

Probability-Deviation Guided Spelling Error Correction for Low-Resource Languages

Shenzhen CyberArray Network Technology Co., Ltd., Shenzhen, 518038, Guangdong, China

Yu Jia, Guanqiu Qin[†], Heng Shi, Yi Shen, Jianjiang Liu

Abstract

Spelling Error Correction (SEC) is a fundamental task in natural language processing, essential for ensuring textual quality across applications from writing assistants to professional publishing. While deep learning approaches have advanced significantly for high-resource languages, low-resource languages face substantial challenges due to a scarcity of annotated data. The recent emergence of large language models (LLMs) pre-trained on vast multilingual corpora offers promising potential for few-shot and zero-shot learning in low-resource SEC. However, their direct application is hindered by issues such as hallucination and overcorrection. To bridge this gap, we propose a unified framework designed to effectively harness LLMs for robust SEC in low-resource settings. Our approach is built on two pillars: reasoning optimization and data enhancement. We introduce a novel pipeline that integrates error detection—leveraging token-level probability deviations within the LLM—with generative correction, using a selective adoption strategy that applies corrections only where both modules agree. Furthermore, we construct reproducible data augmentation pipelines to generate large-scale, task-specific pretraining data for Indonesian and Vietnamese, mitigating data scarcity. Evaluated on the International Conference on Multilingual Intelligent Information Processing (IMLIP 2025) shared task benchmarks, our method achieves state-of-the-art detection performance (85.0% F1 for Vietnamese, 77.8% for Indonesian) and significantly improves correction precision (from 50% to over 86%), demonstrating the framework’s effectiveness and generalizability for low-resource language SEC.

Keywords: Spelling Error Correction; Large Language Models; Low-Resource Languages; Error Detection; Data Augmentation

1. Introduction

Spelling Error Correction (SEC) is a foundational task in natural language processing (NLP), critical for ensuring textual quality in applications from writing assistants to professional publishing. The field has evolved through distinct paradigms: early systems relied on rule-based methods like edit distance[1]; the growth of digital text enabled statistical approaches using n -gram probabilities[2]; and the current era is dominated by deep learning, where pre-trained

[†]Corresponding author: Guanqiu Qin (Email: qgqkelven@gmail.com; ORCID:0000-0002-1923-8591)

language models leverage contextual understanding to correct a wide range of errors [3]. However, significant research efforts and resources remain concentrated on high-resource languages such as English and Chinese.

For low-resource languages, automated spelling and grammatical error correction faces fundamental challenges. The direct transfer of techniques developed for well-supported languages often yields marginal gains, primarily due to the scarcity of large-scale, high-quality annotated corpora [4]. This data paucity hinders the effective training of data-intensive supervised models, including powerful masked language models like BERT [5].

The recent emergence of large language models (LLMs) pre-trained on vast multilingual data offers a promising alternative [6, 7, 8]. These models, with their immense parameter scales and strong generative capabilities, have demonstrated remarkable few-shot and zero-shot learning abilities across diverse tasks. Their implicit multilingual knowledge suggests substantial potential for SEC in low-resource language settings. Yet, applying LLMs directly to this task is challenging. Known issues such as hallucination [9] and overcorrection [10, 11] can degrade performance, allowing smaller, task-specific models to remain competitive on established benchmarks [12].

To bridge this gap, we propose a unified framework designed to more effectively harness the capabilities of LLMs for robust SEC. Our approach centers on two pillars: reasoning optimization and data enhancement. The former refers to a novel training strategy that unifies error detection and correction within a single LLM, using detection signals to refine the generation process. The latter involves constructing tailored pretraining data for low-resource languages to mitigate data scarcity. The principal contributions of this work are:

- **An efficient LLM-based error detection method.** Our approach equips LLMs with precise error detection capability, achieving state-of-the-art performance on both adapted and off-the-shelf models.
- **A detection-guided correction method.** We introduce a correction paradigm where the correction result by LLM generation is explicitly constrained and refined by detection signals, outperforming standard generation-only LLM baselines.
- **Reproducible data augmentation pipelines.** We design and validate systematic pipelines for generating spelling correction pretraining data for Indonesian and Vietnamese, effectively improving model performance in these low-resource settings.

2. Related Work

2.1. SEC for Low-Resource Languages

SEC has evolved from rule-based and statistical methods to neural approaches. While early neural models relied on sequence-to-sequence architectures [13], recent work often adopts more efficient edit-based paradigms (e.g., GECToR [14]) which frame correction as sequence tagging using encoder-only models like BERT, balancing accuracy and speed. A significant shift has been ushered in by LLM pre-trained on vast, multilingual corpora (e.g., Llama [7], Qwen [8]). By learning from trillions of tokens spanning hundreds of languages, they acquire unprecedented cross-lingual knowledge and general-purpose abilities, which naturally extend to strong few-shot performance on downstream tasks [15], including grammatical error correction. However, their application to SEC, particularly for low-resource languages, is underexplored, and issues like overcorrection and hallucination persist [9, 10, 11], making direct application suboptimal.

2.2. Language Modeling in Low-Resource Contexts

The performance of any neural correction model is fundamentally constrained by the quality of its underlying language model (LM). For languages like Indonesian and Vietnamese, the scarcity of large-scale monolingual corpora has driven the development of dedicated pretrained LMs (e.g., IndoBERT [16], BARTpho [17]), which typically outperform general multilingual models (e.g., mBERT [5]) on local tasks. To further alleviate data constraints, recent strategies employ multilingual seq2seq models (e.g., mT5 [6]) for cross-lingual transfer or train regional models (e.g., Sailor [18]) targeting specific language families. Despite these advances, generating high-quality text in extremely low-resource settings remains challenging, highlighting the need for data-efficient adaptation methods.

3. Problem Formulation

The SEC task aims to automatically detect and rectify spelling errors in text, transforming it into linguistically correct form. Formally, we frame it as a seq2seq text generation problem, where the core objective is to learn a mapping from the erroneous sequence to its corrected form. Given an input text sequence $S = \{w_1, w_2, \dots, w_n\}$, where w_i denotes the i -th word and which may contain spelling errors, the goal of an SEC system is to produce a corrected output sequence $\hat{S} = \{\hat{w}_1, \hat{w}_2, \dots, \hat{w}_n\}$. The corrected sequence $\hat{S}$ should preserve the semantic intent of the original input S while conforming to the orthographic norms of the target language.

4. Approach

Building upon the problem formulation in Section 3, an ideal spelling error correction system for low-resource languages should address two core challenges: (1) how to balance the powerful generative capability of LLMs with correction reliability to suppress overcorrection and hallucination, and (2) how to effectively mitigate data scarcity in low-resource settings. To this end, we propose a unified four-stage framework. Its core idea is to constrain the generative correction with a high-precision, probability deviation-guided detection module, while leveraging targeted data augmentation to solidify the model’s linguistic knowledge and task understanding.

As illustrated in Figure 1, the framework comprises the following stages: (1) supervised fine-tuning to align the model with the task, (2) probability deviation-based spelling error detection, (3) generative correction, and (4) a selective adoption strategy. The detection and correction modules work in concert, with the final output modifying the original text only where both modules reach a consensus, thereby significantly enhancing the overall precision and robustness of the system.

4.1. Supervised Fine-Tuning

To align the model with the data distribution of the SEC task and enhance its proficiency in low-resource languages, we perform Supervised Fine-Tuning (SFT) on the base LLM. The objective of this stage is to improve the effectiveness of our subsequent error detection stage (4.2) and correction stage (4.3). Therefore, we format the training data as "Instruction + Input Sentence + Output Label" and optimize the model using causal language modeling:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{|Y|} \sum_{t=1}^{|Y|} \log P(y_t \mid X_{\text{Instruction}}, S, y_{<t}; \Theta) \quad (1)$$

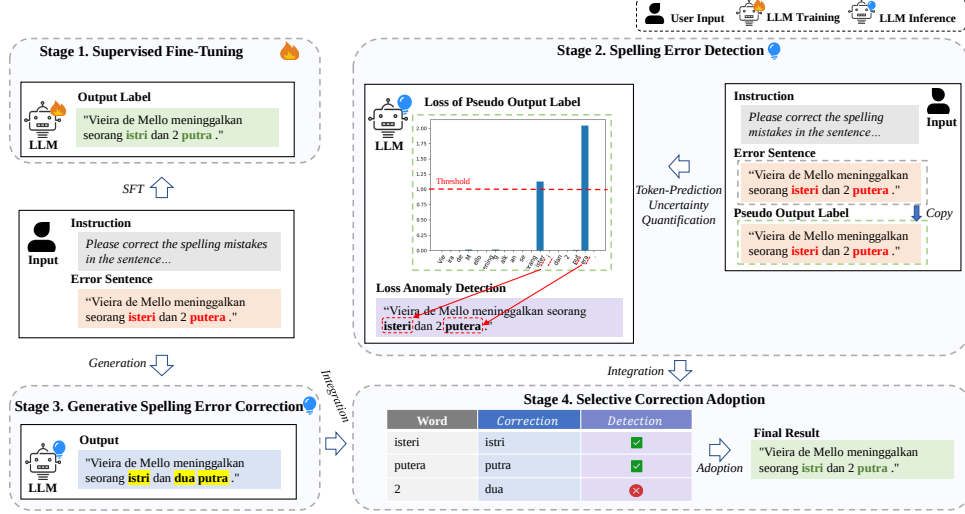


Figure 1: The overall architecture of the proposed Probability-Deviation Guided SEC framework. The pipeline consists of four stages: (1) Model Adaptation via Supervised Fine-Tuning, (2) Spelling Error Detection based on token-level probability deviation, (3) Generative Correction, and (4) Selective Adoption Decision. The detection set D acts as a filter for the correction set C , ensuring that an edit is applied only when both modules agree, thereby enhancing precision.

Here, $X_{\text{Instruction}} = \{x_i\}$ is the token sequence of the task instruction and output specification, S is the token sequence of the input sentence to be corrected (which is included in the input context but excluded from gradient computation for the loss), $Y = \{y_i\}$ is the token sequence of the ground-truth corrected sentence, and Θ represents the LLM parameters.

4.2. Spelling Error Detection

Generative SEC methods based on LLMs are inevitably susceptible to hallucination and over-correction. To mitigate this, inspired by the use of language model perplexity for text fluency evaluation [19], we propose a novel detection method that leverages the causal language modeling capability of LLMs. We focus on token-level probability deviation, where a higher deviation suggests the current token is a candidate for replacement.

For a word w composed of subword tokens $\{x_{w,k}\}_{k=1}^m$, we extract the maximum probability deviation $\max_k(\Delta p_{w,k})$ among its tokens. A word is flagged as potentially erroneous if this maximum exceeds a predefined threshold τ :

$$\text{Detect}(w) = \mathbb{I} \left[\max_{k \in [1, m]} (\Delta p_{w,k}) \geq \tau \right] \quad (2)$$

Applying this check to all words in a sentence yields the detection set D :

$$D = \{w \mid \text{Detect}(w) = 1\} \quad (3)$$

We observe that directly computing probability deviation on the original sentence can be confounded by issues like ungrammaticality or low fluency. To address this, we design a zero-shot

prompting scheme for more targeted deviation calculation, the core of which is the construction and use of a pseudo label.

The pseudo label $\tilde{Y}$ is a copy of the original sentence S (that is, $\tilde{Y} = S$). It is placed within a SEC task context: the model receives input formatted as “Instruction + Original Sentence S + Pseudo Label $\tilde{Y}$ ” and is tasked to autoregressively compute the sequence probability of $\tilde{Y}$ given the instruction and S . This setup leverages the task instruction to frame $\tilde{Y}$ as the corrected version of S for prediction. For tokens containing errors in S , the model assigns a markedly lower probability in this context, resulting in a sharp peak in the negative log-likelihood Δp_t . Compared to directly computing the perplexity of S , this method effectively isolates spelling error signals from general language disfluency through the task context.

Formally, let the pseudo label $\tilde{Y}$ be represented as a sequence of tokens $(\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_L)$, we compute the negative log-likelihood Δp_t for token $\tilde{y}_t$ conditioned on the instruction $X_{\text{Instruction}}$, the original sentence S , and the preceding context $\tilde{y}_{<t}$:

$$\Delta p_t = -\log P(\tilde{y}_t | X_{\text{Instruction}}, S, \tilde{y}_{<t}; \Theta) \quad (4)$$

The optimal probability deviation threshold τ of our method is determined by searching over the sorted probability deviation values of all tokens in the validation set.¹

4.3. Generative Correction

In the correction stage, we use the same “Instruction + Input Sentence” format as in SFT. Given the instruction prefix $X_{\text{Instruction}}$ and the input sentence $S = \{w_1, w_2, \dots, w_n\}$, the model generates the corrected sequence $\hat{S}$ autoregressively:

$$P(\hat{S} | X_{\text{Instruction}}, S; \Theta) = \prod_{t=1}^{|\hat{S}|} P(\hat{y}_t | X_{\text{Instruction}}, S, \hat{y}_{<t}; \Theta) \quad (5)$$

Subsequently, by comparing the original sentence S and the generated sentence $\hat{S}$, we extract all differing words to form the preliminary correction set C :

$$C = \{w_i | w_i \neq \hat{w}_i, w_i \in S, \hat{w}_i \in \hat{S}\} \quad (6)$$

where w_i and $\hat{w}_i$ denote the i -th word in the original and generated sentences, respectively.

4.4. Selective Adoption Strategy

To improve the reliability of the final output, we introduce a selective adoption strategy that addresses common issues in generative models such as overcorrection and error accumulation. The core of this strategy is to use the high-precision detection set D as a filter for the high-recall correction set C . A proposed correction is adopted into the final sentence if and only if the corresponding word was flagged as erroneous by the detector. Formally, the final corrected sentence $S_{\text{corrected}}$ is constructed as:

$$S_{\text{corrected}} = \{\hat{w}_i \text{ if } w_i \in D \text{ and } \hat{w}_i \in C \text{ else } w_i | i = 1, 2, \dots, n\} \quad (7)$$

This selective adoption acts as a consensus mechanism, effectively suppressing the generator’s overcorrection tendency while preserving the recall of the detector, thereby improving the overall precision and robustness of the SEC system.

¹We search τ over the sorted set of unique probability deviation values observed on the validation set (e.g., from ~ 0.3 to the maximum observed value), which is more efficient and data-adaptive than a fixed-range grid search.

5. Data Augmentation for Low-Resource SEC

To improve the performance of LLMs on SEC for low-resource languages, we employ a task-specific Continued Pretraining (CPT) strategy for Vietnamese and Indonesian. This approach enhances the model’s general language understanding and its specific knowledge of SEC patterns. We first constructed raw text corpora by collecting articles and dictionaries from public sources for each language. Following sampling and cleaning, approximately 2 million high-quality sentences were retained per language. Leveraging dictionary information, we then generated pre-training data specifically designed for spelling correction, yielding a final dataset with over 16 million tokens per language.

5.1. Data Sources

The data for continued pretraining was compiled from the following public resources:

- **WanJuan-Vietnamese** [20]: An open-source Vietnamese corpus exceeding 280GB, covering diverse domains such as history, politics, culture, commerce, and encyclopedic knowledge across 7 major and 34 minor categories.
- **IndoSum** [21]: A large-scale Indonesian summarization dataset compiled from online news articles.
- **The Leipzig Corpora Collection** [22]: A multilingual text corpus collection. We integrated contemporary Indonesian news articles from 2023–2025.
- **Vietnamese Dictionary** ²: A dictionary dataset scraped from the Vietnamese dictionary website ³, containing approximately 27,000 words and phrases.
- **Indonesian Dictionary** ⁴: A dictionary derived from the *Kompas* newspaper, comprising approximately 100,000 words and phrases.

5.2. Preprocessing

Separate preprocessing pipelines were designed for article data and dictionary data.

Article Data: Texts were first segmented into sentences. Excessively short or malformed sentences were discarded. MiniHash was then applied at the corpus level to remove near-duplicate sentences.

Dictionary Data: This data served as the basis for generating synthetic errors. In addition to the raw entries, we constructed two auxiliary resources: (1) a *valid split table* mapping compound words to their valid constituent parts (e.g., *otosista* → [*oto*, *sista*]), and (2) a *collocation table* recording frequent word pairs (e.g., *takaran* commonly co-occurs with *kehidupan* or *orambang*).

Specifically, for Vietnamese, we sampled 10,000 articles from each of the 7 major categories in WanJuan-Vietnamese, resulting in 5 million initial sentences. After cleaning, filtering, and deduplication, 2 million high-quality sentences were retained. For Indonesian, we merged data from IndoSum and the Leipzig news corpus (2023–2025) and applied an identical pipeline to obtain a final set of 2 million sentences.

²<https://huggingface.co/datasets/tsdocode/vietnamese-dictionary>

³<https://tudientv.com>

⁴<https://github.com/kata-ai/indosum>

5.3. Augmentation Method

To generate synthetic spelling errors that are both linguistically plausible and aligned with the error distribution of our target datasets (Table 1), we employ three core perturbation operations: **word replacement**, **space insertion**, and **space deletion**.

Each operation can be executed in either a *random* or a *rule-based* manner to balance diversity and authenticity.

- **Word Replacement**

- *Random*: For a selected word, one of the following perturbations is applied: (a) random insertion of a character, (b) random deletion of a character, or (c) replacement with a randomly selected word from the dictionary.
- *Rule-based*: A BK-tree constructed from the dictionary is used to find a candidate word within a Levenshtein distance of 1 or 2 from the original word for replacement.

- **Space Insertion**

- *Random*: A space is inserted at a random position within a word.
- *Rule-based*: For words present in the valid split table, a space is inserted at the boundary between their valid constituent subwords.

- **Space Deletion**

- *Random*: The space between two randomly selected consecutive words is deleted.
- *Rule-based*: The space between two consecutive words is deleted if the resulting bigram exists in the collocation table.

This pipeline generates diverse, controlled, and linguistically plausible spelling error samples, forming a high-quality pretraining corpus tailored for the SEC task in Vietnamese and Indonesian.

6. Experiments

6.1. Datasets

Table 1: Error type distribution in offline validation data.

Language	WR	SR	SM
Vietnamese	97.71%	0.26%	0.03%
Indonesian	88.84%	10.64%	1.37%

We evaluate on the Indonesian and Vietnamese datasets from Evaluated on the International Conference on Multilingual Intelligent Information Processing (IMLIP 2025) shared task. Each language provides 10,000 sentences for offline validation and 50,000 for online testing. We randomly split the validation data into 5,000/2,500/2,500 for training/development/test. As is shown in Table 1, error distribution analysis reveals distinct patterns. Vietnamese is dominated by word replacement errors, appearing in 97.71% of sentences, with minimal space-related issues. Indonesian exhibits more diversity: 88.84% word replacement, 10.64% space redundancy, and 1.37% space missing, reflecting its agglutinative morphology.

6.2. Experimental Setup

We employ Qwen2.5-Instruct-3B⁵ as the base language model. All training is conducted with bfloat16 mixed precision using the AdamW optimizer. A linear warmup schedule with a warmup ratio of 0.05 is applied. The training process consists of two distinct stages: Continued Pretraining (CPT) and Supervised Fine-Tuning (SFT), utilizing full-parameter tuning and rank-stabilized LoRA (rsLoRA)[23], respectively. The detailed hyperparameters for each stage are provided in Table 2.

Table 2: Training hyperparameters for Continued Pretraining (CPT) and Supervised Fine-Tuning (SFT).

Hyperparameter	CPT	SFT(rsLoRA)
Epoch	1	3
Total Batch Size	128	64
Learning Rate	1×10^{-5}	5×10^{-5}
LoRA Rank (r)	-	64
LoRA Alpha (α)	-	8
LoRA Target Modules	-	all-linear

6.3. Benchmark Evaluation

We evaluate the effectiveness of our general-purpose methodology on the official competition leaderboards for Indonesian and Vietnamese spelling error detection and correction. While our exploration was limited in scope (e.g., no extensive base model selection or prompt engineering was conducted) and our pipeline uses a simple ensemble without joint optimization between detection and correction modules, the results are promising. As shown in Table 3, our method (team CArray) achieved top-5 rankings across all four tracks, demonstrating its strong potential and generalizability for low-resource language SEC.

6.4. Error Detection Result

We conduct a comprehensive ablation study on the validation sets of both Indonesian and Vietnamese datasets to validate our proposed spelling error detection method. We use the same LLM for performance evaluation and comparison, with standard generative correction as our baseline. Specifically, we conducted three types of ablation experiments: (1) Zero-shot inference,

⁵<https://huggingface.co/Qwen/Qwen2.5-3B-Instruct>

Table 3: Leaderboard rankings (Top 5) for Indonesian and Vietnamese spelling error detection & correction tracks. Team “CArray” denotes our submission.

Rank	Detection(Ind)	Correction(Ind)	Detection(Viet)	Correction(Viet)
1	Team A	Team A	Team A	Team A
2	Team B	CArray (Ours)	CArray (Ours)	Team B
3	Team C	Team B	Team B	Team F
4	Team D	Team C	Team C	CArray (Ours)
5	CArray (Ours)	Team E	Team D	Team E

(2) Supervised fine-tuning (SFT) based on the evaluation dataset splits, and (3) Continued pre-training (CPT) using additional datasets before SFT. The optimal probability deviation threshold τ of our method is determined via grid search on the validation set.

Detailed results of these experiments for both languages are presented in Table 4 and Table 5, where we compare the performance across all three settings and analyze the impact of each training stage.

Table 4: Performance comparison of spelling error detection by different methods on offline Vietnamese dataset.

Method	Dev				Test		
	P	R	F1	τ	P	R	F1
Baseline _{Zero-Shot}	-	-	-	-	4.6	54.4	8.5
Baseline _{SFT}	-	-	-	-	34.9	71.0	46.8
Baseline _{CPT+SFT}	-	-	-	-	58.8	71.5	64.6
Ours _{Zero-Shot}	72.8	68.2	70.4	1.7	72.7	69.3	71.0
Ours _{SFT}	<u>75.4</u>	<u>76.3</u>	<u>75.9</u>	0.89	<u>76.5</u>	<u>76.3</u>	<u>76.4</u>
Ours _{CPT+SFT}	82.9	84.8	83.9	0.5	85.0	85.0	85.0

Table 5: Performance comparison of spelling error detection by different methods on offline Indonesian dataset.

Method	Dev				Test		
	P	R	F1	τ	P	R	F1
Baseline _{Zero-Shot}	-	-	-	-	5.1	39.0	9.1
Baseline _{SFT}	-	-	-	-	19.4	83.3	31.4
Baseline _{CPT+SFT}	-	-	-	-	44.7	<u>77.6</u>	53.4
Ours _{Zero-Shot}	60.2	50.5	55.0	1.52	58.6	49.5	53.7
Ours _{SFT}	<u>68.5</u>	<u>63.7</u>	<u>66.0</u>	0.87	<u>69.0</u>	62.1	<u>65.2</u>
Ours _{CPT+SFT}	80.1	77.1	78.5	0.5	79.9	75.8	77.8

As shown, our method substantially outperforms the baseline across all settings. With CPT+SFT training, we achieve 85.0% F1 on Vietnamese and 77.8% F1 on Indonesian test sets, surpassing the baseline CPT+SFT by 20.4% and 24.4% absolute respectively. Notably, our zero-shot performance already exceeds the fully trained baseline, delivering 71.0% F1 for Vietnamese and 53.7% for Indonesian.

Ablation reveals three key observations. First, SFT improves our method by 5.4% on Vietnamese and 11.5% on Indonesian over zero-shot. Second, CPT yields further gains of 8.5% and 12.6% F1 respectively, confirming that domain-specific pre-training enhances probability calibration. Third, the optimal threshold τ decreases with training, dropping from approximately 1.5 in zero-shot to 0.5 for CPT+SFT, indicating better-trained models require less conservative margins. The balanced precision and recall of our CPT+SFT trained method confirms its robustness for high-accuracy applications.

6.5. Error Correction Results

We evaluate the final correction performance using the ensemble method that integrates detection outputs with correction results via the consolidation strategy defined in Equation (7) (Section 4.4). Table 6 shows that our approach significantly outperforms the standalone generative baseline. On Vietnamese, Ours_{CPT+SFT} achieves 74.7% F1, a 10.1 point improvement

Table 6: Final spelling error correction performance comparison.

Method	Test _{Viet}			Test _{Ind}		
	P	R	F1	P	R	F1
Baseline _{Zero-Shot}	4.6	54.4	8.5	5.1	39.0	9.1
Baseline _{CPT+ST}	58.8	71.5	64.6	44.7	77.6	53.4
Ours _{CPT+ST}	86.4	<u>65.8</u>	74.7	86.9	<u>73.4</u>	79.6

over Baseline_{CPT+ST}’s 64.6%. The gain is larger on Indonesian, where we reach 79.6% F1 versus 53.4% for the baseline.

The improvement stems primarily from precision gain. Our method attains 86.4% precision on Vietnamese and 86.9% on Indonesian, far exceeding the baseline’s 58.8% and 44.7% respectively. This confirms that our detection module effectively filters overcorrections. Recall is slightly lower than the baseline’s peak of 71.5% (Vietnamese) and 77.6% (Indonesian), yet the precision boost yields a higher overall F1.

The zero-shot baseline performs poorly with F1 scores below 10% on both languages, highlighting the necessity of task-specific training. Our method’s strong performance validates that even a simple ensemble combining detection and correction provides substantial reliability gains.

7. Conclusion

This work presents a unified framework to effectively harness Large Language Models for Spelling Error Correction in low-resource languages. To address the inherent issues of LLMs such as overcorrection, we propose a novel detection-guided correction paradigm. This approach integrates a probability-based error detection module with a generative corrector, applying edits only where both agree, which significantly enhances reliability. Furthermore, we develop reproducible data augmentation pipelines to generate task-specific pretraining data for Indonesian and Vietnamese, effectively mitigating data scarcity. Evaluated on the IMLIP 2025 benchmarks, our framework achieves top-tier competitive rankings and demonstrates substantial gains, particularly in correction precision (over 86%). The results validate that coupling LLMs’ generative capability with constrained, detection-driven reasoning offers a robust and effective pathway for low-resource SEC.

8. Limitations and Future Work.

Despite its effectiveness, this work has limitations that open avenues for future research. Our pipeline employs a sequential, modular design with a post-hoc selective adoption strategy; a more tightly integrated, end-to-end architecture could enhance efficiency and synergy. While the data augmentation method proved successful for Indonesian and Vietnamese, its direct transfer to languages with vastly different morphological or writing systems requires further validation. Future work will explore dynamic thresholding mechanisms, more sophisticated context-aware error simulation, and the extension of the framework to joint spelling and grammatical error correction.

Author Contributions

Yu Jia: Conceptualization, Methodology, Supervision, Writing – review & editing. **Guan-qiu Qin:** Conceptualization, Methodology, Investigation, Data Curation, Formal Analysis, Validation, Visualization, Writing – original draft. **Heng Shi:** Project Administration, Resources. **Yi Shen:** Supervision, Funding Acquisition, Writing – review & editing. **JianJiang Liu:** Supervision, Funding Acquisition, Writing – review & editing.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. U22B2061).

References

- [1] F. J. Damerau, “A technique for computer detection and correction of spelling errors,” *Commun. ACM*, vol. 7, no. 3, pp. 171–176, 1964.
- [2] K. W. Church and W. A. Gale, “Probability scoring for spelling correction,” *Statist. Comput.*, vol. 1, no. 2, pp. 93–103, 1991.
- [3] S. Liu, T. Yang, T. Yue, F. Zhang, and D. Wang, “PLOME: Pre-training with misspelled knowledge for Chinese spelling correction,” in *Proc. 59th Annu. Meet. Assoc. Comput. Linguist. and 11th Int. Joint Conf. Natur. Lang. Process.*, 2021, pp. 2991–3000.
- [4] A. Magueresse, V. Carles, and E. Heetderks, “Low-resource languages: A review of past work and future challenges,” 2020, arXiv:2006.07264.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.*, 2019, pp. 4171–4186.
- [6] L. Xue *et al.*, “mT5: A massively multilingual pre-trained text-to-text transformer,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.*, 2021, pp. 483–498.
- [7] H. Touvron *et al.*, “Llama: Open and efficient foundation language models,” 2023, arXiv:2302.13971.
- [8] A. Yang *et al.*, “Qwen3 technical report,” 2025, arXiv:2505.09388.
- [9] Z. Ji *et al.*, “Survey of hallucination in natural language generation,” *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2023.
- [10] X. Zhang, X. Zhang, C. Yang, H. Yan, and X. Qiu, “Does correction remain a problem for large language models?” 2023, arXiv:2308.01776.
- [11] T. Park, H. Do, and G. Lee, “Leveraging what’s overfitted: Post-correction via LLM grammatical error overcorrection,” in *Proc. Conf. Empirical Methods Natur. Lang. Process.*, 2025, pp. 28183–28195.
- [12] L. Liu, H. Wu, and H. Zhao, “Chinese spelling correction as rephrasing language model,” in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 18662–18670.
- [13] A. Vaswani *et al.*, “Attention is all you need,” in *Adv. Neural Inform. Process. Syst.*, 2017, pp. 5998–6008.
- [14] K. Omelianchuk, V. Atrasevych, A. Chernodub, and O. Skurzhanyskiy, “GECToR—grammatical error correction: Tag, not rewrite,” 2020, arXiv:2005.12592.
- [15] T. Brown *et al.*, “Language models are few-shot learners,” in *Adv. Neural Inform. Process. Syst.*, 2020, pp. 1877–1901.
- [16] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP,” 2020, arXiv:2011.00677.
- [17] N. L. Tran, D. M. Le, and D. Q. Nguyen, “BARTpho: Pre-trained sequence-to-sequence models for Vietnamese,” 2021, arXiv:2109.09701.
- [18] L. Dou *et al.*, “Sailor: Open language models for South-East Asia,” in *Proc. Conf. Empirical Methods Natur. Lang. Process.: Syst. Demonstrations*, 2024, pp. 424–435.
- [19] G. Wenzek *et al.*, “CCNet: Extracting high quality monolingual datasets from web crawl data,” in *Proc. 12th Lang. Res. Eval. Conf.*, 2020, pp. 4003–4012.
- [20] J. Yu *et al.*, “WanJuanSiLu: A high-quality open-source webtext dataset for low-resource languages,” 2025, arXiv:2501.14506.
- [21] K. Kurniawan and S. Louvan, “IndoSum: A new benchmark dataset for Indonesian text summarization,” in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, 2018, pp. 215–220.

- [22] T. Eckart and U. Quasthoff, “Statistical corpus and language comparison on comparable corpora,” in *Building and Using Comparable Corpora*, Berlin, Germany: Springer, 2013, pp. 151–165.
- [23] D. Kalajdziewski, “A rank stabilization scaling factor for fine-tuning with lora,” 2023, arXiv:2312.03732.

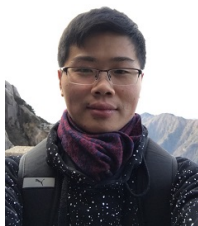
Author Biography



Yu Jia, the Director of Business Department, Shenzhen CyberArray Network Technology Co., Ltd. His research focuses on artificial intelligence, particularly in content & cognitive security, AI & big data, and computer vision analysis. He has led several sub-projects under national key projects.



Guanqiu Qin is a technical R&D staff member in the IT Department at Shenzhen CyberArray Network Technology Co., Ltd. He has participated in multiple key national research projects. His long-term research focuses on content and cognitive security, as well as fundamental studies of LLMs.



Heng Shi, Assistant Engineer, R&D staff at Shenzhen CyberArray Network Technology Co., Ltd., has long been engaged in application-software development and data analytics for artificial-intelligence and big-data solutions.



Yi Shen, Research Fellow, Member of the Science and Technology Committee of CHINA ELECTRONICS TECHNOLOGY CYBER SECURITY CO.,LTD., Deputy Director of Guangdong Engineering Technology Research Center for Cyberspace Cognitive Domain, Expert of Sichuan Provincial Department of Science and Technology, and Expert Consultant of the National Technology Center of Qianjia Technology. He has long been engaged in research related to content and cognitive security, and has presided over and participated in many national key projects.



Jianjiang Liu, the Director of the Guangdong Provincial Engineering Technology Research Center for Cyberspace Cognitive Security, an expert of the Shenzhen Science and Technology Innovation Bureau, and an expert of the Shenzhen Office for Civil-Military Integration. He is a Senior Engineer with extensive experience in information security, communication confidentiality, and content and cognitive security. He has led and participated in multiple national key research projects.