# Reasoning to Find, Restraining to Fix: An Explainable Chinese Spell Checking Framework via Group Relative Policy Optimization

Ruiqi Wang[1], Jindian Su[1*], Jiebin Huang[1], Xiaobin Ye[2], and Dandan Ma[2]
[1] School of Computer Science & Engineering, South China University of Technology, Guangzhou, China
[2] Guangdong Unicomm, Guangzhou, China

***Abstract*—Chinese Spell Checking demands a delicate balance between detection sensitivity and correction faithfulness. Traditional discriminative models often lack the semantic reasoning to detect subtle errors, whereas generative Large Language Models (LLMs) suffer from severe over-correction. To bridge this gap, we propose $R^2$-CSC, a two-stage framework leveraging reasoning to find errors and restraining mechanisms to fix them. We first introduce a Recall-Oriented Supervised Fine-Tuning stage to enhance semantic sensitivity via Chain-of-Thought (CoT) instructions. Subsequently, we implement a Precision-Oriented Reinforcement Learning stage using Reference-Anchored Group Relative Policy Optimization. Unlike standard GRPO, RA-GRPO incorporates an identity baseline into advantage estimation. Coupled with Minimal Edit and Length Constraint rewards, this mechanism strictly penalizes unnecessary modifications by ensuring the model outperforms a do-nothing strategy. Extensive experiments demonstrate that $R^2$-CSC achieves state-of-the-art (SOTA) performance with a lightweight 4B parameter model. Notably, our method breaks the precision-recall trade-off, combining high semantic recall with BERT-level precision while offering interpretable reasoning paths.**

***Index Terms*—Chinese Spell Checking, Large Language Models, Reinforcement Learning, Group Relative Policy Optimization, Explainable AI.**

## I. Introduction

Chinese Spell Checking (CSC) aims at correcting erroneous characters in Chinese sentences [1, 2]. Unlike alphabetic languages, Chinese characters are ideographic and dense in information. A tiny change in a character can completely alter the sentence's meaning. Therefore, a high-quality CSC system must satisfy two conflicting requirements: Sensitivity (high recall) to detect subtle errors, and Faithfulness (high precision) to preserve the user's original intent without unnecessary modifications. However, existing methods are limited in their ability to balance these two objectives, resulting in a precision–recall trade-off. Traditional discriminative models based on Transformer [3], such as BERT [4], treat CSC as a character classification task [5–8]. While they achieve high precision on common phonological errors [9], they often fail to understand complex semantic logic, resulting in low recall for context-dependent errors. Recently, Large Language Models (LLMs) have shown promise due to their strong generation capabilities. Yet, straightforward Supervised Fine-Tuning (SFT) often makes LLMs prone to over-correction, causing them to rewrite sentences based on generation probability rather than correction rules, and sometimes hallucinate content. [10, 11].

To break this trade-off, we argue that an ideal CSC model should mimic a human expert's two-step process: first, use reasoning to analyze the context and find potential errors; second, use restraint to verify if a modification is strictly necessary. Based on this insight, we propose $R^2$-CSC (Reasoning and Restraining for Chinese Spell Checking), a novel framework powered by Reinforcement Learning (RL). Specifically, our method consists of two training stages. In the first stage, we use Recall-Oriented SFT to enhance the model's semantic sensitivity. We teach the model to generate a Chain-of-Thought (CoT) [12] analysis before correcting, enabling it to identify deep logical errors that BERT misses. In the second stage, we introduce Precision-Oriented Reinforcement Learning using a novel Reference-Anchored Group Relative Policy Optimization (RA-GRPO). Unlike standard GRPO [13] which relies solely on group averages, RA-GRPO incorporates an identity baseline into the advantage estimation. This means that if a model's modification yields a lower reward than simply keeping the original text (the identity mapping), it is strictly penalized. Coupled with our specialized Minimal Edit Reward, this mechanism encourage the model to refrain from changing the text unless it is fully certain.

Our contributions are summarized as follows:

- New Paradigm: We propose $R^2$-CSC, one of the first frameworks to apply Reference-Anchored GRPO to the CSC task. By integrating an identity baseline into the optimization objective, we successfully adapt the generative power of LLMs for precise error correction.
- Breaking the Trade-off: Extensive experiments show that our method achieves simultaneously high precision and recall. It maintains the high recall of generative models while matching the high precision of discriminative models, effectively solving the over-correction problem.
- Efficiency & Interpretability: We achieve State-of-the-Art (SOTA) results using a lightweight 4B parameter model, making it resource-efficient. Furthermore, our model provides explicit reasoning paths for every correction, enhancing the reliability of the system.

*Corresponding author: Jindian Su (sujd@scut.edu.cn)

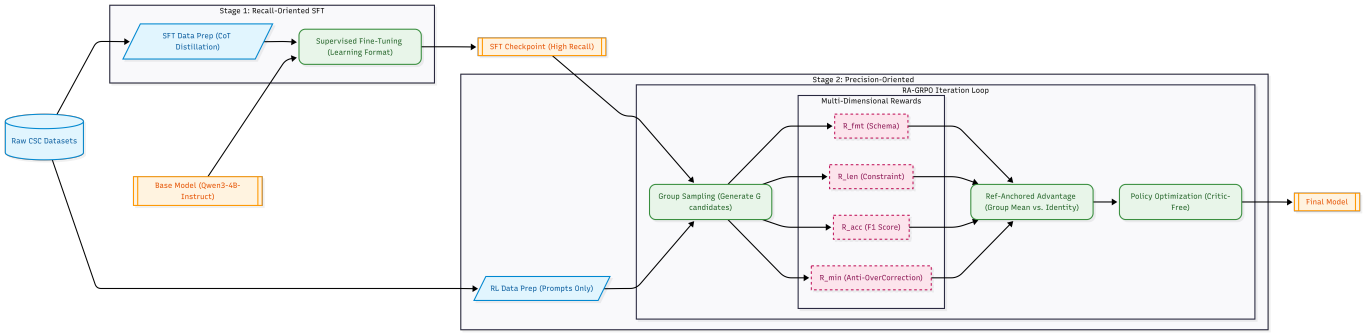



Fig. 1: The overall architecture of the $R^2$-CSC framework

## II. Related Work

### A. Discriminative Models for CSC

For a long time, CSC was treated as a sequence labeling or character classification task. Pre-trained language models like BERT [4] laid the foundation for this paradigm. Early works, such as Soft-Masked BERT [14], introduced a detection network to help the correction network focus on likely error positions. PLOME [15] uses a Gated Recurrent Unit (GRU) [16] to extract phonological and visual embeddings. CRASpell [17] aligns correction models with noisy and original contexts for improved training. PYGC [18] treats Chinese pinyin as an independent language model, employing parallel Transformer encoders to capture pinyin and semantic features. However, these discriminative models inherently lack semantic reasoning capabilities. By predicting characters independently based on local context, they frequently fail to detect subtle logical errors, resulting in limited recall for complex real-world scenarios.

### B. Generative Models for CSC

With the rise of Large Language Models (LLMs), Generative CSC methods have gained attention. Models like Qwen3 [19] typically perform CSC via zero-shot prompting or Instruction Tuning (SFT). Unlike BERT, LLMs model the probability of the entire sentence, giving them superior semantic understanding. Recent studies [9, 20–22] have shown that LLMs can correct complex errors that traditional models miss. Despite this, they suffer from severe over-correction, often altering valid user intent or violating the strict equal-length constraint required for CSC. Since standard Supervised Fine-Tuning often proves insufficient to curb these hallucinations, a specialized mechanism to restrain generation is essential.

### C. Reinforcement Learning and Chain-of-Thought

Standard RLHF typically relies on PPO [23], which is computationally expensive due to the need for a separate Critic model. While DPO [24] simplifies optimization, it lacks the ability to explore dynamic reasoning paths. Chain-of-Thought (CoT) [12] unlocks LLMs' reasoning potential, yet applying it to CSC requires balancing generation with restraint. Standard GRPO [13] eliminates the Critic via group sampling but overlooks the cost of unnecessary edits. To address this, we propose Reference-Anchored GRPO (RA-GRPO). By incorporating an identity baseline into advantage estimation, our method combines CoT reasoning with strict penalties for over-correction, effectively resolving the precision-recall trade-off.

## III. Methodology

In this section, we present $R^2$-CSC (Reasoning and Restraining for Chinese Spell Checking), a unified framework designed to break the precision-recall trade-off. As illustrated in Fig. 1, our method follows a "Coarse-to-Fine" curriculum, consisting of a Reasoning-Driven SFT stage for error detection and a Restraining-Guided GRPO stage for precise correction.

### A. Problem Formulation

Given a potentially erroneous Chinese source sentence $X = \{x_1, x_2, ..., x_n\}$ of length $n$, the goal of CSC is to generate a corrected target sentence $Y = \{y_1, y_2, ..., y_n\}$. A critical constraint in CSC is the Length Constraint, which requires the output length to be strictly equal to the input length ($|Y| = |X|$). In our framework, we extend the output to include a reasoning path. The model is trained to generate a sequence $O = [T; Y]$, where $T$ represents the CoT analysis.

### B. Recall-Oriented Supervised Fine-Tuning

Standard LLMs operate as probabilistic engines, generating text based on surface patterns rather than deep reasoning. To enhance the model's sensitivity to semantic errors, we construct a high-quality instruction tuning dataset $D_{SFT}$. Each sample in $D_{SFT}$ contains a rationale path $T$ distilled from a strong teacher model (e.g., ChatGPT), which explicitly explains the error type (phonological, visual, or semantic) before correction. We fine-tune the base model $\pi_\theta$ (initialized from Qwen3-4B-Instruct) using the Cross-Entropy loss:

$$\mathcal{L}_{SFT}(\theta) = -\sum_{t=1}^{|O|} \log P(o_t | X, o_{<t}; \theta) \tag{1}$$

where $o_t$ is the $t$-th token of the combined output sequence $O$. After this stage, the model gains the ability to follow the specific format and detect obscure errors, achieving High Recall. However, it may still suffer from hallucinations or over-correction due to the lack of negative constraints.

### C. Precision-Oriented GRPO

While standard Reinforcement Learning aligns models with human intent, applying it directly to CSC is non-trivial due to the high risk of over-correction. To address this, we propose Reference-Anchored GRPO (RA-GRPO), a task-specific variant of the Group Relative Policy Optimization algorithm.

*1) Reference-Anchored Advantage Estimation:* Standard GRPO estimates the advantage of an output $o_i$ by normalizing its reward against the group mean. However, in CSC, this relative comparison is flawed when the entire group suffers from hallucination. In such cases, the "least bad" hallucination would incorrectly receive a positive advantage. To strictly suppress over-correction, we introduce the Identity Reward Baseline $r_{id}$, which represents the hypothetical reward the model would receive if it simply copied the input source $X$ as the output. We modify the advantage calculation to be Reference-Anchored:

$$\hat{A}_i = \frac{r_i - \max(\mu_{group}, r_{id})}{\sigma_{group} + \epsilon} \tag{2}$$

where $\mu_{group} = \frac{1}{G}\sum_{j=1}^{G} r_j$ is the group mean reward. By introducing the $\max(\cdot)$ operator, we enforce a hard restraint constraint: If the model's elaborate reasoning and correction ($r_i$) yields a lower score than identity mapping ($r_{id}$), the advantage $\hat{A}_i$ becomes strictly negative. This mechanism mathematically penalizes any modification that fails to improve upon the original text, effectively teaching the model to favor the original sentence in ambiguous cases.

*2) Optimization Objective:* With the reference-anchored advantage $\hat{A}_i$ established, our goal is to optimize the policy $\pi_\theta$ to maximize the expected reward. To enable efficient updates, we employ importance sampling. Let $\rho_i(\theta)$ denote the probability ratio between the current policy and the sampling policy:

$$\rho_i(\theta) = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} \tag{3}$$

The optimization objective is to maximize the following loss function:

$$\mathcal{J}(\theta) = \mathbb{E}_{q,\mathbf{o}}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\rho_i(\theta)\hat{A}_i - \beta\mathbb{D}_{KL}(\pi_\theta||\pi_{ref})\right)\right] \tag{4}$$

where $\mathbb{E}_{q,\mathbf{o}}$ represents the expectation over the query dataset and the sampled output group. The term $\mathbb{D}_{KL}(\pi_\theta||\pi_{ref})$ calculates the KL-divergence between the current policy and the reference SFT model $\pi_{ref}$. This regularization term, controlled by the coefficient $\beta$, is crucial for preventing the model from deviating too far from the correct language distribution, ensuring that the generated text remains fluent and grammatically correct while optimizing for the specific CSC rewards.

*3) Reward Function Design:* The advantage $\hat{A}_i$ relies on a robust reward signal $r_i$. We design a composite reward function $R(o)$ consisting of four components to inject discrete constraints into the optimization: Format Reward ($R_{fmt}$): Ensures the generation of valid format for interpretability.

$$R_{fmt} = \mathbb{I}(o \in \text{ValidFormat}) \cdot \lambda_{fmt} \tag{5}$$

Length Constraint Reward ($R_{len}$): A hard constraint penalizing length mismatches between input $X$ and prediction $Y_{pred}$.

$$R_{len} = \begin{cases} \lambda_{len}^{+} & \text{if } |Y_{pred}| = |X| \\ \lambda_{len}^{-} & \text{otherwise} \end{cases} \tag{6}$$

Accuracy Reward ($R_{acc}$): Computed during the offline evaluation phase, this metric encourages the model to match the reference ground truth $Y_{gt}$ (using sentence-level F1).

$$R_{acc} = \text{SentenceF1}(Y_{pred}, Y_{gt}) \tag{7}$$

Minimal Edit Reward ($R_{min}$): This component works in tandem with our RA-GRPO advantage. Utilizing $Y_{gt}$ as the offline oracle, it explicitly rewards maintaining the original text when the input is correct.

$$R_{min} = \begin{cases} \lambda_{hold} & \text{if } X = Y_{gt} \wedge Y_{pred} = X \\ \lambda_{penalty} & \text{if } X = Y_{gt} \wedge Y_{pred} \neq X \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The total reward is $r_i = R_{fmt} + R_{len} + R_{acc} + R_{min}$.

## IV. Experiment

### A. Datasets and Evaluation Metrics

We evaluate our method on the widely used SIGHAN15 [25] dataset and the CSCD-NS [26] dataset. Following previous works, we use the standard training set from CSCD-NS for Stage 1 SFT. For Stage 2 RL, the model generates corrections using prompts without ground truth labels to simulate self-exploration, strictly reserving the ground truth for offline reward evaluation. For evaluation metrics, we employ sentence-level precision, recall, and F1-score.

### B. Hyperparameters

Our experiments, conducted on 6 NVIDIA RTX 5090 GPUs, use the ms-swift [27] framework. For Stage 1, we fine-tune the model for 1 epoch with a 256 global batch size. We apply a cosine scheduler, setting the initial learning rate to 5e-5 and warmup ratio to 0.003. For Stage 2, to ensure reinforcement learning stability, we reduce the learning rate to 5e-7 (0.01 warmup ratio) for 1 epoch. Given the group sampling memory overhead, per-device batch size is 1 with 8 gradient accumulation steps, and group size $G = 4$.

### C. Baseline

We compare $R^2$-CSC with following baselines:

-**BERT** directly fine-tunes BERT model for CSC task.

-**Soft-Masked BERT** uses a detection module to soft-mask wrong tokens.

-**PLOME** incorporates visual and phonetic features into a BERT-based model to enhance correction accuracy.

-**CRASpell** aligns correction models with noisy and original contexts for improved training.

-**PYGC** captures features through dual parallel encoders, late fusion and a pronunciation prediction task.

-**CSCLoRA** enhances LoRA [28] with layer-wise contrastive learning and masked alignment for efficient CSC.

TABLE I: The performance of different methods for CSCD-NS test dataset. Note that, $*$ indicates the results directly from [15, 18] and † means the results from our implementation.

| Methods | Detection-Level | | | Correction-Level | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| BERT* | 79.2 | 65.9 | 71.9 | 70.6 | 58.7 | 64.1 |
| Soft-Masked Bert* | 80.9 | 64.8 | 72.0 | 74.4 | 59.6 | 66.2 |
| PLOME* | 79.8 | 57.2 | 66.7 | 78.1 | 56.0 | 65.2 |
| CRASpell† | 73.6 | 66.9 | 70.0 | 71.5 | 65.0 | 68.1 |
| PYGC* | 81.3 | 68.3 | 74.4 | 79.2 | 66.3 | 72.2 |
| **$R^2$-CSC (ours)†** | **89.7** | **74.9** | **81.6** | **87.8** | **73.5** | **80.0** |

TABLE II: The performance of different methods for SIGHAN15 test dataset. Note that, $*$ indicates the results directly from [18, 20] and † means the results from our implementation.

| Methods | Detection-Level | | | Correction-Level | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| BERT* | 74.2 | 78.0 | 76.1 | 71.6 | 75.3 | 73.4 |
| Soft-Masked Bert* | 73.7 | 73.2 | 73.5 | 66.7 | 66.2 | 66.4 |
| PLOME* | 77.4 | 81.5 | 79.4 | 75.3 | 79.3 | 77.4 |
| CRASpell† | 83.5 | 89.2 | 86.3 | 81.9 | 81.6 | 81.7 |
| PYGC* | 82.1 | 83.9 | 83.0 | 80.1 | 81.9 | 81.0 |
| CSCLoRA* | 87.8 | 84.3 | 85.1 | 83.9 | 81.1 | 82.5 |
| **$R^2$-CSC (ours)†** | **90.2** | **88.9** | **89.6** | **87.7** | **88.6** | **88.1** |

### *D. Main Results*

On the CSCD-NS dataset (Table I), which is characterized by a high concentration of homophonic errors derived from Pinyin input methods and diverse user-generated content, our method achieves a correction-level F1 score of 80.0%, surpassing the previous SOTA (PYGC) by a remarkable margin of 7.8 points. Crucially, $R^2$-CSC achieves simultaneous improvements in both precision (+8.6%) and recall (+7.2%) compared to PYGC. The significant boost in Recall indicates that our Reasoning-Driven SFT effectively utilizes global context to resolve ambiguous homophones that rely heavily on semantic logic rather than just local character patterns, which discriminative models often fail to capture.

On the SIGHAN15 dataset (Table II), a widely used standard benchmark, our method achieves a Correction-Level F1 score of 88.1%, surpassing the strongest baseline (CSCLoRA, 82.5%) by a significant margin of 5.6 points. This demonstrates that while LoRA-based methods (CSCLoRA) are parameter-efficient, our Reasoning-Driven SFT + RA-GRPO paradigm unlocks a much higher upper bound for correction performance. In addition, the result reveals that CSCLoRA achieves a relatively high Precision (83.9%) but lags in Recall (81.1%), indicating a conservative tendency. In contrast, $R^2$-CSC achieves both superior Precision (87.7%) and exceptional Recall (88.6%). This confirms that our Identity Baseline in RA-GRPO effectively suppresses over-correction without sacrificing the model's ability to detect errors, successfully breaking the precision-recall trade-off that limits prior arts.

### *E. Ablation Study*

Table III validates the necessity of our component design. First, SFT outperforms the base model, proving its role in adapting the LLM to the task schema. Second, compared with the standard GRPO, the RA-GRPO stage drives the most significant gain, it boosts C-F1 by 3.8 points on SIGHAN15, confirming that the Identity Baseline successfully solves the over-correction bottleneck. Third, CoT is indispensable; removing reasoning paths ($R^2$-CSC w/o CoT) drops C-F1 by 4.5 points on CSCD-NS, highlighting its role in deep error detection. Finally, $R^2$-CSC yields best results, validating the synergy of reasoning and restraint.

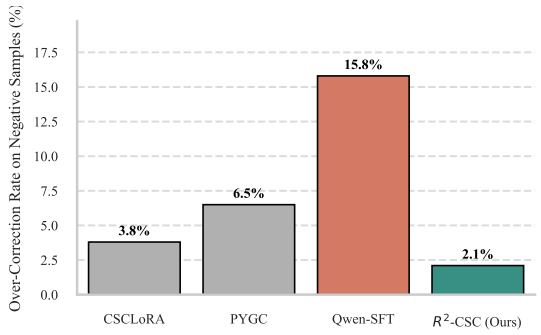
TABLE III: Ablation study results for our proposed method on CSCD-NS and SIGHAN15 test datasets. D-F1 represents the F1 score at the detection-level, and C-F1 represents the F1 score at the correction-level. Qwen3 denotes Qwen3-4B-Instruct.

| Methods | CSCD-NS | | SIGHAN15 | |
|---|---|---|---|---|
| | D-F1 | C-F1 | D-F1 | C-F1 |
| Qwen3 | 66.8 | 64.8 | 68.5 | 67.8 |
| Qwen3+SFT | 69.7 | 73.6 | 71.4 | 72.3 |
| Qwen3+GRPO | 69.9 | 69.3 | 72.7 | 70.2 |
| Qwen3+RA-GRPO | 71.0 | 70.7 | 74.9 | 71.6 |
| $R^2$-CSC w/o CoT | 78.2 | 75.5 | 84.9 | 85.6 |
| **$R^2$-CSC** | **81.6** | **80.0** | **89.6** | **88.1** |

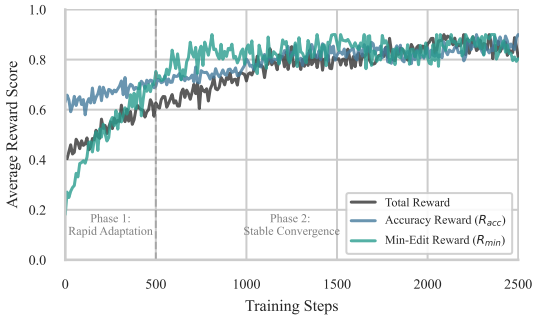
### *F. Analysis and Discussion*

*1) Impact of RA-GRPO on Over-Correction:* The core motivation of introducing RA-GRPO is to mitigate the over-correction problem prevalent in generative models. To quantify this, we define the Over-Correction Rate (OCR) as the percentage of negative samples that are wrongly modified by the model. A lower OCR indicates better faithfulness to the original text. Fig. 2a illustrates the OCR comparison on the SIGHAN15 test set.

- **The SFT Dilemma**: The Qwen3+SFT model exhibits a excessively high OCR. Without negative constraints, the SFT model tends to rewrite sentences based on language modeling probability, often replacing valid words with high-frequency synonyms.




(a) The result of the OCR




(b) Training dynamics of RA-GRPO

- **The RA-GRPO Solution**: After applying Stage 2 training, $R^2$-CSC dramatically reduces the OCR to 2.1%, a level comparable to the discriminative baseline CSCLoRA. This empirical evidence confirms that our Identity Baseline mechanism in RA-GRPO successfully reshapes the model's policy: when the reward of a generated correction does not exceed the reward of keeping the original text, the model learns to choose the latter.

*2) Training Dynamics and Reward Evolution:* To investigate how RA-GRPO enforces the restraint requirement during optimization, we analyze the evolution of different reward components throughout the training process. Fig. 2b visualizes the trajectories of the *Total Reward*, *Accuracy Reward* ($R_{acc}$), and *Minimal Edit Reward* ($R_{min}$).

We observe two distinct phases in the learning process:

- **Phase 1: Rapid Adaptation of Restraint.** In the early stages (Steps 0-500), the $R_{min}$ (Green line) rises sharply. This indicates that the model quickly adapts to the Identity Baseline, realizing that outputting the original text is a reliable strategy for maximizing rewards on ambiguous samples. This validates the efficiency of our Reference-Anchored advantage estimation.
- **Phase 2: Stable Convergence.** Unlike traditional PPO which often suffers from training instability, the RA-GRPO curves remain smooth and converge steadily. Crucially, as $R_{min}$ stabilizes at a high level, $R_{acc}$ (Blue line) maintains its upward trend. This suggests the model is not simply collapsing to a trivial copy-paste policy but is successfully balancing the trade-off: keeping correct sentences unchanged (High $R_{min}$) while precisely correcting erroneous ones (High $R_{acc}$).

This dynamic demonstrates that $R^2$-CSC learns a sophisticated policy boundary: it acts aggressively only when the confidence of error detection outweighs the penalty of the Identity Baseline.

## V. Conclusion and Future Work

In this paper, we introduced $R^2$-CSC, a framework designed to resolve the precision-recall conflict in Chinese Spell Checking. Our core contribution, Reference-Anchored GRPO (RA-GRPO), mathematically enforces restraint by incorporating an Identity Baseline into the optimization objective, effectively penalizing unnecessary edits. Extensive experiments demonstrate that $R^2$-CSC achieves SOTA performance with a lightweight 4B model, offering both accurate corrections and interpretable reasoning. Future work will extend this "Reference-Anchored" paradigm to other restrictive text editing tasks like grammatical error correction.

## References

[1] J. Yu and Z. Li, "Chinese spelling error detection and correction based on language model, pronunciation, and shape," 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. 220–223.

[2] J. Xiong, Q. Zhang, S. Zhang, J. Hou, and X. Cheng, "Hanspeller: a unified framework for chinese spelling correction," in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 20, Number 1, June 2015-Special Issue on Chinese as a Foreign Language*, 2015.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[5] J. Li, G. Wu, D. Yin, H. Wang, and Y. Wang, "Dcspell: A detector-corrector framework for chinese spelling error

correction," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1870–1874.

[6] X. Cheng, W. Xu, K. Chen, S. Jiang, F. Wang, T. Wang, W. Chu, and Y. Qi, "Spellgcn: Incorporating phonological and visual similarities into language models for chinese spelling check," *arXiv preprint arXiv:2004.14166*, 2020.

[7] R. Sun, X. Wu, and Y. Wu, "An error-guided correction model for chinese spelling error correction," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 3800–3810.

[8] 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.

[9] L. Liu, H. Wu, and H. Zhao, "Driving chinese spelling correction from a fine-grained perspective," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 10 727–10 737.

[10] H. Wu, S. Zhang, Y. Zhang, and H. Zhao, "Rethinking masked language modeling for Chinese spelling correction," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 743–10 756.

[11] L. Jiang, H. Wu, H. Zhao, and M. Zhang, "Chinese spelling corrector is just a language learner," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 6933–6943.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[13] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[14] S. Zhang, H. Huang, J. Liu, and H. Li, "Spelling error correction with soft-masked bert," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 882–890.

[15] S. Liu, T. Yang, T. Yue, F. Zhang, and D. Wang, "Plome: Pre-training with misspelled knowledge for chinese spelling correction," 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)*, 2021, pp. 2991–3000.

[16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[17] S. Liu, S. Song, T. Yue, T. Yang, H. Cai, T. Yu, and S. Sun, "Craspell: A contextual typo robust approach to improve chinese spelling correction," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 3008–3018.

[18] H. Chen and X. Wang, "Pygc: A pinyin language model guided correction model for chinese spell checking," in *International Conference on Neural Information Processing*. Springer, 2023, pp. 224–239.

[19] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[20] C. Li, J. Zhou, and T. Li, "Csclora: An efficient fine-tuning method for open-source large language models in chinese spelling correction."

[21] L. Liu, H. Wu, and H. Zhao, "Evolving chinese spelling correction with corrector-verifier collaboration," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 29 010–29 016.

[22] ——, "Chinese spelling correction as rephrasing language model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 18 662–18 670.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[24] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in neural information processing systems*, vol. 36, pp. 53 728–53 741, 2023.

[25] 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*, 2015, pp. 32–37.

[26] Y. Hu, F. Meng, and J. Zhou, "Cscd-ns: a chinese spelling check dataset for native speakers," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 146–159.

[27] Y. Zhao, J. Huang, J. Hu, X. Wang, Y. Mao, D. Zhang, Z. Jiang, Z. Wu, B. Ai, A. Wang *et al.*, "Swift: a scalable lightweight infrastructure for fine-tuning (2024)," *URL https://arxiv. org/abs/2408.05517*.

[28] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.