"When Data is Scarce, Prompt Smarter"...Approaches to Grammatical Error Correction in Low-Resource Settings

BHASHA Task-1 IndicGEC Demonstration

Somsubhra De[1], Harsh Kumar[1] and Arun Prakash A[1,2]

[1] IIT Madras, [2] AI4Bharat

IJCNLP-AACL
2025

BHASHA

Introduction

- Grammatical correctness essential for clear and effective communication ...
- In the NLP field, developing systems that can automatically detect and fix sentence level grammatical errors is a significant area of research
- GEC has seen great success for languages like English, applying these advancements to many other low-resource Indic languages has been challenging, why?

Given a parallel corpus $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, where $x^{(i)}$ is a noisy input sentence and $y^{(i)}$ is its grammatically corrected counterpart, the model is trained to minimize the negative log-likelihood (NLL) of the target sequence conditioned on the input:

$$\mathcal{L}_{\text{GEC}}(\theta) = -\sum_{i=1}^{N} \sum_{t=1}^{|y^{(i)}|} \log P_\theta \left(y_t^{(i)} \mid y_{<t}^{(i)}, x^{(i)} \right) \tag{1}$$

What are we dealing with?

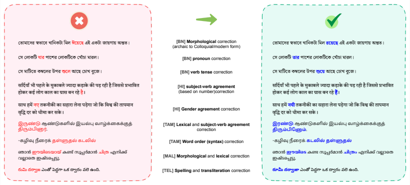
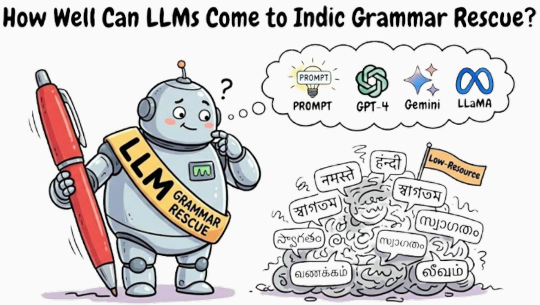


Figure 1: Examples from the GEC task dataset. Input sentence ✗ (with errors in red) - ground truth ✓ (with corrections in blue) pairs. Error types have been mentioned, based on our understanding.

Methodology



We utilize three large language models – GPT4.1 mini, Gemini-2.5-Flash and Llama-4-Maverick 17B-128EInstruct in *zero and few-shot* prompting paradigms. These models are employed as *instruction-following LLMs with role-based prompts*.

Additionally for Hindi, we fine-tuned (we adopt LoRA for PEFT) the Sarvam-M 24B2 multilingual model using Hi-GEC dataset.

*Image generated using Gemini NanoBanana

Prompt

"""You are a <Language> Grammatical Error Correction assistant, in low resource settings. Your task is to accurately identify and correct grammatical errors in the given <Language> sentence. Correct all types of grammatical errors:
Verb usage: Correct conjugation, tense, aspect, and agreement with the subject.,
Pronouns: Usage of proper personal, possessive, and reflexive pronouns.,
Prepositions: Correct use of postpositions or prepositions in context.,
Fix spelling mistakes, diacritic marks (matras), and punctuation errors.,
Gender and number agreement: Ensure adjectives, nouns, and verbs match in gender (masculine/feminine) and number (singular/plural).,
The output should be ONLY the CORRECTED sentence, without any extra text or explanation. *If the input is already correct, return it unchanged.* Please ensure the corrections follow the rules and preserve the intended meaning.

Below are 10 random sentences for your reference."""

Results

We ranked 1st in Tamil (GLEU: 91.57) and Hindi (GLEU: 85.69)
2nd in Telugu (GLEU: 85.22), 4th in Bangla (GLEU: 92.86) and 5th in Malayalam (GLEU: 92.97).

Model	**Tam**			**Mal**			**Hi**		
	GLEU	$F_{0.5}$	**BERT-score**	**GLEU**	$F_{0.5}$	**BERT-score**	**GLEU**	$F_{0.5}$	**BERT-score**
Gemini-2.5-Flash *(fs)*	91.57	87.82	97.83	**92.97**	88.48	97.89	84.61	88.01	95.69
GPT-4.1 mini *(fs)*	86.00	78.97	96.52	91.78	84.72	97.08	85.69	87.86	95.76
GPT-4.1 mini *(zs)*	85.51	78.45	96.38	92.34	84.62	97.24	85.37	87.80	95.53
LLaMA-4 maverick *(zs)*	88.70	81.50	96.84	92.68	85.38	97.20	83.10	86.04	94.64
LLaMA-4 maverick *(fs)*	85.62	77.75	95.98	90.75	83.22	96.65	85.37	87.35	95.56

Model	**Bn**			**Tel**		
	GLEU	$F_{0.5}$	**BERT-score**	**GLEU**	$F_{0.5}$	**BERT-score**
Gemini-2.5-Flash *(fs)*	92.23	89.61	97.10	84.16	76.96	94.92
GPT-4.1 mini *(fs)*	**92.86**	89.27	97.35	**85.22**	77.68	95.28
GPT-4.1 mini *(zs)*	91.62	86.98	96.79	84.74	76.75	95.15
LLaMA-4 maverick *(zs)*	90.39	86.48	96.30	83.01	74.20	94.28
LLaMA-4 maverick *(fs)*	92.00	88.02	97.19	82.02	74.28	94.08

Figure 2: Performance of different approaches on the test set across languages. Highlighted cells indicate the best-performing model for each language, while underlined values denote the overall best score in the task.

The finetuned Sarvam-M significantly failed to capture the correct edits, achieving only 13.81 GLEU in Hindi.

- Experiments revealed that even simple prompting strategies could lead to impressive results, even with limited data.
- LLMs when guided by well-designed prompts, substantially outperform fine-tuned Indic-language models like Sarvam-22B thereby illustrating the exceptional multilingual generalization capabilities of contemporary LLMs for GEC.
- Shows that instead of needing massive amounts of language-specific data to train models from scratch, we can leverage the existing knowledge within large, general-purpose LLMs.
- Where did the LLMs fail? We did some qualitative analysis and found out recurring patterns.
- A high Fertility Score has two direct consequences for GEC: increased latency and Context Window Reduction. To quantify this impact, we looked into the tokenizer fertility.

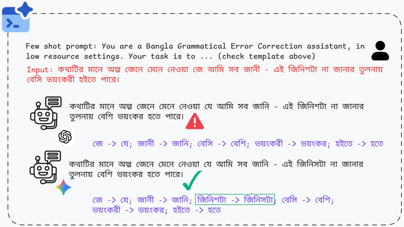


Figure 3: Comparison of model outputs on a **multi-correction** example (Transl. Knowing only a little about something and assuming that I know everything can be more dangerous than not knowing at all.) from test set. *Gemini's output fully aligns with the gold standard, while GPT omits one necessary edit.*

Language	Script Family	GPT-4.1 Mini	Gemini 2.5 Flash	Llama 4 Maverick
HI	Devanagari	1.44	**1.31**	1.55
BN	Eastern Nagari	2.32	**1.76**	2.77
TAM	Dravidian	3.09	**2.54**	5.88
TEL	Dravidian	2.97	**2.87**	4.32
MAL	Dravidian	3.20	**3.10**	4.58

Figure 4: Cross-Model Tokenizer Fertility Comparison

	INPUT SENTENCE	CHARS	WORDS	TOKENS		
HI	हम बस धीडे- धीडे चलते-चलते गलियों में धूम ही रहे थे कि हमने एक लड़की को बहुत सुंडर कपड़े बोचते हुए देखा।	104	22	35	38	39
MAL	ഇന്ന് സഹോദരന്മാർ തമ്മിൽ അടിക്കുന്ന ഈ കാലഘട്ടത്തിൽ നാം ലക്ഷ്മണനിൽ നിന്നും ഒരുപാട് കണ്ടുപിടിക്കേണ്ടതുണ്ട്.	104	11	34	35	48
TAM	திரும்ப ஹோட்டலுக்கு செள்ள டாக்சி கிடைக்காததால், அவர்கள் பெற்றோருடன் என் பெற்றோர்களை அனுப்பி வைத்தேன்.	104	11	28	33	68
BN	উত্তর হিমালয় ভারত পাহারা দিচ্ছে , এই জেনে নিশ্চিত হয়ে প্রায় তিন দিকের উপকূল রয়ে গেল চিরকাল অরক্ষিত।	104	19	24	39	46
TEL	మను చెవి పరిపిత శబ్దం కన్నా ఎక్కువ తీవ్రతతో ఉన్న శబ్దాలు వినటం వల్ల వినికిడి శక్తి కొల్పోయే ప్రమాదం ఉంది	104	16	34	39	60

Figure 5: Tokenization density across the architectures

Overall, the study highlights the immense potential of large language models and prompt-based techniques for grammatical error correction in low-resource settings.

For the paper, data and codes, please scan the QR.