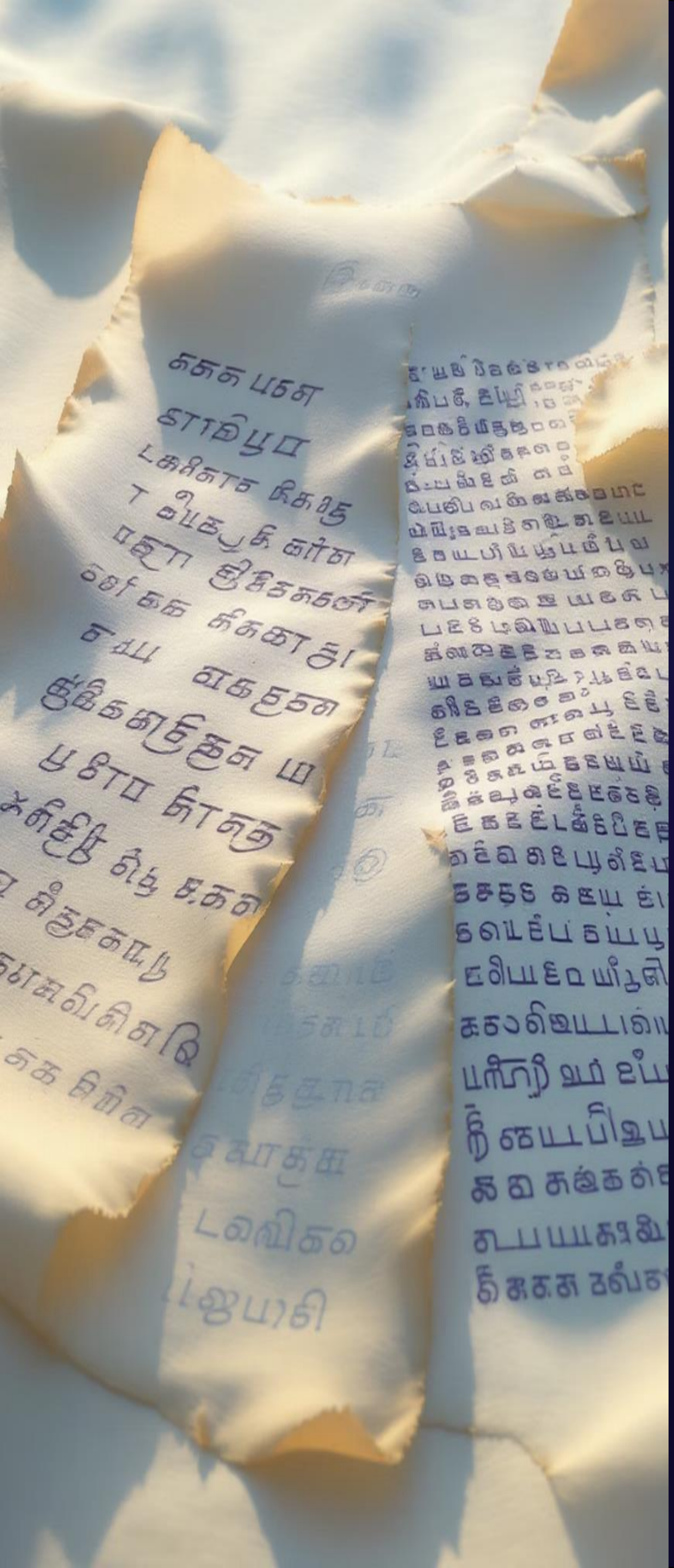
A Hybrid Neurosymbolic Approach for Tamil and Malayalam Grammatical Error Correction

DLRG Team

Akshay Ramesh, Ratnavel Rajalakshmi



Grammatical Error Correction for Low-Resource Indic Languages

Grammatical Error Correction (GEC) aims to automatically detect and correct errors in written text. While significant progress has been made for high-resource languages, GEC for Indic languages faces severe challenges, notably data scarcity and morphological complexity.

Language Context

- **Tamil:** Dravidian language with 75+ million speakers.
- **Malayalam:** Dravidian language with 38+ million speakers.

Key Challenges

- **Extreme Data Scarcity:** IndicGEC provides only 91 training pairs for Tamil, compared to millions for English.
- **Morphological Complexity:** Both languages exhibit agglutinative morphology with rich inflectional systems and complex verb conjugations.
- **Script Complexity:** Unique Unicode challenges, including chillu character variations in Malayalam.

Why Neurosymbolic? The Rationale

Our hybrid neurosymbolic architecture leverages complementary strengths to overcome the limitations of pure neural or rule-based approaches in low-resource settings.

Pure Neural Models

Require millions of training examples, leading to severe overfitting with limited data (e.g., 91 examples for Tamil). Exhibit unpredictable generation behaviours and lack deterministic guarantees.

Pure Rule-Based Systems

Provide perfect accuracy on explicitly encoded patterns but lack generalisation to unseen error types. Cannot correct novel errors not captured in manual rules.

The Neurosymbolic Solution

Combines the generalisation of neural models with the precision of symbolic rules, enhanced by augmented data and intelligent ensemble selection.

System Architectures

We developed language-specific architectures reflecting unique characteristics and dataset constraints.

Tamil: Five-Stage Hierarchical Pipeline

Prioritises correction coverage for complex morphology:

Marker Extraction
Regex-based isolation of formatting elements.

Rule-Based Priority
Exact matching against sentence templates.

Neural Generation
mT5-base with LoRA adaptation, beam search.

Pattern Enhancement
25+ manual Tamil error patterns applied.

Marker Reattachment
Deterministic restoration of original formatting.

Malayalam: Parallel Processing with Safety-First Ensemble

Prioritises output reliability and stability:

Exact Match Check
Validation against learned corrections dictionary.

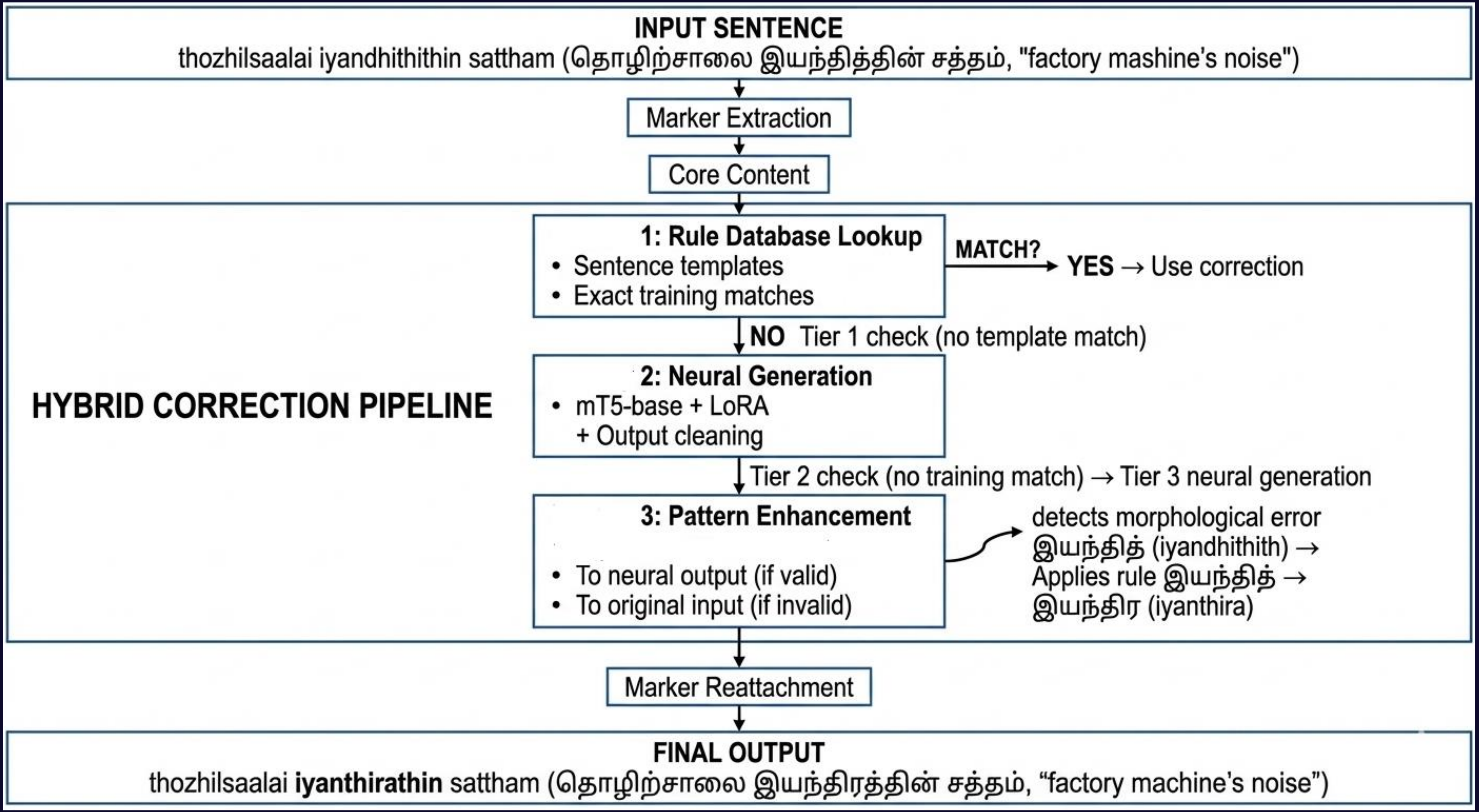
Parallel Processing
Simultaneous neural and rule-based candidate preparation.

Safety Validation
Comprehensive checks on neural outputs (e.g., character presence, length ratios).

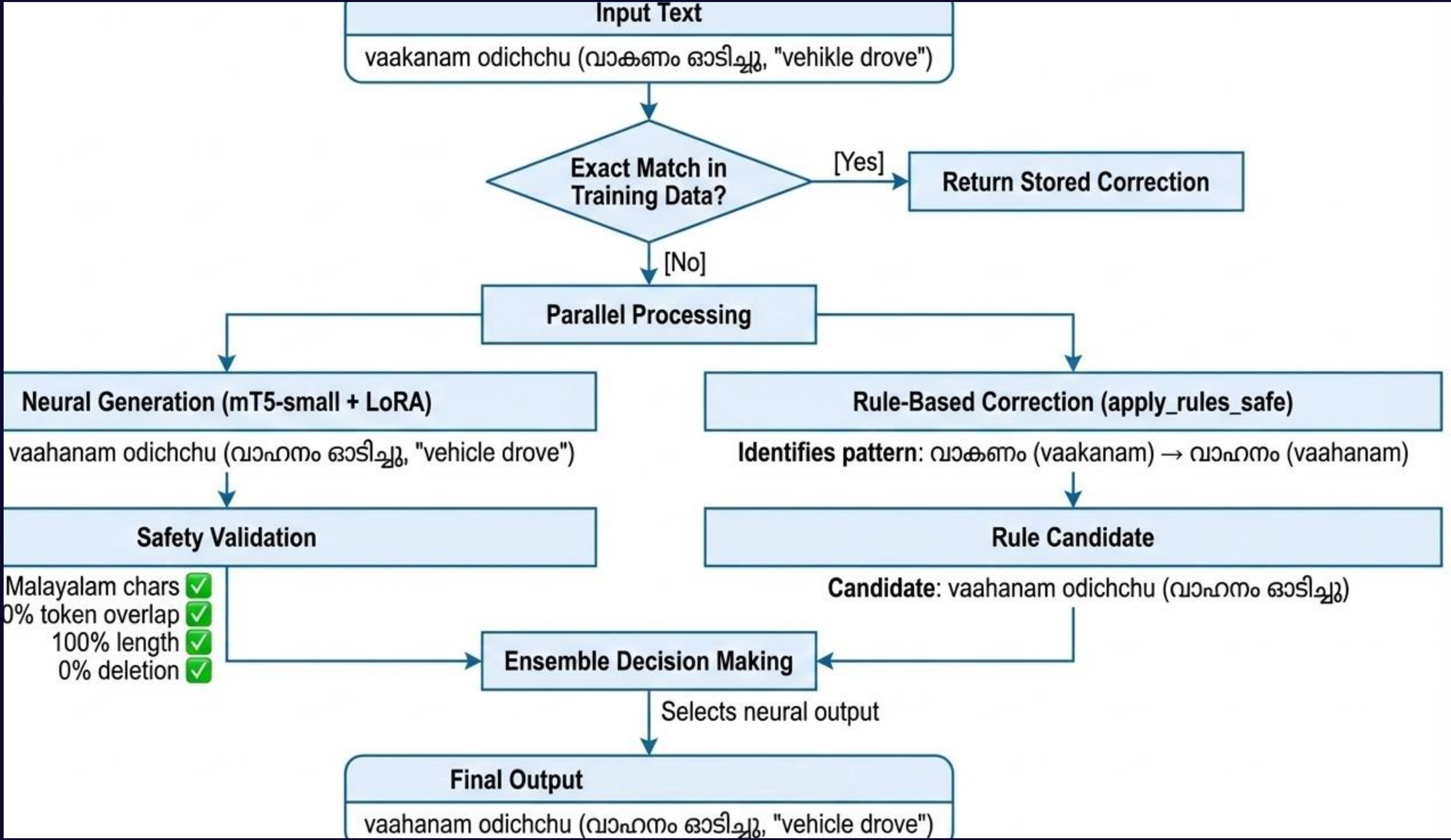
Ensemble Selection
Confidence-based choice or fallback to original input.

Explicit Tracking
Usage statistics for transparency.

Grammatical Error Correction for Tamil



Grammatical Error Correction for Malayalam



Data Augmentation and Training

Language-specific data augmentation strategies were crucial for mitigating data scarcity.

Tamil Augmentation (91 → 5,000 examples)

- **Vowel Dropping:** Targeting 12 Tamil vowels.
- **Character Perturbations:** Duplication and deletion.
- **Structural Changes:** Punctuation perturbation and word order shuffling.
- **Transformation:** Each sentence underwent 1-2 random transformations (55-fold expansion).

Malayalam Augmentation (→ 10,000 examples)

- **Vowel Sign Dropping:** Targeting 12 Malayalam vowel signs.
- **Safe Perturbations:** Avoiding catastrophic truncation.
- **Structural Changes:** Adjacent word swapping, comma spacing removal.
- **Chillu Variation Handling:** Modern-traditional pairs.
- **Quality Filtering:** Similarity filtering (0.6-0.98) and length preservation (≥50%).

Controlled noise injection mimics natural error patterns while maintaining linguistic validity. Quality filtering prevents learning spurious noise patterns.

Training Configuration

Utilised AdamW under FP16 precision, a learning rate of 3e-4, effective batch size of 8, and 10 epochs with early stopping, implemented using Hugging Face Transformers.

Experimental Results

Our hybrid approach demonstrated strong performance in the IndicGEC Shared Task blind evaluation.

Dataset and Evaluation Setup

- **Tamil:** 91 training pairs (augmented to 5,000), 65 test inputs.
- **Malayalam:** Augmented to 10,000 examples, 102 test inputs.

Performance on Test Set

Language	GLEU	Overall Rank
Tamil	85.34	8
Malayalam	95.06	2

Baseline Comparisons: Both hybrid models significantly outperformed individual baselines (e.g., Tamil hybrid 80.47% vs. neural-only 36.21%).

Representative Corrections

- **Tamil Examples:** Corrected morphological errors like iyandhithithin (இயந்தித்தின்) → iyanthirathin (இயந்திரத்தின்) ("machine's"), multi-token errors, and vowel length normalisation.
- **Malayalam Examples:** Corrected spelling (e.g., vaakanam (വാകണം) → vaahanam (വാഹനം)), with conservative preservation of input when no correction was needed.

Comparative Analysis: Our hybrid approach (85.34% Tamil, 95.06% Malayalam) significantly surpasses Czech GEC (approx. 60-70% accuracy) in similar low-resource scenarios.

Neural Component and Ablation Study

Our model capacity selection was empirically validated through ablation experiments.

Neural Architecture Configuration

- **Tamil GEC:** mT5-base (580M parameters), LoRA (Rank 16, Alpha 32), 55-fold augmentation.
- **Malayalam GEC:** mT5-small (300M parameters), LoRA (Rank 8, Alpha 16), 10,000 examples augmentation.

Ablation Study: Model Capacity Analysis

Language	Configuration	GLEU	Delta
Tamil	mT5-base (proposed)	80.47%	Baseline
Tamil	mT5-small	75.17%	-5.30%
Malayalam	mT5-small (proposed)	55.21%	Baseline
Malayalam	mT5-base	55.03%	-0.18%

1

Tamil Requires Higher Capacity

Morphological complexity necessitates higher representational capacity, shown by a 5.30% GLEU degradation with a smaller model.

2

Malayalam Benefits from Conservative Selection

Negligible performance difference with increased capacity (0.18%), validating a lower capacity with strict safety validation for optimal balance.

3

Non-Monotonic Relationship

In extremely low-resource settings, model size and performance is language-dependent and non-monotonic, requiring empirical validation.

Error Analysis: GEC for Tamil and Malayalam

Input Sentence	**Hybrid Output**	**Correction Type**
-தொழிற்சாலை இயந்தித்தின் சத்தம் *thozhilsaalai iyandhithithin sattham* / "factory mashine's noise"	தொழிற்சாலை இயந்திரத்தின் சத்தம் *thozhilsaalai iyanthirathin sattham* / "factory machine's noise"	Morphological இயந்தித் → இயந்திர *iyandhithith → iyanthira*
-போக்குவரத்து வாகணங்களின் ஹாரன் *-pokku varatthu vaakanangalin haaran* / "traffic vehikles' hron"	போக்குவரத்து வாகனங்களின் ஹார்ன் *pokku varatthu vaahanangalin haarn* / "traffic vehicles' horn"	Multi-token . ஹார்ன் -- ஹாரன் ; வாகணம் → வாகனம், *haaran → haarn, vaakanam → vaahanam*
இரயில் பயனத்தில் களைத்துப் போன எங்களுக்கு irayil payanattil kaḷaittup pōna eṅkaḷukku / "train journey in tired gone for us"	ரயில் பயணத்தில் களைத்து போன எங்களுக்கு rayil payaṇattil kaḷaittu pōna eṅkaḷukku / "train journey in tired gone for us"	Multiple Errors இரயில் → ரயில்,பயனம் → பயணம், irayil → rayil, payanam → payaṇam
വാകണം ഓടിച്ചു *vaakanam odichchu* / "vehikle drove"	വാഹനം ഓടിച്ചു *vaahanam odichchu* / "vehicle drove"	Spelling correction വാകണം → വാഹനം / *vaakanam → vaahanam*
ധ്വനി മലിനീകരണത്തിന് കാരണങ്ങൾ *dhvani malineekaranaththinu kāranaṅṅaḷ* / "noise pollution's reasons"	ധ്വനി മലിനീകരണത്തിന് കാരണങ്ങൾ *dhvani malineekaranaththinu kāranaṅṅaḷ* / "noise pollution's reasons"	Token-level preservation

Key Contributions and Insights

Our research provides valuable contributions to low-resource GEC, offering a blueprint for future development.

Novel Hybrid Architecture

Combining neural and symbolic approaches effectively addresses extreme low-resource GEC challenges.

Language-Specific Design

Differentiated architectures for Tamil and Malayalam optimise for correction coverage vs. output reliability.

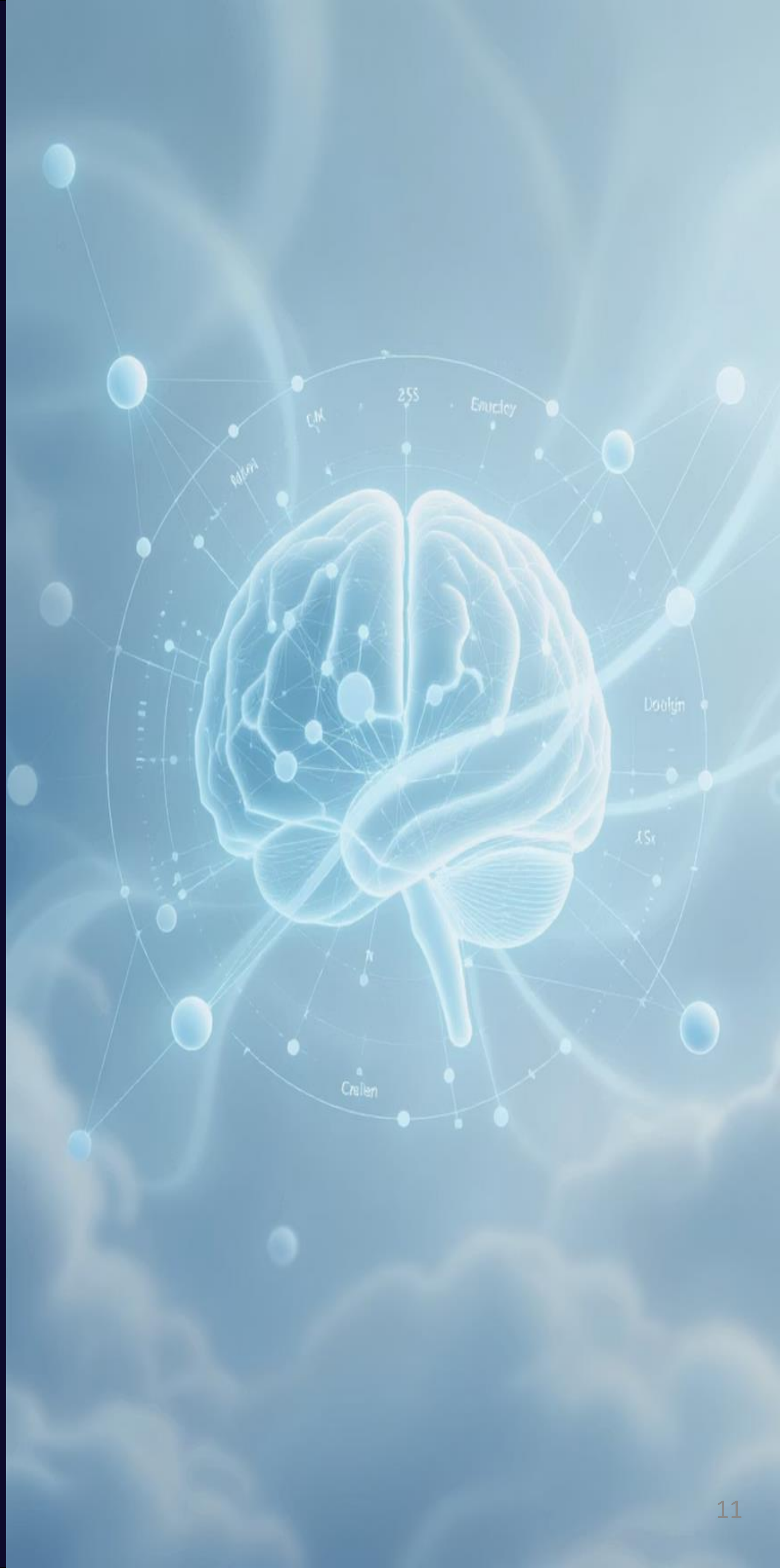
Morphology-Aware Augmentation

Developed synthetic augmentation strategies, achieving significant data expansion for both languages.

Conservative Safety Mechanisms

Multi-layered validation prevents catastrophic failures and over-corrections.

Broader Impact: This work offers a practical approach for developing GEC systems for other low-resource Indic languages, combining modern pre-trained models, parameter-efficient fine-tuning, aggressive augmentation, and linguistic rule engineering.



Limitations and Future Work

We identify current limitations and propose future research directions to further advance low-resource GEC.

Current Limitations

- **Statistical Confidence:** Small datasets limit generalisation confidence.
- **Pattern Coverage Gaps:** Manual patterns are not exhaustive for all error types.
- **Generation Stability:** Observed instability with mT5-base for Malayalam requires investigation.
- **Domain Specificity:** System assumptions may not generalise across text domains.
- **Architectural Limitations:** Ablation only with mT5 variants, other architectures unexplored.

Future Research Directions

- **Adaptive Safety Mechanisms:** Dynamic threshold adjustment based on input characteristics.
- **Cross-Lingual Transfer:** Knowledge transfer between related Dravidian languages.
- **Automated Pattern Discovery:** Explore grammar induction to reduce manual curation.
- **Comprehensive Human Evaluation:** Assess correction quality beyond automatic metrics.
- **Monolingual Model Development:** Address resource gaps through pruning or distillation.

Long-Term Vision: Establish principled guidelines for model selection, safety mechanism design, and architectural choices for low-resource morphologically rich languages.

Thank You