pair-bootstrap CIs. [Tables 6, 11; Apps. C–D]

04 QE pretraining drives the gain

Change only the encoder: **reference-free QE pretraining** shows the largest positive substitution on five-pairs.



Encoder gain: $+0.140$ (Williams $t = 20.5$, $p < 0.001$; $n = 8,160$). Removing multi-layer pooling collapses at all three seeds ($\rho = 0.096$–0.097). Min–max targets add $+0.040$ (0.585→0.625). Within-build contrasts only. [§5; Table 8; App. K]

05 Read quality from hidden states

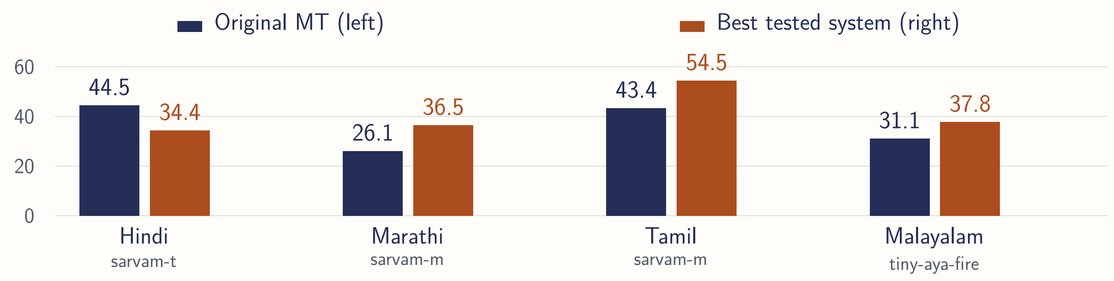
Same tiny-aya-fire backbone: a trained read-out beats prompting on every English→Indic pair. (Macro ρ)

0.202	0.527	0.578
3.4B prompted	3.4B layer probe	GPT-5.5 prompted

Within 0.051 of GPT-5.5, using a frozen 3.4B model and one layer (-20) across six pairs. Length-only $\rho = 0.218$. Telugu: probe 0.282 versus GPT 0.255. The six-pair layer sweep is exploratory on the evaluation split. [§4; Tables 24–25]

06 APE gains depend on the reference

6,518 segments; four pairs. Character edit rate against human post-edits (↓); zero-shot systems, identical rows.



Original MT wins 3/4 pairs. Human edits leave MT unchanged in 0.5% (hi), 12.2% (mr), 29.1% (ta), 30.8% (ml). Post-edit references reward conservative repair; sarvam-t retranslates. [Table 9; App. N]

07 Audit prompting, not just scores

Four-shot Llama: ρ **0.217→0.027**. Jointly scored segments retain the loss, beyond failures to return a usable score.

Model	Paired $\Delta\rho$	Parse
BharatGPT-3B	-0.098	44.7%
Llama-3.2-3B	-0.189	50.1%
tiny-aya-fire	-0.050	86.1%
aya-expanse-32B	$+0.055$	98.7%

All three models ≤3.4B lose correlation and compliance. Paired checks change the losses by ≤0.001. Four-shot eligible pairs: 4/4/7/9; parse is macro over run pairs. Fixed DA template; encoder controls stay unchanged. [Table 7; App. M]

हिन्दी Hindi · मराठी Marathi · ગુજરાતી Gujarati · தமிழ் Tamil · తెలుగు Telugu · മലയാളം Malayalam

Segment-level evaluation; one MT hypothesis per source does not support system-level ranking.

We acknowledge the European Association for Machine Translation for project funding, alongside People-Centred AI & CFILT for their initial financial support.

People-Centred AI UNIVERSITY OF SURREY · CENTRE FOR TRANSLATION STUDIES UNIVERSITY OF SURREY · CFILT भारतीय भाषा प्रौद्योगिकी केन्द्र · UNIVERSITY OF SURREY

Paper · Dataset · Code

d.kanojia@surrey.ac.uk / surrey-nlp/IndicQE-APE