Length-Aware Chinese Spelling Correction with Retrieval and Iterative Refinement

Junhong Liang[1], Junnan Zhu[2], Yupu Liang[2], Bojun Zhang[2], Feifei Zhai[2], Yu Zhou[2]
[1]Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
[2]Institute of Automation, Chinese Academy of Sciences, Beijing, China
junhong.liang@mbzuai.ac.ae

***Abstract*—Chinese Spelling Correction (CSC) traditionally assumes *equal-length* character substitution and relies on pre-trained language models. Recent Large Language Models (LLMs) improve correction fluency but often fail in *domain-specific* settings (e.g., specialized terminology) and are unreliable under strict *length constraints*. Moreover, real-world CSC frequently includes *variable-length* phenomena such as ASR N-best correction and character splitting/merging, which are underexplored. We propose RAIR (Retrieval-Augmented Iterative Refinement), a model-agnostic framework that improves domain adaptation via (i) an adaptive multi-source retrieval corpus built from domain dictionaries and training sentences, (ii) a supervised fine-tuned retriever that is robust to misspellings and captures correction patterns, and (iii) Multi-turn Length Reflection (MLR) that iteratively refines LLM outputs to satisfy task-specific length constraints. Finally, an Adaptive Selection strategy switches between retrieval-augmented and direct generation. Experiments on domain CSC (ECSpell, LEMON) and variable-length settings (ChineseHP/Aishell-1, CSEC) show that RAIR consistently improves multiple LLM backbones and yields strong performance under both equal-length and variable-length scenarios.**

***Index Terms*—LLM, Chinese Error Correction**

I. Introduction

Chinese Spelling Correction (CSC) is a long-established task aimed at correcting misspelled characters in a sentence. Traditional CSC tasks focus on correcting general texts, such as errors in daily writing [1], [2], Optical Character Recognition (OCR) [3], or texts produced by Chinese learners [4]. However, Chinese spelling errors also appear in domain-specific scenarios [5], [6]. Addressing domain-specific spelling errors is crucial because the precision and accuracy of domain documents are essential for communication and avoiding misunderstanding. Traditional CSC methods are based on pre-trained language models (PLMs) such as BERT structure [1], [7] to capture the inherent semantic information of each token, while the complex structure and meticulous design of knowledge fusion of PLMs inherently narrow the scope in addressing domain-specific spelling errors, resulting in less effective results.

LLMs demonstrated excellent performance in several NLP tasks, and using LLMs for CSC has become a research trend. Most research introduces LLM as a generator to correct the equal-length spelling errors, such as by redefining tokenization rules [8] or providing character information using in-context learning [9]. However, challenges still exist for LLM methods. Firstly, it is essential to employ domain-specific knowledge to deal with spelling errors in specific domains, whereas existing approaches often overlook domain-specific insights. Secondly, it is vital to maintain the equal-length requirement in domain spelling correction, while the auto-regressive generation pattern of LLMs may fail to comply with strict length constraints in CSC tasks. Lastly, variable-length correction scenarios, such as ASR correction and character splitting, also exist in daily life and are not adequately investigated currently. The investigated error types are shown in Figure 1.

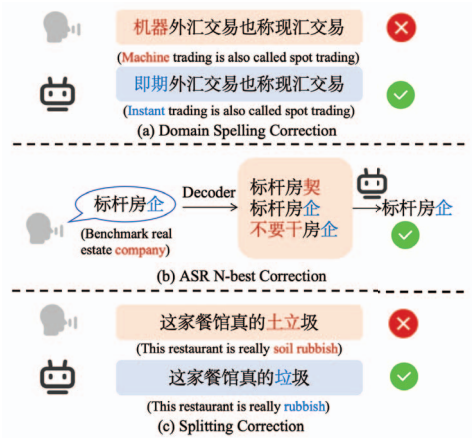

Fig. 1. Examples of Chinese error correction in different scenarios: a) Domain Spelling Correction, where an LLM must rectify semantically incorrect spelling from domain data while preserving sentence length. b) ASR N-best Correction, in which the LLM combines multiple speech recognition hypotheses and outputs the correct sentence from the speaker, and c) Splitting Correction, requiring the LLM to restore erroneously split characters into a single character. The wrong characters and their corrections are marked in red/blue respectively (best viewed in color).

These observations lead to a fundamental question overlooked by previous studies: ***how can LLMs perform reliable Chinese error correction under strict length constraints while remaining effective in domain-specific and variable-length scenarios?*** To address this overlooked challenge, we propose RAIR, a model-agnostic framework for length-constrained Chinese error correction across both equal-length and variable-length settings. This framework not only handles domain spelling errors but also ASR N-best Errors and Splitting Errors. We first dynamically construct a retrieval corpus from various sources, then fine-tune the semantic search retriever to catch the error correction pattern, and incorporate Multi-turn Length Reflection (MLR) to guide the output format for equal

and variable length scenarios, finally using adaptive selection to combine correction prediction from direct and retrieval augmented correction. Our contributions are summarized as follows:

- **Problem formulation:** We extend conventional equal-length CSC to a length-constrained Chinese error correction setting, covering domain-specific spelling correction, ASR N-best correction, and character splitting.
- **Error-pattern-aware retrieval:** We construct a multi-source retrieval corpus and fine-tune a retriever with positive/negative pairs that explicitly model misspellings, improving robustness to noisy inputs.
- **Length-constrained refinement:** We propose MLR, an iterative prompting mechanism that enforces task-specific length constraints and improves exact-match evaluation.
- **Strong empirical results:** RAIR consistently improves several LLM backbones across domain and variable-length benchmarks, with ablations validating the effect of retrieval, MLR, and adaptive selection.

II. Related Work

A. Chinese Error Correction Datasets

Traditional CSC datasets such as SIGHAN13/14/15 focus on the errors made by Chinese learners [10]–[12]; however, these data are unable to reflect the errors that exist in native speakers. Therefore, [2] proposes CSCD-NS, which collects the spelling errors in social media. In order to explore domain error correction abilities, [13] creates ECSpell datasets, creating domain-specific errors in three domains, and [6] proposes the LEMON dataset, including real spelling errors in seven domains. To investigate further variable-length correction, several datasets such as CSEC [14] and ChineseHP [15] are proposed to address the splitting and ASR decoding errors.

B. Chinese Spelling Correction Methods

Traditional CSC approaches mainly rely on transformer-based PLMs and can be broadly categorized into two frameworks. The Information-Learning paradigm integrates phonetic and glyph features for direct correction [16], [17], while the Detection-Correction paradigm first identifies errors and then performs correction [7], [18]. Recent work combines these paradigms by incorporating splitting features or phonetic-aware pretraining to enhance performance [14], [19].

With the emergence of LLMs, CSC has shifted toward generation-based approaches. Prior studies have explored prompt engineering to incorporate semantic context [9], pronunciation and glyph-based calibration for improving output reliability [20], and multimodal alignment to enhance error localization [21]. In addition, architectural adaptations such as character-level tokenization and specialized error detection training have also shown effectiveness [8], [22].

Apart from this, notable advancement comes from RAG approaches. Current solutions demonstrate effectiveness through two key mechanisms: error-robust retrieval that utilizes training data as knowledge sources [23], and few-shot corpus retrieval [24].

III. RAIR Framework

Our proposed framework is displayed in Figure 2, including a retrieval corpus, fine-tuned retriever, multi-turn length reflection, and adaptive selection.

A. Problem Definition

a) Existing Definition of Spelling Correction: Given an input sentence $\boldsymbol{X} = [x_1, \ldots, x_n]$ consisting of n characters, the goal is to generate a corresponding correct sentence $\boldsymbol{Y}$. If $\boldsymbol{X}$ contains a misspelled character x_i with its correct form being x_i', then the output should be $\boldsymbol{Y} = [x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n]$. This task requires the input and output lengths to remain consistent.

Traditional approaches focus only on equal-length spelling error correction; we extend the concepts into variable-length correction.

b) Definition of Splitting Correction: Given an input sentence $\boldsymbol{X} = [x_1, \ldots, x_m]$ consisting of m characters, where there may exist a continuous character sequence $x_i, \ldots, x_{i+l-1}$ of length l formed by splitting a single Chinese character. The original character $\hat{x}$ can be restored by merging these characters. The objective is to generate an output sentence $\boldsymbol{Y} = [x_1, \ldots, x_{i-1}, \hat{x}, x_{i+l}, \ldots, x_n]$. Since this task involves character merging, the input and output lengths may differ.

c) Definition of ASR N-best Correction: Given a set of candidate sentences from the ASR decoder $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, where X_i represents the i-th candidate sentence with length l_i, the goal is to generate the correct target sentence $\boldsymbol{Y}$.

B. Creation of Retrieval Corpus

For each dataset, we construct the retrieval corpus from two perspectives.

- **Domain Terms with Explanations**: Utilizing the THUOCL[1] lexicon, domain-specific terms from the legal and medical fields are extracted and denoted as W_D. These terms are then processed by GPT-4o as a domain expert E to generate detailed explanations for each term. This part is represented as $E(W_D)$.
- **Training Data and Expansion**: We first collect the correct sentences from each error–correction pair in the training set, denoted as S_{train}. To enrich contextual diversity, we further expand this set by prompting GPT-4o to generate coherent paragraphs conditioned on each sentence. These paragraphs are subsequently split into individual sentences, forming an augmented corpus denoted as $E(S_{\text{train}})$.

The final retrieval corpus S_R is the union of these three parts: $S_R = \{S_{\text{train}}, E(S_{\text{train}}), E(W_D)\}$.

For datasets without training data, we cannot build a task-specific retrieval corpus from supervision. In this case, we use GPT-4o to generate a background description conditioned only on the input sentence, which serves as auxiliary contextual

[1] http://thuocl.thunlp.org/

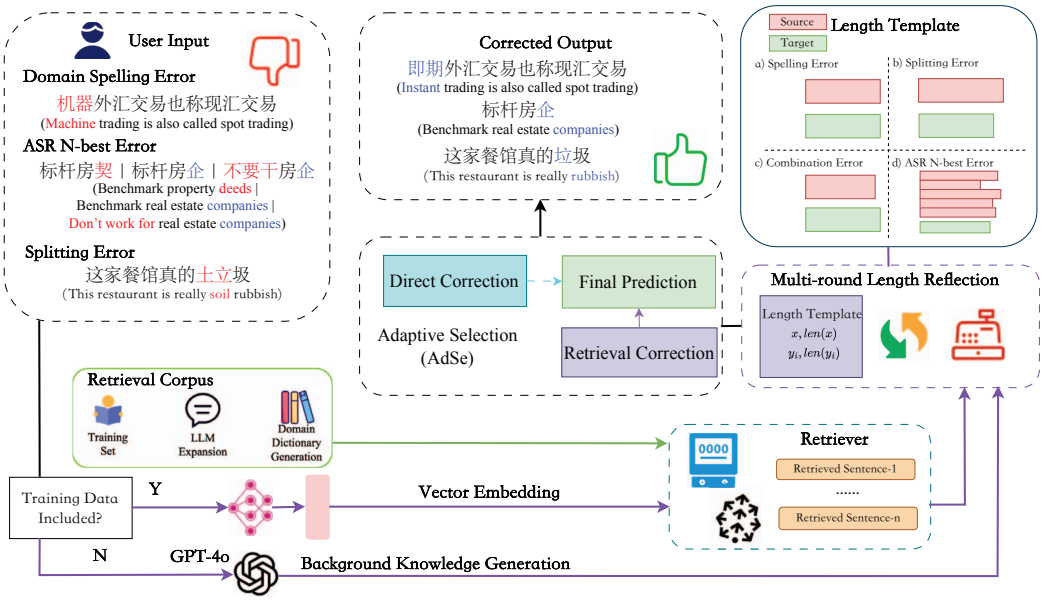

Fig. 2. Structure diagram of RAIR, including several modules such as retrieval corpus, retriever, multi-round length reflection and adaptive selection (best viewed in color).

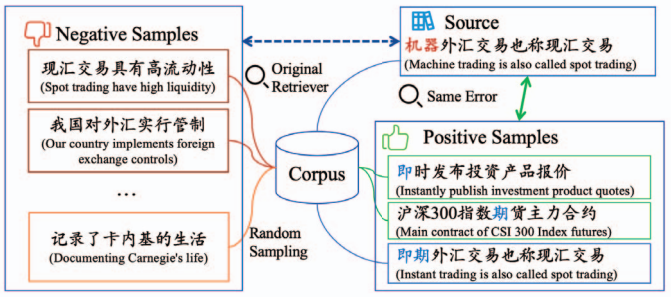

Fig. 3. Construct positive and negative samples of retrievers

evidence rather than retrieval from a fixed corpus. We note that this setting is methodologically different from standard corpus-based retrieval, and we therefore treat it as a fallback knowledge augmentation strategy for low-resource scenarios.

C. Retrieval Module

Given an input sentence x, the base retriever M_{Ret} returns the top-k most similar sentences $\{r_1, \ldots, r_k\} = M_{\text{Ret}}(x)$ based on semantic embeddings. However, when x contains misspelled characters, semantic retrieval alone is often insufficient. For example, in the sentence "机器外汇交易也称现汇交易", the erroneous token 机器 disrupts sentence-level semantics, causing a vanilla retriever to miss relevant domain knowledge for the correct term 即期.

To address this issue, we fine-tune M_{Ret} with supervised contrastive learning to make retrieval robust to character-level errors. As illustrated in Figure 3, we construct training instances using both positive and negative samples. Positive samples include: (i) the ground-truth corrected sentence y (e.g., "即期外汇交易也称现汇交易"), and (ii) corpus sentences that contain the same corrected character(s), such as "即时发布产品投资报价" and "沪深300指数期货主力合约". These samples encourage the retriever to associate noisy inputs with correction-relevant evidence rather than surface-form similarity.

Negative samples are constructed from (i) semantically related sentences that do not contain the correct character(s), and (ii) randomly sampled sentences from the retrieval corpus. This design forces the retriever to distinguish correction-relevant patterns from general topical similarity.

Formally, for a given input x, let $\{r_1^+, \ldots, r_m^+\}$ denote positive samples and $\{r_1^-, \ldots, r_n^-\}$ denote negative samples. We optimize the following contrastive objective:

$$\mathcal{L} = -\frac{1}{m}\sum_{i=1}^{m} \log \frac{\exp\left(s(\mathbf{E}(x), \mathbf{E}(r_i^+))/\tau\right)}{\sum_{r\in\{r^+\cup r^-\}} \exp\left(s(\mathbf{E}(x), \mathbf{E}(r))/\tau\right)},$$

where $\mathbf{E}(\cdot)$ is the embedding function of M_{Ret}, $s(\cdot,\cdot)$ denotes cosine similarity, and τ is a temperature parameter.

By fine-tuning the retriever in this way, M_{Ret} learns to retrieve correction-pattern-relevant sentences even when the query contains noisy or incorrect characters, providing reliable domain evidence for downstream LLM-based correction.

D. Multi-turn Length Reflection (MLR)

To enforce task-specific length constraints, we introduce *Multi-turn Length Reflection* (MLR), which enables the LLM to iteratively revise its output when length requirements are violated. Let $L(\cdot)$ denote the sentence length function. After each generation round i, we compare the output y_i with the input x and construct a *length report* $T(x, y_i, t)$ conditioned on the task label t. This report explicitly describes whether the length constraint is satisfied and, if not, provides corrective feedback to guide the next generation step. For equal-length spelling correction, MLR enforces $L(y_i) = L(x)$. For splitting correction, the constraint requires $L(y_i) \leq L(x)$. For ASR

TABLE I
DETAILED STATISTICS OF DATASETS USED IN THE EXPERIMENTS.

Dataset	Domain	#Sents	Len.	# Errors
ECSpell Train	*law*	1,960	60.4	1,059
	med	3,000	99.5	1,473
	odw	1,728	81.7	1,000
ECSpell Test	*law*	500	58.5	255
	med	500	98.2	226
	odw	500	80.6	266
LEMON	*car*	3,245	85.9	1,577
	con	993	79.2	441
	enc	3,274	78.9	1,591
	gam	393	64.6	148
	mec	1,942	77.4	905
	new	5,887	49.3	2,941
	nov	5,982	71.5	3,006
Aishell-1 Train	*news*	120,099	14.41	85,504
Aishell-1 Test	*news*	7,176	14.60	5,433
CSEC Train	*news*	7,869	30.7	8,344
	social	6,330	28.8	6,401
CSEC Test	*news*	1,174	30.9	1,268
	social	1,108	28.6	1,128

N-best correction, where the target length is ambiguous, we constrain $L(y_i)$ to fall within the minimum and maximum lengths observed in the candidate hypotheses. The LLM is re-prompted with this length feedback until the constraint is satisfied or a maximum number of rounds is reached.

E. Adaptive Selection

We propose an Adaptive Selection (AdSe) method to dynamically choose between retrieval-based and non-retrieval generation strategies based on task labels. If a training set of a certain task is available and included in the retrieval corpus, the retrieval-based method is prioritized; otherwise, the non-retrieval method is used by default. The system automatically switches to the alternative method if it fails to produce a valid sentence within k iterations, ensuring robust correction.

IV. EXPERIMENTS

A. Datasets

Domain Spelling Errors. We evaluate domain CSC on ECSpell [13] and LEMON [6]. ECSpell contains artificially constructed errors from three domains: legal (*law*), medical (*med*), and official document writing (*odw*), and provides a training set. LEMON is a real-world CSC benchmark spanning seven domains: automotive (*car*), contracts (*cot*), encyclopedia (*enc*), gaming (*gam*), medicine (*mec*), news (*new*), and novels (*nov*), and does not include training data. Both datasets consist of parallel error–correction sentence pairs.

ASR N-best Errors. We use the ChineseHP/Aishell-1 dataset [15], a benchmark for ASR correction in news reading speech, where each instance contains the top-10 decoding hypotheses produced by Belle-distilwhisper-large-v2-zh.

Splitting Errors. We adopt the CSEC dataset [14], which simulates character splitting errors in social media and news text using a Chinese splitting dictionary. Dataset statistics are summarized in Table I.

B. Baselines

We evaluate RAIR using both large language models (LLMs) and pretrained language models (PLMs). For LLM baselines, we consider GPT-3.5, Qwen2.5 [25], and DeepSeek-V3 [26], which represent strong general-purpose instruction-following models with different architectural designs and training data. Specifically, we use *gpt-3.5-turbo*, *qwen-plus*, and *deepseek-chat* as the deployed APIs for evaluation. For PLM-based CSC methods, we include MacBERT [27], which introduces CSC-oriented pretraining objectives with masked prediction, T5 [28], which formulates CSC as a text-to-text generation task, and ReLM [6], which treats spelling correction as a rephrasing problem. These baselines cover both conventional CSC architectures and modern LLM-based correction paradigms.

C. Evaluation Metrics

For spelling and splitting correction, we adopt the sentence-level criterion from [6], where a correction $\hat{y}$ is considered correct if and only if it matches the reference y, reporting precision (P), recall (R), and $F_1 = 2 * P * R/(P + R)$.

For ASR N-best correction, we follow the evaluation framework of [15] and report Character Error Rate (CER), which measures the edit distance between the predicted sentence and the reference, normalized by the reference length.

D. Detail Settings

To improve the retriever's robustness to error-prone inputs, we fine-tune *bge-m3* using the FlagEmbedding framework[2], with a negative sampling size $k = 5$, learning rate $lr = 10^{-5}$, warm-up ratio of 0.1, contrastive temperature $\tau = 0.2$, and two training epochs. For ASR N-best error correction, the top 5 predictions are selected. We use the Pinecone vector database to construct retrieval services. All experiments are conducted on two NVIDIA 3090 GPUs.

E. Main Results

Table II reports F_1 scores on ECSpell, LEMON, and CSEC, also CER on ASR N-best correction.

Spelling and Splitting Errors. On ECSpell, ReLM achieves the best performance in the *law* and *med* domains (91.2 and 82.4), while DeepSeek-V3+RAIR attains the highest score on *odw* (92.0). Across all LEMON domains, DeepSeek-V3+RAIR consistently outperforms all baselines, achieving up to 70.6 on *mec* and 70.5 on *new*. All LLMs are evaluated in a zero-shot setting, and RAIR improves performance across GPT-3.5, Qwen2.5, and DeepSeek-V3, with Qwen2.5 seeing the largest gains (e.g., 40.5 vs. 23.5 on ECSpell-*med*). On the CSEC splitting dataset, RAIR substantially boosts F_1 scores, improving DeepSeek-V3 from 54.2 to 66.1 on *News* and from 47.0 to 52.0 on *Social*.

[2]https://huggingface.co/BAAI/bge-m3

TABLE II
PERFORMANCE ON ECSPELL, LEMON, CSEC, AND AISHELL-1. F_1 IS USED FOR CSC AND SPLITTING, AND CER FOR ASR.

Methods	ECSpell (F_1)			LEMON (F_1)							CSEC (F_1)		ASR (CER↓)
	law	med	odw	car	cot	enc	gam	mec	new	nov	News	Social	Aishell-1
MacBERT	38.6	28.7	37.7	31.1	44.2	28.8	12.9	40.7	29.7	12.6	–	–	5.06
BERT-MFT	76.1	58.0	59.2	52.8	64.4	45.6	33.9	51.5	57.0	36.7	**73.1**	45.3	–
ReLM	**91.2**	**82.4**	83.6	53.6	67.7	47.7	34.6	53.9	58.8	38.0	–	–	5.20
GPT-3.5	38.0	26.0	47.4	21.4	30.7	29.6	11.8	29.8	22.8	15.2	13.6	5.8	9.84
GPT-3.5+**RAIR**	45.4	31.9	54.0	27.0	40.9	33.8	14.2	32.0	25.0	16.7	26.2	14.8	6.31
Qwen2.5	46.9	23.5	54.4	26.6	33.5	34.7	20.8	24.3	34.6	24.2	34.3	13.6	5.17
Qwen2.5+**RAIR**	68.0	40.5	73.0	38.8	50.1	49.4	27.1	47.9	45.2	34.1	47.2	26.9	4.37
DeepSeek-V3	82.7	67.6	90.2	56.2	65.6	60.6	40.9	68.7	68.8	57.7	54.2	47.0	4.94
DeepSeek-V3+**RAIR**	87.4	76.6	**92.0**	**60.4**	**69.1**	**61.6**	**43.7**	**70.6**	**70.5**	**59.4**	66.1	**52.0**	**4.15**

TABLE III
CORRECT LENGTH STATISTICS USING DIFFERENT MODELS IN ECSPELL DATASET, WHERE THE NUMBER INDICATES THE NUMBER OF PREDICTIONS WHICH HAVE EQUAL LENGTH WITH THE SOURCE SENTENCE.

Model	Domain			Total
	Law	Med	Odw	
GPT3.5	438	362	447	1247
GPT3.5+MLR	464	399	464	1327
Qwen2.5	337	231	322	890
Qwen2.5+MLR	478	377	420	1275
DeepSeek-V3	494	467	492	1453
DeepSeek-V3+MLR	**500**	**489**	**495**	**1484**
#Sents	500	500	500	1500

TABLE IV
RETRIEVAL PERFORMANCE ON THE ECSPELL TEST SET. #ERRORS DENOTES THE TOTAL NUMBER OF ERRONEOUS CHARACTERS.

Method	Hit@5↑			MRR↑		
	law	*med*	*odw*	*law*	*med*	*odw*
#Errors	390	356	407	–	–	–
bge-m3	315	185	328	0.87	0.76	0.85
bge-m3-ft	**369**	**248**	**352**	**0.88**	**0.83**	**0.90**

ASR N-best Errors. LLMs significantly outperform PLM baselines on ASR correction. Qwen2.5 and DeepSeek-V3 achieve substantially lower CERs than prior PLM results reported in [29]. The CER of top 1 decoding from ChineseHP/Aishell-1 data is 5.84. Although GPT-3.5 performs poorly in the zero-shot setting, RAIR reduces its CER by 35.9%. The best performance is obtained by DeepSeek-V3+RAIR, which yields a 28.9% relative CER reduction over the top 1 decoding.

F. Ablation Study

We perform ablation studies by removing retrieval, Multi-turn Length Reflection (MLR), and adaptive selection. As shown in Table V, removing any component degrades performance on ECSpell, LEMON, and CSEC. Retrieval and adaptive selection primarily support domain adaptation and robustness, while removing MLR causes the largest drop, especially under strict sentence-level evaluation where length violations are penalized. On the CSEC splitting dataset, the impact of MLR is particularly pronounced, confirming that explicit length control is essential for split/merge correction.

G. Analysis

a) Multi-turn Length Reflection: Table III shows that MLR substantially improves length correctness. For DeepSeek, MLR increases the number of length-valid outputs from (494, 467, 492) to (500, 489, 495). Qwen benefits even more from MLR, with gains of +141, +146, and +98 correct-length predictions, indicating that explicit length feedback effectively mitigates instruction-following errors. These results confirm that MLR is essential for enforcing length constraints and improving exact-match performance.

b) Retrieval Module: We evaluate retriever quality on ECSpell using Hit@5 and MRR (Table IV). The fine-tuned *bge-m3-ft* consistently outperforms the base *bge-m3* across all domains, achieving higher Hit@5 and MRR. This indicates that correction-aware supervision enables robust retrieval under noisy inputs and ranks relevant evidence higher, providing more accurate domain context for downstream LLM-based correction.

V. CONCLUSION

This work proposes **RAIR**, a length-aware framework for Chinese error correction that unifies equal-length and variable-length scenarios. By introducing Multi-turn Length Reflection, RAIR explicitly enforces length constraints and improves reliability under strict sentence-level evaluation. Experiments demonstrate consistent gains across multiple LLM backbones in zero-shot settings on domain spelling, splitting, and ASR N-best correction tasks. As a model-agnostic framework, RAIR can be readily applied without task-specific fine-tuning. Overall, our results highlight that explicit length control is crucial for robust and practical LLM-based Chinese spelling correction.

TABLE V
ABLATION RESULTS OF DEEPSEEK-V3 ON ECSPELL, LEMON, AND CSEC (F_1).

Base Model	Strategy	ECSpell			LEMON							CSEC	
		law	*med*	*odw*	*car*	*cot*	*enc*	*gam*	*mec*	*new*	*nov*	*News*	*Social*
DeepSeek-V3	RAIR	87.4	76.6	92.0	60.4	69.1	61.6	43.7	70.6	70.5	59.4	66.1	52.0
	w/o Retrieval	84.0	71.1	90.7	59.7	67.6	61.0	42.4	70.0	70.2	59.2	55.0	47.2
	w/o MLR	85.5	73.0	88.4	49.9	56.6	54.3	42.1	56.1	55.7	50.0	50.7	47.0
	w/o AdSe	84.5	71.7	91.9	53.0	60.4	58.5	41.2	59.4	56.9	51.9	63.8	49.9

REFERENCES

[1] H.-D. Xu, Z. Li, Q. Zhou, C. Li, Z. Wang, Y. Cao, H. Huang, and X.-L. Mao, "Read, listen, and see: Leveraging multimodal information helps chinese spell checking," 2021. [Online]. Available: http://arxiv.org/abs/2105.12306

[2] Y. Hu, F. Meng, and J. Zhou, "Cscd-ime: correcting spelling errors generated by pinyin ime," *arXiv preprint arXiv:2211.08788*, 2022.

[3] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," 2015.

[4] Y.-H. Tseng, L.-H. Lee, L.-P. Chang, and H.-H. Chen, "Introduction to SIGHAN 2015 bake-off for Chinese spelling check," in *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, L.-C. Yu, Z. Sui, Y. Zhang, and V. Ng, Eds. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 32–37. [Online]. Available: https://aclanthology.org/W15-3106/

[5] S. Song, Q. Lv, L. Geng, Z. Cao, and G. Fu, "Rspell: Retrieval-augmented framework for domain adaptive chinese spelling check," in *Natural Language Processing and Chinese Computing*, F. Liu, N. Duan, Q. Xu, and Y. Hong, Eds. Cham: Springer Nature Switzerland, 2023, pp. 551–562.

[6] H. Wu, S. Zhang, Y. Zhang, and H. Zhao, "Rethinking masked language modeling for chinese spelling correction," 2023.

[7] S. Zhang, H. Huang, J. Liu, and H. Li, "Spelling error correction with soft-masked bert," 2020.

[8] K. Li, Y. Hu, L. He, F. Meng, and J. Zhou, "C-llm: Learn to check chinese spelling errors character by character," 2024. [Online]. Available: https://arxiv.org/abs/2406.16536

[9] M. Dong, Y. Chen, M. Zhang, H. Sun, and T. He, "Rich semantic knowledge enhanced large language models for few-shot chinese spell checking," 2024.

[10] S.-H. Wu, C.-L. Liu, and L.-H. Lee, "Chinese spelling check evaluation at SIGHAN bake-off 2013," in *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, L.-C. Yu, Y.-H. Tseng, J. Zhu, and F. Ren, Eds. Nagoya, Japan: Asian Federation of Natural Language Processing, Oct. 2013, pp. 35–42. [Online]. Available: https://aclanthology.org/W13-4406/

[11] L.-C. Yu, L.-H. Lee, Y.-H. Tseng, and H.-H. Chen, "Overview of SIGHAN 2014 bake-off for Chinese spelling check," in *Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, L. Sun, C. Zong, M. Zhang, and G.-A. Levow, Eds. Wuhan, China: Association for Computational Linguistics, Oct. 2014, pp. 126–132. [Online]. Available: https://aclanthology.org/W14-6820/

[12] T.-H. Chang, H.-C. Chen, and C.-H. Yang, "Introduction to a proofreading tool for Chinese spelling check task of SIGHAN-8," in *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, L.-C. Yu, Z. Sui, Y. Zhang, and V. Ng, Eds. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 50–55. [Online]. Available: https://aclanthology.org/W15-3109/

[13] Q. Lv, Z. Cao, L. Geng, C. Ai, X. Yan, and G. Fu, "General and domain-adaptive chinese spelling check with error-consistent pretraining," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 5, May 2023. [Online]. Available: https://doi.org/10.1145/3564271

[14] J. Liang, Z. Junnan, F. Zhai, N. Cheng, C. Zong, and Y. Zhou, *A Hybrid Approach towards Chinese Spelling and Splitting Error Correction*. ECAI, 10 2024, pp. 3883–3890.

[15] Z. Tang, D. Wang, S. Huang, and S. Shang, "Pinyin regularization in error correction for chinese speech recognition with large language models," in *Interspeech 2024*. ISCA, 2024, pp. 1910–1914.

[16] Y. Hong, X. Yu, N. He, N. Liu, and J. Liu, "FASPell: A fast, adaptable, simple, powerful chinese spell checker based on DAE-decoder paradigm," in *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Association for Computational Linguistics, 2019, pp. 160–169. [Online]. Available: https://www.aclweb.org/anthology/D19-5522

[17] Z. Sun, X. Li, X. Sun, Y. Meng, X. Ao, Q. He, F. Wu, and J. Li, "ChineseBERT: Chinese pretraining enhanced by glyph and Pinyin information," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 2065–2075. [Online]. Available: https://aclanthology.org/2021.acl-long.161/

[18] C. Zhu, Z. Ying, B. Zhang, and F. Mao, "MDCSpell: A multi-task detector-corrector framework for Chinese spelling correction," in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1244–1253. [Online]. Available: https://aclanthology.org/2022.findings-acl.98/

[19] Z. Feng and F. Cao, "Cnmbert: A model for hanyu pinyin abbreviation to character conversion task," 2025. [Online]. Available: https://arxiv.org/abs/2411.11770

[20] H. Zhou, Z. Li, B. Zhang, C. Li, S. Lai, J. Zhang, F. Huang, and M. Zhang, "A simple yet effective training-free prompt-free approach to chinese spelling correction based on large language models," 2024. [Online]. Available: https://arxiv.org/abs/2410.04027

[21] Y. Yuhang, P. Yizhou, E. S. Chng, and X. Zhong, "Bridging speech and text: Enhancing asr with pinyin-to-character pre-training in llms," 2024. [Online]. Available: https://arxiv.org/abs/2409.16005

[22] Z. Xu, Z. Zhao, Z. Zhang, Y. Liu, Q. Shen, F. Liu, Y. Kuang, J. He, and C. Liu, "Enhancing character-level understanding in llms through token internal structure learning," 2024. [Online]. Available: https://arxiv.org/abs/2411.17679

[23] X. Yin, X. Hu, J. Jiang, and X. Wan, "Error-robust retrieval for chinese spelling check," 2024. [Online]. Available: https://arxiv.org/abs/2211.07843

[24] M. Dong, Z. Cheng, C. Luo, and T. He, "Retrieval-augmented generation for large language model based few-shot Chinese spell checking," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 10 767–10 780. [Online]. Available: https://aclanthology.org/2025.coling-main.717/

[25] Q. Team, A. Yang, B. Yang, B. Zhang *et al.*, "Qwen2.5 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2412.15115

[26] DeepSeek-AI, A. Liu, B. Feng, B. Xue *et al.*, "Deepseek-v3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2412.19437

[27] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for Chinese natural language processing," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 657–668. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.58/

[28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, Jan. 2020.

[29] J. Liang, "Perl: Pinyin enhanced rephrasing language model for chinese asr n-best error correction," 2024. [Online]. Available: https://arxiv.org/abs/2412.03230