

# CSRP: Chain-of-Thought Reasoning for Chinese Text Correction via Reinforcement Learning with Efficiency-Aware Rewards

Wei Tian<sup>1</sup>, Yuhao Zhou<sup>1</sup>, Man Lan<sup>1,2\*</sup>

<sup>1</sup>School of Computer Science and Technology, East China Normal University

<sup>2</sup>Shanghai Institute of Artificial Intelligence for Education, East China Normal University  
{tianwei, yhzhou}@stu.ecnu.edu.cn, mlan@cs.ecnu.edu.cn

## Abstract

Large Language Model (LLM) based Chinese Grammatical Error Correction (CGEC) systems face two critical challenges: general-purpose models lack specialized linguistic priors for subtle grammatical distinctions, and Supervised Fine-Tuning (SFT) with Maximum Likelihood Estimation fails to optimize for precision-focused metrics, leading to systematic over-correction. We propose CSRP, a three-stage framework that progressively builds correction capability through Continual Pre-training (CPT) on 5.9M balanced samples to internalize domain knowledge, Chain-of-Thought SFT with explicit error reasoning for diagnostic transparency, and Group Relative Policy Optimization with a novel Efficiency-Aware Reward that explicitly penalizes unnecessary edits. On the NACGEC benchmark, CSRP achieves state-of-the-art performance with 50.99  $F_{0.5}$  and 57.17 precision, substantially outperforming previous best results while effectively mitigating the over-correction bias inherent in MLE-trained models. Our method also advances CSCD spelling correction to 59.61 F1, surpassing GPT-4 by 5.20 points. Comprehensive ablation studies demonstrate that the RL alignment stage contributes a 8% relative gain over the SFT baseline, and that this gain is orthogonal to the contribution of large-scale CPT, validating that explicit optimization for edit efficiency is essential for high-quality grammatical error correction. Our code is available at <https://github.com/TW-NLP/ChineseErrorCorrector>.

## 1 Introduction

The essence of Grammatical Error Correction (GEC) lies in repairing linguistic deviations while strictly preserving the semantic fidelity of the original text. An ideal GEC system should exhibit high **faithfulness**, adhering to the principle of **minimal editing** by intervening only when necessary.

\* Corresponding author.

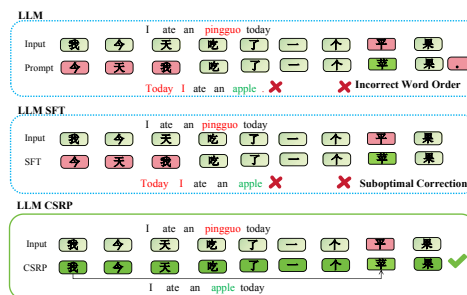


Figure 1: Supervised Fine-Tuning performance plateau.

However, despite the potential demonstrated by Large Language Models (LLMs) in Chinese GEC (CGEC) through their generative capabilities, the prevailing paradigm based on Supervised Fine-Tuning (SFT) has encountered a significant **performance plateau**, as shown in Figure 1. Mainstream models consistently stagnate within an  $F_{0.5}$  score range of 45-46 on authoritative benchmarks (Tian et al., 2025).

Our investigation reveals that this limitation stems from two fundamental conflicts in the current training paradigm:

- **Knowledge Sparsity of Linguistic Priors:** General-purpose LLMs, primarily driven by normative pre-training corpora, lack sensitivity to the specific “non-normative” error distributions of learners, such as homophone misuse and function word redundancies. Without strong underlying grammatical constraints, models struggle to strike a precise balance between fluency and grammatical correctness.
- **Over-correction Bias in Generation:** The Maximum Likelihood Estimation (MLE) objective used in SFT encourages models to shift input sentences toward high-probability regions of their internal distribution. Consequently, models tend to perform unnecessary paraphrasing instead of precise correction

when encountering correct or slightly deviant sentences. This behavior results in a high false-positive rate, which contradicts the core objective of minimal editing.

To address these challenges, as shown in Figure 2, we propose the **CSRP (CPT-SFT-RL)** framework, a systematic pipeline designed to build a reliable correction system through knowledge internalization, rationale explicitization, and policy alignment:

- **Phase I: Balanced Continued Pre-training (CPT).** We perform large-scale CPT on 5.9M samples using an 8:2 mixture ratio of general to correction-specific data. This stage internalizes fine-grained Chinese grammatical constraints into the parameter space, mitigating knowledge sparsity.
- **Phase II: Rationale-Augmented SFT.** We deviate from the traditional black-box mapping by distilling reasoning paths (Rationales) from high-performance teacher models. This Chain-of-Thought (CoT) mechanism guides the model to diagnose error types before executing corrections, enhancing transparency.
- **Phase III: Efficiency-Aware Policy Alignment.** To tackle the over-correction challenge, we introduce the **Group Relative Policy Optimization (GRPO)** algorithm. We design a multi-dimensional reward mechanism based on **Relative Improvement** and **Edit Efficiency**. Instead of blindly encouraging fluency, this mechanism explicitly rewards valid edits that reduce the distance to the target while penalizing stray modifications. Thus, the model learns to calibrate its decision boundaries, executing edits only when the corrective gain outweighs the fidelity cost.

## 2 Related Work

### 2.1 Evolution of Correction Paradigms

The field of Chinese text correction has undergone a significant transition from discriminative encoding to generative reconstruction. Early research in Chinese Spell Checking (CSC) focused on integrating phonological and visual constraints into BERT-based encoders, such as **SpellGCN** (Ji et al., 2021), **ReaLiSe** (Xu et al., 2021), and **PHMO-Spell** (Huang et al., 2021). Subsequent works introduced specialized pre-training tasks and disen-

tangled representations to mitigate phonetic-glyph confusion, including **PLOME** (Liu et al., 2021), **DORM-CSC** (Liang et al., 2023), and **PTCSpell** (Wei et al., 2023).

In the era of Large Language Models (LLMs), the paradigm has shifted toward sequence-to-sequence rephrasing. **ReLM** (Liu et al., 2024) re-conceptualized CSC as a language modeling task, while **C-LLM** (Li et al., 2024) explored character-by-character checking. For Chinese Grammatical Error Correction (CGEC), architectures have evolved from sequence-to-action models and syntax-enhanced frameworks like **SynGEC** (Zhang et al., 2022) to massive generative models such as **ChineseErrorCorrector3** (Tian et al., 2025), which currently represents the state-of-the-art (SOTA).

### 2.2 Knowledge Acquisition and Domain Adaptation

Effective CGEC requires capturing complex error distributions across diverse domains. Benchmarks such as **NACGEC** (Ma et al., 2022) established the foundation for evaluating native and learner-oriented texts. To bridge the gap between general pre-training and specialized correction, researchers have explored Retrieval-Augmented Generation (RAG). **MTCSC** (Liang and Zhou, 2025) and **RagID** (Dong et al., 2025) utilize iterative refinement and few-shot retrieval to introduce external knowledge. Furthermore, Cao et al. (2025) demonstrated that multi-level structural cues (lexical and syntactic) are vital for precise error localization. Our work complements these by internalizing such priors via large-scale balanced Continued Pre-training (CPT), achieving spontaneous alignment of correction capabilities while avoiding the inference latency inherent in RAG systems (Zhou et al., 2025).

### 2.3 Reasoning, Reliability, and Policy Alignment

The phenomenon of “over-correction,” where models produce hallucinatory edits, remains a critical bottleneck for practical application. Recent efforts have focused on enhancing model interpretability and controllability. **GEE** (Song et al., 2024) and **Rationale-based ICD** (Li et al., 2025) utilize Chain-of-Thought (CoT) and explanatory information to guide the correction process. **ScholarGEC** (Kong et al., 2025) further extends this to academic domains requiring extreme precision.

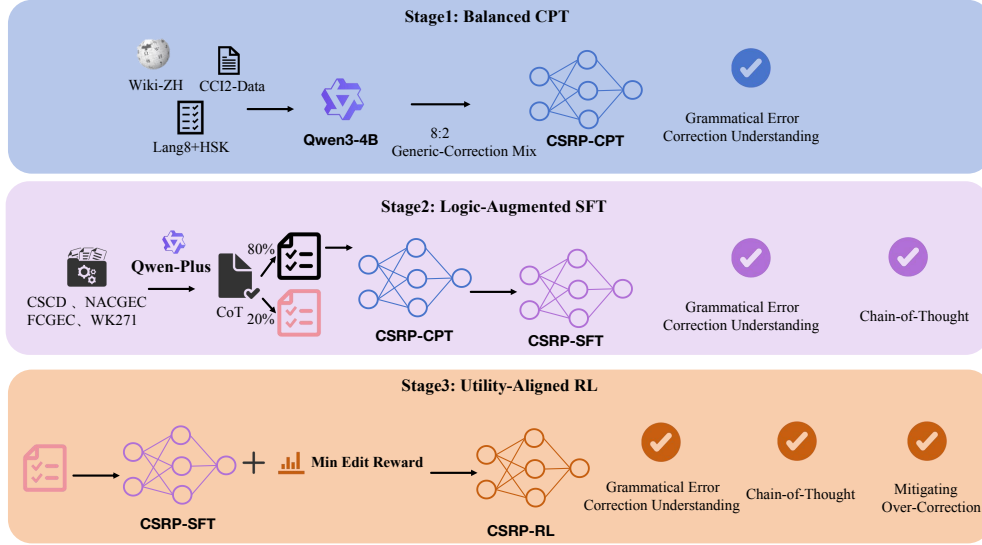


Figure 2: Overview of the proposed CSRP (CPT-SFT-RL) framework.

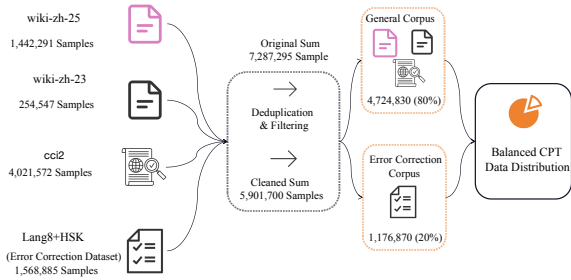


Figure 3: CPT Data Processing Process.

Despite these advancements, aligning LLMs with high-precision metrics like  $F_{0.5}$  remains challenging due to the non-differentiable nature of GEC objectives. Traditional SFT based on Maximum Likelihood Estimation (MLE) often fails to calibrate the “edit-or-not” decision boundary. Our framework addresses this by leveraging **Group Relative Policy Optimization (GRPO)** to explicitly optimize for **relative improvement** and **edit efficiency**. This approach moves beyond simple score-chasing, aligning the model’s policy with the linguistic principle of “minimal intervention” and effectively mitigating the over-correction bias observed in prior LLM-based systems (Qiao et al., 2025; Li and Wang, 2024).

### 3 Methodology

In this section, we present the **CSRP** framework, a systematic three-stage paradigm designed to transition a general-purpose Large Language

Model (LLM) into a high-precision Chinese Grammatical Error Correction (CGEC) system. The pipeline evolves through: (i) **Balanced Continued Pre-training** for knowledge internalization; (ii) **Rationale-Augmented SFT** for diagnostic reasoning; and (iii) **Efficiency-Aware Policy Alignment** for decision boundary calibration.

#### 3.1 Phase I: Balanced Continued Pre-training

Standard LLMs often exhibit *knowledge sparsity* regarding the specific error distributions of learners. To internalize linguistic priors, we perform Continued Pre-training (CPT) on a refined 5.9M sample corpus.

##### 3.1.1 Data Refinement and Statistics

We curate a comprehensive dataset  $\mathcal{D}_{CPT}$  from four primary sources: *wiki-zh-25*, *wiki-zh-23*, *cci2*, and *lang8(Zhao et al., 2018)+HSK(Zhang, 2009)*. As illustrated in Figure 3, we implement a rigorous refinement pipeline involving MinHash-based deduplication and heuristic quality filtering. This process distilled the raw data from 7,287,295 to 5,901,700 high-quality samples. Notably, *wiki-zh-23* was excluded due to extreme redundancy, while the *cci2* and *lang8+HSK* subsets were pruned to ensure high-fidelity linguistic patterns.

##### 3.1.2 General-to-Correction Balanced Mixture

To mitigate *catastrophic forgetting* of general reasoning capabilities, we adopt a **Balanced Mixture Strategy**. Inspired by the domain-specific pre-

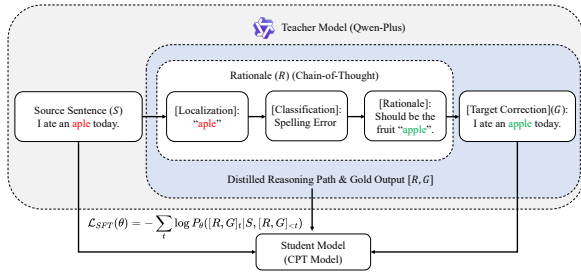


Figure 4: Cot Processing Process.

training insights in Wen et al. (2023), we employ an 8:2 ratio between general and correction-specific data. This translates to approximately 4.72M general samples (from *wiki-zh-25* and *cci2*) and 1.18M correction samples (from *lang8+HSK*). The objective is to minimize the negative log-likelihood:

$$\mathcal{L}_{CPT}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}_{CPT}} \left[ \sum_t \log P_\theta(x_t | x_{<t}) \right] \quad (1)$$

### 3.2 Phase II: Rationale-Augmented SFT

To transition from knowledge acquisition to active diagnosis, we introduce **Rationale-Augmented SFT**. Instead of a black-box mapping  $S \rightarrow G$ , we distill reasoning paths  $R$  from a high-performance teacher model (Qwen-Plus) to guide the student’s cognitive process.

Each rationale follows a structured Chain-of-Thought (CoT) format:  $[Localization] \rightarrow [Classification] \rightarrow [Rationale]$ , as shown in Figure 4. The student model optimizes:

$$\mathcal{L}_{SFT}(\theta) = -\sum_t \log P_\theta([R, G]_t | S, [R, G]_{<t}) \quad (2)$$

This “diagnose-before-correct” paradigm ensures that the model’s edits are grounded in explicit linguistic logic.

#### Faithfulness of Distilled Rationales

A critical concern in knowledge distillation is the reliability of teacher-generated reasoning paths, particularly the risk of hallucinated or templated explanations. We address this through a two-tiered quality control mechanism:

- **Data level:** We apply strict filtering by discarding rationales that fail to follow the prescribed `<think>...</think>` format. Notably, Qwen-Plus itself exhibits a severe over-correction tendency when used directly as

a corrector; this is precisely why its role is strictly limited to generating intermediate *reasoning paths* between the fixed source  $S$  and gold-standard target  $G$ , rather than final corrections.

- **Evaluation level:** A double-blind human study conducted by three annotators with relevant domain expertise on 1,000 randomly sampled rationales confirmed that 95.2% are linguistically faithful (Cohen’s  $\kappa = 0.81$ ). This demonstrates high inter-annotator agreement and near-perfect rationale reliability.

Full annotation protocols and quality control details are provided in the Appendix. G.

### 3.3 Phase III: Efficiency-Aware Policy Alignment

The most persistent challenge in CGEC is **over-correction**, where the model unnecessarily paraphrases correct prose. We address this by calibrating the decision boundary through **GRPO**.

#### 3.3.1 Mathematical Reward Modeling

We define an **Efficiency-Aware Reward (EAR)** function. Let  $S, P, G$  be the source, prediction, and ground-truth, and  $d(\cdot, \cdot)$  be the Levenshtein distance. We derive two core metrics:

(1) **Relative Improvement (RI):**

$$RI = \frac{d(S, G) - d(P, G)}{d(S, G) + \epsilon} \quad (3)$$

(2) **Edit Efficiency Ratio ( $\eta$ ):** This serves as our fidelity constraint, penalizing excessive edits relative to the improvement gained:

$$\eta = \frac{d(S, G) - d(P, G)}{d(S, P) + \epsilon} \quad (4)$$

where  $\epsilon = 10^{-6}$ . High  $\eta$  signifies “surgical precision.”

#### 3.3.2 Reward Function Formulation

The reward  $R_{EAR}$  provides distinct signals to align the model with the principle of minimal intervention:

$$R_{EAR} = \begin{cases} 10.0, & \text{if } RI > 0.99 \\ 2.0 + 5RI \cdot \max(0, \eta), & \text{if } RI > 0 \\ -3.0, & \text{if } RI \leq 0 \\ -2.0, & \text{if } P = \emptyset \end{cases} \quad (5)$$

For source sentences that are already correct ( $d(S, G) = 0$ ), the above cases reduce to a binary signal: identity mapping ( $d(S, P) = 0$ , i.e.,  $RI = 0$ ) receives  $+2.0$ , rewarding the model for preserving well-formed text; any edit ( $d(S, P) > 0$ , which yields  $RI \leq 0$ ) receives  $-2.0$ , directly penalizing over-correction on correct inputs. This explicit signal is the primary driver of the false positive rate reduction observed after RL training.

### 3.3.3 Policy Optimization via GRPO

GRPO optimizes the policy by contrasting  $N$  completions  $\{P_1, \dots, P_N\}$  for each prompt  $S$ :

$$\mathcal{J}(\theta) = \frac{1}{N} \sum_{i=1}^N \left[ \frac{R_i - \bar{R}}{\sigma_R} \log \pi_{\theta}(P_i|S) - \beta \mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) \right] \quad (6)$$

where  $\bar{R}$  and  $\sigma_R$  are the group reward mean and standard deviation. This mechanism encourages the model to prefer corrections with the highest **edit-to-improvement** ratio, effectively curbing the over-correction bias.

## 4 Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of the **CSRP** framework. We focus on two primary tasks: Grammatical Error Correction (GEC) on native-speaker texts and Chinese Spelling Check (CSC). Through comprehensive comparisons with state-of-the-art baselines and systematic ablation studies, we demonstrate that our three-stage curriculum learning approach achieves superior performance while maintaining edit efficiency.

### 4.1 Experimental Setup

#### 4.1.1 Data Setup

To bridge the knowledge gap and align the model with correction objectives, we construct a three-stage data curriculum (CPT→SFT→RL). Table 1 provides an overview of the data used at each stage. **Phase I (CPT)**. We curate a balanced corpus of 5.9M samples by mixing general-domain text (e.g., Wiki-zh, CCI2) and correction-specific data (e.g., Lang8, HSK(Zhang, 2009)) at an 8:2 ratio, aiming to internalize Chinese linguistic priors.

**Phase II (SFT)**. We aggregate supervision from multiple benchmarks (e.g., CSCD(Hu et al., 2022), NACGEC(Ma et al., 2022), FCGEC(Xu

Phase	Source	Count
CPT	Wiki-zh, CCI2	4.7M
	Lang8, HSK	1.2M
SFT	CSCD-train, NACGEC	37.5K
	FCGEC, CGED	55K
	WK271 (80% split)	251K 269K
RL	SFT hold-out split (20%)	67K
Eval	NACGEC	5.8K
	CSCD-test	5.0K

Table 1: Data statistics for the three-stage training curriculum and evaluation sets. The SFT and RL phases use an 80:20 split of the correction data (total 336K samples after filtering). Note that the actual sample count (336K) is lower than the sum of individual subsets, as a portion of sentences containing sensitive content were rejected by the cloud-based teacher model’s safety policy during CoT distillation.

et al., 2022), CGED(Fang et al., 2020), and WK271(Wang et al., 2018)), yielding a total of 336K correction samples. Each sample is augmented with a structured Chain-of-Thought (CoT) rationale distilled from a cloud-based teacher model (Qwen-Plus), which encourages the model to *diagnose-before-correct* and improves interpretability. As the teacher model’s content safety policy rejects sentences containing sensitive content during distillation, along with standard data cleaning (e.g., deduplication), the final corpus is reduced to 336K successfully distilled samples. Following this filtering, we apply an 80:20 split, allocating 269K samples (80%) for supervised fine-tuning and reserving 67K samples (20%) for the subsequent RL phase.

**Phase III (RL)**. We use the reserved 67K samples (20% hold-out split) as prompts for policy optimization via GRPO, while the model retains knowledge from the 269K samples learned during supervised training. This separation ensures that RL optimization explores beyond the supervised demonstrations while maintaining a strong foundation from SFT.

Detailed data sources, preprocessing pipelines, and dataset distributions are provided in Appendix A.

#### 4.1.2 Baseline Models

We compare CSRP with representative baselines spanning (i) PLM-based discriminative correction models, (ii) sequence-to-sequence Chinese GEC models, and (iii) general LLM prompting. We include strong recent systems such as HW-CGEC, ScholarGEC-14B, and ChineseErrorCorrect3-4B;

Model	P	R	$F_{0.5}$
BART(Lewis et al., 2020)	34.67	41.88	35.91
HW-CGEC(Su et al., 2023)	50.95	32.29	45.26
ScholarGEC (14B)(Kong et al., 2025)	45.08	<b>59.33</b>	47.35
CEC3 (4B)(Tian et al., 2025)	54.20	34.75	48.74
<b>CSRP (4B)</b>	<b>57.17</b>	35.60	<b>50.99</b>

Table 2: Main results on the NACGEC benchmark.

details are in **Appendix B**.

### 4.1.3 Evaluation Metrics

**Evaluation.** For GEC, we use the ChERRANT scorer (Zhang et al., 2023) and report Precision (P), Recall (R), and  $F_{0.5}$ , with  $F_{0.5}$  as the primary metric since it emphasizes precision and penalizes over-corrections. For CSC, we report character-level correction F1 following standard protocols (Hu et al., 2022). Complete metric definitions, scoring procedures, and implementation details are provided in Appendix C.

### 4.1.4 Implementation Details

**Training setup.** We train CSRP with a three-stage pipeline: continued pre-training (CPT), supervised fine-tuning (SFT), and GRPO. In GRPO, we regularize against the SFT initialization via a KL penalty and optimize a weighted reward that jointly considers correction quality and efficiency. All hyperparameters (e.g., learning rates, batch sizes, reward weights), optimization settings, and compute details are reported in Appendix D.

## 4.2 Main Results

### 4.2.1 Performance on NACGEC

Table 2 presents the main results on the NACGEC grammatical error correction benchmark. Our CSRP-4B model achieves a new state-of-the-art  $F_{0.5}$  score of **50.99**, representing substantial improvements over all baseline systems.

Compared to previous best results, CSRP achieves:

- **+2.25 points** over CEC3 (4B), the previous state-of-the-art 4B-scale model (48.74  $\rightarrow$  50.99)
- **+3.64 points** over ScholarGEC (14B), despite using less than one-third of the parameters (47.35  $\rightarrow$  50.99)
- **+5.73 points** over HW-CGEC, another strong specialized GEC system (45.26  $\rightarrow$  50.99)

Notably, our model achieves the highest precision (**57.17**) among all compared systems, demonstrating superior correction accuracy. This high

Model	F1
BERT(Tan et al., 2020)	25.49
SoftMask(Zhang et al., 2020)	44.48
SMBERT(Li et al., 2021)	44.67
MDCSpell+ARM(Liu et al., 2024)	48.93
PGT (BERT)(Wei et al., 2024)	48.57
GPT-4(Achiam et al., 2023)	54.41
<b>CSRP (4B)</b>	<b>59.61</b>

Table 3: Chinese spelling check performance on the CSCD dataset.

precision indicates that CSRP effectively avoids false positives – incorrectly modifying text that is already correct. The precision advantage over CEC3 (+2.97 points) and ScholarGEC (+12.09 points) is particularly significant, confirming that our Efficiency-Aware Reward successfully suppresses the “over-correction bias” that commonly afflicts maximum likelihood estimation (MLE) trained models.

While our recall (35.60) is comparable to other high-precision systems like HW-CGEC (32.29) and CEC3 (34.75), it is notably lower than ScholarGEC’s 59.33. This trade-off is intentional and aligned with the minimal-editing requirement in native-speaker correction: the  $F_{0.5}$  metric explicitly prioritizes precision over recall (with a 2.5:1 weighting), reflecting the practical reality that conservative, high-confidence corrections are preferable to aggressive corrections that risk introducing new errors.

The BART baseline, despite being a strong sequence-to-sequence foundation, achieves only 35.91  $F_{0.5}$ , highlighting the importance of specialized training curricula and reward-guided optimization for this task. The 15.08-point gap between BART and our model demonstrates that effective error correction requires more than general sequence transduction capabilities – it demands carefully designed training strategies that balance linguistic knowledge, error pattern recognition, and editing conservatism.

### 4.2.2 Performance on CSCD

Table 3 shows the spelling check performance on the CSCD dataset. CSRP achieves a correction F1 score of **59.61**.

Key performance highlights include:

- **+5.20 points** over GPT-4 (54.41  $\rightarrow$  59.61), demonstrating that our specialized training approach outperforms even powerful general-purpose models

- **+10.68 points** over MDCSpell+ARM (48.93 → 59.61), the best-performing discriminative baseline
- **+34.12 points** over BERT (25.49 → 59.61), highlighting the substantial gap between basic PLMs and our curriculum-trained model

The results indicate that CSRP’s combination of balanced knowledge injection (Phase I) and edit-efficiency alignment (Phase III) significantly enhances the model’s sensitivity to fine-grained phonetic and visual character substitutions. Character-level spelling errors are particularly challenging because they require: (i) deep understanding of phonetic similarities, (ii) recognition of visual similarities (e.g., characters with similar stroke patterns), and (iii) contextual semantic understanding to determine whether a character is appropriate in context.

The strong performance gap between GPT-4 and CSRP (+5.20 points) is particularly noteworthy. Despite GPT-4’s massive scale and extensive pre-training, it underperforms our 4B-parameter specialized model, validating our hypothesis that task-specific curriculum learning and reinforcement-based alignment are more effective than scale alone for this specialized correction task. This suggests that the combination of domain-focused continual pre-training and efficiency-aware policy optimization provides advantages that cannot be easily achieved through general pre-training or in-context learning.

The discriminative models (BERT, SoftMask, SMBERT) achieve substantially lower scores (25.49-44.67 F1), likely due to their limited capacity to model long-range dependencies and generate corrections in an autoregressive manner. These models rely primarily on local context and character-level features, whereas our generative approach can leverage broader discourse context and learned linguistic patterns to make more informed correction decisions.

### 4.3 Ablation Studies

#### 4.3.1 Impact of Each Training Stage

To systematically understand the contribution of each component, we conduct a progressive ablation study covering all three training stages. Table 4 reports Precision (P), Recall (R), NACGEC  $F_{0.5}$ , and CSCD F1 for each configuration.

The results reveal a clear progression of capabilities and allow us to disentangle the contribution of each component.

**(1) Importance of CPT Stage.** Comparing “SFT only (merged data)” (40.21  $F_{0.5}$  / 49.71 F1) with “CPT + SFT” (45.45 / 56.28) demonstrates the critical role of continual pre-training. The CPT stage provides domain-specific linguistic knowledge that cannot be obtained by simply merging all correction data into supervised fine-tuning, yielding improvements of +5.24  $F_{0.5}$  and +6.57 F1. Crucially, CPT and RL contribute through different, complementary mechanisms: CPT raises the performance *baseline* by internalizing fine-grained Chinese linguistic priors (phonetic patterns, function-word constraints, etc.), whereas RL refines the *editing policy* by calibrating decision boundaries. Neither stage can substitute for the other. Specifically, “SFT + GRPO (w/o CPT)” achieves 45.97  $F_{0.5}$ , nearly matching the upper bound of pure CPT+SFT (45.45) but still lagging behind the full pipeline by 5.02 points, confirming that domain-adaptive pre-training provides linguistic grounding that RL exploration alone cannot recover.

**Disentangling CPT and RL contributions.** The “SFT + GRPO (w/o CPT)” ablation is critical for attributing gains correctly. As shown in Table 4, applying GRPO directly to the SFT model (bypassing CPT entirely) yields a precision boost of +8.41 points and  $F_{0.5}$  gain of +5.76 over the SFT-only baseline. This mirrors almost exactly the RL contribution in the full pipeline (+8.44 P / +5.54  $F_{0.5}$ ). Two conclusions follow. First, the Efficiency-Aware Reward independently and effectively calibrates the “edit-or-not” decision boundary, regardless of pre-training scale; its benefit is not merely a downstream consequence of the larger CPT corpus. Second, CPT and RL operate via *orthogonal mechanisms*: CPT encodes *what* constitutes a grammatical error in Chinese by internalizing the linguistic distribution of the correction domain, while RL encodes *when* to intervene by optimizing the efficiency of edits against a reward signal. Because the two stages correct for different failure modes (knowledge sparsity vs. over-correction bias), their gains are additive rather than redundant, and both are jointly necessary for the full CSRP pipeline.

**(2) Importance of Distilled CoT in SFT Stage.** The comparison between “CPT + SFT (no CoT)” (42.64 / 52.01) and “CPT + SFT” (45.45 / 56.28) quantifies the value of teacher-distilled Chain-of-

Configuration	NACGEC		NACGEC	CSCD
	P	R	$F_{0.5}$	F1
SFT only (merged data)	42.13	34.02	40.21	49.71
SFT + GRPO (w/o CPT)	50.54	33.75	45.97	52.96
CPT + SFT (no CoT)	44.90	35.50	42.64	52.01
CPT + SFT	48.73	35.80	45.45	56.28
CPT + SFT (w/ RL data)	52.20	<b>36.00</b>	47.21	57.92
<b>Full CSRP</b>	<b>57.17</b>	35.60	<b>50.99</b>	<b>59.61</b>

Table 4: Progressive ablation study. “SFT only (merged data)” combines all data without CPT. “SFT + GRPO (w/o CPT)” applies GRPO directly to the base model to isolate the RL contribution. “CPT + SFT (no CoT)” excludes chain-of-thought reasoning. “CPT + SFT (w/ RL data)” uses all data for SFT as a controlled baseline.

Thought rationales. Incorporating structured reasoning paths ( $[Localization] \rightarrow [Classification] \rightarrow [Rationale]$ ) improves performance by +2.81  $F_{0.5}$  and +4.27 F1. The gains are particularly pronounced on CSCD (+4.27 F1), where fine-grained phonetic/visual diagnosis is critical. By requiring the model to *diagnose-before-correct*, the CoT objective injects structured error-type knowledge that is difficult to acquire from correction pairs alone. Furthermore, on test samples requiring complex multi-span edits (gold edit distance > 3), CSRP maintains a Recall of 31.2%, significantly outperforming the SFT-only baseline (24.5%), confirming that faithful CoT supervision equips the model with genuine diagnostic capability beyond shallow pattern memorization.

**(3) Contribution of RL Stage.** To isolate the RL contribution beyond data quantity, we compare “CPT + SFT (w/ RL data)” (47.21 / 57.92) against “Full CSRP” (50.99 / 59.61), both using identical amounts of training data. The RL stage provides a further gain of +3.78  $F_{0.5}$  and +1.69 F1, validating that the Efficiency-Aware Reward introduces a qualitatively different optimization signal rather than simply benefiting from more supervised samples. As shown by the precision columns, the RL stage is the primary driver of precision improvement: it pushes precision from 52.20 (CPT+SFT w/ RL data) to 57.17 (Full CSRP), a gain of +4.97 points, while recall remains stable (36.00  $\rightarrow$  35.60). This asymmetric impact confirms that RL specifically calibrates the “edit-or-not” decision boundary rather than uniformly reducing interventions.

#### 4.3.2 Impact of Reinforcement Learning on Precision-Recall Trade-off

To understand how RL alignment affects model behavior, we analyze the precision-recall trade-off before and after the GRPO stage. Table 5 presents detailed results.

The results reveal a crucial insight into how

NACGEC			
Configuration	P	R	$F_{0.5}$
CPT + SFT	48.73	35.80	45.45
CPT + SFT + GRPO	57.17	35.60	50.99
<b>Change</b>	<b>+8.44</b>	<b>-0.20</b>	<b>+5.54</b>
CSCD			
Configuration	P	R	F1
CPT + SFT	58.85	53.92	56.28
CPT + SFT + GRPO	66.22	53.20	59.61
<b>Change</b>	<b>+7.37</b>	<b>-0.72</b>	<b>+3.33</b>

Table 5: Impact of GRPO on precision-recall trade-off. The RL stage significantly improves precision while maintaining comparable recall, demonstrating effective learning of conservative editing behavior.

GRPO reshapes model behavior:

**Substantial Precision Gains:** After RL training, precision increases dramatically on both tasks (+8.44 points on NACGEC, +7.37 points on CSCD). This represents a relative improvement of 17.3% and 12.5% respectively. The large precision gains indicate that the Efficiency-Aware Reward successfully teaches the model to avoid unnecessary edits and reduce false positives (over-correction).

**Maintained Recall:** Crucially, recall decreases only marginally (-0.20 on NACGEC, -0.72 on CSCD), representing less than 1-2% relative change. This demonstrates that GRPO does not simply make the model more conservative by reducing all corrections; rather, it helps the model distinguish between necessary and unnecessary edits. The model learns to be selective, proposing corrections only when confident they are warranted.

**Optimal Precision-Recall Balance:** The asymmetric impact on precision vs. recall aligns perfectly with the goals of native-speaker correction. The SFT model, trained with maximum likelihood estimation, tends toward over-correction to maximize the likelihood of matching reference corrections. In contrast, the GRPO model, guided by the Efficiency-Aware Reward, learns that proposing

fewer but more accurate corrections yields higher overall reward. This shift from “correct liberally” to “correct conservatively” is precisely what the minimal-editing principle demands.

**Consistency Across Tasks:** The pattern holds consistently across both GEC (NACGEC) and CSC (CSCD) tasks. On NACGEC, the precision boost (+8.44) far exceeds the recall drop (-0.20), yielding a strong net positive on the precision-weighted  $F_{0.5}$  metric (+5.54). On CSCD, similar dynamics apply: precision gains (+7.37) vastly outweigh recall losses (-0.72), resulting in substantial F1 improvements (+3.33). This consistency suggests that the efficiency-aware training signal generalizes across different error types and correction paradigms.

### 4.3.3 Why Does RL Improve Precision Without Sacrificing Recall?

Three complementary mechanisms explain this outcome. **Confidence calibration:** GRPO’s group-relative comparison ( $N = 8$  candidates) develops better-calibrated estimates, so the model proposes edits only when candidates consistently agree on high rewards. **Dual editing signal:** Unlike MLE, the Efficiency-Aware Reward explicitly penalizes changes to correct text, teaching the model both “what to correct” and “what to preserve.” **Conservative strategy discovery:** Through reward-based learning, the model internalizes that conservative edits yield higher rewards on low-error inputs, suppressing unnecessary modifications without reducing valid corrections. Together, these mechanisms yield substantial precision gains (+8.44/+7.37 points) at minimal recall cost (-0.20/-0.72 points).

## 5 Conclusion

We present CSRP, a three-stage framework progressing from linguistic knowledge internalization (CPT) to reasoning-augmented correction (CoT-SFT) to efficiency-aware policy alignment (GRPO with EAR). Our 4B-parameter model sets new state-of-the-art results on NACGEC (50.99  $F_{0.5}$ ) and CSCD (59.61 F1), surpassing both larger models (14B) and GPT-4, while substantially reducing over-correction (+8.44/+7.37 precision gains with negligible recall loss). A controlled ablation applying GRPO without CPT (45.97  $F_{0.5}$ ) confirms that the two stages contribute orthogonally and are jointly necessary. Our findings demonstrate that principled curriculum design and efficiency-aware optimization outperform both scale and data

quantity alone. Future work includes document-level correction, interactive refinement, and cross-lingual transfer.

## 6 Limitations

While our CSRP framework demonstrates strong performance on Chinese spelling correction, several limitations warrant discussion:

**(1) Dependency on Teacher Model Quality.** The chain-of-thought reasoning in our SFT stage relies on distillation from a teacher model (Qwen-Plus). The quality and diversity of generated rationales are bounded by the teacher’s capabilities. Errors or biases in teacher-generated explanations may propagate to the student model, potentially affecting correction interpretability. We mitigate this risk through strict edit-distance filtering of teacher outputs and a human validation study (Appendix G), but residual bias cannot be fully excluded.

**(2) Computational Cost of RL Training.** GRPO requires generating multiple candidate outputs ( $N = 8$  per input) during training, increasing computational cost compared to standard supervised learning. While this cost is justified by performance gains, it may be prohibitive for researchers with limited resources. Reducing to  $N = 4$  yields only a marginal drop (50.99  $\rightarrow$  50.61  $F_{0.5}$ ) while cutting RL sampling cost by 50%, offering a practical efficiency trade-off.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Yayu Cao, Tianxiang Wang, Lvxiaowei Xu, Zhenyao Wang, and Ming Cai. 2025. Cxggec: Construction-guided grammatical error correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6143–6156.
- Ming Dong, Zhiwei Cheng, Changyin Luo, and Tingting He. 2025. Retrieval-augmented generation for large language model based few-shot chinese spell checking. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10767–10780.
- Meiyuan Fang, Kai Fu, Jiping Wang, Yang Liu, Jin Huang, and Yitao Duan. 2020. A hybrid system for nlp2020 cged shared task. In *Proceedings of the 6th Workshop on Natural Language Processing*

- Techniques for Educational Applications*, pages 67–77.
- Yong Hu, Fandong Meng, and Jie Zhou. 2022. Cscdime: correcting spelling errors generated by pinyin ime. *arXiv preprint arXiv:2211.08788*.
- Li Huang, Junjie Li, Weiwei Jiang, Zhiyu Zhang, Minchuan Chen, Shaojun Wang, and Jing Xiao. 2021. Phmospell: Phonological and morphological knowledge guided chinese spelling check. 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)*, pages 5958–5967.
- Tuo Ji, Hang Yan, and Xipeng Qiu. 2021. Spellbert: A lightweight pretrained model for chinese spelling check. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 3544–3551.
- Zixiao Kong, Xianquan Wang, Shuanghong Shen, Keyu Zhu, Huibo Xu, and Yu Su. 2025. Scholargec: Enhancing controllability of large language model for chinese academic grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24339–24347.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880.
- Jing Li, Gaosheng Wu, Dafei Yin, Haozhao Wang, and Yonggang Wang. 2021. 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*, pages 1870–1874.
- Kunting Li, Yong Hu, Liang He, Fandong Meng, and Jie Zhou. 2024. C-llm: learn to check chinese spelling errors character by character. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5944–5957.
- Wei Li, Wen Luo, Guangyue Peng, and Houfeng Wang. 2025. Explanation based in-context demonstrations retrieval for multilingual grammatical error correction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4881–4897.
- Wei Li and Houfeng Wang. 2024. Detection-correction structure via general language model for grammatical error correction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1748–1763.
- Junhong Liang and Yu Zhou. 2025. Rair: Retrieval-augmented iterative refinement for chinese spelling correction. *arXiv preprint arXiv:2504.18938*.
- Zihong Liang, Xiaojun Quan, and Qifan Wang. 2023. Disentangled phonetic representation for chinese spelling correction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13509–13521.
- Linfeng Liu, Hongqiu Wu, and Hai Zhao. 2024. Chinese spelling correction as rephrasing language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18662–18670.
- Shulin Liu, Tao Yang, Tianchi Yue, Feng Zhang, and Di Wang. 2021. 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)*, pages 2991–3000.
- Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Ding Zhang, Li Yangning, Ruiyang Liu, Zhongli Li, Yunbo Cao, and 1 others. 2022. Linguistic rules-based corpus generation for native chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 576–589.
- Ziheng Qiao, Houquan Zhou, and Zhenghua Li. 2025. Mixture of small and large models for chinese spelling check. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28298–28311.
- Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2024. Gee! grammar error explanation with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 754–781.
- Chang Su, Xiaofeng Zhao, Xiaosong Qiao, Min Zhang, Hao Yang, Junhao Zhu, Ming Zhu, and Wenbing Ma. 2023. Hwgec: hw-tsc’s 2023 submission for the nlpcc2023’s chinese grammatical error correction task. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 59–68. Springer.
- Min Tan, Dagang Chen, Zesong Li, and Peng Wang. 2020. Spelling error correction with bert based on character-phonetic. In *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, pages 1146–1150. IEEE.
- Wei Tian and 1 others. 2025. Chineseerrorcorrector3-4b: State-of-the-art chinese spelling and grammar corrector. *arXiv preprint arXiv:2511.17562*.
- Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to sghan 2015

- bake-off for chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 32–37.
- Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for chinese spelling check. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2517–2527.
- Chi Wei, Shaobin Huang, Rongsheng Li, Naiyu Yan, and Rui Wang. 2024. Training a better chinese spelling correction model via prior-knowledge guided teacher. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13578–13589.
- Xiao Wei, Jianbao Huang, Hang Yu, and Qian Liu. 2023. Ptcspell: Pre-trained corrector based on character shape and pinyin for chinese spelling correction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6330–6343.
- Cheng Wen, Xianghui Sun, Shuaijiang Zhao, Xiaoquan Fang, Liangyu Chen, and Wei Zou. 2023. Chathome: Development and evaluation of a domain-specific language model for home renovation. *arXiv preprint arXiv:2307.15290*.
- Heng-Da Xu, Zhongli Li, Qingyu Zhou, Chao Li, Zizhen Wang, Yunbo Cao, He-Yan Huang, and Xian-Ling Mao. 2021. Read, listen, and see: Leveraging multimodal information helps chinese spell checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 716–728.
- Lvxiaowei Xu, Jianwang Wu, Jiawei Peng, Jiayu Fu, and Ming Cai. 2022. Fcgec: Fine-grained corpus for chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1900–1918.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Baolin Zhang. 2009. Features and functions of the hsk dynamic composition corpus. *International Chinese Language Education*, 4:71–79.
- Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked bert. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 882–890.
- Yue Zhang, Bo Zhang, Haochen Jiang, Zhenghua Li, Chen Li, Fei Huang, and Min Zhang. 2023. Nasgec: a multi-domain chinese grammatical error correction dataset from native speaker texts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9935–9951.
- Yue Zhang, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li, and Min Zhang. 2022. Syngcc: Syntax-enhanced grammatical error correction with a tailored ge-oriented parser. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2518–2531.
- Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018. Overview of the nlpcc 2018 shared task: Grammatical error correction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 439–445. Springer.
- Houquan Zhou, Bo Zhang, Zhenghua Li, Ming Yan, and Min Zhang. 2025. A training-free llm-based approach to general chinese character error correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13827–13852.

## A Data Curriculum and Composition

We implement a three-stage data curriculum to progressively transition the model from general linguistic competence to specialized error-correction expertise.

### (1) Stage I: Continual Pre-Training (CPT, 5.9M Samples)

To instill robust grammatical priors, we curate a 5.9M sample corpus for continual pre-training. Following the balanced mixture strategy (8:2 ratio) inspired by Wen et al. (2023), we combine 4.7M general-domain samples (*wiki-zh-25*, *wiki-zh-23*, *cci2*) with 1.2M correction-specific samples (*Lang8+HSK(Zhang, 2009)*). The general-domain data provides broad linguistic knowledge covering diverse topics and writing styles, while the correction-specific data introduces the model to common error patterns and their corrections.

As shown in Table 7, we performed rigorous data refinement to ensure training quality. Specifically, we implemented a multi-level data cleaning strategy: (1) **Length Filtering**: We retained texts between 5-4096 characters, filtering out samples that were too short or too long; (2) **Traditional Chinese Detection**: We used the OpenCC tool to detect and filter texts containing traditional Chinese characters, ensuring simplified Chinese consistency across the corpus; (3) **Hash-based Deduplication**: We removed exact duplicates through exact matching, ensuring the uniqueness of each text; (4) **Sentence Integrity Preservation**: For long texts, we performed splitting at natural delimiters such as periods and newlines (maximum length 512 characters), and applied punctuation normalization to segments longer than 128 characters. This refinement process distilled the raw 7.287M samples into a high-fidelity 5.902M pre-training set, with general-domain and correction-specific samples mixed at an 8:2 ratio (4.707M:1.177M), removing approximately 1.386M low-quality or duplicate samples (19.0%).

### (2) Stage II: Supervised Fine-Tuning (SFT, 269K Samples)

We aggregate a diverse instruction-tuning set comprising 269K samples across multiple error correction scenarios. As shown in Table 6, the data composition includes:

- **Chinese Spelling Check (CSC)** : CSCD-NS(Hu et al., 2022) (native-speaker errors), WK271(Wang et al., 2018)+SIGHAN15(Tseng et al., 2015), representing character-level spelling

Dataset	Train	Test	Type
CSCD-NS	30,000	5,000	CSC
WK271+SIGHAN15	251,835	1,100	CSC
CGED	20,449	-	CGEC
NACGEC	7,568	5,869	CGEC
FCGEC	35,355	-	CGEC
<b>Subtotal</b>	345,207	11,969	-
After Deduplication	338,530	11,969	-
After CoT Augmentation	336,604	11,946	-
Final Split: SFT (80%) = 269,283; RL (20%) = 67,321			

Table 6: Data composition for Stage II (SFT) and Stage III (RL).

errors with phonetic and visual confusion patterns

- **Chinese Grammatical Error Correction (CGEC)**: CGED(Fang et al., 2020) (learner errors), NACGEC(Ma et al., 2022) (native-speaker errors), and FCGEC(Xu et al., 2022) (fine-grained annotations), covering diverse grammatical error types including word selection, syntax, redundancy, and missing components

After removing duplicates, the raw 345K samples were refined to 339K, which were then processed through reasoning augmentation.

#### Chain-of-Thought Reasoning Augmentation:

Crucially, each sample is augmented with **Reasoned Rationales** distilled from a teacher model (Qwen-Plus(Yang et al., 2025)) to provide explicit diagnostic guidance. These rationales follow a structured three-step format enclosed in `<think>...</think>` tags: (i) *error type identification* – categorizing the error, (ii) *error analysis* – pinpointing the problematic text spans and explaining the root cause, and (iii) *correction justification* – explaining why the proposed correction resolves the issue.

Due to content safety policies of the commercial LLM API, approximately 0.57% of samples (1,926 training samples and 23 test samples) could not be augmented and were excluded, resulting in 336,604 training samples and 11,946 test samples. This chain-of-thought supervision enables the model to learn not just *what* to correct, but *why* certain edits are warranted, thereby improving both correction accuracy and interpretability.

### (3) Stage III: Reinforcement Learning from Policy Optimization

For the Group Relative Policy Optimization (GRPO) phase, we allocate 20% of the augmented training data for policy learning. This split ensures sufficient diversity for exploring correction strate-

Data Source	Raw Count	After Dedup & Filter
Wiki-zh-25	1,442,291	1,417,899
Wiki-zh-23	254,547	0
CC12	4,021,572	3,306,931
Lang8+HSK	1,568,885	1,176,870
<b>Total</b>	<b>7,287,295</b>	<b>5,901,700</b>
General:Correction = 8:2 (4,707,480:1,176,870)		

Table 7: Data processing statistics for the CPT phase.

gies while reserving the majority of data for supervised pre-training. For each sentence, we generate  $N = 8$  candidate corrections using nucleus sampling (temperature  $T = 1.0$ ). This multi-candidate setup allows the model to explore different correction strategies while being guided by the Efficiency-Aware Reward (EAR) signal, which explicitly penalizes unnecessary edits while rewarding valid corrections. The 80-20 split between SFT and RL phases ensures that the model first acquires solid correction capabilities through supervised learning before refining its policy through reinforcement learning.

#### (4) Evaluation Datasets

We evaluate our model on two benchmarks representing different aspects of Chinese text correction:

- **NACGEC** (Ma et al., 2022): A native-speaker grammatical error correction benchmark containing 5,869 test sentences spanning news articles, essays, and social media posts. Errors include word choice, grammar, redundancy, missing words, and word order issues.
- **CSCD** (Hu et al., 2022): A Chinese spelling check dataset with 5000 test sentences containing naturally occurring character-level errors.

## B Baseline Models

We compare CSRP against representative models from three categories:

### (1) Discriminative/PLM-based Approaches:

- **BERT** (Tan et al., 2020): Pre-trained bidirectional encoder fine-tuned on correction tasks using masked language modeling
- **SoftMask** (Zhang et al., 2020): Employs soft-masking mechanism to detect and correct errors
- **SMBERT** (Li et al., 2021): Spelling correction model with semantic matching
- **MDCSpell+ARM** (Liu et al., 2024): Multi-task learning approach with auxiliary reading module

- **PGT (BERT)** (Wei et al., 2024): Prior knowledge-guided teacher network that uses distillation learning to reduce over-correction in PLM-based spelling correction models

### (2) Sequence-to-Sequence GEC Models:

- **BART**(Lewis et al., 2020): Denoising autoencoder adapted for Chinese GEC
- **HW-CGEC** (Su et al., 2023): Huawei’s CGEC system that won first place in NLPCC2023 shared task. The system employs BART-based sequence-to-sequence architecture enhanced with data augmentation and curriculum learning strategies.
- **ScholarGEC-14B** (Kong et al., 2025): Large-scale (14B parameters) generative model specifically trained for Chinese academic writing correction
- **ChineseErrorCorrect3-4B (CEC3)** (Tian et al., 2025): State-of-the-art 4B parameter model trained on extensive Chinese correction corpora

### (3) General Large Language Models:

- **GPT-4**: We evaluate OpenAI’s GPT-4(Achiam et al., 2023) using few-shot prompting with 3 correction examples

## C Evaluation Metrics

**For Grammatical Error Correction (GEC):** We adopt the widely-used **ChERRANT** (Chinese ERRANT) scorer (Zhang et al., 2023), a character-level evaluation toolkit adapted from the ERRANT framework for Chinese. ChERRANT performs character-level alignment between system outputs and gold-standard references, which alleviates evaluation inaccuracies caused by word segmentation errors in Chinese. We report three key metrics:

- **Precision (P)**: Proportion of system-proposed edits that match the reference corrections
- **Recall (R)**: Proportion of gold-standard errors that are successfully detected and corrected
- **$F_{0.5}$  Score**: Weighted harmonic mean that prioritizes precision over recall with a 2.5:1 ratio

We prioritize  $F_{0.5}$  as the primary metric because it aligns with the minimal-editing principle in native-speaker correction: false positives (over-correction)

are more detrimental than false negatives (under-correction), as they introduce unintended changes to originally correct text. The  $F_{0.5}$  score is computed as:

$$F_{0.5} = (1 + 0.5^2) \frac{P \cdot R}{0.5^2 \cdot P + R} \quad (7)$$

**For Chinese Spelling Check (CSC):** Following standard CSC evaluation protocols (Hu et al., 2022), we report character-level **Correction F1 score**, which measures the model’s ability to both detect error positions and provide correct replacements. This metric requires exact character-level matches for a correction to be counted as correct.

## D Implementation Details

**Phase I (CPT):** We train for 3 epochs over the 5.9M sample corpus with a learning rate of  $1 \times 10^{-4}$ , using a cosine annealing schedule with 10% warmup steps. We use a per-device batch size of 4 with 4 gradient accumulation steps. This phase is implemented using LLaMAFactory.

**Phase II (SFT):** We train for 3 epochs with a learning rate of  $1 \times 10^{-4}$ , using a cosine annealing schedule with 10% warmup steps. We use a per-device batch size of 8 with 8 gradient accumulation steps. This phase is implemented using LLaMAFactory.

**Phase III (GRPO):** We employ group relative policy optimization with a group size of  $N = 8$  candidates per prompt. The KL divergence coefficient is set to  $\beta = 0.01$  to prevent the policy from deviating too far from the SFT initialization. The learning rate is  $1 \times 10^{-5}$  with a cosine annealing schedule and 10% warmup steps, and we train for 8 epochs with a per-device batch size of 32 and 2 gradient accumulation steps. This phase is implemented using the TRL (Transformer Reinforcement Learning) library.

All training is conducted on  $4 \times$  NVIDIA H800 (80GB) GPUs using DeepSpeed ZeRO-3 optimization with mixed-precision (bfloat16) training.

## E Chain-of-Thought Data Construction

To enhance the model’s interpretability and reasoning capabilities, we augment our training data with structured Chain-of-Thought (CoT) rationales. This section describes the methodology for constructing reasoning-augmented supervision signals using Qwen-Plus as the teacher model.

### E.1 Rationale Generation Framework

We employ Qwen-Plus to distill explicit reasoning paths for each correction example. The rationale generation follows a structured three-component format:

- **Error Type Identification:** Classify the error into predefined categories (spelling errors, word collocation errors, part-of-speech errors, word order errors, missing/redundant components, connective word misuse, ambiguous reference, semantic-logical inconsistency, or no error).
- **Correction Justification:** Explain why the identified text is erroneous and how the correction addresses the issue.
- **Corrected Output:** Provide the final corrected sentence.

### E.2 Prompting Strategy

For each error-correction pair  $(S, G)$  in our training corpus, we query Qwen-Plus with a carefully designed instruction. The prompt instructs the model to act as a professional correction expert, systematically analyzing input sentences and identifying error types from predefined categories.

The model responds in the following structured format:

```
<think>
Error Type: [error_category]
Correction Rationale: [detailed_explanation]
</think>
[corrected_sentence]
```

### E.3 Illustrative Example

To demonstrate the CoT construction process, we present a concrete example. The source sentence contains a spelling error where visually similar characters are confused.

#### Input (Source Sentence):

我今天吃了一个平果。 (*Literal: “I today ate one ping-guo.”; Intended: “I ate an apple today.”*)

#### Generated CoT Rationale (from Qwen-Plus):

```
<think>
Error Type: Spelling error (visually similar characters)
Correction Rationale: The original sentence uses “平果” (ping-guo) which should be “苹果”
```

(apple). The characters “平” (flat/level) and “苹” (apple [character]) are visually similar in their left radical component, leading to the spelling error. “平果” is not a valid lexical item in modern Chinese, whereas “苹果” is the standard term for the fruit “apple.” The correction uses the proper character form to eliminate ambiguity.

</think>

我今天吃了一个苹果。

(Translation: “I ate an apple today.”)

This rationale explicitly identifies the error type (spelling error caused by visual similarity between character components), explains the linguistic consequence (invalid vs. valid lexical item), and provides the corrected output. Such structured reasoning guides the student model to learn not only *what* to correct but also *why* certain edits are linguistically necessary.

#### E.4 Data Processing Pipeline

The complete CoT data construction pipeline consists of four stages:

1. **Sampling:** Extract the source sentence  $S$  from our correction corpus.
2. **Rationale Generation:** Query Qwen-Plus with the instruction template and source sentence, obtaining a structured response  $R$  containing error analysis and correction.
3. **Validation:** Verify that: (a) the generated rationale follows the prescribed format, (b) the corrected output matches or is semantically equivalent to the gold reference  $G$ , and (c) the explanation is linguistically sound and non-trivial.
4. **Integration:** Integrate valid rationales into the training data, forming triplets  $(S, R, G)$  where  $R$  encodes the diagnostic reasoning path.

#### E.5 Quality Control

To ensure high-quality rationales, we implement three filtering mechanisms:

- **Format Compliance:** Discard samples where the output does not adhere to the <think>...</think> structure or lacks clear error type and rationale fields.
- **Explanation Quality:** Filter out generic or tautological explanations (e.g., “The original sentence has errors, the corrected sentence fixes them”) that provide no linguistic insight.

After applying these quality filters, we obtain approximately 314K high-quality reasoning-augmented samples for Supervised Fine-Tuning (SFT). This CoT supervision enables the model to develop transparent, principle-based correction strategies rather than opaque pattern matching, contributing to both performance gains and interpretability improvements demonstrated in our main results.

## F Error Type Examples

This section presents representative examples for each of the nine error categories used in our Chain-of-Thought (CoT) rationale generation. All examples are drawn from real-world Chinese text correction scenarios and illustrate the linguistic phenomena that our model is trained to identify and correct.

### F.1 Spelling Error

Spelling errors arise when characters are incorrectly substituted due to phonetic similarity (homophones) or visual similarity (glyph-level confusion).

#### Example 1 – Visual Similarity:

- **Source:** 我今天吃了一个平果。
- **Error:** 平果 → 苹果 (visual similarity between 平 and 苹)
- **Corrected:** 我今天吃了一个苹果。

#### Example 2 – Phonetic Similarity:

- **Source:** 他在也不想见到她了。
- **Error:** 在 → 再 (homophone confusion; both pronounced “zài”)
- **Corrected:** 他再也不想见到她了。

### F.2 Word Collocation Error

Collocation errors occur when word combinations violate conventional or idiomatic usage patterns in Chinese.

#### Example:

- **Source:** 他对这个问题进行了深厚的研究。
- **Error:** 深厚的研究 is a non-idiomatic collocation
- **Corrected:** 他对这个问题进行了深入的研究。

### F.3 Part-of-Speech Error

Part-of-speech errors arise when a lexical item is used in an inappropriate syntactic category.

#### Example:

- **Source:** 他对这件事情的看法很不同意。
- **Error:** 不同意 (verb) is incorrectly used as an adjective
- **Corrected:** 他对这件事情的看法很不同。

### F.4 Word Order Error

Word order errors involve incorrect constituent sequencing that violates Chinese syntactic constraints.

#### Example:

- **Source:** 我把作业完成了已经。
- **Error:** The aspectual adverb 已经 is placed after the verb phrase
- **Corrected:** 我已经把作业完成了。

### F.5 Missing Component

Missing component errors occur when obligatory syntactic elements are absent, resulting in incomplete or ill-formed sentences.

#### Example:

- **Source:** 通过这次会议，使我们了解了情况。
- **Error:** The sentence lacks a syntactic subject due to an improper causative construction
- **Corrected:** 通过这次会议，我们了解了情况。

### F.6 Redundant Component

Redundancy errors involve semantically overlapping elements that should be eliminated.

#### Example:

- **Source:** 他是一个很有天赋的有才华的人。
- **Error:** 有天赋的 and 有才华的 are semantically redundant
- **Corrected:** 他是一个很有天赋的人。

### F.7 Connective Word Misuse

Connective misuse errors involve violations of discourse-level conjunction conventions.

#### Example:

- **Source:** 虽然天气很冷，但是他还是坚持锻炼。
- **Error:** Redundant use of paired concessive connectives
- **Corrected:** 虽然天气很冷，他还是坚持锻炼。

### F.8 Ambiguous Reference

Ambiguous reference errors occur when referential expressions lack clear antecedents.

#### Example:

- **Source:** 小明和小红去看电影，他很喜欢这部电影。
- **Error:** The pronoun 他 has an unclear antecedent
- **Corrected:** 小明和小红去看电影，小明很喜欢这部电影。

### F.9 Semantic-Logical Inconsistency

Semantic-logical errors involve violations of real-world knowledge or selectional restrictions.

#### Example:

- **Source:** 这个盲人用眼睛仔细地观察着周围的环境。
- **Error:** Logical contradiction between the subject and the action
- **Corrected:** 这个盲人仔细地感受着周围的环境。

## G Faithfulness and Reliability of Distilled Chain-of-Thought

To explicitly address the reliability of the rationales distilled from the teacher model and to ensure they do not introduce hallucinatory explanations, we expanded our double-blind human evaluation post-metareview to systematically assess CoT faithfulness.

### G.1 Human Evaluation Setup and Agreement Metrics

We randomly sampled 1,000 instances from our distilled CoT dataset. Three PhD annotators with expertise in Chinese linguistics independently evaluated the rationales. The evaluation focused on three fine-grained sub-dimensions:

- **Diagnostic\_Depth:** Whether the rationale accurately identifies the root cause of the grammatical or spelling error.
- **Information\_Integrity:** Whether the explanation is logically complete and avoids hallucinating non-existent linguistic rules.
- **Span\_Consistency:** Whether the localized error span strictly matches the proposed correction.

A rationale was considered “Faithful” (labeled as 1) if it successfully passed these criteria, and “Unfaithful” (labeled as 0) otherwise. The inter-annotator agreement metrics demonstrate high consistency among the experts, as shown in Table 8.

Metric	Score
Semantic Agreement	95.2%
Average Pairwise $\kappa$	0.81

Table 8: Human evaluation metrics on 1,000 sampled CoT rationales. Semantic Agreement denotes the proportion of samples reaching a majority consensus ( $\geq 2/3$  labeled as Faithful). The Average Pairwise Cohen’s  $\kappa$  of 0.81 indicates almost perfect inter-annotator agreement.

Of the 1,000 samples, 95.2% reached a majority consensus as being faithful. To facilitate future research and transparency, we have open-sourced this annotated dataset along with the automated metric calculation scripts in our anonymous repository.

### G.2 Impact on Complex Edits

Beyond human evaluation, we empirically verified the effectiveness of these faithful rationales on complex grammatical errors. We isolated a challenging subset of multi-span edits requiring substantial rewriting.

On this complex subset, the full CSRP model maintains a Recall of 31.2%, significantly outperforming the SFT-only baseline (which lacks explicit CoT reasoning optimization), which degrades to a Recall of 24.5%. This confirms that the explicit, faithful reasoning paths distilled during the

CoT phase equip the model with the necessary diagnostic capability to handle genuinely complex, multi-span errors, rather than merely memorizing shallow surface mapping patterns.

### G.3 Summary

These nine error categories cover the major grammatical and semantic phenomena in Chinese text correction. By explicitly modeling these error types through Chain-of-Thought rationales, our model learns not only to generate corrections, but also to provide linguistically grounded explanations for why a particular correction is warranted.