

A Multi-stage Framework for Spelling Correction in Low-Resource Languages with Data Augmentation and Prompt Learning

Xinfeng Liao¹, Jiahang Xu², Jie Li¹, Xixuan Huang^{3*}

¹School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510006, China

²School of Economics and Trade, Guangdong University of Foreign Studies, Guangzhou 510006, China

³Piston Intelligence CO. Ltd., Guangzhou 510000, China

Keywords: Low-resource languages; Spelling correction; Data augmentation; Prompt learning; IMLIP 2025

Citation: Liao X.F., Xu J.H., Li J., et al.: A Multi-stage Framework for Spelling Correction in Low-Resource Languages with Data Augmentation and Prompt Learning. *Data Intelligence*, 2026, Vol. 8, Art. No.: 20260092. DOI: <https://doi.org/10.3724/2096-7004.di.2026.0092>

ABSTRACT

Spelling Correction (SC) in low-resource languages is significantly challenged by scarce labeled data and unbalanced error distributions. To address these limitations, this paper presents a multistage framework integrating data augmentation and prompt learning. First, we establish a hybrid rule-LLM augmentation mechanism that generates large-scale pseudo-error corpora from diverse sources, ensuring broad error coverage. Second, we implement a progressive training paradigm—transitioning from augmented data to human annotations—to effectively mitigate distribution biases during error detection. Finally, we introduce a prompt-based fine-tuning strategy with hierarchical templates, guiding lightweight Large Language Models (LLMs) to perform precise correction via error localization signals. Our framework ranked first in all four tracks of the IMLIP 2025 Low-Resource SC evaluation. This work provides a practical data-centric solution and demonstrates that lightweight models, when equipped with effective data strategies, can address data scarcity in low-resource spelling correction.

1. INTRODUCTION

Spelling errors are ubiquitous in human language interaction. Such errors not only compromise text readability but also severely hinder the performance of downstream Natural Language Processing (NLP)

*Corresponding author: Xixuan Huang (E-mail: 736797854@qq.com; ORCID: 0000-0003-4729-6075).

tasks. Spelling Correction (SC) aims to detect and correct spelling errors in text, thereby generating output that conforms to spelling norms. However, existing research predominantly focuses on high-resource languages such as English and Chinese. In contrast, low-resource languages face distinct challenges, including the scarcity of annotated data and imbalanced error types, necessitating the development of targeted solutions.

Current spelling correction methods generally fall into three categories: rule/statistical-based, neural-based, and Large Language Model (LLM)-based approaches. Rule/statistical methods rely on manual features and lack generalization, as seen in confusion set [1] and edit distance [2] studies. Neural methods, such as Seq2Seq [3] and BERT [4], depend on large-scale parallel corpora, limiting their efficacy in low-resource settings. While LLMs offer context-aware corrections, they suffer from high computational costs, strong data dependency, and uncontrollable error generation [5-6]. Specifically, fine-tuning demands high-quality data, which conflicts with the inherent scarcity in low-resource languages. Moreover, LLMs often struggle to precisely simulate specific errors like whitespace redundancy and omission, resulting in insufficient error pattern coverage. Additionally, high deployment costs hinder their adaptation to lightweight scenarios [7]. Consequently, existing works have not yet effectively addressed the collaborative optimization of data scarcity and error type specificity in low-resource contexts.

To address the aforementioned limitations, we propose a multi-stage spelling correction framework for low-resource languages that integrates data augmentation and prompt learning. The framework primarily consists of three core stages: a multi-source corpus augmentation strategy, multi-stage error detection training, and prompt-based error correction fine-tuning. Specifically, the multi-source corpus augmentation strategy first acquires diverse and rich initial corpora from various sources, and then constructs confusion sets based on edit distance via rulebased augmentation methods to controllably inject three types of errors: word substitution (WP), whitespace redundancy (SR), and whitespace omission (SM). Subsequently, an LLM-assisted augmentation module is utilized to generate controllable error types to extend coverage of complex error patterns. In the multi-stage error detection training stage, a three-stage progressive training strategy is adopted to gradually correct data bias. First, approximately 4 million entries of rule-augmented data are used for preliminary training of the detection model; second, about 240,000 entries of LLM-augmented data are introduced for the second stage of training; finally, 8,000 entries of manually annotated data are utilized for the third stage of training, thereby achieving gradual optimization of the model's recognition capability and bias correction. In the prompt-based error correction fine-tuning stage, by fusing detection localization signals with similar word candidate information, a hierarchical prompt template is designed to guide a lightweight LLM (3B parameter scale) to achieve precise correction.

This study verifies the effectiveness of the proposed framework on the Indonesian and Vietnamese tracks of the IMLIP 2025 First Low-Resource Language Spelling Correction Shared Task. In the error detection task, the multi-stage training strategy substantially improves performance, increasing the F1 score from 24.96% to 91.17% for Indonesian and from 56.83% to 90.36% for Vietnamese. In the error correction task, the prompt fusion-based method achieves an F1 score of 80.35% on Indonesian. Introducing similar word candidates improves correction performance more clearly for Indonesian than for Vietnamese, while LLM-augmented data benefits detection but may introduce distributional bias when directly used for correction. The proposed framework ranks first in both the detection and correction tasks of the Indonesian and Vietnamese tracks.

The main contributions of this paper can be summarized as follows:

- We propose a rule- and LLM-enhanced corpus augmentation framework for constructing large-scale spelling correction corpora for Indonesian and Vietnamese;
- We design a multi-stage progressive training strategy to mitigate the bias caused by weakly annotated data;
- We establish a detection-correction fusion mechanism that enables lightweight open-source models to achieve competitive performance in the IMLIP 2025 Low-Resource SC evaluation.

2. RELATED WORK

2.1 Spelling Correction Methods

Spelling correction has evolved from rule-based statistics to deep representation learning. Early methods relied on edit distance [8] and noisy channel models [9], incorporating confusion sets and phonetic features for Chinese [1-2]. While interpretable, these methods suffer from manual feature dependency and limited semantic modeling. Subsequently, neural networks dominated the field, with Seq2Seq [3, 10] and BERT-based models [11-12] significantly improving performance. To leverage prior knowledge, works like SpellGCN [13] and FASpell [14] integrated phonological and visual similarities into deep representations. However, such discriminative models lack flexibility in handling complex, variable-length edits [15]. Recently, Large Language Models (LLMs) have demonstrated strong zero-shot correction capabilities [5, 16-18] but face challenges regarding computational cost, hallucination, and over-correction. Consequently, collaborative strategies among small and large models have emerged [7]. Following this trajectory, this paper proposes a hierarchical framework combining precise detection by discriminative models with fluent correction by generative models.

2.2 Data Augmentation in Low-Resource Scenarios

Data scarcity remains a primary bottleneck for spelling correction in low-resource languages, as most research favors high-resource contexts [19-20]. The lack of large-scale parallel corpora severely hinders model generalization [21-22]. To mitigate this, traditional approaches rely on heuristic rules, such as generating pseudo-data via edit distance and confusion sets [23-25], or mining filtered revision histories from Wikipedia [26-27]. However, these methods often yield data with rigid semantics or residual noise. Recently, model-based data synthesis has gained traction. Techniques like back-translation [28], DAGA [29], and ReLM [30] demonstrate that generative models can significantly enhance performance in sequence labeling and correction tasks. Building on these advances, we propose a collaborative framework integrating rule-based coverage with LLM-based semantic generation to effectively address data scarcity in low-resource settings.

3. THE PROPOSED APPROACH

Figure 1 illustrates the proposed multi-stage framework, comprising three core modules: multi-source corpus construction, multi-stage detection training, and prompt-based correction fine-tuning. This section details their implementation.

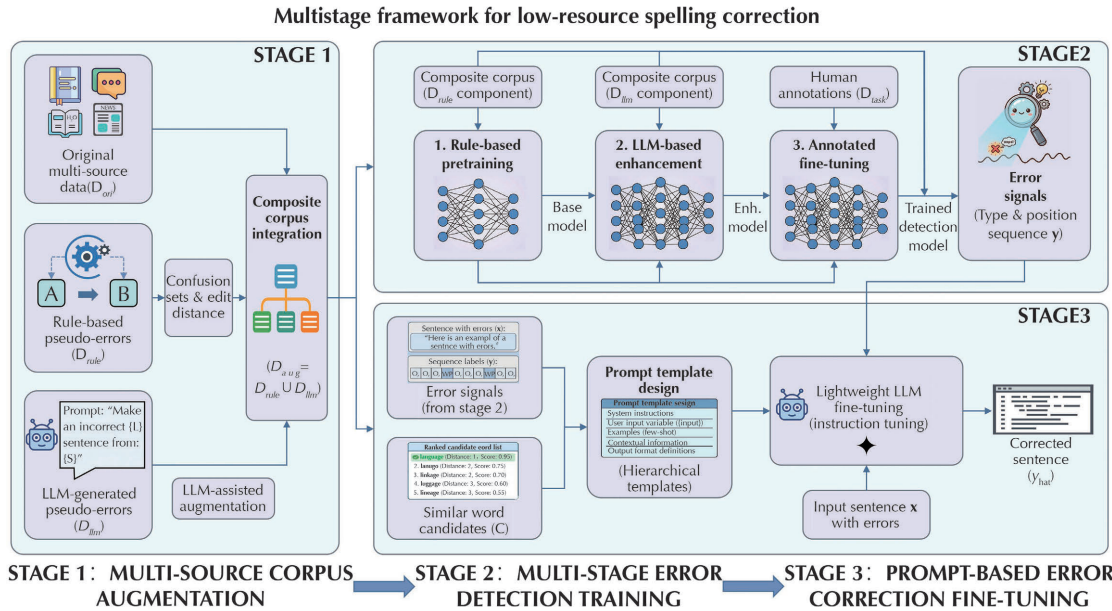


Figure 1. The overall architecture of the proposed multi-stage framework. It comprises three phases: multi-source corpus augmentation, multi-stage error detection training, and prompt-based fine-tuning for final correction.

3.1 Problem Formulation

This study formalizes spelling correction for low-resource languages as a cascaded “detectcorrect” task. Given an input sequence $\mathbf{x} = (w_1, \dots, w_n)$ that may contain (a) word substitution errors, (b) whitespace redundancy errors, and (c) whitespace omission errors, the model first predicts a token-level label sequence $\mathbf{y} = (y_1, \dots, y_n)$, where each label indicates the error type or the absence of an error. Subsequently, it generates a corrected sentence $\tilde{\mathbf{x}}$ conditioned on $\mathbf{x}$ and $\mathbf{y}$, such that it approximates the original clean sentence $\mathbf{x}^*$ in both semantics and structure.

3.2 Multi-Source Corpus Construction and Augmentation

To address data scarcity in low-resource languages, we first construct an original multi-source corpus $\mathcal{D}_{ori}$ and then derive two augmented corpora, namely the rule-based pseudo-error corpus $\mathcal{D}_{rule}$ and the LLM-generated pseudo-error corpus $\mathcal{D}_{llm}$. The resulting augmented training corpus is defined as:

$$\mathcal{D}_{aug} = \mathcal{D}_{rule} \cup \mathcal{D}_{llm} \quad (1)$$

3.2.1 Multi-Source Corpus Construction

Prioritizing coverage and domain relevance, we curate $\mathcal{D}_{ori}$ from diverse sources (see Section 4.1.1). To ensure quality, the corpus undergoes deduplication via text hashing and sampling control to balance source ratios, yielding a standardized dataset conforming to semantic and grammatical norms.

3.2.2 Rule-based Enhancement Mechanism

We simulate common error patterns through confusion sets and random token injection.

(1) *Confusion Set Construction*. For high-frequency words in the vocabulary $\mathcal{V}$, we compute edit distances to construct a candidate set $Conf(w)$. To enhance randomness, a special token $\langle other \rangle$ is included with an assigned probability weight s^{other} :

$$Conf(w) = \{(w'_1, s_1), \dots, (w'_k, s_k), (\langle other \rangle, s^{other})\}, \quad w \in \mathcal{V} \quad (2)$$

where k is the maximum number of candidates, and s_i denotes the sampling weight for candidate w'_i based on its edit distance to w , constrained by a similarity threshold τ .

(2) *Error Injection*. To inject errors dynamically, we apply an error generation function $f_{err}(\cdot)$ to a randomly sampled token w_j^* within a clean sentence $\mathbf{x}^* = (w_1^*, \dots, w_n^*) \in \mathcal{D}_{ori}$. This yields an error-injected input sentence $\mathbf{x}$, formally defined as:

$$\mathbf{x} = (w_1^*, \dots, w_{j-1}^*, f_{err}(w_j^*), w_{j+1}^*, \dots, w_n^*), \quad j \sim U(\{1, \dots, n\}) \quad (3)$$

The function f_{err} encompasses three operations:

- **Word Substitution (WP):** A word is sampled from $Conf(w_j^*)$. If $\langle other \rangle$ is selected, $f_{err}(w_j^*) = v$, where $v \sim U(\mathcal{V})$ represents a word uniformly sampled from the vocabulary.
- **Whitespace Redundancy (SR):** A whitespace character is inserted at a random position within w_j^* (applied only if $|w_j^*| > 2$).
- **Whitespace Omission (SM):** The whitespace between adjacent words (w_j^*, w_{j+1}^*) is removed.

3.2.3 LLM-assisted Enhancement Mechanism

To capture complex error distributions, we employ Qwen2.5-3B-Instruct with LoRA finetuning. We design a prompt function $T(\ell, \mathbf{x}^*) = Instr(\ell) \oplus \mathbf{x}^*$, where $\ell \in \{WP, SM, SR\}$ indicates the targeted error type, and $\oplus$ denotes string concatenation. The generation process for the unsupervised corpus $\mathcal{D}_{ori}$ is defined as:

$$\mathbf{x} \sim M_{LLM}^{finetune}(T(\ell, \mathbf{x}^*)), \quad \mathbf{x}^* \in \mathcal{D}_{ori} \quad (4)$$

This yields $\mathcal{D}_{llm} = \{(\mathbf{x}, \mathbf{x}^*)\}$, enriching the diversity of error patterns.

3.3 Multi-Stage Detection Training

We propose a “coarse-to-fine” curriculum framework to mitigate noise and improve generalization. Formulated as a sequence labeling task, the model optimizes the cross-entropy loss over the training dataset used at each stage:

$$\mathcal{L}_{det}(\theta) = - \sum_{i=1}^{|\mathcal{D}|} \sum_{t=1}^{T_i} \sum_{c \in \mathcal{Y}} \mathbb{I}(y_{i,t} = c) \log P_{\theta}(y_{i,t} = c | \mathbf{x}_i) \quad (5)$$

where θ represents the trainable parameters of the detection model, $\mathcal{D}$ denotes the training set used at the current stage, T_i is the length of the i -th sequence $\mathbf{x}_i$, $\mathcal{Y}$ is the tag set, $\mathbb{I}(\cdot)$ is the indicator function, $y_{i,t}$ is the ground-truth label, and $P_{\theta}(y_{i,t} = c | \mathbf{x}_i)$ represents the predicted probability assigned to class c .

1. **Rule-based Pretraining (Parameter Initialization):** The model initializes weights $\theta^{(1)}$ on $\mathcal{D}_{rule}$, acting as a robust feature extractor for rigid linguistic constraints and basic structural boundaries.
2. **LLM-based Enhancement (Distribution Smoothing):** Updating to $\theta^{(2)}$ on $\mathcal{D}_{llm}$, this stage serves as a critical bridge. Semantically rich LLM pseudo-errors help contextualize patterns, smoothing the rigid decision boundaries established in Stage 1.
3. **Annotated Fine-Tuning (Target Alignment):** Leveraging $\theta^{(2)}$ as a strong prior, minimal parameter updates are performed on the human-annotated downstream target dataset $\mathcal{D}_{task}$ to align the final model ($\theta^{(3)}$) with authentic human distributions while mitigating catastrophic forgetting.

The training trajectory is formalized as:

$$\theta^{(1)} \leftarrow \text{Train}(\theta^{(0)}, \mathcal{D}_{\text{rule}}), \theta^{(2)} \leftarrow \text{Train}(\theta^{(1)}, \mathcal{D}_{\text{llm}}), \theta^{(3)} \leftarrow \text{Train}(\theta^{(2)}, \mathcal{D}_{\text{task}}) \quad (6)$$

3.4 Prompt-Based Correction Instruction Fine-Tuning

We cast error correction as conditional generation ($\mathbf{x} \rightarrow \mathbf{x}^*$), optimized via parameter-efficient instruction tuning. Given predicted error positions $\mathcal{E}$ (derived from the predicted sequence labels $\hat{\mathbf{y}}$) and a candidate set $C = \{c_1, \dots, c_k\}$ extracted from the corresponding confusion sets $\text{Conf}(w)$, the comprehensive prompt template is defined as:

$$\mathcal{T}(\mathbf{x}, \mathcal{E}, C) = \text{instr} \oplus \mathbf{x} \oplus \text{ErrPos}(\mathcal{E}) \oplus \text{Cand}(C) \quad (7)$$

We evaluate four template variants: **Basic**(P_{instr}) (instruction + input), **Candidate**(P_{cand}) (adds Top- k candidates), **Position**(P_{pos}) (incorporates error positions), and **Fusion**(P_{comb}) (combines both). The model minimizes the conditional generation loss on the correction fine-tuning dataset $\mathcal{D}_{\text{corr}}$:

$$\mathcal{L}_{\text{corr}}(\theta) = - \sum_{i=1}^{|\mathcal{D}_{\text{corr}}|} \sum_{t=1}^{T_i} \log P_{\theta}(\mathbf{x}_{i,t}^* | \mathcal{T}_i, \mathbf{x}_{i,<t}^*) \quad (8)$$

where T_i denotes the length of the target corrected sequence.

4. EXPERIMENTS AND ANALYSIS

4.1 Datasets

We utilize the official dataset from the IMLIP 2025 Low-Resource Language Spelling Correction Task, comprising 10,000 labeled entries each for Indonesian and Vietnamese. As shown in Table 1, both languages exhibit significant data scarcity and class imbalance, with word substitution errors dominating. The dominance of word substitution suggests that most errors are local lexical deviations. Notably, the error patterns differ: whitespace redundancy is prevalent in Indonesian, whereas Vietnamese exhibits fewer whitespace redundancy cases and a relatively higher proportion of whitespace omission. These linguistic characteristics directly inform our targeted data augmentation strategies.

Table 1. Comparison of Error Category Distribution between Indonesian and Vietnamese.

Statistical Indicators	WP		SR		SM	
	Ind	Viet	Ind	Viet	Ind	Viet
Number of Errors	8,884	9,771	1,064	26	137	246
Total Words in Error Sentences	225,826	426,422	23,956	1,127	9,771	10,740
Sentence-level Proportion (%)	88.84	97.71	10.64	0.26	1.37	0.25
Word-level Proportion (%)	3.93	2.29	4.44	2.31	3.97	2.29

4.1.1 Unsupervised Data Sources

To ensure linguistic diversity, we constructed a multi-domain unsupervised corpus (detailed in Table 2) comprising: (1) Wikipedia and news datasets (IndoSum/VietNews) for formal patterns; (2) Universal Dependencies (UD) for syntactic structures; and (3) game texts (MNBVC) to capture colloquial expressions. This diverse combination enables the model to adapt across a spectrum of styles, from encyclopedic to conversational.

Table 2. Overview of Unsupervised Corpus Information for Indonesian and Vietnamese.

Language	Data Source	Sentences	Description
Indonesian	IdWiki	4,000,000	Well-structured encyclopedia texts
	IdUD	7,628	Professional-grade syntactic annotations
	Game Corpus	543,000+	Dialogue from Genshin, StarRail, etc.
	IndoSum	2,056,422	Colloquial news and summaries
Vietnamese	ViWiki	1,500,000	Sampled from viwiki dump
	ViUD	3,423	Syntactically annotated data from UD
	Game Corpus	543,000+	Same as Indonesian, Vietnamese portion
	VietNews	2,974,780	VNDS project, news summary data

4.2 Implementation

4.2.1 Dataset Construction

(1) Rule-based Augmentation: We first calculated the edit distances among the top 100,000 frequent words to construct a confusion set containing the top 100 nearest neighbors. Following the strategy described in Section 3.2 (with a similarity threshold $\tau = 0.1$), we configured the augmentation parameters based on a distribution-aware oversampling strategy. As shown in Table 1, word substitution (WP) naturally accounts for approximately 90% of the errors, serving as our baseline. However, to prevent the model from underfitting the long-tail minority classes (whitespace redundancy and whitespace omission), we empirically up-sampled their generation probabilities to 25% and 15%, respectively, which was selected based on validation performance. The intra-sentence error ratio was fixed at 5%. Based on this configuration, we constructed approximately 4.86 million Indonesian and 4.55 million Vietnamese rule-augmented samples.

(2) LLM-assisted Augmentation: We fine-tuned the Qwen2.5-3B-Instruct^① model via LoRA using an 8 : 2 split of the training set. The model was then employed to generate conditional noise for 300,000 clean sentences sampled from the unsupervised corpus. After alignment and filtering for specific error types, we obtained approximately 260,000 Indonesian and 240,000 Vietnamese high-quality LLM-augmented sentence pairs.

4.2.2 Model Architecture and Experimental Setup

(1) Error Detection: We treated error detection as a sequence labeling task and fine-tuned BERT-based^② models for each language. For Indonesian, we utilized IndoBERT^③; for Vietnamese, we employed CafeBERT^④. The tagging scheme followed the BIO format with labels {O, WP, SM, SR-B, SR-I}. Training was conducted using 16-bit floating-point (FP16) precision, with a maximum gradient norm of 2.0, a maximum sequence length of 192, and the AdamW optimizer paired with a linear warmup strategy. The multi-stage training configurations are summarized in Table 3.

Table 3. Multi-stage Error Detection Training Setup.

Stage	Training Data	Description	Learning Rate
Stage 1	$\mathcal{D}_{rule}$ (4.55M – 4.86M augmented samples)	Pretraining stage	4×10^{-5}
Stage 2	$\mathcal{D}_{llm}$ (240k – 260k LLM-augmented samples)	Second-stage training	2×10^{-5}
Stage 3	$\mathcal{D}_{task}$ (8k human-annotated samples)	Final fine-tuning	1×10^{-5}

(2) Error Correction: We formulated spelling correction as a conditional text generation task incorporating prompt engineering. Based on Qwen2.5-3B-Instruct, we performed LoRA fine-tuning (rank $r = 32$) using the LLaMA-Factory framework. Training was executed with FP16 precision, a learning rate of 5×10^{-5} , and a gradient clipping threshold of 1.0.

4.3 Experimental Results

4.3.1 Error Detection Results

For compactness, dataset combinations are denoted by compound subscripts; for example, $\mathcal{D}_{rule + llm}$ represents $\mathcal{D}_{rule} \cup \mathcal{D}_{llm}$.

① <https://huggingface.co/Qwen/Qwen2.5-3B-Instruct>.
 ② <https://huggingface.co/google-bert/bert-base-uncased>.
 ③ <https://huggingface.co/indolem/indobert-base-uncased>.
 ④ <https://huggingface.co/uitnlp/CafeBERT>.

Table 4 presents the error detection performance evaluated on the validation set. Training solely on $\mathcal{D}_{rule}$ yielded suboptimal F1 scores of 24.96% for Indonesian and 56.83% for Vietnamese, indicating that rigid rules alone were insufficient to capture authentic error distributions, although Vietnamese demonstrated better generalization on rule-based patterns compared to Indonesian. However, combining $\mathcal{D}_{rule}$ with $\mathcal{D}_{task}$ significantly boosted performance, improving the F1 score by 6.13 percentage points for Indonesian and 2.90 percentage points for Vietnamese compared to using $\mathcal{D}_{task}$ alone. This suggested that the rule-based dataset served as a robust prior, particularly for Indonesian.

Table 4. Error Detection Results for Indonesian and Vietnamese on 2,000 Validation Samples. All metrics are reported in %.

Training Data	Indonesian			Vietnamese		
	P	R	F1	P	R	F1
$\mathcal{D}_{rule}$	58.68	15.85	24.96	74.44	45.96	56.83
$\mathcal{D}_{task}$	84.25	81.47	82.84	87.46	85.76	86.61
$\mathcal{D}_{rule + llm}$	89.49	83.55	86.42	85.23	86.84	86.03
$\mathcal{D}_{rule + task}$	88.23	89.72	88.97	88.29	90.77	89.51
$\mathcal{D}_{rule + llm + task}$	91.23	91.12	91.17	90.95	89.78	90.36

Incorporating $\mathcal{D}_{llm}$ further enhanced the detection capabilities, yielding additional improvements of 2.20 and 0.85 percentage points for Indonesian and Vietnamese, respectively. Consequently, the full multi-stage approach achieved the best overall F1 scores of 91.17% and 90.36%. Overall, while Vietnamese exhibited stronger generalization under rule-based training, Indonesian benefited more substantially from the combined augmentation strategies. This suggests that multi-source data augmentation is particularly useful when rule-based patterns alone are insufficient to approximate authentic error distributions.

4.4 Error Correction Results

Table 5 presents the generative correction performance. We evaluated four prompt templates: the basic instruction (P_{instr}), the similarity-enhanced candidate template (P_{cand}), the positioninformed template (P_{pos}), and the comprehensive fusion template (P_{comb}). Notably, P_{instr} serves as a strong baseline, representing a direct end-to-end LLM correction approach without explicit detection guidance.

The empirical results reveal three key observations. First, incorporating the top-5 similar candidate words substantially improves correction F1, especially for Indonesian, confirming the importance of candidate reference words. However, increasing the candidate size to 10 slightly degrades performance, suggesting that excessive candidates may introduce contextual noise.

Table 5. Correction and Detection Results on Indonesian and Vietnamese Validation Sets. All metrics are reported in %.

Method	Correction			Detection		
	P	R	F1	P	R	F1
Indonesian						
P_{instr}	75.16	73.44	74.29	87.71	85.70	86.69
P_{cand} (Top- $k = 5$)	81.20	77.80	79.46	89.14	85.41	87.23
P_{cand} (Top- $k = 10$)	81.01	77.58	79.26	89.44	85.66	87.51
P_{pos}	75.94	75.56	75.75	89.12	86.66	88.89
P_{comb} (Top- $k = 5$)	81.99	78.77	80.35	89.92	86.38	88.11
P_{cand} (Top- $k = 5$) + LLM	81.24	77.29	79.21	89.82	86.38	87.58
Vietnamese						
P_{instr}	78.05	76.06	77.04	86.54	84.34	85.43
P_{cand} (Top- $k = 5$)	79.23	77.66	78.44	86.88	85.16	86.01
P_{cand} (Top- $k = 10$)	79.20	77.36	78.27	86.84	84.82	85.82
P_{pos}	77.74	78.65	78.19	88.15	89.17	88.66
P_{comb} (Top- $k = 5$)	79.91	79.60	79.75	88.87	88.52	88.70
P_{cand} (Top- $k = 5$) + LLM	81.46	77.31	79.33	87.82	83.35	85.53

Second, the fusion strategy achieves the best F1 scores among all evaluated prompt variants for both languages, thereby outperforming the isolated P_{cand} group. More importantly, compared with the general P_{instr} baseline, the gains achieved by P_{comb} highlight the advantage of the cascaded “detect-correct” framework over direct instruction-based generation. This demonstrates that explicit error position cues $\mathcal{E}$ derived from the detection module are vital for precise sequence correction.

Third, the effect of LLM-augmented data is mixed during correction. It improves Vietnamese correction mainly through higher precision, but slightly reduces Indonesian performance and lowers recall in both languages. We hypothesize that synthetic generative data may contain error artifacts that differ from human-generated errors, thereby biasing the model toward conservative predictions and reducing coverage of authentic error patterns.

5. CONCLUSION

This paper addresses the critical data scarcity in low-resource language spelling correction by proposing a data-centric multi-stage framework. We constructed large-scale, high-quality corpora for Indonesian and Vietnamese using a hybrid augmentation strategy that synergizes rulebased injection with LLM-driven generation. By employing a coarse-to-fine training paradigm and an instruction-tuned generative mechanism, our approach effectively mitigates data bias and optimizes lightweight model performance. The proposed framework ranked first in all four tracks of the IMLIP 2025 challenge, demonstrating the robustness of combining high-quality synthetic data with prompt learning. Future work will explore more robust filtering strategies for LLMgenerated errors to reduce distributional bias in correction.

AUTHOR CONTRIBUTIONS

Xinfeng Liao: Conceptualization, Methodology, Writing-Original Draft.

Jiahang Xu: Data Curation, Software, Validation.

Jie Li: Investigation, Formal Analysis.

Xixuan Huang: Supervision, Project Administration, Writing-Review & Editing.

ACKNOWLEDGMENTS

This work was supported by the CVTE Employment Capability Enhancement Project (Project No. 2025042398392). The authors also thank the anonymous reviewers for their valuable comments.

REFERENCES

- [1] C. J. Lin, W. C. Chu, "A study on Chinese spelling check using confusion sets and n-gram statistics," *Int. J. Comput. Linguist. & Chinese Lang. Process.*, vol. 20, no. 1, 2015.
- [2] D. H. Li, D.W. Peng, "Spelling correction for Chinese language based on pinyin-soundex algorithm," in *Proc. 2011 Int. Conf. Internet Technol. Appl.*, pp. 1–3, 2011.
- [3] S. Gu, F. Lang, "A Chinese text corrector based on seq2seq model," in *Proc. 2017 Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery (CyberC)*, pp. 322–325, 2017.
- [4] S. Zhang, H. Huang, J. Liu, et al., "Spelling Error Correction with Soft-Masked BERT," in *Proc. 58th Annu. Meet. Assoc. Comput. Linguistics*, pp. 882–890, 2020.
- [5] C. Liu, K. Zhang, J. Jiang, et al., "ARM: An alignment-and-replacement module for Chinese spelling check based on LLMs," in *Proc. 2024 Conf. Empirical Methods Nat. Lang. Process.*, pp. 10156–10168, 2024.
- [6] K. Li, Y. Hu, L. He, et al., "C-LLM: Learn to Check Chinese Spelling Errors Character by Character," in *Proc. 2024 Conf. Empirical Methods Nat. Lang. Process.*, pp. 5944–5957, 2024.
- [7] Z. Qiao, H. Zhou, Z. Li, "Mixture of Small and Large Models for Chinese Spelling Check," *arXiv preprint arXiv:2506.06887*, 2025.

- [8] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.
- [9] E. Brill, R. C. Moore, "An improved error model for noisy channel spelling correction," in *Proc. 38th Annu. Meet. Assoc. Comput. Linguistics (ACL)*, pp. 286–293, 2000.
- [10] H. T. Nguyen, T. B. Dang, L. M. Nguyen, "Deep learning approach for Vietnamese consonant misspell correction," in *Computational Linguistics: 16th Int. Conf. Pacific Assoc. Comput. Linguistics, PACLING 2019, Hanoi, Vietnam, Oct. 11–13, 2019, Revised Selected Papers 16*, pp. 497–504, 2020.
- [11] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguistics (NAACL)*, pp. 4171–4186, 2019.
- [12] Y. Cui, W. Che, T. Liu, et al., "Revisiting pre-trained models for Chinese natural language processing," *arXiv preprint arXiv:2004.13922*, 2020.
- [13] X. Cheng, W. Xu, K. Chen, et al., "SpellGCN: Incorporating Phonological and Visual Similarities into Language Models for Chinese Spelling Check," in *Proc. 58th Annu. Meet. Assoc. Comput. Linguistics (ACL)*, pp. 871–881, 2020.
- [14] Y. Hong, X. Yu, N. He, N. Liu, J. Liu, "FASpell: A Fast, Adaptable, Simple, Powerful Chinese Spell Checker Based On DAE-Decoder Paradigm," in *Proc. 2019 Workshop Noisy User-generated Text (W-NUT)*, pp. 160–169, 2019.
- [15] Z. Sun, X. Li, X. Sun, et al., "ChineseBERT: Chinese pretraining enhanced by glyph and pinyin information," *arXiv preprint arXiv:2106.16038*, 2021.
- [16] H. Wu, W. Wang, Y. Wan, W. Jiao, M. Lyu, "ChatGPT or Grammarly? Evaluating ChatGPT on Grammatical Error Correction Benchmark," *arXiv preprint arXiv:2303.13648*, 2023.
- [17] M. C. Penteado, F. Perez, "Evaluating GPT-3.5 and GPT-4 on Grammatical Error Correction for Brazilian Portuguese," in *LatinX in AI Workshop at ICML*, 2023.
- [18] T. Do, D. P. Tran, A. Vo, et al., "Reference-Based Post-OCR Processing with LLM for Precise Diacritic Text in Historical Document Recognition," in *Proc. AAAI Conf. Artif. Intell.*, pp. 27951–27959, 2025.
- [19] H. Wu, H. Zhang, R. Xuan, et al., "Bi-DCSpell: A Bi-directional Detector-Corrector Interactive Framework for Chinese Spelling Check," in *Findings Assoc. Comput. Linguistics: EMNLP 2024*, pp. 3974–3984, 2024.
- [20] M. Flor, M. Fried, A. Rozovskaya, "A benchmark corpus of English misspellings and a minimally-supervised model for spelling correction," in *Proc. Fourteenth Workshop Innovative Use NLP for Building Educ. Applications*, pp. 76–86, 2019.
- [21] L. Jiang, H. Wu, H. Zhao, et al., "Chinese spelling corrector is just a language learner," in *Findings Assoc. Comput. Linguistics ACL 2024*, pp. 6933–6943, 2024.
- [22] N. Luitel, N. Bekoju, A. K. Sah, et al., "Contextual spelling correction with language model for low-resource setting," in *2024 Int. Conf. Inventive Comput. Technol. (ICICT)*, pp. 582–589, 2024.
- [23] S. Singh, S. Singh, et al., "Handling real-word errors of Hindi language using N-gram and confusion set," in *Amity Int. Conf. Artif. Intell. (AICAI)*, pp. 433–438, 2019.
- [24] A. Jain, M. Jain, et al., "UTTAM: An efficient spelling correction system for Hindi language based on supervised learning," *ACM Trans. Asian Low-Resour. Lang. Inf. Process. (TALLIP)*, vol. 18, no. 1, pp. 1–26, 2018.
- [25] D. N. Tien, T. T. T. Minh, L. L. Vu, et al., "Vietnamese Spelling Error Detection and Correction Using BERT and N-gram Language Model," in *Intelligent Systems and Networks*, pp. 427–436, 2022.
- [26] A. Max, G. Wisniewski, "Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History," in *Proc. Seventh Int. Conf. Language Resources Eval. (LREC'10)*, pp. 827–832, 2010.

- [27] Y. Hu, F. Meng, J. Zhou, “CSCD-IME: correcting spelling errors generated by pinyin IME,” arXiv preprint arXiv:2211.08788, 2022.
- [28] R. Sennrich, B. Haddow, A. Birch, “Neural Machine Translation of Rare Words with Subword Units,” in Proc. 54th Annu. Meet. Assoc. Comput. Linguistics (ACL), pp. 1715–1725, 2016.
- [29] B. Ding, L. Yao, B. Kang, L. Carin, “DAGA: Data Augmentation with a Generation Approach for Low-resource Tagging Tasks,” in Proc. 2020 Conf. Empirical Methods Nat. Lang. Process. (EMNLP), pp. 6045–6057, 2020.
- [30] L. Liu, H. Wu, H. Zhao, “Chinese spelling correction as rephrasing language model,” in Proc. AAAI Conf. Artif. Intell., pp. 18662–18670, 2024.