Team Horizon at BHASHA Task 1: Multilingual IndicGEC

How Linguistically-Guided Synthetic Data Bridges the Low-Resource Gap for Grammatical Error Correction

Manav Dhamecha, Gaurav Damor,
Sunil Choudhary, Pruthwik Mishra

LANGUAGES: Bangla, Hindi, Tamil, Telugu, Malayalam

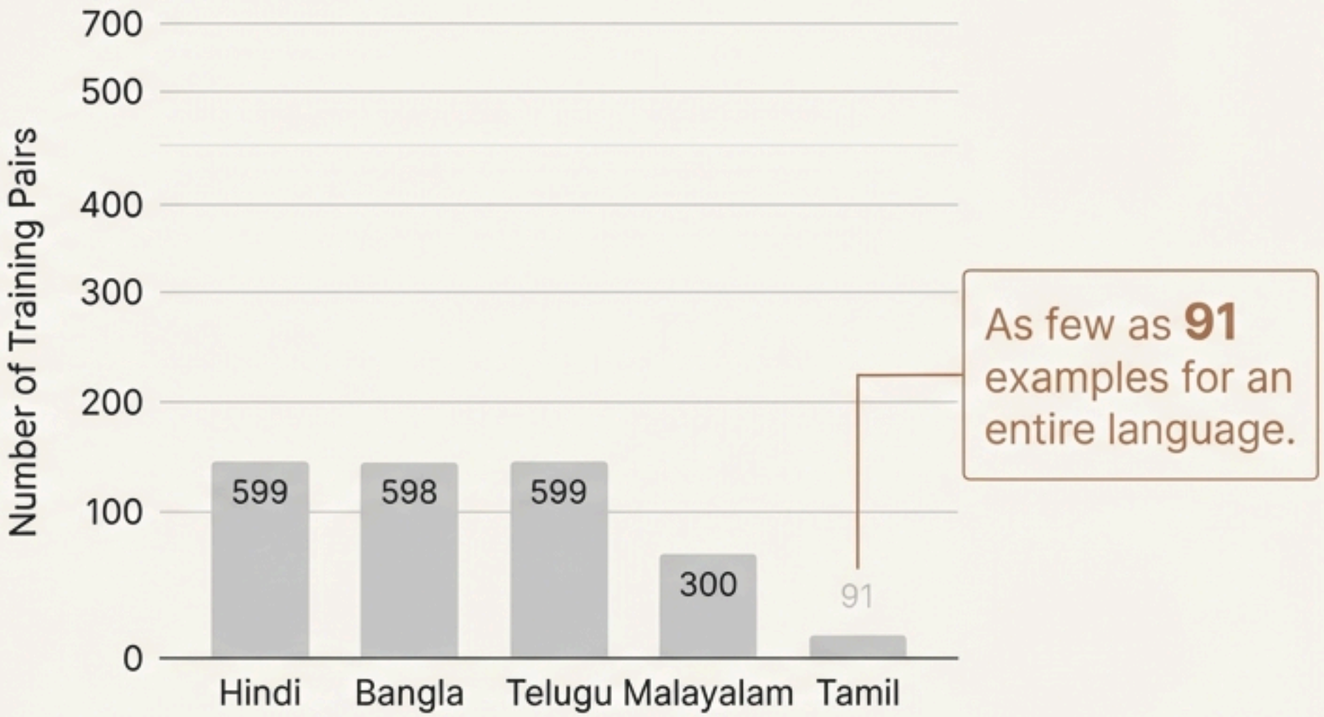
MODELS: mT5-small, IndicBART

CORE CONTRIBUTION: A synthetic data augmentation pipeline that introduces realistic linguistic errors to scale training data.

Grammatical Error Correction for Indic Languages is Crippled Crippled by a Severe Lack of Annotated Data

Indic languages present unique challenges for GEC due to high morphological complexity, rich inflectional patterns, and free word order. However, the primary bottleneck is the extreme scarcity of high-quality, annotated training data, which makes traditional supervised learning insufficient for robust performance.

Official Training Data is Extremely Limited



"Linguistically informed synthetic error-injection erron framework for Indic GEC data augmentation."

Our Solution: A Linguistically-Grounded Pipeline to Synthetically Augment Training Data by Over 10x

To overcome data scarcity, we developed a synthetic data augmentation pipeline. By programmatically injecting realistic grammatical errors into clean sentences from corpora like IndicCorp v2, we expanded the training set from under 1,000 to over 10,000 high-quality pairs per language.

"Clean Sentences (from IndicCorp v2, Wikipedia)"

"Controlled Error Injection Framework (42 linguistic rules)

"10k+ Synthetic 'Incorrect/Correct' Pairs"

Editorial Academia

The Pipeline Injects Controlled Errors Across 10 Linguistic Categories to Simulate Natural Mistakes

Spelling

Random मत्र swaps (ा↔ि) & visually similar consonant substitution (e.g., त → थ).

Word Agreement

Morphological inflection mutations for gender (है → थी), number (हैं → है), and case (ने → को).

Structure

Random omission, duplication, or insertion of postpositions (ने, को, में) and adverbs.

Example: Person Agreement Error (Hindi)

Incorrect: 'मैं स्कूल जाती है।' (maiṁ skūl jātī hai)

Correct: 'मैं स्कूल **जाता हूँ**।' (maiṁ skūl jātā hūṁ)

(I go to school.)

We Fine-Tuned Two Lightweight Multilingual Models to Establish Reproducible Baselines

Models

mT5-small

300M parameters. A general-purpose, massively multilingual text-to-text transformer pre-trained on the mC4 corpus.

IndicBART

A sequence-to-sequence model pre-trained specifically on 11 Indic languages, designed to better capture their linguistic nuances.

Training Configuration

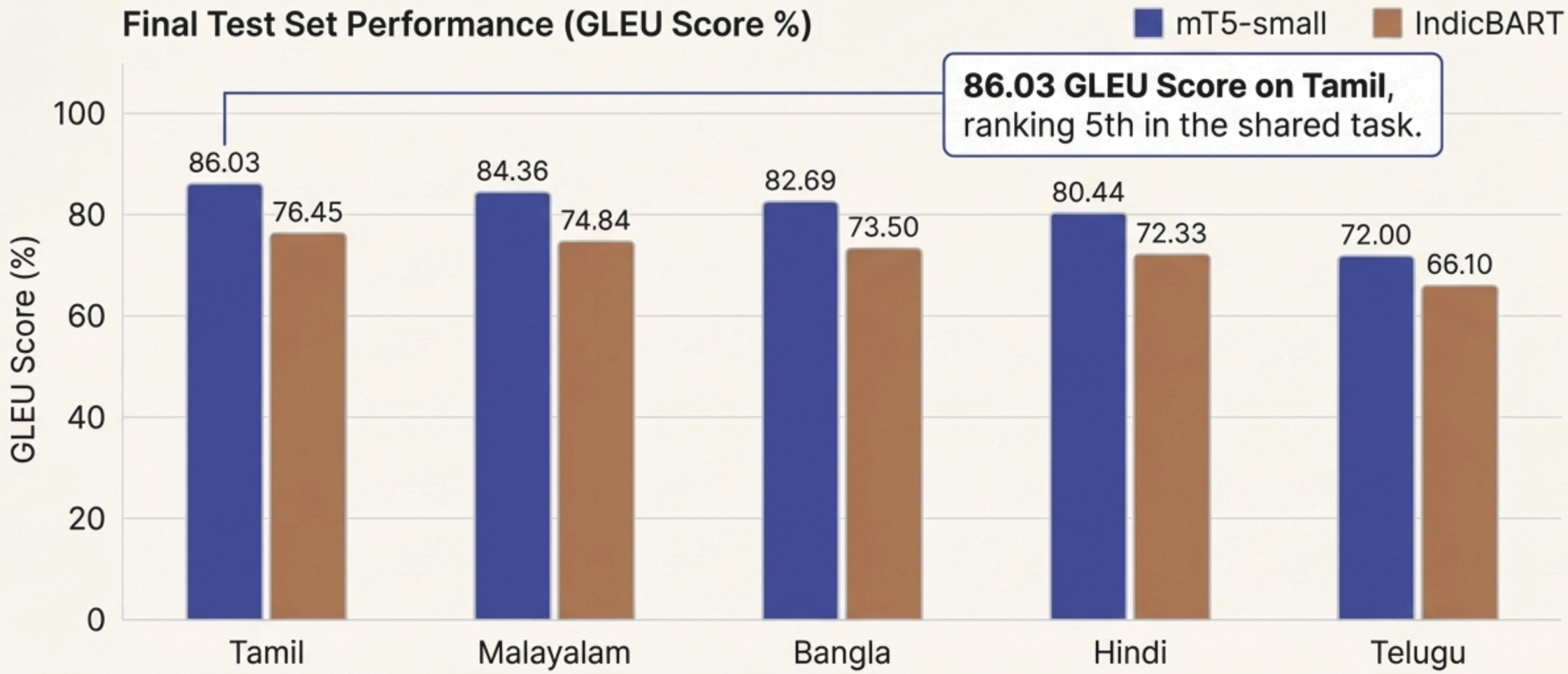
Input Format:
`correct this: <incorrect sentence>`

Language Tags:
`[HI], [BN], [TA], [TE], [ML] prepended for multilingual training.

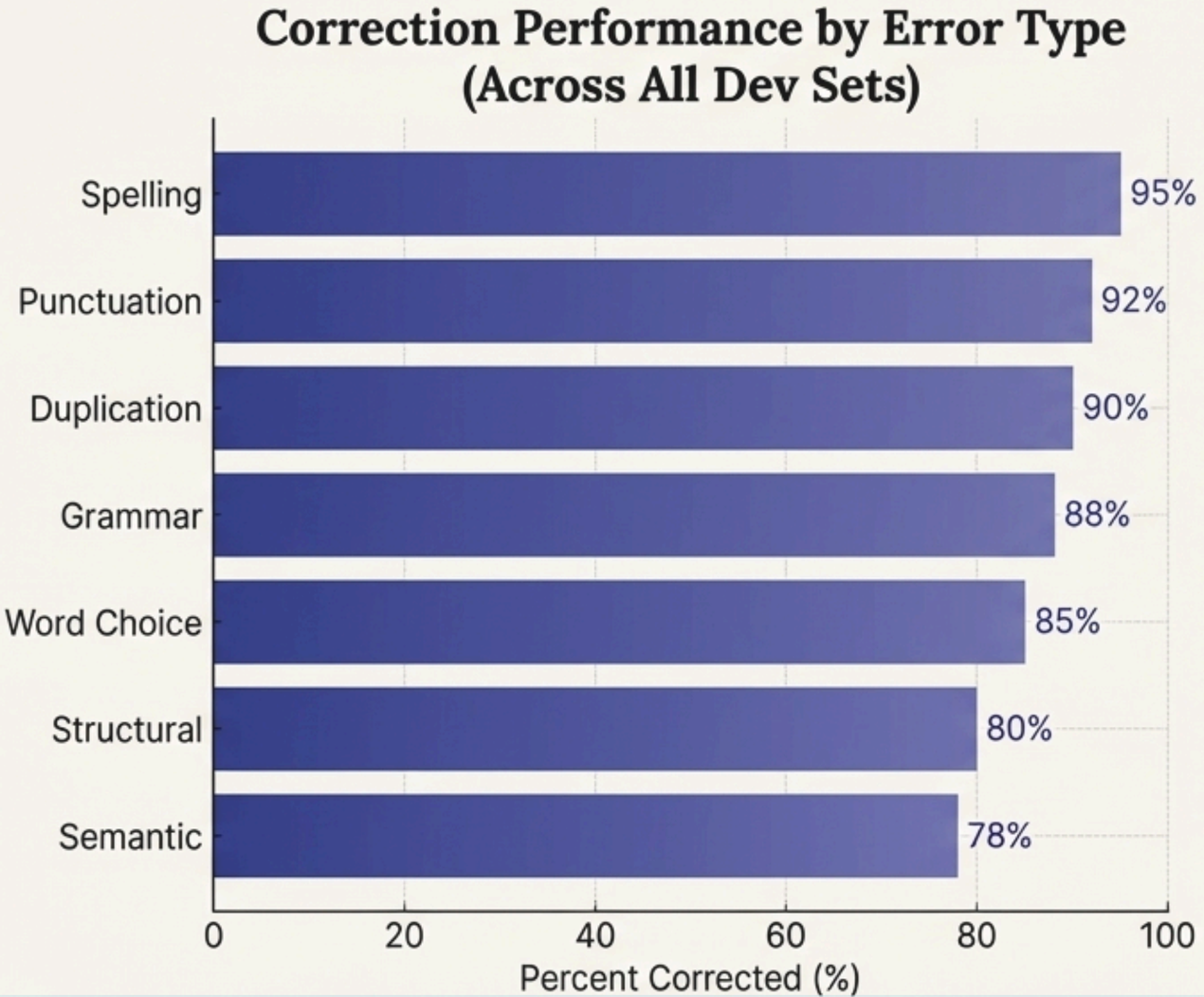
Key Hyperparameters

Optimizer: AdamW
Learning Rate: 5e-5 (mT5) / 3e-5 (IndicBART)
Batch Size: 16-32
Epochs: 10-15 (with early stopping on dev GLEU)

Our Data-Augmented Approach Achieved Competitive Performance, with mT5-small Leading Across All Languages



Error Analysis Reveals High Efficacy on Spelling and Grammar, while Semantic Errors Remain a Challenge



Key Insights

- **"High Performance"**: Models effectively corrected surface-level errors like spelling and punctuation (>90%).
- **"Lingering Challenge"**: Deeper, meaning-based errors (Semantic, Structural) proved more difficult.
- **"Language-Specific Note"**: Morphological and word-order errors were particularly challenging for Dravidian languages (Tamil, Telugu, Malayalam).

Linguistically-Informed Data Augmentation is a Powerful and Scalable Strategy for Low-Resource Indic GEC

Introduced a novel, linguistically-informed framework for synthetic error injection in Indic languages.

Demonstrated its effectiveness in scaling limited annotated data by over 10x, significantly improving model performance.

Established strong, reproducible baselines using lightweight, publicly available multilingual models (mT5-small and IndicBART).

Proved that this approach effectively helps bridge the performance gap caused by data scarcity in morphologically rich, low-resource languages.

Acknowledging Limitations and Charting the Path Forward

Limitations

- **Ecological Validity**: Synthetic errors may not fully capture the diversity of real-world mistakes made by human learners.
- **Model Scope**: Evaluation was limited to
Model Scope: Evaluation was limited to
two multilingual models, excluding stronger language-specific alternatives.
- **Evaluation Metric**: Relied solely on the automatic GLEU metric without human assessment of fluency or meaning preservation.

Future Directions

- Incorporate **real learner corpora** to improve error distribution.
- **Evaluate larger**, more powerful language-specific models (e.g., BanglaT5).
- Perform **human evaluations** to assess practical usability.
- Investigate **advanced cross-lingual** and few-shot learning approaches for ultra-low-resource settings.